Universidade de Brasília

Departamento de Estatística

Aplicação de Métodos Multivariados em Dados *Multi-omics*

Ana Carolina Souto Valente Motta

Brasília

2019

Ana Carolina Souto Valente Motta

Aplicação de Métodos Multivariados em Dados *Multi-omics*

Universidade de Brasília

Instituto de Ciências Exatas

Departamento de Estatística

Trabalho de Conclusão de Curso de Graduação

Orientador: Professora Joanlise Marco de Leon Andrade

Brasília

2019

Agradecimentos

A conclusão no curso de graduação em estatística é de grande satisfação para mim e, por esse motivo, gostaria de agradecer a todos que me auxiliaram nessa jornada.

Agradeço, primeiramente, por ter estudado em uma instituição de excelência como a UnB. Agradeço ao departamento de Estatística, nele há grandes profissionais que realizam um trabalho excepcional e buscam engradecer a profissão. Aos professores que dedicam boa parte de seu tempo e energia para trazer sempre o melhor para os alunos. Em especial à professora Joanlise, que foi uma grande orientadora, por todo apoio, todos ou auxílios e por ter sido uma grande amiga em todo esse processo.

Agradeço à ESTAT por todas os aprendizados, oportunidades e por boa parte do meu crescimento pessoal e profissional. Ainda, por todas as amizades que contribuíram direta ou indiretamente. Em especial, gostaria de agradecer meus amigos Beatriz, Benjamin, Fernando, Gabriela, Gongora, Jady, Karin, Lima, Luíza, Olivia, Paulo, Ravi, Victória, Tatiana e William.

Agradeço à minha família que sempre trouxe leveza e muito amor para minha vida, tornando os momentos menos difíceis. Em especial, agradeço aos meus pais, Rogério e Renata, meus irmãos, Geovana e Rodrigo, e meu namorado, Pedro, que estiveram cada minuto ao meu lado torcendo por mim.

Agradeço também à Deus por todas as boas oportunidades que tive na vida e por ter colocado as pessoas certas para me acompanhar nessa caminhada.

Por fim, gostaria de agradecer a todos que estiveram na minha vida durante esses quatro anos e que me apoiaram de alguma forma.

Resumo

Dados com informações sobre genótipos, transcritos, proteínas e outros componentes moleculares são chamados de informações moleculares. Avanços em biotecnologia têm possibilitado o estudo de tais informações em larga escala. Nesses contextos adiciona-se o sufixo "ômica" (*omics* em Inglês). Com base no dogma da biologia molecular, entende-se a conexão entre as diferentes informações moleculares em um organismo e a necessidade de considerar mais essas interações. Nesse contexto, tem-se como objetivo principal o estudo da técnica multivariada capaz de lidar com análise supervisionada de dados *multi-omics* implementada no pacote *mixomics*. Este trabalho fornece então uma aplicação das técnicas multivariadas: Análise de Componentes Principais, Análise de discriminante por mínimos quadrados parciais e DIABLO. O banco de dados é um conjunto pacientes de câncer de mama e tal aplicação possui o intuito de entender o funcionamento destas técnicas, uma breve comparação entre elas, com diferentes especificações, utilização de diversas ferramentas gráficas e análise dos resultados de predição. Conclui-se que a utilização de técnicas multivariadas mais sofisticadas pode trazer benefícios para a predição em novas observações, melhor interpretação dos resultados biológicos e boas ferramentas visuais para uma melhor compreensão da complexidade de um organismo.

Palavras-chaves: análise multivariada; estatística; *multi-omics*; *mixomics*; DIABLO; predição; classificação; genética; câncer de mama.

Abstract

Genotypes, transcripts, proteins, and other molecular components data is called molecular information. Advances in biotechnology have enabled the study of such information on a large scale. In these contexts we add the suffix “omics”. Based on the dogma of molecular biology, we understand the connection between the different molecular information in an organism and the need to consider these interactions. In this context, the main objective is to study the multivariate technique capable of dealing with supervised analysis of multi-omics data implemented in the mixomics package. Here we provide an application of multivariate techniques: Principal Component Analysis, Partial Least Squares Discriminant Analysis and DIABLO. The database is a set of breast cancer patients and such application aims to understand the operation of these techniques, a brief comparison between them, with different specifications, use of some graphical tools and analysis of prediction results. It is concluded that the use of more sophisticated multivariate techniques can benefit prediction in new observations, better interpretation of biological results and good visual tools for a better understanding of the complexity of an organism.

Palavras-chaves: multivariate analysis; statistic; multi-omics; mixomics; DIABLO; prediction; classification; genetics; breast cancer.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

(s)pca	*(sparse) Principal Component Analysis*
(s)ipca	*(sparse) Independent Principal Component Analysis*
(s)plsda	*(sparse) Partial Least Squares Discriminant Analysis*
(s)pls	*(sparse) Partial Least Squares*
(s)cca	*(sparse) Canonical Correlation Analysis*
(G)cca	*(Generalized) Canonical Correlation Analysis*
rcca	*regularized Canonical Correlation Analysis*
sGCCA	*sparse Generalized Canonical Correlation Analysis*
DIABLO	*Data Integration Analysis for Biomarker discovery using aLatent component method for Omics studies*
B(s)PLSDA	*Block (sparse) Partial Least Squares Discriminant Analysis - DIABLO*
MINT	*Multivariate Integrative method*
mRNA	RNA mensageiro
miRNA	*micro* RNA
VC	*Validação Cruzada*
PC	*Principal Component*
VL	Variável Latente
Comp	Componente: combinação linear das variáveis originais.
LumA	*Luminal A*
BER	*Balanced Error Rate*

Sumário

1 Introdução e Justificativa

O genoma, formado pelo DNA (ácido desoxirribonucleico, do inglês *desoxyribonucleic acid*), representa o chamado código genético de um organismo. Transmite características hereditárias e contém instruções que coordenam o crescimento, o desenvolvimento e o funcionamento do organismo. O DNA é composto por sequências de milhões de bases nitrogenadas (dos tipos adenina, timina, citosina e guanina, representadas por A, T, G e C, respectivamente) que formam os genes e outras regiões. A produção de proteínas é regulada por genes e utiliza os processos de transcrição e de translação. A transcrição envolve a conversão da informação armazenada em um segmento do DNA para a forma de mRNA (ácido ribonucleido mensageiro, do inglês *messenger ribonucleic acid*), processo que ocorre geralmente no núcleo da célula. Segmentos de mRNA se locomovem pelo citoplasma em direção aos ribossomos, onde a tradução ocorre. Nesse processo, o mRNA pode ser decodificado em uma sequência de aminoácidos que formam uma proteína ou pode ser atribuído para outras funções. Proteínas são macromoléculas essenciais e abundantes e desempenham inúmeras funções para o funcionamento do organismo (3). Exitem ainda outros componentes moleculares, como o miRNA que são sequências de RNA não codificadores que regulam a expressão gênica pelo bloqueio da tradução de mRNAs específicos e provocam a sua degradação.

Genótipos, sequências, transcritos, proteínas e outros componentes moleculares trabalham de maneira orquestrada para o desenvolvimento das mais diversas atividades celulares (3). Avanços em biotecnologia têm possibilitado o estudo de tais informações em larga escala em inúmeros organismos (8). Foram desenvolvidas diversas plataformas que podem medir centenas de milhares a milhões de informações moleculares simultaneamente para cada tipo de componente de sistemas biológicos. Cada tipo de dados requer diferentes metodologias de extração, processamento e análise. Estudos que envolvem tais dados recebem o sufixo ômica/ômico (ou *omics* em Inglês), indicando que a medição é simultânea e **de forma global**. Cita-se como exemplo os estudos de genoma ou genômicos (*genomics* do inglês), que avaliam o código genético de um organismo como um todo. Tais estudos podem ser realizados por meio de sequenciamento genético, que, nos seres humanos, envolveria a leitura de todos os mais de 3,3 bilhões de pares de bases nitrogenadas. Uma outra forma de estudar o genoma inclui a coleta de dados de marcadores genéticos do tipo *SNP's* (*Single Nucleotide Polymorphism*), que são as bases do genoma que tendem a variar entre indivíduos de uma mesma espécie. Atualmente, plataformas de genotipagem de marcadores podem medir até aproximadamente 4,3 milhões de SNPs (4).

Exemplos de estudos que avaliam outros tipos de **informações moleculares** incluem os estudos transcritômicos ou *transcriptomics* (estudos de expressão gênica de

transcritos por todo o genoma), proteômicos ou *proteomics* (estudos de proteínas e variantes proteicas produzidas por todo o genoma), os epigenômicos ou *epigenomics* (estudos de níveis de metilação de sequências de todo o genoma), entre outros (18). Não há muita consistência na utilização dos sufixos *ome* e *omics* e informalmente, tais conjuntos de dados ou estudos são chamados, de forma geral, como *'omics*.

Por muitos anos, estudos em genética avaliaram separadamente relações de fenótipos com genótipos, expressões gênicas, produção de proteínas, entre outros elementos de forma global ou não. Com isso, realizavam-se análises com apenas um dos tipos de informação genética do sistema para avaliar alguma característica ou desfecho de interesse (8), ou seja, as interações entre as informações moleculares e a complexidade do sistema eram pouco explorados.

Estudos que integram dados de diferentes tipos de informações moleculares, chamados de *multi-omics* ou de dados *multi-omics*, apresentam grande potencial para elucidar de maneira mais completa as relações entre fenótipos e informações moleculares e suas interações (8). Tal entendimento tem o potencial para, no caso de uma doença como desfecho de interesse, ampliar as opções de diagnóstico, possibilitar a detecção de subgrupos de casos com prognósticos diferenciados, identificar "assinaturas moleculares" *multi-omics* para melhor entendimento de mecanismos biológicos que levam a fenótipos de interesse e, assim, desenvolver formas de tratamentos mais personalizados. Porém, tais análises apresentam diversos desafios incluindo o gerenciamento e análise de dados superdimensionados com naturezas heterogêneas, dificuldades computacionais e interpretações biológicas (11).

Nesse contexto, técnicas multivariadas e de aprendizagem de máquinas vêm sendo aplicadas e aprimoradas para uma avaliação global de sistemas biológicos compostos por diferentes entidades moleculares que agem conjuntamente.

O presente trabalho tem por objetivo o estudo e a aplicação de algumas técnicas multivariadas em análises *multi-omics* implementadas no pacote `mixOmics` (11) do *software* R (15) em um banco de dados.

2 Metodologia

Métodos multivariados são comumente utilizados em análises de dados *multi-omics*, pois são uma solução para casos nos quais se deseja avaliar um grande número de variáveis (10). O pacote `mixOmics`, do *software* R, inclui atualmente 19 métodos de análise de dados *multi-omics*. Tais métodos foram implementados principalmente para análises de integração desses dados e seleção das principais variáveis como apresentado na Figura 1.

Tabela 1 – Métodos multivariados para análise de dados *multi-omics* incluídos no pacote `mixOmics`.

Banco de dados		Função	Esparsa	Predição
Um 'omics	Não supervisionado	pca	-	-
		ipca	-	-
		sipca	✓	-
		spca	✓	-
	Supervisionado	plsda	-	✓
		splsda	✓	✓
N-integrado	Supervisionado (2 'omics)	rcca	-	-
		plsda	-	✓
		splsda	✓	✓
	Não supervisionado	wrapper.rgcca	-	-
		wrapper.sgcca	✓	✓
		block.pls	-	✓
		block.spls	✓	✓
	Supervisionado (DIABLO)	block.plsda	-	✓
		block.splsda	✓	✓
P-integrado (MINT)	Não supervisionado	mint.pls	-	✓
		mint.spls	✓	✓
	Supervisionado	mint.plsda	-	✓
		mint.splsda	✓	✓

Fonte: Rohart F, Gautier B, Singh, M, Lê Cao K-A. mixOmics: an R package for 'omics feature selection and multiple data integration. PLoS Comp Biol 13 (16).

O pacote `mixOmics` permite a utilização de análises supervisionadas ou não. O caso supervisionado é utilizado quando a variável resposta é categorizada, como, por exemplo, um estudo do tipo caso-controle. Análises não-supervisionadas são utilizadas principalmente quando não se tem interesse em comparar diferentes grupos, sendo então

mais utilizadas para análises exploratórias dos dados (11).

Além disso, dois tipos de integração de dados moleculares estão disponíveis. Bancos de dados P-integrados são formados por amostras diferentes para um mesmo tipo de variável (mesmo tipo de informação molecular como mRNA, miRNA, sequências de metilação, proteínas entre outros), ou envolvendo apenas um *'omics*. Já os dados N-integrados (ou *multi-omics*) envolvem a junção de bancos de dados *'omics* distintos para uma mesma amostra (mesmos indivíduos). Cada banco de dados *'omics* distinto será aqui referido também como **bloco**.

A escolha do método a ser aplicado é determinada pelos objetivos e características do trabalho (10). Ainda, é interessante que cada banco de dados passe por processos de redução de variáveis (filtragem). Além disso, utiliza-se análises gráficas para se entender o comportamento das observações (comumente denominadas amostras, nesse contexto) e das variáveis (11).

Este trabalho concentrou-se na utilização de métodos exploratórios de análise multivariada e no estudo da metodologia desenvolvida por Amrit Singh et al. (1) e implementada no pacote `mixOmics` para dados supervisionados com a integração do tipo "N" (para dados *multi-omics*).

Todas as análises estatísticas foram realizadas pelo *software* R, versão 3.6.0 (15).

2.1 Análise de componentes principais

O método de análise de componentes principais (ACP, do inglês *Principal Component Analysis* ou PCA) envolve a decomposição de um conjunto de variáveis em componentes resultantes de combinações lineares ortogonais das variáveis de modo a extrair o máximo de sua variabilidade. Utiliza-se uma quantidade k de componentes principais (CP), menor que o número de variáveis, que expliquem um percentual alto da variabilidade do conjunto de variáveis originais.

No contexto de dados *multi-omics*, a ACP pode ser utilizada para descrever a estrutura dos dados, reduzir a dimensão do banco de dados tentando não perder muita informação (utilizando-se poucas componentes principais no lugar das variáveis), verificar se existe algum agrupamento natural de condições biológicas, analisar o quanto da variância é explicada por cada componente e analisar a importância de cada variável em uma dada componente (11).

Mais detalhadamente, a ACP é baseada na obtenção de combinações lineares das p variáveis aleatórias $X_1, X_2, \ldots, X_p$ originais. Geometricamente, isso representa a seleção de um novo sistema de coordenadas em que os novos eixos representam as direções com a maior variabilidade e fornecem uma descrição mais simples e mais parcimoniosa da

estrutura de covariância (10).

A técnica depende somente da matriz de covariâncias $\mathbf{\Sigma}$ (ou de correlações ρ, quando há muita diferença nas escalas das variáveis) de $X_1, X_2, \ldots, X_p$ e, à princípio, não é necessário que tais variáveis possuam normalidade multivariada (10).

Considera-se o vetor aleatório $\mathbf{X}' = \{X_1, X_2, \ldots, X_p\}$ com matriz de covariâncias $\mathbf{\Sigma}$, com pares de autovalores-autovetores $(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), \ldots, (\lambda_p, \mathbf{e}_p)$, em que $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$, e com as combinações lineares da seguintes forma:

$$Y_i = \mathbf{a_i'X} = a_{i1}X_1 + a_{i2}X_2 + \cdots + a_{ip}X_p.$$

As componentes principais serão as combinações lineares não correlacionadas, as quais terão a maior $Var(Y_i) = \mathbf{a_i'\Sigma a_i}$ possível. Esse critério é atendido quando $\mathbf{a_i} = \mathbf{e_i}$ (10). As CPs são melhor representadas em função dos autovetores, da seguinte maneira:

$$\begin{cases} Y_1 = & \mathbf{e_1'X} = e_{11}X_1 + e_{12}X_2 + \cdots + e_{1p}X_p, \\ \vdots & \\ Y_p = & \mathbf{e_p'X} = e_{p1}X_1 + e_{p2}X_2 + \cdots + e_{pp}X_p. \end{cases}$$

Assim, tem-se que:

$$Var(Y_i) = \mathbf{e_i'\Sigma e_i} = \lambda_i, \ i = 1, 2, \ldots, p$$

e

$$Cov(Y_i, Y_k) = \mathbf{e_i'\Sigma e_k} = 0, \ i \neq k.$$

Se alguns autovalores λ_i forem iguais, então a escolha dos autovetores $\mathbf{e_i}$ e, consequentemente, das componentes Y_i não será única.

