CARACTERIZAÇÃO COMPUTACIONAL DE NEOANTÍGENOS E SUA IMPLICAÇÃO PROGNÓSTICA

ALEXANDRE DEFELICIBUS

Tese apresentada à Fundação Antônio Prudente para obtenção do Título de Doutor em Ciências

Área de concentração: Oncologia

Orientador: Dr. Israel Tojal da Silva
Co-orientador: Dr. Vladmir Cláudio Cordeiro de Lima

São Paulo
2021

FICHA CATALOGRÁFICA

Preparada Ensino Apoio ao aluno da Fundação Antônio Prudente*

D313p Defelicibus, Alexandre
Caracterização computacional de neoantígenos e sua implicação prognóstica / Alexandre Defelicibus - São Paulo, 2021.
92p.
Tese (Doutorado)-Fundação Antônio Prudente.
Curso de Pós-Graduação em Ciências - Área de concentração: Oncologia.
Orientador: Israel Tojal da Silva

Descritores: 1. Biologia Computacional/Computational Biology. 2. Expressão Gênica/ Gene Expression. 3. Imunoterapia/Immunotherapy. 4. Melanoma/Melanoma. 5. Prognóstico/ Prognosis.

Elaborado por Suely Francisco CRB 8/2207

A simplicidade é o último grau de sofisticação.

Leonardo Da Vinci.

A meus pais, Celso e

Sônia

AGRADECIMENTOS

A Deus.

À minha família, pelo carinho, dedicação e incentivo para continuar sempre estudando.

À minha noiva Érica, pela amor e companheirismo.

Ao meu orientador, Dr. Israel, pela oportunidade, confiança e pelo tempo dedicado a esse projeto.

Ao meu co-orientador, Dr. Vladmir, pelos ensinamentos sobre câncer e sistema imune, e pela colaboração nesse projeto.

Aos meus amigos do LBCB, Bruno, Rodrigo Borges, Rodrigo Drummond, Monize, Daniel, Jaqueline, Felipe, Renan e Gabriela, pela companhia e aprendizado.

À CAPES, pelo apoio financeiro inicial.

E à Fundação Antônio Prudente e ao Departamento de Pós-graduação, pelo acolhimento e pelo trabalho realizado na orientação dos alunos.

RESUMO

Defelicibus A. **Caracterização computacional de neoantígenos e sua implicação prognóstica**. [Tese]. São Paulo; Fundação Antônio Prudente; 2021.

Mutações somáticas não sinônimas podem iniciar a tumorigênese e, também, uma resposta citotóxica antitumoral. Com o desenvolvimento das tecnologias de sequenciamento, tornou-se possível identificar as mutações em todos os genes humanos e, consequentemente, as variantes que induzem uma resposta imune (neoantígenos), representando uma oportunidade para pacientes que possam se beneficiar de imunoterapias, mas também um desafio com a necessidade de várias camadas de informações e a integração computacional de vários tipos de dados. Neste trabalho, foi desenvolvido o pipeline de identificação de neoantígeno neo2P, o qual realiza a integração completa de todos os passos necessários para a detecção e neoantígenos e apresentou uma eficiência computacional superior de até seis vezes em comparação com outro método. Além disso, foi proposto um score para priorizaração das mutações somáticas a partir da distribuição dos níveis da expressão gênica de 9.679 pacientes de 32 projetos do TCGA, o qual apresentou um poder de discriminação (AUC) próximo ou superior a 0.7 na maioria das coortes avaliadas. O neo2P foi aplicado em um conjunto de dados de pacientes com melanoma e foram identificados aspectos adicionais da relação de neoantígenos e aspectos imunes, como a expressão de alguns genes marcadores que podem estar relacionados com a resposta ao tratamento. Adicionalmente, a carga de neoantígenos detectados pelo neo2P estratificou, de maneira significativa, pacientes respondedores (R) e não respondedores (NR) quando comparado com o marcador TMB. O pipeline neo2P está publicamente disponível no github: https://github.com/adefelicibus/neo2P.

Descritores: Biologia Computacional. Expressão Gênica. Imunoterapia. Melanoma. Prognóstico.

ABSTRACT

Defelicibus A. **[Computational characterization of neoantigens and their prognostic implications].** [Tese]. São Paulo; Fundação Antônio Prudente; 2021.

Somatic non-synonymous mutations can initiate tumorigenesis and, conversely, anti-tumor cytotoxic T cell (CTL) responses. With the development of next-generation sequencing, it has become feasible to detect mutation-derived neoantigens within exome and thereby predict potential neoantigens, which represents an opportunity to patients that may be treated with immunotherapies, but also a challenge due to multiple layers of information and a computational integration of several types of data. In this work, it was developed a neoantigen identification pipeline called neo2P, which integrates all the necessary steps involved for neoantigen detection and presented a six-times superior computational efficiency compared to another method. In addition, a score was proposed to prioritize somatic mutations based on the distribution of gene expression levels in 9,679 patients from 32 TCGA projects, which showed a stratification ability (AUC) close to or greater than 0.7 in most evaluated cohorts.neo2P was applied to a dataset of patients with melanoma and additional aspects of the relationship between neoantigens and immune aspects were identified, such as the expression of some marker genes that may be related to the treatment response. Additionally, the neoantigen load detected by neo2P significantly stratified responders (R) and non-responders (NR) patients when compared to the TMB marker. The neo2P pipeline is publicly available on github: https://github.com/adefelicibus/neo2P.

Key-words: Computational Biology. Gene Expression. Immunotherapy. Melanoma. Prognosis.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SÍGLAS E ABREVIATURAS

APC	Antigen-presenting cells
API	Application Programming Interface
AUC	Area Under The Curve
BAM	Binary Alignment Map
BLAST	Basic Local Alignment Search Tool
BWA	Burrows-Wheeler Aligner
CPU	Central Processing Unit
ECDF	Empirical Cumulative Distribution Function
FPKM	Fragments per kilo base per million mapped reads
GATK	Genome Analysis Toolkit
GB	Gigabase
GDC	Genomic Data Commons
GTF	Gene Transfer Format
HLA	Human Leukocyte Antigen
IC50	Inhibitory Concentration
ICGC	International Cancer Genome Consortium
IGV	Integrative Genomics Viewer
INDEL	Insertion or Deletion mutation
IMGT	International ImMunoGeneTics Information System
MAF	Mutation Annotation Format
MHC	Major Histocompatibility Complex
PON	Panel of Normals
RAM	Random Access Memory
RNA-Seq	RNA Sequencing
ROC	Receiver Operating Characteristics
SNV	Single Nucleotide Variants
TAP	Transporter Associated with Antigen Processing
TB	Terabase
TCGA	The Cancer Genome Atlas
TCR	T-cell receptor

TIL	Tumor-Infiltrating Lymphocytes
TMB	Tumor Mutation Burden
TME	Tumor microenvironment
TPM	Transcripts per Million
VAF	Variant Allelic Frequency
VCF	Variant Call Format
WES	Whole Exome Sequencing

SUMÁRIO

1 INTRODUÇÃO

O câncer é a segunda principal causa de morte no mundo ocidental. O número de casos de câncer no mundo em 2018, para todos os sexos e idades, foi maior do que 18 milhões. Em relação ao número de mortes por câncer, foram registradas mais de 9 milhões também em 2018 (Bray et al. 2018). No Brasil, a mortalidade de câncer registrada em 2018 foi maior do que 200 mil mortes (Ministério da Saúde 2018). Para 2020, a estimativa é de 600 mil novos casos de câncer no Brasil (Ministério da Saúde 2020). A Figura 1 mostra a estimativa de casos no Brasil em 2020 para os dez tipos de câncer mais incidentes, separado por sexo e tipo de tumor. A Figura 2 mostra a estimativa de incidência e mortalidade mundial por câncer, separados por tipo de tumor em 2020.

Localização primária	Casos	%
Próstata	65.840	29,2%
Cólon e Reto	20.540	9,1%
Traqueia, Brônquio e Pulmão	17.760	7,9%
Estômago	13.360	5,9%
Cavidade Oral	11.200	5,0%
Esôfago	8.690	3,9%
Bexiga	7.590	3,4%
Linfoma não Hodgkin	6.580	2,9%
Laringe	6.470	2,9%
Leucemias	5.920	2,6%

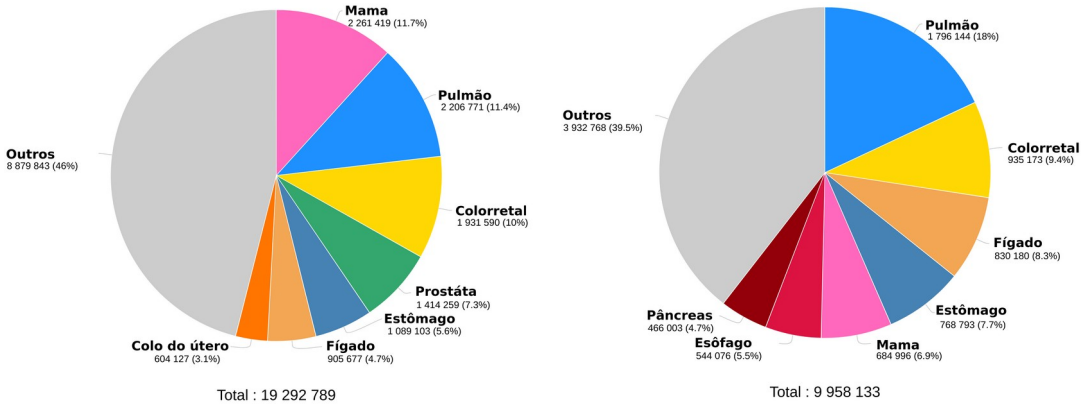
Localização primária	Casos	%
Mama feminina	66.280	29,7%
Cólon e Reto	20.470	9,2%
Colo do útero	16.710	7,5%
Traqueia, Brônquio e Pulmão	12.440	5,6%
Glândula Tireoide	11.950	5,4%
Estômago	7.870	3,5%
Ovário	6.650	3,0%
Corpo do útero	6.540	2,9%
Linfoma não Hodgkin	5.450	2,4%
Sistema Nervoso Central	5.230	2,3%

Fonte: Ministério da Saúde (2020).

Figura 1 – Distribuição proporcional dos dez tipos de câncer mais incidentes estimados para 2020 por sexo, exceto pele não melanoma.

O câncer pode ser caracterizado pelo acúmulo de variações genéticas e epigenéticas que não são reparadas pelo organismo. Durante o envelhecimento do organismo, as células são expostas a eventos endógenos e exógenos que são prejudiciais ao DNA. Essa exposição pode resultar em alterações no material genético, as quais podem ocorrer em regiões do genoma que controlam a divisão celular podendo levar à proliferação desordenada de células. Isso permite uma evolução somática que atua para selecionar células capazes de contornar os mecanismos de

defesa do organismo que controlam a proliferação celular.



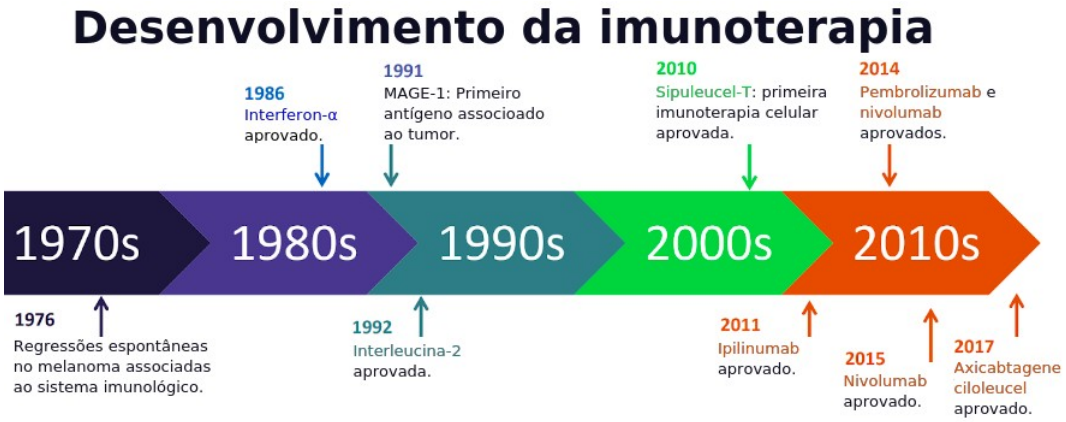
Fonte: Adaptado de: Ferlay et al. (2020).

Figura 2 – Estimativa de número de casos e morte por câncer no mundo, em 2020, separado por tipo tumoral.

Atualmente, existem muitas abordagens estabelecidas no tratamento do câncer, variando de acordo com o tipo e o estadiamento do tumor, sendo as mais utilizadas a cirurgia, terapias baseadas em radiação e a quimioterapia, e as mais específicas, como terapia hormonal, transplante de medula óssea e terapia-alvo (Ortiz et al. 2012). A cirurgia é o tratamento mais antigo aplicado ao câncer e, por ser um procedimento invasivo, muitas vezes compromete a qualidade de vida do paciente, fazendo com que esse método seja mais usado em tumores sólidos localizados (Ortiz et al. 2012; Wyld et al. 2015).

Com as limitações dos tratamentos clássicos do câncer, novas alternativas foram surgindo, como a terapia gênica e a imunoterapia. A imunoterapia está revolucionando o tratamento dessa doença e encontra-se na primeira linha da medicina moderna. Sua ação compreende em modular a atividade de componentes do sistema imune favorecendo o reconhecimento de células tumorais, com o uso de vacinas, terapia celular e anticorpos, contribuindo para benefícios clínicos a longo prazo (Coulie et al. 2014). Atualmente, o uso de várias drogas foi aprovado para vários tipos de tumores, aumentando um pouco mais a abrangência desse trata-

mento. Embora as estratégias disponíveis beneficiam um número ainda reduzido de pacientes, o tratamento resulta em pouco dano colateral quando comparado aos métodos convencionais. A Figura 3 ilustra a evolução da imunoterapia.



Fonte: Adaptado de: Walko e Kudchadkar (2018).

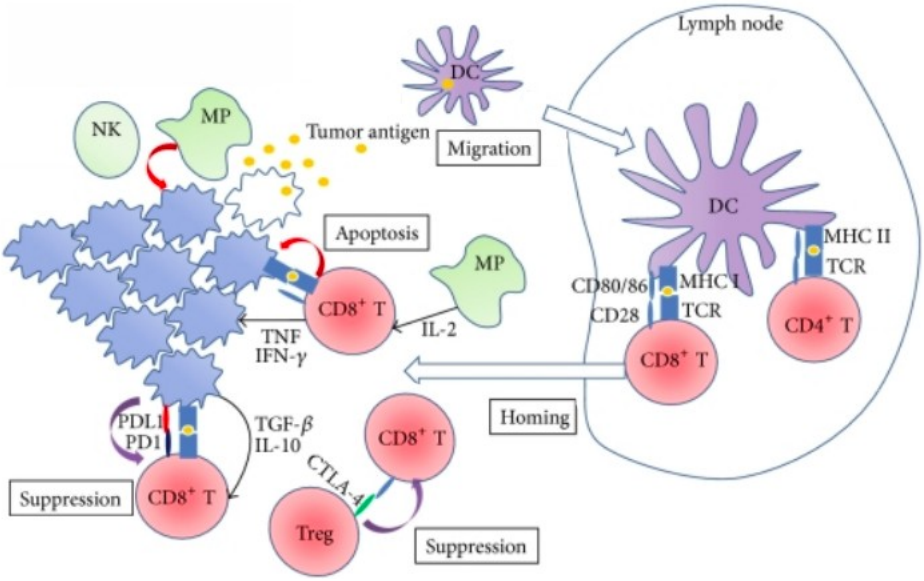
Figura 3 – Marcos históricos durante os últimos 40 anos no tratamento do câncer baseado em imunoterapias.

No microambiente celular, a discriminação do que é conhecido (*self*) e o que é considerado estranho (*non-self*) pelo sistema imune é determinado por um conjunto de proteínas de membranas que formam o complexo principal de histocompatibilidade (MHC, do inglês, *Major Histocompatibility Complex*) (Janeway e Medzhitov 2002). Em câncer, esta discriminação decorre de diferenças fundamentais entre as células cancerosas e suas homólogas normais que, ao adquirirem mutações, modificam os genes codificadores de proteínas e, assim, levam à expressão de proteínas aberrantes, podendo resultar na formação de antígenos tumorais (Lu e Robbins 2016).

Atualmente, são reconhecidos três classes de antígenos tumorais: antígenos tumor-específicos (TSA, do inglês *tumor-specific antigens*); antígenos associados ao tumor (TAA, do inglês *tumor-associated antigens*); e antígenos do tipo câncer-testis (CTA, do inglês *cancer-germline/cancer testis antigens*). Os TSAs são antígenos que não são codificados no genoma normal do hospedeiro e podem representar tanto proteínas virais quanto proteínas anormais que surgem como

consequência de mutações somáticas (isto é, neoantígenos) (Gubin et al. 2015).

Wood e colaboradores, em uma análise computacional (*in silico*), utilizando dados de sequenciamento de exoma completo (WES, do inglês, *Whole Exome Sequencing*) de amostras de tumor de mama, detectou mutações que formavam neoantígenos específicos reconhecidos por células T CD8+ (Wood et al. 2007). A Figura 4 mostra o processo de apresentação de antígeno tumoral para às células do sistema imune e a atuação dessas células no combate às células do câncer.



Fonte: Adaptado de: Uehara et al. (2015).

Figura 4 – Processo de apresentação e reconhecimento de neoantígenos pelo sistema imune.

1.1 CÂNCER

Atualmente, é amplamente aceito que o câncer é uma doença que resulta de diversas mudanças no genoma das células, e essas alterações produzem oncogenes com funções dominantes ou fazem com que genes supressores tumorais não atuem como protetores do genoma (Hanahan e Weinberg 2000). O processo de tumorigênese em humanos pode ser caracterizado como um processo de várias

etapas, as quais se refletem em mutações gênicas que direcionam a transformação de células normais em células malignas (Hanahan e Weinberg 2000). Esse processo, ilustrado na Figura 5, promove características evolutivas vantajosas que, complementarmente, propiciam a progressão do fenótipo maligno (Fouad e Aanei 2017).

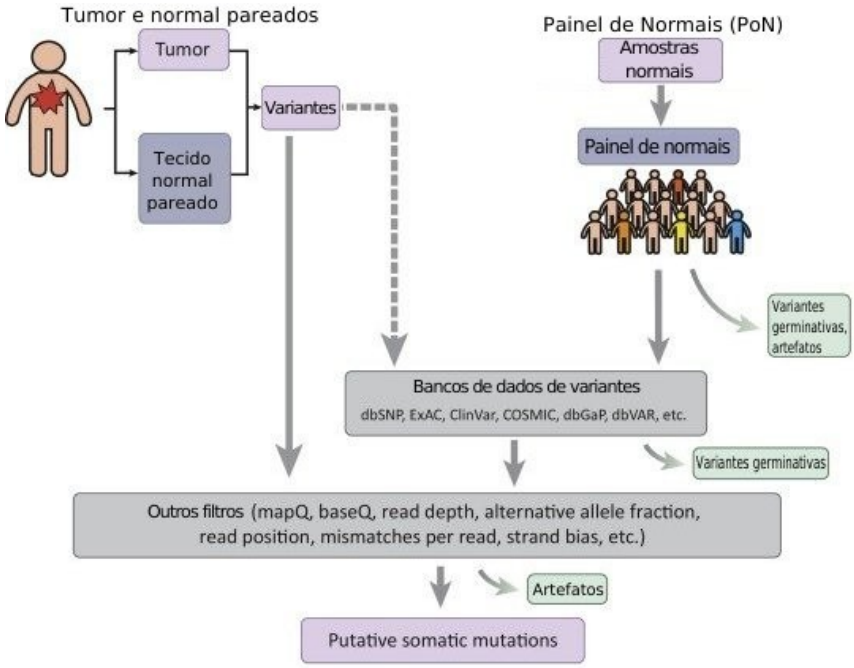
Fonte: Adaptado de: Fouad e Aanei (2017).

Figura 5 – Processo de transformação celular. Diferentes processos agem continuamente nas células levando a mutações genéticas, rearranjos cromossômicos e interações heterotípicas que, ao longo do processo evolucional, apresentam características malignas.

As células tumorais apresentam características específicas e bem diferentes das células normais. Uma dessas características fundamentais está relacionada com a capacidade de manter uma proliferação celular crônica. Um órgão ou tecido normal controla a proliferação e o crescimento celular através de sinais específicos, com o objetivo de manter a homeostase e a sua função correta no organismo. As células tumorais perdem esses sinais e, consequentemente, têm desregulado o processo de crescimento celular (Hanahan e Weinberg 2011).

Os avanços nas análises moleculares e genéticas do câncer, e a melhora das tecnologias de sequenciamento genômico, permitiram a realização de análises mais robustas em relação a diversos tipos de mutações no DNA (Hanahan e Weinberg 2011; Goodwin et al. 2016). Além disso, várias estratégias computacionais diferentes para identificação de variantes somáticas no DNA, utilizando dados de sequenciamento de nova geração, vêm sendo propostas. A Figura 6 ilustra duas

estratégias amplamente utilizadas: (i) amostras de tumor e de tecido normal [1] pareadas e (ii) painel de amostras normais para filtro de variantes germinativas, quando não é possível ter uma amostra normal pareada (Dou et al. 2018).



Fonte: Adaptado de: Dou et al. (2018).

Figura 6 – Identificação de variantes somáticas. Para realizar a chamada de variante somática para um amostra tumoral, é necessário uma amostra normal do mesmo indivíduo, que pode ser gerada a partir de um tecido normal adjacente ao tumor ou do próprio sangue do indivíduo. Caso a amostra normal não esteja disponível, as variantes germinativas, bem como alguns artefatos, podem ser removidos a partir de filtros retirados de bancos de dados públicos com variantes frequentes nas populações, ou a criação de um 'painel de normais', a partir de amostras não tumorais de outros indivíduos. Demais filtros são aplicados para remoção de variantes com baixa qualidade ou provenientes de erros no processo de sequenciamento.

As células tumorais apresentam, além da importante característica de sustentar uma proliferação celular crônica, outras cinco características em relação à biologia do câncer, propostas por Hanahan e Weinberg (2000): (i) não sensibilidade a supressores de crescimento; (ii) resistência à morte celular; (iii) imortalidade replicativa; (iv) indução da angiogênese e (v) invasão de tecido e metástase.

Adicionalmente à instabilidade genômica nas células do câncer, outra característica que confere às células tumorais a capacidade de sobreviver, proliferar

[1] A amostra normal pode ser de um tecido adjacente ao tumor ou amostra de sangue.

e se disseminar, é a possibilidade de se proteger dos ataques das células do sistema imune, como linfócitos T e B, macrófagos e células *natural-killer*.

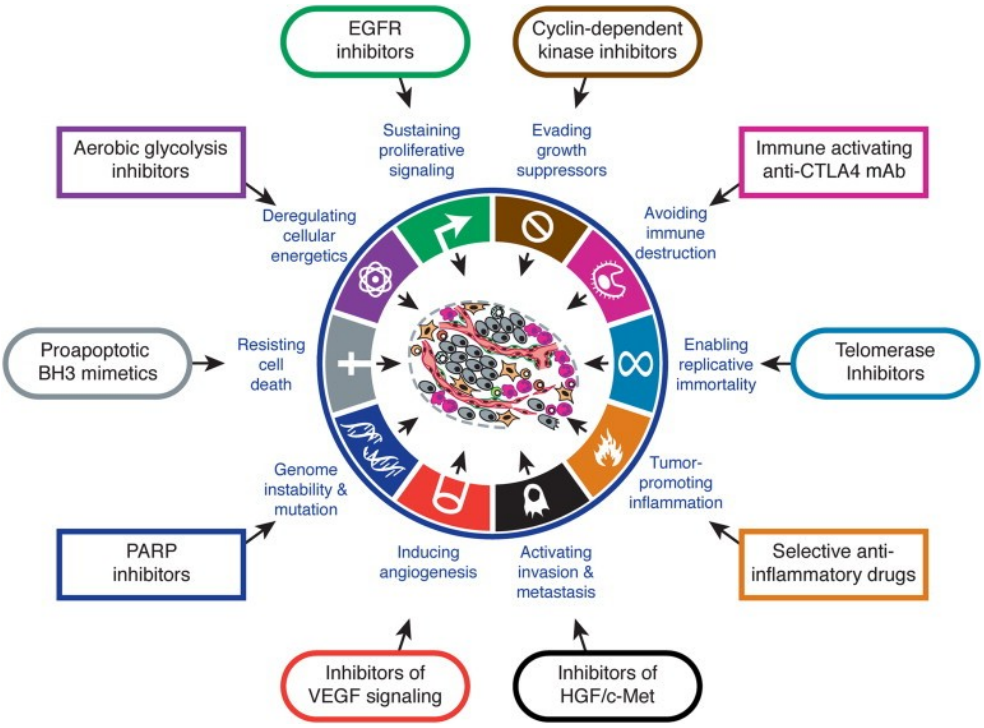
Apesar dessa característica, várias evidências científicas indicam a existência de uma resposta imune antitumoral em alguns tipos de câncer, agindo como uma barreira ao crescimento e progressão tumoral. Por exemplo, pacientes com tumores de colo uterino e ovário com muito infiltrado de linfócitos citotóxicos T CD8 e células *natural-killer* apresentaram um melhor prognóstico em relação aos pacientes com tumores com baixa infiltração por linfócitos (Nelson 2008; Pagès et al. 2010).

Com isso, a partir dessas novas evidências, os tumores podem apresentar um ambiente composto por diversos tipos de células, como células normais, tumorais e do sistema imune, constituindo assim o microambiente tumoral (TME, do inglês *Tumor microenvironment*).A Figura 7 ilustra o TME, as novas características do câncer propostas por Hanahan e Weinberg (2011) e novas terapias que vem sendo desenvolvidas para interferir em cada um dos mecanismos envolvidos nessas características.

1.2 SISTEMA IMUNE E O RECONHECIMENTO DE ANTÍGENOS

O conhecimento a respeito da proteção que o sistema imune realiza no organismo contra as células do câncer é muito antigo (Decker e Safdar 2009) e, atualmente, é amplamente aceito que o sistema imune possui uma importante atuação contra o desenvolvimento e o controle dos vários tipos de câncer, com diversos relatos de casos de regressão tumoral espontânea (Challis e Stam 1990; Chida et al. 2017).

O sistema imune é constituído por diversas células que têm a função de proteger o organismo contra algo desconhecido (*non-self*), como vírus, bactérias e das células do câncer e pode ser caracterizado, de forma simplista, como



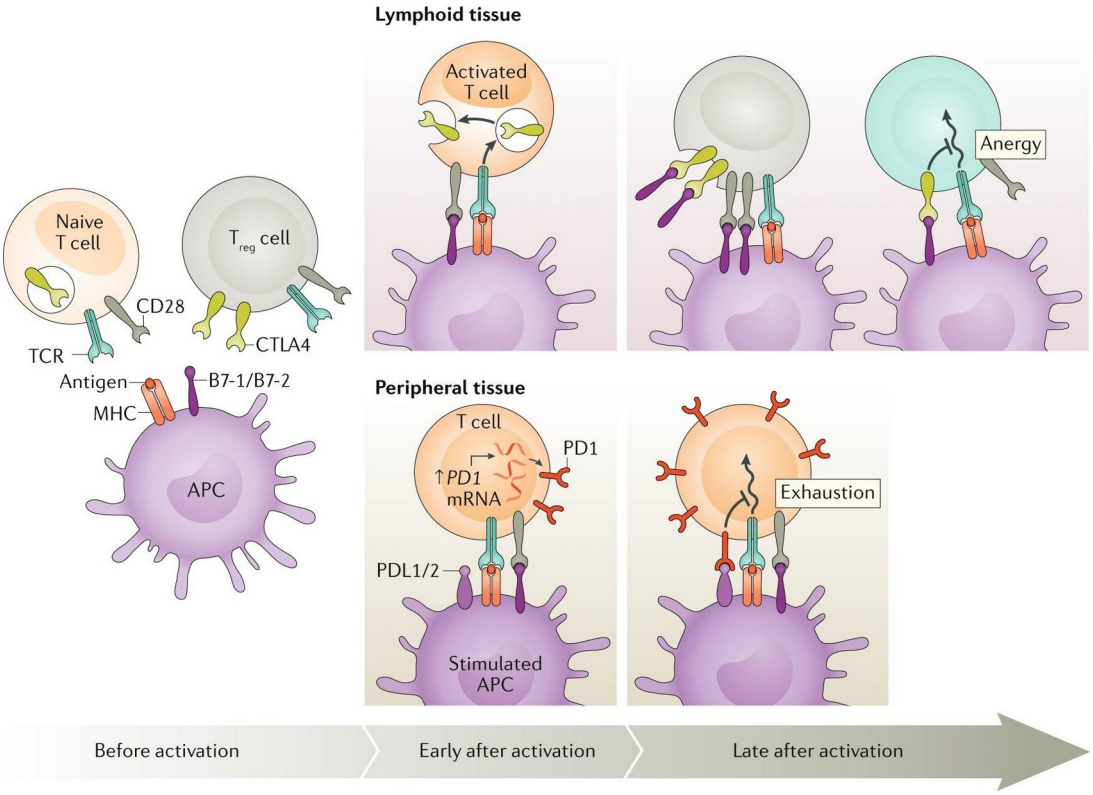
Fonte: Adaptado de: Hanahan e Weinberg (2011).

Figura 7 – *Hallmarks* do câncer e terapia-alvo. Novas características do câncer foram identificadas e, com isso, outras opções de tratamento, como terapia-alvo, vem sendo utilizadas no tratamento do câncer.

tendo duas linhas de defesa: (i) imunidade inata e (ii) imunidade adaptativa. A imunidade inata representa a primeira linha de defesa contra um organismo invasor, não é dependente de antígeno para ser ativada e não possui memória. Por outro lado, a imunidade adaptativa precisa da apresentação de um antígeno para ser ativada, o que faz com que sua resposta seja mais lenta, porém possui a capacidade reter memória do antígeno para que, em uma exposição posterior, o organismo possa responder mais rapidamente (Marshall et al. 2018).

A função principal da imunidade adaptativa é reconhecer os antígenos estranhos ao organismo, também chamados de *non-self*, e diferenciá-los do antígenos conhecidos do organismo, também chamados de *self*. As principais células do sistema imune que atuam na imunidade inata são as células apresentadoras de antígenos (APC, do inglês *Antigen-presenting cells*), como macrófagos e células dendríticas, e na imunidade adaptativa são os linfócitos T e B. As célu-

las T expressam um receptor em sua membrana (TCR, do inglês *T-cell receptor*), enquanto as células APC expressam um grupo de proteínas que compõem o complexo principal de histocompatibilidade (MHC), que também é chamado de Antígeno Leucocitário Humano (HLA, do inglês *Human Leukocyte Antigen*). A Figura 8 ilustra o processo de ativação das células T pelas células APC para reconhecimento de um antígeno.

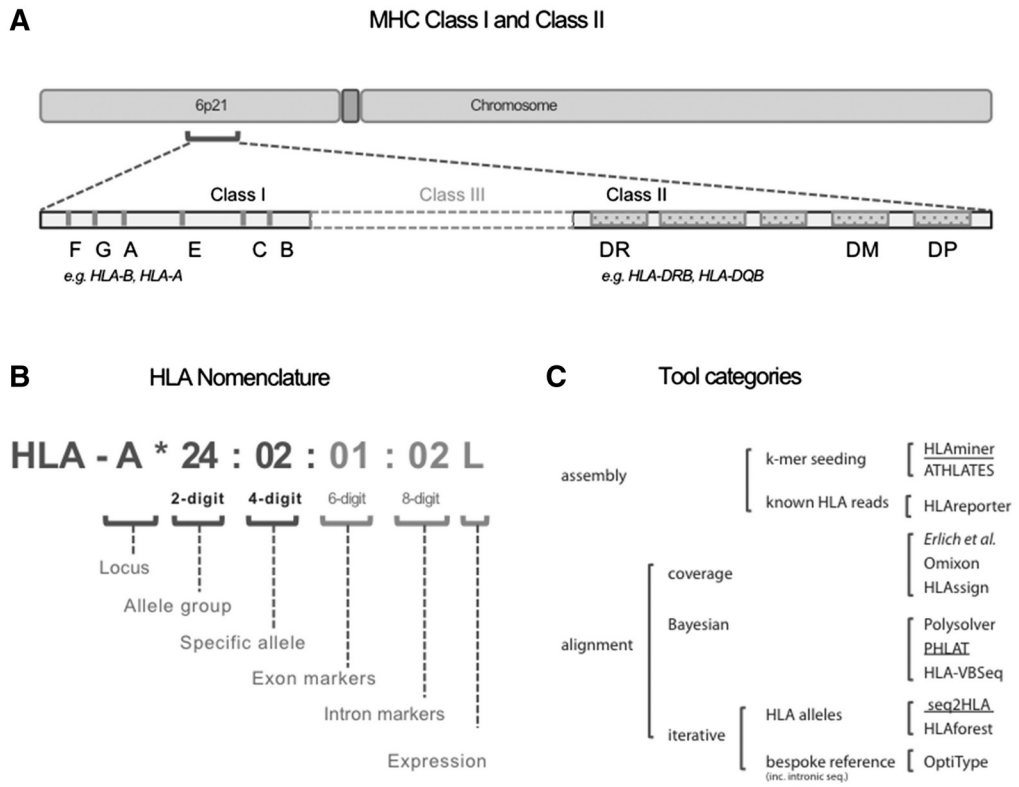


Fonte: Adaptado de: Waldman et al. (2020).

Figura 8 – Mecanismo de ativação e regulação das células TAntes da ativação, as APCs carregam o antígeno no MHC para se ligar ao TCR. As células T são ativadas quando seus TCRs se ligam ao antígeno conjugados com o MHC das APCs. Mais tarde após a ativação, as células T ainda expressam CTLA4 e PD1 nas suas superfícies, promovendo assim a exaustão de células T em locais onde há apresentação crônica de antígeno.

As principais classes do MHC são as classes I e II. A Classe I é representada principalmente pelos genes HLA A, B e C, que são encontrados em todas as células nucleadas e apresentam ao sistema imune peptídeos (fragmentos de antígenos) endógenos (intracelulares), e a Classe II possui os genes HLA DP, DQ e DR, além de outros que não são expressos, os quais são encontrados somente em

algumas células do sistema imune, como células B e dendríticas, e apresentam peptídeos exógenos (extracelulares) para as células T (Marshall et al. 2018). As sequências dos genes HLA estão no cromossomo 6 (6p21.3) (Figura 9A), uma região altamente polimórfica ainda não conhecida completamente e que pode ter um número diferente de alelos com muitas variações nas sequências de DNA (Figura 9B) (Bauer et al. 2018).



Fonte: Adaptado de: Bauer et al. (2018).

Figura 9 – Localização genômica e nomenclatura dos genes HLA. (A) Localização genômica dos genes HLA de Classe I e de Classe II. (B) Nomenclatura dos genes HLA. Para identificação de neoantígeno, utiliza-se somente até os quatro primeiros dígitos. (C) Algumas ferramentas para tipagem de HLA agrupadas por tipo de abordagem utilizada nos algoritmos. As ferramentas sublinhadas aceitam, como dado de entrada, tanto RNA quanto DNA.

Cada indivíduo expressa diferentes alelos HLA e a identificação individual desses alelos (*HLA-typing*) é amplamente utilizada para evitar rejeição no transplante de órgãos (Gourraud et al. 2014). No contexto de câncer e imunoterapia, identificar corretamente os alelos HLA é importante porque esses genes possuem importantes funções na regulação do sistema imune (Shiina et al. 2009), e esse

processo de tipagem de HLA tem se beneficiado com as últimas tecnologias de sequenciamento genético, permitindo análises de integração de dados com alto potencial em aplicação clínica (HLA-*omics*) (Hosomichi et al. 2015).

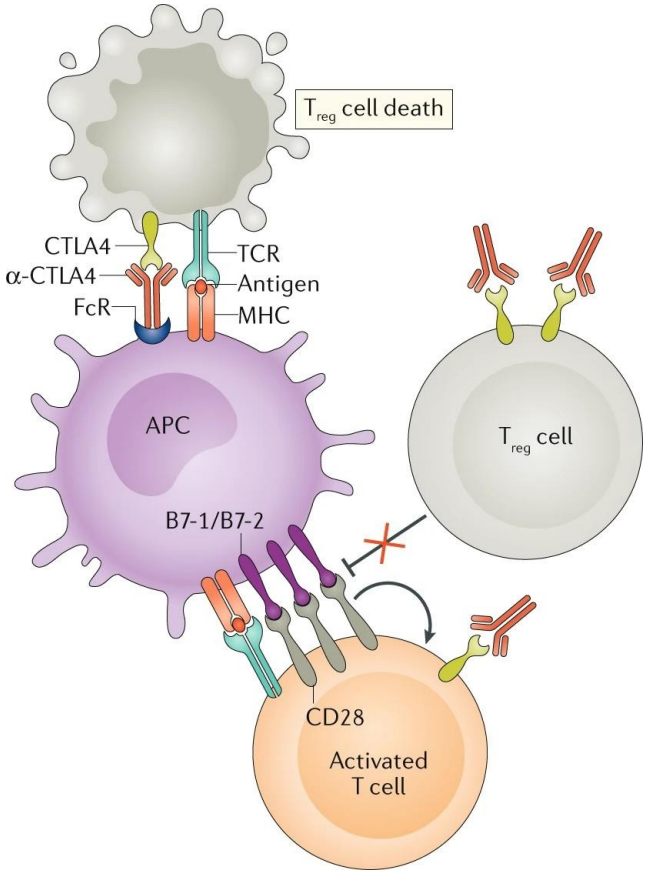
1.3 CÂNCER E IMUNOTERAPIA

Os pacientes com câncer têm, atualmente, diversas opções de tratamento contra essa doença, como cirurgia, quimioterapia, radioterapia, terapia-alvo e, mais recentemente, a imunoterapia, que tem sido considerada o mais novo pilar nas terapias contra o câncer (Oiseth e Aziz 2017).Além disso, uma grande parte dos pacientes com câncer em estágio avançado que respondem a tratamentos como quimioterapia, apresentam recorrência com resistência a esse tratamento (Nelson 2008).

