

**IMPACTO DA ANCESTRALIDADE NA PROGRESSÃO E
BIOLOGIA DO MELANOMA E CARCINOMA DE PULMÃO
DE CÉLULAS NÃO PEQUENAS**

FELIPE DE AZEVEDO OLIVEIRA

**Dissertação apresentada à Fundação Antônio Prudente para
obtenção do Título de Mestre em Ciências.**

Área de concentração: Oncologia

Orientador: Dr. Israel Tojal da Silva

Coorientador: Dr. Kenneth John Gollob

São Paulo

2021

FICHA CATALOGRÁFICA

Preparada pelo Ensino Apoio ao aluno da Fundação Antônio Prudente*

O48i Oliveira, Felipe de Azevedo

Impacto da ancestralidade na progressão e biologia do melanoma e carcinoma de pulmão de células não pequenas / Felipe de Azevedo Oliveira – São Paulo, 2021.

106p.

Dissertação (Mestrado)-Fundação Antônio Prudente.

Curso de Pós-Graduação em Ciências - Área de concentração: Oncologia.

Orientador: Israel Tojal da Silva

Descritores. 1. Carcinoma de Pulmão de Células Não Pequenas/Non-Small Cell Lung Carcinoma. 2. Biologia Computacional/Computational Biology. 3. Melanoma/Melanoma. 4. Pesquisa em genética/Genetic Research.

Elaborado por Suely Francisco CRB 8/2207

*Todos os direitos reservados à FAP. A violação dos direitos autorais constitui crime, previsto no art. 184 do Código Penal, sem prejuízo de indenizações cabíveis, nos termos da Lei nº 9.610/08.

SUPORTE À PESQUISA POR AGÊNCIA DE FOMENTO

Este trabalho recebeu apoio da Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP), através de auxílio à Pesquisa – Mestrado (nº 2018/26006-9) e do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), através de auxílio à Pesquisa – Mestrado (nº 132359/2019-7).

AGRADECIMENTOS

A minha mãe e a minha irmã, Waldeli e Fernanda, por todo o apoio, carinho e paciência durante este período.

Aos meus avós, Waldir e Sueli, pelos ensinamentos, carinho e compreensão.

Aos meus amigos, pelas lembranças, pela companhia e por todo o apoio.

A todos os colegas do Laboratório de Biologia Computacional e Bioinformática, pela parceria, por todo aprendizado e ajuda.

Ao meu orientador, Israel, pela oportunidade, pelo tempo dedicado à este trabalho e pela participação na minha formação acadêmica.

RESUMO

Oliveira FA. **Impacto da ancestralidade na progressão e biologia do melanoma e carcinoma de pulmão de células não pequenas.** [Dissertação]. São Paulo; Fundação Antônio Prudente; 2021.

Um grande número de estudos indica que a diversidade genética entre populações pode levar a diferenças em múltiplos aspectos como: inflamação crônica, resposta a antígenos e no padrão de mutações. Estas diferenças seriam parcialmente responsáveis por alterações em fenótipos e na suscetibilidade a doenças. O carcinoma de pulmão de células não pequenas (CPCNP) possui fortes componentes genéticos e ambientais, sendo o seu principal fator de risco o tabagismo. Segundo dados de 2012, o câncer de pulmão foi o tipo de câncer mais incidente no mundo, e foi observado nos Estados Unidos, uma maior incidência em afrodescendentes. Já o melanoma tem como seu principal fator de risco a exposição excessiva à radiação solar ultravioleta (UV), além do histórico familiar. Neste cenário, torna-se relevante avaliar a relação entre ancestralidade e o seu impacto na saúde. Para isso, o presente trabalho busca elucidar esta relação em casos de melanoma e CPCNP e contribuir para a identificação de possíveis marcadores. Dados de transcriptoma, sequenciamento completo de exoma e clínicos foram obtidos a partir do *The Cancer Genome Atlas* (TCGA). Com o intuito de inferir ancestralidade individual foi utilizado o *software* *Admixture* e foi construído um *pipeline* automatizado através do gerenciador *Snakemake*. Além disso, também foi avaliado o infiltrado inflamatório, expressão gênica, perfil mutacional somático e assinatura mutacional. A integração destes dados foi realizada através da aplicação dos modelos de predição, LASSO e Elastic Net. Em relação ao infiltrado inflamatório, assinatura mutacional, perfil mutacional somático e expressão gênica, foram observadas diferenças importantes entre os grupos avaliados, independentemente do tipo de tumor. Constatou-se, em indivíduos com ancestralidade africana, uma maior frequência de mutações em TP53 e maiores níveis de linfócitos T e B. Em relação à ancestralidade europeia, observou-se maiores níveis de monócitos e macrófagos M2. A aplicação dos modelos de predição foi capaz de identificar genes importantes, destacando-se: MASTL, TP73, FAAH, PWP2, AHRR, COL22A1, ACBD3, PLEK2, EEPD1 e ASH2L. Como um todo, este estudo reforça a importância da inferência de ancestralidade individual e sugere diferenças importantes em cada um dos grupos, podendo impactar diretamente na sobrevida de pacientes.

Descritores: Carcinoma de Pulmão de Células Não Pequenas. Melanoma. Biologia Computacional. Pesquisa em Genética.

ABSTRACT

Oliveira FA. **[Impact of ancestry on the progression and biology of melanoma and non-small cell lung carcinoma]**. [Dissertation]. São Paulo; Fundação Antônio Prudente; 2021

A large number of studies indicates that genetic diversity between populations can lead to differences in multiple aspects such as: chronic inflammation, response to antigens and pattern of mutations. These differences would be partly responsible for changes in phenotypes and disease susceptibility. The non-small cell lung carcinoma (NSCLC). has strong genetic and environmental components, and its main risk factor is smoking. According to data from 2012, lung cancer was the most common type of cancer in the world, and it was observed in the United States a higher incidence in Afro-descendants. On the other hand, melanoma's main risk factor is excessive exposure to ultraviolet (UV) solar radiation, in addition to the family history. In this scenario, it becomes relevant to assess the relationship between ancestry and its impact on health. To this end, the present study seeks to elucidate this relationship in cases of melanoma and NSCLC and to contribute to the identification of possible markers. Transcriptome, whole exome sequencing and clinical data were obtained from The Cancer Genome Atlas (TCGA). In order to infer individual ancestry, the software Admixture was used, and an automated pipeline was built using Snakemake. In addition, inflammatory infiltrate, gene expression and somatic mutational profile and mutational signature were also evaluated. The integration of these data was accomplished through the application of regression models, LASSO and Elastic Net. Regarding the inflammatory infiltrate, mutational signature, somatic mutation profile and gene expression, important differences were observed between the groups evaluated, regardless of the type of tumor. It was found, in individuals with African ancestry, a higher frequency of mutations in TP53 and higher levels of T and B lymphocytes. In relation to European ancestry, higher levels of monocytes and macrophages M2 were observed. The use of regression models was able to identify important genes, highlighting: MASTL, TP73, FAAH, PWP2, AHRR, COL22A1, ACBD3, PLEK2, EEPD1 e ASH2L. As a whole, this study reinforces the importance of the inference of individual ancestry and suggests important differences in each of the groups, which may directly impact on patient survival.

Keywords: Non-Small Cell Lung Carcinoma. Melanoma. Computational Biology. Genetic Research.

LISTA DE FIGURAS

Figura 1	Desenho do estudo	20
Figura 2	<i>Workflow</i> para identificação de ancestralidade em população mista a partir do banco com as populações de referência	31
Figura 3	<i>Workflow</i> do <i>pipeline</i> proposto para ancestralidade individual	32
Figura 4	Ancestralidade individual em pacientes TCGA	33
Figura 5	Comparação entre os dados obtidos a partir do <i>pipeline</i> de ancestralidade individual e o TCGAA, para pacientes de adenocarcinoma gástrico do TCGA pelo método <i>Bland-Altman</i>	35
Figura 6	Comparação da ancestralidade individual do banco de referência com 534.734 variantes e a ancestralidade individual do banco de referência com variantes selecionadas por LEI, em conjuntos de 5000, 4000, 3000, 2000, 1000, 500 e 100 variantes.....	38
Figura 7	Comparação dos níveis de células imunes e estromais entre os clusters 1, 2 e 3 em pacientes de adenocarcinoma gástrico, aplicando os softwares <i>xCell</i> , <i>CIBERSORT</i> e <i>ESTIMATE</i>	40
Figura 8	Comparação dos níveis de células imunes e estromais entre os clusters 1, 2 e 3 em pacientes de adenocarcinoma de pulmão, aplicando os softwares <i>xCell</i> , <i>CIBERSORT</i> e <i>ESTIMATE</i>	41
Figura 9	Comparação dos níveis de células imunes e estromais entre os clusters 1, 2 e 3 em pacientes de carcinoma de células escamosas de pulmão, aplicando os softwares <i>xCell</i> , <i>CIBERSORT</i> e <i>ESTIMATE</i>	42
Figura 10	Genes significativamente mutados em pacientes de adenocarcinoma gástrico utilizando <i>MutSi</i>	44

Figura 11	Genes significativamente mutados em pacientes de adenocarcinoma de pulmão utilizando <i>MutSig</i>	45
Figura 12	Genes significativamente mutados em pacientes de carcinoma de células escamosas de pulmão utilizando <i>MutSig</i>	46
Figura 13	Genes significativamente mutados em pacientes de melanoma no <i>cluster 1</i> (ancestralidade europeia), utilizando <i>MutSig</i>	47
Figura 14	Assinaturas mutacionais em pacientes de adenocarcinoma gástrico	48
Figura 15	Assinaturas mutacionais em pacientes de melanoma.....	50
Figura 16	Assinaturas mutacionais em pacientes de carcinoma de pulmão.....	52
Figura 17	Assinaturas mutacionais em pacientes de carcinoma de células escamosas de pulmão.....	53
Figura 18	Representação gráfica da comparação entre <i>cluster 2</i> (ancestralidade europeia) e <i>cluster 3</i> (ancestralidade asiática) em pacientes de adenocarcinoma gástrico do TCGA, através do <i>software Cytoscape</i> para as vias identificadas por <i>Gene Set Enrichment Analysis</i>	56
Figura 19	Representação gráfica da comparação entre <i>cluster 1</i> (ancestralidade asiática) e <i>cluster 3</i> (ancestralidade africana) em pacientes de adenocarcinoma de pulmão do TCGA, através do <i>software Cytoscape</i> para as vias identificadas por <i>Gene Set Enrichment Analysis</i>	57
Figura 20	<i>Kaplan-Meier</i> para pacientes de adenocarcinoma gástrico e de pulmão, carcinoma de células escamosas de pulmão e melanoma de acordo com estadiamento.....	60
Figura 21	<i>Hazard-ratio</i> (HR) para amostras de adenocarcinoma gástrico.....	63
Figura 22	<i>Hazard-ratio</i> (HR) para amostras de adenocarcinoma de pulmão.....	65

Figura 23	<i>Hazard-ratio</i> (HR) para amostras de carcinoma de células escamosas de pulmão.....	67
Figura 24	<i>Hazard-ratio</i> (HR) para amostras de melanoma.....	69

LISTA DE TABELAS

Tabela 1	Populações, superpopulações e bancos de dados utilizados como referência para inferência de ancestralidade	21
Tabela 2	Número de pacientes por tipo de tumor no TCGA e por <i>cluster</i>	24
Tabela 3	<i>Gene Set Enrichment Analysis</i> de pacientes de adenocarcinoma gástrico do TCGA.....	55
Tabela 4	<i>Gene Set Enrichment Analysis</i> de pacientes de adenocarcinoma de pulmão do TCGA, comparando <i>cluster 2</i> (ancestralidade asiática) e <i>cluster 3</i> (ancestralidade africana)	57
Tabela 5	<i>Gene Set Enrichment Analysis</i> de pacientes de carcinoma de células escamosas de pulmão do TCGA	58

LISTA DE ABREVIATURAS E SIGLAS

1000G	The 1000 Genomes Project
ACBD3	<i>Acyl-CoA binding domain containing 3</i>
AFR	Ancestralidade africana
AHRR	<i>Aryl-hydrocarbon receptor repressor</i>
AIM	Marcadores informativos de ancestralidade
ALK	<i>Anaplastic lymphoma kinase</i>
ALM	<i>Acral letiginous melanoma</i>
AMR	Ancestralidade ameríndia
BAM	<i>Binary Alignment Map</i>
BER	<i>Base excision repair</i>
BRAF	Proto-oncogene B-raf
COSMIC	<i>Catalogue of Somatic Mutations in Cancer</i>
COL22A1	<i>Collagen Type XXII alpha 1 chain</i>
CP	Câncer de pulmão
CPCNP	Carcinoma de pulmão de células não pequenas
CRYBB2	Cristalina Beta B2
DepMap	<i>Dependecy Map</i>
DNA	<i>Deoxyribonucleic acid</i>
EAS	Ancestralidade do leste asiático
EEPDI	<i>Endonuclease/exonuclease/phosphatase family domain containing 1</i>
EGFR	<i>Epidermal growth factor receptor</i>
ESTIMATE	<i>Estimation of stromal and immune cells in malignant tumor using expression data</i>
ERBB4	<i>Erb-b2 receptor tyrosine kinase 4</i>
EUR	Ancestralidade europeia
FAAH	<i>Fatty acid amide hydrolase</i>
FDR	<i>False discovery rate</i>
FGF	<i>Fibroblast growth factor</i>
FPKM	<i>Fragments per kilobase of transcript per million</i>
GATK	<i>Genome Analysis Toolkit</i>

GLUT1	Transportador de glicose 1
GNA11	<i>Guanine nucleotide-binding protein subunit alpha-11</i>
GO	<i>Gene Ontology</i>
GSEA	<i>Gene set enrichment analysis</i>
GRIN2A	<i>Glutamate ionotropic receptor NMDA type subunit 2A</i>
HapMap	International HapMap Project
HGDP	<i>Human Genome Diversity Project</i>
HR	<i>Hazard-ratio</i>
HRR	<i>Homologous recombination repair</i>
HIF	<i>Hypoxia-inducible factor</i>
ICB	<i>Immune checkpoint blocker</i>
IFN- γR	Receptor de interferon gama
IL	Interleucina
KRAS	<i>Kristen rat sarcoma viral oncogene homolog</i>
LD	<i>Linkage disequilibrium</i>
LEI	<i>Lancaster estimator of independence</i>
LMM	<i>Lentigo maligna melanoma</i>
LT- αR	Receptor de linfotoxina alfa
LUAD	<i>Lung adenocarcinoma</i>
LUSC	<i>Lung squamous cell carcinoma</i>
MAF	<i>Mutation annotation format</i>
MASTL	<i>Microtubule-associated serine/threonine kinase-like</i>
MAPK	<i>Mitogen-activated protein kinase</i>
MDSC	<i>Monocytic myeloid-derived suppressor cell</i>
MHC	<i>Major histocompatibility complex</i>
MLH1	<i>MutL homolog 1</i>
MM	<i>Mucosal melanoma</i>
MMR	<i>Mismatch repair</i>
MSH2	<i>MutS Homolog 2</i>
MSH6	<i>MutS Homolog 6</i>
MSI	Instabilidade de microssatélites
NER	<i>Nucleotide excision repair</i>
NHEJ	<i>Nonhomologous end joining</i>
NK	<i>Natural killer</i>

NM	<i>Nodular melanoma</i>
NRAS	<i>Neuroblastoma RAS viral oncogene homolog</i>
NSCLC	<i>Non-small cell lung carcinoma</i>
PD-1	<i>Programmed cell death protein 1</i>
PD-L1	<i>Programmed death-ligand 1</i>
PLEK2	<i>Pleckstrin 2</i>
POLE	<i>DNA Polymerase epsilon</i>
PSPHL	<i>Phosphoserine phosphatase pseudogene 1</i>
PWP2	<i>PWP2 small subunit processome component</i>
RNA	<i>Ribonucleic acid</i>
ROS1	<i>ROS proto-oncogene 1</i>
SAS	<i>Ancestralidade do sul da Ásia</i>
SNP	<i>Single nucleotide polymorphism</i>
SNV	<i>Single nucleotide variant</i>
SSM	<i>Superficial spreading melanoma</i>
STAT3	<i>Signal transducer and activator of transcription 3</i>
TCGA	<i>The Cancer Genome Atlas</i>
TCGAA	<i>The Cancer Genetic Ancestry Atlas</i>
TP53	<i>Tumor protein P53</i>
TP73	<i>Tumor protein P73</i>
T_{reg}	<i>Linfócitos T reguladores</i>
USH2A	<i>Usherin</i>
UV	<i>Ultravioleta</i>
VCF	<i>Variant Call Format</i>
VEGF	<i>Vascular endotelial growth factor</i>
VIM	<i>Vimentin</i>
VQSR	<i>Variant quality score recalibration</i>
XDH	<i>Xanthine dehydrogenase</i>

SUMÁRIO

1	INTRODUÇÃO.....	1
1.1	Genética	1
1.1.1	Genética de populações	2
1.2	Ancestralidade e etnia	4
1.2.1	Inferência de ancestralidade	5
1.3	Câncer	6
1.3.1	Instabilidade genômica e mutações	9
1.3.2	Sistema imune e o câncer	11
1.4	Câncer de pulmão	15
1.5	Melanoma	16
1.6	Observações	18
2	OBJETIVOS	19
2.1	Objetivo geral	19
2.2	Objetivos específicos	19
3	MATERIAIS E MÉTODOS	20
3.1	Desenho do estudo	20
3.2	Dados públicos de melanoma, CPCNP e adenocarcinoma gástrico	20
3.3	Dados públicos para populações de referência	21
3.4	Deteccção de variantes germinativas	22
3.5	<i>Pipeline</i> para ancestralidade individual e gerenciamento de <i>workflows</i>	22
3.6	Inferência de ancestralidade individual, validação e clusterização	23
3.7	Análise do infiltrado inflamatório	25
3.8	Análise do perfil mutacional somático	26
3.9	Análise da assinatura mutacional	26
3.10	Análise de expressão gênica diferencial	23
3.11	Análise de sobrevida	28
3.12	<i>Pipeline</i> para modelos de predição	28

4	RESULTADOS	30
4.1	Identificação de ancestralidade individual em amostras do TCGA e estruturação do <i>pipeline</i>	30
4.1.1	<i>Pipeline</i> para ancestralidade individual	30
4.1.2	Ancestralidade individual em dados de adenocarcinoma gástrico do TCGA e validação	34
4.1.3	Otimizando o número de variantes	36
4.2	Infiltrado inflamatório	39
4.3	Perfil mutacional somático	43
4.4	Assinatura mutacional	47
4.5	Expressão gênica diferencial	54
4.6	Sobrevida e modelos de regressão	59
5	DISCUSSÃO	70
6	CONCLUSÕES.....	82
7	REFERÊNCIAS.....	83

ANEXO

Anexo 1 Carta de aprovação do Comitê de Ética em Pesquisa-CEP

1 INTRODUÇÃO

1.1 GENÉTICA

Genética representa um ramo da biologia responsável pelo estudo da hereditariedade em organismos (World Health Organization-WHO 2002). A sua origem pode ser traçada até a publicação de Gregor Mendel sobre hibridização de plantas (Mendel 1866), onde, pela primeira, vez foram apresentadas evidências científicas de padrões de herança genética. Segundo Mendel, cada característica hereditária seria controlada por dois fatores, que segregariam de maneira independente e seriam transmitidos através de células reprodutivas (Mendel 1866). Apesar disso, a palavra "genética" foi utilizada pela primeira vez somente em 1905 por William Bateson, em carta ao colega Adam Sedgwick, ao descobrir os trabalhos de Gregor Mendel (Gayon 2016; Winchester 2020).

Ainda que não houvesse um conhecimento claro sobre genética, em registros históricos de civilizações antigas, era possível notar que estes já haviam constatado diversos padrões reprodutivos, como por exemplo: animais gerando descendentes da mesma espécie, filhos se parecendo com seus pais e plantas dando origem a plantas semelhantes (Wexler 2009). Também existem evidências de que humanos tinham conhecimento implícito de genética de pelo menos 15.000 anos a.C (Carey 2010; Winchester 2020). Estas evidências residem nos resquícios das primeiras tentativas de engenharia genética, chamada hoje de "seleção artificial", com a reprodução deliberada de plantas e animais para obtenção das características desejadas (Carey 2010; Winchester 2020).

As primeiras especulações sobre hereditariedade documentadas começaram na Grécia Antiga, principalmente com Hipócrates e Aristóteles (Carey 2010; Winchester, 2020). Hipócrates, também conhecido como o pai da medicina, acreditava na herança de características adquiridas. Postulava que todos os órgãos do corpo dos pais emitiam "sementes" invisíveis, que eram como componentes de construção miniaturizados e que eram transmitidos durante a relação sexual (Carey 2010; Winchester 2020). Mais tarde, Aristóteles refutou esta teoria, observando que animais e indivíduos que tinham sofrido mutilação ou perda de partes do corpo não passavam esta informação para a prole (Wexler 2009; Carey 2010).

No século 19, assim como os estudos de Gregor Mendel, também originaram-se importantes teorias da evolução, com Jean Baptiste Lamarck e, principalmente, Charles Darwin

(Avisé e Ayala 2009; Burkhardt 2013; Kovác 2019). Lamarck, em 1809, postulou em seu livro *Philosophie Zoologique* que o ambiente criaria necessidades às quais os organismos responderiam, usando suas características, que seriam acentuadas ou atenuadas de acordo com o seu uso e desuso (Burkhardt 2013; Kovác 2019). Já Charles Darwin publicou, em 1859, o livro *On the Origin of Species*, destacando a sua teoria e as suas observações realizadas nas ilhas Galápagos (Avisé e Ayala 2009; Burkhardt 2013). Segundo Darwin, a evolução ocorreria através de um processo nomeado "seleção natural", onde seres vivos que não estivessem bem adaptados ao ambiente tinham uma maior probabilidade de perecer, enquanto aqueles mais aptos teria uma maior probabilidade de sobreviver, transmitindo estas características para as gerações seguintes (Avisé e Ayala 2009; Burkhardt 2013).

1.1.1 Genética de populações

As teorias postuladas por Charles Darwin e Gregor Mendel representam a base da genética moderna e a síntese entre Mendel e Darwin culminou na genética de populações (Crow 1987; Brookfield 1996; Okasha 2016). Apesar de hoje ser facilmente observável a complementariedade entre as duas teorias, isso não era tão claro no início do século 20 (Okasha 2016). Muitos dos primeiros mendelianos não aceitaram a explicação "gradual" de Darwin para evolução, assim como os primeiros darwinistas acreditavam que a herança mendeliana era incompatível com o processo descrito por Darwin (Okasha 2016). Isto mudou quando modelos matemáticos propostos por Ronald Aylmer Fisher, John Burdon Sanderson Haldane e Sewall Wright passaram a surgir. Estes modelos buscavam avaliar o processo de seleção natural em populações que obedecessem às regras mendelianas (Crow 1987; Brookfield 1996; Okasha 2016).

O estudo de genética de populações é algo bastante quantitativo e a descrição de populações ocorre, em grande parte, através da frequência relativa de alelos e genótipos (Gillespie 2004a). Alelo se refere a uma variante de um dado gene. Já o genótipo pode ser definido, em organismos diploides, como o par de alelos para um gene específico, herdados a partir de seus progenitores (Gillespie 2004a). A relação entre estas frequências originou uma das principais leis do campo da genética de populações: a lei do equilíbrio de *Hardy-Weinberg*. Esta descoberta surgiu em 1908, de forma independente, com G.H. Hardy e W. Weinberg (Gillespie 2004a; Okasha 2016). Este princípio teórico descreve, matematicamente, a relação entre a frequência alélica e genotípica em populações infinitas, na qual não há pressão da seleção natural, migração, mutações, e que estão em panmixia, isto é, quando todos os cruzamentos ocorrem de forma aleatória (Gillespie 2004a; Okasha 2016). Este conceito conclui

que, nestas circunstâncias, as proporções alélicas e genotípicas permanecem constantes, de geração para geração. Considerando um indivíduo diploide e um gene com dois alelos distintos, sendo p e q a frequência de cada um destes alelos, p^2 a frequência para um genótipo homozigoto para p , q^2 a frequência para um genótipo homozigoto para q e $2pq$ a frequência do genótipo heterozigoto, a lei do equilíbrio de *Hardy-Weinberg* pode ser definida pela seguinte fórmula (Brookfield 1996; Gillespie 2004a; Okasha 2016):

$$p^2 + 2pq + q^2 = 1$$

A existência deste equilíbrio é importante, pois fornece um ponto de partida e de referência para estudos, avaliando os efeitos de processos como: seleção natural, mutação e consanguinidade (Hartl e Clark 2006; Mayo 2008).

A partir destes conceitos, alguns fundamentos teóricos e estatísticos passaram a fornecer informações importantes sobre como o processo evolutivo influencia a estrutura da variação genética dentro e entre populações. Entre estes fundamentos, podem ser citadas as estatísticas F de Wright (Wright 1949, 1965; Holsinger e Weir 2009) e *linkage disequilibrium* (LD) (Holsinger e Weir 2009).

As estatísticas F apresentadas por Wright envolvem três coeficientes F distintos: F_{IS} , F_{IT} e F_{ST} (Wright 1951, 1965; Holsinger e Weir 2009). F_{ST} representa a correlação entre alelos escolhidos aleatoriamente dentro da mesma subpopulação, relativo a toda a população (Wright 1949, 1965; Holsinger e Weir 2009). Já F_{IS} e F_{IT} tratam-se de medidas de desvio das proporções de *Hardy-Weinberg* dentro das subpopulações e da população total, respectivamente (Wright 1951, 1965; Holsinger e Weir 2009). Estas medidas, em especial F_{ST} , possuem um papel extremamente relevante para o estudo da história demográfica das populações, além de diversas aplicações, desde o mapeamento por associação de doenças até a ciência forense (Holsinger e Weir 2009).

LD pode ser definido como uma associação não aleatória entre alelos em loci diferentes dentro de uma população (Nordborg e Tavaré 2002; Gillespie 2004b). Este conceito foi inicialmente introduzido por Hilda Geiringer, em 1944 (Geiringer 1944; Nordborg e Tavaré 2002). Posteriormente, a primeira grande estatística para LD, conhecida como D , foi proposta por Lewontin e Kojima (Lewontin e Kojima 1960; Nordborg e Tavaré 2002). Hoje, esta estatística é raramente aplicada devido a sua escala ser dependente da frequência alélica (Nordborg e Tavaré 2002). Uma estatística mais comumente utilizada para LD foi introduzida por Hill e Robertson, conhecida como r^2 (Hill e Robertson 1968; Nordborg e Tavaré 2002). LD

é um componente essencial para o mapeamento de doenças no contexto de genética humana, reduzindo o custo para a avaliação do genoma humano (Nordborg e Tavaré 2002).

1.2 ANCESTRALIDADE E ETNIA

Disparidades já foram observadas na prevalência, mortalidade e risco de diversas doenças entre populações (Rotimi e Jorde 2010). Estas disparidades podem ser explicadas por fatores genéticos, ambientais e sociais, incluindo diferenças socioeconômicas e no acesso a tratamento (Özdemir e Dotto 2017; Gurdasani et al. 2019). Porém, é observado que, mesmo em condições iguais de acesso a tratamento, as discrepâncias entre populações ainda são predominantes (Andaya et al. 2013; Schreiber et al. 2014).

Em estudos avaliando a relação entre etnia e o câncer de estômago, por exemplo, observou-se que aproximadamente 50% de todos os casos do mundo ocorriam na Ásia Oriental, em especial na China (McLean e El-Omar 2014). Avaliando estes números nos Estados Unidos, também se percebeu uma maior incidência em indivíduos com descendência asiática, seguidos por afrodescendentes e brancos, respectivamente (Lui et al. 2014). Apesar destes aspectos epidemiológicos, asiáticos ainda apresentam uma melhor taxa de sobrevida em comparação a outros grupos (Merchant et al. 2014).

Diferenças biológicas em diversos aspectos também já foram observadas, incluindo, mas não limitando-se a: perfil mutacional, obesidade, inflamação crônica e sistema imune (inato e adaptativo) (Özdemir e Dotto 2017). Estudos anteriores destacam a prevalência (>3x) de tuberculose, esclerose sistêmica, septicemia e lúpus eritematoso sistêmico em afrodescendentes quando comparados a indivíduos com descendência europeia (Nédélec et al. 2016). Apesar disso, observou-se que afrodescendentes possuem uma melhor resposta inflamatória contra infecções e apresentam uma redução no crescimento bacteriano intracelular (Nédélec et al. 2016).

Em um estudo sobre câncer de cólon, investigou-se como a imunidade citotóxica atua no tumor, entre diferentes etnias (Basa et al. 2016). Foram detectadas diferenças na diversidade de células imunes no ambiente tumoral (infiltrado inflamatório), registrando uma menor infiltração de granzima B em indivíduos afrodescendentes, o que poderia indicar um pior prognóstico (Basa et al. 2016). Granzimas são uma família distinta de serinas-protease e estão presentes em grânulos de linfócitos citotóxicos, como linfócitos T citotóxicos e células *Natural Killer* (NK) (Cullen et al. 2010; Voskoboinik et al. 2015). A sua principal função consiste em

desencadear o processo de morte celular em células alvo (Cullen et al. 2010; Voskoboinik et al. 2015).

Nos Estados Unidos, ao analisar câncer de mama, nota-se uma maior taxa de mortalidade em mulheres afrodescendentes, quando comparadas às mulheres brancas (Quan et al. 2014; Daly e Olopade 2015). Além disso, a frequência de casos de câncer de mama triplo-negativo (não apresentam receptor do fator de crescimento epidérmico-2 e receptores para estrogênio e progesterona), considerado o tipo mais agressivo, é mais frequente em mulheres afrodescendentes (20- 50%) do que em outras etnias (5-15%) (Özdemir e Dotto 2017).

Análises mais aprofundadas sobre o sistema imune adaptativo em pacientes com câncer de mama indicaram que variantes gênicas em genes relacionados à resposta imune, como: receptor de interleucina-4 (IL-4R), interleucina-15 (IL-15R), linfotóxina alfa (LT- α R) e receptor de interferon gama (IFN- γ R), estão associadas a um aumento do risco de câncer em pacientes com ancestralidade africana (Quan et al. 2014). Em um outro estudo, avaliando expressão genética em câncer de mama, foram observados níveis de expressão consideravelmente elevados em mulheres afrodescendentes em dois genes específicos: homólogo de L-3-fosfatase fosfoserina (*L-3-phosphoserine phosphatase homolog - PSPHL*) e cristalina Beta B2 (*crystallin beta B2 - CRYBB2*). Além disso, estes genes puderam ser utilizados para clusterizar mulheres de descendência europeia e afrodescendentes de forma eficiente (Martin et al. 2009; Field et al. 2012). Este padrão de CRYBB2 também foi constatado em outros tipos de tumor, como em câncer de próstata e colorretal (Huo et al. 2017). CRYBB2 e PSPHL foram associados, respectivamente, à formação de catarata e pterígio, uma deposição patológica de matriz extracelular no tecido conectivo do olho, mas a função destes genes no contexto tumoral ainda não foi constatada (Özdemir e Dotto 2017).

1.2.1 Inferência de ancestralidade

Apesar das disparidades populacionais e étnicas observadas em diversos aspectos, a etnia autoatribuída se apresenta como um possível fator de confusão e não está, necessariamente, corretamente correlacionada à ancestralidade genética (Kasaniit et al. 2020). Com isto em mente, diversas técnicas e metodologias passaram a ser desenvolvidas com o intuito de inferir a ancestralidade, obtendo informações com maior acurácia.

Este tipo de análise tem sido empregada em diversos campos da pesquisa, incluindo genealogia, antropologia e epidemiologia, possibilitando a classificação de organismos em diferentes grupos populacionais (Royal et al. 2010; Padhukasahasram 2014). Este processo, em geral, é baseado na análise de grandes conjuntos de *single nucleotide polymorphisms* (SNPs)

ou marcadores informativos de ancestralidade (AIMs) presentes em populações ou bancos de referência (Royal et al. 2010; Padhukasahasram 2014), como, por exemplo *The 1000 Genomes Project* (1000G) (The 1000 Genomes Project Consortium 2015), *Human Genome Diversity Project* (HGDP) (Rosenberg et al. 2002) e *International HapMap Project* (HapMap) (The International HapMap Consortium 2003).

Atualmente, a inferência de ancestralidade pode ser dividida em três tipos: ancestralidade populacional, individual e local (ou cromossômica) (Padhukasahasram 2014; Bansal e Libiger 2015). Para os dois primeiros tipos, são aplicados programas como: *STRUCTURE* (Pritchard et al. 2000), *Admixture* (Alexander et al. 2009) e *FRAPPE* (Tang et al. 2005). Em câncer, esta abordagem já foi aplicada, por exemplo, no consórcio TCGA (*The Cancer Genome Atlas*) pelo *The Cancer Genetic Ancestry Atlas* (TCGAA) (Yuan et al. 2018).

No caso da ancestralidade local, busca-se avaliar a ancestralidade em cada fragmento de cada cromossomo, como um mosaico (Padhukasahasram 2014) e, por isso, outros métodos são aplicados. Entre eles podem ser citados: LAMP (Sankararaman et al. 2008), HAPMIX (Price et al. 2009) e RFMix (Maples et al. 2013).

1.3 CÂNCER

Câncer pode ser definido como um conjunto de doenças que podem surgir em quase qualquer órgão ou tecido do corpo e tem início quando células anormais passam a crescer de forma descontrolada (WHO 2020a). A sua origem está associada a uma grande variedade de fatores endógenos e exógenos, entre eles o componente hereditário e genético, dieta, falta de exercício físico, consumo de tabaco, exposição a alguns tipos de vírus, entre outros fatores (Vineis e Wild 2014).

Em 2018 foram estimados cerca de 18,1 milhões de novos casos e 9,6 milhões de óbitos no mundo (WHO 2020b). Segundo dados de 2015, esta seria a primeira ou segunda doença com o maior número de mortes no mundo em 91 dos 172 países, em indivíduos com menos de 70 anos (Bray et al. 2018). No Brasil, a estimativa para cada ano do triênio 2020- 2022 aponta para 625 mil novos casos pelo Instituto Nacional do Câncer-INCA (Ministério da Saúde 2020). Em mulheres, destacam-se os tumores de mama, cólon e reto, colo do útero, pulmão e glândula tireoide (Ministério da Saúde 2020). Já em homens, destacam-se o câncer de próstata, cólon e reto, pulmão, estômago e cavidade oral (Ministério da Saúde 2020).

Biologicamente, esta doença pode ser definida por dez características predominantes: evasão do sistema imune, potencial replicativo ilimitado, promoção de processos inflamatórios,

ativação de metástase e invasão tecidual, indução de angiogênese, instabilidade genômica e mutacional, resistência a mecanismos de morte celular programada ou apoptose, reprogramação do metabolismo energético, sinalização de proliferação celular e escape de processos inibitórios de proliferação (Hanahan e Weinberg 2000; Hanahan e Weinberg 2011).

Em tecidos normais, o controle da produção e liberação de fatores de crescimento é extremamente importante para a progressão do ciclo celular e para a manutenção da homeostase no número de células, e, portanto, mantendo a arquitetura e função normal do tecido (Hanahan e Weinberg 2011; Thermo Fisher Scientific 2017). Já em células tumorais, essa regulação não ocorre. Seja pela exposição a fatores exógenos ou por mutações adquiridas ao longo do tempo, estas células adquirem a capacidade de sustentar sinais de proliferação por diversos métodos (Hanahan e Weinberg 2011; Thermo Fisher Scientific 2017). Isto pode acontecer através da estimulação autócrina para proliferação (Hanahan e Weinberg 2011; Xie et al. 2020b), assim como através da estimulação de células do estroma associado ao tumor para a produção de fatores de crescimento (Bussard et al. 2016).

Além da desregulação da sinalização para o início do ciclo celular, em relação a tecidos normais, a capacidade proliferativa em células tumorais é também ilimitada (Hanahan e Weinberg 2011; Jesus e Blasco 2013; Akincilar et al. 2016). Normalmente, com a idade, gradualmente ocorre a deterioração dos telômeros, levando a célula à apoptose ou senescência (Shammas 2011). Esta região é especialmente importante para proteção do genoma da degradação nucleolítica, recombinação e fusão intercromossômica (Shammas 2011). Em células tumorais, essa regulação não ocorre, visto que, diferentemente de células normais, há uma grande expressão da enzima telomerase, a DNA polimerase especializada em adicionar segmentos repetidos no final do telômero, mantendo-o intacto (Hanahan e Weinberg 2011; Shammas 2011).

Outra importante característica em células tumorais é a reprogramação do metabolismo energético. Essas alterações possuem um impacto direto no desbalanço proliferativo e em outras disfunções já conhecidas (DeBerardinis et al. 2008; Hanahan e Weinberg 2011; DeBerardinis e Chandel 2016). Otto Warburg foi o primeiro a observar uma característica anômala do metabolismo energético: mesmo na presença de oxigênio, estas células seriam capazes obter energia, principalmente por meio da glicólise (Warburg 1956a e b; Hanahan e Weinberg 2011). Apesar do uso da glicólise ser energeticamente muito menos eficiente do que a fosforilação oxidativa, isto é compensado pela sua velocidade e pelo grande consumo de glicose (DeBerardinis et al. 2008). Em grande parte, esta mudança é possível, visto que estas células expressam, em maior quantidade, o transportador de glicose GLUT1, que aumenta a captação

de glicose pela célula (DeBerardinis et al. 2008; Hanahan e Weinberg 2011; DeBerardinis e Chandel 2016; Kim e DeBerardinis 2019; Faubert et al. 2020). Apesar da grande importância da observação de Otto Warburg e do seu impacto na área clínica, hoje está cada vez mais claro que há uma grande heterogeneidade dentro do tumor, com células apresentando disfunções metabólicas e energéticas distintas e que evoluem constantemente de acordo com a progressão da doença (Kim e DeBerardinis 2019; Faubert et al. 2020).

