

UNIVERSIDADE ESTADUAL PAULISTA
"Júlio de Mesquita Filho"
INSTITUTO DE BIOCIÊNCIAS DE BOTUCATU

AVALIAÇÃO DA PERFORMANCE DE FERRAMENTAS DE
ALINHAMENTO DE SEQUÊNCIAS CURTAS NA ANÁLISE DA
DIVERSIDADE GENÉTICA DO COMPLEXO NKG2

LAYLA NECHY RODRIGUES DOS SANTOS

PROF. DR. ERICK DA CRUZ CASTELLI

Orientador

Trabalho de conclusão de curso apresentado ao Instituto de Biociências, Campus de Botucatu, UNESP, para obtenção de Bacharel em Ciências Biomédicas.

Botucatu - SP

2023

Layla Nechy Rodrigues dos Santos

**AVALIAÇÃO DA PERFORMANCE DE FERRAMENTAS DE
ALINHAMENTO DE SEQUÊNCIAS CURTAS NA ANÁLISE
DA DIVERSIDADE GENÉTICA DO COMPLEXO NKG2**

Trabalho de conclusão de curso apresentado ao Instituto de Biociências, Campus de Botucatu, UNESP, para obtenção de Bacharel em Ciências Biomédicas.

Orientador: Prof. Dr. Erick da Cruz Castelli.

Botucatu - SP

2023

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉC. AQUIS. TRATAMENTO DA INFORM.
DIVISÃO TÉCNICA DE BIBLIOTECA E DOCUMENTAÇÃO - CÂMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: ROSANGELA APARECIDA LOBO-CRB 8/7500

Santos, Layla Nechy Rodrigues dos.

Avaliação da performance de ferramentas de alinhamento de sequências curtas na análise da diversidade genética do complexo NKG2 / Layla Nechy Rodrigues dos Santos. - Botucatu, 2023

Trabalho de conclusão de curso (bacharelado - Ciências Biomédicas) - Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências de Botucatu
Orientador: Erick da Cruz Castelli
Capes: 20202008

1. Alinhamento de sequências (Bioinformática). 2. Antígenos HLA. 3. Seqüenciamento de nucleotídeo. 4. Receptores de células.

Palavras-chave: Alinhamento; Bioinformática; HLA; NGS; NKG2.

SUMÁRIO

INTRODUÇÃO.....	04
METODOLOGIA.....	07
RESULTADOS E DISCUSSÃO.....	09
CONCLUSÃO.....	14
REFERÊNCIAS BIBLIOGRÁFICAS.....	15

RESUMO

As células *Natural Killer* (NK) são células com capacidade citolítica e fazem parte do sistema imunológico inato. A atividade das células NK é regulada por uma série de receptores de superfície, que incluem os receptores da família NKG2. Os receptores NKG2 possuem ação inibitória ou ativadora, a depender de sua conformação e de seus ligantes. Os principais ligantes dos receptores NKG2 são as moléculas do Sistema de Antígenos Leucocitários Humano (HLA) expressas na maioria das células somáticas. Os genes que codificam os receptores NKG2 são codificados em um cluster no cromossomo 12 e apresentam elevada similaridade entre si. Esta similaridade de sequência, alinhada com a provável presença de polimorfismos, pode prejudicar o alinhamento de sequências originadas em estudos de sequenciamento de segunda geração, ou NGS. Neste estudo, avaliamos a capacidade dos alinhadores tradicionais, como o BWA MEM, de alinhar corretamente sequências oriundas dos genes NKG2, por meio de simulação de sequenciamentos do tipo *paired-end* e rastreamento do local de alinhamento de cada *read* e do gene que deu origem a esta sequência. Os resultados apontam problemas de alinhamento principalmente nos genes *KLRC2/NKG2C* e *KLRC3/NKG2E* (alinhamento cruzado) e *KLRC4/NKG2F* (perda de *reads*). Portanto, para uma correta genotipagem desses genes, será necessário utilizar uma metodologia que considera todas as sequências conhecidas para cada gene (alinhamento multi-referenciado), assim como a utilizada para os genes HLA.

Palavras chave: NKG2, HLA, NGS, alinhamento, bioinformática

INTRODUÇÃO

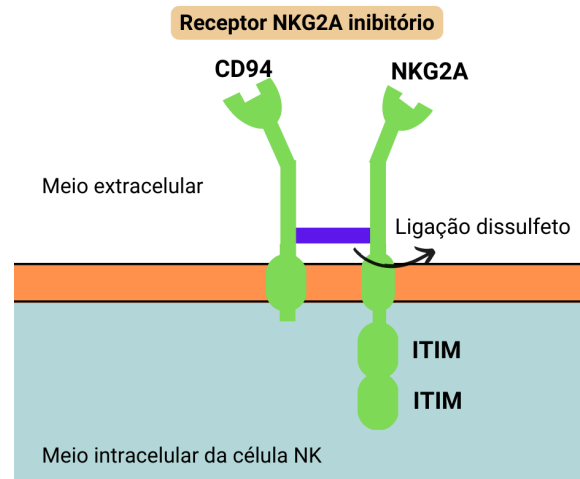
As células *Natural Killer* (NK) são células envolvidas tanto na imunidade inata quanto na adaptativa e apresentam ação citotóxica notável. Essas células são muito importantes no controle de infecções. Representam, junto com outros componentes da imunidade inata, a primeira linha de defesa do organismo. Dentro da imunidade inata, as células NK desempenham um papel chave na regulação e indução da morte celular de células tumorais e no reconhecimento e eliminação de células infectadas por microrganismos intracelulares - como vírus (MALE, 2021).

