UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

Avaliação *in silico* de sítios de inserção em *Genomic Safe Harbors* para uso do sistema CRISPR/Cas9 em diferentes populações humanas

Paola Carneiro

Dissertação submetida ao Programa de Pós-Graduação em Genética e Biologia Molecular da UFRGS como requisito parcial para obtenção do título de **Mestre em Genética e Biologia Molecular** .

Orientador(a): Ursula Matte

Porto Alegre, Julho, 2022

Instituição e suporte financeiro

Este trabalho foi desenvolvido no Laboratório de Células, Tecidos e Genes do Centro de Pesquisa Experimental (CTG-CPE) e no Núcleo de Bioinformática do Hospital de Clínicas de Porto Alegre (HCPA), financiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). Aspectos metodológicos e dados utilizados nesta dissertação compreendem apenas análise de dados públicos proveniente de bancos de dados, não necessitando de aprovação pelo Comitê de Ética e Pesquisa do Hospital de Clínicas de Porto Alegre.

Agradecimentos

A professora Dra. Ursula Matte por me incentivar a aprender, pelas contribuições nesta dissertação e para meu desenvolvimento profissional, por ter me feito as perguntas que precisava para conduzir pesquisa e ter me ensinado a responder e, principalmente, a também perguntar.

A todos os colegas do Laboratório de Células, Tecidos e Genes (CTG) pelo companheirismo, pelo conhecimento trocado e pelas boas risadas.

A Ágnis, Cristal, Felipe, Hallana, Marina, Martiela e Pâmella pelos conselhos, conversas científicas e vários "SERÁ?" ou "Não sei,vamos pesquisar", mas principalmente por terem feito essa jornada acadêmica e a vida dentro e fora do laboratório mais divertida.

Aos meus pais pelo incentivo e apoio na minha educação, pelos ensinamentos na vida pessoal e por me darem tudo que eu precisava durante este mestrado e para seguir os meus sonhos.

Ao meu namorado por me apoiar e compartilhar alguns dos meus sonhos como também pelos abraços necessários.

A mim por ter sido resiliente em momentos não tão fáceis.

A UFRGS, PPGBM e as agências de fomento o Fundo de Incentivo à Pesquisa e Eventos (FIPE) do Hospital de Clínicas de Porto Alegre e o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) por darem suporte neste estudo e incentivaram a pesquisa brasileira.

"Curiouser and curiouser!"

***Alice's Adventures in Wonderland*- Lewis Carroll**

SUMÁRIO

RESUMO

A inserção de transgenes com expressão não transiente e com minimização de efeitos adversos na funcionalidade celular do organismo tem grande importância no âmbito clínico, fornecendo recurso terapêutico aos pacientes com tratamento de eficácia limitada ou tratamentos que dependem de doadores compatíveis. Nesse cenário, a inserção de produtos terapêuticos dentro de *Genomic Safe Harbors* (GSH) usando o sistema CRISPR/Cas9 promove uma alternativa terapêutica para aplicação a distintas doenças genéticas, prevendo antecipadamente a interação do transgene no genoma humano. Contudo, avaliações sobre a arquitetura tridimensional do genoma e efeitos *off-target* considerando a variabilidade genética levantam o questionamento da segurança do tratamento em diferentes tipos celulares e populações humanas. Portanto, o objetivo deste trabalho foi avaliar a organização tridimensional cromossômica em três tipos celulares como também efeitos *off-targets* a partir da probabilidade de clivagem dessas regiões para sgRNAs desenhados em GSH usando ferramentas desenvolvidas para acesso aos dados de variantes genéticas contidos em diferentes bancos. A investigação de sítios de inserção em 4 *loci* GSH (*AAVS1, H11, SH6* e *THUMPD3-AS1*) demonstrou que *THUMPD3-AS1* está dentro de um potencial *Topological Associate Domains* (TAD) conservado ao longo dos diferentes tipos celulares. Entretanto, TADs foram apenas estabelecidos para demais *loci* para tipos celulares específicos devido à baixa resolução de dados. Ainda, a avaliação de *off-target* questiona a segurança da aplicação do mesmo sgRNA ao longo de diferentes populações humanas em GSH devido a *off-target* privados e compartilhados adjacentes a existentes ou criadas PAM canônicas. Dentre todas as populações, a população Africana (n = 256) seguindo da população Brasileira (n = 220) demonstraram o maior número de sítios polimórficos que criam PAMs canônicas considerando todos os sgRNAs. Com intuito de minimização de efeitos *off-target*, o sgRNA5 para *AAVS1* com menor número dessas regiões demonstrou que sítios polimórficos diminuem a probabilidade de clivagem, demonstrando menor efeitos *off-target* para população Latina que apresenta o maior número de sítios polimórficos (n = 302). Finalmente, ferramentas para acesso de dados variantes genéticas foram desenvolvidas para auxiliar futuras investigações sobre a variabilidade genética das populações nos bancos gnomAD e ABraOM. Espera-se que a análise conduzida para sítios de inserção em GSH ajudem na aplicação de tratamentos terapêuticos na clínica para doenças genéticas distintas em diferentes populações.

