

UNIVERSIDADE ESTADUAL PAULISTA JÚLIO DE MESQUITA FILHO
INSTITUTO DE BIOCIÊNCIAS DE BOTUCATU
Departamento de Ciências Químicas e Biológicas

**OsteoCLUST: framework para clusterização da resposta de
biomateriais ósseos**

Matheus Amaral Silva
Mestre em Biotecnologia

BOTUCATU
2025

Matheus Amaral Silva

**OsteoCLUST: framework para clusterização da resposta de
biomateriais ósseos**

Dissertação apresentada ao Programa de
Pós-Graduação em Biotecnologia como
requisito para obtenção do título de mestre.

Orientador: Prof. Dr. Willian Fernando
Zambuzzi

Coorientador: Dr. Marcel Rodrigues
Ferreira

BOTUCATU
2025

S586o	Silva, Matheus Amaral OsteoCLUST: framework para clusterização da resposta de biomateriais ósseos / Matheus Amaral Silva. -- Botucatu, 2025 124 p. : il., tabs. Dissertação (mestrado) - Universidade Estadual Paulista (UNESP), Instituto de Biociências, Botucatu Orientador: Willian Fernando Zambuzzi Coorientador: Marcel Rodrigues Ferreira 1. Biomateriais. 2. Bioinformática. 3. Transcriptoma. 4. Aprendizado de máquina. 5. Tecido ósseo. I. Título.
-------	---

IMPACTO POTENCIAL DESTA PESQUISA

Esta pesquisa apresenta impactos significativos ao propor uma abordagem computacional inovadora para a análise da resposta celular a diferentes superfícies de biomateriais. A integração de ferramentas de bioinformática, geração de dados sintéticos e visualização interativa resultou em um framework capaz de oferecer subsídios relevantes para a avaliação e seleção de materiais com potencial biomédico, promovendo maior agilidade e precisão no desenvolvimento de soluções terapêuticas. Para além de seu mérito técnico-científico, a proposta atende a uma demanda ética e regulatória contemporânea ao explorar metodologias que contribuem para a redução do uso de animais em experimentação, o que fortalece sua aplicabilidade em contextos acadêmicos e industriais. Adicionalmente, os resultados deste trabalho contribuem de forma expressiva para o fortalecimento da capacidade nacional em ciência, tecnologia e inovação na área da saúde. Ao promover o desenvolvimento de uma base de dados biológicos preditiva e de ferramentas computacionais aplicáveis à análise de biomateriais, esta pesquisa amplia o potencial do país para gerar conhecimento estratégico, fomentar a autonomia tecnológica e impulsionar a bioindústria nacional, oferecendo suporte ao desenvolvimento de soluções mais eficazes e sustentáveis no campo da medicina regenerativa.



UNIVERSIDADE ESTADUAL PAULISTA

Câmpus de Botucatu



ATA DA DEFESA PÚBLICA DA DISSERTAÇÃO DE MESTRADO DE MATHEUS AMARAL SILVA, DISCENTE DO PROGRAMA DE PÓS-GRADUAÇÃO EM BIOTECNOLOGIA, DO INSTITUTO DE BIOCIÊNCIAS - CÂMPUS DE BOTUCATU.

Aos 28 dias do mês de abril do ano de 2025, às 9h, por meio de Videoconferência, realizou-se a defesa de DISSERTAÇÃO DE MESTRADO de MATHEUS AMARAL SILVA, intitulada **OsteoCLUST: framework para clusterização da resposta de biomateriais ósseos**. A Comissão Examinadora foi constituída pelos seguintes membros: Prof. Dr. WILLIAN FERNANDO ZAMBUZZI (Orientador(a) - Participação Virtual) do(a) Departamento de Ciências Químicas e Biológicas / Instituto de Biociências de Botucatu Unesp, Profa. Dra. SARAH SANTILONI CURY (Participação Virtual) do(a) Departamento de Ciências Químicas e Biológicas / Instituto de Biociências de Botucatu - UNESP, Prof. Dr. RAFAEL JUNQUEIRA BORGES (Participação Virtual) do(a) Centro de Química Medicinal / Universidade Estadual de Campinas. Após a exposição pelo mestrando e arguição pelos membros da Comissão Examinadora que participaram do ato, de forma presencial e/ou virtual, o discente recebeu o conceito final: Aprovado. Nada mais havendo, foi lavrada a presente ata, que após lida e aprovada, foi assinada pelo(a) Presidente(a) da Comissão Examinadora.

Prof. Dr. WILLIAN FERNANDO ZAMBUZZI

Prof. Titular: Willian Fernando Zambuzzi
Instituto de Biociências
Unesp Botucatu

DEDICATÓRIA

Dedico este trabalho ao meu avô, que sempre me apoiou com cuidado, dedicação e orgulho. Sua presença foi fundamental na minha formação pessoal, e seu exemplo de caráter e generosidade continua a me inspirar todos os dias. Tenho plena certeza de que, onde quer que ele esteja, está feliz e orgulhoso pelas conquistas que alcancei. Carrego comigo muito do que ele foi, nos valores que me ensinou, na maneira como enxergava a vida e na forma única que ele tinha de amar. A ele, minha eterna gratidão e o mais sincero agradecimento por tudo.

AGRADECIMENTOS

Agradeço, em primeiro lugar, ao meu orientador e ao meu coorientador, pela orientação, confiança e apoio contínuo ao longo desta trajetória acadêmica. Estendo minha gratidão aos meus pais e ao meu irmão, pelo suporte incondicional, presença constante e encorajamento nos momentos mais desafiadores. À minha namorada, registro um agradecimento especial pelo apoio emocional durante os períodos mais intensos deste processo e pela valiosa contribuição na revisão dos textos desta dissertação. Reconheço que esta conquista não seria possível sem a dedicação, incentivo e parceria dessas pessoas fundamentais na minha vida, às quais sou profundamente grato. Por fim, agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro concedido durante o período do mestrado.

RESUMO

As lesões ósseas constituem um dos problemas mais comuns na área médico-odontológica. A reconstrução de tecidos perdidos é um dos grandes desafios enfrentados pelos profissionais dessas áreas, exigindo muitas vezes a aplicação de biomateriais capazes de uma boa regeneração tecidual. Apesar da relevância médica, social e econômica de tais biomateriais, pouco se avançou em metodologias de bioanálise capazes de hierarquizar novos biomateriais. Nos últimos 10 anos o grupo LaBIO vem atuando no desenvolvimento de ferramentas que possam contornar esta problemática. O objetivo deste projeto foi avaliar diferentes *scores* de bioinformática computados sob um experimento (disponível em base de dados aberta) de sequenciamento do transcriptoma de células cultivadas sobre 21 diferentes biomateriais ósseos. Estes diferentes *scores* foram então utilizados para a realização de agrupamentos (clusterização) dos biomateriais e ao final foram disponibilizados os resultados na forma de um aplicativo (ou dashboard) para que usuários possam comparar seus resultados com a base de dados gerada. O aplicativo tem como nome: OsteoCLUST. Concluindo, pretende-se que o produto gerado sirva de parâmetro de referência para novas análises, focando sua difusão no setor produtivo na área de biomateriais.

PALAVRAS-CHAVES: Biomateriais; Bioinformática; Transcriptoma; Aprendizado de máquina; Tecido ósseo

ABSTRACT

Bone injuries are one of the most common problems in the medical and dental fields. The reconstruction of lost tissue is one of the major challenges faced by professionals in these areas, often requiring the application of biomaterials capable of good tissue regeneration. Despite the medical, social and economic relevance of such biomaterials, little progress has been made in bioanalysis methodologies capable of ranking new biomaterials. Over the last 10 years, the LaBIO group has been working to develop tools that can overcome this problem. The aim of this project was to evaluate different bioinformatics scores computed under an experiment (available in an open database) sequencing the transcriptome of cells grown on 21 different bone biomaterials. These different scores were then used to cluster the biomaterials and, at the end, the results were made available in the form of an application (or dashboard) so that users could compare their results with the database generated. The application is called OsteoCLUST. In conclusion, the aim is for the product generated to serve as a benchmark for new analyses, with a view to disseminating it in the biomaterials production sector.

KEYWORDS: Biomaterials; Bioinformatics; Transcriptome; Machine Learning; Bone Tissue

LISTA DE ILUSTRAÇÕES

Figura 1 – Métodos de Aprendizado de Máquina.	11
Figura 2 – Representação esquemática do comportamento dos dados nos principais métodos de aprendizado de máquina.	12
Figura 3 – Representação esquemática das distâncias.	13
Figura 4 – Representação esquemática do comportamento dos dados quando aplicado o método da PCA. FONTE: (Claus O. Wilke, 2019)	16
Figura 5 – Representação por meio de histogramas, das distribuições dos scores GSVA (com e sem o grupo controle POLY), GSEA e ssGSEA, gerados para as assinaturas moleculares presentes no conjunto Hallmarks.	29
Figura 6 - Representação por meio de boxplot, das distribuições dos scores GSVA (com e sem o grupo controle POLY), GSEA e ssGSEA, gerados para as assinaturas moleculares presentes no conjunto Hallmarks.	30
Figura 7 - Mapa de calor correlacionando os 21 biomateriais analisados com os scores GSVA (com o grupo POLY) gerados para cada uma das 50 assinaturas moleculares do conjunto gênico Hallmarks. Juntamente com um segundo mapa representando os scores de expressão gênica ao longo de 7 dias (D0, D3 e D7), utilizados como controle biológico da progressão osteogênica para comparação com os perfis dos biomateriais.	33
Figura 8 - Mapa de calor correlacionando os 21 biomateriais analisados com os scores GSEA gerados para cada uma das 50 assinaturas moleculares do conjunto gênico Hallmarks. Juntamente com um segundo mapa representando os scores de expressão gênica ao longo de 7 dias (D0, D3 e D7), utilizados como controle biológico da progressão osteogênica para comparação com os perfis dos biomateriais. Os valores de todas as assinaturas em D0 foram manualmente definidas como sendo zero.	34
Figura 9 - Mapa de calor correlacionando os 21 biomateriais analisados com os scores ssGSEA gerados para cada uma das 50 assinaturas moleculares do conjunto gênico Hallmarks. Juntamente com um segundo mapa representando os scores de expressão gênica ao longo de 7 dias (D0, D3 e D7), utilizados como controle biológico da progressão osteogênica para comparação com os perfis dos biomateriais.	35
Figura 10 - Scree plot dos 21 biomateriais analisados para o conjunto Hallmarks, com base nos scores GSVA (com POLY), GSVA (sem POLY), GSEA e ssGSEA, respectivamente.	39

Figura 11 - Análise de componentes principais (PCA), analisando a distribuição dos 21 biomateriais para o conjunto Hallmarks, com base na presença ou não de TCP na composição, sendo avaliados para os scores GSVA (com POLY), GSVA (sem POLY), GSEA e ssGSEA, respectivamente. Estudo feito para 1ª e 2ª dimensão..... 40

Figura 12 - Análise de componentes principais (PCA), analisando a distribuição dos 21 biomateriais para o conjunto Hallmarks, com base na presença ou não de Hidroxiapatita na composição, sendo avaliados para os scores GSVA (com POLY), GSVA (sem POLY), GSEA e ssGSEA, respectivamente. Estudo feito para 1ª e 2ª dimensão. 41

Figura 13 - Análise de componentes principais (PCA), analisando a distribuição dos 21 biomateriais para o conjunto Hallmarks, com base na presença ou não de PLA na composição, sendo avaliados para os scores GSVA (com POLY), GSVA (sem POLY), GSEA e ssGSEA, respectivamente. Estudo feito para 1ª e 2ª dimensão..... 42

Figura 14 - Análise de componentes principais (PCA), analisando a distribuição dos 21 biomateriais para o conjunto Hallmarks, com base na presença ou não de Titânio na composição, sendo avaliados para os scores GSVA (com POLY), GSVA (sem POLY), GSEA e ssGSEA, respectivamente. Estudo feito para 1ª e 2ª dimensão. 43

Figura 15 - Contagem de dados gerados sinteticamente para score GSVA (sem POLY), avaliando a presença ou não dos compósitos TCP, Hydroxyapatite, PLA e Titanium 47

Figura 16 - Contagem de dados gerados sinteticamente para score GSVA (com POLY), avaliando a presença ou não dos compósitos TCP, Hydroxyapatite, PLA e Titanium. 48

Figura 17 - Contagem de dados gerados sinteticamente para score GSEA, avaliando a presença ou não dos compósitos TCP, Hydroxyapatite, PLA e Titanium. 49

Figura 18 - Contagem de dados gerados sinteticamente para score ssGSEA, avaliando a presença ou não dos compósitos TCP, Hydroxyapatite, PLA e Titanium..... 50

Figura 19 - Mapa de calor correlacionando os 150 biomateriais gerados sinteticamente com base no score GSVA (com POLY) para cada uma das 50 assinaturas moleculares presentes no conjunto gênico Hallmarks. 51

Figura 20 - Mapa de calor correlacionando os 150 biomateriais gerados sinteticamente com base no score GSVA (sem POLY) para cada uma das 50 assinaturas moleculares presentes no conjunto gênico Hallmarks. 52

Figura 21 - Mapa de calor correlacionando os 150 biomateriais gerados sinteticamente com base no score GSEA para cada uma das 50 assinaturas moleculares presentes no conjunto gênico Hallmarks. 53

Figura 22 - Mapa de calor correlacionando os 150 biomateriais gerados sinteticamente com base no score ssGSEA para cada uma das 50 assinaturas moleculares presentes no conjunto gênico Hallmarks.	54
Figura 23 - Visão geral da primeira versão do aplicativo.	61
Figura 24 - Opções de escolha da aba "Score".	62
Figura 25 - Opções de escolha da aba "Distance".	62
Figura 26 - Identificação do input para análise de novos biomateriais.	63
Figura 27 - Identificação da quantidade de clusters desejados e possível exportação dos resultados.	63
Figura 28 - Mapa de calor correlacionando os 21 biomateriais analisados, incluindo os biomateriais do estudo de Murali et al. (2024), com os scores GSVA (com o grupo POLY) calculados para cada uma das 50 assinaturas moleculares pertencentes ao conjunto gênico Hallmarks.	65
Figura 29 - Mapa de calor correlacionando os 21 biomateriais analisados, incluindo os biomateriais do estudo de Murali et al. (2024), com os scores ssGSEA (com o grupo POLY) calculados para cada uma das 50 assinaturas moleculares pertencentes ao conjunto gênico Hallmarks.	66

LISTA DE TABELAS

Tabela 1 - Relação de biomateriais contidos no array E-MTAB-7219, com a descrição do número de replicatas utilizadas para cada biomaterial (n). As colunas TCP, Titanium, PLA e Hydroxyapatite contém informações sobre o biomaterial conter, ou não, este composto em sua composição que serão utilizadas para a construção dos modelos de classificação.	22
Tabela 2 - Relação dos números de genes Up e Down regulados e o total de genes diferencialmente expressos para cada biomaterial contidos no array E-MTAB-7219, com exceção do Poliestireno (POLY).	27
Tabela 3 - Relação dos conjuntos de genes, disponíveis no MSigDB, com as quantidades de assinaturas moleculares presentes utilizadas para a avaliação da pontuação de enriquecimento.	32
Tabela 4 - Desempenho das métricas de machine learning nos conjuntos de treino e teste do GSEA sem POLY, exibido separadamente para as composições PLA, Hydroxyapatite, TCP e Titanium, respectivamente. Cores verdes indicam maiores valores e cores vermelhas indicam valores menores.	56
Tabela 5 - Desempenho das métricas de machine learning nos conjuntos de treino e teste do GSVA sem POLY, exibido separadamente para as composições PLA, Hydroxyapatite, TCP e Titanium, respectivamente. Cores verdes indicam maiores valores e cores vermelhas indicam valores menores.	57
Tabela 6 - Desempenho das métricas de machine learning nos conjuntos de treino e teste do GSVA com POLY, exibido separadamente para as composições PLA, Hydroxyapatite, TCP e Titanium, respectivamente. Cores verdes indicam maiores valores e cores vermelhas indicam valores menores.	58
Tabela 7 - Desempenho das métricas de machine learning nos conjuntos de treino e teste do ssGSEA com POLY, exibido separadamente para as composições PLA, Hydroxyapatite, TCP e Titanium, respectivamente. Cores verdes indicam maiores valores e cores vermelhas indicam valores menores.	59

SUMÁRIO

CAPÍTULO I: INTRODUÇÃO GERAL E OBJETIVOS.....	1
1 INTRODUÇÃO.....	1
1.1 Tecido e regeneração óssea	1
1.2 Biomateriais.....	3
1.3 Bioinformática	7
1.4 Machine Learning.....	9
1.5 Justificativa.....	18
2 OBJETIVOS	20
CAPÍTULO II: CLUSTERIZAÇÃO, DADOS SINTÉTICOS E SHINY	21
1 RESUMO DO CAPÍTULO	21
2 MATERIAIS E MÉTODOS.....	22
2.1 Dados utilizado	22
2.2 Análise de dados	23
2.3 Análise de expressão diferencial e enriquecimento.....	24
2.4 Análise de clusterização hierárquica	24
2.5 Análises de redução de dimensionalidade	25
2.6 Geração de dados sintéticos.....	25
2.7 Modelos de predição.....	25
3 RESULTADOS E DISCUSSÃO	26
3.1 Genes diferencialmente expressos.....	27
3.2 Estudo de distribuição dos dados.....	27
3.3 Scores e mapa de calor das assinaturas moleculares	32
3.4 Análises de redução de dimensionalidade	38
3.5 SynthPop: Dados sintético do conjunto Hallmarks	45
3.6 Modelos de predição.....	55
3.7 ShinyApp: OsteoCLUST	60
4 CONSTATAÇÕES.....	67

5	CONCLUSÃO GERAL	69
	REFERÊNCIAS	70
	APÊNDICE e ANEXO	79

CAPÍTULO I: INTRODUÇÃO GERAL E OBJETIVOS

1 INTRODUÇÃO

Defeitos ósseos constituem um dos problemas mais presentes na área médico-odontológica. A reconstrução de um tecido perdido é um dos grandes desafios para os profissionais dessas áreas, necessitando, muitas vezes, da aplicação de um biomaterial que permita uma boa regeneração tecidual. Nesta área, os materiais osteo-substitutos constituem uma das mais comuns e antigas formas de terapia regenerativa (Kao & Scott, 2007; McAllister & Haghighat, 2007). Até o presente momento, não foi desenvolvido nenhum material sintético que reúna as propriedades e capacidades de um material autógeno, denominado também como material ideal. Apesar da relevância clínica, social e econômica de tais biomateriais, poucos são os avanços quanto a metodologias de bioanálise capazes de ranquear novos biomateriais. Consequentemente, universidades e empresas recorrem, quase sempre, a testes tradicionais, muitas vezes com experimentação animal como modelo. Diante desta problemática, torna-se evidente a necessidade de explorar ferramentas de bioinformática, para que assim possam auxiliar a experimentação de novos biomateriais, otimizando mecanismos de bioanálise.

1.1 Tecido e regeneração óssea

O tecido ósseo é um tipo especializado de tecido conjuntivo formado por células e matriz extracelular calcificado conferindo-lhe rigidez e mantendo um certo grau de elasticidade (Junqueira & Carneiro, 2004). É componente principal do esqueleto, cumprindo a função de suporte para os tecidos moles, protegendo órgãos vitais, dando apoio aos músculos esqueléticos e constituindo um sistema de alavancas que amplia as forças geradas na contração muscular (Junqueira & Carneiro, 2004). Além dessas funções, o tecido ósseo é a maior reserva primária de cálcio no organismo, íon que participa intensamente da manutenção do pH interno no corpo, tal como da transmissão e condução de impulso elétrico nos nervos e músculos, incluindo o músculo cardíaco (Dempster & Zhou, 1999).

A matriz óssea é composta por elementos orgânicos, principalmente por colágeno tipo I, proteoglicanos e glicoproteínas, e por elementos inorgânicos. A porção inorgânica é constituída por sais de cálcio e fosfato de cristais de hidroxiapatita (Junqueira & Carneiro, 2004; Lirani, 2004). Os principais tipos de células ósseas são os osteoblastos, osteoclastos e osteócitos, sendo todas derivadas de uma mesma linha de células osteoprogenitoras (Lirani, 2004). Os osteoblastos são responsáveis pela osteogênese (síntese e secreção da matriz

orgânica), além de armazenar minerais e revestirem a maior parte das superfícies ósseas. Quando ativos, os osteoblastos se hipertrofiam e o núcleo (com nucléolo evidente) se desloca para as extremidades da superfície óssea (Banks, 1992; Junqueira & Carneiro, 2004). Os osteoclastos são células gigantes, móveis e multinucleadas, derivadas dos macrófagos, que reabsorvem o tecido ósseo, participando dos processos de remodelação dos ossos. Os osteócitos são osteoblastos maduros incorporados à matriz óssea mineralizada, participando da sua manutenção (Lirani, 2004).

A consolidação é um processo que pode ser dividido em três fases principais: inflamatória, reparo e remodelação (Banks, 1992; Dyce et al., 2004; Junqueira & Carneiro, 2004; Sousa, 2003). A fase inflamatória tem início imediatamente após a fratura no tecido e ocorre aproximadamente nas primeiras duas a três semanas. No período de 48 horas o exsudato (líquido com alto teor de proteínas e leucócitos) da lesão contém mediadores inflamatórios, fatores angiogênicos e de crescimento liberados pelas plaquetas, células locais, mastócitos, macrófagos, neutrófilos e linfócitos (Dyce et al., 2004; Sousa, 2003).

Ao final da fase inflamatória, temos o processo de angiogênese, sendo esse marcado pela formação de novos vasos, que é induzida pelo fator de crescimento vascular endotelial e esse processo ocorre antes da etapa de osteogênese. A angiogênese, estágio essencial na osteogênese, é um processo importante para a cicatrização do local lesionado, visto que as células dos capilares possuem uma relação direta na formação óssea devido à sua proximidade com os osteoblastos e células osteoprogenitoras nos sítios da nova formação óssea (Pelissier et al., 2004). A formação de uma rede capilar no início do processo de reparação é fundamental para que o tecido ósseo possa ser regenerado e, dessa forma, tenha nutrientes para o seu desenvolvimento (Dyce et al., 2004; Junqueira & Carneiro, 2004).

Nos estágios iniciais do processo de reparo, a característica mais evidente é a proliferação de células do periósteo próximas à fratura, sendo essas células as precursoras de osteoblastos (Crocì et al., 2003). Após o amadurecimento do tecido celular formado devido a fratura, as células originam os osteoblastos. Por sua vez, essas células maduras depositam uma matriz intercelular de colágeno e polissacarídeos impregnada de sais de cálcio, formando dessa maneira o osso primário, que gradualmente, pela ação dos osteoblastos, se transformará em osso amadurecido, com uma estrutura lamelar típica conhecida como calo ósseo (Betti, 2004; Crocì et al., 2003). O processo de remodelação óssea tem início com a reabsorção e remodelação do calo formado (Betti, 2004). Esse processo é realizado pelos osteoclastos e

osteoblastos, removendo e depositando osso respectivamente (Sousa, 2003). Essas células juntas são denominadas como Unidade Óssea Multicelular.

1.2 Biomateriais

Pode-se definir biomaterial como sendo qualquer substância ou combinação de substâncias, naturais ou sintéticas, que não sejam drogas ou fármacos, que podem ser utilizadas em aplicações biomédicas e que interagem com sistemas biológicos, tratando, aumentando ou substituindo quaisquer tecidos, órgãos ou funções do corpo (Gil & Ferreira, 2006; Von Recum & Laberge, 1995). Ainda, o Conselho Consultivo para Biomateriais da Universidade de Clemson definiu formalmente biomateriais como sendo “substâncias sistemicamente e farmacologicamente inertes projetadas para implantação ou incorporação em sistemas vivos” (Park & Lakes, 2017).

Historicamente, o uso de biomateriais para substituir a perda óssea tem sido uma prática comum (Chow, 2009). Inicialmente, para restituir perdas ósseas eram utilizados autoenxertos, dado como ideal por ser constituído do material do próprio indivíduo. No entanto, devido à maior incidência de complicações, outras técnicas foram desenvolvidas ao longo do tempo. Diante das limitações dos autoenxertos, os aloenxertos (compostos de materiais de outro indivíduo da mesma espécie) e os xenoenxertos (materiais obtidos de outra espécie) surgiram como alternativas. Contudo, essas opções também apresentam limitações, como o risco de rejeição ou de transmissão de doenças (Precheur, 2007).

Nesse contexto, destacam-se os enxertos aloplásticos, que correspondem aos biomateriais propriamente ditos. Eles podem ser classificados conforme sua origem (natural ou sintética) ou sua composição química (metálicos, cerâmicas, polímeros ou compósitos) (Oliveira et al., 2009). Atualmente, diversos biomateriais estão disponíveis para aplicação como substitutos do tecido ósseo, variando não apenas quanto à origem e composição, mas também quanto à ação mecânica e configuração espacial (blocos sólidos, lâminas, esponjas porosas e hidrogéis) (Abukawa et al., 2006; Giannoudis et al., 2005).

Com o avanço na área, os biomateriais passaram a ser classificados em três gerações. A primeira geração compreende os materiais bioinertes, cujo objetivo era evitar qualquer tipo de reação indesejada no organismo (Hench, 1980). A segunda geração introduziu os conceitos de biocompatibilidade – que diz respeito à interação material-organismo - e biodegradabilidade – a capacidade do material de se degradar ou solubilizar no organismo (Pereira et al., 1999;

Tabata, 2009). Esses conceitos conferem ao biomaterial uma interação favorável com o organismo em que foi implantado, desde que: (i) o material não incite uma resposta inflamatória ou tóxica; (ii) o tempo estimado de degradação do material permita a ocorrência de regeneração ou cura do sítio afetado; (iii) o material apresente propriedades mecânicas adequadas à aplicação para que foi designado; (iv) a degradação gere produtos que sejam facilmente metabolizados e liberados pelo corpo, sem gerar resíduos tóxicos (Cao & Wang, 2009). Por fim, a terceira geração inclui materiais capazes de estimular respostas celulares específicas no nível molecular (Hench & Polak, 2002).

Em 2024 o mercado global de biomateriais foi estimado em 205 bilhões de dólares. Espera-se para os anos de 2024 a 2030 que registre uma taxa de crescimento anual composta (CAGR) de 15,6%, impulsionada pela crescente prevalência de doenças musculoesqueléticas e crônicas, além do envelhecimento populacional — fatores que aumentam a demanda por soluções associadas a osteoartrite, osteoporose e outras condições relacionadas (Grand View Research, 2024).

A seleção adequada de biomateriais representa um elemento fundamental no desenvolvimento de estratégias terapêuticas voltadas à regeneração óssea. Neste contexto, destacam-se materiais amplamente reconhecidos tanto na prática clínica quanto na literatura científica, como o titânio (Ti), a hidroxiapatita (HA), o ácido polilático (PLA) e o fosfato tricálcico (TCP). A escolha desses compostos como foco de análise nesta pesquisa fundamenta-se, portanto, tanto na sua relevância no mercado de biomateriais quanto na robustez das evidências que sustentam seu desempenho biológico e físico-químico.

O titânio é um metal amplamente utilizado como biomaterial devido às suas excelentes propriedades mecânicas, elevada resistência à corrosão e biocompatibilidade. A formação espontânea de uma camada de óxido na superfície do titânio proporciona elevada estabilidade química, permitindo sua integração ao tecido ósseo por meio do processo de osseointegração (união estável e funcional entre o osso e uma superfície de titânio) (Long & Rack, 1998). Seu uso é predominante em implantes ortopédicos e dentários, como próteses articulares, placas, parafusos e implantes endósseos (Geetha et al., 2009).

A hidroxiapatita (HA) é um fosfato de cálcio que compõe a matriz mineral dos ossos e dentes humanos. Sua elevada bioatividade e semelhança com o tecido ósseo natural a tornam um material amplamente utilizado em recobrimentos de implantes metálicos, substitutos ósseos e *scaffolds* empregados em engenharia tecidual (Dorozhkin, 2010). Entretanto, sua baixa

resistência mecânica limita seu uso estrutural isolado, sendo frequentemente utilizada em combinação com outros materiais, como o titânio ou polímeros biodegradáveis.

O ácido polilático (PLA) é um polímero biodegradável derivado de fontes renováveis, como o ácido láctico. Apresenta biocompatibilidade, propriedades mecânicas ajustáveis e degrada-se por via hidrolítica em subprodutos atóxicos, características que o tornam particularmente adequado para aplicações temporárias, como suturas reabsorvíveis, sistemas de liberação controlada de fármacos e *scaffolds* para regeneração tecidual (Middleton & Tipton, 2000). Além disso, sua capacidade de ser moldado em diferentes formatos favorece seu uso em impressões 3D para criação de estruturas personalizadas.

Já o fosfato tricálcico (TCP), especialmente na forma β -TCP, é um fosfato de cálcio bioabsorvível, também utilizado como substituto ósseo. Sua taxa de reabsorção, superior à da hidroxiapatita, favorece a substituição progressiva pelo tecido ósseo regenerado (LeGeros, 2002). O TCP é empregado em cirurgias ortopédicas e odontológicas, frequentemente combinado com HA ou PLA para formar compósitos que associam bioatividade, biodegradação controlada e resistência mecânica adequada.

A escolha dos biomateriais deve considerar não apenas suas propriedades físico-químicas e biológicas, mas também o contexto clínico de aplicação, a forma de processamento e as exigências regulatórias para uso em seres humanos.

A avaliação da biocompatibilidade de novos biomateriais exige a aplicação de métodos *in vitro* robustos e reprodutíveis, capazes de prever a resposta celular frente a diferentes superfícies, composições e topografias. Esses ensaios são amplamente utilizados como etapa inicial em pesquisas de engenharia tecidual, pois permitem investigar parâmetros como adesão, proliferação, diferenciação e expressão gênica, sem a necessidade imediata de experimentação animal (Anselme, 2000).

Na engenharia óssea, uma das linhagens celulares mais utilizadas é a hFOB 1.19 (CRL-3602), derivada de osteoblastos fetais humanos. Essa linhagem apresenta um fenótipo osteoblástico compatível com o ambiente fisiológico, incluindo a expressão de marcadores relacionados à mineralização da matriz extracelular, como fosfatase alcalina e osteocalcina (Dr. Harris et al., 1995). No entanto, seu cultivo apresenta desafios técnicos significativos, uma vez que exige condições específicas de temperatura, sendo idealmente cultivada a 33,5 °C (Dr. Harris et al., 1995). Essa particularidade impõe limitações em termos de infraestrutura

laboratorial e pode afetar a reprodutibilidade dos experimentos entre diferentes centros de pesquisa.

Diante dessas limitações impostas por linhagens celulares mais exigentes, alternativas viáveis têm sido amplamente empregadas, como as linhagens celulares imortalizadas derivadas de osteossarcoma humano, destacando-se, entre elas, as MG-63 e Saos-2. Essas células são de fácil cultivo em condições padrão (37 °C, 5% CO₂), crescimento acelerado, manutenção estável em cultura e expressão de marcadores osteoblásticos em níveis comparáveis aos de osteoblastos primários (Smeets et al., 2009; Tang et al., 2021). A linhagem MG-63, em especial, tem se mostrado eficaz em ensaios de adesão e proliferação celular, sendo amplamente utilizada na avaliação de diferentes biomateriais, incluindo superfícies metálicas, cerâmicas e compósitos bioativos (Lu & Leng, 2003).

Apesar das vantagens operacionais, o uso da MG-63 como substituto de osteoblastos humanos normais apresenta limitações. Por se tratar de uma linhagem tumoral, sua resposta biológica pode divergir da observada em osteoblastos primários, especialmente em aspectos como sinalização intracelular e metabolismo ósseo (Czekanska et al., 2012). Por isso, os dados obtidos com MG-63 devem ser interpretados com cautela e, sempre que possível, validados por meio de modelos celulares primários. Ainda assim, sua origem humana confere uma representatividade clínica relevante (Smeets et al., 2009).

Nos ensaios *in vitro*, é comum o uso de superfícies tratadas com substratos que favorecem a adesão celular, como o poliestireno (POLY). Essas superfícies funcionam como controles positivos, nos quais as células aderem e proliferam antes de serem expostas ao biomaterial de interesse. A transição da superfície padrão para o novo material marca o início efetivo do experimento, permitindo a avaliação das interações célula-material em condições controladas.

Dessa forma, a seleção criteriosa da linhagem celular, aliada à padronização dos protocolos de cultivo, é fundamental para assegurar a confiabilidade dos dados obtidos em sistemas *in vitro*, contribuindo para o desenvolvimento de biomateriais voltados à regeneração óssea.

1.3 Bioinformática

A Bioinformática pode ser definida como um sistema de informação gerencial para a biologia, ou seja, pela aplicação de técnicas computacionais para compreender e organizar informações associadas às macromoléculas biológicas. Sendo esse um campo interdisciplinar que aplica técnicas derivadas de disciplinas como matemática aplicada, estatística e informática no tratamento de dados biológicos. Englobando o desenvolvimento e o uso de ferramentas computacionais com o objetivo de expandir e tornar mais acessível o uso de dados biológicos por meio da capacidade de coletar, armazenar, analisar e disseminar tais dados, de acordo com o NIH - NCBI (*National Institutes of Health – National Center for Biotechnology Information*).

Os processos biológicos geralmente envolvem proteínas, que por sua vez são sintetizadas a partir do fluxo de informação codificada no DNA. No entanto, o mecanismo pelo qual uma via bioquímica ou uma proteína interage com outros componentes biológicos é um problema de escala muito maior e mais difícil de ser abordado. São necessárias técnicas para realizar experimentos em larga escala e ferramentas para armazenar e analisar os dados gerados por esses experimentos. Sendo assim, desde o surgimento da bioinformática, a área de conhecimento da biologia molecular é a que está mais intimamente ligada a essas técnicas, principalmente quando se trata de análises de sequências genéticas e de DNA, do funcionamento de estruturas e funções de proteínas e da regulação e expressão gênica.

A análise da expressão gênica pode ser aplicada e estudada para diferentes temáticas. O que se sabe atualmente é que há grandes diferenças de expressão gênica de osteoblastos face às diferentes superfícies químicas (Sohn et al., 2006; Song et al., 2008). Portanto, explorar o efeito de diferentes substratos no fenótipo diferencial de osteoblastos pode ter importantes implicações no entendimento dos mecanismos pelos quais as células respondem a diferentes substratos, auxiliando no desenvolvimento de novos biomateriais como eficazes substitutos artificiais do osso.

Estudos de expressão gênica, geralmente, se concentram na elaboração de métodos para agrupar os genes por semelhança nos perfis de expressão. Os métodos mais comuns utilizados nessas análises são a clusterização hierárquica, mapas auto-organizadores e agrupamento *k-means*. Métodos hierárquicos agrupam os genes de forma “*bottom-up*”, assim os genes com perfis de expressão mais próximos serão agrupados primeiro, e aqueles com perfis diversos serão incluídos de maneira iterativa (Alon et al., 1999; Eisen et al., 1998; Wen et al., 1998). Em contrapartida, o mapa auto-organizador (Tamayo et al., 1999; Törönen et al., 1999) e o método *k-means* (Tavazoie et al., 1999) trabalham com a abordagem “*top-down*” em que o usuário pré-

determina o número de *clusters* para o conjunto de dados desejado. Inicialmente os *clusters* são determinados de maneira aleatória até encontrar uma maneira em que os genes serão agrupados de forma ideal para o estudo em questão. O uso dos métodos de clusterização para esse tipo de análise é interessante, pois dessa maneira se torna possível relacionar os dados de expressão com outros atributos biológicos, tais como estrutura e função de determinado produto.

Antes de aplicarmos os métodos de *machine learning* propriamente ditos, um processo de extrema importância é o manuseio dos *datasets* (conjunto de dados). Os *datasets* são componentes fundamentais para qualquer projeto na área de ciência de dados ou de *machine learning*. São bases de dados específicas utilizadas para que os algoritmos consigam treinar, aprender e exibir resultados consistentes.

Dependendo da problemática analisada, pode ser necessário utilizar métodos para a geração de dados complementares. A estes, damos o nome de dados sintéticos. O uso desse tipo de dado tem como objetivo aumentar os dados rotulados disponíveis, aumentando assim, o conjunto de dados e consequentemente a efetividade dos algoritmos e dos modelos de *machine learning* (Dahmen & Cook, 2019). Os dados gerados sinteticamente consistem em observações completamente novas e artificiais, que possuem as mesmas propriedades estatísticas que os dados reais (El Emam et al., 2020).

Temos três tipos de dados sintéticos. O primeiro tipo de dado é gerado a partir de conjuntos de dados reais, o segundo tipo não utiliza dados reais, e o terceiro tipo é um híbrido do primeiro e do segundo.

Dados sintéticos gerados com base em regras são dados gerados com um conjunto predefinido de especificações, sintetizados a partir de conjuntos de dados reais. São gerados por meio de modelos de aprendizado de máquina com base nos dados iniciais para capturar qualquer característica, relação ou padrão estatístico desejado. Se o modelo criado para a síntese desses dados for uma boa representação dos dados originais, então os dados gerados sinteticamente terão propriedades estatísticas similares às dos dados reais (El Emam et al., 2020).