As células NK são reguladas por uma série de receptores de superfície, entre eles os receptores da família NKG2. Os genes da família de NKG2 - que vão do NKG2A ao NKG2H - são normalmente expressos em células NK desde a fase embrionária até a fase adulta e são considerados como pouco polimórficos (ASENJO et al., 2022), embora não haja estudos detalhando a diversidade desses genes em populações reais. A expressão das moléculas de NKG2 depende do ambiente celular. Esses receptores apresentam como principal ligante as moléculas do Sistema de Antígenos Leucocitários Humanos (HLA) expressas na superfície da maioria das células nucleadas. Esta interação entre a molécula HLA e os receptores NKG2, auxilia a célula NK a identificar células contaminadas com patógenos ou cancerígenas e executar a resposta adequada para cada célula alvo (MALE, 2021).

A maior parte da família NKG2 forma heterodímeros com CD94 através de uma ligação dissulfeto e apenas o NKG2D forma homodímeros. Os receptores formados por essas junções podem ser do tipo ativadores ou inibidores, a depender de sua conformação, onde o receptor inibidor apresenta ITIMs celulares (motivo de inibição do imunoreceptor baseado em tirosina) (Figura 1) e o receptor ativador apresenta resíduo de lisina carregado (K) em sua porção transmembrana, que o permite fazer comunicação com os ITAM (motivo de ativação do imunorreceptor baseado em tirosina) (MALE, 2021). Apenas NKG2A e NKG2B (isoforma de NKG2A) se configuram como receptores inibitórios, todos os demais são receptores ativadores (Figura 2). Os receptores inibitórios funcionam porque os ITIMs atraem fosfatases inibitórias que impedem a fosforilação de receptores ativadores, impedindo a ativação da célula NK, isso porque a fosforilação facilita a ativação das fosfatases de tirosina SHP-1 e SHP-2, o que leva a supressão do comportamento da célula NK (BORREGO et al., 2006). Já no caso dos receptores ativadores, o que ocorre é que eles se associam a proteínas intracelulares com ITAMs e, a partir deles, fosforilam e resgatam cinases de tirosina, o que leva à ativação celular. Apesar disso, ainda não se sabe como exatamente é feito o balanceamento entre os receptores ativadores e inibidores (MALE, 2021).

Células NK formam resposta contra células que não tenham o HLA em sua superfície e identificam essas células através de seus receptores inibitórios e ativadores. Caso esses receptores das células NK reconheçam e se liguem a molécula de HLA na superfície da célula alvo, é formada uma resposta de inibição de sua citotoxicidade, resultando na preservação da célula alvo (MALE, 2021). Entretanto, caso não ocorra a identificação do HLA na superfície dessa célula por parte do receptor NKG2, então a célula NK segue com uma resposta de eliminação da célula alvo, o que demonstra que aquela célula está passando por um processo de infecção viral, bacteriana ou por processo cancerígeno.

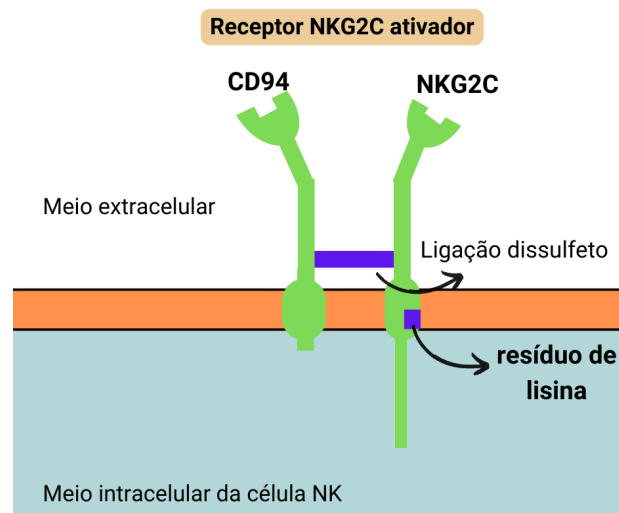
Figura 1 - Esquema de representação dos ITIMs celulares



Representação da formação de um receptor inibitório e dos ITIMs celulares presentes na célula NK e seu receptor CD94/NKG2A, ligados por uma ponte dissulfeto.

Fonte: Adaptado de MALE, 2021.

Figura 2 - Esquema de representação de um receptor ativador



Representação da formação de um receptor ativador e seu resíduo de lisina transmembrana, a qual o permite fazer comunicação com os ITAMs.

Fonte: Adaptado de MALE, 2021.

Os receptores NKG2 são codificados em um cluster no cromossomo 12 (12p13.2). Há uma elevada similaridade de sequência entre os genes da família, que provavelmente se originaram por duplicação gênica a partir de um ancestral comum. Esta similaridade de sequência, acoplada com um eventual polimorfismo elevado (embora não se conheça a diversidade desses genes) poderia atrapalhar as análises de sequenciamento de segunda geração, principalmente quando se avalia sequências curtas (*short reads*). Por exemplo, o excesso de polimorfismo poderia impedir o alinhamento de sequências, acarretando em uma grande quantidade de sequências não mapeadas. A consequência disso seria que apenas sequências mais próximas do genoma de referência seriam alinhadas, levando a genotipagens incorretas, sobretudo a detecção de homozigotes para o alelo presente no genoma de referência. Essa similaridade de sequência entre os genes acarretaria alinhamentos cruzados, com ênfase em estratégias de sequenciamento que produzem sequências curtas. Estes alinhamentos cruzados levariam a erros de genotipagem, principalmente à detecção de heterozigotes que não são verdadeiras. Estes problemas são bastante conhecidos para os genes HLA, que codificam os ligantes principais dos receptores NKG2 (BRANDT et al., 2015; CASTELLI et al., 2018).

