UNIVERSIDADE DE SÃO PAULO
Faculdade de Ciências Farmacêuticas
Programa de Pós-Graduação em Farmácia (Fisiopatologia e Toxicologia)
Área de Análises Clínicas

Desenvolvimento e validação de um *score* de desregulação das vias de reparo de DNA na predição do prognóstico em diferentes tipos de câncer

Jefferson Leandro Jimenez Restrepo

Tese apresentada à Faculdade de Ciências Farmacêuticas Departamento de Análises clínicas para obtenção do título de doutorado em Ciências.
Orientador: Prof. Dr. Helder I. Nakaya.

São Paulo
2023

UNIVERSIDADE DE SÃO PAULO
Faculdade de Ciências Farmacêuticas
Programa de Pós-Graduação em Farmácia (Fisiopatologia e Toxicologia)
Área de Análises Clínicas

Desenvolvimento e validação de um *score* de desregulação das vias de reparo de DNA na predição do prognóstico em diferentes tipos de câncer

Jefferson Leandro Jimenez Restrepo

Versão corrigida

Tese apresentada à Faculdade de Ciências Farmacêuticas Departamento de Análises clínicas para obtenção do título de doutorado em Ciências.
Orientador: Prof. Dr. Helder I. Nakaya.

São Paulo
2023

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletronico, para fins de estudo e pesquisa, desde que citada a fonte.

Ficha Catalográfica elaborada eletronicamente pelo autor, utilizando o programa desenvolvido pela Seção Técnica de Informática do ICMC/USP e adaptado para a Divisão de Biblioteca e Documentação do Conjunto das Químicas da USP

Bibliotecária responsável pela orientação de catalogação da publicação:
Marlene Aparecida Vieira - CRB - 8/5562

j61d

jimenez, jefferssson
Desenvolvimento e validação de um score de desregulação das vias de reparo de DNA na predição do prognóstico em diferentes tipos de câncer / jefferssson jimenez. - São Paulo, 2023.
64 p.

Tese (doutorado) - Faculdade de Ciências Farmacêuticas da Universidade de São Paulo. Departamento de Farmácia.
Orientador: Nakaya, Helder

1. Machine learning. 2. expressão gênica. 3. RNA-seq. 4. Mecanismo de reparo. 5. Sobrevida global. I. T. II. Nakaya, Helder, orientador.

Jefferson Leandro Jimenez Restrepo

Desenvolvimento e validação de um *score* de desregulação das vias de reparo de DNA na predição do prognóstico em diferentes tipos de câncer.

Comissão Julgadora
da
Dissertação apresentada à Faculdade de Ciências Farmacêuticas Departamento de Análises clínicas para obtenção do título de doutorado em Ciências.

Prof. Dr.
orientador/presidente

1o. Examinador

2o. Examinador

3o. Examinador

4o. Examinador

São Paulo, _____ de _____ de 2023.

AGRADECIMENTOS

Ao estimado Dr. Helder Nakaya, que nos anos de nossa convivência compartilhou seu conhecimento e sabedoria comigo, contribuindo grandemente para o meu crescimento científico e intelectual.

A estimada Dra. Anamaria Camargo, minha primeira orientadora, que nos anos do mestrado me mostrou o caminho certo para a ciência.

À minha amada família, que sempre me proporcionou atenção e apoio durante todo o processo de formação.

Ao Brasil, minha segunda pátria, tão magnífica e grandiosa, que abriu seus braços para me acolher. À Colômbia, minha terra natal, que me proporcionou uma educação pública de qualidade, moldando-me no ser que sou hoje. Com profunda gratidão, lhes agradeço imensamente. Muchisimas gracias Colombia!

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001

RESUMO

Jimenez, Jeffersson. **Desenvolvimento e validação de um *score* de desregulação das vias de reparo de DNA na predição do prognóstico em diferentes tipos de câncer**. 2023. 63f. Tese (Doutorado em Ciências) Faculdade de Ciências Farmacêuticas Departamento de Análises clínicas, Universidade de São Paulo, São Paulo, Brasil. 2023

As técnicas de aprendizado de máquina têm se destacado na identificação de padrões em dados de RNA-seq, particularmente no reconhecimento de padrões de expressão gênica associados à sobrevivência em casos de câncer. No entanto, um desafio persistente é a validação destes padrões em novos conjuntos de amostras, visto que a acurácia frequentemente diminui. A razão primária para tal é que muitos desses modelos se baseiam em estruturas matemáticas, sem um embasamento biológico. Para contornar esta limitação, focamos em processos biológicos naturalmente associados com a sobrevida dos pacientes com câncer. Estudos recentes apontam para uma associação entre genes de reparo de DNA e a sobrevivência global em vários tipos de câncer. Com base nisso, nosso trabalho inicial visou criar e validar um score que levasse em conta a expressão destes genes em 32 coortes de pacientes com tumores primários do banco de dados TCGA. Em consequência, a estratégia simples de conformação de 7 scores de desregulação, composto com 10 genes de reparo ao DNA mostrou uma associação com a sobrevida em 31 coortes diferentes, englobando mais de 10,000 pacientes. Além disso, foi possível validar um destes *score* em dados scRNA-seq de células tumorais de amostras de pacientes com câncer de ovário. Os modelos de aprendizado de máquina na análise de sobrevivência mostraram-se bem ajustados aos conjuntos dados de expressão onde foram gerados. Identificamos nos dois algoritmos *surv.blackboost* e *surv.ranger*, junto ao método de composição, as melhores estratégias para a análise de sobrevivência. Em síntese, os scores de desregulação utilizando genes envolvidos nos mecanismos de reparo ao DNA estão associados com a sobrevida global em diferentes tipos de câncer enquanto os métodos de aprendizados tem um overfitting ao conjunto de genes analisados.

Palavras-chave: RNA-seq, expressão gênica. Sobrevida, Machine learning. Câncer.

ABSTRACT

Jimenez, Jeffersson. **Development and validation of a DNA repair pathway deregulation score in predicting prognosis in different types of cancer.** 2023. 63f. Tese (Doutorado em Ciências) Faculdade de Ciências Farmacêuticas Departamento de Análises clínicas, Universidade de São Paulo, São Paulo, Brasil. 2023

Machine learning techniques have been prominent in identifying patterns in RNA-seq data, particularly in recognizing gene expression patterns associated with survival in cancer cases. However, a persistent challenge is the validation of these patterns in new sample sets, as accuracy often decreases. The primary reason for this is that many of these models rely on mathematical structures without a biological meaning. To address this limitation, we focused on biological processes naturally associated with the survival of cancer patients. Recent studies point to an association between DNA repair genes and overall survival in various types of cancer. Based on this, our initial work aimed to create and validate a score that took into account the expression of these genes in 32 cohorts of patients with primary tumors from the TCGA database. As a result, the simple strategy of conforming 7 deregulation scores, composed of 10 DNA repair genes, showed an association with survival in 31 different cohorts, encompassing over 10,000 patients. Furthermore, it was possible to validate one of these scores in scRNA-seq data from tumor cells from ovarian cancer patient samples. Machine learning models in survival analysis were well-suited to the expression data sets where they were generated. We identified in the two algorithms, surv.blackboost and surv.ranger, along with the composition method, the best strategies for survival analysis. In summary, the deregulation scores using genes involved in DNA repair mechanisms are associated with overall survival in different types of cancer, while learning methods are highly tailored to the algorithm and the set of genes analyzed.

Keywords: RNA-seq, gene expression, survival, machine learning, cancer.

Sumário

1. INTRODUÇÃO

1.1. Expressão gênica, sobrevida e aprendizado de máquina

A avaliação dos níveis de expressão gênica pode, sem dúvida, ser utilizada para prever a sobrevida global de pacientes com câncer (Chen et al., 2009). Diversos estudos apontaram a maior eficácia deste tipo de informação na associação com a sobrevivência, em comparação a dados clínicos e outros fatores prognósticos (Ishibashi et al., 2003; Pennathur et al., 2013). A superioridade dos dados de expressão gênica foi evidenciada ao proporcionar tratamentos individualizados adequados em pacientes com tumores malignos (Guo et al., 2006; Sato et al., 2005). A análise de perfis genéticos mostra-se útil para aprimorar a precisão na estimativa da sobrevida (Guo et al., 2006; Sato et al., 2005), sendo o uso de ferramentas preditivas o passo mais crucial dessa análise (Wang et al., 2012). Por exemplo, o Oncotype DX, um teste para certos tipos de câncer de mama. Este analisa a expressão de 21 genes em uma amostra tumoral para determinar o benefício da quimioterapia e a probabilidade de recorrência em 10 anos (Shreshtha et al., 2014; Dhillon et al., 2018; Tsukada et al., 2022). Na mesma linha, MammaPrint é outro teste para o câncer de mama que avalia 70 genes e classifica os pacientes em grupos de alto ou baixo risco de recorrência. (Cusumano et al., 2014)

Métodos de aprendizado de máquina tem a capacidade de desenvolver modelos de previsão de sobrevivência com base em dados de expressão gênica. A aprendizagem de máquina é um subcampo das ciências da computação baseado em algoritmos matemáticos que buscam a criação de modelos para reconhecimento de padrões, classificação ou predição. Aprendizagem supervisionada e não supervisionada são duas principais facetas da mecanização na aprendizagem de máquina que tem aplicações em biologia. A aprendizagem supervisionada é caracterizada pelo conhecimento a priori do desfecho das amostras analisadas, tais como óbito, infetado e outros. Por outro lado, na aprendizagem não supervisionada o algoritmo não tem conhecimento do desfecho e busca classificar as amostras somente pelos padrões de entrada. Alguns exemplos de técnicas de aprendizado de máquinas supervisionadas são as redes neurais artificiais (ANN), as Árvores de Decisão e o SVM (do inglês *Support vector machine*) (Guo et al., 2006; Kourou et al., 2014; Sato et al., 2005; Tarca et al., 2007).

Ao analisar a expressão gênica usando técnicas de aprendizado de máquina, durante a

primeira etapa de pré-processamento é necessário extrair e analisar dados de expressão gênica. Nesta etapa, as entradas são identificadas, caracterizadas por ser um conjunto de genes previamente selecionado, e em seguida, utilizando algoritmos de aprendizado de máquina, modelos preditivos são construídos e testados. Depois, este modelo, baseado em entradas (dados de expressão gênica) podem prever saídas (tempo de sobrevivência) em pacientes com câncer (Guo et al., 2006; Kourou et al., 2014; Sato et al., 2005; Tarca et al., 2007). Diferentes estudos mostraram consistentemente a poderosa capacidade de técnicas de aprendizagem de máquina em identificar padrões, processo das interações de expressão gênica e melhorar a precisão no diagnóstico do câncer, suscetibilidade e recorrência (Chen et al., 2009).

Modelos preditivos que foram criados usando esses recursos analíticos e com base em dados de expressão gênica, podem ajudar os médicos a otimizar a tomada de decisão clínica, fornecer tratamento individualizado, gerenciar e reduzir o custo a pacientes e nos sistemas de saúde (Bashiri et al., 2017). Não obstante, o uso de técnicas de aprendizagem de máquina não tem mostrado a mesma acurácia, sensibilidade e especificidade quando se utiliza a mesma assinatura ou *score* em conjuntos de amostras independentes. A grande dependência das assinaturas gênicas ou *scores* em relação às amostras utilizadas para a sua identificação e o insucesso na aplicação clínica dos resultados obtidos em conjuntos de amostras independentes, vêm sendo atribuídos a utilização de técnicas de aprendizado de máquina para detectar padrões de expressão gênica sem embasamento biológico na escolha dos genes que compõem as assinaturas dos *scores* (Domany, 2014).

Para contornar esta limitação, precisamos definir um embasamento biológico relacionado com a sobrevivência dos pacientes, para aumentar a acurácia das assinaturas desenvolvidas quando avaliadas em conjuntos de amostras independentes. Estudos recentes têm identificado vários genes de reparo associados com a sobrevivência global em câncer (Hosoya & Miyagawa, 2014). Portanto, a utilização de técnicas de aprendizado será baseada em entradas de dados de expressão gênica ou *scores* de genes envolvidos nos mecanismos de reparo.

1.2. Mecanismos de reparo de DNA

Para assegurar a homeostase celular, as células humanas possuem diferentes sistemas de reparo de danos causados na molécula de DNA. Esses sistemas são altamente interconectados

entre si, sendo capazes de reparar vários tipos de lesões que incluem as quebras de fita de dupla e simples do DNA, modificações das bases nitrogenadas e ligações cruzadas entre as duas fitas do DNA. As rupturas de fita simples (SBB, do inglês *Single-strand break*) ou reparo de bases são reparadas pelos sistemas de BER (do inglês *Base Excision Repair*), NER (do inglês *Nucleotide Excision Repair*) e MMR (do inglês *Mismatch Repair*). As rupturas de fita dupla (DSB, do inglês *Double-strand breaks*) são reparadas principalmente pelos sistemas de NHEJ (do inglês *Non-homologus End Joining*) e HR (do inglês *Homologus Recombination*). O reparo de bases modificadas e oxidadas envolve tanto a via MGMT (do inglês *O6-alkylguanine DNA alkyltransferase-O-6-metilguanina-DNA metiltransferase*) quanto a via de BER. O emparelhamento incorreto de bases e as ligações cruzadas na molécula de DNA ativam as vias MMR, NER e BER, que também estão envolvidas no reparo de SSB. As ligações cruzadas também podem ser reparadas pelos sistemas de TLS (do inglês Translesion synthesis), FA (do inglês *Fanconi anemia*), NER e HR (Hosoya & Miyagawa, 2014).

Alterações genéticas e epigenéticas em genes pertencentes aos diferentes sistemas de reparo do DNA são frequentemente encontradas em tumores. Mutações germinativas em genes de reparo conferem maior predisposição ao desenvolvimento de diferentes tipos de tumores, incluindo tumores colorretais, enquanto mutações somáticas e o silenciamento epigenético de genes de reparo promovem instabilidade cromossômica e uma maior frequência de mutações.

Evidências iniciais também sugerem que alterações nos diferentes sistemas de reparo possam estar associadas ao prognóstico do câncer. Tais alterações podem ser usadas como biomarcadores de prognóstico dos pacientes com câncer e utilizadas para definir a conduta terapêutica e também podem ser exploradas com propósito terapêutico (Hosoya & Miyagawa, 2014).

1.3. Scores para genes de reparo de DNA e sobrevida

Estudos recentes, com diversos tipos de tumores, incluindo ovário (Kang et al., 2012), mama (Pitroda et al., 2017), pulmão (Pitroda et al., 2014) , entre outros, têm mostrado a aplicabilidade desses escores na determinação do prognóstico da doença. Autores (Kang et al., 2012) desenvolveram um escore capaz de prever o prognóstico e a evolução da doença em pacientes com câncer de ovário, baseado no padrão de expressão de genes envolvidos no reparo de danos ao DNA mediados por compostos à base de platina. Para isso, os pesquisadores utilizaram os dados de expressão de 151 genes de diferentes vias de reparo do DNA em 511

tumores de ovário, disponibilizados pelo consórcio internacional TCGA (do inglês *The Cancer Genome Atlas*).

A partir da análise do padrão de expressão desses 151 genes, os autores conseguiram identificar 23 genes envolvidos no reparo de danos ao DNA e cujo padrão de expressão estava associado a uma maior ou menor sobrevida. Para cada paciente, foi criado um *score*, atribuindo o valor +1 quando o nível de expressão de um determinado gene associado a uma melhor sobrevida estava acima da média da expressão desse mesmo gene no conjunto de amostras, e o valor -1 quando o nível de expressão de um determinado gene associado a uma pior sobrevida estava acima da média da expressão desse gene no conjunto das amostras. O *score* de cada amostra foi calculado a partir da soma dos valores atribuídos ao nível de expressão dos diferentes genes analisados. Os pacientes foram então agrupados de acordo com o valor do escore em um grupo com baixo escore (valores entre 1 e 10) e outro grupo de alto escore (valores entre 11 e 20). Pacientes com baixo escore apresentaram pior sobrevida global e sobrevida livre de doença, sendo esses resultados validados em dois grupos independentes de pacientes.

Utilizando uma estratégia semelhante, os autores (Pitroda et al., 2014) criaram um *score* para determinar a eficiência de reparo de quebras de fita dupla na molécula de DNA em tumores de mama e de pulmão. O *score*, denominado *score* de proficiência de recombinação (RPS do inglês *Recombination proficiency score*), foi calculado com base nos níveis de expressão de quatro genes: Rif1, PARI, RAD51 e Ku80 envolvidos nos dois principais sistemas de reparo de quebras de fita dupla: o HR e o NHEJ. A expressão aumentada desses genes resulta em um valor baixo de RPS que, por sua vez, reflete a supressão da via HR e ativação da via NHEJ. Tumores com baixos RPS apresentaram uma maior taxa de mutagênese e menores taxas de sobrevida dos pacientes, indicando que a supressão da via de HR contribui para a instabilidade genômica e a progressão tumoral. Curiosamente, o efeito negativo relacionado a um baixo RPS foi neutralizado nos pacientes com câncer de pulmão que foram tratados com compostos à base de platina, sugerindo que a inibição da via de HR pode estar ligada a uma maior sensibilidade a esses tipos de agentes quimioterápicos e, portanto, o RPS pode ser usado para prever a resposta e orientar o tratamento (Pitroda et al., 2014) .