Assim como células normais, para a sobrevivência do tumor são necessários nutrientes e oxigênio, bem como a eliminação de resíduos metabólicos e dióxido de carbono. Por esta razão, uma importante característica em tumores é a formação de uma neovasculatura, através do processo chamado de angiogênese (Hanahan e Weinberg 2011; Alberts et al. 2014). Este mecanismo é estimulado com o crescimento do tumor, aumento da hipóxia e com a privação de nutrientes (Weis e Cheresh 2011; Alberts et al. 2014), e acontece principalmente através da sinalização de algumas proteínas, como o fator de crescimento endotelial vascular (*vascular endothelial growth factor*, VEGF) e o fator de crescimento de fibroblastos (*fibroblast growth factor*, FGF) (Hanahan e Weinberg 2011; Weis e Cheresh 2011).

Os novos vasos, formados através do mecanismo de angiogênese, tendem a apresentar uma maior fragilidade, facilitando a invasão de células tumorais à corrente sanguínea e permitindo o seu deslocamento para outros tecidos (Hanahan e Weinberg 2011; Alberts et al. 2014; Coban et al. 2021). Este processo é chamado de metástase e pode ser resumido em três etapas: escape do tecido de origem, deslocamento através da circulação e colonização de um novo sítio (Alberts et al. 2014). Durante a primeira etapa, é observada principalmente a transição de células com alta adesão entre si para células com características mais migratórias. Este mecanismo é chamado de transição epitélio-mesenquimal (Wan et al. 2013; Alberts et al. 2014; Coban et al. 2021). Isto ocorre, em grande parte, devido à baixa expressão de E-caderina, que é responsável por manter a adesão célula-célula, permitindo assim que células se destaquem do tumor primário (Wan et al. 2013; Alberts et al. 2014; Coban et al. 2021). Já a segunda etapa é caracterizada pela sobrevivência das células tumorais na circulação e a sua saída para um novo tecido (Alberts et al. 2014; Freitas et al. 2018) e por último, a célula tumoral invade o novo sítio e busca sobreviver neste novo meio e colonizá-lo (Alberts et al. 2014; Freitas et al. 2018). Devido ao grande número de etapas necessárias para a conclusão da metástase, poucas são as células capazes de sobreviver. Quando isso acontece, no entanto, há um enorme impacto no prognóstico do paciente (Hanahan e Weinberg 2011; Alberts et al. 2014).

1.3.1 Instabilidade genômica e mutações

O surgimento das diversas *hallmarks*, apontadas anteriormente, está, em grande parte, atrelada a uma sucessão de alterações nos genomas das células neoplásicas (Hanahan e Weinberg 2011). Algumas dessas alterações são extremamente relevantes, visto que conferem vantagens seletivas a subclones de células, permitindo o seu crescimento e eventual dominância do ambiente do tecido local (Hanahan e Weinberg 2011).

A instabilidade genômica pode ocorrer em vários níveis, resultando em diferentes lesões genéticas, desde um aumento na frequência de mutações pontuais no genoma, inserções e deleções, até rearranjos cromossômicos em grande escala e alterações na ploidia (Burrell et al. 2013). Essa diversidade não só impacta diretamente no prognóstico, como também tem um importante papel na formação de um tumor heterogêneo. (Burrell et al. 2013; Andor et al. 2017).

Em células normais, o genoma é replicado e dividido com uma alta fidelidade, apresentando uma taxa de mutação pontual em células somáticas relativamente baixa (0.77×10^{-9} a cada local por divisão celular) (Lynch 2010; Burrell et al. 2013). Este equilíbrio é mantido através de diversas vias de reparo de DNA, como por exemplo: reparo por excisão de bases (*base excision repair*, BER) e nucleotídeos (*nucleotide excision repair*, NER), reparo de mal pareamento de DNA (*mismatch repair*, MMR), reparo por recombinação homóloga (*homologous recombination repair*, HRR) e ligação de extremidades não-homólogas (*nonhomologous end joining*, NHEJ) (Negrini et al. 2010; Burrell et al. 2013; Chatterjee e Walker 2017).

BER e NER são as duas principais vias de reparo por excisão. Este mecanismo é conservado em células eucarióticas e um protótipo deste mecanismo existe em células procarióticas (Izumi e Mellon 2016). Ambos os mecanismos podem ser resumidos em quatro etapas simples: 1) reconhecimento do dano; 2) excisão do dano; 3) síntese de DNA, a fim de preencher a lacuna de nucleotídeo(s); 4) selamento dos cortes no DNA (Izumi e Mellon 2016). BER é capaz de reparar danos pequenos em bases, quebras de fita simples do DNA, danos oxidativos e por deaminação (Izumi e Mellon 2016). Já NER pode reparar trechos relativamente grandes de DNA, incluindo danos a subprodutos de radiação ultravioleta e da exposição a carcinógenos (Izumi e Mellon 2016).

Erros de pareamento de DNA podem ser resultado de erros de incorporação pela DNA polimerase, recombinação entre DNAs parentais heteroalélicos e danos químicos ou físicos a nucleotídeos (Fishel 2015). Estes tipos de erros são corrigidos através da via de reparo de mal pareamento. Este mecanismo é evolutivamente conservado e mutações em genes associados a esta via contribuem para o surgimento de doenças como a síndrome de Lynch (Li e Martin

2016). Entre estes genes, podem ser citados: MSH2, MLH1, MSH6 e PMS2 (Li e Martin 2016). Mutações de genes relacionados a MMR resultam em um fenótipo conhecido como instabilidade de microssatélites (MSI), que é caracterizado por alterações no comprimento das repetições de nucleotídeos em tandem localizadas em todo o genoma (Li e Martin 2016).

Também existem vias importantes para o reparo de quebras de fita dupla de DNA. Este tipo de erro é extremamente prejudicial, podendo resultar em inserções, deleções ou em translocações cromossômicas que são passos importantes para o surgimento de células tumorais (Pannunzio et al. 2018; Wright et al. 2018). Este tipo de erro pode ser corrigido através de duas vias principais: recombinação homóloga e ligação de extremidades não-homólogas. Para que haja a recombinação homóloga, é necessária uma sequência homóloga como referência para o reparo, sendo um processo muito preciso (Ranjha et al. 2018). Já a via de ligação de extremidades não-homólogas, em grande parte, é um processo independente de homologia e constitui um processo mais rápido em comparação à recombinação homóloga (Pannunzio et al. 2018; Ranjha et al. 2018).

Estas alterações no genoma podem surgir tanto no tecido somático quanto no germinativo, e são chamadas de mutações somáticas e mutações germinativas, respectivamente (Griffiths et al. 2000). Mutações somáticas ocorrem em uma única célula, e esta é progenitora de células mutagênicas idênticas. Por outro lado, mutações germinativas ocorrem na linhagem germinativa (células reprodutivas, como espermatozoide e óvulo) e estas mutações são passadas para as gerações seguintes (Griffiths et al. 2000). Esta distinção é importante, pois as diferentes mutações podem dar origem a tipos distintos de câncer: hereditário e esporádico (Chatterjee e Walker 2017). Enquanto o primeiro está associado a mutações germinativas, predispondo o indivíduo a desenvolver um tumor, o segundo está associado, principalmente, a mutações somáticas. No caso de câncer hereditário, muitas vezes, está associado a síndromes como: síndrome de Lynch e Li Fraumeni, que surgem devido a mutações germinativas em genes importantes para vias de reparo e em TP53 (*tumor protein P53*) (Correa 2016; Boland et al. 2018).

Mutações somáticas podem ocorrer de diversas formas e com diferentes origens. Estas alterações podem decorrer de erros na replicação do DNA, ou pela exposição a fatores endógenos e exógenos (Martincorena e Campbell 2015; Chatterjee e Walker 2017). A maior parte dos danos causados por fatores endógenos surge pela reação do DNA com reações hidrolíticas e oxidativas com a água e espécies reativas de oxigênio (ROS), que estão naturalmente presentes nas células (Chatterjee e Walker 2017). Além disso, vírus e retrotransposons endógenos também podem causar a inserção de sequências de DNA

(Martincorena e Campbell 2015). Por outro lado, danos ao DNA por fatores exógenos ocorrem através da exposição do DNA a agentes químicos, físicos ou ambientais (Chatterjee e Walker 2017). Entre estes agentes, podem ser citados: radiação ultravioleta, radiações ionizantes e agentes alcalinos (Martincorena e Campbell 2015; Chatterjee e Walker 2017).

Processos mutacionais associados a mutações somáticas, como os citados anteriormente, possuem características próprias e padrões mutacionais distintos, gerando o que é chamado de assinatura mutacional (Alexandrov et al. 2020). Danos a partir de, por exemplo, radiação ultravioleta ou por deficiências em vias de reparo, possuem padrões específicos, ou assinaturas, que podem ser detectados e medidos de forma a avaliar o impacto destes fatores sob um tipo específico de tumor (Alexandrov et al. 2020). Em câncer de mama, por exemplo, assinaturas mutacionais já foram observadas como preditores para deficiências em BRCA1 e BRCA2 (Davies et al. 2017).

1.3.2 Sistema imune e o câncer

O sistema imune pode ser definido como um mecanismo de defesa contra, principalmente, agentes infecciosos, podendo atuar também contra substâncias estranhas não infecciosas e produtos de células danificadas (Parkin e Cohen 2001; Abbas et al. 2017a). Este sistema pode ser dividido em dois grupos diferentes, que atuam de maneira e com especificidades diferentes. O primeiro deles consiste no sistema imune inato, que é mediado por mecanismos que estão presentes mesmo antes da infecção, oferecendo uma resposta quase imediata contra agentes infecciosos (Parkin e Cohen 2001; Abbas et al. 2017a). Os receptores presentes nestas células são específicos para estruturas comuns a grupos de agentes relacionados e não distinguem diferenças sutis entre eles (Parkin e Cohen 2001; Abbas et al. 2017a). Os principais componentes deste sistema podem ser descritos como: (1) barreiras físicas e químicas, como epitélio e compostos produzidos pela superfície epitelial; (2) células fagocíticas (neutrófilos e macrófagos), células dendríticas, mastócitos, células NK e outras células inatas linfóides; e (3) proteínas sanguíneas e outros mediadores inflamatórios (Parkin e Cohen 2001; Abbas et al. 2017a).

Já o segundo grupo se refere ao sistema imune adaptativo, que é mediado por linfócitos e os seus produtos (Abbas et al. 2017a; Paludan et al. 2020). Ao contrário do sistema imune inato, a imunidade adaptativa é capaz de reagir de maneira mais intensa, específica e diversa, porém a sua resposta leva mais tempo para ocorrer (Abbas et al. 2017a; Paludan et al. 2020). A imunidade adaptativa é capaz de reconhecer e reagir a um grande número de substâncias microbianas e não microbianas, chamadas de antígenos (Abbas et al. 2017a). Estes antígenos

possuem regiões complexas, ou epítomos, que são reconhecidas por linfócitos, desencadeando a resposta imune (Abbas et al. 2017a). Isso ocorre devido a presença de receptores em linfócitos, capazes de distinguir diferenças sutis entre epítomos distintos (Abbas et al. 2017a).

Os linfócitos podem ser divididos em duas populações principais: B e T (Yatim e Lakkis 2015; Abbas et al. 2017a). Linfócitos B são responsáveis pela imunidade humoral, atuando através de moléculas que são produzidas por estas células, chamadas de anticorpos, e que estão presentes no sangue e em secreções da mucosa. Estas moléculas são capazes de reconhecer antígenos e neutralizar agentes infecciosos, marcando-os para serem eliminados por fagocitose e pelo sistema complemento (Yatim e Lakkis 2015; Abbas et al. 2017a). Este tipo de resposta é o principal mecanismo contra micróbios e suas toxinas localizadas fora das células (Yatim e Lakkis 2015; Abbas et al. 2017a). Já os linfócitos T são responsáveis por mediar a imunidade celular. Frequentemente, agentes infecciosos são capazes de infectar células do hospedeiro e sobreviver dentro de fagócitos, impedindo, portanto, o contato com anticorpos circulantes. Nestes casos, a imunidade celular se torna necessária. Os receptores de linfócitos T reconhecem peptídeos que são expostos pelas células hospedeiras, através do complexo principal de histocompatibilidade (*major histocompatibility complex*, MHC) (Yatim e Lakkis 2015; Abbas et al. 2017a).

Apesar da divisão entre sistema imune inato e adaptativo, a comunicação entre eles é essencial. A resposta inicial do sistema imune inato fornece sinais precoces de perigo que estimulam a resposta imune adaptativa, através da apresentação de antígenos, principalmente, pelas células dendríticas (Yatim e Lakkis 2015; Abbas et al. 2017a). Além disso, o sistema imune adaptativo é capaz de estimular e amplificar a resposta a agentes infecciosos pelo sistema imune inato (Yatim e Lakkis 2015; Abbas et al. 2017a).

Atualmente, a biologia do câncer está constantemente mudando de uma visão centrada especificamente nas células tumorais, para um conceito mais abrangente, incluindo o ambiente do tumor e as células a ele associadas, compreendendo, portanto, o sistema imune (Greten e Grivennikov 2019). Essa interação pode ser modulada por diversos fatores, como: tecido de origem, localização do tumor, o espectro de mutações genéticas, mutações epigenéticas, composição do microbioma do hospedeiro e fatores exógenos (Galon e Bruni 2020).

Historicamente, os campos da oncologia e da imunologia têm sido estudados de forma relativamente independente, até as últimas duas décadas (Galon e Bruni 2020). Ainda sim, hipóteses de grande importância foram elaboradas. Paul Ehrlich, em 1909, foi o primeiro a sugerir que tumores em desenvolvimento poderiam ser identificados e eliminados através do sistema imune, antes que ocorresse qualquer manifestação clínica (Ehrlich 1909; Muenst et al.

2016; Galon e Bruni 2020). Este conceito foi posteriormente aprimorado por Frank MacFarlane Burnet em 1957 e nomeado como *immunosurveillance* (Burnet 1957; Muenst et al. 2016; Galon e Bruni 2020).

Com a evolução e o aumento dos estudos nas áreas de oncologia e imunologia, em 2001, o conceito de *immunosurveillance* foi adaptado, ampliado e nomeado como *immunoediting* (Shankaran et al. 2001; Dunn et al. 2002). Este modelo propõe uma interação mais dinâmica e constante entre tumor e o sistema imune (Galon e Bruni 2020). Essa interação poderia resultar na eliminação do tumor, em um equilíbrio ou a evasão do tumor contra o sistema imune (Muenst et al. 2016; Galon e Bruni 2020). Na fase de equilíbrio, haveria um contínuo processo de proliferação e mutação das células tumorais, buscando a sua evasão do sistema imune (Galon e Bruni 2020). Apesar dessa separação, estas etapas não ocorrem separadamente, mas sim de forma contínua, deslocando de uma para a outra, dependendo do próprio sistema imune e das características das células tumorais (Muenst et al. 2016).

Diversos mecanismos imunológicos já foram observados e estabelecidos. Um dos principais efetores para a eliminação do tumor são os linfócitos T CD8⁺ citotóxicos. Devido a sua ação citotóxica estas células conferem uma proteção duradoura e protege recorrentemente contra a ação de células tumorais (Elinav et al. 2013; Greten e Grivennikov 2019; Galon e Bruni 2020). Células NK, linfócitos T gama/delta e macrófagos também são importantes mediadores para a lise de células tumorais (Galon e Bruni 2020). Em especial, as duas primeiras são produtoras de IFN- γ , uma importante citocina que aumenta a expressão de MHC e imunoproteassomas em células tumorais, sensibilizando-as para a ação de linfócitos T CD8⁺ citotóxicos (Greten e Grivennikov 2019; Galon e Bruni 2020). Já os macrófagos possuem um papel mais complexo, podendo atuar tanto a favor do tumor, auxiliando através de diversos mecanismos o seu desenvolvimento, ou contra (Galon e Bruni 2020).

Macrófagos podem ser divididos em três fenótipos distintos, M0, um estado de repouso e M1 e M2, fenótipos polarizados (Zhang et al. 2015). M1 é um fenótipo pró-inflamatório e possui uma atividade antitumoral, enquanto M2 possui uma função imunossupressora, que contribui para a progressão do tumor (Najafi et al. 2019). A obtenção destes fenótipos está associada à liberação de fatores por células tumorais e por células do microambiente tumoral (Najafi et al. 2019). O fenótipo M1 é induzido na presença de fatores como: TNF- α , IFN- γ e interleucinas (ILs 8, 12 e 21) (Najafi et al. 2019). A atuação destas células favorece a estimulação da apresentação de antígenos via MHC de classe I e faz com que células tumorais tornem-se mais sensíveis a lise (Najafi et al. 2019). Além disso, possuem uma interação importante com linfócitos T CD8⁺ citotóxicos, auxiliando no seu recrutamento (Najafi et al.

2019). Já o fenótipo M2 ocorre na presença de HIF-1 α e HIF-2 α , fatores produzidos por células em hipóxia (Najafi et al. 2019). Além disso, há um importante papel das interleucinas 4, 6, 10 e 33 (Zhang et al. 2015; Najafi et al. 2019). A interação destas células com as células tumorais possui grande importância, pois potencializam a sua migração e sobrevivência (Najafi et al. 2019; Porrello et al. 2018). Além disso, possuem uma ação inibitória a linfócitos T CD8⁺ citotóxicos, através de STAT3, de forma a impedir a sua resposta (Najafi et al. 2019).

Além dos macrófagos M2, outras células também são extremamente relevantes no contexto tumoral, ajudando no seu desenvolvimento. Neste caso destacam-se, principalmente, linfócitos T reguladores (T_{Reg}) e células supressoras derivadas da linhagem mieloide (*monocytic myeloid-derived suppressor cells*, MDSCs) (Galon e Bruni 2020). T_{Regs} suprimem as respostas imunológicas através de diversos mecanismos e previnem reações autoimunes (Abbas et al. 2017b). No contexto tumoral, a sua atuação é extremamente relevante, visto que a sua associação a células tumorais resulta em um pior prognóstico em múltiplos tumores sólidos, como pâncreas, ovário, gástrico e mama (Paijens et al. 2020). Estas células são conhecidas por produzirem citocinas anti-inflamatórias como: IL-10, IL-35 e TGF- β , além de suprimirem a resposta imune através da privação de nutrientes e citólise (Paijens et al. 2020). Já MDSCs representam uma população heterogênea de células mielóides imaturas, consistindo em precursores de granulócitos, macrófagos e células dendríticas, que acumulam durante inflamações crônicas e durante a progressão tumoral (Umansky et al. 2016). A sua presença no tumor resulta em um pior prognóstico e estas células produzem fatores como VEGF e FGF, promovendo neoangiogênese tumoral, auxiliando em seu desenvolvimento (Umansky et al. 2016).

Outro importante mecanismo para a sobrevivência de células tumorais está no processo de exaustão de linfócitos T CD8⁺. Em respostas imunes prolongadas, com a presença contínua de fatores como IFN- γ , estas células passam a expressar receptores inibitórios como PD-1 (Abbas et al. 2017b; Paijens et al. 2020). Devido a expressão do ligante de PD-1 pelas células tumorais, estas células induzem linfócitos T CD8⁺ à exaustão, levando a uma resposta imune deficiente (Paijens et al. 2020). A importância desta interação levou ao desenvolvimento de terapias importantes tendo como alvo PD-1, como é o caso de imunoterapias.

1.4 CÂNCER DE PULMÃO

Câncer de pulmão (CP) é uma doença agressiva, de grande impacto na saúde pública. É o tipo de câncer mais comumente diagnosticado, representando 11,6% de todos os casos, com a maior incidência (2,1 milhões de novos casos) e o maior número de óbitos, totalizando 1,8 milhão (18,4% de todos os casos) (Bray et al. 2018). O fator de risco mais relevante para o surgimento desta neoplasia é o tabagismo. Mundialmente, 80% dos homens e 50% das mulheres diagnosticadas com CP são ou foram fumantes. Outros fatores, como tabagismo passivo, poluição e predisposição genética também parecem contribuir para o surgimento da doença (Gridelli et al. 2015).

No Brasil, estimam-se 17.760 novos casos para homens e 12.440 para mulheres para cada ano do triênio 2020-2022 (Ministério da Saúde 2020). Esta neoplasia, entre os homens, é o segundo tipo de câncer mais frequente no País, sendo o segundo na região Sul (31,07/100 mil) e Nordeste (11,01/100 mil), enquanto nas regiões Sudeste (18,10/100 mil), Centro-Oeste (15,11/100 mil) e Norte (9,24/100 mil) corresponde ao terceiro. Entre as mulheres, corresponde ao quarto tipo de câncer mais frequente, sendo o terceiro nas regiões Sul (18,66/100 mil) e Sudeste (12,09/100 mil). Nas demais regiões, Nordeste (8,86/100 mil), Centro-Oeste (10,87/100 mil) e Norte (6,47/100 mil) ocupa a quarta posição (Ministério da Saúde 2020).

Histologicamente, os casos de CP são, em sua maior parte, carcinomas de pulmão de células não pequenas (CPCNP), representando cerca de 85% dos casos (Herbst et al. 2018). CPCNP pode ser dividido, principalmente, em dois subtipos: adenocarcinoma de pulmão (*lung adenocarcinoma* - LUAD) e carcinoma de células escamosas do pulmão (*lung squamous cell carcinoma* - LUSC). Apesar de ambos os subtipos poderem manifestar-se a partir do tabagismo, LUAD é o principal entre não-fumantes, indicando que outros fatores são relevantes para o seu surgimento (Gridelli et al. 2015).

Sendo esta uma doença molecularmente heterogênea, entender a sua biologia é crucial para o desenvolvimento de terapias efetivas. Pacientes classificados como fumantes possuem uma alta quantidade de mutações somáticas, cerca de dez vezes maior do que em não-fumantes (Gridelli et al. 2015). Neste sentido, podemos destacar o receptor do fator de crescimento epidérmico (*epidermal growth factor receptor* - EGFR) e a quinase do linfoma anaplásico (*anaplastic lymphoma kinase* - ALK), como dois dos principais alvos de anomalias mutacionais em NSCLC (Rotow e Bivona et al. 2017). Outras mutações, apesar de menos frequentes, também se destacam, como: oncogene homólogo do sarcoma viral de camundongo de Kirsten

(*kirsten rat sarcoma viral oncogene homolog* - KRAS), ROS proto-oncogene 1 (ROS1), TP53 (Rotow e Bivona et al. 2017).

Em termos de tratamento, nos últimos anos, o uso de terapias citotóxicas, baseadas na preferência do médico, estão sendo substituídas por terapias mais personalizadas, incluindo um subconjunto de pacientes sendo tratados de acordo com alterações genéticas em seus tumores e o status do ligante de morte programada-1 (*programmed death-ligand 1* - PD-L1) (Herbst et al. 2018). Este modelo de análise beneficia terapias baseadas em alvos específicos e em bloqueadores de checkpoint imune (*immune checkpoint blockers* - ICB) (Herbst et al. 2018). Por estas razões, o estudo e otimização da imunoterapia tem se tornado extremamente relevante no panorama de tratamento do CPCNP.

Em estudos avaliando CPCNP, observou-se que a incidência é maior em afrodescendentes, do gênero masculino, nos Estados Unidos, com uma taxa de 87,9 novos casos a cada 100.000 pessoas. Já em indivíduos brancos, asiáticos e ameríndios, as taxas são de 75,9, 45,2 e 71,9 a cada 100.000 indivíduos, respectivamente (Siegel et al. 2018). Além disso, diferenças genéticas também já foram observadas. Indivíduos com ancestralidade europeia possuem, em média, uma menor quantidade de casos de CPCNP com mutações em EGFR e ALK (10%), enquanto em indivíduos com descendência asiática, esta taxa chega a 60% (Gridelli et al. 2015).

1.5 MELANOMA

Melanoma refere-se a um tipo de câncer, cujo surgimento se dá a partir da malignidade em melanócitos, células de origem neuroectodérmica e produtoras de pigmentos. Estas células podem ser encontradas na pele, íris e no reto. O principal fator de risco para a doença é a exposição excessiva à radiação solar ultravioleta (UV), porém outros fatores como cor da pele, dos olhos e histórico familiar também podem aumentar o risco de desenvolvimento da doença.

Em 2020, foi estimado, globalmente, um total de 324.635 novos casos de melanoma da pele (International Agency for Research on Cancer-IARC 2020). Porém, uma característica bastante relevante para o melanoma é a grande disparidade na taxa de incidência entre países (Bray et al. 2018; Schadendorf et al. 2018; Ferlay et al. 2019). Os países com os maiores números de casos, em geral, possuem uma população majoritariamente branca (ancestralidade europeia). Em conjunto, Austrália e Nova Zelândia possuem uma taxa de incidência de 73,0 para homens e 52,2 novos casos a cada 100.000 indivíduos para mulheres (IARC 2020). Já na

região sul-central da Ásia, esta taxa é de 0.32 para homens e de 0.3 novos casos a cada 100.000 indivíduos para mulheres (IARC 2020).

No Brasil, as maiores taxas de incidência em homens e mulheres encontra-se na região Sul, com 6,49 e 6,61 a cada 100.000 indivíduos, respectivamente (Ministério da Saúde 2020). Entre os homens, a região Centro-Oeste possui uma taxa de 2,09/100 mil, a região Sudeste 5,80/100 mil, a região Nordeste 1,64/100 mil, a região Norte 0,83/100 mil. Entre as mulheres, a região Sudeste possui uma taxa de 5,10/100 mil, a região Centro-Oeste 3,38/100 mil, a região Nordeste 1,94/100 mil, a região Norte 0,78/100 mil (Ministério da Saúde 2020).

Histologicamente, o melanoma pode ser dividido em cinco principais subtipos: disseminativo superficial (*superficial spreading melanoma* - SSM), nodular (*nodular melanoma* - NM), lentigo maligno (*lentigo maligna melanoma* - LMM), lentiginoso acral (*acral lentiginous melanoma* - ALM) e mucoso (*mucosal melanoma* - MM) (Schadendorf et al. 2015). Enquanto o melanoma cutâneo ocorre predominantemente em indivíduos com pele clara, populações com pele mais escura, na Ásia e na África, possuem uma maior tendência de desenvolver MM e ALM (Schadendorf et al. 2015).

O melanoma é uma doença altamente agressiva e possui um prognóstico ruim em seu estágio metastático. A sua progressão é tipicamente descrita como linear e gradual, em que a metástase ocorre como um evento tardio. Porém, o processo de metástase também pode ocorrer em estágios iniciais, mesmo durante a formação do tumor primário, levando à formação de um modelo de progressão paralela (Schadendorf et al. 2015).

Em termos genéticos, esta doença possui mutações em 4 oncogenes clássicos, observadas frequentemente em pacientes, sendo estes: proto-oncogene B-Raf (BRAF), proto-oncogene NRAS (NRAS), gene da proteína ligadora de nucleotídeos de guanina (*guanine nucleotide-binding protein subunit alpha-11* - GNA11) e proto-oncogene KIT (KIT) (Kunz 2014). Porém, mutações em BRAF, NRAS e GNA11 também estão presentes, com frequência, em neoplasias melanocíticas benignas, indicando que estas mutações representam os primeiros estágios de progressão do melanócito para a senescência induzida por oncogene. Portanto, para o surgimento do melanoma, seriam necessárias outras alterações para superar a senescência (SCHADENDORF et al. 2015). Mutações em genes, como gene da proteína quinase ativada por mitógeno 1 e 2 (*mitogen-activated protein kinase 1/2* - MAPK1/2), gene do receptor Tirosina Quinases Erb-B2 4 (*Erb-B2 Receptor Tyrosine Kinase 4* - ERBB4) e gene da subunidade 2 do receptor de glutamato ionotrópico tipo NMDA (*glutamate Ionotropic Receptor NMDA Type Subunit 2A* - GRIN2A) fazem parte de um número de novos candidatos a oncogenes, que estão frequentemente associados ao melanoma (Kunz 2014).

Inicialmente, o tratamento de mieloma era restrito à terapia convencional, com drogas citotóxicas, porém, devido à resistência do melanoma em estágios metastáticos e com o aumento do conhecimento envolvendo o surgimento e progressão da doença, novos métodos estão começando a ser utilizados e estudados (Mattia et al. 2018). Entre estes métodos o que mais se destaca é a imunoterapia (Mattia et al. 2018).

1.6 OBSERVAÇÕES

O CPCNP e o melanoma ainda são considerados problemas oncológicos graves e, muitas vezes, o diagnóstico e o tratamento de neoplasias são feitos sem levar em conta a ancestralidade ou etnia. Como visto anteriormente, existem diversas diferenças genéticas e biológicas, entre populações, que podem impactar no surgimento e desenvolvimento de doenças (Özdemir e Dotto 2017). Por esta razão, torna-se importante entender melhor a relação entre ancestralidade e câncer, buscando um diagnóstico mais personalizado, específico e efetivo para tratamento.

2 OBJETIVOS

2.1 OBJETIVO GERAL

Este projeto visa identificar as diferenças genéticas e biológicas entre diferentes grupos, de acordo com a sua ancestralidade, que possam ser relevantes no desenvolvimento e evolução do melanoma e do carcinoma de pulmão de células não pequenas.

2.2 OBJETIVOS ESPECÍFICOS

- Detectar e analisar ancestralidade individual em pacientes do TCGA;
- Avaliação e validação do *pipeline* para ancestralidade individual;
- Avaliar expressão gênica diferencial entre os grupos e determinar principais vias biológicas por *Gene Set Enrichment*;
- Definir as diferenças na composição de células imunes entre os grupos;
- Avaliar o perfil mutacional somático e detecção de assinaturas mutacionais entre os grupos;
- Avaliar possíveis associações entre estes fatores e o prognóstico dos pacientes entre os grupos.

3 MATERIAIS E MÉTODOS

3.1 DESENHO DO ESTUDO

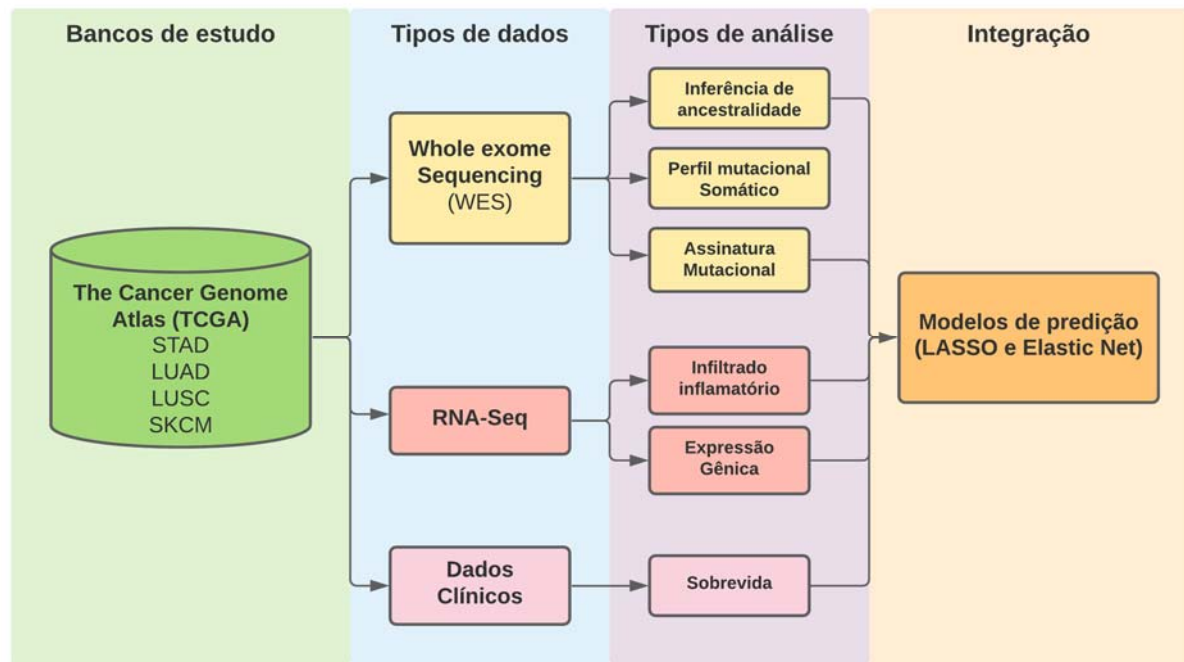


Figura 1 – Desenho do estudo. Para este projeto, foram utilizados dados obtidos a partir do *The Cancer Genome Atlas (TCGA)*, para adenocarcinoma gástrico (STAD) e de pulmão (LUAD), carcinoma de células escamosas de pulmão (LUAD) e melanoma (SKCM). Estes dados podem ser divididos em *whole exome sequencing* (WES), RNA-seq e dados clínicos. A partir destes dados foram realizadas análises para inferência de ancestralidade, perfil mutacional somático, assinatura mutacional, infiltrado inflamatório, expressão gênica e sobrevida. Estas análises foram então, avaliadas, de forma integrada, através de modelos de predição, como LASSO e *Elastic Net*.

3.2 DADOS PÚBLICOS PARA MELANOMA, CPCNP E ADENOCARCINOMA GÁSTRICO

Para este estudo, foram coletados dados do consórcio *The Cancer Genome Atlas-TCGA* para os casos de adenocarcinoma de pulmão, carcinoma de células escamosas de pulmão e melanoma. Além disso, também utilizamos como modelo de testes, dados para casos de adenocarcinoma gástrico. O TCGA processou, analisou e caracterizou molecularmente mais de 20.000 amostras de tumor primário, obtidas a partir de diversos centros ao redor do mundo para 33 tipos tumor, gerando dados clínicos, dados de sequenciamento completo de exoma e

transcriptoma. O conjunto de pacientes aqui extraídos representa as nossas amostras de interesse para o resto do estudo. Ao todo, foram avaliadas 443 amostras para adenocarcinoma gástrico, 519 amostras para adenocarcinoma de pulmão, 504 amostras para carcinoma de células escamosas de pulmão e 470 amostras de melanoma.

3.3 DADOS PÚBLICOS PARA POPULAÇÕES DE REFERÊNCIA

Com o intuito de identificar a variabilidade genética em diversas populações e definir ancestralidade, foram selecionadas populações de referência a partir dos bancos *1000 Genomes Project – Phase 3* (1000G) (The 1000 Genomes Consortium 2015) e *Human Genome Diversity Project* (HGDP) (Rosenberg et al. 2002; Rosenberg 2006) (Tabela 1). Ao todo, foram selecionadas oito populações a partir do 1000G e, posteriormente, foram agrupadas em quatro superpopulações: europeia (EUR), africana (AFR), sul asiática (SAS) e do leste asiático (EAS) (Tabela 1). Essa classificação é a mesma aplicada pelo próprio 1000G. A ancestralidade para estas populações foi inferida utilizando o *software Admixture* em seu modo não-supervisionado (Alexander et al. 2009) (com $K=4$), extraíndo as que apresentaram maior homogeneidade. Já em HGDP, foram selecionadas cinco populações, classificadas como ameríndios (AMR) pelo próprio banco (Tabela 1). As variantes em comum entre os bancos foram filtradas para menor frequência alélica (*minor allele frequency*, MAF) $< 1\%$. Também foram removidas para o teste de equilíbrio de *Hardy-Weinberg* $< 10^{-3}$. Ao todo, foram selecionadas 534.734 variantes.

Tabela 1 - Populações, superpopulações e bancos de dados utilizados como referência para inferência de ancestralidade.

População	Superpopulação	Banco Referência
Yoruba, Ibadan, Nigeria (YRI)	AFR	1000G
Luhya, Webuye, Quênia (LWK)	AFR	1000G
Britânico, Inglaterra e Escócia (GBR)	EUR	1000G
Toscani, Itália (TSI)	EUR	1000G
Japoneses, Tóquio, Japão (JPT)	SAS	1000G
Chinês Han, Pequim, China (CHB)	SAS	1000G
Indiano Telugu, Reino Unido (ITU)	EAS	1000G
Tamil, Sri Lanka, Reino Unido (STU)	EAS	1000G
Colombiano, Colômbia	AMR	HGDP
Surui, Brasil	AMR	HGDP
Karitiana, Brasil	AMR	HGDP
Maya, México	AMR	HGDP
Pima, México	AMR	HGDP

3.4 DETECÇÃO DE VARIANTES GERMINATIVAS

Para a detecção das variantes germinativas, foram recuperados o sequenciamento do conjunto de todos os genes (exoma completo) das amostras de adenocarcinoma gástrico, pulmão, melanoma e carcinoma de células escamosas de pulmão. Esses arquivos foram obtidos diretamente do TCGA e encontram-se no formato BAM (*Binary Alignment Map*). Como parte da detecção de variantes germinativas, as leituras (*reads*) do DNA, contidas nesse arquivo, foram pré-processadas pelos seguintes métodos: remoção de duplicatas de PCR via Picard v2.6.0 (Broad Institute 2018) realinhamento com *Genome Analysis Toolkit* (GATK) v3.8-1 e recalibração de *base quality score* com GATK v3.8-1 (McKenna et al. 2010) e dbSNP versão 144.

O arquivo de target aplicado neste processo foi obtido a partir da interseção do target utilizado pelo TCGA com a região das variantes do banco de referência com 1000G e HGDP. Com este intuito, foi utilizado o software *BEDTools* (Quinlan e Hall 2010).