Para a análise de genes HLA, foram necessárias estratégias computacionais que corrigem e minimizam esses erros de alinhamento. Uma das linhas de pesquisa do laboratório de Genética Molecular e Bioinformática (GeMBio), onde este estudo foi realizado, visa criar estratégias computacionais que minimizem esses erros de alinhamento nos genes HLA (CASTELLI et al., 2018; CASTELLI et al., 2022).

Dentro deste contexto, para viabilizar estudos de associação entre os genes NKG2 e infecções - por exemplo Covid-19 - e definir o grau de confiabilidade dos métodos tradicionais de análise de variantes para detectar polimorfismos em genes da família NKG2, exploramos a performance de ferramentas de alinhamento de sequências curtas (*short reads*) por meio de simulação de sequenciamentos, com foco especificamente nos genes da família NKG2. Portanto, o objetivo do trabalho foi avaliar se há falhas de alinhamento ou alinhamento cruzado entre os genes da família NKG2, bem como detectar quais são as regiões onde estes erros ocorrem.

METODOLOGIA

Inicialmente, coletamos as coordenadas genômicas de cada gene do complexo NKG2, que incluem os genes *KLRC1*, *KLRC2*, *KLRC3*, *KLRC4* e *KLRK1*, os quais codificam respectivamente as proteínas *NKG2A*, *NKG2C*, *NKG2E*, *NKG2F* e *NKG2D*. Essas coordenadas consistem na posição de início e fim do transcrito canônico de cada gene no cromossomo 12, no genoma humano de referência hg19. Em seguida, utilizamos os dados de genotipagem do projeto 1000 Genomas fase III (GIBBS et al, 2015) para inferir sequências conhecidas destes genes. Esses dados consistem em SNPs e indels já faseados, ou seja, com suas mudanças de base que ocorrem em cada cromossomo. Para isso, inferimos quais seriam as duas sequências completas de cada indivíduo para cada um dos genes do complexo NKG2. Utilizamos a ferramenta *bcftools view* para extrair os SNPs e indels presentes na região de cada gene e criamos as sequências completas utilizando a ferramenta *vcfx fasta* (Molecular Genetics and Bioinformatics Laboratory, 2015). Esta ferramenta utiliza uma referência de cromossomo, neste caso o cromossomo 12, e substitui cada nucleotídeo na posição correta, recriando a sequência do cromossomo. Ao final desta etapa, obtivemos as duas sequências existentes em cada gene, em cada indivíduo incluído no projeto 1000 Genomas.

Avaliamos a frequência observada para cada sequência de um determinado gene, selecionando aquelas que apresentaram ao menos 1% de frequência global considerando os mais de 2500 indivíduos do projeto 1000 Genomas. Essas sequências (uma cópia de cada) foram depositadas em arquivos no formato FASTA, um para cada gene, para identificar as sequências mais frequentes encontradas mundo afora. Além disso, as sequências foram identificadas contendo números e o nome do gene que deu origem a elas.

Em seguida, simulamos sequenciamentos do tipo Illumina, com sequências curtas pareadas, cada qual utilizando como referência dez sequências escolhidas aleatoriamente (duas de cada gene em avaliação). Esses sequenciamentos virtuais foram criados pela ferramenta *Art-illumina* (HUANG et al., 2012). Repetimos este procedimento 50 vezes, simulando 50 sequenciamentos virtuais. Utilizamos os seguintes parâmetros para gerar os sequenciamentos virtuais: `-ss HS25 -p -l 150 -f 30 -m 450 -s 150`. Estes parâmetros simulam um sequenciamento do tipo *paired-end*, com tamanho de *reads* em torno de 150 bases e tamanho médio de fragmento de 450 bases, com desvio padrão de 150. Este padrão é semelhante ao usado em projetos de avaliação de genomas completos. A simulação também introduz erros de sequenciamento, incluindo trocas de base e *indels* em uma proporção semelhante à observada em sequenciamentos reais. O produto gerado aqui são dois arquivos FASTQ por simulação.

Os reads presentes nos arquivos FASTQ foram alinhados ao genoma humano de referência em sua versão hg 38, que é mais recente utilizando o algoritmo **BWA MEM** (*Burrows-Wheeler Alignment*) (LI; DURBIN, 2009), com os parâmetros sugeridos pelo programa. O produto desta etapa é um arquivo no formato SAM, que contém as informações de cada sequência e como estas sequências estão alinhadas ao genoma de referência. Estes arquivos SAM foram processados pela ferramenta **samtools sort** e **samtools index** para criar arquivos BAM, um para cada amostra virtual.

Como as sequências de referência utilizadas para criar o sequenciamento virtual foram nomeadas com a identificação do gene que deu origem a elas, todos os *reads* do sequenciamento virtual também possuem esta identificação em seus nomes. Assim, no arquivo BAM, é possível rastrear onde cada *read* foi alinhado e qual gene deu origem a este *read*. Para fazer este rastreamento, utilizamos a ferramenta **samtools mpileup**, com os seguintes parâmetros: `-aa --output-QNAME`. O arquivo de saída, no formato de texto plano, contém cada posição do genoma de referência e quais *reads* estão alinhados naquela posição. Esta estratégia foi usada para cada uma das simulações, rastreando o número total de *reads* em cada posição e a origem dos mesmos, ou seja, qual gene deu origem a esta sequência. Estes resultados foram plotados em gráficos utilizando a linguagem R e ggplot2 e demonstram as regiões onde ocorrem erros de alinhamento e eventuais cruzamentos entre os genes de uma mesma família gênica.

