UNIVERSIDADE DO ESTADO DA BAHIA
CAMPUS I - SALVADOR
DEPARTAMENTO DE CIÊNCIAS EXATAS E DA TERRA
CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO

JESSICA TEIXEIRA NOGUEIRA DE JESUS

CARACTERIZAÇÃO DE MARCADORES GENÉTICOS RELACIONADOS À EXPRESSÃO DE ANTÍGENOS DE GRUPOS SANGUÍNEOS ERITROCITÁRIOS UTILIZANDO PROCESSAMENTO DE DADOS BRUTOS DE SEQUENCIAMENTO DE NOVA GERAÇÃO

Salvador
2023

UNIVERSIDADE DO ESTADO DA BAHIA
CAMPUS I - SALVADOR
DEPARTAMENTO DE CIÊNCIAS EXATAS E DA TERRA
CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO

JESSICA TEIXEIRA NOGUEIRA DE JESUS

CARACTERIZAÇÃO DE MARCADORES GENÉTICOS RELACIONADOS À EXPRESSÃO DE ANTÍGENOS DE GRUPOS SANGUÍNEOS ERITROCITÁRIOS UTILIZANDO PROCESSAMENTO DE DADOS BRUTOS DE SEQUENCIAMENTO DE NOVA GERAÇÃO

Trabalho de Conclusão de Curso apresentado para obtenção do grau de Bacharel em Sistemas de Informação.

Orientador: Prof. Dr. Diego Gervasio Frías Suárez
Coorientador: Prof. Dr. Vagner de Souza Fonseca
Coorientadora: Prof. Dra. Evandra Strazza Rodrigues Sandoval

Salvador
2023

Teixeira Nogueira de Jesus, Jéssica
CARACTERIZAÇÃO DE MARCADORES GENÉTICOS RELACIONADOS À EXPRESSÃO DE ANTÍGENOS DE GRUPOS SANGUÍNEOS ERITROCITÁRIOS UTILIZANDO PROCESSAMENTO DE DADOS BRUTOS DE SEQUENCIAMENTO DE NOVA GERAÇÃO/ JESSICA TEIXEIRA NOGUEIRA DE JESUS. – Salvador, 2023.
42 p. : il. (algumas color.) ; 30 cm.

Orientador: Prof. Dr. Diego Gervasio Frías Suárez

Coorientador: Prof. Dr. Vagner de Souza Fonseca

Coorientadora: Prof. Dra. Evandra Strazza Rodrigues Sandoval

Monografia – UNIVERSIDADE DO ESTADO DA BAHIA
CAMPUS I - SALVADOR
CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO, 2023.

1. Bioinformática. 2. Grupos sanguíneos. 3. Marcadores genéticos 3. Sequenciamento de Nova Geração. 4. Pipeline de Genotipagem I. Título.

Termo de Anuência do Orientador

Declaro para os devidos fins que li e revisei este trabalho e atesto sua qualidade como resultado final desta monografia. Confirmo que o referencial teórico apresentado é completo e suficiente para fundamentar os objetivos propostos e que a metodologia científica utilizada e os resultados finais são consistentes e com qualidade suficiente para submissão à banca examinadora final do Trabalho de Conclusão de Curso II do curso de Bacharelado em Sistemas de Informação.

Prof. Dr. Diego Gervasio Frías Suárez

JESSICA TEIXEIRA NOGUEIRA DE JESUS

CARACTERIZAÇÃO DE MARCADORES GENÉTICOS RELACIONADOS À EXPRESSÃO DE ANTÍGENOS DE GRUPOS SANGUÍNEOS ERITROCITÁRIOS UTILIZANDO PROCESSAMENTO DE DADOS BRUTOS DE SEQUENCIAMENTO DE NOVA GERAÇÃO

Trabalho de Conclusão de Curso apresentado para obtenção do grau de Bacharel em Sistemas de Informação.

Data da Defesa: 11/12/2023

Conceito: Aprovada

Banca Examinadora

Prof. Dr. Diego Gervasio Frías Suárez
UNEB – Orientador

Prof. Dr. Vagner de Souza Fonseca
UNEB – Coorientador

Prof. Dra. Evandra Strazza Rodrigues
Fundação Hemocentro de Ribeirão Preto - Coorientadora

Prof. Dr. Alexandre Rafael Lenz
UNEB

Prof. Dra. Maria Inés Valderrama Restovic
UNEB

Salvador
2023

Este trabalho é dedicado ao povo brasileiro, em especial à classe trabalhadora, que através do seu suor e impostos, sustenta a universidade pública e a pesquisa científica no país.

AGRADECIMENTOS

Agradeço aos meus pais, Nelma e Edgar, por me criarem com amor e compreensão, por estarem ao meu lado em todos os momentos da minha vida e pelo constante esforço para me proporcionar uma educação de qualidade.

Agradeço aos meus amados irmãos Gabriel, Sophia, Guilherme e Lucas, por existirem e me incentivarem à ser um exemplo.

Aos meus primos Diego, Pedro Felipe, Gustavo e Alice, pelo incentivo e pelas boas conversas.

Aos meus tios, André, que sempre me socorre nos momentos difíceis e Simone, que sempre me ajuda e que me ensinou a dar os primeiros passos nesse mundo.

Agradeço aos meus amigos Juliana, Renan, Caio, Thales, Samantha, Marina, Cléber, Lucas, Ozzy, Lívia e Rebeca, por estarem comigo e me apoiarem em todos os momentos, sejam eles bons ou ruins.

Agradeço aos meus orientadores Diego Gervasio Frías Suárez, Vagner de Souza Fonseca e Evandra Strazza Rodrigues, pela dedicação, paciência, pela vontade de ensinar, pela orientação fantástica e por me darem a oportunidade de realizar uma pesquisa interdisciplinar.

Agradeço aos grandes mestres e doutores que ensinaram, na graduação, não somente ciência, mas cidadania. Em especial aos professores Cláudio Amorim, Alexandre Lenz, Maria Inés, Mônica Massa, Ernesto Massa e Ana Patrícia.

Aos colegas de curso, em especial Marcelo, Maurício, Ramon, Ricardo, Evelyn e Jefferson, com os quais compartilhei conhecimento, as alegrias e as agonias desta etapa final.

Por último, mas não menos importante, agradeço à minha psicóloga, Nice Lago, por cuidar da minha saúde mental e por me fazer acreditar que é possível superar qualquer obstáculo.

"O sucesso consiste em seguir de fracasso em fracasso sem perder o entusiasmo."
(Autor desconhecido)

RESUMO

Os sistemas ABO e RH desempenham um papel crucial na classificação sanguínea, determinando tipos como A, B, O e AB. A presença ou ausência do antígeno RhD, principalmente em casos RhD positivo/negativo, é um detalhe clínico discreto, mas de grande importância devido ao potencial perigo de reações hemolíticas. A complexidade e heterogeneidade desses sistemas ressaltam a necessidade urgente de aprimorar nosso entendimento. Este estudo focou na caracterização detalhada de marcadores genéticos relacionados à expressão de antígenos sanguíneos eritrocitários. Para atingir esse objetivo, utilizamos tecnologias de sequenciamento de nova geração baseadas em imuno-hematologia e biologia molecular. Desenvolvemos um protótipo inovador de *backend* web para análise de marcadores genéticos e grupos sanguíneos. O método tecnológico inovador, baseado em dados brutos de sequenciamento genético de próxima geração (NGS), proporciona compreensão profunda e acessível. Foi utilizada a metodologia de *Design Science Research (DSR)*, comprometemo-nos não apenas com a teoria, mas também com a aplicação prática de soluções. O ciclo adaptativo de testes, validações e correções demonstra nosso compromisso em alcançar e superar resultados esperados, proporcionando compreensão mais profunda da expressão genética nos grupos sanguíneos. Os resultados antecipados são promissores: identificamos 130 e 152 SNPs em duas amostras RHD, 233 SNPs nas amostras RHCE (incluindo um indel) e 167 SNPs evidenciando riqueza genética subjacente. Os resultados revelaram três mutações em uma amostra RHD, ampliando a compreensão das variações genéticas. Até agora, esses resultados não são apenas preliminares, mas indicativos concretos do progresso científico. Contribuímos para a compreensão da expressão de antígenos sanguíneos, acreditando que tais descobertas podem transformar a saúde. Além disso, esses resultados abrem caminho para futuras investigações, ressaltando a importância de uma abordagem médica mais individualizada.

Palavras-chave: Bioinformática. Grupos sanguíneos. Marcadores genéticos. Sequenciamento de Nova Geração. Pipeline de Genotipagem.

ABSTRACT

The ABO and RH systems play a crucial role in blood classification, determining types such as A, B, O, and AB. The presence or absence of the RhD antigen, especially in RhD positive/negative cases, is a discreet but clinically significant detail due to the potential danger of hemolytic reactions. The complexity and heterogeneity of these systems highlight the urgent need to enhance our understanding. This study focused on the detailed characterization of genetic markers related to the expression of erythrocyte blood antigens. We used next-generation sequencing technologies based on immuno-hematology and molecular biology to achieve this goal. We developed an innovative web backend prototype to analyze genetic markers and blood groups. The innovative technological method, based on raw data from next-generation genetic sequencing platforms (NGS), provides a profound and accessible understanding. Following the Design Science Research (DSR) methodology, we are committed to the theory and the practical application of solutions. The adaptive cycle of tests, validations, and corrections demonstrates our commitment to achieving and surpassing expected results, providing a deeper understanding of genetic expression in blood groups. The preliminary results are promising: we identified 130 and 152 SNPs in two RHD samples, 233 SNPs in RHCE samples (including an indel), and 167 SNPs indicating underlying genetic richness. The results revealed three mutations in an RHD sample, expanding our understanding of genetic variations. So far, these results are preliminary and concrete indicators of scientific progress. We have contributed to understanding the expression of blood antigens, believing that such discoveries can transform healthcare. Furthermore, these results pave the way for future investigations, emphasizing the importance of a more individualized medical approach.