ABSTRACT

Integrating a gene with no transient expression and minimization of adverse effects on cell functionality is important in the clinical field, providing treatment to individuals with limited treatment effects or that require a compatible donor. In this scenario, inserting therapeutic gene products inside Genomic Safe Harbors (GSH) using the CRISPR/Cas9 system offers a relatively less-cost therapeutic alternative for different human genetic diseases with previous knowledge of transgene into the human genome. However, the three-dimensional genome architecture and possible off-target effects regarding the genetic variability raise safety concerns in the application of treatment in different cell types and human populations. Thus, this investigation aims to provide a-priori in silico analysis of three-dimensional chromosome structure among three cell types and off-target effects considering the probability cleavage for sgRNA designed into GSH by using the development of bioinformatics tools to access data from gnomAD and ABraOM database. The gene insertion investigation into 4 GSH (AAVS1, H11, SH6 and THUMPD3-AS1) showed THUMPD3-AS1 is inside a potential conserved TAD across different cell types. For other loci, it only observed TADs for specific cell types due to poor data quality. In another scenario, the off-target prediction raises concerns about applying the same sgRNA designed into GSH among human populations due to private and shared off-target region investigation among existing and creating canonical PAMs. Among all populations, African (n =256) followed by Brazilian (n = 220) populations demonstrated the majority number of polymorphic sites that created PAM among human populations when considering all sgRNA. To minimize off-target effects among human populations, the sgRNA5 from AAVS1, which showed the smallest number of off-target regions using the reference human genome, demonstrated that probability cleavage decreases considering the majority of polymorphic sites in the off-target region among human populations. It also revealed that Latino could benefit from this sgRNA since this population has the most significant number of polymorphic sites (n =302). Finally, bioinformatics tools were developed to access and facilitate vast human genetic variant data, thus aiding future research concerning genetic variability investigation on gnomAD and ABraOM databases. This assessment a-priori analysis is expected to assist therapeutic treatment applications in clinics for different human genetic diseases among human populations.

1 INTRODUÇÃO GERAL

1.1 Terapia gênica e adição de material genético

A terapia gênica é o mecanismo de modificação genética em células de um organismo visando efeito terapêutico para o tratamento, cura ou prevenção de doenças humanas (Kaufmann e*t al.*, 2013). Em 1972, o potencial da inserção de DNA dentro no genoma humano foi discutido quanto aos principais desafios e projeções no tratamento das doenças genéticas (Friedmann *et al.*, 1972). Atualmente, a terapia gênica é destacada como uma promissora alternativa terapêutica na qual novas estratégias, usando recentes tecnologias de edição, vêm sendo impulsionadas e aprovadas (Walters *et al.*, 2022) conjuntamente com outros produtos terapêuticos para aplicação em âmbito clínico.