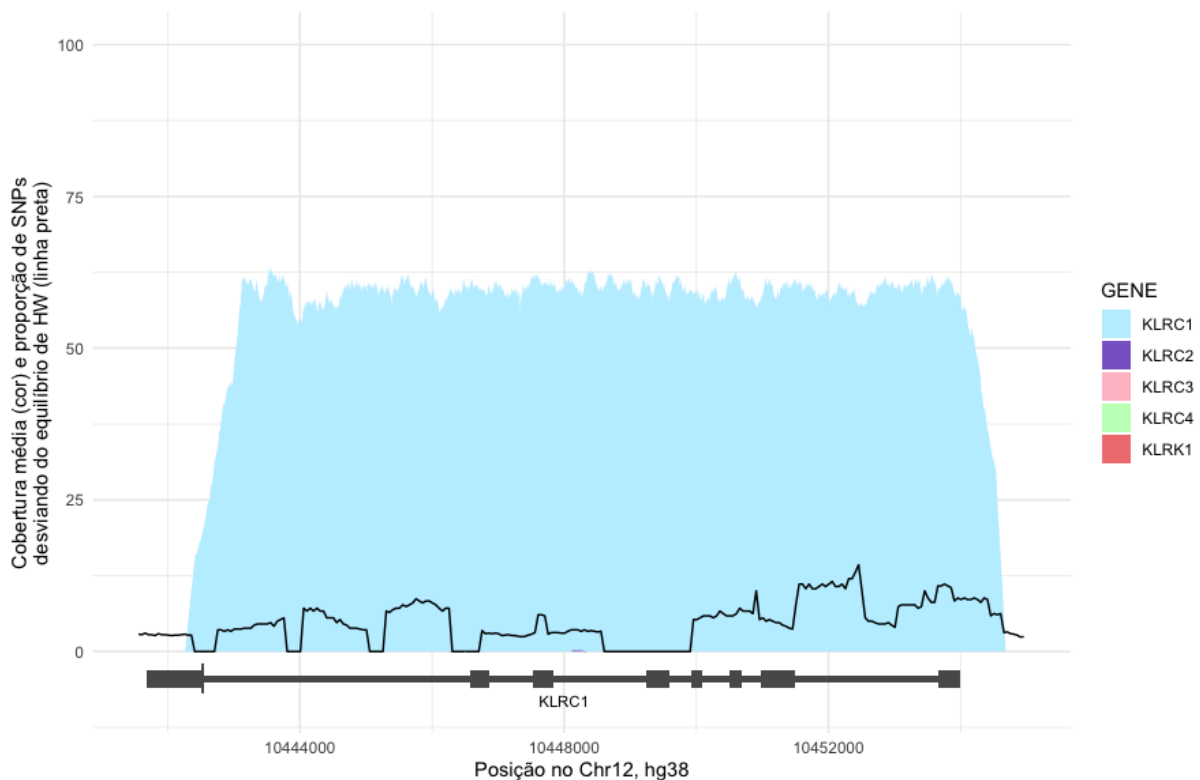
Como uma forma adicional de avaliar erros de alinhamento, avaliamos a distribuição ao longo de cada gene de variantes que fogem ao equilíbrio de Hardy-Weinberg. Esta estratégia parte da premissa de que erros de alinhamento acarretam erros de genotipagem, que por sua vez modificam as frequências genotípicas. Para esta etapa, utilizamos o programa **vcftools --hardy**, que gera um arquivo de texto plano contendo o probabilidade de aderência ao equilíbrio de Hardy-Weinberg (EHW), com base nas frequências genotípicas, para cada variante bi-alelérica. Vindo da lei de mesmo nome, o equilíbrio de Hardy-Weinberg avalia se os genótipos estão em uma proporção adequada para a população, p^2 , $2pq$, q^2 (PIERCE, 2011). Uma população dentro do equilíbrio qualifica que seus genótipos estão dentro do esperado em proporção e variedade genética. Em seguida, avaliamos a proporção de variantes que fogem ao EHW ao longo de cada gene, em janelas de 1000 bases, avançando 50 bases por análise e incluindo todas as variantes existentes dentro de cada janela. Calculamos em cada janela a proporção de variantes que fogem ao equilíbrio, consideramos apenas variantes cujo valor de P é menor que 0.005%. Estes dados foram plotados juntamente com os dados de origem de cada *read*. Todos os procedimentos descritos foram realizados usando scripts em linguagem

de programação Python e Perl desenvolvidos para automatizar cada análise e projetados para articular com softwares baseados em Unix.

RESULTADOS E DISCUSSÃO

Os resultados obtidos mostram a estrutura do gene em análise, a cobertura observada ao longo do gene e a qual membro da família NKG2 pertencem esses *reads*, além de regiões onde se acumulam variantes que fogem ao EHW. A configuração usada na simulação gera um sequenciamento virtual com cobertura próxima a 65X, ou seja, cada base é sequenciada 65 vezes.