Keywords: Bioinformatics. Blood groups. Genetic markers. Next-Generation Sequencing. Genotyping pipeline.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

DSR	*Design Science Research*
ISBT	*International Society of Blood Transfusion*
NGS	*Next-Generation Sequencing*
NJ	*Neighbour Joining*
PCR	*Polymerase Chain Reaction*
RAPD	*Random Amplification of Polymorphic DNA*
RFLP	*Restriction Fragment Length Polymorphism*
SNP	*Single Nucleotide Polymorphism*

SUMÁRIO

1 INTRODUÇÃO

1.1 Contexto

A Bioinformática é um campo interdisciplinar, que possui como objetivo o desenvolvimento de técnicas e ferramentas computacionais, visando auxiliar a compreensão de dados biológicos. Uma das formas de estudar um organismo é a análise genômica, que se caracteriza como um processo de "leitura" do genoma do organismo em questão. Essa leitura, apesar de ser uma tarefa aparentemente simples, é computacionalmente complexa, devido ao tamanho dos genomas e à dificuldade de encontrar informações relevantes rapidamente (LESK, 2008). O genoma humano é, fundamentalmente complexo, e as informações sobre seu sequenciamento podem ser comparadas ao programa espacial que enviou pessoas à lua, como um dos maiores avanços em realizações tecnológicas do último século (LESK, 2008). Nesse contexto, é indispensável a busca pelo aperfeiçoamento das tecnologias da bioinformática, para atender à demanda crescente de processamento de dados biológicos, pois isso tem permitido avanços significativos em diversas áreas da biologia e da medicina. Um exemplo disso é a nossa compreensão acerca dos grupos sanguíneos.

Os grupos sanguíneos são proteínas expressas na superfície das hemácias que apresentam grande importância clínica devido ao alto risco de reações hemolíticas. Foram descritos no início do século XX, sendo os sistemas ABO e RH os mais imunogênicos e polimórficos. Os antígenos do sistema ABO são responsáveis por classificar o sangue em quatro tipos diferentes: A, B, O e AB, e o antígeno Rh, conhecido como fator Rh, é expresso em indivíduos classificados com RhD positivo e não é expresso em indivíduos RhD negativos. Avanços nos estudos de grupos sanguíneos eritrocitários mostram que existe uma enorme complexidade e heterogeneidade entre os grupos sanguíneos. A *International Society of Blood Transfusion (ISBT)* atualmente reconhece mais de 360 antígenos eritrocitários, que estão classificados em 43 grupos sanguíneos (JADHAO et al., 2022). O método padrão usado para determinar o status Rh em receptores e doadores de sangue é a triagem sorológica com anticorpos monoclonais anti-D, no entanto, na maioria dos casos, esse método revela apenas a presença de uma variante RhD, mas não determina exatamente qual variante RhD parcial ou RhD fraco está presente. Neste sentido, para a confirmação e correta identificação de amostras RhD são indicados os testes moleculares. A genotipagem de grupos sanguíneos eritrocitários oferece suporte para um procedimento de transfusão sanguínea seguro e eficaz, além disso, auxilia na detecção de variantes e marcadores genéticos moleculares de grande complexidade (RODRIGUES et al., 2021).

Marcadores genéticos moleculares são sequências de DNA ou polimorfismos, que permitem a caracterização e diferenciação de um indivíduo, e que são reproduzidas em seus descendentes. O DNA mitocondrial é um exemplo de marcador genético, assim como os

antígenos de grupos sanguíneos eritrocitários. Os sistemas de grupos sanguíneos consistem em marcadores clinicamente essenciais em transfusões de sangue, transplantes de órgãos, obstetrícia, na incompatibilidade materno-fetal. Além disso, são usados em medicina legal e genética forense, na identificação individual e na investigação de paternidade (BORGES-OSóRIO; ROBINSON, 2013). Existem diversos tipos de marcadores genéticos e cada um deles possui vantagens e desvantagens. O que determina a escolha de um marcador é o estudo das diferenças entre os tipos, a aplicação para a qual se destina e os recursos, técnicos e financeiros, disponíveis. Com o surgimento de plataformas de sequenciamento de alto desempenho, atualmente denominadas plataformas de *NGS*, tornou-se possível a identificação de marcadores em escala genômica. Essas técnicas incorporam aspectos vantajosos de várias outras, aumentando a sensibilidade e resolução.

Os métodos mais utilizados para a genotipagem de antígenos eritrocitários são *Polymerase Chain Reaction (PCR)* seguida de *Restriction Fragment Length Polymorphism (RFLP)*, o *PCR* alelo específico, sequenciamento de DNA e o *PCR* em tempo real (FUKUMORI et al., 1995; POLIN et al., 2008; RODRIGUES et al., 2015). Essas técnicas moleculares trouxeram grandes benefícios para imuno-hematologia, porém a capacidade de automação e de processamento de um grande número de amostras ainda é uma limitação. O *Next-Generation Sequencing (NGS)*, é uma tecnologia que se refere às segunda e terceira gerações de técnicas de sequenciamento genômico capaz de superar as limitações de genotipagem de grupos sanguíneos, pois é capaz de sequenciar genomas humanos inteiros com rapidez. Esta tecnologia fornece informações detalhadas sobre a sequência de nucleotídeos a partir de várias leituras gênicas e pode ser aplicada para avaliar um grande número de indivíduos simultaneamente.

Todas as técnicas de Nova Geração compartilham a característica de realizar leituras curtas ou longas e a possibilidade de sequenciar muitos milhões de fragmentos simultaneamente. O *NGS* diminuiu os custos por *megabase* e aumentou, drasticamente, a quantidade de dados processados (FÜRST et al., 2020).

Com o objetivo de caracterizar marcadores genéticos relacionados à expressão de antígenos de grupos sanguíneos eritrocitários utilizando processamento de dados brutos de *NGS*, este trabalho visa compreender os grupos sanguíneos eritrocitários, os marcadores genéticos, as diferentes técnicas de genotipagem existentes e o processamento de dados brutos provenientes de *NGS*; e à partir desse estudo, desenvolver um protótipo de *backend* de ambiente web que permita a caracterização de grupos sanguíneos e marcadores genéticos provenientes de dados brutos de sequenciamento genético em plataformas de *Next-Generation Sequencing (NGS)*.

1.2 Justificativa

As transfusões de sangue no Brasil aumentaram de 2016 até 2019, passando de 2,8 milhões para aproximadamente 2,95 milhões de transfusões realizadas (CASSIANO, 2020). Dentre os pacientes que realizam o procedimento estão portadores de anemia falciforme, talassemia e outras doenças, que geralmente necessitam de transfusões sanguíneas recorrentes. Equívocos na identificação de qualquer uma dessas variantes pode acarretar em aloimunização desses pacientes, tornando desafiadora a tarefa de encontrar doadores com fenótipo compatível, podendo atrasar o procedimento de transfusão e afetar a saúde dos pacientes de forma negativa (ORZIńSKA et al., 2018). Podem também trazer riscos de efeitos adversos como Reação Hemolítica Transfusional, complicações na gestação, reações alérgicas, entre outros (JADHAO et al., 2022). A genotipagem de grupos sanguíneos eritrocitários pode mitigar o risco de aloimunização e outras complicações que possam decorrer do procedimento de transfusão sanguínea, uma vez que será possível analisar o sangue coletado e, à partir dessa análise, selecionar a melhor opção entre doadores. O conceito de genotipagem de antígenos eritrocitários foi introduzido a partir da década de 1990, e ao longo dos anos a análise molecular se mostrou precisa e muito útil na identificação de amostras RHD variantes. Em países desenvolvidos, a genotipagem é utilizada com propósito assistencial para melhorar a prática hemoterápica e também com o intuito investigativo para a caracterização molecular dos genes que codificam os grupos sanguíneos. A seleção mais aprimorada de unidades transfusionais pode otimizar o aproveitamento e gerenciamento dos estoques de bolsas de sangue, especialmente quando pensamos em tipos mais raros de grupos sanguíneos.

Nos últimos anos, as plataformas moleculares para identificação de grupos sanguíneos ajustados para uma escala de alto rendimento têm sido usadas por alguns países para triagem em massa de doadores de sangue. No caso das plataformas de genotipagem baseadas em *microarray* são consideradas com grande sensibilidade e especificidade e algumas possuem acreditação pelo FDA para serem usadas com propósito clínico. No entanto, os ensaios de *microarray* possuem limitações que devem ser consideradas, pois detectam apenas alterações de nucleotídeo único com base na suposição de que a sequência próxima ou ao redor do *SNP* alvo/*SNP* de referência conhecida e complementar ao iniciador/sonda utilizado (ORZIńSKA et al., 2018). Assim, apenas os alelos incorporados na lâmina de *microarray* podem ser avaliados e novos alelos não são considerados na análise. Outra limitação é que as plataformas de *microarray* comerciais foram desenvolvidas para uma população específica como, por exemplo, para Europeus ou Americanos. E no caso da população brasileira, onde existe uma grande miscigenação e representação étnica diversificada, pode ser muito diferente quando comparada com indivíduos Europeus ou Americanos. O *NGS* é uma tecnologia que contorna essas limitações, combinando sequenciamento complexo com a capacidade de sequenciar genomas inteiros (ORZIńSKA et al., 2018).

Do ponto de vista imuno-hematológico, *NGS* de regiões *SNPs* permite a triagem de antígenos clinicamente importantes para muitos indivíduos em um único teste e é um método promissor para genotipagem massiva de grupos sanguíneos em doadores de sangue (ORZIńSKA et al., 2018) Além disso, seu custo tem diminuído consideravelmente ao longo dos anos (FÜRST et al., 2020). Podemos concluir que a caracterização e a identificação de grupos sanguíneos eritrocitários utilizando plataformas de *NGS* pode, portanto, contribuir para a melhoria da qualidade de vida de pacientes que necessitam de transfusões sanguíneas, transplantes de órgãos e de outros procedimentos médicos que dependam da análise desses grupos; e do sistema de saúde público, a um baixo custo e em larga escala.

1.3 Objetivos

1.3.1 Objetivo Geral

Desenvolver um protótipo de *backend* de ambiente web que permita a caracterização de marcadores genéticos de grupos sanguíneos provenientes de dados brutos de sequenciamento genético em plataformas de *Next-Generation Sequencing (NGS)*.

1.3.2 Objetivos Específicos

- Avaliar e compreender as técnicas no processo de caracterização de marcadores genéticos de grupos sanguíneos.
- Desenvolver um pipeline para leitura rápida dos dados brutos de plataformas de *NGS* para caracterização de marcadores genéticos de grupos sanguíneos.
- Validar os resultados do pipeline por especialista humano, em um conjunto selecionado de amostras.