Nesse cenário, a utilidade do RPS na previsão de resposta à quimioterapia neoadjuvante foi recentemente avaliada em pacientes com câncer de mama (Pitroda et al., 2017). Neste estudo, foram incluídas 513 pacientes com câncer de mama de diferentes subtipos submetidas à quimioterapia neoadjuvante com antraciclinas. Pacientes com tumores apresentando valores

baixos de RPS tiveram uma taxa de resposta patológica completa ou presença de doença residual mínima duas vezes maior do que pacientes com tumores com valores altos de RPS. Em geral, tumores dos subtipos basal, HER2 positivos e luminal B exibiram valores menores de RPS e o valor preditivo do RPS foi mantido mesmo após o ajuste em relação às outras variáveis clínicas e patológicas e ao subtipo molecular (Pitroda et al., 2014, 2017).

Ainda em 2014, Kassambara e colaboradores (Kassambara et al., 2014; Pitroda et al., 2017) empregaram dados de expressão de 84 genes envolvidos em diversas vias de reparo do DNA em 206 amostras de pacientes com mieloma múltiplo, visando gerar um índice de desregulação associado a um prognóstico desfavorável. Ao analisar o padrão de expressão desses 84 genes, os pesquisadores conseguiram identificar 17 genes cuja expressão estava relacionada a um pior prognóstico e 5 genes cuja expressão estava vinculada a um prognóstico mais favorável. Para cada paciente, foi estabelecido um índice, adotando a mesma abordagem utilizada por Kang *et al. (Kang et al., 2012)*. Os pacientes foram divididos em dois grupos conforme o índice obtido: índice > 07 (24,8% dos pacientes) e índice ≤ 07 (75,2% dos pacientes). Aqueles com índice menor apresentaram menor sobrevida livre de doença e sobrevida global em comparação com pacientes com índices acima de 08.

Em suma, esses quatro estudos revelam o potencial dos índices de desregulação baseados na expressão diferencial e na relevância biológica de genes de reparo na previsão da progressão da doença em diversos tipos de câncer, surgindo como uma possível ferramenta para a previsão do diagnóstico em diferentes neoplasias.

Com base nessas evidências e no uso de métodos de aprendizado de máquina no prognóstico de doenças, o presente trabalho buscou utilizar os genes envolvidos nos mecanismos de reparo do DNA para desenvolver um índice utilizando aprendizado de máquina para prever a sobrevida de pacientes em diferentes coortes de câncer a partir de dados de RNA-seq disponíveis em repositórios públicos. Isso ampliará nossa compreensão sobre os genes de reparo na carcinogênese e como as alterações nesses genes podem ser utilizadas como biomarcadores prognósticos para pacientes com câncer.

2. Objetivos

2.1. Objetivos principais

Desenvolvimento de *scores* de desregulação das vias de reparo de DNA em diferentes tipos de câncer a partir de dados *RNA-seq* utilizando inteligência artificial a fim de identificar e predizer o prognóstico da doença.

Identificar os padrões de expressão global em diferentes agrupamentos de células tumorais a partir de dados *scRNA-seq* de pacientes com câncer e validar os *scores* de desregulação gene.

2.2. Objetivos específicos

a) Obter e processar os dados de *RNA-seq* (do inglês *RNA sequencing*) e *scRNA-seq* (do inglês *single cell RNA sequencing*) de diferentes tipos de tumor disponibilizados em bancos de dados públicos com dados de Sobrevivências Globais.

b) Utilizar métodos de aprendizagem de máquina supervisionada para desenvolver *scores* que permitam predizer o prognóstico dos pacientes utilizando genes envolvidos nos mecanismos de reparo ao DNA.

c) Validação dos *scores* de desregulação em diferentes coortes de tumores utilizando dados *RNA-seq*.

d) Validação dos *scores* de desregulação em diferentes coortes de tumores utilizando dados *scRNA-seq*.

3. Materiais e métodos

3.1. Pacientes, dados clínicos e de expressão.

Foram incluídos neste estudo pacientes maiores de 18 anos com diagnóstico de 32 tipos de tumores diferentes (Tabela 1), com dados de expressão gênica de RNA-seq pré processados e dados clínicos correspondentes: i) o tempo até o evento acontecer e o ii) tipo de evento: censura ou morte. Os dados clínicos e de expressão gênica pré-processados (expressão gênica) dos tumores primários de pacientes foram obtidos da base de dados de acesso público TCGA (https://portal.gdc.cancer.gov/) em 2021 utilizando a ferramenta TCGAbiolikns (Colaprico et al., 2015) (Tabela 2).

3.2. Levantamento de genes pertencentes aos diferentes sistemas de reparo dos danos ao DNA

O levantamento do conjunto de genes pertencentes aos diferentes sistemas de reparo de danos ao DNA foi realizado através da consulta em dois bancos de dados distintos: KEGG (do inglês *Kyoto Encyclopedia of Genes and Genomes*) (http://www.genome.jp/kegg/) e REPAIRtoire (http://repairtoire.genesilico.pl/). Após consulta em ambos bancos de dados foram identificados 285 genes pertencentes aos diferentes sistemas de reparo do DNA. Foram excluídos genes envolvidos na ubiquitinação de proteínas de reparo ou genes cuja função nas diferentes vias de reparo ainda não estava bem caracterizada (Tabela 3). Para determinar aos processos biológicos que pertence cada gene foi utilizado base de dados gene ontology Biological process (Carbon & Mungall, 2018)

Tabela 1. Tipos de tumores com dados de expressão gênica e dados clínicos analisados neste estudo.

CHOL	Colangiocarcinoma
COAD	Adenocarcinoma de cólon
DLBC	Neoplasma linfóide linfoma difuso de grandes células B
ESCA	Carcinoma esofágico
GBM	Glioblastoma multiforme
HNSC	Carcinoma escamoso de cabeça e pescoço
KICH	Cromofobo renal
KIRC	Carcinoma renal de células claras
KIRP	Carcinoma papilar renal de células
LGG	Glioma cerebral de grau inferior

LIHC	Carcinoma hepatocelular
LUAD	Adenocarcinoma de pulmão
LUSC	Carcinoma escamoso de pulmão

MESO	Mesotelioma
OV	Câncer de ovário
PAAD	Adenocarcinoma pancreático
PCPG	Feocromocitoma e Paraganglioma
PRAD	Adenocarcinoma de próstata
READ	Adenocarcinoma de reto
SARC	Sarcoma
SKCM	Melanoma cutâneo
STAD	Adenocarcinoma gástrico
TGCT	Tumores germinativos testiculares
THCA	Carcinoma da tireoide
THYM	Timoma
UCEC	Carcinoma endometrial do corpo uterino
UCS	Carcinosarcoma uterino
UVM	Melanoma uveal

Tabela 2. Características clínicas e número dos pacientes incluídos nas análises de sobrevida e métodos de aprendizado.

Tumor primário	Pacientes	Status vital (% vivos)	Média idade (max-min)	Gênero (% feminino)
ACC	92	61 (66.3)	47.16 (83-14)	60 (65.2)
BLCA	412	303 (73.5)	68.1 (90-34)	108 (26.2)
BRCA	1097	993 (90.5)	58.46 (90-26)	1085 (98.9)
CESC	307	247 (80.5)	48.27 (88-20)	307 (100)
CHOL	48	28 (58.3)	63.64 (82-29)	27 (56.2)
COAD	459	402 (87.6)	66.92 (90-31)	216 (47.1)
DLBC	48	43 (89.6)	56.27 (82-23)	26 (54.2)
ESCA	185	128 (69.2)	62.45 (90-27)	27 (14.6)
GBM	599	147 (24.5)	57.83 (89-10)	230 (38.4)
HNSC	528	358 (67.8)	60.91 (90-19)	142 (26.9)
KICH	113	103 (91.2)	51.2 (86-17)	51 (45.1)
KIRC	537	372 (69.3)	60.59 (90-26)	191 (35.6)
KIRP	291	259 (89)	61.49 (88-28)	77 (26.5)
LGG	515	423 (82.1)	42.94 (87-14)	230 (44.7)
LIHC	377	286 (75.9)	59.45 (90-16)	122 (32.4)
LUAD	522	395 (75.7)	65.33 (88-33)	280 (53.6)
LUSC	504	343 (68.1)	67.28 (90-39)	131 (26)
MESO	87	29 (33.3)	62.99 (81-28)	16 (18.4)
OV	587	282 (48)	59.74 (89-26)	587 (100)
PAAD	185	119 (64.3)	64.86 (88-35)	83 (44.9)
PCPG	179	173 (96.6)	47.33 (83-19)	101 (56.4)
PRAD	500	492 (98.4)	61.01 (78-41)	0 (0)
READ	171	161 (94.2)	64.46 (90-31)	78 (45.6)
SARC	261	185 (70.9)	60.87 (90-20)	142 (54.4)
SKCM	470	313 (66.6)	58.22 (90-15)	180 (38.3)

STAD	443	356 (80.4)	65.68 (90-30)	158 (35.7)
TGCT	134	131 (97.8)	31.99 (67-14)	0 (0)
THCA	507	493 (97.2)	47.26 (89-15)	371 (73.2)
THYM	124	118 (95.2)	58.16 (84-17)	60 (48.4)
UCEC	548	503 (91.8)	63.93 (90-31)	548 (100)

UCS	57	25 (43.9)	69.74 (90-51)	57 (100)
UVM	80	67 (83.8)	61.65 (86-22)	35 (43.8)

Tabela 3. Lista de genes pertencentes aos diferentes sistemas de reparo ao dano de DNA utilizados para o desenvolvimento do score e aprendizado de máquinas, composição e benchmark.

ABH2	ENDOV	HMCES	PARP2	RAD51	SPIDR	DMC1
ABRAXAS1	ERCC1	HUS1	PARP3	RAD51B	SPIDR	DNA2
ADPRT	ERCC2	KIAA1530	PARPBP	RAD51C	SPO11	DNPH1
ADPRTL2	ERCC3	KIAA179	PAXIP1	RAD51D	SPRTN	DNTT
ADPRTL3	ERCC4	LIG1	PCNA	RAD52	SSB1	DSS1
ALKBH2	ERCC5	LIG3	PDS5B	RAD52B	SWI5	DUT
ALKBH3	ERCC6	LIG4	PDS5B	RAD54B	SWS1	EBE2N
APE1	ERCC8	MAD2L2	PER1	RAD54B	SWSAP1	EME1
APEX1	EXO1	MBD4	PMS1	RAD54L	TDG	EME2
APEX2	EXO5	MDC1	PMS1	RAD54L	TDP1	H2AFX
APLF	FAAP100	METNAS E	PMS2	RAD6A	TDP2	H2AX
APTX	FAAP20	MGMT	PMS2L3	RAD6B	TFIIH	HCNP
ATM	FAAP24	MLH1	PMS2L3	RAD9	TOP3A	HEL308
ATR	FAM35A	MLH3	PMS2P3	RBBP8	TOPBP1	HELQ
ATRIP	FAN1	MMS19	PNKP	RDM1	TP53	HERC2
ATRX	FANCA	MMS2	POL4P	RECQ1	TP53BP1	HEX1
BARD1	FANCB	MMS4L	POLA1	RECQL	TREX1	HFM1
BLM	FANCC	MNAT1	POLB	RECQL4	TREX2	HLTF
BRCA1	FANCD1	MPG	POLD1	RECQL5	TTDA	NTHL1
BRCA2	FANCD1	MPLKIP	POLD2	REV1	TTDN1	NUDT1
BRCA2	FANCD2	MRE11A	POLD3	REV1L	TTRAP	NUDT15
BRIP1	FANCE	MSH2	POLD3	REV3L	UBC13	NUDT18
C19orf40	FANCF	MSH3	POLD4	REV7	UBE2A	OGG1
C1orf86	FANCG	MSH4	POLE	RIF1	UBE2B	PALB2
CAF1	FANCI	MSH5	POLE1	RMI1	UBE2N	PARG
CCNH	FANCJ	MSH6	POLE2	RNF168	UBE2T	PARK7
CDK7	FANCL	MTH1	POLE3	RNF4	UBE2V2	PARP1
CETN2	FANCM	MTH2	POLE4	RNF8	UNG	RAD1
CETN2	FANCN	MTH3	POLG	RPA1	USP1	RAD17
CHAF1A	FANCO	MTMR15	POLH	RPA2	UVSSA	RAD18
CHEK1	FANCP	MUS81	POLI	RPA3	WDR48	RAD23A
CHEK2	FANCT	MUS81	POLK	RPA4	WRN	RAD23B
CLK2	FEN1	MUTYH	POLL	RRM2B	XAB2	RAD23B
CSA	GEN1	MYH	POLM	SEM1	XLF	RAD24
CSB	GIYD1	NABP2	POLN	SETMAR	XPA	RAD30B
DCLRE1A	GIYD2	NBN	POLQ	SHFM1	XPB	RAD50

DCLRE1B	GTF2E2	NBS1	POLZ	SHLD1	XPC	PRKDC
DCLRE1C	GTF2H1	NEIL1	PRIMPOL	SHLD2	XPD	PRPF19
SLX1B	SMC6	XRCC3	XRCC9	PSQ4	GTF2H5	GTF2H2
XPF	SMUG1	SHPRH	ZSWIM7	PTIP	DEPC1	GTF2H3
SLX4	SNM1A	XRCC4	XPG	NHEJ1	DINB1	NEIL2
SMARCA3	SNM1B	XRCC5	XRCC1	NTH1	DDB1	NEIL3
SMC5	XRCC2	XRCC6	SLX1A	GTF2H4	DDB2	SHLD3

3.3. Dados de expressão gênica utilizados neste estudo

Para o desenvolvimento do *score* de desregulação baseado na expressão de genes pertencentes às vias de reparo de DNA e a identificação de métodos de aprendizado para predição do tempo de sobrevida dos pacientes com câncer, foram utilizados os dados de expressão gênica de *RNA-Seq* obtidos da base de dados TCGA (https://portal.gdc.cancer.gov/). Nas 32 coortes de pacientes incluídas neste estudo, para cada um dos pacientes analisados nesta etapa do projeto, foram geradas matrizes de expressão a partir de sequências de bibliotecas de *RNA-Seq*. Resumidamente, as sequências foram alinhadas a ferramenta STAR (Dobin et al., 2013) contra o genoma de referência hg38, e foram contadas, para cada gene, quantas reads alinhadas (*counts*) se sobrepuseram a éxons utilizando *HTseq-count* (Anders et al., 2015). As *counts* foram obtidas da base de dados TCGA e posteriormente normalizadas pelo tamanho total da biblioteca para obter os dados expressão gênica em valores de CPM (*Counts Per Million*), para cada tumor primário utilizando o pacote *edgeR* (Robinson et al., 2010) na linguagem de programação R.

3.4. Seleção dos conjuntos de genes

Para identificar os conjuntos de genes a avaliar o seu poder preditivo através dos métodos de aprendizado de máquinas, foram utilizados 3 abordagens baseados em 200 genes envolvidos no reparo de danos ao DNA que apresentarem valores de expressão ao longo das 32 coortes:

i) Selecionamos genes para cada coorte através de algoritmos de seleção de variáveis do pacote pec (Mogensen et al., 2012) na linguagem da programação R.

ii) Selecionamos genes sabidamente associados com a sobrevida dos pacientes analisados.

iii) Selecionamos todos os genes de reparo que tem valores de expressão maior do que 0 em todas as 32 coortes analisadas e os genes baseados em processos biológicos reportados na base de dados gene ontology associados ao reparo de DNA.

3.5. Desenvolvimento de um *score* de desregulação baseado nos genes envolvidos nas vias de reparo de DNA

Para calcular os *scores* de desregulação baseados em vias de reparo ao DNA, foram selecionados genes com valor de expressão maior que 0 CPM nas 32 matrizes de expressão dos tumores primários analisados. No total, 200 dos 285 genes envolvidos em diferentes mecanismos de reparo foram incluídos nas análises. Os valores de expressão em CPM de dois genes de reparo foram somados para compor o *score* de desregulação (Gene A + Gene B = *score* de desregulação Gene AB). No total, 19900 *score* de desregulação utilizando genes pertencentes ao reparo do DNA foram criadas para cada matriz de expressão dos tumores primário. Para categorizar os pacientes baseado no *score* de desregulação em cada matriz de expressão, foram utilizados os percentis 33.33% e 66.66% dos valores do score de todos os pacientes, para classificar os pacientes como: i) valores do *score* maior do que 66.66% (*score* alto), ii) valores menor ou igual do que 66.66 % ou maior ou igual do que 33.33 (*score* intermediário) e iii) valores menor do que 33.33% (*score* baixo). Um exemplo da categorização do *score* baseado no gene A e B e esquematizado na Tabela 4.

3.6. Análise de sobrevida dos pacientes com câncer após diagnóstico da doença

Para determinar se o *score* de desregulação das vias de reparo está associado com a sobrevida dos pacientes com câncer, foram utilizados os eventos de óbito e o número de dias dos pacientes desde diagnóstico inicial até i) a data do óbito, ou ii) a última data de seguimento clínico para os pacientes sobreviventes em um período de 5 anos. Foi calculada a probabilidade de sobrevida para cada *score* de desregulação comparando dois grupos dos pacientes com alto e intermediário/baixo valor do *score* pelo método de Kaplan-Meier (Kaplan & Meier, 1992). As curvas de probabilidade de sobrevida foram computadas e os Hazard ratios foram calculados usando a regressão de Cox.