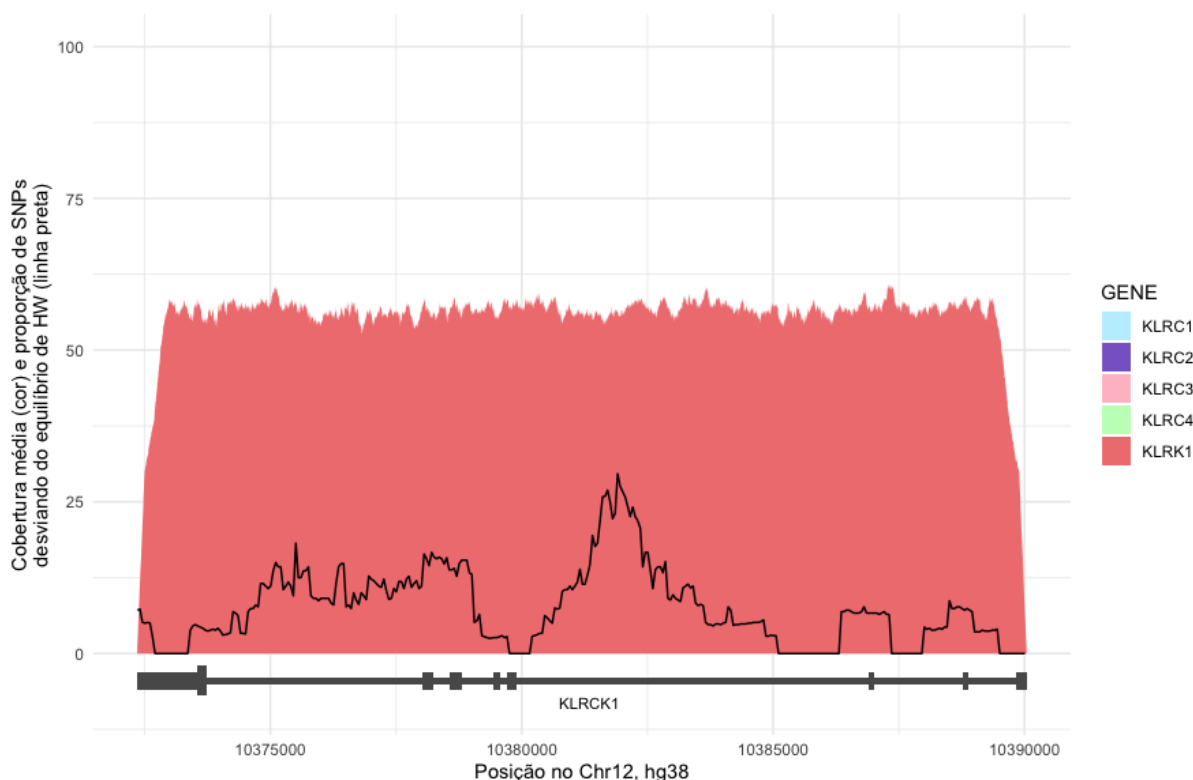
Figura 3 - Gráfico da cobertura do gene *KLRC1/NKG2A*



A cobertura observada ao longo do gene *KLRC1/NKG2A* (cor azul), a origem de cada sequência, a estrutura do gene (exons e íntrons) e a proporção de variantes que fogem ao equilíbrio de Hardy-Weinberg (linha preta contínua), em janelas de 1000 bases com *step-size* de 50. Na estrutura do gene, as linhas finas representam os íntrons e as linhas espessas representam os éxons.

Fonte:Elaborado pelo autor, 2023.

Figura 4 - Gráfico da cobertura do gene *KLRK1/NKG2D*



Cobertura (cor salmão) observada ao longo do gene *KLRK1/NKG2D*, a origem de cada sequência, a estrutura do gene (exons e íntrons) e a proporção de variantes que fogem ao equilíbrio de Hardy-Weinberg (linha preta contínua), em janelas de 1000 bases com *step-size* de 50. Na estrutura do gene, as linhas finas representam os íntrons e as linhas espessas representam os éxons.

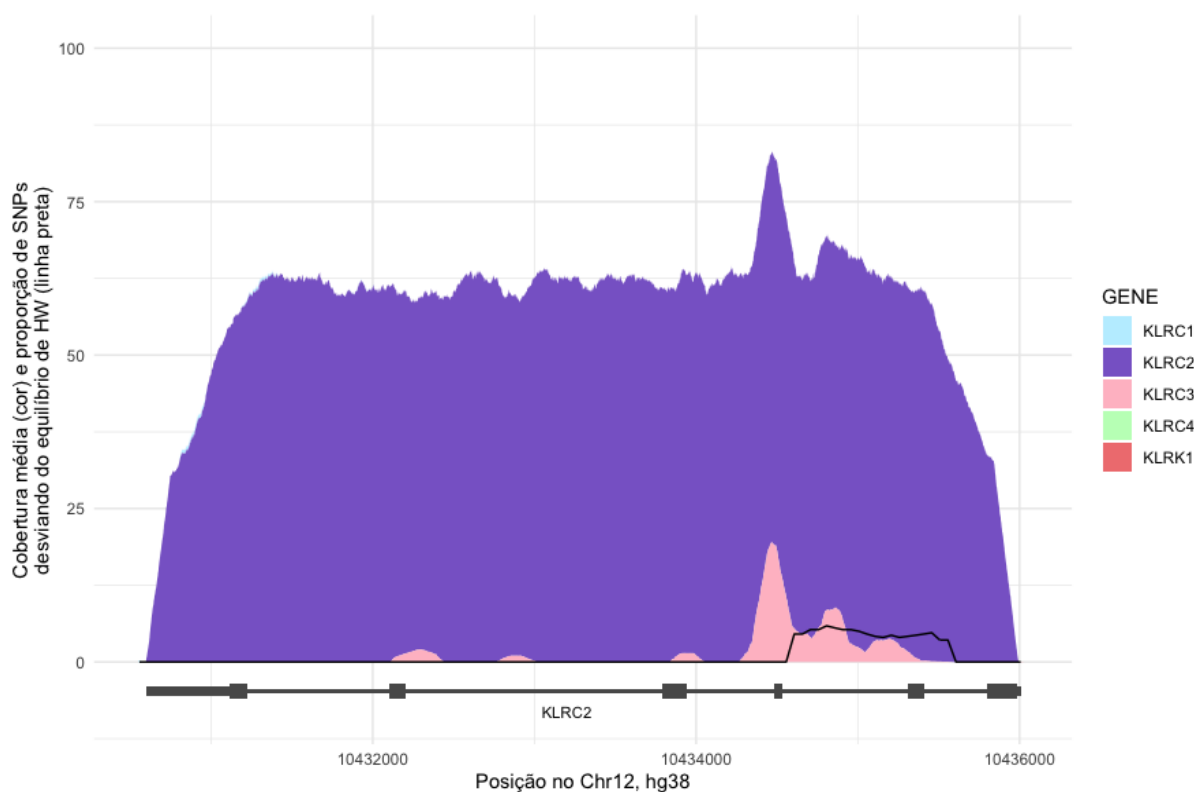
Fonte:Elaborado pelo autor, 2023.