A imunoterapia pode ser definida como um método de tratamento do câncer que gera ou aumenta a resposta do sistema imunológico contra a doença, vem sendo estudada há mais de um século (Coley 1910) e tem resultado em melhora da sobrevida global de pacientes com estágio avançado de câncer (Hodi et al. 2010; Robert et al. 2015). Em um estudo conduzido por Schadendorf e colaboradores em 2015, mais de 20% dos pacientes com melanoma metastático que receberam tratamento com ipilimumabe estavam vivos em 10 anos e sem evidências da doença (Schadendorf et al. 2015).

Uma das abordagens em imunoterapia é o uso de imunoterapêuticos chamados de bloqueadores de *immune checkpoints* que têm como objetivo reiniciar respostas imunes mediadas por células T. As moléculas Antígeno de linfócito T citotóxico 4 (CTLA4, do inglês *Cytotoxic T lymphocyte antigen 4*) e morte celular programada 1 (PD1, do inglês *programmed celldeath 1*) são os alvos para os quais existem drogas disponíveis atualmente. Leach et al. (1996) demonstraram que ao se utilizar um anticorpo anti-CTLA4 em um modelo animal, ativava-se uma resposta imune antitumoral, abrindo caminho para o uso dessa abordagem

no tratamento do câncer (Grosso e Jure-Kunkel 2013). A Figura 10 ilustra o processo e os efeitos do uso de bloqueadores de CTLA4.

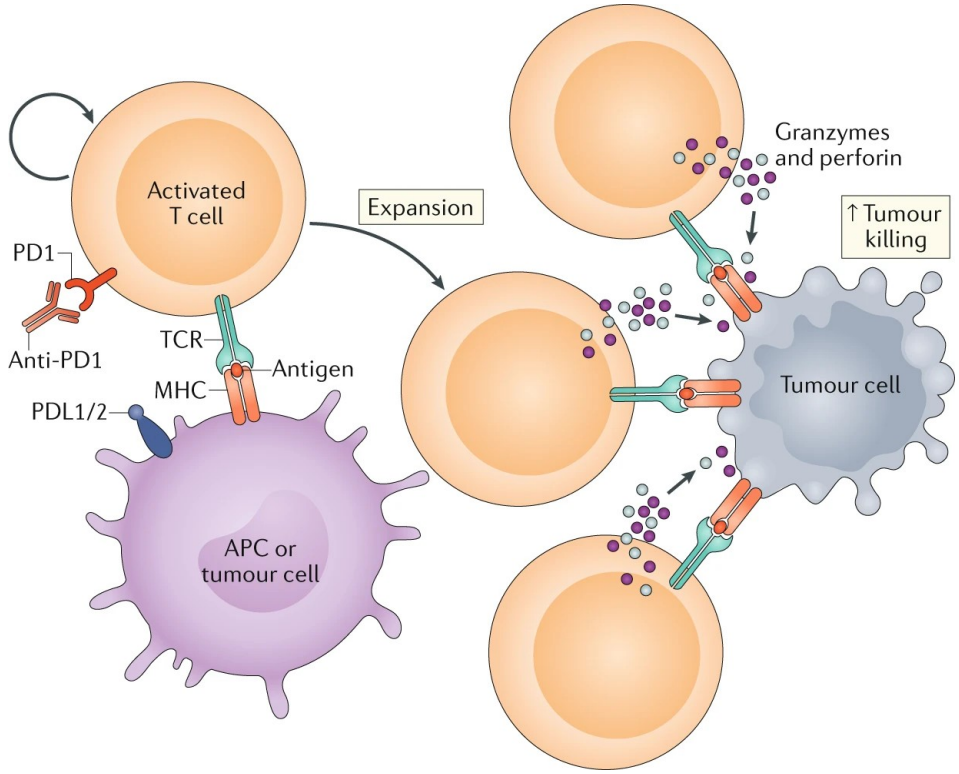


Fonte: Adaptado de: Waldman et al. (2020).

Figura 10 – Efeitos de bloqueadores de CTLA4. Os anticorpos que bloqueiam CTLA4, chamados de α-CTLA4, quando ligados a um receptor Fc nas APCs, podem promover citotoxicidade celular. Ao mesmo tempo, α-CTLA4 também pode promover respostas de células T bloqueando CTLA4 na superfície de células T convencionais à medida que sofrem ativação.

A molécula PD1 (ou PDL1) tem uma função semelhante ao CTLA4, uma vez que a PD1 restringe as respostas imunológicas principalmente por meio de sinalização intracelular inibitória em células T efetoras e reguladoras (Keir et al. 2008). Nivolumabe e pembrolizumabe são alguns dos imunoterapêuticos anti-PD1 utilizados no tratamento de diversos tipos de tumores, tendo promovido aumento de sobrevida global dos pacientes (Ding e Chen 2019). A Figura 11 ilustra o mecanismo de inibição do PD1.

Entretanto, o uso de anticorpos monoclonais no tratamento do câncer não



Fonte: Adaptado de: Waldman et al. (2020).

Figura 11 – Mecanismos de inibição de PD1As células T ativadas expressam PD1, o qual se liga aos seus ligantes PD-L1 ou PD-L2 para inibir a ativação do linfócito. Os bloqueadores anti-PD1 (ou PD-L1 ou PD-L2) bloqueia essa interação, promovendo a atividade antitumoral e proliferação de células T. Com mais células T se ligando aos antígenos tumorais, acontece a liberação de enzimas que atuam na morte celular.

apresenta efetividade para todos os tipos de tumores (Waldman et al2020). Em um estudo conduzido por Snyder e colaboradores, foi demonstrado que existe uma correlação entre a melhora da sobrevida do paciente, ao ser tratado com imunoterapia, com a quantidade de linfócitos no sangue e com o microambiente tumoral (Snyder et al. 2014). Além disso, a carga de neoantígenos se apresenta como um fator determinante para respostas à imunoterapia (Rooij et al. 2013).

1.4 IDENTIFICAÇÃO DE NEOANTÍGENOS

No processo de progressão do câncer, as células tumorais adquirem mutações e parte dessas variações podem gerar proteínas mutantes que são reconhecidas pelo sistema imune como proteínas (epítopos) desconhecidas (*non-self*)

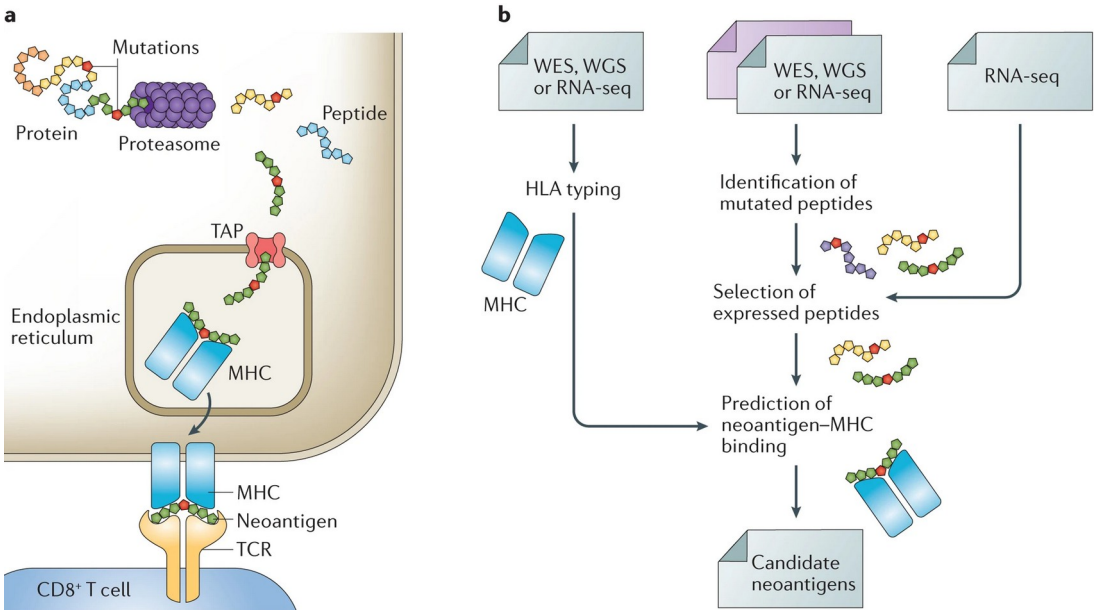
(Gubin et al. 2015). Esses novos epítopos são chamados de neoantígenos.Os neoantígenos derivados do tumor são utilizados para a criação de vacinas e terapias personalizadas contra o câncer, e podem induzir uma resposta imune robusta e duradoura (Waldman et al. 2020).

Em estudos com modelos com camundongos, Monach e seu grupo demonstraram que proteínas mutantes derivadas de células tumorais também funcionavam como neoantígenos altamente imunogênicos (Monach et al1995). Lennerz e seu grupo reportaram um caso de um paciente com melanoma que apresentou uma resposta imune antitumoral espontânea e esta resposta estava relacionada a neoantígenos formados por mutações somáticas em alguns genes distintos (Lennerz et al. 2005). Em outro estudo, Zhou e colaboradores mostraram que um paciente com melanoma apresentou regressão completa do tumor após a transferência de linfócitos infiltrantes de tumor (TIL, do inglês *tumor-infiltrating lymphocytes*), os quais continham células T específicas para dois antígenos tumorais (Zhou et al. 2005).

Inicialmente, os neoantígenos eram identificados através de estímulos nas células T de amostras de pacientes ou de doadores de sangue. Porém, além de ser um processo lento, a maioria das células T reativas aos neoantígenos estavam relacionadas a mutações que não eram compartilhadas entre os pacientes com câncer (Lu e Robbins 2016). Com as tecnologias de sequenciamentos genômico, que geram um grande volume de dados genéticos, é possível identificar mutações somáticas nas amostras tumorais através de análises *in silico* e verificar, através de algoritmos computacionais, quais neoantígenos seriam imunogênicos (Waldman et al. 2020).

A identificação e predição de neoantígeno, no contexto computacional, pode ser dividida em várias etapas, como identificação de variantes somáticas, tipagem de HLA e a verificação de ligação e afinidade entre o peptídeo e o MHC (pMHC). Com isso, é necessário a integração de várias camadas de dados genômicos, como DNA e RNA, e de várias ferramentas e algoritmos para processar e

analisar os diversos tipos de dados (Hackl et al. 2016). A Figura 12 ilustra os processos biológico e computacional para a identificação de neoantígeno.

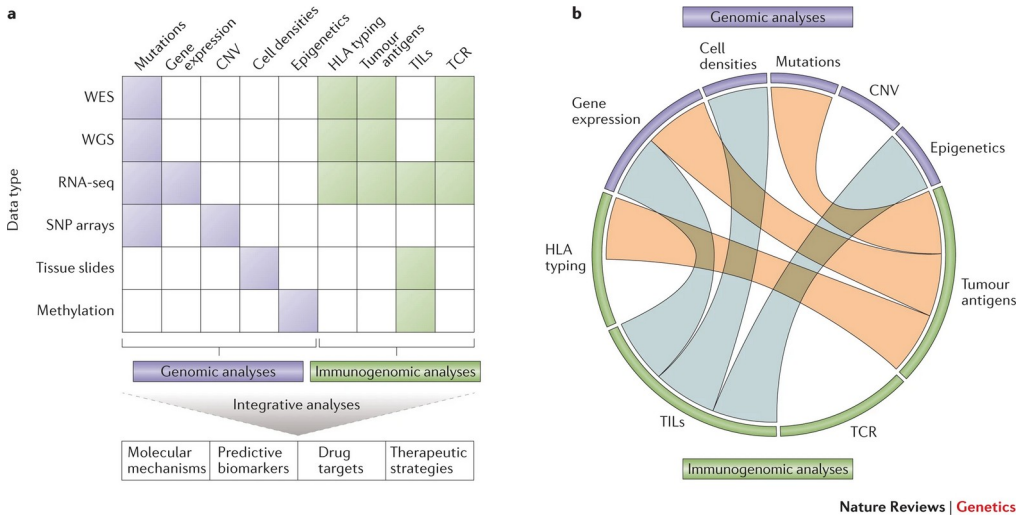


Fonte: Adaptado de: Hackl et al. (2016).

Figura 12 – Identificação de neoantígeno. (a) Os neoantígenos se originam de proteínas modificadas expressadas em células tumorais. Essa proteína é dividida em vários pequenos peptídios pelo proteassoma e transportada pelo Transportador associado com o processamento de antígeno (TAP, do inglês *transporter associated with antigen processing*) até o retículo endoplasmático, onde o peptídio se liga ao MHC (pMHC) e é expresso na superfície da APC. (b) Etapas computacionais para predição de neoantígenos utilizando dados de sequenciamento de nova geração: identificação de variantes a partir de dados de exoma completo (WES); seleção de peptídios expressos a partir de dados de sequenciamento de RNA; tipagem de HLA a partir de dados de WES; e predição de afinidade do pMHC com alelos HLA.

A predição e identificação de neoantígeno ainda enfrenta diversos desafios. Primeiro, o processo de tipagem de HLA (*HLA typing*) é uma etapa crucial para garantir que os neoantígenos preditos vão ser apresentados às APCs e reconhecidos pelas células T. Existem diversos algoritmos para verificar a afinidade dos peptídeos gerados pelas mutações somáticas com o MHC, apresentando acurácia de predição superior a 95%, mas ainda é um problema computacional em aberto (Richters et al. 2019). Segundo, a necessidade de várias camadas de dados (*omics*) é uma limitação no processo de predição de neoantígenos. Por exemplo, para identificar variantes somáticas, é recomendado o sequenciamento de amostras tumorais e normais (sangue ou tecido adjacente ao tumor); e para verificar

se uma mutação somática que ocorre em um gene está realmente expressa, é necessário a informação de expressão gênica gerada pelo sequenciamento de RNA. Por fim, a necessidade de métodos computacionais eficientes que integre as análises genômicas e imunogenômicas é extremamente necessário para tornar viável o processo de identificação de neoantígeno (Blass e Ott 2021). A Figura 13 ilustra a relação entre os diversos tipos de *omics* e como eles podem ser utilizados.



Fonte: Adaptado de: Hackl et al. (2016).

Figura 13 – Integração de dados genômicos para análise genômicas e imunogenômicas. (a) Visão geral de como diferentes tipos de dados ômicos (linhas) são utilizados por ferramentas computacionais para análises genômicas e imunogenômicas (colunas). (b) As análises imunogenômicas são baseadas na análise genômica de dados ômicos e podem ser divididas em tipagem de HLA, quantificação de linfócitos infiltrantes de tumor (TILs, do inglês *tumour-infiltrating lymphocytes*), identificação de TCR e predição de antígenos tumorais. Por exemplo, a predição de neoantígenos requer a integração da análise de expressão gênica, mutações e tipagem HLA.

Por isso, esse trabalho propõe algumas ferramentas que visam facilitar esse processo de identificação de neoantígeno e contribuir para o avanço do conhecimento científico no aspecto computacional do problema de predição de neoantígeno.

2 OBJETIVOS

2.1 OBJETIVO GERAL

Estabelecer um *pipeline* de análise de neoantígenos com afinidade aos alelos HLA do MHC Classe I e a aplicação em um conjunto de dados inicial de cerca de 15.000 pacientes disponíveis no banco de dados do TCGA.

2.2 OBJETIVOS ESPECÍFICOS

- Realizar uma revisão do estado da arte dos métodos de predição, identificação e anotação de HLA e neoantígenos;
- Propor um *score* de expressão gênica para ser utilizado quando não exista a informação de expressão das amostras;
- Relacionar, quando possível, a informação de neoantígeno com aspectos clínicos;
- Estabelecer um *workflow* de análise no contexto de MHC-I.

3 METODOLOGIA

Este capítulo tem como objetivo detalhar os dados públicos (Seção 3.1) utilizados neste trabalho, e o desenvolvimento do score de expressão gênica (Seção 3.2). Além disso, apresenta os métodos computacionais (Seção 3.3) utilizados no desenvolvimento de um pipeline para identificação de neoantígeno, a estimação do infiltrado inflamatório (Seção 3.4) e as análises estatísticas realizadas no projeto (Seção 3.5).

3.1 DADOS CLÍNICOS E MOLECULARES

Nas diferentes etapas deste trabalho, foram utilizados quatro conjuntos de dados públicos: (i) para a criação do *score* de expressão gênica (Seção 3.2) e validação do método foram utilizados os dados do TCGA e do TCGA-PanCancerAtlas, detalhados na Seção 3.1.1; (ii) para a avaliação do *score* de expressão (Seção 4.2.1) em uma coorte diferente foi utilizado os dados disponibilizados por Karasaki et al. (2017), detalhado na Seção 3.1.2; (iii) para o *benchmarking* do pipeline de tipagem de HLA (Seção 4.3) foi utilizado os dados do 1000 Genomes, detalhado na Seção 3.1.4; e, para a aplicação do pipeline neo2P desenvolvido neste projeto foi utilizado a coorte de Hugo et al. (2016), detalhado na Seção 3.1.3.

3.1.1 TCGA

Os dados de expressão gênica (FPKM) e clínicos do TCGA foram baixados utilizando a ferramenta *GADO*, desenvolvida *in house* e que permitiu o download e processamento de todos os arquivos necessários (ver Seção 4.1)No total, foram analisados os dados de 32 projetos e um total de 9.679 amostras com dados clínicos e de expressão gênica disponíveis. Foi utilizado o *Data Release* 27.0 do portal GDC, com o último download realizado em 13 de Novembro de 2020.

Além dos dados clínicos disponíveis no GDC, outros dados foram recuperados utilizando o software TCGAbiolinks (Colaprico et al. 2016). As informações de subtipo imune foram retiradas do PanImmune (Thorsson et al. 2018) e as classificações dos subtipos moleculares foram retiradas de várias análises integradas realizadas para o projeto Pan-Cancer Atlas[1] (Cancer Genome Atlas Research Network et al. 2013b).

Para a análise de AUC e ponto de corte do *score* de expressão gênica, foi utilizado a lista completa de neoantígenos gerada por Thorsson et al. (2018) e, como referência e conjunto-verdade (*ground truth*), foram selecionado todos os neoantígenos com *score* de afinidade IC50$<$ 500 e genes com expressão (TPM)$>$ 1.6). O download dos dados públicos e restritos[2] também foi realizado utilizando a ferramenta *GADO* e o processamento foi realizado na linguagem R[3]. O download dos dados foi atualizado pela última vez em Dezembro de 2020. A Tabela 1 apresenta um sumário completo dos dados utilizados para todos os 32 projetos do TCGA. Foi necessário a utilização dos dados de expressão gênica nas duas métricas, FPKM para os dados do GDC e TPM para os dados do PanImmune, uma vez que os dados foram disponibilizados somente nessas métricas.

3.1.2 Karasaki et al

A avaliação do *score* de expressão gênica em uma coorte diferente foi realizada utilizando os neoantígenos preditos pelo trabalho de Karasaki e colaboradores para quatro amostras de pacientes com câncer de pulmão de não pequenas células (Karasaki et al. 2017). Foi considerado no trabalho as variantes e neoantígenos descritos na tabela suplementar 3.

Esse *dataset* foi escolhido porque ele apresenta diferentes estratégias de identificação de neoantígeno e a contribuição que a informação de expressão gê-

[1] A lista completa de arquivos está disponível em https://gdc.cancer.gov/about-data/publications/pancanatlas

[2] A lista de arquivos está disponível em https://gdc.cancer.gov/about-data/publications/panimmune.

[3] https://www.r-project.org/

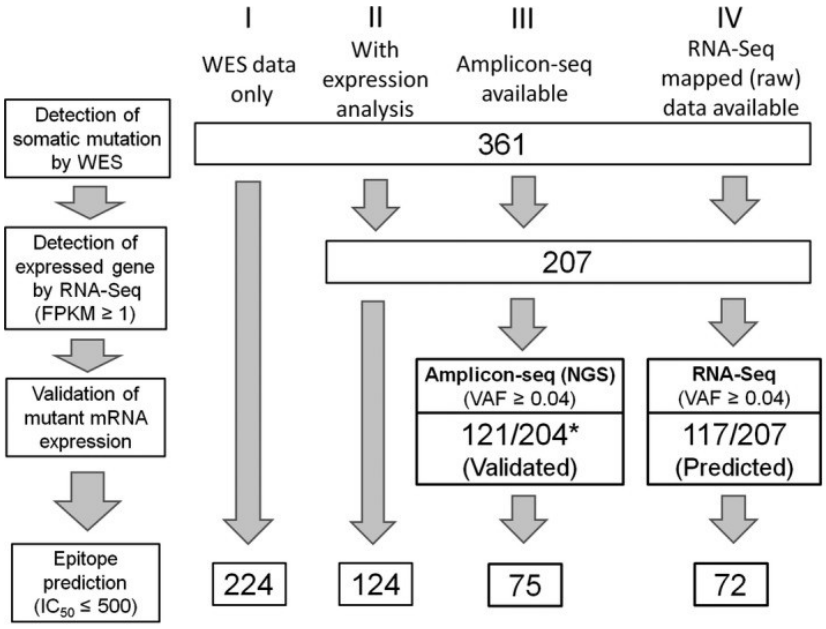
Tabela 1 – Sumário completo do número de amostras (casos) utilizados em todas as análises para todos os 32 projetos do TCGA. Os subtipos moleculares não foram detalhados nessa tabela uma vez que cada tumor tem um subtipo diferente. (exp=expressão gênica; clin=dados clínicos; neo=neoantígenos; -=NA)

	# de amostras por categoria clínica																
	# por tipo de dado			estadiamento					subtipo imune							subtipo molecular	
projeto	exp.	clín.	neo.	I	II	III	IV	total	C1	C2	C3	C4	C5	C6	total	ref. subtipo	total
ACC	79	92	78	9	44	19	18	90	1	1	23	49	3	1	78	(Zheng et al. 2016)	79
BLCA	408	412	404	-	-	-	-	410	-	-	-	-	-	-	397	(Robertson et al. 2017)	407
BRCA	1091	1098	993	203	691	276	22	1073	407	434	212	104	-	46	1083	(Berger et al. 2018)	1096
CESC	304	307	282	163	70	46	21	300	77	217	-	6	-	-	300	-	-
CHOL	36	51	35	20	14	4	10	48	7	2	17	8	-	1	35	-	-
COAD	456	461	369	76	178	129	65	448	332	85	9	12	-	3	441	(Cancer Genome Atlas Network 2012)	341
DLBC	48	58	37	8	17	5	12	42	-	-	-	-	-	-	-	-	-
ESCA	161	185	140	18	79	56	9	162	71	87	7	6	-	2	173	(Cancer Genome Atlas Research Network et al. 2017)	169
GBM	154	617	143	-	-	-	-	-	3	-	-	150	1	-	154	(Ceccarelli et al. 2016)	365
HNSC	500	528	482	27	74	82	270	453	128	379	2	2	-	3	514	(Cancer Genome Atlas Network 2015)	279
KICH	65	113	65	54	33	19	7	113	2	-	38	12	13	-	65	(Davis et al. 2014)	66
KIRC	530	537	350	269	57	125	83	534	7	20	445	27	3	13	515	-	-
KIRP	288	291	267	180	25	52	16	273	3	4	202	66	2	2	279	-	-
LGG	511	516	507	-	-	-	-	-	-	-	10	147	356	1	514	(Ceccarelli et al. 2016)	513
LIHC	371	377	351	175	87	86	5	353	22	45	135	159	-	1	362	-	-
LUAD	513	585	499	279	124	85	26	514	83	147	179	20	-	28	457	-	-
LUSC	501	504	461	245	163	85	7	500	275	182	8	7	-	14	486	(Cancer Genome Atlas Research Network 2012)	178
MESO	86	87	79	10	16	45	16	87	32	21	8	11	-	11	83	-	-
OV	374	608	160	17	30	446	89	582	46	159	3	61	-	-	269	(Thorsson et al. 2018)	487
PAAD	177	185	161	21	152	4	5	182	57	32	40	1	-	21	151	-	-
PCPG	178	179	176	-	-	-	-	-	-	1	121	66	5	2	178	(Fishbein et al. 2017)	173
PRAD	495	500	488	3	74	241	80	398	35	18	307	45	-	-	405	(Cancer Genome Atlas Research Network 2015)	333
READ	166	172	123	33	51	52	25	161	127	18	9	1	-	1	156	(Cancer Genome Atlas Network 2012)	118
SARC	259	261	230	-	-	-	-	-	-	-	-	-	-	-	-	-	-
SKCM	103	470	102	77	140	172	24	411	42	28	14	19	-	2	103	-	-
STAD	375	443	95	59	130	183	44	416	129	210	36	9	-	7	391	(Cancer Genome Atlas Research Network 2014)	416
TGCT	150	150	142	55	12	14	-	81	42	104	2	1	-	-	149	-	-
THCA	502	507	481	285	52	113	55	505	2	13	460	22	-	3	500	-	-
THYM	119	124	116	38	61	15	8	122	-	-	-	-	-	-	-	-	-
UCEC	543	560	509	342	52	124	30	548	247	212	52	16	-	1	528	(Cancer Genome Atlas Research Network et al. 2013a)	507
UCS	56	57	56	22	5	20	10	57	41	14	-	2	-	-	57	-	-
UVM	80	80	76	-	39	36	4	79	-	-	30	48	2	-	80	-	-

nica adiciona no processo de priorização dos peptídeos preditos. A abordagem I utiliza somente as variantes encontradas no exoma das amostras, sem demais filtros; a abordagem II adiciona um filtro de expressão, selecionando somente os neoantígenos relacionados aos genes com FPKM $>$ 1; a abordagem III adiciona o filtro de frequência alélica das variantes identificadas por amplicon (VAF, do inglês *Variant Allelic Frequency*) (VAF $\geq$ 0.04); e, por fim, a abordagem IV adiciona um filtro de frequência alélica das variantes identificadas nos dados de RNA-Seq (VAF$\geq$ 0.04). A Figura 14 ilustra as quatro abordagens utilizadasEm relação aos dados clínicos, a Tabela 2 apresenta algumas características clínicas das amostras utilizadas.

Tabela 2 – Características clínicas das amostras de Karasaki et al. (2017).

Paciente	Idade	Sexo	Tipo de tumor	Estadiamento
LK029	78	M	Carcinoma de células escamosas	T1bN0M0-IA
LK047	41	M	Adenocarcinoma	T1bN2M0-IIIA
LK070	67	M	Carcinoma de células escamosas	T3N0M0-IIB
LK073	67	M	Adenocarcinoma	T2aN0M0-IB



Fonte: Karasaki et al. (2017).

Figura 14 – Diagrama com as abordagens de priorização de neoantígenos realizadas por Karasaki e colaboradores. Ao incluir a informação de expressão gênica na abordagem II, praticamente 50% dos neoantígenos encontradas foram filtrados, demonstrando a necessidade dessa informação na análise de priorização de neoantígenos.

3.1.3 Hugo et al

A aplicação do pipeline desenvolvido nesse projeto foi realizada com os dados gerados por Hugo e colaboradores, que disponibilizaram os dados brutos de exoma completo, RNA-Seq e clínicos de pacientes com melanoma e tratados com pembrolizumabe (Hugo et al. 2016). Algumas amostras também foram utilizadas no *benchmarking* de tipagem de HLA.

Os dados brutos de 20 amostras de pacientes com melanoma foram obtidos através do *download* do SRA [4], com os números de projetos SRP067938, SRP090294 e SRP070710, e correspondem a 134 arquivos FASTQ *paired-end* de exoma completo pareado (tumor e sangue) e RNA-Seq tumoral, com um total de ~780GB de dados.

Além dos dados brutos, foi coletado também os dados clínicos disponíveis para os pacientes, com informações de sobrevida, estadiamento e resposta ao

[4] https://www.ncbi.nlm.nih.gov/sra

tratamento. A Tabela 3 sumariza as informações clínicas das amostras.

Tabela 3 – Informações clínicas das amostras da coorte de Hugo et al.

paciente	sexo	idade	estad.	status	sobrevida	grupo
Pt2	M	55	M1c	vivo	927	R
Pt4	M	62	M1c	vivo	948	R
Pt5	M	61	M1c	vivo	439	R
Pt12	M	59	M1c	morto	327	NR
Pt20	F	63	M1c	morto	337	NR
Pt22	M	55	M1c	morto	182	NR
Pt23	M	63	M1c	morto	103	NR
Pt25	M	74	M1c	morto	262	NR
Pt29	M	84	M1c	morto	269	NR
Pt35	F	65	M1c	vivo	427	R
Pt37	F	70	M1c	vivo	364	R
Pt38	F	57	M1c	vivo	448	R
Pt8	M	69	M1a	vivo	-	R
Pt9	M	68	M0	vivo	1054	R
Pt10	M	60	M1a	vivo	387	NR
Pt13	F	53	M1c	vivo	917	R
Pt14	F	27	M1c	vivo	54	NR
Pt15	M	70	M1b	morto	980	R
Pt28	M	82	M1c	morto	439	R
Pt31	M	47	M1c	vivo	704	NR

3.1.4 1000 Genomes

Os dados brutos de 92 amostras do projeto 1000 Genomes (1000G) (1000 Genomes Project Consortium et al. 2010) foram obtidos utilizando um *script* desenvolvido *in house* a partir do endereço https://ftp-trace.ncbi.nih.gov/1000genomes/ftp/phase3/data, e corresponde a 184 arquivos FASTQ *paired-end* de exomas de amostras de sangue, com um total de 1TB de dados. Essas amostras são de população Britânica e foram escolhidas aleatoriamente para a realização do *ben-*

chmarking de tipagem HLA, uma que a ancestralidade da amostra pode interferir no processo de tipagem do HLA (Abi-Rached et al. 2018).

Além dos dados brutos, as informações dos alelos HLA das amostras foram utilizadas no processo de *benchmarking* de tipagem de HLA para o pipeline desenvolvido (ver Seções 3.3.3 e 4.3). Os dados foram baixados em Janeiro de 2019 e estão disponíveis no endereço: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/.

A tipagem de HLA de todas essas amostras foi realizada por método Sanger. A Figura 15 mostra a diversidade e frequência dos alelos nas amostras identificadas no projeto 1000 Genomes (1000G). O alelo HLA-A*02:01 é o mais frequente, presente em mais de 50% das amostras.

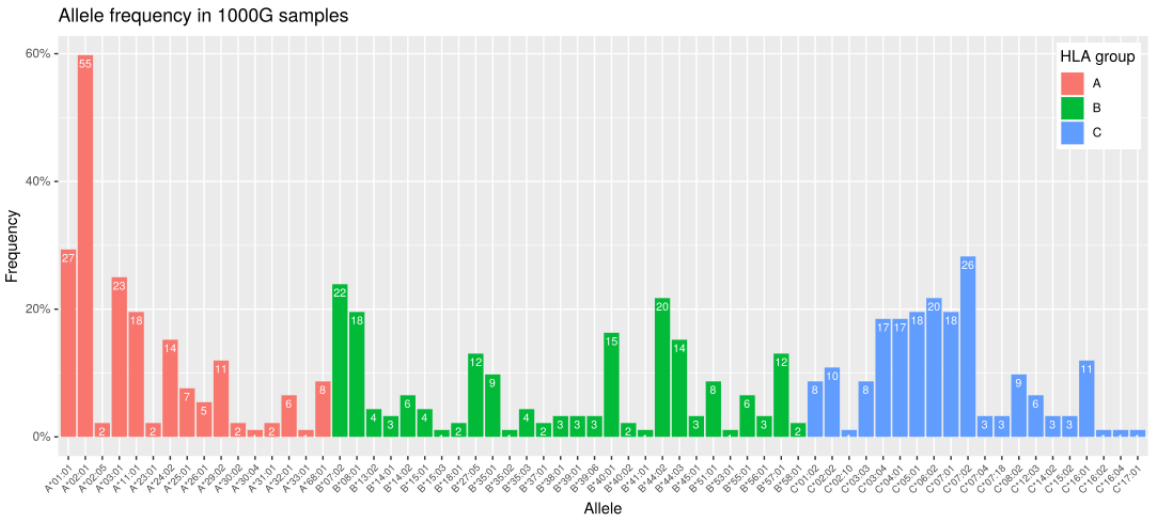


Figura 15 – Diversidade e frequência alélica dos genes HLAs das 92 amostras do 1000G.

3.2 *SCORE* DE EXPRESSÃO GÊNICA

O score de expressão gênica proposto neste trabalho possui duas etapas, as quais serão detalhadas a seguir: a primeira etapa consiste na estratificação dos níveis de expressão (Seção 3.2.1 e a segunda no cálculo do score para cada gene (Seção 3.2.2). Neste projeto foi utilizada a métrica FPKM como medida de expressão gênica.

3.2.1 Estratificação dos níveis de expressão

A partir do valor de expressão cada gene, são definidos 5 níveis de expressão: *high*, *medium*, *low*, *trace* e *silent*, os quais correspondem à estratificação do $\log_{10}$ dos valores de expressão gênica de acordo com quatro quartis possíveis. Com isso, os valores de expressão são normalizados e classificados nas cinco categorias. A Tabela 4 é um exemplo de arquivo de entrada para a estratificação da expressão nos níveis.

Tabela 4 – Exemplo de uma matriz de expressão gênica utilizada como *input* para a estratificação dos níveis de expressão.

	gene_id	**fpkm**
1	ENSG00000000003	7.116
2	ENSG00000000005	1.264
3	ENSG00000000419	32.910
...	...	...
N	ENSG00000281921	0.116

Resumidamente, para cada arquivo de expressão gênica das amostras, os genes que possuem valor de expressão zero é removido do cálculo inicial. Com os genes restantes, é calculado o $\log_{10}$ do valor de expressão gênica e este é dividido em dois clusters. Os genes do cluster 2 são considerados como expressos e divididos em quatro quartis, com as probabilidades $0, \frac{1}{3}$, $\frac{2}{3}$ e 1. Caso o $\log_{10}$ do gene seja maior do que o primeiro quartil, é definido o nível *low*, caso seja maior que o segundo quartil, é definido o nível *medium*, caso seja maior que o terceiro quartil, é definido o nível *high*, do contrário o gene recebe o nível *trace*. Os genes com valor de expressão zero recebem o nível *silent*. O Algoritmo 1 ilustra esse processo e o Apêndice A apresenta o código em linguagem *R* para a estratificação dos níveis de expressão.

3.2.2 Cálculo do *score* de expressão gênica

O *score* de expressão gênica é calculado a partir de uma matriz com os níveis de expressão dos genes (linhas) para cada amostra (colunas), separado por

Algoritmo 1 Definir nível de expressão

```
1: para arquivo_expressao = s1.fpkm, s2.fpkm, ..., sN.fpkm faça
2:     Lê matriz de expressão
3:     Filtra linhas com gene_id duplicados
4:     Cria nova matriz com fpkm>0
5:     Gera dois clusters a partir do log10 fpkm
6:     Divide em quatro quartis os gene_id expressos no cluster 2
7:     para gene_id = 1, 2, ..., N faça
8:         nivel = "silent"
9:         se log10 fpkm > quartil[3] então
10:            nivel = "high"
11:        senão
12:            se log10 fpkm > quartil[2] então
13:                nivel = "medium"
14:            senão
15:                se log10 fpkm > quartil[1] então
16:                    nivel = "low"
17:                senão
18:                    nivel = "trace"
19:                fim se
20:            fim se
21:        fim se
22:    fim para
23: fim para
```

tipo tumoral. Depois de calculado o score, cada gene terá um valor que demonstra um nível de expressão nas amostras do tipo tumoralA Tabela 5 possui os valores de exemplos utilizados como entrada (*input*) para o cálculo do *score*.

Tabela 5 – Exemplo de uma matriz com o FPKM de todas as amostras por projeto do TCGA para input no cálculo do score de nível de expressão gênica.

	gene_id	amostra1	amostra2	amostra3	...	amostraN
1	ENSG00000000003	high	high	medium	...	high
2	ENSG00000000005	low	trace	silent	...	medium
3	ENSG00000000419	high	high	high	...	high
...	...	...	...	...	...	...
N	ENSG00000281921	silent	silent	silent	...	trace

Resumidamente, para cada gene da matriz de níveis de expressão, é contado o número de ocorrências, ou seja, nas amostras, de cada nível e, utilizando esse número de ocorrências, é calculado primeiro a frequência relativa para cada nível e, depois, a média aritmética da frequência relativa. A média aritmética é

utilizada como peso para penalizar os níveis e, assim, aumentando o seu *score*. Portanto, os genes com menor valor de *score* são os com maiores potenciais de estar expresso no contexto to tipo tumoral. O Algoritmo 2 detalha os passos realizados no cálculo. Este algoritmo foi desenvolvido na linguagem R na versão 4.0.

Algoritmo 2 Cálculo do *score* de expressão gênica

1: Carrega matriz de expressão
2: **para** $gene\_id = 1, 2, \ldots, N$ **faça**
3: Conta o número de ocorrências (F_i) para cada $level$
4: Calcula a frequência relativa (F_{R_i}) para cada $level$, onde

$$F_{R_i} = \frac{F_i}{n} \text{ , e } n = \text{total de amostras}$$

5: **fim para**
6: **para** $level = high, medium, low, trace, silent$ **faça**
7: Calcula a média aritmética ($\bar{x}_l$), onde

$$\bar{x}_l = \frac{1}{N} \sum_{i=1}^{N} F_{R_i} \text{ , e } N = \text{total de genes}$$

8: **fim para**
9: **para** $gene\_id = 1, 2, \ldots, N$ **faça**
10: Calcula o *score* (s_i) de expressão gênica, onde

$$s_i = (F_{R_i} * \bar{x})_{high} + (F_{R_i} * \bar{x})_{medium} + (F_{R_i} * \bar{x})_{low} + (F_{R_i} * \bar{x})_{trace} + (F_{R_i} * \bar{x})_{silent}$$

11: **fim para**

A média aritmética foi utilizada como peso no cálculo do *score* porque foi verificado que os níveis *high*, *medium* e *low* apresentavam uma variação menor na frequência relativa, portanto, essa média menor faz com o *score* resultante seja menor, em relação aos níveis *trace* e *silent*, pois eles apresentam uma média na frequência maior e, assim, apresentarão um *score* maior. O objetivo é selecionar os genes com o menor *score*.