Tabela 4. Exemplo do cálculo de *score* baseado no gene A e B utilizado os percentis.

Valor de expressão gene A	Valor de expressão gene B	score (gene A + gene B)	Percentil do score (%)	Categorização do score
1	2	3	0	Baixo
2	3	5	11,11	Baixo
3	4	7	22,22	Baixo
4	5	9	33,33	intermediários
5	6	11	44,44	intermediários
6	7	13	55,55	intermediários
7	8	15	66,66	Alto
8	9	17	77,77	Alto

9	10	19	88,88	Alto

3.7. Métodos de aprendizado de máquinas para a predição da sobrevida em pacientes com câncer.

Alguns dos *learners* (algoritmos de aprendizado) podem ter vários tipos de predição (Tabela 5). Neste estudo, algoritmos de predição de classificação contínua de risco (do inglês crank) foram desconsiderados, ao passo que para os *learners surv.blackboost*, *surv.cv_glmnet*, *surv.mboost*, *surv.parametric* foram considerados como método de predição *lp* (do inglês *linear predictor*) e para métodos *surv.akritas*, *surv.blackboost*, *surv.cforest*, *surv.kaplan*, *surv.mboost*, *surv.nelson*, *surv.penalized*, *surv.ranger*, *surv.rpart* o método predição *distr* (do inglês *distribution*). As semelhanças e diferenças entre esses dois métodos de predição são explicadas (Anexo 1, 2 e 3).

Tabela 5. Lista dos *Learners* para predizer o tempo de sobrevida dos pacientes com câncer utilizados en este estudo e o tipo de predição

Iniciais	Algoritmo	Tipo de predição
surv.cv_glmnet	*Cross-Validated GLM with Elastic Net Regularization Survival Learner (Friedman et al., 2010)*	crank, lp
surv.mboost	*Boosted Generalized Additive Survival Learner (González-Recio et al., 2010)*	distr, crank, lp
surv.parametric	*Survival Fully Parametric Learner (Kalbfleisch & Prentice, 2002)*	distr, lp, crank
surv.akritas	*Survival Akritas Estimator Learner (Akritas, 1994)*	crank, distr
surv.cforest	*Survival Conditional Random Forest Learner* (Hothorn & Zeileis, 2014)	distr, crank
surv.kaplan	*Kaplan-Meier Estimator Survival Learner (Hothorn & Zeileis, 2014; Kaplan & Meier, 1992)*	crank, distr
surv.nelson	*Survival Nelson-Aalen Estimator Learner* (Aalen, 1978)	crank, distr
surv.penalized	*L1 and L2 Penalized Estimation in GLMs Survival Learner (Goeman, 2010)*	distr, crank
surv.ranger	*Ranger Survival Learner* (Breiman, 1984; Goeman, 2010)	distr, crank
surv.rpart	*Rpart Survival Trees Survival Learner* (Breiman, 2017)	crank, distr
surv.blackboost	*Gradient Boosting with Regression Trees Survival Learner (Bühlmann & Yu, 2003)*	distr, crank, lp

Legenda: crank; classificação contínua de risco distr; distribuição de sobrevivência lp; predição linear

Para todos os *learners* foram utilizados os parâmetros sugeridos pelo pacote autores do pacote mlr3 (Lang et al., 2019) segundo o tipo de predição, e o método de reamostragem incluído neste estudo foi cinco vezes o cv (do inglês *cross validation resampling*), 70% para o treina e

30% para o grupo teste.

A capacidade do modelo preditivo, utilizando os algoritmos selecionados e genes envolvidos no reparos dos danos ao DNA, para fornecer corretamente uma classificação confiável dos tempos de sobrevivência com base nas pontuações de risco individuais, foi determinada através do índice de concordância ou índice C (Harrell's C-index) (Harrell et al., 1982; "On the Use of Harrell's C for Clinical Risk Prediction via Random Survival Forests," 2016)). Valores do índice C próximos a 0,5 indicam que as previsões do modelo preditivo não são melhores do que uma seleção randômica para determinar qual paciente terá maior tempo de sobrevida após o diagnóstico da doença. Valores próximos a 1 indicam que os modelos preditivos são bons para determinar qual dos pacientes terá maior tempo ou menor tempo ao óbito após diagnóstico da doença. A seleção dos métodos de aprendizado e o cálculo do Harrell's C-index para cada modelo predição utilizaram o pacote mlr3 (Lang et al., 2019) na linguagem de programação R.

3.8. Estratégia da composição dos métodos de aprendizado para a predição do tempo de sobrevida em paciente com câncer após diagnóstico da doença

Para aumentar o valor de prognóstico dos métodos de aprendizado de máquina foi utilizada a estratégia da composição, que junta a predição *lp* e a predição *distr* de dois diferentes tipos de *learners* como descrito nos pipelines do pacote mlr3 (Lang et al., 2019) na linguagem de programação R. Resumidamente, um *learner* do tipo de predição *distr* e um *learner* do tipo de predição *lp* são treinados com o mesmo grupo de treinamento e o mesmo conjunto de genes, depois a avaliação no grupo teste é utilizando uma única métrica do resultado de uma nova predição linear (*lp*) utilizando como a base a distribuição de sobrevivência. A capacidade do modelo preditivo da composição de dois *Learners* foi determinada através do índice de concordância ou índice C (Harrell's C-index) (Harrell et al., 1982; "On the Use of Harrell's C for Clinical Risk Prediction via Random Survival Forests," 2016). O método de composição utilizou a função *compose_distr* e cálculo do Harrell's C-index para cada modelo preditivo de composição utilizou o pacote mlr3 (Lang et al., 2019) na linguagem de programação R.

3.9. Comparação dos genes de reparo na predição de sobrevida dos pacientes com câncer através com outras vias envolvidas em carcinogênese

Para comparar o desempenho do gene de reparo na predição da sobrevida, foi utilizando o genes de apoptose GO:0006915), cell cycle (GO:0007049) do *gene ontology* e o

transcriptoma Resumidamente, foram selecionados o método de composição *surv.ranger* e *surv.balckboost* mensuráveis com o do índice de concordância ou índice C (Harrell et al., 1982; “On the Use of Harrell’s C for Clinical Risk Prediction via Random Survival Forests,” 2016), e incluídos o mesmo tipo de reamostragem e conjuntos de genes. Para cada *via* foi calculado o valor índice C (Harrell’s C-index) (Harrell et al., 1982; “On the Use of Harrell’s C for Clinical Risk Prediction via Random Survival Forests,” 2016) por separado (Lang et al., 2019) utilizando três vezes a reamostragem cv do pacote mlr3 (Lang et al., 2019) na linguagem de programação R.

3.10. Pacientes, dados clínicos e de expressão em dados *scRNA-seq*

Foram incluídos neste estudo pacientes maiores de 18 anos com diagnóstico de câncer de ovário, com dados de expressão gênica de sc*RNA-seq* pré-processados e dados clínicos correspondentes: i) o tempo até o evento acontecer e o ii) tipo de evento: censura ou morte. Os dados clínicos e de expressão gênica pré-processados (expressão gênica) dos tumores primários de pacientes foram obtidos da base de dados de acesso público TISCH2 (*Tumor Immune Single-cell Hub 2*) (*http://tisch.comp-genomics.org*). Para este estudo, foram incluídos pacientes com dados de sobrevida. Apenas 9 de 24 pacientes com câncer de ovário pertencentes ao GEO (do inglês *Gene Expression Onmibus*) (*https://www.ncbi.nlm.nih.gov/gds*) GSE158722 apresentaram dados clínicos e de expressão (Tabela 6) .

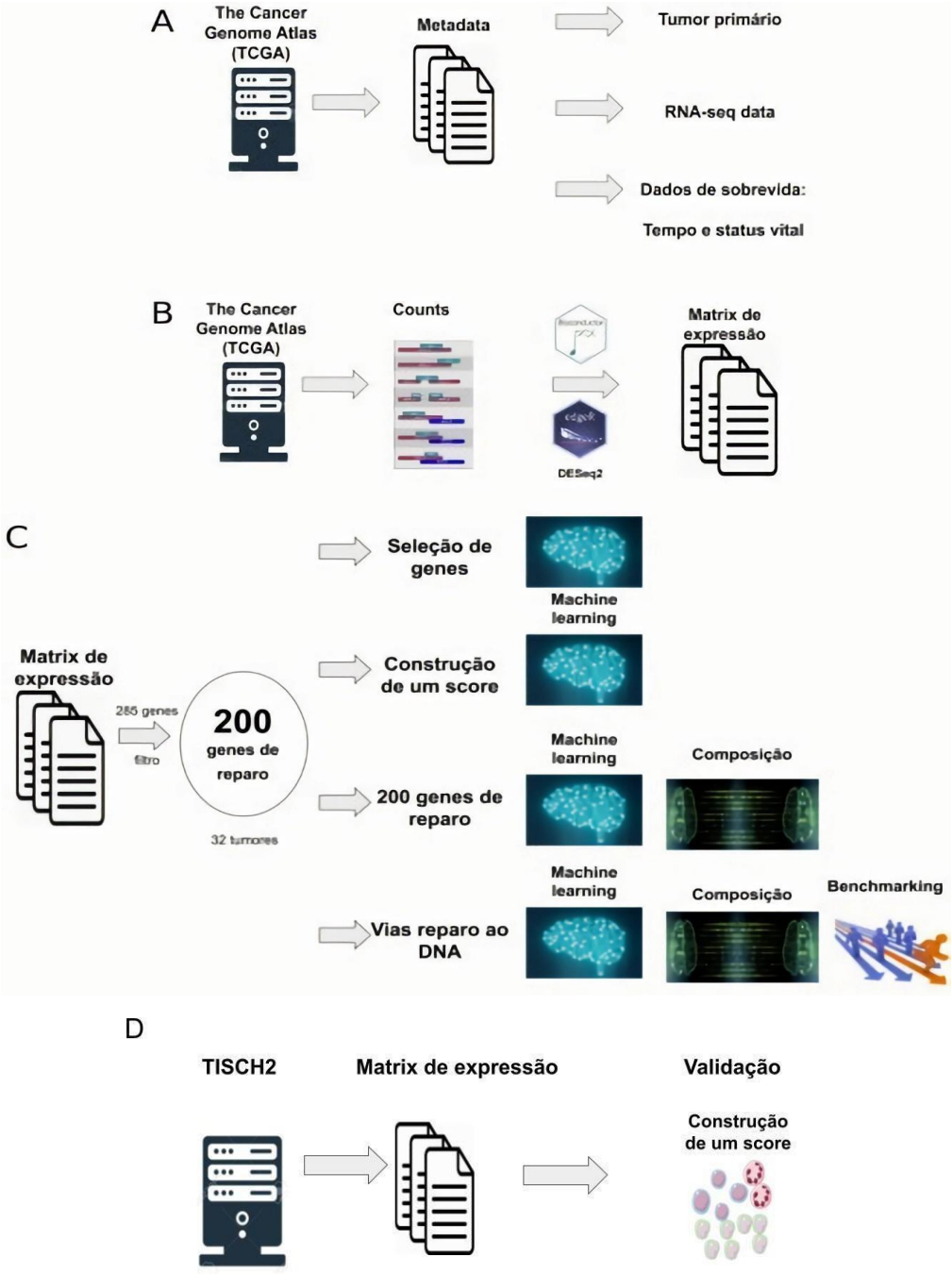


Figura 1. Representação da estratégia utilizada para escolha de genes e desenvolvimento de *scores* de desregulação das vias de reparo de DNA em diferentes tipos de câncer a partir de dados *RNA-seq*, utilizando inteligência artificial a fim de identificar e predizer o prognóstico da doença. A) Metadatos. Critérios de inclusão dos pacientes neste estudo; B) Dados de expressão. Obtenção das matrizes de expressão para cada tumor/coorte; C) Genes de reparo e métodos de aprendizado. Métodos de aprendizado para identificar os genes de reparo para predizer o tempo de vida após o diagnóstico da doença: i) seleção de genes por algoritmos de seleção, ii) construção de um *score* baseado nos genes de reparo, iii) utilizando todos os genes expressos em todas as coortes analisadas e iv) genes por vias de repaso segundo o *Gene Ontology*. D) Validação do *score* utilizando dados de single cell RNA-seq em pacientes com

câncer de ovário.

Tabela 6. Lista de pacientes com dados clínicos e de expressão gênica do tumor.

Paciente	Sobrevivência Global (Cirurgia Primária até a Morte, Dias)
Paciente 1	1700
Paciente 2	1049
Paciente 3	929
Paciente 4	2011
Paciente 5	448
Paciente 6	789
Paciente 7	1430
Paciente 8	966
Paciente 9	1131

Para Identificar os padrões de expressão global em diferentes agrupamentos de células tumorais baseados no padrão de expressão genes de reparo em dados *scRNA-seq* e validar a capacidade preditiva dos *scores* para determinar sobrevida dos pacientes com câncer, foram utilizados os dados de expressão gênica de sc*RNA-Seq* obtidos da base de dados de acesso público TISCH2. Na coorte de pacientes incluídas neste estudo, para cada um dos pacientes analisados nesta etapa do projeto, as sequências foram alinhadas contra o genoma de referência hg38, e foram contadas, para cada gene, quantas reads alinhadas (*counts*) se sobrepuseram a éxons utilizando *ICELL8 CellSelect software* (https://www.takarabio.com/) ou a função *cellranger count* de 10X *genomics* (www.10xgenomics.com). As *counts* foram obtidas da base de dados TISCH2 e posteriormente normalizadas foram identificados os clusters baseados no transcriptoma ou genes de reparo utilizando o pacote Seurat (Butler et al., 2018) na linguagem de programação R. Para classificar as células do fibroblasto, foram utilizados os marcadores ACTA1, FAP, THY1, FGF2. Já para monocitos e macrófagos, utilizou-se CD68, F4, F80, NOS2, CD206, CD14, CD16, CCR2, CD115 e CSF1. Para as células CD8, foram usados CD2 e CD3. Para as células tumorais de ovário, foram utilizados EpCAM, CA125, WT1 e PAX8.

4. Resultados

4.1. Seleção de genes para predição do prognóstico de pacientes com câncer após diagnóstico da doença

Para o desenvolvimento de *scores* de desregulação das vias de reparo de DNA em diferentes tipos de câncer a partir de dados *RNA-seq*, utilizando métodos de aprendizado, para predizer o prognóstico da doença é preciso definir conjuntos de genes. Para tanto, como esquematizado na figura 1C, utilizamos 3 abordagens diferentes utilizando genes envolvidos no reparo de danos ao DNA.

4.1.1. Seleção de genes através de algoritmos de seleção de variáveis

Para a seleção inicial de genes que permitam predizer o prognóstico de pacientes com câncer, utilizamos 200 genes envolvidos no reparo de dano ao DNA (Tabela 3) com valores de expressão maior do que 0 CPM em todas as coortes analisadas através de algoritmo de seleção do modelo de regressão de Cox para dados de sobrevivência do pacote pec (Mogensen et al., 2012). Este algoritmo selecionou genes para análises de sobrevida em apenas 19 das 32 coortes. Além disso, uns altos números de genes foram selecionados por coorte. Mais de 50 variáveis foram selecionadas utilizando o algoritmo, exceto em UVM e UCS (Figura 2).

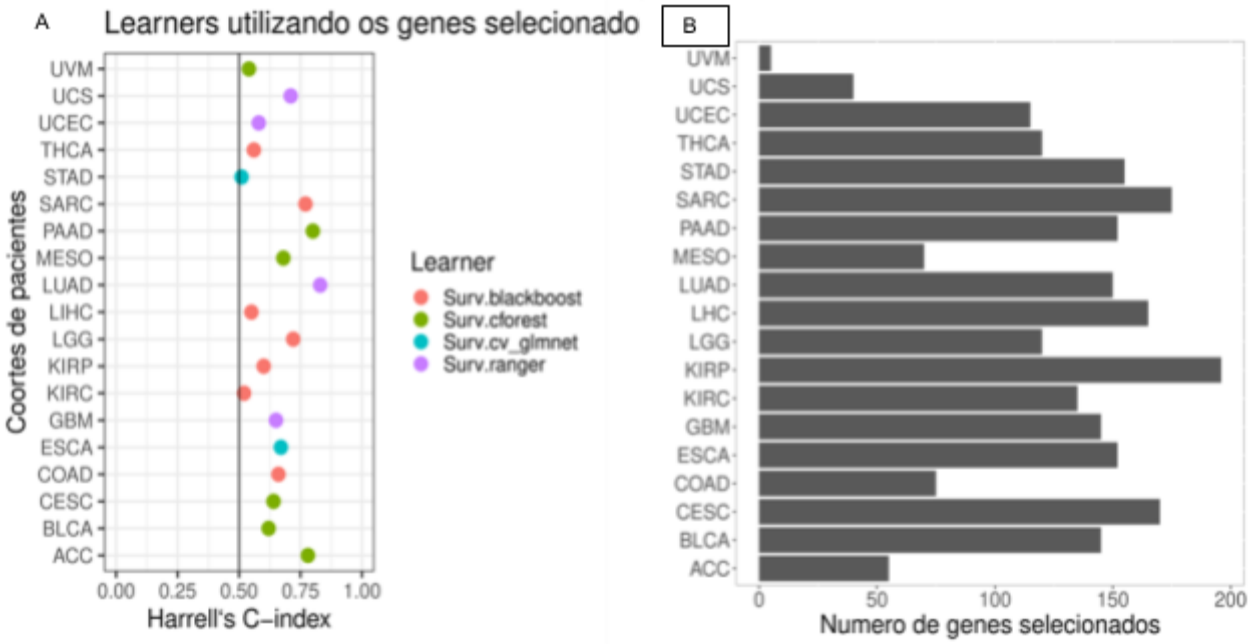


Figura 2. Seleção de genes através de um algoritmo. A) Valor preditivo dos *Learners* utilizando os genes selecionados. B) Número de genes selecionados pelo algoritmo de seleção de variáveis por coorte de pacientes.