A cobertura observada para os genes *KLRC1* e *KLRK1* ((Figura 3 e Figura 4, respectivamente) foi homogênea e em torno do valor esperado de 65X. Não observamos picos, os quais indicam excesso de *reads* ou vales que indiquem falta de *reads* ou falha de alinhamento em nenhum trecho dos genes. Não foram observados também cruzamento com outros genes da família, visto que todos os alinhamentos de *reads* identificados nesses genes estão alinhados corretamente. Portanto, não é esperado erros de genotipagem por conta de erros de alinhamento nos genes *KLRC1* e *KLRK1*.

A análise da proporção de variantes que fogem ao equilíbrio de Hardy-Weinberg, indicada pela linha preta contínua (Figuras 3 e 4) demonstrou uma elevada proporção de variantes fugindo ao equilíbrio, principalmente nas regiões intrônicas, onde há uma maior similaridade entre os genes NKG2. Salvo raras exceções, as janelas que incluem os éxons apresentam proporções menores de variantes fugindo ao equilíbrio. Mesmo sua maior quantidade estando presente em regiões não codificadoras de íntrons, por estarem envolvidos no processo de splicing, atingem também a expressão gênica, podendo afetar a produção de

proteínas e apresentar leituras de dados inadequadas para trabalhos que tenham como alvo estudar tais proteínas.

Figura 5 - Gráfico da cobertura do gene *KLRC2/NKG2C*



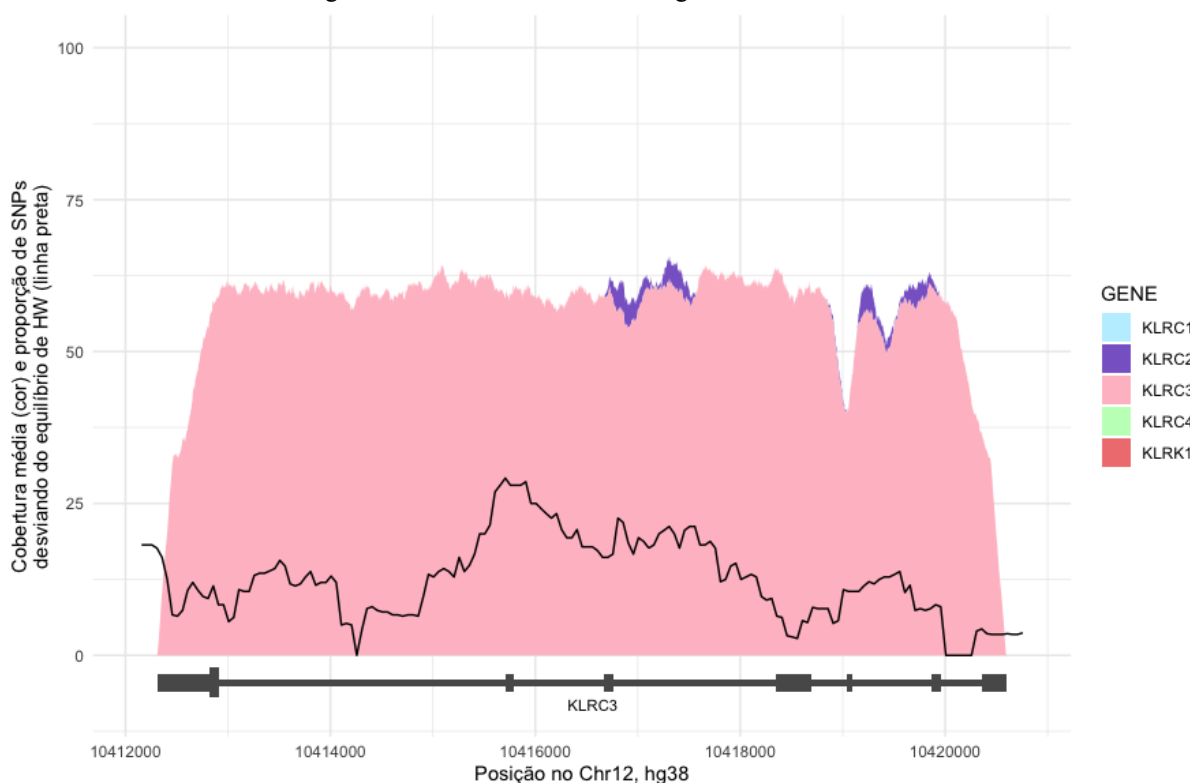
Cobertura (cor roxa) observada ao longo do gene *KLRC2/NKG2C*, a origem de cada sequência, a estrutura do gene (exons e íntrons) e a proporção de variantes que fogem ao equilíbrio de Hardy-Weinberg (linha preta contínua), em janelas de 1000 bases com *step-size* de 50. Na estrutura do gene, as linhas finas representam os íntrons e as linhas espessas representam os éxons.

Fonte:Elaborado pelo autor, 2023.