Os dados sintéticos do segundo tipo não são gerados a partir de dados reais. Sendo assim, são obtidos por algoritmos com modelos existentes. Os modelos utilizados podem ser modelos estatísticos ou também simulações. Nesse caso, provavelmente, os dados gerados não possuirão as mesmas propriedades dos dados reais, entretanto podem ser úteis para alguns tipos de problemas, como, por exemplo, testes de desempenho de um *software* (El Emam et al., 2020).

Os benefícios que encontramos com esses tipos de dados são: não há restrições de uso como há em dados reais, visto que são isentos de regulamentações de privacidade; permite aumentar e simular dados completamente novos, no caso de escassez de dados, por exemplo; coleta mais rápida do que quando comparada a dados do mundo real; podem não conter problemas estatísticos que prejudicam a análise; aumentam a heterogeneidade de um conjunto de dados de treinamento para resultar em um modelo de *machine learning* mais robusto (El Emam et al., 2020).

Nos últimos anos, o grupo LaBIO busca desenvolver ferramentas de bioinformática que auxiliem no desenvolvimento de biomateriais. A primeira iniciativa nesse sentido foi batizada de *OsteoBLAST*, na qual dados do perfil de fosforilação global de células cultivadas sob diferentes biomateriais eram utilizados para a comparação destes utilizando uma métrica de distância simples desenvolvida durante o mestrado do aluno Marcel Ferreira (Ferreira et al., 2020). Outra iniciativa desenvolvida foi a aplicação da métrica desenvolvida no trabalho do *OsteoBLAST* em dados de transcriptoma, porém utilizando-se do enriquecimento de conjuntos gênicos pelo método GSVA em assinaturas moleculares curadas da base de dados MSigDB (Ferreira et al., 2021).

1.4 Machine Learning

O aprendizado de máquina, também conhecido como *machine learning*, constitui uma subárea fundamental da inteligência artificial (IA), caracterizando-se pela capacidade de sistemas computacionais aprenderem automaticamente padrões e relações a partir de dados, sem a necessidade de programação explícita. Essa abordagem possibilita a análise de múltiplos tipos de dados, como imagens, textos e números, com ampla aplicabilidade em diferentes domínios (Bishop, 2006; Goodfellow et al., 2016; Murphy, 2022).

A IA compreende qualquer técnica que permita aos computadores imitar o comportamento humano e reproduzir ou se sobrepor à tomada de decisões humanas para a resolução de tarefas complexas de forma independente ou com o mínimo de intervenção humana (Russell & Norvig, 2021). Como tal, ela se preocupa com uma variedade de problemas centrais, incluindo representação do conhecimento, raciocínio, aprendizagem, planejamento, percepção e comunicação, e se refere a uma variedade de ferramentas e métodos (por exemplo, raciocínio baseado em casos, sistemas baseados em regras, algoritmos genéticos, modelos difusos, sistemas multi-agentes) (Chen et al., 2008).

A eficiência do aprendizado de máquina em resolver problemas complexos reside na construção de modelos analíticos, que geram previsões, regras, recomendações ou outros resultados similares. Esses modelos são desenvolvidos por meio de algoritmos iterativos, capazes de identificar padrões complexos e informações ocultas nos dados, dispensando a necessidade de instruções programáticas explícitas (Bishop, 2006). O progresso tecnológico recente, marcado pela maior disponibilidade de dados e pelo aumento exponencial do poder computacional, impulsionou o uso do aprendizado de máquina em áreas como saúde, finanças, ciência dos dados e bioinformática (Brynjolfsson & McAfee, 2017).

Considerando que o aprendizado de máquina permite a análise de diversos tipos de dados, podemos nos deparar com situações em que existe uma variável de interesse no conjunto de dados, ou seja, uma variável que desejamos analisar e utilizar para prever resultados. Por outro lado, também há casos em que o conjunto de dados não contém uma variável de interesse, o que exige uma abordagem diferente. Diante dessas variações nos dados e nos objetivos da análise, surgem três principais tipos de aprendizado: supervisionado, não supervisionado e por reforço.

O aprendizado supervisionado aplica-se a contextos em que o objetivo é prever uma variável-alvo a partir de variáveis preditoras. Essa abordagem exige conjuntos de dados rotulados, nos quais os valores esperados são conhecidos. Os algoritmos mais utilizados nesse contexto incluem Regressão Linear, Regressão Logística, Máquinas de Vetores de Suporte (*Support Vector Machines*), Árvores de Decisão, Florestas Aleatórias (*Random Forests*), Redes Neurais Artificiais, Naive Bayes e Gradient Boosting Machines (Janiesch et al., 2021). A escolha do método depende do tipo de variável-alvo: métodos de regressão são utilizados para variáveis contínuas, enquanto métodos de classificação são adequados para variáveis categóricas (Angermueller et al., 2016). A construção de modelos supervisionados requer a divisão do conjunto de dados em subconjuntos de treinamento e teste, de modo a calibrar os parâmetros do modelo e validar sua aplicabilidade. Essa abordagem é amplamente utilizada em problemas de classificação, análise preditiva e previsão de séries temporais.

O aprendizado não supervisionado, por outro lado, aplica-se a situações em que não há rótulos nos dados de entrada, sendo o objetivo principal explorar a estrutura intrínseca dos dados (Janiesch et al., 2021). Dessa forma, os dados de treinamento consistem apenas de variáveis de entrada (variáveis independentes) com o objetivo de encontrar informações estruturais de interesse. As técnicas mais comumente empregadas incluem K-médias (k-means); Clusterização Hierárquica; DBSCAN (*Density-Based Spatial Clustering of*

Applications with Noise); Análise de Componentes Principais (PCA); *t-Distributed Stochastic Neighbor Embedding* (t-SNE); *Autoencoders*; Modelos de Mistura Gaussiana (GMM); Mapas Auto-organizados (SOM) e Latent Dirichlet Allocation (LDA). Esses métodos têm aplicação em problemas como redução de dimensionalidade, agrupamento e descoberta de padrões em grandes volumes de dados (Bishop, 2006; Lever et al., 2017). Por não dispor de uma variável-alvo pré-definida, os resultados obtidos são interpretados de maneira mais subjetiva, sendo necessária uma análise criteriosa para determinar sua aplicabilidade ao contexto específico do problema.

De forma geral, os métodos de aprendizado de máquina podem ser classificados em grandes categorias, tais como Classificação, Regressão, Sumarização, Agrupamento e Associação. Cada uma dessas categorias pode ser abordada por meio de técnicas e métodos específicos, como ilustrado na **Figura 1**. O comportamento dos dados tende a ser semelhante dentro de cada um desses grupos, como demonstrado na **Figura 2**, que destaca os principais padrões observados.

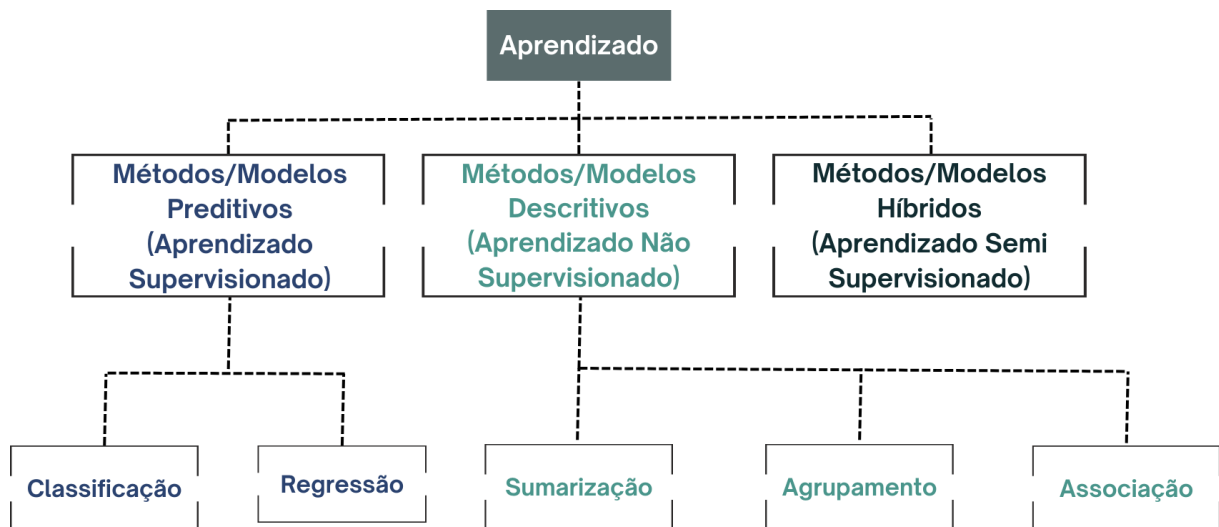


Figura 1 – Métodos de Aprendizado de Máquina.

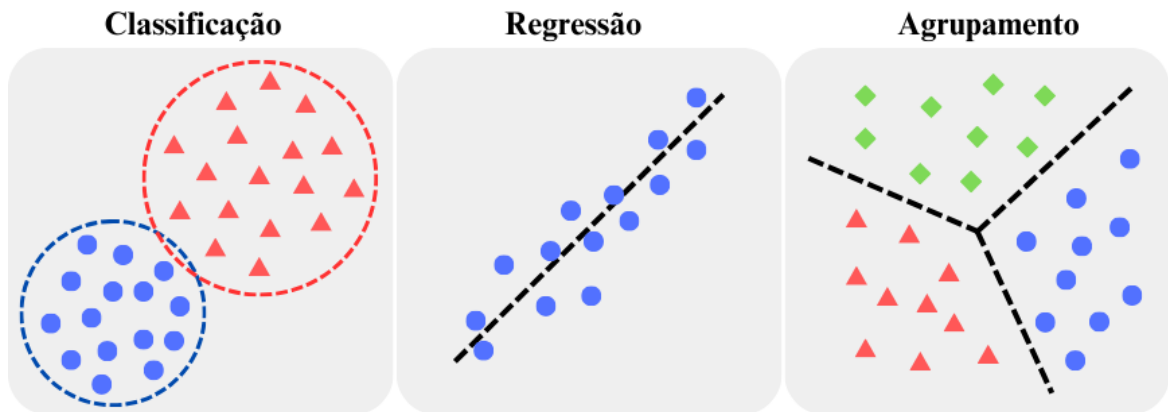


Figura 2 – Representação esquemática do comportamento dos dados nos principais métodos de aprendizado de máquina.

Dentre as técnicas de aprendizado de máquina apresentadas, no projeto em questão foram utilizados os métodos de Clusterização Hierárquica e as Análises de Componentes Principais (PCA).

O método de clusterização hierárquica é uma técnica de aprendizado não supervisionado amplamente utilizada em análise de dados, que consiste em agrupar dados em uma estrutura hierárquica na forma de árvore, conhecida por dendrograma. Esta técnica pode ser dividida em dois tipos: clusterização aglomerativa (*Bottom-Up*), em que inicia-se a análise com cada dado tratando sendo um *cluster* individual e, em seguida, combina de maneira iterativa os *clusters* mais próximos, até que todos os dados tratados sejam agrupados em um único *cluster*; e a clusterização divisiva, que diferentemente da primeira técnica, tem início com todos os dados fazendo parte de um único *cluster* e gradualmente os divide em *clusters* menores (Murtagh & Legendre, 2014). O método *Bottom-Up* tem como base a construção de uma matriz de distância, em que as distâncias entre os dados analisados são calculadas. É válido ressaltar, que a distância entre os *clusters* pode ser dimensionada de várias maneiras, como por exemplo, utilizando a distância *Euclidiana*, de *Manhattan*, *Chebyshev*, entre outras, como podemos ver esquematicamente, por meio da **Figura 3**, como estas distâncias são representadas. De tal forma que a estrutura final dos *clusters* pode ser influenciada de acordo com o método escolhido para o cálculo.

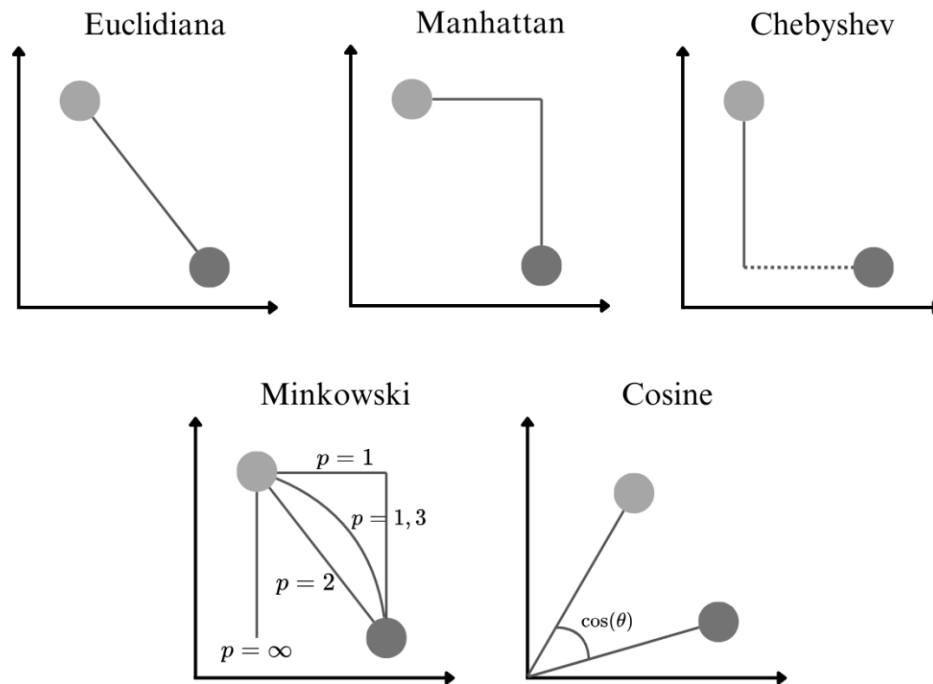


Figura 3 – Representação esquemática das distâncias.

Após essa primeira etapa, os *clusters* são então criados e a matriz de distância gerada é atualizada após cada combinação de *clusters*. Para recalcular a matriz, temos quatro métodos conhecidos: ligação simples, que se trata da distância entre os pontos mais próximos de cada *cluster*; ligação completa, sendo a distância entre os pontos mais distantes de cada *cluster*; ligação média, a média das distâncias entre todos os pontos de dois *clusters* e o método *Ward*, utilizado neste projeto, que tem como objetivo minimizar a variância dentro dos *clusters* (Jain, 2010). Todas essas etapas se repetem até que todos os dados analisados sejam agrupados. Por fim, há a criação de dendrogramas, que facilitam a visualização dos *clusters* gerados, ajudando assim, na identificação da estrutura hierárquica dos dados. Como desvantagens desse método, temos a dificuldade computacional em lidar com grandes conjuntos de dados e a sensibilidade a *outliers* (dados que desviam significativamente de um padrão geral em um conjunto de dados), que podem afetar a formação dos *clusters* (Hastie et al., 2009). A clusterização hierárquica foi um dos métodos escolhidos para as análises feitas neste projeto, pois se trata de uma técnica amplamente usada para análises de dados biológicos (Everitt et al., 2011). Para o dimensionamento dos *clusters* na análise realizada, foram empregados diferentes métodos de cálculo de distâncias, incluindo as métricas *Euclidiana*, *Máxima* (ou *Chebyshev*), *Manhattan*, *Canberra* e *Minkowski*. Adotando x_i e y_i como as coordenadas dos pontos avaliados, essas métricas podem ser então definidas pelas expressões matemáticas apresentadas a seguir:

Distância *Euclidiana*:

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (1)$$

Distância *Máxima* ou *Chebyshev*:

$$d(x, y) = \max_{i=1, \dots, n} |x_i - y_i| \quad (2)$$

Distância de *Manhattan*:

$$d(x, y) = \sum_{i=1}^n |x_i - y_i| \quad (3)$$

Distância de *Canberra*:

$$d(x, y) = \sum_{i=1}^n \frac{|x_i - y_i|}{|x_i| + |y_i|} \quad (4)$$

Distância de *Minkowski*:

$$d(x, y) = (\sum_{i=1}^n |x_i - y_i|^p)^{\frac{1}{p}} \quad (5)$$

A distância de *Minkowski* (ou distância generalizada, com *p-norm*) é determinada pelo parâmetro p , que define o tipo de métrica utilizada. Quando $p = 2$, a distância corresponde à distância *Euclidiana*, enquanto para $p = 1$, ela se torna equivalente à distância *Manhattan*.

A Análise de Componentes Principais (PCA) configura-se como uma técnica estatística amplamente empregada para a redução da dimensionalidade de conjuntos de dados em diversas áreas do conhecimento, como aprendizado de máquina, bioinformática e estatística multivariada. O objetivo primordial dessa abordagem consiste em reduzir o número de variáveis do conjunto de dados, preservando a maior proporção possível da variabilidade inerente aos dados originais (Jolliffe, 2002; Jolliffe & Cadima, 2016). A PCA realiza a transformação de variáveis possivelmente correlacionadas em um novo sistema de coordenadas, denominado componentes principais, que são ortogonais e, portanto, não correlacionados entre si. Essa

transformação promove a identificação de padrões, simplifica a visualização dos dados e facilita a interpretação dos mesmos. Os componentes principais são ordenados de maneira que o primeiro componente principal explica a maior parte da variância total dos dados, o segundo componente principal retém a segunda maior parcela da variância, e assim sucessivamente. De modo que a PCA orienta os eixos principais de acordo com as direções que apresentam a maior variação nos dados, como podemos verificar pelo meio da **Figura 4**. A implementação da PCA geralmente inicia-se com a etapa de padronização dos dados, de modo a assegurar que estes apresentem média igual a zero e variância unitária. Esse procedimento é essencial para garantir a comparabilidade entre variáveis, especialmente quando estas se encontram em escalas distintas (Abdi & Williams, 2010). Em seguida, procede-se ao cálculo da matriz de covariância ou de correlação, conforme o objetivo da análise, para avaliar as relações entre as variáveis e a forma como estas variam conjuntamente. Com base nessa matriz, são extraídos os autovalores e autovetores. Os autovalores indicam a proporção da variância explicada por cada componente principal, enquanto os autovetores definem as direções desses componentes. A seleção dos componentes principais é realizada considerando-se os autovalores e a proporção acumulada da variância total que se deseja preservar. Posteriormente, os dados originais são projetados no novo sistema de coordenadas formado pelos componentes principais, gerando um conjunto de variáveis reduzido que retém a maior parte da variabilidade, eliminando redundâncias e ruídos (Jolliffe, 2002; Lever et al., 2017). Embora apresente diversas vantagens, como a simplificação estrutural dos dados e a redução de ruídos, a PCA possui algumas limitações. Entre estas, destacam-se a dificuldade de interpretação dos componentes principais, uma vez que estes consistem em combinações lineares das variáveis originais, frequentemente sem uma correspondência semântica direta. Além disso, a técnica assume linearidade nas relações entre as variáveis, o que pode não ser adequado para determinados tipos de dados (Shlens, 2014).

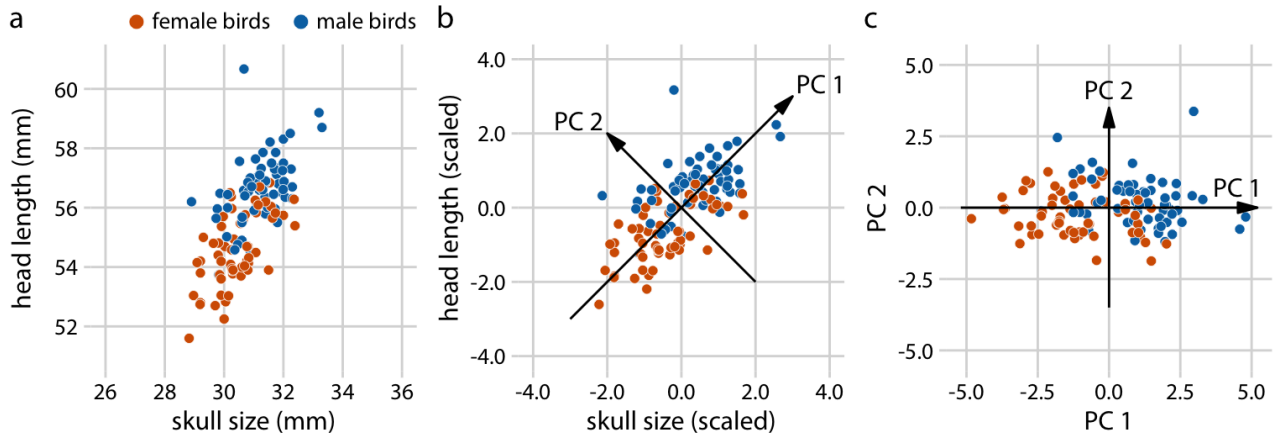


Figura 4 – Representação esquemática do comportamento dos dados quando aplicado o método da PCA. FONTE: (Claus O. Wilke, 2019)

A versatilidade da PCA é evidenciada por sua aplicação em uma ampla gama de áreas, incluindo a identificação de padrões genômicos em bioinformática, redução de ruído em sinais e imagens, e como etapa de pré-processamento em algoritmos de aprendizado de máquina. Apesar do avanço de outras técnicas de aprendizado não supervisionado, a PCA mantém-se como uma ferramenta essencial para a exploração, simplificação e compreensão de conjuntos de dados complexos (Izenman, 2008; Lever et al., 2017).

A avaliação do desempenho de modelos de classificação é uma etapa crucial no desenvolvimento de soluções baseadas em aprendizado de máquina. Essa avaliação não apenas determina a qualidade do modelo, mas também orienta decisões quanto à sua aplicabilidade prática, confiabilidade e robustez. Diversas métricas podem ser utilizadas para esse fim, cada uma com características específicas que as tornam mais ou menos adequadas a depender do contexto do problema. Métricas como acurácia, precisão, sensibilidade, especificidade, F1-score e área sob a curva ROC (ROC AUC) são particularmente relevantes para problemas de classificação, principalmente quando há desbalanceamento entre as classes (Géron, 2019).

No contexto deste trabalho, a escolha das métricas foi orientada pela necessidade de uma avaliação abrangente e robusta, além da necessidade de avaliar não apenas a proporção total de acertos, mas também a capacidade do modelo em identificar corretamente as classes positivas e negativas, além do equilíbrio entre falsos positivos e falsos negativos. Tal abordagem é particularmente relevante em contextos nos quais os erros têm diferentes implicações práticas, como nas áreas médica, financeira ou de segurança (Fawcett, 2006; Han

& Pei, 2012). Dessa forma, foram escolhidas as seguintes métricas para a realização do estudo: acurácia, precisão, sensibilidade, especificidade, F1-score e área sob a curva ROC (ROC AUC).

A acurácia (*accuracy*) é uma das métricas mais utilizadas em tarefas de classificação e representa a proporção de predições corretas – tanto positivas quanto negativas – em relação ao total de predições realizadas. Sua fórmula é expressa por:

$$Acurácia = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

Onde TP refere-se aos verdadeiros positivos, TN aos verdadeiros negativos, FP aos falsos positivos e FN aos falsos negativos. Apesar de ser amplamente utilizada, essa métrica pode induzir a interpretações equivocadas em cenários com conjuntos desbalanceados, pois tende a superestimar o desempenho quando a classe majoritária domina a base de dados (Powers, 2020).

A precisão (*precision*), por sua vez, mede a proporção de exemplos classificados como positivos que, de fato, pertencem à classe positiva.

$$Precisão = \frac{TP}{TP+FP} \quad (7)$$

Essa métrica é particularmente útil em situações nas quais os falsos positivos devem ser minimizados, como em sistemas de detecção de fraudes ou diagnósticos médicos, nos quais um alarme falso pode gerar consequências indesejadas (Han & Pei, 2012).

A sensibilidade (*sensitivity* ou *recall*) avalia a capacidade do modelo de identificar corretamente os exemplos positivos. Em outras palavras, trata-se da razão entre os verdadeiros positivos e o total de exemplos efetivamente positivos.

$$Sensibilidade = \frac{TP}{TP+FN} \quad (8)$$

Essa métrica é essencial quando o foco está em minimizar a quantidade de falsos negativos, como em diagnósticos de doenças graves, nos quais deixar de identificar um caso positivo pode ter sérias implicações (Géron, 2019).

Por sua vez, a especificidade (*specificity*) mede a proporção de exemplos negativos corretamente identificados, ou seja, a capacidade do modelo de evitar falsos positivos.

$$Especificidade = \frac{TN}{TN+FP} \quad (9)$$

Embora menos frequentemente utilizada de forma única, a especificidade é importante para avaliar a capacidade do modelo em discriminar corretamente entre as duas classes (Fawcett, 2006).

A medida F, ou F1-score, é a média harmônica entre precisão e sensibilidade, sendo particularmente indicada quando há necessidade de equilibrar os erros falso-positivo e falso-negativo.

$$F1score = 2 \cdot \frac{Precisão \cdot Sensibilidade}{Precisão + Sensibilidade} \quad (10)$$

Se trata de uma métrica amplamente utilizada em problemas de classificação com avaliados conjuntos de dados desbalanceados (Powers, 2020).

Fechando o conjunto de métricas selecionadas, temos a área sob a curva ROC (ROC AUC), que constitui uma métrica baseada na curva ROC (*Receiver Operating Characteristic*), em que relaciona a taxa de verdadeiros positivos (sensibilidade) com a taxa de falsos positivos (1 - especificidade). A AUC quantifica a capacidade do modelo de distinguir entre as classes e varia de 0 a 1, sendo que valores próximos de 1 indicam excelente desempenho. Sendo uma métrica fortemente usada quando se trata de conjunto de dados desbalanceados (Fawcett, 2006).

A utilização conjunta dessas métricas permite uma avaliação mais completa e confiável dos modelos desenvolvidos, conferindo maior transparência e robustez à análise dos resultados obtidos.

1.5 Justificativa

O entendimento da biologia dos tecidos tornou-se uma necessidade crescente na atualidade, especialmente diante das transformações demográficas em curso. Nas próximas décadas, o Brasil deverá enfrentar um aumento expressivo na sua população idosa, o que acarretará um crescimento significativo no custo social da saúde. Esse cenário demanda a busca por

estratégias que garantam qualidade de vida com custos e impactos sociais mais sustentáveis. Nesse contexto, um dos principais desafios reside na regeneração funcional de tecidos mineralizados, cuja perda compromete diretamente a mobilidade, a autonomia e a saúde global dos indivíduos.

É nesse panorama que a medicina regenerativa tem se destacado, oferecendo alternativas terapêuticas avançadas e multifatoriais, com o objetivo de restaurar, de forma funcional, tecidos lesados ou perdidos. As abordagens atuais visam recriar um microambiente favorável ao reparo tecidual, fornecendo elementos celulares essenciais, fatores de crescimento e diferenciação, além de estruturas supramoleculares (arcabouços) capazes de guiar a organização espacial dos tecidos regenerados. Essas estratégias caracterizam o escopo da engenharia de tecidos, que, atualmente, se expande em direção à biofabricação e ao desenvolvimento de ferramentas de bioanálise para acelerar a inovação em novos materiais biomédicos.

Paralelamente, cresce mundialmente a necessidade de métodos alternativos ao uso de animais em pesquisa. Essa tendência é reforçada por iniciativas como o programa dos 3Rs – *Replacement, Reduction e Refinement* (Substituição, Redução e Refinamento) – concebido em 1959 por William Russell e Rex Burch (Tannenbaum & Bennett, 2015). O programa propõe a diminuição do uso de animais em experimentação, bem como o aperfeiçoamento dos métodos existentes, com vistas à redução do sofrimento animal e ao desenvolvimento de alternativas *in vitro* e *in silico* mais éticas e eficazes.

Diante da crescente demanda por tecnologias biomédicas e da necessidade de fortalecer a capacidade nacional de inovação, este projeto propõe o desenvolvimento de uma plataforma de bioanálise com foco na clusterização de dados moleculares, visando contribuir com o avanço da bioengenharia e da medicina regenerativa no Brasil. Essa iniciativa busca, ainda, viabilizar o desenvolvimento de novos biomateriais por meio da integração entre biotecnologia, bioinformática e análise computacional.

2 OBJETIVOS

O objetivo geral deste projeto foi desenvolver um *framework*¹ de bioinformática que utilize ferramentas biotecnológicas e conceitos de bioengenharia para comparar a resposta celular, por meio de assinaturas moleculares do transcriptoma, em células aderidas a diferentes superfícies de materiais biomédicos. Os objetivos específicos foram:

- a) Avaliação de diferentes *scores* de enriquecimento em resposta aos diferentes materiais;
- b) Avaliação de distâncias para clusterização hierárquica no contexto proposto;
- c) Geração de dados sintéticos para validação do *framework* desenvolvido;
- d) Construção de *dashboard/app* utilizando *Shiny*;

Meta: *Estabelecer um banco de dados biológicos capaz de prever a resposta biológica de novos biomateriais, contribuindo para a triagem e desenvolvimento de materiais com alta biocompatibilidade.*

Hipótese: *"Assinaturas moleculares do transcriptoma podem revelar diferenças específicas na resposta celular a diferentes superfícies de biomateriais, permitindo a criação de um framework de bioinformática capaz de comparar e prever a biocompatibilidade e eficiência de adesão celular, favorecendo o desenvolvimento de biomateriais com propriedades otimizadas para aplicações biomédicas."*

¹ Defini-se por *framework* como um conjunto de ferramentas/programas computacionais (softwares) que podem ser reutilizados para análises específicas. Exemplos de *frameworks* são os pacotes tidyverse, data.table e tidymodels no R e pacotes pandas, numpy, scikit-learn no Python. Dentre outros.

CAPÍTULO II: CLUSTERIZAÇÃO, DADOS SINTÉTICOS E SHINY

1 RESUMO DO CAPÍTULO

O capítulo delinea os materiais e métodos empregados nas análises de dados, partindo dos dados iniciais até a aplicação das técnicas de *machine learning*, além de abordar o método utilizado para geração de dados sintéticos por meio da biblioteca *synthpop*. Os resultados gerados iniciam-se com estudos de distribuição que visam compreender a variabilidade e os padrões nos dados analisados. São descritos os procedimentos para o cálculo de *scores*, os quais quantificam características específicas das moléculas, e que são posteriormente usados para as análises subsequentes. A seguir, aborda-se a construção de mapas de calor, que visualizam a intensidade e a correlação entre variáveis, proporcionando uma interpretação mais clara dos dados. O capítulo também explora a análise de redução de dimensionalidade, com ênfase na técnica de Análise de Componentes Principais (PCA), permitindo assim simplificar os conjuntos de dados complexos enquanto preservam suas estruturas. Além disso, aborda-se a criação de mapas de calor a partir dos dados sintéticos gerados, permitindo assim visualizar as interações e padrões resultantes com as assinaturas moleculares. Essa abordagem evidencia a relevância dos dados sintéticos nos processos de teste e validação de modelos de *machine learning*, cuja performance é avaliada por meio de métricas consolidadas em tabelas comparativas, permitindo uma análise detalhada de cada *score* obtido. Por fim, é apresentada a versão inicial de um aplicativo *Shiny*, que oferece uma interface interativa para a exploração e visualização dos dados, tornando as análises acessíveis a usuários e promovendo uma compreensão mais aprofundada da correlação dos biomateriais com as assinaturas moleculares.

2 MATERIAIS E MÉTODOS

2.1 Dados utilizado

Neste trabalho, foram utilizados dados públicos provenientes do experimento disponível no repositório *ArrayExpress* (<https://www.ebi.ac.uk/arrayexpress/>), sob código de acesso E-MTAB-7219. A escolha deste banco de dados se justifica por ser, até o momento, o que reúne o maior número de diferentes biomateriais avaliados em um mesmo experimento, com dados amplamente disponibilizados. Essa característica é particularmente relevante para o presente trabalho, que buscou realizar análises comparativas da resposta celular a múltiplos substratos biomédicos. O experimento original foi conduzido utilizando células da linhagem MG-63 (fibroblasto de osteossarcoma humano), cultivadas em meio de cultura alfa-MEM (alpha-MEM, Life Technologies), suplementado com 10% soro fetal bovino, 0,2 mM ácido ascórbico (Asap, Life Technologies), 2 mM L-glutamina (Life Technologies), 100 U/ml penicilina e 100 ug/ml estreptomicina. O meio de cultura foi renovado duas vezes por semana, e todas as etapas experimentais foram realizadas sob condições controladas de 37°C e 5% de CO₂. As células foram semeadas por 150 µl de material previamente umedecido com meio, na densidade 2×10^6 células/ml. Após um período inicial de 4 horas (tempo de fixação), um volume adicional de meio foi adicionado ao poço. No total, 21 biomateriais distintos foram testados, conforme detalhado na **Tabela 1**. Após 48 horas de incubação, as células foram coletadas seguindo o protocolo do kit *Nucleospin* RNA (Macherey-Nagel) para isolamento de RNA. A partir de 275 ng de RNA, foi realizada a síntese de cRNA utilizando o kit de amplificação de RNA Illumina TotalPrep (Ambion), conforme protocolo do fabricante. A qualidade do RNA e do cRNA foi avaliada com o uso do *Bioanalyser* 2.100 (Agilent). *Microarrays* foram realizados usando *Beadchips* de expressão Illumina HT-12 v4. Resumidamente, 750 ng de cRNA foram hibridizados no *array* durante a noite, após o que o *array* foi lavado e bloqueado. Então, com adição de estreptavidina Cy-3, um sinal fluorescente foi desenvolvido. Os *arrays* foram escaneados em um leitor Illumina Beadarray e os valores brutos de intensidade foram corrigidos em segundo plano no BeadStudio (Illumina). Posteriormente, os dados normalizados foram transformados por estabilização de variância para adequação às análises subsequentes.

Tabela 1 - Relação de biomateriais contidos no array E-MTAB-7219, com a descrição do número de replicatas utilizadas para cada biomaterial (n). As colunas TCP, Titanium, PLA e Hydroxyapatite contém informações sobre o biomaterial conter, ou não, este composto em sua composição que serão utilizadas para a construção dos modelos de classificação.

Biomaterial	Code	n	TCP	Titanium	PLA	Hydroxyapatite	Classificação
Biphasic Calcium Phosphate	BCP1150_HAP60TCP40	3	Yes	No	No	Yes	Biocompatíveis

Biphasic Calcium Phosphate	BCP1150_HAP70TCP30	3	Yes	No	No	Yes	Biocompatíveis
Biphasic Calcium Phosphate	BCP1150_HAP80TCP20	6	Yes	No	No	Yes	Biocompatíveis
Biphasic Calcium Phosphate	BCP1300_HAP80TCP20	3	Yes	No	No	Yes	Biocompatíveis
Decellularized Bone Extracellular Matrix	DBEM	3	No	No	No	No	Biocompatíveis
Demineralized Bone Matrices	DBM	3	No	No	No	No	Biocompatíveis
Hydroxyapatite	HAP1100	3	No	No	No	Yes	Biocompatíveis
Hydroxyapatite	HAP1150	3	No	No	No	Yes	Biocompatíveis
Hydroxyapatite	HAP1250	3	No	No	No	Yes	Biocompatíveis
Polylactic Acid	PLA	3	No	No	Yes	No	Biocompatíveis
Polylactic Acid HydroxyapatiteÂ Composite	PLAHA	3	No	No	Yes	Yes	Biocompatíveis
Polylactic Acid HydroxyapatiteÂ Composite	PLAHA0.125MNaOH	3	No	No	Yes	Yes	Biocompatíveis
Polylactic Acid HydroxyapatiteÂ Composite	PLAHA0.25MNaOH	3	No	No	Yes	Yes	Biocompatíveis
Polylactic Acid Hydroxyapatite Composite	PLAHA0.5MNaOH	3	No	No	Yes	Yes	Biocompatíveis
Polylactic Acid Hydroxyapatite Composite	PLAHA0MNaOH	3	No	No	Yes	Yes	Biocompatíveis
Polylactic Acid Hydroxyapatite Composite	PLAHA1MNaOH	3	No	No	Yes	Yes	Biocompatíveis
Polylactic Acid Hydroxyapatite Composite With Octacalcium Phosphate Coating	PLAOCTA	3	No	No	Yes	Yes	Biocompatíveis
Polystyrene	POLY	3	No	No	No	No	Biocompatíveis
Porous Titanium	TI	3	No	Yes	No	No	Biocompatíveis
Porous Titanium With Octacalcium Phosphate Coating	TIOCTA	3	No	Yes	No	No	Biocompatíveis
Tricalcium Phosphate	TCP1050	3	Yes	No	No	No	Biocompatíveis
Tricalcium Phosphate	TCP1100	3	Yes	No	No	No	Biocompatíveis

2.2 Análise de dados

Os dados foram analisados utilizando a linguagem de programação R na interface de desenvolvimento *RStudio* (RStudio Team, 2022). Os principais pacotes utilizados foram: *tidyverse* (Wickham et al., 2019), que é uma coleção de pacotes para ciência de dados; pacotes do *bioconductor* para análises de bioinformática, tais como *AnnotationDBI* (Pagès et al., 2021), *msigdb* (Igor Dolgalev, 2022), *fgsea* (Korotkevich et al., 2021), *GSVA* (Hänzelmann et al., 2013a), *limma* (Ritchie et al., 2015a) e *Biobase* (Huber et al., 2015); pacote *Shiny* (Chang et al., 2022) para criação de *dashboards* e *app* com o *framework* proposto.