Para identificar o valor preditivo dos genes selecionados através do algoritmo de seleção de variáveis nas 19 das 32 coortes , foram utilizados 11 diferentes *learners* (Tabela 5) com os parâmetros sugeridos pelos autores e com o método de reamostragem cv. Apenas 4 (*surv.blackboost, surv.ranger, surv.cforest, surv.glmnet*) dos 11 *learners* utilizados apresentaram valores Harrell's C-index superiores a 0.5. Poucas coortes de pacientes apresentarem altos valores preditivos. Quatro (SARC, ACC, PAAD, LUAD) das 19 tiveram valores preditivos maiores a 0.75. Os algoritmos *surv.blackboost, surv.cforest* foram os learners como melhor desempenho na predição em 12 das 19 coortes enquanto o *Surv.ranger* apresentou o maior valor índice C de todos os algoritmos (Figura 2A).

4.1.2. Seleção de genes sabidamente associados com a sobrevida dos pacientes e desenvolvimento de um *score* de desregulação da expressão de genes pertencentes às vias de reparo de DNA

O primeiro passo para o desenvolvimento de um *score* de desregulação da expressão de genes pertencentes às de vias reparo de DNA capaz de predizer o prognóstico de pacientes com

câncer consistiu na seleção dos genes para compor o *score*. Para tanto, genes associados (com um valor $p< 0.05$) ao reparo do DNA foram selecionados nas 32 coortes de pacientes com diferentes tipos de câncer (Tabela 1). Em total, 200 dos 285 (Tabela 2) genes avaliados apresentaram uma expressão maior do que 0 CPM em todas as coortes incluídas neste estudo. Como descrito em detalhes na seção 1.4, os valores de expressão em CPM de dois genes associados ao reparo de DNA foram somados para compor o *score* de desregulação (Gene A + GeneB = score de desregulação GeneAB). Como resultado destas permutações, utilizando os 200 genes foi possível construir 19,900 *scores* de desregulação de genes pertencentes ao reparo do DNA em cada coorte de pacientes. Posteriormente cada paciente foi classificado como *score* alto, intermediário ou baixo baseado no percentil 33.33% e 66.66% do para cada um dos 19900 *scores* em cada coorte.

Associações com a sobrevida utilizando a classificação do *score* no grupo de pacientes com alto *score* e no grupo de pacientes com intermediário/baixo *score* foram avaliadas através do teste de Log-rank. Com essa abordagem, identificamos 7 *scores* de genes do reparo do DNA associados com a sobrevida nas 32 coortes de pacientes analisados (Figura 3). Dentro desses 7 *scores* identificados, foram compostos por 10 dos 200 genes utilizados nesta análise. Dos 10 genes, 5 pertencem ao sistema recombinação homóloga ou Anemia de Falconi de reparo das quebras de fitas (BRCA1, BRCA2, PDS5B, RAD1, RAD52), 3 genes envolvidos nas DNA helicases e DNA polimerases que participam em diferentes vias envolvidas no reparo ao DNA (RECQL, POLB, POLD3), o restante pertence aos sistemas BER e NER de reparo de quebras de fitas simples (LIG3, RPA1). Alguns dos *scores* foram associados à sobrevida dos pacientes em mais de 7 coortes diferentes. Por exemplo, o *score* composto pelo gene POLB estava associado com os pacientes em 10 das 32 coortes de pacientes. Surpreendentemente, o gene RAD52 compôs 4 dos 7 escores. Esse gene e seus respectivos pares RECQL, POLB, RAD1 e LIG3, tinham a capacidade de prever a sobrevida dos pacientes em mais de 20 coortes (Figure 3).

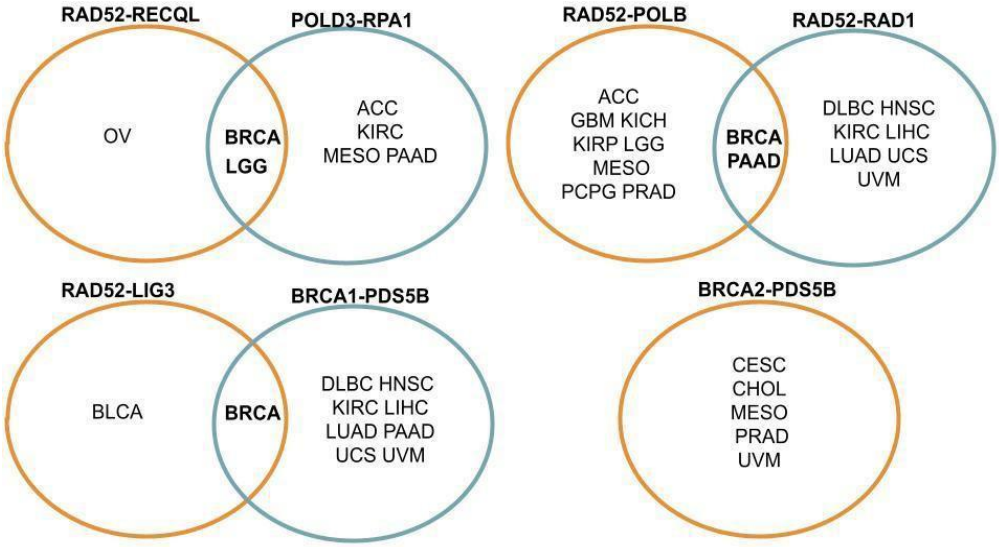


Figura 3. Score de desregulação das vias de reparo ao DNA associados com o tempo da sobrevida de pacientes com câncer após diagnóstico da doença nas 32 de coortes de pacientes analisadas.

Dentro desses *scores*, 6 *scores* com valores intermediários\baixos mostraram uma associação com o aumento das probabilidades de sobreviver após o diagnóstico da doença (Figura 4A-4E). Já os altos valores do score POLD3-RPA1 mostraram favorecer a sobrevida nos pacientes da coorte KIRC (Figura 4E).

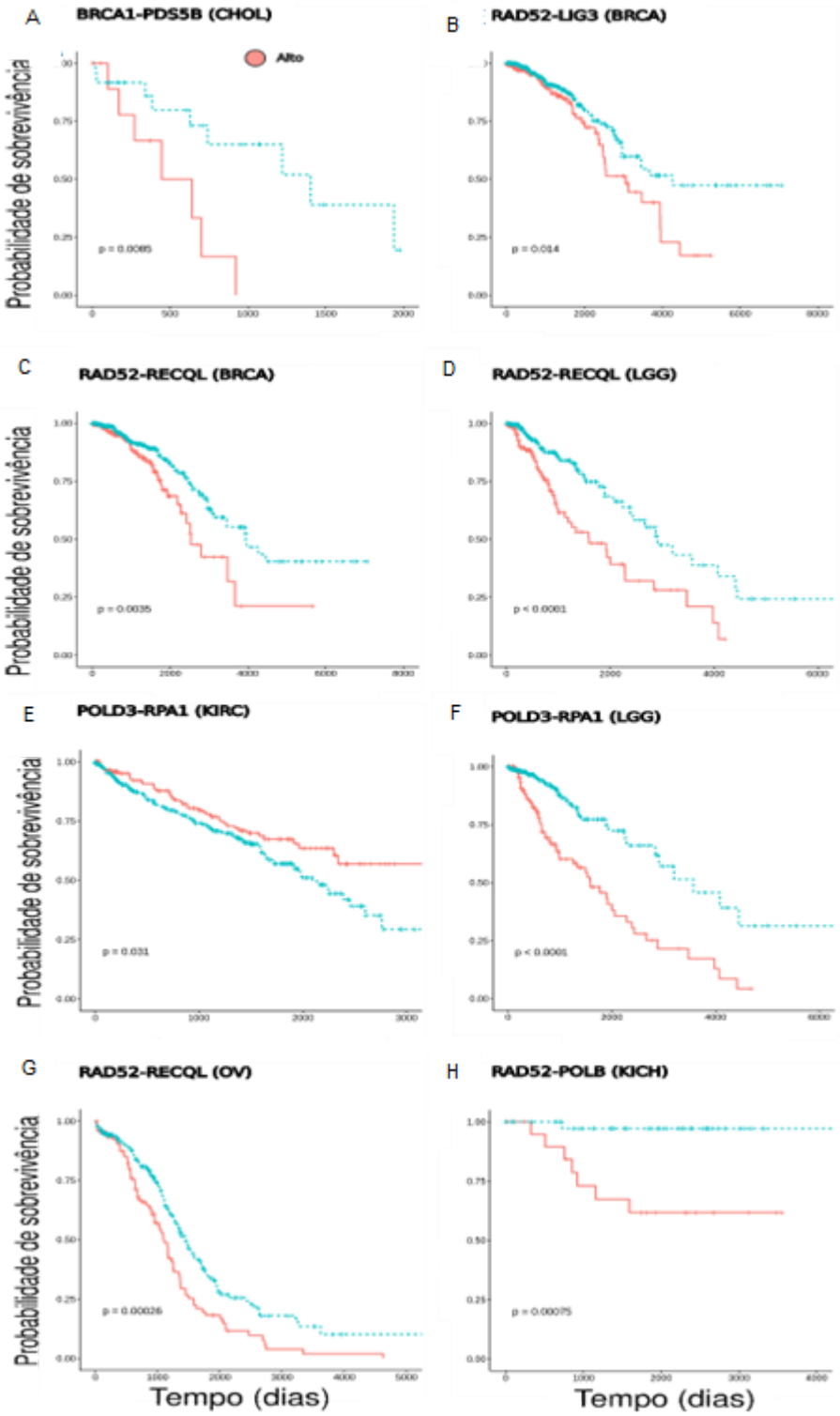


Figura 4. Avaliação dos *scores* na sobrevida das pacientes. A-E) Cinco *scores* dos 31 *scores* que apresentam valores intermédios/baixos dos scores associados com um aumento da sobrevida e F) o único score que apresentou valores altos em favor do aumento da sobrevida dos pacientes avaliados na coorte KIRC.

Para identificar o valor preditivo destes genes que compõem o *score* de desregulação de genes de reparo ao DNA, que foram associados significativamente ($p< 0.05$) com a sobrevida dos pacientes, utilizamos 11 diferentes *learners* (Tabela 5) com os parâmetros sugeridos pelos autores e com o método de reamostragem cv. Os pares de genes foram avaliados com o valor CPM em vez da categoria do *score* (alto, intermediário, baixo). Apenas 6 dos 32 *scores* das coortes apresentaram 2 ou mais *learners* com valores Harrell's C-index superior a 0.5. Apenas o *score* composto pelos genes RAD52 e POLB na coorte KICH apresentou valores do Harrell's C-index superiores a 0.75 (Figura 5A). Também avaliamos o valor preditivo dos genes pertencentes a 6 *scores* que não estão associados com a sobrevida dos pacientes ($p>0.98$). Surpreendentemente, alguns dos pares de genes apresentarem *learners* com valor preditivo superior a 0.5 (Figura 5B).

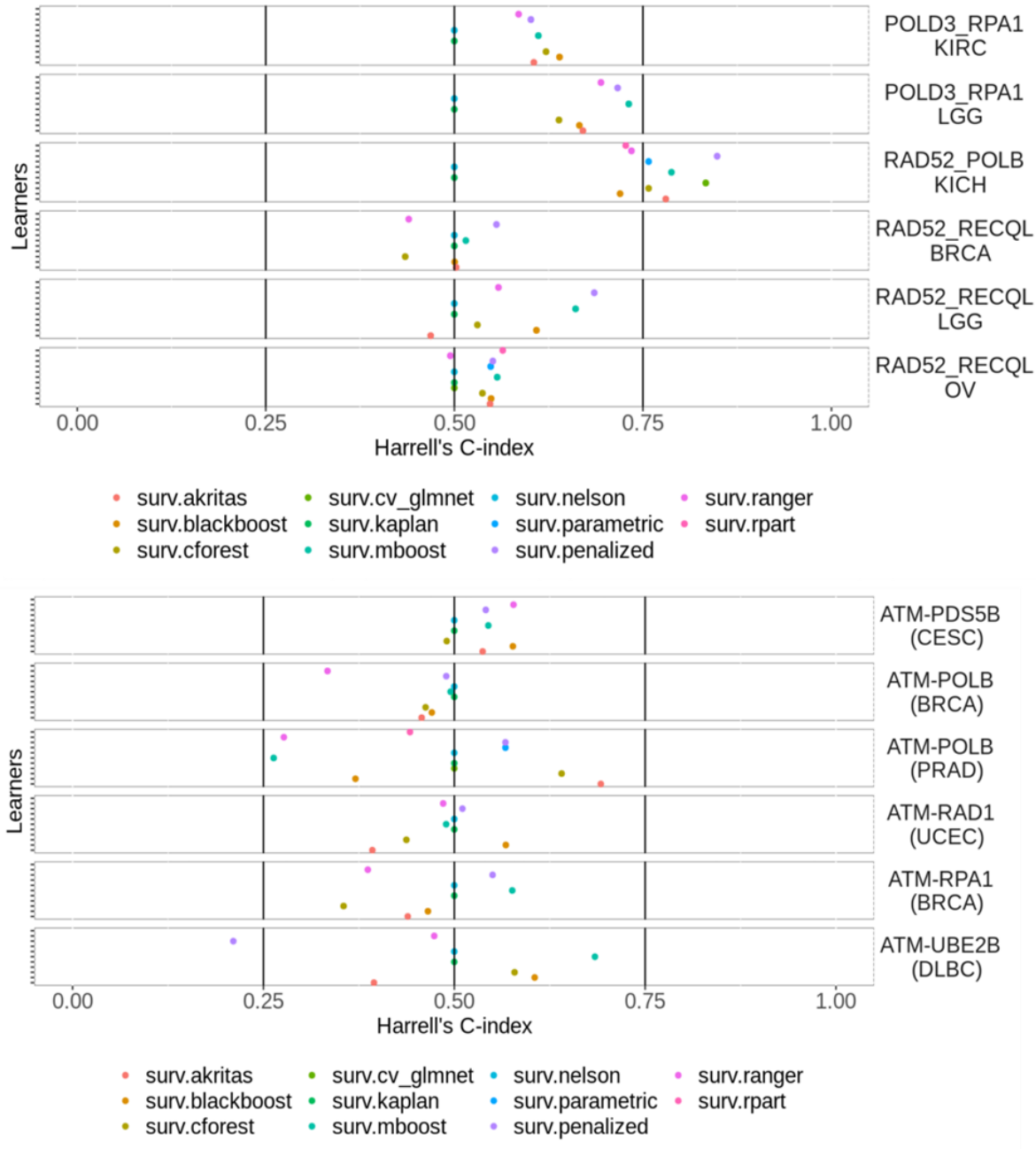


Figura 5. *Scores* associados com a sobrevida e *learners*. Painel superior: Avaliação do valor preditivo dos *learners* utilizando genes de reparo que compõe os *scores* sabidamente associados com a sobrevida. Painel inferior: Avaliação do valor preditivo dos *learners* utilizando genes reparo que não tem associação com o tempo de vida dos pacientes com câncer após diagnóstico da doença.