3.3 PREDIÇÃO DE NEOANTÍGENOS

Como foi apresentado na introdução deste trabalho, o processo de predição de neoantígeno exige a integração de vários tipos de dados (*omics*) e muitas etapas. As próximas seções detalham os passos que são executados no pipeline apresentado na Seção 4.4.

A Seção 3.3.1 detalha a etapa de chamada de variante somática. O passo de quantificação da expressão gênica, a partir dos dados brutos de RNA-Seq, é detalhada na Seção 3.3.2. Os métodos para tipagem de HLA, para caracterização do repertório de TCR e para o cálculo de afinidade com o TAP são apresentados nas Seções 3.3.3, 3.3.4 e 3.3.5. A Seção 3.3.6 apresenta o método pVAC-Seq, o qual é responsável por gerar os peptídeos a partir das variações somáticas e, por fim, a Seção 3.3.7 apresenta as ferramentas utilizadas para a implementação do pipeline de predição de neoantígenos (Seção 4.4).

3.3.1 Identificação de variantes genéticas

A análise de variante somática para os dados de sequenciamento de exoma foi desenvolvida de acordo com as boas práticas do GATK (McKenna et al. 2010; Van der Auwera et al. 2013), as quais compreende as etapas de alinhamento, pré-processamento e chamada de variante. Resumidamente, o software BWA (Li e Durbin 2009; Li 2013) versão 0.7.17-r1188, no modo *MEM*, é utilizado para alinhamento dos arquivos FASTQ (reads) contra a sequência de referência do genoma humano, que pode ser a versão GRCh37/hg19 ou GRCh38/hg38. O BWA foi configurado para executar com os parâmetros padrões, exceto para o número de *threads* que foi definido para 8 para um processamento mais rápido.

Os arquivos alinhados (BAMs) são pré-processados utilizando o GATK versão 4.0 (Van der Auwera, GA and O'Connor, DB 2020), seguindo os passos: *MarkDuplicates*, *BaseRecalibrator* e *ApplyBSQR*. O programa *MarkDuplicates* consiste na identificação de pares de reads que podem ter sido geradas a partir

de duplicatas dos mesmos fragmentos de DNA. Os programas *BaseRecalibrator* e *ApplyBSQR* são para a aplicação de um algoritmo de *machine learning* para detectar e corrigir erros nos valores da qualidade de cada base gerado pelo sequenciador.

Para identificação de SNVs e Indels, o pipeline utiliza o método *MuTect2*, o qual pode ser executado em vários modosi) amostras de tumor e normal pareado; ii) somente amostras de tumor; e iii) qualquer um dos modos anteriores mais um painel de amostras normais (PoN, do inglês *Panelof Normals*). Em conjunto, os métodos *LearnReadOrientationModel*e *CalculateContamination* utiliza as *reads* para identificar possíveis artefatos de sequenciamento e/ou contaminação que possam ter ocorrido entre amostras no preparo de biblioteca.

A seguir, são aplicados dois filtros nas variantes encontradas: contaminação e artefatos, os quais anotam as variantes que não passaram nesses filtros. Em seguida, as variantes são anotadas com vários bancos de dados, como 1000G (1000 Genomes Project Consortium et al. 2015), dbNSFP (Liu et al. 2020), gnomAD (Lek et al. 2016) e ABRaOM (Naslavsky et al. 2017), pelos métodos *snpEff* (Cingolani et al. 2012b) e *SnpSift* (Cingolani et al. 2012a). Por fim, uma anotação funcional é realizada pelo método *Funcotator* do GATK. O resultado final do pipeline é um arquivo VCF e um arquivo MAF anotados por par de amostras, tumor e normal.

A partir do VCF anotado, os filtros a seguir foram aplicados antes de identificar os neoantígenos: VAF do tumor ≥ 0.1 e VAF do normal < 0.04 e número de alterações no normal < 1 e número de alterações no tumor > 3 e cobertura (*depth*) no tumor e no normal > 10. Não foi aplicado nenhum filtro extra, como variantes encontradas no banco de dados ABRaOM, uma vez que as amostras não são da população brasileira, e variantes identificadas em um PoN porque a análise foi realizada utilizando amostras tumor e normal pareadas.

3.3.2 Expressão gênica

A análise de expressão gênica para os dados de sequenciamento de RNA-Seq foi desenvolvida de acordo com as boas práticas disponíveis na literatura, as quais compreende as etapas de filtro de contaminantes (ribossomal), alinhamento, contagem e expressão diferencial. Em resumo, os arquivos FASTQ são alinhados contra a sequência de referência de ribossomal utilizando o software BWA versão 0.7.17-r1188, no modo *MEM*, com os parâmetros padrões. As reads não mapeadas nessa etapa seguem para a etapa de alinhamento contra a sequência de referência do genoma humano, que pode ser a versão GRCh37/hg19 ou GRCh38/hg38, utilizando o software STAR (Dobin et al. 2013), na versão 2.6.1a_08-27, com parâmetros padrões, exceto os parâmetros *–chimSegmentMin 20*, *–limitIObufferSize 62500000* e *–runThreadN 8*.

A contagem dos reads por gene é calculada utilizando o software HTSeq-count versão 0.11.2 (Anders et al.2015) com todos os parâmetros padrões, exceto *-r pos -s no*. O banco de dados de anotação é um arquivo GTF[5] (do inglês, *Gene Transfer Format*) do Ensembl na versão 92 no mesmo *build* da referência do genoma humano (GRCh37 ou GRCh38). Para normalização dos *counts* é utilizada a métrica FPKM (do inglês, Fragments Per Kilobase Million) utilizando um script em python 3 desenvolvido no laboratório de Bioinformática do A.C.Camargo.

A análise de expressão diferencial é realizada utilizando o software DEseq2 versão 1.18.1 (Love et al.2014), disponível no pacote *Bioconductor* para R versão 3.6, e segue as boas práticas listadas no *Standard Workflow* disponível na literatura[6]. Os parâmetros utilizados nessa análise estão definidos como: *test_method Wald, fit_method parametric, padj BH* e *normalization vst*. Para o cálculo da saturação, é considerado uma cobertura mínima de 1(uma) read. Os gráficos são gerados na ferramenta R versão 3.6, utilizando as bibliotecas *cowplot*, *ggbiplot* e

[5] GTF é um formato de arquivo que contem informações sobre a estrutura dos genes.

[6] Manual do Deseq2: http://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html

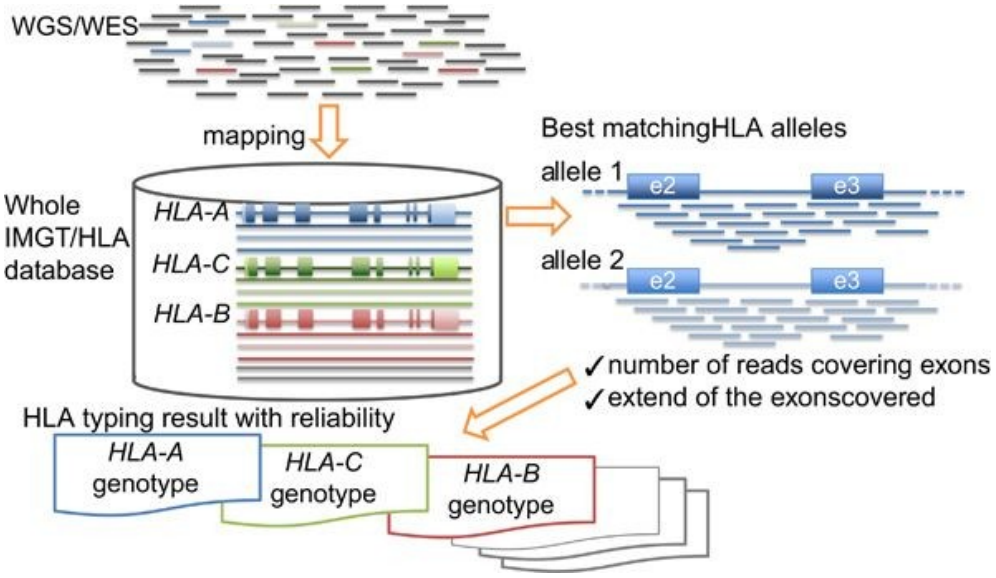
pheatmap.

3.3.3 HLA-*typing*

Uma etapa importante para a predição de neoantígeno é o processo de identificação dos genes da família do HLA (do inglês, *Human Leukocytes Alleles*) do indivíduo. Esses genes fazem parte do Complexo Principal de Histocompatibilidade (MHC, do inglês, *Major Histocompatibility Complex*), que é responsável por codificar proteínas do sistema imune que reconhecem e apresentam antígenos às células T. Estes genes são encontrados no cromossomo 6 e são altamente polimórficos, o que dificulta o processo de predição (Hosomichi et al. 2015).

Para definir um método de tipagem de HLA, foi realizado um *benchmarking* das principais ferramentas e algoritmos disponíveis na literatura, com o objetivo de testar, além dos diversos tipos de algoritmos capazes de predizer corretamente os HLAs, os tipos de amostras de NGS que podem ser utilizadas nesse processo. Atualmente, os principais métodos utilizam amostras de sangue (ou tecido normal) ou RNA-Seq e identificam os alelos HLA com até 90% de acurácia através de variantes germinativas (Hosomichi et al. 2015). Entretanto, no contexto de pesquisa de câncer, nem sempre há disponibilidade de amostras de tumor e normal pareadas para o mesmo indivíduo. Assim, será que, ao utilizar amostras de tumor, os softwares seriam capazes de predizer os alelos com a mesma acurácia?. A Figura 16 ilustra o processo canônico de identificação de HLA utilizando dados de NGS.

Nesse *benchmarking*, foram utilizados quatro softwares com abordagens distintas em relação ao processo de mapeamento e identificação de *reads* na região do cromossomo 6, *score* e ordenação dos melhores HLAs encontrados.Os quatro softwares são: Polysolver (Shukla et al. 2015), Optitype (Szolek et al. 2014), SOAP (Cao et al.2013) e seq2HLA (Boegel et al.2012), os quais serão detalhados nas Seções 3.3.3.1, 3.3.3.2, 3.3.3.3 e 3.3.3.4 respectivamente. Para este teste, foram utilizados os *datasets* públicos descritos nas Seções 3.1.3 e 3.1.4, os quais



Fonte: Hosomichi et al. (2015).

Figura 16 – *Overview* do processo de análise de dados para tipagem de HLA utilizando dados de NGS. As reads geradas no sequenciamento de genoma, exoma ou RNA-Seq são alinhadas ao banco de dados de HLA (IMGT/HLA) para encontrar os alelos com maior identidade de alinhamento e cobertura. O alelo escolhido é o que tem mais reads mapeadas e menos erros de sequenciamento (do inglês, *mismatch*).

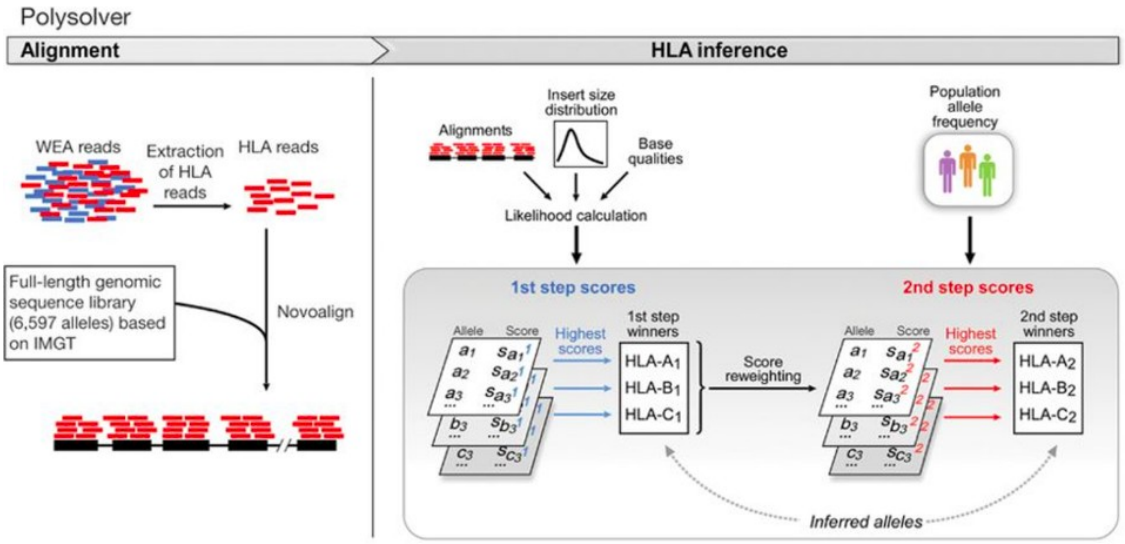
possuem amostras pareadas de DNA, tumoral e normal, e amostras tumorais de RNA-Seq, bem como dados de HLA identificados por meio de PCR-Seq, padrão ouro de tipagem de HLA (Hosomichi et al. 2015).

3.3.3.1 Polysolver

O software Polysolver é um programa escrito em linguagem Bash que executa um pipeline de predição de HLA através do algoritmos desenvolvidos em linguagem Perl para identificação e priorização de alelos HLA Classe I. O software utiliza arquivos no formato BAM contendo as leituras (reads) previamente alinhadas contra o genoma humano de referência (hg19). A primeira etapa compreende em recuperar somente as reads alinhadas na região do cromossomo 6, onde estão os genes do MHC. A partir das reads da região do HLA, é realizado um alinhamento local pelo software Novoalign (NovoCraft 2020), o qual corrige possíveis erros de alinhamento, uma vez que o software identifica variações com

cobertura baixa ou que são resultados de erros do sequenciamento.Com as reads da região do HLA alinhadas, o software calcula um *score* para cada alelo, ou seja, uma probabilidade de que as reads tenham identificado o HLA correto (Shukla et al. 2015).

Com essa matriz de HLAs e o seu score correspondente, o Polysolver aplica um segundo passo de priorização através de um filtro de frequência alélica populacional, caso a etnia da amostra seja informada. Na versão atual (v4), foi incorporado um pipeline para identificação da ancestralidade da amostra, mas esse recurso não foi utilizado nesse trabalho. Os dois alelos para cada grupo de HLA-A, HLA-B e HLA-C com maiores *scores* são o *output* do software. O Polysolver está publicamente disponível pelo link: https://software.broadinstitute.org/cancer/cga/polysolver. A Figura 17 ilustra o algoritmo de predição do programa.



Fonte: Shukla et al. (2015).

Figura 17 – Algoritmo de predição e priorização de alelos do HLA Classe I pelo método Polysolver. Após a extração das reads alinhadas na região dos alelos HLA (cromossomo 6), essas reads são alinhadas no banco de dados de HLAs conhecidos (IMGT/HLA). As reads corretamente alinhadas são utilizadas para inferência dos alelos HLA a partir de um método de classificação Bayesiano.

3.3.3.2 Optitype

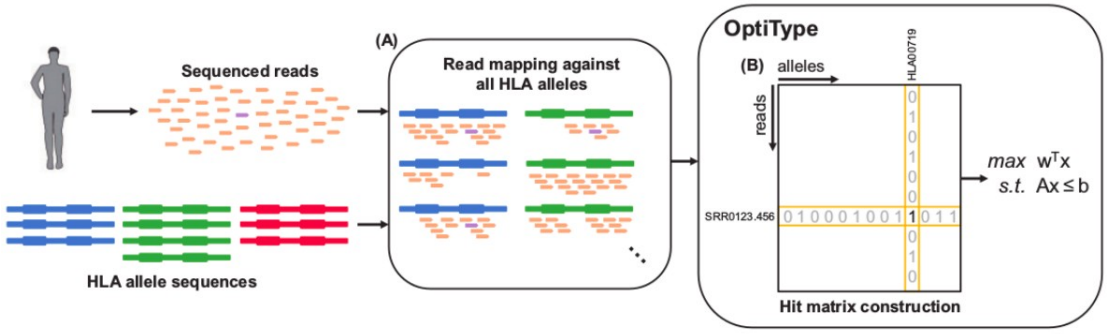
O método Optitype é um software escrito em linguagem Python que realiza a predição e priorização de alelos HLA Classe I e Classe II a partir de reads de exoma de amostra de sangue. O processo é divido em três passos: i) primeiro, as reads são alinhadas contra um banco de dados de sequências criado para garantir a melhor acurácia no processo de alinhamento. Esse processo é realizado pelo software RazerS3 (Weese et al. 2012), que é um algoritmo para alinhamento de um grande volume de sequências pequenas, mesmo com a presença de erros ou *gaps*[7]; ii) segundo, é gerado uma matriz binária indicando em qual alelo uma read foi alinhada, considerando um número de *mismatches*[8]; iii) por fim, com base na matriz gerada no passo anterior, um algoritmo seleciona dois alelos por grupo, maximizando o número de reads que pode explicar o genótipo predito (Szolek et al. 2014).

O *output* do método é uma lista com, no máximo, dois alelos para cada grupo de HLA-A, HLA-B e HLA-C, no caso dos alelos da Classe I. Além disso, a tabela gerada possui o *score* e o número de reads para cada alelo. A Figura 18 ilustra o algoritmo utilizado do método Optitype. O código-fonte está disponível no github: https://github.com/FRED-2/OptiType.

3.3.3.3 seq2HLA

O software seq2HLA é um sistema escrito em linguagem Python para realizar a predição de alelos HLA Classe I e Classe II, inicialmente desenvolvido para utilizar somente dados de de RNA-Seq no processo de predição, mas que também pode ser utilizado com dados de DNA-Seq, como foi mostrado no *benchmarking* realizado por (Kiyotani et al. 2017). Iniciando a partir de arquivos de reads FASTQ, o software utiliza o algoritmo Bowtie (Langmead et al2009) para alinhar as reads contra uma referência de sequências dos alelos HLA. O primeiro

[7] *Gaps* são inserções ou deleções que ocorrem na sequência
[8] *Mismatch* pode ser interpretado como uma mutação pontual na sequência



Fonte: Szolek et al. (2014).

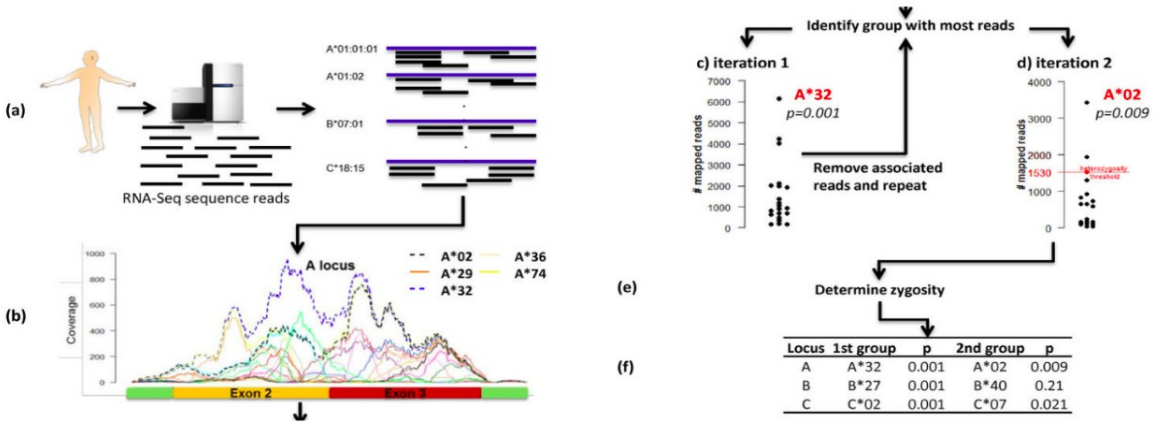
Figura 18 – Processo de identificação de alelos HLA do Optitype. As reads geradas no sequenciamento NGS são alinhadas contra uma referência própria de alelos HLA. A partir do resultado do mapeamento, uma matriz binária é construída com as reads que mapearam em pelo menos um alelo na referência. Com base nessa matriz, um algoritmo é formulado para maximizar o número de reads que define um par de alelos HLA.

alelo é definido de acordo com o maior número de reads alinhadas; para o segundo alelo, essas reads são filtradas e o alelo que tiver mais reads alinhadas é escolhido. Para cada alelo escolhido, um *score* é definido e reflete a probabilidade de que os alelos escolhidos para cada grupo é o correto (Boegel et al. 2012).

O método também considera a zigosidade para cada grupo de alelos preditos. Essa informação influencia no *score* de cada aleloOs alelos com melhor *score* é escolhido como os alelos "vencedores". A Figura 19 mostra o pipeline realizado pelo software para realizar a tipagem dos alelos HLA. O código fonte do método seq2HLA está disponível no github: https://github.com/TRON-Bioinformatics/seq2HLA.

3.3.3.4 SOAP-hla

O software SOAP-hla é uma ferramenta escrita na linguagem de programação Perl para predição de alelos HLA de Classe I e Classe II utilizando dados de exoma de amostras de sangue. A partir de um arquivo no formato BAM contendo as leituras (reads) previamente alinhadas ao genoma humano de referência, o método utiliza as reads alinhadas e não alinhadas com boa qualidade ($\geq$ Q30) para alinhá-las contra um banco de dados de referência de genes do MHC, usando



Fonte: Boegel et al. (2012).

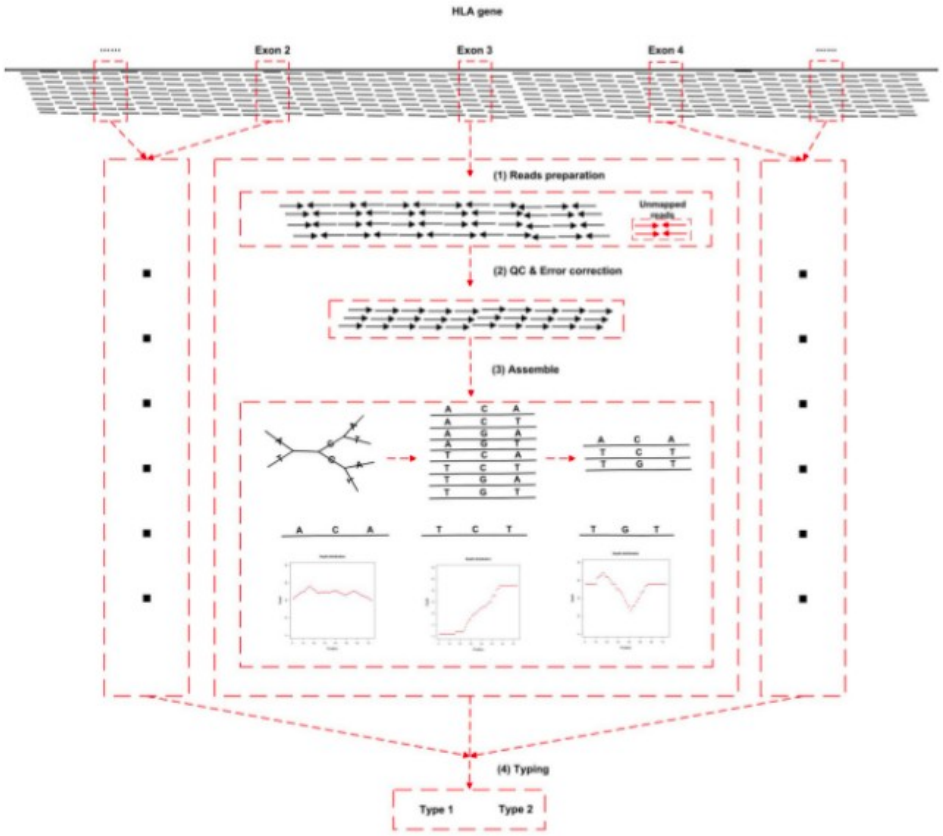
Figura 19 – *Workflow* realizado pelo método seq2HLA para predição de alelos HLA. (a) As reads são mapeadas contra a referência de sequências de HLA. (b) Para cada locus, o alelo com o maior número de reads alinhadas é escolhido. (c) As reads associadas ao primeiro grupo são removidas. (d) O segundo digito do haplótipo é definido. (e) Definição da zigosidade. (f) Os genótipos identificados e o valores de p são gerados.

o software BLAST (Altschul et al. 1990). Os alinhamentos com no máximo dois *mismatches* são mantidos para o passo de montagem dos haplótiposCada haplótipo recebe um *score* calculado pelo algoritmo e os alelos com os maiores escores são mantidos (Cao et al. 2013).

O *output* do método SOAP-hla é uma tabela com os alelos preditos e o *score* associado a cada alelo. A Figura 20 ilustra o pipeline realizado pelo software SOAP-hla para realizar a predição dos alelos dos grupos HLA-A, HLA-B e HLA-C, no caso dos alelos de Classe I. O código fonte do software está disponível no github: https://github.com/adefelicibus/soap-hla, onde o método é disponibilizado através de um pacote que pode ser instalado no ambiente conda pelo canal Bioconda (Grüning et al. 2018),

3.3.4 TCR

Um outro passo importante para identificar neoantígenos como possíveis alvos terapêuticos é verificar a resposta imune ao antígeno. Uma maneira de encontrar essa atividade de células T é através da caracterização do repertório do receptor de células T (TCR, do inglês, *T-Cell Receptor*)O TCR é uma molécula



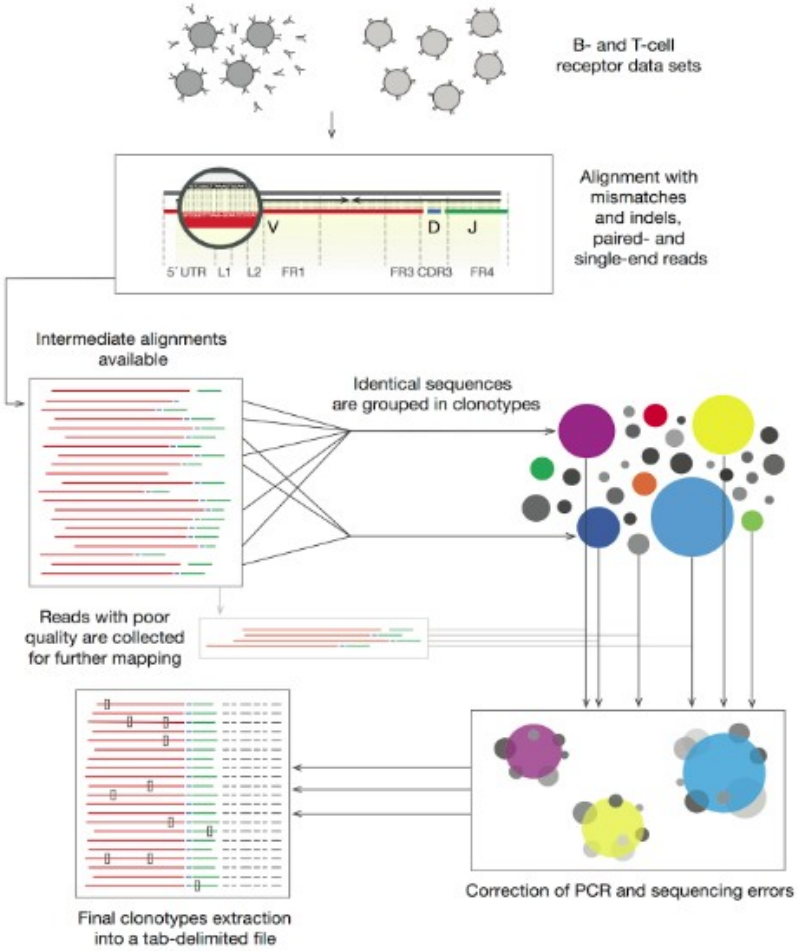
Fonte: Cao et al. (2013).

Figura 20 – Pipeline para predição de alelos HLA realizado pelo método SOAP-hla. As reads mapeadas nas regiões dos alelos HLA e as não mapeadas com alta qualidade são extraídas do arquivo BAM e alinhadas em um banco de dados de sequência de HLA. As reads com menos erros de alinhamento são mantidas para o processo de montagem dos haplótipos e os alelos com maior score de qualidade são escolhidos.

presente na superfície dos linfócitos T e são responsáveis pelo reconhecimento de antígenos ligados ao complexo do MHC.

Neste trabalho, foi escolhido o método MiXCR para identificar o perfil do TCR das amostras e verificar se os antígenos preditos podem ser reconhecidos pelo receptor. O MiXCR (Bolotin et al.2015) é um software escrito na linguagem Java e recebe como *input* arquivos de sequência genômica em formato FASTQ, tanto de RNA quanto de DNA-Seq, e identifica o repertório de TCR e também BCR (do inglês, *B-Cell Receptor*). As análises realizadas com o MiXCR não foram consideradas nos resultados apresentados nesse trabalho, mas esse método está disponível no *software* desenvolvido.A Figura 21 mostra o pipeline de análise

do software MiXCR.



Fonte: Bolotin et al. (2015).

Figura 21 – *Workflow* de análise do software MiXCRA partir das reads de NGS de DNA ou RNA, o software alinha reads nas regiões de interesse e as reads alinhadas são agrupadas em clonotipos e, caso exista erros, são corrigidas com um algoritmo de *clustering*. No final, uma tabela com os clonotipos selecionados é gerada.

3.3.5 TAP

Mesmo que um antígeno possua afinidade com o MHC, essa afinidade pode não ser importante se o processamento do peptídeo não chegar até o MHC. Esse transporte é realizado pela proteína transportadora associada ao processamento de antígeno (TAP, do inglês, *Transporter associated with antigen processing*). Os fragmentos de peptídeos são gerados dos antígenos que estão no citoplasma

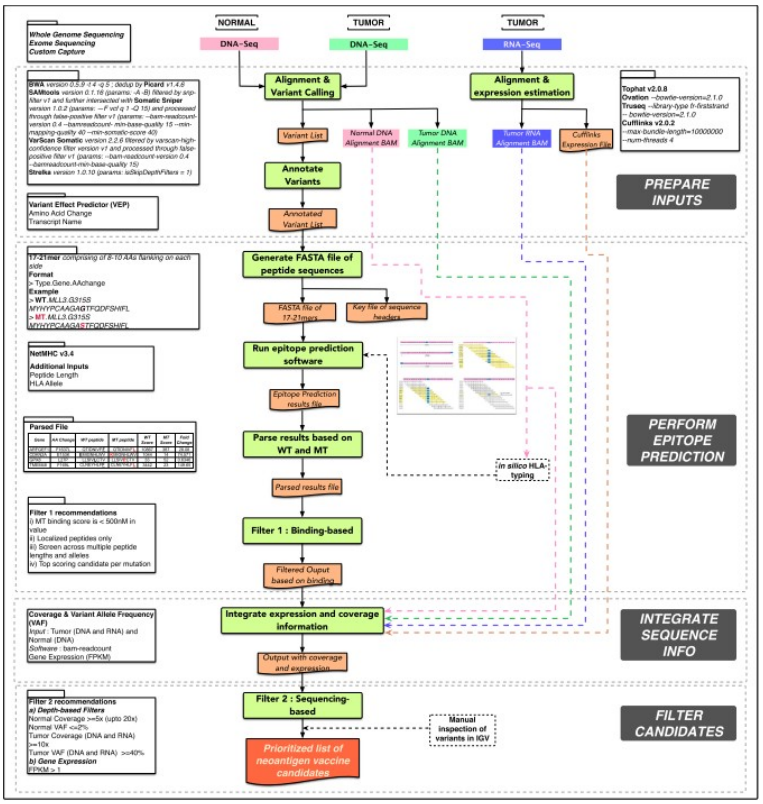
e são transportados pelo TAP até o retículo endoplasmático e apresentados às moléculas do MHC.

Para fazer a predição de afinidade dos peptídeos que podem ser gerados pelos neoantígenos e transportados pelo TAP, o método escolhido para esse projeto foi o NetChop-3.0, que é um algoritmo baseado em rede neural e amplamente utilizado na literatura para predizer os alvos de quebra do peptídeo e a afinidade com o TAP (Keşmir et al. 2002; Nielsen et al. 2005). Esse método recebe uma lista de antígenos e o alelo HLA correspondente, e o *output* do método é um *score* que representa a afinidade do peptídeo para o determinado alelo HLA ser transportado pelo TAP.

3.3.6 pVACseq

Para identificar os neoantígenos, o método escolhido nesse projeto foi o pVAC-Seq, um dos métodos disponíveis no pacote PVACtools (Hundal et al. 2020). A partir das variantes somáticas anotadas com informações do efeito da alteração na proteína resultante, e com uma frequência alélica $\leq 2\%$, o software pVACseq (Hundal et al. 2016) gera diversas combinações possíveis de peptídeos que podem ser gerados por essa variante. Para cada combinação, é calculado um *score* de afinidade com o alelo do MHC. Esse método também integra informações de expressão gênica para realizar o filtro dos neoantígenos preditos e selecionar o algoritmo que verifica a afinidade do alelo ao MHC. Por padrão, no pipeline desenvolvido, o método definido é o NetMHCpan4.1 (Nielsen et al. 2007; Jurtz et al. 2017), por ter apresentado um melhor resultado nos nossos testes (Seção 4.3).

O *output* desse software é uma lista de peptídeos preditos, a sequência de aminoácidos, os scores de predição e a afinidade com o MHC, entre outras informações. Com essa tabela, é possível realizar demais análises, como calcular a afinidade dos peptídeos com o TAP, como descrito na Seção 3.3.5.A Figura 22 demostra o pipeline de execução do software pVAC-Seq.



Fonte: Hundal et al. (2016).

Figura 22 – *Workflow* de análise de identificação de neoantígeno realizado pelo método pVAC-Seq. O pipeline de identificação de neoantígeno do pVACseq possui basicamente três passos: predição dos epítopos, integração de informações de NGS (exoma, RNA-Seq, etc) e filtragem dos neoantígenos candidatos. O processo depende de uma preparação anterior dos arquivos de *input* para ser executado.

3.3.7 Pipeline de identificação de neoantígeno

O pipeline integrado de identificação de neoantígeno criado nesse trabalho foi desenvolvido utilizando como *framework* base o software Snakemake (Köster e Rahmann 2018), que é um sistema escrito em Python para desenvolvimento de *workflows*, e oferece, através de regras escritas com sintaxe Python, um ambiente de execução de processos robusto e com suporte para vários tipos de ambientes, como computador local, supercomputadores (*clusters*) e na nuvem.

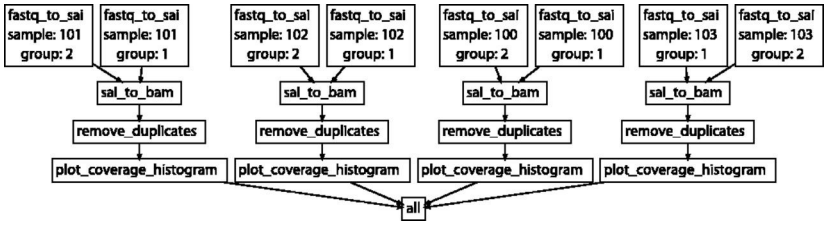
Cada etapa do pipeline é definida em uma regra (*rule*, na sintaxe do Snakemake), e essas regras estão conectadas através do *input* e *output* de cada uma delas. Além dessas opções, as regras possuem várias outras propriedades, como

log, parâmetros e um comando de execução, o qual pode ser a chamada para um software externo ou até mesmo um script R ou Python. As regras desenvolvidas no pipeline utilizam o ambiente Conda ou o Singularity (Kurtzer et al. 2017) para gerenciamento das dependências de software. O código abaixo é um exemplo de uma regra no Snakemake.

```
rule  pvacseq :
    input :
        vcf  = ' results / tmp / vcf /{ x }/{ c }. vcf ',
        allele  = " results / tmp /{ hla }/{ x }/ alleles /{ a }. txt "
    output : directory ( ' results / tmp / pvacseq /{ hla }/{ x }/{ c }/{ a }/ ') ,
    threads : 2
    resources :  mem  = 4
    log : ' results / logs /{ hla }/{ x }.{ c }.{ a }. pvacseq . log '
    params : jobname  = ' pvacseq .{ x } ' ,
        ... ,
    shell :
        ' pvacseq  run  { input . vcf }{ wildcards . x }{ wildcards . a }{
    params . pvacalgorithm }  { output } -e1  { params . epitopes_length }  --
    normal - sample - name { params . normal }-- iedb - install - directory {
    params . mhcpath } { params . pvacparams } '
```

Uma vez que todas as regras do *workflow* estão criadas e conectadas, o Snakemake gera um grafo acíclico demonstrando a relação de dependência de todas as regras, e essa relação é utilizada para determinar a ordem de execução das regras. A partir desse grafo, o Snakemake gerencia a execução de cada regra, independente do ambiente que está sendo utilizado e, em caso de erro, paralisa a execução de todos os processos. A Figura 23 ilustra um grafo de regras do Snakemake.

Uma das vantagens do uso de um gerenciador de *workflow* e a possibilidade de paralelizar a execução das etapas de um pipeline. Como cada regra é um processo que pode ser executado independente, o Snakemake é capaz de gerar vários processos ao mesmo tempo, acelerando a execução de análises com muitas



Fonte: Hundal et al. (2016).

Figura 23 – Grafo de execução de um pipeline de exemplo. Ao ser executado, o Snakemake gera um grafo acíclico que representa o plano de execução. Os nós do grafo são a execução de uma regra e uma conexão direta entre as regras A e B significa que o processo gerado pela regra B depende do *output* da regra A como seu *input*. O caminho do grafo representa a sequência de processos que precisa ser executada de forma serial.

etapas. Para gerenciar os processos em um cluster, o pipeline utiliza o software SLURM (Yoo et al. 2003) como gerenciador de fila. Por esses motivos, e pelo conhecimento prévio em Python, o Snakemake foi escolhido para ser o *framework* utilizado nesse trabalho.