Variantes germinativas foram detectadas através do módulo *HaplotypeCaller* do GATK v3.8-1 (Auwera et al. 2013) para cada amostra. Posteriormente, o arquivo GVCF destas amostras foi combinado em um único arquivo GVCF com o módulo *CombineGVCFs* do GATK v3.8-1 e foi realizada a genotipagem, em conjunto, com o módulo *GenotypeGVCFs*. Para anotação, foram utilizados *SnpEff* v5.0 (GRCh38.99) (Cingolani et al. 2012b) e *SnpSift* v4.3 (Cingolani et al. 2012a), com os bancos *ABRaOM* v1.0.2 (Naslavsky et al. 2020) e *dbNSFP* v3.5 (Liu et al. 2016). Para o processo de *Variant quality score recalibration* (VQSR) foram utilizados os módulos *VariantRecalibrator* e *ApplyRecalibration* do GATK v3.8-1 (Auwera et al. 2013). Para estes dados foram aplicados os parâmetros padrão de cada *software*.

Ao todo, foram identificadas 7.880 variantes para amostras de adenocarcinoma gástrico, 7.887 variantes para amostras de adenocarcinoma de pulmão, 7.881 variantes para amostras de carcinoma de células escamosas de pulmão e 7.749 variantes para amostras de melanoma, em comum com as regiões do banco de referência.

3.5 PIPELINE PARA ANCESTRALIDADE INDIVIDUAL E GERENCIAMENTO DE WORKFLOWS

A ancestralidade individual corresponde à proporção de ancestralidade em cada indivíduo. Para a sua inferência, foram utilizados dados genômicos obtidos a partir das

populações de referência, extraídas do 1000G e HGDP. Estes dados correspondem a variações genéticas germinativas, mais especificamente ligadas à *single nucleotide variants* (SNVs). Além disso, foi aplicado o *software Admixture* que utiliza um modelo baseado na máxima verossimilhança para estimar a ancestralidade de indivíduos não relacionados.

A fim de automatizar este processo, um *pipeline* foi estruturado. Para isso foi aplicado o *software Snakemake* (Köster e Rahmann 2012). Esta ferramenta tem como função gerenciar o *workflow* do *pipepline*. É empregada uma linguagem própria, muito próxima da linguagem padrão de *python*, fornecendo um ambiente com escalabilidade automática, otimizando o número de processos paralelos, não precisando alterar o *workflow* (Köster e Rahmann 2012). Seu sistema funciona através de regras que indicam como criar arquivos de saída a partir de arquivos de entrada. O *workflow* é deduzido pela dependência entre cada uma das regras (Köster e Rahmann 2012).

3.6 INFERÊNCIA DE ANCESTRALIDADE INDIVIDUAL, VALIDAÇÃO E CLUSTERIZAÇÃO

O *pipeline* foi aplicado para as amostras de adenocarcinoma gástrico e de pulmão, melanoma e carcinoma de células escamosas de pulmão, utilizando dados de sequenciamento completo de exoma disponibilizados pelo TCGA. Os dados obtidos a partir da chamada de variantes germinativas foram processados através do *software* PLINK (Purcell et al. 2007) e um conjunto de critérios para controle de qualidade foram aplicados. Primeiramente SNVs com taxas de chamadas em amostras $< 95\%$ e menor frequência alélica (*minor allele frequency*, MAF) $< 1\%$ foram removidos. SNVs também foram removidos para o teste de equilíbrio de Hardy-Weinberg $< 10^{-6}$ e *pruned* para *linkage disequilibrium* para *window size* = 50, *step size* = 5 e *r² threshold* = 0.2. Estes critérios reduziram o número de variantes utilizadas para a inferência de ancestralidade individual.

A partir das estimativas obtidas para as amostras do TCGA, para cada tipo de tumor, os pacientes foram agrupados utilizando o método de agrupamento *kmeans*. A fim de obter um número ótimo de agrupamentos (*clusters*), foram utilizados os métodos *elbow*, *average silhouette* e *gap statistic* (Tibshirani et al. 2002). Ao final, o número otimizado de clusters para todos os tipos de tumor foi de três. Cada *cluster* é definido, principalmente, por uma das seguintes ancestralidades: asiática (EAS), europeia (EUR) e africana (AFR) (Tabela 2).

Tabela 2 - Número de pacientes por tipo de tumor no TCGA e por *cluster*.

Tipo de Tumor (TCGA)	Cluster	Número de pacientes	Ancestralidade
Adenocarcinoma Gástrico	1	39	AFR
Adenocarcinoma Gástrico	2	315	EUR
Adenocarcinoma Gástrico	3	89	EAS
Adenocarcinoma de Pulmão	1	9	EAS
Adenocarcinoma de Pulmão	2	456	EUR
Adenocarcinoma de Pulmão	3	54	AFR
Carcinoma de Células Escamosas de Pulmão	1	12	EAS
Carcinoma de Células Escamosas de Pulmão	2	32	AFR
Carcinoma de Células Escamosas de Pulmão	3	460	EUR
Melanoma	1	450	EUR
Melanoma	2	8	AFR
Melanoma	3	12	EAS

Com o intuito de validar as estimativas obtidas a partir da metodologia anteriormente descrita, utilizamos os dados apresentados pelo *The Cancer Genetic Ancestry Atlas* (TCGAA) (Yuan et al. 2018). Este estudo utilizou o *software STRUCTURE* (Pritchard et al. 2000) e dados brutos a partir de arquivos em formato CEL de *Affymetrix Human SNP Array 6.0* (YUAN et al. 2018). Este tipo de dado contém informação de marcadores genéticos, incluindo SNPs, cobrindo todo o genoma. Estes dados foram então combinados, calibrados e normalizados para cada tipo de tumor, utilizando o *software Affymetrix Power Tools 1.18.2* (Yuan et al. 2018). Para a chamada de genótipo de SNPs foi utilizado o *software Birdseed version 2* (Korn et al. 2008; Yuan et al. 2018). O estudo utilizou somente quatro superpopulações de referência (população africana, do leste asiático, europeia e ameríndia), a partir dos bancos *1000 Genomes Project* (The 1000 Genomes Consortium 2015) e *International HapMap Project* (The International HapMap Consortium 2003).

Para comparar as estimativas de ancestralidade entre o *pipeline* aqui descrito e o TCGAA foi utilizado o método gráfico de *Bland-Altman* (Altman e Bland 1983). Esta metodologia tem como objetivo quantificar a concordância entre dois métodos, construindo limites de concordância (Giavarina 2015). Estes limites estatísticos são calculados utilizando a média e o desvio padrão das diferenças entre duas medições (Giavarina 2015). O gráfico resultante é um gráfico de dispersão com eixos X e Y, nos quais Y indica a diferença entre duas medidas pareadas, enquanto o eixo X representa a média entre as mesmas duas medidas pareadas (Giavarina 2015).

3.7 ANÁLISE DO INFILTRADO INFLAMATÓRIO

Com o intuito de avaliar o infiltrado inflamatório de cada um dos pacientes de interesse, os softwares *CIBERSORT* (Newman et al. 2015), *xCell* (Aran et al. 2017) e *Estimation of STromal and Immune cells in MAlignant Tumours using Expression data* (ESTIMATE) (Yoshihara et al. 2013) foram aplicados. Esses métodos são amplamente utilizados e apresentam abordagens distintas. *ESTIMATE* tem como propósito estimar a pureza do tumor, células estromais e imunes infiltradas no tumor (Yoshihara et al. 2013). Já *xCell* e *CIBERSORT* inferem a quantidade de tipos de células imunes e/ou estromais individualmente. *xCell* utiliza um painel de assinaturas com 64 tipos celulares estromais e imunes (Aran et al. 2017), enquanto *CIBERSORT* possui um painel de assinaturas para 22 tipos de células imunes (LM22) (Newman et al. 2015).

Para os três métodos foram utilizados dados de expressão gênica normalizados, de cada paciente, em *Fragments Per Kilobase of transcript per Million* (FPKM). Estes arquivos são gerados a partir do *workflow* descrito pelo TCGA (National Cancer Institute-NCI 2020). A partir de um arquivo no formato BAM, já alinhado com o genoma de referência, é utilizado o software HTSeq (Anders et al. 2015), com o objetivo de obter o número absoluto de *reads* de cada gene. Posteriormente este dado é normalizado pelo método de FPKM. Essa normalização ocorre dividindo o número de *reads* pelo tamanho do gene e pelo número total de *reads* mapeadas em genes codificantes para proteínas (NCI 2020).

Para comparação dos grupos, os dados foram avaliados para testes de normalidade através de: *Shapiro-Wilk* (Shapiro e Wilk 1965) e *Kolmogorov-Smirnov* (Kolmogorov 1933) e para testes de igualdade de variâncias, com o teste de *Levene* (Olkin 1960). Nestas duas análises, nenhuma das variáveis apresentaram normalidade e/ou apresentaram igualdade de variâncias entre os grupos. Portanto, foi utilizado o teste de *Kruskal-Wallis* (Kruskal e Wallis 1952) com o pós-teste de *bonferroni*, com o valor de p ajustado. Foi considerado como significativo valor de p abaixo de 0.05.

3.8 ANÁLISE DO PERFIL MUTACIONAL SOMÁTICO

A fim de avaliar se existe diferença de genes frequentemente mutados entre os grupos, foi utilizado um conjunto de variantes somáticas para os pacientes do TCGA, obtidos a partir de um arquivo no formato *Mutation Annotation Format* (MAF).

Os pacientes foram subdivididos de acordo com a identificação da ancestralidade e foram utilizados o *software MutSig* (Lawrence et al. 2013) e o pacote *maftools* (Mayakonda et al. 2018). A ferramenta *MutSig* analisa um conjunto de mutações somáticas a fim de identificar genes que foram mutados com mais frequência do que o esperado, considerando a probabilidade de uma base sofrer mutação por acaso. Em seguida, utilizamos o pacote *maftools* para sumarizar esses genes com o perfil mutacional de cada um dos grupos, além de criar gráficos utilizando os resultados gerados pelo *MutSig*.

3.9 ANÁLISE DA ASSINATURA MUTACIONAL

A exposição das assinaturas mutacionais foi avaliada considerando o catálogo de assinaturas previamente disponibilizado pelo *Catalogue of Somatic Mutations in Cancer* (COSMIC) versão 3 (Tate et al. 2019; Wellcome Sanger Institute 2020) como referência. Este catálogo é baseado na análise de 19.184 exomas e 1865 genomas completos em 40 tipos distintos de câncer humano (Wellcome Sanger Institute 2020).

A contribuição de cada assinatura descrita no COSMIC foi estimada por meio de três pacotes para linguagem R: *MutationalPatterns* (Blokzijl et al. 2018), *deconstructSigs* (Rosenthal et al. 2016) e *signeR* (Rosales et al. 2017). Ambos encontram-se disponíveis no repositório *Bioconductor* (Gentleman et al. 2004).

Para comparação dos grupos, os dados foram avaliados para testes de normalidade através de: *Shapiro-Wilk* (Shapiro e Wilk 1965) e *Kolmogorov-Smirnov* (Kolmogorov 1933) e para testes de igualdade de variâncias, com o teste de *Levene* (Olkin 1960). Nestas duas análises, nenhuma das variáveis apresentaram normalidade e/ou apresentaram igualdade de variâncias entre os grupos. Portanto, foi utilizado o teste de *Kruskal-Wallis* (Kruskal e Wallis 1952) com o pós-teste de *bonferroni*, com o valor de p ajustado. Foi considerado como significativo valor de p abaixo de 0.05.

3.10 ANÁLISE DE EXPRESSÃO GÊNICA DIFERENCIAL

Esta análise foi realizada aplicando o *software DESeq2* (Love et al. 2014). Este método consiste na análise diferencial de dados com número absoluto de *reads* por gene, através de estimativas de encolhimento para dispersões e *fold changes*, melhorando a estabilidade e a interpretação das estimativas.

Os arquivos com o total de transcritos para cada gene foram obtidos a partir do TCGA, para cada um dos pacientes de interesse. Estes arquivos são gerados a partir do *workflow* descrito pelo TCGA (NCI 2020). Foi utilizado *software Biobambam* para converter as leituras obtidas do sequenciamento, para arquivos no formato FASTQ. Estes foram alinhados segundo o genoma de referência (GRCh38), utilizando *STAR 2.4.0.1* (Dobin et al. 2013). Os fragmentos foram anotados segundo às coordenadas de cada gene na lista de *targets* via *HTSeq v0.6.1* (Anders et al. 2015). Ao final, para cada gene é atribuído um valor correspondente ao total de *reads* no respectivo gene, que, por sua vez, representa sua atividade transcricional. Esses resultados são então utilizados para análise de expressão diferencial entre os grupos de interesse por meio do *software DESeq2* (Love et al. 2014). Todos os genes foram ranqueados de acordo com o seu *p-value* ajustado e foram selecionados até 1000 genes, desde que todos possuísem um *p-value* ajustado abaixo de 0.05.

Com o objetivo de avaliar o impacto funcional das diferenças transcricionais dos genes diferencialmente expressos, foi realizada uma análise de enriquecimento (GSEA, *Gene Set Enrichment Analysis*) (Mootha et al. 2003; Subramanian et al. 2005). Este método permite a identificação de vias biológicas que estariam enriquecidas em diferentes condições (Mootha et al. 2003; Subramanian et al. 2005). Para esta análise, foi aplicada a nomenclatura adotada pelo banco de dados *Gene Ontology* (GO) para processos biológicos.

A fim de compilar a informação obtida de algumas comparações, foi utilizado o *software Cytoscape* (Shannon et al. 2003; Reimand et al. 2019). Juntamente, foram utilizadas as seguintes extensões: *AutoAnnotate* (Kucera et al. 2016), *Enrichment Map* (Merico et al. 2010), *ClusterMaker* (Morris et al. 2011) e *WordCloud* (Oesper et al. 2011). *Enrichment Map* foi aplicado com o objetivo de representar visualmente as principais redes e vias encontradas através do uso de GSEA, utilizando um *q-value* menor do que 0.05. Em seguida, foi utilizado *AutoAnnotate* em conjunto com *ClusterMaker* e *WordCloud*, a fim de agrupar e adicionar *tags* a cada um dos clusters encontrados. Clusters formados por uma única via foram removidos.

3.11 ANÁLISE DE SOBREVIDA

A sobrevida foi analisada utilizando o estimador de *Kaplan-Meier* (Goel et al. 2010) para os pacientes obtidos a partir do TCGA, considerando o estadiamento definido para adenocarcinoma gástrico e de pulmão, melanoma e carcinoma de células escamosas de pulmão. Nesta etapa, foram utilizados os pacotes *survival* e *survminer* disponíveis na linguagem R. O *p-value* foi calculado a partir do teste de *log-rank*.

3.12 PIPELINE PARA MODELOS DE PREDIÇÃO

Buscando avaliar o impacto do infiltrado inflamatório, da assinatura mutacional e expressão gênica na sobrevida, foram utilizados dois modelos de predição: *least absolute shrinkage and selection operator* (Lasso) (Tibshirani 1996) e *Elastic Net* (Zou e Hastie 2005). Para isto, estes modelos foram aplicados através do pacote *glmnet* (Friedman et al. 2010; Simon et al. 2011) para linguagem R.

Inicialmente, cerca de 30.000 variáveis foram avaliadas, incluindo dados para expressão gênica, infiltrado inflamatório e assinatura mutacional. Devido às diferenças de escalas, estas foram previamente normalizadas. As variáveis avaliadas também foram filtradas antes de serem aplicadas aos modelos de regressão. Este processo foi realizado através das seguintes etapas:

1. Remoção de amostras com tempo = 0 ou sem informação;
2. Remoção de variáveis com intervalo interquantil (IQR) = 0;
3. Remoção de variáveis com *p-value* > 0.05 utilizando modelo univariado de riscos proporcionais de Cox (Cox 1972), através do pacote *Survival* (Therneau 2020) para linguagem R;
4. Agrupamento de variáveis por *k-medoids* através do pacote *kmed* (Budiaji 2021). O número de *clusters* otimizado foi selecionado a partir do gráfico de *silhouette*;
5. Remoção de *centroids* com *p-value* > 0.05, através do modelo multivariado de riscos proporcionais de Cox;
6. Recuperação de variáveis relacionadas aos *centroids* significativos;
7. Teste e seleção das variáveis recuperadas juntamente com os *centroids* significativos através do modelo multivariado de riscos proporcionais de Cox. Foram consideradas significativas variáveis com *p-value* < 0.05;

Após o processo de seleção de variáveis, estas foram aplicadas aos modelos de regressão

de Lasso e *Elastic Net*, juntamente com as informações para idade e estadiamento do tumor. A qualidade dos modelos foi avaliada segundo o gráfico de *cross-validation* e como medida de ajuste foram utilizados *Harrell's concordance* (Harrell e Mark 1996) e *deviance*. O λ para *one-standard-error* foi aplicado. Foram selecionadas as variáveis consideradas significativas tanto pelo modelo de Lasso quanto por *Elastic Net*. A partir disso foi calculado o *hazard-ratio*.

Os genes considerados significativos foram avaliados utilizando os seguintes bancos de dados: *Cancer Gene Census* (Sondka et al. 2018), *Dependency Map* (DepMap) (Broad Institute 2021), *Protein Cancer Atlas* (Pontén 2008; Uhlen et al. 2017) e *Catalogue of Cancer Genes* (Cheng et al. 2014).

4 RESULTADOS

4.1 IDENTIFICAÇÃO DE ANCESTRALIDADE INDIVIDUAL EM AMOSTRAS DO TCGA E ESTRUTURAÇÃO DO PIPELINE

4.1.1 Pipeline para ancestralidade individual

Um passo importante para avaliar a relação entre ancestralidade e câncer é identificar a diversidade populacional em nossas amostras de interesse. Com este objetivo, utilizamos o *software Admixture* (Alexander et al. 2009).

Também foram definidas as populações de referência, necessárias para a inferência da ancestralidade. Ao todo foram selecionadas quatro superpopulações a partir de 1000G (The 1000 Genomes Project Consortium 2015) e uma extraída do banco HGDP (Rosenberg et al. 2002; Rosenberg 2006), totalizando treze populações.

Como modelo de teste para desenvolvimento do *workflow*, utilizamos dados já previamente disponíveis, de amostras normais de pacientes de adenocarcinoma gástrico do TCGA. Filtros para *minor allele frequency*, *missingness per individual*, *missingness per marker*, *linkage disequilibrium* e *hardy-weinberg equilibrium*, foram utilizados como procedimentos para controle de qualidade (Turner et al. 2011; Marees et al. 2018), tanto para as amostras de interesse, quanto para os bancos de referência. Estes procedimentos foram realizados utilizando o *software* PLINK (Purcell et al. 2007) e, ao final, os resultados foram carregados no *Admixture* (Figura 2).

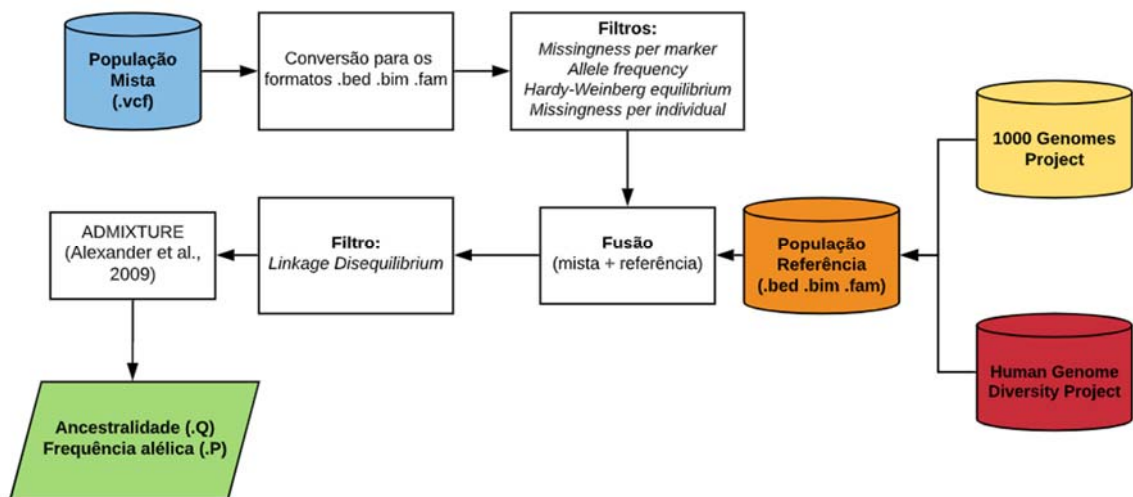


Figura 2 - Workflow para identificação de ancestralidade em população mista a partir do banco com as populações de referência. O arquivo no formato VCF, com todas as amostras de interesse, foi utilizado como input. Em seguida, foi convertido para o formato binário do PLINK e filtrado segundo parâmetros de *minor allele frequency*, *missingness per individual*, *missingness per marker*, *linkage disequilibrium* e *hardy-weinberg equilibrium*. Os dados das amostras de interesse foram então fundidos com as populações de referência e foi aplicado o parâmetro de *linkage disequilibrium*. Como último passo, foi utilizado o software *Admixture*. Ao final, o software fornece as proporções das ancestralidades e a frequência alélica de cada uma das variantes utilizadas.

Hoje não existe, publicamente disponível, um *pipeline* já estruturado para o pré-processamento de dados, inferência da ancestralidade individual e controle de qualidade. Por isso, como parte do projeto, decidimos automatizar este *workflow* e disponibilizá-lo para comunidade científica. Com este intuito, utilizamos um sistema (Figura 3) de gerenciamento de *workflows* baseado na linguagem python, denominado *Snakemake* (Köster e Rahmann 2012). A utilização deste sistema permite que regras sejam criadas, a fim de personalizar e configurar a execução de códigos de forma bastante eficiente, além de estabelecer uma estrutura de diretórios que seja organizada e intuitiva. Em sua versão atual, é esperado como *input* um único arquivo no formato VCF (*Variant Call Format*) com todas as amostras de interesse (Figura 3). São realizadas etapas de pré-processamento do VCF, conversão para o formato necessário ao software PLINK (Purcell et al. 2007) e, em seguida, são aplicados filtros para i) ausência de dados por marcador, ii) menor frequência alélica, iii) equilíbrio de *Hardy-Weinberg*, iv) ausência de dados por indivíduo e v) desequilíbrio de ligação (Figura 3B). Ao todo, passaram por estes filtros 5.831 variantes para amostras de adenocarcinoma gástrico, 5.837 variantes para amostras de adenocarcinoma de pulmão, 5.690 variantes para amostras de carcinoma de células escamosas de pulmão e 5.683 variantes para amostras de melanoma.

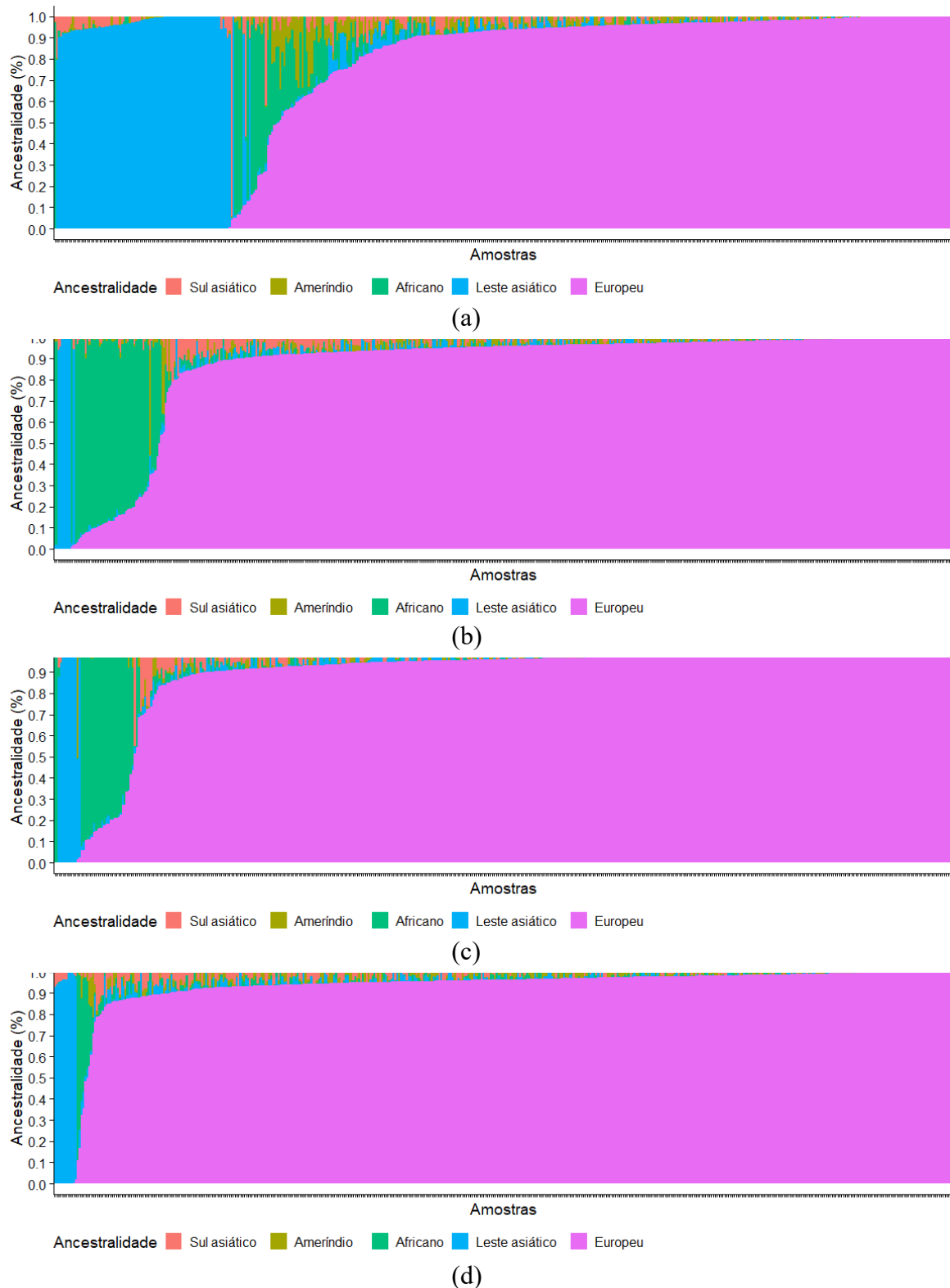


Figura 4 - Ancestralidade individual em pacientes do TCGA. Os pacientes foram analisados de acordo com o banco de referência, com cinco superpopulações. No eixo X são representadas as amostras, enquanto o eixo Y corresponde à proporção de ancestralidade de cada população parental. Cada uma das cores representa uma das cinco ancestralidades, de acordo com as superpopulações de referência, incluindo (A) adenocarcinoma gástrico (N=443); (B) adenocarcinoma de pulmão (N=519); (C) carcinoma de células escamosas de pulmão (N=504); (D) melanoma (N=470).

4.1.2 Ancestralidade individual em dados de adenocarcinoma gástrico do TCGA e validação

Com o objetivo de validar os resultados do *pipeline*, como modelo de teste, foram utilizados os dados de ancestralidade dos pacientes de adenocarcinoma gástrico (Figura 4A). Pudemos observar que a maior parte dos indivíduos possui uma grande proporção de ancestralidade europeia.

Os resultados obtidos através do *pipeline* foram então comparados com dados de ancestralidade inferida pelo consórcio TCGAA (Yuan et al. 2018). O grupo utilizou dados de genoma completo disponíveis no TCGA. Nessa comparação, reduzimos o número de superpopulações para quatro (AFR, AMR, EUR e EAS) a fim de utilizar as mesmas superpopulações utilizadas pelo estudo.

A comparação foi realizada com a análise gráfica de *Bland-Altman* (Altman e Bland 1983) e separando as proporções de AFR (Figura 5A), AMR (Figura 5B), EAS (Figura 5C) e EUR (Figura 5D). Essa abordagem permite observar, de forma bastante específica, se o número reduzido de variantes disponíveis poderia ter algum impacto para distinguir cada um dos grupos. Percebemos, porém, que, em sua maior parte, os resultados se mostraram bastante semelhantes. Na maior parte dos casos, foi observada uma diferença de, no máximo, 0.05 em relação aos dados do TCGAA (Figura 5). Além disso, a média das diferenças entre os métodos foi de aproximadamente zero em todos os casos (Figura 5). Vale destacar, ainda, que a estimativa da ancestralidade pelo consórcio foi realizada pelo *software Structure* (Pritchard et al. 2000), desta forma, concluímos que, mesmo com metodologias distintas, os resultados obtidos foram semelhantes. Esses resultados endossam o *pipeline* proposto no presente trabalho, bem como a parametrização dos softwares nas diferentes etapas de análise.

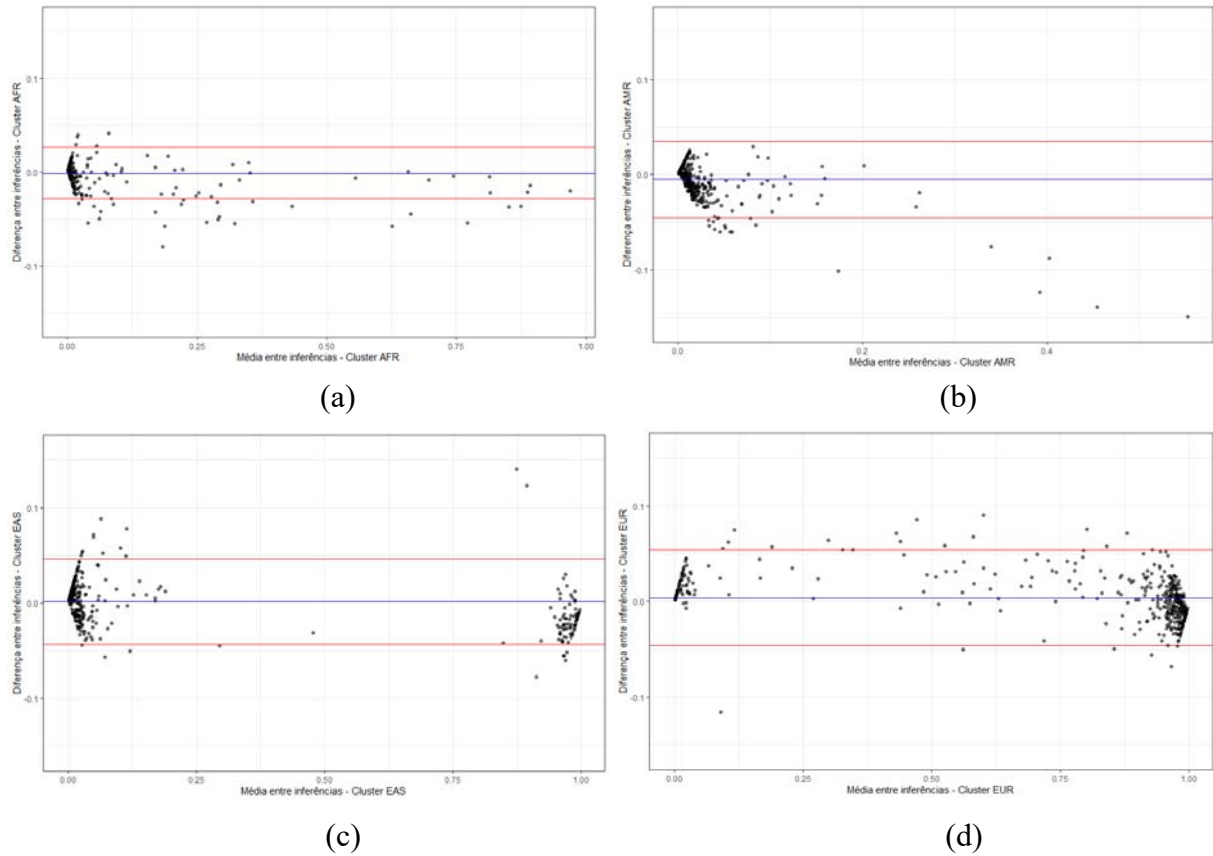


Figura 5 - Comparação entre os dados obtidos a partir do *pipeline* de ancestralidade individual e o TCGAA, para pacientes de adenocarcinoma gástrico do TCGA pelo método *Bland-Altman*. Os dados de ancestralidade de todos os pacientes de adenocarcinoma gástrico do TCGA, produzidos pelo TCGAA, foram comparados com os resultados obtidos com o *pipeline* de ancestralidade individual para 4 superpopulações (AFR, AMR, EAS e EUR). Os dados de todos os pacientes foram divididos de acordo com as proporções de cada superpopulação. Os pontos representam cada um dos pacientes. No eixo X são representadas as médias das inferências entre os dois métodos, enquanto o eixo Y corresponde à diferença das inferências entre os métodos. As linhas vermelhas representam os limites de concordância superior e inferior, e a linha azul o viés; **(A)** Proporções para AFR; **(B)** Proporções para AMR; **(C)** Proporções para EAS; **(D)** Proporções para EUR.

4.1.3 Otimizando o número de variantes

O número de variantes é um importante fator para inferir ancestralidade com confiança, porém, não é o único. Cada variante possui um grau de relevância para este tipo de análise e, por isso, é interessante selecionar aquelas que são mais informativas e encontrar um equilíbrio entre quantidade e eficiência. Para alcançar este objetivo, alguns métodos foram avaliados.

Primeiro, utilizamos o valor de F_{st} (Weir e Cockerham 1984), calculado através do *software* PLINK (Purcell et al. 2007), *Infocalc* (Rosenberg et al. 2003), *Lancaster Estimator of Independence* (LEI) (Wathen et al. 2019) e, finalmente, realizamos *random sampling* por 100 vezes e comparamos com os resultados obtidos a partir dos outros métodos.

Os testes foram realizados com todas as variantes do banco de referência (N=534.734), com as cinco superpopulações e foram selecionadas as 5000, 4000, 3000, 2000, 1000, 500 e 100 variantes mais bem ranqueadas para cada um dos métodos. Em seguida, essas amostragens de variantes foram utilizadas para inferir ancestralidade através do modo não supervisionado do *Admixture* e os resultados foram comparados com os valores obtidos a partir das 534.734 variantes. Como resultado, F_{st} , *Infocalc* e *random sampling* não foram capazes de identificar e separar as cinco superpopulações em nenhum dos conjuntos. Já LEI foi capaz de identificar as cinco superpopulações com o conjunto de 5000 variantes, porém, abaixo deste número o método não foi capaz de reproduzir os mesmos resultados. A redução de variantes, em todos os métodos, dificultou a distinção das superpopulações AMR, EUR e SAS. Isto se deve à proximidade genética entre estas populações, prejudicando a análise.

Devido à vantagem apresentada pelo método LEI, decidimos ainda melhorar o seu desempenho e utilizá-lo em todas as etapas. Este método tem como objetivo identificar as variantes mais relevantes para a inferência de ancestralidade (Wathen et al. 2019). Para isso, é atribuído um *score* para cada uma das variantes, baseando-se na frequência dos alelos de referência e alternativo (Wathen et al. 2019). Porém, esta abordagem mostrou-se limitada ao número de populações e pela proximidade entre elas.

Observamos, ainda, que as variantes mais relevantes para distinguir as superpopulações AMR, EUR e SAS eram removidas por apresentarem um *score* intermediário. Por esta razão, ao invés de utilizarmos todas as superpopulações em conjunto, dividimos as cinco superpopulações em dez pares (AFR-EUR, AFR-SAS, AFR-EAS, AFR-AMR, EUR-SAS, EUR-EAS, EUR-AMR, SAS-EAS, SAS-AMR, EAS-AMR) e as analisamos separadamente. Desta forma, pudemos selecionar, de maneira uniforme, as variantes mais relevantes. Com esta abordagem, fomos capazes de discriminar todas as superpopulações, mesmo com um conjunto de 100 variantes.

A análise gráfica de *Bland-Altman* foi realizada a fim de comparar os valores obtidos utilizando cada conjunto de variantes (Figura 6). Ao final, observamos que foram separadas as proporções de AFR (Figura 6A), AMR (Figura 6B), EAS (Figura 6C), EUR (Figura 6D) e SAS (Figura 6E). Como destacado anteriormente, o número de variantes também é importante, e por esta razão, apesar de o conjunto de 100 variantes ser capaz de distinguir as populações, há uma perda de acurácia, principalmente quando observamos o cluster para SAS (Figura 6E). Nesse cenário, o uso de conjuntos menores do que 500 parece prejudicar muito na inferência da ancestralidade e não é recomendado.

Concluimos que a avaliação do total de SNPs para estimação da ancestralidade foi necessária para discriminar, de maneira acurada, as populações de interesse.