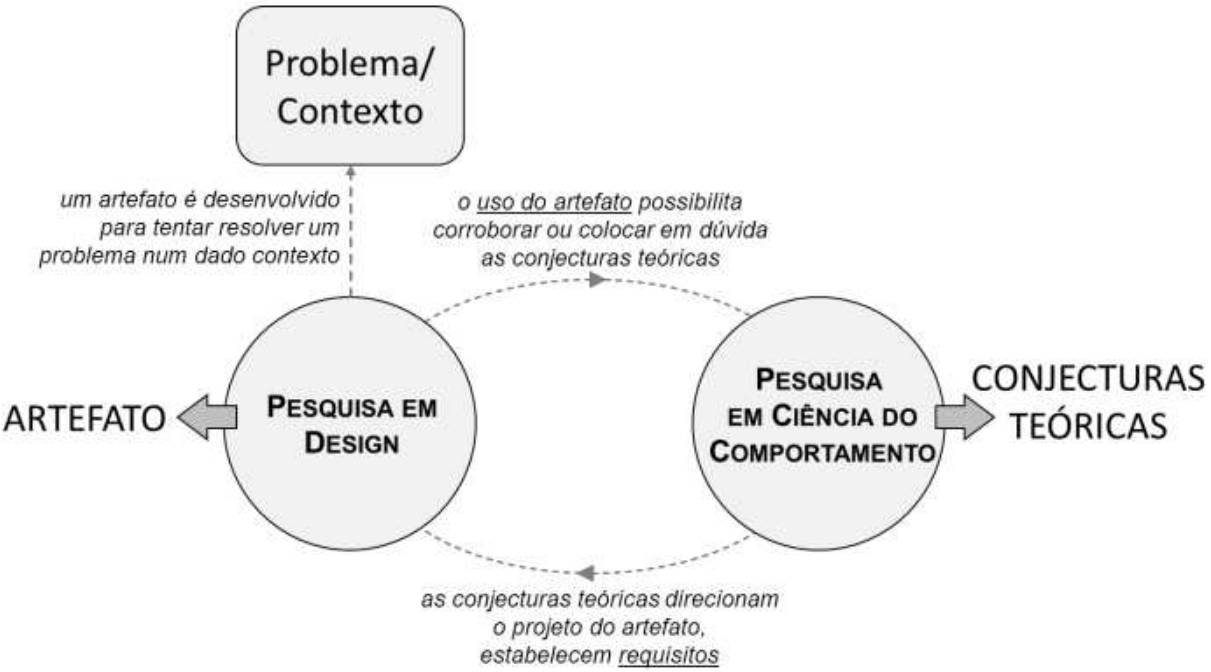
1.4 Metodologia

A metodologia selecionada para este estudo foi a DSR (*Design Science Research*) pois esta se concentra na criação de artefatos, soluções ou sistemas que abordem problemas práticos do mundo real.

Ao utilizar essa metodologia o pesquisador possui dois objetivos: resolver um problema prático no contexto específico do artefato e gerar um novo conhecimento, criando dois principais ciclos de pesquisa inter-relacionados, chamados Ciclo de Design e Ciclo de Engenharia. Evidenciado o problema, o artefato é desenvolvido para afirmar ou colocar

em dúvida as conjecturas teóricas que geram um ciclo de melhoria/direcionamento, como na Figura 1.

Figura 1 – Ciclos em Design Science Research



Fonte: (PIMENTEL; FILIPPO; SANTORO, 2019)

O Ciclo de Engenharia tem como principais características as avaliações do artefato, buscando melhorias e refinamento do projeto (PIMENTEL; FILIPPO; SANTORO, 2019) ao seguir cinco etapas: investigação do problema, design da solução, validação da solução, implementação da solução e avaliação dos resultados.

O Ciclo Empírico baseia-se em teorias e métodos científicos para garantir que a condução da pesquisa seja realizada a rigor teórico e metodológico (PIMENTEL; FILIPPO; SANTORO, 2019) em sete etapas: contexto da pesquisa, análise do problema de pesquisa, pesquisa e inferência do design, validação, execução da pesquisa, análise dos dados e contribuição da pesquisa.

Por fim, o Ciclo de Relevância relaciona o contexto ao artefato projetado para atingir o objetivo da pesquisa, levando em consideração o ambiente, pessoas, problemas e oportunidades. Desta maneira, identificou-se ser necessário realizar três avaliações: se o artefato satisfaz aos requisitos; se o problema foi resolvido de forma satisfatória; e se as conjecturas teóricas parecem válidas (PIMENTEL; FILIPPO; SANTORO, 2019).

Desta maneira, foi constatada a necessidade de realizar três avaliações em pesquisas concebidas no paradigma da DSR na figura 2: se o artefato satisfaz aos requisitos; se o problema foi resolvido satisfatoriamente; e se as conjecturas teóricas parecem válidas (PIMENTEL; FILIPPO; SANTORO, 2019). A figura 2 ilustra o mapa de elementos esperados da pesquisa na abordagem DSR.

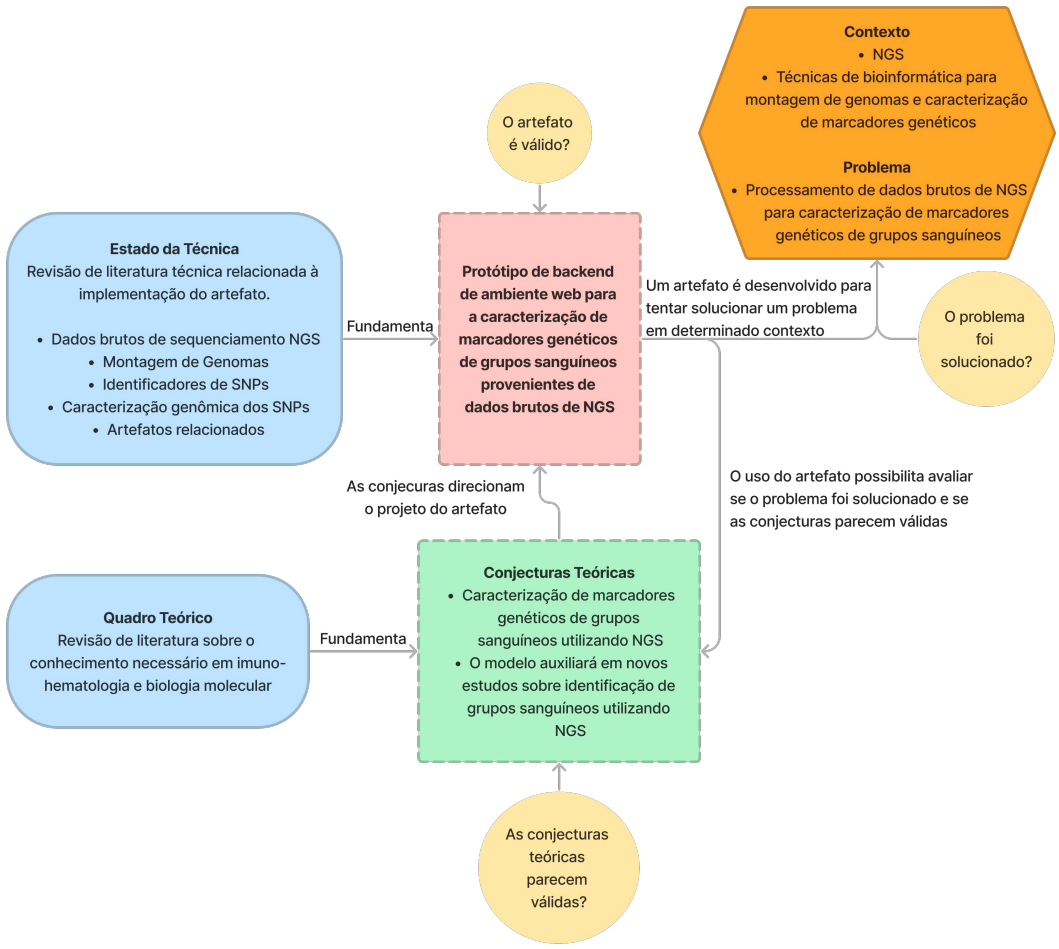


Figura 2 – Diagrama da pesquisa adaptado de (PIMENTEL; FILIPPO; SANTORO, 2019) para este projeto de pesquisa.

Com base no diagrama, definiu-se o objetivo de desenvolver um protótipo de *backend* de ambiente web que permita a caracterização de marcadores genéticos de grupos sanguíneos provenientes de dados brutos de sequenciamento genético em plataformas de *NGS*, baseando-se nas conjecturas teóricas das áreas de imuno-hematologia e biologia molecular; e nas técnicas para a criação do artefato. A solução será desenvolvida de forma cíclica adaptativa, realizando testes, validações e correções, quando necessário, até alcançar os objetivos elencados neste trabalho.

1.5 Estrutura do Trabalho

O restante do texto está organizado da seguinte maneira: O capítulo 2 apresenta o referencial teórico necessário para a compreensão deste trabalho, bem como técnicas consolidadas para a identificação de marcadores genéticos relacionados à expressão de antígenos

de grupos sanguíneos eritrocitários. O capítulo 3 descreve o ambiente de desenvolvimento necessário para a execução do estudo. O capítulo 4 apresenta os resultados alcançados e o capítulo 5 apresenta as considerações finais e trabalhos futuros.

2 REFERENCIAL TEÓRICO

Neste capítulo são apresentados os conceitos que embasam o desenvolvimento desta pesquisa.

2.1 Grupos Sanguíneos

No início do século XX ocorreu a descoberta dos antígenos eritrocitários, sendo considerada um dos avanços mais importantes na área médica transfusional. Eritrócitos são as hemácias, ou glóbulos vermelhos. São as células mais numerosas do sangue e se caracterizam por terem vida média de 120 dias, não possuírem núcleo nem organelas, além de possuir uma membrana celular rica em enzimas. Seu principal componente é a Hemoglobina (Hb), que pela presença de Ferro (Fe) em sua composição, torna o sangue avermelhado (BORGES-OSóRIO; ROBINSON, 2013). O primeiro sistema de grupos sanguíneos foi descoberto em 1900 pelo médico e cientista austríaco Karl Landsteiner, que se baseou na expressão dos antígenos presentes na superfície das hemácias. A maioria dos antígenos de grupos sanguíneos são formados por polimorfismos de um único nucleotídeo e portanto podem ser identificados por técnicas moleculares.

Até os dias atuais, o sistema "ABO" é comumente utilizado para classificar o sangue humano. Entretanto, recentes avanços em estudos de grupos sanguíneos eritrocitários mostram que existe uma enorme complexidade e heterogeneidade entre os grupos sanguíneos. A *International Society of Blood Transfusion (ISBT)* atualmente reconhece mais de 360 antígenos eritrocitários, distribuídos em 43 grupos sanguíneos, codificados por 48 genes e definidos por mais de 1500 alelos (JADHAO et al., 2022).

Os objetos de estudo deste trabalho são os genes do complexo Rh, denominados RHD e RHCE.