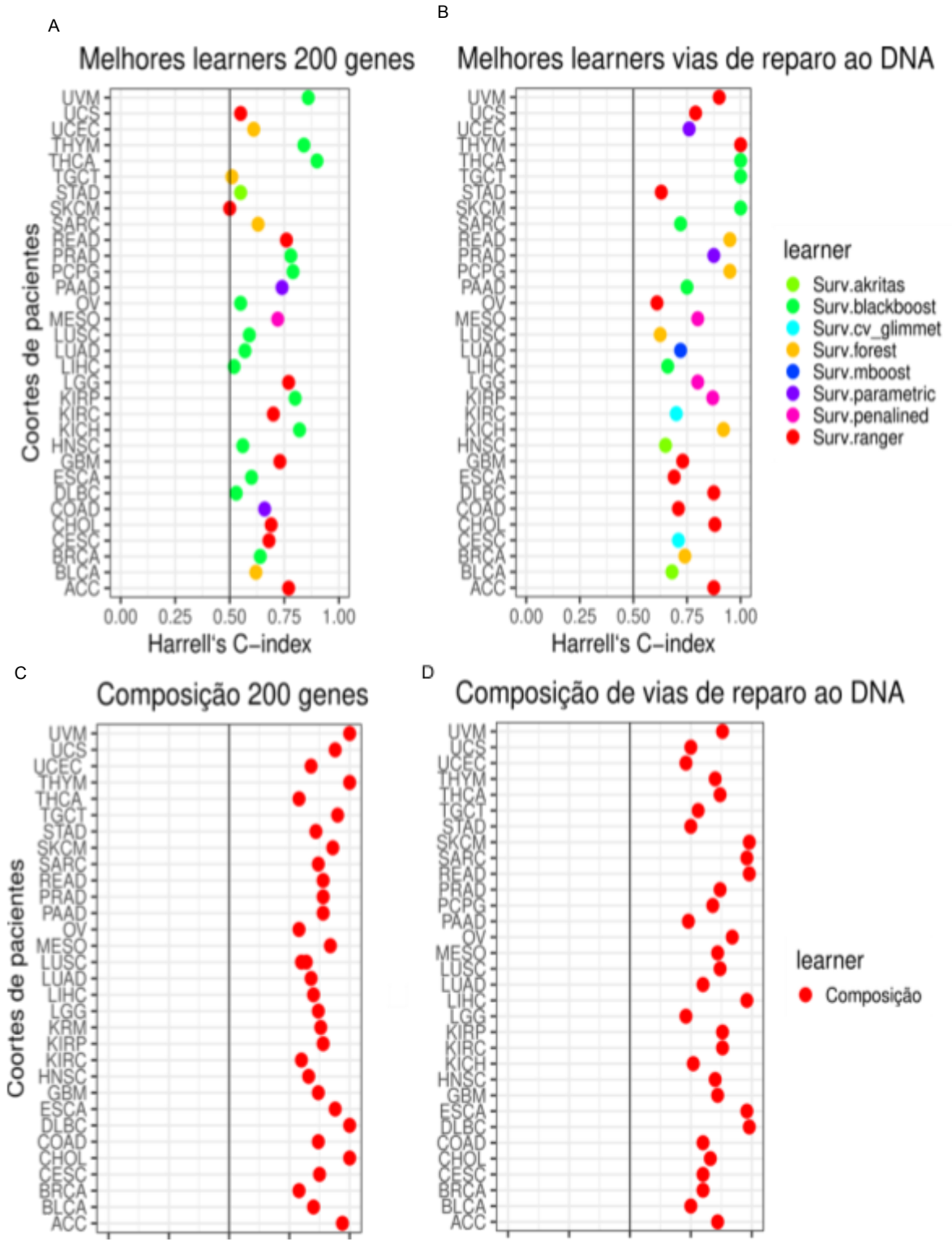
4.1.3. Genes dos processos biológicos reportados no *gene ontology* associados ao reparo de DNA no prognóstico de pacientes com câncer após diagnóstico da doença

Para a identificação dos conjuntos de genes que permitam predizer o prognóstico de pacientes com câncer baseado num embasamento biológico, utilizamos 49 grupos de genes divididos em 2 análises: i) um grupo de 200 genes de reparo ao DNA que apresentaram uma expressão maior do que 0 CPM em todas as coortes analisadas e, utilizando esses mesmos genes, ii) 48 grupos de genes baseado nas vias de reparo anotadas na base de dados *gene ontology* (Tabela 6).

Para identificar o poder preditivo dos 200 genes nas 32 coortes de pacientes, foram utilizados 11 diferentes *learners* (Tabela 5) com os parâmetros sugeridos pelos autores e com o método de reamostragem cv. Como pode ser verificado na figura 6A, observamos um valor predictivo superior a 0.5 em 30 das coortes. A coorte SKMC e TGCT não apresentou valores preditivos diferentes de 0.5, enquanto a coorte THAC apresentou um valor do Harrell's C-index próximo a 1. Para aumentar o poder preditivo dos 200 genes, foi utilizado o método de composição "lp-distr". Para tanto, as composições foram feitas entre 2 *learners* com diferentes tipo de predição: um *learner* lp (*surv.blackboost*, *surv.cv_glmnet*, *surv.mboost*, *surv.parametric*) e um *learner* distr (*surv.akritas*, *surv.blackboost*, *surv.cforest*, *surv.kaplan*, *surv.mboost*, *surv.nelson*, *surv.penalized*). Posteriormente selecionamos os maiores valores preditivos das composições. Como pode ser verificado figura 6B, a composição baseada com *learner surv.blackboost* com qualquer outro dos *learner* distr, apresentou aumento do poder preditivo dos 200 genes em todas as coortes, exceto em THAC.

Em seguida, avaliamos o poder preditivo das 48 vias de reparo. Identificamos a via com maior valor do Harrell's C-index para cada coorte (Figura 6C). Entre as 32 coortes, 18 mostraram ao menos uma via de reparo com valor preditivo superior a 0.75. Como pode ser verificado na figura 6D, a composição baseada no *learner surv.blackboost* aumentou o valor preditivo das coortes com menor valor de index. Inesperadamente, alguns *learners* tiveram maior valor preditivo que a estratégia de composição. Por exemplo, quando comparado o máximo valor predito obtido na coorte THAC utilizando os 200 genes (Figura 6A) com a estratégia de composição para a mesma coorte (figura 6B), há uma 31 diminuição do valor preditivo. Esta mesma diminuição do poder preditivo na estratégia da composição, também foi achada utilizando as vias de reparo para as coortes SKMC, THCA, THYM (Figura 6C e 6D). As vias de reparo ao

DNA com maiores valores do Harrell's C-index estão relacionadas aos sistemas de reparo de quebra de fitas duplas (Figura 6E) como a recombinação homóloga seguida dos mecanismos nas quebras de fita simples, NER ou BER. Diferente da composição, o *learner surv.blackboost* não foi o método dos valores preditivos mais altos utilizando vias de reparo ao DNA.


A
Melhores learners 200 genes
Coortes de pacientes
UVM
UCS
UCEC
THYM
THCA
TGCT
STAD
SKCM
SARC
READ
PRAD
PCPG
PAAD
OV
MESO
LUSC
LUAD
LIHC
LGG
KIRP
KIRC
KICH
HNSC
GBM
ESCA
DLBC
COAD
CHOL
CESC
BRCA
BLCA
ACC
0.00
0.25
0.50
0.75
1.00
Harrell's C−index
B
Melhores learners vias de reparo ao DNA
learner
Surv.akritas
Surv.blackboost
Surv.cv_glimmet
Surv.forest
Surv.mboost
Surv.parametric
Surv.penalined
Surv.ranger
C
Composição 200 genes
Coortes de pacientes
KRM
D
Composição de vias de reparo ao DNA
Composição

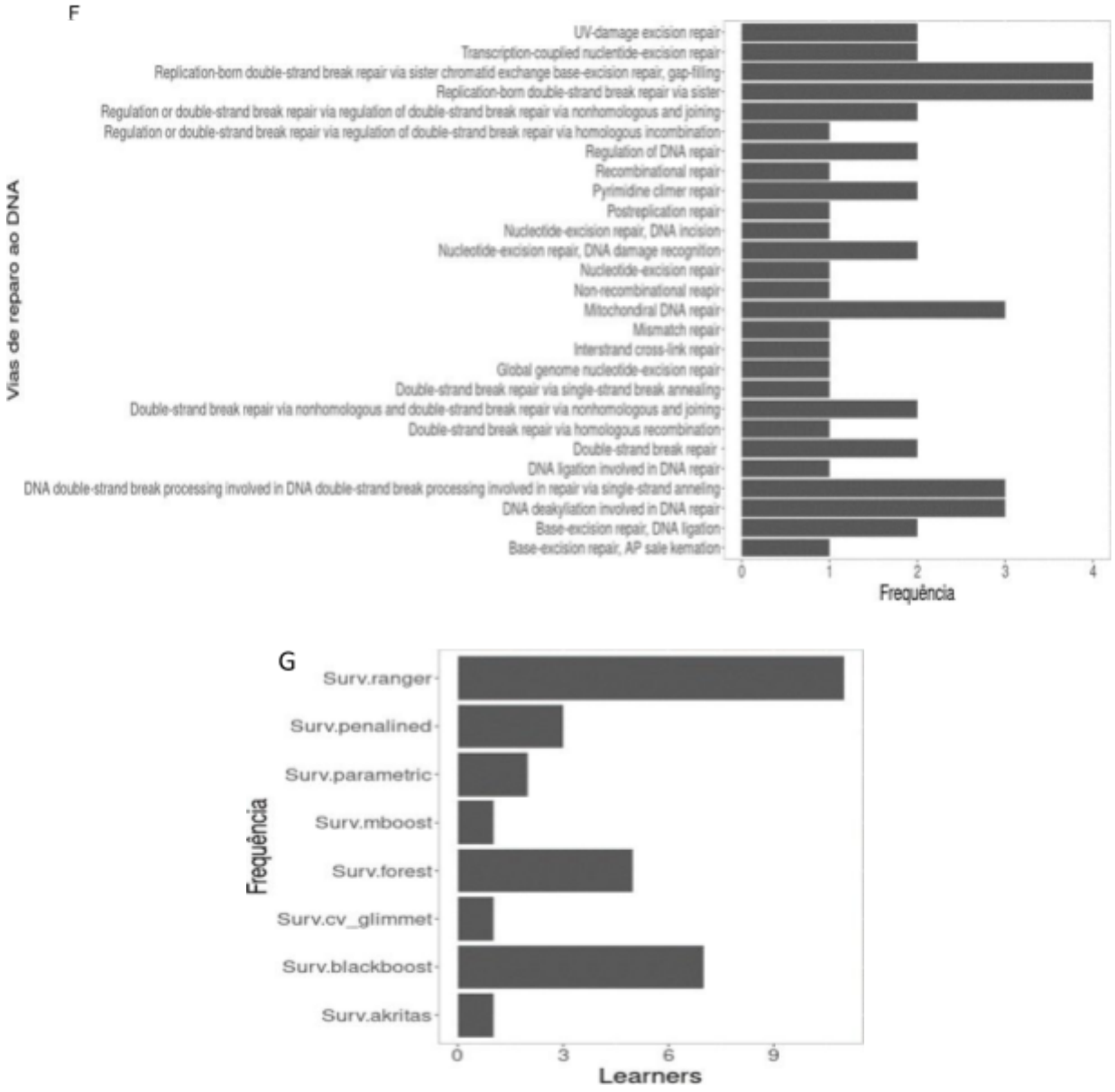


Figura 6. Avaliação dos modelos preditivos baseados nas vias de reparo do DNA. A) O maior valor preditivo entre os *learners* utilizando os 200 genes para cada coorte. B) Maior valor preditivo para cada learner e coorte. C) Maior valor preditivo com a composição utilizando uma das 48 vias de reparo para cada coorte. D) Maior valor preditivo entre os *learners* utilizando uma das 48 vias de reparo. E) Vias com maior valor preditivo nas 32 coortes. F) *Learners* mais frequentes com maior valor nas vias de reparo.

Tabela 6. Vias associadas ao reparo do DNA anotadas no gene ontology e o número de genes correspondentes utilizados neste estudo nas 32 coortes.

Vias	Número de genes
meiotic mismatch repair	5
base-excision repair, DNA ligation	5
DNA double-strand break processing involved in repair via single-strand annealing	5
replication-born double-strand break repair via sister chromatid exchange	5
DNA dealkylation involved in DNA repair	6
double-strand break repair via single-strand annealing	6
double-strand break repair via classical nonhomologous end joining	6
double-strand break repair via alternative nonhomologous end joining	6
mitochondrial DNA repair	7
double-strand break repair via synthesis-dependent strand annealing	7
base-excision repair, gap-filling	8
DNA ligation involved in DNA repair	8
UV-damage excision repair	8
nucleotide-excision repair, DNA damage recognition	9
pyrimidine dimer repair	9
single strand break repair	10
regulation of double-strand break repair via nonhomologous end joining	10
base-excision repair, AP site formation	12
nucleotide-excision repair, DNA duplex unwinding	16
nucleotide-excision repair, DNA gap filling	16
regulation of double-strand break repair via homologous recombination	16
global genome nucleotide-excision repair	18
nucleotide-excision repair, preincision complex stabilization	19
nucleotide-excision repair, DNA incision, 3′-to lesion	19
regulation of double-strand break repair	22
nucleotide-excision repair, preincision complex assembly	23
DNA synthesis involved in DNA repair	24

mismatch repair	24
nucleotide-excision repair, DNA incision, 5'-to lesion	26
postreplication repair	28
nucleotide-excision repair, DNA incision	28
base-excision repair	33
transcription-coupled nucleotide-excision repair	34

Em suma, este estudo identificou 22 vias de reparo que predizem o tempo de sobrevida dos pacientes com câncer após diagnóstico em 32 coortes de pacientes diferentes, utilizando um único *Learner* ou método de composição (*surv.blackboost*). O poder preditivo para todas as coortes é igual ou superior ao valor Harrell's C-index de 0.75 (Tabela 7).

Tabela 7. Conjuntos de genes representados em vias de processos biológicos relacionados ao reparo do DNA e *learners* melhor predizem o tempo de sobrevida para os pacientes com câncer em cada coorte analisada.

Coorte	Vias de reparo	Valor preditivo	Learner	Estratégia
SKCM	base-excision repair, gap-filling	1	*surv.cforest*	Learners
TGCT	DNA double-strand break processing involved in repair via single-strand annealing	1	*surv.blackboost*	Learners
THCA	nucleotide-excision repair, DNA incision	1	*surv.blackboost*	Learners
THYM	non-recombinational repair	1	*surv.parametric*	Learners
DLBC	DNA dealkylation involved in DNA repair	0.99	*surv.blackboost*	Composição
ESCA	double-strand break repair via homologous recombination	0.98	*surv.blackboost*	Composição
LIHC	global genome nucleotide-excision repair	0.98	*surv.blackboost*	Composição
SARC	transcription-coupled nucleotide-excision repair	0.98	*surv.blackboost*	Composição
READ	postreplication repair	0.97	*surv.blackboost*	Composição
PCPG	base-excision repair, gap-filling	0.95	*surv.blackboost*	Learners
KICH	base-excision repair, AP site formation	0.92	*surv.mboost*	Learners
UVM	UV-damage excision repair	0.92	*surv.ranger*	Learners
OV	UV-damage excision repair	0.92	*surv.blackboost*	Composição
KIRC	double-strand break repair	0.91	*surv.blackboost*	Composição
KIRP	regulation of DNA repair	0.9	*surv.blackboost*	Composição
CHOL	regulation of DNA repair	0.9	*surv.ranger*	Learners
ACC	mitochondrial DNA repair	0.9	*surv.cv_glmnet*	Learners
PRAD	base-excision repair, gap-filling	0.9	*surv.ranger*	Learners
MESO	DNA ligation involved in DNA repair	0.88	*surv.blackboost*	Composição
GBM	regulation of double-strand break repair via homologous recombination	0.87	*surv.blackboost*	Composição
HNSC	transcription-coupled nucleotide-excision repair	0.86	*surv.blackboost*	Composição
LGG	regulation of double-strand break repair via nonhomologous end joining	0.84	*surv.ranger*	Learners
UCS	pyrimidine dimer repair	0.82	*surv.penalized*	Learners

4.2. Comparação dos genes de reparo na predição de sobrevida dos pacientes com câncer através com outras vias envolvidas em carcinogênese.

Para avaliar o desempenho dos genes de reparo e utilizados ao longo deste estudo, foram comparadas vias de envolvidas com na carcinogênese como apoptose e ciclo celular. Utilizamos o método de composição com os algoritmos surv.ranger e *surv.blackboost* para avaliar a capacidade preditiva desses conjuntos de genes.

Os genes de reparo foram divididos no core (os genes utilizados neste estudo) e os genes core e genes associados com vias de reparo associados a ubiquitinas e outros processos também importante no reparo do DNA. Estes foram misturados com genes de apoptose ou ciclo celular.

Como pode ser observado na figura 7, os genes de reparo apresentaram o valores mais baixos na predição do prognóstico de pacientes com câncer após diagnóstico da doença utilizando machine learning com o método de composição em quando o global da expressão gênica mostrou se como com os valores mais altos na maioria das coortes.

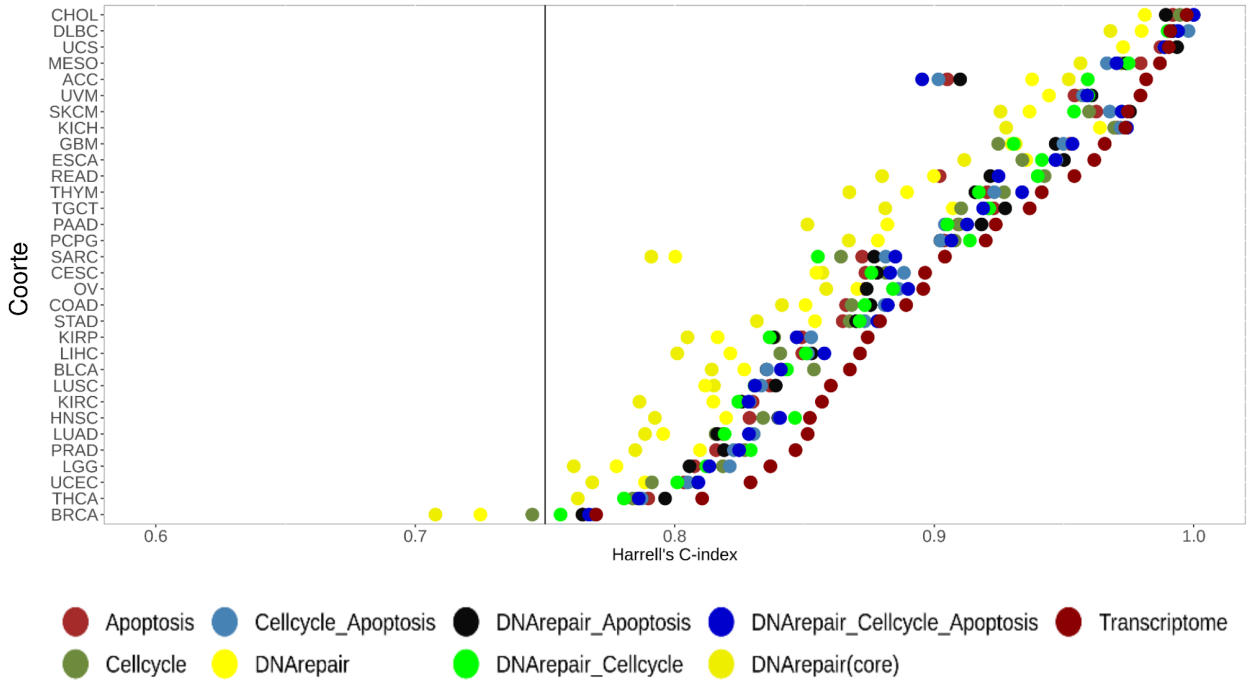


Figura 7. Comparação dos genes de reparo ao DNA com genes de ciclo celular e apoptose utilizando a estratégia de composição de métodos de aprendizado automático.