3.4 CARACTERIZAÇÃO DO INFILTRADO INFLAMATÓRIO

A estimação do infiltrado inflamatório foi realizado pelo método CIBERSORTx e foi executado usando um servidor público online: https://cibersortx.stanford.edu/. O CIBERSORTx é um método baseado em deconvolução para a caracterização da composição celular a partir dos perfis de expressão gênica das amostras (Steen et al. 2020), e utiliza uma matriz validada de assinatura de genes de leucócitos chamada de LM22, a qual contêm 547 genes que diferencia 22 tipos de células, incluindo sete tipos de células T, células de memória e plasma, entre outras.

A análise com o CIBERSORTx foi realizada para a caracterização dos valores absolutos e relativos para a composição das células imunes das amostras da coorte de Hugo et al, utilizando os seguintes parâmetros: 1000 permutações, normalização dos quantis desabilitada e correção *batch* aplicada com o modo *B-mode*.

3.5 ANÁLISES ESTATÍSTICAS

As comparações das médias das variáveis quantitativas e grupos foram realizadas utilizando o teste de Wilcoxon. Para correlação entre variáveis quantitativas foram utilizados os métodos Spearman ou Pearson, como apropriado e destacado em cada análise, utilizando a função star_cor do pacote ggpubr do R. Para clusterização no gráfico de *heatmap* foi utilizado o método ward.D2 e a distância manhattan do pacote pheatmap no R.

Para o cálculo da AUC, curva ROC e demais métricas de performance do modelo, foram utilizadas as funções roc e auc com os parâmetros padrões do pacote pROC (Turck et al. 2011).

Para a análise de sobrevida global foi utilizado o estimador de Kaplan-Meier e a comparação entre os grupos foi realizada utilizando o teste de log-rank, ambos do pacote survminer do R.

Todas as análises foram realizadas no ambiente R versão 4.0 e foram utilizados vários pacotes, como ggplo2, cowplot, tidyverse, RColorBrewer, readr, survminer, ggpubr, entre outros.

4 RESULTADOS E DISCUSSÕES

Este capítulo apresenta os resultados e discussões deste projeto.Na Seção 4.1 é detalhado o software GADO utilizado para download dos dados do TCGA; na Seção 4.2 é apresentado o *score* de expressão gênica, bem como a sua validação (Seção 4.2.1). Por fim, as Seções 4.3 e 4.4 apresentam os testes de tipagem de HLA e o pipeline de neoantígeno desenvolvido neste projeto, bem como a aplicação do pipeline em uma coorte com 20 amostras de pacientes com melanoma.

4.1 DOWNLOAD E PROCESSAMENTO DE DADOS PÚBLICOS

Para realizar o download de um grande volume de dados, foi desenvolvida a ferramenta GADO: GDC API DOwnload, utilizando a linguagem de programação Python3. Essa ferramenta utiliza a API (do inglês, *Application Programming Interface*) disponibilizada por três grandes bancos de dados para pesquisa e download de arquivos: GDC[1], ICGC[2] e cBioPortal[3].

A ferramenta GADO é um software de linha de comando e possui várias opções de parâmetros, os quais possibilitam especificar qual o tipo de informação a ser recuperada e qual(is) o(s) projeto(s) de interesse. Além disso, é possível especificar, com mais detalhes, a estrutura de diretórios a ser gerada pelos diversos tipos de arquivos, dando mais liberdade e controle no processamento do que foi baixado. Abaixo tem um exemplo de execução do GADO para download de arquivos do TCGA:

```
# !/ bin / bash
python  gdc_download . py  gdc  " bronchus  and  lung " -c  " transcriptome
    profiling "  -s  RNA - Seq -w  HTSeq - FPKM -- sample_type  tumor  --
    threads  40  -o  output /
```

[1] https://gdc.cancer.gov/developers/gdc-application-programming-interface-api
[2] https://docs.icgc.org/portal/api/
[3] https://docs.cbioportal.org/6.-web-api-and-clients/api-and-api-clients

O comando acima vai pesquisar todos os arquivos de expressão gênica (FPKM) das amostras de tumor do tipo tumoral *bronchus and lung*, o que corresponde aos projetos TCGA-LUAD e TCGA-LUSC, e vai fazer o download de 40 arquivos simultaneamente no diretório *output/*.O Apêndice B apresenta mais detalhes dos parâmetros disponíveis na ferramenta e abaixo tem uma lista com as principais funcionalidades:

- *Download* de arquivos de três bancos de dados:TCGA, ICGC e cBioPortal;
- *Download* de dados de alinhamento (BAM) completo ou por regiões (*slicing*);
- *Download* de vários arquivos em paralelo;
- Várias opções de filtros, como projetos, amostra, tipo ou formato de arquivo;
- Continuação do download de onde parou, em caso de falha de conexão;
- Validação da integridade do arquivo baixado.

Esta ferramenta é utilizada no Laboratório de Bioinformática e Biologia Computacional (LBCB) do Centro Internacional de Pesquisa (CIPE) para *download* e controle de atualizações de informações clínicas e moleculares que são utilizados em vários projetos. Para esse projeto, essa ferramenta foi utilizada para *download* dos dados clínicos e de expressão gênica de todos os 32 projetos do TCGA, o que corresponde a ~11.500 arquivos e ~6GB de dados, e demorou aproximadamente 4 horas.

Atualmente, esse *software* está na versão 1.1.0 e está disponível publicamente no github: https://github.com/adefelicibus/GADO. Novas funcionalidades estão sendo desenvolvidas e serão incorporadas no *software* nas próximas versões, como por exemplo, processamento automático dos arquivos para deixá-los em um formato específico para análises em outros *softwares*, como R.

4.2 *SCORE* DE EXPRESSÃO GÊNICA

Com o objetivo de criar um banco de dados de *score* de expressão gênica, foi aplicado o método desenvolvido nesse projeto (3.2) em 32 projetos do TCGA e, com isso, definiu-se um *threshold* ótimo para cada tipo tumoral, o qual foi utilizado para priorização de neoantígenos nos casos onde não existe informação de RNA-Seq disponível.

A partir dos dados de expressão gênica de 32 projetos do TCGA (3.1) baixados utilizando a ferramenta GADO (4.1), foi calculado o *score* de expressão gênica para cada tipo tumoral. Para cada amostra de cada projeto, um arquivo de duas colunas, contendo o código do gene no formato Ensembl e o FPKM, foi utilizado como entrada (*input*) para o algoritmo de estratificação por níveis de expressão (3.2.1), classificando cada gene em um nível de expressão. A Tabela 6 apresenta um exemplo do resultado do algoritmo.

Tabela 6 – Exemplo de uma matriz resultante do algoritmo de estratificação de nível de expressão gênica. Nesse arquivo, a coluna um é o código do gene em formato Ensembl e a coluna dois é o nível (*level*) de expressão identificado no algoritmo.

	gene_id	**level**
1	ENSG00000000003	medium
2	ENSG00000000005	trace
3	ENSG00000000419	high
...	...	...
N	ENSG00000001084	high

Após a estratificação dos genes de todas amostras e de todos os projetos, foi criada uma matriz para a aplicação do cálculo do *score* de expressão gênica (3.2) por tipo tumoral, utilizando todas as amostras de cada projeto. Como resultado do algoritmo de cálculo do *score*, para cada gene dessa matriz é identificada a contagem de ocorrências para cada nível de expressão, a frequência relativa dessa contagem em relação ao total de amostra e, por fim, o *score*, o qual utiliza, como 'penalidade' no *score*, a média aritmética de cada nível. A Tabela 7 apresenta um exemplo dessa matriz e a Figura 24 ilustra a distribuição do *score* em cada um dos 32 projetos do TCGA.

Tabela 7 – Exemplo de uma matriz com o resultado do cálculo do *score* de expressão gênica. Para cada gene nas linhas, é exibida a contagem de ocorrência para o nível de expressão e, em seguida, a frequência relativa dessa contagem em relação ao número de amostras. Por fim, é apresentado o *score* calculado para cada gene com base nas informações anteriores.

		contagem acumulada por nível					frequência relativa por nível					
	gene_id	high	medium	low	trace	silent	high.freq	medium.freq	low.freq	trace.freq	silent.freq	score
1	ENSG00000000003	161	192	20	2	0	0.429	0.512	0.053	0.005	0	0.093
2	ENSG00000000005	0	1	12	250	112	0	0.002	0.032	0.666	0.298	0.348
3	ENSG00000000419	375	0	0	0	0	1	0	0	0	0	0.091
4	ENSG00000003096	3	28	133	211	0	0.008	0.074	0.354	0.562	0	0.244
5	ENSG00000047648	5	82	151	137	0	0.013	0.218	0.402	0.365	0	0.192
...	...	...	...	...	...	...	...	...	...	...	...	...
N	ENSG00000281921	1	4	28	337	5	0.002	0.010	0.074	0.898	0.013	0.334

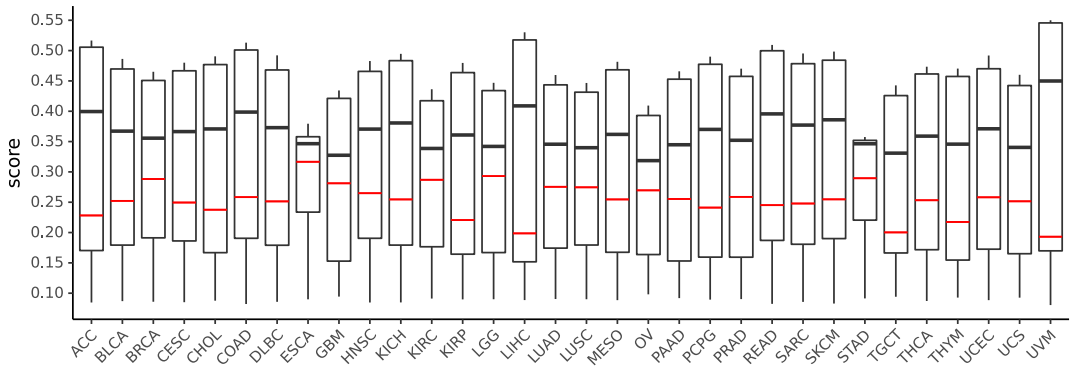


Figura 24 – Distribuição do *score* por projeto do TCGA. Cada projeto do TCGA (eixo x) apresenta uma distribuição do *score* (eixo y) diferente, demonstrando a variabilidade de expressão que existe nos vários tipos tumoraisA linha vermelha destaca o *threshold* identificado para cada projeto (análise detalhada a seguir).

A fim de identificar o melhor ponto de corte (*threshold*) do *score* para selecionar os genes com maior potencial de estar expressos no tipo tumoral, foi realizada uma análise de AUC aplicando o *score* identificado nos genes dos neoantígenos disponíveis do TCGA (3.1.1). Para isso, foram selecionados os potenciais neoantígenos identificados para os 32 projetos do TCGA e utilizados como *ground truth* na análise de AUC.

Como é ilustrado na Figura 25, o resultado da aplicação do *score* apresentou uma AUC diferente para cada projeto. O projeto Carcinoma Endometrial do Corpo Uterino (TCGA-UCEC, do inglês *Uterine Corpus Endometrial Carcinoma*) apresentou uma AUC de 0.91 enquanto o projeto Cistadenocarcinoma Seroso de Ovário (TCGA-OV, do inglês *Ovarian Serous Cystadenocarcinoma*) apresentou uma AUC de 0.59, e os demais projetos apresentaram uma AUC entre 0.6 e 0.8. A Tabela 8 apresenta demais métricas para os projetos com melhores

AUCs.

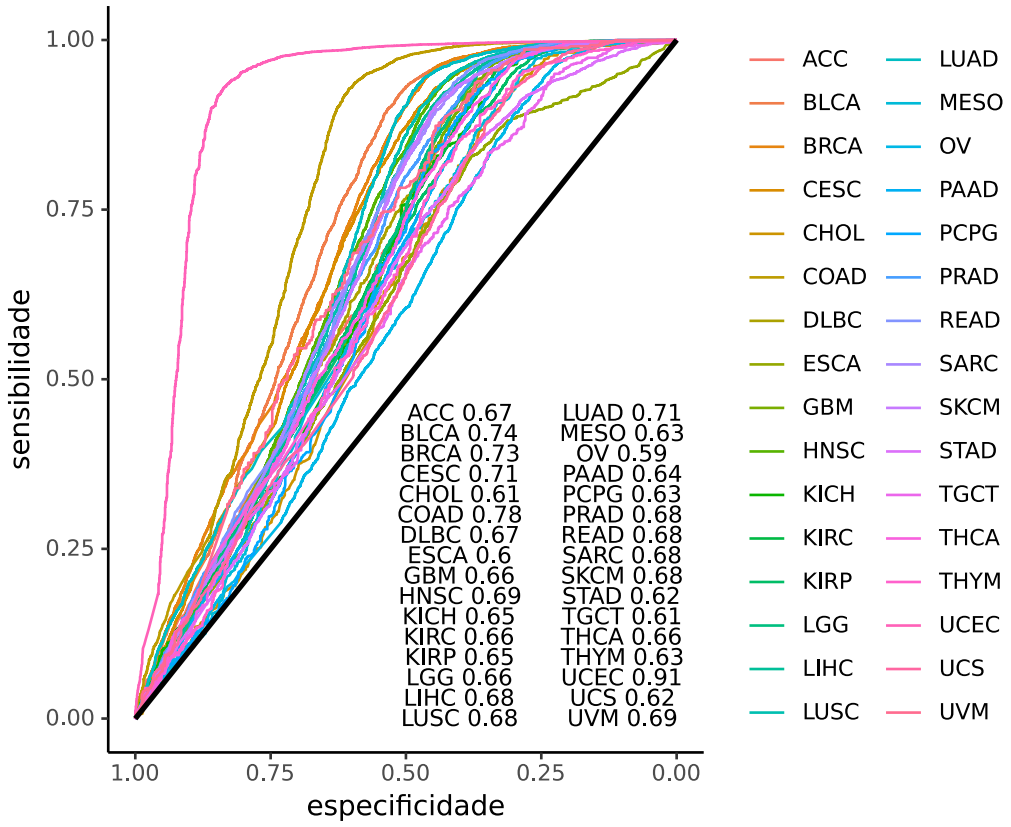


Figura 25 – Análise de AUC do *score* nos projetos do TCGA. Para cada projeto do TCGA, o gráfico apresenta a curva ROC e as métricas de AUC. A lista de projetos está ordenada por ordem alfabética.

Tabela 8 – Métricas de alguns projetos com as melhores AUC. Tabela com algumas métricas da análise de AUC nos projetos do TCGA. Nota-se que o *score* apresenta sensibilidade e precisão altas e um Fscore acima de ~0.8.

Projeto	AUC (CI: 95%)	threshold	sensibilidade	acurácia	precisão	f-score
UCEC	0.91 (0.90-0.92)	0.258	0.930	0.920	0.978	0.953
COAD	0.78 (0.77-0.79)	0.258	0.928	0.861	0.900	0.914
BLCA	0.74 (0.72-0.75)	0.252	0.919	0.822	0.858	0.887
BRCA	0.73 (0.72-0.74)	0.288	0.952	0.816	0.822	0.882
LUAD	0.71 (0.70-0.72)	0.275	0.933	0.833	0.863	0.897
CESC	0.71 (0.70-0.72)	0.249	0.893	0.767	0.795	0.842
LUSC	0.68 (0.67-0.69)	0.275	0.899	0.802	0.853	0.875
SKCM	0.68 (0.66-0.69)	0.254	0.893	0.710	0.693	0.780

É importante destacar que a acurácia do *score* é influenciado pelo número de amostras e também pela carga de neoantígeno do tipo tumoral. No caso do número de amostras, a influência está na frequência relativa e na média de

cada um dos níveis de expressão. Por exemplo, no projeto UCEC, os genes com menores *scores* estão praticamente nos níveis *high* e *medium*, com alguns *outliers* no nível *trace*. Esse mesmo padrão é observado nos tumores COAD e BLCA, porém nos projetos seguintes, como BRCA e LUAD, o número de genes de outros níveis abaixo do *threshold* parece ser maior. A Figura 26 ilustra a distribuição do *score* dos projetos com melhor AUC nos cinco níveis de expressão gênica.

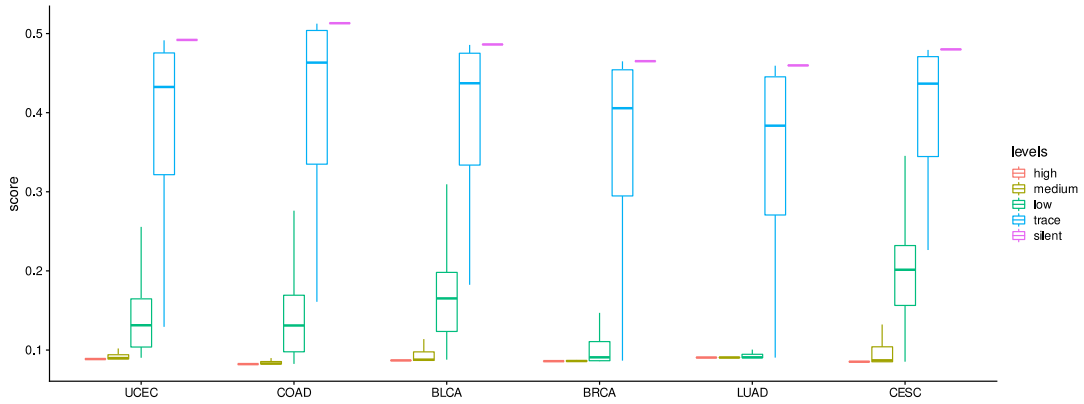


Figura 26 – Distribuição do *score* nos cinco níveis de expressão gênica. Com destaque para os projetos com melhores AUCs (eixo x), essa figura ilustra a distruição do *score* (eixo y) nas cinco classes de níveis de expressão gênica.

Em relação à carga de neoantígenos, outras métricas são influenciadas, como o *threshold* definido na análise de AUC. Como pode-se observar na Figura 27, o ponto de corte até apresenta uma correlação positiva fraca com o F-Score, mas ele não determina o resultado do método, uma vez que os *scores* na faixa de 0.24 a 0.28 apresentam um F-Score que varia de 0.7 a 0.95Portante, se houvesse uma relação direta, poderia ser adotado um único ponto de corte para todos os projetos. Além disso, a Figura 27 ilustra a relação do *threshold* com o F-Score e o valor da AUC para cada tumor.

Essas comparações foram realizadas utilizando os potenciais neoantígenos identificados no projeto PanCancerAtlas do TCGA, e a metodologia utilizada para priorização dos peptídios foram os filtros de afinidade com o MHC (Ic50 $<$ 500) e a expressão gênica (TPM $>$ 1.6). Porém, será que todos os genes com TPM abaixo do valor utilizado realmente não está expresso? E em relação à va-

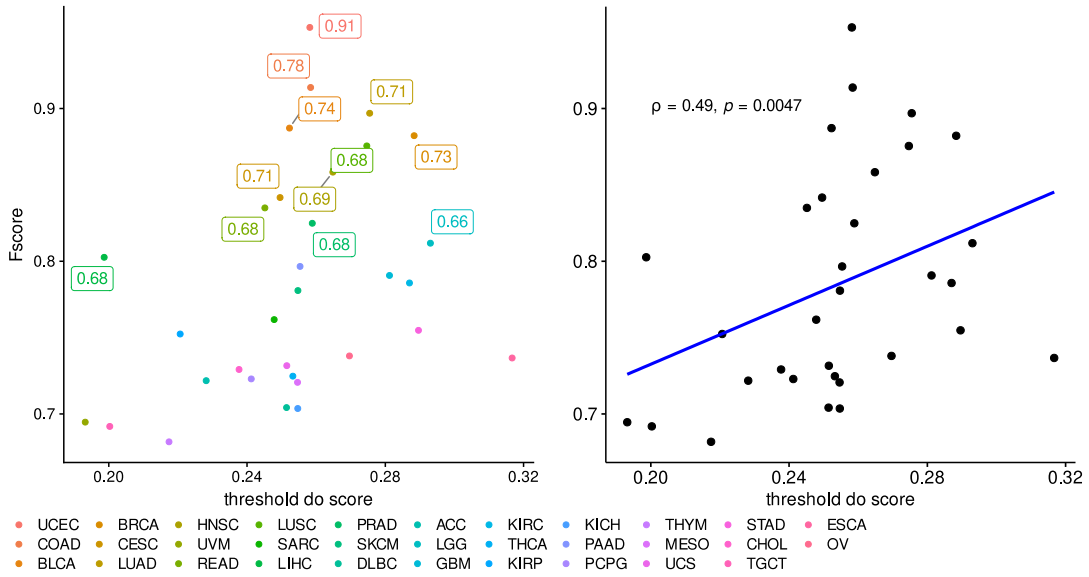


Figura 27 – Relação do *threshold* com a métrica F-Score. No gráfico da esquerda, observa-se que o *threshold* não influencia no F-Score e nem na AUC, mesmo apresentando uma leve correlação positiva (figura da direita), o que evidencia a variabilidade entre os tipos tumorais.

riabilidade genética dos tipos tumorais, ao se utilizar o mesmo valor de expressão para todos os tumores não estaríamos ignorando essa variabilidade? Um teste ideal exige uma outra metodologia de priorização de neoantígeno, que valide a expressão gênica ao invés de um *cutoff* arbitrário, como é utilizado em várias abordagens ao definir o filtro de genes expressos os que apresentam um FPKM $>$ 1, mas não foi encontrado esse tipo de dado.

Portanto, uma limitação é a falta de outro *dataset* como *ground truth* que avalie outros filtros de expressão no processo de identificação de neoantígeno. Entretanto, para validar o *score* em uma coorte diferente do TCGA, o *score* calculado para os projetos TCGA-LUAD e TCGA-LUSC foi aplicado no *dataset* de Karasaki et al. (2017) (3.1.2) e foi avaliada a acurácia do método. O cálculo do *score* nos projetos LUAD e LUSC são abaixo e a análise com o teste do *score* é apresentada na Seção 4.2.1.

O *score* calculado para todos os 32 projetos do TCGA está disponível publicamente no github: https://github.com/adefelicibus/neo2P e, como trabalho futuro, será criada uma ferramenta *online* para a consulta, visualização e

download do *score* por gene ou por projeto.

TCGA-LUAD

Na análise do *score* com os dados de expressão do projeto TCGA-LUAD, o ponto de corte identificado na análise de AUC foi de 0.275 e a AUC identificada foi de 0.71. A Figura 28 ilustra a distribuição da frequência relativa por nível de expressão (superior esquerdo), um histograma do *score* calculado para todos os genes (inferior esquerdo), um ECDF[4] (do inglês, *Empirical cumulative distribution function*) do *score* com destaque para o *threshold* identificado (superior direito) e a curva ROC do projeto.

O *score* também foi calculado considerando algumas informações clínicas, como o estadiamento, subtipo imune e subtipo molecular. A Figura 29 ilustra o cálculo do *score* estratificando por estadiamento.

Observa-se que a estratificação do cálculo do *score* por estadiamento não adicionou nenhuma melhora na curva ROC, muito pelo contrário, indicando que essa característica não influencia a expressão gênica. Esse mesmo resultado foi observado em outras duas características clínicas testadas, como o subtipo imune e o subtipo molecular. Alguns estudos indicam que a idade e sexo interferem na expressão de uma porcentagem dos genes em diferentes órgãos (Oliva et al. 2020), e também entre diferentes populações (Li et al. 2010), mas nesse trabalho essas variáveis não foram testadas.

TCGA-LUSC

Na análise do *score* com os dados de expressão do projeto TCGA-LUSC, o ponto de corte identificado na análise de AUC também foi de 0.275 e a AUC identificada foi de 0.68. A Figura 30 ilustra a distribuição da frequência relativa por nível de expressão (superior esquerdo), um histograma do *score* calculado

[4] Esse gráfico ilustra a distribuição da frequência acumulada de uma variável.

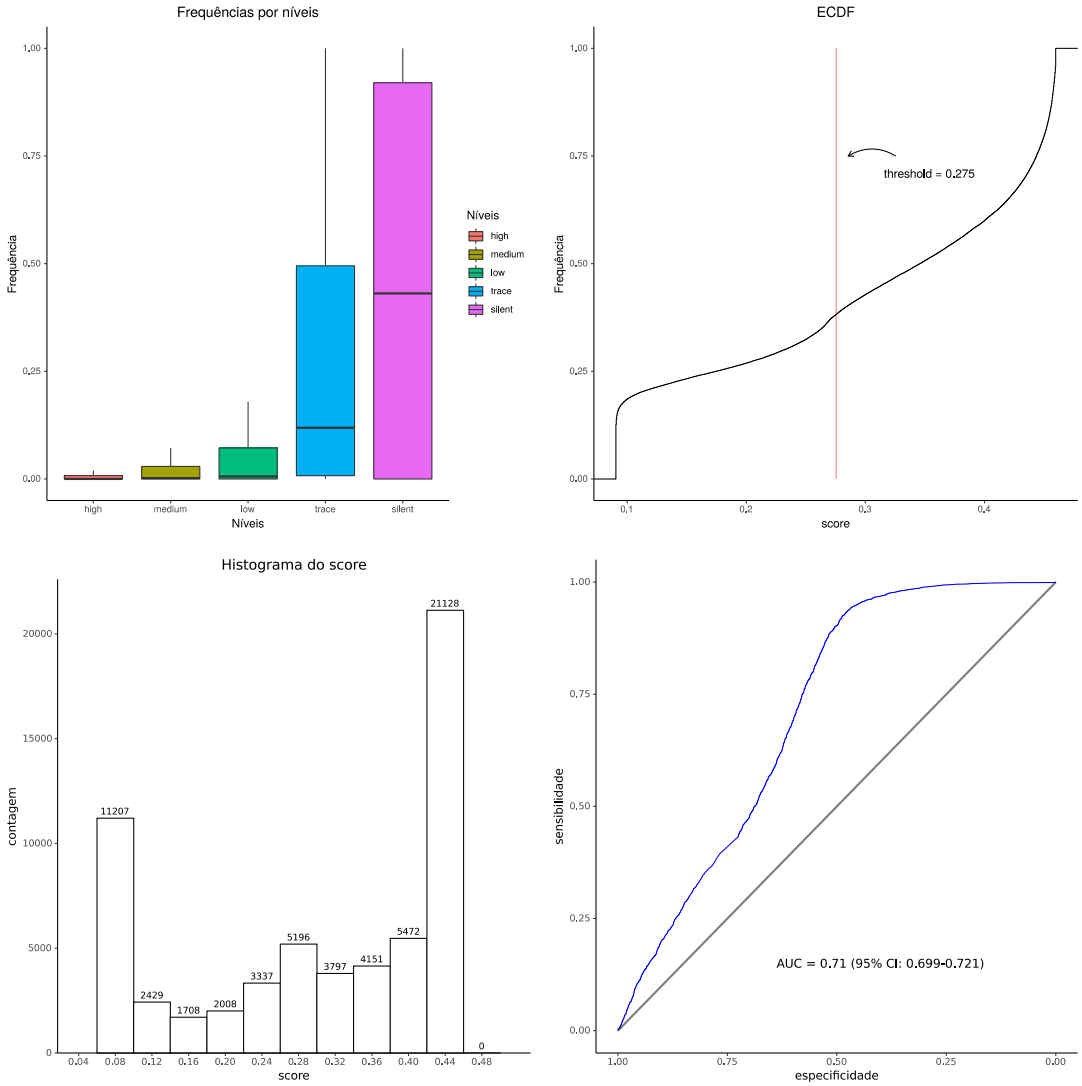


Figura 28 – Análise do cálculo do *score* para o projeto TCGA-LUAD. O primeiro gráfico à esquerda ilustra a distribuição da frequência relativa para cada nível de expressão, onde os níveis *high*, *medium* e *low* apresentam as frequências mais baixas, o contrário do nível *silent*. Abaixo, tem o histograma de ocorrências por valor de *score*, mostrando que a maior parte dos genes possuem *score* alto, e à direita tem o ECDF do *score*, indicando que 50% dos genes tem um *score* de ~0.35. Por fim, abaixo tem o gráfico com a curva ROC e a AUC de 0.71 identificada para esse tipo tumoral.

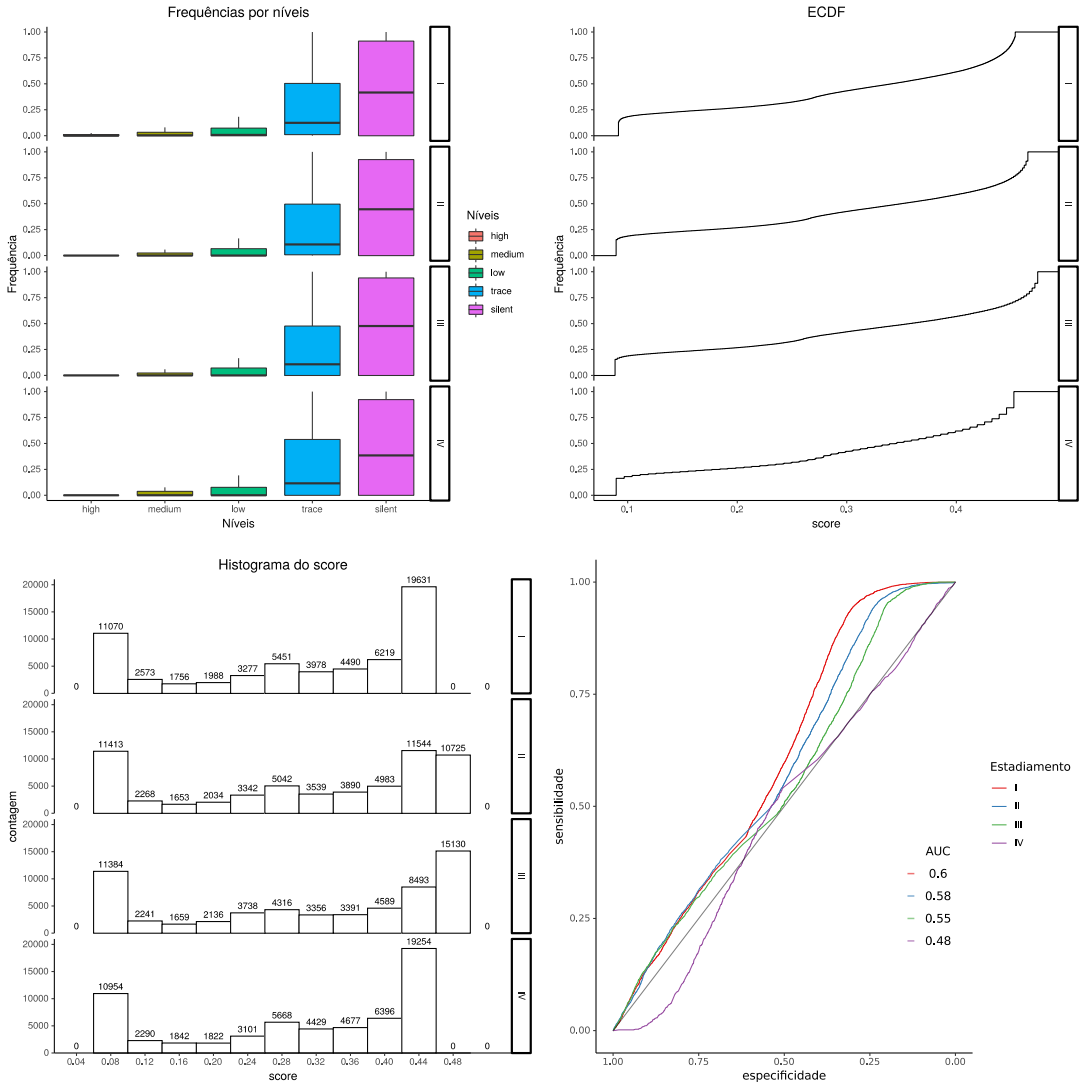


Figura 29 – Análise do cálculo do *score* para o projeto TCGA-LUAD por estadiamento. O primeiro gráfico à esquerda ilustra a distribuição da frequência relativa para cada nível de expressão e estadiamento, onde os níveis *high* e *medium* apresentam as frequências mais baixas, o contrário do nível *silent*. Abaixo, tem o histograma de ocorrências por valor de *score* e estadiamento, mostrando que a maior parte dos genes possui *score* alto, e à direita tem o ECDF do *score* por estadiamento. Por fim, abaixo tem o gráfico com uma curva ROC e uma AUC para cada estadiamento.

para todos os genes (inferior esquerdo), um ECDF do *score* com destaque para o *threshold* identificado (superior direito) e a curva ROC do projeto.

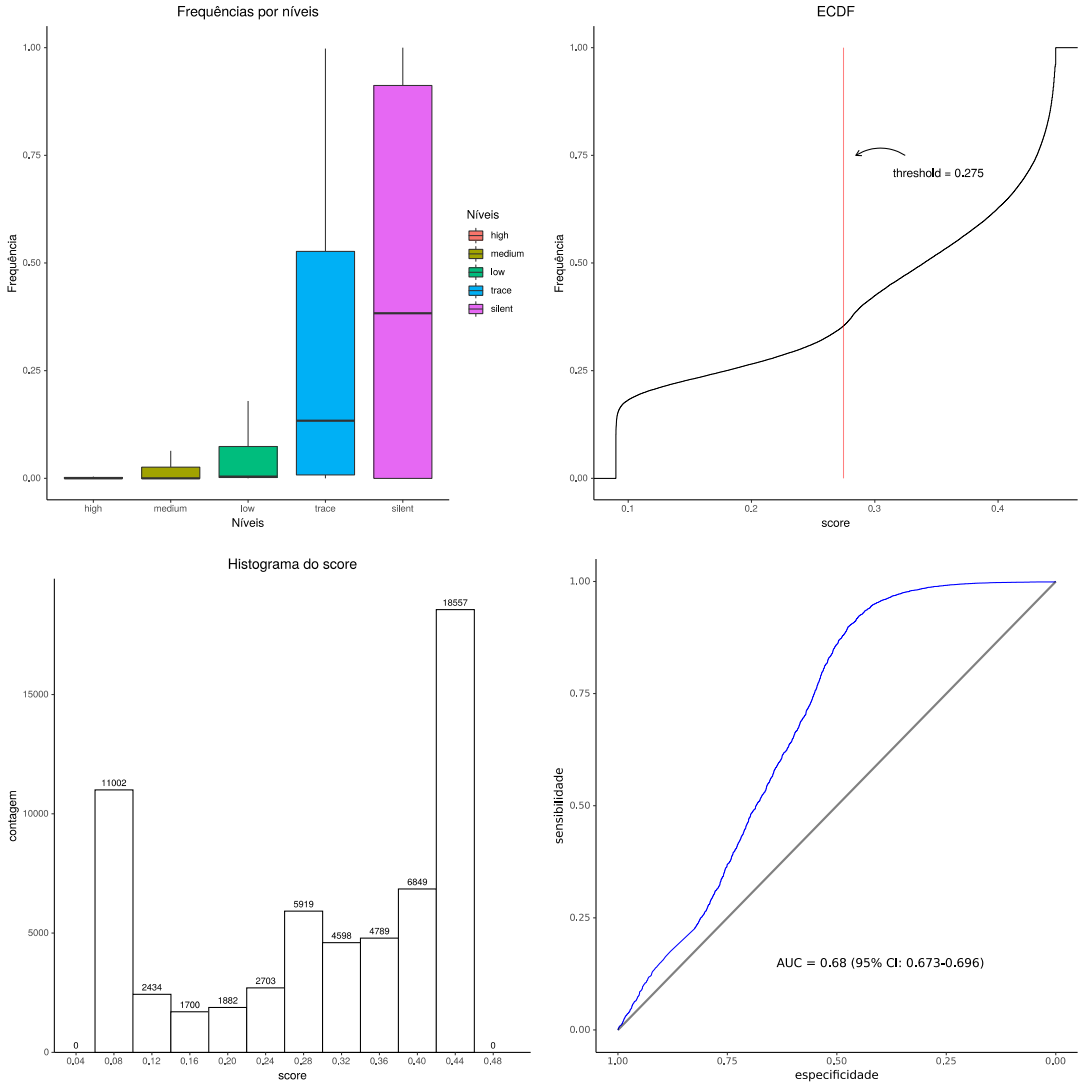


Figura 30 – Análise do cálculo do *score* para o projeto TCGA-LUSC. O primeiro gráfico à esquerda ilustra a distribuição da frequência relativa para cada nível de expressão, onde os níveis *high* e *medium* apresentam as frequências mais baixas, o contrário dos níveis *trace* e *silent*. Abaixo, tem o histograma de ocorrências por valor de *score*, mostrando que a maior parte dos genes possuem *score* alto, e à direita tem o ECDF do *score*, indicando que 50% dos genes tem um *score* de ~0.32. Por fim, abaixo tem o gráfico com a curva ROC e a AUC de 0.68 (IC: 0.67-0.69) identificada para esse tipo tumoral.

O *score* também foi calculado considerando algumas informações clínicas, como o estadiamento, subtipo imune e subtipo molecular. A Figura 31 ilustra a análise do cálculo do *score* estratificando por estadiamento.