2.1.1 Estrutura do complexo Rh nas hemácias

O grupo sanguíneo Rh (ISBT 004) é o mais complexo, polimórfico e imunogênico sistema de grupo sanguíneo já conhecido em humanos. Após o ABO, é o mais importante na medicina transfusional. Os genes homólogos RHD e RHCE são responsáveis pela expressão de cinco principais antígenos, D(RH1), C(RH2), E(RH3), c(RH4) e e(RH5), que representam a maioria dos anticorpos clinicamente significantes. Com mais de 49 diferentes antígenos caracterizados, é o maior de todos os sistemas sanguíneos. As hemácias Rh positivo e Rh negativo referem-se à presença ou ausência do antígeno D, porém ambas expressam os antígenos C\c e E\e. O antígeno C é antitético, isto é, contrário ao c enquanto que o antígeno E ao e. Cada cromossomo contém os genes C ou c e E ou e. Os antígenos estão localizados em duas proteínas expressas na membrana dos eritrócitos e seus precussores

imediatos: RhD (CD240D) e a RhCE (CD240CE), que carregam respectivamente os antígenos D(Rh1) e os C, c, E, e (Rh2-Rh5) em várias combinações (ce, cE, Ce e CE) (NARDOZZA et al., 2010). Ambas as proteínas RHD e RHCE são hidrofóbicas e não glicolisadas, cada uma com peso molecular de 30 a 32 KD, compostas de 417 aminoácidos que se distribuem em seis segmentos extracelulares, responsáveis diretos pela resposta imune; 12 transmembranosos e 7 intracelulares. As porções N-terminal e C-terminal são intracelulares. As proteínas RhD e RhCE apresentam 92% de homologia e se diferenciam em 35/36 aminoácidos (8,4% de divergência), sugerindo que os genes correspondentes são resultado de uma duplicação de um gene ancestral comum. Este conceito é baseado no fato da identificação de genes Rh-*like* em primatas não-humanos. As diferenças entre RHD e RHCE ocorrem na região extracelular, nas quais são restritas as alças (porções extracelulares) 2, 3, 4 e 6 (Figura 3) (NARDOZZA et al., 2010).

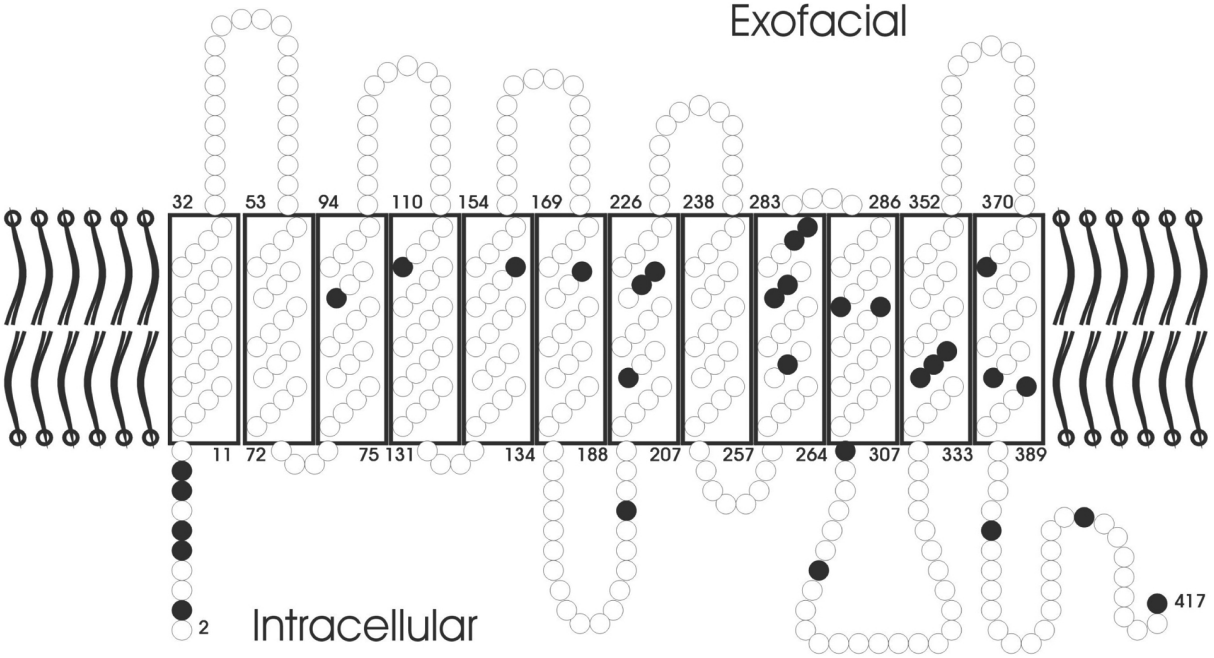


Figura 3 – Topologia prevista de RhD na membrana de hemácias. Os aminoácidos estão representados como círculos. Os círculos pretos indicam substituições de aminoácidos, cada uma das quais foi correlacionada com um tipo D fraco molecularmente distinto. (WAGNER; FLEGEL, 2004)

2.1.2 Base Molecular do sistema Rh

O gene RHD foi descoberto em 1992, dois anos após o RHCE (o sistema Rh foi descoberto em 1940). Apesar da existência de mais de 170 alelos RHD descritos, este gene ainda não foi completamente caracterizado. Experimentos feitos utilizando a técnica *Southern Blot* com sonda cDNA Rh demonstraram que somente três espécies carregam mais de um gene Rh: chimpanzés, gorilas e humanos. Os dois genes do sistema Rh (RHD e RHCE) estão localizados no braço curto do cromossomo 1, lócus 34-36 (NARDOZZA et al., 2010). (Ver figura 4).

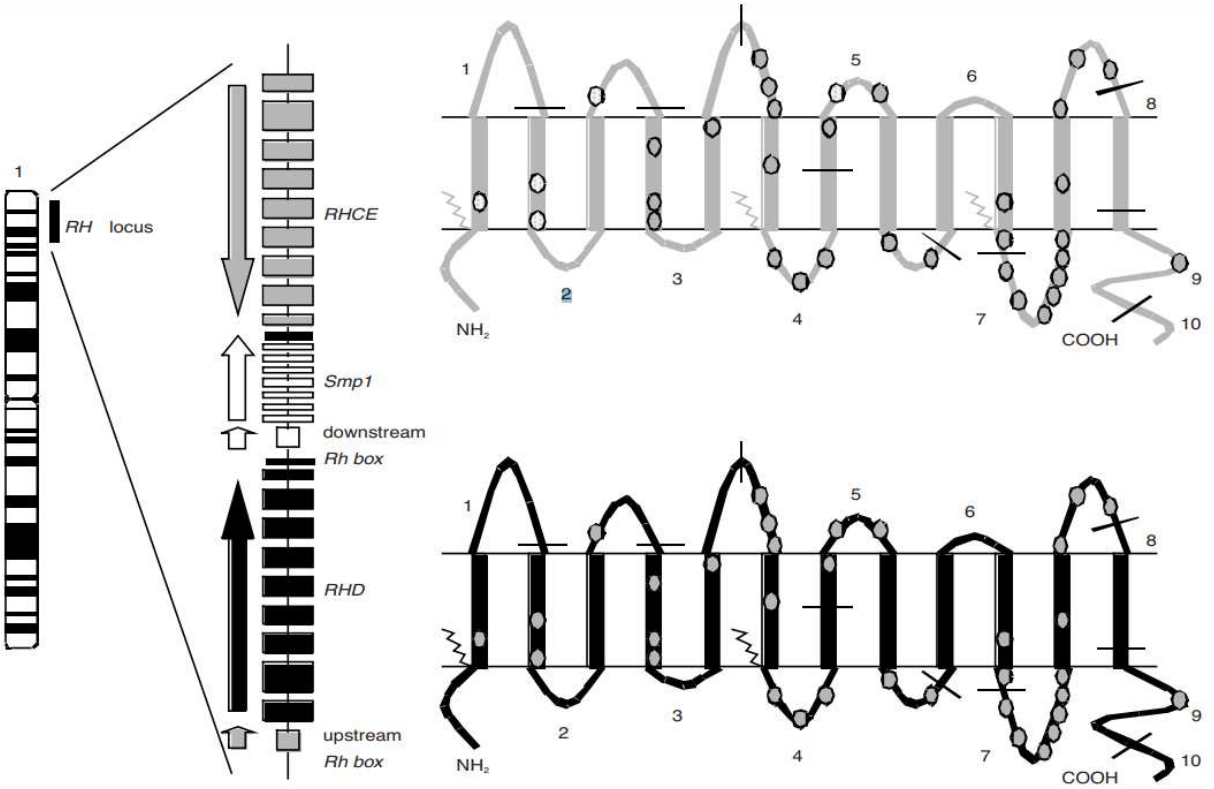


Figura 4 – Representação do arranjo cromossômico do locus Rh, os genes RHD e RHCE e seus polipeptídeos (TAX, 2005)

São genes altamente similares (93,8%), contendo cada um 10 éxons, com uma sequência total de aproximadamente 60.000 pares de base. A maior diferença está no íntron 4, onde o RHD contém uma deleção de 600 pares de base em relação ao RHCE. Eles estão em orientação opostas pelos terminais 3´ e separados por uma sequência de 30.000 pares de base. O gene SMP1(Small Membrane Protein 1) está localizado entre o RHD e o RHCE. Ele é funcionalmente relacionado ao Rh com relação à expressão deste na membrana eritrocitária. O RHD é cercado por dois segmentos de DNA com tamanho de 9000 pares de base, similaridade de 98,6% e orientação idêntica, denominados *Rhesus boxes* (NARDOZZA et al., 2010).

2.1.3 Variantes RhD

Embora a maioria das pessoas sejam classificadas com fenótipo RhD positivo ou RhD negativo, também existem variantes do antígeno RhD denominadas RhD Fraco, RhD Parcial e DEL.

Os três principais mecanismos moleculares responsáveis pelo fenótipo RhD negativo são: deleção total do gene RHD, pseudogene RHD e gene híbrido, que variam de frequência de acordo com a raça em questão (NARDOZZA et al., 2010). O gene RHD pode ainda não ser expresso devido a um códon de parada prematuro, inserções de nucleotídeos, pontos de mutação ou RHD/CE híbrido. Atualmente existem diferentes alelos D- descritos, com

uma frequência de 1/1500 na população caucasiana (NARDOZZA et al., 2010).

As variantes RhD na maioria das vezes são atribuídas a uma expressão enfraquecida do antígeno D na superfície da hemácia através da reação com soro anti-D. Uma reação forte é esperada na presença de uma expressão de antígeno D normal. Entretanto, reatividade fraca inesperada pode ocorrer e geralmente está relacionada à variantes fracas ou parciais, que podem causar expressão de D anormal (RODRIGUES et al., 2021).