4.3. Validação de um *score* de desregulação das vias de reparo de DNA na predição do prognóstico em diferentes tipos de câncer utilizando dados *scRNA-seq*

A identificação de genes para previsão do prognóstico de pacientes com câncer após o diagnóstico da doença, juntamente com a seleção de i) genes através de algoritmos de seleção de variáveis, ii) genes conhecidos por estarem associados à sobrevivência dos pacientes para o desenvolvimento de um score de desregulação da expressão de genes pertencentes às vias de reparo de DNA e iii) genes dos processos biológicos relatados no *gene ontology* associados ao reparo de DNA no prognóstico de pacientes com câncer após o diagnóstico da doença, utilizando dados *scRNA-seq*, foi descartada devido à escassez de estudos bancos de dados públicos, com número suficiente de células, pacientes e dados clínicos adequados para estabelecer a sobrevivência dos pacientes ou utilizar métodos de aprendizado disponíveis em bancos de dados públicos. Portanto, utilizamos os resultados dos 7 scores de desregulação desenvolvidos em RNA-seq para determinar o prognóstico do pacientes no único estudo com dados de sobrevivência. No entanto, os métodos de aprendizado também não foram utilizados devido ao baixo número de pacientes.

4.4. Validação dos *scores* de genes do reparo do DNA associados com a sobrevida em pacientes com câncer de ovário em dados scRNA-seq

Para a validação dos *scores* de genes reparo vias de reparo de DNA em câncer de ovário a partir de dados sc*RNA-seq*, e determinar o valor preditivo no prognóstico da doença é preciso os processamentos dos dados de expressão. Em total, foram analisadas 84742 células de pacientes com tumor de ovários, gerando 25 clusters diferentes (Figura 8A). Foram anotadas células T CD8, fibroblastos, células mieloides e 68387 células malignas (Figura 8B) pertencentes aos nove pacientes (Figura 8C). Foram removidos os efeitos de lote das duas tecnologias (10X genomics e iCell 8) utilizadas para a obtenção as *counts* com o pacote seguindo os parâmetros definidos pelos autores e os dados de expressão foram obtidos com a mesma ferramenta (Figura 8D).

Para selecionar os genes envolvidos no reparo do DNA, foram selecionados genes com expressão maior que 0 em 10% das células TU em cada paciente incluído neste trabalho, partindo dos 200 genes utilizados nas analise de RNA-seq. Após a filtragem, foi definido um conjunto de 100 genes. Em seguida, conforme esquematizado na tabela 4, foi calculado o *os scoresI,* analisados nos dados RNA-seq, de cada célula em todos os pacientes com uma modificação no final do cálculo. Antes da categorização dos *scores*, a soma dos *scores* de cada célula maligna foi ponderada para obter os resultados em cada paciente.

Finalmente, identificamos uma alta expressão do gene RAD52 específica das células tumorais de pacientes com câncer de ovário (Figura 8D). Também, o *score* formado pelos genes RAD52 e RECQL, desenvolvido utilizando dados *RNA-seq* (Figura 3), apresentou altos valores em pacientes com menor sobrevivência após o diagnóstico da doença. Da mesma forma, os valores máximos deste *score* nas células malignas dos pacientes com câncer de ovário apresentam uma menor sobrevivência geral, embora não seja estatisticamente significativo ($p<0,065$). Destacadamente, utilizando os *scores* desenvolvidos em dados de *RNA-seq* de 32 coortes de tumores, foi possível identificar que 1 dos 7 *scores* com a capacidade de determinar a sobrevivência dos pacientes com câncer de ovário após cirurgia. Os valores altos do score formado pelos genes RAD52 e POLB nas células malignas mostraram uma associação com menor sobrevivência dos pacientes com câncer em dados RNA-seq (Figura 3).

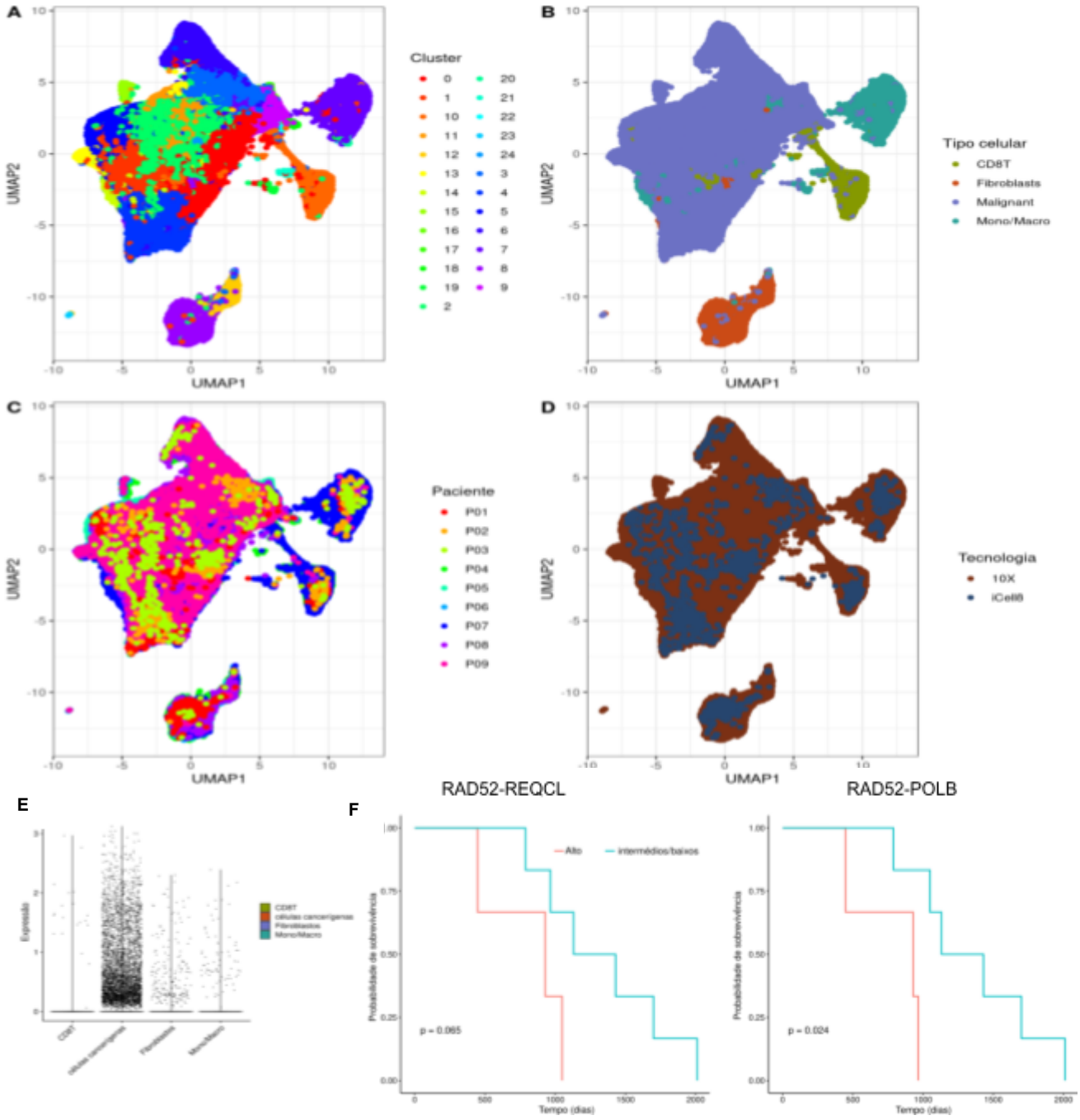


Figura 8. Avaliação do *score* de desregulação em pacientes com câncer de ovário em dados *scRNA-seq*. A) Cluster das células em amostras de nove pacientes com câncer de ovário. B) Anotação das células mielóides, linfóides, fibroblastos e células tumorais. C) Células por paciente incluídas neste estudo. D) Tecnologias de s*cRNA-seq* em 10x e iCell8. E) Perfil de expressão gênica de RAD52 em células tumorais de câncer de ovário, CD8T, fibroblastos e células mielóides. F)

Curvas de sobrevida do *score* de desregulação gênica gerado com a expressão dos pares de genes RAD52/REQCL e RAD52/POLB em células tumorais do câncer de ovários.

5. Discussão

5.1. Os *scores* de desregulação baseado em gene de reparo ao DNA

Os mecanismos de reparo do DNA são essenciais para a manutenção da integridade genômica e a sobrevivência celular. A resposta ao dano ao DNA promove a transmissão fiel dos genomas em células em divisão, revertendo o dano extrínseco e intrínseco ao DNA e é necessária para a sobrevivência celular durante a replicação. A eficácia dos medicamentos antitumorais é altamente influenciada pela capacidade de reparo do DNA celular, sendo um fator chave na resposta em diferentes esquemas de tratamento do câncer (Helleday et al., 2008). Alterações nos mecanismo de reparo ao DNA podem ser usadas como biomarcadores de prognóstico dos pacientes com câncer (Kang et al., 2012; Kassambara et al., 2014; Pitroda et al., 2014).

Estudos recentes, com diversos tipos de tumores, incluindo ovário, mama, pulmão, entre outros, têm mostrado a aplicabilidade desses *scores* na determinação do prognóstico da doença (Kang et al., 2012; Kassambara et al., 2014; Pitroda et al., 2014). Esses *scores* utilizam abordagem similares às utilizadas neste estudo, mostrando que a somatória da expressão gênica de genes sabidamente relacionados, podem ser utilizadas para prever o possível desfecho de uma doença de paciente que não receberam esquemas de tratamento ou com pré ou pós tratamento. Também, nosso grupo de pesquisa participou do desenvolvimento de um *score* de desregulação com genes pertencentes às vias de reparo ao DNA. Em 2020, Jimenez at el., apresentou uma abordagem para a previsão de resposta insatisfatória à quimiorradiação neoadjuvante em pacientes com câncer retal usando 8 genes das vias BER, NER, NHEJ e HR. Interessantemente, a previsão de resposta à quimiorradiação foi realizada em dados de expressão, como RNA-seq, qPCR ou microarray, utilizando a mesma assinatura gênica e atingindo valores de acurácia na predição da resposta entre 70% e 90% (Jimenez et al., 2020),

Com base em toda essa evidência e na experiência no desenvolvimento de *scores* para a previsão de desfechos clínicos, este estudo buscou identificar padrões de expressão gênica

envolvidos nos mecanismos de reparo do DNA, que permitam determinar a sobrevida geral em coortes de pacientes com diferentes tipos de tumores. Para tanto foram utilizados os dados de expressão gênica de tumores primários e os dados clínicos de desfecho clínico do repositório público TCGA.

A primeira abordagem utilizada neste trabalho foi a utilização do algoritmo de seleção de variáveis do pacote pec (Mogensen et al., 2012) na linguagem de programação R. As variáveis selecionadas através do método de seleção de variáveis baseado em árvores de decisão foram diferentes para cada coorte. O método de seleção encontrou variáveis apenas em 19 dos 32 coortes, o que impediu uma comparação geral dos dados disponíveis e selecionou poucos genes para coortes com menos pacientes, como UVM ou UCS (Figura 2). Inexplicavelmente, não selecionou genes para a coorte com maior número de pacientes, o BRCA (Figura 2). Ao predizer a sobrevida global com os conjuntos de genes selecionados os valores Harrel's C-Index foram baixos. Nenhum dos 19 conjuntos de genes selecionados apresentou valores acima de 90% no Harrel's C-Index (Figura 2A) e em coortes como KIRC, LIHC ou STAD, apesar de ter selecionado mais de 100 genes, serem datasets bem anotados e com centos de amostras, apresentou valores de 50% no Harrell's C-Index, o que indica que a seleção de variáveis nesses estudos não apresenta capacidade predizer o prognóstico da doença. Este método de seleção de variáveis apresenta vantagens como a facilidade de interpretação dos resultados, a capacidade de lidar com dados categóricos e numéricos, além de ser capaz de lidar com dados faltantes. No entanto, essa abordagem pode apresentar algumas desvantagens, como a sensibilidade a dados com um pequeno número de amostras e outliers, além de poder ser computacionalmente intensivo e levar a modelos super ajustados (overfitting) (Sanchez-Pinto et al., 2018). Devido a essas limitações, a escolha de genes por algoritmos pode ser diferente entre as coortes, o que diminui a chance de encontrar padrões de expressão gênica para predição da sobrevida global entre diferentes tipos de câncer.

Em uma segunda tentativa, avaliamos pares de genes para compor o *score* de desregulação com genes associados aos mecanismos de reparo ao DNA. Surpreendentemente, determinamos 7 scores com a capacidade de prever a sobrevivência global em 31 coortes de pacientes analisados (Figura 3). Esses *scores* eram compostos por 10 genes. Em linha com nossos resultados, alguns desses genes estão relacionados ao prognóstico do câncer. Por exemplo, o gene PDSB5 (também conhecido como APRIN) pertence à família PDS5, conhecidas por estarem

envolvidas em interações proteína-proteína e estão alterados no câncer de ovário (Couturier et al., 2016) e no carcinoma de esôfago (Gan et al., 2022). Além disso, o gene PDSB5 foi reportado como preditor de resposta à terapia em câncer de mama (Brough et al., 2012). O gene PDSB5 compõe dois *scores* de desregulação que têm a capacidade de prever a sobrevivência de pacientes em 14 das coortes, incluindo a BRCA. Outro exemplo de genes envolvidos no prognóstico do câncer é o LIG3, sendo um biomarcador de prognóstico em câncer de mama (Sun et al., 2021), e sua superexpressão relacionada a fenótipos agressivos, resistência à platina e baixa sobrevida livre de progressão em câncer de mama (Ali et al., 2021). Embora apenas componha um *score* de desregulação com valor preditivo nas coortes BLCA e BRCA, este gene está sabidamente alterado em múltiplos tumores, como câncer de mama, leucemias, neuroblastoma e outros (Tomkinson et al., 2020). Nosso resultado não unicamente aponta na mesma direção do reportado na literatura, mas também revela o potencial de alguns genes como proto oncogenes em outros carcinomas e seu valor como biomarcador de prognóstico até agora pouco explorado.

Esta estratégia simples de conformação dos *scores* de desregulação demonstrou genes sabidamente associados ao tempo de sobrevida de pacientes com câncer. No entanto, esses genes não são biomarcadores prognósticos das doenças devido ao perfil de expressão. No caso de BRCA1 e BRCA2, são amplamente conhecidos pelas mutações e seu diagnóstico e prognóstico em câncer de mama, tanto nos casos com história familiar quanto esporádicos (Baretta et al., 2016). No mesmo sentido, alguns estudos mostraram evidências de uma associação entre a mutação germinal no gene RECQL e um mau prognóstico em câncer de mama (Bowden & Tischkowitz, 2019). Outros genes relacionados com a mutação e carcinogênese são POLB em câncer gástrico (Tan et al. 2015) e POLD3, que codifica para a subunidade p66 da enzima ADN polimerase δ, em câncer colorretal (Alvizo-Rodríguez et al. 2022). Este trabalhou revelou que os genes BRCA1, BRCA2, RECQL, POLB e POLD3, os quais apresentam mutações em diferentes tipos de câncer, compõem diferentes *scores* de desregulação, e baseado na expressão gênica, podem determinar a sobrevida em mais de 75% das coortes de pacientes analisados (figura 3).

Surpreendentemente, o gene RAD52, que compõe 4 dos 7 *scores* de desregulação que podem prever a sobrevida geral em 20 coortes LGG, OV, BRCA, ACC, KIRC, MESO, PAAD, GMB, PCPG, PRAD, PAAD, DLBC, HSNC, KIRP, LIHC, LUAD, UVM, UCS, BLCA e KICH (Figura 3 e 4). Estudos recentes mostraram que o RAD52 tem um papel importante na manutenção da estabilidade genômica e na supressão do câncer. A proteína sensível à radiação 52 (RAD52), foi identificada pela primeira vez em S. cerevisiae e ainda é um gene de recombinação

homóloga pouco caracterizado. Esta proteína desempenha um papel na troca de fitas de DNA e medeia a interação DNA-DNA necessária para a ligação de fitas de DNA complementares durante a recombinação homóloga (HR) em células de mamíferos (Lieberman & You, 2017). O papel proto-oncogênico do gene RAD52 está claramente elucidado em câncer de ovário, fígado (Lieberman & You, 2017), mama (Huang et al., 2016; Liang et al., 2012; Palmieri et al., 2009), pulmão (Lieberman et al., 2016; J. Shi et al., 2012), leucemia (Chandramouly et al., 2015; Cramer-Morales et al., 2013), linfoma (Treuner et al., 2004), câncer cervical (T.-Y. Shi et al., 2012) e colorretal (Naccarati et al., 2016). No entanto, a literatura apenas mostra que o RAD52 está relacionado com o prognóstico de sobrevida em pacientes com câncer de ovário (Diao et al., 2022), câncer cervical (T.-Y. Shi et al., 2012) e câncer retal (Ho et al., 2020). Nosso estudo indica que RAD52 tem valor preditivo prognóstico em coortes de pacientes com câncer diferentes tipos de câncer que ainda não foram analisadas (Figura 3). Além disso, identificamos um padrão de expressão de RAD52 específico das células tumorais de pacientes com câncer de ovário em os *scores* de desregulação RAD52-POLB desenvolvidos em dados de RNA-seq foram replicados em dados de *scRNA-seq* de pacientes com câncer de ovário (Figura 8F). Este *score* mostrou capacidade de previsão utilizando muitas células tumorais e poucos pacientes (Figura 8).