2.2 Análise de Correlação Canônica

A análise de correlação canônica procura identificar e quantificar as associações entre dois conjuntos de variáveis com base em correlações entre suas combinações lineares. Não é um método destinado à seleção de variáveis (10), mas procura transformar um conjunto muito grande de interações que existe entre dois conjuntos de variáveis em alguns poucos pares de combinações lineares e, então, maximizar a correlação entre tais elementos. Os pares a serem selecionados e suas correlações são conhecidos, respectivamente, como variáveis canônicas e correlações canônicas.

No contexto de dados *multi-omics*, esta análise é utilizada para a obtenção de correlações entre as combinações lineares das variáveis de cada *'omics*, buscando explicar de

maneira resumida as interações ("assinaturas") entre os diferentes conjuntos de informações moleculares.

Na análise de correlação canônica, considera-se dois grupos de variáveis que são representados pelos vetores aleatórios $\mathbf{X}_1$ $(p \times 1)$ e $\mathbf{X}_2$ $(q \times 1)$, com $p \leq q$, indicando, respectivamente, o menor e o maior conjunto (com mais variáveis entre os dois conjuntos) com base nos quais calcula-se algumas medidas:

$$E(\mathbf{X}_i) = \mu_i, \text{ i=1,2;}$$

$$Var(\mathbf{X}_i) = Cov(X_i, X_i) = \mathbf{\Sigma}_{ii}, \text{ i=1,2;}$$

$$Cov(\mathbf{X}_1, \mathbf{X}_2) = \mathbf{\Sigma}_{12} = \mathbf{\Sigma}'_{21}.$$

O aspecto de maximização da correlação na técnica representa uma tentativa de se obter combinações lineares U e V, que irão resumir da melhor maneira possível essa associação entre os blocos de variáveis em poucas componentes, considerando os vetores constantes, $\mathbf{a}$ e $\mathbf{b}$, da maneira a seguir:

$$U = \mathbf{a}'\mathbf{X}_1 \text{ e}$$

$$V = \mathbf{b}'\mathbf{X}_2,$$

nas quais os resultados subsequentes são apresentados, considerando Σ_{ij} como a matriz de covariâncias entre $\mathbf{X_i}$ e $\mathbf{X_j}$, como:

$$Var(U) = \mathbf{a}'\mathbf{\Sigma}_{11}\mathbf{a};$$

$$Var(V) = \mathbf{b}'\mathbf{\Sigma}_{22}\mathbf{b};$$

$$Cov(U, V) = \mathbf{a}'\mathbf{\Sigma}_{12}\mathbf{b}.$$

As estimativas para $\mathbf{a}$ e $\mathbf{b}$ são dadas de tal forma que maximizem a função de correlação observada a seguir:

$$Corr(U, V) = \frac{\mathbf{a}'\mathbf{\Sigma}_{12}\mathbf{b}}{\sqrt{\mathbf{a}'\mathbf{\Sigma}_{11}\mathbf{a}}\sqrt{\mathbf{b}'\mathbf{\Sigma}_{22}\mathbf{b}}}. \tag{2.1}$$

O próximo passo envolve encontrar o primeiro par de combinações lineares U_1 e V_1, tendo variâncias unitárias (padronização) (10), que maximizem a correlação em (2.1) e, caso seja de interesse, encontra-se o segundo par, entre todas as escolhas não correlacionadas com o par escolhido inicialmente, da mesma maneira e assim por diante.

O k-ésimo par de variáveis canônicas $(k = 1, 2, \ldots, min(p, q))$ que maximiza $Corr(U_k, V_k) = \rho_k$ é dado por:

$$U_k = \mathbf{e}'_k\mathbf{\Sigma}_{11}^{-1/2}\mathbf{X}_1,$$

$$V_k = \mathbf{f}_k' \mathbf{\Sigma}_{22}^{-1/2} \mathbf{X}_2,$$

em que:

- $\rho_1^2 \geq \rho_2^2 \geq \ldots \rho_p^2$ são autovalores de $\mathbf{\Sigma}_{11}^{-1/2} \mathbf{\Sigma}_{12} \mathbf{\Sigma}_{22}^{-1} \mathbf{\Sigma}_{21} \mathbf{\Sigma}_{11}^{-1/2}$;
- $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p$ são os correspondentes autovetores $(p \times 1)$;
- $\rho_1^2 \geq \rho_2^2 \geq \ldots \rho_p^2$ são também autovalores de $\mathbf{\Sigma}_{22}^{-1/2} \mathbf{\Sigma}_{21} \mathbf{\Sigma}_{11}^{-1} \mathbf{\Sigma}_{12} \mathbf{\Sigma}_{22}^{-1/2}$;
- $\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_p$ são os correspondentes autovetores $(q \times 1)$;
- Cada $\mathbf{f}_i$ é proporcional a $\mathbf{\Sigma}_{22}^{-1/2} \mathbf{\Sigma}_{21} \mathbf{\Sigma}_{11}^{-1/2} e_i$.

Além disso, o coeficiente de correlação canônico pode ser invariante em casos de padronização das variáveis originais.

Essa análise pode ser estendida para um caso generalizado, no qual considera-se correlacionar mais de dois conjuntos de variáveis. Ainda, pode-se utilizá-la de maneira esparsa, buscando selecionar apenas algumas das variáveis que forem mais importantes na associação com os outros blocos (17).

2.3 Análise de discriminante

A análise de discriminante é uma técnica multivariada supervisionada que visa a identificação de funções que podem sem utilizadas para a classificação de novas observações nos diferentes grupos, de acordo com as características apresentadas (10). Busca-se então para a construção de tais funções a minimização do erro de classificação para essas novas observações.

Esse método tem por objetivo a criação de uma regra de classificação estatística, linear ou não, utilizando as variáveis disponíveis no banco de dados e a classificação feita previamente. O problema consiste em se obter uma combinação linear de características observadas que apresente maior poder de classificar ou de alocar novas observações em um dos diferentes grupos previamente definidos.

Uma técnica muito conhecida é a decomposição linear de Fisher que equivale a decomposição da variabilidade em variâncias dentro e entre grupos. Porém, além de possuir fortes suposições, pode apresentar problemas computacionais em casos de muitas variáveis preditoras correlacionadas. Já a análise discriminante por mínimos quadrados parciais, por exemplo, é mais utilizada por possuir melhor desempenho computacional (10).

No contexto *multi-omics*, tem-se o objetivo de encontrar um conjunto de regras que ajudem a identificar um grupo de informações moleculares que mais influenciem

na discriminação de um fenótipo, para que estas auxiliem na predição do fenótipo de interesse(16).

2.4 Projeção para a estrutura latente (PLS) com abordagem de Análise de Discriminante (PLS-DA)

O método de projeção para a estrutura latente (PLS, do inglês *Projection on Latent Structure*), também conhecido como regressão de mínimos quadrados parciais (PLS regression, do inglês *Partial Least Squares regression*), possui características de análise de regressão linear múltipla e análise de componentes principais (PCA), além de suas generalizações (9).

O objetivo de tal método envolve extrair dos preditores um conjunto de fatores ortogonais, chamados de variáveis latentes (VL), com melhor poder para a previsão das variáveis dependentes (9). Para isso, considera-se uma matriz $\mathbf{Y}_{IxK}$ com I observações para as K variáveis dependentes e, com os J preditores com I observações, tem-se a matriz $\mathbf{X}_{IxJ}$. Em outras palavras, o objetivo é prever a matriz $\mathbf{Y}$ a partir da $\mathbf{X}$ e descrever a estrutura comum entre elas (9).

A ACP busca decompor a matriz $\mathbf{X}$ em componentes que expliquem da melhor maneira a sua variação. Já no método PLS, obtém-se componentes que não só explicam a maior parte da variação mas que também forneçam boa predição de $\mathbf{Y}$. Então, realiza-se a decomposição simultânea das duas matrizes, $\mathbf{X}$ e $\mathbf{Y}$, de forma que as variáveis latentes cumpram, ao mesmo tempo, os dois objetivos citados acima, ou seja, expliquem o máximo da variação das variáveis preditoras e da covariância entre as duas matrizes (9). O próximo passo, em geral, envolve a estimação da matriz $\mathbf{Y}$.

Isso equivale a encontrar dois conjuntos de pesos, denotados $\mathbf{a}$ e $\mathbf{b}$, para criar, respectivamente, uma combinação linear das colunas de $\mathbf{X}$ e de $\mathbf{Y}$, de modo que essas duas combinações lineares tenham covariância máxima (9). Depois de obtido, o primeiro par de vetores é retirado das matrizes e essa etapa se repete até que $\mathbf{X}$ seja uma matriz nula (9).

O PLS com uma abordagem de análise de discriminante (PLS-DA, do inglês *Projection on Latent Structure - Discriminant Analysis*) é uma extensão natural do PLS que utiliza como bloco de variável resposta uma matriz dummy de variáveis indicadoras das classes de uma variável explicativa qualitativa, ou seja, é uma versão supervisionada do PLS (16). Nesse caso, os objetivos envolvem a discriminação das classes da variável resposta e a maximização da covariância entre a variável resposta e as variáveis preditoras, como no PLS.

O método PLS-DA constrói H sucessivas VLs, que são combinações lineares da forma $\mathbf{t}_h = \mathbf{X}_h\mathbf{a}_h$ e $\mathbf{u}_h = \mathbf{Y}_h\mathbf{b}_h$, das variáveis de $\mathbf{X}$ e $\mathbf{Y}$, respectivamente, para $h = 1, \ldots, H$.

Tais VLs representam projeções dos dados. Os vetores de coeficientes (ou pesos),

O método DIABLO (*Data Integration Analysis for Biomarker discovery using a Latent component method for Omics studies*) consiste em uma técnica nova desenvolvida por Amrit Singh et al. (1) e implementada no pacote `mixOmics`.

Conhecido também como Block-sPLSDA (BsPLSDA), foi proposto para bancos de dados N-integrados e supervisionados, baseando-se a técnica em dois métodos estendidos: Análise de Correlação Canônica Generalizada Esparsa para uma abordagem de Análise de Discriminante (sGCCA) e Análise de Discriminante por Projeção de Estrutura Latente Esparsa (sPLS-DA) (1). Isso é realizado de tal forma que a técnica seleciona as variáveis mais correlacionadas entre os blocos (bancos de dados de elementos moleculares ou *'omics* distintos), utilizando a sGCCA, buscando ao mesmo tempo associar tais variáveis latentes à variável resposta categorizada por meio do sPLS-DA. O termo esparso nesse contexto implica que tais métodos permitem que sejam utilizadas apenas algumas das variáveis (e não todas) de cada bloco.

A maximização da informação correlacionada entre os vários blocos de variáveis busca identificar redes moleculares, elegendo de maneira ótima as principais variáveis que explicam e classificam entre os grupos de interesse (por exemplo: doentes ou não doentes). Com isso, tem-se modelos preditivos que podem ser aplicados a novas observações para classifica-las. Além disso, a ferramenta permite a especificação de alguns argumentos como uma matriz de delineamento (*design*), o número de componentes, a seleção de um número de variáveis, bem como a avaliação do desempenho preditivo do modelo e a produção de gráficos e diagramas para visualização dos dados, que auxiliam na interpretação dos complexos resultados.

Inicialmente, é necessária a realização de métodos para controle de qualidade, normalização e o pré-processamento dos dados (a nível de *probe* ou sequência). Vários métodos existem para tais tratamentos e variam conforme o tipo de dados *'omics* a ser utilizado. Análises exploratórias também devem ser realizadas para avaliar o comportamento das observações e das variáveis.

A integração e análise dos dados *multi-omics* pode ser então realizada e ser precedida por alguns procedimentos:

1. O primeiro passo envolve a definição da matriz de delineamento (*design*) de relação entre os blocos, aqui referida como **matriz *design* de blocos** com valores podendo variar entre 0 e 1. Tais valores podem ser escolhidos com base em conhecimento prévio do pesquisador ou de acordo com o que os dados mostraram em outras análises não supervisionadas. A escolha dos valores visa um equilíbrio entre a maximização da correlação entre os blocos e a maximização da discriminação entre as classes

(fenótipos) (16). Se o maior interesse for maximizar a relação entre os blocos, então os valores para as relações entre os blocos devem estar entre $0,5$ e 1. Caso seja discriminar melhor as classes, então esses valores devem estar entre 0 e $0,5$ (16). Os valores referentes as relações da variável resposta com os blocos será igual a 1.

2. Um modelo Block-PLSDA pode ser implementado utilizando todas as componentes possíveis ou pelo menos com uma quantidade maior do que a idealizada para o modelo final. Estas componentes incluem todas as variáveis. Este modelo será chamado de **BPLSDA1**.

3. Define-se o número de componentes para cada bloco. Pode-se utilizar valores arbitrários, sugeridos por analises exploratórias como APC ou PLS-DA, ou obtidos por um processo de calibração. Tal processo se baseia no modelo BPLSDA1 e pode envolver um método de validação cruzada para avaliar a performance do modelo. O método *k-fold* é o padrão utilizado para esse caso no pacote *mixomics*. Este processo é feito de maneira análoga à **VC1-PLSDA**. O procedimento de validação cruzada utilizada para definir o número de componentes será denominado **validação cruzada 1** ou **VC1-BPLSDA**;

4. Define-se então o número de variáveis que serão selecionadas para cada componente de cada bloco. Para tanto pode-se utilizar um processo de validação cruzada análogo ao do passo anterior, porém já considerando o número de componentes obtido (ou estabelecido arbitrariamente). Considera-se também um vetor de quantidade de variáveis a serem testadas. Esta etapa é denominada de **validação cruzada 2** ou **VC2-BsPLSDA**.

5. Um modelo BsPLSDA é implementado aos dados, com base na matriz *design* de relação entre os blocos, no número de componentes e no número de variáveis de cada componente, definidos nos passos anteriores. Tal modelo será considerado como **modelo final BsPLSDA**.

6. O modelo pode ser avaliado por meio de gráficos e pela taxa de erro de classificação para predição nos dados de teste.

A metodologia BsPLSDA assume K conjuntos de dados normalizados, centrados e padronizados, $\mathbf{X}_1(n \times p_1), \ldots, \mathbf{X}_K(n \times p_K)$, com dados das respectivas $p_1, \ldots, p_K$ variáveis de cada bloco, todas medidas em uma mesma amostra com n indivíduos. Com o intuito de identificar as variáveis altamente correlacionadas entre e dentro de cada bloco, utiliza-se

sGCCA para encontrar uma solução ótima, de acordo com a equação a seguir (1) (referente à primeira dimensão $h = 1$):

$$\begin{aligned} &\max_{\mathbf{a}_1^{(h)},\ldots,\mathbf{a}_K^{(h)}} \quad \sum_{k,j=1,k\neq j}^{K} c_{k,j} Cov(\mathbf{X}_k^{(h)}\mathbf{a}_k^{(h)}, \mathbf{X}_j^{(h)}\mathbf{a}_j^{(h)}) \\ &\text{sujeito a} \quad ||\mathbf{a}_k^{(h)}||_2 = 1 \text{ e } ||\mathbf{a}_k^{(h)}||_1 < \beta_k^{(h)}, \end{aligned} \tag{2.2}$$

em que

- $c_{k,j}$ é referente aos valores da matriz *design* de relação entre os blocos e indica se deve maximizar a correlação entre $\mathbf{X}_k$ e $\mathbf{X}_j$;
- $\mathbf{a}_k$ é o vetor de coeficientes para cada bloco $\mathbf{X}_k$;
- β_k é um parâmetro de penalização não-negativo que controla a quantidade de coeficientes não nulos em $\mathbf{a}_k$, responsáveis pela redução de dimensão;

A solução dessa função é obtida utilizando um algoritmo de convergência monótona (16). A Equação 2.2 é referente a uma dimensão (h) e, uma vez que se possui o vetor de coeficientes $\mathbf{a}_k^{(h)}$ e a componente de escores associada $\mathbf{t}_k^{(h)} = \mathbf{X}_k^{(h)}\mathbf{a}_k^{(h)}$, matrizes residuais são calculadas para a próxima dimensão, como $\mathbf{X}_k^{(h+1)} = \mathbf{X}_k^{(h)} - \mathbf{t}_k^{(h)}\mathbf{a}_k^{(h)}$. Utiliza-se então a nova matriz $\mathbf{X}_k^{(h+1)}$ na equação 2.2 e repete-se o processo até se atingir a quantidade de componentes necessárias, quantidade esta decidida na etapa 3. A penalização do tipo l_1 no vetor de escores $\mathbf{a}_k^{(h)}$ que permite interpretação das componentes $\mathbf{t}_k = \mathbf{X}_k\mathbf{a}_k$ de escores (1).