Normalmente hemácias RhD positivas possuem uma densidade antigênica variando entre 15.000 a 33.000 antígenos por célula. Contudo, alguns fenótipos foram identificados com densidade variando entre 70 e 5.200 antígenos RhD. Esses fenótipos são denominados de D fracos e são causados pela substituição de aminoácidos nas porções transmembranosas e intracelulares da proteína RhD, devido a uma única mutação *missense* no gene RHD. Hemácias com fenótipo D fraco expressam um antígeno RhD intacto, ocorrendo em 0,2% a 1% dos caucasianos (NARDOZZA et al., 2010).

As variantes RhD parciais são definidas pela ausência de um ou mais epítopos causados pelos rearranjos dos genes RHD e RHCE. Essa configuração genética propicia microconversões e trocas unidirecionais de fragmentos de gene RHD e RHCE, ou parte deles, levando à formação de alelos RHD-CE-D ou RHCE-D-CE, respectivamente (NARDOZZA et al., 2010).

Um estudo conduzido por (RODRIGUES et al., 2021) afirma que estudos que estimam a frequência de fenótipos incomuns RhD em diferentes regiões são importantes para determinar protocolos para identificação correta de grupos RhD, bem como conduzir políticas públicas para assegurar transfusões sanguíneas seguras. No estudo em questão, foi possível determinar alelos RHD em uma população de doadores de sangue do sudeste do Brasil que possuía fenótipos D atípicos.

2.2 Marcadores Genéticos, Genotipagem e Árvores Filogenéticas

Com o avanço das técnicas de biologia molecular, a manipulação do DNA em laboratório tornou-se uma técnica recorrente. No início dos anos 1980, o uso de marcadores moleculares passou a integrar rotineiramente a análise do DNA das mais diversas espécies. Desde então, eles vêm sendo aperfeiçoados e evoluídos juntamente com os avanços nas técnicas de sequenciamento em larga escala (TURCHETTO-ZOLET et al., 2017).

Existem diversos tipos de marcadores genéticos e cada um deles possui vantagens e desvantagens. O que determina a escolha de um marcador é o estudo das diferenças entre os tipos, a aplicação para a qual se destina e os recursos, técnicos e financeiros, disponíveis. Os marcadores de DNA são divididos em três categorias principais: os baseados

em hibridização, os baseados em Reação em Cadeia da Polimerase (em inglês, *Polymerase Chain Reaction*) e por fim, marcadores baseados em sequenciamento (TURCHETTO-ZOLET et al., 2017).

2.2.1 Genotipagem

Os métodos de genotipagem visam estudar o genoma de determinado organismo, analisando as características do polimorfismo genético concomitante dos mesmos. Baseiam-se na localização do material genético do organismo, o que permite gerar novas alterações no padrão de expressão genética e fornece alternativas mais estáveis e reprodutíveis (MERCHáN; CAICEDO; TORRES, 2017). São divididos em dois tipos:

- *in vitro*: O processo biológico é realizado em ambiente fechado e controlado, utilizando reagentes químicos.
- *in silico*: O processo biológico é realizado através de simulação computacional.

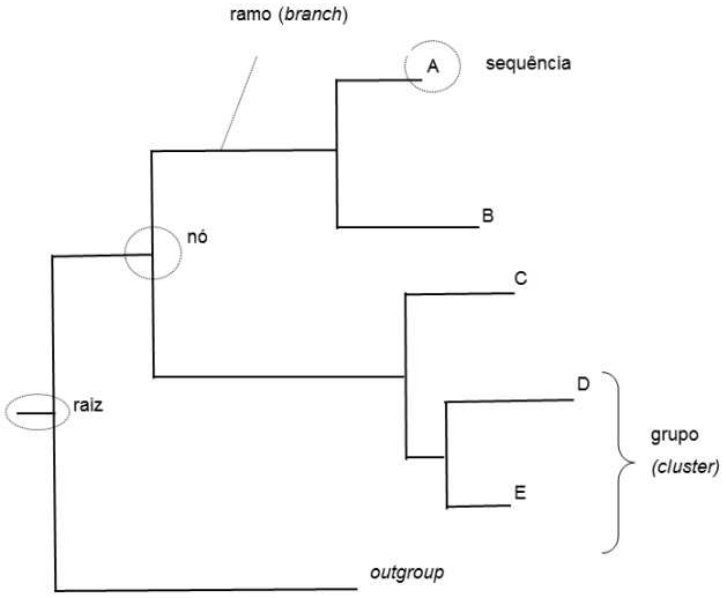
Métodos *in vitro*:

- *PCR*: Amplifica um fragmento de DNA específico a partir de uma ou poucas cópias de DNA. Baseia-se no processo de replicação do DNA que ocorre *in vitro* (MERCHáN; CAICEDO; TORRES, 2017).
- Hibridização: Análise para detectar a presença de ácidos nucleicos (DNA/RNA). Utilizada uma sonda para detectar uma molécula alvo à partir de uma sequência complementar à ela (MERCHáN; CAICEDO; TORRES, 2017).
- PRTF (em inglês, *RFLP*): Método molecular comumente utilizado por sua rapidez na identificação de *SNPs* específicos, baixo custo e especificidade. É baseado nas diferenças entre as sequências específicas de DNA de cada indivíduo que são reconhecidas por diferentes enzimas de restrição que reconhecem e clivam as sequências de DNA[1] (MERCHáN; CAICEDO; TORRES, 2017).

Métodos *in silico*:

- Árvore Filogenética: Uma árvore filogenética (Figura 5) é uma representação gráfica de relações ancestral-descendente entre organismos ou sequências genéticas e deve ser considerada como uma hipótese de um relacionamento evolutivo entre um grupo de organismos. (CALDART et al., 2016)

Figura 5 – Exemplo de Árvore Filogenética



Fonte: (CALDART et al., 2016)

2.2.2 Árvores Filogenéticas

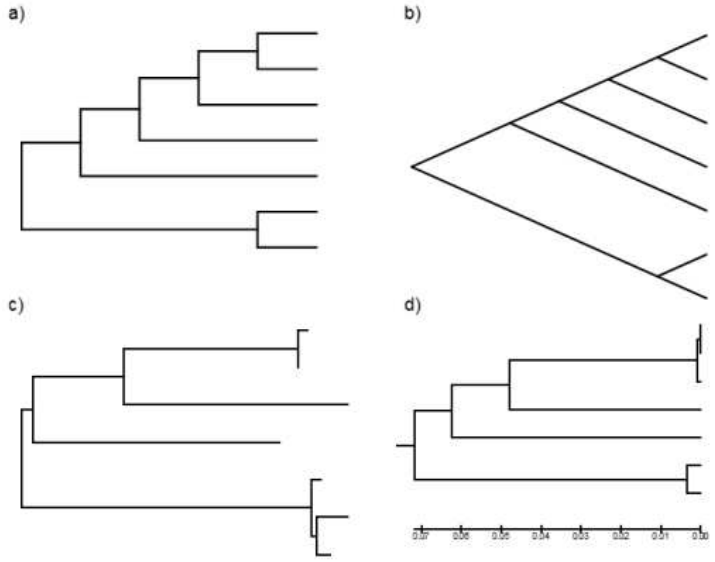
As árvores filogenéticas podem ser enraizadas ou não. As que possuem raíz contemplam uma sequência chamada grupo externo (*outgroup*), cuja função é dar à árvore uma direção evolutiva, mostrando quem são as sequências mais ancestrais (CALDART et al., 2016). As espécies de interesse se encontram nas extremidades de linhas chamadas de ramos da árvore (*tips*). Nos ramos (*branch*) encontram-se as sequências de proteínas ou nucleotídeos. Em árvores não enraizadas não existe nenhuma indicação de qual nó representa o ancestral comum das demais sequências. Cada linha horizontal na árvore representa uma série de ancestrais, levando até a espécie em seu final.

Duas sequências estão mais relacionadas se possuem um ancestral comum mais recente e menos relacionadas se o ancestral comum estiver mais distante. Se observarmos a Figura 5, podemos concluir que o nó que liga A e B é o ancestral comum mais recente. A está mais relacionada com B, do que com E, pois é necessário "percorrer" a árvore em 2 (dois) nó para A e 3 (três) nós para E para ligar as duas espécies à um ancestral comum, enquanto que A e B estão separados por apenas 1 (um) nó. Existem várias formas de representar uma árvore filogenética, como está evidente na figura 6, e essa representação vai depender de sua finalidade.

2.3 Sequenciamento de Nova Geração

O *NGS* é uma tecnologia que se refere às segunda ou terceira gerações de técnicas de sequenciamento genômico, que sucederam a primeira geração, baseada em Método de Sanger. Todas as técnicas de Nova Geração (2ª ou 3ª gerações) compartilham a

Figura 6 – Diferentes representações de uma árvore filogenética



Fonte: (CALDART et al., 2016) a) cladograma retangular construído com método NJ; b) cladograma clássico construído com método NJ; c) filograma representando relacionamento filogenético construído com método NJ; d) árvore construída pelo método de distância UPGMA representando uma árvore ultra métrica

característica de realizar leituras curtas ou longas e a possibilidade de sequenciar muitos milhões de fragmentos simultaneamente. As técnicas de terceira geração se caracterizam pelo sequenciamento direto de fragmentos de única sequência. Este último método permite leituras mais longas, porém a qualidade da sequência é inferior, se comparada com técnicas de 2ª geração. O *NGS* diminuiu os custos por *megabase* e aumentou, drasticamente, a quantidade de dados processados (FÜRST et al., 2020).

O *NGS* permite uma avaliação sem precedentes de mudanças genéticas, incluindo novas mudanças e Variações Estruturais (VEs) complexas, representando o novo melhor padrão para genotipagem de antígenos de hemácias (LANE, 2021).