Em síntese, a estratégia simples de conformação de 7 *scores* de desregulação, composto com 10 genes de reparo ao DNA mostrou uma associação com a sobrevida de mais de 10,000 pacientes em 31 coortes diferentes.

5.2. Métodos de aprendizado de máquinas e sobrevida global

A literatura sobre a previsão da sobrevida global em pacientes com câncer utilizando aprendizado de máquina tem crescido e evoluído nos últimos anos. Os pesquisadores têm explorado diferentes algoritmos e técnicas de aprendizado de máquina, como redes neurais artificiais, máquinas de vetores de suporte (SVM), florestas aleatórias (do inglês *random forests*) e algoritmos de aprendizado profundo, entre outros, para melhorar a precisão e a eficácia das previsões. Estes estudos geralmente envolvem a análise de grandes conjuntos de dados clínicos e moleculares, incluindo informações demográficas, histórico médico, dados genômicos, expressão gênica e proteômica, e dados de imagem, como radiômica. O objetivo é identificar padrões e características que possam ser correlacionados com a sobrevida e a resposta ao tratamento em

pacientes com câncer. Os resultados desses estudos têm sido promissores e sugerem que os modelos de aprendizado de máquina podem desempenhar um papel significativo na previsão da sobrevida global em pacientes com câncer. (Donizy et al., 2022; Mosquera et al., 2021).

Neste estudo, realizamos uma análise extensiva dos algoritmos que permitem prever a sobrevida global utilizando a expressão gênica de genes sabidamente relacionados entre si. No final, identificamos o melhor algoritmo e estratégia de método de aprendizado para a previsão da sobrevida global e comparamos genes de reparo ao DNA com outros genes sabidamente relacionados à tumorigênese.

Inicialmente, os métodos de aprendizado de máquina apresentarem baixos valores do índice de concordância utilizando diferentes conjuntos de genes selecionados por diferentes estratégias como i) a selecção de variáveis por algoritmos (Figura 2), ii) genes sabidamente associados com a sobrevida (Figura 5), iii) genes sem nenhuma relação a sobrevida dos pacientes com câncer (Figura 5). Depois, partimos para nova abordagem de seleção de genes para melhorar os modelos preditivos, aumentando o número de genes (até 200) (Tabela 3) ou agrupá-los em vias de reparo segundo o reportado as vias *gene ontology* (Tabela 6). Como resultado disto, foi possível identificar que os algoritmos de predição, com um mesmo conjunto de genes, têm diferentes valores de concordância segundo o tipo da coorte. Os algoritmos *surv.ranger* e *surv.blackboost* apresentarem o melhor desempenho na predição utilizando conjuntos de genes selecionados por i) variáveis (Figura 2), ii) 200 genes de reparo ao DNA analisados neste estudo (Figura 6) ou ii) genes pertencentes às mecanismos de reparo (Figura 6).

O surv.ranger (do inglês Ranger Survival Learner) do pacote mlr3 (Lang et al., 2019), é um algoritmo de aprendizagem de máquina de código aberto projetado para análise de tempo de sobrevivência ou eventos. Essa técnica utiliza uma implementação eficiente de *Random Survival Forest*s (RSF), que é uma técnica de modelagem preditiva usada para prever o tempo de sobrevivência de um paciente ou a duração de um evento. No entanto, até o momento, a literatura científica não relata o uso dessa técnica ou dados de expressão ou dados clínicos para a previsão de sobrevida. Por outro lado, o algoritmo *surv.blockbuster* é amplamente utilizado na predição de sobrevida. O *surv.blackboost* (do inglês *Gradient Boosting with Regression Trees*) do pacote mlr3 (Lang et al., 2019) é uma técnica de aprendizado de máquina para problemas de regressão amplamente utilizada na previsão de sobrevida. Ele constrói cada árvore de regressão

sequencialmente, utilizando uma função de perda predefinida para medir o erro em cada etapa e corrigi-lo na próxima. Dessa forma, o modelo de previsão é, na verdade, um conjunto de modelos de previsão mais fracos. Utilizando dados clínicos, essa técnica de aprendizado de máquina pode prever a mortalidade entre etapas após a paliação do ventrículo único, um tratamento complexo e em estágios para doenças cardíacas congênitas que afetam a estrutura e a função do coração (Sunthankar et al., 2023). Estudos recentes mostraram a capacidade de prever a sobrevida utilizando a expressão gênica de 900 linhas de câncer do projeto DepMap (Rosenski et al., 2023).

Os algoritmos *surv.blackboost* e *surv.ranger* têm suas desvantagens. Tanto o Ranger Survival Learner quanto o Gradient Boosting com árvores de regressão podem ser sensíveis a ruído e outliers nos dados. Isso pode afetar negativamente o desempenho dos modelos em alguns casos. Para contornar estas limitações, propusemos a estratégia de composição. Esta refere-se à combinação de vários modelos de aprendizado de máquina para criar um modelo final mais preciso e robusto. A ideia básica é que, ao combinar vários modelos diferentes, os pontos fortes de um modelo possam compensar as fraquezas de outro, resultando em um modelo geral melhor. Essas estratégias de composição em machine learning podem ajudar a melhorar a precisão e a robustez dos modelos de aprendizado de máquina, tornando-os mais adequados para uso em uma ampla gama de aplicações. Como resultado dessa estratégia, atingimos valores preditivos de sobrevida superior a 85% até 100% utilizando menor quantidade de genes (Tabela 7).

Finalmente, comparamos os genes de reparo ao DNA com outras vias sabidamente associadas à carcinogênese utilizando o método de composição (Figura 7). Nosso estudo mostrou que os métodos de aprendizado de máquina melhoram com um aumento da quantidade de genes nos modelos. O valor Harrell`s C-index nos modelos com dos genes de reparo foi inferior quando comparado com apoptose, ciclo celular ou todo o transcriptoma. Destacadamente, as vias de reparo atingiram valores de concordância superiores a 95% nos modelos preditivos das coortes SKCM (*base-excision repair, gap-filling*, 8 genes), TGCT *(DNA double-strand break processing involved in repair via single-strand annealing*, 5 genes), THCA (*nucleotide-excision repair, DNA incision*, 28 genes), LHIC (*DNA dealkylation involved in DNA repair, 6 genes*) e SARC (*global genome nucleotide-excision repair*, 18 genes) (Tabela 6 e 7) comparado ao valores utilizando centos de genes das vias reparo, apoptose, ciclo celular ou todo o transcriptoma (Figura 7).

Em resumo, os modelos de aprendizado de máquina na análise de sobrevivência mostraram- se bem ajustados ao conjunto de genes, aos dados de expressão, aos algoritmos e às estratégias utilizadas. No entanto, ainda há desafios a serem enfrentados, como a qualidade e a consistência dos dados, a escolha do algoritmo apropriado e a interpretabilidade dos modelos para avaliar se a expressão gênica é necessária e/ou suficiente para a predição da sobrevida.

Nosso estudo apresenta uma abordagem com um alcance inesperado, mas há várias limitações que têm que ser levadas em conta. A estratégia utilizada para a conformação dos escores de desregulação é simples e pode não capturar a complexidade da relação entre os genes de reparação do ADN e a sobrevivência em pacientes com câncer. Seria benéfico explorar abordagens analíticas mais avançadas que possam abordar melhor as interações entre genes e seu impacto no prognóstico do câncer. No entanto, seria útil replicar nos nossos *scores* os resultados em outros tipos de câncer em conjuntos de dados de *scRNA-seq* para garantir a validade e a generalização dos achados. Embora alguns genes identificados nos *scores* de desregulação tenham sido relatados como biomarcadores de prognóstico em certos tipos de câncer, é possível que esses genes não sejam biomarcadores prognósticos das doenças devido ao perfil de expressão unicamente. Seria necessário investigar mais a fundo a relação entre a expressão gênica, mutações e o prognóstico em diferentes tipos de câncer.

6. Conclusões

- A seleção de variáveis através de algoritmos está superajustada aos dados.
- Os métodos de aprendizado de máquina estão superajustados aos conjuntos de dados e esses modelos podem não ser replicáveis em diferentes coortes.
- As vias de reparo ajudam a melhorar a predição dos modelos gerados pelo aprendizado de máquinas.
- Os algoritmos *surv.blackboost* e *surv.ranger*, juntamente com o método de composição atingem valores mais altos do valor de concordância.
- Genes associados a outros processos carcinogênicos podem ajudar na predição da sobrevida global.
- Os scores de desregulação gênica com genes de mecanismo de reparo ao DNA estão associados à sobrevida geral.
- Os genes de reparo do DNA estão associados à sobrevida global.
- Genes que estão associados ao prognóstico de uma doença por suas mutações, em *scores* de desregulação, têm sua expressão gênica associada ao prognóstico.
- Sete *scores* de desregulação compostos por 10 genes de reparo do DNA estão associados à sobrevida global de 31 coortes de pacientes.

7. Referências

Aalen, O. (1978). Nonparametric inference for a family of counting processes. *Annals of Statistics*, *6*(4), 701–726.

Akritas, M. G. (1994). Nearest Neighbor Estimation of a Bivariate Distribution Under Random Censoring. In *The Annals of Statistics* (Vol. 22, Issue 3).

Ali, R., Alabdullah, M., Algethami, M., Alblihy, A., Miligy, I., Shoqafi, A., Mesquita, K. A., Abdel-Fatah, T., Chan, S. Y., Chiang, P. W., Mongan, N. P., Rakha, E. A., Tomkinson, A. E., & Madhusudan, S. (2021). Ligase 1 is a predictor of platinum resistance and its blockade is synthetically lethal in XRCC1 deficient epithelial ovarian cancers. *Theranostics*, *11*(17).

Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* , *31*(2), 166–169.

Baretta, Z., Mocellin, S., Goldin, E., Olopade, O. I., & Huo, D. (2016). Effect of BRCA germline mutations on breast cancer prognosis: A systematic review and meta-analysis. *Medicine*, *95*(40).

Bashiri, A., Ghazisaeedi, M., Safdari, R., Shahmoradi, L., & Ehtesham, H. (2017). Improving the Prediction of Survival in Cancer Patients by Using Machine Learning Techniques: Experience of Gene Expression Data: A Narrative Review. *Iranian Journal of Public Health*, *46*(2).

Bowden, A. R., & Tischkowitz, M. (2019). Clinical implications of germline mutations in breast cancer genes: RECQL. *Breast Cancer Research and Treatment*, *174*(3).

Breiman, L. (1984). *Classification and Regression Trees Regression Trees*.

Breiman, L. (2017). *Classification and Regression Trees*. Routledge.

Brough, R., Bajrami, I., Vatcheva, R., Natrajan, R., Reis-Filho, J. S., Lord, C. J., & Ashworth, A. (2012). APRIN is a cell cycle specific BRCA2-interacting protein required for genome integrity and a predictor of outcome after chemotherapy in breast cancer. *The EMBO Journal*, *31*(5). https://doi.org/10.1038/emboj.2011.490

Bühlmann, P., & Yu, B. (2003). Boosting With the*L*2Loss. In *Journal of the American Statistical Association* (Vol. 98, Issue 462, pp. 324–339). https://doi.org/10.1198/016214503000125

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., & Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, *36*(5), 411–420.

Carbon, S., & Mungall, C. (2018). *Gene Ontology Data Archive* [Data set]. https://doi.org/10.5281/zenodo.1899458

Chandramouly, G., McDevitt, S., Sullivan, K., Kent, T., Luz, A., Glickman, J. F., Andrake, M., Skorski, T., & Pomerantz, R. T. (2015). Small-Molecule Disruption of RAD52 Rings as a Mechanism for Precision Medicine in BRCA-Deficient Cancers. *Chemistry & Biology*, *22*(11), 1491–1504.

Chen, Y.-C., Yang, W.-W., & Chiu, H.-W. (2009, June). Artificial neural network prediction for cancer survival time by gene expression data. *2009 3rd International Conference on Bioinformatics and Biomedical Engineering*. 2009 3rd International Conference on Bioinformatics and Biomedical Engineering (iCBBE), Beijing, China. https://doi.org/10.1109/icbbe.2009.5162409

Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, Sabedot T, Malta TM, Pagnotta SM, Castiglioni I, Ceccarelli M, Bontempi G, Noushmehr H (2015). " TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data." Nucleic

Acids Research. doi:10.1093/nar/gkv1507, http://doi.org/10.1093/nar/gkv1507.

Couturier, A. M., Fleury, H., Patenaude, A.-M., Bentley, V. L., Rodrigue, A., Coulombe, Y., Niraj, J., Pauty, J., Berman, J. N., Dellaire, G., Di Noia, J. M., Mes-Masson, A.-M., & Masson, J.-Y. (2016). Roles for APRIN (PDS5B) in homologous recombination and in ovarian cancer prediction. *Nucleic Acids Research*, *44*(22), 10879–10897.

Cramer-Morales, K., Nieborowska-Skorska, M., Scheibner, K., Padget, M., Irvine, D. A., Sliwinski, T., Haas, K., Lee, J., Geng, H., Roy, D., Slupianek, A., Rassool, F. V., Wasik, M. A., Childers, W., Copland, M., Müschen, M., Civin, C. I., & Skorski, T. (2013). Personalized synthetic lethality induced by targeting RAD52 in leukemias identified by gene mutation and expression profile. *Blood*, *122*(7), 1293–1304.

Cusumano P. (2014), European inter-institutional impact study of MammaPrint. Daniele Generali, Eva Ciruelos 31 Jul -The Breast (Churchill Livingstone)-Vol. 23, Iss: 4, pp 423-428.

Diao, Y., Li, Y., Wang, Z., Wang, S., Li, P., & Kong, B. (2022). SF3B4 promotes ovarian cancer progression by regulating alternative splicing of RAD52. *Cell Death & Disease*, *13*(2). https://doi.org/10.1038/s41419-022-04630-1

Domany, E. (2014). Using high-throughput transcriptomic data for prognosis: a critical overview and perspectives. *Cancer Research*, *74*(17). https://doi.org/10.1158/0008-5472.CAN-13-3338

Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, Thomas R Gingeras (2013). STAR: ultrafast universal RNA-seq aligner. Jan 1;29(1):15-21.doi: 10.1093/bioinformatics/bts635.

Donizy, P., Krzyzinski, M., Markiewicz, A., Karpinski, P., Kotowski, K., Kowalik, A.,

Orlowska-Heitzman, J., Romanowska-Dixon, B., Biecek, P., & Hoang, M. P. (2022). Machine learning models demonstrate that clinicopathologic variables are comparable to gene expression prognostic signature in predicting survival in uveal melanoma. *European Journal of Cancer* , *174*. https://doi.org/10.1016/j.ejca.2022.07.031

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, *33*(1). https://pubmed.ncbi.nlm.nih.gov/20808728/

Gan, W., Wang, W., Li, T., Zhang, R., Hou, Y., Lv, S., Zeng, Z., Yan, Z., & Yang, M. (2022). Prognostic Values and Underlying Regulatory Network of Cohesin Subunits in Esophageal Carcinoma. *Journal of Cancer*, *13*(5), 1588.

Goeman, J. J. (2010). L1 penalized estimation in the Cox proportional hazards model.*Biometrical Journal. Biometrische Zeitschrift*, *52*(1), 70–84.

González-Recio, O., Weigel, K. A., Gianola, D., Naya, H., & Rosa, G. J. M. (2010). L2-Boosting algorithm applied to high-dimensional problems in genomic selection. *Genetics Research*, *92*(3), 227–237.

Guo, L., Ma, Y., Ward, R., Castranova, V., Shi, X., & Qian, Y. (2006). Constructing molecular classifiers for the accurate prognosis of lung adenocarcinoma. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, *12*(11 Pt 1). https://doi.org/10.1158/1078-0432.CCR-05-2336

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982). Evaluating the Yield of Medical Tests. *JAMA: The Journal of the American Medical Association*, *247*(18), 2543–2546.

Helleday, T., Petermann, E., Lundin, C., Hodgson, B., & Sharma, R. A. (2008). DNA repair pathways as targets for cancer therapy. *Nature Reviews. Cancer*, *8*(3), 193–204.

Hosoya, N., & Miyagawa, K. (2014). Targeting DNA damage response in cancer therapy. *Cancer Science*, *105*(4). https://doi.org/10.1111/cas.12366

Hothorn, T., & Zeileis, A. (2014). *Partykit: A Modular Toolkit for Recursive Partytioning in R*.