Na terapia gênica, a inserção de gene no genoma como produtos terapêutico para entregar uma cópia funcional para restaurar a função perdida de um transcrito endógeno, vem apresentando relevância na medicina. Nesse cenário, a terapia gênica é uma opção atraente e com resultados encorajadores para o tratamento de pessoas com doenças genéticas como as do sistema nervoso central (Cartier *et a*l., 2009; Kaplitt *et al*., 2007) e como tratamento para doenças imunológicas que apresentam alternativas terapêuticas com eficácia limitada devido a efeitos colaterais(Kohn *et al*. 2021). De fato, em 1990, a primeira demonstração promissora dessa tecnologia no tratamento de pacientes com imunodeficiência severa combinada (ADA-SCID), causada pela perda funcional da enzima adenosina desaminase (ADA, EC 3.5.4.4) (Blaese *et al*,1995), mostrou o potencial da terapia gênica no tratamento clínico. Contudo, apesar do sucesso no tratamento de alguns pacientes (Hacein-Bey-Abina *et a*l., 2002), o método terapêutico apresentou efeito limitado e indesejado em outros como, por exemplo, expressão transiente ou mutagênese insercional com ativação de proto-oncogenes (Hacein-Bey-Abina *et al*., 2003; Hacein-Bey-Abina *et al*., 2008). Com isso ficou evidente a relevância do desenvolvimento de terapias que avaliassem a segurança e eficiência da integração do transgene no genoma.

Com a finalidade da entrega de transgenes, algumas abordagens terapêuticas compreendem expressão episssomal usando vetores virais representantes da família do Adenovírus ou Vírus Adeno-associado (Bulcha *et al*., 2021) e vetores não-virais como plasmídeos (Mulia *et al*.,2021), integração mediada por transposase (Tipanee *et a*l., 2017; Sandoval-Villegas *et al*.,2021) e inserção de genes funcionais dentro de regiões seguras do genoma. Nesse cenário, embora a expressão de genes sem inserção no genoma seja atrativa

no que compreende evitar mutagênese e aumento da capacidade do inserto promovidas por modificações em vetores virais, a rejeição imune-mediada por vetores virais na terapia gênica apresenta substanciais limitações e desafios para o vasto uso no tratamento de pacientes (Shirley *et al*. 2020). Ainda, a perda de episssomas não-virais plasmidiais em divisões celulares, bem como efeitos de silenciamento gênico decorrentes do aumento dos níveis de dinucleotídeos CpG no vetor e no inserto já foram descritas (Mulia *et al*.,2020; Bruter *et al*., 2018). Adicionalmente, avaliações experimentais demonstraram complicações na utilização de terapias genéticas baseadas em transposons, com a desregulação de genes relacionados a câncer e perfis de integração quase aleatório ou próximo a sítios de início de transcrição (Huang *et al*. 2010). Nesse sentido, regiões no genoma previamente avaliadas experimentalmente quanto a aspectos de segurança e eficiência na inserção do transgene, se tornam atrativas Ainda, essas regiões, denominadas *Genomic Safe Harbors* (GSH), possibilitam a ampliação de tratamento a doenças distintas ocasionadas pela perda da expressão gênica decorrente de uma ou mais variantes genéticas patogênicas.

1.2 *Genomic Safe Harbors*

GSH são sítios cromossômicos, localizados dentro ou fora de unidades transcricionais do genoma, usados para inserção de transgenes que apresentam expressão previsível e estabilidade após integração sem desencadear efeitos adversos na função celular (Sadelain *et al*.,2012). A integração de cópias funcionais em GSH permite apropriada interação do gene exógeno com o genoma hospedeiro, pois possibilita a expressão de transgenes sem desregulação ou perda funcional de produtos gênicos endógenos (Papetrou *et al.*, 2016). Conjuntamente, a utilização desses sítios na terapia gênica permite a inserção de diferentes transgenes, conferindo uma alternativa de tratamento a doenças distintas.

Investigações experimentais descreveram os critérios desejados para a inserção de transgenes nas regiões genômicas seguras. A partir da avaliação mediada por lentivirus, cinco critérios foram sugeridos para os sítios de inserção como: distância mínima de 50 kb da extremidade 5’ de qualquer gene, 300 kb de oncogenes, 300 kb de microRNAs (miRNA) e localização fora de unidades transcricionais bem como de regiões ultra-conservadas (UCR) do genoma humano (Papapetrou *et al*., 2011). Ainda, o desenvolvimento científico sobre a constituição do genoma permitiu que outros trabalhos propusessem a adição de aspectos moleculares aos GSH como: não localização dentro de regiões com variação de número de