A principal suposição deste modelo é que a maior parte da fonte de variação comum entre os blocos pode ser extraída pelas componentes de escores $\mathbf{t}_k$, enquanto que as heterogeneidades naturais não desejadas entre os tipos de *omics* não causem impacto no modelo (1).

Uma vez que um dos objetivos da técnica é abordar uma perspectiva de classificação (discriminação), utiliza-se uma extensão do sGCCA substituindo um dos blocos pelas informações de uma variável resposta qualitativa, a qual é transformada em uma matriz ($n \times F$) de variáveis indicadoras (*dummy*[1]) dos F subgrupos desta variável. Substitui-se ainda o parâmetro $\beta_k^{(h)}$, da penalização l_1, pelos números de variáveis que serão selecionadas de cada bloco (k) para cada componente (h), os quais foram definido na etapa 4. Isto pode ser feito já que há um ligação direta entre os dois parâmetros.

Uma nova observação j, a qual é medida para os K diferentes blocos $\widetilde{\mathbf{X}_k}$, é pertencente a uma classe predita pelo modelo final, utilizando os vetores de coeficientes

[1] Cada categoria da variável resposta $\mathbf{Y}$ é expressa como uma coluna, na qual as linhas indicam se a observação pertence ou não a categoria em questão

estimados $\hat{\mathbf{a}}_k$ para as variáveis, para que se obtenha os escores preditos $\mathbf{t}_{k;j} = \widetilde{\mathbf{X}_k^j}\hat{\mathbf{a}}_k$, $k = 1, \ldots, K$.

Para cada bloco é obtida uma classe predita (considerada como um voto) para a nova observação j, a partir dos escores preditos, usando uma das distâncias: Máxima, Centróide (ou Euclidiana) ou Mahalanobis. Há estudos indicando que, na prática, distâncias baseadas no centróide, como as distâncias centróide e mahalanobis, levam a resultados mais precisos que a distância máxima para a integração do tipo "N" (16). As três estão disponíveis no pacote *mixomics*.

Figura 1 – Resumo do método

*Predição por maioria ou média dos votos

Fonte: Amrit Singh et al. (2016) (1)

A classe final é decidida pela maioria dos votos ou pela média dos escores ($\mathbf{t}_{k;j}$) de todos os blocos. Quando ocorrem empates a utilização da maioria dos votos, irá atribuir valor faltante (NA) à classe predita e esta será considerada como um erro de classificação na avaliação da performance do modelo. Por esta razão, optou-se pela média dos escores para a classificação final das novas observações. Como a classe é baseada no voto de cada bloco separadamente, o método BsPLSDA permite a predição para novas amostras mesmo se faltar algum dos blocos.

A metodologia Block-sPLSDA diferencia-se do método sPLS-DA principalmente pelo fato de envolver uma técnica capaz de lidar com a integração de blocos e suas interações. A Figura 1 apresenta um resumo desta metodologia.

2.5 Banco de Dados

O câncer de mama é um dos tipos de câncer mais comuns, com quase 17 mil mortes em 2017 (SIM) e com uma previsão de mais de 59 mil novos casos em 2018 (INCA)[2]. É uma doença considerada heterogênea, causada pela multiplicação desordenada

[2] https://www.inca.gov.br/numeros-de-cancer

de células anormais da mama, formando um tumor. Nesse processo, alguns elementos genéticos (hereditários ou não) são considerados como possíveis fatores de risco para o desenvolvimento da doença[3].

Para uma aplicação do método BsPLSDA implementado no pacote *mixOmics* (11), utilizou-se um banco de dados (1) proveniente de um subconjunto de dados gerados pelo "Cancer Genome Atlas Network" (Network, 2012 (5)). O programa "*The Cancer Genome Atlas* (TCGA)" é tradicional em estudos *genomics* para o câncer. Construído por um esforço conjunto entre o Instituto Nacional do Câncer e o Instituto Nacional de Pesquisa do Genoma Humano desde 2006. Hoje, o TCGA gerou mais de 2,5 petabytes de dados genômicos, epigenômicos, transcriptômicos e proteômicos, com contribuições de mais de 11 mil pacientes e muito esforço de milhares de pesquisadores, os quais já levaram a melhorias na capacidade de diagnosticar, tratar e prevenir o câncer, e estão disponíveis para comunidade de pesquisa (14). Portanto representa uma excelente fonte para estudos *multi-omics*.

O conjunto de dados completo e original, contendo 5 subtipos de câncer e 5 diferentes conjuntos de variáveis, está em "*level 3 TCGA*" (versão 2015_11_01) e foi recuperado do site que é mantido pelo *Broad Institute*, `firebrowse.org`. O banco de dados foi utilizado por Amrit Singh et al. (1), sendo pré-processado e normalizado, além de ter sido separado em treinamento e teste. Porém, para o presente estudo foi utilizado apenas um subconjunto desses dados, com três subtipos de câncer e três *'omics*.

Tal subconjunto do banco de dados é dividido em duas partes: 150 observações para treinamento e 70, para testes, classificando-as em três subtipos da doença, que são diferenciados principalmente por características do tumor e que levam a diferentes prognósticos e tratamentos. São eles:

- **Luminal A:** tumor com receptor hormonal positivo (estrogênio e/ou progesterona), HER2[4] negativo e baixos níveis da proteína Ki-67, com crescimento mais lento (13);

- **HER2-enriquecido**: tumor com receptor hormonal negativo (estrogênio e progesterona) e HER2 positivo. Tendem a crescer mais rápido que o tipo anterior, porém os tratamentos são frequentemente bem sucedidos por atingirem e bloquearem a proteína HER2 (13);

- **Basal/Triplo negativo**: tumor com receptor hormonal negativo (estrogênio e progesterona) e HER2 negativo. Esse é o tipo mais comum em mulheres com a mutação no gene *BRCA1* (13).

[3] https://www.inca.gov.br/tipos-de-cancer/cancer-de-mama
[4] Proteína conhecida como *"human epidermal growth factor receptor 2 (HER2)"*

Nos dados de treinamento, estão disponíveis 3 conjuntos de variáveis de elementos moleculares (*'omics*) distintos: *mRNA* (200 variáveis), *miRNA* (184 variáveis) e *proteomics* (142 variáveis). Já para os dados de teste, tem-se apenas informações sobre os 2 primeiros *'omics* (200 e 184 variáveis, respectivamente). Tais variáveis foram previamente normalizadas e pré-filtradas (1). Esses dados foram apresentados de forma N-integrada e podem ser melhor compreendidos na Figura 2.

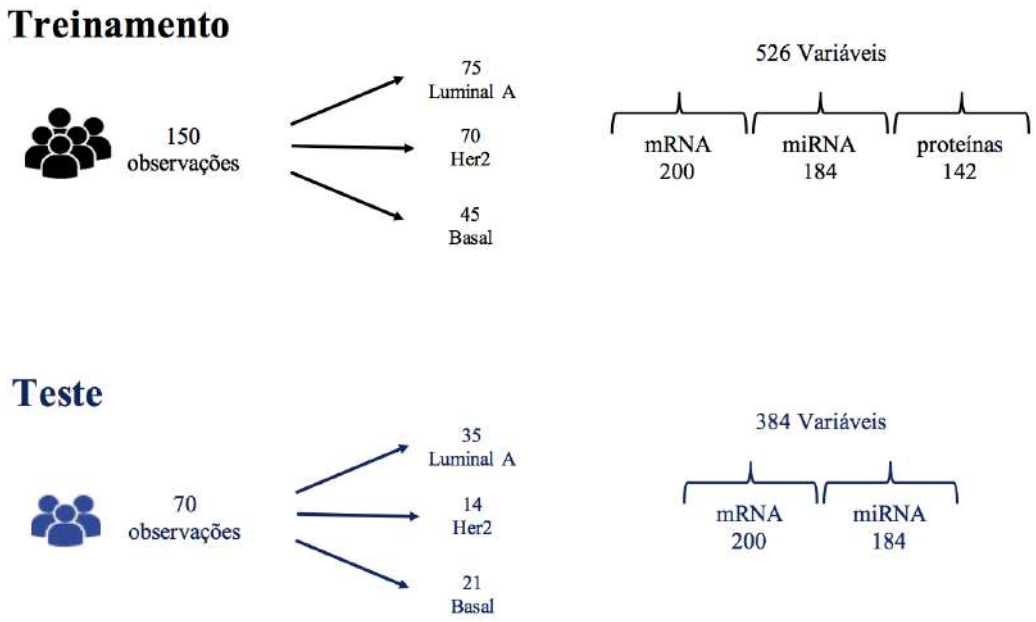

Figura 2 – Diagrama com a composição de grupos de treinamento e teste, tipos e número de variáveis incluídos no banco de dados utilizado no presente trabalho (Dados provenientes do estudo Amrit Singh et. al (1))

O *download* dos dados clínicos originais (Merge_Clinical) foram da tabela original do BRCA Clinical Archives. Os dados originais de cada conjunto de dados *'omics* utilizados neste subconjunto foram obtidos da seguinte maneira: o banco mRNA RSEM normalizado (illuminahiseq_rnaseqv2-RSEM_genes_normalized) foi baixado da tabela original do BRCA mRNASeq Archives; os dados de miRNA (illuminahiseq_ mirnaseqmiR_gene_ expression and illuminaga_mirnaseq-miR_gene_expression) foram obtidos pelo *download* da tabela original do BRCA miRSeq Archives; e a matriz de dados de proteína de fase reversa (mda_rppa_core-protein_normalization) são da tabela original do BRCA RPPA Archives.

3 Resultados

Uma variedade de metodologias podem ser utilizadas para o estudo de dados com informações moleculares. Técnicas multivariadas são muito utilizadas nesse contexto devido ao grande número de variáveis incluídas e suas relações de interdependência. O avanço de plataformas para a análise de informações moleculares bem como a possibilidade de coleta de dados de diferentes naturezas tem fornecido motivação para o desenvolvimento extensões e generalizações de diversas técnicas.

Nessas condições, a abordagem *multi-omics* ganhou bastante espaço e, com ela, além de técnicas tradicionais, utiliza-se novas metodologias, como a estudada neste trabalho que visa analisar uma N-integração de dados de diferentes naturezas de forma supervisionada.

Assim, inicialmente analisou-se de maneira exploratória separadamente cada um dos blocos. Essas análises podem ser feitas de maneira supervisionada ou não, buscando uma melhor visualização do comportamento das amostras, das variáveis e suas fontes de maior variação. Com esse intuito, análises são realizada por meio de ACP e PLS-DA.

Ainda, utiliza-se a metodologia BsPLSDA, que consiste em uma combinação de técnicas multivariadas, para estudar um caso mais geral. Busca-se, com isso, avaliar as interações entre os diferentes blocos com o objetivo de melhor prever algum desfecho de interesse. Ao mesmo tempo em que se tenta ter uma redução de dimensões, espera-se resultados de predição semelhantes aos que se obteria utilizando todos os dados originais.

Para comparação dos métodos utilizou-se algumas medidas de maneira padronizada. A Taxa de Erro Balanceada foi escolhida, na maior parte das vezes gerou resultados melhores que a Geral. A medida de distância centróide, pela comparação dos resultados obtidos com os da Mahalanobis, já que ambas são as mais indicadas para o caso do BsPLSDA (16). Utilizou-se 5-fold com 10 repetições para todos os resultados de validação cruzada deste trabalho.

3.1 Análise de Componentes Principais

Análises de componentes principais foram aplicadas nos blocos (mRNA, miRNA e proteínas) de treinamento separadamente, com o intuito de explorar as variáveis e identificar as maiores fontes de variação. Para isso, a informação sobre o subtipo de câncer dos pacientes foi desconsiderada para os cálculos, já que esta é uma análise não supervisionada.

Analisou-se então as dez primeiras componentes, calculadas de forma centrada, com base na matriz de covariâncias, a partir das contribuições individuais de cada uma

na variação total dos dados (Figura 3) e percentuais acumulados explicados da variância total explicada (Tabela 2).

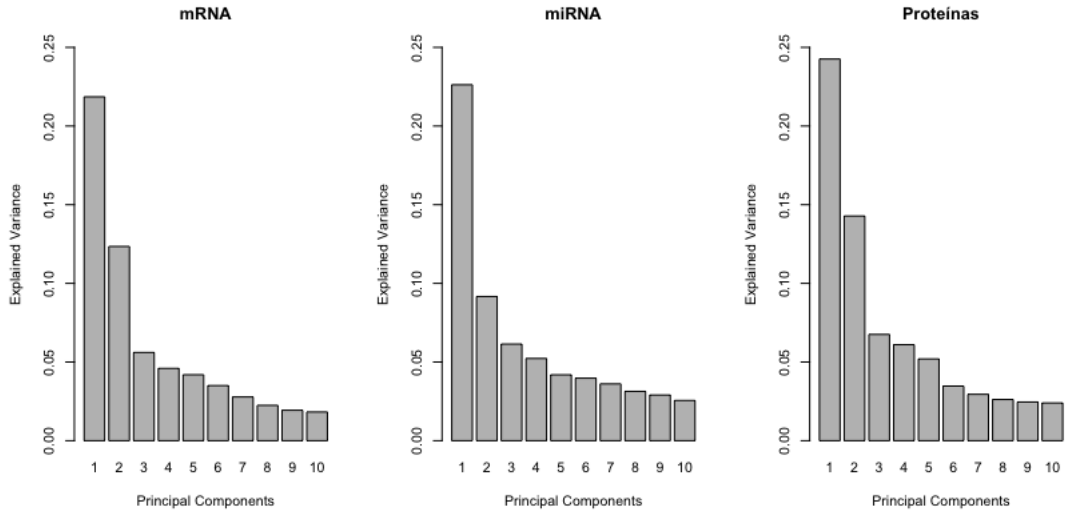

Figura 3 – Gráfico do percentual da variância explicada por CP

Tabela 2 – Percentual acumulado da variância explicada por CP

	mRNA	miRNA	Proteínas
PC1	0,22	0,23	0,24
PC2	0,34	0,32	0,39
PC3	0,40	0,38	0,45
PC4	0,44	0,43	0,51
PC5	0,49	0,47	0,57
PC6	0,52	0,51	0,60
PC7	0,55	0,55	0,63
PC8	0,57	0,58	0,66
PC9	0,59	0,61	0,68
PC10	0,61	0,64	0,70

Nos três blocos, obteve-se que a primeira componente de cada explicou entre 20% e 25% da variância. As duas primeiras CPs, nos três casos, mesmo sendo as que mais explicaram, acumularam um percentual de variação explicada muito baixo. A Tabela 2 evidenciou a necessidade de se utilizar pelo menos seis CPs para conseguir explicar mais que 50% da variância nos dados de mRNA e miRNA e 4 CPs no caso das proteínas. Nessas situações, nas quais se analisa dados de informações moleculares, é comum que as componentes expliquem pouco, pois há um número grande de variáveis de efeito pequeno a moderado, sendo que estas ainda podem ter comportamentos bastante heterogêneos.

A Figura 4 apresenta a relação entre as duas primeiras componentes de cada um dos blocos do conjunto de treinamento. Destaca-se que, tais componentes foram construídas sem levar em consideração os subtipos de câncer, mas é possível observar que há diferenças

no comportamento entre tais grupos, principalmente para mRNA e Proteínas. A primeira componente parece ser fortemente influenciada pelos subtipos de câncer, sugerindo uma associação desses elementos moleculares com esse grupos considerados. Porém, não se observa diferenças entre os grupos na segunda componente. A contribuição de tais componentes entre os blocos foi próxima (entre 22% e 24% para as primeiras componentes e entre 9% e 14% para as segundas componentes.