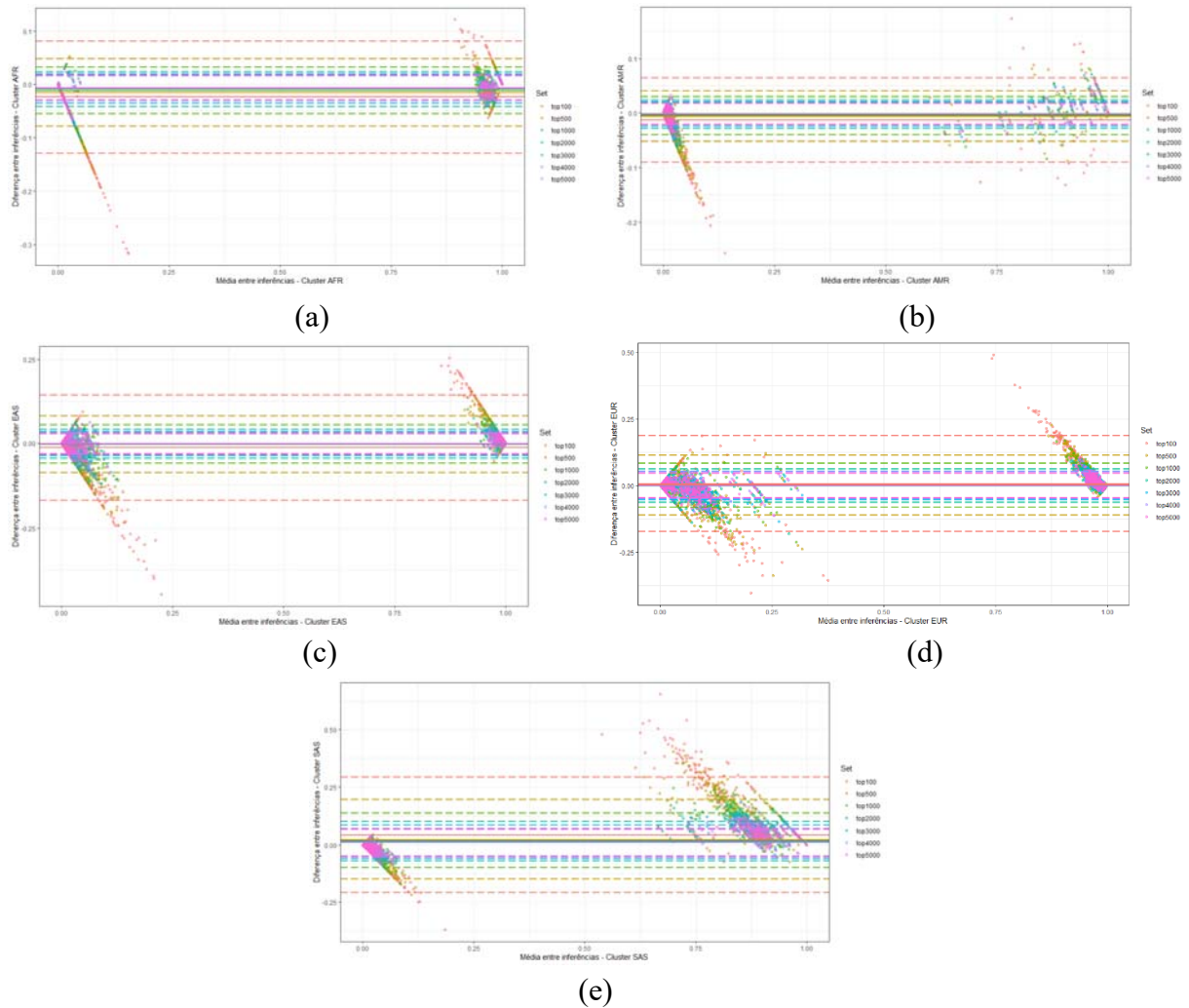


Figura 6 - Comparação da ancestralidade individual do banco de referência com 534.734 variantes e a ancestralidade individual do banco de referência com variantes selecionadas por LEI, em conjuntos de 5000, 4000, 3000, 2000, 1000, 500 e 100 variantes. Todos os indivíduos do banco de referência foram analisados utilizando o *pipeline* para ancestralidade individual. As proporções esperadas foram obtidas utilizando todas as variantes do banco de referência (534.734) e foram comparadas com conjuntos de 5000, 4000, 3000, 2000, 1000, 500 e 100 variantes obtidos com LEI. Seus dados foram separados de acordo com as proporções de AFR, AMR, EAS, EUR e SAS, e analisados por *Bland-Altman*. Os pontos representam cada um dos indivíduos do banco. No eixo X são representadas as médias das inferências entre o valor esperado e cada um dos conjuntos de variantes. O eixo Y corresponde à diferença das inferências entre o valor esperado e o valor obtido a partir de cada um dos conjuntos de variantes. As linhas tracejadas representam os limites de concordância superior e inferior, e as linhas sólidas o viés; **(A)** Proporções para AFR; **(B)** Proporções para AMR; **(C)** Proporções para EAS; **(D)** Proporções para EUR; **(E)** Proporções para SAS.

4.2 INFILTRADO INFLAMATÓRIO

Com o objetivo de investigar a composição do infiltrado inflamatório e sua associação com ancestralidade, utilizamos três programas: *ESTIMATE* (Yoshihara et al. 2013), *CIBERSORT* (Newman et al. 2015) e *xCell* (Aran et al. 2017). Esses métodos são amplamente utilizados e apresentam abordagens distintas. *ESTIMATE* tem como propósito estimar a pureza do tumor, células estromais e imunes infiltradas no tumor (Yoshihara et al. 2013). Já *xCell* e *CIBERSORT* inferem a quantidade de tipos de células imunes e/ou estromais individualmente. *xCell* utiliza um painel de assinaturas com 64 tipos celulares estromais e imunes (Aran et al. 2017), enquanto *CIBERSORT* possui um painel de assinaturas para 22 tipos de células imunes (Newman et al. 2015).

Devido à não normalidade e/ou à heteroscedasticidade dos resultados foi utilizado o teste de *Kruskal-Wallis* (Kruskal e Wallis 1952) com o pós-teste de *bonferroni*, com o valor de *p* ajustado. Foi considerado como significativo valor de *p* abaixo de 0.05. Para *xCell* e *CIBERSORT* somente resultados com valor de *p* significativo para o teste de *Kruskal-Wallis* são aqui apresentados.

Em adenocarcinoma gástrico, analisando os resultados obtidos a partir dos softwares *xCell* (Figura 7A) e *CIBERSORT* (Figura 7B), observamos diferenças significativas em múltiplos tipos de células T, como: células T Natural Killer (NKT), T regulatórias, T CD4 positivas de memória em repouso e T foliculares auxiliares. Nestes quatro casos, as diferenças foram detectadas, principalmente em indivíduos de ancestralidade asiática (cluster 3) e africana (cluster 1). Além disso, foram identificadas diferenças significativas em pacientes com ancestralidade europeia (cluster 2) para monócitos (Figura 7A) e macrófagos M2 (Figura 7B), um fenótipo muito frequente em macrófagos associados a tumores (Li et al. 2019).

Também foram identificadas diferenças significativas entre células estromais, como fibroblastos e condrócitos, indicando menores níveis em indivíduos com ancestralidade africana (cluster 1). Isto pôde ser confirmado pela utilização do *software ESTIMATE*, indicando uma grande diferença nos valores para células estromais, e, portanto, na pureza do tumor (Figura 7C).

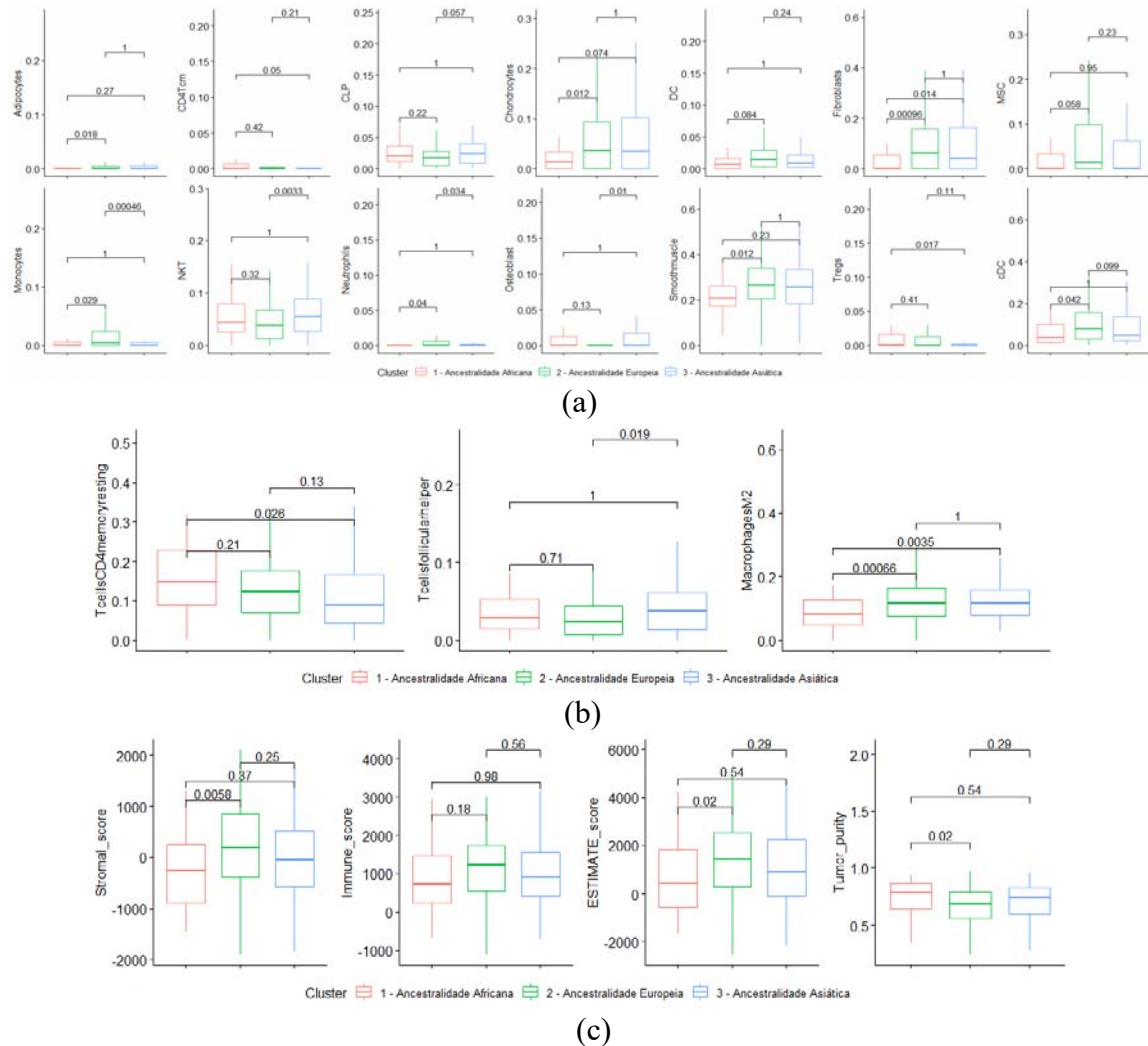


Figura 7 - Comparação dos níveis de células imunes e estromais entre os clusters 1,2 e 3 em pacientes de adenocarcinoma gástrico, aplicando os softwares *xCell*, *CIBERSORT* e *ESTIMATE*. Os grupos foram avaliados com o pós-teste de *bonferroni*. Foi considerado significativo valor de $p < 0.05$; (A) *xCell*; (B) *CIBERSORT*; (C) *ESTIMATE*.

Avaliando os pacientes para adenocarcinoma de pulmão, os dados de *xCell* (Figura 8A) e *CIBERSORT* (Figura 8B), indicam diferenças significativas entre os grupos, principalmente em: macrófagos M2, monócitos e células dendríticas. Indivíduos com ancestralidade europeia (*cluster* 2) apresentaram maiores níveis de monócitos e macrófagos M2 em comparação a indivíduos com ancestralidade africana (*cluster* 3). Já células dendríticas, os resultados indicam diferenças significativas em células dendríticas plasmocitóides e ativadas. Enquanto, no primeiro, indivíduos com ancestralidade asiática (*cluster* 1) apresentaram menores níveis em relação aos outros dois grupos, em células dendríticas ativadas foi observado o oposto. O uso do software *Estimate* não apresentou valores significativos em nenhuma das comparações (Figura 8C)

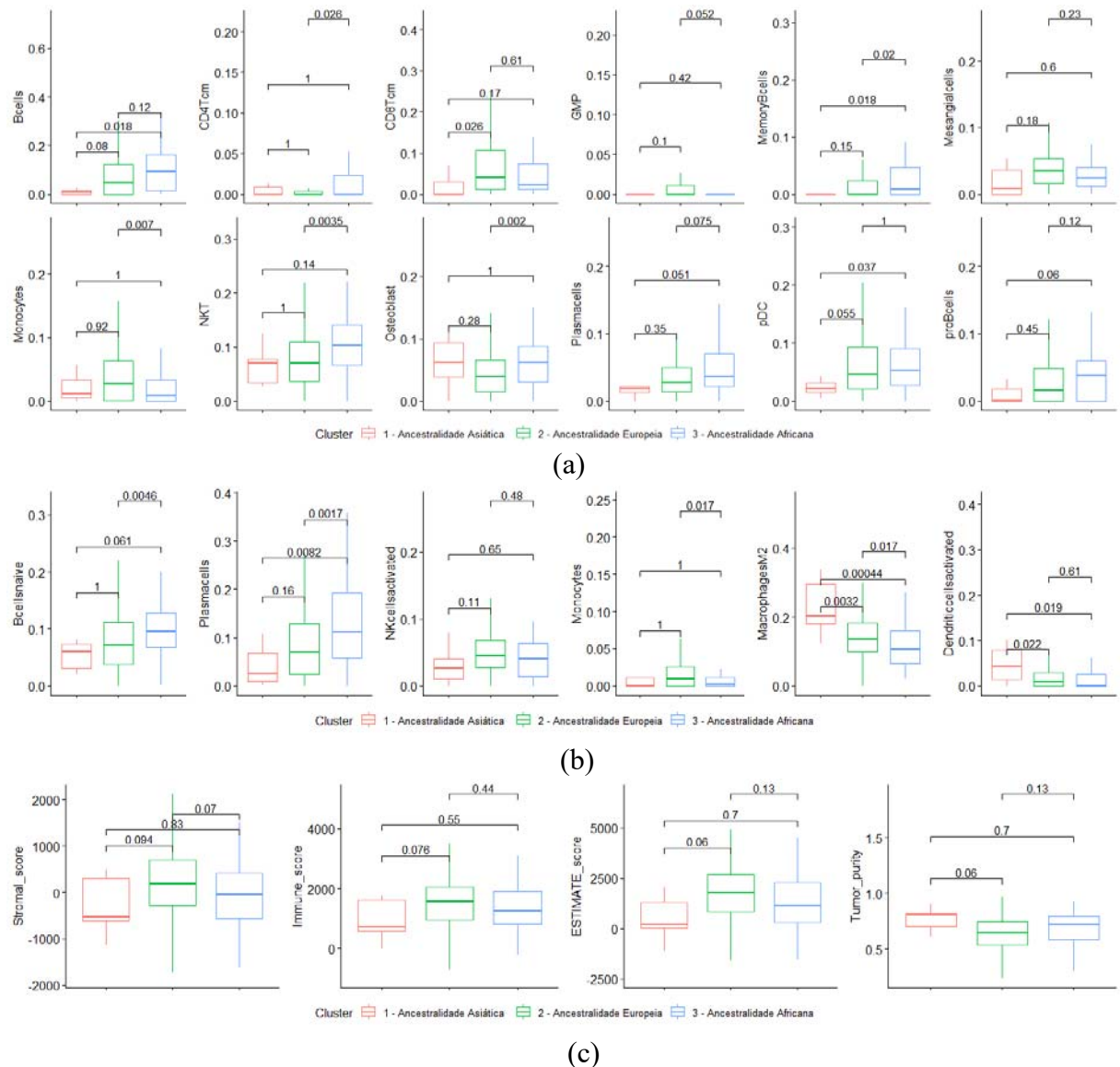


Figura 8 - Comparação dos níveis de células imunes e estromais entre os clusters 1,2 e 3 em pacientes de adenocarcinoma de pulmão, aplicando os softwares *xCell*, *CIBERSORT* e *ESTIMATE*. Os grupos foram avaliados com o pós-teste de *bonferroni*. Foi considerado significativo valor de $p < 0.05$; (A) *xCell*; (B) *CIBERSORT*; (C) *ESTIMATE*.

Em carcinoma de células escamosas de pulmão, tanto *xCell* (Figura 9A) e *CIBERSORT* (Figura 9B) sugerem diferenças para macrófagos M2, entre indivíduos com ancestralidade europeia (*cluster* 3) e africana (*cluster* 2), com os últimos apresentando os menores valores, assim como foi percebido nos casos de adenocarcinoma gástrico e de pulmão. Além disso, *xCell* sugere que existem diferenças nos níveis de macrófagos, apresentando o mesmo perfil do que foi observado a partir de macrófagos M2 (Figura 9A). Os três métodos também foram aplicados para as amostras de melanoma, mas não foram encontradas diferenças significativas em nenhum dos casos.

Com estes dados, foi possível concluir que existem diferenças claras em relação ao infiltrado inflamatório, principalmente em relação a monócitos, macrófagos e macrófagos M2. Foi possível observar que, nestas três células, indivíduos com ancestralidade africana apresentaram uma tendência a menores valores, em relação aos outros dois grupos. Isso pôde ser percebido em adenocarcinoma gástrico e de pulmão e em carcinoma de células escamosas de pulmão.

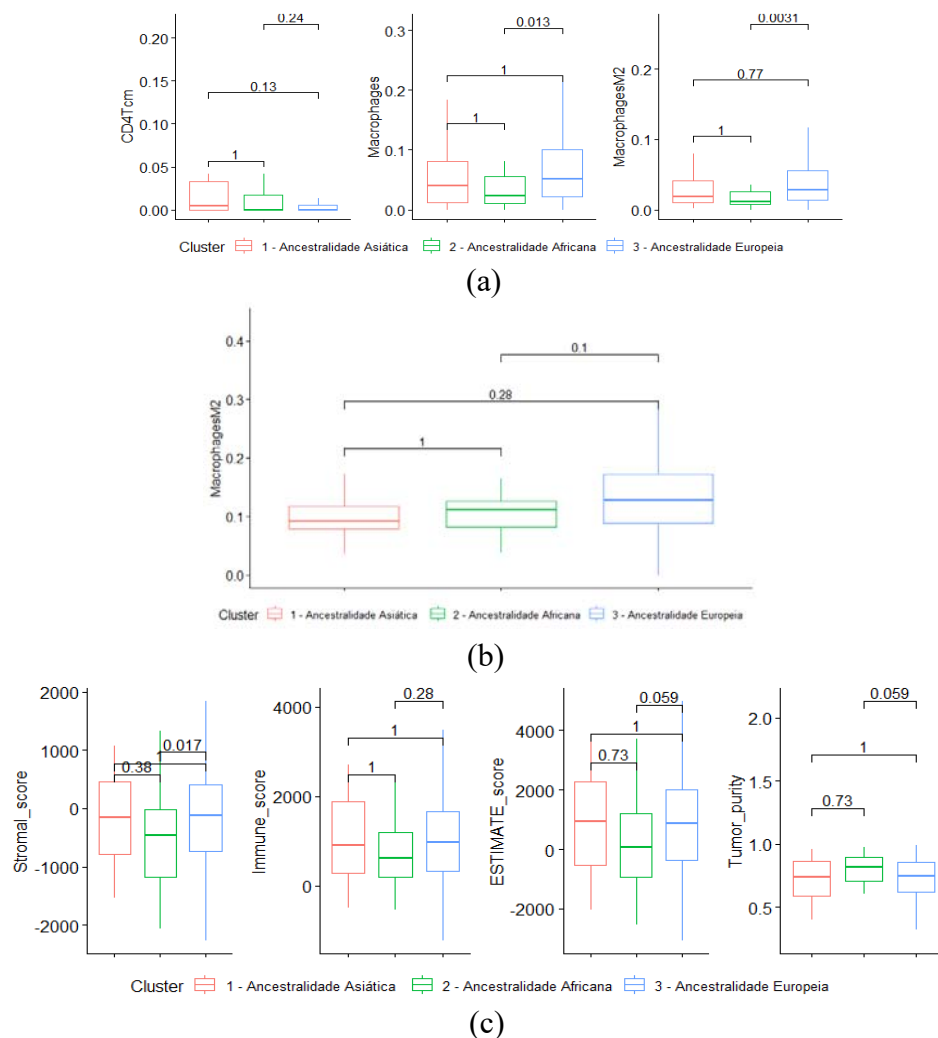


Figura 9 – Comparação dos níveis de células imunes e estromais entre os clusters 1,2 e 3 em pacientes de carcinoma de células escamosas de pulmão, aplicando os softwares *xCell*, *CIBERSORT* e *ESTIMATE*. Os grupos foram avaliados o pós-teste de bonferroni. Foi considerado significativo valor de $p < 0.05$; **(A)** *xCell*; **(B)** *CIBERSORT*; **(C)** *ESTIMATE*.

4.3 PERFIL MUTACIONAL SOMÁTICO

A fim de avaliar se existe diferença de genes frequentemente mutados entre indivíduos com diferentes ancestralidades, foi utilizado um conjunto de variantes somáticas dos pacientes com adenocarcinoma gástrico, adenocarcinoma de pulmão e carcinoma de pulmão de células escamosas.

Cada um dos grupos foi, então, avaliado utilizando o software *MutSig* (Lawrence et al. 2013) e o pacote para a linguagem R, *maftools* (Mayakonda et al. 2018). *MutSig* analisa um conjunto de mutações somáticas a fim de identificar genes que foram mutados com mais frequência do que o esperado, considerando a probabilidade de uma base sofrer mutação por acaso. Já o pacote *maftools* foi utilizado com o intuito de sumarizar os resultados observados pelo *MutSig*, utilizando a função *oncoplplot*.

Em adenocarcinoma gástrico, para o *cluster* 1 (ancestralidade africana) foram encontrados somente os genes TP53 e ARID1 como significativos, devido ao número reduzido de pacientes (Figura 10A). No *cluster* 2 (ancestralidade europeia) foram encontrados um total de 36 genes (Figura 10B). Já no *cluster* 3 (ancestralidade asiática) foram encontrados 6 genes (TP53, DCLK1, EDNRB, KLF3, ACVR2A e MXRA8) (Figura 10C).

O gene TP53 foi o mais frequentemente mutado nos três clusters, porém apresentou porcentagens mais altas em indivíduos do *cluster* 1 (Figura 10A) em relação ao *cluster* 2 (Figura 10B) e *cluster* 3 (Figura 10C) (62% contra 46% e 52%, respectivamente). DCLK1, apesar de ser o segundo gene mais frequentemente mutado no *cluster* 3 (15%) (Figura 10C), não aparece como sendo significativo em indivíduos do *cluster* 2 (Figura 10B). Este gene é descrito como um marcador de células-tronco tumorais, em casos de câncer de cólon e também de rim (Nakanishi et al. 2013; Chandrakesan et al. 2017; Mohammadia et al. 2018). Mutações em KLF3 foram pouco frequentes em indivíduos do *cluster* 2 (4%) (Figura 10B), enquanto no *cluster* 3 chegou-se a 11% (Figura 10C).

Outra diferença muito interessante foi observada no gene ARID1A, que apesar de frequentemente mutado nos clusters 1 e 2, com frequência de 31% e 25%, respectivamente (Figuras 10A e 10B), não foi observado em indivíduos no *cluster* 3 como sendo um gene significativamente mutado. Mutações neste gene podem estar correlacionadas com vias de reparo prejudicadas e com sensibilidade à inibidores de PARP (Shen et al. 2015).

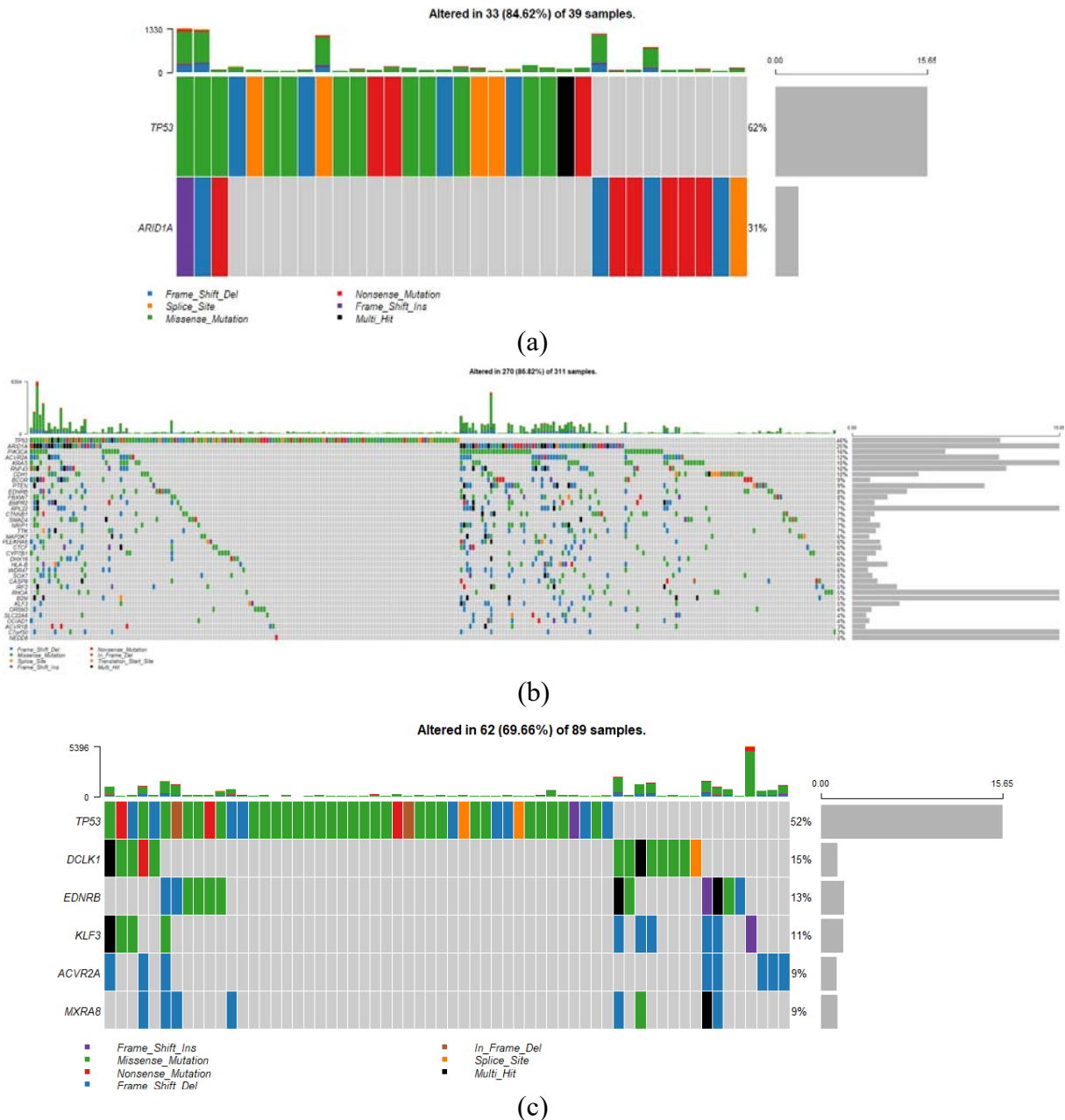


Figura 10 - Genes significativamente mutados em pacientes de adenocarcinoma gástrico utilizando *MutSig*. Foi utilizado o software *MutSig* para a identificação dos genes significativamente mutados, de acordo com a sua ancestralidade; **(A)** Cluster 1 (ancestralidade africana); **(B)** Cluster 2 (ancestralidade europeia); **(C)** Cluster 3 (ancestralidade asiática).

Em adenocarcinoma de pulmão, foram encontrados um conjunto de 25 genes significativamente mutados em indivíduos do *cluster 2* (ancestralidade europeia) (Figura 11A) e 4 no *cluster 3* (ancestralidade africana) (Figura 11B). Genes como KRAS, KEAP1 e SKT11 apresentaram frequências muito similares entre os dois grupos, enquanto mutações em TP53 mostraram-se muito mais frequentes em indivíduos do *cluster 3* (67% contra 49%) (Figura 11).

Em indivíduos do *cluster* 1 (ancestralidade asiática) não foram observados genes significativamente mutados.

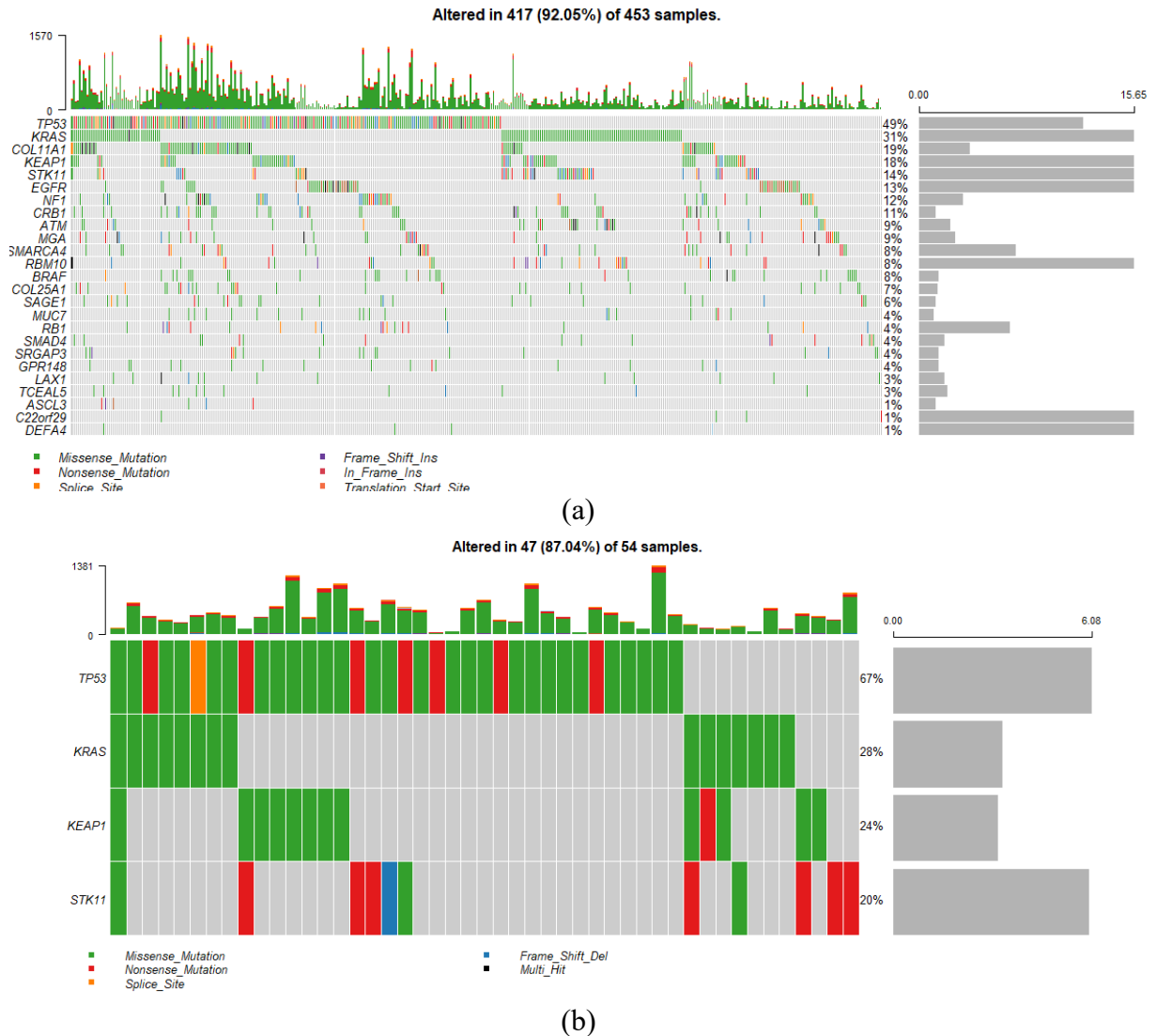
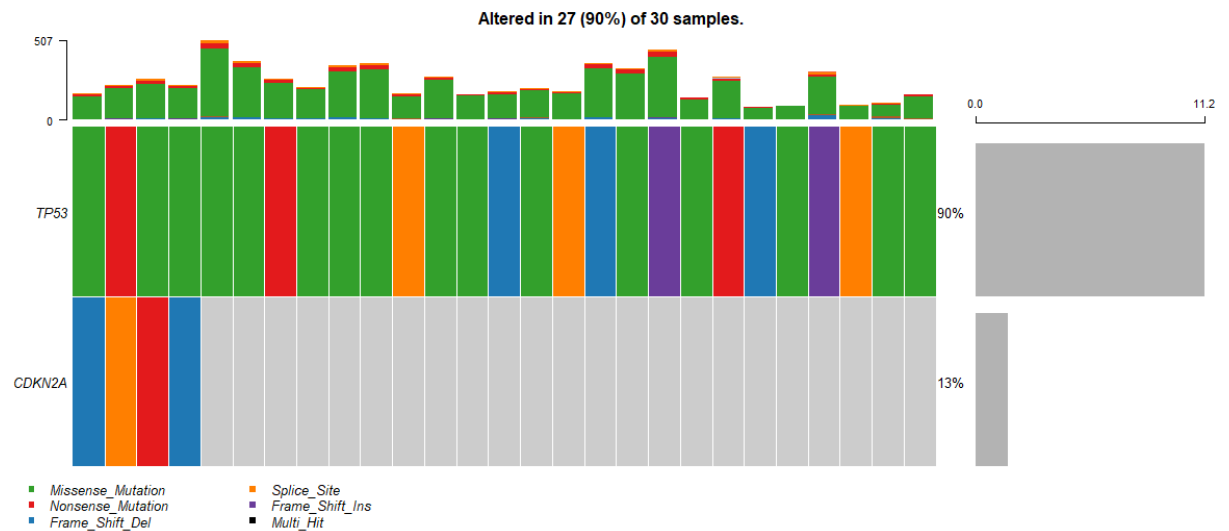


Figura 11 - Genes significativamente mutados em pacientes de adenocarcinoma de pulmão utilizando *MutSig*. Foi utilizado o software *MutSig* para a identificação dos genes significativamente mutados, de acordo com a sua ancestralidade; **(A)** *Cluster* 2 (ancestralidade europeia); **(B)** *Cluster* 3 (ancestralidade africana). Em indivíduos do *cluster* 1 (ancestralidade asiática) não foram encontrados genes significativamente mutados.

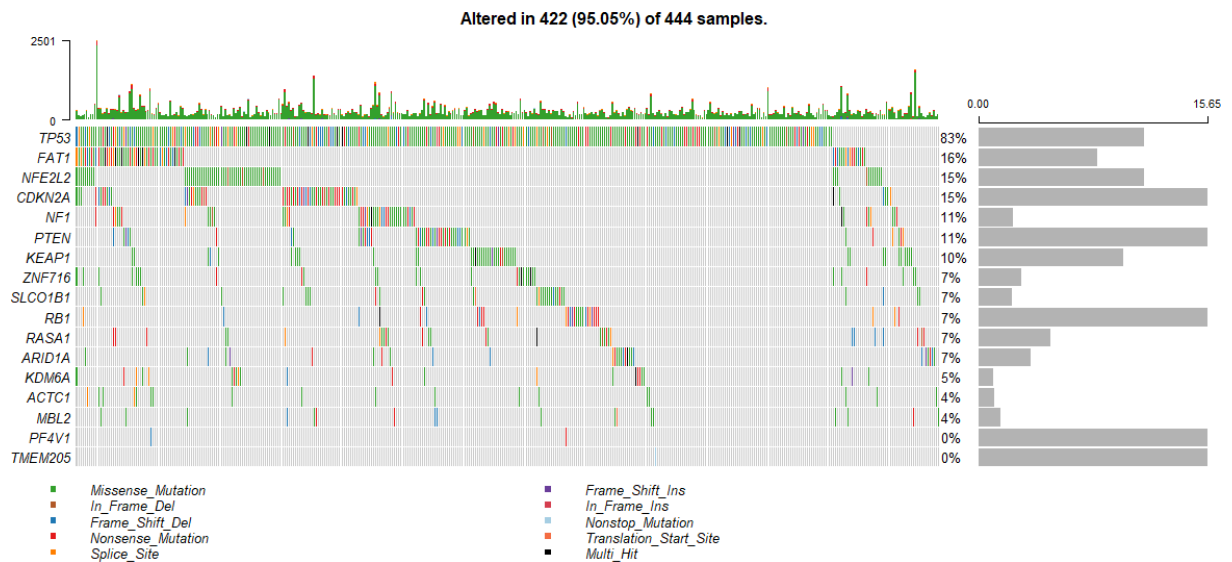
Em carcinoma de células escamosas de pulmão foram observados um conjunto de 2 genes significativamente mutados, TP53 e CDKN2A, em indivíduos do cluster 2 (ancestralidade africana) (Figura 12A) e 17 genes no cluster 3 (ancestralidade europeia) (Figura 12B). Tanto TP53, quando CDKN2A, apresentaram frequências muito semelhantes entre os dois grupos

(Figura 12). Em indivíduos do cluster 1 (ancestralidade asiática) não foram observados genes significativamente mutados.

Em melanoma foi possível encontrar genes significativamente mutados somente em indivíduos do cluster 1 (ancestralidade europeia), devido ao número reduzido de indivíduos nos outros dois grupos (Figura 13).



(a)



(b)

Figura 12 - Genes significativamente mutados em pacientes de carcinoma de células escamosas de pulmão utilizando *MutSig*. Foi utilizado o *software MutSig* para a identificação dos genes significativamente mutados, de acordo com a sua ancestralidade; **(A)** Cluster 2 (ancestralidade africana); **(B)** Cluster 3 (ancestralidade europeia). Em indivíduos do *cluster* 1 (ancestralidade asiática) não foram observados genes significativamente mutados.