cópias, inseridas dentro de um contexto de eucromatina e uma única cópia no genoma (Pellenz *et al*., 2019). Entretanto a consideração de uma interação não linear entre componentes genéticos é uma importante perspectiva na avaliação da determinação de sítios de inserção, posto que a regulação gênica de transcritos endógenos está intimamente relacionada com a organização estrutural do genoma nuclear. A identificação de subcompartimentos cromossômicos que delimitam interações entre elementos genéticos, denominados domínios topológicos (*Topological Associated Domains*, TADs) foram reconhecidos como sendo compreendidos por sítios de ligação de proteínas CTCF (Dixon *et al*., 2012). Na investigação, 15% de todos os sítios de ligação de CTCF identificados foram localizados em TADs. Ainda, a co-ocupação de sítios de ligação de coesina promoveu a identificação de outro marcador da organização tridimensional do genoma (Tang *et al., 2015),* e forneceu maior detalhamento sobre a organização tridimensional do genoma nuclear. Figura 1 demonstra uma representação da estrutura tridimensional cromossômica no nível de TADs no qual níveis de organização hierárquica são demonstrados.

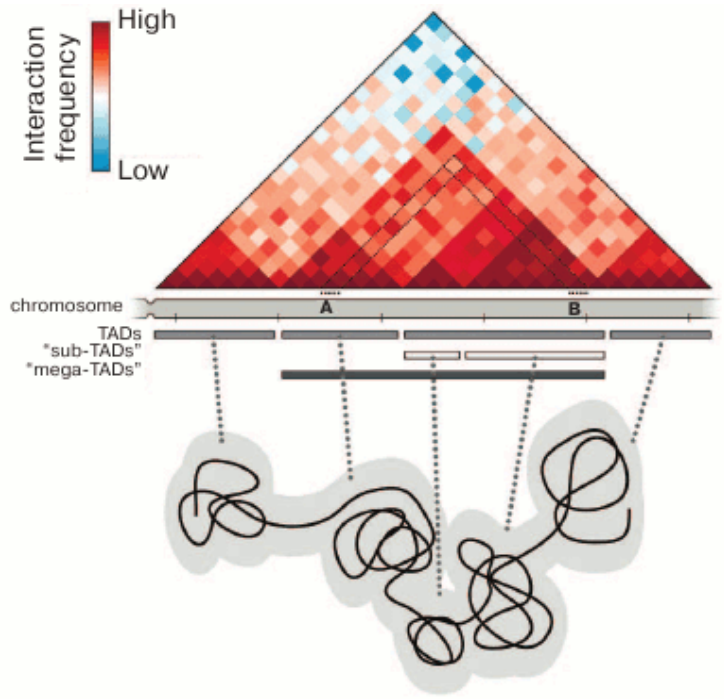

Figura 1: Representação ilustrativa demonstrando os diferentes níveis de organização cromossômica no nível de TAD. Frequência de contato é demonstrado no Hi-C seguindo de anotação de TAD para uma região hipotética. Nível da interação são demonstrados para regiões com menor (azul) para as com maior (vermelho) frequência de contato. Imagem obtida do artigo Razin *et al.* 2018.

Nesse cenário, a utilização de GSH amplifica a possibilidade de tratamento de doenças genéticas fornecendo previamente informações sobre segurança e eficiência do gene

exógeno no genoma hospedeiro. Conjuntamente, a investigação da aplicabilidade da arquitetura organizacional em GSH confere maior refinamento dos dados para a perspectiva desses *loci* no que concerne à utilização para inserção de transgenes ao longo de diferentes tipos celulares usando novas tecnologias de edição genômica.