2.3 Análise de expressão diferencial e enriquecimento

A expressão gênica diferencial foi computada utilizando o pacote *limma* (Ritchie et al., 2015b). Consideraram-se genes diferencialmente expressos aqueles com $\log_2(FC) > 0,50$ e p-valor ajustado (FDR) $< 0,05$. A partir desse conjunto de genes diferencialmente expressos, foram calculadas as pontuações de enriquecimento para diferentes coleções de assinaturas gênicas. Todas as coleções disponíveis no *MSigDB*, foram utilizadas para avaliar os *scores* de enriquecimento nos diferentes conjuntos de dados (Liberzon et al., 2011, 2015; Subramanian et al., 2005a). A análise de enriquecimento de conjuntos gênicos (*GSEA*) foi calculada utilizando o pacote *fgsea* (Korotkevich et al., 2021), utilizando os genes ranqueados com base nos resultados da expressão diferencial. Para essa análise, não foram impostas restrições quanto ao número mínimo ou máximo de genes por assinatura. Além disso, foram aplicadas abordagens complementares: a análise de variação de conjuntos gênicos (*GSVA*) e o *GSEA* de amostra única (*ssGSEA*) – este último, sendo uma extensão da abordagem *GSEA* -, ambas conduzidas com o pacote *GSVA* (Hänzelmann et al., 2013). A análise *GSVA* foi realizada em duas etapas: primeiramente incluindo o grupo controle (Poliestireno), permitindo comparações diretas com os grupos experimentais; posteriormente, excluindo o controle, com o objetivo de obter uma visão mais ampla das assinaturas gênicas em diferentes contextos experimentais. De forma semelhante, o *ssGSEA* foi realizado com a inclusão do controle, enquanto o *GSEA* foi aplicado apenas aos grupos experimentais, sem o grupo Poliestireno.

2.4 Análise de clusterização hierárquica

As matrizes de distâncias foram calculadas utilizando-se a função *dist* dos pacotes *stats*. Sendo avaliadas as seguintes distâncias: "*euclidiana*", "*manhattan*", "*canberra*", "*maximum*", "*binary*" e "*minkowski*". Clusterização hierárquica foi realizada com a função *hclust* do pacote *stats*, avaliando pelo método de ligação "*ward.D2*". Para visualização dos dados foram construídos mapas de calor com o pacote *ComplexHeatmap* (Gu et al., 2016) e dendrogramas com o pacote *ggdendro* (de Vries & Ripley, 2022).

2.5 Análises de redução de dimensionalidade

A análise de redução de dimensionalidade foi realizada a partir dos diferentes *scores* obtidos (*GSVA*, *GSEA* e *ssGSEA*) para as assinaturas moleculares testadas através do pacote *stats* (R Core Team, 2021) para análise de componentes principais (PCA).

2.6 Geração de dados sintéticos

O pacote *Synthpop* do R foi utilizado para geração de dados sintéticos através do método *CART* (árvore de classificação e regressão) (Nowok et al., 2016), tendo como entrada os dados dos *scores* gerados de *GSVA*, *GSEA* e *ssGSEA* para as assinaturas moleculares do conjunto *Hallmarks* (Liberzon et al., 2015).

2.7 Modelos de predição

Os modelos de predição foram baseados em métodos de classificação para as condições apresentadas na **Tabela 1**. Foi realizado um benchmarking de diferentes algoritmos como: Regressão Logística, Árvore de Decisão, Random Forest, XGBoost, SVM e Naive-Bayes. Foram escolhidos esses algoritmos por serem métodos de classificação consolidados presentes no framework *tidymodels* (Kuhn & Wickham, 2020), possuindo assim, diversos conteúdos publicados sobre estratégias para otimização (tunagem) de seus hiperparâmetros. As métricas de avaliação dos modelos serão acurácia, precisão, sensibilidade, especificidade, F1-score, área sob a curva ROC. Os pacotes utilizados serão do framework *tidymodels* (Kuhn & Wickham, 2020), *parsnip* (Kuhn & Vaughan, 2022) será utilizado como interface para os diferentes pacotes dos modelos de machine learning que serão utilizados, *yardstick* (Kuhn, 2022) para o cálculo das métricas e finalmente *tune* (Kuhn, 2022) para a tunagem dos modelos.

3 RESULTADOS E DISCUSSÃO

Tendo conhecimento que as lesões ósseas representam um dos problemas mais frequentes na área médico-odontológica, com uma tendência crescente ao longo dos anos, impulsionada pela maior incidência de condições músculo-esqueléticas e crônicas, além do aumento da população idosa, que enfrenta riscos elevados de osteoartrose, osteoporose e outras doenças relacionadas. O projeto *OsteoCLUST* foi criado com a finalidade de integrar ferramentas biotecnológicas e princípios de bioengenharia, visando desenvolver um *framework* de bioinformática que possibilite análises da resposta celular em células aderidas a diversas superfícies de materiais biomédicos. Isso resultará em uma base de dados que permitirá a comparação com novos biomateriais avaliados pelos usuários do aplicativo. Assim, o tratamento das lesões ósseas poderá ser realizado de forma personalizada, otimizada e com custo reduzido. Dessa forma, para desenvolvimento do aplicativo, era necessário calcular e avaliar *scores* de enriquecimento, encontrar maneiras eficazes de visualizar esses dados, definir os melhores modelos de aprendizado de máquina para classificação, determinar um método adequado para geração dos dados sintéticos e encontrar um melhor modelo para predição de resultados. Para que então seja compilada todas essas informações e verificada a viabilidade da proposta final do aplicativo.

Ao se realizar análises de experimentos de sequenciamento de alto rendimento (*high-throughput*), comumente são feitas análises de vias biológicas na busca de contextualização biológicas dos resultados obtidos na expressão gênica/proteica (Khatri et al., 2012). Dentre as ferramentas utilizadas para tal destacam-se o *GSVA* (*Gene Set Variation Analysis*), *GSEA* (*Gene Set Enrichment Analysis*) e *ssGSEA* (*Single-sample Gene Set Enrichment Analysis*). O *GSVA* requer valores brutos de expressão como entrada, não sendo necessária a comparação entre dois grupos. O cálculo é realizado utilizando distribuições normais, e o método não se baseia em diferenças de regulação positiva ou negativa, nem atribui diferenças estatísticas ao resultado (Hänzelmann et al., 2013a). Por outro lado, o *GSEA* utiliza dados estatísticos, como o *fold change*, e utiliza todos os genes sem a necessidade de um limiar de corte, analisando o conjunto por grupos de amostras. Os cálculos do *GSEA* envolvem permutações, o que aumenta o custo computacional (Subramanian et al., 2005b). O *ssGSEA*, por sua vez, é uma técnica que permite a avaliação da expressão gênica em amostras individuais, avaliando assim, de uma maneira independente. Dessa forma, a pontuação de enriquecimento é definida para conjuntos de genes específicos, possibilitando identificar padrões em nível de amostra (Barbie et al., 2009).

3.1 Genes diferencialmente expressos

A expressão gênica diferencial foi calculada analisando-se aproximadamente 32200 genes para cada um dos 21 biomateriais do conjunto de dados, sendo o Poliestireno (*POLY*) considerado como o grupo controle para a análise. Pela **Tabela 2** pode-se observar o valor total de genes diferencialmente expressos para cada biomaterial em questão.

Tabela 2 - Relação dos números de genes Up e Down regulados e o total de genes diferencialmente expressos para cada biomaterial contidos no array E-MTAB-7219, com exceção do Poliestireno (*POLY*).

Biomaterial	Up regulados	Down regulados	Total diferencialmente expressos
BCP1150_HAP60TCP40	176	12	188
BCP1150_HAP70TCP30	181	9	190
BCP1150_HAP80TCP20	166	22	188
BCP1300_HAP80TCP20	207	23	230
DBEM	106	11	117
DBM	255	83	338
HAP1100	173	13	186
HAP1150	123	19	142
HAP1250	176	16	192
PLA	204	17	221
PLAHA	596	828	1424
PLAHA0.125MNaOH	1950	2515	4465
PLAHA0.25MNaOH	138	17	155
PLAHA0.5MNaOH	157	20	177
PLAHA0MNaOH	1958	2421	4379
PLAHA1MNaOH	154	29	183
PLAOCTA	159	13	172
TI	123	51	174
TIOCTA	49	19	68
TCP1050	155	20	175
TCP1100	167	11	178

3.2 Estudo de distribuição dos dados

A análise das distribuições de dados por meio de histogramas e *boxplots* é essencial na estatística, visto que, por meio desses estudos, é possível visualizar padrões e avaliar a forma das distribuições, podendo variar entre normais ou assimétricas. Com os histogramas temos uma facilidade com a identificação de características gerais dos dados, enquanto com os *boxplots* é possível obter características da estatística descritiva, como medianas e quartis, obter uma representação da dispersão dos dados, além de identificarem outliers presentes nos dados.

Essas análises gráficas iniciais são cruciais para entender a variabilidade dos dados, realizar comparações entre diferentes grupos, com o intuito de verificar possíveis diferenças significativas. Sendo formas visuais de apresentar resultados de uma maneira mais clara e, consequentemente, facilitando a interpretação do conjunto analisado.

Neste contexto, foram avaliadas as distribuições dos dados resultantes dos cálculos dos *scores* (*GSVA*, *GSEA* e *ssGSEA*) para as assinaturas moleculares presentes no conjunto gênico *Hallmarks*. A **Figura 5** ilustra essas distribuições por meio de histogramas, enquanto a **Figura 6** apresenta os *boxplots* correspondentes.

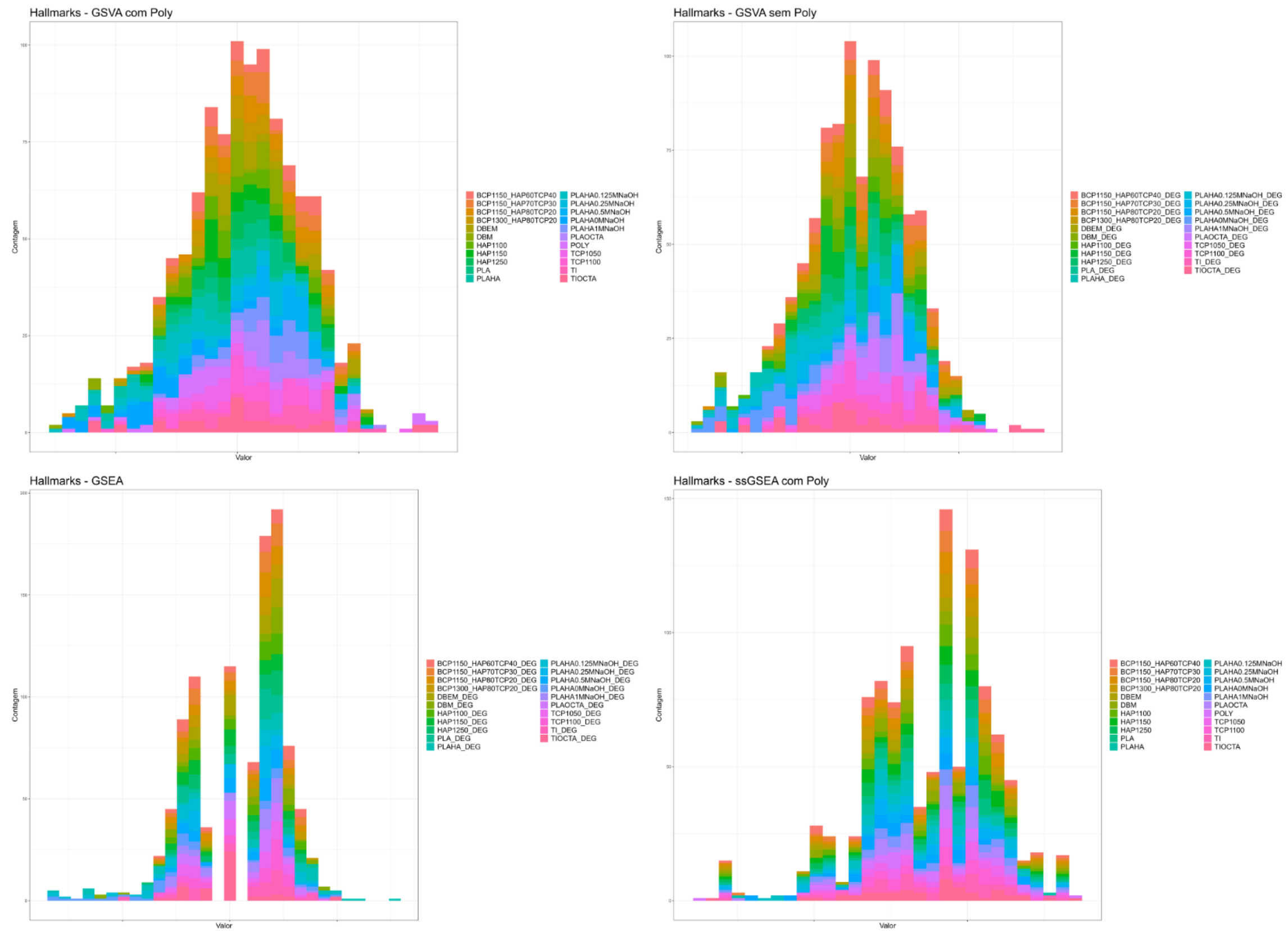


Figura 5 – Representação por meio de histogramas, das distribuições dos scores GSV (com e sem o grupo controle POLY), GSEA e ssGSEA, gerados para as assinaturas moleculares presentes no conjunto Hallmarks.

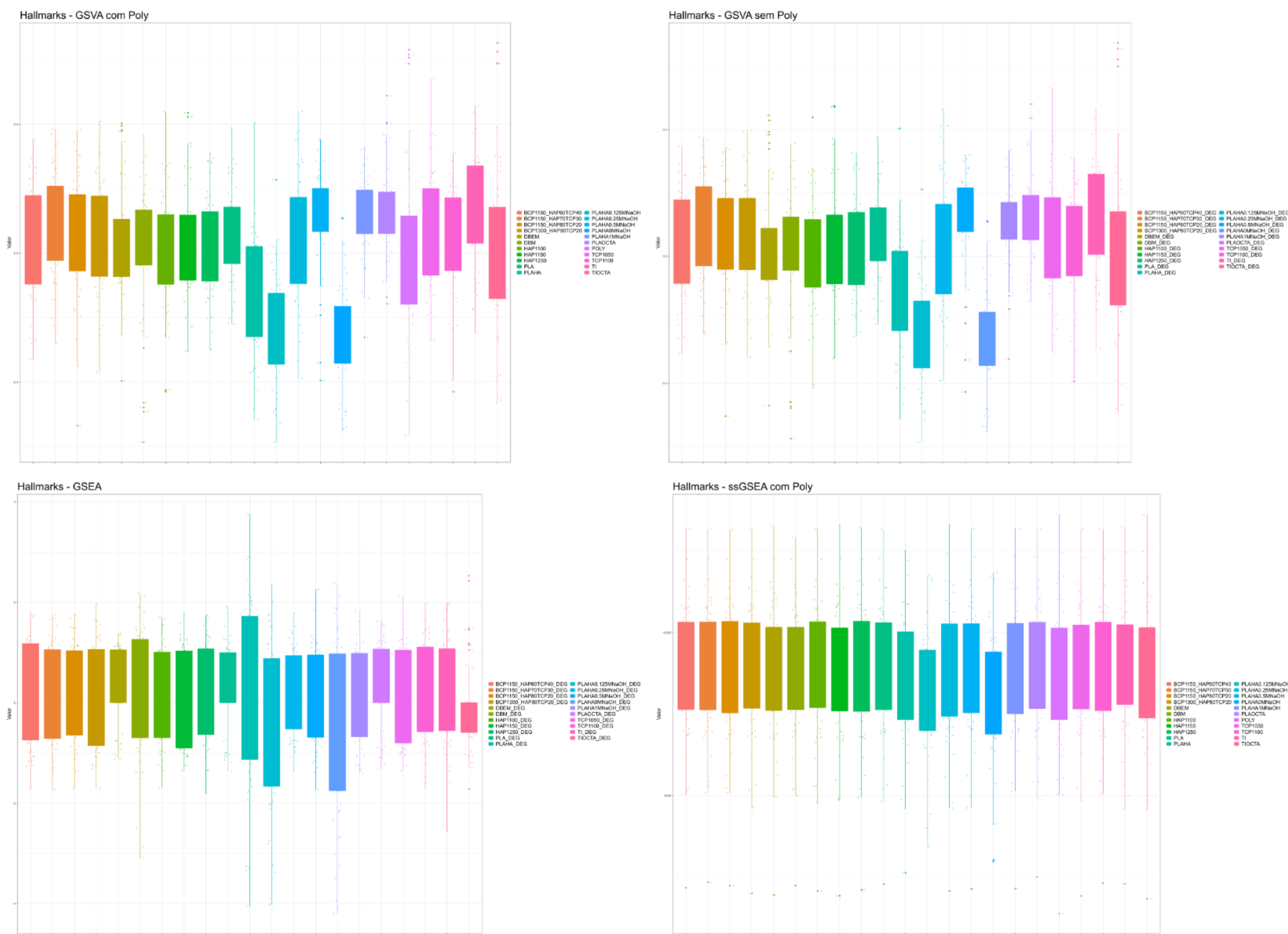


Figura 6 - Representação por meio de boxplot, das distribuições dos scores GSVA (com e sem o grupo controle POLY), GSEA e ssGSEA, gerados para as assinaturas moleculares presentes no conjunto Hallmarks.

A análise dos histogramas gerados revela que a distribuição dos dados relativos ao *score GSVa*, tanto na presença quanto na ausência de POLY, apresenta simetria em relação ao eixo central. Esta simetria sugere uma aproximação a uma distribuição normal. Adicionalmente, observa-se que ambas as distribuições compartilham padrões semelhantes, o que pode indicar uma homogeneidade na variabilidade dos dados. Essa conformidade com a normalidade é relevante, pois permite a aplicação de testes estatísticos que pressupõem essa característica, facilitando a interpretação e a comparação dos resultados.

A avaliação das distribuições dos scores *GSEA* e *ssGSEA* revela uma similaridade entre os conjuntos de dados, embora com uma tendência de deslocamento para a direita. Este deslocamento sugere uma distribuição assimétrica, o que pode implicar na presença de *outliers* ou na natureza dos dados analisados. Adicionalmente, no histograma correspondente ao *score GSEA*, observa-se a presença de lacunas, que podem ser atribuídas às limitações do *software* utilizado para o cálculo. Especificamente, valores que se aproximam de zero são interpretados pelo método como dados ausentes e, conseqüentemente, substituídos por zeros. Essa peculiaridade no processamento dos dados pode impactar a análise, resultando em uma subestimação dos scores em regiões centrais da distribuição. É fundamental considerar essa característica ao interpretar os resultados, uma vez que a ausência de dados em pontos críticos pode influenciar a validade das conclusões tiradas a partir do *score GSEA*.

A análise dos *boxplots* indica que, para o *score GSVa*, os biomateriais *PLAHA*, *PLAHA0.125MNaOH* e *PLAHA0MNaOH* apresentam deslocamento inferior em relação aos demais biomateriais analisados. Essa redução nos *scores* sugere uma menor eficácia ou atividade biológica associada a esses materiais considerando as assinaturas moleculares contidas na coleção *Hallmarks*, o que pode ser relevante para aplicações em engenharia de tecidos ou biomateriais.

No que diz respeito ao *score GSEA*, esses mesmos biomateriais revelam uma alta dispersão dos dados, caracterizada por intervalos interquartílicos maiores quando comparados aos outros biomateriais. Essa maior dispersão, juntamente com a presença de longas caudas nos *boxplots*, particularmente acentuadas no caso do *PLAHA*, indica uma variabilidade considerável nas respostas dos materiais, sugerindo que esses biomateriais podem estar associados a comportamentos não homogêneos em suas interações.

Por outro lado, a avaliação do método *ssGSEA* revela um padrão distinto, onde a distribuição dos dados se mostra mais homogênea. Os intervalos interquartílicos e o

comprimento das caudas são semelhantes entre todos os biomateriais, sugerindo uma distribuição mais uniforme dos *scores*. Essa uniformidade pode ser indicativa de um desempenho consistente dos biomateriais em relação ao *score ssGSEA*, o que poderia ser explorado em estudos futuros para investigar as características que contribuem para essa estabilidade.

Essas observações destacam a importância de utilizar múltiplos métodos de análise para uma compreensão mais abrangente do comportamento dos biomateriais, permitindo uma interpretação mais robusta dos resultados experimentais. A combinação de diferentes abordagens analíticas pode proporcionar *insights* valiosos sobre a relação entre as propriedades dos biomateriais e suas respectivas atividades biológicas.

3.3 Scores e mapa de calor das assinaturas moleculares

As assinaturas moleculares que foram utilizadas para a avaliação das pontuações de enriquecimento (*GSVA*, *GSEA* e *ssGSEA*) fazem parte dos conjuntos gênicos de humanos presentes no banco de dados *MSigDB* (versão 7.4). Pela **Tabela 3**, podemos visualizar os diferentes conjuntos com suas respectivas assinaturas moleculares.

Tabela 3 - Relação dos conjuntos de genes, disponíveis no *MSigDB*, com as quantidades de assinaturas moleculares presentes utilizadas para a avaliação da pontuação de enriquecimento.

Sigla	Conjunto de genes	Assinaturas presentes
C1	posicionais	278
C2	curados	6290
C3	reguladores	3731
C4	computacionais	858
C5	de ontologia	14998
C6	oncogênicos	189
C7	imunológicos	5219
C8	de assinaturas de células	671
-	Hallmarks	50
Total		32284

A representação escolhida para a visualização dos dados obtidos por meio das avaliações das pontuações de enriquecimento, foram os mapas de calor. Os mapas apresentados nesta seção são referentes ao conjunto gênico *Hallmarks*, visto que se trata de um conjunto construído com base nos outros 8 conjuntos presentes no banco de dados *MSigDB*. Os dados utilizados dos *scores GSVA*, *GSEA* e *ssGSEA*, foram previamente escalonados, utilizando-se o

método *Z-score*². Para essa primeira análise, a construção dos mapas de calor foi feita utilizando a distância "euclidiana" para os dendrogramas e o método de clusterização usado foi o "ward.D2". Assim, pode-se visualizar o resultado da análise para o *score GSEA* (com o grupo controle *POLY*) através da **Figura 7**, *GSEA* com a **Figura 8** e para o *ssGSEA* por meio da **Figura 9**.

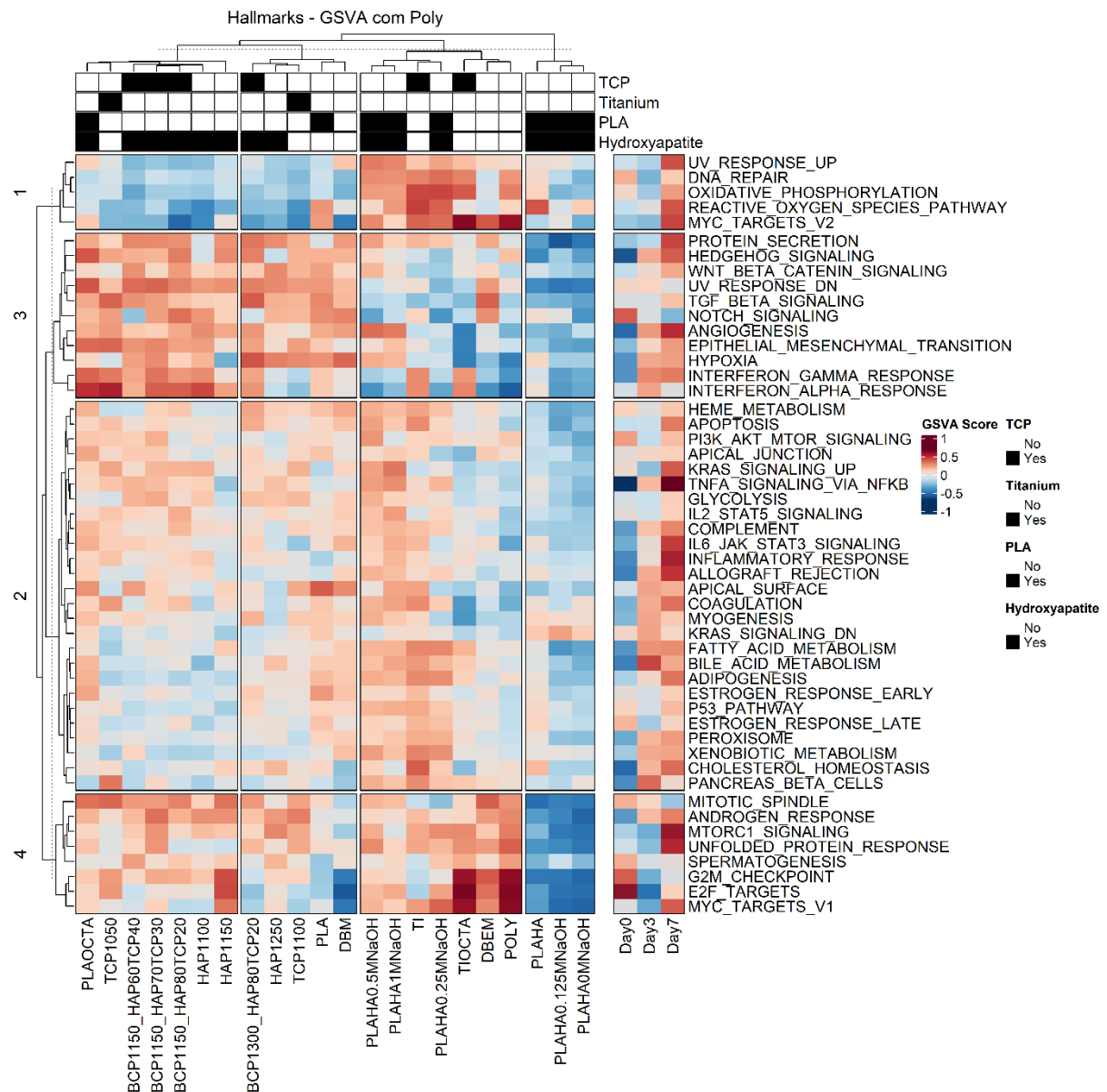


Figura 7 - Mapa de calor correlacionando os 21 biomateriais analisados com os scores *GSEA* (com o grupo *POLY*) gerados para cada uma das 50 assinaturas moleculares do conjunto gênico *Hallmarks*. Juntamente com um segundo mapa representando os scores de expressão gênica ao longo de 7 dias (*D0*, *D3* e *D7*), utilizados como controle biológico da progressão osteogênica para comparação com os perfis dos biomateriais.

²Calculado como $z = \frac{x - \mu}{\sigma}$, em que x é o valor bruto, μ a média e σ o desvio-padrão da amostra.

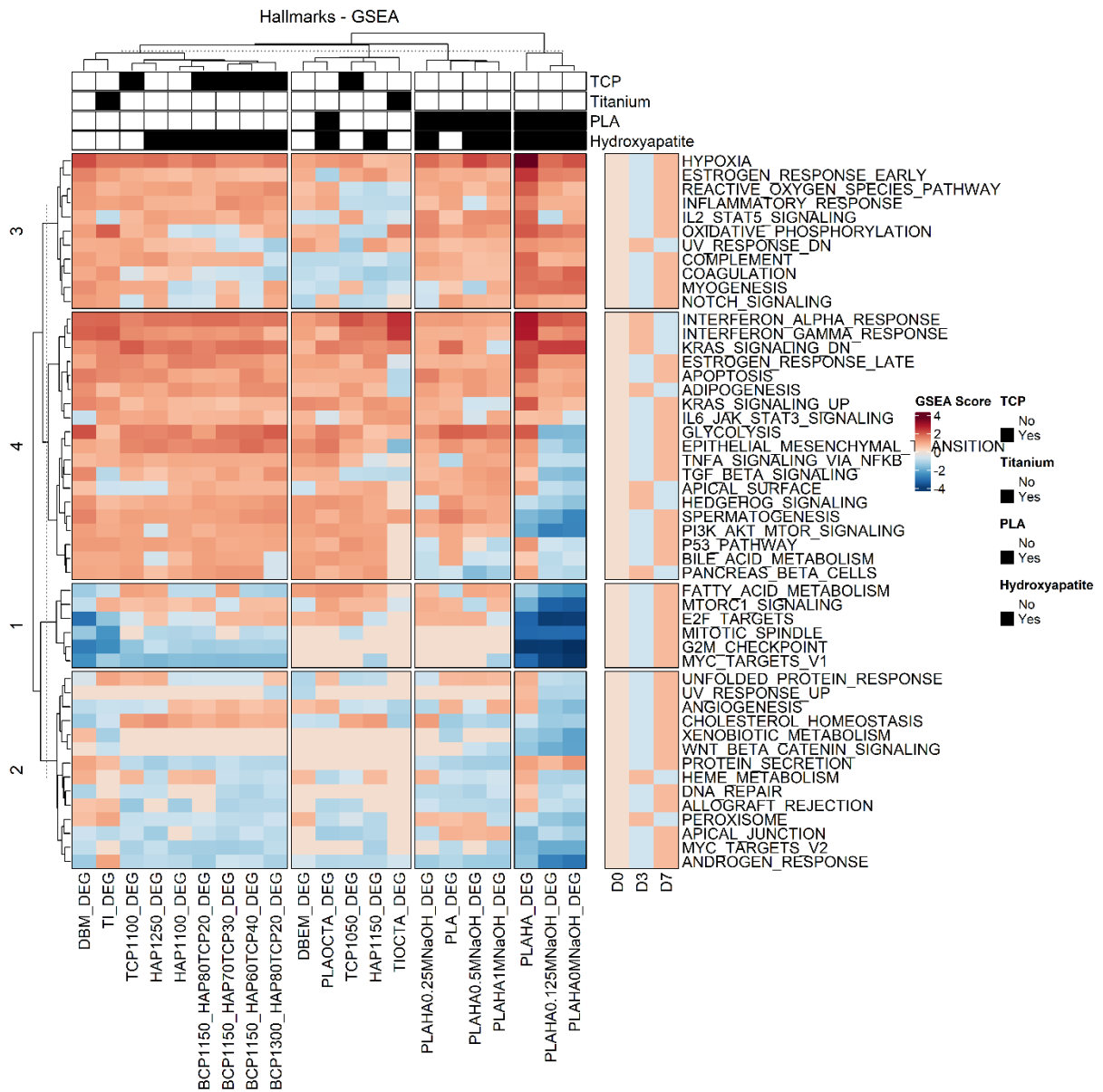


Figura 8 - Mapa de calor correlacionando os 21 biomateriais analisados com os scores GSEA gerados para cada uma das 50 assinaturas moleculares do conjunto gênico Hallmarks. Juntamente com um segundo mapa representando os scores de expressão gênica ao longo de 7 dias (D0, D3 e D7), utilizados como controle biológico da progressão osteogênica para comparação com os perfis dos biomateriais. Os valores de todas as assinaturas em D0 foram manualmente definidas como sendo zero.

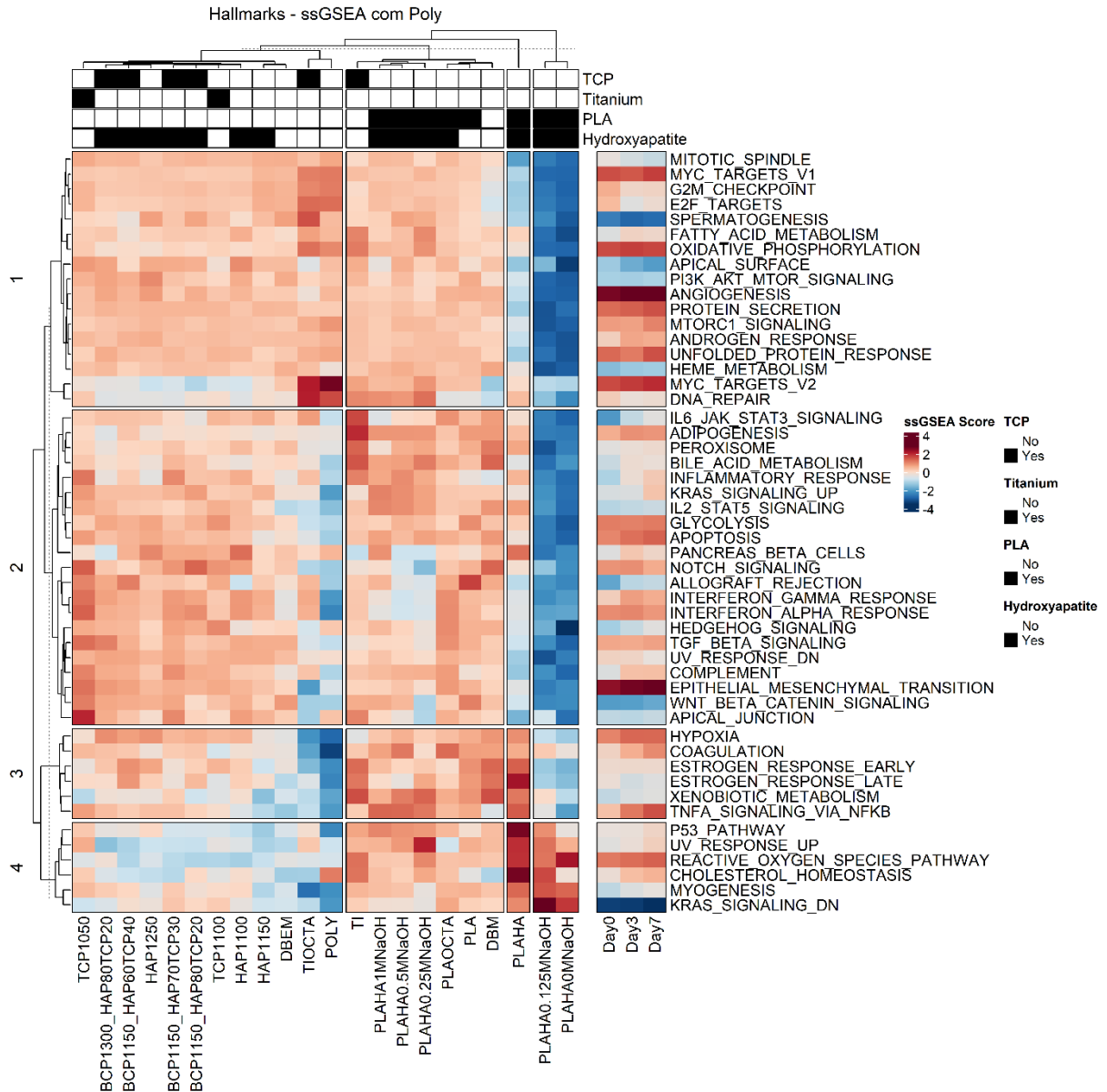


Figura 9 - Mapa de calor correlacionando os 21 biomateriais analisados com os scores ssGSEA gerados para cada uma das 50 assinaturas moleculares do conjunto gênico Hallmarks. Juntamente com um segundo mapa representando os scores de expressão gênica ao longo de 7 dias (D0, D3 e D7), utilizados como controle biológico da progressão osteogênica para comparação com os perfis dos biomateriais.

Com base nas análises de enriquecimento realizadas pelos scores *GSVA*, *GSEA* e *ssGSEA*, foram gerados mapas de calor para visualizar a correlação entre os biomateriais e as assinaturas moleculares presentes no conjunto *Hallmarks*. Para essa visualização, diversos métodos de cálculo de distância serão utilizados na construção dos dendrogramas, permitindo observar o comportamento dos dados analisados. Até o momento, as análises foram realizadas utilizando a distância euclidiana, um método mais simples, o que possibilitou verificar a

viabilidade dos modelos de classificação estudados. No entanto, a versão final do aplicativo possui a opção de escolher uma métrica de distância desejada para a análise.

Os mapas de calor apresentados nesta análise destacam o grau de interação entre os biomateriais, representados no eixo das abscissas, e as assinaturas moleculares, dispostas no eixo das ordenadas. Este grau de interação é ilustrado por meio de diferentes tonalidades de cores, utilizando uma escala uniforme: cores mais frias correspondem aos valores mais baixos, enquanto cores mais quentes representam os valores mais altos. Além da visualização da intensidade dessas interações, os gráficos também evidenciam a clusterização dos biomateriais, realizada por meio de diferentes métodos de cálculo de distância para a construção dos dendrogramas, conforme descrito anteriormente.