Assim como no projeto TCGA-LUAD, a informação do estadiamento clínico não melhorou a análise, apresentando uma AUC inferior em relação à en-

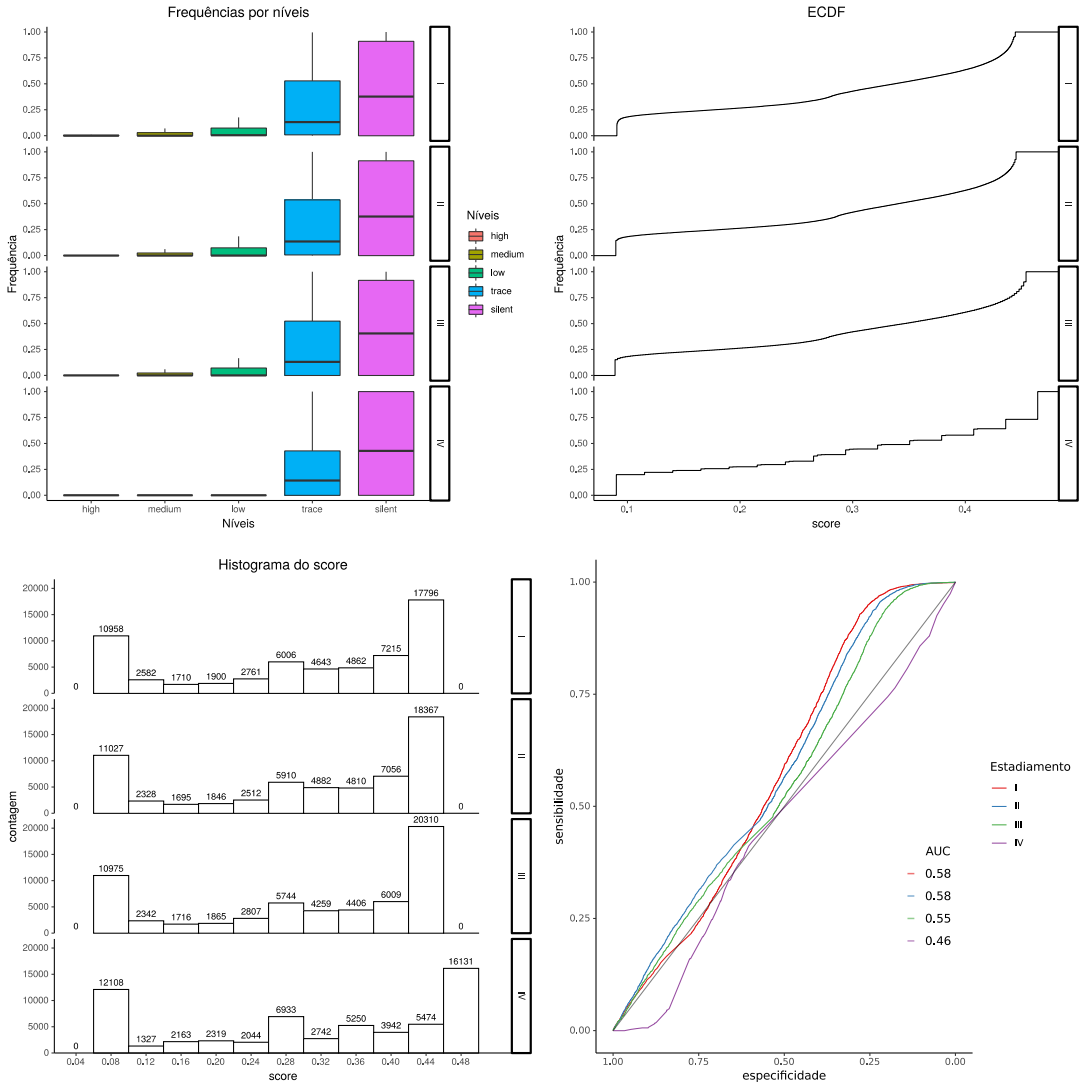


Figura 31 – Análise do cálculo do *score* para o projeto TCGA-LUSC por estadiamento. O primeiro gráfico à esquerda ilustra a distribuição da frequência relativa para cada nível de expressão e estadiamento, onde os níveis *high*, *medium* e *low* apresentam as frequências mais baixas, o contrário do nível *silent*. Abaixo, tem o histograma de ocorrências por valor de *score* e estadiamento, mostrando que a maior parte dos genes possuem *score* alto, e à direita tem o ECDF do *score* por estadiamento. Por fim, abaixo tem o gráfico com uma curva ROC e uma AUC para cada estadiamento.

contrada sem a estratificação. As análises do cálculo das outras características clínicas são apresentadas no Apêndice C.

4.2.1 Validação do *score* de expressão

O *score* de expressão gênica identificado com os dados do TCGA foi aplicado em um *dataset* diferente para avaliar a sua acurácia.O *dataset* de Karasaki et al. (2017) contém 4 amostras de pacientes com câncer de pulmão (2) e informações clínicas e moleculares, além da lista de neoantígenos candidatos identificados em três abordagens diferentes relacionando as vantagens da integração de RNA-Seq no processo de priorização de neoantígenos.

Para cada uma das abordagens apresentadas por Karasaki et al. (2017), foi verificada a acurácia do *score* de expressão gênica em filtrar os genes com *score* abaixo do *threshold* dos projetos LUAD e LUSC e comparado aos filtros realizados nas abordagens II a IV. O objetivo dessa validação foi verificar qual a contribuição do *score* na priorização de neoantígeno quando não se tem a informação de expressão gênica disponível. A Tabela 9 apresenta as métricas dessa validação separadas por metodologias.

Tabela 9 – Métricas de validação do *score* no *dataset* de Karasaki et al. (2017). A aplicação do *score* apresentou uma acurácia boa nas amostras para a abordagem II e um FScore acima de 0.8. Para as demais abordagens, o *score* não apresentou métricas satisfatórias.

	II					III					IV				
amostra	acurácia	precisão	sensib.	especif.	fscore	acurácia	precisão	sensib.	especif.	fscore	acurácia	precisão	sensib.	especif.	fscore
LK029	0.865	0.816	1	0.667	0.898	0.596	0.447	1	0.4	0.618	0.635	0.5	1	0.424	0.667
LK047	0.706	0.630	1	0.412	0.773	0.382	0.222	1	0.25	0.364	0.382	0.222	1	0.25	0.364
LK070	0.791	0.707	0.976	0.614	0.82	0.589	0.396	1	0.435	0.568	0.612	0.431	1	0.45	0.602
LK073	0.778	0.756	1	0.286	0.861	0.511	0.463	1	0.154	0.633	0.511	0.463	1	0.154	0.633

O *score* apresentou uma sensibilidade muita alta para todas as amostras na análise e em todas as abordagens, com muitos poucos genes falso positivos. Porém, a especificidade ficou muito abaixo nas amostras LK047 e LK073 na abordagem II e em todas as amostras nas demais abordagens. De forma geral, para a abordagem II, que é a que utiliza o filtro de FPKM, o método apresentou uma acurácia alta, maior do que 0.7 para todas as amostras, não sendo maior por conta da presença de alguns genes com expressão abaixo de 1 FPKM mas com

score acima do *threshold* de 0.275, como ilustrado na Figura 32 com a distribuição do FPKM das amostras em relação ao *score*.

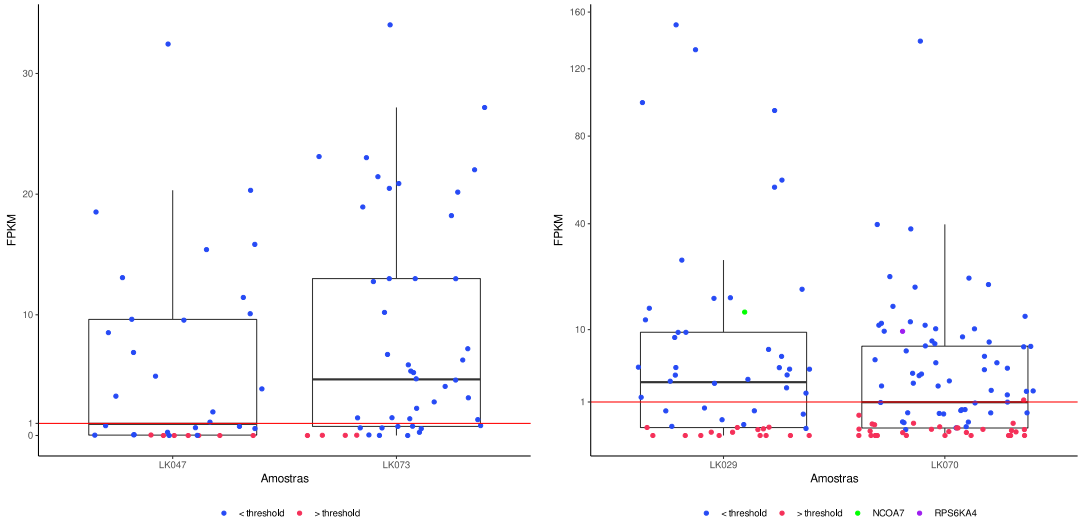


Figura 32 – Distribuição do FPKM das amostras em relação ao *score*. Cada ponto do gráfico é um gene e os pontos em azul são aqueles com valor do *score* abaixo do *threshold*, enquanto que os pontos em vermelho são os com *score* acima. Os pontos verde e roxo correspondem a dois genes onde foram identificados neoantígenos com resposta citotóxicas (Karasaki et al. 2017).

Entretanto, ao analisar os genes que sobrepõem as abordagens e os selecionados pelo filtro do *score*, é possível observar que nas amostras LK029 e LK073, quase 50% dos genes em comuns nas três abordagens estão incluídos no filtro do *score* e somente a amostra LK070 teve dois genes excluídos pelo filtro. As Figuras 33 e 34 ilustram o *overlap* entre os genes selecionados pelas abordagens II, III e IV e o filtro do *score* para os projetos LUAD e LUSC. Como observado na Tabela 9, o filtro do *score* se mostrou mais permissivo do que as demais abordagens, o que não é exatamente um problema, visto que o filtro de expressão é um dos filtros que são aplicados no processo de priorização de neoantígeno, mas também é uma oportunidade para futuras melhorias no cálculo do *score* de expressão gênica.

O *score* de expressão gênica estratificando por estadiamento também foi aplicado nas quatro amostras e foi verificado a acurácia do método. Exceto para a métrica de sensibilidade para a amostra LK070 que apresentou uma melhora, para todas as amostras em todas as abordagens as demais métricas não apresentaram

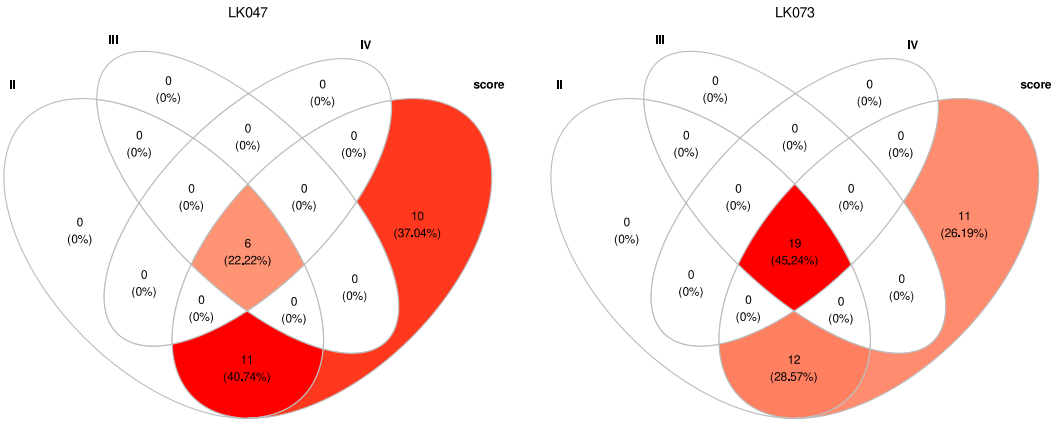


Figura 33 – Diagrama de Venn com os genes selecionados pelo *score* e pelas três abordagens realizadas por Karasaki et al no projeto LUAD. Para a amostra LK047, o *score* apresentou 37% de genes exclusivos, 40% de *overlap* com a abordagem II e 22% com todas as abordagens juntas. Já para a amostra LK073, o filtro do *score* apresentou 26% de genes exclusivos, 28% de genes em comum com a abordagem II e 45% com todas as abordagens.

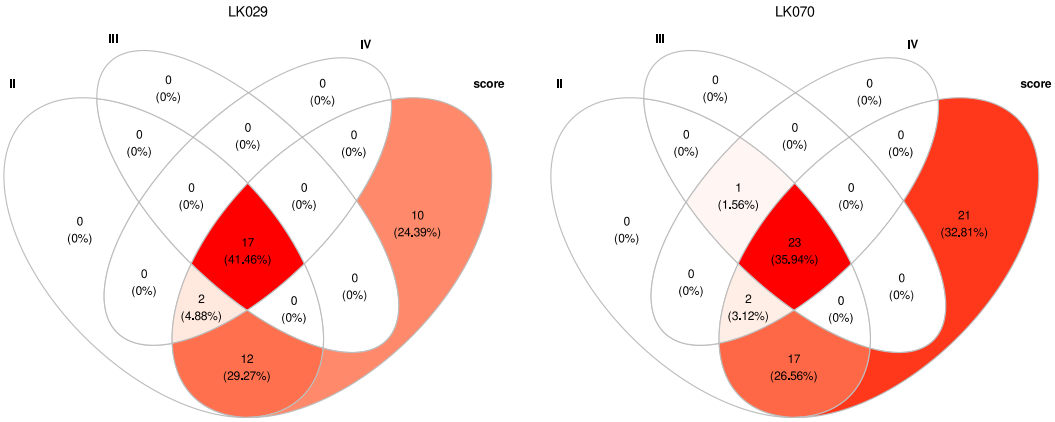


Figura 34 – Diagrama de Venn com os genes selecionados pelo *score* e pelas três abordagens realizadas por Karasaki et al no projeto LUSC. Para a amostra LK029, o *score* apresentou 24% de genes exclusivos, 29% de *overlap* com a abordagem II e 41% com todas as abordagens juntas. Já para a amostra LK070, o filtro do *score* apresentou 32% de genes exclusivos, 26% de genes em comum com a abordagem II e 35% com todas as abordagens.

melhoras ao utilizar o *score* estratificado por estadiamentoA Tabela 10 sumariza as métricas dessa análise.

Tabela 10 – Métricas de validação do *score* no *dataset* de Karasaki et al. (2017) por estadiamento. A aplicação do *score* por estadiamento não apresentou melhor acurácia para essa análise, com exceção da métrica sensibilidade para a amostra LK070.

		II					III					IV				
amostra	estad.	acurácia	precisão	sensib.	espec.	fscore	acurácia	precisão	sensib.	espec.	fscore	acurácia	precisão	sensib.	espec.	fscore
LK029	I	0.808	0.756	1	0.524	0.861	0.538	0.415	1	0.314	0.586	0.577	0.463	1	0.333	0.633
LK047	III	0.706	0.630	1	0.412	0.773	0.382	0.222	1	0.25	0.364	0.382	0.222	1	0.25	0.364
LK070	II	0.744	0.656	1	0.5	0.792	0.529	0.365	1	0.355	0.535	0.553	0.397	1	0.367	0.568
LK073	I	0.756	0.738	1	0.214	0.849	0.489	0.452	1	0.115	0.623	0.489	0.452	1	0.115	0.623

Como conclusão parcial, o *score* apresenta uma boa acurácia ao filtrar os genes com maior potencial de expressão, possibilitando a priorização de neoantígeno nos casos onde não existe a informação de expressão gênica das amostras estudadas. Além disso, existem duas oportunidades de otimização do cálculo do *score*, que são o acréscimo de amostras de outros bancos de dados públicos, como o cBioPortal e ICGC, e a otimização do peso aplicado aos níveis de expressão, com a utilização de algoritmo genético, por exemplo.

4.3 *BENCHMARKING* DOS MÉTODOS DE TIPAGEM DE HLA

Foi desenvolvido, neste trabalho, um *pipeline* para a realização da tipagem do HLA utilizando amostras de DNA, tumoral e normal, e os resultados obtidos de cada *software* foram comparados com a tipagem realizada no estudo (Hugo et al. 2016) (ver Seção 3.1.3). Nessa etapa, os softwares Optitype (Seção 3.3.3.2, Polysolver (Seção 3.3.3.1), soap-HLA (Seção 3.3.3.4) e seq2HLA (Seção 3.3.3.3) foram executados utilizando os parâmetros padrões em seis pares de amostras, tumor e normal, escolhidas aleatoriamente.

Os resultados deste teste mostraram uma concordância alta entre os métodos e com a referência, mas ainda assim possuem vários erros de tipagem. Em relação à acurácia, o método Optitype identificou, para as 6 amostras, todos os HLAs corretamente, tanto em relação a 2 dígitos como para 4 dígitos, utilizando as amostras de sangue. Entretanto, ao utilizar as amostras tumorais, o mesmo método apresentou a acurácia de 97.22%, indicando que as variantes, no con-

texto somático, podem adicionar um efeito negativo no processo de predição de HLA e o método ainda é sensível a essas variações, mas ainda apresentando uma acurácia alta. Os demais métodos apresentaram erros independente do tipo de amostra.

Apesar dos métodos apresentarem erros no processo de tipagem dos HLAs, todos eles apresentaram uma concordância alta nos resultados. Alguns aspectos podem estar relacionados com essa diferença nos métodos e com os erros apresentados, como a cobertura da região dos HLAs nas amostras, a taxa mutacional, o contexto das mutações, entre outros. Esses aspectos não foram avaliados nesse trabalho, uma vez que não existia o objetivo identificar possíveis falhas no processo de tipagem, mas sim escolher um método para o pipeline.Outros trabalhos já avaliaram esses aspectos com alguns dos métodos que foram testados neste *benchmarking* (Kiyotani et al. 2017; Matey-Hernandez et al. 2018). A Figura 35 ilustra a concordância entre os métodos executados.

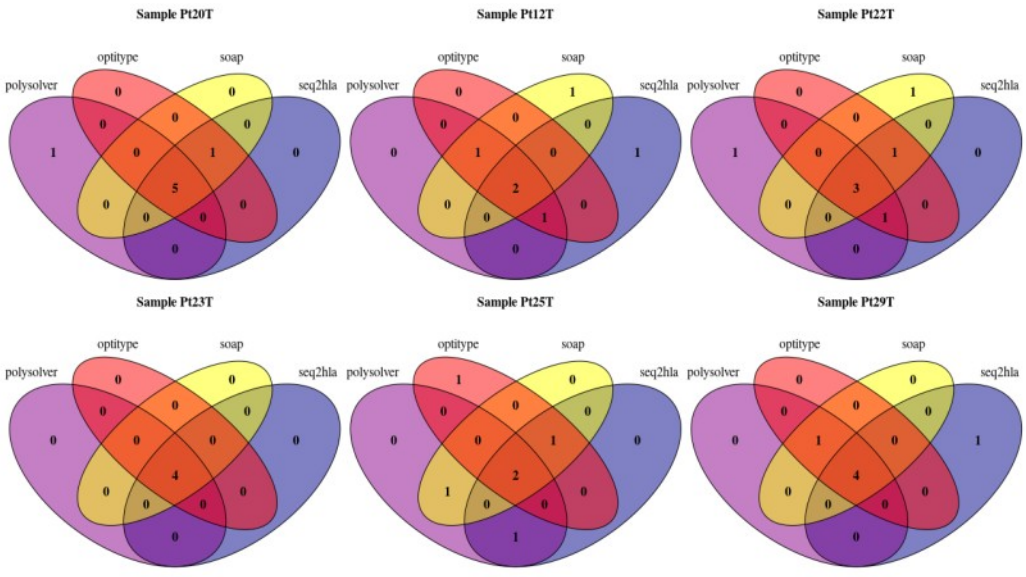


Figura 35 – Concordância entre os métodos de *hla-typing* utilizando amostras tumorais. Os resultados preliminares sugerem que existe um consenso entre os algoritmos e os erros de cada um podem ser filtrados selecionando o alelo mais frequentemente predito.

Conforme descrito por Bauer et al. (2018), a união das diferentes abordagens dos softwares poderia favorecer a criação de um método *ensemble* com alta acurácia na predição de HLA. Essa hipótese foi testada para verificar se é possível

melhorar a acurácia utilizando essa abordagem e entender os erros apresentados pelos softwares. Além disso, se esse método apresentar uma acurácia maior, ele pode ser utilizado em amostras tumorais de exoma. Para isso, foram utilizadas amostras do projeto 1000 Genomes (ver Seção 3.1.4).

A partir dos arquivos FASTQ do projeto 1000G, os quatro softwares foram executados para cada amostra, utilizando os parâmetros padrões, em paralelo, em um cluster Linux com 7 nós de processamento, totalizando 472 CPUs e 608GB de RAM através de um gerenciador de fila SLURM (Yoo et al. 2003). Após a execução dos quatro softwares para todas as amostras, foi realizada a comparação da acurácia de cada método utilizando a tipagem realizada com Sanger e demais análises.

Os quatro métodos foram escolhidos para criarmos um método *ensemble* com uma acurácia melhor do que um único método. Cada método utiliza abordagens diferentes, desde o processo de alinhamento até o algoritmo para escolha do melhor alelo. Com essas diferenças, enquanto um método é mais sensível para um grupo de HLA, outro método pode apresentar uma acurácia melhor, e a união desses métodos pode apresentar uma correção no processo de predição dos alelos do HLA. A Figura 36 mostra a diferença entre os métodos. Para os métodos Polysolver e SOAP que recebem como *input* um arquivo BAM, foi utilizado o software BWA (Li e Durbin 2009; Li 2013) para alinhamento contra o genoma de referência na versão hg19, e o software GATK3 (McKenna et al. 2010) para pré-processamento e recalibração das bases alinhadas.

O resultado de cada método foi comparado com a tipagem realizada por Sanger, para verificar a acurácia de cada método. Como cada grupo de HLA tem características distintas, a acurácia foi separada por grupo de HLA (HLA-A, HLA-B e HLA-C) e a acurácia geral (HLA-ABC). O método Optitype apresentou a melhor acurácia em todos os grupos e no geral, seguido pelo método Polysolver. Os métodos seq2HLA e SOAP apresentaram muitos erros no processo de predição, principalmente no grupo HLA-B. O software seq2HLA foi originalmente

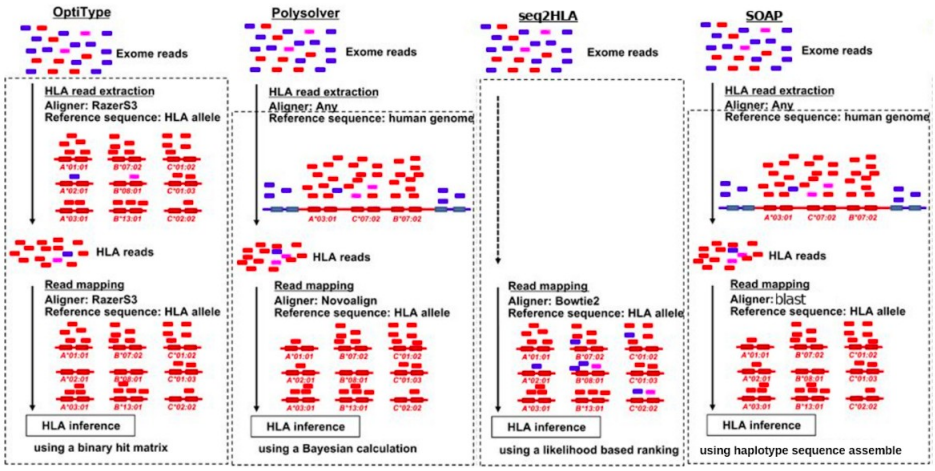
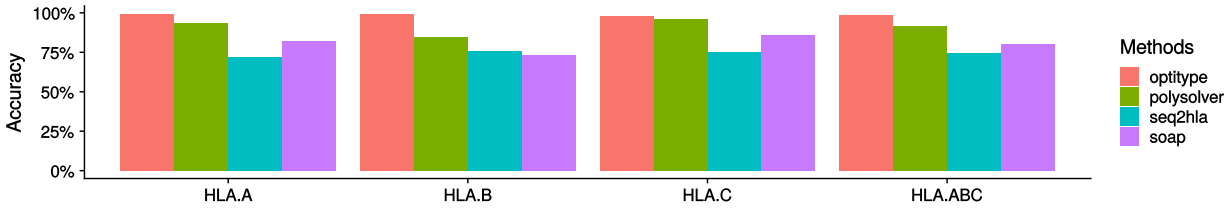


Figura 36 – Comparação entre os quatro métodos utilizados para predição de HLA. Os métodos utilizados diferem em praticamente todas as etapas de identificação dos alelos HLA.

desenvolvido para predição de HLA utilizando RNA-Seq, mas já foi utilizado com dados de DNA, o que pode indicar a baixa acurácia (Bauer et al. 2018). A Figura 37 mostra a acurácia de cada um dos métodos para os grupos de HLA, e a Figura 38 mostra quais são os alelos com mais erros para cada método.



class	optitype	polysolver	seq2hla	soap
HLA.A	0.9891304	0.9347826	0.7173913	0.8206522
HLA.B	0.9945652	0.8478261	0.7582418	0.7292818
HLA.C	0.9782609	0.9565217	0.7500000	0.8586957
HLA.ABC	0.9873188	0.9130435	0.7418777	0.8028765

Figura 37 – Acurácia de cada método separado por grupos de alelos.

Para uma análise dos erros identificados pelos métodos, foi escolhido o método Optitype, por ter apresentado a melhor acurácia. O alelo com mais erros foi o HLA-C*07:18 que foi incorretamente definido como HLA-C*07:01.Ao verificar as sequências desses alelos, observa-se uma semelhança muito grande nas bases e no alinhamento das amostras. Esses alelos possuem cerca de 720 bases de tamanho, e somente 6 bases são diferentes entre eles. Isso demonstra a

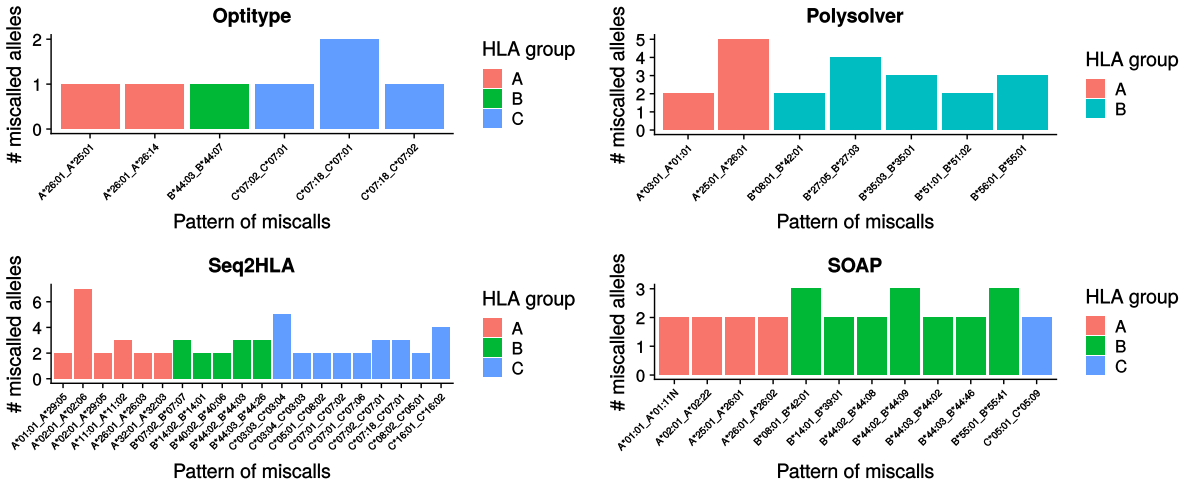


Figura 38 – Frequência de alelos com mais erros por método.

dificuldade e limitação dos métodos de alinhamento de alinhar as sequências com o HLA correto.

Uma conclusão parcial é que a informação de que alguns alelos são muito similares poderia ser utilizada pelos métodos de predição de HLA ou, em um futuro próximo, com a melhoria das tecnologias de sequenciamento genômico sendo capazes de gerar sequências cada vez mais longas, esse tipo de erro pode ser evitado (Amarasinghe et al. 2020). A Figura 39 mostra uma imagem do software IGV (Robinson et al. 2011) com os alinhamentos de sequências da amostra HG00253 e os alelos citados acima. A Figura 40 mostra as bases desses mesmos alelos, com destaque para as diferenças entre eles.

Em relação aos outros erros apresentados pelo Optitype, os alelos HLA-A*26:14 e HLA-B*44:07 apresentam diferenças na frequência encontrada na população Britânica. Os alelos identificados erroneamente para ambos os alelos mencionados, os alelos HLA-A*26:01 e HLA-B*44:03, respectivamente, apresentam uma frequência maior na população de origem da amostra. Como mostrado por (Matey-Hernandez et al. 2018), a informação referente à população e à frequência alélica poderia ser um fator a ser adicionado no processo de predição de HLA. A Figura 41 ilustra a frequência alélica dos HLAs corretos, HLA-A*26:14 e HLA-B*44:07 e dos incorretos, HLA-A*26:01 e HLA-B*44:03 com destaque para

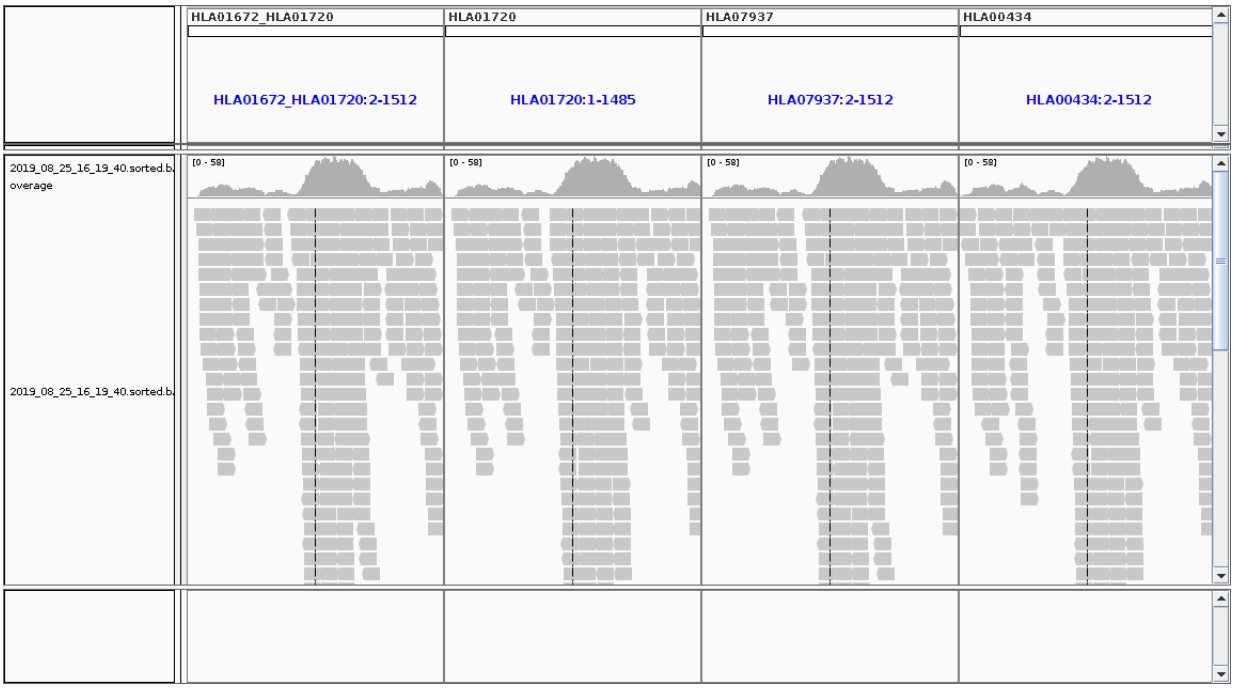


Figura 39 – Comparação dos alinhamentos das reads da amostra HG00253 em quatro alelos diferentes. É possível comparar a cobertura da amostras nas regiões dos alelos X, Y, Z e as reads alinhadas.

```
HLA07937              TGAGTGCGGGGTTGGGAGGGAAGCGGCCTCTATGGAGAGGAGCGAGGGGCCCGGCCCGGCG     60
HLA00434              TGAGTGCGGGGTTGGGAGGGAAGCGGCCTCTGCGGAGAGGAGCGAGGGGCCCTCCCGGCG     60
HLA01672_HLA01720     TGAGTGCGGGGTTGGGAGGGAAGCGGCCTCTGCGGAGAGGAGCGAGGGGCCCGGCCCGGCG     60
HLA01720              TGAGTGCGGGGTTGGGAGGGAAGCGGCCTCTGCGGAGAGGAGCGAGGGGCCCGGCCCGGCG     60
                      ********************************* ****************** *******

HLA07937              AGTATTGGGACCGGGAGACACAGAACTACAAGCGCCAGGCACAGGCTGACCGAGTGAGCC     360
HLA00434              AGTATTGGGACCGGGAGACACAGAAGTACAAGCGCCAGGCACAGGCTGACCGAGTGAGCC     360
HLA01672_HLA01720     AGTATTGGGACCGGGAGACACAGAACTACAAGCGCCAGGCACAGGCTGACCGAGTGAGCC     360
HLA01720              AGTATTGGGACCGGGAGACACAGAACTACAAGCGCCAGGCACAGGCTGACCGAGTGAGCC     360
                      ************************* ********************************

HLA07937              CGCAGGTCACGACCCCTCCCCATCCCCCACGGACGGCCCGGGTCGCCCCGAGTCTCCCCG     480
HLA00434              CGCAGGTCACGACCCCTCCCCATCCCCCACGGACGGCCCGGGTCGCCCAGAGTCTCCCCG     480
HLA01672_HLA01720     CGCAGGTCACGACCCCTCCCCATCCCCCACGGACGGCCCGGGTCGCCCAGAGTCTCCCCG     480
HLA01720              CGCAGGTCACGACCCCTCCCCATCCCCCACGGACGGCCCGGGTCGCCCAGAGTCTCCCCG     480
                      ************************************************ ***********

HLA07937              CTCCAGAGGATGTAATGGCTGCGACCTGGGGCCCGACGGGCGCCTCCTCCGCGGGTATGAC     720
HLA00434              CTCCAGAGGATGTCTGGCTGCGACCTGGGGCCCGACGGGCGCCTCCTCCGCGGGTATGAC     720
HLA01672_HLA01720     CTCCAGAGGATGTAATGGCTGCGACCTGGGGCCCGACGGGCGCCTCCTCCGCGGGTATGAC     720
HLA01720              CTCCAGAGGATGTCTGGCTGCGACCTGGGGCCCGACGGGCGCCTCCTCCGCGGGTATGAC     720
                      ************ ************************************************
```

Figura 40 – Sequência das bases de quatro alelos HLA-C: 07:18, 07:19, 07:01 e 07:02. Os quatro alelos possuem o comprimento de cerca de 720 bases e somente 6 bases são diferentes entre os alelos.

a população Britânica.

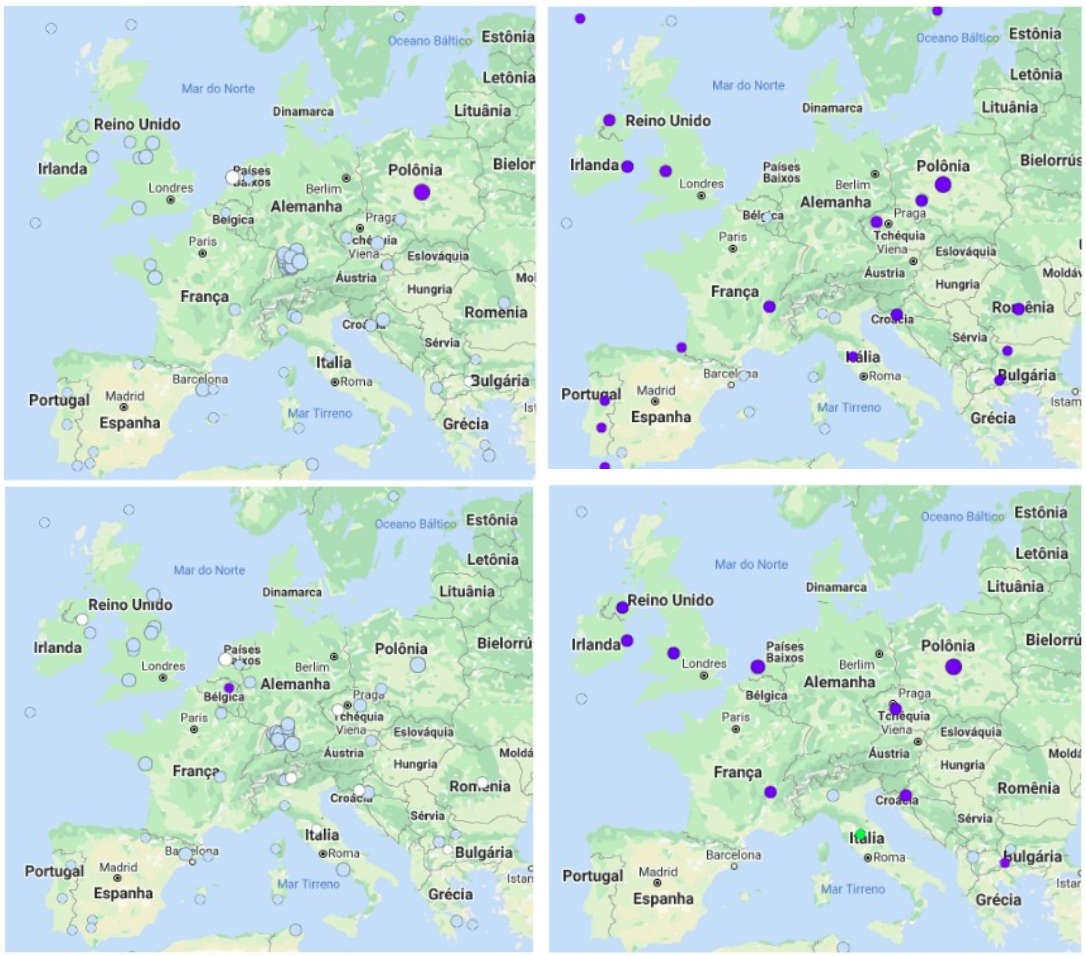
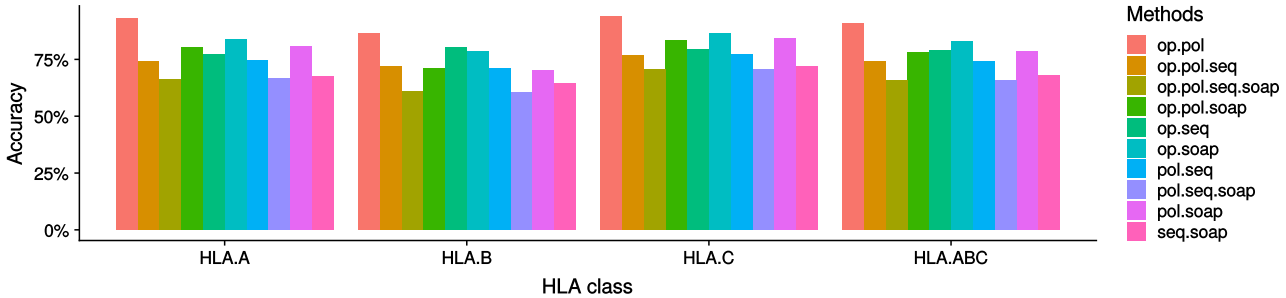


Figura 41 – Frequência dos alelos com erros na população GBR. Os dois primeiros gráficos correspondem à frequência dos alelos HLA-A*26:14 e HLA-A*26:01. Com destaque na região do Reino Unido, o alelo A*26:01 apresenta uma frequência maior que o A*26:14. Nos dois gráficos abaixo correspondem à frequência dos alelos HLA-B*44:07 e HLA-B*44:03, onde o alelo B*44:03 apresenta uma frequência maior que o B*44:07 na população GBR. Imagens retiradas do site http://www.allelefrequencies.net/.