Ho, V., Chung, L., Singh, A., Lea, V., Abubakar, A., Lim, S. H., Chua, W., Ng, W., Lee, M., Roberts, T. L., de Souza, P., & Lee, C. S. (2020). Aberrant Expression of RAD52, Its Prognostic Impact in Rectal Cancer and Association with Poor Survival of Patients. *International Journal of Molecular Sciences*, *21*(5). https://doi.org/10.3390/ijms21051768

Huang, F., Goyal, N., Sullivan, K., Hanamshet, K., Patel, M., Mazina, O. M., Wang, C. X., An, W. F., Spoonamore, J., Metkar, S., Emmitte, K. A., Cocklin, S., Skorski, T., & Mazin, A. V. (2016). Targeting BRCA1- and BRCA2-deficient cells with RAD52 small molecule inhibitors. *Nucleic Acids Research*, *44*(9), 4189–4199.

Ishibashi, Y., Hanyu, N., Nakada, K., Suzuki, Y., Yamamoto, T., Yanaga, K., Ohkawa, K., Hashimoto, N., Nakajima, T., Saito, H., Matsushima, M., & Urashima, M. (2003). Profiling gene expression ratios of paired cancerous and normal tissue predicts relapse of esophageal squamous cell carcinoma. *Cancer Research*, *63*(16). https://pubmed.ncbi.nlm.nih.gov/12941848/

Jimenez, L., Perez, R. O., Gp, S. J., Vailati, B. B., Fernandez, L. M., Gama-Rodrigues, J., Habr-Gama, A., DeVecchio, J., Kalady, M. F., & Camargo, A. A. (2020). Prediction of Poor Response to Neoadjuvant Chemoradiation in Patients With Rectal Cancer Using a DNA Repair Deregulation Score: Picking the Losers Instead of the Winners. *Diseases of the Colon and Rectum*, *63*(3). https://doi.org/10.1097/DCR.0000000000001564

Kalbfleisch, J. D., & Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. John Wiley & Sons.

Kang, J., D'Andrea, A. D., & Kozono, D. (2012). A DNA repair pathway-focused score for prediction of outcomes in ovarian cancer treated with platinum-based chemotherapy.

Journal of the National Cancer Institute, *104*(9). https://doi.org/10.1093/jnci/djs177

Kaplan, E. L., & Meier, P. (1992). Nonparametric Estimation from Incomplete Observations. *Breakthroughs in Statistics*, 319–337.

Kassambara, A., Gourzones-Dmitriev, C., Sahota, S., Rème, T., Moreaux, J., Goldschmidt, H., Constantinou, A., Pasero, P., Hose, D., & Klein, B. (2014). A DNA repair pathway score predicts survival in human multiple myeloma: the potential for therapeutic strategy. *Oncotarget*, *5*(9). https://doi.org/10.18632/oncotarget.1740

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2014). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, *13*. https://doi.org/10.1016/j.csbj.2014.11.005

Lang, M., Binder, M., Richter, J., Schratz, P., Pfisterer, F., Coors, S., Au, Q., Casalicchio, G., Kotthoff, L., & Bischl, B. (2019). mlr3: A modern object-oriented machine learning framework in R. *Journal of Open Source Software*, *4*(44), 1903.

Liang, Z., Ahn, J., Guo, D., Votaw, J. R., & Shim, H. (2012). MicroRNA-302 Replacement Therapy Sensitizes Breast Cancer Cells to Ionizing Radiation. *Pharmaceutical Research*, *30*(4), 1008–1016.

Lieberman, R., Xiong, D., James, M., Han, Y., Amos, C. I., Wang, L., & You, M. (2016). Functional characterization of RAD52 as a lung cancer susceptibility gene in the 12p13.33 locus. *Molecular Carcinogenesis*, *55*(5), 953–963.

Lieberman, R., & You, M. (2017). Corrupting the DNA damage response: a critical role for Rad52 in tumor cell survival. *Aging*, *9*(7), 1647.

Mogensen, U. B., Ishwaran, H., & Gerds, T. A. (2012). Evaluating Random Forests for Survival Analysis using Prediction Error Curves. *Journal of Statistical Software*, *50*(11). https://doi.org/10.18637/jss.v050.i11

Mosquera, O. A., Ms, G. P., Já, D. A., Antelo, R. B., Alonso, V. N., Á, B. L., Abuín, B. A., Bao, P. L., Peleteiro, R. A., Cid, L. M., Mm, P. E., Jl, B. L., & Mv, M. M. (2021). Survival prediction and treatment optimization of multiple myeloma patients using machine-learning models based on clinical and gene expression data. *Leukemia*, *35*(10). https://doi.org/10.1038/s41375-021-01286-2

Naccarati, A., Rosa, F., Vymetalkova, V., Barone, E., Jiraskova, K., Di Gaetano, C., Novotny, J., Levy, M., Vodickova, L., Gemignani, F., Buchler, T., Landi, S., Vodicka, P., & Pardini, B. (2016). Double-strand break repair and colorectal cancer: gene variants within 3′ UTRs and microRNAs binding as modulators of cancer risk and clinical outcome. In *Oncotarget* (Vol. 7, Issue 17, pp. 23156–23169). https://doi.org/10.18632/oncotarget.6804 On the use of Harrell's C for clinical risk prediction via random survival forests. (2016). *Expert Systems with Applications*, *63*, 450–459.

Palmieri, D., Lockman, P. R., Thomas, F. C., Hua, E., Herring, J., Hargrave, E., Johnson, M., Flores, N., Qian, Y., Vega-Valle, E., Taskar, K. S., Rudraraju, V., Mittapalli, R. K., Gaasch, J. A., Bohn, K. A., Thorsheim, H. R., Liewehr, D. J., Davis, S., Reilly, J. F., … Steeg, P. S. (2009). Vorinostat Inhibits Brain Metastatic Colonization in a Model of Triple-Negative Breast Cancer and Induces DNA Double-Strand Breaks. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, *15*(19), 6148–6157.

Pennathur, A., Xi, L., Litle, V. R., Gooding, W. E., Krasinskas, A., Landreneau, R. J., Godfrey, T. E., & Luketich, J. D. (2013). Gene expression profiles in esophageal adenocarcinoma predict survival after resection. *The Journal of Thoracic and Cardiovascular Surgery*, *145*(2). https://doi.org/10.1016/j.jtcvs.2012.10.031

Pitroda, S. P., Bao, R., Andrade, J., Weichselbaum, R. R., & Connell, P. P. (2017). Low Recombination Proficiency Score (RPS) Predicts Heightened Sensitivity to DNA-Damaging

Chemotherapy in Breast Cancer. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, *23*(15), 4493–4500.

Pitroda, S. P., Pashtan, I. M., Logan, H. L., Budke, B., Darga, T. E., Weichselbaum, R. R., & Connell, P. P. (2014). DNA repair pathway gene expression score correlates with repair proficiency and tumor sensitivity to chemotherapy. *Science Translational Medicine*, *6*(229), 229ra42.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* , *26*(1), 139–140.

Rosenski, J., Shifman, S., & Kaplan, T. (2023). Predicting gene knockout effects from expression data. *BMC Medical Genomics*, *16*(1). https://doi.org/10.1186/s12920-023-01446-6

Sanchez-Pinto, L. N., Venable, L. R., Fahrenbach, J., & Churpek, M. M. (2018). Comparison of variable selection methods for clinical predictive modeling. *International Journal of Medical Informatics*, *116*. https://doi.org/10.1016/j.ijmedinf.2018.05.006

Sato, F., Shimada, Y., Selaru, F. M., Shibata, D., Maeda, M., Watanabe, G., Mori, Y., Stass, S. A., Imamura, M., & Meltzer, S. J. (2005). Prediction of survival in patients with esophageal carcinoma using artificial neural networks. *Cancer*, *103*(8).

Shreshtha M. (2014) Epidemiology of breast cancer in Indian women. Asia Pac J Clin Oncol. Vol. 13, Iss: 4, pp 289-295.

Shi, J., Chatterjee, N., Rotunno, M., Wang, Y., Pesatori, A. C., Consonni, D., Li, P., Wheeler, W., Broderick, P., Henrion, M., Eisen, T., Wang, Z., Chen, W., Dong, Q., Albanes, D., Thun, M., Spitz, M. R., Bertazzi, P. A., Caporaso, N. E., … Landi, M. T. (2012). Inherited variation at chromosome 12p13.33, including RAD52, influences the risk of squamous cell lung carcinoma. *Cancer Discovery*, *2*(2), 131–139.

Shi, T.-Y., Yang, G., Tu, X.-Y., Yang, J.-M., Qian, J., Wu, X.-H., Zhou, X.-Y., Cheng, X., & Wei, Q. (2012). RAD52 variants predict platinum resistance and prognosis of cervical cancer. *PloS One*, *7*(11), e50461.

Sun, L., Liu, X., Song, S., Feng, L., Shi, C., Sun, Z., Chen, B., & Hou, H. (2021). Identification of LIG1 and LIG3 as prognostic biomarkers in breast cancer. *Open Medicine: A Peer-Reviewed, Independent, Open-Access Journal*, *16*(1).

Sunthankar, S. D., Zhao, J., Wei, W. Q., Hill, G. D., Parra, D. A., Kohl, K., McCoy, A., Jayaram, N. M., & Godown, J. (2023). Machine Learning to Predict Interstage Mortality Following Single Ventricle Palliation: A NPC-QIC Database Analysis. *Pediatric Cardiology*. https://doi.org/10.1007/s00246-023-03130-z

Tsukada H. (2022) . Radiological predictive factors on preoperative multimodality imaging are related to Oncotype DX recurrence score in estrogen-positive/human epidermal growth factor receptor 2-negative invasive breast cancer: a cross-sectional study. Ann Nucl Med. 2022 Oct;36(10):853-864.

Tarca, A. L., Carey, V. J., Chen, X. W., Romero, R., & Drăghici, S. (2007). Machine learning and its applications to biology. *PLoS Computational Biology*, *3*(6).

Tomkinson, A. E., Naila, T., & Bhandari, S. K. (2020). Altered DNA ligase activity in human disease. *Mutagenesis*, *35*(1), 51.

Treuner, K., Helton, R., & Barlow, C. (2004). Loss of Rad52 partially rescues tumorigenesis and T-cell maturation in Atm-deficient mice. *Oncogene*, *23*(27), 4655–4661.

Wang, C. Y., Lee, T. F., Fang, C. H., & Chou, J. H. (2012). Fuzzy logic-based prognostic score for outcome prediction in esophageal cancer. *IEEE Transactions on Information Technology in Biomedicine: A Publication of the IEEE Engineering in Medicine and Biology Society*, *16*(6).

8. Anexos

8.1. Anexo 1. Semelhancas e diferencas dos métodos de predição *lp* (linear predictor) e Métodos de predição *distr* (distribuição)

O pacote mlr (Machine Learning in R) é uma biblioteca em R para aprendizado de máquina que oferece uma interface unificada para vários algoritmos de aprendizado. Os métodos de predição *lp* e *distr* são usados em diferentes contextos e para diferentes tipos de modelos. Abaixo, as características, semelhanças e diferenças entre esses dois métodos de predição são explicadas sumarizadas em 12 pontos.

Método de predição *lp* (linear predictor):

- Comumente usados em modelos de regressão, como regressão linear ou regressão logística.
- Resulta da combinação linear das características (variáveis independentes) ponderadas pelos coeficientes do modelo.
- Fornece uma estimativa do valor esperado de uma variável dependente, com base nas características e nos coeficientes do modelo.
- Modelos de sobrevivência, como Cox *Proportional Hazards* e modelos paramétricos de sobrevivência, podem usar *lp* para fornecer estimativas de risco.

Método de predição distr (distribuição):

- O método de predição *distr* é geralmente usado em modelos probabilísticos que estimam a distribuição da variável dependente.
- Ao contrário do *lp*, que fornece uma estimativa pontual, o distr fornece uma estimativa de toda a distribuição.

- O *distr* é útil em cenários em que é importante conhecer a incerteza associada às estimativas, como em modelos de risco ou sobrevivência.
- Modelos de sobrevivência, como Kaplan-Meier, podem usar *distr* para fornecer estimativas de probabilidade de sobrevivência.

Semelhanças:

- Ambos os métodos de predição *lp* e *distr* são aplicados no contexto de aprendizado supervisionado e podem ser usados em modelos de sobrevivência ou risco.
- O pacote mlr suporta ambos os métodos de predição, permitindo a implementação e comparação de vários modelos de aprendizado de máquina.

Diferenças:

- O *lp* é geralmente usado em modelos de regressão e fornece uma estimativa pontual do valor esperado da variável dependente, enquanto o *distr* é usado em modelos probabilísticos e fornece estimativas de toda a distribuição.
- O *distr* é útil para avaliar a incerteza associada às estimativas, enquanto o *lp* fornece uma estimativa única sem informações sobre a incerteza.

8.2. Anexo 2. Descrição dos métodos de predição *lp* (linear predictor):

O pacote mlr3 (Machine Learning in R) é uma biblioteca em R para aprendizado de máquina que fornece uma interface unificada para várias técnicas de aprendizado e tarefas. No contexto de análise de sobrevivência, os métodos *surv.blackboost*, *surv.cv_glmnet*, *surv.mboost* e *surv.parametric* são learners que abordam a tarefa de previsão de tempo até a ocorrência de um evento de interesse.

- *surv.blackboost*: O método *surv.blackboost* é uma implementação do algoritmo de aprendizado boosting com árvores de regressão para a análise de sobrevivência. O *boosting* combina iterativamente modelos simples para criar um modelo mais forte,

ajustando-se aos resíduos do modelo anterior. O algoritmo é aplicado em árvores de decisão para modelar a relação entre as variáveis explicativas e a resposta de sobrevivência.

- *surv.cv_glmnet*: O método *surv.cv_glmnet* é uma implementação do modelo de regressão de Cox regularizado com penalidades de *Lasso* e *Ridge* (*elastic net*). Este método combina as vantagens da regularização Lasso e Ridge para selecionar variáveis e controlar a multicolinearidade. O modelo de Cox é a abordagem mais utilizada na análise de sobrevivência para modelar o risco relativo dos eventos.
- *surv.mboost*: O método *surv.mboost* também é uma implementação do algoritmo de aprendizado *boosting*, mas com a flexibilidade de trabalhar com outros tipos de *base-learners* além de árvores de regressão. A ideia é semelhante à do *surv.blackboost*, mas permite que a escolha e combinação de diferentes *learners* para criar um modelo mais forte.
- *surv.parametric*: O método *surv.parametric* se refere a modelos de regressão de sobrevivência que assumem uma distribuição de tempo de ocorrência do evento específica, como exponencial, *Weibull* ou *Gama*. Esses modelos estimam os parâmetros da distribuição de tempo de ocorrência do evento e a relação entre as covariáveis e a resposta de sobrevivência.

8.3. Anexo 3. Descrição dos Métodos de predição *distr* (distribuição):

O pacote mlr3 (Machine Learning in R) é uma biblioteca em R para aprendizado de máquina que fornece uma interface unificada para várias técnicas de aprendizado e tarefas. No contexto de análise de sobrevivência, os métodos *surv.akritas, surv.cforest*, *surv.kaplan*, *surv.nelson*, *surv.penalized*, *surv.ranger e surv.rpart* são learners que abordam a tarefa de previsão de tempo até a ocorrência de um evento de interesse.

- *surv.akritas:* O método *surv.akritas* é um algoritmo de estimação de sobrevivência não paramétrico baseado no estimador de *Kaplan-Meier* e na estatística *log-rank*. É usado para comparar grupos de pacientes com base na sua função de sobrevivência.
- *surv.cforest:* O método *surv.cforest* é uma implementação de florestas condicionais aleatórias, uma extensão do algoritmo de florestas aleatórias adaptada para análise de sobrevivência. As florestas aleatórias são conjuntos de árvores de decisão que fazem previsões por meio de votação ou agregação dos resultados de cada árvore individual.
- *surv.kaplan*: O método *surv.kaplan* se refere ao estimador de *Kaplan-Meier*, uma técnica não paramétrica amplamente utilizada para estimar a função de sobrevivência a partir de dados de tempo de ocorrência do evento.
- *surv.nelson*: O método *surv.nelson* se refere ao estimador de Nelson-Aalen, um método não paramétrico para estimar a função cumulativa de risco a partir de dados de tempo de ocorrência do evento.
- *surv.penalized*: O método *surv.penalized* é uma implementação de modelos de regressão de sobrevivência penalizados, como a regressão de Cox com penalidades *Lasso* ou *Ridge*. Esses modelos aplicam regularização para selecionar variáveis e evitar o ajuste excessivo (overfitting).
- *surv.ranger*: O método *surv.ranger* é uma implementação rápida e eficiente do algoritmo

de florestas aleatórias para análise de sobrevivência, usando o modelo de Cox proporcional aos riscos.

- *surv.rpart*: O método *surv.rpart* é uma implementação de árvores de decisão para análise de sobrevivência. Este algoritmo é baseado na abordagem de árvores de decisão recursiva para classificação e regressão (CART), adaptada para a análise de sobrevivência. As árvores de decisão são estruturas de aprendizado de máquina que dividem os dados em subconjuntos com base nos valores das variáveis de entrada, criando uma árvore com nós internos (decisões) e nós folhas (previsões). No contexto da análise de sobrevivência, a árvore de decisão é treinada para prever o tempo até a ocorrência de um evento de interesse, como a ocorrência do evento de um componente ou a morte de um paciente.