Os genes *KLRC2* e *KLRC3* (Figuras 5 e 6, respectivamente) mostram um padrão diferente do observado anteriormente. Há regiões com alinhamentos cruzados, ou seja, sequências de *KLRC2* alinhando-se em *KLRC3* e vice-versa. Essas regiões envolvem vários éxons e íntrons. Os erros de alinhamento são enviesados na direção do *KLRC2*, ou seja, há notoriamente mais reads de *KLRC3* alinhados em *KLRC2* do que o contrário. Esses erros acarretam em uma cobertura muito superior ao esperado em parte do íntron II e na totalidade do Exon III do *KLRC2*, formando um pico (Figura 5). Além disso, os *reads* de *KLRC3* que se deslocaram para *KLRC2* causaram uma queda brusca na cobertura em *KLRC3* próximo ao seu Exon III, formando um vale (Figura 6). Este alinhamento cruzado demonstra o quão similar são esses dois genes, principalmente nesta região do Exon III. Esses erros de alinhamento

acarretarão erros de genotipagem, em especial a identificação de heterozigotes que não são verdadeiras, principalmente no *KLRC2*, e identificação de homozigotes no *KLRC3* devido à perda de *reads*. A análise de fuga ao EHW também mostrou um padrão diferente para o *KLRC2*, onde as variantes que fogem ao EHW se acumulam justamente na região com erros de alinhamento.

Figura 6 - Gráfico da cobertura do gene *KLRC3/NKG2E*



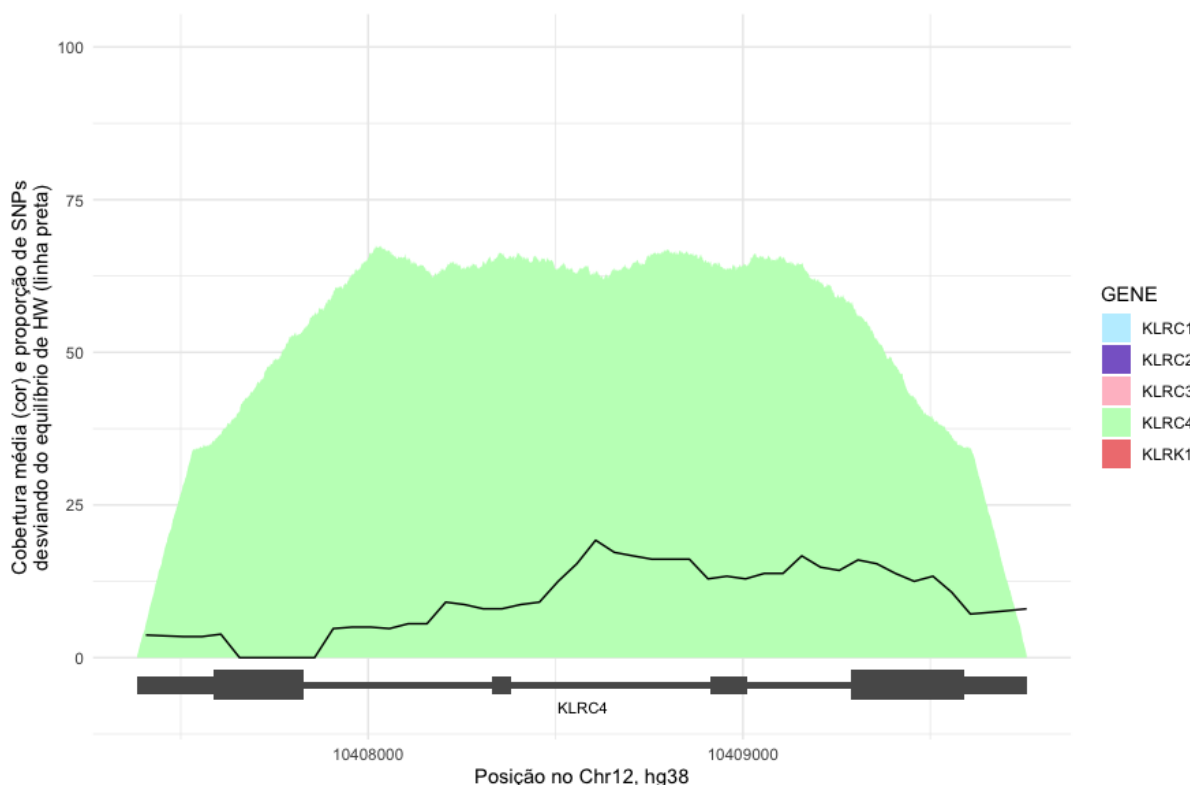
Cobertura (cor rosa) observada ao longo do gene *KLRC3/NKG2E*, a origem de cada sequência, a estrutura do gene (exons e íntrons) e a proporção de variantes que fogem ao equilíbrio de Hardy-Weinberg (linha preta contínua), em janelas de 1000 bases com *step-size* de 50. Na estrutura do gene, as linhas finas representam os íntrons e as linhas espessas representam os éxons.

Fonte: Elaborado pelo autor, 2023.

O gene *KLRC4* (Figura 7) apresentou um padrão distinto daquele observado nos anteriores. É possível notar que não há alinhamentos cruzados com outros membros da família NKG2. No entanto, observa-se uma perda de cobertura considerável nas extremidades do gene, em suas regiões 3' e 5' não traduzidas. O gráfico demonstra que os *reads* dessas regiões não estão adequadamente alinhados ou podem ter se alinhado em outras regiões do genoma. Novas análises serão necessárias para entender o motivo desta falha de alinhamento e identificar a problemática envolvida, podendo se estender a ferramenta de alinhamento utilizada, a população estudada, interações não identificadas entre os membros da família,

transcritos e genes de referência usados e por assim em diante. Nesse caso, como consequência são esperados erros de genotipagem nas pontas desse gene, assim como erros de leitura de dados.

Figura 7 - Gráfico da cobertura do gene *KLRC4/NKG2F*



Cobertura (cor verde) observada ao longo do gene *KLRC4/NKG2F*, a origem de cada sequência, a estrutura do gene (exons e íntrons) e a proporção de variantes que fogem ao equilíbrio de Hardy-Weinberg (linha preta contínua), em janelas de 1000 bases com *step-size* de 50. Na estrutura do gene, as linhas finas representam os íntrons e as linhas espessas representam os éxons.