Foi verificado também se ao criar o método *ensemble* seria possível melhorar a acurácia no processo de predição de HLA. Nós fizemos todas as combinações de softwares possíveis e verificamos a acurácia de cada combinação. Em todos os casos, as combinações não apresentaram uma melhor acurácia em relação ao método Optitype quando executado sozinho. A Figura 42 mostra a acurácia de cada combinação testada.

Como conclusão desse *benchmarking* utilizando amostras do projeto 1000 Genomes, podemos destacar que os métodos para predição de HLA são sensíveis



class	op.pol	op.seq	op.soap	pol.seq	pol.soap	seq.soap	op.pol.seq	op.pol.soap	pol.seq.soap	op.pol.seq.soap
HLA.A	0.9292929	0.7711864	0.8387097	0.7439024	0.8063063	0.6732283	0.7389558	0.8034934	0.6654545	0.6618705
HLA.B	0.8632075	0.8034934	0.7854077	0.7113821	0.6991870	0.6461538	0.7187500	0.7093023	0.6047297	0.6092715
HLA.C	0.9384615	0.7929515	0.8632075	0.7721519	0.8425926	0.7172131	0.7689076	0.8325792	0.7049808	0.7049808
HLA.ABC	0.9103207	0.7892105	0.8291083	0.7424788	0.7826953	0.6788651	0.7422045	0.7817917	0.6583884	0.6587076

Figura 42 – Acurácia de todas as combinações dos métodos de predição de HLA.

ao fato de existir alelos com sequências muito parecidas, evidenciando um problema computacional clássico de comparação de texto.Além disso, não levam em consideração a frequência do alelo na população da amostra. Entretanto, para o pipeline desenvolvido por esse trabalho, foi selecionado, como método padrão, o software Optitype para tipagem de HLA utilizando dados de DNA-Seq, uma vez que o método apresentou a melhor acurácia em relação aos demais métodos. O pipeline desenvolvido será detalhado na Seção 4.4.

4.4 NEO2P: NEOANTIGEN PREDICTION PIPELINE

Neste trabalho foi desenvolvido um pipeline para predição e priorização de neoantígenos usando como base o *framework* Snakemake, chamado de neo2P: **NEO**antigen **P**rediction **P**ipeline. O neo2P é um pipeline que integra todas as etapas para identificação de neoantígenos: chamada de variante somática, expressão gênica (opcional) e a identificação de novos epítopos.

As variantes somáticas são a principal informação necessária para iniciar o processo de identificação de neoantígeno e, para o neo2P, apesar de disponível como um módulo no pipeline, não é necessário executar esse passo caso já exista um VCF com variantes previamente identificadas, o que dá mais liberdade ao

usuário de escolher a própria metodologia de chamada de variante. Como detalhado na Seção 3.3.1, o pipeline de identificação de variante disponível no neo2P foi criado com base nas boas práticas do GATK4 e seu *workflow* é ilustrado na Figura 43.

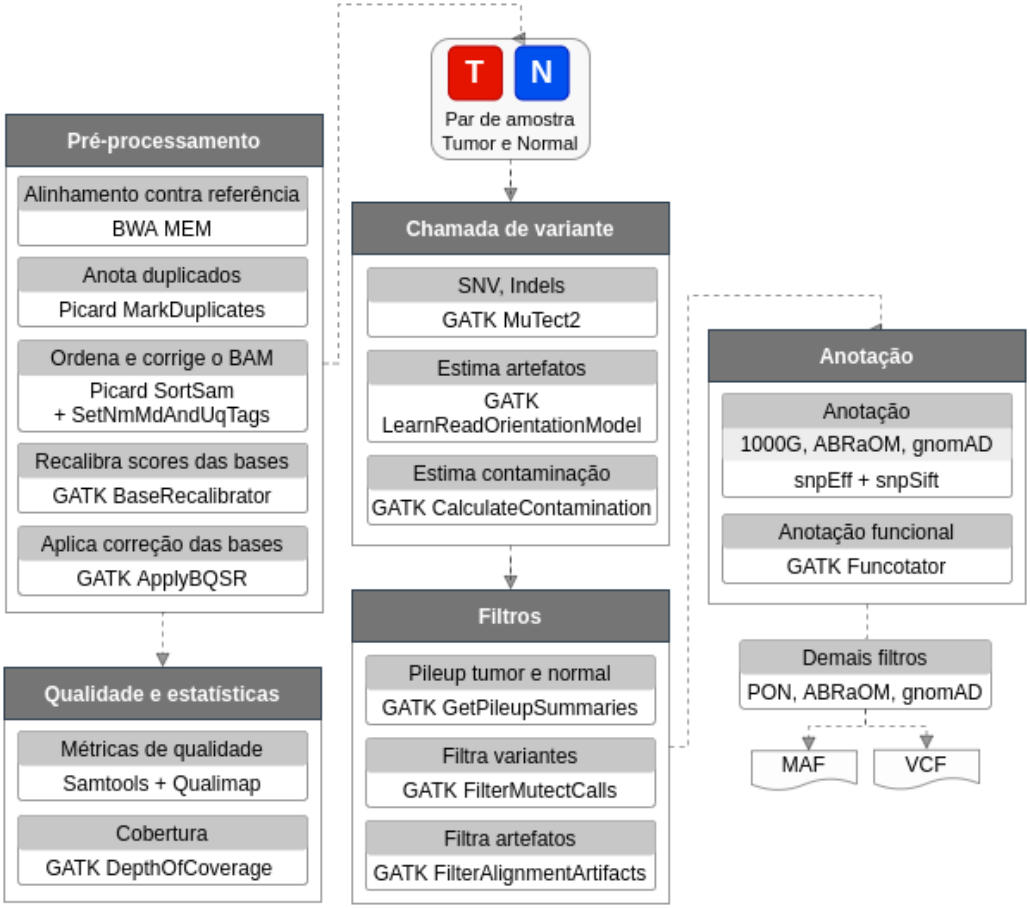


Figura 43 – *Workflow* de chamada de variante somática do neo2P. A partir dos dados brutos (FASTQ), as reads passam por alinhamento e recalibração, seguindo pela identificação de variantes SNV e indels. Após essa etapa, filtros de contaminação e artefatos são aplicados e as variantes restantes são anotadas com vários bancos de dados. Por fim, são gerados um arquivo VCF e um arquivo MAF por par de amostra, tumor e normal.

Outra informação importante para a identificação de neoantígeno é a expressão gênica, a qual é utilizada para filtrar os genes que estão expressos na amostra. No pvactools e em vários outros pipelines (Zhou et al. 2019; Lazdun et al. 2021), é definido um *cutoff* padrão de expressão como 1 FPKM, ou seja, genes com FPKM $\geq$ 1 são mantidos na análise.

No neo2P, a execução do pipeline de expressão é opcional e, caso não

exista a informação de expressão gênica disponível, o *score* de expressão gênica é utilizado para filtragem dos genes. Ademais, os genes são filtrados antes da identificação de neoantígeno, diminuindo o número de variantes e, portanto, o tempo de execução do pipeline. O *workflow* do pipeline de expressão gênica disponível no neo2P é ilustrado na Figura 44.

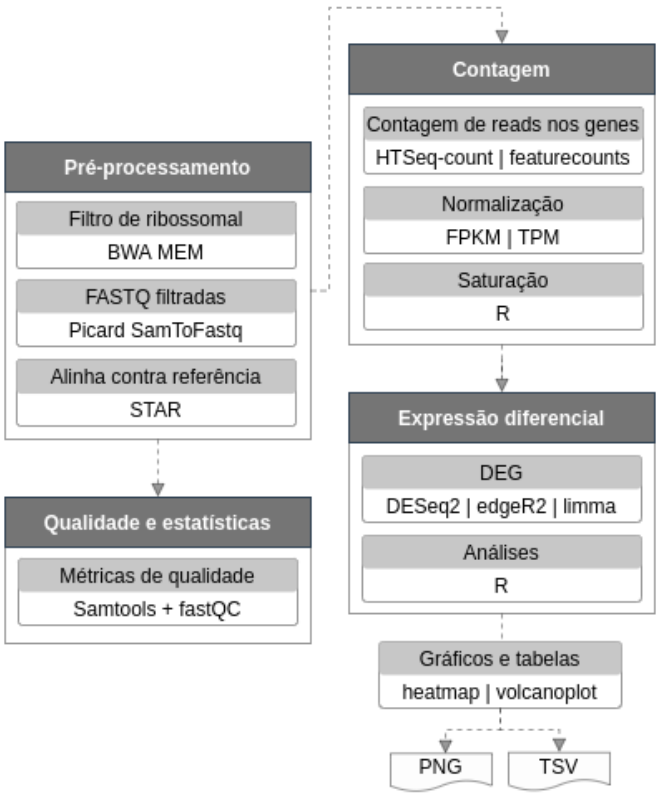


Figura 44 – *Workflow* de chamada de variante somática do neo2P. A partir dos dados brutos de RNA-Seq (FASTQ), possível contaminantes são filtradas das reads e alinhadas contra a referência do genoma humano.O arquivo com as reads alinhadas é utilizado para quantificação por gene e normalização por FPKM ou TPM. Caso necessário, é realizado uma análise de expressão gênica diferencial.

Integrando as informações de variantes e expressão gênica, o pipeline de identificação de neoantígeno pode ser executado. Uma outra etapa importante precisa ser executada antes da identificação dos epítopos, que é a tipagem de HLA, responsável por identificar os alelos do complexo MHC para cada amostra. No neo2P existem quatro softwares disponíveis para tipagem de HLA, e o método padrão é o Optitype, o qual utiliza das reads brutas (FASTQ) da amostra não-tumoral para identificar os alelos HLAs.

Resumidamente, o neo2P realiza a integração de variantes somática, expressão gênica, tipagem de HLA, identificação de epítopos, verificação de afinidade com o MHC, predição de clivagem do proteassoma e TCR, encapsulado em um estrutura de *workflow* no Snakemake. A Figura 45 ilustra o *workflow* dessa parte do pipeline.

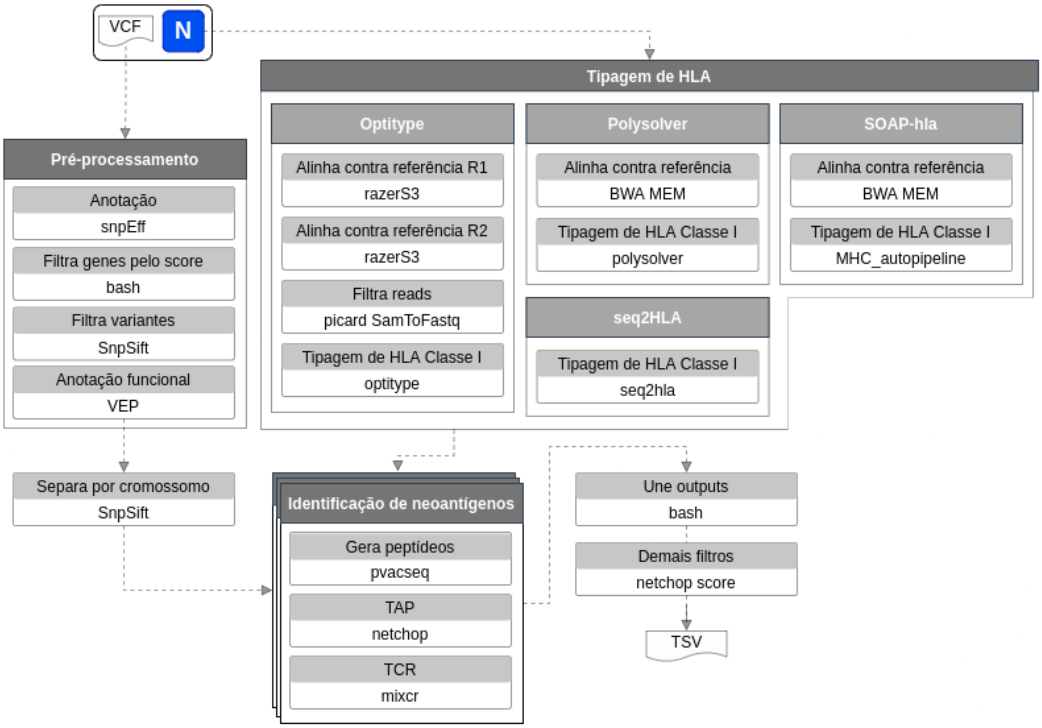


Figura 45 – *Workflow* de identificação de neoantígeno do neo2PA partir de um arquivo VCF com as variantes e dos arquivos FASTQ de amostra não-tumoral, o VCF é pré-processado para remoção de variantes filtradas e anotação funcional e os arquivos FASTQ são utilizados para tipagem de HLA. Para cada amostra, as variantes são divididas por cromossomos e, junto com cada alelo HLA, é verificada a afinidade com o MHC. Por fim, os resultados são unidos e é gerado um arquivo final com os neoantígenos candidatos.

O pipeline neo2P está disponível publicamente no github:https://github.com/adefelicibus/neo2P e as atualizações são realizadas pela Facility de Bioinformática do A.C.Camargo Cancer Center.

4.4.1 Aplicação do neo2P na coorte Hugo et al

O pipeline neo2P foi testado utilizando em uma coorte com 20 amostras de pacientes com melanoma e tem disponível informação de DNA-Seq e RNA-Seq pareados para todas os casos(3.1.3). Esse é um importante *dataset* de teste

porque os pacientes foram submetidos à imunoterapia e foram avaliados quanto à resposta ao tratamento. Esses dados clínicos foram utilizados para a avaliar a implicação dos neoantígenos no prognóstico dos pacientes.

O pipeline neo2P foi executado de maneira integral (*end-to-end*), ou seja, foi realizada a chamada de variante somática, a quantificação da expressão gênica a tipagem de HLA e a identificação de neoantígeno. Além disso, foi avaliada a acurácia do *score* de expressão gênica também nesses dados. Por isso, foram executadas duas análises: i) filtro dos genes pelo *score* de expressão; ii) filtro dos genes com FPKM $\geq$ 1. Para verificação de acurácia do *score*, a análise *ii* foi utilizada como conjunto verdade (*ground truth*), uma vez que não está disponível na literatura os neoantígenos identificados para essas amostras. A Tabela 11 sumariza o número de variantes encontradas, o TMB e carga de neoantígeno.

Tabela 11 – Informações sobre as variantes, TMB e carga de neoantígeno para as amostras de Hugo et al.

paciente	# variantes não sin.	TMB	grupo TMB	carga de neoant.	grupo carga neoant.
Pt2	1443	22.529	TMB-high	1090	high
Pt4	2977	46.479	TMB-high	1312	high
Pt5	236	3.684	TMB-low	239	high
Pt12	31	0.484	TMB-low	5	low
Pt20	332	5.183	TMB-low	6	low
Pt22	91	1.421	TMB-low	2	low
Pt23	175	2.732	TMB-low	29	low
Pt25	84	1.311	TMB-low	14	low
Pt29	153	2.389	TMB-low	20	low
Pt35	88	1.374	TMB-low	178	high
Pt37	276	4.309	TMB-low	136	high
Pt38	11	0.172	TMB-low	-	-
Pt8	7101	110.866	TMB-high	2054	high
Pt9	187	2.919	TMB-low	-	-
Pt10	170	2.654	TMB-low	74	low
Pt13	317	4.949	TMB-low	97	low
Pt14	665	10.382	TMB-low	243	high
Pt15	405	6.323	TMB-low	224	high
Pt28	33	0.515	TMB-low	-	-
Pt31	315	4.918	TMB-low	124	low

Para avaliar a *acurácia* do *score* de expressão gênica, foi verificado os genes utilizados nas duas análises realizadas, depois dos filtros específicos de cada análise. A Tabela 12 sumariza as métricas de performance da aplicação do *score*. O resultado obtido foi muito próximo da validação realizada com a coorte de Karasaki et al (9), com o *score* apresentando uma sensibilidade alta mas uma especificidade baixa. Porém, nesse *dataset*, em algumas amostras, o *score* se aproximou mais do resultado obtido pela análise *ii*, com a acurácia variando de 0.6 a 0.95 e o Fscore variando de 0.6 a 0.96.

Tabela 12 – Métricas de priorização de neoantígeno usando o *score* de expressão

amostra	acurácia	precisão	sensibilidade	especificidade	fscore
Pt10T	0.756	0.719	0.958	0.471	0.821
Pt12T	-	-	-	-	-
Pt13T	0.79	0.686	0.96	0.656	0.8
Pt14T	0.627	0.445	0.96	0.483	0.608
Pt15T	0.692	0.559	0.95	0.531	0.704
Pt20T	0.833	0.75	1	0.667	0.857
Pt22T	-	-	-	-	-
Pt23T	0.944	0.923	1	0.833	0.96
Pt25T	0.928	0.923	1	0.5	0.96
Pt28T	-	-	-	-	-
Pt29T	0.857	0.833	1	0.5	0.909
Pt2T	0.668	0.537	0.938	0.501	0.683
Pt31T	0.618	0.410	0.961	0.493	0.575
Pt35T	0.482	0.328	0.909	0.328	0.482
Pt37T	0.605	0.353	0.947	0.507	0.514
Pt38T	-	-	-	-	-
Pt4T	0.725	0.598	0.979	0.552	0.742
Pt5T	0.742	0.595	1	0.584	0.746
Pt9T	-	-	-	-	-
Pt8T	0.722	0.669	0.99	0.387	0.799

Em relação aos erros (falsos positivos e falsos negativos) apresentados pelo *score*, muitos genes em algumas amostras apresentaram uma expressão inferior ao que foi identificado no TCGA no cálculo do *score*A Figura 46 ilustra a distribuição do FPKM por amostra e observa-se que o filtro pelo *score*, nessas amostras, está considerando muitos genes abaixo do *threshold* definido como ponto de corte ótimo (bolinhas azuis), mas que apresentam um FPKM menor do que 1Em con-

tra partida, os genes com FPKM acima de 1 estão também classificados abaixo do *threshold* do *score*, salvo para alguns genes nas amostras Pt2T e Pt14T, assim como foi observado na maioria das amostras do projeto SKCM do TCGA.

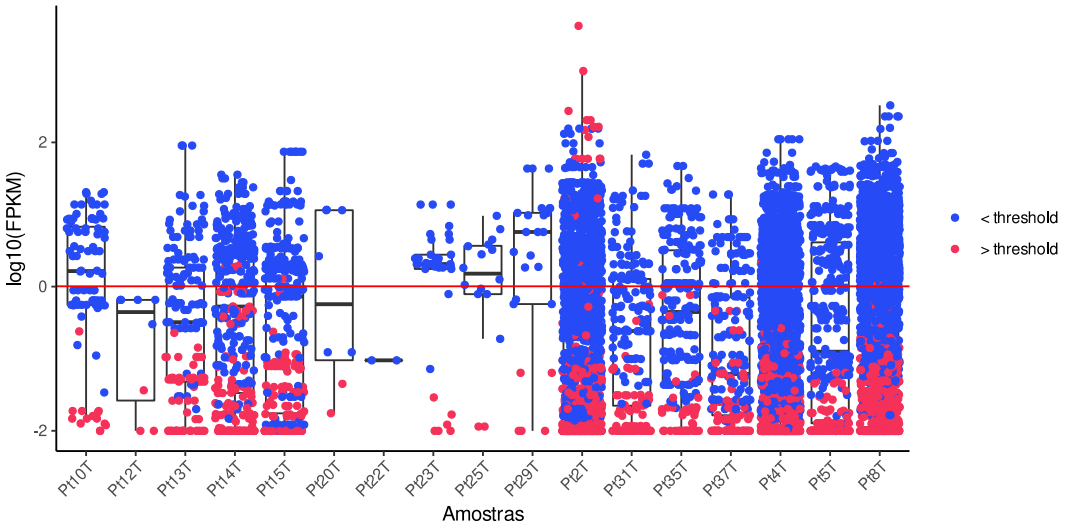


Figura 46 – Distribuição do FPKM nas amostras da coorte de Hugo et al. Cada bolinha do gráfico representa o FPKM de um gene onde foi identificado pelo menos um neoantígeno. Os genes filtrados pelo *score* (análise *i*) são destacados em vermelho, e os genes que continuaram na análise estão em azul. A linha vermelha no gráfico destaca o ponto que corresponde a log 10(1), o filtro realizado na análise *ii*.

Um ponto importante para destacar a respeito do *score* de expressão gênica proposto é que ele se aproxima dos resultados obtidos pelo padrão-ouro e, para casos onde a expressão gênica tumoral não está disponível, ele poderia ser usado como filtro de genes expressos. Como pontuado na revisão realizada por Richters et al. (2019), a expressão gênica identificada para o mesmo tipo tumoral deve ser utilizada nos casos onde não existe a informação identificada por amostra, e os resultados obtidos nesses testes mostram que o *score* poderia ser utilizado nesses casos.

Foi avaliado também nesse trabalho, utilizando as informações clínicas disponíveis das amostras (3.1.3), a relação entre a carga de variantes (TMB, do inglês *Tumor mutation burden*) e carga de neoantígenos, e como essas informações poderiam influenciar na resposta ao tratamento e na sobrevida global dos pacientes.

O TMB pode ser definido como o número de variantes somáticas identificadas por megabase do genoma que foi sequenciado, e tem sido considerado como um biomarcador preditivo de resposta ao tratamento com inibidores de *checkpoints* (Wu et al. 2019; Kang et al. 2020) e utilizado na prática de oncologia clínica (Chan et al. 2019). Além disso, tem sido relacionado positivamente a melhores prognósticos ao comparar pacientes com um alto valor de TMB em diversos tipos de cânceres (Riviere et al. 2020; Shao et al. 2020), inclusive pacientes com melanoma que receberam tratamento com anti-PD1 ou anti-CTLA4 (Snyder et al. 2014; Rizvi et al. 2015; Van Allen et al. 2015). Por fim, a carga de neoantígeno também já foi relacionada com bom prognóstico e maior sobrevida (Miller et al. 2017).

Recentemente, o FDA (do inglês *Food and Drug Administration* dos Estados Unidos aprovou o tratamento com pembrolizumab para pacientes com um TMB ≥ 10, caracterizados como *TMB-high*, baseado nos resultados obtidos pelo estudo clínico fase-2 KEYNOTE-158 (Marabelle et al. 2020). Como não existe um consenso de qual *cutoff* utilizar para estratificar um amostra como TMB-high ou TMB-low (Stenzinger et al. 2019), neste projeto foi utilizado o seguinte critério: amostras com a carga de mutação superior à média de todas as amostras foi caracterizada como TMB-high, e as demais como TMB-low. Para estratificar as amostras em relação à carga de neoantígeno em *high* e *low*, foi utilizada a mediana de todas as amostras por conta da variância e *outliers*. As análises realizadas considerando a estratificação adotada pelo FDA também foi realizada e está detalhada no Apêndice D.

Nos resultados obtidos, ao comparar o TMB com os grupos de pacientes Respondedor (R) e Não-Respondedor (NR), não foi encontrada nenhuma diferença estatisticamente significativa entre os dois grupos. Em contra partida, ao avaliar a carga de neoantígenos entre os grupos, houve uma diferença estatisticamente significativa entre o grupo R e NR, mostrando que os pacientes com melhor desfecho ao tratamento apresentaram uma carga maior de neoantígeno.

Um ponto a ser destacado é que, mesmo que exista uma correlação positiva entre o TMB e a carga de neoantígeno, outros aspectos como cobertura e frequência alélica da variante, expressão gênica e afinidade com o MHC interferem nos neoantígenos candidatos. A Figura 47 ilustra esses resultados.

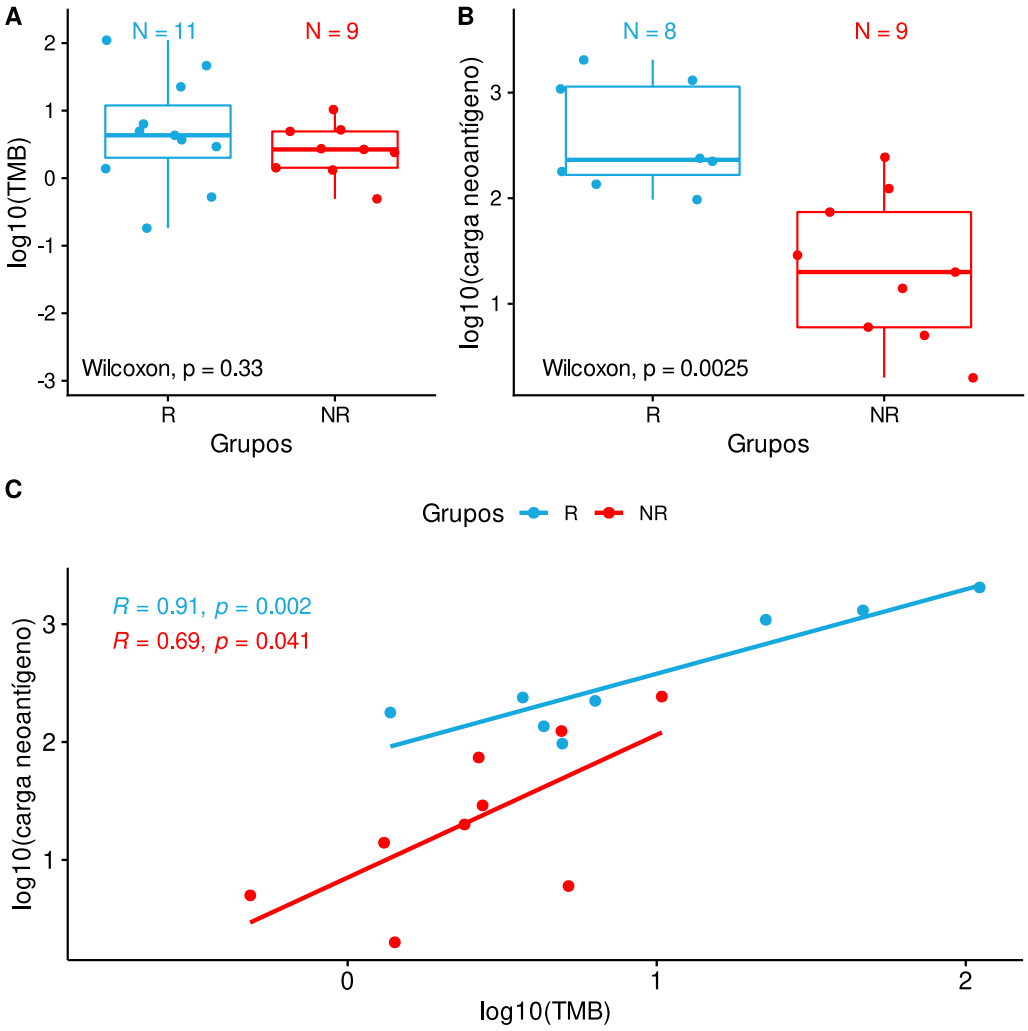


Figura 47 – Relação de TMB e carga de neoantígenos entre os grupos R e NR. (A) Distribuição do TMB (em $log10$) para as amostras do grupo R (n=11) e do grupo NR (n=9), sem diferença estatisticamente significativa (*p-value*=0.33); (B) Distribuição da carga de neoantígeno (em $log10$) para as amostras do grupo R (n=8) e do grupo NR (n=9), com diferença estatisticamente significa (*p-value*=0.0025); (C) Correlação entre TMB e carga de neoantígeno para os grupos R e NR, com o grupo R apresentando uma correlação positiva forte (R=0.91, *p-value*=0.002), enquanto que o grupo NR apresentou uma correlação positiva moderada (R=0.69, *p-value*=0.041).

Em relação ao desfecho clínico dos pacientes e sobrevida global, como ilustrado na Figura 48 e já descrito em Hugo et al. (2016), os pacientes do grupo R apresentaram uma sobrevida global superior e estatisticamente significativa

em relação aos pacientes do grupo NR. Adicionalmente, ao avaliar o TMB, os pacientes com TMB-high apresentaram também uma sobrevida global melhor, porém não estatisticamente significativa. Por outro lado, ao verificar a carga de neoantígeno, os pacientes classificados como *high* apresentaram uma sobrevida global superior e estatisticamente significativa em relação aos pacientes classificados como *low*. Novamente, foi observado que a carga de neoantígeno pode ser mais informativa que somente o TMB, mas essa conclusão pode estar enviesada pelo tamanho da coorte.

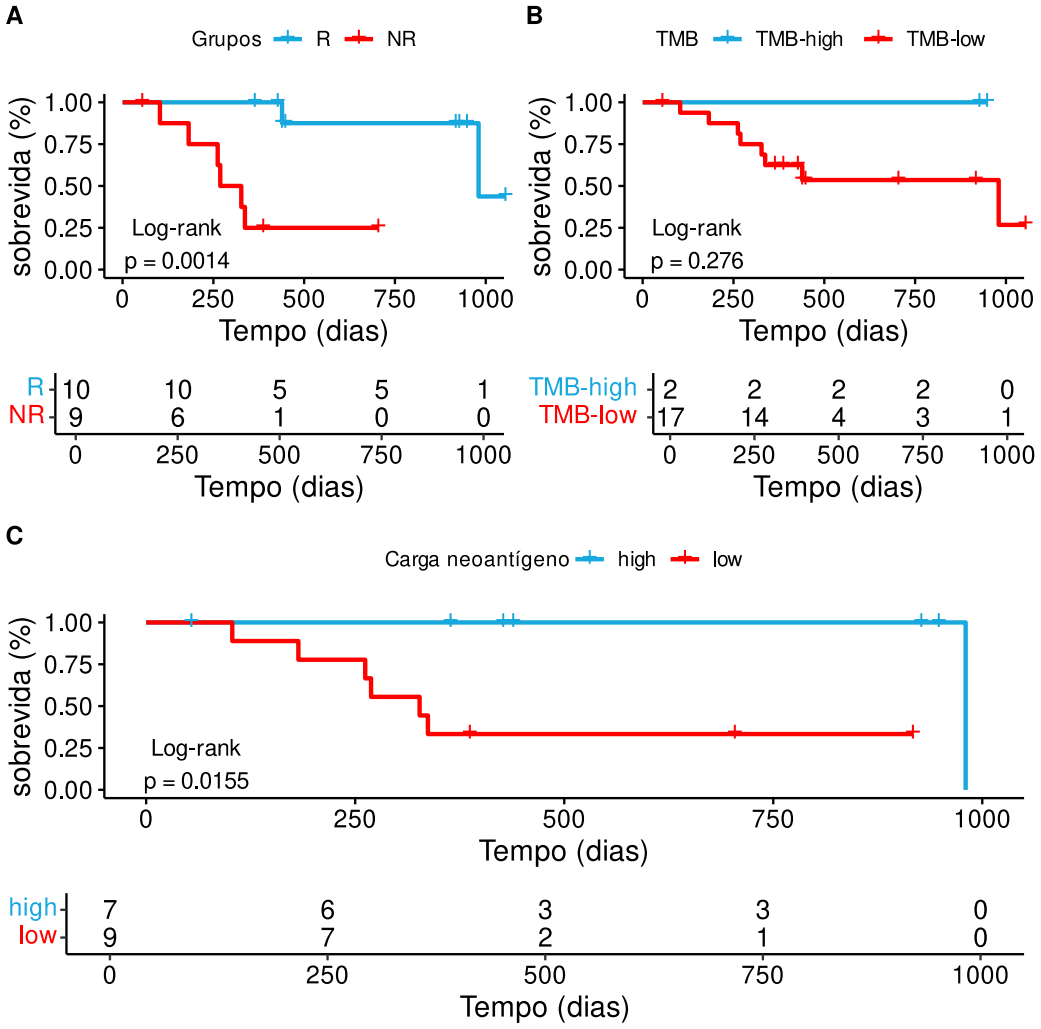


Figura 48 – Análise de sobrevida global da coorte Hugo et al. (A) Sobrevida global entre os pacientes dos grupos R (n=10) e NR (n=9), com diferença estatisticamente significativa (*p-value*=0014); (B) Sobrevida global entre os pacientes dos grupos TMB-high (n=2) e TMB-low(n=17), sem diferença estatisticamente significativa (*p-value*=0.276); (C) Sobrevida global entre os pacientes dos grupos *high* (n=7) e *low* (n=9) da carga de neoantígeno, com diferença estatisticamente significativa (*p-value*=0.0155).

Uma vez que os dados de RNA-Seq estavam disponíveis para essas amostras, foi possível avaliar também a expressão de alguns genes marcadores de resposta imune: inflamatória/citotóxica, supressão/exaustão, coestimulação/apresentação de antígeno e outros (Buttura et al.2021). Ao comparar a expressão desses genes marcadores entre os grupos R e NR, foi identificado um nível de expressão maior do gene IL10 no grupo NR. Apesar de não estar claro qual o papel da interleucina-10 no ambiente tumoral, muitos trabalhos têm caracterizado a IL-10 como um agente supressor do sistema imune (Smith et al. 2018) e também com a progressão tumoral em melanoma (Itakura et al.2011). A Figura 49 ilustra a expressão dos marcados em quatro listas diferentes.

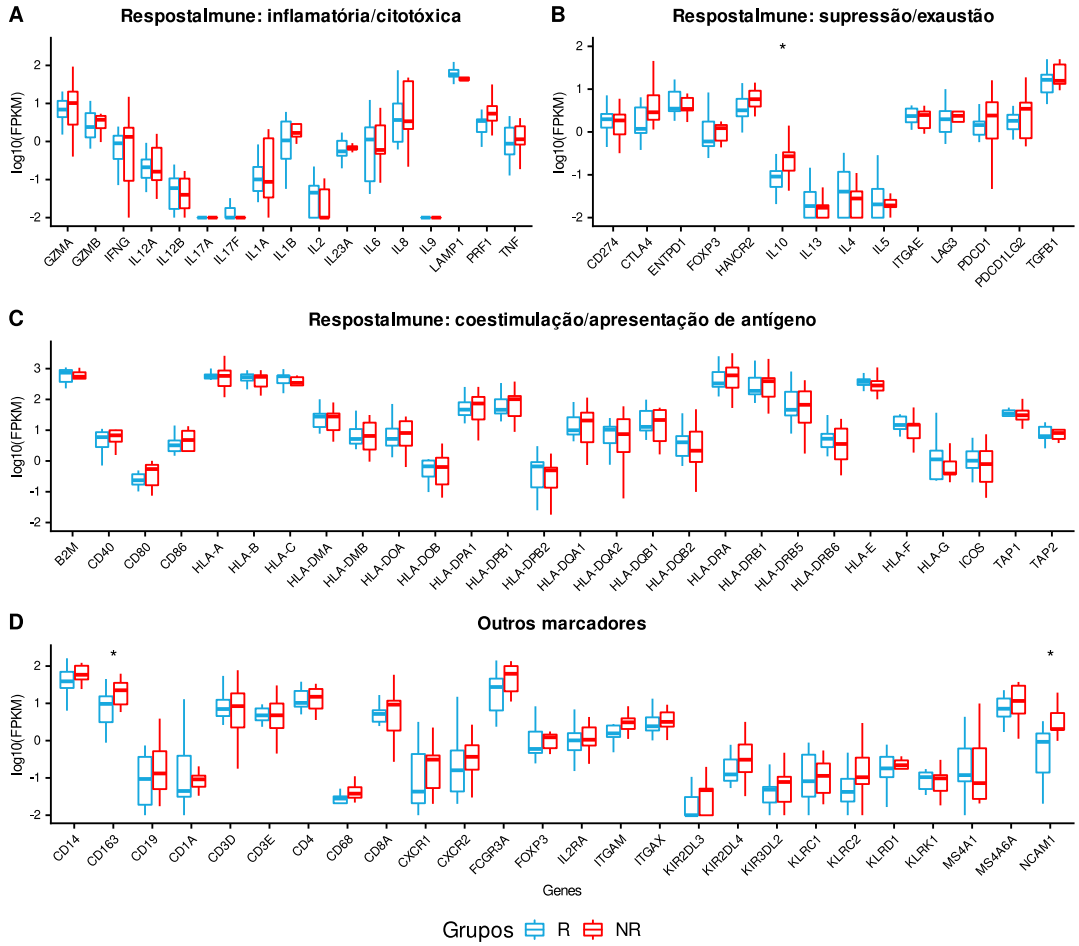


Figura 49 – Expressão de genes marcadores entre os grupos R e NR. São quatro listas com marcadores de resposta imune diferentes, comparando a expressão (em $log10$) dos genes entre os grupos R e NR: (A) inflamatória/citotóxica; (B) supressão/exaustão; (C) coestimulação/apresentação de antígeno; (D) outros marcadores. * $p \leq 0.05$.

A comparação de expressão gênica dos marcadores também foi realizada considerando os grupos TMB-high e TMB-low. Conforme ilustrado na Figura 50, foi verificada uma expressão maior do gene IL9 para o grupo R em relação ao grupo NR. A interleucina-9 pode regular a função de vários tipos de células e, recentemente, tem sido estudada sua atividade antitumoral (Wan et al. 2020) e o seu papel no infiltrado tumoral no câncer de melanoma (Parrot et al. 2016). Alem do gene IL9, outros genes como IL5 e CTLA4, caracterizados pelo papel de supressão do sistema imune, apresentaram expressão maior nas amostras do grupo TMB-low.