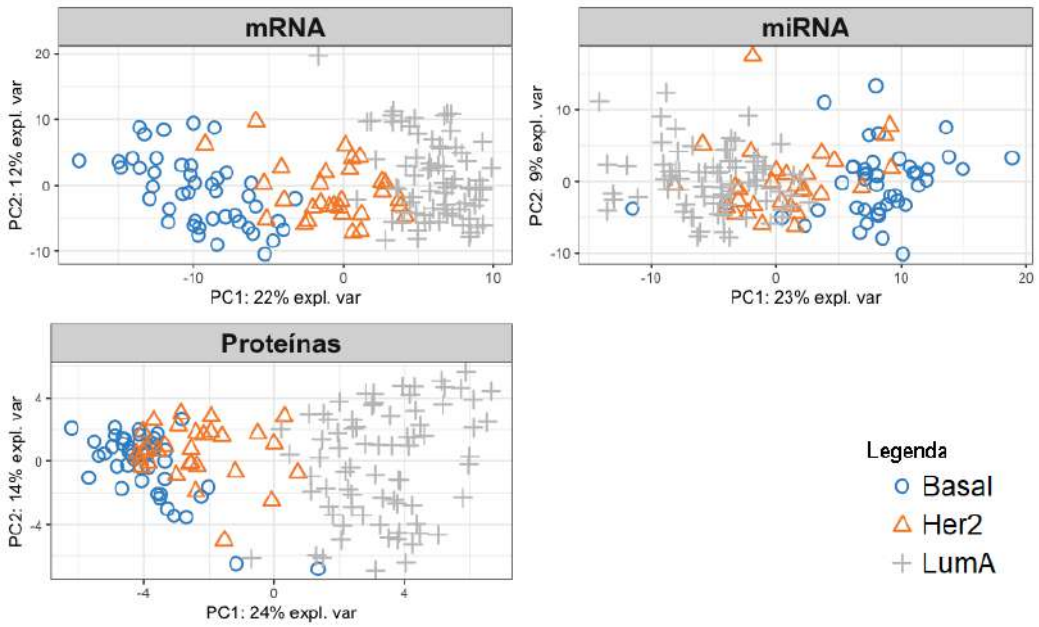

Figura 4 – Diagramas de dispersão das primeiras duas CPs dos conjuntos de dados de treinamento para mRNA, miRNA e Proteínas

3.2 PLS-DA e sPLS-DA

A metodologia PLS-DA foi utilizada separadamente para os conjuntos de dados de treinamento (150 observações) para mRNA, miRNA e Proteínas visando explorar a associação das variáveis de cada bloco com os três subtipos de câncer (variável resposta categorizada).

Essa análise começa com o ajustes de um modelo PLS-DA1 com 10 VLs e considerando todas as variáveis, para cada bloco do conjunto de treinamento. As amostras foram projetadas em um subespaço formado pelas duas primeiras variáveis latentes. Obteve-se assim, como na seção anterior, uma análise de cada conjunto de dados separadamente, porém agora de forma supervisionada. Observou-se uma ligeira mudança no comportamento das amostras quando se buscou a combinação de variáveis levando em consideração os subtipos de câncer (Figura 5).

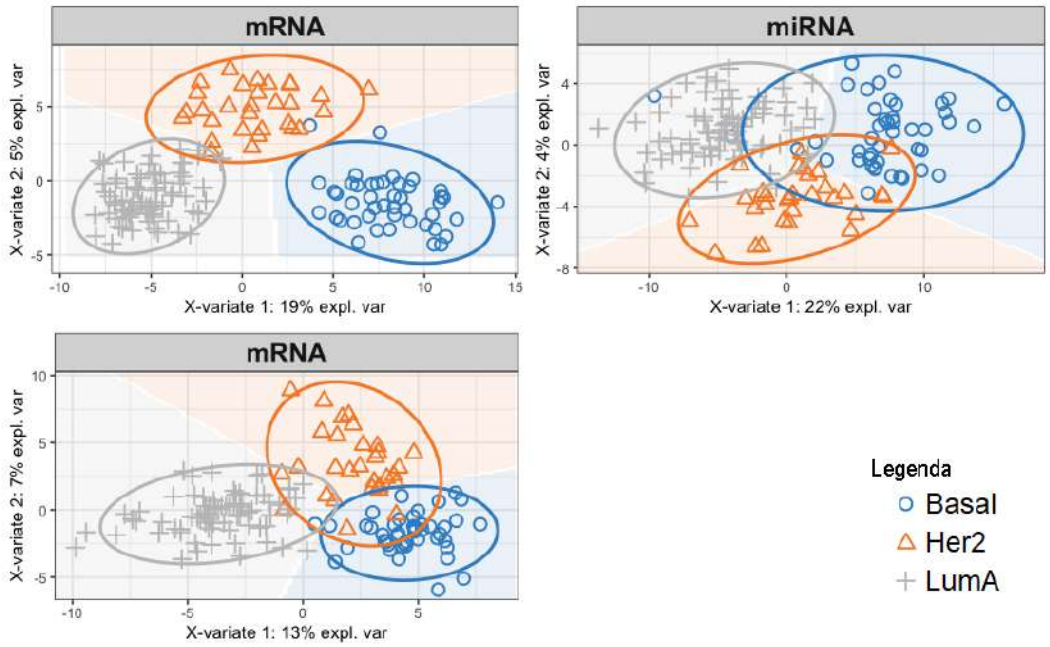

Figura 5 – Diagramas de dispersão das primeiras duas VLs do modelo PLS-DA para os conjuntos de dados de treinamento para mRNA, miRNA e Proteínas

Observou-se que os percentuais de variância explicada por cada VL diminuíram em relação aos valores da ACP. Isso ocorreu porque, no caso de PLS-DA, buscou-se as coordenadas capazes de discriminar melhor os indivíduos dentre os três subtipos de câncer, mesmo não sendo a direção de maior variação dos dados. Os três pares de VLs apresentados na Figura 5 parecem separar bem entre os três subtipos de câncer.

No caso do mRNA, por exemplo, considerando o primeiro eixo, obteve-se alguma separação entre os três subtipos de câncer e, nos outros dois casos parece que apenas as diferenças entre os tipos Basal e Luminal A são representadas. Já para a segunda VL, nos três blocos, o tipo Her2 foi mais discriminado dos outros dois.

A primeira análise de performance, VC1-PLSDA, sugeriu que se utilizasse 2 VLs para o mRNA, 3 para o miRNA e 2 para as proteínas. A VC2-sPLSDA, que já considera essa quantidade de VLs, foi utilizada já que deseja-se obter um modelo esparso. O teste considerou todas as quantidade de 1 até 150 e o resultado da calibração da quantidade ótima de variáveis, que devem compor cada uma das VLs, está apresentado na Tabela 3. Tais resultados mostram que a quantidade a ser utilizada ainda é muito grande em alguns casos, como nas VLs do miRNA, por exemplo.

Tabela 3 – Resultado da calibração do número de variáveis para os modelos finais sPLS-DA

	Comp 1	Comp 2	Comp 3
mRNA	32	105	×
miRNA	44	145	120
Proteínas	11	1	×

Ajusta-se modelos finais do tipo sPLS-DA (um para cada bloco) de acordo com as quantidades calibradas acima. Com base nesses modelos, novos resultados foram obtidos, inclusive os de previsão para os dados de teste. A Figura 6 evidencia uma diminuição no percentual de variação explicada por cada VL devido à retirada das informações contidas nas variáveis excluídas.

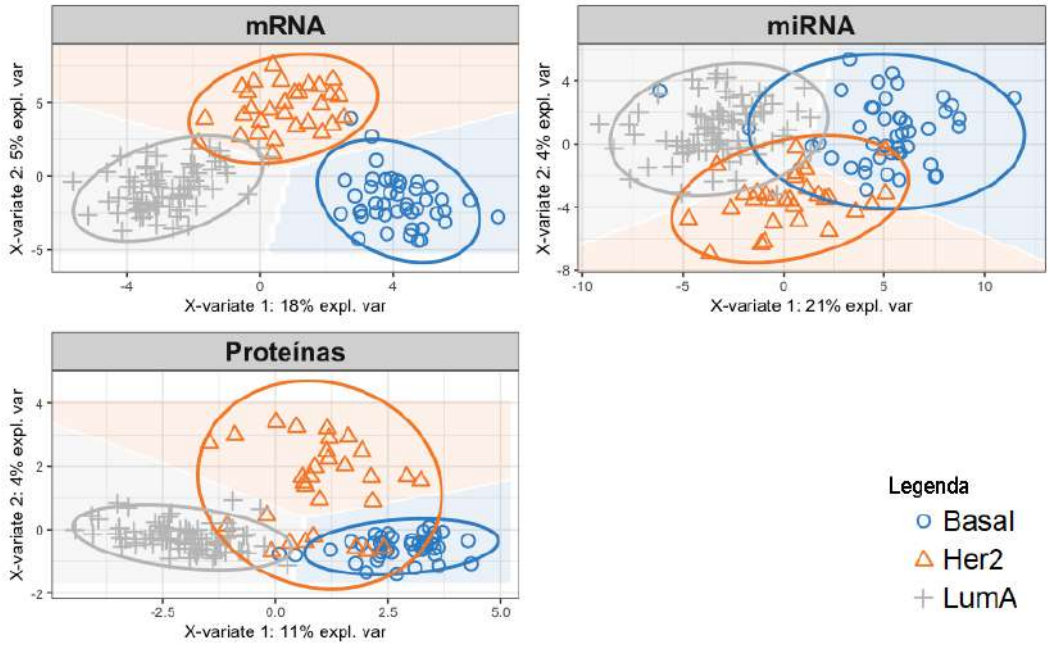

Figura 6 – Diagramas de dispersão das primeiras duas VLs do modelo sPLS-DA para os conjuntos de dados de treinamento para mRNA, miRNA e Proteínas

As Figuras 7, 8 e 9 apresentam gráficos de barras horizontais que exibem os vetores de coeficientes das variáveis para cada VL de cada bloco. Por ser uma supervisionada, essa análise fornece também a visualização do subtipo de câncer que obteve a maior média para aquela variável, pela cor da barra, ou seja, a classe que é mais influenciada por aquela variável. Para os blocos de mRNA e do miRNA exibiu-se apenas as 20 primeiras, com o objetivo de verificar apenas as variáveis mais influentes e futuramente compará-las com as dos outros métodos.

A primeira VL de cada bloco parece diferenciar melhor entre os subtipos de câncer Luminal A e Basal. Já a segunda VL parece diferenciar a classe Her2 dos outros dois. Isso acontece principalmente no caso do mRNA, no qual destaca-se alguns transcritos mais influentes como o ZNF552 e o KDM4B para o Luminal A e o TP53INP2 para o Her2. Na Figura 7, representando o bloco mRNA, observou-se que a primeira VL possui variáveis que diferenciam os tipos Luminal A e Basal, enquanto na segunda, os três tipos foram representados e há diferenciação do grupo HER2 com os outros dois.

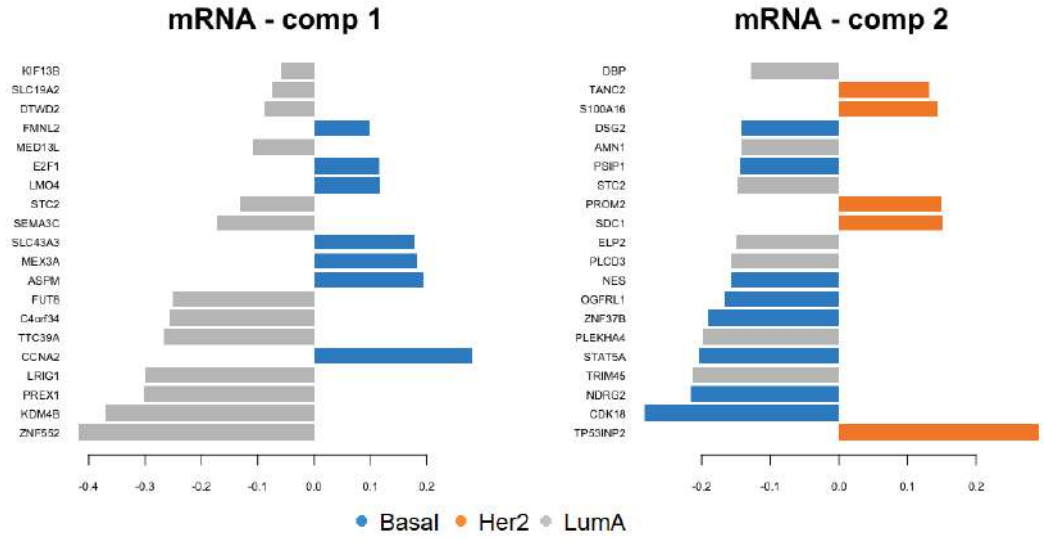

Figura 7 – Gráficos dos coeficientes das 20 variáveis mais influentes nas duas primeiras VLs do sPLS-DA para o mRNA

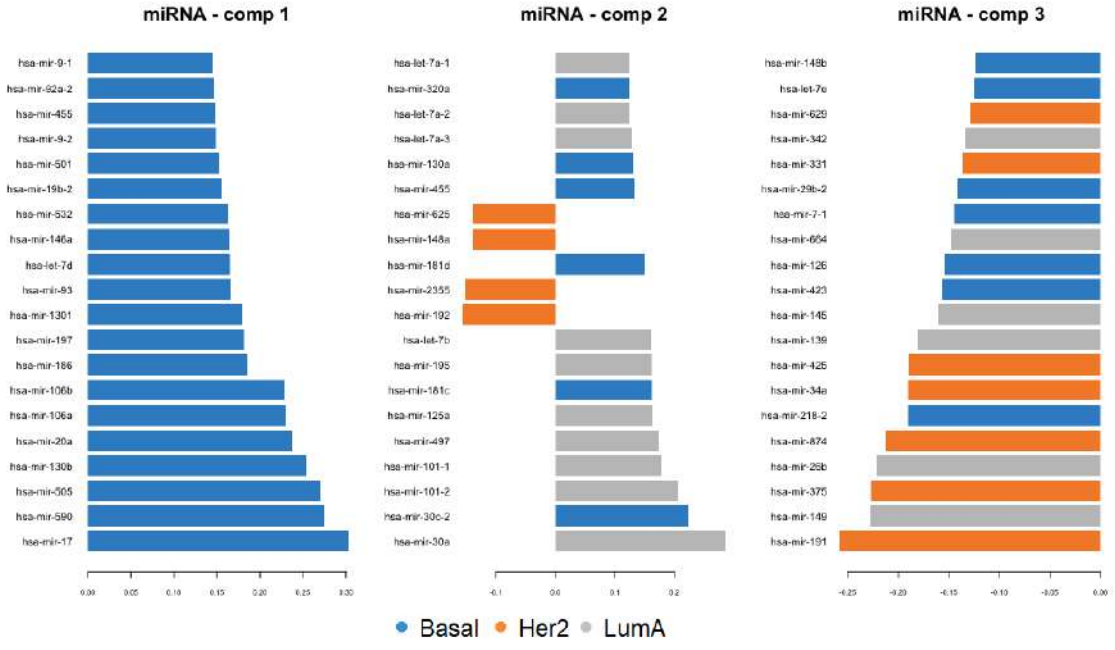

Figura 8 – Gráficos dos coeficientes das 20 variáveis mais influentes nas VLs do sPLS-DA para o miRNA

Ainda, foi possível perceber que o bloco miRNA não apresentou uma diferenciação tão clara entre os três grupos, pois precisou selecionar uma VL a mais e um número relativamente grande de variáveis, ou seja, foi necessário buscar mais informações para conseguir diferenciar bem os três subtipos de câncer. Ainda, percebeu-se que as informações fornecidas pela Figura 8 não geram interpretações tão claras e diretas. A primeira VL pareceu ser bastante influente para o grupo Basal, enquanto a segunda parece fazer um

contraste entre variáveis que expliquem bem o tipo Her2 com as dos outros dois blocos e, finalmente, a terceira parece ter variáveis muito influentes para todos os subtipos de câncer. Cada uma das variáveis mais influentes das três componente, hsa-mir-17, hsa-mir-30a e hsa-mir-191, relacionou-se com um subtipo diferente do câncer: Basal, Luminal A e HER2, respectivamente.

No caso das proteínas em particular, há poucas variáveis em cada VL. A segunda VL é formada por apenas uma proteína conhecida como HER2 que é um dos principais fatores a ser considerado para definir o tipo de câncer que nomeia o grupo. Destaca-se também as proteínas ER-alpha e GATA3 para o fenótipo Luminal A, e para o tipo Basal tem-se a ASNS e a Cyclin_B1.