2.3.1 Dados brutos de *NGS*

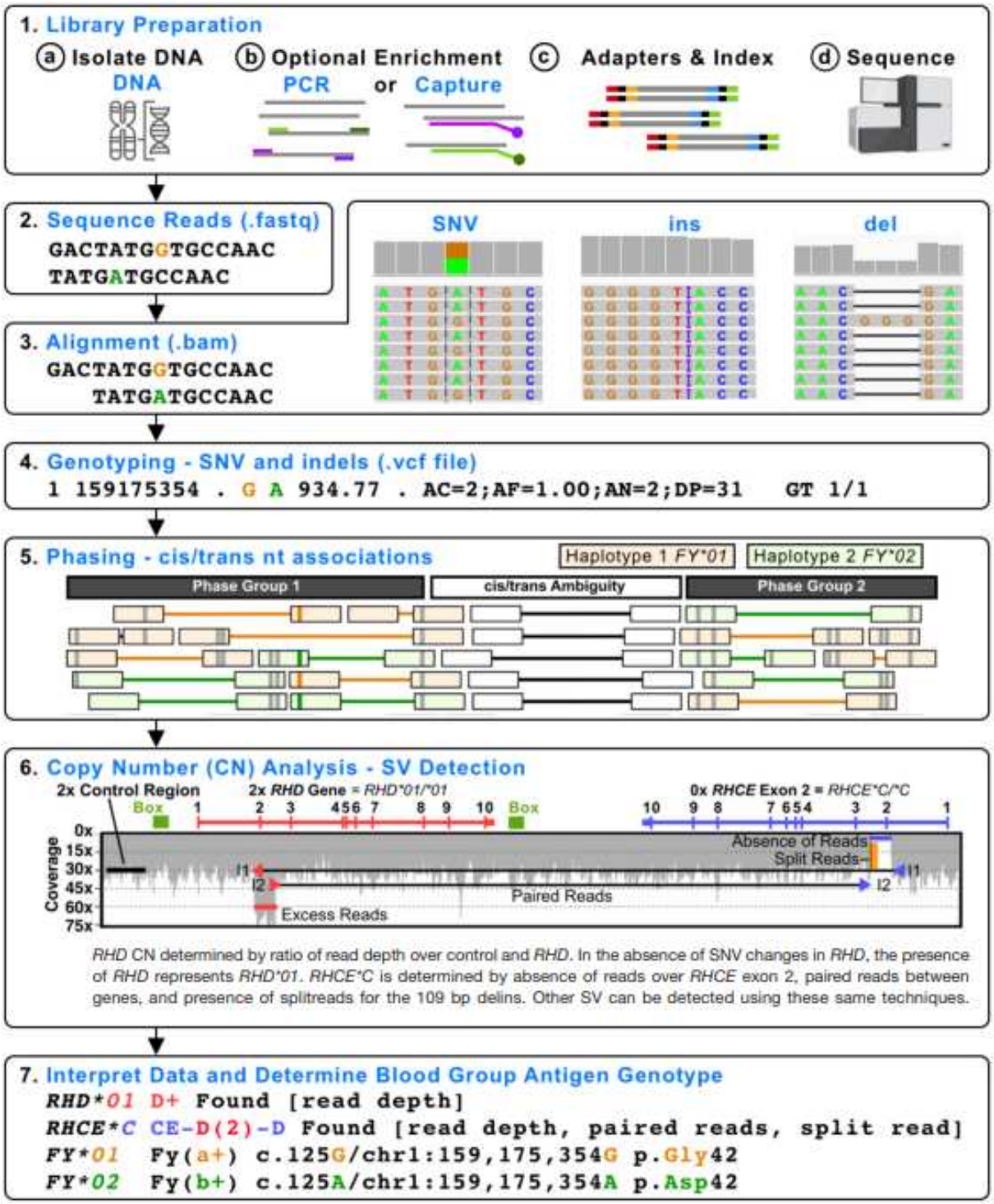
Para atingir o objetivo de caracterizar marcadores genéticos relacionados à expressão de antígenos de grupos sanguíneos eritrocitários através do processamento dos dados de *NGS*, é necessário, primeiramente, fazer o tratamento desses dados.

Em artigo recente, (LANE, 2021) propõe um *workflow* para a genotipagem de grupos sanguíneos baseado em *NGS*, representado na figura 7.

O *workflow*, representado na figura 7, consiste em 7 (sete) passos:

1. Preparação
2. Leitura das sequências: O DNA é sequenciado e armazenado como um arquivo .fastq

Figura 7 – Proposta de *workflow* de genotipagem baseado em *NGS*



Fonte: (LANE, 2021)

3. Alinhamento: É feito o alinhamento das sequências ao genoma de referência, que em seguida é armazenado em um arquivo .bam

4. Genotipagem, *SNPs e indels*: Os métodos de genotipagem são usados para determinar variações de nucleotídeo único (*SNPs)* e inserção e deleção de nucleotídeos, que então são salvos em um arquivo .vcf

5. Faseamento
6. Análise de número de cópias para detecção de *Structural Variations (SVs)*
7. Interpretar os dados e determinar o genótipo de grupo sanguíneo da amostra

2.3.1.1 FASTQ

O formato FASTQ (representado na figura 8) é um formato baseado em texto utilizado para armazenar os dados do sequenciamento realizado. O arquivo FASTQ possui 4 linhas por sequência (ILLUMINA, 2021):

1. A primeira linha começa com o caractere "@", seguido do identificador da sequência e uma descrição (opcional)
2. A segunda linha contém a sequência propriamente dita, representada através de letras (A, C, T, G e N)
3. A terceira linha contém um separador, que é simplesmente o caractere "+"
4. A quarta linha contém os *quality scores*

Figura 8 – Exemplo de representação de uma sequência em um arquivo FASTQ

```
@ML-P2-14:9:000H003HG:1:11102:17290:1073 1:N:0:TCCTGAGC+GCGATCTA
TTTGGTAACAGCATGAATTATTCTAGCCACTAAAACTCTATGAACATCTTGTGAAGGTTTCAGATAGAGCCTGAAGTACACAGAGAACAATTCTTAAAAAA
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE<AEEEEEEE
```

Fonte: (ILLUMINA, 2021)

2.3.1.2 BAM

Um arquivo .bam é a versão binária comprimida de um arquivo .sam e é utilizado para representar sequências alinhadas. Um arquivo .bam contém uma seção de cabeçalho e uma seção de alinhamento. A seção de cabeçalho contém informações sobre todo o arquivo, como nome da amostra, tamanho da amostra e método de alinhamento. Alinhamentos na seção de alinhamentos são associados com informações específicas na seção de cabeçalho (ILLUMINA, 2022). A seção de alinhamentos inclui as seguintes informações para cada par de leitura:

- *Read group*, que indica o número de leituras para uma amostra específica
- Etiqueta de código de barras, que indica o ID de amostra demultiplexada associada à leitura
- Qualidade de alinhamento *single-end*

- Qualidade de alinhamento *paired-end*
- Tag de distância de edição, que registra a distância de *Levenshtein* entre a leitura e a referência.
- Marca de nome do *amplicon*, que registra o ID do *amplicon* associado à leitura.

2.3.1.3 VCF

O *Variant Call Format (VCF)* é um formato de arquivo de texto utilizado para armazenar variações de sequência gênica. Um arquivo .vcf contém linhas de meta-informação, uma linha de cabeçalho e linhas contendo dados, onde cada linha contém informação sobre uma posição no genoma (Ver figura 9).

Figura 9 – Exemplo de arquivo VCF

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

Fonte: (THE. . . , 2022)

O formato possui uma documentação extensa e complexa. Uma descrição mais detalhada sobre esse tipo de arquivo pode ser encontrada em <https://samtools.github.io/hts-specs/VCFv4.2.pdf>

2.4 Técnicas consolidadas e trabalhos correlatos

As análises de genotipagem atualmente existentes se baseiam em amplificação por *PCR* para detectar variantes únicas de nucleotídeos (*SNPs*) seja diretamente (*primers* específicos de alelos) ou por detecção secundária (hibridação de sonda ou Método de Sanger) (LANE, 2021). Enquanto a técnica baseada em *PCR* pode ser utilizada para genotipar a maioria dos antígenos, os formatos atuais de análise limitam o número de *SNPs* e amostras que podem ser testadas de forma simultânea. Além disso, análises baseadas em *PCR* detectam variações estruturais (*SVs*) com *PCR primers* abrangendo a região de ponto de interrupção, o que requer que esses pontos de interrupção estejam bem caracterizados, e muitos não estão (LANE, 2021). Isso significa que as atuais análises de genotipagem,

baseadas em *PCR*, não conseguem caracterizar todos os antígenos geneticamente conhecidos. Ademais, as atuais análises de genotipagem baseadas em *PCR* muitas vezes representam um custo elevado e investimento de tempo por amostra (LANE, 2021).

Em Identificação das variantes dos genes RHD e RHCE em pacientes com doença falciforme usando estratégia de sequenciamento de nova geração (RIBEIRO, 2020), por exemplo, concluiu-se que o *NGS* foi mais assertivo na interpretação de polimorfismos inerentes à alelos híbridos do que a metodologia de *Sanger*. Nesse estudo, foi demonstrado que a interpretação sorológica de polimorfismos Rh inesperados foi imprecisa em 62% dos casos, representando um potencial fator de risco para pacientes que necessitam de transfusão.

Em *Blood Group Testing* (LI; GUO, 2022) trazem os diferentes métodos de testagem de grupos sanguíneos existentes, fazendo comparações entre eles. Ademais, o estudo conclui que o desenvolvimento e validação de métodos de identificação de grupos sanguíneos que sejam rápidos, de baixo custo, simples e que sejam possíveis de serem realizados no ponto de atendimento irão beneficiar a testagem de compatibilidade pré-transfusão, bem como determinar rapidamente o tipo sanguíneo em cenários emergenciais ou em áreas remotas onde o acesso rápido à laboratórios não é possível.

Em *Sequence-Based Typing of Human Blood Groups* (SELTSAM; DOESCHER, 2009) destaca-se que sequenciamento de ácido nucleico provê informação detalhada da sequência genética através da determinação da ordem dos 4 diferentes nucleotídeos em uma molécula de DNA e que a vantagem deste médodo comparado à outras abordagens de genotipagem *DNA-based* é que é possível detectar não somente polimorfismos conhecidos, como também novas mutações.

Em *Molecular typing of blood group genes in diagnostics* (CASTILHO, 2021) traz informações de estudo com 48 pacientes brasileiros portadores de anemia falciforme e genotipados serologicamente como D fraco. Esse estudo mostrou que 40 deles possuíam genes RHD codificando D parcial e 4 pacientes haviam desenvolvido anti-D, demonstrando a importância de diferenciar D fraco e D parcial em pacientes que precisam de transfusões recorrentes, a fim de estabelecer uma política de recomendações para transfusão.

3 AMBIENTE DE DESENVOLVIMENTO

Ao preparar o ambiente de desenvolvimento, foi escolhida a ferramenta *Docker*, por sua versatilidade, conveniência e isolamento. Através do *Docker* é possível escolher, entre inúmeras imagens pré-existentes, a que melhor se adequa às necessidades de desenvolvimento. É também possível criar sua própria imagem personalizada (DOCKER, 2023). No caso deste trabalho, a opção escolhida foi criar um contêiner à partir da imagem 'continuumio/anaconda3', disponível no *Docker Hub*. O contêiner gerado à partir dessa imagem roda o sistema operacional Debian GNU/Linux 11 e já possui a plataforma Anaconda 3 instalada. Essa plataforma possui inúmeras ferramentas para as áreas de Bioinformática e Ciência de Dados, o que motivou a escolha da imagem mencionada acima. A IDE escolhida foi o *Visual Studio Code*, por possuir uma quantidade enorme de extensões, que dão suporte e agilidade ao desenvolvimento; e por sua fácil integração com o *Docker*.