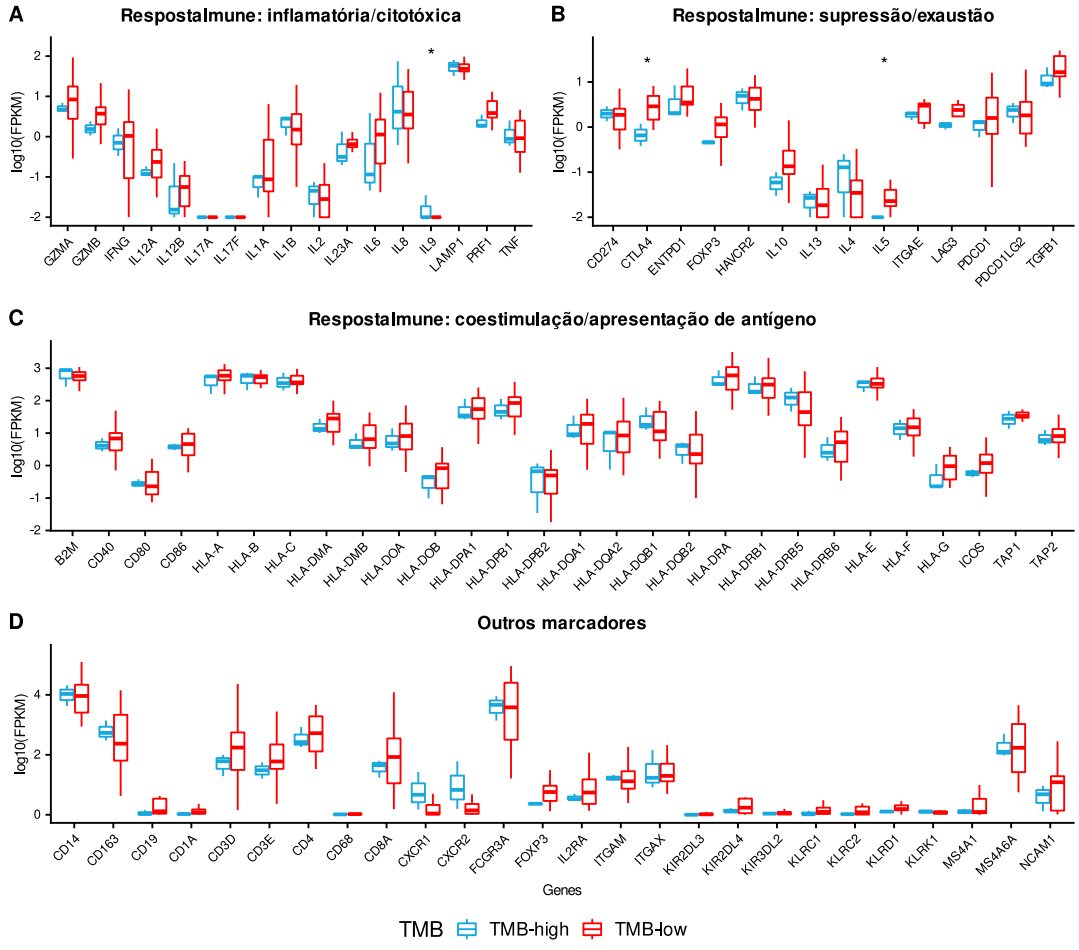


Figura 50 – Expressão de genes marcadores entre os grupos TMB-high e TMB-low. São quatro listas com marcadores de resposta imune diferentes, comparando a expressão (em $log10$) dos genes entre os grupos TMB-high e TMB-low: (A) inflamatória/citotóxica; (B) supressão/exaustão; (C) coestimulação/apresentação de antígeno; (D) outros marcadores. * $p \leq 0.05$.

A mesma análise anterior foi realizada no contexto da carga de neoantí-

geno, comparando a expressão dos genes marcadores entre os grupos *high* e *low*. Assim como nas análises anteriores, foi possível observar uma expressão maior de genes supressores, como IL10, CTLA4 e FOXP3 no grupo *low* em relação ao grupo *high*. Ademais, outros marcadores também apareceram com diferença estatisticamente significativa, como CD1A e NCAM1.A Figura 51 ilustra a expressão dos genes marcadores para os dois grupos.

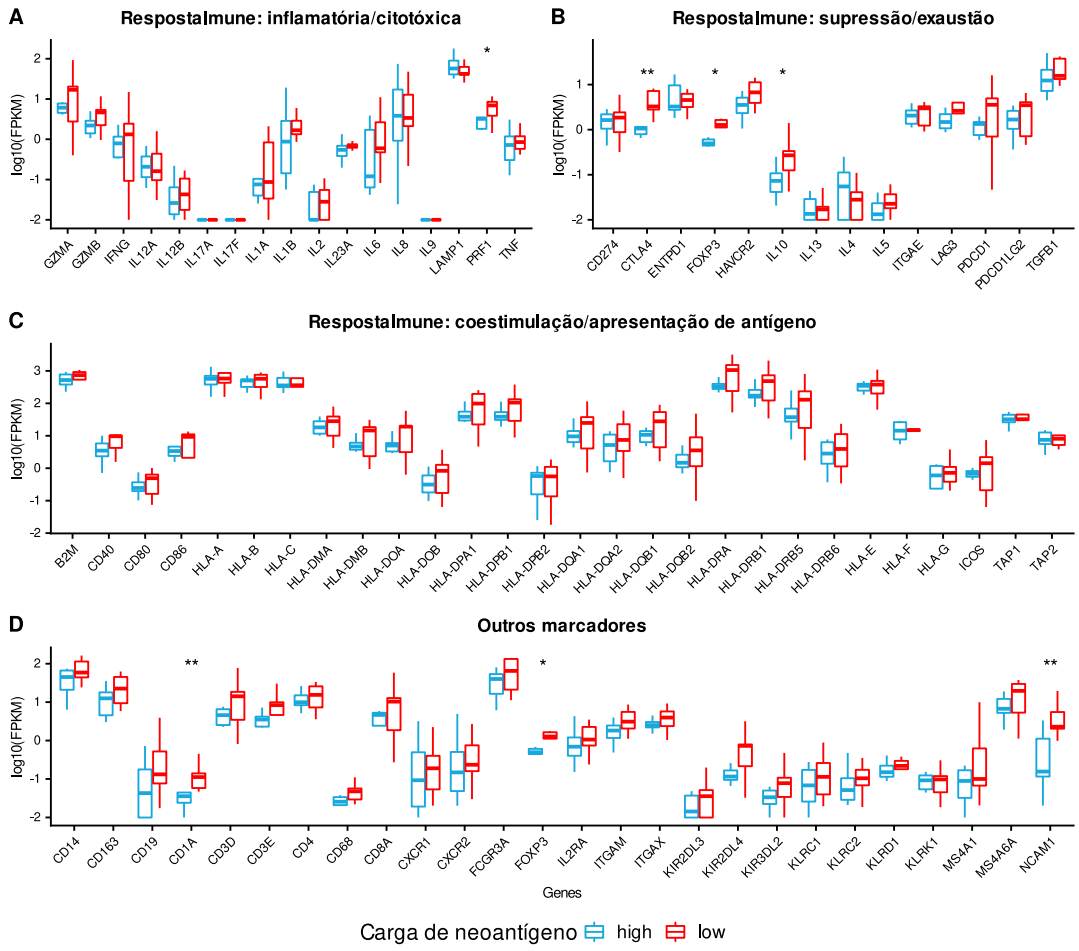


Figura 51 – Expressão de genes marcadores entre os grupos *high* e *low* da carga de neoantígeno. São quatro listas com marcadores de resposta imune diferentes, comparando a expressão (em $log10$) dos genes entre os grupos *high* e *low*: (A) inflamatória/citotóxica; (B) supressão/exaustão; (C) coestimulação/apresentação de antígeno; (D) outros marcadores. * $p \leq 0.05$, ** $p \leq 0.01$.

Em relação à composição do microambiente tumoral estimado pelo CIBERSORTx para todas as amostras utilizando os dados de RNA-Seq, não foi encontrado nenhuma diferença estatisticamente significativa na composição das células do infiltrado inflamatório nas três comparações realizadas:R e NR, TMB-

high e TMB-low e *high* e *low*, exceto para a célula *B.memory*, com diferença identificada somente para os grupos Respondedor e com carga de neoantígeno *high*. Uma composição parecida foi encontrada com os dados do TCGA para o projeto SKCM (Wang et al.2020). Como já foi dito antes, a carga de neoantígeno pode ser mais informativa do que somente o uso do TMB. A Figura 52 ilustra a composição das células imunes nas três comparações realizadas.

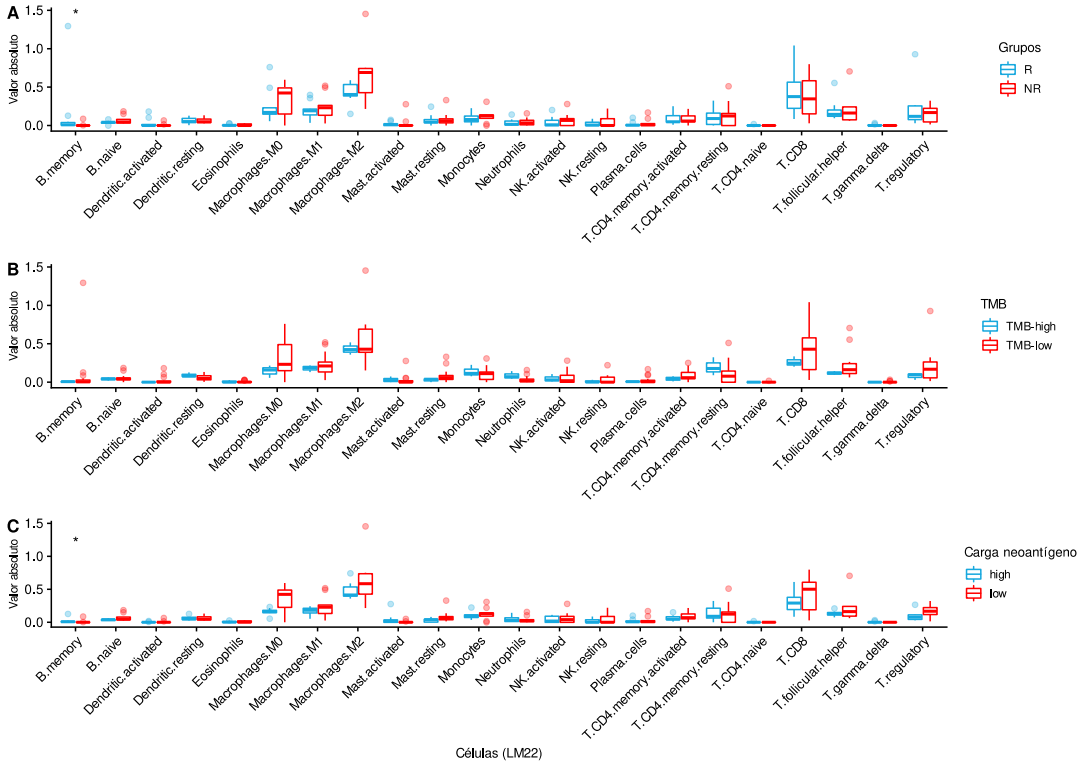


Figura 52 – Distribuição da composição de células imunes na coorte Hugo et al em três comparações de grupos diferentes estimado pelo CIBERSORTx. (A) Comparação da composição de 22 células do sistema imune nos grupos R e NR; (B) Comparação da composição de 22 células do sistema imune nos grupos TMB-high e TMB-low; (C) Comparação da composição de 22 células do sistema imune nos grupos *high* e *low*. * $p \leq 0.05$.

Apesar de não ter sido encontrado diferença estatisticamente significativa na composição das células do sistema imune, é possível visualizar que as células T CD8, macrófagos M2 e M0 e Treg apresentaram mais abundância dos que as demais células. Essa mesma conclusão também pode ser visualizada na Figura 53, onde é possível ver dois *clusters* nas colunas (células), separando as células T CD8 e Macrófagos M2 das demais células. Em relação às amostras, também é possível visualizar dois *clusters* nas linhas, o primeiro formado basicamente por

amostras dos grupos NR, TMB-low e *low* (carga de neoantígeno) e com menor abundância das células T CD8 e M2, e o segundo *cluster* com praticamente todas as amostras dos grupos R, TMB-high e *high* (carga de neoantígeno).

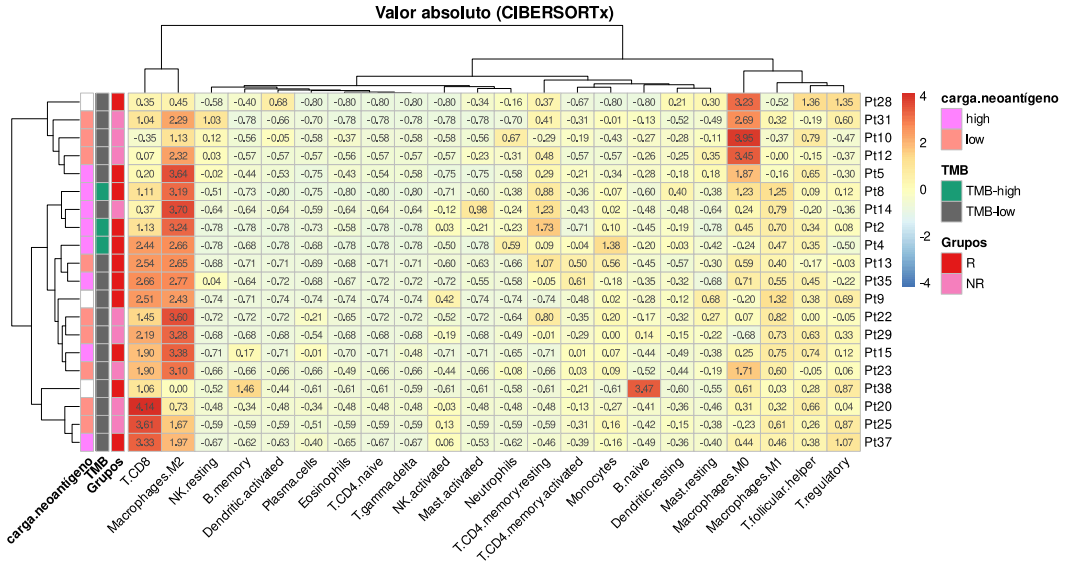


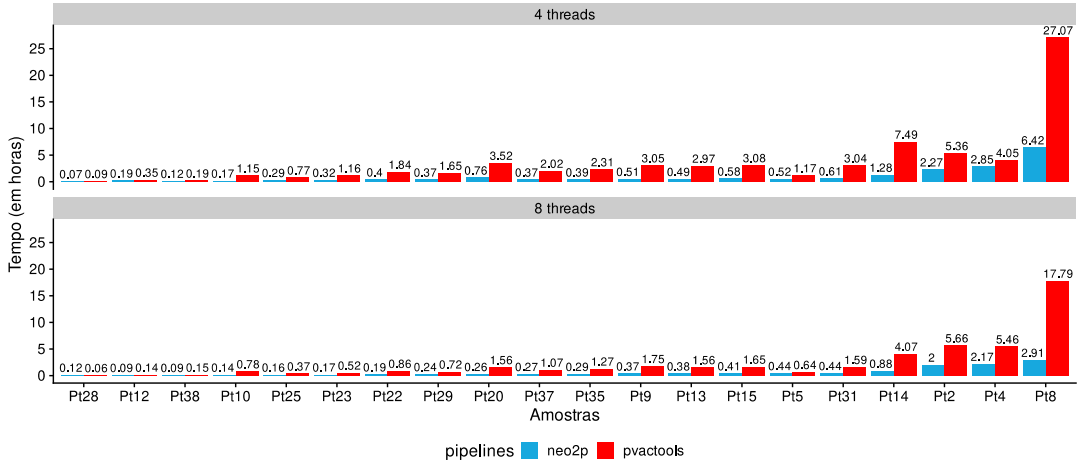
Figura 53 – *Heatmap* com a composição de células imunes na coorte Hugo et al. Composição do valor absoluto estimado pelo CIBERSORTx para cada uma das amostras, clusterizando pelas células (colunas) e pelas amostras (linhas).

4.4.2 Performance computacional

Um importante aspecto no processo de identificação de neoantígeno é a eficiência computacional. Além da necessidade de vários tipos de *omics*, o pipeline de predição e priorização de neoantígeno tem um custo computacional alto.Portanto, foi avaliada a performance computacional do neo2p, somente do pipeline de predição de neoantígeno, e do pvactools, ao aplicar o algoritmo de predição pvacseq.

Os dois métodos foram executados em um computador com processador AMD Opteron 6262 HE 1600 MHz de 64 bits, 2MB de cache L2 com 64 núcleos e 2 *threads* por núcleo, totalizando 128 *threads*, e 64GB de RAM DDR4. Duas configurações de paralelismo dos métodos foram testadas: 4 e 8 *threads*, para avaliar o ganho de performance ao aumentar o poder computacional disponível. A Figura 54 ilustra o tempo de execução dos dois pipelines para cada uma das

20 amostras de pacientes com melanoma, além de uma tabela com o TMB e a carga de neoantígeno de cada amostra.



amostra	Pt28	Pt12	Pt38	Pt10	Pt25	Pt23	Pt22	Pt29	Pt20	Pt37	Pt35	Pt9	Pt13	Pt15	Pt5	Pt31	Pt14	Pt2	Pt4	Pt8
# neoantigeno	NA	5	NA	74	14	29	2	20	6	136	178	NA	97	224	239	124	243	1090	1312	2054
TMB	0.52	0.48	0.17	2.65	1.31	2.73	1.42	2.39	5.18	4.31	1.37	2.92	4.95	6.32	3.68	4.92	10.38	22.53	46.48	110.87

Figura 54 – Análise de tempo de execução dos pipelines neo2p e pvactools para cada amostra. Em geral, os resultados mostram que o pipeline neo2p apresenta uma eficiência computacional de 2 a 6 vezes superior ao pvactools. Além disso, o número de variantes e o número de *threads* disponíveis influenciam o tempo de execução do pvactools, mas influenciam pouco a eficiência do neo2p.

Observa-se que o pipeline neo2p apresenta uma eficiência computacional de 2 a 6 vezes superior ao pvactools, independente do número de *threads*. Além disso, o tempo computacional utilizado pelo neo2p com 4 *threads* foi inferior ao tempo necessário que o pvactools levou para finalizar a identificação de neoantígenos com 8 *threads*. Ademais, a eficiência computacional é muito influenciada pelo número de variantes apresentado pela amostra, mas nos resultados encontrados essa relação não é linear, uma vez que o número de alelos HLA e o tamanho dos peptídeos de interesse também impactam no tempo computacional.

Essa eficiência computacional do neo2p é caracterizada por duas otimizações realizadas no pipeline desenvolvido nesse trabalho em relação ao pvactools. Primeiro, o neo2p utiliza o snakemake, o qual disponibiliza ferramenta para dividir o problema em partes menores e executa vários processos em paralelo. Segundo, o VCF de cada amostra é dividido por cromossomo e a execução do pvacseq é realizada por cromossomo e por alelo HLA, em paralelo. Impor-

tante destacar que essas otimizações não alteraram o código fonte dos programas utilizados, foram realizadas utilizando as vantagens de uma ferramenta de gerenciamento de *workflow* como o snakemake.

O pvactools realiza a execução em paralelo, mas somente no processo de cálculo de afinidade do peptídeo ao MHC. Após a geração do FASTA para cada peptídeo, o pvacseq divide os peptídeos em pequenos arquivos e realiza a verificação de afinidade para cada arquivo, mas repete o processo para cada alelo HLA e tamanho de peptídeo a ser predito, e para uma amostra com muitas variantes, o processo pode ser muito demorado, como no caso da amostra Pt8.

Outra vantagem do neo2p é a execução dos processos em paralelo em um ambiente de cluster, com muitos computadores conectados. A mesma análise anterior foi realizada em um cluster com sete computadores com configurações similares ao do teste anterior, e o nep2p executou cerca de 3000 processos em aproximadamente ~6 horas.

5 CONCLUSÕES

Através da integração de ferramentas computacionais foi desenvolvido um pipeline para detecção de neoantígenos denominado *neo2P*. Quando realizado um *benchmarck* da abordagem proposta, o neo2P apresentou uma eficiência superior quanto aos recursos computacionais, além de uma integração completa de todos os passos necessários para a detecção de neoantígenos. Diante da falta de dados de expressão gênica, necessários para correta detecção de neoantígenos, foi proposto um *score* para priorizar as mutações somáticas a partir da distribuição esperada dos níveis de expressão gênica de aproximadamente 10,000 pacientes disponíveis em 32 projetos do TCGA. De importância, foi mostrado que o *score* apresenta um poder de discriminação (AUC) próximo ou superior a 0.7 na maioria das coortes avaliadas. Ao aplicar a abordagem proposta em uma coorte de 20 pacientes com melanoma, foram pontuados aspectos adicionais da relação de neoantígenos e aspectos imunes, como a expressão de alguns genes marcadores que podem estar relacionados com a resposta ao tratamento.Adicionalmente, foi mostrado ainda que a carga de neoantígenos detectados pelo neo2P estratifica, de maneira significativa, pacientes respondedores (R) e não respondedores (NR) quando comparado com o marcador TMB. Diante da disponibilidade dos dados genômicos e a importância do valor prognóstico dos neoantígenos, o pipeline neo2P será útil nos estudos orientados a neoantígenos e imunoterapia.

6 REFERÊNCIAS

1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, et al. A map of human genome variation from population-scale sequencing. Nature. 2010 Oct; 467:1061–73.

1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015 Oct; 526:68–74.

Abi-Rached L, Gouret P, Yeh JH, Di Cristofaro J, Pontarotti P, Picard C, et al. Immune diversity sheds light on missing variation in worldwide genetic diversity panels. PLoS One. 2018 Oct; 13:e0206512.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct; 215:403–10.

Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and challenges in long-read sequencing data analysis. Genome Biol. 2020 Dec; 21: 30.

Anders S, Pyl PT, Huber W. HTSeq–a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015 Jan; 31:166–9.

Bauer DC, Zadoorian A, Wilson LOW, Melbourne Genomics Health Alliance, Thorne NP. Evaluation of computational programs to predict HLA genotypes from genomic sequencing data. Brief Bioinform. 2018 Nov; 19:179–187.

Berger AC, Korkut A, Kanchi RS, Hegde AM, Lenoir W, Liu W, et al. A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers. Cancer Cell. 2018 Apr; 33:690–705.e9.

Blass E, Ott PA. Advances in the development of personalized neoantigen-based therapeutic cancer vaccines. Nat Rev Clin Oncol. 2021 Apr; 18:215–229.

Boegel S, Löwer M, Schäfer M, Bukur T, de Graaf J, Boisguérin V, et al. HLA typing from RNA-Seq sequence reads. Genome Med. 2012 Dec; 4:102.

Bolotin DA, Poslavsky S, Mitrophanov I, Shugay M, Mamedov IZ, Putintseva EV, et al. MiXCR: software for comprehensive adaptive immunity profiling. Nat Methods. 2015 May; 12:380–1.

Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018 Nov; 68:394–424.

Buttura JR, Provisor Santos MN, Valieris R, Drummond RD, Defelicibus A, Lima JP, et al. Mutational Signatures Driven by Epigenetic Determinants Enable the Stratification of Patients with Gastric Cancer for Therapeutic Intervention. Cancers (Basel). 2021 Jan; 13:490.

Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012 Jul; 487:330–7.

Cancer Genome Atlas Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature. 2015 Jan; 517:576–82.

Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012 Sep; 489:519–25.

Cancer Genome Atlas Research Network. Comprehensive molecular characterization of gastric adenocarcinoma. Nature. 2014 Sep; 513:202–9.

Cancer Genome Atlas Research Network. The Molecular Taxonomy of Primary Prostate Cancer. Cell. 2015 Nov; 163:1011–25.

Cancer Genome Atlas Research Network, Analysis Working Group:Asan University, BC Cancer Agency, Brigham and Women's Hospital, Broad Institute, Brown University, et al. Integrated genomic characterization of oesophageal carcinoma. Nature. 2017 Jan; 541:169–175.

Cancer Genome Atlas Research Network, Kandoth C, Schultz N, Cherniack AD, Akbani R, Liu Y, et al. Integrated genomic characterization of endometrial carcinoma. Nature. 2013 May; 497:67–73.

Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013 Oct; 45:1113–20.

Cao H, Wu J, Wang Y, Jiang H, Zhang T, Liu X, et al. An integrated tool to study MHC region: accurate SNV detection and HLA genes typing in human MHC region using targeted high-throughput sequencing. PLoS One. 2013 Jul; 8: e69388.

Ceccarelli M, Barthel FP, Malta TM, Sabedot TS, Salama SR, Murray BA, et al. Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma. Cell. 2016 Jan; 164:550–63.

Challis GB, Stam HJ. The spontaneous regression of cancer. A review of cases from 1900 to 1987. Acta Oncol. 1990 Jan; 29:545–50.

Chan T, Yarchoan M, Jaffee E, Swanton C, Quezada S, Stenzinger A, et al. Development of tumor mutation burden as an immunotherapy biomarker:utility for the oncology clinic. Ann Oncol. 2019 Jan; 30:44–56.

Chida K, Nakanishi K, Shomura H, Homma S, Hattori A, Kazui K, et al. Spontaneous regression of transverse colon cancer: a case report. Surg case reports. 2017 Dec; 3:65.

Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM, et al. Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. Front Genet. 2012 Mar; 3:35.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly (Austin). 2012 Apr; 6:80–92.

Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res. 2016 May; 44:e71.

Coley WB. The Treatment of Inoperable Sarcoma by Bacterial Toxins (the Mixed Toxins of the Streptococcus erysipelas and the Bacillus prodigiosus). Proc R Soc Med. 1910 Aug; 3:1–48.

Coulie PG, Van den Eynde BJ, van der Bruggen P, Boon T. Tumour antigens recognized by T lymphocytes: at the core of cancer immunotherapy. Nat Rev Cancer. 2014 Feb; 14:135–46.

Davis CF, Ricketts CJ, Wang M, Yang L, Cherniack AD, Shen H, et al. The somatic genomic landscape of chromophobe renal cell carcinoma. Cancer Cell. 2014 Sep; 26:319–330.

Decker WK, Safdar A. Bioimmunoadjuvants for the treatment of neoplastic and infectious disease: Coley's legacy revisited. Cytokine Growth Factor Rev. 2009 Aug; 20:271–81.

Ding L, Chen F. Predicting Tumor Response to PD-1 Blockade. N Engl J Med. 2019 Aug; 381:477–479.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan; 29:15–21.

Dou Y, Gold HD, Luquette LJ, Park PJ. Detecting Somatic Mutations in Normal Cells. Trends Genet. 2018 Jul; 34:545–557.

Ferlay J, Ervik M, Lam F, Colombet M, Mery L, Piñeros M, et al. *Global Cancer Observatory: Cancer Today*. 2020. Disponível em:https://gco.iarc.fr/today [2021 02 02].

Fishbein L, Leshchiner I, Walter V, Danilova L, Robertson AG, Johnson AR, et al. Comprehensive Molecular Characterization of Pheochromocytoma and Paraganglioma. Cancer Cell. 2017 Feb; 31:181–193.

Fouad YA, Aanei C. Revisiting the hallmarks of cancer. Am J Cancer Res. 2017 May; 7:1016–1036.

Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016 Jun; 17:333–51.

Gourraud PA, Khankhanian P, Cereb N, Yang SY, Feolo M, Maiers M, et al. HLA diversity in the 1000 genomes dataset. PLoS One. 2014 Jul; 9:e97282.

Grosso JF, Jure-Kunkel MN. CTLA-4 blockade in tumor models:an overview of preclinical and translational research. Cancer Immun. 2013 Jan; 13:5.

Grüning B, Dale R, Sjödin A, Chapman BA, Rowe J, Tomkins-Tinch CH, et al. Bioconda: sustainable and comprehensive software distribution for the life sciences. Nat Methods. 2018 Jul; 15:475–476.

Gubin MM, Artyomov MN, Mardis ER, Schreiber RD. Tumor neoantigens: building a framework for personalized cancer immunotherapy. J Clin Invest. 2015 Sep; 125:3413–21.

Hackl H, Charoentong P, Finotello F, Trajanoski Z. Computational genomics tools for dissecting tumour-immune cell interactions. Nat Rev Genet. 2016 Aug; 17:441–58.

Hanahan D, Weinberg RA. The Hallmarks of Cancer. Cell. 2000 Jan; 100:57–70.

Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011 Mar; 144:646–74.

Hodi FS, O'Day SJ, McDermott DF, Weber RW, Sosman JA, Haanen JB, et al. Improved survival with ipilimumab in patients with metastatic melanoma. N Engl J Med. 2010 Aug; 363:711–23.

Hosomichi K, Shiina T, Tajima A, Inoue I. The impact of next-generation sequencing technologies on HLA research. J Hum Genet. 2015 Nov; 60:665–73.

Hugo W, Zaretsky JM, Sun L, Song C, Moreno BH, Hu-Lieskovan S, et al. Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. Cell. 2016 Mar; 165:35–44.

Hundal J, Carreno BM, Petti AA, Linette GP, Griffith OL, Mardis ER, et al. pVAC-Seq: A genome-guided in silico approach to identifying tumor neoantigens. Genome Med. 2016 Jan; 8:11.

Hundal J, Kiwala S, McMichael J, Miller CA, Xia H, Wollam AT, et al. pVACtools: A Computational Toolkit to Identify and Visualize Cancer Neoantigens. Cancer Immunol Res. 2020 Jan; 8:409–420.

Itakura E, Huang RR, Wen DR, Paul E, Wünsch PH, Cochran AJ. IL-10 expression by primary tumor cells correlates with melanoma progression from radial to vertical growth phase and development of metastatic competence. Mod Pathol. 2011 Jun; 24:801–809.

Janeway CA, Medzhitov R. Innate immune recognition. Annu Rev Immunol. 2002 Apr; 20:197–216.

Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M. NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. J Immunol. 2017 Nov; 199:3360–3368.

Kang K, Xie F, Mao J, Bai Y, Wang X. Significance of Tumor Mutation Burden in Immune Infiltration and Prognosis in Cutaneous Melanoma. Front Oncol. 2020 Sep; 10:573141.

Karasaki T, Nagayama K, Kuwano H, Nitadori Ji, Sato M, Anraku M, et al. Prediction and prioritization of neoantigens: integration of RNA sequencing data with whole-exome sequencing. Cancer Sci. 2017 Feb; 108:170–177.

Keir ME, Butte MJ, Freeman GJ, Sharpe AH. PD-1 and its ligands in tolerance and immunity. Annu Rev Immunol. 2008 Apr; 26:677–704.

Keşmir C, Nussbaum AK, Schild H, Detours V, Brunak S. Prediction of proteasome cleavage motifs by neural networks. Protein Eng. 2002 Apr; 15:287–96.

Kiyotani K, Mai TH, Nakamura Y. Comparison of exome-based HLA class I genotyping tools: identification of platform-specific genotyping errors. J Hum Genet. 2017 Mar; 62:397–405.

Köster J, Rahmann S. Snakemake-a scalable bioinformatics workflow engine. Bioinformatics. 2018 Oct; 34:3600.

Kurtzer GM, Sochat V, Bauer MW. Singularity: Scientific containers for mobility of compute. PLoS One. 2017 May; 12:e0177459.

Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009 Mar; 10:R25.

Lazdun Y, Si H, Creasy T, Ranade K, Higgs BW, Streicher K, et al. A New Pipeline to Predict and Confirm Tumor Neoantigens Predict Better Response to Immune Checkpoint Blockade. Mol Cancer Res. 2021 Mar; 19:498–506.

Leach DR, Krummel MF, Allison JP. Enhancement of antitumor immunity by CTLA-4 blockade. Science. 1996 Mar; 271:1734–6.

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016 Aug; 536:285–91.

Lennerz V, Fatho M, Gentilini C, Frye RA, Lifke A, Ferel D, et al. The response of autologous T cells to a human melanoma is dominated by mutated neoantigens. Proc Natl Acad Sci U S A. 2005 Nov; 102:16013–8.

Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv. 2013 May; 00:1–3.

Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009 Jul; 25:1754–60.

Li J, Liu Y, Kim T, Min R, Zhang Z. Gene expression variability within and between human populations and implications toward disease susceptibility. PLoS Comput Biol. 2010 Aug; 6:e1000910.

Liu X, Li C, Mou C, Dong Y, Tu Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. Genome Med. 2020 Dec; 12:103.

Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014 Dec; 15:550.

Lu YC, Robbins PF. Cancer immunotherapy targeting neoantigens. Semin Immunol. 2016 Feb; 28:22–7.

Marabelle A, Fakih M, Lopez J, Shah M, Shapira-Frommer R, Nakagawa K, et al. Association of tumour mutational burden with outcomes in patients with advanced solid tumours treated with pembrolizumab:prospective biomarker analysis of the multicohort, open-label, phase 2 KEYNOTE-158 study. Lancet Oncol. 2020 Oct; 21:1353–1365.

Marshall JS, Warrington R, Watson W, Kim HL. An introduction to immunology and immunopathology. Allergy Asthma Clin Immunol. 2018 Sep; 14:49.

Matey-Hernandez ML, Danish Pan Genome Consortium, Brunak S, Izarzugaza JMG. Benchmarking the HLA typing performance of Polysolver and Optitype in 50 Danish parental trios. BMC Bioinformatics. 2018 Dec; 19:239.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep; 20:1297–303.

Miller A, Asmann Y, Cattaneo L, Braggio E, Keats J, Auclair D, et al. High somatic mutation and neoantigen burden are correlated with decreased progression-free survival in multiple myeloma. Blood Cancer J. 2017 Sep; 7:e612–e612.

Ministério da Saúde. Instituto Nacional do Câncer José Alencar Gomes da Silva. Estimativa 2018/incidência de câncer no Brasil. Rio de Janeiro: INCA; 2019.

Ministério da Saúde. Instituto Nacional do Câncer José Alencar Gomes da Silva. Estimativa 2020/incidência de câncer no Brasil. Rio de Janeiro: INCA; 2020. Disponível em: https://bit.ly/3cgPAWx. [2021 jan 12].

Monach PA, Meredith SC, Siegel CT, Schreiber H. A unique tumor antigen produced by a single amino acid substitution. Immunity. 1995 Jan; 2:45–59.

Naslavsky MS, Yamamoto GL, de Almeida TF, Ezquina SAM, Sunaga DY, Pho N, et al. Exomic variants of an elderly cohort of Brazilians in the ABraOM database. Hum Mutat. 2017 Jul; 38:751–763.

Nelson BH. The impact of T-cell immunity on ovarian cancer outcomes. Immunol Rev. 2008 Apr; 222:101–16.

Nielsen M, Lundegaard C, Blicher T, Lamberth K, Harndahl M, Justesen S, et al. NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. PLoS One. 2007 Aug; 2:e796.

Nielsen M, Lundegaard C, Lund O, Keşmir C. The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage. Immunogenetics. 2005 Apr; 57:33–41.

NovoCraft. *Novoalign*. 2020. Http://www.novocraft.com/products/novoalign/ [Online; acessado em 03/07/2020].

Oiseth SJ, Aziz MS. Cancer immunotherapy: a brief review of the history, possibilities, and challenges ahead. J Cancer Metastasis Treat. 2017 Oct; 3:250.

Oliva M, Muñoz-Aguirre M, Kim-Hellmuth S, Wucher V, Gewirtz ADH, Cotter DJ, et al. The impact of sex on gene expression across human tissues. Science. 2020 Sep; 369:eaba3066.

Ortiz R, Melguizo C, Prados J, J Alvarez P, Caba O, Rodriguez-Serrano F, et al. New Gene Therapy Strategies for Cancer Treatment: A Review of Recent Patents. Recent Pat Anticancer Drug Discov. 2012 Jul; 7:297–312.

Pagès F, Galon J, Dieu-Nosjean MC, Tartour E, Sautès-Fridman C, Fridman WH. Immune infiltration in human tumors: a prognostic factor that should not be ignored. Oncogene. 2010 Feb; 29:1093–102.

Parrot T, Allard M, Oger R, Benlalam H, Raingeard de la Blétière D, Coutolleau A, et al. IL-9 promotes the survival and function of human melanoma-infiltrating CD4 + CD8 + double-positive T cells. Eur J Immunol. 2016 Jul; 46:1770–1782.

Richters MM, Xia H, Campbell KM, Gillanders WE, Griffith OL, Griffith M. Best practices for bioinformatic characterization of neoantigens for clinical utility. Genome Med. 2019 Dec; 11:56.

Riviere P, Goodman AM, Okamura R, Barkauskas DA, Whitchurch TJ, Lee S, et al. High Tumor Mutational Burden Correlates with Longer Survival in Immunotherapy-Naïve Patients with Diverse Cancers. Mol Cancer Ther. 2020 Oct; 19:2139–2145.

Rizvi NA, Hellmann MD, Snyder A, Kvistborg P, Makarov V, Havel JJ, et al. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. Science. 2015 Apr; 348:124–8.

Robert C, Schachter J, Long GV, Arance A, Grob JJ, Mortier L, et al. Pembrolizumab versus Ipilimumab in Advanced Melanoma. N Engl J Med. 2015 Jun; 372: 2521–32.

Robertson AG, Kim J, Al-Ahmadie H, Bellmunt J, Guo G, Cherniack AD, et al. Comprehensive Molecular Characterization of Muscle-Invasive Bladder Cancer. Cell. 2017 Oct; 171:540–556.e25.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011 Jan; 29:24–6.

van Rooij N, van Buuren MM, Philips D, Velds A, Toebes M, Heemskerk B, et al. Tumor exome analysis reveals neoantigen-specific T-cell reactivity in an ipilimumab-responsive melanoma. J Clin Oncol. 2013 Nov; 31:e439–42.