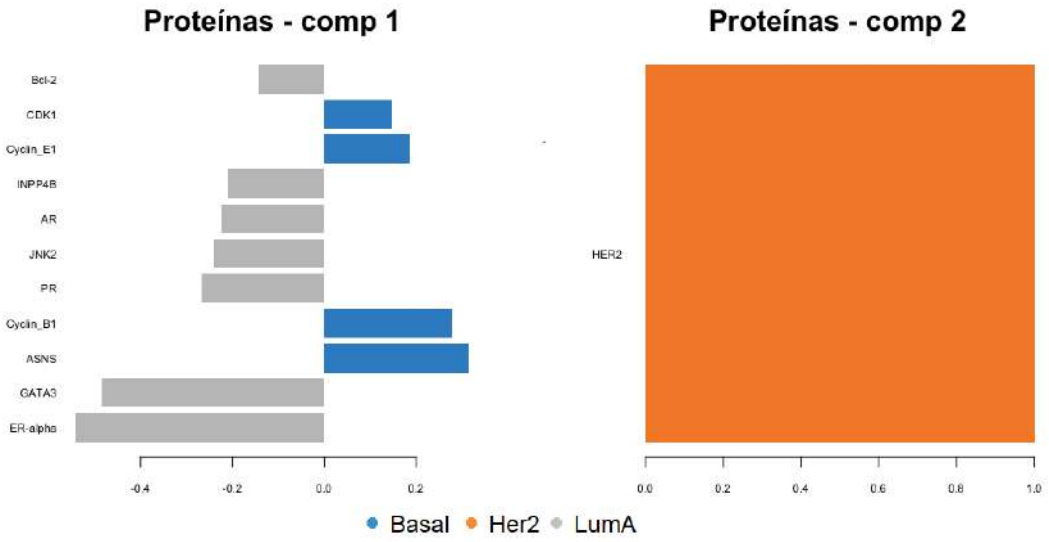

Figura 9 – Gráfico dos coeficientes das variáveis influentes nas VLs do sPLS-DA para as proteínas

Para se avaliar o poder de previsão do método, utilizou-se os dados de teste com informações de mRNA e miRNA de 70 novos pacientes. Obteve-se os resultados para os blocos mRNA e miRNA, apresentados nas matrizes de confusão nas Tabelas 4 e 5, com taxas de erro balanceadas aproximadamente iguais a $2,54\%$ e $7,62\%$, respectivamente. Por serem valores baixos indicam que o modelo parece estar bem ajustado.

Tabela 4 – Matriz de confusão das categorias preditas em relação as verdadeiras para o bloco mRNA pelo método sPLS-DA

mRNA	Predito como Basal	Predito como Her2	Predito como LumA
Basal	20	1	0
Her2	0	14	0
LumA	0	1	34

Tabela 5 – Matriz de confusão das categorias preditas em relação as verdadeiras para o bloco miRNA pelo método sPLS-DA

miRNA	Predito como Basal	Predito como Her2	Predito como LumA
Basal	18	3	0
Her2	0	14	0
LumA	0	3	32

As análises anteriores indicam que este modelo parece ser adequado para a discriminação destes grupos apresentados e os resultados da predição para os dados de teste também indicam isso. Foram obtidos poucos erros, todos classificando como Her2 quem na verdade pertencia a outra classe. As variáveis do mRNA apresentaram um melhor desempenho que as do miRNA. Como para os dados de teste não estavam disponíveis informações sobre as proteínas, não foi possível determinar o desempenho do modelo referente a tal bloco. Destaca-se que esse bloco de dados teria o maior potencial de economizar recursos em termos de quantidade de variáveis.

3.3 Block-sPLSDA

O método DIABLO, também conhecido como Block-sPLSDA (BsPLSDA), é utilizado para a análise supervisionada de dados N-integrados e tem como objetivo a busca de variáveis que observem as interações entre os blocos e que discriminem bem o fenótipo de interesse (subtipo de câncer de mama). Esta técnica se diferencia da anterior, principalmente por permitir a análise de mais de um bloco de variáveis explicativas levando em conta suas relações. Para a realização da técnica, considera-se cada um dos *'omics* como um bloco, tendo ainda a variável resposta qualitativa, com informações sobre o fenótipo, que é transformada em uma matriz do tipo "*dummy*", caracterizando mais um bloco.

A matriz *design* de relação entre os blocos determina as relações entre os blocos com valores entre 0 e 1. Utilizou-se então dois tipos extremos de matrizes de delineamento de blocos: uma com todas as interações entre os pares de blocos iguais a $0,1$ (Matriz 3.1) e outra uma matriz de relações máximas entre os blocos (*full design matrix*), apresentada em 3.2, que considera todas interações entre os blocos iguais a 1.

$$D_1 = \begin{array}{c} \begin{array}{ccc} mRNA & miRNA & Proteína \end{array} \\ \begin{pmatrix} 0 & 0,1 & 0,1 \\ 0,1 & 0 & 0,1 \\ 0,1 & 0,1 & 0 \end{pmatrix} \end{array} \tag{3.1}$$

$$D_2 = \begin{array}{c} \begin{array}{ccc} mRNA & miRNA & Proteína \end{array} \\ \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \end{array} \tag{3.2}$$

Para cada uma das matrizes apresentadas, (Matrizes 3.1 e 3.2), utilizou-se os passos descritos na sessão 3.2 para gerar seus respectivos modelos finais BsPLSDA. Os resultados

obtidos estão apresentados nas subseções a seguir e possibilitam melhor entendimento do comportamento das amostras, das variáveis e da performance destes modelos em predições para novas amostras. Para a construção do modelo BSPLSDA1 optou-se por utilizar 5 VLs, tendo em vista os resultados obtidos por ACP e PLS-DA, que identificaram no máximo 3 componentes por bloco.

3.3.1 *Design* com relação igual a 0,1

Nesta seção, utiliza-se a métrica apresentada por Amrit Singh et al. (1), a qual considera relação medida assumida de $0,1$ entre os diferentes blocos de informações moleculares e igual a 1 entre cada bloco e a variável resposta Y, como apresentado na Matriz 3.1. Sendo assim, o primeiro passo foi construir o modelo BPLSDA1, considerando esta matriz *design* de blocos, com 5 VLs e todas as variáveis disponíveis nos três blocos.

Realizou-se então duas análises de performance consecutivas, com validação cruzada. A VC1-BPLSDA calibrou a quantidade de VLs necessárias, como apresentado na Figura 10. Os resultados definiram duas VLs para cada bloco como uma quantidade apropriada para este caso. Já utilizando essa informação, a VC2-BsPLSDA foi realizada com o intuito de verificar quantas variáveis deveriam compor cada VL de cada bloco.

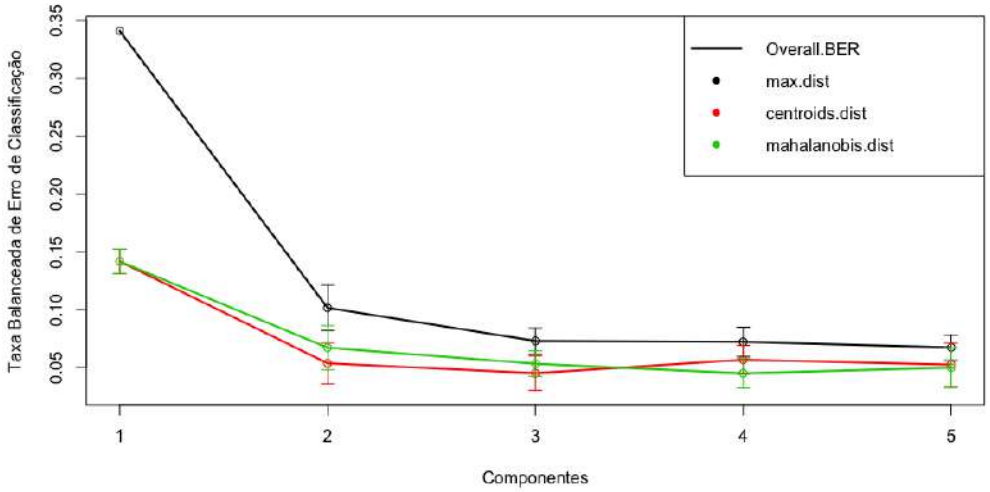

Figura 10 – Taxa de Erro de Classificação Balanceada por quantidade de componentes (VLs) para a matriz *design* de relações iguais a $0,1$

O modelo final BsPLSDA foi construído a partir dos resultados apresentados acima, com duas VLs para cada bloco e respectivas quantidades de variáveis apresentadas na Tabela 6. Assim, na Figura 11, observou-se que a explicação da variação dos dados foi muito parecida com os resultados dos outros métodos, mesmo com uma grande redução

na dimensão. A quantidade de variáveis selecionadas foi bem menor que quando utilizado o sPLS-DA para cada bloco separadamente (Tabela 3).

Tabela 6 – Resultado da calibração do número de variáveis para o modelo final BsPLSDA com matriz *design* de relações iguais a $0,1$

	Comp 1	Comp 2
mRNA	30	16
miRNA	10	35
Proteínas	8	5

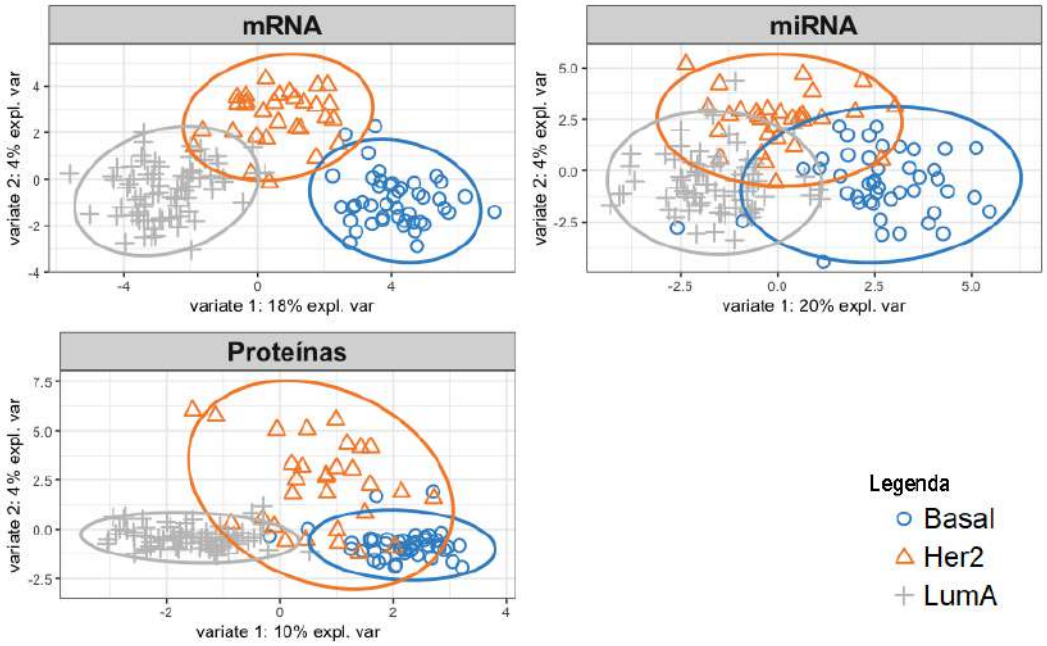

Figura 11 – Diagramas de dispersão das primeiras duas VLs do modelo BsPLSDA com matriz *design* de relações iguais a $0,1$ para os dados de treinamento

Apresenta-se nas Figuras 12, 13 e 14 o comportamento e a influência das 20 principais variáveis selecionadas nas VLs de cada bloco. Com isso, tem-se indícios de quais são as variáveis mais importantes na classificação dos três subtipos de câncer estudados.

No caso do mRNA, tem-se que a primeira VL busca informações para discriminar os tipos Luminal A e Basal, os quais destacam os genes *ZNF552*, *KDM4B* e *CCNA2*. Enquanto isso, para a segunda componente a principal variável é o gene *TP53INP2*, relacionado ao tipo Her2, que parece diferenciá-lo dos outros dois subtipos de câncer. Além disso, observa-se o gene *CDK18* em relação ao tipo Basal com uma influência relativamente forte. Todas estas variáveis citadas aqui foram também observadas com os modelos PLS-DA e sPLS-DA (Seção 3.2).

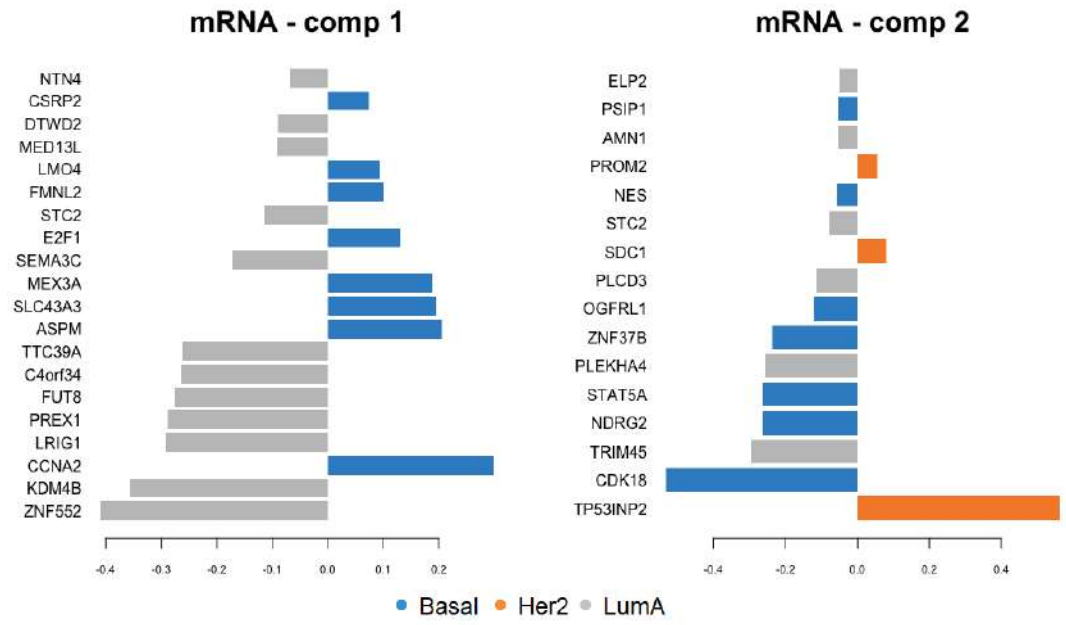

Figura 12 – Gráfico dos coeficientes de até 20 variáveis mais influentes nas VLs do mRNA no modelo BsPLSDA com a matriz *design* de relações iguais a $0,1$

Para o bloco do miRNA, a primeira componente parece discriminar apenas o tipo Basal, enquanto a segunda parece explicar um pouco dos três tipos, mas principalmente o Luminal A. Tais resultados sugerem que estas VLs não foram boas preditoras do tipo Her2.

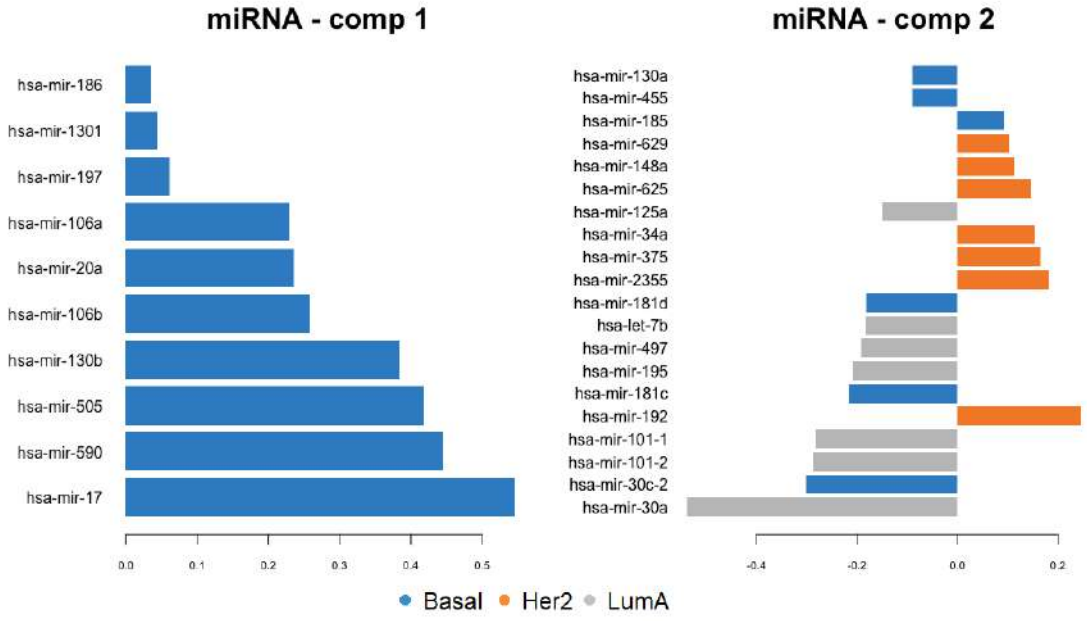

Figura 13 – Gráfico dos coeficientes de até 20 variáveis mais influentes nas VLs do miRNA no modelo BsPLSDA com a matriz *design* de relações iguais a $0,1$

Observou-se que este modelo selecionou um número maior de proteínas no total

para compor as duas VLs que o anterior, o que pode ter acontecido porque não foram testada as quantidades mais baixas no processo de calibração VC2-BsPLSDA. Destaca-se ainda as proteínas HER2 e ER-alpha como muito informativas sobre os tipos HER2 e Luminal A, respectivamente.