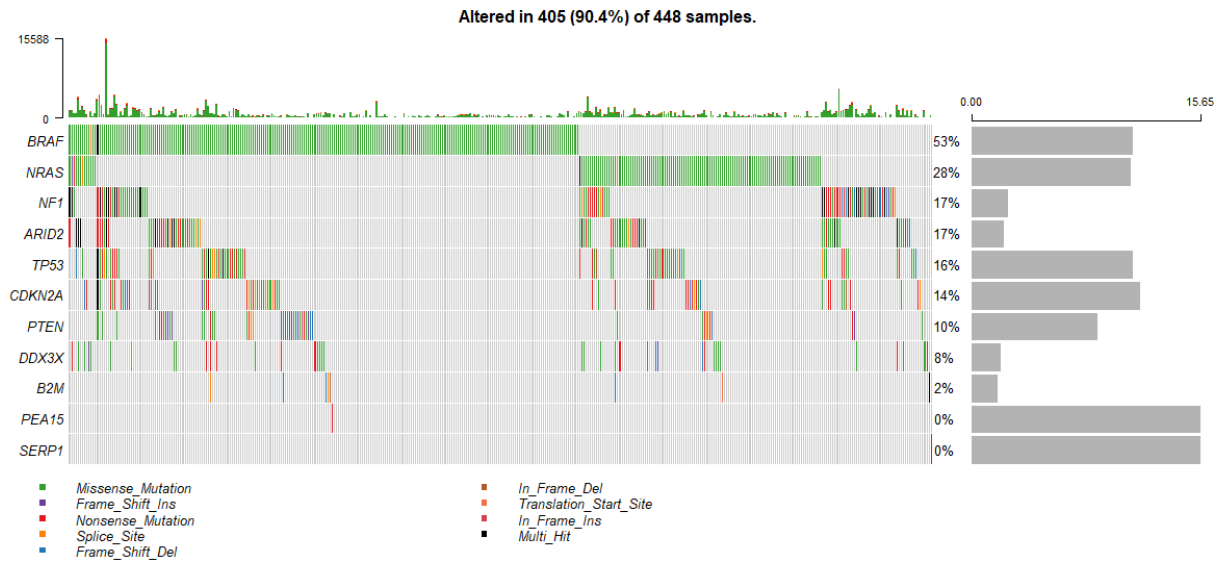


Figura 13 - Genes significativamente mutados em pacientes de melanoma no *cluster* 1 (ancestralidade europeia), utilizando *MutSig*. Foi utilizado o *software* MutSig para a identificação dos genes significativamente mutados em indivíduos do *cluster* 1 (ancestralidade europeia).

Coletivamente, esses resultados ilustram a diversidade de genes mutados associada a cada grupo de indivíduos.

4.4 ASSINATURA MUTACIONAL

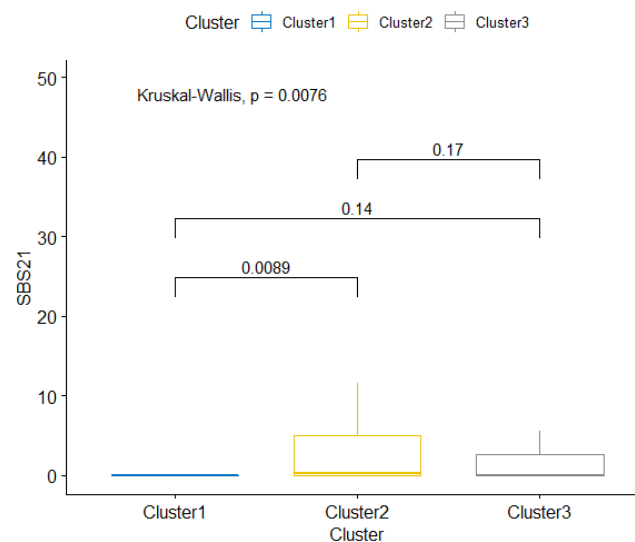
Com o intuito de avaliar assinaturas mutacionais, utilizamos o catálogo de assinaturas identificadas pelo *Catalogue of Somatic Mutations in Cancer* (COSMIC) (Tate et al. 2019; Wellcome Sanger Institute 2020). Para a extração das assinaturas nas amostras de interesse, foram utilizados três métodos: *signeR* (Rosales et al. 2017), *deconstructSigs* (Rosenthal et al. 2016) e o *MutationalPatterns* (Blokzijl et al. 2018). Estes métodos foram aplicados para todas as amostras de adenocarcinoma gástrico e de pulmão, melanoma e carcinoma de células escamosas de pulmão.

Devido à não normalidade e/ou à heteroscedasticidade das exposições identificadas por *MutationalPatterns* e *deconstructSigs*, foi utilizado o teste de *Kruskal-Wallis* (Kruskal e Wallis 1952) com o pós-teste de *bonferroni*, com o valor de *p* ajustado. Foi considerado como significativo valor de *p* abaixo de 0.05. Somente resultados com valor de *p* significativo para o teste de *Kruskal-Wallis* são aqui apresentados. Com o *signeR*, foi utilizada uma função nativa do pacote para exposição diferencial (*DiffExp*).

Avaliando os pacientes para adenocarcinoma gástrico, não foram encontradas assinaturas significativas de acordo com *signeR* (Figura 14A) e *deconstructSigs*. Com *MutationalPatterns* uma assinatura foi identificada: SBS21 (Figura 14B). Esta assinatura está associada a defeitos na via de reparo de mal pareamento do DNA (Tate et al. 2019; Wellcome Sanger Institute 2020) e foram observados valores mais altos em indivíduos do *cluster 2* (ancestralidade europeia).



(a)



(b)

Figura 14 - Assinaturas mutacionais em pacientes de adenocarcinoma gástrico. Com o intuito de comparar as principais assinaturas presentes entre cada um dos clusters (cluster 1 - ancestralidade africana, cluster 2 - ancestralidade europeia, cluster 3 - ancestralidade asiática), foram utilizados os métodos **(A)** *signeR*. Gráfico com cálculo da exposição diferencial a partir da função *DiffExp*. O eixo X representa cada uma das assinaturas obtidas a partir do COSMIC e o eixo Y o *p-value* entre os grupos. A linha vermelha indica o *threshold* (*p-value* = 0.05) para a identificação das assinaturas com exposição diferencial; **(B)** *MutationalPatterns*. Para *MutationalPatterns* foi considerado significativo valor de $p < 0.05$. Somente assinaturas com valor de *p* significativo para o teste de *Kruskal-Wallis* foram selecionadas.

Já em pacientes para melanoma, a aplicação do *signeR* e *MutationalPatterns* identificou, respectivamente, três (Figuras 15A e 15B) e cinco assinaturas significativas (Figura 15C). Com *signeR* foram identificadas as assinaturas SBS7a, SBS7b, SBS10b (Figura 15A e 15B). SBS7a e SBS7b estão relacionadas à exposição a raios ultravioleta (Tate et al. 2019; Wellcome Sanger Institute 2020) e apresentam valores altos em indivíduos com ancestralidade europeia (*cluster* 1) em relação aos outros dois grupos (Figuras 15A e 15B). Já a assinatura SBS10b é marcada por um grande número de mutações somáticas e está relacionada com mutações à DNA polimerase épsilon (Tate et al. 2019; Wellcome Sanger Institute 2020), também com valores mais altos em indivíduos do *cluster* 1 (Figuras 15A e 15B). Com *MutationalPatterns* foram identificadas as assinaturas: SBS1, SBS7a, SBS7b, SBS10b e SBS87 (Figura 15C). A diferença entre os grupos para as assinaturas SBS7a, SBS7b e SBS10b se mantiveram em relação ao observado com o *signeR* (Figura 15C). SBS1 está relacionada ao processo mutacional endógeno espontâneo ou enzimático de deaminação de 5-metilcitosina para timina, levando a erros de pareamento (Tate et al. 2019; Wellcome Sanger Institute 2020). Esta assinatura foi observada em maior quantidade em indivíduos do *cluster* 3 (Figura 15C). Já a assinatura SBS87 está associada à quimioterapia por tiopurinas e foi mais observada em indivíduos do *cluster* 1 (Figura 15C). O uso de *deconstructSigs* não identificou assinaturas significantes para os casos de melanoma.

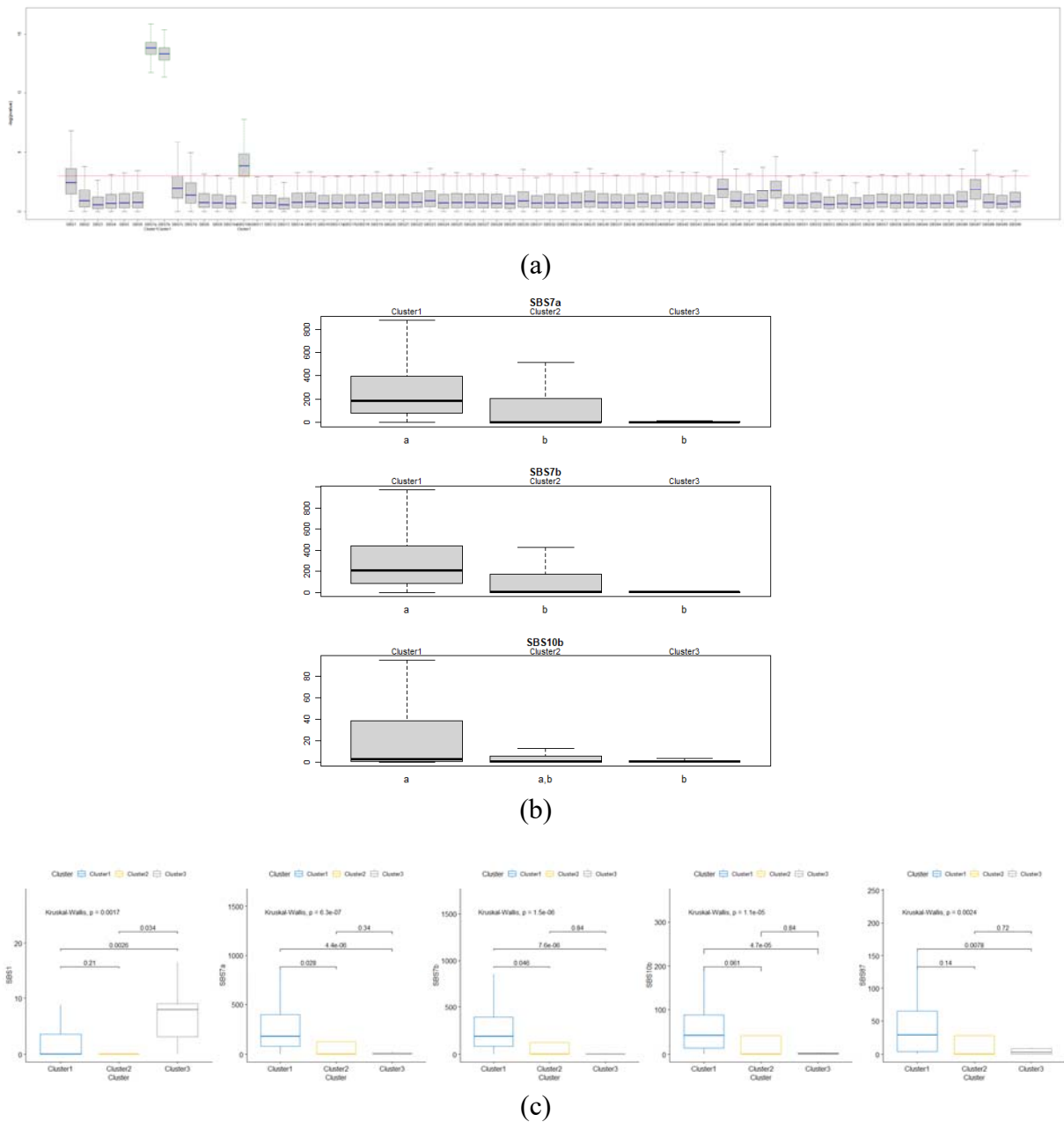


Figura 15 - Assinaturas mutacionais em pacientes de melanoma. Com o intuito de comparar as principais assinaturas presentes entre cada um dos clusters (*cluster 1* - ancestralidade europeia, *cluster 2* - ancestralidade africana, *cluster 3* - ancestralidade asiática), foram utilizados os métodos *signeR* e *MutationalPatterns* (A) Gráfico com cálculo da exposição diferencial, a partir dos dados de exposição do *signeR*, a partir da função *DiffExp*. O eixo X representa cada uma das assinaturas obtidas a partir do COSMIC e o eixo Y o *p-value* entre os grupos. A linha vermelha indica o *threshold* (*p-value* = 0.05) para a identificação das assinaturas com exposição diferencial; (B) Assinaturas classificadas como significativas para exposição diferencial; (C) *MutationalPatterns*. Para *MutationalPatterns* foi considerado significativo valor de $p < 0.05$. Somente assinaturas com valor de p significativo para o teste de *Kruskal-Wallis* foram selecionadas.

Em casos de adenocarcinoma de pulmão, destacaram-se as assinaturas SBS4 e SBS24 (Figura 16). Estas assinaturas foram identificadas por todos os métodos aplicados e com o mesmo padrão entre os grupos, com ambas apresentando uma maior exposição em indivíduos do *cluster* 3 (ancestralidade africana) (Figura 16). A assinatura SBS4 está associada ao uso de tabaco, enquanto SBS24 está relacionada à exposição por aflatoxina (Tate et al. 2019; Wellcome Sanger Institute 2020). Outras assinaturas também foram encontradas pelo *MutationalPatterns* e *deconstructSigs*, como: SBS7d (Figura 16B), SBS13 (Figura 16C) e SBS15 (Figura 16B), porém nenhuma delas foi identificada por mais de um método. Estas assinaturas estão relacionadas a dano por exposição a luz ultravioleta, pela atividade de APOBEC e por defeitos na via de reparo de mal-pareamento, respectivamente (Tate et al. 2019; Wellcome Sanger Institute 2020).

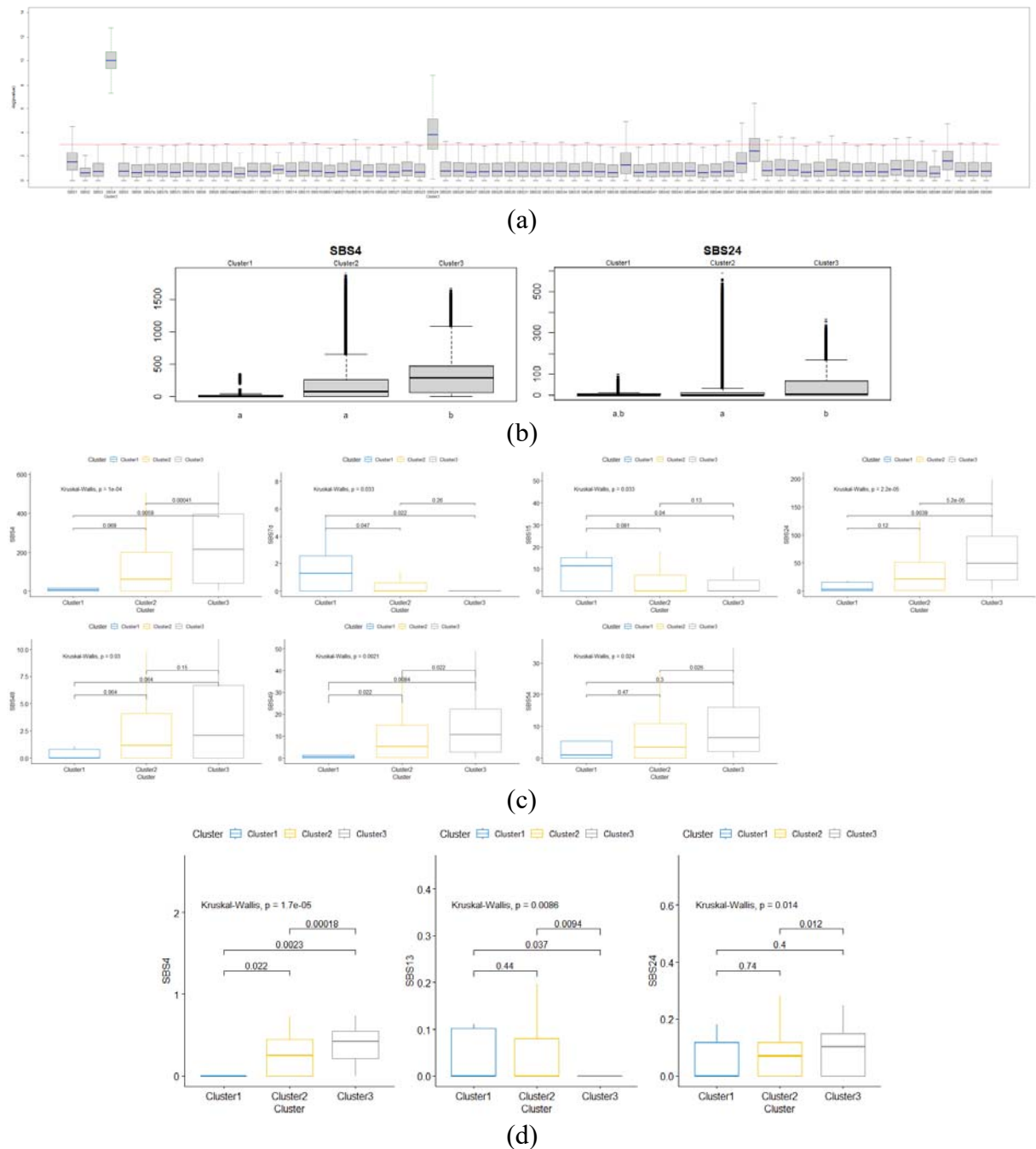


Figura 16 - Assinaturas mutacionais em pacientes de adenocarcinoma de pulmão. Com o intuito de comparar as principais assinaturas presentes entre cada um dos *clusters* (*cluster 1* - ancestralidade asiática, *cluster 2* - ancestralidade europeia, *cluster 3* - ancestralidade africana), foram utilizados os métodos *signeR*, *MutationalPatterns* e *deconstructSigs* **(A)** Gráfico com cálculo da exposição diferencial, a partir dos dados de exposição do *signeR*, a partir da função *DiffExp*. O eixo X representa cada uma das assinaturas obtidas a partir do COSMIC e o eixo Y o *p-value* entre os grupos. A linha vermelha indica o *threshold* (*p-value* = 0.05) para a identificação das assinaturas com exposição diferencial; **(B)** Assinaturas classificadas como significativas para exposição diferencial; **(C)** *MutationalPatterns*; **(D)** *deconstructSigs*. Para *MutationalPatterns* e *deconstructSigs* foi considerado significativo valor de $p < 0.05$. Somente assinaturas com valor de p significativo para o teste de *Kruskal-Wallis* foram selecionadas.

Já para os pacientes de carcinoma de células escamosas de pulmão, foram encontradas assinaturas somente através do método *MutationalPatterns* (Figura 17B). Entre elas foram identificadas: SBS13, SBS16, SBS24 e SBS53 (Figura 17B). Como dito anteriormente, SBS13 está relacionada à atividade de APOBEC e SBS24 à exposição a aflatoxina (Tate et al. 2019; Wellcome Sanger Institute 2020). Já SBS16 possui etiologia desconhecida e SBS53 está associada a possíveis artefatos de sequenciamento (Tate et al. 2019; Wellcome Sanger Institute 2020).

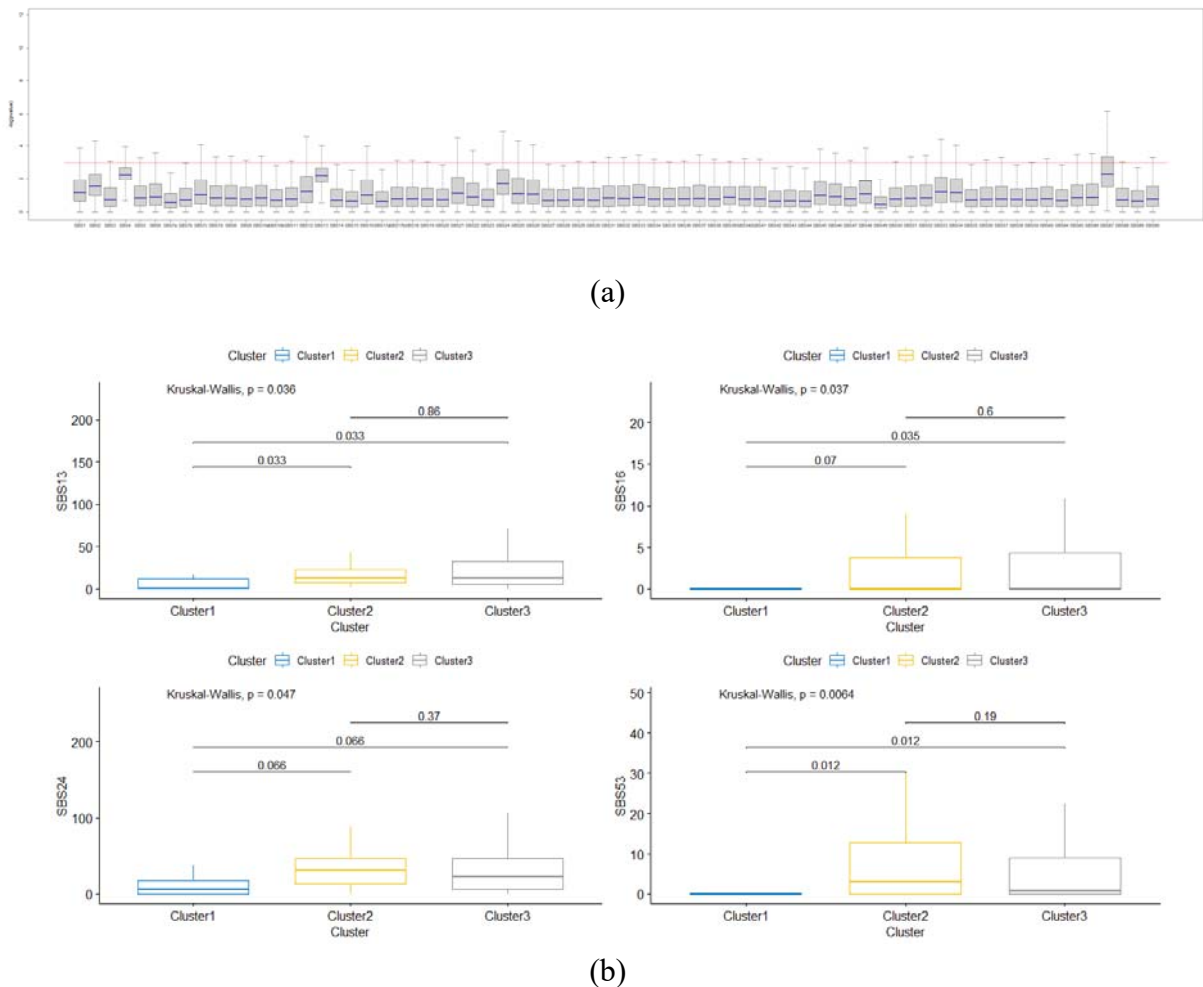


Figura 17 - Assinaturas mutacionais em pacientes de carcinoma de células escamosas de pulmão. Com o intuito de comparar as principais assinaturas presentes entre cada um dos clusters (cluster 1 - ancestralidade asiática, cluster 2 - ancestralidade africana, cluster 3 - ancestralidade europeia), foram utilizados os métodos (A) *signeR*. Gráfico com cálculo da exposição diferencial a partir da função *DiffExp*. O eixo X representa cada uma das assinaturas obtidas a partir do COSMIC e o eixo Y o p -value entre os grupos. A linha vermelha indica o *threshold* (p -value = 0.05) para a identificação das assinaturas com exposição diferencial.

4.5 EXPRESSÃO GÊNICA DIFERENCIAL

Buscando avaliar as diferenças na expressão gênica entre os grupos, utilizamos dados disponíveis no TCGA e previamente pré-processados pelo *software HTseq* (Anders et al. 2015). Em seguida, aplicamos o *software Deseq2* (Love et al. 2014) com o intuito de obter a expressão gênica diferencial. Este processo foi realizado analisando os grupos em pares. Os genes foram ranqueados de acordo com o seu *p-value* ajustado e foram selecionados até 1000 genes, desde que todos possuísem um *p-value* ajustado abaixo de 0.05. Estes genes por sua vez, foram utilizados na análise de enriquecimento denominada GSEA (*Gene Set Enrichment Analysis*) (Mootha et al. 2003; Subramanian et al. 2005), com o objetivo de identificar quais vias biológicas estariam mais enriquecidas entre os grupos. Para esta análise foi aplicada a nomenclatura adotada pelo banco de dados *Gene Ontology* (GO) para processos biológicos.

A fim de compilar a informação obtida de algumas comparações, foi utilizado o *software Cytoscape* (Shannon et al. 2003; Reimand et al. 2019). Juntamente, foram utilizadas as seguintes extensões: *AutoAnnotate* (Kucera et al. 2016), *Enrichment Map* (Merico et al. 2010), *ClusterMaker* (Morris et al. 2011) e *WordCloud* (Oesper et al. 2011). *Enrichment Map* foi aplicado com o objetivo de representar visualmente as principais redes e vias encontradas através do uso de GSEA, utilizando um *q-value* menor do que 0.05. Em seguida, foi utilizado *AutoAnnotate* em conjunto com *ClusterMaker* e *WordCloud*, a fim de agrupar e adicionar *tags* a cada um dos *clusters* encontrados. *Clusters* formados por uma única via foram removidos.

Para os quatro tipos de tumores foram realizadas comparações na seguinte ordem: *cluster 1 - cluster 2*, *cluster 1 - cluster 3* e *cluster 2 - cluster 3*. Nas amostras de adenocarcinoma gástrico (Tabela 3), observou-se que em indivíduos de ancestralidade africana (*cluster 1*) há uma menor expressão de genes relacionados à proteólise, quando comparados com indivíduos de ancestralidade europeia (Tabela 3A). Além disso, ao comparar indivíduos de ancestralidade africana (*cluster 1*) e asiática (*cluster 3*), observou-se que os primeiros são caracterizados por uma maior expressão de genes relacionados ao desenvolvimento da epiderme e epitélio e diferenciação de células epiteliais (Tabela 3B), e uma menor expressão de genes envolvidos com modificação pós-traducional de proteínas e regulação do progresso de sistemas (Tabela 3C). Na comparação entre indivíduos com ancestralidade europeia (*cluster 2*) e asiática (*cluster 3*), foram encontradas diversas vias e por isso foram compiladas utilizando o *software Cytoscape* (Figura 18). Todas as vias encontradas se mostraram positivamente enriquecidas em indivíduos com ancestralidade europeia (*cluster 2*). Estas vias foram classificadas em cinco

clusters: *development and differentiation, migration leukocyte, response organism, biological adhesion cell e activation immune process* (Figura 18).

Tabela 3 - Gene Set Enrichment Analysis de pacientes de adenocarcinoma gástrico do TCGA.

Gene Ontology (GO) Set	Size	Enrichment Score (ES)	Normalized Enrichment Score (NES)	FDR q-val
Proteolysis	34	-0.66646844	-2.3588839	0

(a)

Gene Ontology (GO) Set	Size	Enrichment Score (ES)	Normalized Enrichment Score (NES)	FDR q-val
Epidermis Development	17	0.8892135	2.8998108	0
Epithelial Cell Differentiation	33	0.7081038	2.7076976	0
Epithelium Development	57	0.50729567	2.1831803	0.0028427674

(b)

Gene Ontology (GO) Set	Size	Enrichment Score (ES)	Normalized Enrichment Score (NES)	FDR q-val
Post Translational Protein Modification	29	-0.5691001	-2.5073476	0.0044874074
Regulation of System Progress	38	-0.4975304	-2.333559	0.007496441

(c)

A análise por GSEA foi realizada comparando os três grupos e utilizando a nomenclatura do banco de dados *Gene Ontology* para processos biológicos em amostras de adenocarcinoma gástrico do TCGA; (a) Vias enriquecidas negativamente para o cluster 1 (ancestralidade africana) em comparação com o cluster 2 (ancestralidade europeia); (b) Vias enriquecidas positivamente para o cluster 1 (ancestralidade africana) em comparação com o cluster 3 (ancestralidade asiática); (c) Vias enriquecidas negativamente para o cluster 1 (ancestralidade africana) em comparação com o cluster 3 (ancestralidade asiática).

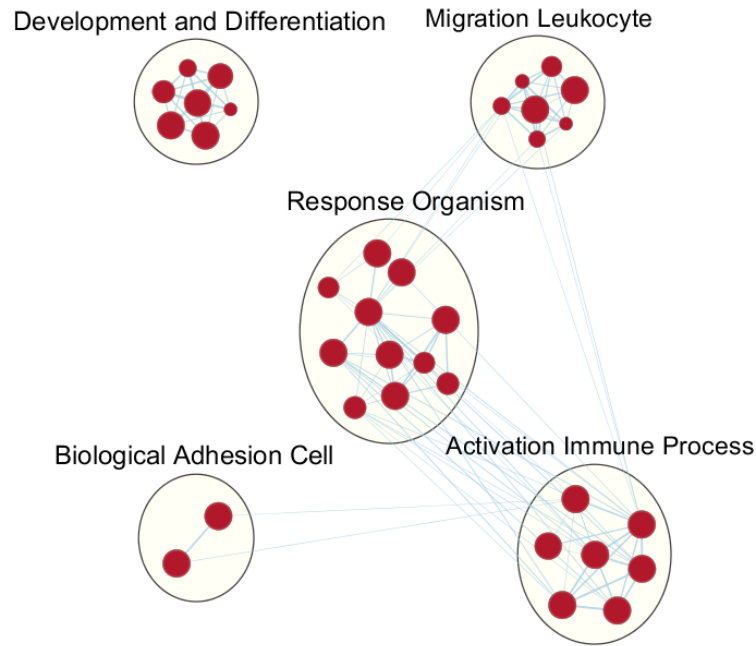


Figura 18 - Representação gráfica da comparação entre *cluster 2* (ancestralidade europeia) e *cluster 3* (ancestralidade asiática) em pacientes de adenocarcinoma gástrico do TCGA, através do *software Cytoscape* para as vias identificadas por *Gene Set Enrichment Analysis*. Análise por GSEA foi realizada comparando o *cluster 2* (ancestralidade europeia) e *cluster 3* (ancestralidade asiática) e aplicando a nomenclatura do banco de dados *Gene Ontology* para processos biológicos em amostras de adenocarcinoma gástrico do TCGA. Somente vias com $q\text{-value} < 0.05$ foram consideradas e então complicadas e agrupadas utilizando *Cytoscape* e as extensões *AutoAnnotate*, *Enrichment Map*, *WordCloud* e *ClusterMaker*.

Já em casos de adenocarcinoma de pulmão, a análise entre indivíduos de ancestralidade de ancestralidade asiática (*cluster 1*) e europeia (*cluster 2*) não encontrou vias com valor de FDR significativo. Já entre indivíduos de ancestralidade asiática (*cluster 1*) e africana (*cluster 3*), foram observadas diversas vias, e estas foram compiladas e agrupadas utilizando *Cytoscape* (Figura 19). É importante destacar que todas as vias encontradas apresentaram enriquecimento para ancestralidade asiática. Entre os principais *clusters* identificados, podemos destacar: *Regulation of cellular protein localization*, *splicing RNA*, *mitochondrial translation and biosynthetic process* e *chromatin organization* (Figura 19). Entre indivíduos com ancestralidade europeia (*cluster 2*) e africana (*cluster 3*), foi encontrada somente a via de *epithelium development*, enriquecida positivamente para a ancestralidade europeia (Tabela 4).

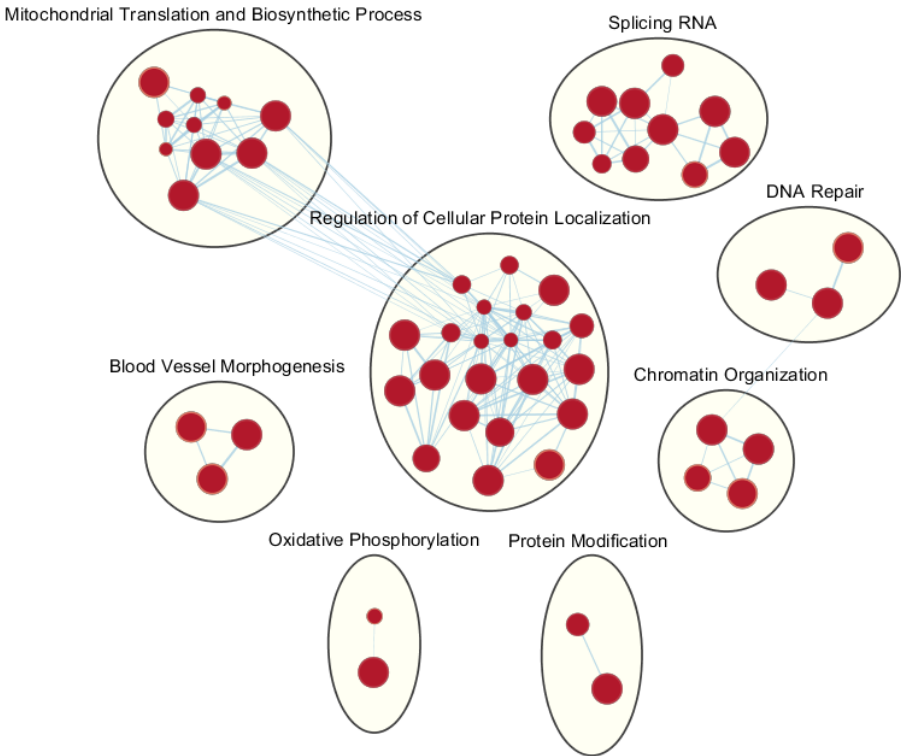


Figura 19 - Representação gráfica da comparação entre *cluster* 1 (ancestralidade asiática) e *cluster* 3 (ancestralidade africana) em pacientes de adenocarcinoma de pulmão do TCGA, através do software *Cytoscape* para as vias identificadas por *Gene Set Enrichment Analysis*. Análise por GSEA foi realizada comparando o *cluster* 1 (ancestralidade asiática) e *cluster* 3 (ancestralidade africana) e aplicando a nomenclatura do banco de dados *Gene Ontology* para processos biológicos em amostras de adenocarcinoma de pulmão do TCGA. Somente vias com *q-value* < 0.05 foram consideradas e então complicadas e agrupadas utilizando *Cytoscape* e as extensões *AutoAnnotate*, *Enrichment Map*, *WordCloud* e *ClusterMaker*.

Tabela 4 - *Gene Set Enrichment Analysis* de pacientes de adenocarcinoma de pulmão do TCGA, comparando *cluster* 2 (ancestralidade europeia) e *cluster* 3 (ancestralidade africana).

Gene Ontology (GO) Set	Size	Enrichment Score (ES)	Normalized Enrichment Score (NES)	FDR q-val
Epithelium Development	39	0.5845797	2.4765599	0

A análise por GSEA foi realizada comparando *cluster* 2 (ancestralidade europeia) e *cluster* 3 (ancestralidade africana) e aplicando a nomenclatura do banco de dados *Gene Ontology* para processos biológicos em amostras de adenocarcinoma de pulmão do TCGA.

Em casos de carcinoma de células escamosas de pulmão, indivíduos com ancestralidade asiática (*cluster 1*) apresentaram um *score* negativo para a via *regulation of intracellular signal transduction* (Tabela 5A), em relação a indivíduos com ancestralidade africana (*cluster 2*). Entre indivíduos com ancestralidade africana (*cluster 2*) e europeia (*cluster 3*) múltiplas vias foram identificadas com um valor de FDR significativo (Tabelas 4B e 4C). Entre elas, destacam-se algumas vias mitocondriais com enriquecimento em indivíduos com ancestralidade africana, como: *mitochondrion organization* e *mitochondrial transport* (Tabela 5B). Com *score* negativo, foram encontradas as vias: *reponse to extracellular stimulus*, *maintenance of location* e *inflammatory response* (Tabela 5C). Já em melanoma não foram observadas diferenças significativas.

Tabela 5 - Gene Set Enrichment Analysis de pacientes de carcinoma de células escamosas de pulmão do TCGA.

Gene Ontology (GO) Set	Size	Enrichment Score (ES)	Normalized Enrichment Score (NES)	FDR q-val
Regulation of Intracellular Signal Transduction	18	-0.441355	-2.3269029	0.009455884

(a)

Gene Ontology (GO) Set	Size	Enrichment Score (ES)	Normalized Enrichment Score (NES)	FDR q-val
Mitochondrion Organization	24	0.5549868	2.683937	0
Mitochondrial Transport	15	0.62673426	2.5052197	0
Protein Targeting	17	0.5510444	2.4114559	0.00091239315
Regulation of Chromosome	18	0.46868256	2.183885	0.0027335035

(b)

Gene Ontology (GO) Set	Size	Enrichment Score (ES)	Normalized Enrichment Score (NES)	FDR q-val
Response to Extracellular Stimulus	26	-0.6732701	-2.1197844	0.002009636
Maintenance of Location	20	-0.6721731	-2.038202	0.004041341
Inflammatory Response	19	-0.66949433	-2.0119915	0.006097228

(c)

A análise por GSEA foi realizada comparando os três grupos e aplicando a nomenclatura do banco de dados *Gene Ontology* para processos biológicos em amostras de carcinoma de células escamosas de pulmão do TCGA; **(A)** Vias enriquecidas negativamente para o cluster 1 (ancestralidade asiática) em comparação com o cluster 2 (ancestralidade africana); **(B)** Vias enriquecidas positivamente para o cluster 2 (ancestralidade africana) em comparação com o cluster 3 (ancestralidade europeia); **(C)** Vias enriquecidas negativamente para o cluster 2 (ancestralidade africana) em comparação com o cluster 3 (ancestralidade europeia).

4.6 SOBREVIDA E MODELOS DE REGRESSÃO

O estudo e avaliação do infiltrado inflamatório, da expressão gênica e de assinaturas mutacionais revelou diferenças significativas entre ancestralidades em tipos distintos de tumor. Com isto em mente, o próximo passo consistiu em avaliar a sobrevida destes pacientes e o impacto destes fatores. Nesta etapa, foi utilizado *Kaplan-Meier* afim de medir e comparar a curva de sobrevida de cada *cluster* em cada tumor (Figura 20). Para esta análise cada tipo de tumor foi avaliado de duas formas: 1) utilizando todas as amostras e 2) removendo amostras de estágio III e IV (Figura 20).