Schadendorf D, Hodi FS, Robert C, Weber JS, Margolin K, Hamid O, et al. Pooled Analysis of Long-Term Survival Data From Phase II and Phase III Trials of Ipilimumab in Unresectable or Metastatic Melanoma. J Clin Oncol. 2015 Jun; 33:1889–94.

Shao C, Li G, Huang L, Pruitt S, Castellanos E, Frampton G, et al. Prevalence of High Tumor Mutational Burden and Association With Survival in Patients With Less Common Solid Tumors. JAMA Netw open. 2020 Oct; 3:e2025109.

Shiina T, Hosomichi K, Inoko H, Kulski JK. The HLA genomic loci map:expression, interaction, diversity and disease. J Hum Genet. 2009 Jan; 54:15–39.

Shukla SA, Rooney MS, Rajasagi M, Tiao G, Dixon PM, Lawrence MS, et al. Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. Nat Biotechnol. 2015 Nov; 33:1152–8.

Smith LK, Boukhaled GM, Condotta SA, Mazouz S, Guthmiller JJ, Vijay R, et al. Interleukin-10 Directly Inhibits CD8+ T Cell Function by Enhancing N-Glycan Branching to Decrease Antigen Sensitivity. Immunity. 2018 Feb; 48:299–312.e5.

Snyder A, Makarov V, Merghoub T, Yuan J, Zaretsky JM, Desrichard A, et al. Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma. N Engl J Med. 2014 Dec; 371:2189–2199.

Steen CB, Liu CL, Alizadeh AA, Newman AM. Profiling cell type abundance and expression in bulk tissues with CIBERSORTx. Methods Mol Biol. 2020 Nov; 2117:135–157.

Stenzinger A, Allen JD, Maas J, Stewart MD, Merino DM, Wempe MM, et al. Tumor mutational burden standardization initiatives: Recommendations for consistent tumor mutational burden assessment in clinical samples to guide immunotherapy treatment decisions. Genes, Chromosom Cancer. 2019 Aug; 58:578–588.

Szolek A, Schubert B, Mohr C, Sturm M, Feldhahn M, Kohlbacher O. OptiType: precision HLA typing from next-generation sequencing data. Bioinformatics. 2014 Dec; 30:3310–6.

Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou Yang TH, et al. The Immune Landscape of Cancer. Immunity. 2018 Apr; 48:812–830.e14.

Turck N, Vutskits L, Sanchez-Pena P, Robin X, Hainard A, Gex-Fabry M, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011 Mar; 8:12–77.

Uehara T, Fujiwara T, Takeda K, Kunisada T, Ozaki T, Udono H. Immunotherapy for Bone and Soft Tissue Sarcomas. Biomed Res Int. 2015 Nov; 2015:820813.

Van Allen EM, Miao D, Schilling B, Shukla SA, Blank C, Zimmer L, et al. Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. Science. 2015 Oct; 350:207–211.

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinforma. 2013 Oct; 43: 11.10.1–11.10.33.

Van der Auwera, GA and O'Connor, DB. *Genomics in the Cloud:Using Docker, GATK, and WDL in Terra (1st Edition)*. [S.l.]: O'Reily Media, 2020.

Waldman AD, Fritz JM, Lenardo MJ. A guide to cancer immunotherapy: from T cell basic science to clinical practice. Nat Rev Immunol. 2020 Nov; 20:651–668.

Walko C, Kudchadkar R. *Evolving Role of Immunotherapy in Cancer Treatment.* 2018. Disponível em: https://www.pharmacytimes.com/conferences/ashpmidyear2018/immunotherapy-managing-immune-related-toxicities-in-patients [2020 07 03].

Wan J, Wu Y, Ji X, Huang L, Cai W, Su Z, et al. IL-9 and IL-9-producing cells in tumor immunity. Cell Commun Signal. 2020 Dec; 18:50.

Wang P, Zhang X, Sun N, Zhao Z, He J. Comprehensive Analysis of the Tumor Microenvironment in Cutaneous Melanoma associated with Immune Infiltration. J Cancer. 2020 Apr; 11:3858–3870.

Weese D, Holtgrewe M, Reinert K. RazerS 3faster, fully sensitive read mapping. Bioinformatics. 2012 Oct; 28:2592–9.

Wood LD, Parsons DW, Jones S, Lin J, Sjöblom T, Leary RJ, et al. The genomic landscapes of human breast and colorectal cancers. Science. 2007 Nov; 318:1108–13.

Wu HX, Wang ZX, Zhao Q, Chen DL, He MM, Yang LP, et al. Tumor mutational and indel burden: a systematic pan-cancer evaluation as prognostic biomarkers. Ann Transl Med. 2019 Nov; 7:640.

Wyld L, Audisio RA, Poston GJ. The evolution of cancer surgery and future perspectives. Nat Rev Clin Oncol. 2015 Feb; 12:115–24.

Yoo AB, Jette MA, Grondona M. SLURM: Simple Linux Utility for Resource Management. In: *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*. Springer Berlin Heidelberg, 2003. 2862, 44–60. ISBN 9783540397274. Disponível em: http://link.springer.com/10.1007/10968987_3.

Zheng S, Cherniack AD, Dewal N, Moffitt RA, Danilova L, Murray BA, et al. Comprehensive Pan-Genomic Characterization of Adrenocortical Carcinoma. Cancer Cell. 2016 May; 29:723–736.

Zhou C, Wei Z, Zhang Z, Zhang B, Zhu C, Chen K, et al. pTuneos: prioritizing tumor neoantigens from next-generation sequencing data. Genome Med. 2019 Dec; 11:67.

Zhou J, Dudley ME, Rosenberg SA, Robbins PF. Persistence of multiple tumor-specific T-cell clones is associated with complete tumor regression in a melanoma patient receiving adoptive cell transfer therapy. J Immunother. 2005 Jan; 28: 53–62.

APÊNDICES

Apêndice 1 – Código para estratificação de nível de expressão

```
library ( mclust )

load _ fpkmr <- function ( fn ) {

  expr <- read . table (fn , header =T , stringsAsFactors = F )

  expr <- expr [ ! duplicated ( expr $ gene _ id ) ,]

  silent <- expr $ gene _ id [ expr $ FPKM 0]
  notSilent <- subset ( expr ,! gene _ id % in % silent )
  notSilent $ log10FPKM <- log10 ( notSilent $ FPKM )

  m <- Mclust ( notSilent $ log10FPKM , G =2)

  expressed <- notSilent $ gene _ id [ m $ classification == 2]

  trace <- notSilent $ gene _ id [ m$ classification == 1]

  tert <- quantile (
    notSilent $ log10FPKM [ notSilent $ gene _ id% in %expressed ],
    probs =c (0 ,1 / 3 ,2 / 3 ,1)
  )

  expr $ log10FPKM <- log10 ( expr $ FPKM )

  expr $ class <- " silent "
  expr $ class [ expr $ FPKM 0] <- " trace "
  expr $ class [ expr $ log10FPKM> tert [1]] <- " low "
  expr $ class [ expr $ log10FPKM> tert [2]] <- " medium "
  expr $ class [ expr $ log10FPKM> tert [3]] <- " high "
```

```
31   expr $ class <- factor (
32     expr $ class ,
33     levels =c(" silent " ," trace " ," low " ," medium " ," high ")
34   )
35
36   return  ( expr )
37 }
38
39 args  = commandArgs ( trailingOnly =T )
40
41 files  <- list . files ( args [1] ,full . names = T )
42 txtfiles  <- files [ grep ( glob2rx ( " * . FPKM . txt " ) , files )]
43 output _ path = paste ( args [1] ,'/ levels / ' ,sep = "")
44
45 for ( fin txtfiles ) {
46   expr  <- load _ fpkmr (f)
47   arquivo  <- tail ( strsplit (f ,'/ ') [[1]] ,1)
48   expr [ ' sample ']<- sub (
49     pattern = " (. * ?) \\.. * $" ,replacement = ' \\1 ', x= arquivo )
50   dir . create ( output _ path ,showWarnings =F ,  recursive =T )
51   write . table (
52     expr , file = paste ( output _ path , arquivo , sep = ' ') , sep="\ t" ,
      names =F , quote = F)
53 }
```

Apêndice 2 – Parâmetros da ferramenta GADO

```
usage : gdc_download . py  [ - h][-- version ] { gdc , icgc , cbioportal , venn - diagram } ...

Download  files  from  several  data  portals  using  API

optional  arguments :
  -h , -- help              show  this  help  message  and  exit
  -- version                show  program 's  version  number  and  exit

Commands :
  Commands  available

  { gdc , icgc , cbioportal , venn - diagram }
                            Use  a  subcommand  to  show  its  options .
    gdc                     GDC  Data  Portal  API  options
    icgc                    ICGC  Data  Portal  API  options
    cbioportal              cBioPortal  API  options
```

```
# !/ bin / bash
usage : gdc_download . py  gdc  [ -h] [-- category  CATEGORY  [ CATEGORY ...]]  [ -- type TYPE  [ TYPE ...]]  [ -- strategy  STRATEGY  [ STRATEGY ...]]  [ -- workflow  WORKFLOW  [ WORKFLOW  ...]]  [-- data_format DATA_FORMAT  [ DATA_FORMAT  ...]]
                            [ -- projects  PROJECTS  [ PROJECTS  ...]]  [-- access  { open , controlled , all }][ - - sample_type { all , tumor , normal }]  [ - - overwrite ][- - depth { category , type , strategy , sample_type , workflow }]  [ -- token TOKEN ]
                            [ -- user - projects USER_PROJECTS ]  [- - slicing  SLICING ]  [ - - cases - file CASES_FILE ]  [- - md5checksum ] [ -- by - project ] [ -- threads  THREADS ]  [- - output OUTPUT ]  [-- verbose ]
```

```
                              P [P ...]

positional arguments :
  P                       Names of Primary Tissue from GDC
    separated by comma . Use " all "to get all projects from GDC .

optional arguments :
  -h , -- help            show this help message and exit
  -- category CATEGORY [ CATEGORY ...] , -c CATEGORY [ CATEGORY ...]
                          Data category of GDC ( default : [ ' Clinical
    ' ])
  -- type TYPE [ TYPE ...] , -t TYPE [ TYPE ...]
                          Data Type of GDC ( default : None )
  -- strategy STRATEGY [ STRATEGY ...] , -s STRATEGY [ STRATEGY ...]
                          Experimental Strategy of GDC ( default :
    None )
  -- workflow WORKFLOW [ WORKFLOW ...] , -w WORKFLOW [ WORKFLOW ...]
                          Workflow Type of GDC ( default : None )
  -- data_format DATA_FORMAT [ DATA_FORMAT ...] , -f DATA_FORMAT [
    DATA_FORMAT ...]
                          Data Format of files ( default : None )
  -- projects PROJECTS [ PROJECTS ...] , -p PROJECTS [ PROJECTS ...]
                          Projects ID to filter ( default : None )
  -- access { open , controlled , all } -a { open , controlled , all }
                          Access level of file on GDC ( default :
    open )
  -- sample_type { all , tumor , normal }
                          Type of sample to download . ( default : all
    )
  -- overwrite , -e       Overwrite existing files ( default : False )
  -- depth { category , type , strategy , sample_type , workflow } -d {
    category , type , strategy , sample_type , workflow }
                          Depth of path where files will be saved .
    ( default : category )
```

```
  -- token TOKEN            Token file to download controlled - access files from GDC .
  -- user - projects USER_PROJECTS
                            A JSON file with all projects accessible by the user on GDC . See : https :// docs . gdc . cancer . gov / Data / Data_Security / Data_Security
  -- slicing SLICING , -l SLICING
                            A JSON file with regions to download . See : https :// docs . gdc . cancer . gov / API / Users_Guide / BAM_Slicing /
  -- cases - file CASES_FILE
                            A file with cases_id to be download .
  -- md5checksum , -k       MD5 checksum on downloaded file ( default : False )
  --by - project , -j        Save files by project ( default : False )
  -- threads THREADS        Number of threads to download files ( default : 2)
  -- output OUTPUT , -o OUTPUT
                            Output path for files ( default : / mnt / data4 / lbcb / students / Alexandre . Defelicibus / gdc - api )
  -- verbose , -v           Run with verbose logging ( default : False )
```

```
# !/ bin / bash
usage : gdc_download . py icgc [-h ] [- - type TYPE [ TYPE ...]] [ - - strategy STRATEGY [ STRATEGY ...]] [ -- analysis ANALYSIS [ ANALYSIS ...]] [ - - access{ open , controlled }] [ - - sample_type { all , tumor , normal }][ - - overwrite ]
                                [ - - depth{ type , strategy , sample_type , workflow }] [ - - md5checksum ][ -- project ][ - - threadsTHREADS ] [ - - output OUTPUT ] [ - - verbose ]
                                P [ P ...]

positional arguments :
  P                         Names of Primary Tissue from ICGC separated by comma . Use " all "to get all projects from ICGC .

```

```
optional arguments :
  -h , -- help              show this help message and exit
  -- type TYPE [ TYPE ...] , -t TYPE [ TYPE ...]
                            Data type of ICGC ( default : [ ' Clinical
    Data '])
  -- strategy STRATEGY [ STRATEGY ...] , -s STRATEGY [ STRATEGY ...]
                            Experimental Strategy of ICGC ( default :
    None )
  -- analysis ANALYSIS [ ANALYSIS ...] , -a ANALYSIS [ ANALYSIS ...]
                            Analysis Type of ICGC ( default : None )
  -- access { open , controlled } ,-l { open , controlled }
                            Access level of file on ICGC ( default :
    open )
  -- sample_type { all , tumor , normal }
                            Type of sample to download . ( default : all
    )
  -- overwrite , -e         Overwrite existing files ( default : False )
  -- depth { type , strategy , sample_type , workflow } ,-d { type , strategy
    , sample_type , workflow }
                            Depth of path where files will be saved .
    ( default : type )
  -- md5checksum , -k       MD5 checksum on downloaded file ( default :
     False )
  -- project , -j           Save files by project ( default : False )
  -- threads THREADS        Number of threads to download files (
    default : 30)
  -- output OUTPUT , -o OUTPUT
                            Output path for files ( default : / mnt /
    data4 / lbcb / students / Alexandre . Defelicibus / gdc - api )
  -- verbose , -v           Run with verbose logging ( default : False )
```

```
# !/ bin / bash
usage : gdc_download . py cbioportal [ -h ][ - - config - init[]-- config -
    file CONFIG_FILE ] [ - - category CATEGORY [ CATEGORY ...]] [ --
    overwrite ] [ -- project ] [-- threads THREADS ] [- - output OUTPUT ]
```

```
    [-- verbose ] C [C ...]

positional arguments :
  C                     Codes of Cancer Studies from cBioPortal
    separated by comma . Use " all "to get all Cancer Studies from
    cBioPortal .

optional arguments :
  -h , -- help          show this help message and exit
  -- config - init ,-i  Create a config file with code of cancer
    studies and tissue from cBioPortal using cancer_studies . yaml .
    sample ( default : False )
  -- config - file CONFIG_FILE , -f CONFIG_FILE
                        Config file with information about cancer
     studies ( default : cancer_studies . yaml . sample )
  -- category CATEGORY [ CATEGORY ...] , -c CATEGORY [ CATEGORY ...]
                        Data category of cBioPortal [ Clinical ,
    Mutation , Expression ] ( default : [ ' Clinical ' ])
  -- overwrite , -e     Overwrite existing files ( default : False )
  -- project , -j       Save files by project ( default : False )
  -- threads THREADS    Number of threads to download files (
    default : 30)
  -- output OUTPUT , -o OUTPUT
                        Output path for files ( default : / mnt /
    data4 / lbcb / students / Alexandre . Defelicibus / gdc - api )
  -- verbose , -v       Run with verbose logging ( default : False )
```

Apêndice 3 – Cálculo do *score* por subtipo imune e molecular

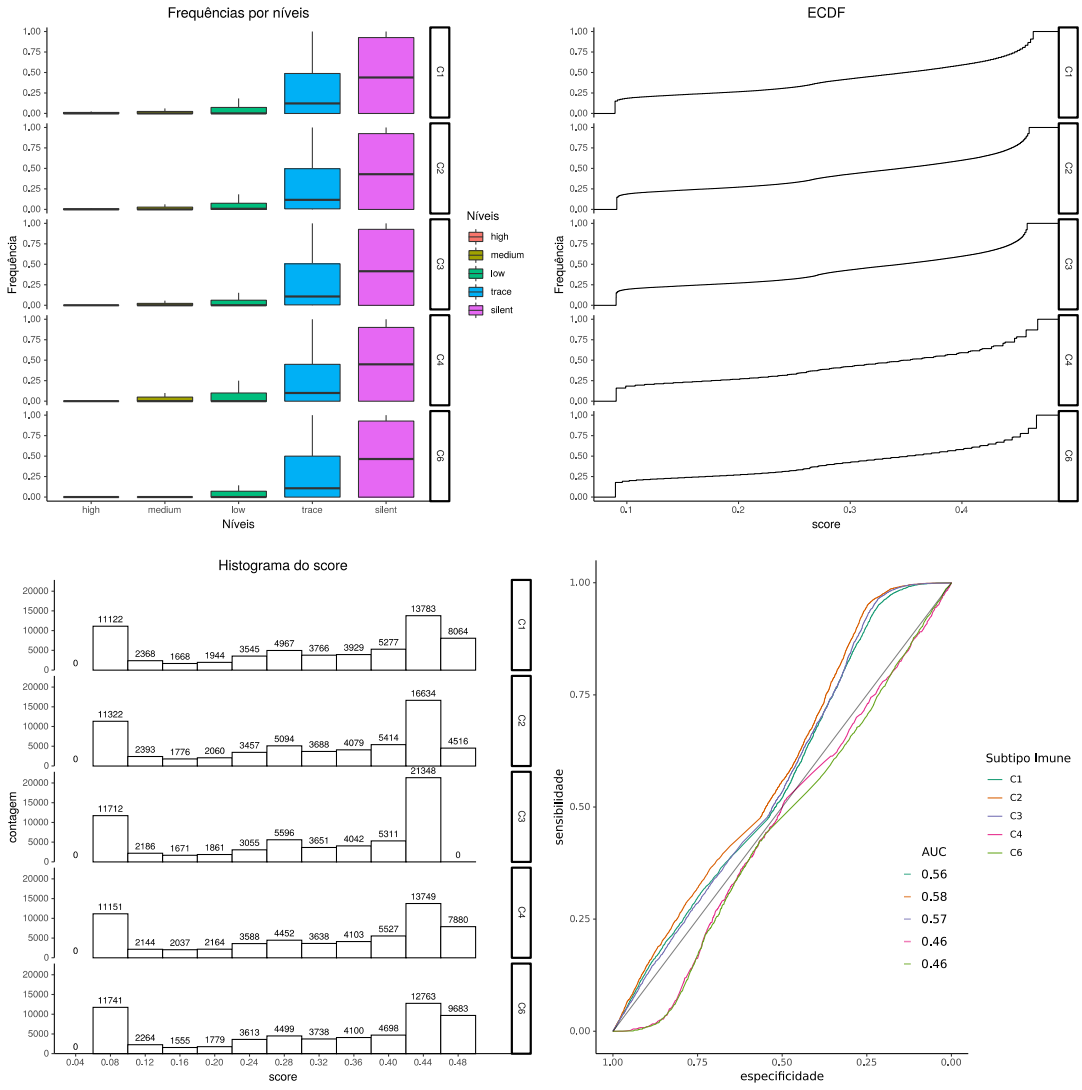


Figura 55 – Análise do cálculo do *score* para o projeto TCGA-LUAD por subtipo imune. O primeiro gráfico à esquerda ilustra a distribuição da frequência relativa para cada nível de expressão e subtipo imune. Abaixo, tem o histograma de ocorrências por valor de *score* e subtipo imune e à direita tem o ECDF do *score* por subtipo imune. Por fim, abaixo tem o gráfico com uma curva ROC e uma AUC para cada subtipo imune.

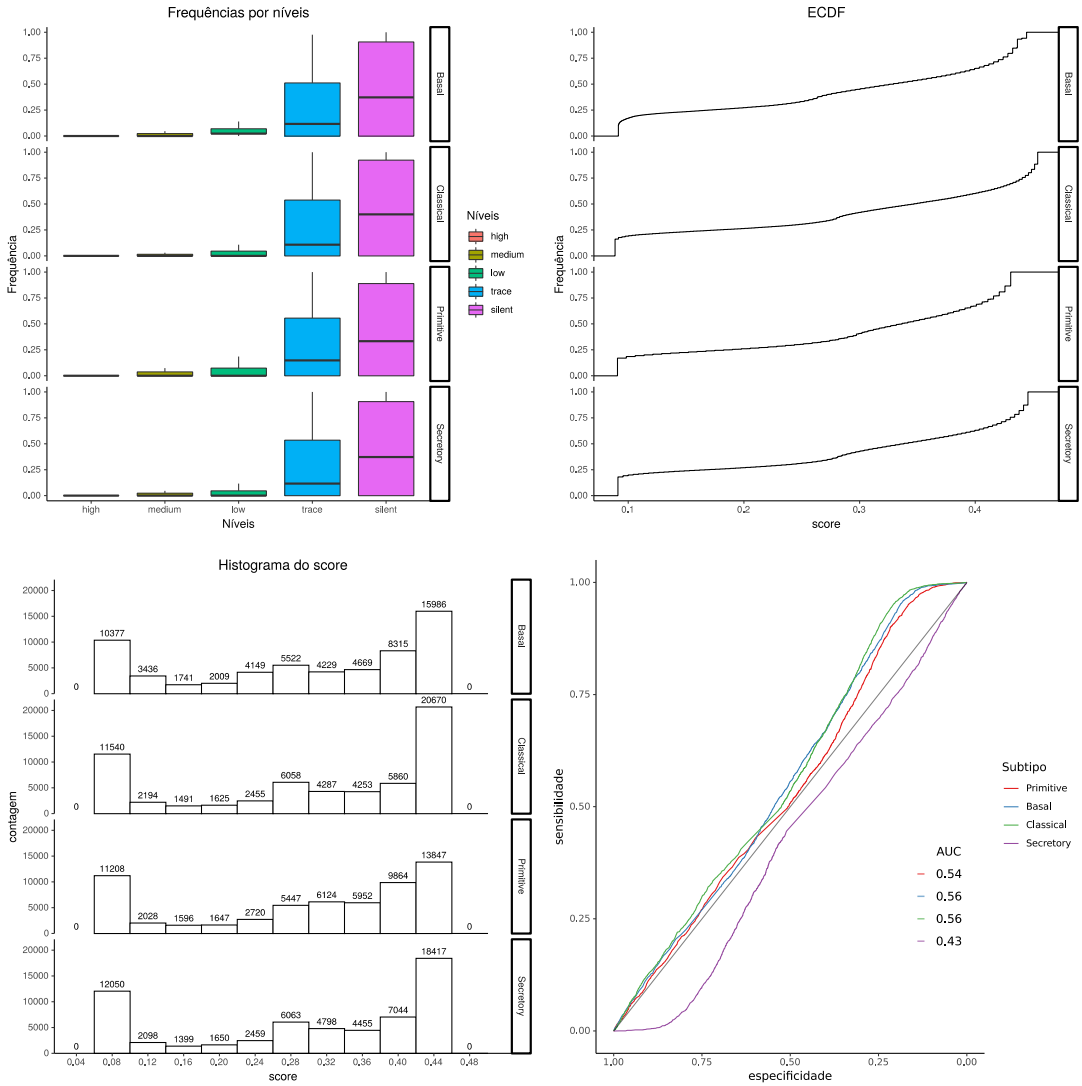


Figura 56 – Análise do cálculo do *score* para o projeto TCGA-LUSC por subtipo. O primeiro gráfico à esquerda ilustra a distribuição da frequência relativa para cada nível de expressão e subtipo. Abaixo, tem o histograma de ocorrências por valor de *score* e subtipo e à direita tem o ECDF do *score* por subtipo. Por fim, abaixo tem o gráfico com uma curva ROC e uma AUC para cada subtipo.

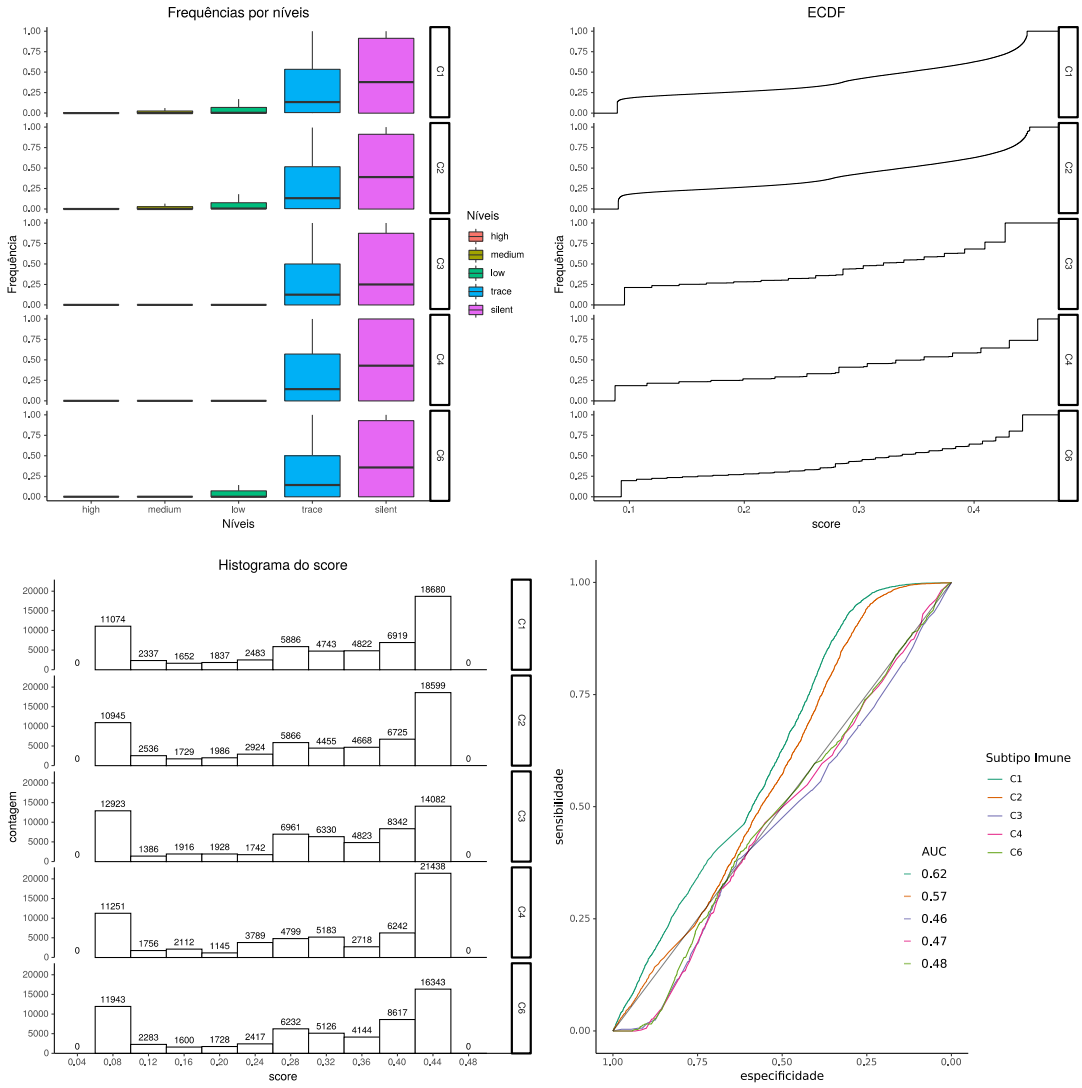


Figura 57 – Análise do cálculo do *score* para o projeto TCGA-LUSC por subtipo imune. O primeiro gráfico à esquerda ilustra a distribuição da frequência relativa para cada nível de expressão e subtipo imune. Abaixo, tem o histograma de ocorrências por valor de *score* e subtipo imune e à direita tem o ECDF do *score* por subtipo imune. Por fim, abaixo tem o gráfico com uma curva ROC e uma AUC para cada subtipo imune.

Apêndice 4 – Outras análises entre TMB-high e TMB-low

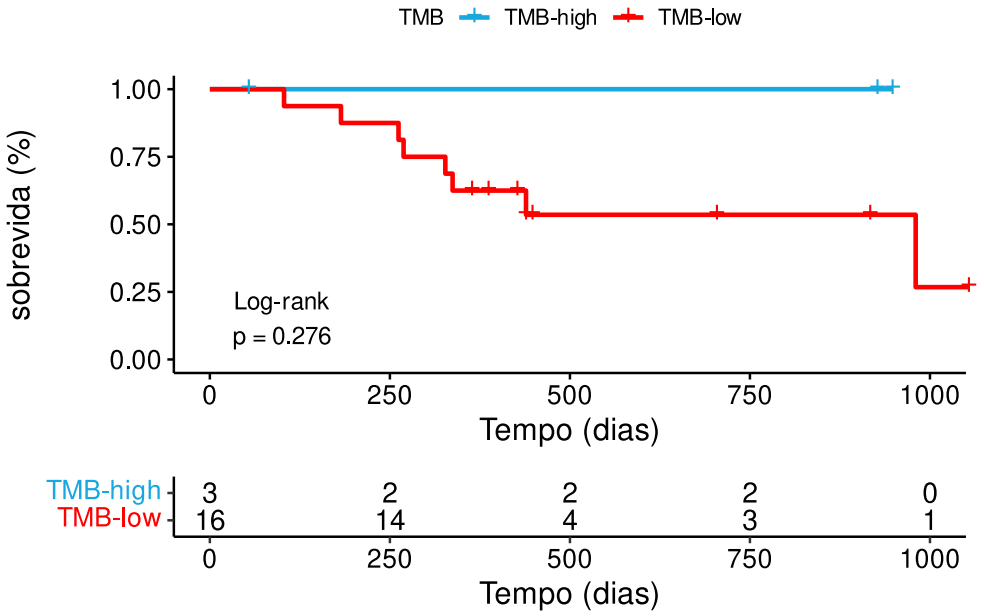


Figura 58 – Análise de sobrevida global da coorte Hugo et al. (Sobrevida global entre os pacientes dos grupos TMB-high (n=3) e TMB-low(n=16), sem diferença estatisticamente significativa (*p-value*=0.276).

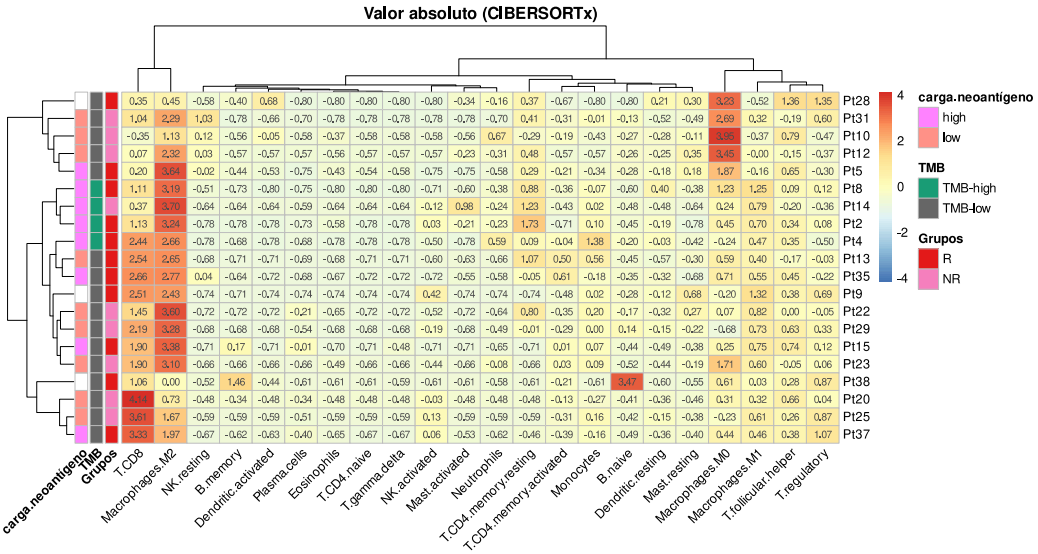


Figura 59 – *Heatmap* com a composição de células imunes na coorte Hugo et al. Composição do valor absoluto estimado pelo CIBERSORTx para cada uma das amostras, clusterizando pelas células (colunas) e pelas amostras (linhas).

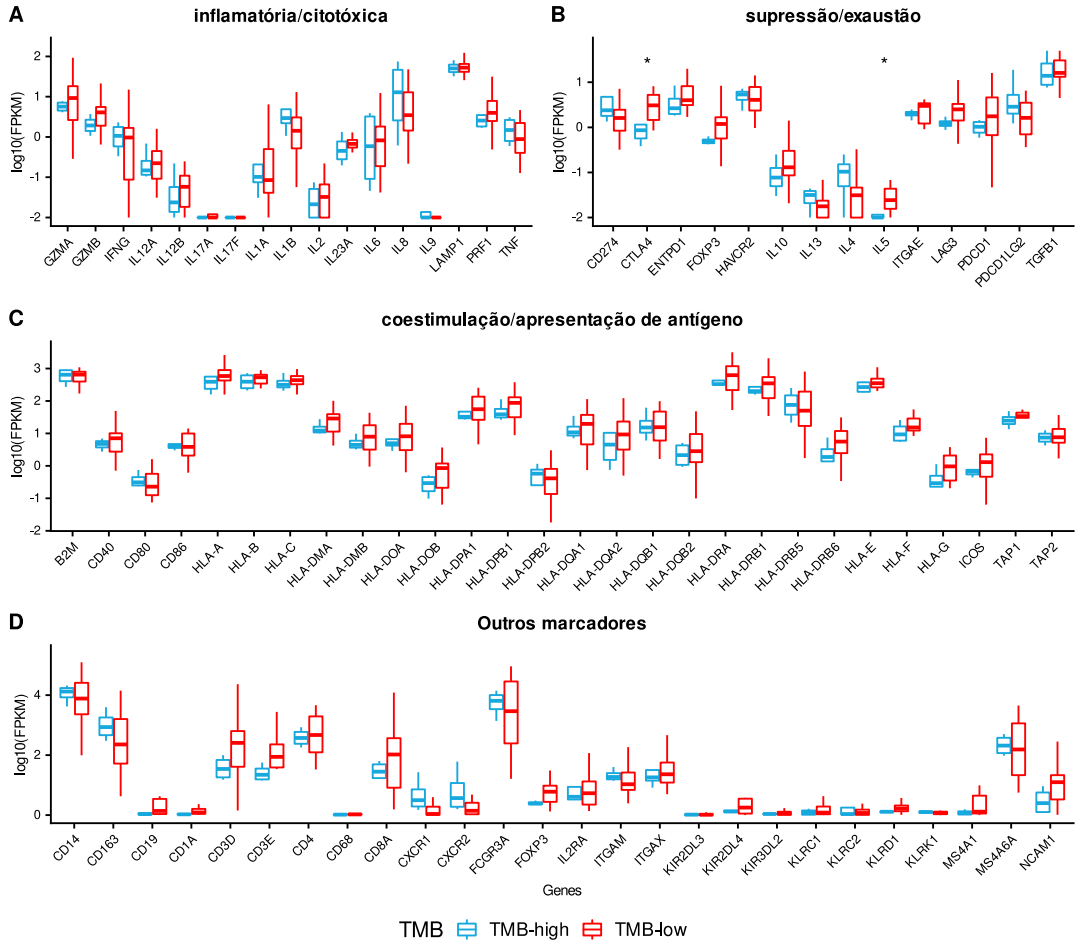


Figura 60 – Expressão de genes marcadores entre os grupos TMB-high e TMB-low. São quatro listas com marcadores de resposta imune diferentes, comparando a expressão (em $log10$) dos genes entre os grupos TMB-high e TMB-low:(A) inflamatória/citotóxica; (B) supressão/exaustão; (C) coestimulação/apresentação de antígeno; (D) outros marcadores. * $p \leq 0.05$.

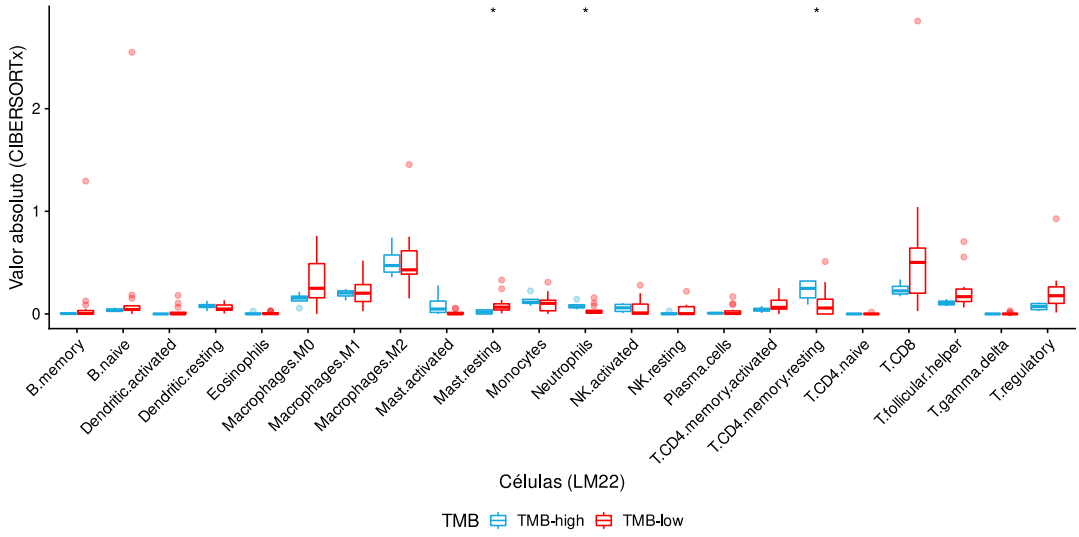


Figura 61 – Distribuição da composição de células imunes na coorte Hugo et al para os grupos TMB-high e TMB-low estimado pelo CIBERSORTxComparação da composição de 22 células do sistema imune nos grupos TMB-high e TMB-low. * $p \leq 0.05$.