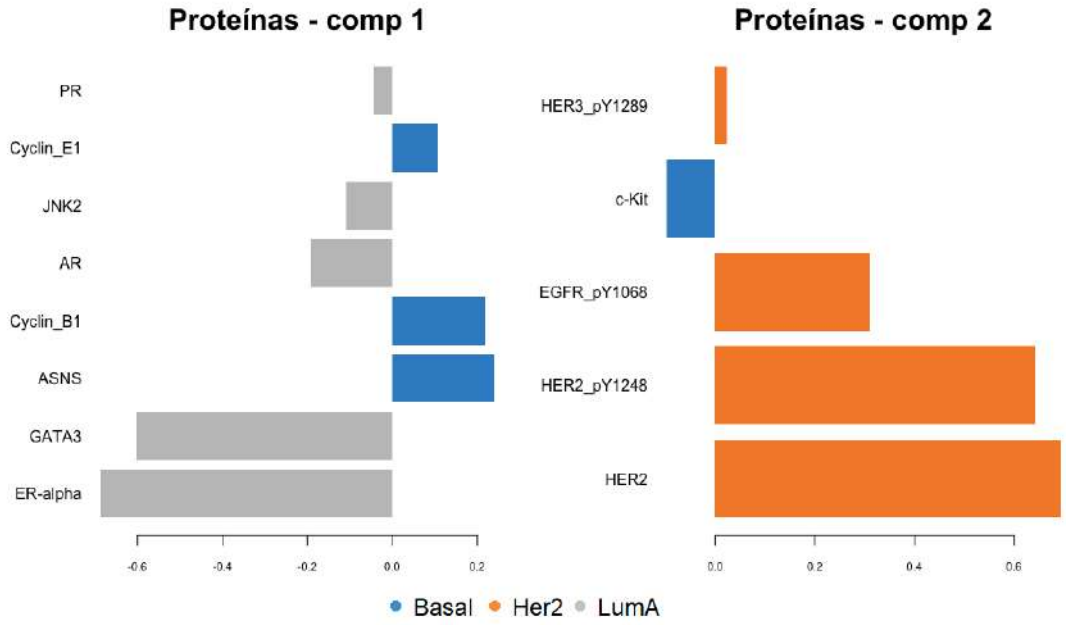

Figura 14 – Gráfico dos coeficientes das variáveis selecionadas nas VLs das proteínas no modelo BsPLSDA com a matriz *design* de relações iguais a $0,1$

Esta metodologia também possibilita uma visualização da estrutura geral das correlação entre os blocos. A Figura 15 mostra que foi possível extrair a maior parte da correlação entre os blocos na primeira dimensão. Considerando-se as correlações dos 3 pares de blocos, a correlação entre proteínas e mRNAs foi a maior e entre proteínas e o miRNA, a menor.

A Figura 16 apresenta uma das ferramentas gráficas mais interessantes disponíveis no pacote *mixomics*. O gráfico apresenta no círculo todas as variáveis selecionadas para compor as 2 VLs do modelo, separadas por bloco com diferentes cores (roxo para mRNA, verde para miRNA e amarelo para proteínas). As linhas fora do círculo indicam os níveis médios de expressão de cada um dos grupos dos subtipos de câncer em relação a cada uma das variáveis apresentadas. Finalmente, cada linha dentro do círculo representa a correlação das duas variáveis ligadas por ela, podendo ser uma correlação positiva (linha vermelha) ou negativa (linha preta).

Foram selecionadas apenas as correlações com valores, em módulo, acima de $0,8$, para que fosse possível visualizar e analisar as interações mais importantes. Foram também ocultadas as relações dentro dos blocos. Com isso, tem-se um conjunto de poucas interações, todas entre uma proteína, ER-alpha ou GATA3, e uma variável de outro tipo de elemento

molecular.

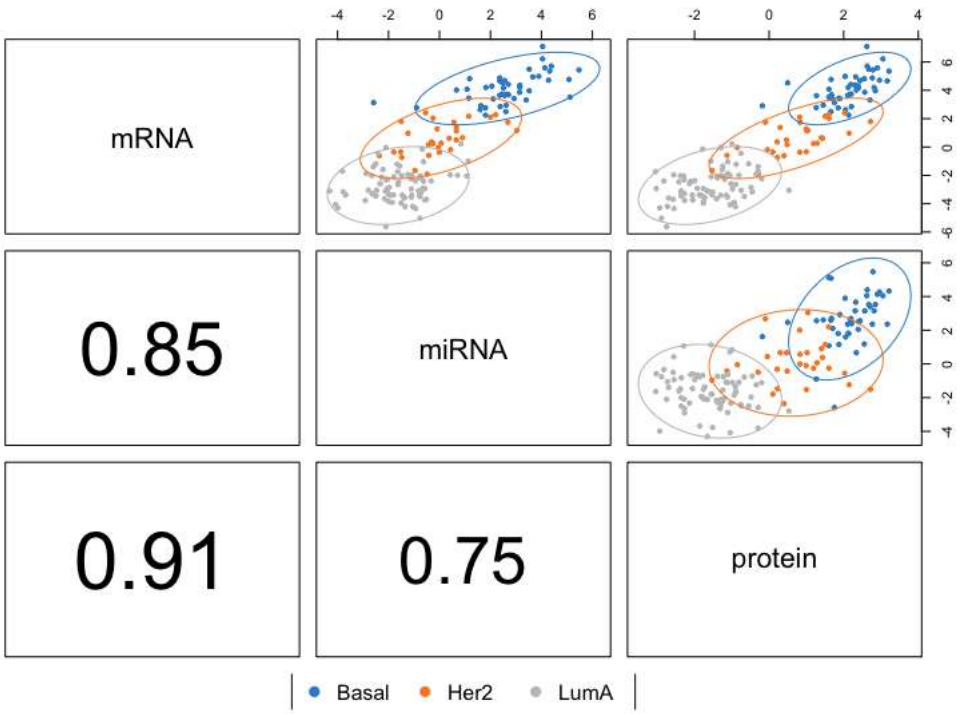

Figura 15 – Gráficos de correlação entre as primeiras VLs de cada bloco geradas pelo BsPLSDA que utiliza a matriz *design* de relações iguais a $0,1$

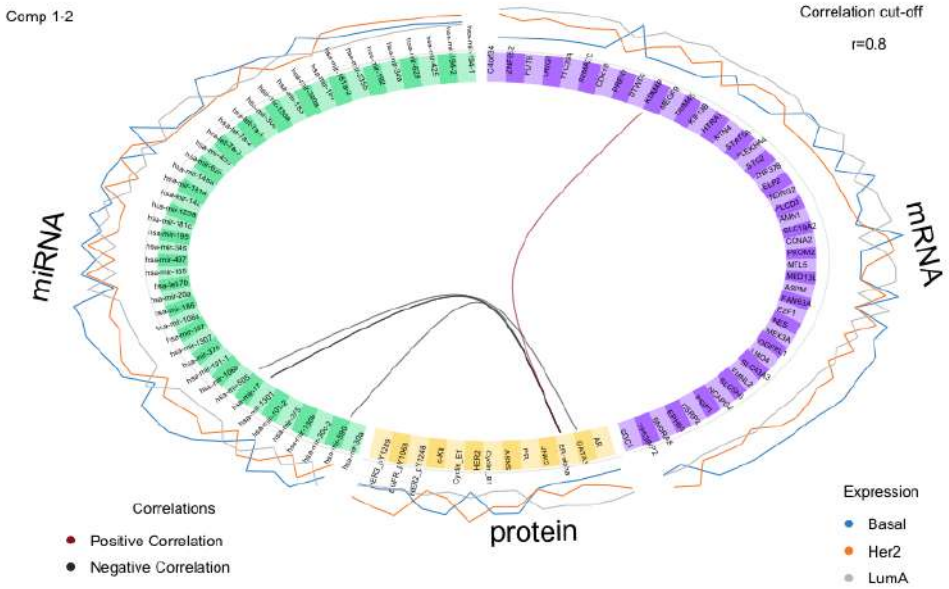

Figura 16 – Gráfico de "Circo" com os resultados do modelo BsPLSDA com matriz *design* de relações iguais a $0,1$

Tais proteínas também se destacaram na composição das VLs. Ambas apresentaram níveis mais baixos para o tipo Basal, intermediário para o tipo Her2 e mais alto para o tipo

Luminal A. Enquanto isso, o mRNA *KDM4B* apresenta a única correlação positiva, uma vez que os níveis dos tipos de câncer são exatamente os mesmos do ER-alpha. Tal relação já foi descrita na literatura (7) evidenciando que este mRNA é um dos reguladores para os receptores de estrogênio. É interessante notar que os maiores níveis são exatamente do tipo que é receptor hormonal positivo (LumA). Na Figura 14 também percebe-se a influência desse gene para o subtipo Luminal A.

Todas as correlações restantes são negativas e sempre entre uma proteína e um miRNA. Percebe-se que realmente os níveis de expressão dos subtipos de câncer se apresentam de forma contrária nas variáveis correlacionadas negativamente. As análises acima são exemplos de como esses resultados podem ser muito importantes para este tipo de estudo. As outras relações apresentadas podem ser exploradas da mesma maneira.

O diagrama de redes (*network*) apresenta redes de associações relevantes, consideradas pelo método BsPLSDA. O valor de similaridade para cada par de variáveis é calculado pela soma das correlações entre as variáveis originais e cada VL do modelo. A intensidade de tais valores é representada pelas cores das linhas que conectam duas variáveis. A vantagem dessa ferramenta gráfica é a habilidade de exibir simultaneamente se as correlações são positivas ou não.

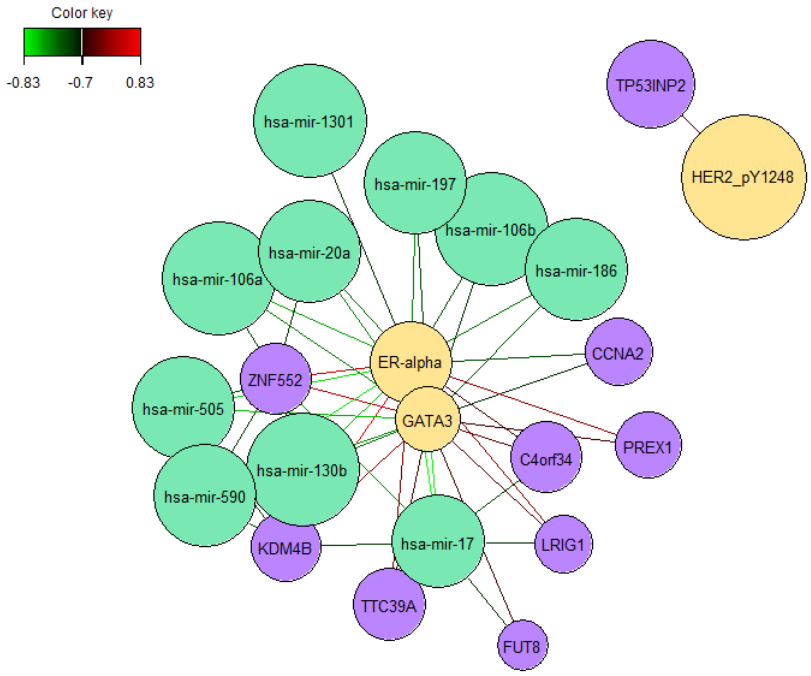

Figura 17 – Diagrama de rede de relações entre as variáveis do modelo BsPLSDA que utiliza a matriz *design* de relações iguais a $0,1$

Apresentado na Figura 17, com um corte de $0,7$, em módulo, o diagrama de redes apresenta a possibilidade de visualizar mais relações que as apresentadas no Gráfico

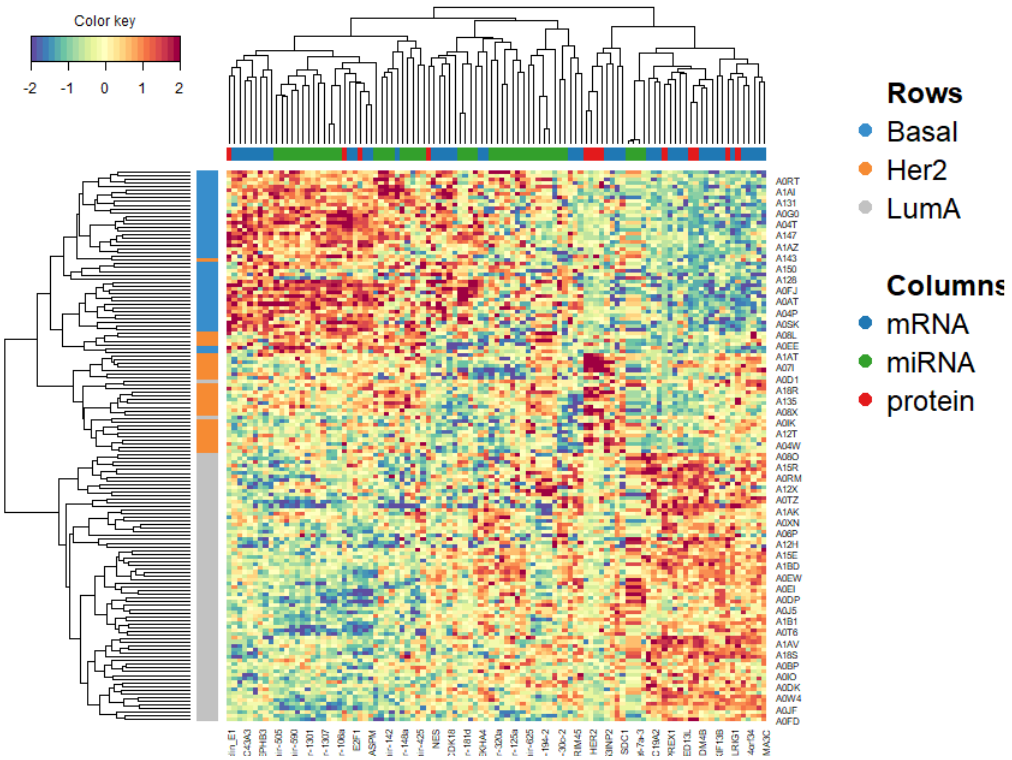

Figura 18 – Clustered Image Map com indivíduos nas linhas e variáveis selecionadas para todas as componentes nas colunas

de Circo. Porém, ambos evidenciam que as proteínas ER-alpha e GATA3 são muito importantes pois fazem as ligações mais fortes entre as proteínas e os outros dois blocos.

O gráfico de calor (*heatmap*) possibilita a identificação de perfis de expressão em dados com grandes dimensões. A Figura 18 mostra que para cada subtipo de câncer há um grupo de variáveis que segue determinado padrão. Entre as proteínas, por exemplo, o subtipo de câncer Her2 foi bem influenciado pela proteína que leva este mesmo nome, mesmo que tenha acontecido alguns casos em que o individuo não foi tão bem diferenciado por essas variáveis. Ainda, a maior parte das variáveis de miRNA selecionadas estão sendo utilizadas na discriminação do tipo Basal, informação esta que também pode ser observada na Figura 13.

Tabela 7 – Matriz de confusão das categorias preditas por votos com peso em relação as verdadeiras para o modelo BsPLSDA com matriz design de relações iguais a $0,1$

	Predito como Basal	Predito como Her2	Predito como LumA
Basal	20	1	0
Her2	0	14	0
LumA	0	1	34

A predição realizada pelo modelo final BsPLSDA, com matriz design de relações

iguais a $0,1$, obteve uma taxa de erro de classificação balanceada igual a $2,54\%$, ou seja, mesmo com um bloco faltante, obteve-se um bom ajuste aos dados. A Tabela 7 mostra que a predição por este modelo resultou em apenas 2 erro, evidenciando uma boa habilidade preditiva.

Os resultados apresentados na Figura 15 indicam que é possível ter muitas informações sobre um dos blocos mesmo que ele esteja ausente para a predição em novas observações. Mesmo que não tenha sido possível coletar dados sobre as proteínas do grupo de teste, o modelo considerou as influência deste bloco faltante para realizar a predição para uma nova amostra.