Fonte: Elaborado pelo autor, 2023.

Os resultados apresentados anteriormente demonstram que, assim como observado nos genes HLA (CASTELLI et al., 2018), são esperados erros de genotipagem por conta de erros de alinhamento em ao menos três dos genes da família de NKG2 estudados. Para o *KLRC4* em específico, se faz necessário compreender o que ocasionou a perda da cobertura nas extremidades do gene já que a análise não indicou um caminho alternativo para os *reads* que deveriam ser alinhados nas pontas do *KLRC4*, ou seja, não foi possível identificar visualmente um ponto para onde seus *reads* poderiam fazer seu alinhamento de forma inadequada em qualquer um dos outros quatro genes avaliados. Portanto, é possível que esses *reads* não tenham se alinhado a nenhuma região do genoma estudado, provavelmente por conta de uma

grande divergência entre as sequências e o genoma de referência, demonstrando que esses dados podem ter se perdido. Isso já foi observado quando se trata dos genes de HLA em que há muitas trocas de base e a sequência final do cromossomo é muito diferente daquela representada no genoma de referência.

Para os genes *KLRC2* e *KLRC3*, o alinhamento cruzado acarreta erros de genotipagem, principalmente na região do Exon III de ambos os loci. Segundo BRENNAN et al. (1996, apud GLIENK, et al., 1998), o Exon III codifica para uma região importante para a especificidade com moléculas como o HLA. Assim, polimorfismos nesta região poderiam influenciar a interação entre essas moléculas, principalmente quando consideramos que as moléculas HLA estão entre as mais polimórficas do genoma. No entanto, os erros de alinhamento que detectamos inviabilizam a análise de genótipos na região do Exon III e sequências adjacentes.

Assim como foi realizado com genes HLA, será necessário desenvolver uma estratégia computacional para corrigir ou minimizar esses alinhamentos nos genes NKG2. Esta estratégia deverá utilizar sequências conhecidas de cada loci para realizar um alinhamento multi-referenciado, em que são consideradas todas as sequências conhecidas de cada loci para definir a origem de cada *read* e onde ele deve ser alinhado na única sequência de referência presente no genoma hg 38. Sem uma estratégia para corrigir os alinhamentos, as genotipagens relacionadas aos genes *KLRC2*, *KLRC3* e *KLRC4* não são confiáveis e poderão levar a resultados enviesados, como falsas associações com doenças infecciosas, identificação de homo e heterozigotes errôneas na população estudada e conclusões precipitadas sobre a qualidade de seu alinhamento genético.

CONCLUSÃO

A análise utilizando simulação de sequenciamentos de segunda geração mostrou a necessidade de desenvolvimento de ferramentas computacionais para correção de alinhamentos com foco nos genes do complexo NKG2 de receptores. Em especial, foram detectados alinhamentos cruzados envolvendo parte dos genes *KLRC2* e *KLRC3*, principalmente e de forma acentuada ao redor do Exon III. Apenas uma análise multi-referenciada dos genes dessa família como a usada em programas que avaliam famílias gênicas e genes hiper-variáveis, como os genes HLA, onde todas as sequências conhecidas de cada gene são consideradas durante a fase de alinhamento, poderá minimizar esses problemas de alinhamento. A próxima etapa deste trabalho será o desenvolvimento desta ferramenta, que será aplicada em estudos envolvendo sequenciamentos reais.

REFERÊNCIAS BIBLIOGRÁFICAS

- ASENJO, J. et al. Diversity of NKG2C genotypes in a European population: Conserved and recombinant haplotypes in the coding, promoter, and 3'-untranslated regions. **HLA**, v.100, p. 469- 478, 2022.
- BORREGO, F. et al. The CD94/NKG2 family of receptors: from molecules and cells to clinical relevance. **Immunologic Research**, v.35, p. 263-277, 2006.
- BRANDT, D. Y. C. et al. Mapping Bias Overestimates Reference Allele Frequencies at the *HLA* Genes in the 1000 Genomes Project Phase I Data. **G3 Genes| Genome | Genetics**, v. 5, p. 931-941, 2015.
- CASTELLI, E. C. et al. MUC22, HLA-A, and HLA-DOB variants and COVID-19 in resilient super-agers from Brazil. **Frontiers in Immunology**, v.13, e. 975918, 2022.
- CASTELLI, E. C. et al. Hla-mapper: An application to optimize the mapping of HLA sequences produced by massively parallel sequencing procedures. **Human Immunology**, v. 79, p. 678-684, 2018.
- GIBBS, R. A. et al. A global reference for human genetic variation. **Nature**, v. 526, p. 68–74, 2015.
- GLINKE, J. et al. The genomic organization of *NKG2C*, *E*, *F*, and *D* receptor genes in the human natural killer gene complex. **Immunogenetics**, v. 48, p. 163–173, 1998.
- HUANG W. et al. ART: a next-generation sequencing read simulator. **Bioinformatics**, v. 28, p. 593-594, 2012.
- LI, H.; DURBIN, R. Fast and accurate short read alignment with Burrows–Wheeler transform. **Bioinformatics**, v. 25, p. 1754–1760, 2009.
- MALE, D. **Imunologia**. Guanabara Koogan, 2021: Grupo GEN, 2014. E-book. ISBN 9788595151451. Disponível em: <https://integrada.minhabiblioteca.com.br/#/books/9788595151451/>. Acesso em: 10 out. 2022.
- Molecular Genetics and Bioinformatics Laboratory. **VCFx**. Molecular Genetics and Bioinformatics Laboratory, 2015. Disponível em: <http://www.castelli-lab.net/vcfx.html> Acesso em: 26 Dez. 2022.
- PIERCE, B. A. **Genética: um enfoque conceitual**. Guanabara Koogan, 2011.