Em amostras de adenocarcinoma gástrico (Figuras 20A e 20B), não foram observadas diferenças significativas entre os grupos, porém, inicialmente, indivíduos do *cluster* 1 (ancestralidade africana) apresentaram uma tendência a um pior prognóstico em relação aos outros dois grupos (Figura 20A). Já indivíduos do *cluster* 2 (ancestralidade asiática) apresentaram uma tendência a um melhor prognóstico a partir de cerca de 500 dias, independentemente do estadiamento (Figuras 20A e 20B).

Em adenocarcinoma de pulmão, não foram observadas diferenças significativas entre os grupos, independentemente do estadiamento (Figuras 20C e 20D). Em carcinoma de células escamosas de pulmão, foi possível observar uma tendência a um pior prognóstico em indivíduos do *cluster* 2 (ancestralidade africana), independentemente do estadiamento, principalmente em relação ao *cluster* 3 (ancestralidade europeia) (Figuras 20E e 20F). Em melanoma, nota-se uma tendência a um pior prognóstico em indivíduos do *cluster* 3 (ancestralidade asiática) em relação aos outros dois grupos (Figura 20G). Devido ao baixo número de amostras em alguns dos grupos neste tipo de tumor, não foi possível fazer a análise sem as amostras de estágio III e IV.

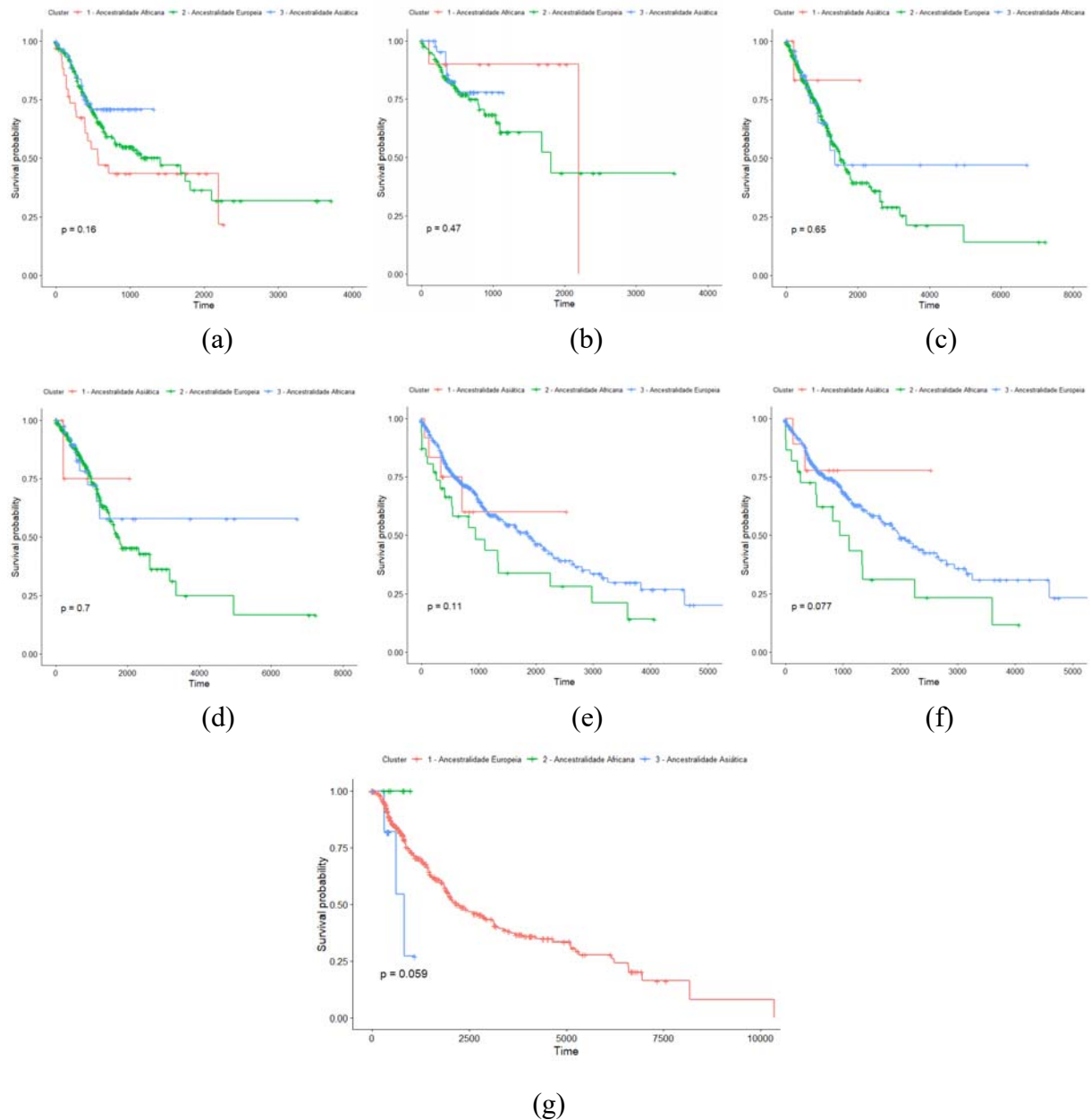


Figura 20 - Kaplan-Meier para pacientes de adenocarcinoma gástrico e de pulmão, carcinoma de células escamosas de pulmão e melanoma de acordo com estadiamento. Com o objetivo de avaliar a sobrevida entre cada um dos clusters, foi utilizado o método *Kaplan-Meier*. Esta análise foi realizada com amostras de **(A)** adenocarcinoma gástrico considerando estágios I, II, III e IV; **(B)** adenocarcinoma gástrico somente para estágios I e II; **(C)** adenocarcinoma de pulmão considerando estágios I, II, III e IV; **(D)** adenocarcinoma de pulmão somente para estágios I e II; **(E)** carcinoma de células escamosas de pulmão considerando estágios I, II, III e IV; **(F)** carcinoma de células escamosas de pulmão somente para estágios I e II; **(G)** melanoma considerando estágios I, II, III e IV.

O passo seguinte consistiu em avaliar o impacto do infiltrado inflamatório, da expressão gênica e das assinaturas mutacionais na sobrevida dos pacientes de adenocarcinoma gástrico e de pulmão, melanoma e de carcinoma de células escamosas de pulmão. Com este objetivo, dois modelos de predição foram aplicados: *least absolute shrinkage and selection operator* (Lasso) (Tibshirani et al. 1996) e *Elastic Net* (Zou e Hastie 2005).

Para selecionar as variáveis a serem aplicadas aos modelos, diversos filtros foram estabelecidos previamente. Foram consideradas significativas variáveis que passaram tanto pelo modelo de Lasso quanto para *Elastic Net*. Informações para idade e estadiamento do tumor também foram adicionadas. Os genes encontrados foram avaliados utilizando os seguintes bancos de dados: *Cancer Gene Census* (Sondka et al. 2018), *Dependency Map* (DepMap) (Broad Institute 2021), *Protein Cancer Atlas* (Pontén 2008; Uhlen et al. 2017) e *Catalogue of Cancer Genes* (Cheng et al. 2014).

A partir deste processo, foi calculado o valor de *hazard-ratio* para cada uma das variáveis significativas para os casos de adenocarcinoma gástrico (Figura 21) e de pulmão (Figura 22), carcinoma de células escamosas de pulmão (Figura 23) e melanoma (Figura 24). Os *clusters* foram separados e avaliados de forma independente através dos modelos.

Nos casos de adenocarcinoma gástrico, múltiplas variáveis foram observadas, tanto para o *cluster 1* (ancestralidade africana) quanto para o *cluster 2* (ancestralidade europeia). Em ambos os casos "estadiamento do tumor" (*stage_event_pathologic_stage*) foi considerado significativo, sendo um importante controle para a análise (Figura 21). Entre estas variáveis, somente dados para expressão gênica foram encontrados.

Em pacientes do *cluster 1*, alguns genes se destacam: MASTL (*microtubule-associated serine/threonine kinase-like*) e TP73 (*tumor protein P73*) (Figura 21A). MASTL codifica para uma proteína essencial para o funcionamento do ciclo celular (Hermida et al. 2020) e, segundo *DepMap*, é avaliado como um gene essencial para linhagens celulares tumorais (Broad Institute 2021). Nestes pacientes, foi identificado um *hazard-ratio* favorável (Figura 21A), porém, de acordo com *Protein Cancer Atlas* a alta expressão deste gene estaria relacionada a um pior prognóstico em casos de câncer renal, fígado e pancreático (Pontén 2008; Uhlen et al. 2017). Além disso, este gene está incluso no *Catalogue of Cancer Genes*, reforçando a sua relevância (Cheng et al. 2014). TP73 é um gene membro da família que compreende TP63 e TP53, e de maneira similar à TP53, é ativado em resposta a danos ao DNA e regula a expressão de genes envolvidos com a parada do ciclo celular e apoptose (Yao Z et al. 2019). Apesar disso, a função deste gene na tumorigenese ainda é pouco conhecida (Yao Z et al. 2019). De acordo com o *DepMap*, este gene não é visto como essencial para linhagens celulares tumorais (Broad

Institute 2021). Assim, como foi observado com MASTL, este gene apresentou um *hazard-ratio* favorável (Figura 21A), corroborando com o que foi identificado pelo *Protein Cancer Atlas* em carcinoma cervical, de endométrio e de cabeça e pescoço (Pontén 2008; Uhlen et al. 2017). TP73 também está presente no *Catalogue of Cancer Genes* (Cheng et al. 2014).

Em pacientes do *cluster 2*, diversos genes foram considerados significativos (Figura 21B). Entre eles, alguns genes se destacam segundo os bancos de referência: AHRR (*Aryl-Hydrocarbon Receptor Repressor*), FAAH (*Fatty Acid Amide Hydrolase*) e PWP2 (*PWP2 Small Subunit Processome Component*). AHRR está envolvido na regulação do crescimento celular e no processo de diferenciação (The UniProt Consortium 2019). Nestes pacientes foi notado um *hazard-ratio* favorável (Figura 21B) e está presente no *Catalogue of Cancer Genes* (Cheng et al. 2014). De acordo com *DepMap*, este gene não é essencial em células tumorais (Broad Institute 2021) e segundo o *Protein Cancer Atlas* a alta expressão deste gene não foi observada como um fator significativo para prognóstico (Pontén 2008; Uhlen et al. 2017).

FAAH é responsável por catalisar a hidrólise de lípidios endógenos (The UniProt Consortium 2019). Nestes pacientes apresenta um *hazard-ratio* favorável (Figura 21B), em concordância com o percebido pelo *Protein Cancer Atlas* em casos de carcinoma gástrico, rim e de endométrio (ATLAS 2020). De acordo com o *DepMap*, este não é considerado um gene essencial em células tumorais (Broad Institute 2021) e FAAH está presente no *Catalogue of Cancer Genes* (Cheng et al. 2014). Já PWP2 está associada à subunidade UTP-B, que se liga ao RNA pré-ribossomal para formar o processoma SSU (*small subunit*), importante para biogênese ribossomal (Boissier et al. 2017). Segundo *DepMap*, este gene é essencial em células tumorais (DEPMAP 2020) e nestes pacientes possui um *hazard-ratio* não favorável (Figura 21B). Este gene não está incluso no *Catalogue of Cancer Genes* (Cheng et al. 2014) e não foi identificado como um fator para prognóstico em tumores, segundo *Protein Cancer Atlas* (Pontén 2008; Uhlen et al. 2017).

Em relação ao *cluster 3* (ancestralidade asiática), não foram encontradas variáveis significativas, tanto para expressão gênica, como para infiltrado inflamatório e assinatura mutacional.

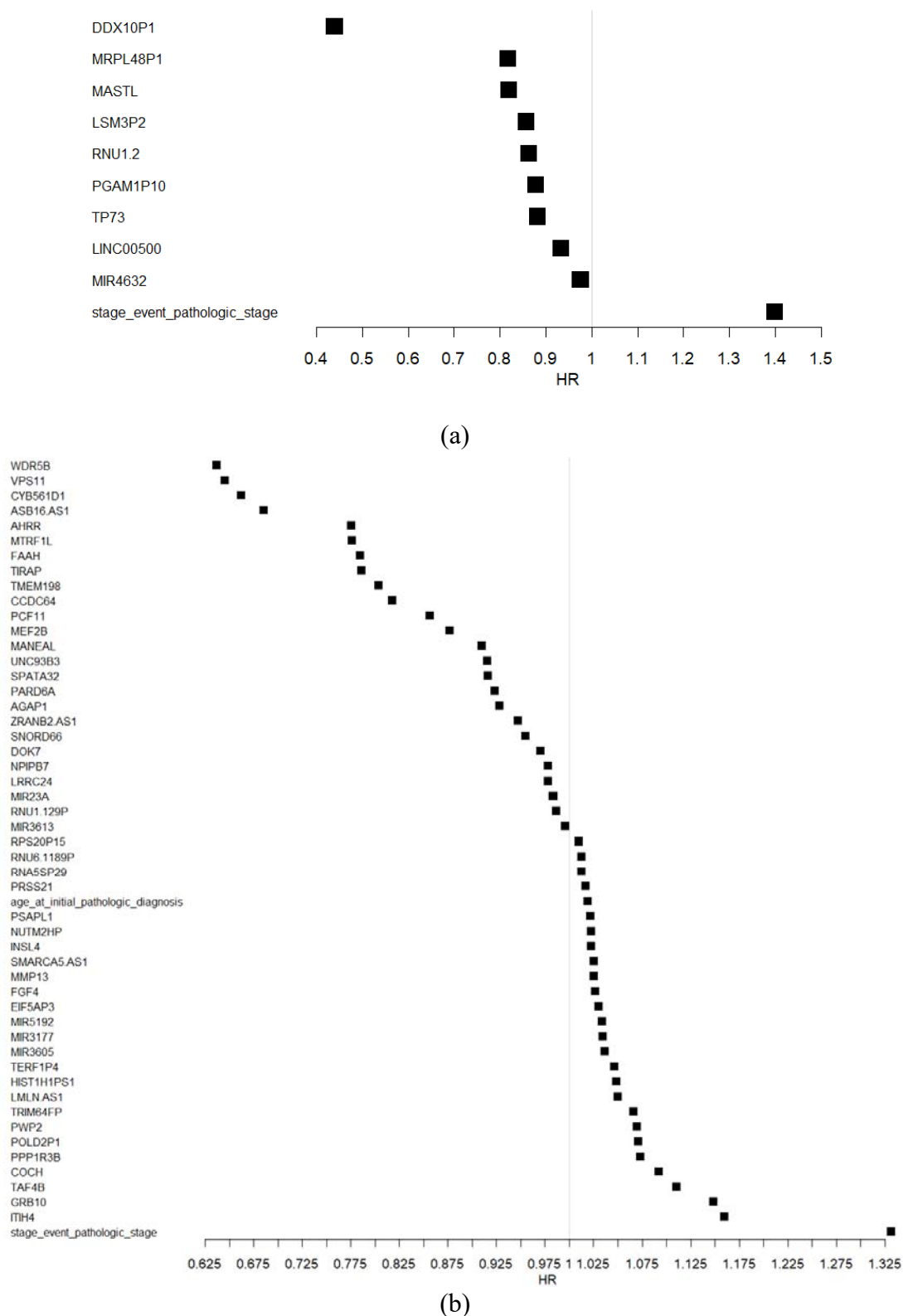


Figura 21 - Hazard-Ratio (HR) para amostras de adenocarcinoma gástrico. *Hazard-ratio* (HR) calculado para amostras de adenocarcinoma gástrico, através de modelos de predição: Lasso e *Elastic Net*. Foram utilizados dados para expressão gênica, assinatura mutacional e infiltrado inflamatório, além da idade (*age_at_initial_pathologic_diagnosis*) e estadiamento do tumor (*stage_event_pathologic_stage*); **(A)** Análise para amostras do *cluster 1* (ancestralidade africana); **(B)** Análise para amostras do *cluster 2* (ancestralidade europeia).

Em adenocarcinoma de pulmão foram identificados múltiplos genes significativos (Figura 22), tanto para amostras do *cluster* 2 (ancestralidade europeia) (Figura 22B) quanto para o *cluster* 3 (ancestralidade africana) (Figura 22A).

Nas amostras do *cluster* 3, sete variáveis foram selecionadas como significativas, incluindo estadiamento do tumor (*stage_event_pathologic_stage*). Entre as outras variáveis, somente dados para expressão gênica foram encontrados. Nesta análise, dois genes merecem destaque. O primeiro deles é COL22A1 (Collagen Type XXII Alpha 1 Chain), apresentando um *hazard-ratio* não favorável (Figura 22A). Este gene codifica para um específico tipo de colágeno (colágeno XXII), que contribui para a estabilização de junções miotendinosas e fortalecimento de ligações do músculo esquelético durante a atividade contrátil (Ton et al. 2018). Assim como observado nestas amostras, segundo *Protein Cancer Atlas*, a alta expressão deste gene está relacionada a um prognóstico não favorável em tumores de pulmão e em glioma. De acordo com *DepMap*, este gene não seria essencial para células tumorais (Broad Institute 2021) e está inserido no *Catalogue of Cancer Genes* (Cheng et al. 2014). O segundo gene de destaque foi ACBD3 (*Acyl-CoA Binding Domain Containing 3*), apresentando um valor de *hazard-ratio* muito alto em relação às outras variáveis (Figura 22A). Este gene possui poucas informações nos bancos de referência analisados, em relação à sua importância e impacto no prognóstico. ACBD3 está associada a diversas funções, como: esteroidogênese, embriogênese, neurogênese e tráfico de membrana (Yue et al. 2019). A sua alta expressão também foi reportada, tendo um impacto importante no prognóstico de pacientes com câncer de mama (Huang et al. 2019).

Já nas amostras do *cluster* 2, foram identificadas nove variáveis, entre elas estadiamento do tumor (*stage_event_pathologic_stage*) (Figura 22B). Assim como os grupos anteriores, dados para infiltrado inflamatório e assinatura mutacional não foram significativos. Entre os genes identificados, PLEK2 (*pleckstrin 2*) foi o que mais se destacou de acordo com os bancos de referência, apresentando um *hazard-ratio* não favorável (Figura 22B). A proteína expressa por este gene é capaz de auxiliar na redistribuição da actina dentro da célula, levando a um ampliamiento da célula (Shen et al. 2019). Também já foi constatado um papel importante de PLEK2 no processo de metástase de células de câncer de mama do tipo luminal A (Shen et al. 2019). Segundo *DepMap*, este gene é essencial em células tumorais (Broad Institute 2021) e está inserido no *Catalogue of Cancer Genes* (Cheng et al. 2014). Além disso, os dados do *Protein Cancer Atlas* corroboram com o resultado observado, visto que a sua alta expressão impactou de forma negativa e significativa o prognóstico de pacientes de câncer de pulmão e pâncreas (Pontén 2008; Uhlen et al. 2017).

Em relação ao *cluster* 1 (ancestralidade asiática), não foram encontradas variáveis significativas, tanto para expressão gênica, como para infiltrado inflamatório e assinatura mutacional.

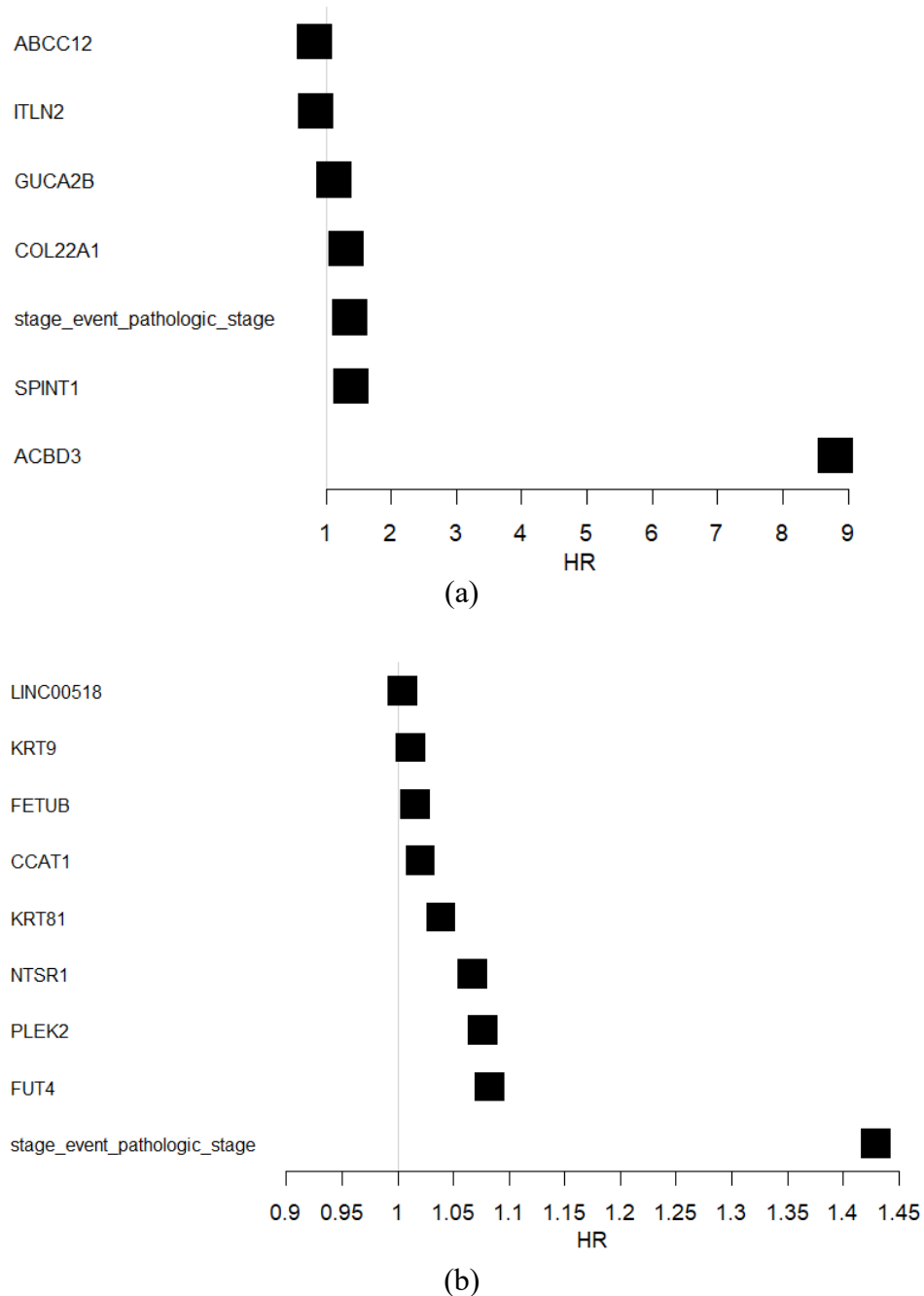


Figura 22 - Hazard-Ratio (HR) para amostras de adenocarcinoma de pulmão. *Hazard-ratio* (HR) calculado para amostras de adenocarcinoma gástrico, através de modelos de modelos de predição: Lasso e *Elastic Net*. Foram utilizados dados para expressão gênica, assinatura mutacional e infiltrado inflamatório, além da idade (*age_at_initial_pathologic_diagnosis*) e estadiamento do tumor (*stage_event_pathologic_stage*); **(A)** Análise para amostras do *cluster* 3 (ancestralidade africana); **(B)** Análise para amostras do *cluster* 2 (ancestralidade europeia).

Avaliando casos de carcinoma de células escamosas de pulmão, foi possível identificar múltiplos genes significativos, tanto para amostras do *cluster* 2 (ancestralidade africana) (Figura 23A) quanto para amostras do *cluster* 3 (ancestralidade europeia) (Figura 23B). Variáveis para assinatura mutacional e infiltrado inflamatório não foram significativas. Estadiamento do tumor (*stage_event_pathologic_stage*) foi considerado significativo somente para as amostras do *cluster* 3 (Figura 23B).

Em amostras do *cluster* 2 três genes foram selecionados e observando as suas informações nos bancos de referência destaca-se EEPD1 (*endonuclease/exonuclease/phosphatase family domain containing 1*). Este gene ainda é pouco conhecido, porém já foi descrito possuindo um papel importante no controle da replicação e sendo importante para a via de reparo por recombinação homóloga (Kim et al. 2017). EEPD1 não foi classificado como um gene essencial em células tumorais, segundo *DepMap* (Broad Institute 2021) e está incluso no *Catalogue of Cancer Genes* (Cheng et al. 2014). Nestas amostras foi observado um *hazard-ratio* favorável (Figura 23A).

Já em amostras do *cluster* 3 diversos genes foram selecionados como significativos (Figura 23B). Entre eles, destaca-se ASH2L. Este gene possui um *hazard-ratio* favorável (Figura 23B) e é descrito como essencial para células tumorais, de acordo com o *DepMap* (Broad Institute 2021). Além disso, a sua alta expressão também já foi associada a um prognóstico favorável (Pontén 2008; Uhlen et al. 2017) e este está inserido no *Catalogue of Cancer Genes* (Cheng et al. 2014). Este gene é um componente importante de um complexo multimérico de histonas metiltransferase, que estão envolvidas com a manutenção da transcrição ativa (Zeng et al. 2020).

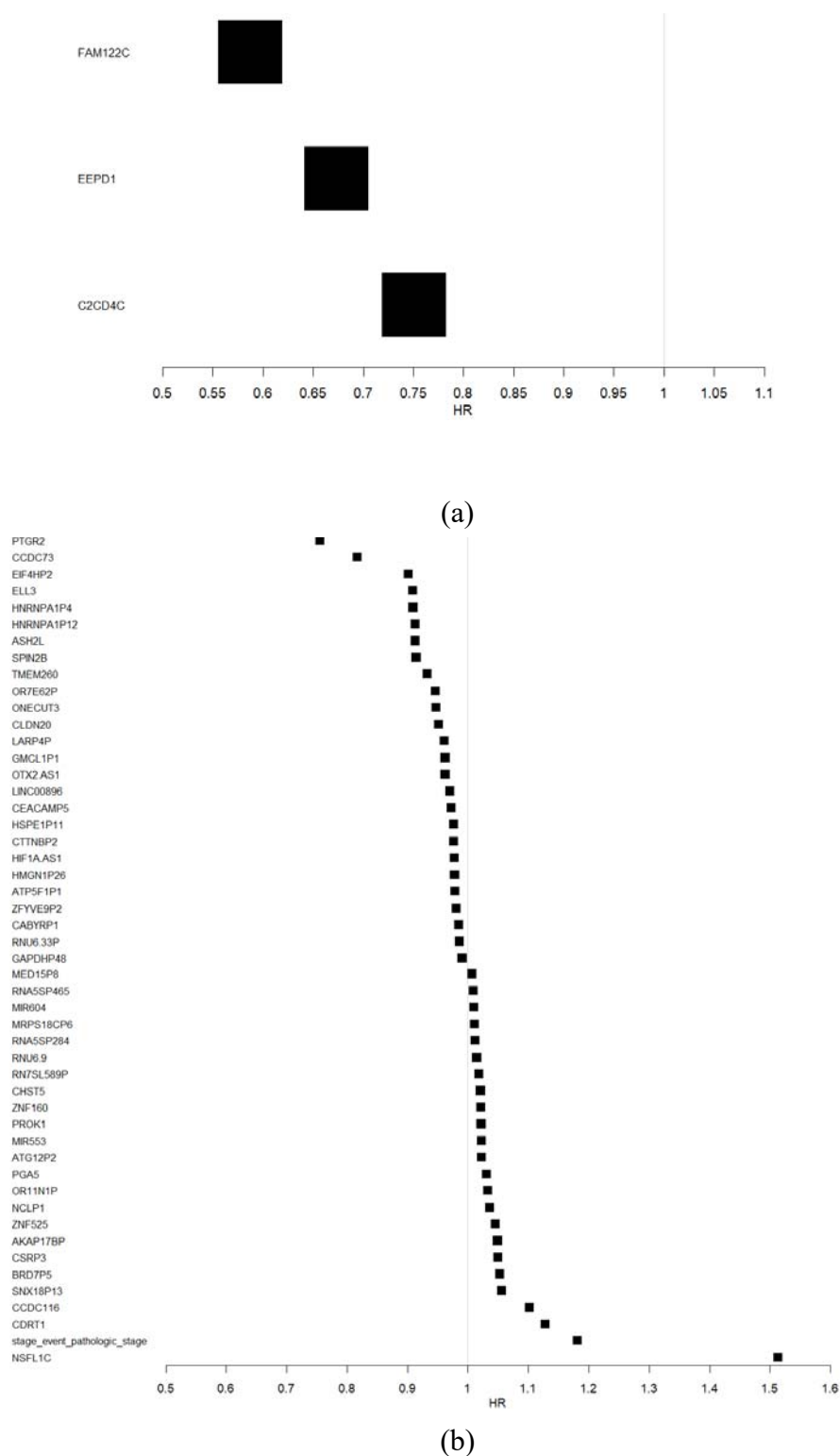


Figura 23 - Hazard-Ratio (HR) para amostras de carcinoma de células escamosas de pulmão. Hazard-ratio (HR) calculado para amostras de carcinoma de células escamosas de pulmão, através de modelos de modelos de predição: Lasso e *Elastic Net*. Foram utilizados dados para expressão gênica, assinatura mutacional e infiltrado inflamatório, além da idade (*age_at_initial_pathologic_diagnosis*) e estadiamento do tumor (*stage_event_pathologic_stage*); **(A)** Análise para amostras do *cluster 2* (ancestralidade africana); **(B)** Análise para amostras do *cluster 3* (ancestralidade europeia).

Em melanoma, foram observadas variáveis significativas somente para as amostras do *cluster* 1 (ancestralidade europeia) (Figura 24). Entre elas, foram identificados múltiplos genes além das informações para idade (*age_at_initial_pathologic_diagnosis*) e para estadiamento do tumor (*stage_event_pathologic_stage*) (Figura 24). Diferente dos tipos de tumor analisados anteriormente, para os casos de melanoma as assinaturas mutacionais SBS7a e SBS7b do COSMIC também foram identificadas (Figura 24). Estas assinaturas são predominantemente encontradas em melanomas e estão associadas a danos pela exposição à raios ultravioleta (Wellcome Sanger Institute 2020).

Entre os genes encontrados, cinco deles podem ser destacados de acordo com os bancos de referência analisados: VIM (*vimentin*), XDH (*xanthine dehydrogenase*) e USH2A (*usherin*). O gene VIM codifica para a proteína vimentina, um membro da família de filamentos intermediários e que é encontrada em diversas células não epiteliais, especialmente células mesenquimais (The UniProt Consortium 2019). A vimentina está normalmente associada ao núcleo, retículo endoplasmático e mitocôndria e é extremamente importante para a manutenção da integridade celular e na resistência ao estresse (Satelli e Li 2011). Nestas amostras, foi observado um *hazard-ratio* favorável (Figura 24), o que, em parte, também é observado pelo *Protein Cancer Atlas* (Pontén 2008; Uhlen et al. 2017). Neste banco foi constatado um prognóstico favorável em amostras de câncer de pulmão, pâncreas e mama quando há uma alta expressão deste gene. No entanto, também foi notado um prognóstico não favorável em outros tipos de tumor, como em casos de câncer de endométrio (Pontén 2008; Uhlen et al. 2017). Este gene não é considerado essencial em células tumorais segundo *DepMap* (Broad Institute 2021) e está incluso no *Catalogue of Cancer Genes* (Cheng et al. 2014).

Assim, como VIM, XDH também apresentou um *hazard-ratio* favorável (Figura 24), corroborando com os dados observados pelo *Protein Cancer Atlas* para tumores de urotélio (Pontén 2008; Uhlen et al. 2017). Este gene possui um papel importante na degradação de purinas, além de ter um papel importante no processo inflamatório e na imunidade inata (Liu et al. 2018). Ele está inserido no *Catalogue of Cancer Genes* (Cheng et al. 2014) e não é considerado um gene essencial para células tumorais (Broad Institute 2021). Outro gene a ser destacado é: USH2A. Diferente dos casos anteriores, este gene apresentou um *hazard-ratio* não favorável em casos de melanoma para o *cluster* 1 (Figura 24). USH2A um papel importante na visão e na audição e é essencial em células tumorais, segundo *DepMap* (Broad Institute 2021), além de estar inserido no *Catalogue of Cancer Genes*.

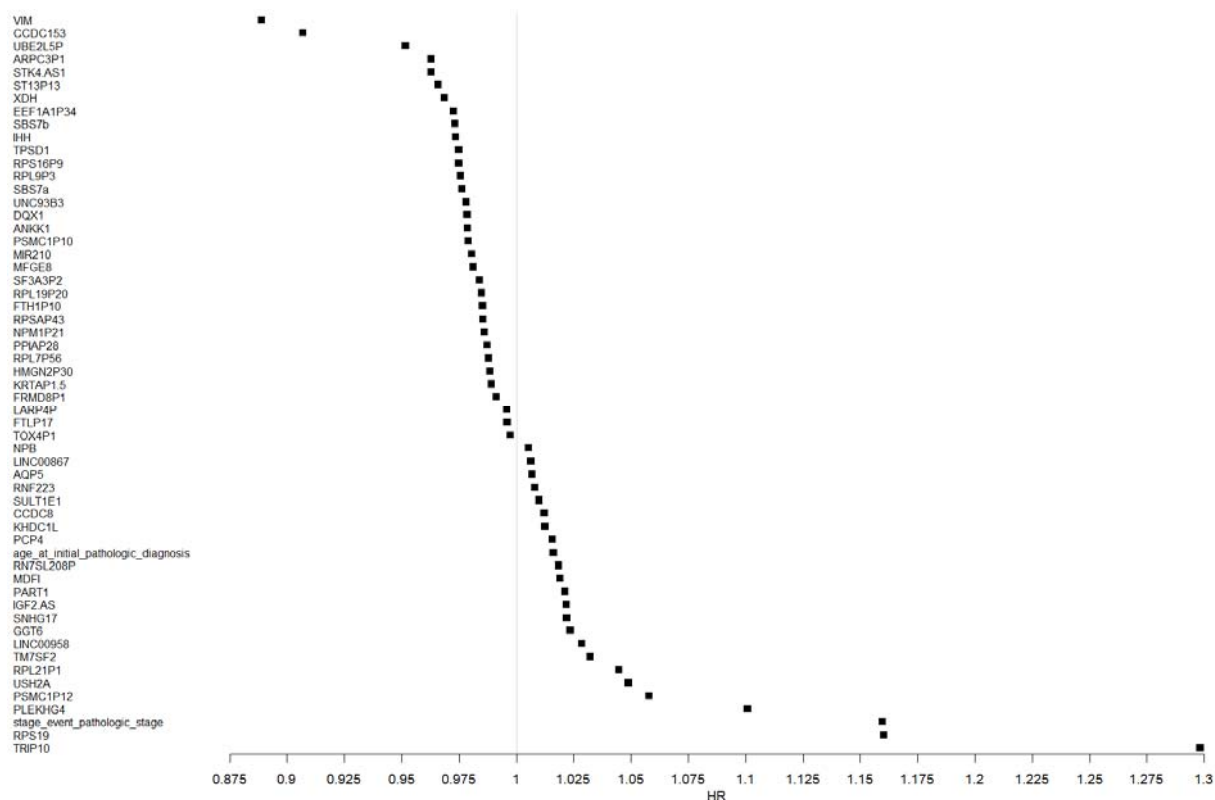


Figura 24 - Hazard-Ratio (HR) para amostras de melanoma. *Hazard-ratio* (HR) calculado para amostras de melanoma do *cluster 1* (ancestralidade europeia), através de modelos de modelos de predição: Lasso e *Elastic Net*. Foram utilizados dados para expressão gênica, assinatura mutacional e infiltrado inflamatório, além da idade (*age_at_initial_pathologic_diagnosis*) e estadiamento do tumor (*stage_event_pathologic_stage*).

5 DISCUSSÃO

O câncer é definido como um grupo de doenças caracterizadas pelo crescimento descontrolado e pela propagação de células anormais (American Cancer Society-ACS 2020). Embora os mecanismos da tumorigênese de vários tipos de câncer ainda não sejam conhecidos em sua totalidade, sabe-se que diversos fatores podem contribuir para este processo (ACS 2020). Entre eles, estão fatores passíveis de serem modificados e balanceados, como dieta, consumo de tabaco, exposição à radiação ultravioleta, sedentarismo, entre outros (Vineis e Wild 2014; ACS 2020). Além disso, existem fatores genéticos, intrínsecos ao indivíduo, que elevam o risco de desenvolvimento de tumores (Vineis e Wild 2014; ACS 2020).

Segundo dados de 2018, mundialmente, foram estimados cerca de 18,1 milhões de novos casos e 9,6 milhões de óbitos (WHO 2020b). Em comparação a outras doenças, o câncer ocupa a segunda colocação em número de óbitos no mundo, atrás somente de doenças cardiovasculares (Our World in Data 2017). Em termos de tratamento, diversos tipos de câncer ainda possuem opções muito limitadas, impactando diretamente na sobrevida destes pacientes (ACS 2020; Persi et al. 2020). Apesar disso, nas últimas décadas, ocorreram grandes avanços na área da oncologia, tanto na prevenção, quanto na detecção e no tratamento, reduzindo a taxa de mortalidade, cujo pico, em 1991, nos Estados Unidos, chegou a 215 óbitos a cada 100.000 pessoas. Hoje tem-se uma taxa próxima a 152 óbitos a cada 100.000 pessoas (ACS 2020).