Complementando essa análise, foi incorporada uma visualização adicional nos mapas de calor, representando os valores de *Z-score* da expressão gênica em células ao longo de um período de 7 dias (com D0 correspondendo ao tempo inicial, D3 ao terceiro dia e D7 ao sétimo dia). Os dados utilizados nessa etapa foram tratados e analisados a partir do conjunto disponibilizado no estudo de Barter (Barter et al., 2024). Essa abordagem foi adotada com o objetivo de atuar como controle biológico da progressão osteogênica nas assinaturas moleculares em análise. A lógica dessa comparação baseia-se na premissa de que o padrão observado no dia 7 (D7) representa o perfil esperado de diferenciação osteogênica. Assim, quando uma assinatura apresenta correlação positiva no D7 (indicada por tonalidades avermelhadas), espera-se que o biomaterial em estudo também demonstre comportamento semelhante para essa mesma assinatura, o que permite avaliar a sua potencial eficácia na promoção da osteogênese.

As análises foram feitas todas com a inclusão do grupo POLY, justificando-se pelo fato de que, no estudo conduzido por Barter et al. (2024), a metodologia empregada foi o RNA-seq. Diferentemente de outras abordagens, como a utilização do *limma*, que realiza um ajuste linear seguido da normalização dos dados antes da identificação dos genes diferencialmente expressos, a técnica de RNA-seq possibilita o cálculo direto dos *scores* a partir das contagens brutas de transcritos ou de valores previamente normalizados, como os TPM (Transcripts Per Million). Em razão dessa característica metodológica, o grupo controle POLY é mantido de forma constante nas análises realizadas com dados derivados de RNA-seq. Por esse motivo, para esse momento, foi optado por retirar a análise por meio dos mapas de calor referente ao *GSVA* sem o POLY.

Com a combinação dos elementos mencionados, é possível avaliar como os diferentes *scores* de enriquecimento selecionados para análise podem influenciar a relação entre biomateriais e assinaturas moleculares. Com base nas representações gráficas apresentadas, as análises de dissimilaridade revelam que alguns conjuntos de biomateriais permanecem agrupados em um mesmo *cluster*, mesmo diante das variações nos *scores*. Esse comportamento é particularmente evidente nos biomateriais à base de *Hidroxiapatita*, especificamente nos conjuntos formados pelas variações do *PLAHA* e do *BCP*. Tal comportamento corrobora a conclusão de que o modelo de clusterização proposto é viável, configurando-se como uma ferramenta eficaz para verificar a correlação entre materiais com composições semelhantes.

Adicionalmente, com base nas figuras geradas, destaca-se como um resultado relevante a variação na expressão gênica em função do tratamento aplicado a um mesmo biomaterial. Um exemplo ilustrativo dessa relação pode ser observado na comparação entre os materiais *PLAHA0.5MNaOH* e *PLAHA1MNaOH*, que compartilham a mesma composição baseada em *PLA* e hidroxiapatita, diferenciando-se unicamente pela concentração de *NaOH* utilizada no processo de tratamento superficial (0,5 M e 1 M, respectivamente).

Mesmo estando agrupados em um mesmo cluster, nota-se que os perfis de correlação com as assinaturas moleculares podem apresentar variações consideráveis em função da concentração de *NaOH*, tanto nas análises *GSVA* quanto nas análises *GSEA*. Em alguns casos, essas variações são expressivas a ponto de exibirem comportamentos opostos. No caso específico do *GSVA*, destacam-se as assinaturas "*Cholesterol Homeostasis*" e "*Apical Surface*" como exemplos dessas divergências. Já na análise *GSEA*, assinaturas como "*Angiogenesis*" e "*Myogenesis*" apresentam comportamentos distintos entre os dois materiais. A análise *ssGSEA* também revelou variações relevantes, especialmente nas assinaturas "*Pancreas Beta Cells*" e "*Apical Surface*".

Outra análise relevante extraída dos mapas de calor diz respeito à avaliação do potencial osteoindutivo dos biomateriais investigados. Conforme discutido anteriormente, para que essa relação seja considerada positiva, é necessário que tanto o biomaterial quanto o perfil de expressão gênica no dia 7 (D7) do tratamento apresentem escores positivos para a mesma assinatura molecular, sugerindo um comportamento compatível com a progressão da osteogênese.

Com base nesse critério, é possível observar esse comportamento de correlação em diferentes contextos analíticos. Na análise *GSVA*, por exemplo, os materiais *HAPI150*,

PLAHA0.25MNaOH, *TIOCTA* e *POLY* demonstraram *scores* positivos consistentes com o perfil do D7 para a assinatura "*MYC Targets*". Na análise *GSEA*, os biomateriais *PLAHA*, *TCP1050* e *TI* apresentaram resultados compatíveis com a assinatura "*Interferon Alpha Response*". Já na abordagem *ssGSEA*, os materiais *PLAHA*, *PLAHA1MNaOH* e *PLAHA0.5MNaOH* exibiram correlação positiva com a assinatura "*TNFA Signaling via NFkB*", reforçando o potencial dessas formulações na promoção de processos associados à osteogênese.

Por fim, para uma avaliação abrangente do modelo de clusterização empregado, foram gerados mapas de calor com base nas diferentes distâncias dos dendrogramas propostos neste estudo. Esses mapas estão disponíveis na primeira versão do *ShinyApp*: *OsteoCLUST*

3.4 Análises de redução de dimensionalidade

Foi empregado o método de Análise de Componentes Principais (PCA) para avaliar os diferentes *scores* obtidos em relação aos biomateriais envolvidos. Para determinar os componentes principais (PCs) do conjunto analisado, foram gerados *Scree Plots* para os *scores* *GSVA*, *GSEA* e *ssGSEA*, conforme ilustrado na **Figura 10**. Por meio dos *Scree Plots* é possível identificar visualmente a quantidade de componentes principais que capturam a maior parcela de variância dos dados analisados, bem como a porcentagem de captura de cada componente. Dessa forma, a análise desse tipo de gráfico ajuda a definir a quantidade de componentes principais para que haja uma interpretação significativa dos dados, de modo a selecionar apenas os componentes que contribuam significativamente para a variabilidade.

Com base na análise inicial feita por meio dos *Scree Plots*, foram elaborados os gráficos de PCA, apresentados de forma comparativa entre os *scores* e, adicionalmente, analisados individualmente em relação à composição de cada biomaterial. Os gráficos variam entre TCP (**Figura 11**), Hidroxiapatita (**Figura 12**), PLA (**Figura 13**) e Titânio (**Figura 14**). Essa abordagem permite não apenas visualizar as relações entre os *scores*, mas também compreender como cada biomaterial se comporta individualmente dentro do conjunto analisado.

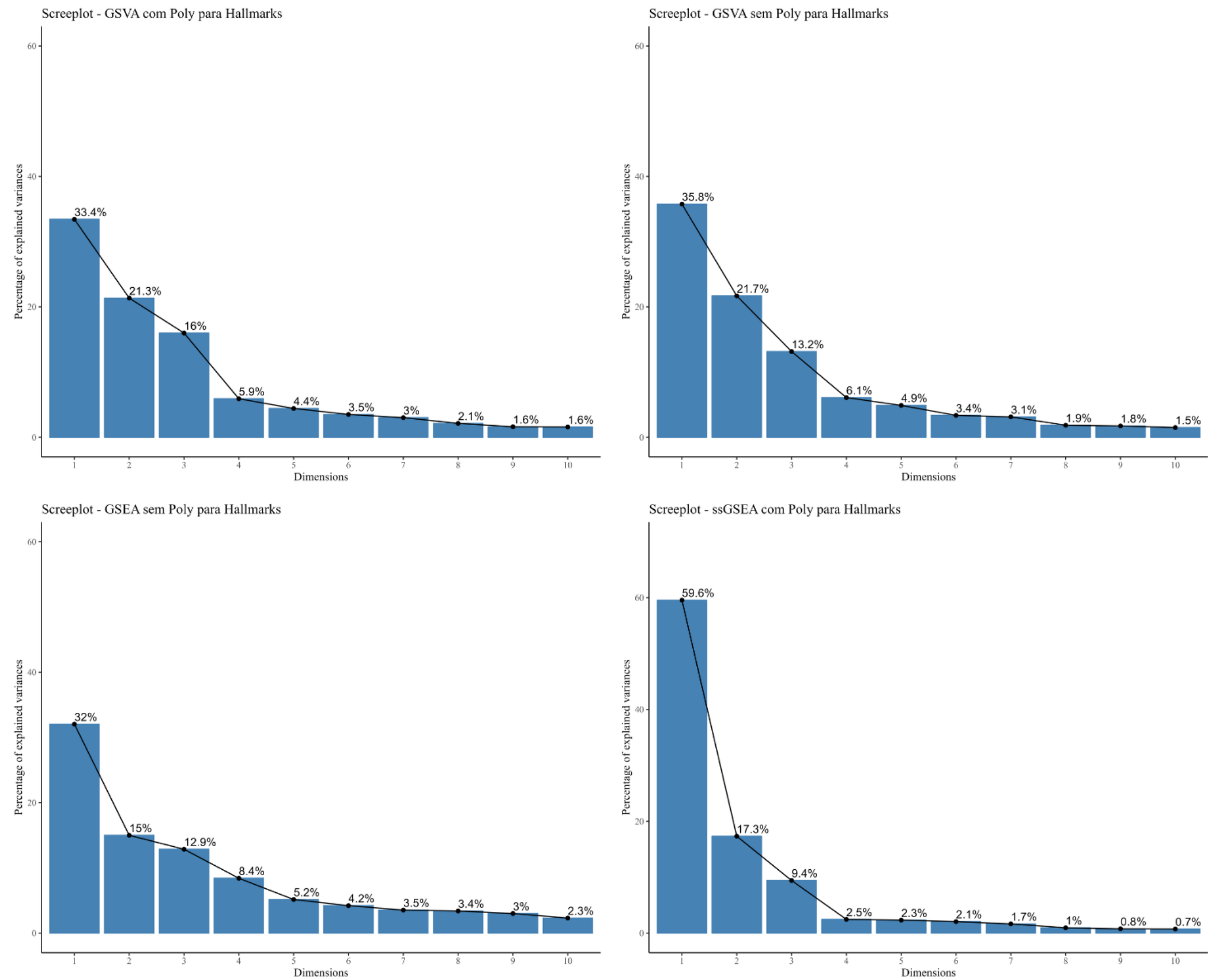


Figura 10 - Scree plot dos 21 biomateriais analisados para o conjunto Hallmarks, com base nos scores GSVA (com POLY), GSVA (sem POLY), GSEA e ssGSEA, respectivamente.

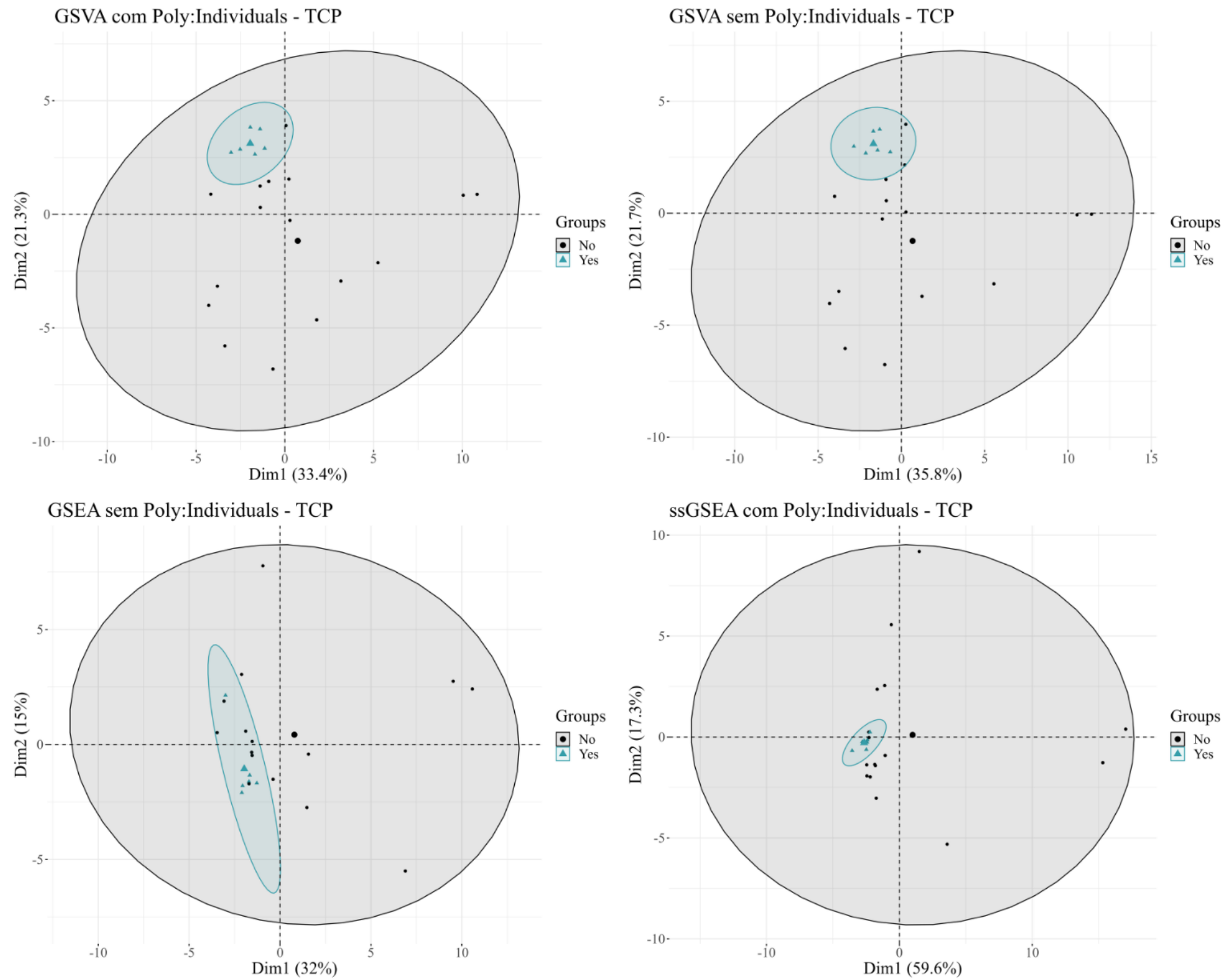


Figura 11 - Análise de componentes principais (PCA), analisando a distribuição dos 21 biomateriais para o conjunto Hallmarks, com base na presença ou não de TCP na composição, sendo avaliados para os scores GSVA (com POLY), GSVA (sem POLY), GSEA e ssGSEA, respectivamente. Estudo feito para 1ª e 2ª dimensão.

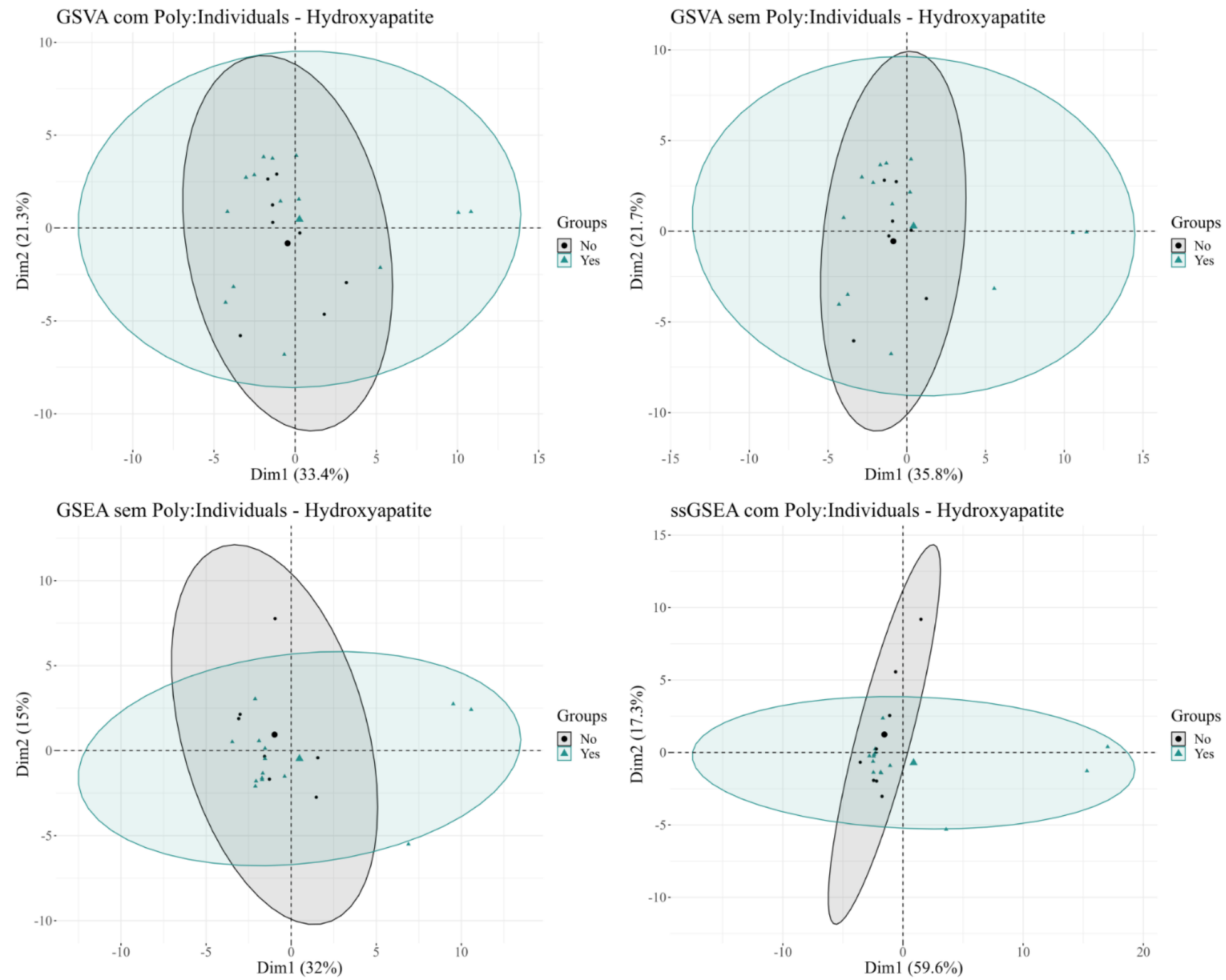


Figura 12 - Análise de componentes principais (PCA), analisando a distribuição dos 21 biomateriais para o conjunto Hallmarks, com base na presença ou não de Hidroxiapatita na composição, sendo avaliados para os scores GSVA (com POLY), GSVA (sem POLY), GSEA e ssGSEA, respectivamente. Estudo feito para 1ª e 2ª dimensão.

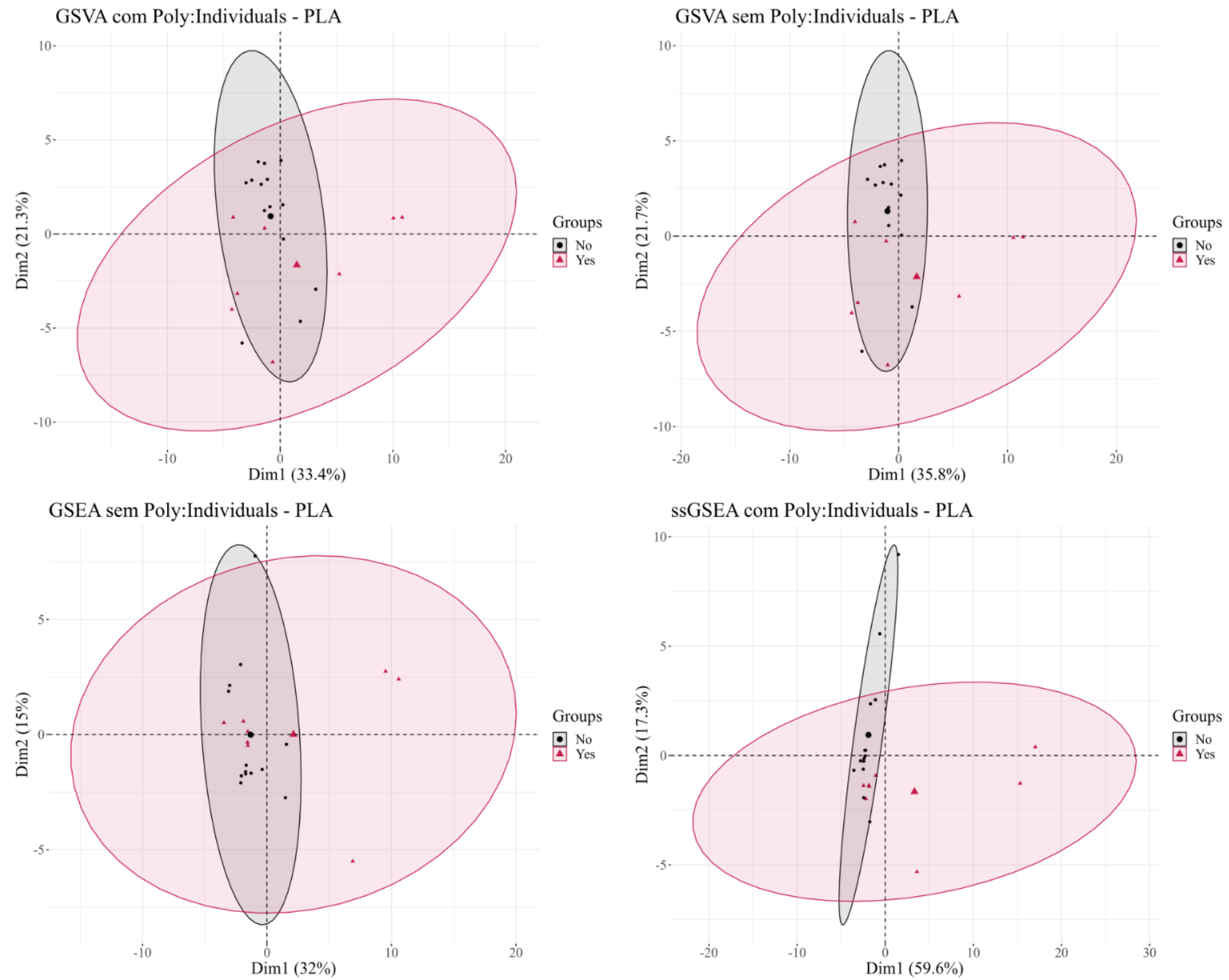


Figura 13 - Análise de componentes principais (PCA), analisando a distribuição dos 21 biomateriais para o conjunto Hallmarks, com base na presença ou não de PLA na composição, sendo avaliados para os scores GSVA (com POLY), GSVA (sem POLY), GSEA e ssGSEA, respectivamente. Estudo feito para 1ª e 2ª dimensão.

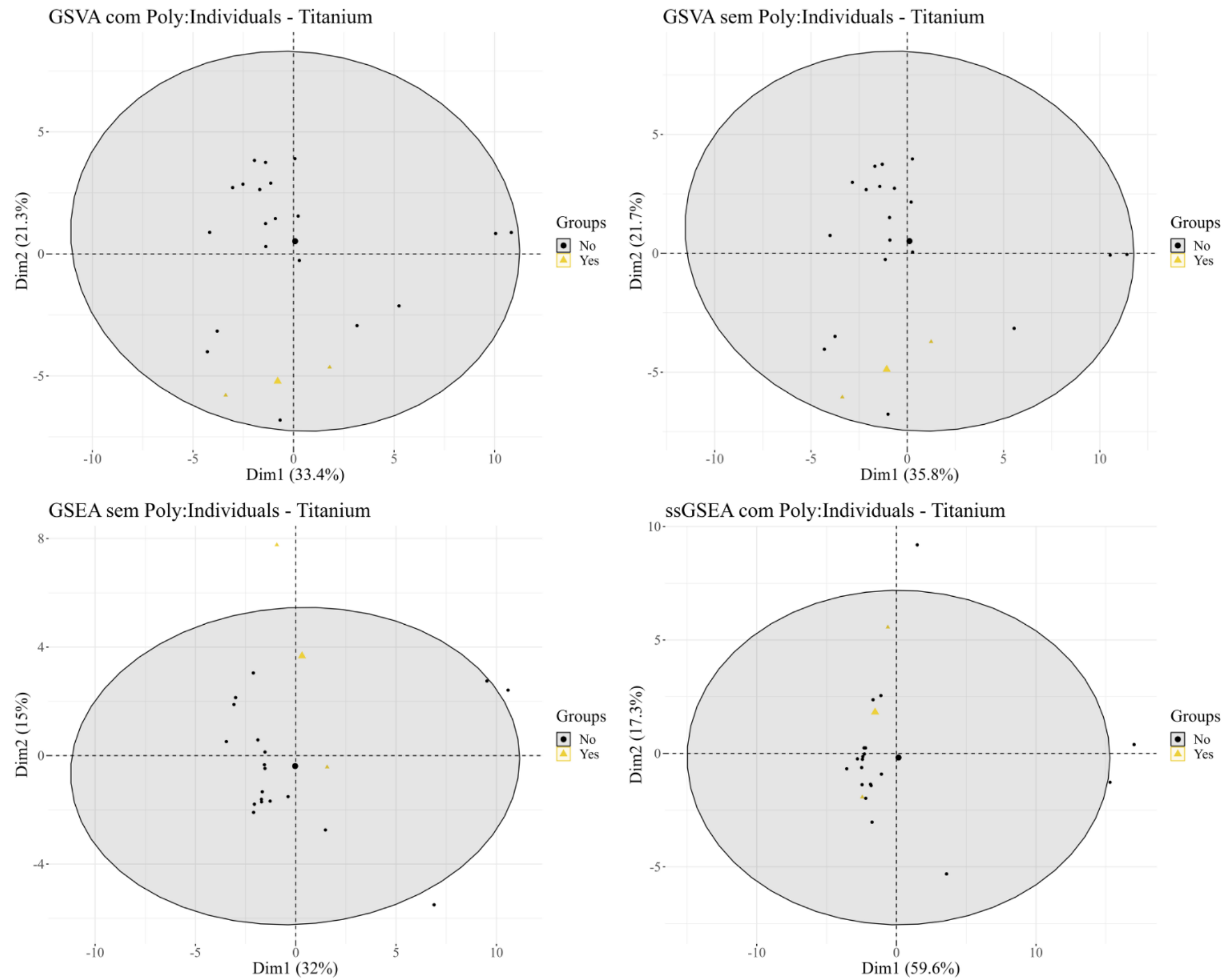


Figura 14 - Análise de componentes principais (PCA), analisando a distribuição dos 21 biomateriais para o conjunto Hallmarks, com base na presença ou não de Titânio na composição, sendo avaliados para os scores GSVA (com POLY), GSVA (sem POLY), GSEA e ssGSEA, respectivamente. Estudo feito para 1ª e 2ª dimensão.

Uma abordagem complementar para avaliar o comportamento dos dados apresentados nesta monografia é por meio da análise de gráficos *scree plot*. No contexto do *GSVA* com *POLY*, as análises unidimensionais e bidimensionais explicam aproximadamente 55% da variabilidade dos dados. Para o *GSVA* sem *POLY*, essa porcentagem aumenta para quase 58%, mantendo as dimensões analíticas. No caso do *GSEA*, a explicação dos dados em 1 e 2 dimensões é cerca de 47%. Em contrapartida, a análise *ssGSEA* demonstra uma explicação significativamente maior, com aproximadamente 77% da variabilidade dos dados sendo capturada nas dimensões consideradas. A avaliação da variabilidade dos dados é fundamental: uma distribuição de pontos mais concentrada indica uma menor variabilidade na dimensão analisada. Após a realização dessa análise inicial, empregando tratamento estatístico por meio da Análise de Componentes Principais (PCA), foram elaborados gráficos que representam os scores calculados, os quais são fundamentados na composição dos biomateriais investigados, variando entre a presença e a ausência de *TCP*, *Titânio*, *PLA* e *Hidroxiapatita*.

Para a análise com o *TCP*, observamos que, para os *scores* de *GSVA* (tanto com quanto sem *POLY*), existe um conjunto bem definido que representa a presença de *TCP* na composição dos biomateriais analisados. A análise gráfica indica que a variabilidade dos pontos associados à presença de *TCP* é reduzida, tanto no primeiro componente principal (Dim1), quanto no segundo componente (Dim2). No que se refere ao score *GSEA*, observa-se que os pontos apresentam uma distribuição mais espaçada na Dimensão 2 (Dim2), enquanto se mantêm agrupados na Dimensão 1 (Dim1). Em contrapartida, para o score *ssGSEA*, identificamos um conjunto bem definido e coeso, sugerindo que este método é eficaz na diferenciação de materiais que contêm *TCP* em sua composição. Ao analisar a ausência de *TCP*, os quatro métodos de cálculo de *scores* demonstram resultados semelhantes, evidenciando uma alta variabilidade nos dados, com os pontos bem dispersos, o que indica uma menor consistência nas medições.

Com base nas figuras apresentadas, observa-se que as análises realizadas tanto para a *Hidroxiapatita* quanto para o *PLA* resultaram em gráficos com estruturas semelhantes ao se comparar os mesmos *scores*. Essa congruência sugere que o método utilizado para a análise pode não ser adequado para este tipo de dado, uma vez que não permite a distinção clara dos pontos em questão. Uma possível justificativa para essa limitação reside nas características físico-químicas dos materiais estudados. Assim, a presença simultânea de *PLA* e *Hidroxiapatita* nos materiais analisados dificulta a diferenciação dos biomateriais com base nesse critério.

Por fim, no que se refere à presença do titânio na composição dos biomateriais analisados, observou-se que os pontos correspondentes a esse composto não foram suficientemente representativos para uma análise estatística robusta, uma vez que não atingiram a quantidade mínima recomendada de dados (ao menos três amostras). Esse aspecto merece destaque, considerando que o titânio é amplamente utilizado no setor de biomateriais, mas encontra-se subrepresentado no conjunto de dados utilizado. Ainda assim, os pontos referentes ao titânio nos gráficos de *GSVA* (com e sem *POLY*) apresentaram uma distribuição relativamente próxima, sugerindo certa consistência interna. Em contraste, os *scores* obtidos pelas abordagens *GSEA* e *ssGSEA* revelaram maior dispersão entre os pontos, indicando variações que podem estar associadas à escassez de dados ou a particularidades metodológicas dessas análises.

3.5 SynthPop: Dados sintético do conjunto Hallmarks

Com o auxílio do pacote *Synthpop*, foram gerados dados sintéticos por meio do método de árvore de classificação e regressão (CART), tendo como *input* os dados calculados dos *scores* *GSVA*, *GSEA* e *ssGSEA* para as assinaturas moleculares do conjunto gênico *Hallmarks*. Foram gerados um total de 600 exemplares sintéticos para cada *score* analisado, sendo 150 para cada condição da **Tabela 1** (*TCP*, *Hydroxyapatite*, *PLA* e *Titanium*). Pode-se verificar a contagem desses dados divididos pela presença ou não dos compósitos por meios das seguintes figuras: **Figura 15**, para *GSVA* com ausência de *POLY*; **Figura 16**, *GSVA* com *POLY*; **Figura 17**, para o *score* *GSEA* e, por fim, para o *ssGSEA* a **Figura 18**.

A análise preliminar, fundamentada nos mapas de calor apresentados (*GSVA* com *POLY*: **Figura 19**; *GSVA* sem *POLY*: **Figura 20**, *GSEA*: **Figura 21**, *ssGSEA*: **Figura 22**) e nas assinaturas moleculares selecionadas, indicou que o método empregado para a geração de dados sintéticos mostrou-se eficaz. A distribuição dos valores gerados apresentou uma proximidade significativa em relação aos valores calculados previamente, o que corrobora para a utilização desses dados para testar e validar o modelo de classificação proposto.

Com a validação do método de geração usado neste trabalho, a intenção é utilizar os dados gerados como treinamento para modelos de *machine learning* focado em predição de resultados, que serão os tópicos desenvolvidos nas próximas etapas do projeto. Para uma validação completa do método selecionado, serão gerados dados sintéticos para os *scores* que envolvem os outros 8 conjuntos de assinaturas moleculares apresentadas na **Tabela 3**.

A utilização de dados sintéticos, neste contexto, apresenta-se como uma estratégia para contornar a escassez de informações disponíveis na literatura. A limitação no número de observações pode comprometer a robustez e a generalização dos modelos de análise, especialmente em abordagens como a clusterização, que dependem da densidade e representatividade dos dados. Assim, a geração de dados sintéticos visa ampliar o conjunto de dados original de forma controlada. Após a devida validação, essa expansão permitirá o aprimoramento da qualidade do modelo de clusterização proposto, contribuindo para resultados mais consistentes e confiáveis.

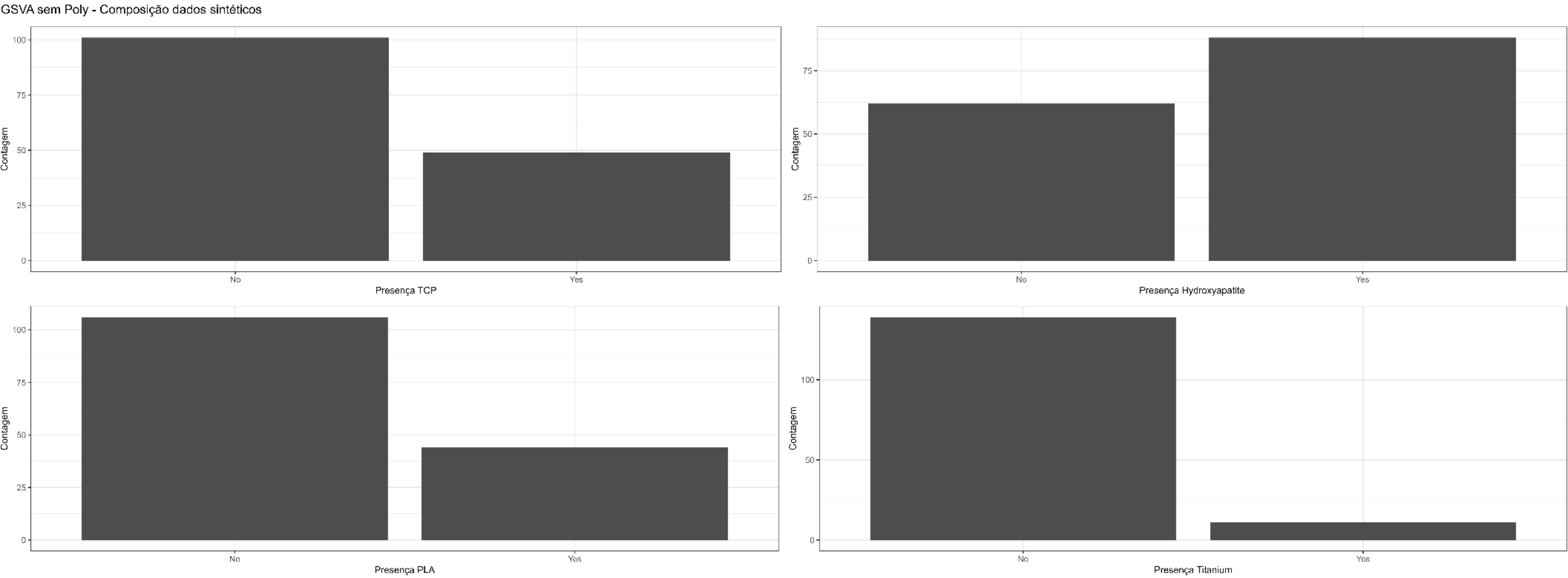


Figura 15 - Contagem de dados gerados sinteticamente para score GSVA (sem POLY), avaliando a presença ou não dos compósitos TCP, Hydroxyapatite, PLA e Titanium

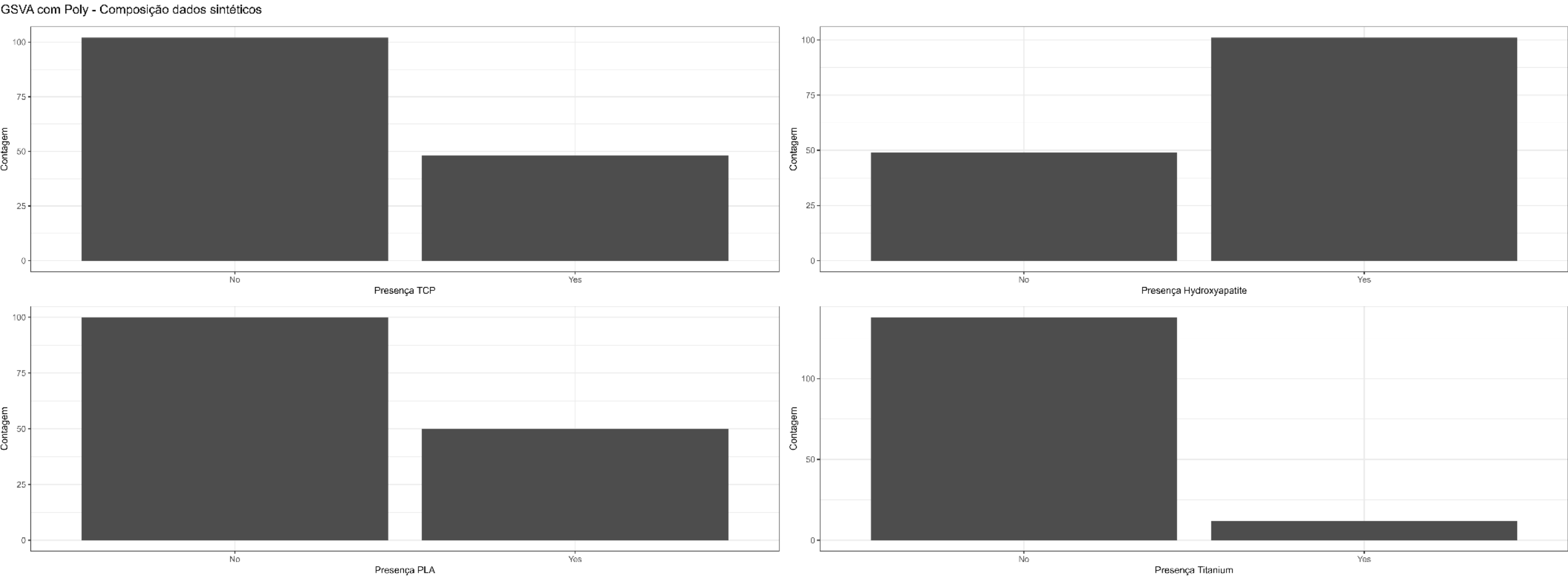


Figura 16 - Contagem de dados gerados sinteticamente para score GSVA (com POLY), avaliando a presença ou não dos compósitos TCP, Hydroxyapatite, PLA e Titanium.