1.3 Inserção de transgenes em *Genomic Safe Harbors* na terapia gênica

A avaliação de sítios de inserção para GSH usando novas tecnologias de edição é um campo em desenvolvimento. Nesse cenário, em adição aos critérios sugeridos para escolha dos GSH (ver seção 1.2), há regiões identificadas como GSH que não satisfazem todos pressupostos elencados, como falta de dados de análises de genotoxicidade e anormalidades no funcionamento nos tipos celulares avaliados após a integração. Dentre os sítios identificados, destaca-se o sítio de inserção do vírus adeno-associado (Kotin *et al*.,1992) que se localiza no éxon 1 e íntron 1 do gene *PPP1R12C* no cromossomo 19. Além desse sítio, o locus *hROSA26* (*THUMPD3-AS1)* , homólogo ao locus *rosa26* em murinos, amplamente empregado em avaliações de inserção, é localizado no cromossomo 3 (Irion *et a*l., 2007) no genoma humano. Além disso, a perspectiva de GSH para integração de transgenes em ambos *loci* foram apresentadas a partir de avaliação em células tronco embrionárias e ao longo da sua diferenciação em outros tipos celulares, mostrando estabilidade e contínua expressão dos genes inseridos (Ramachandra *et al*., 2011, Irion *et al*., 2007). Apesar dos sítios anteriores serem os mais utilizados nas avaliações de inserção gênica em ensaios experimentais no genoma de humanos e murinos, nesta ordem, outros sítios estão sendo reconhecidos com potencial de GSH como *H11* (Zhu *et al*., 2014, Turan *et al*., 2016, Park *et al*.,2019) e *SH6* (Eyquem *et al*., 2003; Rodriguez-Fornes *et al*., 2020) localizados no cromossomo 22 e 21, respectivamente. Similarmente aos dois primeiros GSH, ambos os *loci* apresentam possibilidade de utilização na terapia gênica. Ortólogos de *h11* foram identificados e avaliados experimentalmente em outros mamíferos para integração de transgenes, não apresentando efeitos adversos aos organismos (Ruan *et al*., 2015; Li *et al*., 2019). Ainda, a facilidade de manipulação do genoma a partir de novas tecnologias de edição impulsionam os estudos de avaliação de inserção de transgenes em GSH e em outros locais no genoma humano. A tabela 1 apresenta critérios desejáveis satisfeitos por GSH descritos anteriormente (seção 1.2), e a tabela 2 apresenta quais desses e o total de critérios por cada *locus* com a respectiva fonte da informação.

Tabela 1: Critérios desejáveis para GSH no genoma propostos por Papapetrou *et al*. e Pellenz *et al.*

ID	Critérios desejáveis para potenciais GSH
1	distantes 300 kb de oncogenes
2	diantantes 300 kb de miRNAs ou pequenos RNAs funcional
3	distantes 50 kb de região 5' de um gene
4	distantes 50 kb distante de origem de replicação
5	distantes 50 kb distante de elementos ultra-conservados no genoma
6	baixa atividade transcricional (nenhum mRNA ±25 kb)
7	não presente em regiões com variação de número de cópias
8	localizado em região de cromatina aberta
9	cópia única no genoma

Tabela 2: Critérios desejáveis a potenciais GSH satisfeitos pelos *loci* investigados nessa pesquisa. A tabela mostra ID do critério conforme tabela 1, total computado de critérios satisfeitos por *locus* e fonte da informação sobre os critérios satisfeitos. Travessões (-) representam informações não fornecidas na fonte.

Locus	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Total**	**Fonte**
AAVS1	0	1	0	1	1	0	1	1	5	Pellenz *et al.* 2019
H11	0	0	0	0	1	0	1	1	3	Dados do trabalho
SH6	1	1	1	1	1	-	-	-	5	Rodriguez-Fornes *et al.* 2020
THUMPD3-AS1	0	1	0	0	1	0	1	0	3	Dados do trabalho

1.4 Edição gênica em *Genomic Safe Harbors* usando o sistema CRISPR/Cas9

Desde o conhecimento das estruturas do DNA em 1953 (Watson *et al.*, 1953), a capacidade de realizar mudanças pontuais no genoma humano tem sido um dos alvo na medicina para o tratamento de doenças genéticas. Nos últimos anos, estratégias de edição gênica têm se tornado facilitadas a partir dos avanços em novas ferramentas e tecnologias para manipulação de genomas. O desenvolvimento dessas novas tecnologias de edição baseada em nucleases programáveis como Meganuclease, *Zinc Fingers*, TALENS e CRISPR ampliaram a capacidade de promover mudanças no genoma de diferentes espécies, incluindo células humanas (Zhang *et a*l. 2019). Com isso, promovem a perspectiva de tratamentos a doenças incuráveis. Recentemente, terapias gênicas empreendidas pelas novas tecnologias de

edição genômica em ensaios clínicos foram empregadas no tratamento de doenças como Mucopolissacaridose tipo II [NCT04628871] e Amaurose Congênita de Leber [NCT03872479] usando os sistemas *Zinc Finger* e CRISPR/Cas9, respectivamente (Ledford, 2018; Ledford, 2020) .