Como mencionado anteriormente, um dos principais fatores de risco para o desenvolvimento de câncer é o fator genético e hereditário (Vineis e Wild 2014; ACS2020). Por esta razão, nos últimos anos, diversos estudos têm buscado avaliar a relação entre ancestralidade e o câncer. Várias discrepâncias, principalmente em relação à frequência do tumor, já foram constatadas entre grupos étnicos distintos. Em câncer de estômago, por exemplo, observou-se que aproximadamente 50% de todos os casos do mundo ocorriam na Ásia Oriental, em especial na China (McLean e El-Omar 2014). Apesar disso, asiáticos apresentam uma melhor taxa de sobrevida em comparação a outros grupos (Merchant et al. 2014).

Em parte, estas distinções podem ser explicadas (mas não limitando-se a) por fatores ambientais e sociais, incluindo diferenças socioeconômicas e acesso a tratamento (Özdemir e Dotto 2017; Gurdasani et al. 2019). Porém, já foi observado que, mesmo em condições iguais de acesso a tratamento, as discrepâncias observadas entre populações e grupos ainda são

predominantes (Andaya et al. 2013; Schreiber et al. 2014). Algumas diferenças genéticas e biológicas já foram descritas, como distinções no perfil mutacional, obesidade, inflamação crônica e sistema imune (inato e adaptativo) (Özdemir e Dotto 2017).

Por esta razão, no presente estudo, foram avaliados dois tipos de tumores: melanoma e carcinoma de pulmão de células não pequenas. Em termos populacionais, o melanoma aparece com maior frequência em países com população majoritariamente branca. Por exemplo, nações como Austrália e Nova Zelândia apresentam taxas de incidência entre 63,4 e 58,1 a cada 100.000 indivíduos em comparação com a região sul-central da Ásia, cuja taxa de incidência fica abaixo de 0.4 (IARC 2020). Além disso, observou-se que, em populações majoritariamente brancas, o melanoma cutâneo é predominante, enquanto em países da Ásia e África há uma maior tendência de desenvolvimento de MM e ALM (Schadendorf et al. 2015).

Já em relação ao carcinoma de pulmão de células não pequenas, observa-se, nos Estados Unidos, uma maior incidência em afrodescendentes em relação a indivíduos de descendência europeia, asiáticos ou ameríndios (Siegel et al. 2018). Também foram identificadas distinções na frequência de mutações de EGFR e ALK, com indivíduos com descendência asiática apresentando taxas de até 60%, em relação a indivíduos com descendência europeia que tiveram uma incidência de até 10% (Gridelli et al. 2015).

É importante salientar que, apesar das diferenças étnicas observadas em diversos estudos, esta abordagem não possui um respaldo genético bem definido. Portanto, pesquisas com o intuito de investigar a relação entre a variação genética em nível populacional, com o seu impacto na saúde, tornam-se extremamente relevantes (Gurdasani et al. 2019). Com isso, o primeiro passo do presente trabalho foi realizar a inferência da ancestralidade para as amostras de interesse. Com este intuito, foi criado um *pipeline*, a fim de automatizar este processo. Para isso foi aplicado o software *Snakemake* (Köster e Rahmann 2012), cuja função é a de gerenciar todo o *workflow* do *pipeline*.

Atualmente o *pipeline* consiste em etapas de pré-processamento de dados, onde diversos filtros são aplicados e cujos valores podem ser prontamente modificados, e etapas de análise, onde são realizadas quatro tarefas distintas. A primeira delas se refere à inferência da ancestralidade individual com o software *Admixture* (Alexander et al. 2009). Para esta tarefa, também foram testados outros programas, como *STRUCTURE* (Pritchard et al. 2000), porém, não foram observadas discrepâncias nos resultados, e com *Admixture* foi observada uma grande redução no tempo de análise. Enquanto com *STRUCTURE* algumas análises levavam algumas horas, com *Admixture* isto era realizado dentro de alguns poucos minutos.

A segunda tarefa consiste no controle de qualidade da análise. Previamente, valores de ancestralidade foram definidos para as amostras dos bancos de referência. Isto foi realizado utilizando *Admixture* em seu modo não-supervisionado. Essa comparação é feita utilizando o método gráfico de *Bland-Altman* (Altman e Bland 1983). Para esta etapa, a inferência de ancestralidade individual para as amostras de adenocarcinoma gástrico e de pulmão, melanoma e carcinoma de células escamosas de pulmão apresentaram, como um todo, um intervalo de concordância entre 0.05 e -0.05, indicando que o conjunto de variantes obtido pelo *pipeline* conseguiu resgatar, de forma eficiente, as ancestralidades do banco de referência.

A terceira tarefa se refere à construção de um gráfico para análise dos componentes principais ou PCA. Neste gráfico são utilizados os dados de variantes, através do software PLINK (Purcell et al. 2007) das amostras de interesse, juntamente com os dados do banco de referência. Isto é feito a fim de observar como as amostras de interesse se agrupam em relação às populações extraídas dos bancos 1000G e HGDP.

Por último, é utilizado o método *Lancaster Estimator of Independence* (LEI) (Wathen et al. 2019) para seleção das variantes mais informativas para ancestralidade. Neste caso, o usuário deve fornecer o número de variantes desejadas a serem selecionadas. Ao fim deste processo, é novamente realizado um teste de qualidade, com o intuito de verificar o quanto de informação este conjunto consegue recuperar, em relação ao banco de referência.

O método LEI aqui aplicado é uma versão adaptada do método original. Em vez de ter um único *score* para cada variante, levando todas as superpopulações em consideração, estas superpopulações foram avaliadas em dez pares. O uso deste método, com todas as superpopulações, mostrou-se limitado pela proximidade entre elas, e resultou em um viés no *score* obtido, e, portanto, no processo de seleção. Observou-se que variantes com *score* mais alto, em geral, eram relevantes principalmente para a superpopulação AFR e EAS, enquanto variantes relevantes para a distinção entre AMR, EUR e SAS eram removidas por apresentarem um *score* intermediário. Isto se deve, principalmente, à maior proximidade entre estas superpopulações. Esta modificação e análise por pares possibilitou identificar, de maneira mais equilibrada e eficiente, as variantes relevantes para todas as superpopulações. Outros métodos também foram testados além do LEI, como F_{st} (Weir e Cockerham 1984) e *Infocalc* (Rosenberg et al. 2003). Porém, a performance de ambos ficou abaixo do LEI em sua versão original, e por isso, seguimos com este método.

É importante salientar que, ao nosso conhecimento, não existe um *pipeline* publicamente disponível com todas estas características. Porém, alguns aspectos ainda podem ser

aprimorados. O próximo passo envolveria o desenvolvimento de uma metodologia, a fim de selecionar o conjunto otimizado de variantes de forma automática.

Como um todo, as análises de ancestralidade com as amostras do adenocarcinoma gástrico e de pulmão, melanoma e carcinoma de células escamosas de pulmão do TCGA confirmaram as disparidades em termos étnicos, já observadas dentro do próprio TCGA (Spratt et al. 2016). Em nossas análises, foram avaliados 1936 pacientes relacionados a esses quatro tipos de tumor. Entre eles, um total de 86,82% representa indivíduos com altos índices para ancestralidade europeia, enquanto somente 6,87% representam indivíduos com ancestralidade africana e 6,30% para ancestralidade asiática.

A baixa representatividade e diversidade étnica representa uma limitação no presente estudo e é uma questão não somente observada no TCGA, mas também em diversos outros bancos de dados (Xie et al. 2020a). Apesar da grande importância destes bancos para o estudo e avanço dos tratamentos de diversas doenças, é necessário avaliar o impacto deste viés e questionar se os resultados obtidos a partir destes bancos podem ser generalizados a nível populacional. O atual trabalho mostra que existem diferenças significativas entre populações, em relação ao infiltrado inflamatório, perfil mutacional e expressão gênica e que, em parte, estas diferenças podem impactar na sobrevida do paciente. Estes dados sugerem que a diversidade populacional é um fator importante a ser levado em consideração e que pode ter um impacto direto no resultado.

Em relação ao infiltrado inflamatório, diferentes perfis foram observados entre os grupos. Somente para os casos de melanoma não foram identificadas diferenças significativas, provavelmente devido ao número limitado de pacientes em cada um dos grupos. Já em adenocarcinoma gástrico e de pulmão e em carcinoma de células escamosas de pulmão, foram constatadas diferenças importantes e que ocorrem apesar do tipo de tumor. Nos três casos, verificou-se que indivíduos com ancestralidade africana possuem uma tendência a menores níveis de macrófagos M2 no tumor em relação aos outros grupos. Este tipo de célula possui uma associação importante com o desenvolvimento tumoral, pois está geralmente associada a uma atividade pró-tumoral (Najafi et al. 2019; Casseta et al. 2019; Li et al. 2019; Eum et al. 2020). Estas células possuem uma característica imunossupressiva e capacidade de influenciar diretamente no desenvolvimento, na sobrevivência e na proliferação do tumor (Najafi et al. 2019), levando a um possível pior prognóstico em carcinoma gástrico, por exemplo, promovendo a metástase e transição epitélio mesenquimal (Casseta et al. 2019; Li et al. 2019). Além disso, existem indícios de que a presença de células com este perfil no tumor contribui para a resistência à quimioterapia (Castells et al. 2012; Nakasone et al. 2012).

Entre adenocarcinoma gástrico e de pulmão, também foram identificadas outras semelhanças entre si. Em ambos foram constatadas diferenças significativas entre os grupos para monócitos e células T *Natural Killer*. No caso dos monócitos, o mesmo perfil foi observado, com indivíduos com ancestralidade europeia apresentando os maiores valores em relação aos outros dois grupos. O papel dos monócitos no tumor e o seu impacto ainda são pouco compreendidos, porém, observou-se que tumores de câncer de mama e endométrio induzem a expansão de populações não-clássicas de monócitos, levando a uma alteração do transcriptoma de populações de monócitos (Cassetta et al. 2019). Esta associação do tumor com populações não-clássicas de monócitos mostrou ter um impacto negativo no tamanho e estadiamento de tumores de câncer de mama (Feng et al. 2011).

No caso de células T *Natural Killer*, observou-se um maior destaque em populações com ancestralidade africana, apresentando valores superiores, principalmente, aos indivíduos com ancestralidade europeia. Estas células possuem uma alta capacidade de modular o sistema imune através de citocinas e, por isso, têm um papel importante na *immunosurveillance* do tumor (Krijgsman et al. 2018). No início do desenvolvimento do tumor, estas células podem atuar estimulando outros subtipos de células T e células natural killer efetoras, possibilitando a eliminação do tumor (Krijgsman et al. 2018). Ao mesmo tempo, caso ocorra a progressão do tumor, pode haver uma redução da população deste tipo celular, devido a uma resposta exacerbada, levando à morte celular e as células restantes tornam-se hiporesponsivas ou transitam para um outro subtipo, auxiliando no desenvolvimento do tumor (Krijgsman et al. 2018).

Diferenças mais específicas, associadas ao tipo de tumor, também foram identificadas. No caso de adenocarcinoma gástrico, foram identificadas células T de memória, tanto através da utilização do *CIBERSORT*, quanto com *xCell*, associadas principalmente ao grupo de ancestralidade africana. Além disso, também foi observada uma diferença significativa entre células estromais como fibroblastos e condrócitos, indicando menores níveis em indivíduos com ancestralidade africana. Isto foi confirmado pela utilização do software *ESTIMATE*, indicando uma grande diferença nos valores para células estromais e, portanto, na pureza do tumor.

Em adenocarcinoma de pulmão, os métodos utilizados identificaram diferenças significativas em células dendríticas plasmocitóides e ativadas. Enquanto, no primeiro, indivíduos com ancestralidade asiática apresentaram menores níveis em relação aos outros dois grupos, em células dendríticas ativadas foi observado o oposto. Células dendríticas possuem um importante papel na resposta tumoral, induzindo a resposta de células T, porém sabe-se que

no tumor a sua atuação é limitada e, quando associadas ao tumor, apresentam defeitos em seu processo de diferenciação e ativação, levando a um baixo estímulo do sistema imune (Veglia e Gabrilovich 2017). Células dendríticas plasmocitóides constituem um subtipo único de células dendríticas, sendo especializadas na produção de interferon do tipo I (Swiecki e Colonna 2015; Veglia e Gabrilovich 2017). O seu recrutamento pelo tumor, em geral, está associado a prognóstico negativo (Swiecki e Colonna 2015).

O próximo passo foi a avaliação do perfil mutacional somático. Mutações desse tipo possuem grande importância no contexto tumoral, fornecendo vantagens para o desenvolvimento e sobrevivência de células tumorais (Martincorena et al. 2017). Esta avaliação constatou que indivíduos com ancestralidade africana apresentaram uma maior frequência de mutações em TP53, principalmente em relação a indivíduos de ancestralidade europeia em adenocarcinoma gástrico e de pulmão. Esta diferença em TP53 também foi observada em múltiplos estudos, principalmente em casos de câncer de mama (Keenan et al. 2015; Zhu et al. 2016; Yuan et al. 2018). TP53 é o gene mais frequentemente mutado em tumores, indicando uma pressão para a inativação deste gene durante o desenvolvimento do tumor (Bykov et al. 2018). Normalmente, o transcrito deste gene funciona como um sensor para estresse celular, sendo ativado em condições de dano ao DNA e estresse oncogênico, levando a célula à apoptose (Bykov et al. 2018). Esta diferença observada poderia explicar, em parte, a tendência a um pior prognóstico observada em casos de adenocarcinoma gástrico e de pulmão para indivíduos com ancestralidade africana. Utilizando a plataforma *Xena* (Goldman et al. 2020), com estas amostras é possível observar que a presença de mutações em TP53 leva a um pior prognóstico de maneira significativa em adenocarcinoma de pulmão. O mesmo não foi constatado em casos de adenocarcinoma gástrico.

Também em adenocarcinoma gástrico, em indivíduos de ancestralidade asiática, foi identificado como significativo o gene DCLK1. Isto não ocorreu nos outros dois grupos avaliados. Este gene é descrito como um marcador de células-tronco tumorais, em casos de câncer de cólon e de rim (Nakanishi et al. 2013; Chandrakesan et al. 2017; Mohammadia et al. 2018; Liu et al. 2019). Além disso, a ação deste gene é capaz de regular sinais oncogênicos, através de uma via não canônica dependente de beta-catenina em carcinoma hepatocelular (Ali et al. 2020). Também foi observado que a alta expressão deste gene é capaz de promover o processo de metástase em células de câncer de mama (Liu et al. 2019).

Como parte do estudo do impacto de mutações somáticas, também foram avaliadas assinaturas mutacionais. Este tipo de análise nos permite avaliar quais processos mutacionais estão mais associados a cada um dos grupos e a cada tipo de tumor. Em adenocarcinoma

gástrico, não foram observadas diferenças significativas entre os grupos, em nenhuma das assinaturas avaliadas, em 2 dos 3 métodos avaliados. Este *dataset* contém amostras oriundas do Brasil, especificamente do Hospital do Câncer de Barretos, e por isso também tentamos verificar se haveria diferença entre o observado nestas amostras, com o restante do *dataset*. Também não foram encontradas, porém, diferenças significativas entre estes conjuntos.

Em melanoma, foram significativas SBS7a, SBS7b e SBS10b, segundo os métodos *signeR* e *MutationalPatterns*. Estas assinaturas apresentaram valores mais altos em indivíduos com ancestralidade europeia. Enquanto as duas primeiras estão relacionadas à exposição a raios ultravioleta, SBS10b é marcada por um grande número de mutações somáticas e está associada com mutações à DNA polimerase subunidade épsilon (POLE) (Tate et al. 2019; Wellcome Sanger Institute 2020). POLE é responsável pela replicação da fita líder de DNA e, portanto, defeitos podem resultar em um grande número de mutações somáticas (Hatakeyama et al. 2018). Além disso, mutações neste gene já foram observadas como sendo benéficas para imunoterapia, principalmente em câncer de endométrio, glioblastoma e colorretal (Mehnert et al. 2016; Gong et al. 2017; Yao Y et al. 2019). Para estas amostras, isto poderia indicar um possível potencial terapêutico. SBS7a e SBS7b são assinaturas conhecidamente associadas principalmente a melanomas (Alexandrov et al. 2020). Já SBS10b é observada em poucos tumores, como em adenocarcinomas de útero e colorretal (Alexandrov et al. 2020).

Em adenocarcinoma de pulmão, foram identificadas as assinaturas SBS4 e SBS24, com os três métodos aplicados e com o mesmo padrão entre os grupos. Nos dois casos, foi observada uma maior exposição em indivíduos com ancestralidade africana. SBS4 está associada ao uso de tabaco, e está conhecidamente associada a adenocarcinoma de pulmão (Tate et al. 2019; Wellcome Sanger Institute 2020). Sabe-se que o consumo de tabaco aumenta o risco de desenvolver diversos tipos de câncer, como câncer de pulmão, laringe, oral e fígado, por exemplo (U.S. Department of Health and Human Services-HHS 2020). Este composto pode contribuir para carcinogênese, através de diferentes mecanismos, sendo diretamente genotóxico, causando a hipermetilação de promotores e inflamação (HSS 2020). Esta assinatura também é observada em casos de carcinoma de células escamosas de pulmão e em câncer de fígado (Wellcome Sanger Institute 2020). Já SBS24 está relacionada à exposição por aflatoxina (Tate et al. 2019; Wellcome Sanger Institute 2020). Esta assinatura é pouco observada em tumores, sendo detectada somente em amostras de câncer biliar e de fígado (Wellcome Sanger Institute 2020). Este composto é um tipo de metabólito secundário produzido por certos tipos de fungos, como *Aspergillus flavus* e *A. parasiticus*, e é amplamente encontrado como um contaminante em uma grande quantidade de plantações (Marchese et al. 2018). Além disso,

aflatoxinas são classificadas como substâncias carcinogênicas para humanos (Grupo 1) (Marchese et al. 2018).

Em carcinoma de células escamosas de pulmão, foram encontradas assinaturas com diferenças significativas somente com o uso do pacote *MutationalPatterns*. Apesar deste tipo de câncer apresentar a maior proporção de tumores com esta assinatura (Wellcome Sanger Institute 2020), não foram observadas diferenças significativas entre os grupos, como foi constatado em casos de adenocarcinoma de pulmão. Porém, assim como ocorre em adenocarcinoma de pulmão, segundo *MutationalPatterns*, foram notadas diferenças significativas entre os grupos para a assinatura SBS24 e o mesmo padrão foi observado com indivíduos com ancestralidade europeia e africana, apresentando valores mais altos em relação aos indivíduos com ancestralidade asiática.

Com as diferenças observadas no infiltrado inflamatório e em processos mutacionais, o próximo passo foi avaliar a expressão gênica. Com este objetivo, foi realizada a análise de expressão diferencial. Portanto, foi utilizado o software *Deseq2*, juntamente com o método de *Gene Set Enrichment Analysis*, a fim de identificar as vias biológicas enriquecidas em cada um dos grupos. Aqui, foi aplicada a nomenclatura do banco de dados *Gene Ontology* para processos biológicos. Também foram utilizados o software *Cytoscape* e as extensões *Autoannotate*, *Enrichment Map*, *ClusterMaker* e *WordCloud* para compilar os dados obtidos.

Em adenocarcinoma gástrico, foram observadas diferenças importantes, principalmente entre pacientes de ancestralidade europeia e asiática. Neste caso, todas as vias encontradas mostraram-se positivamente enriquecidas para indivíduos com ancestralidade europeia. Além disso, nota-se que a maior parte das vias identificadas está relacionada ao sistema imune, com genes atuando na promoção da resposta imune, inflamação e migração de células como linfócitos e neutrófilos. Estas vias podem ajudar a explicar alguns dos dados obtidos a partir, principalmente, do software *CIBERSORT*, com neutrófilos, monócitos e células T regulatórias aparecendo em maior quantidade no tumor, principalmente em relação a indivíduos com ancestralidade asiática. Isto é importante pois a presença de neutrófilos e células T regulatórias no tumor indicam um pior prognóstico (Abbas et al. 2017b; Shaul e Fridlender 2019; Paijens et al. 2020) e podem explicar, em parte, a diferença na sobrevida observada entre estas duas populações.

Assim como observado em adenocarcinoma gástrico, em carcinoma de células escamosas de pulmão, uma via para promoção da resposta inflamatória também foi identificada como diferencialmente expressa. Neste caso, a diferença foi observada entre indivíduos com ancestralidade africana e europeia, apresentando um *score* negativo em relação ao primeiro

grupo. Como um todo, estes resultados sugerem que, nestes dois tipos de tumor, o processo inflamatório esteja mais ativo em pacientes com ancestralidade europeia em relação aos outros grupos.

Em adenocarcinoma de pulmão e melanoma, não foram identificadas vias diferencialmente expressas relacionadas ao sistema imune. No primeiro caso, foram observadas vias, principalmente, enriquecidas positivamente em pacientes de ancestralidade asiática em relação àqueles com ancestralidade africana. Em sua maior parte, estas vias estão associadas principalmente à regulação do transporte celular, da localização de proteínas na célula e do processo de biossíntese e tradução mitocondrial. Já em melanoma não foram encontradas vias significativas.

Como um todo, os múltiplos campos avaliados pelo presente trabalho sugeriram diferenças importantes entre os grupos e que possuem conhecido impacto na sobrevivência de pacientes. Por esta razão, buscamos uma forma de avaliar todas estas variáveis, em conjunto e o seu impacto direto na sobrevivência destes pacientes. Foram utilizados modelos de predição, Lasso e *Elastic Net*. Estes modelos têm se tornado muito importantes para o estudo e avaliação, em casos onde há um grande número de variáveis preditoras e, por isso, têm sido especialmente aplicados na área da genômica (Ogutu et al. 2012; Waldmann et al. 2013; Mostafaei et al. 2018). Aqui, foram implementadas as variáveis para infiltrado inflamatório (*CIBERSORT* e *xCell*), assinatura mutacional (*signeR*) e expressão gênica. Para os genes significativos, também foram utilizados os bancos *Cancer Gene Census*, *Dependency Map* e *Catalogue of Cancer Genes*, a fim de avaliar o conhecimento já existente. É importante destacar que não foram identificadas variáveis significativas para pacientes com ancestralidade asiática em todos os tumores.

Em adenocarcinoma gástrico, em indivíduos com ancestralidade africana dois genes significativos se destacaram: *MASTL* e *TP73*. Ambos apresentaram um *hazard-ratio* favorável. *MASTL* possui um papel importante na mitose, especialmente no checkpoint de G2/M. A sua atividade previne a entrada prematura na mitose, minimizando, portanto, erros na segregação cromossômica (Uppada et al. 2018; Fatima et al. 2020; Hermida et al. 2020). Este gene está incluído no *Catalogue of Cancer Genes* e é considerado essencial, de acordo com o *Dependency Map* (Cheng et al. 2014; Broad Institute 2021). Em câncer, foi observado que o aumento da expressão desta quinase está associado a um maior risco, pior prognóstico e resistência à quimioterapia (Pontén 2008; Uhlen et al. 2017; Uppada et al. 2018). Além disso, este gene também tem sido visto como um potencial alvo terapêutico (Fatima et al. 2020). *TP73* é um gene membro da família que compreende *TP63* e *TP53* e, de maneira similar a *TP53*, é ativado em resposta a danos ao DNA e regula a expressão de genes envolvidos com a parada do ciclo

celular e apoptose (Yao Z et al. 2019). Apesar disso, a função deste gene na tumorigênese ainda é pouco conhecida. Em casos de câncer cervical, endométrio e cabeça e pescoço, a sua expressão foi visto como um prognóstico favorável (Pontén 2008; Uhlen et al. 2017).

Em pacientes com ancestralidade europeia, três genes tiveram maior destaque: AHRR, FAAH e PWP2. Enquanto os dois primeiros apresentaram um *hazard-ratio* favorável, o último apresentou um *hazard-ratio* não favorável. Em concordância com o resultado aqui apresentado, a expressão de FAAH também foi observada como um prognóstico positivo em casos carcinoma gástrico, de rim e de endométrio (Pontén 2008; Uhlen et al. 2017). Já AHRR e PWP2 não foram identificados como fatores prognósticos significativos, segundo *Protein Cancer Atlas* (Pontén 2008; Uhlen et al. 2017).

Em adenocarcinoma de pulmão, para casos com ancestralidade africana, dois genes se destacaram: COL22A1 e ACBD3. COL22A1 apresentou um *hazard-ratio* não favorável e isto coincide com o observado pelo *Protein Cancer Atlas*, para tumores de pulmão e em glioma (Pontén 2008; Uhlen et al. 2017). Este gene codifica para a um específico tipo de colágeno (colágeno XXII), que contribui para a estabilização de junções miotendionosas e fortalecimento de ligações do músculo esquelético durante a atividade contrátil (Ton et al. 2018). Também existem indícios de que os níveis de mRNA para COL22A1 estejam significativamente associados à metástase para o linfonodo, ao estadiamento e à redução da sobrevida em casos de câncer de cabeça e pescoço (Misawa et al. 2014). ACBD3 apresentou valores muito altos de *hazard-ratio* em relação as outras variáveis, porém possui poucas informações nos bancos de referência analisados. A sua função está associada à: esteroidogênese, embriogênese, neurogênese e tráfico de membrana (Yue et al. 2019). Além disso, a sua expressão também foi reportada como um importante fator prognóstico em pacientes de câncer de mama (Huang et al. 2019). A alta expressão deste gene foi capaz de promover a renovação de células tumorais e tumorigênese em câncer de mama (Huang et al. 2019).

Em casos de pacientes com ancestralidade europeia, um gene se destacou: PLEK2. Nestas amostras foi observado um *hazard-ratio* não favorável. Isto é corroborado pelos dados do *Protein Cancer Atlas*, impactando negativamente o prognóstico de pacientes de câncer de pulmão e pâncreas (Pontén 2008; Uhlen et al. 2017). Este gene, segundo *DepMap* é considerado essencial em células tumorais (Broad Institute 2021). A proteína expressa por este gene é capaz de auxiliar na redistribuição de actina dentro da célula, levando a uma ampliação da célula (Shen et al. 2019). Em câncer, já foi observado tendo um papel importante no processo de metástase de células de câncer de mama do tipo luminal A (Shen et al. 2019).

Em carcinoma de células escamosas de pulmão, dois genes se destacaram em pacientes de ancestralidade africana e europeia, EEPD1 e ASH2L, respectivamente. EEPD1 apresentou um *hazard-ratio* favorável, mas ainda é pouco conhecido e possui pouca informação nos bancos de referência analisados. Apesar disso, está incluído no *Catalogue of Cancer Genes*. Este gene é descrito como um fator importante no controle da replicação e na via de reparo por recombinação homóloga (Kim et al. 2017). ASH2L também apresentou um *hazard-ratio* favorável e é descrito como um gene essencial para células tumorais (Broad Institute 2021). Corroborando com este resultado, o *Protein Cancer Atlas* também associa este gene a um prognóstico favorável.

Em melanoma, foram encontradas variáveis significativas somente para indivíduos com ancestralidade europeia. Além disso, diferentemente dos outros tumores, uma assinatura do COSMIC também foi considerada significativa (assinaturas SBS7a e SBS7b). Ambas as assinaturas apresentaram um *hazard-ratio* favorável. Estas são conhecidamente encontradas, principalmente, em melanomas e estão associadas à exposição a raios ultravioletas (Wellcome Sanger Institute 2020). Sabe-se que este tipo de exposição aumenta o risco de desenvolvimento de tumores (Vineis e Wild 2014; ACS 2020), porém, este resultado parece indicar que tumores oriundos destas assinaturas apresentam uma menor gravidade.

Além disso, alguns genes relevantes também foram considerados significativos, como, por exemplo: VIM. Este gene é um membro da família de filamentos intermediários e está normalmente associada ao núcleo, retículo endoplasmático e mitocôndria e é extremamente importante para a manutenção da integridade celular e na resistência ao estresse (Satelli e Li 2011). Nestas amostras foi observado um *hazard-ratio* favorável. Este mesmo padrão também foi notado no Protein Cancer Atlas em casos de câncer de pulmão, mama e pâncreas. Já foi constatado que VIM possui um papel importante no processo migratório e na manutenção da morfologia em células gigantes multinucleadas de câncer mama (Xuan et al. 2020).

Como um todo, estes resultados mostram a contribuição do *pipeline* para estudos futuros e traz discussões importantes sobre metodologia. Constatou-se também a importância de *datasets* mais representativos, em termos populacionais. Esta questão se mostrou um fator limitante para o atual estudo, e as diferenças aqui observadas mostram que a ancestralidade é um fator que deve ser levado em consideração, ao extrapolar resultados para um nível populacional.

O uso de modelos de predição possibilitou encontrar genes relevantes e que merecem aprofundamento sobre o seu papel no câncer e na ancestralidade. Entre eles, podem ser citados: MASTL, TP73, FAAH, PWP2, AHRR, COL22A1, ACBD3, PLEK2, EEPD1 e ASH2L.

Apesar de a maior parte das diferenças observadas no infiltrado inflamatório, e nas assinaturas mutacionais, não ter apresentado um impacto direto na sobrevida, ainda pode estar associada a outros mecanismos e pode estar envolvida com o aumento do risco para surgimento do tumor.

Com os dados para expressão diferencial e infiltrado inflamatório, principalmente em relação aos casos de adenocarcinoma gástrico e carcinoma de células escamosas de pulmão, foi observada uma maior resposta imune em indivíduos com ancestralidade europeia, associada, principalmente à linfócitos, neutrófilos e monócitos.

Com dados do perfil mutacional, também foi possível identificar diferenças relevantes na frequência de mutações em TP53, acentuando-se em indivíduos com ancestralidade africana e ocorrendo, principalmente, em casos de adenocarcinoma gástrico e de pulmão. Esta distinção também foi percebida em outros estudos, principalmente em casos de câncer de mama (Keenan et al. 2015; Yuan et al. 2018).

6 CONCLUSÕES

Os resultados reforçam a importância do *pipeline* utilizado para inferência de ancestralidade individual, tornando-se uma importante contribuição para estudos futuros. Constatou-se ainda, a necessidade de *datasets* mais representativos, em termos populacionais, visto que a grande maioria das amostras no TCGA apresentou ancestralidade europeia.

Em relação a expressão gênica diferencial, observou-se uma maior atividade de vias imunes relacionada a ancestralidade europeia, principalmente em relação à macrófagos e monócitos. Este padrão está em concordância com os dados obtidos para o infiltrado inflamatório, onde foi observada uma maior quantidade de monócitos e macrófagos M2 em indivíduos com ancestralidade europeia.

Em assinaturas mutacionais, os resultados indicaram a existência de atividade mutacional nas populações avaliadas, principalmente em casos de adenocarcinoma de pulmão e melanoma. Em melanoma, foi observada uma maior exposição de assinaturas, em pacientes com ancestralidade europeia, relacionadas a mutações em POLE e exposição à UV. Enquanto em adenocarcinoma de pulmão, constatou-se uma maior exposição de assinaturas relacionadas ao uso de tabaco e exposição a aflatoxina em indivíduos com ancestralidade africana.

Também foi avaliado o perfil mutacional somático, e assim como observado pelo estudo TCGAA, constatou-se uma maior frequência de mutações em TP53 em indivíduos com ancestralidade africana, principalmente em casos de adenocarcinoma gástrico e de pulmão.

Apesar de a maior parte das diferenças observadas no infiltrado inflamatório e nas assinaturas mutacionais não ter apresentado um impacto direto na sobrevida, ainda pode estar associada a outros mecanismos e podem estar envolvidas com o aumento do risco para surgimento do tumor. Estudos mais aprofundados, a fim de validar e para um melhor entendimento dos aspectos observados, ainda são necessários.

Embora a compreensão acerca de vários mecanismos moleculares em câncer tenha avançado substancialmente nos últimos anos, ainda temos uma visão muito limitada sobre a interação desses mecanismos em um contexto amplo envolvendo ancestralidade. Este trabalho buscou descrever vários aspectos dessa interação.

7 REFERÊNCIAS

Abbas A, Lichtman A, Pillai S. Cellular and molecular immunology. 9th ed. New York: Elsevier; 2017a. Properties and overview of immune responses; p.1-12.

Abbas A, Lichtman A, Pillai S. Cellular and molecular immunology. 9th ed. New York: Elsevier; 2017b. Immunity to tumors; p.397-416.

Akincilar S, Unal B, Tergaonkar V. Reactivation of telomerase in cancer. Cell Mol Life Sci. 2016 Apr; 73(8):1659-70.

Alberts B, Johnson A, Lewis J, Morgan D, Raff M, Roberts K, Walter P. Molecular Biology of the Cell. 6th ed. New York: Garland Science; 2014. Cancer; p1091-1144.

Alexander D, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009 Sep; 19(9):1655-64.

Alexandrov L, Kim J, Haradhvala N, Huang M, Ng A, Wu Y. et al. The repertoire of mutational signatures in human cancer. Nature. 2020 Feb; 578(7793):94-101.

Ali N, Nguyen C, Chandrakesan P, Wolf R, Qu D, May R. et al. Sci Rep. 2020 Jun 29. 10(1):10578.

Altman D, Bland M. Measurement in medicine: the analysis of method comparison studies. The Statistician. 1983 Sep; 32(3):307-17.

American Cancer Society-ACS. Cancer Facts & Figures 2020. American Cancer Society; 2020. Disponível em: <https://bit.ly/35UTx2e>. [2021 Jan 13]

Andaya A, Enewold L, Zahm S, Shriver C, Stojadinovic A, McGlynn K. et al. Race and Colon Cancer Survival in an Equal-Acess Healthcare System. Cancer Epidemiol Biomarkers Prev. 2013 Jun; 22(6):1030-36.

Anders S, Pyl P, Huber W. HTseq - a Python framework to work with high-throughput sequencing data. *Bioinformatics*. 2015 Jan 15; 31(2):166-69.

Andor N, Maley C, Ji H. Genomic Instability in Cancer: Teetering on the limit of tolerance. *Cancer Res*. 2017 May 1; 77(9):2179-85.

Aran D, Hu Z, Butte A. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol*. 2017 Nov 15;18(1):220.

Auwerda G, Carneiro M, Hartl C, Poplin R, Angel G, Levy-Moonshine A. et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics*. 2013; 43(1110):11.10.1-11.10.33.

Avise JC, Ayala FJ. In the light of evolution III: Two centuries of Darwin. *Proc Natl Acad Sci U S A*. 2009 Jun 16; 106 Suppl 1:9933-38.

Bansal V, Libiger O. Fast individual ancestry inference from DNA sequence data leveraging allele frequencies for multiple populations. *BMC Bioinformatics*. 2015 Jan 16; 16:4.

Basa R, Davies V, Li X, Murali B, Shah J, Yang B. et al. Decreased anti-tumor cytotoxic immunity among microsatellite-stable colon cancers from African Americans. *PLoS One*. 2016 Jun 16; 11(6):e0156660.

Blokzijl F, Janssen R, Boxtel R, Cuppen E. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. *Genome Med*. 2018 Apr 25; 10(1):33.

Boissier F, Schmidt C, Linnermann J, Fribourg S, Perez-Fernandez J. Pwp2 mediated UTP-B assembly via two structurally independent domains. *Sci Rep*. 2017 Jun 9; 7(1):3169.

Boland P, Yurgelun M, Boland C. Recent progress in Lynch syndrome and other familial colorectal cancer syndromes. *CA Cancer J Clin*. 2018 May; 68(3):217-31.

Bray F, Ferlay J, Soerjomataram I, Siegel R, Torre L, 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(6):394-424.

Broad Institute. Picard. 2018. Disponível em: <https://bit.ly/3cJEL2F>. [2020 Nov 20]

Broad Institute. Depmap portal. 2021. Disponível em: <https://bit.ly/36CJUp8>. [2021 Jan 10]

Brookfield J. Population genetics. *Curr Biol*. 1996 Apr 1; 6(4):354-56.

Budiaji W. kmed: Distance-Based K-Medoids. R package version 0.4.0. 2021. Disponível em: <https://bit.ly/3sARmKW>. [2021 Jan 10]

Burkhardt R. Lamark, evolution, and the inheritance of acquired characters. *Genetics*. 2013 Aug; 194(4):793-805.

Burnet M. Cancer; a biological approach. I. The processes of control. *Br Med J*. 1957 Apr 6; 1(5022):779:86.

Burrell R, McGranahan N, Bartek J, Swanton C. The causes and consequence of genetic heterogeneity in cancer evolution. *Nature*. 2013 Sep 19; 501(7467):338-45.

Bussard K, Mutkus L, Stumpf K, Gomez-Manzano C, Marini F. Tumor-associated stromal cells as key contributors to the tumor microenvironment. *Breast Cancer Res*. 2016 Aug 11; 18(1):84.

Bykov V, Eriksson S, Bianchi J, Wiman K. Targeting mutant p53 for efficient cancer therapy. *Nat Rev Cancer*. 2018 Feb; 18(2):89-102.

Carey G. History of genetics. 2010. Disponível em: <https://bit.ly/2MztdnE>. [2020 Nov 10]

Cassetta L, Fragkogianni S, Sims A, Swierczak A, Forrester L, Zhang H. et al. Human tumor-associated macrophage and monocyte transcriptional landscapes reveal cancer-specific reprogramming, biomarkers, and therapeutic targets. *Cancer Cell*. 2019 Apr 15; 35(4):588-602.

Castells M, Thibault B, Delord J, Couderc B. Implication of tumor microenvironment in chemoresistance: tumor-associated stromal cells protect tumor cells from cell death. *Int J Mol Sci*. 2012. 13(8):9545-71.

Chandrakesan P, Yao J, Qu D, May R, Weygant N, Ge Y. et al. Dclk1, a tumor stem cell marker, regulates pro-survival signaling and self-renewal of intestinal tumor cells. *Mol Cancer*. 2017 Feb 1; 16(1):30.

Chatterjee N, Walker G. Mechanisms of DNA damage, repair, and mutagenesis. *Environ Mol Mutagen*. 2017 Jun; 58(5):235-63.