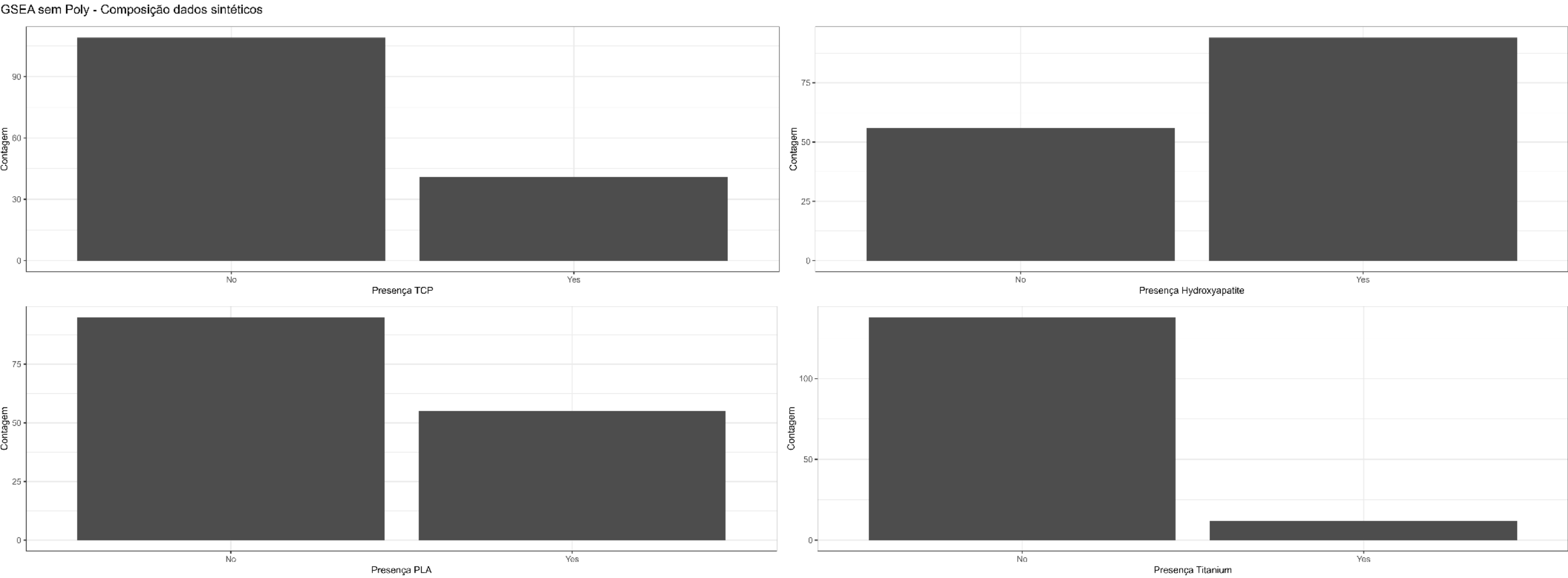


Figura 17 - Contagem de dados gerados sinteticamente para score GSEA, avaliando a presença ou não dos compósitos TCP, Hydroxyapatite, PLA e Titanium.

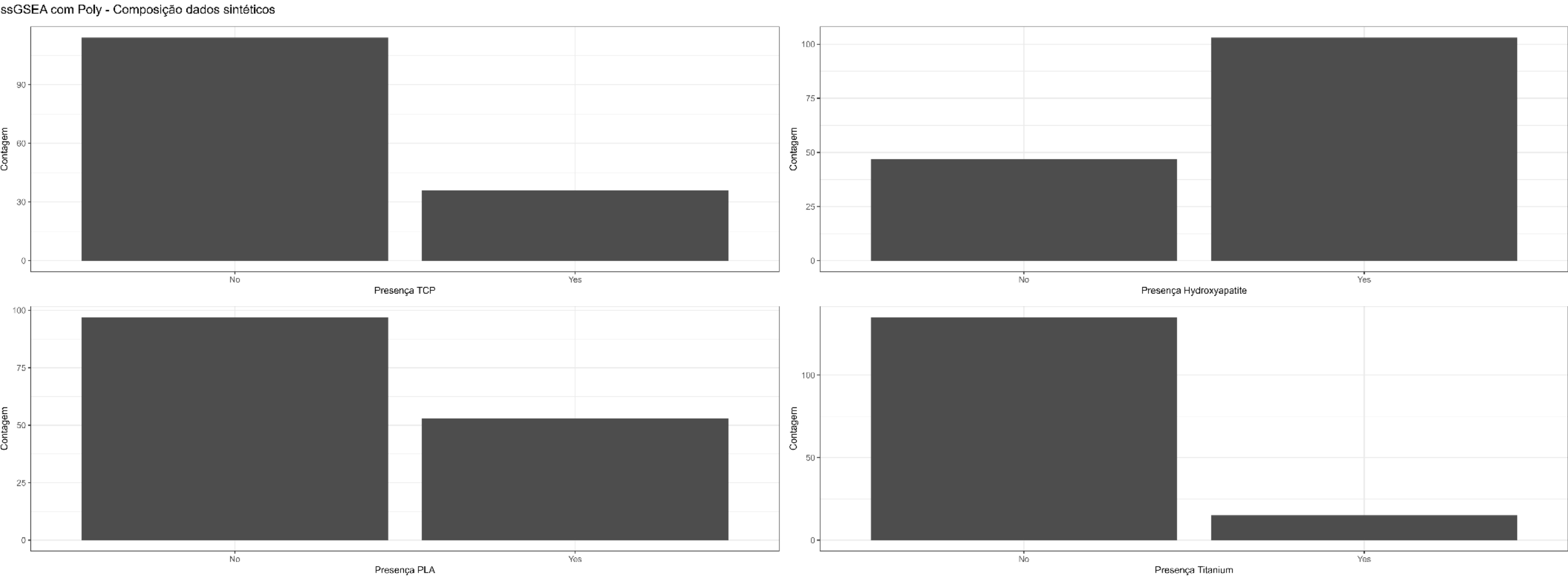


Figura 18 - Contagem de dados gerados sinteticamente para score ssGSEA, avaliando a presença ou não dos compósitos TCP, Hydroxyapatite, PLA e Titanium.

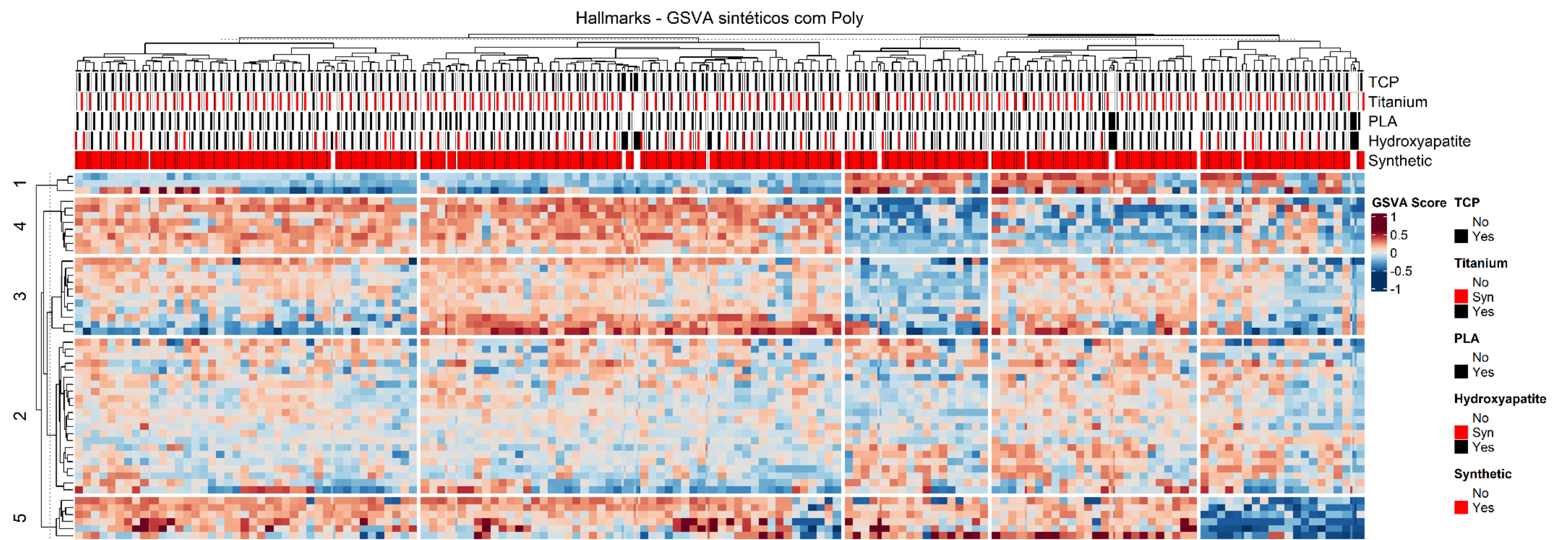


Figura 19 - Mapa de calor correlacionando os 150 biomateriais gerados sinteticamente com base no score GSVA (com POLY) para cada uma das 50 assinaturas moleculares presentes no conjunto gênico Hallmarks.

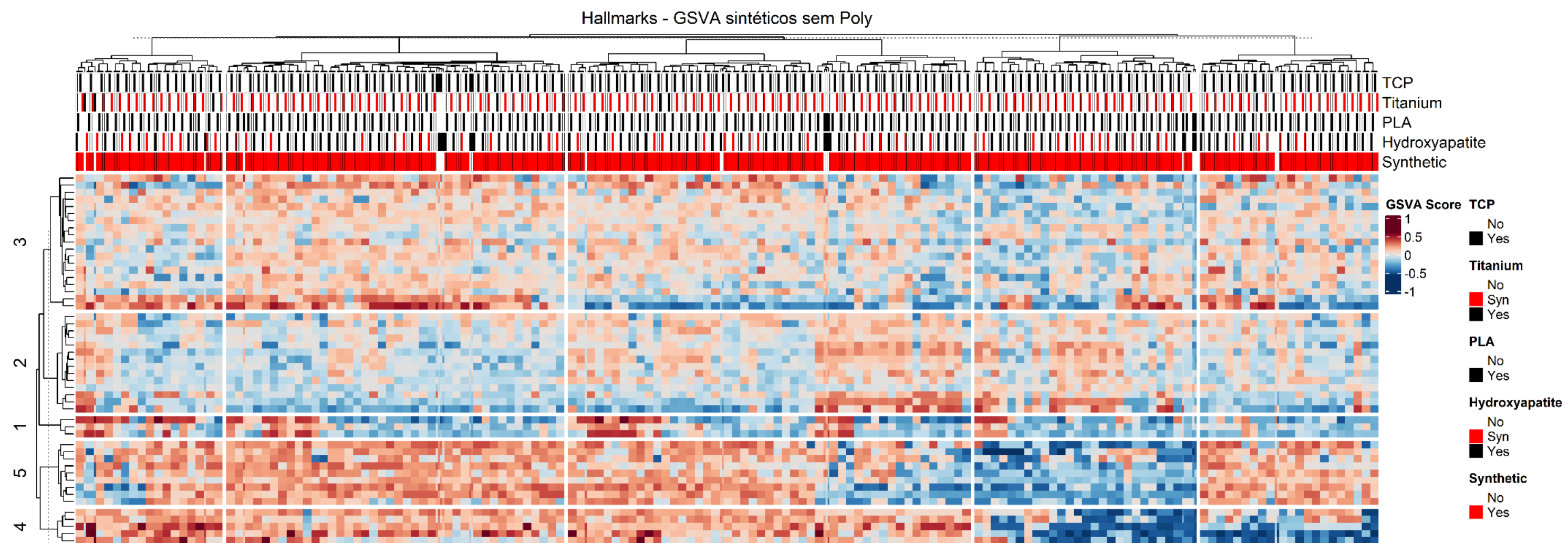


Figura 20 - Mapa de calor correlacionando os 150 biomateriais gerados sinteticamente com base no score GSVA (sem POLY) para cada uma das 50 assinaturas moleculares presentes no conjunto gênico Hallmarks.

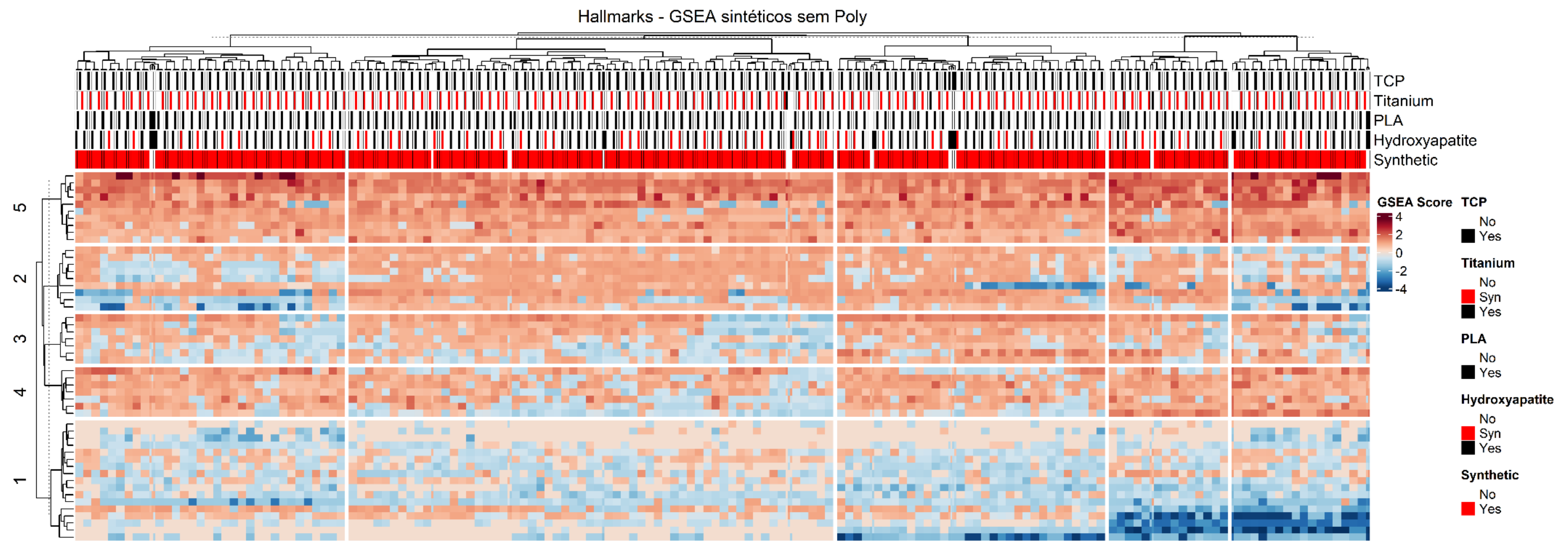


Figura 21 - Mapa de calor correlacionando os 150 biomateriais gerados sinteticamente com base no score GSEA para cada uma das 50 assinaturas moleculares presentes no conjunto gênico Hallmarks.

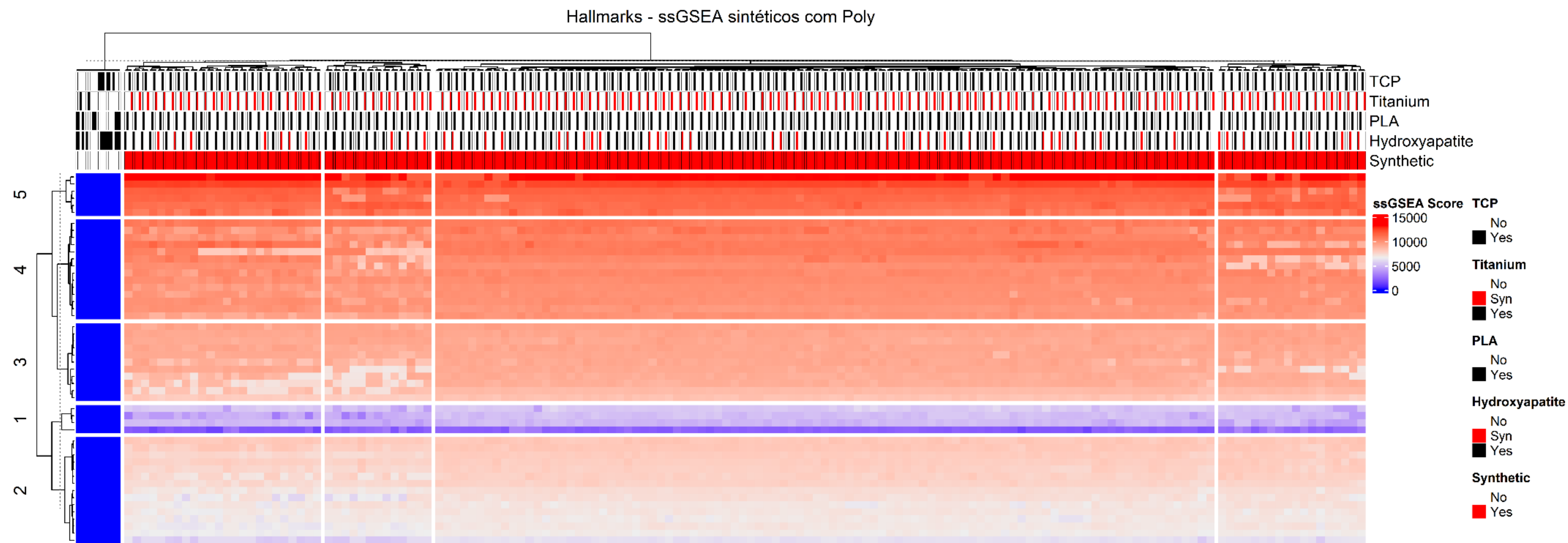


Figura 22 - Mapa de calor correlacionando os 150 biomateriais gerados sinteticamente com base no score ssGSEA para cada uma das 50 assinaturas moleculares presentes no conjunto gênico Hallmarks.

3.6 Modelos de predição

Avaliando o desempenho preditivo de diferentes abordagens de aprendizado de máquina, foram aplicados e comparados diversos algoritmos de classificação aos conjuntos de *scores* obtidos. Os modelos empregados nesta etapa incluíram: Regressão Logística, Árvore de Decisão (*Decision Tree*), Floresta Aleatória (*Random Forest*), XGBoost, Máquina de Vetores de Suporte (*Support Vector Machine* – SVM) e Naive Bayes. A avaliação da performance dos modelos foi conduzida a partir de um conjunto abrangente de métricas de classificação, tais como *f-score*, *precision*, *sensitivity (recall)*, *specificity*, área sob a curva ROC (ROC AUC) e *accuracy*. Os resultados obtidos foram organizados em formato tabular, de modo a permitir uma análise detalhada tanto dos *scores* individualmente quanto da composição de cada biomaterial avaliado.

A aplicação dos modelos foi estruturada em duas etapas: treinamento e teste. O treinamento dos modelos foi realizado utilizando dados sintéticos gerados e previamente validados no trabalho, enquanto a etapa de teste foi conduzida com os dados reais disponíveis. Essa estratégia visou avaliar a capacidade de generalização dos modelos em um cenário mais próximo à aplicação prática, além de verificar a coerência entre os dados sintéticos e os dados experimentais reais.

Ao se considerar os conjuntos de *scores* *GSEA* e *GSVA* sem a presença do POLY (**Tabela 4 e Tabela 5**), bem como *GSVA* e *ssGSEA* com POLY (**Tabela 6 e Tabela 7**), observou-se que, de maneira geral, os modelos com melhor desempenho preditivo foram a Árvore de Decisão e o XGBoost. Esses modelos apresentaram resultados superiores tanto nas etapas de treinamento quanto de teste, e para as diferentes composições de biomateriais analisadas. Em contrapartida, o modelo baseado no classificador Naive Bayes apresentou o pior desempenho global, sendo, em alguns casos, inviável sua aplicação em razão de limitações associadas ao próprio algoritmo frente à estrutura dos dados utilizados.

É importante destacar que, para os quatro *scores* analisados, a maioria dos modelos testados não pôde ser aplicada de forma adequada aos dados referentes a biomateriais contendo Titânio em sua composição. Essa limitação se deve à escassez de amostras disponíveis no conjunto de dados original. Embora a etapa de treinamento dos modelos com dados sintéticos contendo Titânio tenha sido executada com sucesso, a quantidade insuficiente de dados reais impossibilitou a validação externa e o teste adequado desses modelos. Dessa forma, não foi possível concluir as análises preditivas para esse subconjunto, restringindo a generalização dos resultados para materiais com essa composição.

Tabela 4 - Desempenho das métricas de machine learning nos conjuntos de treino e teste do GSEA sem POLY, exibido separadamente para as composições PLA, Hydroxyapatite, TCP e Titanium, respectivamente. Cores verdes indicam maiores valores e cores vermelhas indicam valores menores.

[illegible]

Tabela 5 - Desempenho das métricas de machine learning nos conjuntos de treino e teste do GSVA sem POLY, exibido separadamente para as composições PLA, Hydroxyapatite, TCP e Titanium, respectivamente. Cores verdes indicam maiores valores e cores vermelhas indicam valores menores.

[illegible]

Tabela 7 - Desempenho das métricas de machine learning nos conjuntos de treino e teste do ssGSEA com POLY, exibido separadamente para as composições PLA, Hydroxyapatite, TCP e Titanium, respectivamente. Cores verdes indicam maiores valores e cores vermelhas indicam valores menores.

		Treino						Teste					
		f_meas	precision	sensitivity	specificity	roc_auc	accuracy	f_meas	precision	sensitivity	specificity	roc_auc	accuracy
PLA	Modelo Logístico	0.912	0.924	0.906	0.849	0.942	0.887	0.897	0.867	0.929	0.75	0	0.864
	Modelo Árvore de Decisão	0.938	1	0.886	1	0.943	0.927	0.923	1	0.857	1	0	0.909
	Modelo Randon Forest	-	-	-	-	-	-	-	-	-	-	-	-
	Modelo XGBoost	0.938	1	0.886	1	0.943	0.927	0.923	1	0.857	1	0	0.909
	Modelo SVM	0.925	0.959	0.897	0.922	0.948	0.906	0.857	0.857	0.857	0.75	0	0.818
	Modelo Naive Bayes	0.854	0.807	0.918	0.56	0.938	0.792	0.774	0.706	0.857	0.375	0	0.682
		Treino						Teste					
		f_meas	precision	sensitivity	specificity	roc_auc	accuracy	f_meas	precision	sensitivity	specificity	roc_auc	accuracy
Hydroxyapatite	Modelo Logístico	0.567	0.645	0.509	0.865	0.73	0.754	0.571	0.667	0.5	0.857	0	0.727
	Modelo Árvore de Decisão	0.807	0.714	0.94	0.826	0.886	0.861	0.8	0.667	1	0.714	0	0.818
	Modelo Randon Forest	-	-	-	-	-	-	-	-	-	-	-	-
	Modelo XGBoost	0.807	0.714	0.94	0.826	0.887	0.861	0.8	0.667	1	0.714	0	0.818
	Modelo SVM	-	-	-	-	-	-	-	-	-	-	-	-
	Modelo Naive Bayes	-	-	-	-	-	-	-	-	-	-	-	-
		Treino						Teste					
		f_meas	precision	sensitivity	specificity	roc_auc	accuracy	f_meas	precision	sensitivity	specificity	roc_auc	accuracy
TCP	Modelo Logístico	0.912	0.89	0.939	0.607	0.933	0.86	0.824	0.778	0.875	0.333	0	0.727
	Modelo Árvore de Decisão	0.935	0.99	0.886	0.971	0.929	0.907	0.933	1	0.875	1	0	0.909
	Modelo Randon Forest	-	-	-	-	-	-	-	-	-	-	-	-
	Modelo XGBoost	0.94	0.991	0.895	0.971	0.957	0.914	0.848	0.824	0.875	0.5	0	0.773
	Modelo SVM	0.864	0.76	1	0	0.675	0.76	0.842	0.727	1	0	0	0.727
	Modelo Naive Bayes	0.822	0.884	0.771	0.668	0.787	0.745	0.857	1	0.75	1	0	0.818
		Treino						Teste					
		f_meas	precision	sensitivity	specificity	roc_auc	accuracy	f_meas	precision	sensitivity	specificity	roc_auc	accuracy
Titanium	Modelo Logístico	0.947	0.9	1	0	0.415	0.9	0.952	0.909	1	0	0	0.909
	Modelo Árvore de Decisão	-	-	-	-	-	-	-	-	-	-	-	-
	Modelo Randon Forest	-	-	-	-	-	-	-	-	-	-	-	-
	Modelo XGBoost	-	-	-	-	-	-	-	-	-	-	-	-
	Modelo SVM	-	-	-	-	-	-	-	-	-	-	-	-
	Modelo Naive Bayes	0.944	0.899	0.993	0	0.36	0.893	0.952	0.909	1	0	0	0.909

3.7 ShinyApp: OsteoCLUST

O *ShinyApp*³ denominado *OsteoCLUST*, registrado sob o certificado de programa de computador (nº BR512024004865-0) aprovado pelo Instituto Nacional de Propriedade Industrial, tem a sua primeira versão finalizada e validada. A aplicação tem como objetivo principal fornecer, como *output*, um mapa de calor interativo para a avaliação comparativa dos biomateriais analisados, facilitando a interpretação visual dos resultados obtidos por meio das análises realizadas.

Embora a implementação de modelos de predição ainda represente um desafio técnico em aberto, esta primeira versão do *OsteoCLUST* estabelece uma base sólida para futuras expansões. Entre as funcionalidades previstas para versões posteriores, destacam-se a visualização dos dados por meio de Análise de Componentes Principais (PCA) e a incorporação de uma aba dedicada à predição de resultados utilizando modelos de aprendizado de máquina. Tais aprimoramentos visam ampliar o potencial analítico do aplicativo, tornando-o uma ferramenta mais robusta para apoio à pesquisa em biomateriais.

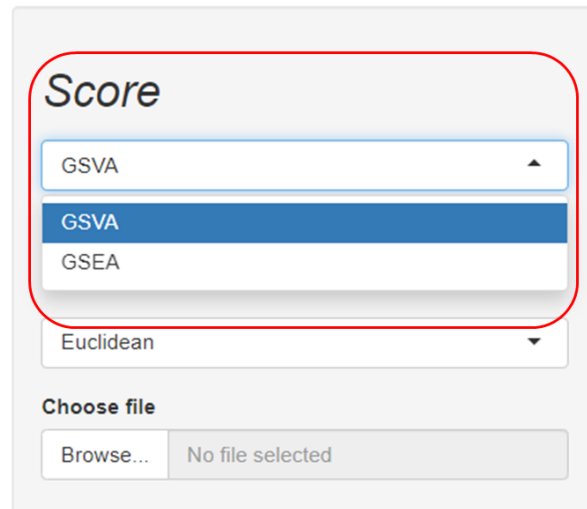
Quanto à funcionalidade do aplicativo desenvolvido, a **Figura 23** apresenta os principais elementos que compõem sua interface na versão atual. A aba “*Score*” permite ao usuário selecionar o score de enriquecimento desejado para a análise, conforme ilustrado na **Figura 24**. Nesta primeira versão, optou-se por não incluir o *score ssGSEA*, em razão do comportamento atípico de seus mapas de calor, os quais podem sugerir inconsistências na análise ou até mesmo limitações inerentes ao próprio método. Em “*Distance*”, é possível definir a métrica de distância a ser utilizada na construção dos dendrogramas apresentados no mapa de calor (**Figura 25**). Por meio do campo “*Choose file*”, o usuário pode inserir arquivos no formato csv, contendo os *scores* de um novo material a ser avaliado, como demonstrado na **Figura 26**. Além dessa funcionalidade, o usuário também pode determinar o número de *clusters* a serem considerados na análise, bem como exportar os resultados gerados por meio do botão “*Download Results*” (**Figura 27**).

Todas as informações fornecidas pelo usuário são processadas por um conjunto de códigos que compõem a base do aplicativo, permitindo a realização das análises de forma automatizada. O resultado dessas análises é apresentado na forma de mapas de calor (**Figura**

³ *ShinyApps* são um conjunto recursos desenvolvidos pela Posit (<https://posit.co/>) para a construção de aplicativos com interface gráfica utilizando linguagem R.

Figura 23 - Visão geral da primeira versão do aplicativo.

App



Score

GSVA

GSVA

GSEA

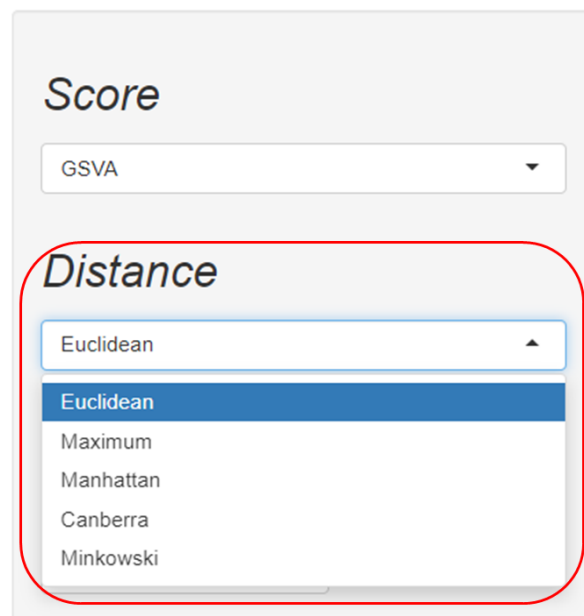
Euclidean

Choose file

Browse... No file selected

Figura 24 - Opções de escolha da aba "Score".

App



Score

GSVA

Distance

Euclidean

Euclidean

Maximum

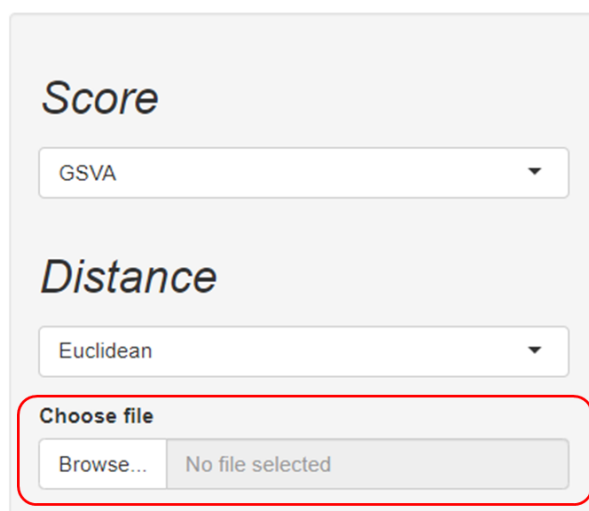
Manhattan

Canberra

Minkowski

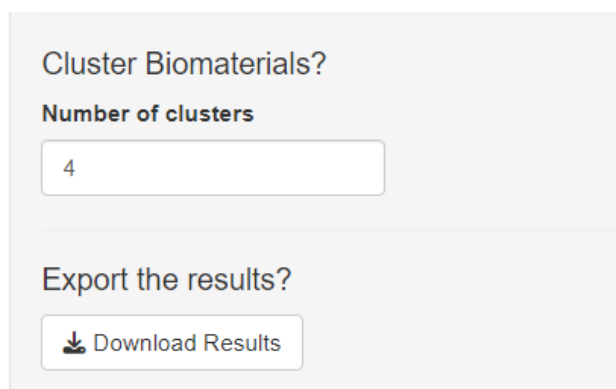
Figura 25 - Opções de escolha da aba "Distance".

App



The screenshot shows a web application interface with a light gray background. At the top, the word "App" is written in a large, dark font. Below it, there are two sections. The first section is titled "Score" in a bold, italicized font. It contains a dropdown menu with "GSVA" selected. The second section is titled "Distance" in a bold, italicized font. It contains a dropdown menu with "Euclidean" selected. Below these, there is a section titled "Choose file" in a bold font. This section is highlighted with a red rectangular border. It contains a "Browse..." button and a text box that says "No file selected".

Figura 26 - Identificação do input para análise de novos biomateriais.



The screenshot shows a web application interface with a light gray background. It contains two sections. The first section is titled "Cluster Biomaterials?" in a bold font. Below it, there is a label "Number of clusters" and a text input field containing the number "4". The second section is titled "Export the results?" in a bold font. Below it, there is a button with a download icon and the text "Download Results".

Figura 27 - Identificação da quantidade de clusters desejados e possível exportação dos resultados.

Com o objetivo de validar o funcionamento do aplicativo desenvolvido, foi utilizado um conjunto de dados oriundo do estudo conduzido por Murali et al. (Murali et al., 2024). Nesse trabalho, os autores investigaram o impacto de diferentes nanomateriais inorgânicos sobre o perfil transcriptômico de células-tronco mesenquimais humanas (hMSCs), com ênfase na indução da diferenciação endocondral.

Foram avaliados quatro tipos de nanomateriais: nano-hidroxiapatita (nHA), nano-whitlockita (nWH), nanosilicatos (nSi) e nano-dióxido de silício (nSiO₂). As células foram expostas a esses materiais e analisadas em três diferentes pontos temporais — dias 7, 14 e 21 — com o intuito de monitorar as alterações na expressão gênica ao longo do tempo.

A metodologia adotada incluiu o sequenciamento de RNA (RNA-seq) para identificação dos genes diferencialmente expressos, permitindo a análise das respostas celulares induzidas por cada biomaterial. Os resultados apontaram os nanosilicatos (nSi) como os mais eficazes na promoção da diferenciação osteogênica, devido à ativação de vias de sinalização relacionadas à ossificação endocondral.

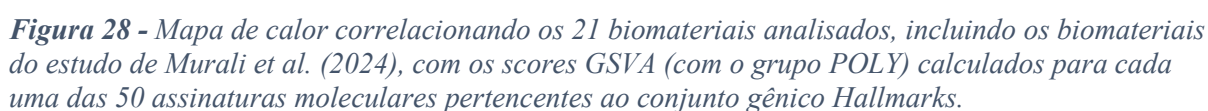
Para a realização do teste no aplicativo, foram utilizados os dados referentes aos biomateriais analisados nesse estudo, com exceção do grupo controle, cujos dados não foram disponibilizados pelos autores.

A **Figura 28** e **Figura 29** apresenta a clusterização dos biomateriais analisados em conjunto com o banco de dados gerado neste trabalho. Como o novo conjunto de dados foi utilizado para validar o aplicativo desenvolvido, essas visualizações foram obtidas por meio da funcionalidade “*Download Results*”. Para a análise foi aplicado o cálculo dos scores *GSVA* e *ssGSEA*, ambos incluindo o grupo controle *POLY*.

A partir da análise da figura correspondente ao método *GSVA*, observa-se uma proximidade significativa entre os biomateriais à base de silício e o grupo controle *POLY*, os quais se agrupam em um mesmo cluster. Além da organização espacial, esses materiais compartilham comportamentos semelhantes para determinadas assinaturas moleculares, como é o caso de “*Oxidative Phosphorylation*” e “*Hypoxia*”, sugerindo similaridades funcionais em suas respostas biológicas.

Adicionalmente, a formação de um *cluster* com maior concentração dos novos biomateriais, juntamente com os materiais DBM e PLA, pode indicar uma semelhança nos perfis transcriptômicos desses grupos. Tal proximidade pode estar relacionada à presença comum de componentes cerâmicos com propriedades físico-químicas similares. Na figura gerada a partir da análise *ssGSEA*, essa tendência de agrupamento se mostra ainda mais evidente, com todos os novos biomateriais organizados dentro de um único *cluster*.