No que tange o processamento dos dados brutos de *NGS*, é necessário instalar algumas ferramentas extras, sendo elas: bedtools, bcftools, minimap2, samtools, pilon, bwa, trimmomatic, mafft, selenium (com chromedriver), SPAdes, hashable, pysam e pyvcf. A linguagem utilizada neste projeto foi Python na versão 3.11.

3.1 Ferramentas utilizadas

A solução compreende um *pipeline* para *trimming*, montagem e alinhamento. As ferramentas utilizadas nesse *pipeline* estão descritas abaixo, na exata ordem em que são utilizadas.

- TrimmomaticPE -threads 8
  $NAMEFILE"_L001_R1_001.fastq.gz"
  $NAMEFILE"_L001_R2_001.fastq.gz"
  $NAMEFILE"_L001_R1_trimmed_001.fastq.gz"
  $NAMEFILE"_L001_R1_unpaired_001.fastq.gz"
  $NAMEFILE"_L001_R2_trimmed_001.fastq.gz"
  $NAMEFILE"_L001_R2_unpaired_001.fastq.gz"
  ILLUMINACLIP:adapters.fasta:2:30:10:2:true HEADCROP:15
  SLIDINGWINDOW:6:15 MINLEN:50

  – TrimmomaticPE: é uma ferramenta para remoção de adaptadores e qualidade de corte em dados de sequenciamento *paired-end*.

  – threads 8: especifica que o processo pode utilizar 8 *threads* para realizar o processamento.

  - $NAMEFILES: especificam os arquivos de entrada e saída.
  - ILLUMINACLIP:adapters.fasta:2:30:10:2:true: especifica o caminho do arquivo de adaptadores, o número máximo de *mismatches*, valores de *score* e outros.
  - HEADCROP:15: remove os primeiros 15 nucleotídeos de cada leitura.
  - SLIDINGWINDOW:6:15: realiza uma média de qualidade deslizante com janela de tamanho 6 e exigência de qualidade 15.
  - MINLEN:50: define o comprimento mínimo aceitável para as leituras de saída.

- spades.py -1 $NAMEFILE“_L001_R1_trimmed_001.fastq.gz”
  -2 $NAMEFILE“_L001_R2_trimmed_001.fastq.gz”
  -t 8 –only-assembler -k 21,33,55,77 -o assembly

  - spades.py: é uma ferramenta que realiza montagem de genoma à partir de sequências de DNA.
  - -1 $NAMEFILE e -2 $NAMEFILE: especificam os arquivos de leituras *paired-end*.
  - -t 8: número de *threads*.
  - –only-assembler: realiza apenas a montagem, sem etapas de correção.
  - -k 21,33,55,77: especifica tamanhos de *k-mer* para a montagem.
  - -o assembly: define o diretório de saída para os resultados da montagem

- minimap2 -x sr –frag=yes –secondary=yes -N 5 -p 0.8 -a refseq.fasta assembly/scaffolds.fasta | samtools view -bS -F 4 - | samtools sort -o $NAMEFILE“_sorted.bam"

  - minimap2: alinha as sequências da montagem resultante do comando anterior com uma referência genômica (refseq.fasta).
  - -x sr: especifica o modo de alinhamento sensível para sequências pareadas.
  - –frag=yes: considera fragmentos desordenados.
  - –secondary=yes: mantém informações sobre alinhamentos secundários.
  - -N 5: especifica a quantidade máxima de alinhamentos múltiplos
  - -p 0.8: define a pontuação mínima para relatar um alinhamento.
  - -a refseq.fasta: usa refseq.fasta como sequência referência.
  - | samtools view -bS -F 4 -: redireciona a saída do minimap2 para samtools view para conversão de formato e filtragem.
  - samtools sort: classifica os resultados e gera um arquivo BAM.

- samtools index: gera um índice para o arquivo BAM gerado no passo anterior.

- SEQ_NAME, REF_NAME e LENGTH: extrai informações do arquivo refseq.fasta e cria um arquivo my.genome.
- bedtools bamtobed -i $NAMEFILE“_sorted.bam» reads.bed: converte o arquivo BAM em um formato BED chamado reads.bed.
- bedtools genomecov -bga -i reads.bed -g my.genome | awk ‘$4 < 1’ > zero.bed: calcula a cobertura do genoma à partir do arquivo BED e filtra as regiões com cobertura inferior a 1, salvando o resultado no arquivo zero.bed.
- maskFastaFromBed -fi refseq.fasta -bed zero.bed -fo masked.fasta: usa o arquivo zero.bed para mascarar regiões do arquivo refseq.fasta, gerando um arquivo de referência mascarado chamado masked.fasta.
- bcftools mpileup -Ou -f masked.fasta $NAMEFILE“_sorted.bam” | bcftools call –ploidy 1 -mv -Oz -o test.vcf.gz: realiza o mapeamento de SNPs à partir das leituras alinhadas e gera um arquivo VCF chamado test.vcf.gz.
- bcftools index test.vcf.gz: cria um índice para o arquivo VCF gerado.
- cat masked.fasta | bcftools consensus test.vcf.gz > new_consensus.fasta: usa o arquivo VCF para criar uma sequência consenso à partir do arquivo masked.fasta, gerando um novo arquivo chamado new_consensus.fasta.
- echo “>$SEQ_NAME” > $NAMEFILE“_draft.fasta”: cria um arquivo NAMEFILE”_draft.fasta” com um cabeçalho que contém o nome da sequência (extraído de SEQ_NAME).
- tail -n +2 new_consensus.fasta » $NAMEFILE“_draft.fasta”: anexa o conteúdo de new_consensus.fasta ao arquivo NAMEFILE“_draft.fasta”.
- sed ‘s/>“‘$SEQ_NAME”’/>“‘$SEQ_NAME”’/g’ $NAMEFILE“_draft.fasta” > $NAMEFILE“_consensus.fasta”: substitui o cabeçalho da sequência no arquivo NAMEFILE“_draft.fasta” pelo valor de SEQ_NAME, gerando o arquivo final NAMEFILE“_consensus.fasta”.

4 RESULTADOS

4.1 Pipeline da solução

A solução desenvolvida para a caracterização de marcadores genéticos de grupos sanguíneos, a partir de dados brutos de sequenciamento genético em plataformas de *NGS*, pode ser representada em um diagrama de atividades, conforme ilustrado na Figura 10 para facilitar a compreensão.

O *pipeline* proposto compreende as seguintes etapas:

1. Coleta de Dados do Usuário: Envolve a obtenção de amostras, sequências de referência e tabelas de mutações fornecidas pelo usuário, para cada gene estudado.
2. Alinhamento *Paired-End*: Utilizando a ferramenta Trimmomatic em conjunto com outras ferramentas, realiza-se a etapa de *trimming*, montagem e alinhamento das amostras com a sequência de referência.
3. Alinhamento das Amostras com a Ferramenta MAFFT.
4. Identificação de *SNPs* nas Amostras: Os arquivos .BAM e .VCF, gerados nas etapas 2 e 3, são carregados. Neste passo, o algoritmo itera em cada registro do arquivo .VCF, verificando cada base nitrogenada em relação à referência. Se uma base nitrogenada difere da referência, um *SNP* é identificado e registrado em uma tabela dedicada.
5. Comparação dos *SNPs* com a Tabela de Mutações: A lista de *SNPs* é analisada individualmente para encontrar correspondências com a tabela de mutações. Este processo pode resultar na identificação ou não de mutações.
6. Geração de Tabela com *SNPs* Comuns: Uma tabela é gerada contendo apenas os *SNPs* presentes tanto no arquivo de amostra quanto na tabela de mutações, caracterizando assim a amostra.

Na elaboração desta solução, buscou-se simplificar o processo de caracterização ao máximo, através do estudo e da interpretação dos *outputs* gerados nos passos 2 e 3.

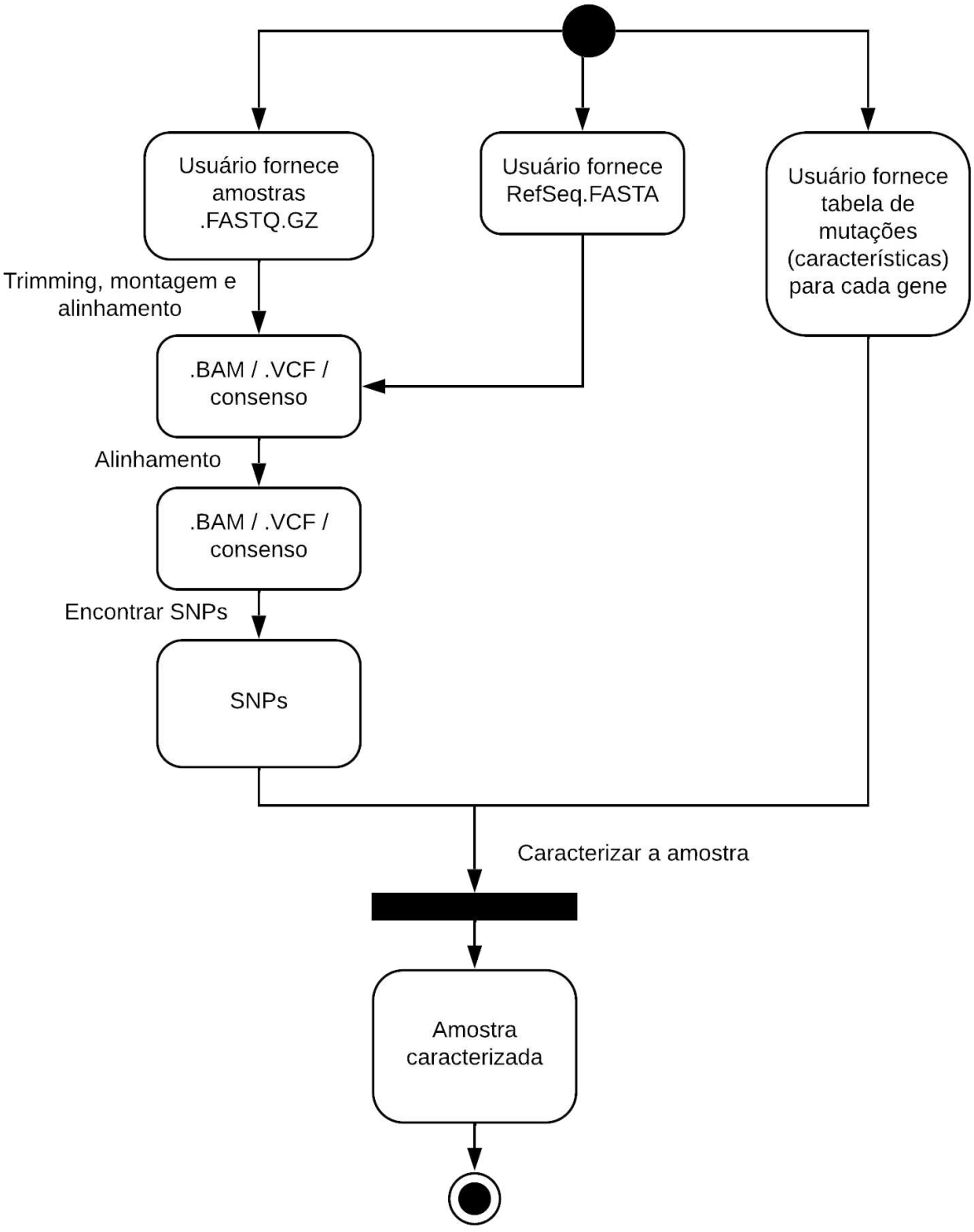


Figura 10 – Pipeline implementado

4.2 Tabelas de mutações e análise das amostras

As sequências referência utilizadas neste estudo foram extraídas do site do *National Center for Biotechnology Information (NCBI)*, que pode ser acessado através da url <https://www.ncbi.nlm.nih.gov>. O *accession number* da sequência referência do gene RHD é NG_007494.1 e o da sequência referência do gene RHCE é NG_009208.3.