3.3.2 *Design* de máxima relação

Esta seção apresenta resultados de um modelo com a metodologia BsPLSDA utilizando uma matriz *design* de blocos considerando relação máxima entre os blocos. Primeiramente, foi construído um modelo com 5 VLs, as quais utilizaram todas as variáveis.

Realizou-se então as etapas de calibração do número de VLs (VC1-BPLSDA) e da quantidade de variáveis selecionadas para cada uma delas (Vc2-BsPLSDA), de maneira análoga ao caso da seção anterior. Observou-de 3 VLs como suficientes (Figura 19) pois a adição de mais uma não traria uma grande diminuição na taxa de erro de classificação balanceada.

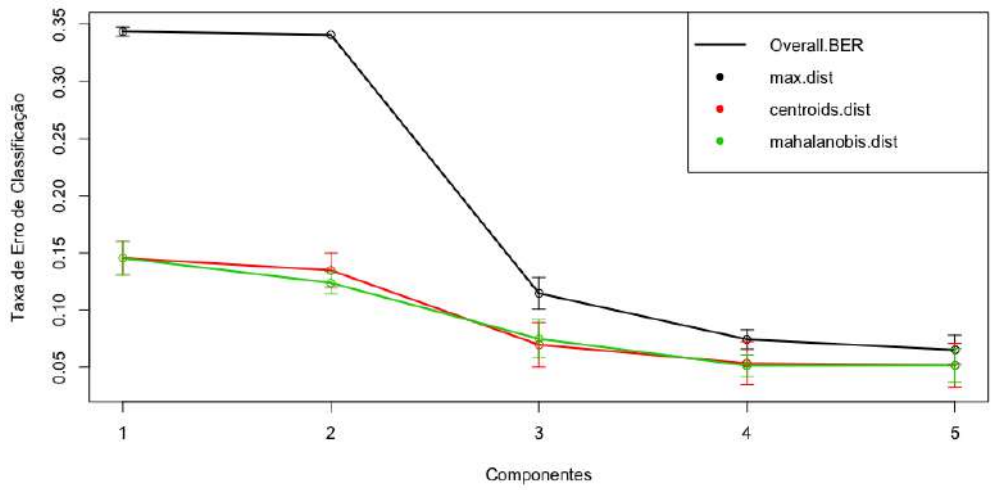

Figura 19 – Taxa de Erro de Classificação Balanceada por quantidade de componentes (VLs) para a matriz *design* de relações iguais a 1

O modelo final (BsPLSDA) foi obtido com base nos números de VLs e variáveis de cada bloco, apresentados na Tabela 8. Ainda, destaca-se que utilizou-se 70 variáveis na primeira VL do bloco miRNA e todas foram necessárias.

As duas primeiras VLs foram utilizadas então para visualizar o comportamento das amostras, na Figura 20. O resultado não foi muito diferente das análises anteriores. Na verdade, o bloco de miRNA aumentou em um ponto percentual na primeira VL e o bloco de proteínas na segunda VL. As observações continuaram bastante misturadas no bloco miRNA.

Tabela 8 – Resultado da calibração do número de variáveis para o modelo final do BsPLSDA com matriz *design* de relações iguais a 1

	Comp 1	Comp 2	Comp 3
mRNA	55	16	5
miRNA	70	14	5
Proteínas	5	25	5

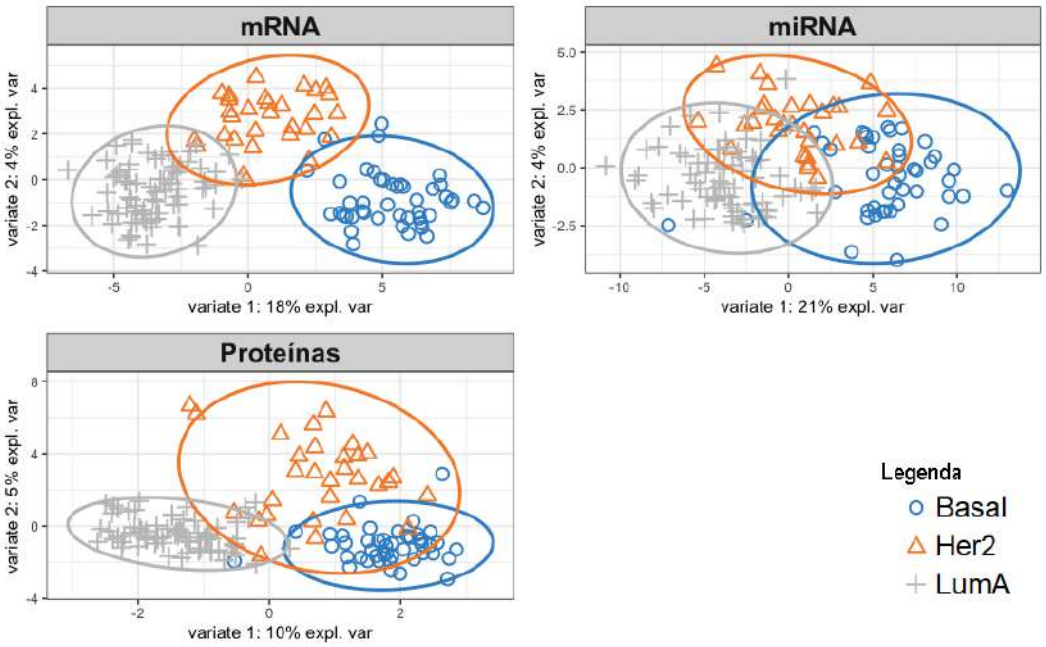

Figura 20 – Diagramas de dispersão das primeiras duas VLs do modelo BsPLSDA com matriz *design* de relações iguais a 1 para os dados de treinamento

A contribuição das variáveis mais influentes para cada VL pode ser visualizado nas Figuras 21, 22 e 23, sendo que a cor de cada uma das barras indica a classe na qual cada variável tem o maior nível de contribuição média.

Para o bloco de mRNA, a segunda VL apresentou variáveis importantes para os três tipos de câncer, enquanto isso a primeira componente está envolveu mais variáveis para os tipos Basal e Luminal A e a terceira, para o Basal e o Her2. São destaques as variáveis que possuem maiores escores: referentes ao tipo BASAL tem-se NCF4 e *CDK18*; ao tipo Luminal A, *ZNF552*; e ao Her2 tem-se *TP53INP2* e *FLI1*.

No caso do miRNA, a primeira VL destacou a variável hsa-mir-17, influenciando principalmente o tipo Basal. Enquanto isso, as outras duas possuem a variável mais

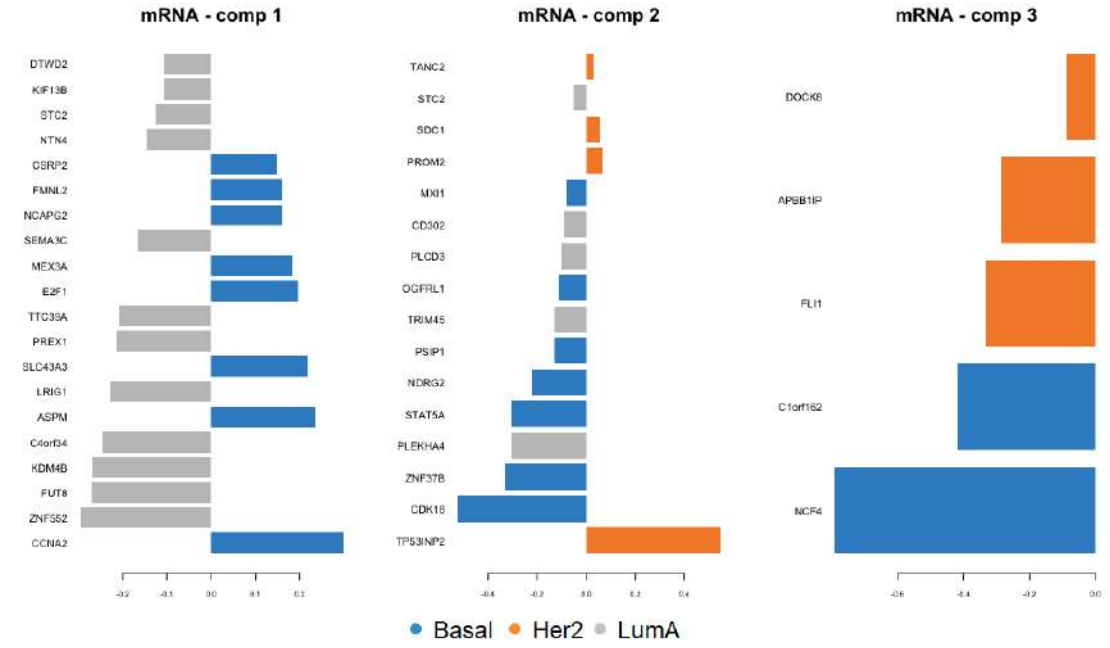

Figura 21 – Gráfico dos coeficientes das variáveis selecionadas nas VLs do mRNA no modelo BsPLSDA com a matriz *design* de relações iguais a 1

influente para os outros dois tipos de câncer, sendo hsa-mir-30a para o tipo Luminal A e hsa-mir-150 para o Her2.

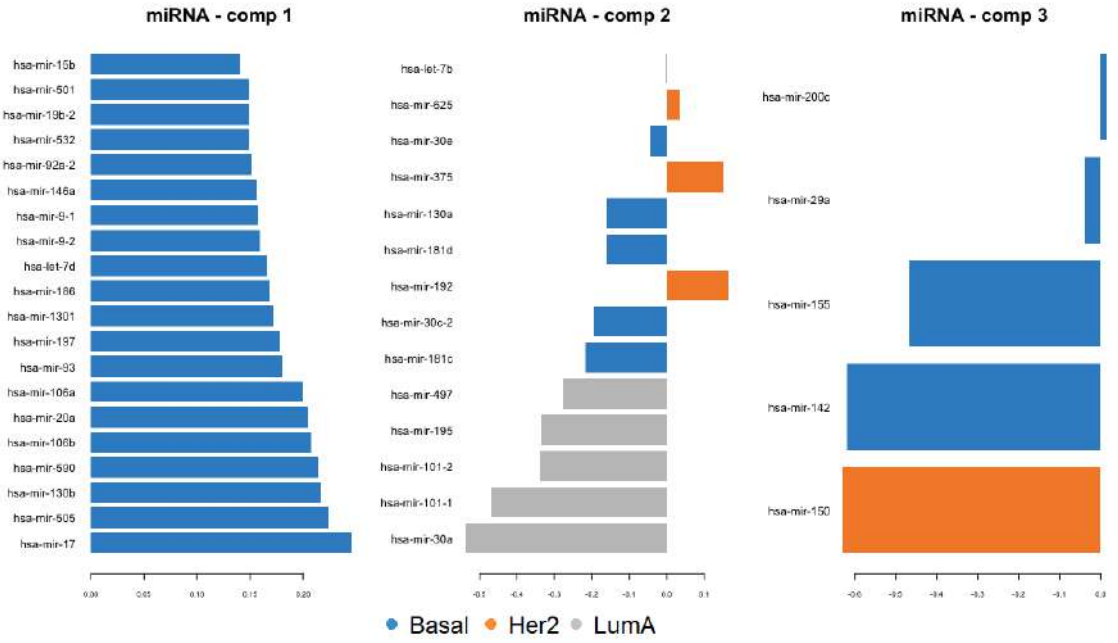

Figura 22 – Gráfico dos coeficientes das variáveis selecionadas nas VLs do miRNA no modelo BsPLSDA com a matriz *design* de relações iguais a 1

O bloco de proteínas apresentou um comportamento parecido com os outros dois bloco, ou seja, a primeira VL não diferenciou bem o tipo HER2, a segunda explicou os

três tipos e a última foi referente ao Basal e HER2. Destacam-se as variáveis ER-alpha, HER2 e Lck para os tipos Luminal A, HER2 e Basal, respectivamente.

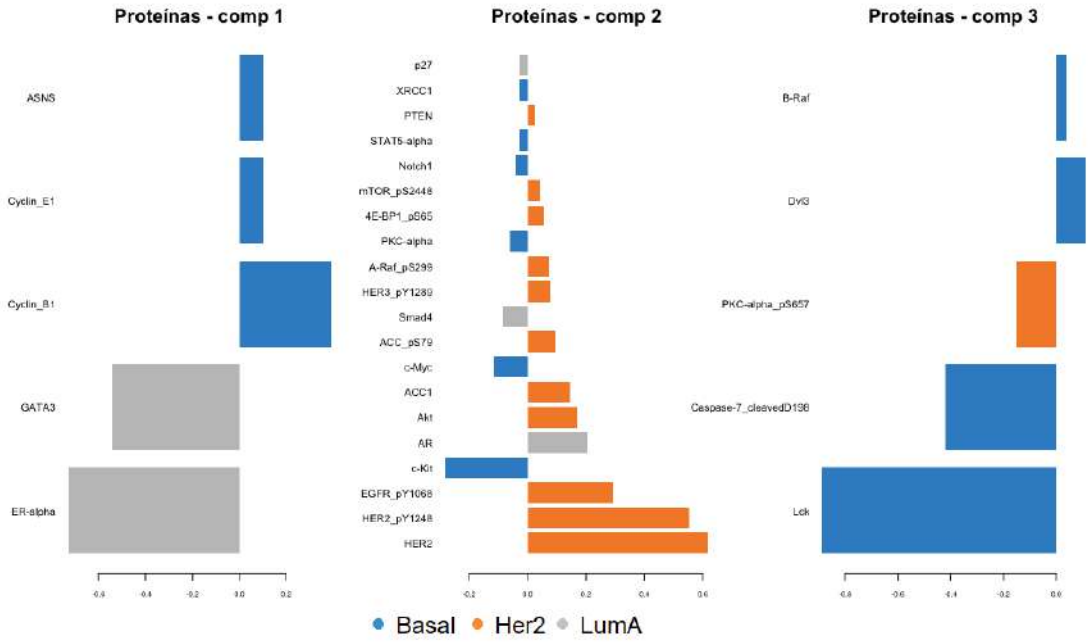

Figura 23 – Gráfico dos coeficientes das variáveis selecionadas nas VLs das proteínas no modelo BsPLSDA com a matriz *design* de relações iguais a 1

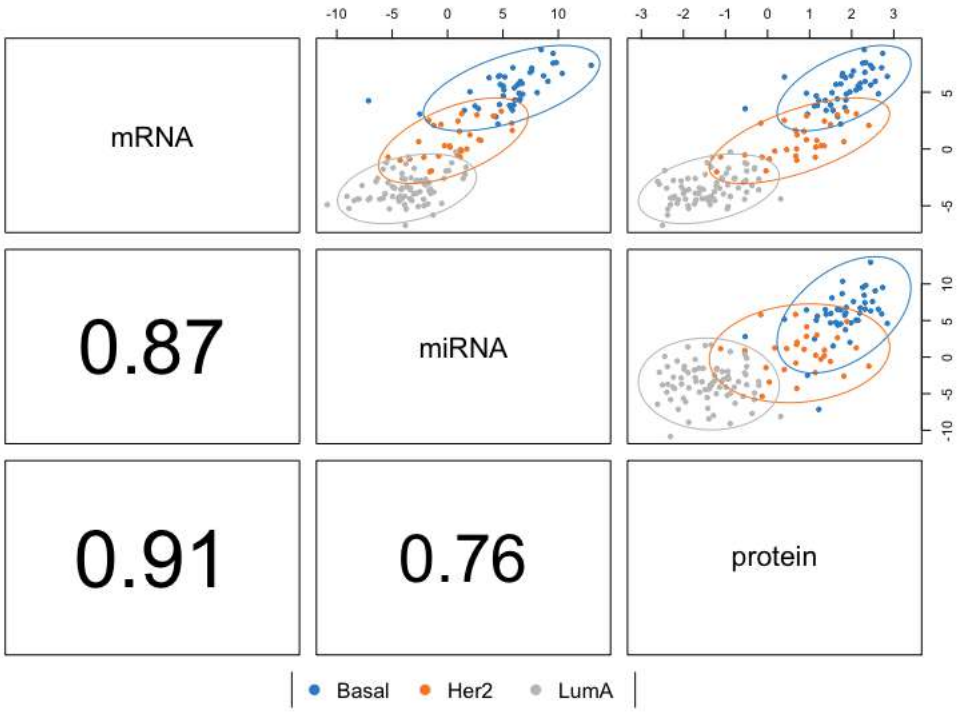

Figura 24 – Gráficos de correlação entre as primeiras componentes de cada bloco geradas pelo BsPLSDA que utiliza a matriz *design* de relações iguais a 1

As quantidades de VLs e variáveis aumentaram em relação à seção anterior, possivelmente porque assumiu-se maior relação entre os blocos, já que foi utilizado o *design* de relação máxima. Os resultados das correlações entre as primeiras componentes já foram altos quando a relação entre os blocos era igual a $0,1$, como é indicado na Figura 15, então a mudança de $0,1$ para 1, não teve muito efeito no quanto as primeiras VLs são correlacionadas, como é possível verificar na Figura 24. A maior relação continuou sendo entre os blocos mRNA e proteínas.