Com base nessa análise, foi possível verificar a funcionalidade do aplicativo desenvolvido na avaliação de novos biomateriais, demonstrando sua capacidade de processar e interpretar dados transcriptômicos de forma eficiente. Os resultados obtidos evidenciam o potencial da ferramenta para identificar padrões biológicos relevantes e agrupar biomateriais com comportamentos similares, o que reforça a importância de sua aplicação em estudos futuros. Dessa forma, o uso do aplicativo pode contribuir significativamente para a análise



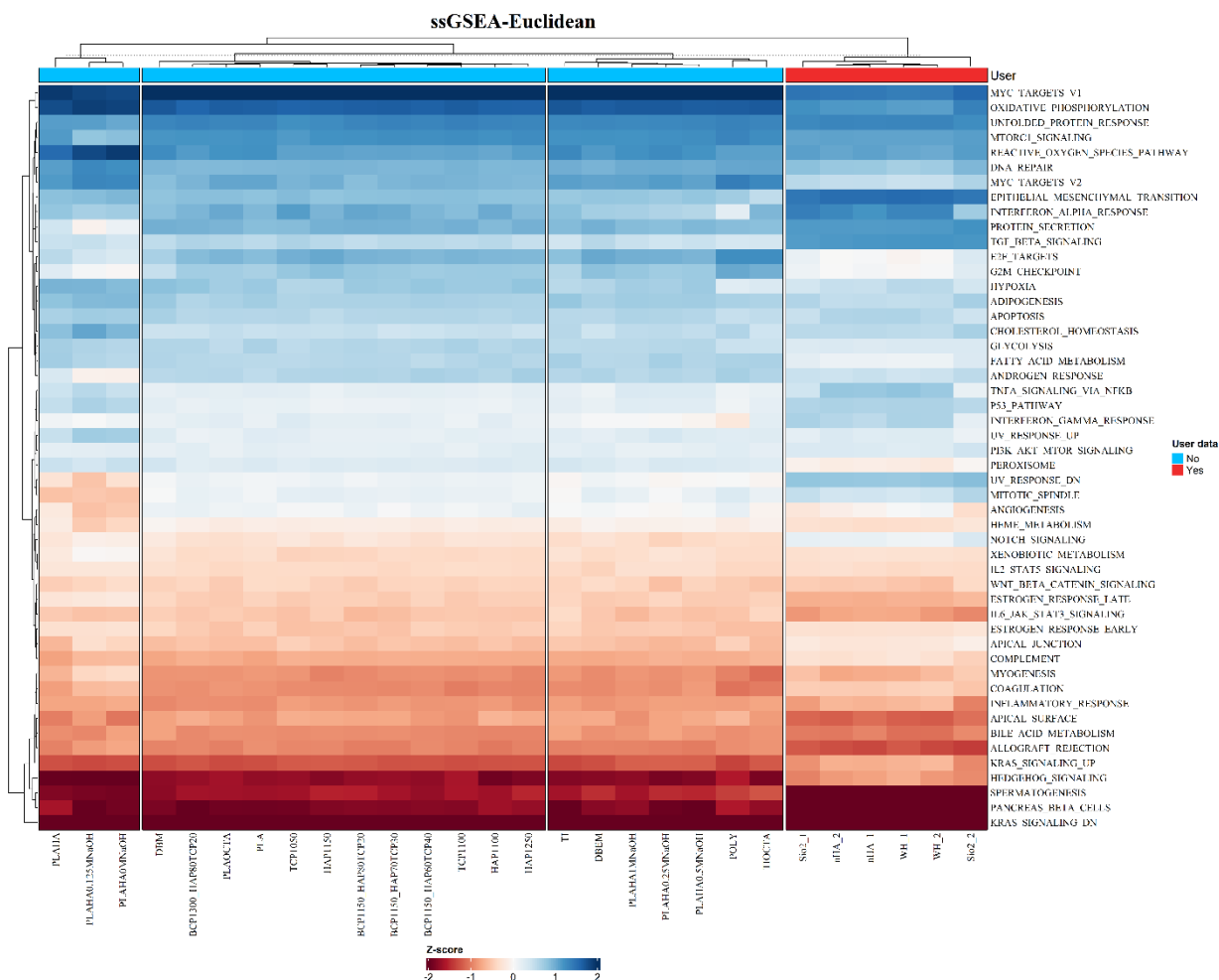


Figura 29 - Mapa de calor correlacionando os 21 biomateriais analisados, incluindo os biomateriais do estudo de Murali et al. (2024), com os scores ssGSEA (com o grupo POLY) calculados para cada uma das 50 assinaturas moleculares pertencentes ao conjunto gênico Hallmarks.

4 CONSTATAÇÕES

Este capítulo abordou uma variedade de metodologias para a análise da distribuição de dados e a avaliação de biomateriais, enfatizando a relevância dos cálculos de *scores* de enriquecimento na identificação de padrões significativos, o que facilitou a interpretação dos resultados obtidos. A construção de mapas de calor possibilitou uma visualização clara das interações entre os biomateriais analisados e o conjunto gênico *Hallmarks*, evidenciando a expressão diferencial sob diversas condições. Além disso, a implementação da técnica de redução de dimensionalidade, com a Análise de Componentes Principais (PCA), foi essencial para a simplificação da complexidade dos dados, permitindo a identificação de agrupamentos e tendências.

Por meio das análises dos dados gerados sinteticamente, é possível destacar a sua relevância pela capacidade de criar conjuntos de dados que replicam as características estatísticas dos dados reais, aumentando assim a disponibilidade de dados e, acima de tudo, preservem a integridade dos dados originais. A visualização desses dados sintéticos, utilizando mapas de calor, revelou-se uma ferramenta eficaz na identificação de padrões, sendo fundamentais para a interpretação dos resultados. Essa abordagem gráfica não apenas facilitou a comparação direta entre os dados sintéticos e os dados originais, mas também evidenciou a eficácia do processo de síntese.

A aplicação dos modelos de classificação permitiu validar as abordagens desenvolvidas, destacando dois algoritmos que obtiveram desempenho satisfatório na classificação dos biomateriais com base em assinaturas moleculares. Esses resultados evidenciam o potencial das técnicas de aprendizado supervisionado para análises transcriptômicas, mesmo diante de limitações amostrais, e contribuem para futuras estratégias preditivas na área de biomateriais.

Contudo, algumas limitações devem ser consideradas no trabalho. Em primeiro lugar, destaca-se o tamanho reduzido do conjunto de dados original, composto por apenas 21 biomateriais. Esse número limitado de amostras compromete a generalização dos resultados e dificulta uma avaliação mais precisa da viabilidade do método proposto. Idealmente, seria necessário validar a abordagem em um conjunto mais amplo e diversificado de biomateriais, a fim de obter conclusões mais robustas e abrangentes.

Adicionalmente, a utilização da linhagem celular MG-63, derivada de osteossarcoma humano, representa outra limitação relevante. Embora essa linhagem seja amplamente utilizada em estudos com biomateriais devido à sua facilidade de cultivo, suas características tumorais

podem introduzir vieses, uma vez que diferem biologicamente das células osteoblásticas normais. Tais diferenças podem impactar diretamente os resultados obtidos nas análises de expressão gênica.

Outra restrição metodológica diz respeito à utilização de microarranjos de DNA (*Microarray*) para a análise transcriptômica. Embora ainda seja uma técnica válida e consolidada, acredita-se que abordagens baseadas em RNA-Seq poderiam fornecer resultados mais detalhados e com maior sensibilidade. No entanto, a implementação do RNA-Seq, considerando os mesmos 21 biomateriais e associando a análise ao uso de uma linhagem celular mais representativa, implicaria em um investimento estimado em cerca de 70 mil reais, o que inviabilizou sua aplicação neste estudo. Ainda assim, essa alternativa configura uma perspectiva promissora para investigações futuras.

Por fim, a apresentação do aplicativo *Shiny* ressaltou a importância da interatividade na exposição dos resultados, permitindo aos usuários explorar os dados de forma dinâmica e acessível, o que contribui para a otimização de futuras pesquisas na área. As metodologias discutidas neste capítulo, portanto, configuram um arcabouço robusto para a análise de dados moleculares, com potencial significativo para aplicações práticas, especialmente no setor produtivo voltado aos biomateriais. Ainda que os resultados obtidos indiquem a aplicabilidade do produto desenvolvido, destaca-se a necessidade de ampliação do conjunto de dados utilizado, com o objetivo de melhorar ainda mais os modelos e as análises realizadas.

5 CONCLUSÃO GERAL

Podemos concluir que os processos selecionados, que incluem o cálculo dos *scores* de enriquecimento, a aplicação de métodos de clusterização, a visualização dos dados por meio de mapas de calor e a geração de dados sintéticos para a avaliação do modelo, mostraram-se eficazes e adequados. Esses elementos forneceram uma base sólida para a construção da primeira versão do *ShinyApp: OsteoCLUST*. Além disso, a integração dessas metodologias não só otimizou a análise de dados, mas também garantiu uma apresentação visual intuitiva e acessível, facilitando a interpretação dos resultados e a tomada de decisões informadas pelos usuários. Essa abordagem multifacetada reforça a robustez do aplicativo proposto e sua utilidade em pesquisas relacionadas ao setor produtivo na área de biomateriais.

REFERÊNCIAS

- Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4), 433–459.
- Abukawa, H., Papadaki, M., Abulikemu, M., Leaf, J., Vacanti, J. P., Kaban, L. B., & Troulis, M. J. (2006). The Engineering of Craniofacial Tissues in the Laboratory: A Review of Biomaterials for Scaffolds and Implant Coatings. *Dental Clinics of North America*, 50(2), 205–216. <https://doi.org/10.1016/j.cden.2005.11.006>
- Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., & Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, 96(12), 6745–6750. <https://doi.org/10.1073/pnas.96.12.6745>
- Angermueller, C., Pärnamaa, T., Parts, L., & Stegle, O. (2016). Deep learning for computational biology. *Molecular Systems Biology*, 12(7), 878. <https://doi.org/10.15252/msb.20156651>
- Anselme, K. (2000). Osteoblast adhesion on biomaterials. *Biomaterials*, 21(7), 667–681. [https://doi.org/10.1016/S0142-9612\(99\)00242-2](https://doi.org/10.1016/S0142-9612(99)00242-2)
- Banks, W. J. (1992). Tecidos de sustentação-osso. Em *Histologia veterinária aplicada* (2º ed, p. 137–166). São Paulo: Ed. Manole.
- Barbie, D. A., Tamayo, P., Boehm, J. S., & et al. (2009). Systematic requirement for the mitogen-activated protein kinase pathway in the growth of human cancer cell lines. *Journal of Clinical Investigation*, 119(8), 2646–2656.
- Barter, M. J., Turner, D. A., Rice, S. J., Hines, M., Lin, H., Falconer, A. M. D., McDonnell, E., Soul, J., Arques, M. del C., Europe-Finner, G. N., Rowan, A. D., Young, D. A., & Wilkinson, D. J. (2024). SERPINA3 is a marker of cartilage differentiation and is essential for the expression of extracellular matrix genes during early chondrogenesis. *Matrix Biology*, 133, 33–42. <https://doi.org/10.1016/j.matbio.2024.07.004>
- Betti, L. V. (2004). *Análises microscópica e radiográfica do reparo de defeitos confeccionados em fêmures de coelhos preenchidos com matriz óssea bovina*

- medular em bloco ou cortical em microgrânulos* [Universidade de São Paulo].
<https://doi.org/10.11606/T.25.2004.tde-21032005-104116>
- Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer Science+Business Media.
- Brynjolfsson, E., & McAfee, A. (2017). The Business of Artificial Intelligence—What It Can and Cannot Do for Your Organization. *Harvard Business Review*, 1–20.
- C. Chow, L. (2009). Next generation calcium phosphate-based biomaterials. *Dental Materials Journal*, 28(1), 1–10. <https://doi.org/10.4012/dmj.28.1>
- Cao, Y., & Wang, B. (2009). Biodegradation of Silk Biomaterials. *International Journal of Molecular Sciences*, 10(4), 1514–1524. <https://doi.org/10.3390/ijms10041514>
- Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2022). *shiny: Web Application Framework for R*. R package version 1.7.3. <https://CRAN.R-project.org/package=shiny>
- Chen, S. H., Jakeman, A. J., & Norton, J. P. (2008). Artificial Intelligence techniques: An introduction to their use for modelling environmental systems. *Mathematics and Computers in Simulation*, 78(2–3), 379–400. <https://doi.org/10.1016/j.matcom.2008.01.028>
- Claus O. Wilke. (2019). *Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures* (1st Edition). O'Reilly Media, Inc.
- Croci Tesconi, A., Camargo Pires de, O., Bitar, G., Pereira Borges, S. L., Moreira, M., & Freitas, S. de. (2003). Efeito do Concentrado de plasma em falhas ósseas provocadas em fêmures de camundongos como estimulação de formação óssea. Estudo experimental. *Acta Ortopédica Brasileira*, 230–239.
- Czekanska, E., Stoddart, M., Richards, R., & Hayes, J. (2012). In search of an osteoblast cell model for in vitro research. *European Cells and Materials*, 24, 1–17. <https://doi.org/10.22203/eCM.v024a01>
- Dahmen, J., & Cook, D. (2019). SynSys: A Synthetic Data Generation System for Healthcare Applications. *Sensors*, 19(5), 1181. <https://doi.org/10.3390/s19051181>

- de Vries, A., & Ripley, B. D. (2022). *ggdendro: Create Dendrograms and Tree Diagrams Using “ggplot2”*. <https://CRAN.R-project.org/package=ggdendro>
- Dempster, D. W., & Zhou, H. (1999). New concepts in bone remodeling. Em *Dynamics of bone and cartilage metabolism* (p. 261–273). San Diego: Academic Press.
- Dorozhkin, S. V. (2010). Bioceramics of calcium orthophosphates. *Biomaterials*, 31(7), 1465–1485. <https://doi.org/10.1016/j.biomaterials.2009.11.050>
- Dr. Harris, S. A., Enger, R. J., Riggs, L. B., & Spelsberg, T. C. (1995). Development and characterization of a conditionally immortalized human fetal osteoblastic cell line. *Journal of Bone and Mineral Research*, 10(2), 178–186. <https://doi.org/10.1002/jbmr.5650100203>
- Dyce, K. M., Sack, W. O., & Wensing, C. J. G. (2004). *Tratado de anatomia Veterinária* (3º ed). Rio de Janeiro: Guanabara Koogan.
- Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25), 14863–14868. <https://doi.org/10.1073/pnas.95.25.14863>
- el Emam, K., Mosquera, L., & Hoptroff, R. (2020). *Practical Synthetic Data Generation*. O'Reilly Media, Inc.
- Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster Analysis*. Wiley.
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874. <https://doi.org/10.1016/j.patrec.2005.10.010>
- Ferreira, M. R., Milani, R., Rangel, E. C., Peppelenbosch, M., & Zambuzzi, W. (2020). OsteoBLAST: Computational Routine of Global Molecular Analysis Applied to Biomaterials Development. *Frontiers in Bioengineering and Biotechnology*, 8. <https://doi.org/10.3389/fbioe.2020.565901>
- Ferreira, M. R., Santos, G. A., Biagi, C. A., Silva Junior, W. A., & Zambuzzi, W. F. (2021). GSVA score reveals molecular signatures from transcriptomes for biomaterials comparison. *Journal of Biomedical Materials Research - Part A*, 109(6), 1004–1014. <https://doi.org/10.1002/jbm.a.37090>

- Geetha, M., Singh, A. K., Asokamani, R., & Gogia, A. K. (2009). Ti based biomaterials, the ultimate choice for orthopaedic implants – A review. *Progress in Materials Science*, 54(3), 397–425. <https://doi.org/10.1016/j.pmatsci.2008.06.004>
- Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2nd Edition). O'Reilly Media, Inc.
- Giannoudis, P. v., Dinopoulos, H., & Tsiridis, E. (2005). Bone substitutes: An update. *Injury*, 36(3), S20–S27. <https://doi.org/10.1016/j.injury.2005.07.029>
- Gil, M. H., & Ferreira, P. (2006). Polissacarídeos como biomateriais. *Boletim Sociedade Portuguesa de Química*, 72–74.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. The MIT Press.
- Grand View Research. (2022). *Biomaterials Market Size, Share & Trends Analysis Report By Product (Natural, Metallic, Polymer), By Application (Cardiovascular, Orthopedics, Plastic Surgery), By Region, And Segment Forecasts, 2022 - 2030*. <https://www.grandviewresearch.com/industry-analysis/biomaterials-industry>
- Gu, Z., Eils, R., & Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Em *Bioinformatics*.
- Han, J., & Pei, J. (2012). *Data Mining*. Elsevier. <https://doi.org/10.1016/C2009-0-61819-5>
- Hänzelmann, S., Castelo, R., & Guinney, J. (2013a). GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics*, 14(1), 7. <https://doi.org/10.1186/1471-2105-14-7>
- Hänzelmann, S., Castelo, R., & Guinney, J. (2013b). GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics*, 14(1), 7. <https://doi.org/10.1186/1471-2105-14-7>
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. *Springer*.
- Hench, L. L. (1980). Biomaterials. *Science*, 208(4446), 826–831. <https://doi.org/10.1126/science.6246576>

- Hench, L. L., & Polak, J. M. (2002). Third-Generation Biomedical Materials. *Science*, 295(5557), 1014–1017. <https://doi.org/10.1126/science.1067404>
- Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K. D., Irizarry, R. A., Lawrence, M., Love, M. I., MacDonald, J., Obenchain, V., Oleš, A. K., ... Morgan, M. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, 12(2), 115–121. <https://doi.org/10.1038/nmeth.3252>
- Igor Dolgalev. (2022). *msigdbr: MSigDB Gene Sets for Multiple Organisms in a Tidy Data Format* (R package version 7.5.1). <https://CRAN.R-project.org/package=msigdbr>.
- Izenman, A. J. (2008). Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning. *Springer*.
- Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), 651–666.
- Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets*, 31(3), 685–695. <https://doi.org/10.1007/s12525-021-00475-2>
- Jolliffe, I. T. (2002). Principal Component Analysis. *Springer*, (2nd ed.).
- Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065).
- Junqueira, L. C., & Carneiro, J. (2004). *Histologia Básica* (10^o ed). Rio de Janeiro: Guanabara Koogan.
- Kao, S. T., & Scott, D. D. (2007). A Review of Bone Substitutes. *Oral and Maxillofacial Surgery Clinics of North America*, 19(4), 513–521. <https://doi.org/10.1016/j.coms.2007.06.002>
- Khatrri, P., Sirota, M., & Butte, A. J. (2012). Ten years of pathway analysis: Current approaches and outstanding challenges. Em *PLoS Computational Biology* (Vol. 8, Número 2). Public Library of Science. <https://doi.org/10.1371/journal.pcbi.1002375>
- Korotkevich, G., Sukhov, V., & Sergushichev, A. (2021). Fast gene set enrichment analysis. *bioRxiv*.

- Kuhn, M. (2022). *tune: Tidy Tuning Tools*. <https://CRAN.R-project.org/package=tune>
- Kuhn, M., & Vaughan, D. (2022). *parsnip: A Common API to Modeling and Analysis Functions*. <https://CRAN.R-project.org/package=parsnip>
- Kuhn, M., & Wickham, H. (2020). *Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles*. <https://www.tidymodels.org>
- LeGeros, R. Z. (2002). Properties of Osteoconductive Biomaterials: Calcium Phosphates. *Clinical Orthopaedics and Related Research*, 395, 81–98. <https://doi.org/10.1097/00003086-200202000-00009>
- Lever, J., Krzywinski, M., & Altman, N. (2017). Principal component analysis. *Nature Methods*, 14(7), 641–642.
- Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., & Tamayo, P. (2015). The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Systems*, 1(6), 417–425. <https://doi.org/10.1016/j.cels.2015.12.004>
- Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdottir, H., Tamayo, P., & Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, 27(12), 1739–1740. <https://doi.org/10.1093/bioinformatics/btr260>
- Lirani, A. P. R. (2004). *Estudo comparativo dos efeitos do ultra-som e do laser de baixa intensidade no reparo ósseo de tibia de rato* [Dissertação (Mestrado)]. Universidade de São Paulo.
- Long, M., & Rack, H. J. (1998). Titanium alloys in total joint replacement—a materials science perspective. *Biomaterials*, 19(18), 1621–1639. [https://doi.org/10.1016/S0142-9612\(97\)00146-4](https://doi.org/10.1016/S0142-9612(97)00146-4)
- Lu, X., & Leng, Y. (2003). Quantitative analysis of osteoblast behavior on microgrooved hydroxyapatite and titanium substrata. *Journal of Biomedical Materials Research Part A*, 66A(3), 677–687. <https://doi.org/10.1002/jbm.a.10022>
- McAllister, B. S., & Haghighat, K. (2007). Bone Augmentation Techniques. *Journal of Periodontology*, 78(3), 377–396. <https://doi.org/10.1902/jop.2007.060048>

- Middleton, J. C., & Tipton, A. J. (2000). Synthetic biodegradable polymers as orthopedic devices. *Biomaterials*, 21(23), 2335–2346. [https://doi.org/10.1016/S0142-9612\(00\)00101-0](https://doi.org/10.1016/S0142-9612(00)00101-0)
- Murali, A., Brokesh, A. M., Cross, L. M., Kersey, A. L., Jaiswal, M. K., Singh, I., & Gaharwar, A. (2024). Inorganic Biomaterials Shape the Transcriptome Profile to Induce Endochondral Differentiation. *Advanced Science*, 11(29). <https://doi.org/10.1002/adv.202402468>
- Murphy, K. P. (2022). Probabilistic Machine Learning: An Introduction. *MIT Press*.
- Murtagh, F., & Legendre, P. (2014). Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *Journal of Classification*, 31(3), 274–295.
- Nowok, B., Raab, G. M., & Dibben, C. (2016). synthpop : Bespoke Creation of Synthetic Data in R. *Journal of Statistical Software*, 74(11). <https://doi.org/10.18637/jss.v074.i11>
- Oliveira, S. M., Mijares, D. Q., Turner, G., Amaral, I. F., Barbosa, M. A., & Teixeira, C. C. (2009). Engineering Endochondral Bone: *In Vivo* Studies. *Tissue Engineering Part A*, 15(3), 635–643. <https://doi.org/10.1089/ten.tea.2008.0052>
- Pages, H., Carlson, M., Falcon, S., & Li, N. (2021). *AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor*. <https://bioconductor.org/packages/AnnotationDbi>
- Park, J., & Lakes, R. S. (2017). *Biomaterials: An Introduction* (3^o ed). Springer Science+Business Media.
- Pelissier, P. H., Masquelet, A. C., Bareille, R., Pelissier, S. M., & Amedee, J. (2004). Induced membranes secrete growth factors including vascular and osteoinductive factors and could stimulate bone regeneration. *Journal of Orthopaedic Research*, 22(1), 73–79. [https://doi.org/10.1016/S0736-0266\(03\)00165-7](https://doi.org/10.1016/S0736-0266(03)00165-7)
- Pereira, A. P. v., Vasconcelos, W. L., & Oréfice, R. L. (1999). Novos biomateriais: híbridos orgânico-inorgânicos bioativos. *Polímeros*, 9(4), 104–109. <https://doi.org/10.1590/S0104-14281999000400018>

- Powers, D. M. W. (2020). *Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation*.
- Precheur, H. v. (2007). Bone Graft Materials. *Dental Clinics of North America*, 51(3), 729–746. <https://doi.org/10.1016/j.cden.2007.03.004>
- R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015a). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7), e47–e47. <https://doi.org/10.1093/nar/gkv007>
- Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015b). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7), e47–e47. <https://doi.org/10.1093/nar/gkv007>
- RStudio Team. (2022). *RStudio: Integrated Development Environment for R*. RStudio, PBC. <http://www.rstudio.com/>
- Russell, S., & Norvig, P. (2021). *Artificial Intelligence: A Modern Approach* (4th edition). Pearson.
- Shlens, J. (2014). A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*.
- Smeets, R., Kolk, A., Gerressen, M., Driemel, O., Maciejewski, O., Hermanns-Sachweh, B., Riediger, D., & Stein, J. M. (2009). A new biphasic osteoinductive calcium composite material with a negative Zeta potential for bone augmentation. *Head & Face Medicine*, 5(1), 13. <https://doi.org/10.1186/1746-160X-5-13>
- Sohn, S.-H., Jun, H.-K., Kim, C.-S., Kim, K.-N., Chung, S.-M., Shin, S.-W., Ryu, J.-J., & Kim, M.-K. (2006). Biological responses in osteoblast-like cell line according to thin layer hydroxyapatite coatings on anodized titanium. *Journal of Oral Rehabilitation*, 33(12), 898–911. <https://doi.org/10.1111/j.1365-2842.2006.01643.x>
- Song, Y., Cheon, S. A., Lee, K. E., Lee, S.-Y., Lee, B.-K., Oh, D.-B., Kang, H. A., & Kim, J.-Y. (2008). Role of the RAM Network in Cell Polarity and Hyphal Morphogenesis in *Candida albicans*. *Molecular Biology of the Cell*, 19(12), 5456–5477. <https://doi.org/10.1091/mbc.e08-03-0272>

- Sousa, V. L. (2003). *Efeitos do ultra som de baixa intensidade sobre a consolidação óssea em fratura de ossos longos (rádio e ulna, tíbia e fíbula) em cães (Canis familiaris)* [Dissertação (Mestrado)]. Universidade de São Paulo.
- Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005a). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43), 15545–15550. <https://doi.org/10.1073/pnas.0506580102>
- Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005b). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43), 15545–15550. <https://doi.org/10.1073/pnas.0506580102>
- Tabata, Y. (2009). Biomaterial technology for tissue engineering applications. *Journal of The Royal Society Interface*, 6(suppl_3). <https://doi.org/10.1098/rsif.2008.0448.focus>
- Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S., & Golub, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences*, 96(6), 2907–2912. <https://doi.org/10.1073/pnas.96.6.2907>
- Tang, Y., Rajendran, P., Veeraraghavan, V. P., Hussain, S., Balakrishna, J. P., Chinnathambi, A., Alharbi, S. A., Alahmadi, T. A., Rengarajan, T., & Mohan, S. K. (2021). Osteogenic differentiation and mineralization potential of zinc oxide nanoparticles from *Scutellaria baicalensis* on human osteoblast-like MG-63 cells. *Materials Science and Engineering: C*, 119, 111656. <https://doi.org/10.1016/j.msec.2020.111656>
- Tannenbaum, J., & Bennett, B. T. (2015). Russell and Burch's 3Rs then and now: the need for clarity in definition and purpose. *Journal of the American Association for Laboratory Animal Science : JAALAS*, 54(2), 120–132.

- Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., & Church, G. M. (1999). Systematic determination of genetic network architecture. *Nature Genetics*, 22(3), 281–285. <https://doi.org/10.1038/10343>
- Törönen, P., Kolehmainen, M., Wong, G., & Castrén, E. (1999). Analysis of gene expression data using self-organizing maps. *FEBS Letters*, 451(2), 142–146. [https://doi.org/10.1016/S0014-5793\(99\)00524-4](https://doi.org/10.1016/S0014-5793(99)00524-4)
- von Recum, A., & Laberge, M. (1995). Educational Goals for Biomaterials Science and Engineering: Prospective View. *Journal of Applied Biomaterials*, 6, 137–144.
- Wen, X., Fuhrman, S., Michaels, G. S., Carr, D. B., Smith, S., Barker, J. L., & Somogyi, R. (1998). Large-scale temporal gene expression mapping of central nervous system development. *Proceedings of the National Academy of Sciences*, 95(1), 334–339. <https://doi.org/10.1073/pnas.95.1.334>
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., ... Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, 4(43), 1686. <https://doi.org/10.21105/joss.01686>

APÊNDICE e ANEXO

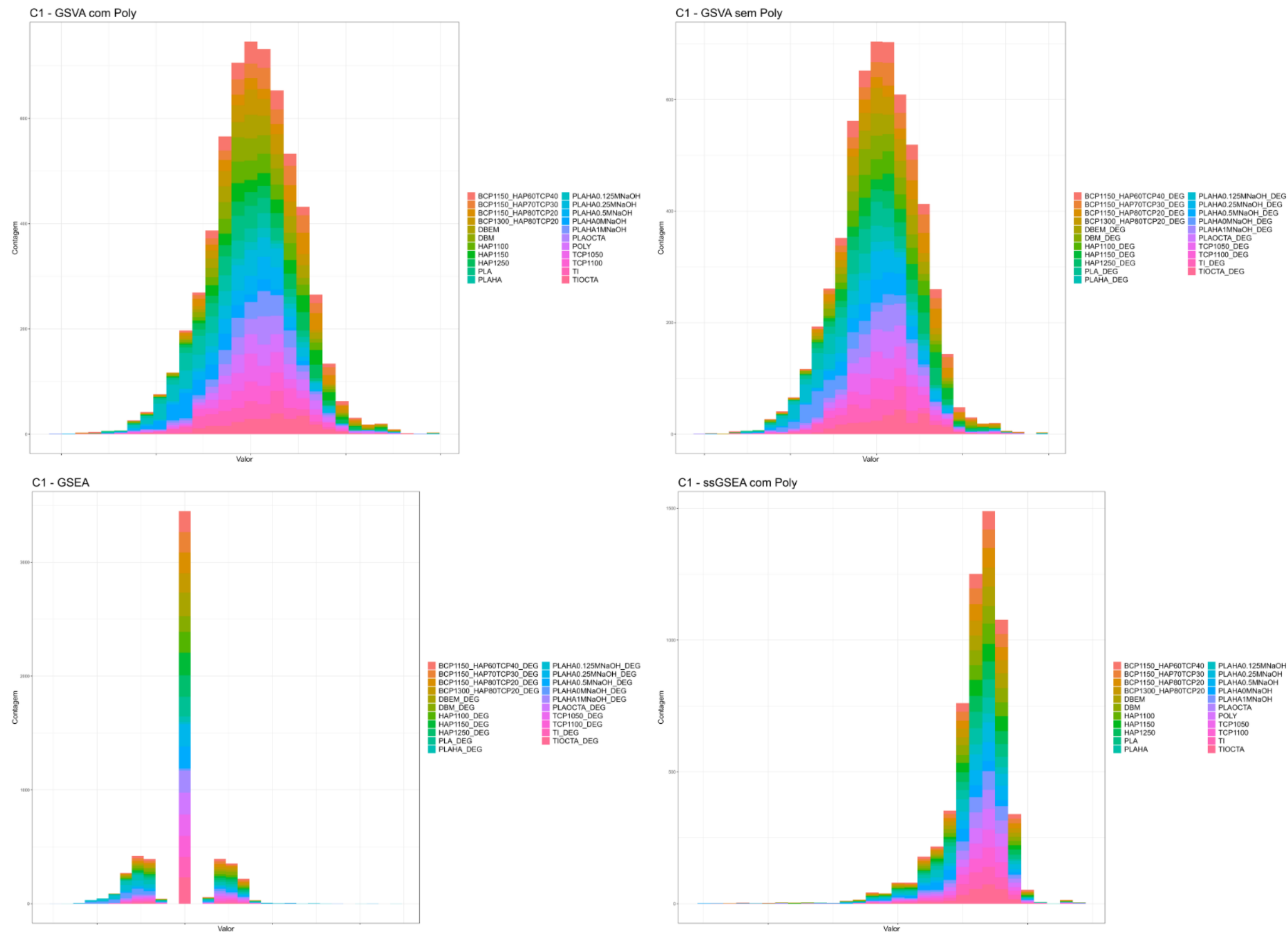


Figura Suplementar 1 - Representação por meio de histogramas, das distribuições dos scores GSVA (com e sem o grupo controle POLY), GSEA e ssGSEA, gerados para as assinaturas moleculares presentes no conjunto C1.

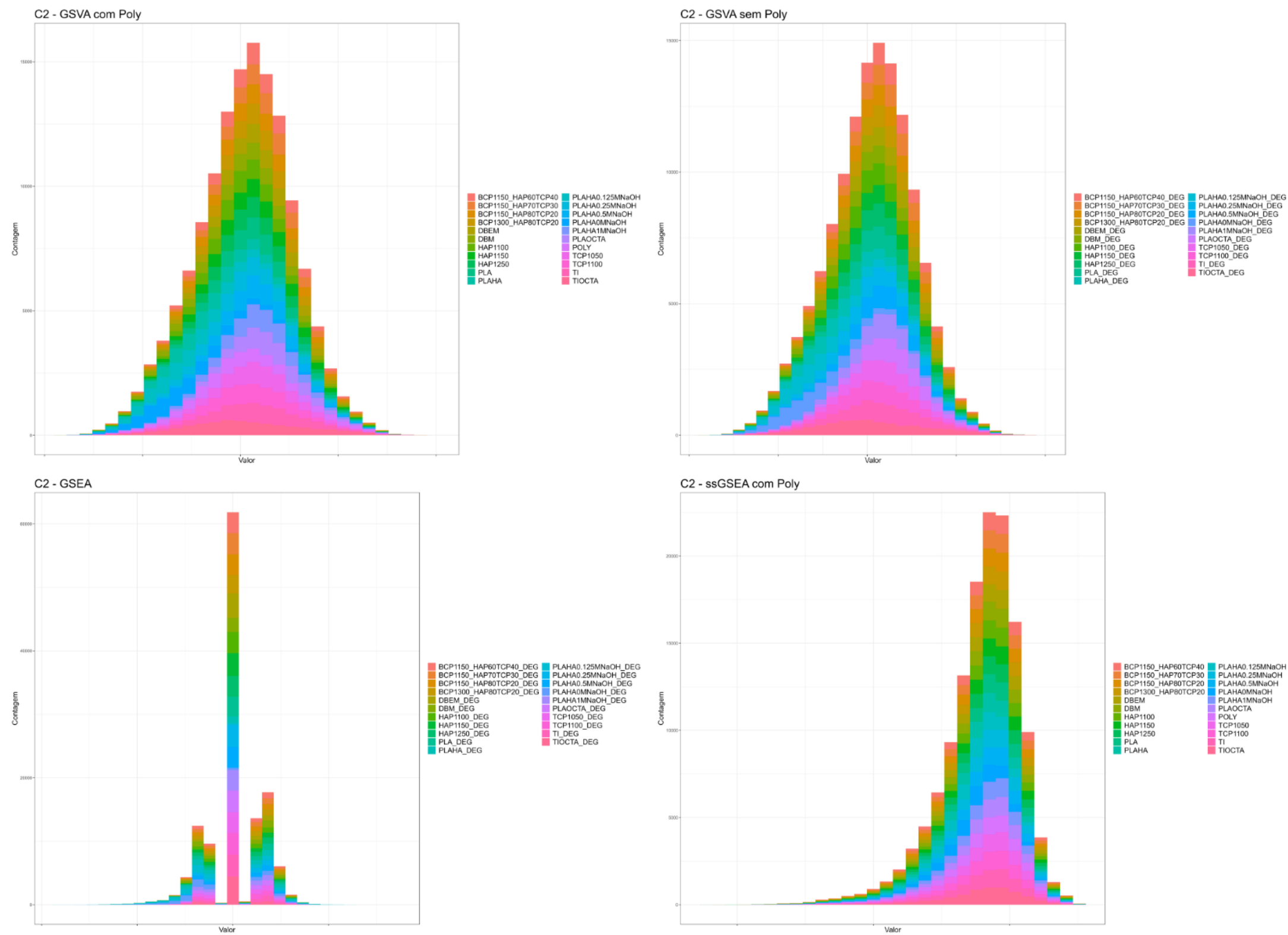


Figura Suplementar 2 - Representação por meio de histogramas, das distribuições dos scores GSVA (com e sem o grupo controle POLY), GSEA e ssGSEA, gerados para as assinaturas moleculares presentes no conjunto C2.