No cenário de novas tecnologias de edição, quando comparado às nucleases programáveis, CRISPR/Cas9 (do inglês *Clustered Regularly Interspaced Short Palindromic Repeats - CRISPR associated with Cas9 protein*) se tornou uma ferramenta de edição onipresente na manipulação de genomas. Alguns dos principais fatores que promoveram seu emprego compreendem principalmente sua acessibilidade de aplicação e relativo baixo custo como alternativa terapêutica na clínica. Isso se dá por apenas requisitar uma nuclease (Cas9) guiada mediante a complementaridade de bases entre um RNA (sgRNA) e uma sequência de DNA alvo adjacente ao motivo adjacente ao protoespaçador (PAM) NGG para *Streptococcus pyogenes*.

Entretanto, a detecção de edição de diferentes regiões concomitantes com o sítio alvo (*off-targets*) é destacada e avaliada no que compreende aspectos de segurança do sistema na sua aplicação no tratamento de pacientes (Fu *et al*. 2013; Schleidgen et al., 2020). Assim, embora a integração de transgenes em GSH mediada por CRISPR/Cas9 amplie profundamente a utilização para fins terapêuticos, aspectos de segurança do sistema de edição em GSH também precisam ser compreendidos e avaliados.

Acurácia e precisão são fatores críticos para a viabilidade utilidade da edição gênica na clínica no que a acurácia se refere a proporção da edição do sítio alvo e *off-targets*, e precisão na taxa de edição do sítio alvo para alcançar o efeito genético esperado (Doudna, 2020). Neste contexto, o desenho personalizado de sgRNA permite avaliar propriedades relacionadas a segurança e eficiência previamente à sua aplicação. Com esse intuito, avaliações *in silico* possibilitam realizar uma triagem de potenciais *off-target*, permitindo prever outros locais além do alvo que possam ser submetidos a edição. Alguns parâmetros avaliados em potenciais *off-targets* compreendem o número de incompatibilidades no pareamento de bases (*mismatches*) ou no tamanho da sequência alvo (*indels*). Apesar de regiões com menor número de incompatibilidades serem mais propensas a clivagem, avaliações experimentais conduzidas em sequências *off-target* usando GUIDE-Seq demonstraram que sequências com até 6 mismatches (Tsai *et al*., 2016) e 2 indels (Lin *et al*., 2014) são possíveis de serem clivadas. Ainda, investigações sobre a probabilidade de clivagem de *off-targets* com relação ao número e posição de incompatibilidades do tipo *mismatches* e *indels* também foram avaliadas experimentalmente, conjuntamente com o

desenvolvimento de algoritmos para computação *in silico* de sequências apenas com *mismatches* como Determinação da Frequência de Corte (CFD) (Doench *et al*., 2016). Nesse contexto, a existência de dados experimentais de avaliação da probabilidade de sequências com *indels* fomenta a criação de algoritmos que permitam avaliar a probabilidade de edição desse tipo de incompatibilidades. Adicionalmente, diferenças no que compreende variabilidade genética entre as populações possibilita o refinamento e segurança do desenvolvimento de alternativas terapêuticas a partir da inserção de transgenes em GSH usando CRISPR/Cas9.

1.5 Implicações de polimorfismos genéticos em *off-targets* para GSH

A variabilidade genética humana é um dos fatores determinantes nos desfechos de resposta terapêuticas na clínica. Investigações conduzidas sobre a variabilidade genética na aplicação de tratamentos usando CRISPR/Cas9 é um campo em desenvolvimento, demonstrando possíveis implicações da variabilidade genética na segurança de tratamentos mediados pelo sistema (Canver *et al*., 2018; Carneiro, *et al*. 2022) .