O Gráfico de "Circo" apresentado utilizou o mesmo corte da seção anterior (acima de $0,8$), porém desta vez apresentou bem mais relações e, ainda, com a maior parte sendo positiva. A proteína ER-alpha apresentou novamente algumas relações que já tinham sido vistas com as duas variáveis do miRNA. Outra proteína que teve destaque e não havia sido identificada antes foi a Lck. Isso provavelmente aconteceu porque agora é interessante também maximizar a correlação entre os blocos e esta variável exibe vários relacionamentos com valores altos.

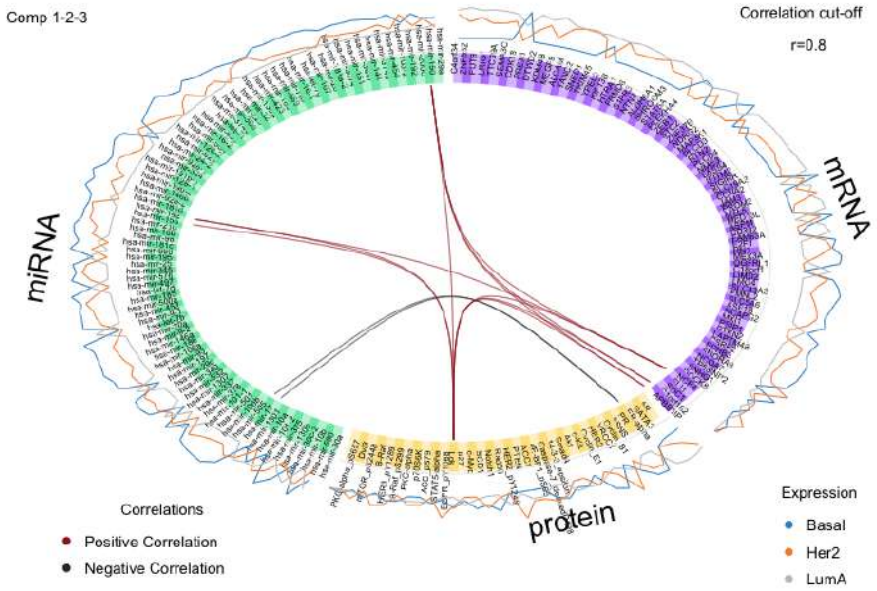

Figura 25 – Gráfico de "Circo" com os resultados do modelo BsPLSDA com matriz *design* de relações iguais a 1

O diagrama de redes, apresentado na Figura 26, reforçou que a diferença deste caso para o anterior envolve apenas algumas ligações adicionais na relação entre os blocos. A maior parte dos principais relacionamentos parecem ligar variáveis do mRNA e do miRNA com algumas poucas proteínas, como ER-alpha , GATA3, Lck e Capase-7_cleavedD198.

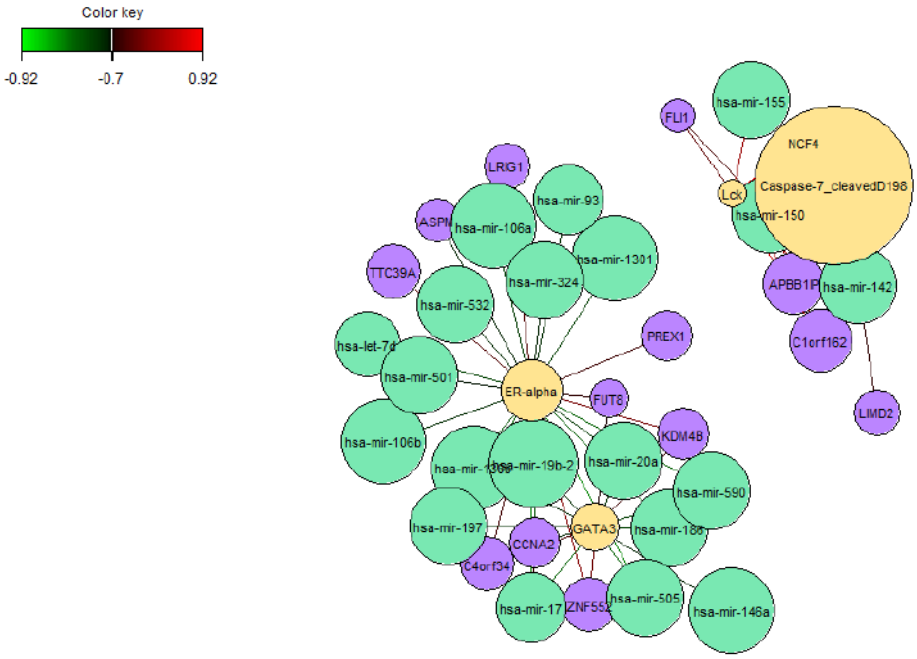

Figura 26 – Diagrama de rede de relações entre as variáveis do modelo BsPLSDA que utiliza a matriz *design* de relações iguais a 1

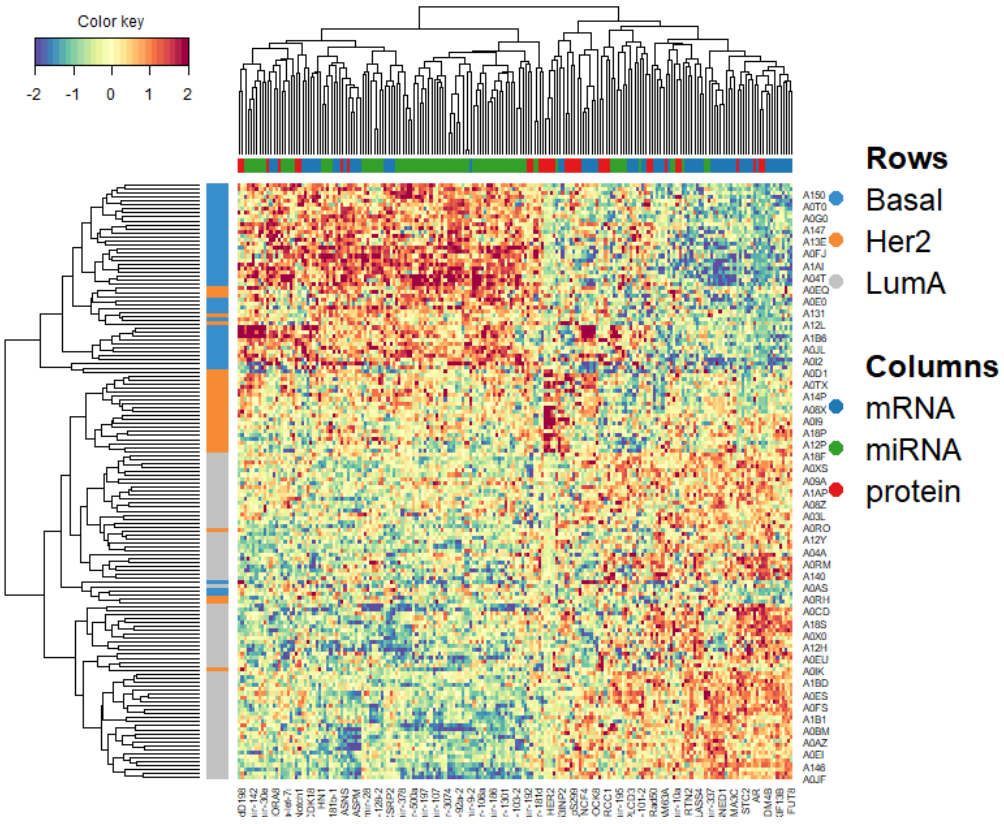

Figura 27 – Clustered Image Map com indivíduos nas linhas e variáveis selecionadas para todas as componentes nas colunas

Ao analisar a Figura 27, percebe-se que mais indivíduos não foram discriminados por esta técnica. Isso pode ter ocorrido pelo fato de que o foco não foi maior na discriminação. Maximizar as principais relações de variáveis entre os diferentes blocos foi fundamental também. O restante das análises foram muito semelhantes à seção anterior.

Tabela 9 – Matriz de confusão das categorias preditas por votos com peso em relação as verdadeiras para o modelo BsPLSDA com matriz *design* de relações iguais a 1

	Predito como Basal	Predito como Her2	Predito como LumA
Basal	19	2	0
Her2	0	14	0
LumA	0	0	35

Finalmente, utilizou-se o modelo para realizar as predições para o banco de dados de teste, para verificar a eficácia do modelo. O resultado, apresentado na Tabela 9, gerou indícios da qualidade do modelo uma vez que foram observadas apenas 2 predições erradas (o que pode ter ocorrido pela falta do bloco de proteínas neste conjunto de dados). A taxa de erro de classificação balanceada foi igual a $3,17\%$, um pouco maior que a do caso anterior. Além disso, ainda que este modelo não tenha sido tão bom em discriminar como o anterior, os resultados obtidos foram bastante satisfatórios.

4 Discussão e Conclusões

As metodologias multivariadas estão apresentando um papel cada vez mais importante nos avanços em biotecnologia. Existe um tratamento específico no caso do subtipo de câncer de mama, Her2, que é muito efetivo, pois trata da proteína que leva esse mesmo nome diretamente, sendo assim, este é um grande exemplo de sucesso desses estudos de componentes moleculares no câncer e um dos motivos que leva os pesquisadores a investirem grandes esforços nas buscas de outras informações relevantes como essa (5). A metodologia Block-sPLSDA exibe um crescimento e aprofundamento em análises de casos complexos, como é o caso do *multi-omics*, exigindo um empenho da comunidade científica que possui o objetivo de entender cada vez melhor o funcionamento de um organismo biológico e como trazer benefícios para este.

Foram testadas algumas metodologias do pacote *mixomics* (ACP, sPLS-DA e Block-sPLSDA), as quais apresentaram fácil aplicação e uma grande variedade de ferramentas gráficas. Problemas computacionais observados foram o longo tempo de processamento de algumas análises de performance e dificuldade de trabalhar com os gráficos do tipo "*network*" e "*heatmap*".

Na análise de componentes principais observou-se que são necessárias muitas componentes para se obter uma grande representação da variação dos dados. Ainda, mesmo que a análise não seja supervisionada, o subtipo de câncer parece ter pelo menos um pouco de influência na formação das primeiras componentes de cada bloco.

A metodologia sPLS-DA gerou resultados satisfatórios para os dados, já que para os blocos de mRNA e miRNA realizou predições com baixas taxas de erro de classificação balanceadas, respectivamente iguais a 2,54% e 7,62%, porém não foi possível avaliar o modelo de proteínas, uma vez que não tinha esse bloco para os dados de teste. Com isso, observou-se a primeira vantagem do método BsPLSDA, pois o modelo pode ser utilizado mesmo quando as novas observações possuem blocos faltantes. Nesse contexto, esse tipo de redução de dimensão pode poupar milhares de dólares.

Os dois modelos BsPLSDA gerados, com relações da matriz *design* de blocos iguais a $0,1$ ou iguais a 1, forneceram quantidades muito parecidas, sendo que no segundo caso a redução de dimensão foi menor. Com estes, realizou-se predições com baixas taxas de erro de classificação balanceadas, respectivamente iguais a 2,54% e 3,17%. Esses resultados envolveram a decisão de algumas quantidade e métricas para a modelagem dos dados. Alguns pontos foram definidos como padronizações, como a distância centróide e a taxa de erro de classificação balanceada. É possível optar também entre dois tipos de voto (por maioria ou com pesos) para realizar a predição em novas observações, porém observou-se

que o método é muito sensível a esta escolha, já que quando utilizada a maioria dos votos é muito mais provável que ocorra empates e, consequentemente, erros de classificação, logo a análise não foi feita como este tipo. Logo, estas análises consistem em várias etapas, as quais foram padronizadas com o intuito de realizar comparações.

Referências

1 Amrit Singh, Benoît Gautier, Casey P. Shannon, Michaël Vacher, Florian Rohart, Scott J. Tebbutt, Kim-Anh Lê Cao (2016) *DIABLO – an integrative, multi-omics, multivariate method for multi-group classification.* bioRxiv 067611; doi: https://doi.org/10.1101/067611 7, 18, 23, 25, 26, 27, 28, 37

2 Amrit Singh, Casey P. Shannon, Benoît Gautier, Florian Rohart, Michaël Vacher, Scott J. Tebbutt, Kim-Anh Lê Cao (2018) *DIABLO: from multi-omics assays to biomarker discovery, an integrative approach* bioRxiv 067611; doi: https://doi.org/10.1101/067611

3 Bruce Alberts et al. (2017) *Biologia molecular da célula* - 6. ed. – Porto Alegre : Artmed. 15

4 CAETANO, Alexandre Rodrigues (2009) *Marcadores SNP: conceitos básicos, aplicações no manejo e no melhoramento animal e perspectivas para o futuro.* R. Bras. Zootec., Viçosa , v. 38, n. spe, p. 64-71, July 2009. Disponível em: $<$`http://www.scielo.br/scielo.php?script=sci`$_a rttextpid = S1516 - 35982009001300008 lng = en nrm = iso > .Acessoem : 23Nov.2019.http : //dx.doi.org/10.1590/S1516 - 35982009001300008.$ 15

5 Cancer Genome Atlas Network. *Comprehensive molecular portraits of human breast tumours.* Nature. 2012;490(7418):61–70. doi:10.1038/nature11412 27, 51

6 David Martino, Rym Ben-Othman, Danny Harbeson and Anthony Bosco (2019) *Multiomics and Systems Biology Are Needed to Unravel the Complex Origins of Chronic Disease.* Challenges 2019, 10, 23; doi:10.3390/challe10010023 `www.mdpi.com/journal/challenges`

7 Gaughan L, Stockley J, Coffey K, et al. (2013) *KDM4B is a master regulator of the estrogen receptor signalling cascade.* Nucleic Acids Res. 2013;41(14):6892–6904. doi:10.1093/nar/gkt469 42

8 Hasin et al. Genome Biology (2017) *Multi-omics approaches to disease.* 18:83 DOI 10.1186/s13059-017-1215-1 15, 16

9 Hervé Abdi (2010) *Partial least squares regression and projection on latent structure regression (PLS Regression).* https://doi.org/10.1002/wics.51 22

10 JOHNSON, Richard Arnold; WICHERN, Dean W (2007) *Applied multivariate statistical analysis.* 6th ed. Upper Saddle River: Pearson Prentice Hall. 17, 18, 19, 20, 21

11 Kim-Anh Le Cao, Florian Rohart, Ignacio Gonzalez, Sebastien Dejean with key contributors Benoit Gautier, Francois Bartolo, contributions from Pierre Monget, Jeff Coquery, FangZou Yao and Benoit Liquet (2016) *mixOmics: Omics Data Integration Project*. R package version 6.1.1. `https://CRAN.R-project.org/package=mixOmics` 16, 18, 27

12 Lê Cao, K., Boitard, S. Besse, P. (2011) *Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems*. BMC Bioinformatics 12, 253 doi:10.1186/1471-2105-12-253

13 Molecular Subtypes of Breast Cancer. *Breastcancer.org*, 2019. Disponível em: <`https://www.breastcancer.org/symptoms/types/molecular-subtypes`>. Acesso em: 20 de nov. de 2019 27

14 The Cancer Genome Atlas Program. *National Cancer Institute*. Disponível em: <`https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga`>. Acesso em: 24 de nov. de 2019 27

15 R Development Core Team (2009) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org. 16, 18

16 Rohart F, Gautier B, Singh A, Le Cao K-A (2017) *mixOmics: An R package for 'omics feature selection and multiple data integration*. PLoS Comput Biol 13(11): e1005752. https://doi.org/ 10.1371/journal.pcbi.1005752 17, 22, 24, 25, 26, 29

17 T. Hastie, R. Tibshirani and M. Wainwright (2015) *Statistical Learning with Sparsity: The Lasso and Generalizations* (Chapman Hall/CRC Monographs on Statistics Applied Probability). 21

18 -Omes and -omics glossary taxonomy. *Cambridge Healthtech Institute*, 2019. Disponível em: <`http://www.genomicglossaries.com/content/omes.asp`>. Acesso em: 10 de ago. de 2019. 16