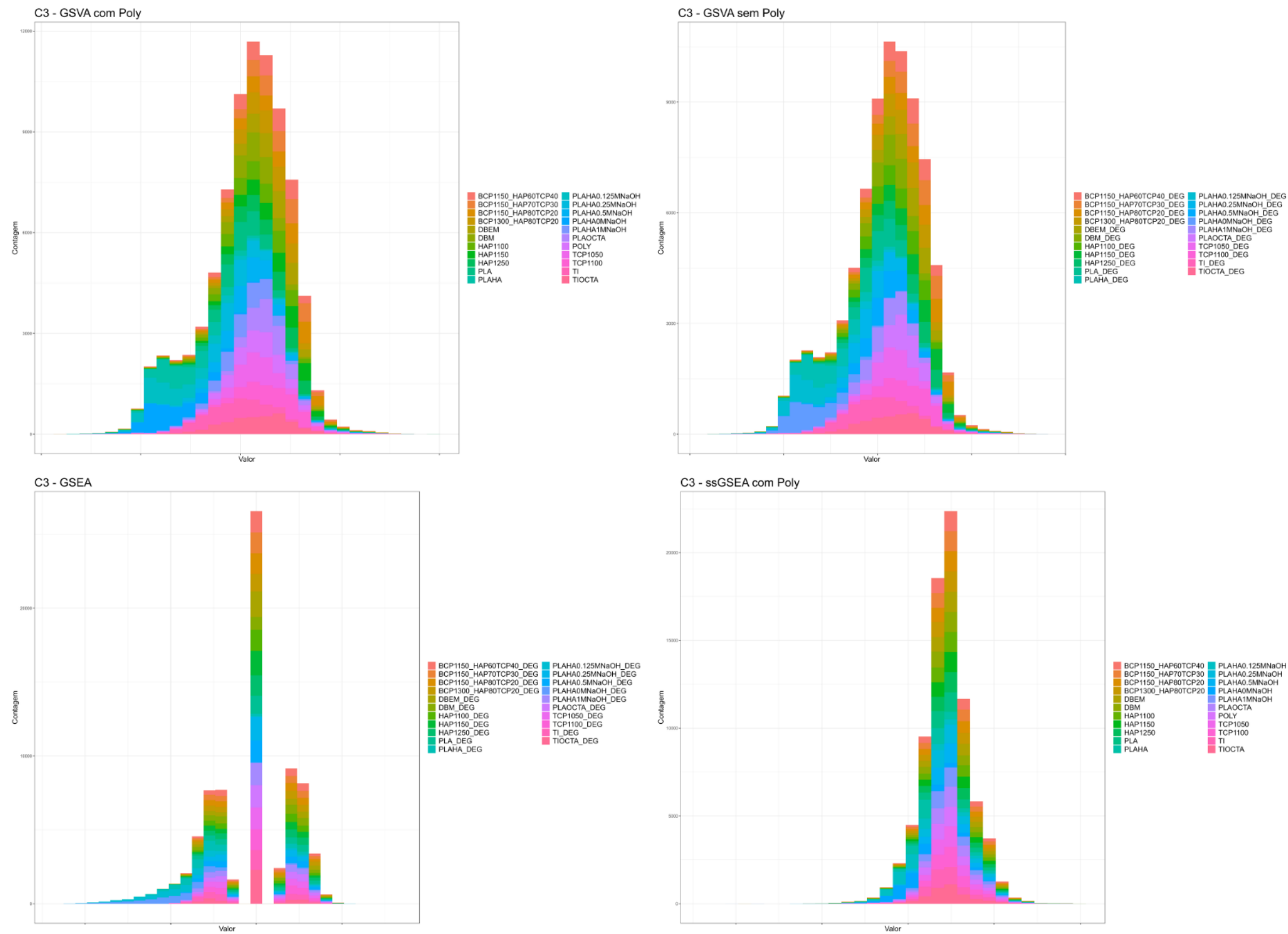


Figura Suplementar 3 - Representação por meio de histogramas, das distribuições dos scores GSVA (com e sem o grupo controle POLY), GSEA e ssGSEA, gerados para as assinaturas moleculares presentes no conjunto C3.

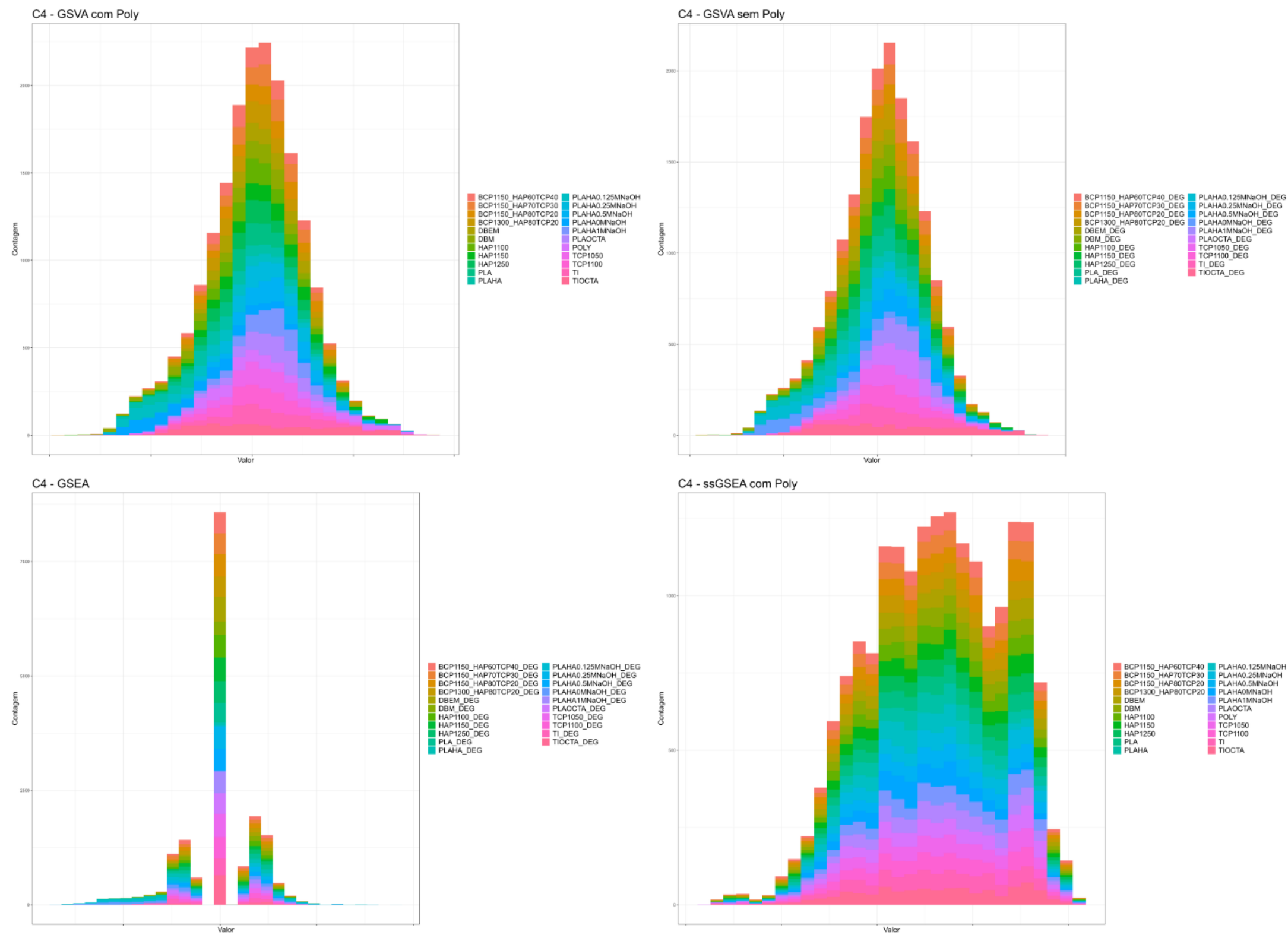


Figura Suplementar 4 - Representação por meio de histogramas, das distribuições dos scores GSVA (com e sem o grupo controle POLY), GSEA e ssGSEA, gerados para as assinaturas moleculares presentes no conjunto C4.

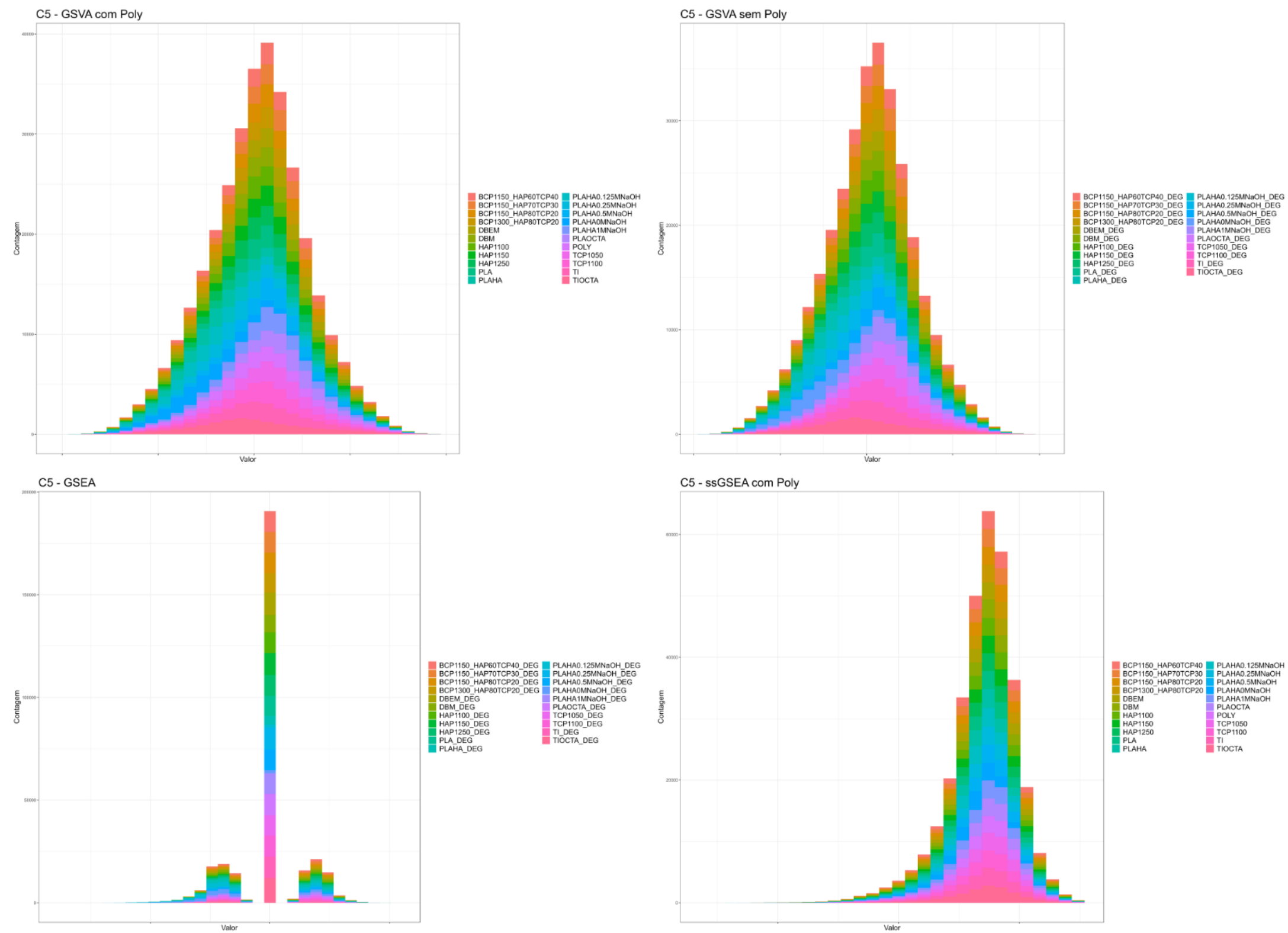


Figura Suplementar 5 - Representação por meio de histogramas, das distribuições dos scores GSVA (com e sem o grupo controle POLY), GSEA e ssGSEA, gerados para as assinaturas moleculares presentes no conjunto C5.

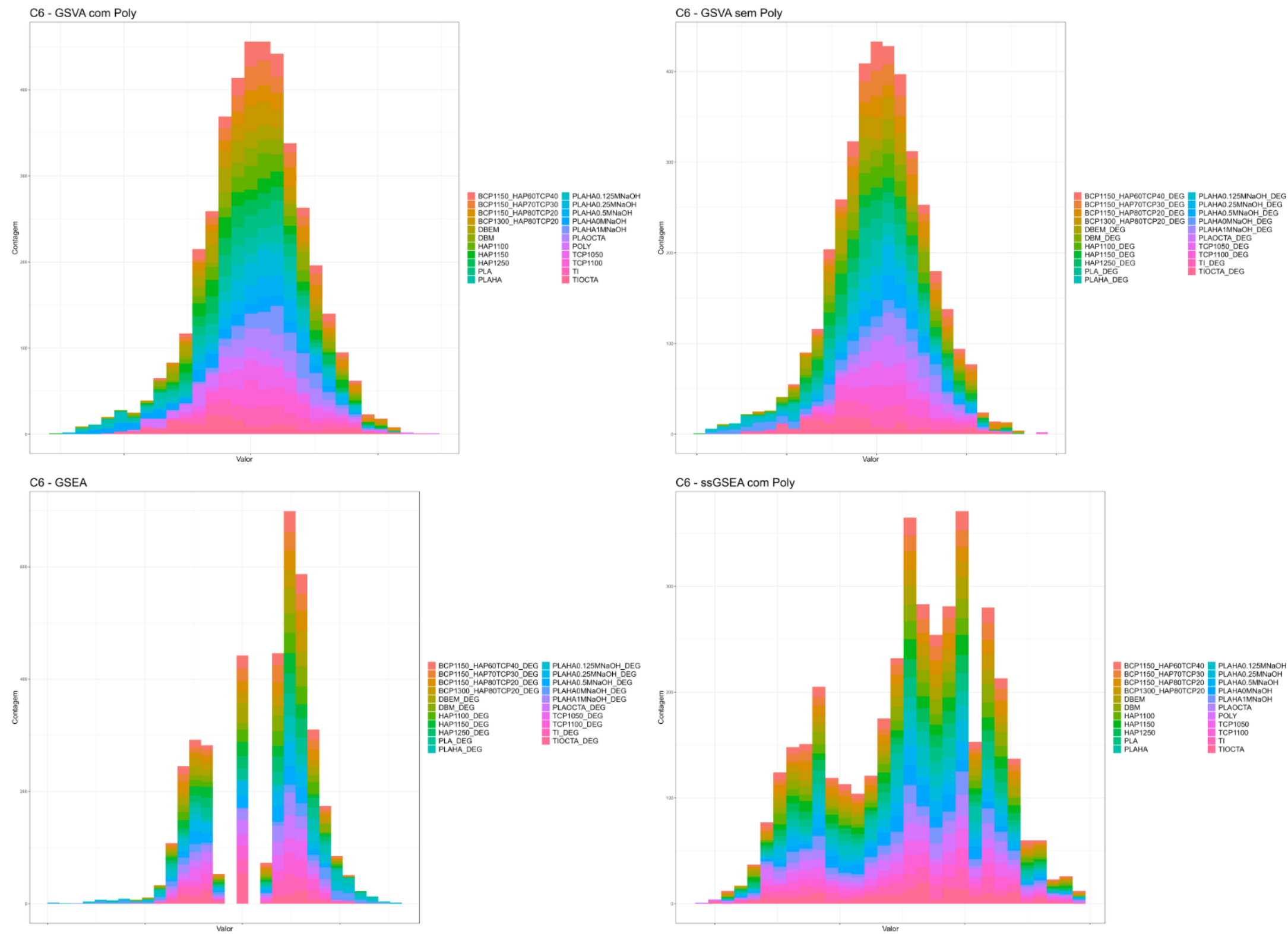


Figura Suplementar 6 - Representação por meio de histogramas, das distribuições dos scores GSVa (com e sem o grupo controle POLY), GSEA e ssGSEA, gerados para as assinaturas moleculares presentes no conjunto C6.

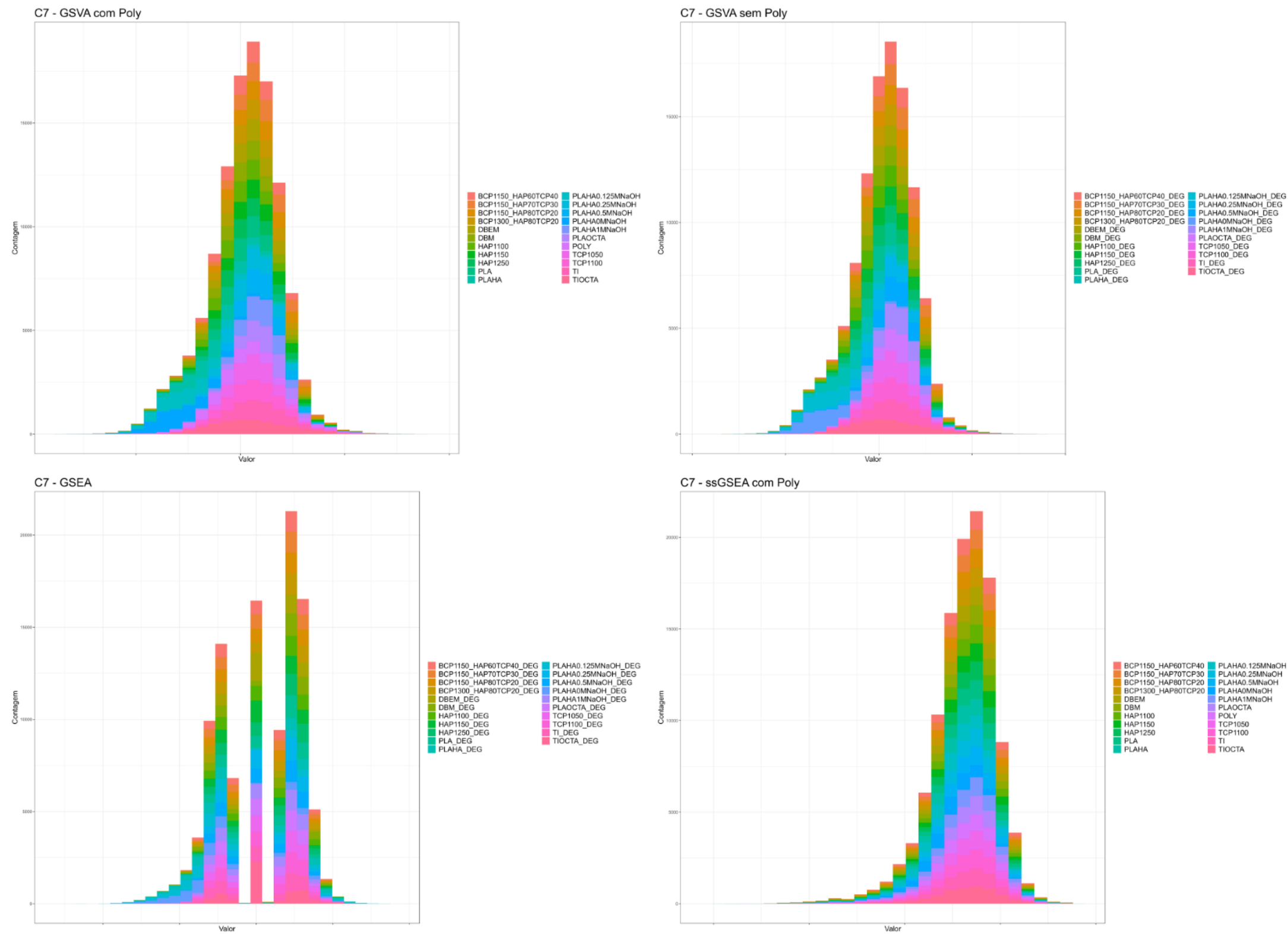


Figura Suplementar 7 - Representação por meio de histogramas, das distribuições dos scores GSVA (com e sem o grupo controle POLY), GSEA e ssGSEA, gerados para as assinaturas moleculares presentes no conjunto C7.

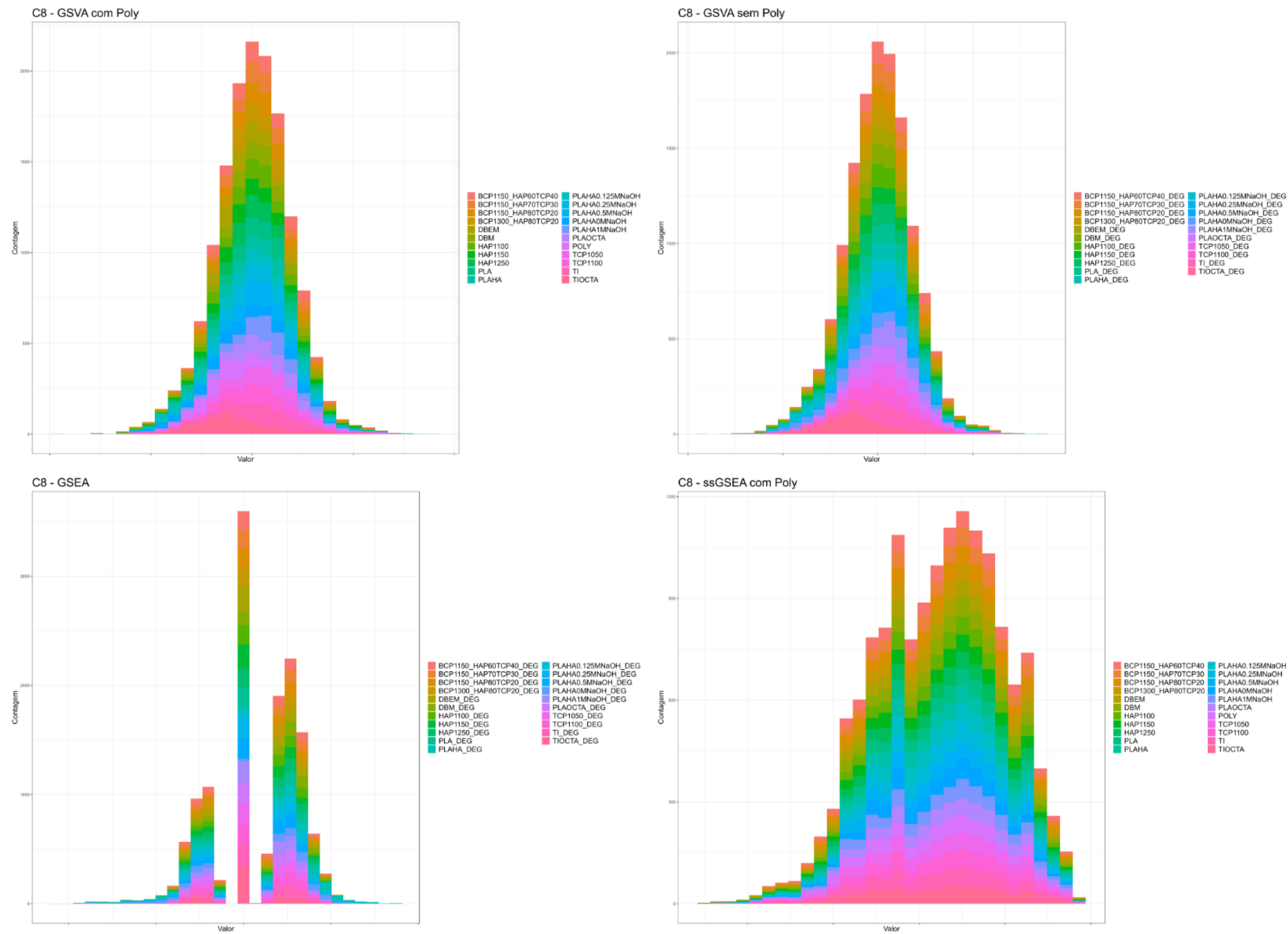


Figura Suplementar 8 - Representação por meio de histogramas, das distribuições dos scores GSV (com e sem o grupo controle POLY), GSEA e ssGSEA, gerados para as assinaturas moleculares presentes no conjunto C8.

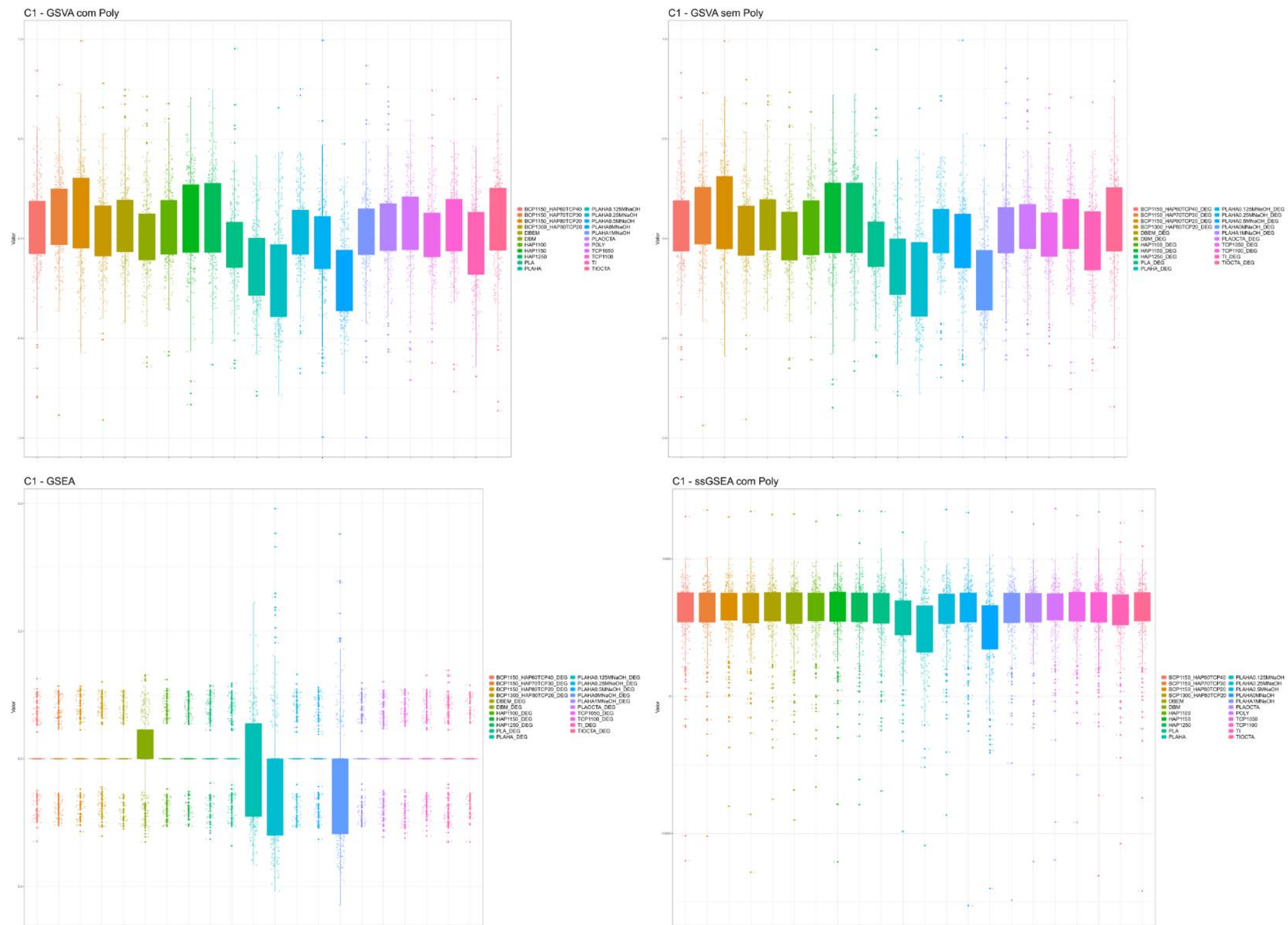


Figura Suplementar 9 - Representação por meio de boxplot, das distribuições dos scores GSVA (com e sem o grupo controle POLY), GSEA e ssGSEA, gerados para as assinaturas moleculares presentes no conjunto C1.

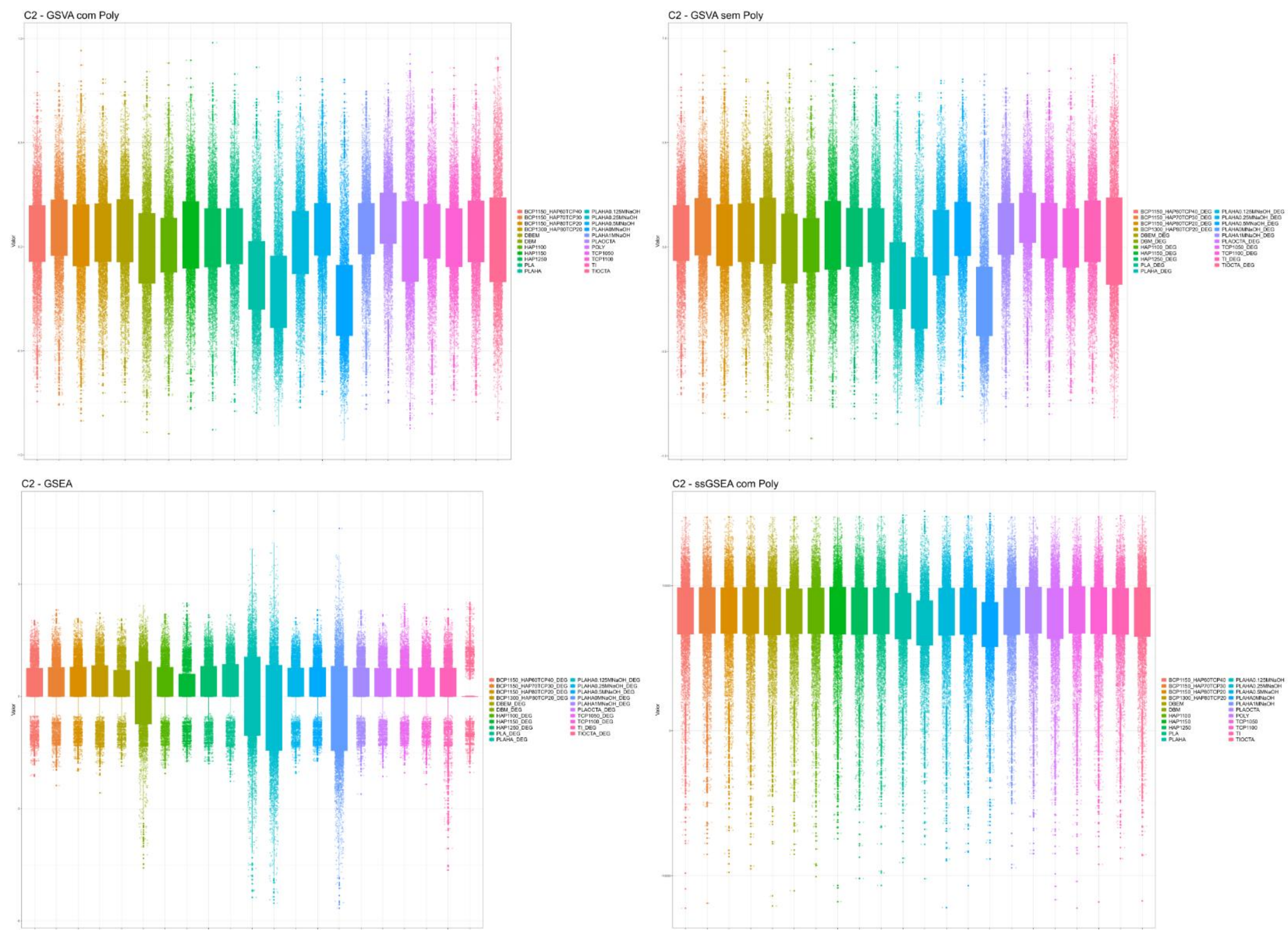


Figura Suplementar 10 - Representação por meio de boxplot, das distribuições dos scores GSVA (com e sem o grupo controle POLY), GSEA e ssGSEA, gerados para as assinaturas moleculares presentes no conjunto C2.

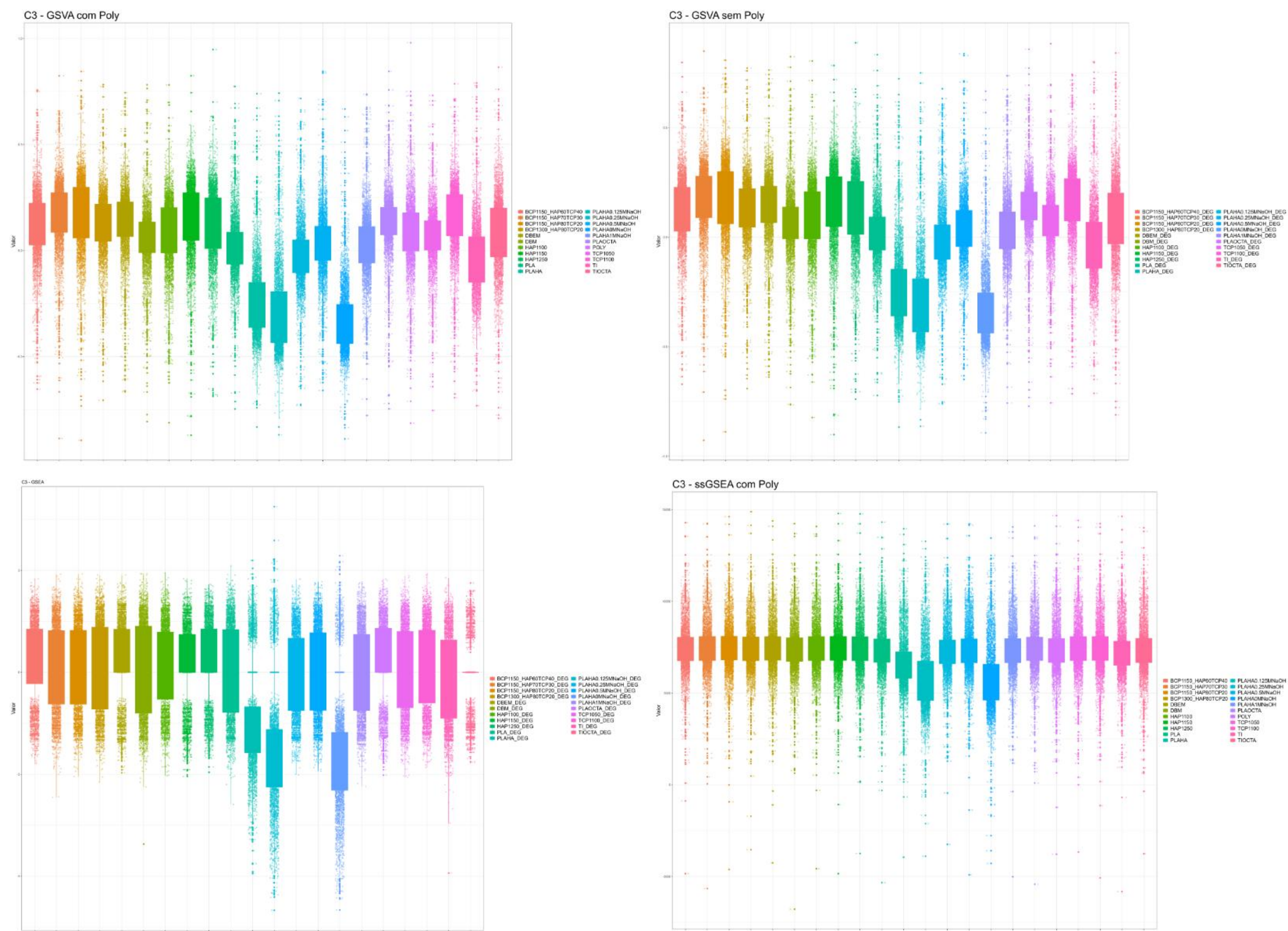


Figura Suplementar 11 - Representação por meio de boxplot, das distribuições dos scores GSVA (com e sem o grupo controle POLY), GSEA e ssGSEA, gerados para as assinaturas moleculares presentes no conjunto C3.

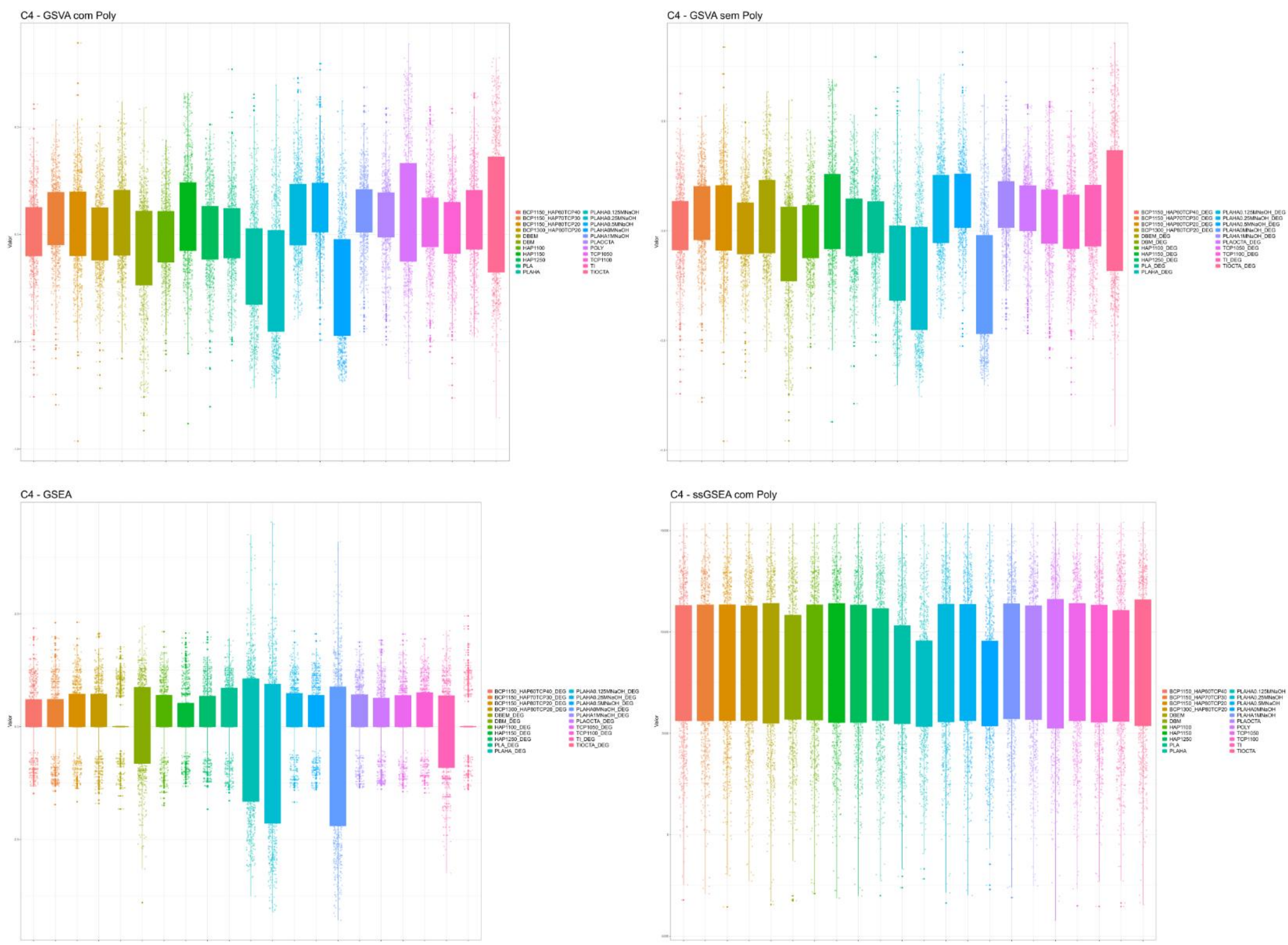


Figura Suplementar 12 - Representação por meio de boxplot, das distribuições dos scores GSVA (com e sem o grupo controle POLY), GSEA e ssGSEA, gerados para as assinaturas moleculares presentes no conjunto C4.