Neste cenário, a descrição abrangente de variantes genéticas comuns encontradas em *off-targets* para sgRNA otimizados para sítios de integração em GSH promove o reconhecimento de possíveis respostas em pacientes submetidos ao mesmo tratamento terapêutico com o sistema CRISPR/Cas9. Essa abordagem considera que tal resposta possa ser diferente nas diversas populações humanas. Isso, possibilita a adaptação e o melhoramento da proposta de tratamento ao contexto genético de indivíduos a partir da análise de variantes genéticas depositadas em bancos de dados. Ainda, o desenvolvimento de uma abordagem computacional que facilite a recuperação de dados nesses bancos é um recurso que permite facilitar a implementação de informações sobre polimorfismos genéticos nas investigações sobre edição gênica conduzidas na pesquisa e âmbito clínico.

1.6 Integração e facilitação na obtenção de dados sobre variantes genéticas

O crescente volume de dados genômicos obtido a partir das modernas tecnologias de sequenciamento beneficiaram o aumento de estudos no campo biológico. A aplicação de dados de variantes genéticas de populações humanas apresenta valioso recurso na clínica para alvos terapêuticos, relação causal entre doença e variantes genéticas, bem como nos estudos de ancestralidade. Esforços em armazenamento e disponibilização de acesso público aos

dados sobre variabilidade genética para populações humanas é um campo em constante desenvolvimento e vêm fomentando o conhecimento sobre biologia humana.

Projetos sobre a variabilidade genética humana como *1000 Genomes* (The 1000 Genomes Project Consortium, 2015), gnomAD (Karczewski *et al*. 2020) e ABraOM (Naslavsky *et al*. 2022), são alguns dos grandes esforços desenvolvidos para o acesso público aos dados de variantes genéticas em diferentes populações humanas. Dentre os bancos de variantes genéticas, a compreensão de dados de diferentes populações e na população brasileira presente nos bancos de dados gnomAD e ABraOM, respectivamente, demonstram vasta aplicabilidade para investigações sobre implicações da variabilidade genética na clínica e na pesquisa.

O entendimento da variabilidade genética promove o conhecimento da história, estrutura e relações das populações, como também fornece suporte em investigações genéticas na clínica (Bergström *et al.* 2020). Nesse campo, a sub-representação da diversidade genética de etnias em bancos de dados pode levar à perda de informação sobre os aspectos genéticos que compreendem determinado fenótipo como também à interpretação incorreta ou parcial de investigações e aplicações clínicas que utilizem esses dados. Recentemente, a iniciativa do projeto brasileiro BIPMed identificou que 809.589 do total de 1.626.829 variantes encontradas na população brasileira não estavam presentes no banco de dados *1000 Genomes* (Rocha *et al*. 2020). Assim, embora gnomAD apresente dados sobre variantes genéticas de diferentes populações, a variabilidade genética de algumas etnias pode também estar sendo sub-representada neste banco de dados. Nesse cenário, destaca-se ainda que em torno de 2 mil variantes presentes no banco ABraOM não foram encontradas em outros bancos de dados (Naslavsky *et al*. 2022). Tal informação apresenta fundamental relevância na clínica, por exemplo, quando consideramos efeitos adversos e eficiência em resposta a um fármaco gerado por variantes presentes em enzimas implicadas no metabolismo de xenobióticos (Sim *et al*. 2012) e susceptibilidade de pacientes a patógenos, como a avaliação *in silico* de variantes genéticas em *ACE2* e *TMPRSS2* conduzidas na a COVID19 (Hou *et al*., 2020) para a população brasileira. Ainda, o entendimento da sub-representação da diversidade genética da população brasileira torna evidente a necessidade de novas estratégias integrativas de dados de bancos brasileiros com outros projetos como também a facilitação no acesso aos dados.

Neste cenário, o desenvolvimento de um método computacional promove a facilitação e integração de dados genômicos sobre variantes genéticas aplicáveis na saúde humana. O manejo e integração de dados sobre variantes genéticas em bancos de dados díspares pode ser

alcançado a partir do desenvolvimento de abordagens de bioinformática que sejam capazes de recuperar, concatenar e sintetizar as informações depositadas em diferentes bases. Com esse propósito, a centralização de informações conjuntamente com a redução da necessidade de trabalho manual para obtê-las proporcionam ferramentas úteis tanto em âmbito clínico quanto investigativo.