As amostras fornecidas foram obtidas de pacientes com anemia falciforme atendidos no Hemocentro de Ribeirão Preto. Foram utilizadas amostras de sangue de pacientes falciformes que apresentaram anticorpos anti-RhD ou anti-RhCE. As amostras foram processadas para extração de DNA e em seguida o DNA foi utilizado para amplificação de regiões específicas do genes RHD e RHCE e à partir dos *amplicons* foram preparadas as bibliotecas de NGS. As amostras foram colhidas após assinatura do termo de consentimento.

Na etapa de caracterização das amostras foram utilizadas 2 tabelas de mutações, 1 contendo 48 mutações do gene RHD (ver tabela 1) e outra contendo 1 mutação do gene RHCE (ver tabela 2). Esta solução possui limitações, pois as tabelas não contém todas as mutações RHD e RHCE descritas.

Os resultados antecipados são promissores: foram encontrados 130 *SNPs* em uma das amostras RHD (AF2RHD_S55_L001) e 152 em outra amostra (AF6RHD_S56_L001), totalizando 2 amostras analisadas. Entre as 2 amostras RHCE analisadas, foram encontrados 233 (S7RHCE_S58_L001) e 167 *SNPs* (S20RHCE_S59_L001), respectivamente, evidenciando riqueza genética subjacente. Esses *SNPs* foram então comparados à tabela de mutações do gene da amostra em questão (passo 5 do *pipeline*). Na amostra AF6RHD_S56_L001 foram localizadas três mutações (tabela 3), sendo uma delas no exon 3 e as outras duas no exon 10. Nas outras três amostras não foram encontradas alterações genéticas nos exons.

Ao concluir essa etapa, o pipeline proposto mostrou-se, então, capaz de caracterizar as amostras analisadas. O projeto desenvolvido neste trabalho está disponível em um repositório no GitHub e pode ser acessado através do link <https://github.com/Jehcky/TCC>, ficando à disposição do Grupo de Pesquisa em Bioinformática e Biologia Computacional da Universidade do Estado da Bahia (G2BC) para estudos futuros.

Mutações RHD (posição em NG_007494.1)			
Posição	Ref	Alt	Exon
5066	C	G	1
5106	G	C	1
5110	C	?	1
17096	T	C	2
17113	A	C	2
17121	G	T	2
17136	G	A	2
17144	G	?	2
17236	T	A	2
17264	T	C	2
23226	C	T	3
23262	C	A	3
23271	A	C	3
33470	T	C	4
33495	C	?	4
33505	T	A	4
33564	A	T	4
33572	C	G	4
33579	G	A	4
34050	C	G	5
34063	T	G	5
34085	G	T	5
34093	G	A	5
34114	A	?	5
34140	C	T	5
35840	T	G	6
35850	G	A	6
35864	G	A	6
35866	G	A	6
35876	G	A	6
35903	C	G	6
39124	G	A	7
39165	G	A	7
39192	T	C	7
39230	G	?	7
39232	C	?	7
49573	C	T	8
49586	G	C	8
54400	G	C	9
61528	A	G	10
62300	A	G	10

Tabela 1 – Tabela de mutações RHD

Mutações RHCE (posição em NG_009208.3)			
Posição	Ref	Alt	Exon
14407	G	C	1

Tabela 2 – Tabela de mutações RHCE

Posição	Mutação	Exon
23271	A>C	3
61528	A>G	10
62300	A>G	10

Tabela 3 – Mutações encontradas na amostra AF6RHD_S56_L001 (posição em NG_007494.1)

5 CONSIDERAÇÕES FINAIS

À partir da detecção ou não de mutações nos exons, é possível caracterizar e classificar os grupos sanguíneos em grupos já descritos na literatura ou encontrar novos polimorfismos. Este projeto alcançou resultados promissores, pois o pipeline proposto e implementado mostrou-se capaz de caracterizar as amostras analisadas. Atingir o objetivo de caracterizar essas amostras utilizando *NGS* representa um avanço significativo na direção de uma identificação mais precisa e de uma compreensão mais profunda da expressão genética nos grupos sanguíneos. Além disso, esses resultados abrem caminho para futuras investigações, ressaltando a importância de uma abordagem médica mais individualizada e possibilitando descobertas que podem transformar a saúde pública.

Em resumo, este trabalho oferece uma base para estudos futuros em soluções aplicadas à bioinformática para identificação de marcadores genéticos e grupos sanguíneos, como por exemplo melhorias na ferramenta (*back-end*), desenvolvimento de um *front-end*, identificação das amostras, bem como a extensão do estudo à outros genes. Espera-se que este estudo contribua de alguma forma para melhorar a compreensão sobre esses grupos.

REFERÊNCIAS

BORGES-OSóRIO, M. R.; ROBINSON, W. M. **Genética Humana - 3a Edição**. [S.l.]: ARTMED, 2013.

CALDART, E. T. et al. Análise filogenética: conceitos básicos e suas utilizações como ferramenta para virologia e epidemiologia molecular. **Acta Scientiae Veterinariae**, 2016.

CASSIANO, B. **Brasil consegue ampliar transfusões de sangue, mas coleta diminui**. 2020. <https://www.gov.br/saude/pt-br/assuntos/noticias/2020/junho/brasil-consegue-ampliar-transfusoes-de-sangue-mas-coleta-diminui>. Accessado em: 06/12/2022.

CASTILHO, L. Molecular typing of blood group genes in diagnostics. **Annals of Blood**, 2021.

DOCKER. **Get Started Guide**. [S.l.], 2023. Accessado em: 21/11/2023.

FUKUMORI, Y. et al. Genotyping of abo blood groups by pcr and rflp analysis of 5 nucleotide positions. **International Journal of Legal Medicine**, 1995.

FÜRST, D. et al. Next-generation sequencing technologies in blood group typing. **Transfusion Medicine and Hemotherapy**, Karger Publishers, v. 47, n. 1, p. 4–13, 2020.

ILLUMINA. **Common File Formats Used by the ENCODE Consortium**. 2021. <https://support.illumina.com/bulletins/2016/04/fastq-files-explained.html>. Acessado em: 06/12/2022.

ILLUMINA. **BAM File Format**. 2022. <https://support.illumina.com/help/BS_App_RNASeq_Alignment_OLH_1000000006112/Content/Source/Informatics/BAM-Format.htm>. Acessado em: 07/12/2022.

JADHAO, S. et al. Using whole genome sequencing to characterize clinically significant blood groups among healthy older australians. **Blood advances**, Elsevier, 2022.

LANE, W. J. Recent advances in blood group genotyping. **Ann Blood**, v. 6, p. 31, 2021.

LESK, A. M. **Introdução à Bioinformática**. [S.l.]: ARTMED, 2008.

LI, H.-Y.; GUO, K. Blood group testing. **Frontiers in Medicine**, 2022.

MERCHáN, M. A.; CAICEDO, M. I. T.; TORRES, A. K. D. **Técnicas de Biología Molecular en el desarrollo de la investigación. Revisión de la literatura.** 2017. <http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1729-519X2017000500012&nrm=iso>. Acessado em: 11/12/2022.

NARDOZZA, L. M. M. et al. Bases moleculares do sistema rh e suas aplicações em obstetrícia e medicina transfusional. Rev Assoc Med Bras, 2010.

ORZIńSKA, A. et al. A preliminary evaluation of next-generation sequencing as a screening tool for targeted genotyping of erythrocyte and platelet antigens in blood donors. **Blood Transfusion**, SIMTI Servizi, v. 16, n. 3, p. 285, 2018.

PIMENTEL, M.; FILIPPO, D.; SANTORO, F. M. **Metodologia de Pesquisa em Informática na Educação: Concepção da Pesquisa**. [S.l.]: SBC, 2019.

POLIN, H. et al. Introduction of a real-time based blood group genotyping approach. **Vox Sanguinis**, 2008.

RIBEIRO, I. H. **Identificação das variantes dos genes RHD e RHCE em pacientes com doença falciforme usando estratégia de sequenciamento de nova geração**. 2020.

RODRIGUES, E. S. et al. T cell receptor signaling pathway is overexpressed in cd4+ t cells from ham/tsp individuals. **The Brazilian Journal of Infectious Diseases**, 2015.

RODRIGUES, E. S. et al. Frequency and characterization of rhd variant alleles in a population of blood donors from southeastern brazil: Comparison with other populations. **Transfusion and Apheresis Science**, 2021.

SELTSAM, A.; DOESCHER, A. Sequence-based typing of human blood groups. **Transfusion Medicine and Hemotherapy**, 2009.

TAX, M. G. Rh variability in multi-ethnic perspective. Thieme Mediacenter Rotterdam, 2005.

THE Variant Call Format (VCF) Version 4.2 Specification. 2022. <https://samtools.github.io/hts-specs/VCFv4.2.pdf>. Accessado em: 07/12/2022.

TURCHETTO-ZOLET, A. C. et al. **Marcadores Moleculares na Era genômica: Metodologias e Aplicações**. [S.l.]: Sociedade Brasileira de Genética, 2017.

WAGNER, F. F.; FLEGEL, W. A. Review: the molecular basis of the rh blood group phenotypes. Immunohematology, 2004.