Cheng F, Jia P, Wang Q, Lin C, Li W, Zhao Z. Studying tumorigenesis through network evolution and somatic mutational perturbations in the cancer interactome. *Mol Biol Evol*. 2014 Aug; 31(8):2156-69.

Cingolani P, Patel V, Coon M, Nguyen T, Land S, Ruden D. Using drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift. *Front Genet*. 2012a Mar 15; 3:35.

Cingolani P, Platts A, Wang L, Coon M, Nguyen T, Wang L. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)*. 2012b; 6(2):80-92.

Coban B, Bergonzini C, Zweemer A, Danen E. Metastasis: crosstalk between tissue mechanics and tumour plasticity. *Br J Cancer*. 2021 Jan; 124(1):49-57.

Correa H. Li-Fraumeni syndrome. *J Pediatr Genet*. 2016 Jun; 5(2):84-88.

Cox D. Regression models and life-tables. *J R Stat Soc Series B Stat Methodol*. 1972; 34(2):187-220.

Crow J. Population genetics history: a personal view. *Annu Rev Genet*. 1987; 21:1-22.

Cullen S, Brunet M, Martin S. Granzymes in cancer and immunity. *Cell Death Differ.* 2010 Apr; 17(4):616-23.

Daly B, Olapade O. A Perfect storm: how tumor biology, genomics, and health care delivery patterns collide to create a racial survival disparity in breast cancer and proposed interventions for change. *CA Cancer J Clin.* 2015; 65(3):221-38.

Davies H, Glodzik D, Morganella S, Yates L, Staaf J, Zou X. et al. HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nat Med.* 2017 Apr; 23(4):517-25.

DeBerardinis R, Chandel, N. Fundamentals of cancer metabolism. *Sci Adv.* 2016 May 27; 2(5):e1600200.

DeBerardinis R, Lum J, Hatzivassiliou G, Thompson C. The biology of cancer: metabolic reprogramming fuels cell growth and proliferation. *Cell Metab.* 2008 Jan; 7(1):11-20.

Dobin A, Davis C, Schlesinger F, Drenkow J, Zaleski C, Jha S. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.* 2013 Jan 1; 29(1):15-21.

Dunn G, Bruce A, Ikeda H, Old L, Schreiber R. Cancer immunoediting: from immunosurveillance to tumor escape. *Nat Immunol.* 2002 Nov; 3(11):991-98.

Ehrlich P. Uber den jetzigen stand der karzinomforschung. *Ned Tijdschr Geneesk.* 1909; 5:273-290.

Elinav E, Nowarski R, Thaiss C, Hu B, Jin C, Flavell R. Inflammation-induced cancer: crosstalk between tumours, immune cells and microorganisms. *Nat Rev Cancer.* 2013 Nov; 13(11):759-71.

Eum H, Kwon M, Ryu D, Jo A, Chung W, Kim N. et al. Tumor-promoting macrophages prevail in malignant ascites of advanced gastric cancer. *Exp Mol Med* 2020 Dec; 52(12):1976-1988.

Fatima I, Singh A, Dhawan P. MASTL: a novel therapeutic target for cancer malignancy. *Cancer Med.* 2020 Sep; 9(17):6322-29.

Faubert B, Solmonson A, DeBerardinis R. Metabolic reprogramming and cancer progression. *Science.* 2020 Apr 10; 368(6487):eaaw5473.

Feng A, Zhu J, Sun J, Yang M, Neckenig M, Wang X. et al. CD16⁺ monocytes in breast cancer patients: expanded by monocyte chemoattractant protein-1 and may be useful for early diagnosis. *Clin Exp Immunol.* 2011 Apr; 164(1):57-65.

Ferlay J, Colombet M, Soerjomataram I, Mathers C, Parkin D, Piñeros M. et al. Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. *Int J Cancer.* 2019 Apr 15; 144(8):1941-53.

Field L, Love B, Deyarmin B, Hooke J, Shriver C, Ellsworth R. et al. Identification of differentially expressed genes in breast tumors From African American compared with caucasian women. *Cancer.* 2012 Mar 1; 118(5):1334-44.

Fishel R. Mismatch repair. *J Biol Chem.* 2015 Oct 30; 290(44):26395-26403.

Freitas V, Hilfenhaus G, Iruela-Arispe M. Metastasis of circulating tumor cells: speed matters. *metastasis of circulating tumor cells: speed matters. Dev Cell.* 2018 Apr 9; 45(1):3-5.

Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw.* 2010; 3(1):1-22.

Galon J, Bruni D. Tumor immunology and tumor evolution: intertwined histories. *Immunity.* 2020 Jan 14; 52(1):55-81.

Gayon J. From mendel to epigenetics: history of genetics. *C R Biol* 2016; 339(7-8):225-230.

Geiringer H. On the probability theory of linkage in Mendelian heredity. *Ann Math Statist.* 1944 Mar; 15(1):25-57.

Gentleman R, Carey V, Bates D, Bolstad B, Dettling M, Dudoit S. et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* 2004; 5(10):R80.

Giavarina D. Understanding bland altman analysis. *Biochem Med (Zagreb).* 2015 Jun 5; 25(2):141-51.

Gillespie JH. Population genetics: a concise guide. 2nd ed. Maryland: Johns Hopkins University Press; 2004a. Genetic variation; p.1-20.

Gillespie JH. Population genetics: a concise guide. 2nd ed. Maryland: Johns Hopkins University Press; 2004b. Genetic drift; p.21-58.

Goel M, Khanna P, Kishore J. Understanding survival analysis: Kaplan-Meier estimate. *Int J Ayurveda Res.* 2010 Oct; 1(4):274-78.

Goldman M, Craft B, Hastie M, Repečka K, McDade F, Kamath A. et al. Visualizing and interpreting cancer genomics data via the Xena platform. *Nat Biotechnol.* 2020 Jun; 38(6):675-78.

Gong J, Wang C, Lee P, Chu P, Fakih M. Response to PD-1 blockade in microsatellite stable metastatic colorectal cancer harboring a POLE Mutation. *J Natl Compr Canc Netw.* 2017 Feb; 15(2):142-47.

Greten F, Grivennikov S. Inflammation and cancer: triggers, mechanisms, and consequences. *Immunity.* 2019 Jul 16; 51(1):27-41.

Gridelli C, Rossi A, Carbone D, Guarize J, Karachaliou N, Mok T. et al. Non-small-cell lung cancer. *Nat Rev Dis Primers.* 2015 May 21; 1:15009.

Griffiths AJF, Wessler SR, Lewontin RC, Gerbart WM, Suzuki DT, Miller JH. An introduction to genetic analysis. In: Freeman WH, editor. Somatic versus germinal mutation. 7th ed. New York: 2000. Disponível em: <https://www.ncbi.nlm.nih.gov/books/NBK21894/>. [2020 Nov 15]

Gurdasani D, Barroso I, Zeggini E, Sandhu M. Genomics of disease risk in globally diverse populations. *Nat Rev Genet.* 2019 Sep; 20(9):520-35.

Hanahan D, Weinberg R. The hallmarks of cancer. *Cell.* 2000 Jan 7; 100(1):57-70.

Hanahan D, Weinberg R. Hallmarks of cancer: the next generation. *Cell.* 2011 Mar 4; 144(5):646-74.

Harrel Jr F, Lee K, Mark D. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med.* 1996 Feb 28; 15(4):361-87.

Hartl DL, Clark AG. Principles of population genetics. 4th ed. New York: OUP USA; 2006. Organization of Genetic Variation; p45-94.

Hatakeyama K, Ohshima K, Nagashima T, Ohnami S, Serizawa M, Shimoda Y. et al. Molecular profiling and sequential somatic mutations shift in hypermutator tumours harbouring POLE mutations. *Sci Rep.* 2018 Jun 7; 8(1):8700.

Herbst R, Morgensztern D, Boshoff C. The biology and management of non-small cell lung cancer. *Nature.* 2018 Jan 24; 553(7689):446-54.

Hermida D, Mortuza G, Pedersen A, Pozdnyakova I, Nguyen T, Maroto M. et al. Molecular basis of the mechanisms controlling MASTL. *Mol Cell Proteomics.* 2020 Feb; 19(2):326-43.

Hill W, Robertson A. Linkage disequilibrium in finite populations. *Theor Appl Genet.* 1968 Jun; 38(6):226-31.

Holsinger K, Weir B. Genetics in geographically structured populations: defining, estimating and interpreting F_{ST} . *Nat Rev Genet.* 2009 Sep; 10(9):639-50.

Huang Y, Yang L, Pei Y, Wang J, Wu H, Yuan J. et al. Overexpressed ACBD3 has prognostic value in human breast cancer and promotes the self-renewal potential of breast cancer cells by activating Wnt/beta-catenin signaling pathway. *Exp Cell Res.* 2018 Feb 1; 363(1):39-47.

Huo D, Hu H, Rhie S, Gamazon E, Cherniack A, Liu J. et al. Comparison of breast cancer molecular features and Survival by African and European Ancestry in The Cancer Genome Atlas. *JAMA Oncol.* 2017 Dec 1; 3(12):1654-62.

International Agency for Research on Cancer-IARC. Global cancer observatory. 2020. Disponível em: <https://bit.ly/3cEQXBA>. [2021 jan 10]

Izumi T, Mellon I. Base excision repair and nucleotide excision repair. In: Kovalchuk I, Kovalchuk O, editors. *Genome stability: from virus to human application*. New York: Academic Press; 2016. p.275-302.

Jesus B, Blasco M. Telomerase at the intersection of cancer and aging. *Trends Genet.* 2013 Sep; 29(9):513-20.

Kaseniit KE, Haque IS, Goldberg JD, Shulman LP, Muzzey D. Genetic ancestry analysis on >93,000 individual undergoing expanded carrier screening reveals limitations of ethnicity-based medical guidelines. *Genet Med.* 2020 Oct; 22(10):1694-1702.

Keenan T, Moy B, Mroz E, Ross K, Niemierko A, Rocco J. et al. Comparison of the genomic landscape between primary breast cancer in African American versus white women and the association of racial differences with tumor recurrence. *J Clin Oncol.* 2015 Nov 1; 33(31):3621-27.

Kim H, Nickoloff J, Wu Y, Williamson E, Sidhu G, Reinert B. et al. Endonuclease EEPD1 Is a gatekeeper for repair of stressed replication forks. *J Biol Chem.* 2017 Feb 17; 292(7):2795-2804.

Kim J, DeBerardinis R. Mechanisms and implications of metabolic heterogeneity in cancer. *Cell Metab.* 2019 Sep 3; 30(3):434-46.

Kolmogorov A. Sulla determinazione empirica di na legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuar.* 1933; 4:84-91.

Korn J, Kuruvilla F, McCarroll S, Wysoker A, Nemesh J, Cawley S. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet.* 2008 Oct; 40(10):1253-60.

Kováč L. Lamark and Darwin revisited. *EMBO Rep.* 2019 Mar 6; 20:e47922.

Krijgsman D, Hokland M, Kuppen P. The role of natural killer t cells in cancer - a phenotypical and functional approach. *Front Immunol.* 2018 Feb 27;9:367.

Kruskal W, Wallis W. Use of ranks in one-criterion variance analysis. *J Am Stat Assoc.* 1952 Dec; 47(260):583-621.

Kucera M, Isserlin R, Arkhangorodsky A, Bader G. A Cytoscape app for summarizing networks with semantic annotations. *F1000Res.* 2016 Jul 15; 5:1717.

Kunz M. Oncogenes in melanoma: An update. *Eur J Cell Biol.* 2014; 93(1-2):1-10.

Köster J, Rahmann S. Snakemake - a scalable bioinformatics workflow engine. *Bioinformatics.* 2012 Oct 1; 28(19):2520-22.

Lawrence M, Stojanov P, Polak P, Kryukov G, Cibulskis K, Sivachenko A. et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature.* 2013 Jul 11; 499(7457):214-218.

Lewontin R, Kojima K. The evolutionary dynamics of complex polymorphisms. *Evolution.* 1960 Dec; 14:458-72.

Li S, Martin A. Mismatch repair and colon cancer: mechanisms and therapies explored. *Trends Mol Med.* 2016 Apr; 22(4):274-89.

Li W, Zhang X, Wu F, Zhou Y, Bao Z, Li H. Gastric cancer-derived mesenchymal stromal cells trigger M2 macrophage polarization that promotes metastasis and EMT in gastric cancer. *Cell Death Dis.* 2019 Dec 4; 10(12):918.

Liu H, Wen T, Zhou Y, Fan X, Du T, Gao T. et al. DCLK1 plays a metastatic-promoting role in human breast cancer cells. *Biomed Res Int*. 2019 May 15; 2019:1061979.

Liu H, Zhu X, Shi W, Lin Y, Ma G, Tao G. et al. Genetic variants in XDH are associated with prognosis for gastric cancer in a Chinese population. *Gene*. 2018 Jul 15;663:196-202.

Liu X, Wu C, Li C, Boerwinkle E. dbNSFP v3.0: a one-stop database of functional predictions and annotations for human non-synonymous and splice site SNVs. *Hum Mutat*. 2016 Mar; 37(3):235-41.

Love M, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014; 15(12):550.

Lui F, Tuan B, Swenson S, Wong R. Ethnic disparities in gastric cancer incidence and survival in the USA: an updated analysis of 1992–2009 SEER Data. *Dig Dis Sci*. 2014 Dec; 59(12):3027-34.

Lynch M. Rate, molecular spectrum, and consequences of human mutation. *Proc Natl Acad Sci U S A*. 2010 Jan 19; 107(3):961-68.

Maples BK, Gravel S, Kenny EE, Bustamante CD. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am J Hum Genet*. 2013 Aug 8; 93(2):278-88.

Marchese S, Polo A, Ariano A, Velotto S, Constantini S, Severino L. Aflatoxin B1 and M1: biological properties and their involvement in cancer development. *Toxins (Basel)*. 2018 May 24; 10(6):214.

Marees A, Kluvier H, Stringer S, Vorspan F, Curis E, Marie-Claire C. et al. A tutorial on conducting genome-wide association studies: quality control and statistical analysis. *Int J Methods Psychiatr Res*. 2018 Jun; 27(2):e1608.

Martin D, Boersma B, Yi M, Reimers M, Howe T, Yfantis H. et al. Differences in the tumor microenvironment between African-American and European-American Breast Cancer Patients. *PLoS One*. 2009; 4(2):e4531.

Martincorena I, Campbell P. Somatic mutation in cancer and normal cells. *Science*. 2015 Sep 25; 349(6255):1483-89.

Martincorena I, Raine K, Gerstung M, Dawson K, Haase K, Loo P. et al. Universal patterns of selection in cancer and somatic tissues. *Cell*. 2017 Nov 16; 171(5):1029-41.

Mattia G, Puglisi R, Ascione B, Malorni W, Carè A, Matarrese P. Cell death-based treatments of melanoma: conventional treatments and new therapeutic strategies. *Cell Death Dis*. 2018 Jan 25; 9(2):112.

Mayakonda A, Lin D, Assenov Y, Plass C, Koeffler H. Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Res*. 2018 Nov; 28(11):1747-56.

Mayo O. A century of hardy-weinberg equilibrium. *Twin Res Hum Genet*. 2008 Jun; 11(3):249-56.

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(9):1297-1303.

McLean M, El-Omar E. Genetics of gastric cancer. *Nat Rev Gastroenterol Hepatol*. 2014 Nov; 11(11):664-74.

Mehnert J, Panda A, Zhong H, Hirshfield K, Damare S, Lane K. et al. Immune activation and response to pembrolizumab in POLE-mutant endometrial cancer. *J Clin Invest*. 2016 Jun 1; 126(6):2334-40.

Mendel G. Versuche Über Pflanzen-Hybride. *Verhandlungen des naturforschenden Vereines zu Brünn* 1866; 4:3-47.

Merchant S, Li L, Kim J. Racial and ethnic disparities in gastric cancer outcomes: more important than surgical technique? *World J Gastroenterol*. 2014 Sep 7; 20(33):11546-51.

Merico D, Isserlin R, Stueker O, Emili A, Bader G. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS One*. 2010 Nov 15; 5(11):e13984.

Ministério da Saúde. Instituto Nacional do Câncer José Gomes de Alencar. Estimativa 2020: Incidência de câncer no Brasil. INCA; 2020. Disponível em: <https://bit.ly/39KreFV>. [2021 Jan 10].

Misawa K, Kanazawa T, Imai A, Endo S, Mochizuki D, Fukushima H. et al. Prognostic value of type XXII and XXIV collagen mRNA expression in head and neck cancer patients. *Mol Clin Oncol*. 2014 Mar; 2(2):285-91.

Mohammadi Y, Tavangar S, Saidijam M, Amini R, Etemadi K, Dermani F. et al. DCLK1 plays an import role in colorectal cancer tumorigenesis through regulation of miR-200c. *Biomed Pharmacother*. 2018 Jul; 103:301-07.

Mootha V, Lindgren C, Eriksson K, Subramanian S, Sihag S, Lehar J. et al. PGC-1-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet*. 2003 Jul; 34(2):268-73.

Morris J, Apeltsin L, Newman A, Baumbach J, Wittkop T, Su G. et al. clusterMaker: a multi-algorithm clustering plugin for Cytoscape. *BMC Bioinformatics*. 2011 Nov 9; 12:436.

Mostafaei S, Kazemnejad A, Jamalkandi S, Amirhashchi S, Donnelly S, Armstrong M. et al. Identification of novel genes in human airway epithelial cells associated with chronic Obstructive Pulmonary Disease (COPD) using Machine-Based Learning Algorithms. *Sci Rep*. 2018 Oct 25; 8(1):15775.

Muenst S, Läubli H, Soysal S, Zippelius A, Tzankov A, Hoeller S. The immune system and cancer evasion strategies: therapeutic concepts. *J Intern Med*. 2016 Jun; 279(6):541-62.

Najafi M, Goradel N, Farhood B, Salehi E, Nashtaei M, Khanlarkhani N. et al. Macrophage polarity in cancer: a review. *J Cell Biochem*. 2019 Mar; 120(3):2756-65.

Nakanishi Y, Seno H, Fukuoka A, Ueo T, Yamaga Y, Maruno T. et al. Dclk1 distinguishes between tumor and normal stem cells in the intestine. *Nat Genet.* 2013 Jan; 45(1):98-103.

Nakasone E, Askautrud H, Kees T, Park J, Plaks V, Ewald A. et al. Imaging tumor-stroma interactions during chemotherapy reveals contributions of the microenvironment to resistance. *Cancer Cell.* 2012 Apr 17; 21(4):488-503.

Naslavsky MS, Sciliar MO, Yamamoto GL, Wang JYT, Zverinova S, Karp T. et al. Whole-genome sequencing of 1,171 elderly admixed individuals from the largest Latin American metropolis (São Paulo, Brazil). *bioRxiv.* 2020 Sep 16.

National Cancer Institute-NCI. mRNA analysis pipeline. Disponível em: <https://bit.ly/3azOaqy>. [2020 Nov 20]

Negrini S, Gorgoulis V, Halazonetis T. Genomic instability - an evolving hallmark of cancer. *Nat Rev Mol Cell Biol.* 2010 Mar; 11(3):220-28.

Newman A, Liu C, Green M, Gentles A, Feng W, Xu Y. et al. Robust enumeration of cell subsets from tissue expression profile. *Nat Methods.* 2015 May; 12(5):453-57.

Nordborg M, Tavaré S. Linkage disequilibrium: what history has to tell us. *Trends Genet.* 2002 Feb; 18(2):83-90.

Nédélec Y, Sanz J, Baharian G, Szpiech Z, Pacis A, Dumaine A. et al. Genetic ancestry and natural selection drive population differences in immune responses to pathogens. *Cell.* 2016 Oct 20; 167(3):657-669.e21.

Oesper L, Merico D, Isserlin R, Bader G. WordCloud: a Cytoscape plugin to create a visual semantic summary of networks. *Source Code Biol Med.* 2011 Apr 8; 6:7.

Ogutu J, Schulz-Streeck T, Piepho H. Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. *BMC Proc.* 2012 May 21; 6:S10.

Okasha S. Population Genetics. The stanford encyclopedia of philosophy. 2016. Disponível em: <https://stanford.io/2LB6PtW>. [2020 Nov 01]

Olkin I. Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling. Stanford: Stanford University Press; 1960.

Özdemir B, Dotto G. Racial differences in cancer susceptibility and survival: more than the color of the skin. Trends Cancer. 2017 Mar; 3(3):181-97.

Our World in Data. Number of deaths by cause. 2017. Disponível em: <https://bit.ly/3oRFfpS>. [2021 Jan 10].

Padhukasahasram B. Inferring ancestry from population genomic data and its applications. Front Genet. 2014 Jul 3; 5:204.

Paijens S, Vledder A, Bruyn M, Nijman H. Tumor-infiltrating lymphocytes in the immunotherapy era. Cell Mol Immunol. 2020 Nov 2.

Paludan S, Pradeu T, Masters S, Mogensen T. Constitutive immune mechanisms: mediators of host defense and immune regulation. Nat Rev Immunol. 2020 Aug 11.

Pannunzio N, Watanabe G, Lieber M. Nonhomologous DNA end-joining for repair of DNA double-strand breaks. J Biol Chem. 2018 Jul 6; 293(27):10512-23.

Parkin J, Cohen B. An overview of the immune system. Lancet. 2001 Jun 2; 357(9270):1777-89.

Persi E, Wolf Y, Horn D, Ruppin E, Demichelis F, Gatenby R. et al. Mutation-selection balance and compensatory mechanisms in tumour evolution. Nat Rev Genet. 2020 Nov 30.

Pontén F, Jirstrom K, Uhlen M. The Human Protein Atlas – a tool for pathology. J Pathol. 2008 Dec; 216(4):387-393.

Porrello A, Leslie P, Harrison E, Gorentla B, Kattula S, Ghosh S. et al. Factor XIIIa—expressing inflammatory monocytes promote lung squamous cancer through fibrin cross-linking. *Nat Commun.* 2018 May 18; 9(1):1988.

Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, Ruczinski I. et al. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet.* 2009 Jun; 5(6):e1000519.

Pritchard J, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics.* 2000 Jun; 155(2):945-959.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M, Bender D. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 2007 Set; 81(3):559-75.

Quan L, Gong Z, Yao S, Bandera E, Zirpoli G, Hwang H. et al. Cytokine and cytokine receptor genes of the adaptive immune response are differentially associated with breast cancer risk in American women of African and European ancestry. *Int J Cancer.* 2014 Mar 15; 134(6):1408-21.

Quinlan A, Hall I. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics.* 2010 Mar 15; 26(6):841-42.

Ranjha L, Howard S, Cejka P. Main steps in DNA double-strand break repair: an introduction to homologous recombination and related processes. *Chromosoma.* 2018 Jun; 127(2):187-214.

Reimand J, Isserlin R, Voisin V, Kucera M, Tannus-Lopes C, Rostamianfar A. et al. Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nat Protoc.* 2019 Feb; 14(2):482-517.

Rosales R, Drummond R, Valieris R, Dias-Neto E, Silva I. signeR: an empirical Bayesian approach to mutational signature discovery. *Bioinformatics.* 2017 Jan 1; 33(1):8-16.

Rosenberg N. Standardized Subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, Accounting for Atypical and Duplicated Samples and Pairs of Close Relatives. *Ann Hum Genet.* 2006 Nov; 70:841-47.

Rosenberg N, Ward L, Pritchard J. Informativeness of Genetic Markers for Inference of Ancestry. *Am J Hum Genet.* 2003 Dec; 73(6):1402-22.

Rosenberg N, Pritchard J, Weber J, Cann H, Kidd K, Zhivotovsky L. et al. Genetic structure of human populations. *Science.* 2002 Dec 20; 298(5602):2381-85.

Rosenthal R, McGranahan N, Herrero J, Taylor B, Swanton C. deconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.* 2016 Feb 22; 17:31.

Rotimi C, Jorde L. Ancestry and Disease in the Age of Genomic Medicine. *N Engl J Med.* 2010 Oct 14; 363(16):1551-58.

Rotow J, Bivona T. Understanding and targeting resistance mechanisms in NSCLC. *Nat Rev Cancer.* 2017 Oct 25; 17(11):637-58.

Royal CD, Novembre J, Fullerton SM, Goldstein DB, Long JC, Bamshad MJ. et al. Inferring Genetic Ancestry: Opportunities, Challenges, and Implications. *Am J Hum Genet.* 2010 May 14; 86(5): 661-73.

Sankararaman S, Sridhar S, Kimmel G, Halperin E. Estimating local ancestry in admixed populations. *Am J Hum Genet.* 2008 Feb; 82(2):290-303.

Satelli A, Li S. Vimentin in cancer and its potential as a molecular target for cancer therapy. *Cell Mol Life Sci.* 2011 Sep; 68(18):3033-46.

Schadendorf D, Akkooi A, Berking C, Griewank K, Gutzmer R, Hauschild A. et al. Melanoma. *Lancet.* 2018 Sep 15; 392(10151):971-84.

Schadendorf D, Fisher D, Garbe C, Gershenwald J, Grob J, Halpern A. et al. Melanoma. *Nat Rev Dis Primers*. 2015 Apr 23; 1:15003.

Schreiber D, Levy E, Schwartz D, Rineer J, Wong A, Rotman M. et al. Impact of race in predominantly African-American population of patients with low/intermediate risk prostate cancer undergoing radical prostatectomy within an equal access care institution. *Int Urol Nephrol*. 2014 Oct; 46(10):1941-46.

Shammas M. Telomeres, lifestyle, cancer, and aging. *Curr Opin Clin Nutr Metab Care*. 2011 Jan; 14(1):28-34.

Shankaran V, Ikeda H, Bruce A, White J, Swanson P, Old L. IFN γ and lymphocytes prevent primary tumour development and shape tumour immunogenicity. *Nature*. 2001 Apr 26; 410(6832):1107-11.

Shannon P, Markiel A, Ozier O, Baliga N, Wang J, Ramage D. et al. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res*. 2003 Nov; 13(11):2498-2504.

Shapiro S, Wilk M. An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*. 1965 Dec; 52:591-611.

Shaul M, Friedlender Z. Tumour-associated neutrophils in patients with cancer. *Nat Rev Clin Oncol*. 2019 Oct; 16(10):601-20.

Shen H, He M, Lin R, Zhan M, Xu S, Huang X. et al., PLEK2 promotes gallbladder cancer invasion and metastasis through EGFR/CCL2 pathway. *J Exp Clin Cancer Res*. 2019 Jun 10; 38(1):247.

Shen J, Peng Y, Wei L, Zhang W, Yang L, Lan L. et al. ARID1A Deficiency Impairs the DNA Damage Checkpoint and Sensitizes Cells to PARP Inhibitors. *Cancer Discov*. 2015 Jul; 5(7):752-67.

Siegel R, Miller K, Jemal A. Cancer Statistics, 2018. *CA Cancer J Clin*. 2018 Jan; 68(1):7-30.

Simon N, Friedman J, Hastie T, Tibshirani R. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *J Stat Softw.* 2011 Mar; 39(5):1-13.

Sondka Z, Bamford S, Cole C, Ward S, Dunham I, Forbes S. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat Rev Cancer.* 2018 Nov; 18(11):696-705.

Spratt D, Chan T, Waldron L, Speers C, Feng F, Ogunwobi O. et al. Racial/Ethnic Disparities in Genomic Sequencing. *JAMA Oncol.* 2016 Aug 1; 2(8):1070-74.

Subramanian A, Tamayo P, Mootha V, Mukherjee S, Ebert B, Gillette M. et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profile. *Proc Natl Acad Sci U S A.* 2005 Oct 25; 102(43):15545-50.

Swiecki M, Colonna M. The multifaceted biology of plasmacytoid dendritic cells. *Nat Rev Immunol.* 2015 Aug; 15(8):471-85.

Tang H, Peng J, Wang P, Risch NJ. Estimation of individual admixture: analytical and study design considerations. *Genet Epidemiol.* 2005 May; 28(4):289-301.

Tate J, Bamford S, Jubb H, Sondka Z, Beare D, Bindal N. et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* 2019 Jan 8; 47(D7):D9412-D947.

The 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 1; 526(7571):68-74.

The International HapMap Consortium. The International HapMap Project. *Nature.* 2003 Dec 18; 426(6968):789-96.

Thermo Fisher Scientific. Exploring cancer proliferative signaling pathways. Thermo Fisher Scientific; 2017. Disponível em: <https://bit.ly/3isoLD1>. [2021 Jan 10].

Therneau TM. survival: Survival Analysis. R package version 3.2-7. 2020. Disponível em: <https://bit.ly/3cLf8y6>. [2020 Nov 20].

The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 2019 Jan 8; 47(D1):D506-D515.

Tibshirani R. Regression Shrinkage and Selection via the Lasso. *J R Stat Soc Series B Stat Methodol.* 1996; 58(1):267-88.

Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via gap statistic. *J R Stat Soc Series B Stat Methodol.* 2002 Jan 6; 63:411-23.

Ton Q, Leino D, Mowery S, Bredemeier N, Lafontant P, Lubert A. et al. Collagen COL22A1 maintains vascular stability and mutations in COL22A1 are potentially associated with intracranial aneurysms. *Dis Model Mech.* 2018 Dec 12; 11(12):dmm033654.

Turner S, Armstrong L, Bradford Y, Carlson C, Crawford D, Crenshaw A. et al. Quality Control Procedures for Genome-Wide Association Studies. *Curr Protoc Hum Genet.* 2011 Jan; 68:1.19.1-1.19.18.

Uhlen M, Zhang C, Lee S, Sjöstedt E, Fagerberg L, Bidkhori G. A pathology atlas of the human transcriptome. *Science.* 2017 Aug 18; 357(6352):eaan2507.

Umansky V, Blattner C, Gebhardt C, Utikal J. The Role of Myeloid-Derived Suppressor Cells (MDSC) in Cancer Progression. *Vaccines (Basel).* 2016 Nov 3; 4(4):36.

Uppada S, Gowrikumar S, Ahmad R, Kumar B, Szeglin B, Chen X. et al. MASTL induces Colon Cancer progression and Chemoresistance by promoting Wnt/-catenin signaling. *Mol Cancer.* 2018 Aug 1; 17(1):111.

U.S. Department of Health and Human Services. Smoking Cessation. A Report of the Surgeon General. HHS. 2020. Disponível em: <https://bit.ly/3oV0mq5>. [2020 Nov 23]

Veglia F, Gabrilovich D. Dendritic cells in cancer: the role revisited. *Curr Opin Immunol*. 2017 Apr; 45:43-51.

Vineis P, Wild C. Global Cancer Patterns: causes and prevention. *Lancet*. 2014 Feb 8; 383(9916):549-57.

Voskoboinik I, Whisstock J, Trapani J. Perforin and granzymes: function, dysfunction and human pathology. *Nat Rev Immunol*. 2015 Jun; 15(6):388-400.

Waldmann P, Mészáros G, Gredler B, Fuerst C, Sölkner, J. Evaluation of the lasso and the elastic net in genome-wide association studies. *Front Genet*. 2013 Dec 4; 4:270.

Wan L, Pantel K, Kang Y. Tumor metastasis: moving new biological insights into the clinic. *Nat Med*. 2013 Nov; 19(11):1450-1464.

Warburg O. On respiratory impairment in cancer cells. *Science*. 1956 Aug 10; 124(3215):269-70.

Warburg O. On the Origin of Cancer Cells. *Science*. 1956 Feb 24; 123(3191):309-14.

Wathen M, Gautam Y, Ghandikota S, Rao M, Mersha T. LEI: A Novel Allele Frequency-Based Feature Selection Method for Multi-ancestry Admixed Populations. *Sci Rep*. 2019 Jul 31; 9(1):11103.

Weir B, Cockerham C. Estimating F-Statistics for Analysis of Population Structure. *Evolution*. 1984 Nov; 38(6):1358-70.

Weis S, Cheresh D. Tumor angiogenesis: molecular pathways and therapeutic targets. *Nat Med* 2011 Nov 7; 17(11):1359-1370.

Wellcome Sanger Institute. Mutational Signatures (v3.1). 2020. Disponível em: <https://bit.ly/3asp25i>. [2020 Nov 01]

Wexler B. Genetics and Genetic Engineering. Information Plus; 2009. The History of Genetics; p1-13.

Winchester AM. Genetics. Encyclopædia Britannica 2020 May 15. Disponível em: <https://bit.ly/3nX7keR>. [2020 Nov 10]

World Health Organization-WHO. Genomics and World Health: Report of the Advisory Committee on Health Research. WHO. 2002. Disponível em: <https://bit.ly/2YUIFyJ>. [2020 Nov 23]

World Health Organization-WHO. Cancer. Disponível em: <https://bit.ly/3pNsFJd>. [2020a Dec 01]

World Health Organization-WHO. WHO report on cancer: setting priorities, investing wisely and providing care for all. World Health Organization. Disponível em: <https://bit.ly/3bRqwIm>. [2020b Dec 01]

Wright S. The genetical structure of populations. Ann Eugen. 1949 Jan; 15:323-54.

Wright S. The interpretation of population structure by F-statistics with special regard to systems of mating. Evolution. 1965 Sep; 19(3):395-420.

Wright W, Shah S, Heyer W. Homologous recombination and the repair of DNA double-strand breaks. J Biol Chem. 2018 Jul 6; 293(27):10524-35.

Xie Y, Lynn B, Moir N, Cameron D, Figueroa J, Sims A. Breast cancer gene expression datasets do not reflect the disease at the population level. NPJ Breast Cancer. 2020a Aug 25; 6:39.

Xie Y, Su N, Yang J, Tan Q, Huang S, Jin M. et al. FGF/FGFR signaling in health and disease. Signal Transduct Target Ther. 2020b Sep 2; 5(1):181.

Yao Y, Gong Y, Zhao W, Han Z, Guo S, Liu H. et al. Comprehensive analysis of POLE and POLD1 Gene Variations identifies cancer patients potentially benefit from immunotherapy in Chinese population. Sci Rep. 2019 Oct 31; 9(1):15767.

Yao Z, Poto C, Mavodza G, Oliver E, Ressim H, Sherif Z. DNA Methylation Activates TP73 Expression in Hepatocellular Carcinoma and Gastrointestinal Cancer. *Sci Rep*. 2019 Dec 18; 9(1):19367.

Yatim K, Lakkis F. A Brief Journey through the Immune System. *Clin J Am Soc Nephrol*. 2015 Jul 7; 10(7):1274-81.

Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring tumour and stromal and immune cell admixture from expression data. *Nat Commun*. 2013; 4:2612.

Yuan J, Hu Z, Mahal B, Zhao S, Kensler K, Pi J. et al. Integrated Analysis of Genetic Ancestry and Genomic Alterations across Cancers. *Cancer Cell*. 2018 Oct 8; 34(4):549-560.e9.

Yue X, Qian Y, Gim B, Lee I. Acyl-CoA-Binding Domain-Containing 3 (ACBD3; PAP7; GCP60): a Multi-Functional Membrane Domain Organizer. *Int J Mol Sci*. 2019 Apr 24; 20(8):2028.

Zeng K, Wu Y, Wang C, Wang S, Sun H, Zou R. ASH2L is involved in promotion of endometrial cancer progression via upregulation of PAX2 transcription. *Cancer Sci*. 2020 Jun; 111(6):2062-77.

Zhang K, Zheng J, Bian G, Liu L, Xue Q, Liu F. et al. Polarized Macrophages Have Distinct Roles in the Differentiation and Migration of Embryonic Spinal-cord-derived Neural Stem Cells After Grafting to Injured Sites of Spinal Cord. *Mol Ther*. 2015 Jun; 23(6):1077-91.

Zhu Y, Wu J, Zhang C, Sun S, Zhang J, Liu W. et al. BRCA mutations and survival in breast cancer: an updated systematic review and meta-analysis. *Oncotarget*. 2016 Oct 25; 7(43):70113-27.

Zou H, Hastie T. Regularization and variable selection via elastic net. *J R Stat Soc Series B Stat Methodol*. 2005; 67(2):301-20.

Anexo 1 – Carta de aprovação do Comitê de Ética em Pesquisa-CEP



A.C. Camargo Cancer Center

Centro Integrado de Diagnóstico, Tratamento, Ensino e Pesquisa

**COMITÊ DE ÉTICA
EM PESQUISA - CEP**

São Paulo, 13 de fevereiro de 2019.

Ao

Dr. Kenneth John Gollob

Orientador: Dr. Israel Tojal da Silva

Aluno: Felipe de Azevedo Oliveira (Mestrado)

Ref.: Projeto de Pesquisa nº. 2463/17D

"IMPACTO DO FATOR RAÇA NA PROGRESSÃO E BIOLOGIA DO MELANOMA E CARCINOMA DE PULMÃO DE CÉLULAS NÃO PEQUENAS."

Os membros do Comitê de Ética em Pesquisa da Fundação Antonio Prudente – A.C. Camargo Cancer Center, em sua última reunião de 12/02/2019, após analisarem as respostas aos questionamentos realizados em reunião de 11/12/2018, tomaram conhecimento e aprovaram o seguinte documento:

- Solicitação de dispensa da submissão da documentação obrigatória e análise ética do projeto acima mencionado por se tratar de um projeto afiliado ao temático " *Um estudo prospectivo para elucidar redes imunorreguladoras sistêmicas e microambientes tumorais relacionadas à falência do tratamento com pembrolizumab em pacientes com melanoma em estágio IV e NSCLC.*", registrado neste CEP sob nº. 2463/17. O projeto afiliado em referência será Mestrado do aluno Felipe de Azevedo Oliveira sob orientação do Dr. Israel Tojal da Silva.

Atenciosamente,


Dra. Sandra Caires Serrano
1ª. Vice-Coordenadora do Comitê de Ética em Pesquisa

1/1