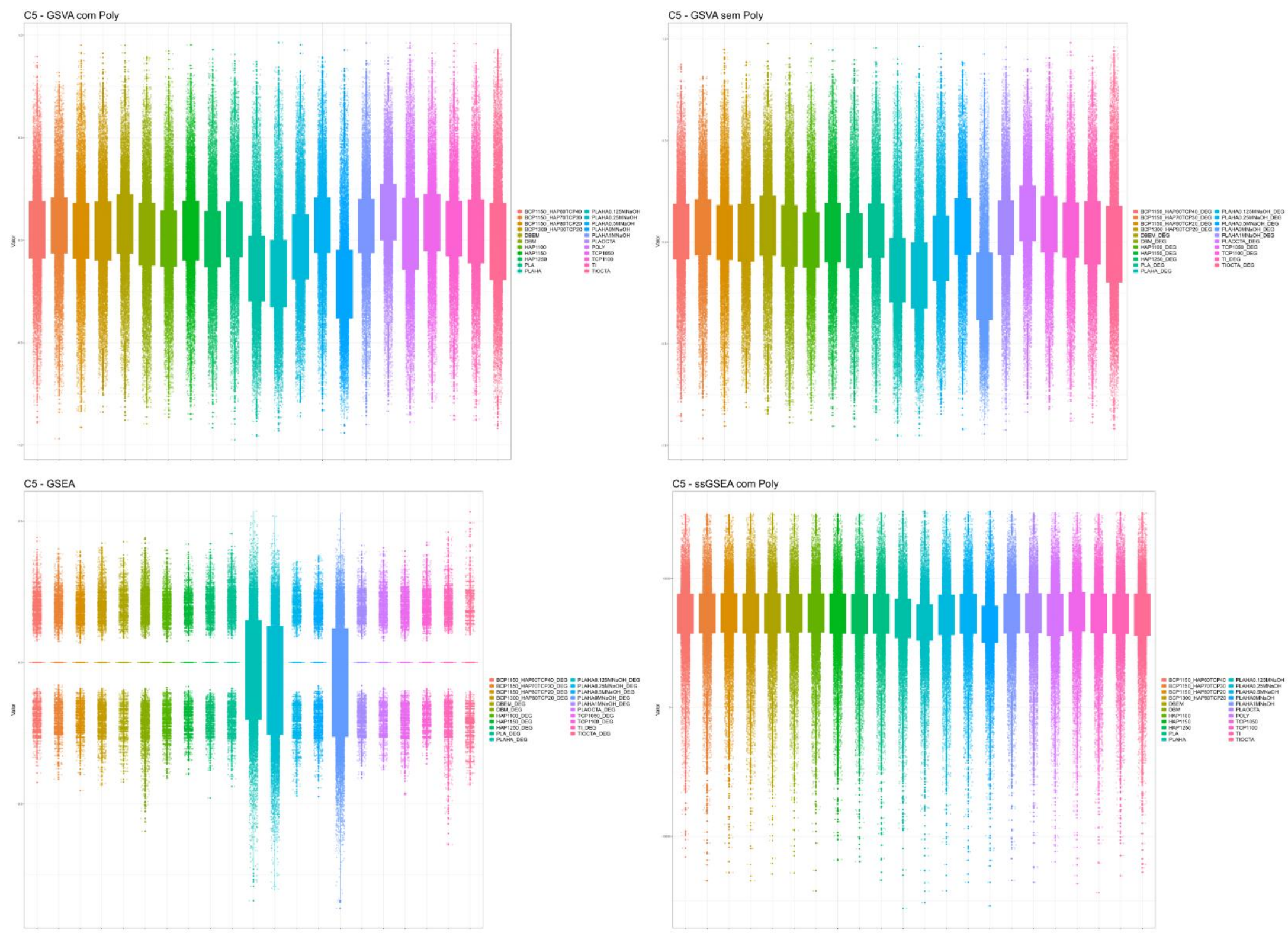


Figura Suplementar 13 - Representação por meio de boxplot, das distribuições dos scores GSVA (com e sem o grupo controle POLY), GSEA e ssGSEA, gerados para as assinaturas moleculares presentes no conjunto C5.

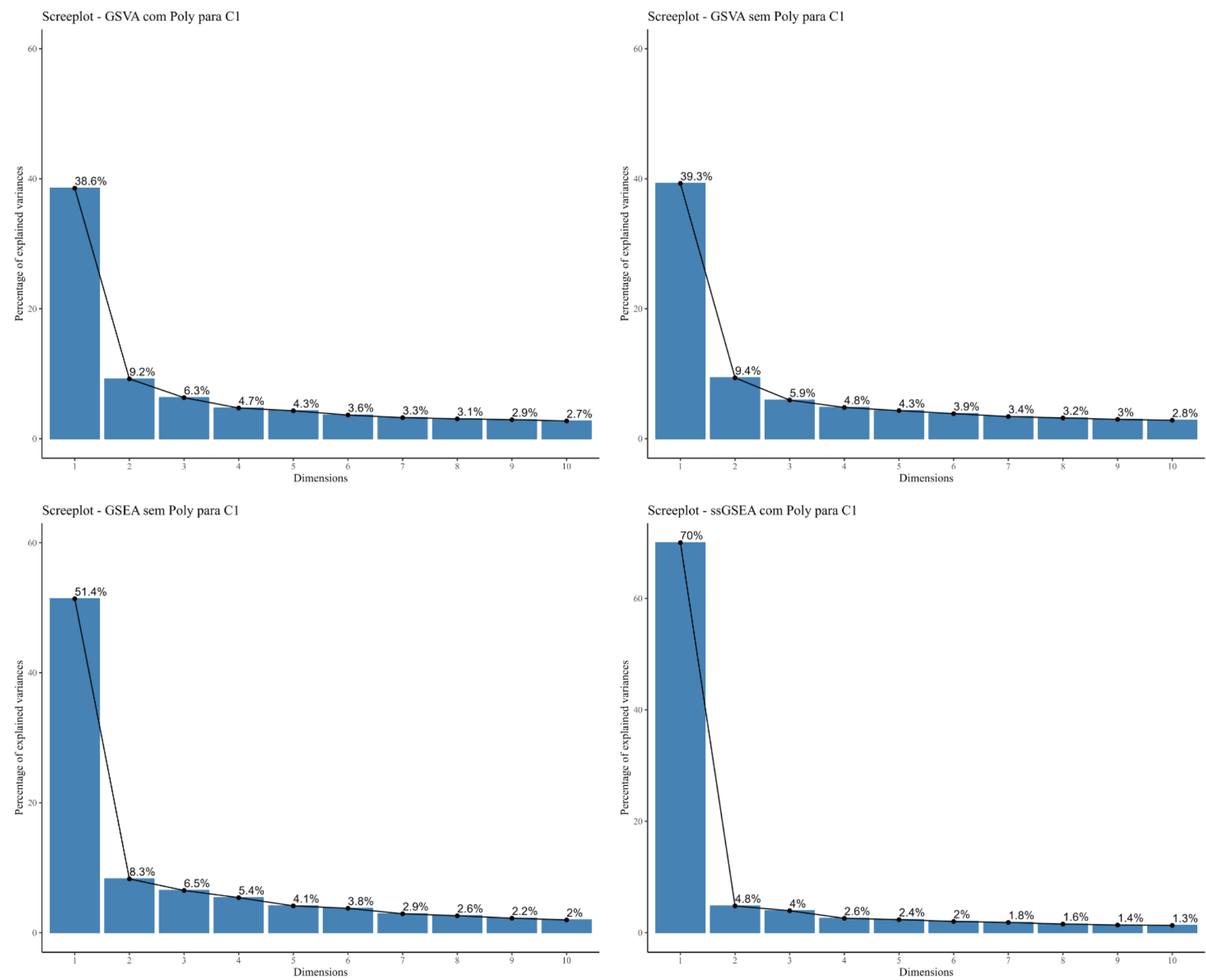


Figura Suplementar 17 - Scree plot dos 21 biomateriais analisados para o conjunto C1, com base nos scores GSVA (com POLY), GSVA (sem POLY), GSEA e ssGSEA, respectivamente.

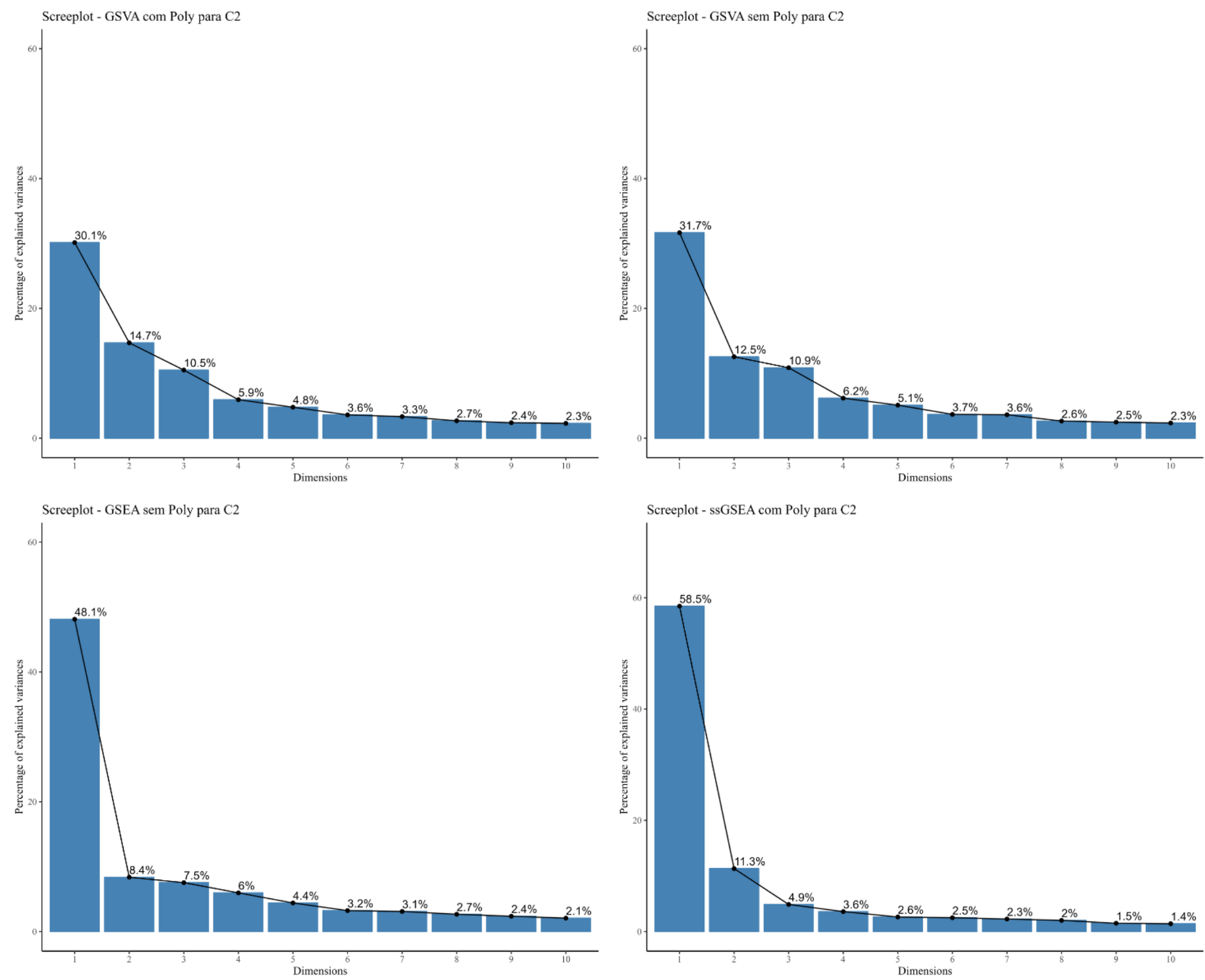


Figura Suplementar 18 - Scree plot dos 21 biomateriais analisados para o conjunto C2, com base nos scores GSVA (com POLY), GSVA (sem POLY), GSEA e ssGSEA, respectivamente.

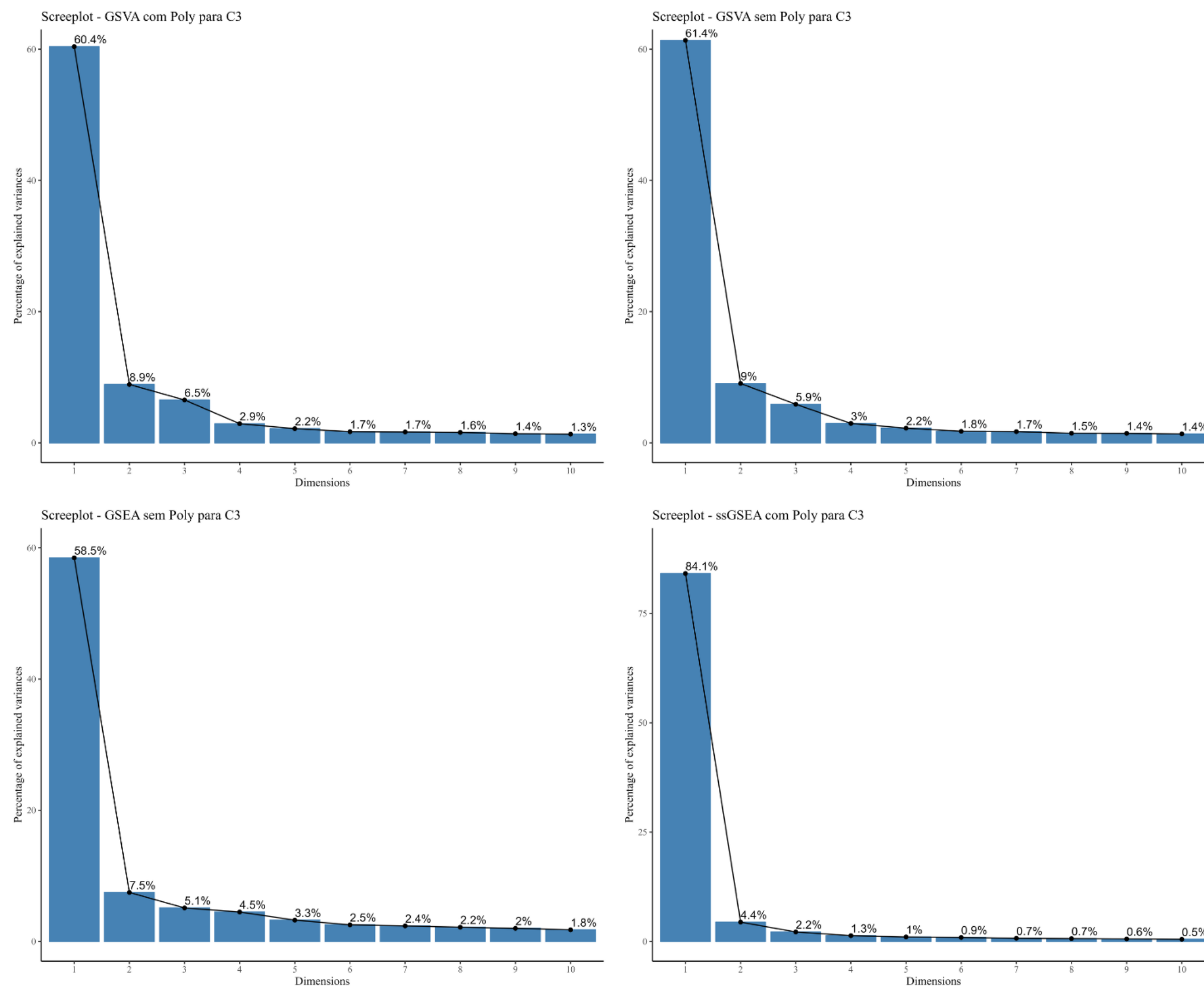


Figura Suplementar 19 - Scree plot dos 21 biomateriais analisados para o conjunto C3, com base nos scores GSVA (com POLY), GSVA (sem POLY), GSEA e ssGSEA, respectivamente.

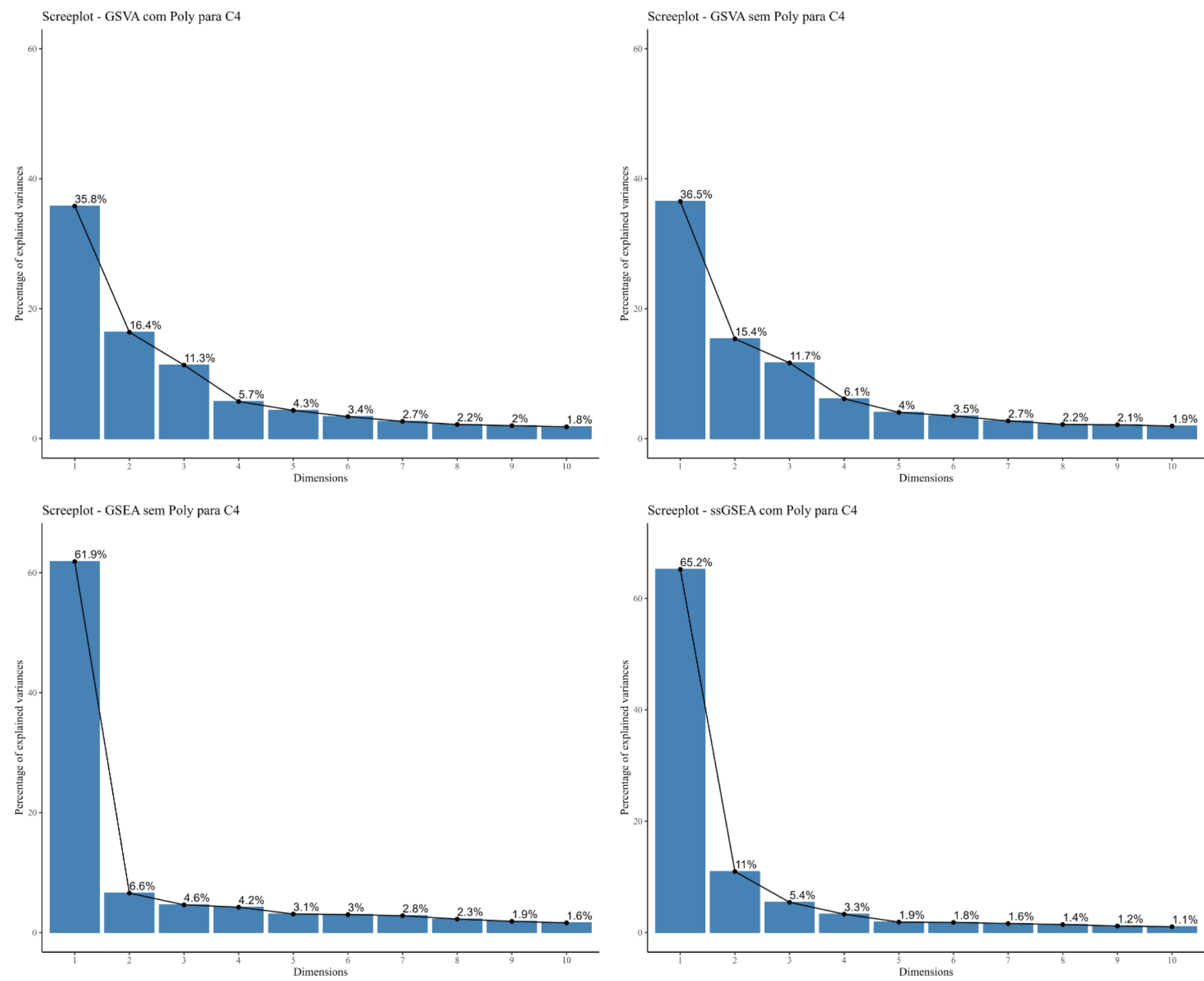


Figura Suplementar 20 - Scree plot dos 21 biomateriais analisados para o conjunto C4, com base nos scores GSVA (com POLY), GSVA (sem POLY), GSEA e ssGSEA, respectivamente.

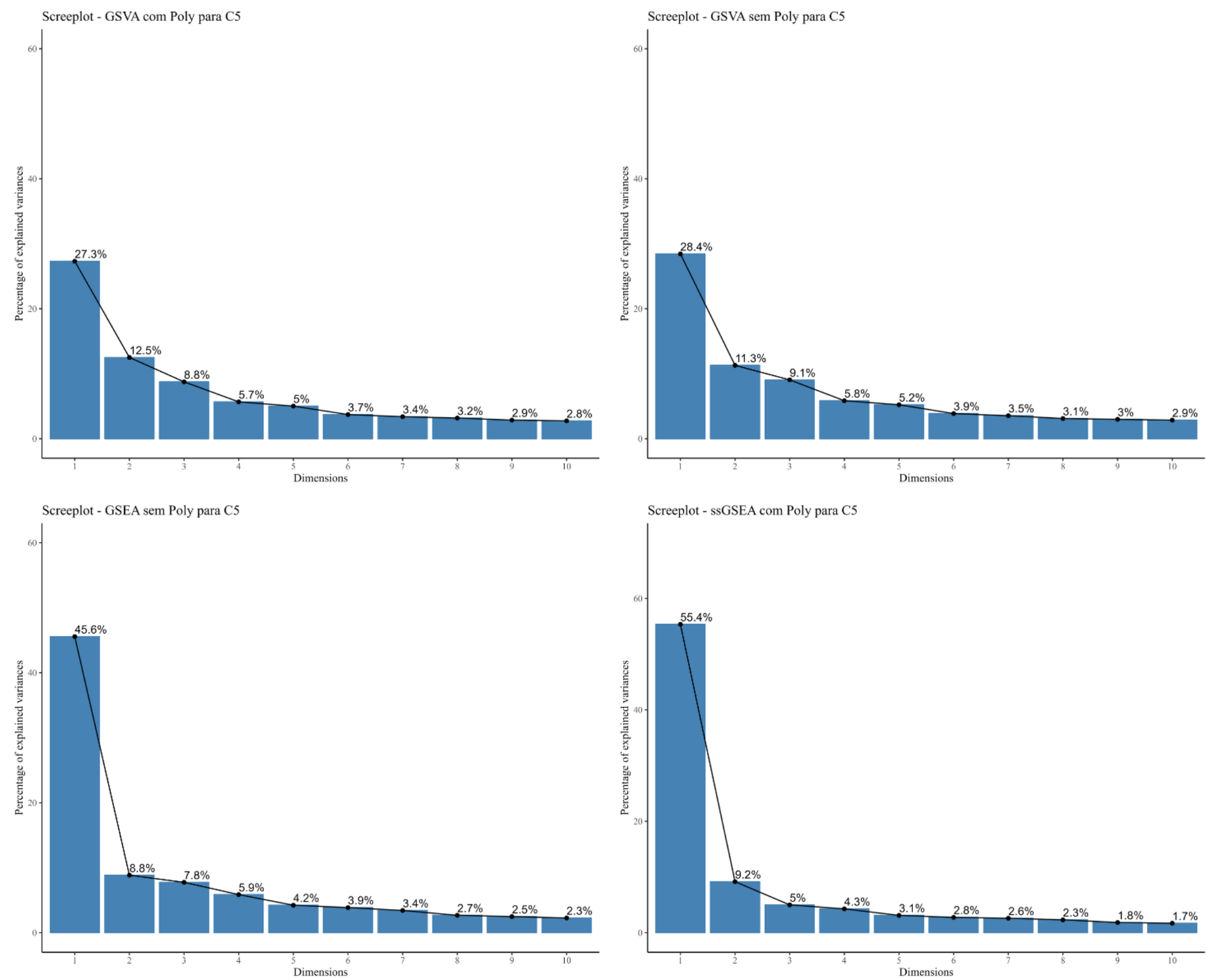


Figura Suplementar 21 - Scree plot dos 21 biomateriais analisados para o conjunto C5, com base nos scores GSVA (com POLY), GSVA (sem POLY), GSEA e ssGSEA, respectivamente.

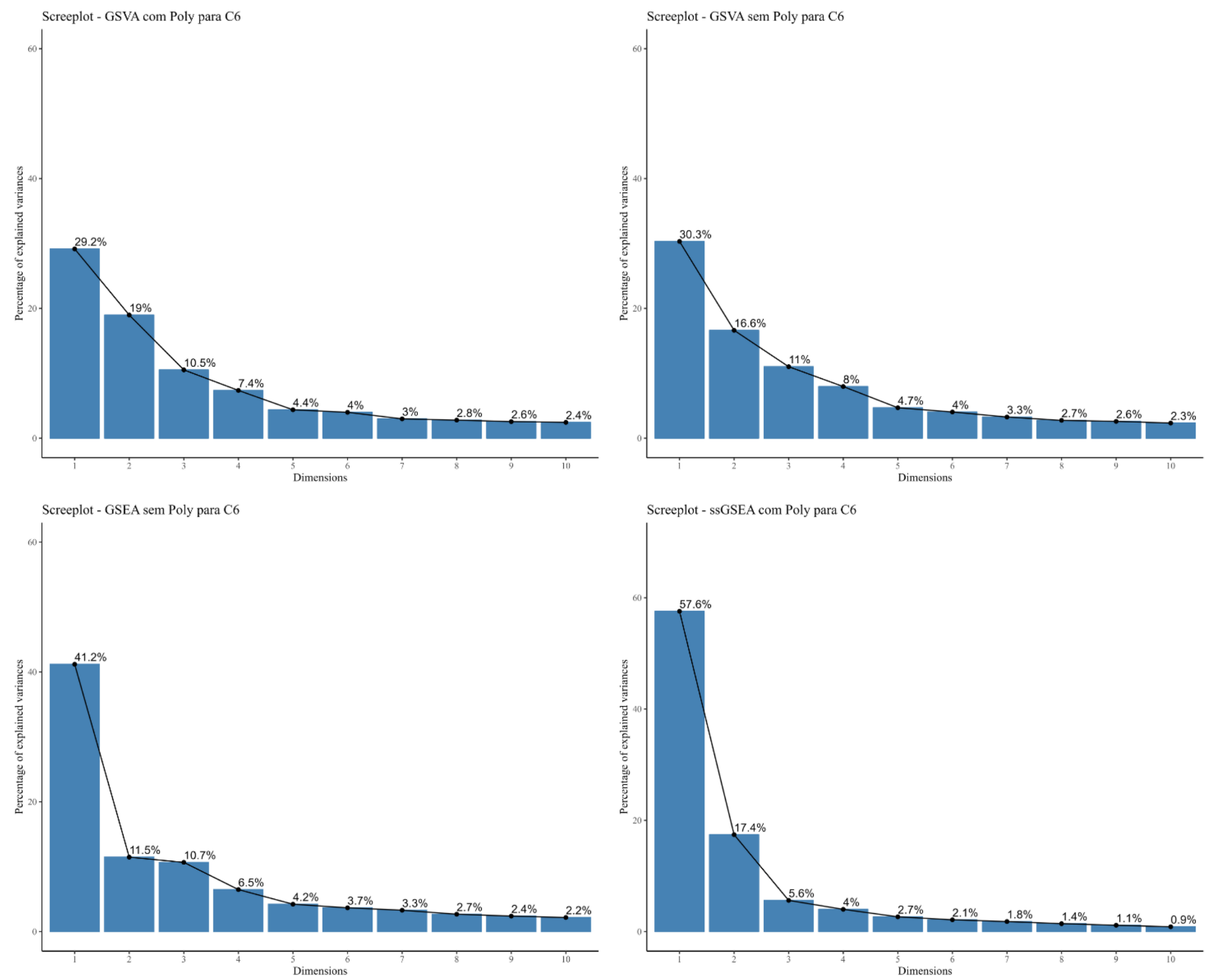


Figura Suplementar 22 - Scree plot dos 21 biomateriais analisados para o conjunto C6, com base nos scores GSVA (com POLY), GSVA (sem POLY), GSEA e ssGSEA, respectivamente.

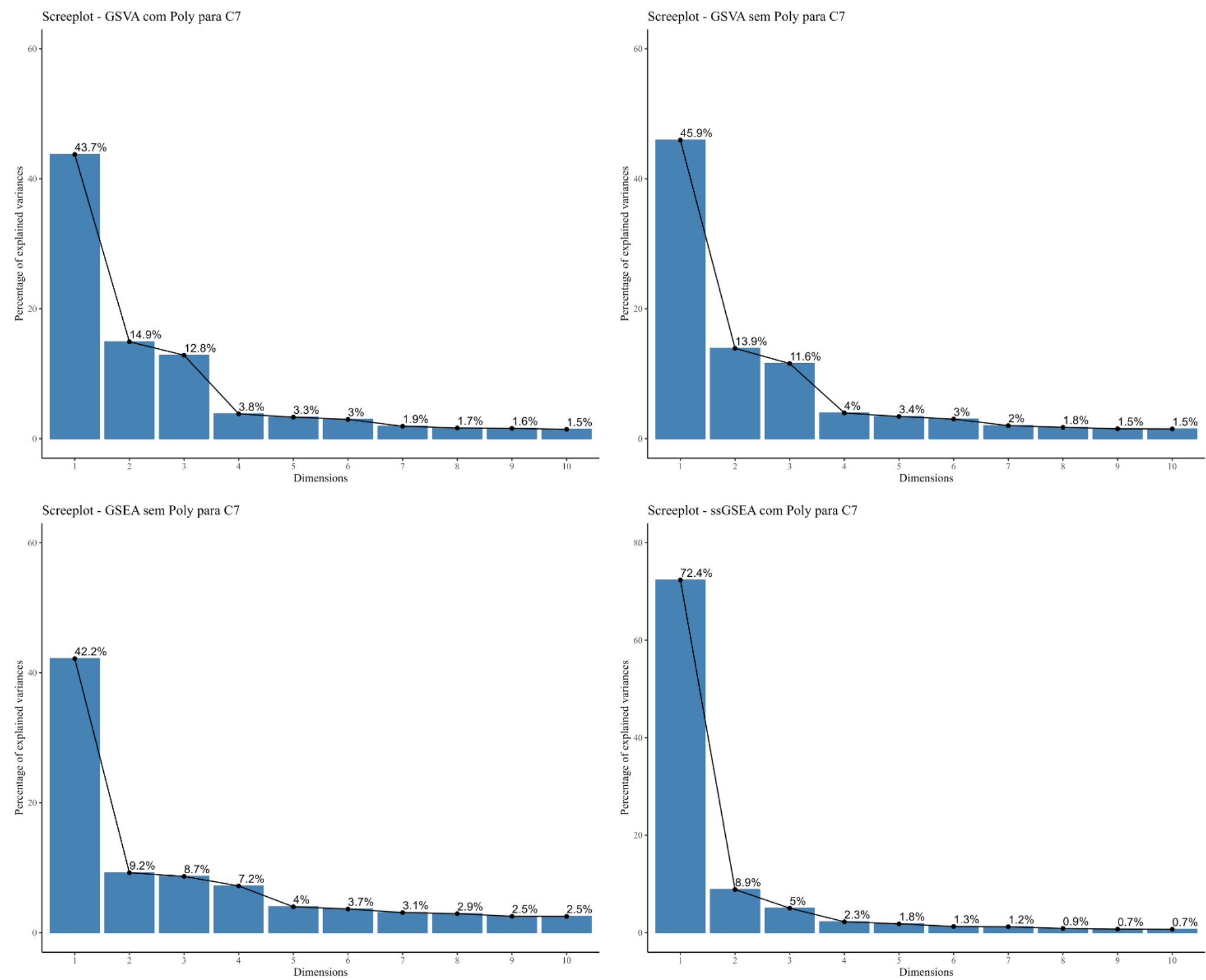


Figura Suplementar 23 - Scree plot dos 21 biomateriais analisados para o conjunto C7, com base nos scores GSVA (com POLY), GSVA (sem POLY), GSEA e ssGSEA, respectivamente.

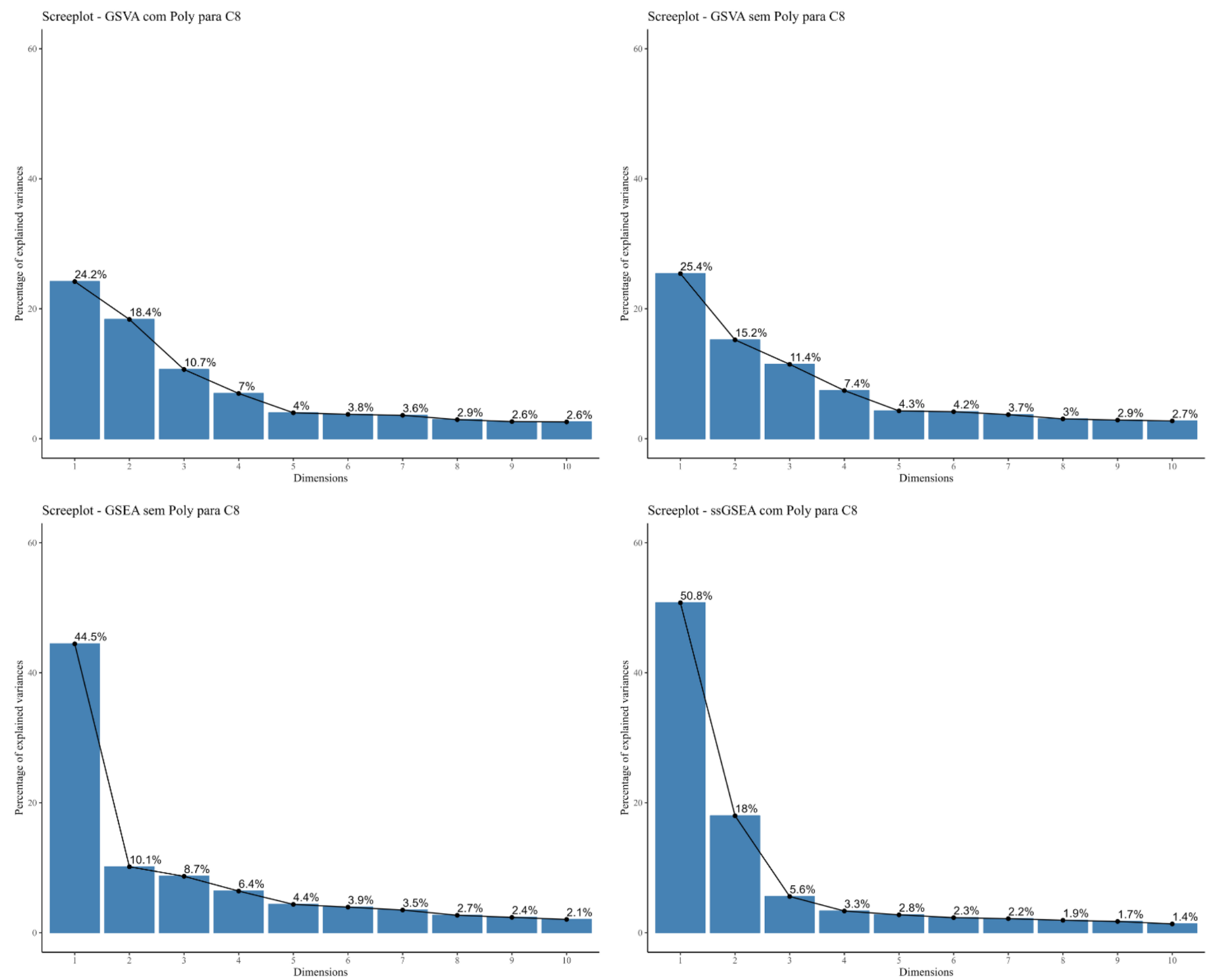


Figura Suplementar 24 - Scree plot dos 21 biomateriais analisados para o conjunto C8, com base nos scores GSVA (com POLY), GSVA (sem POLY), GSEA e ssGSEA, respectivamente.

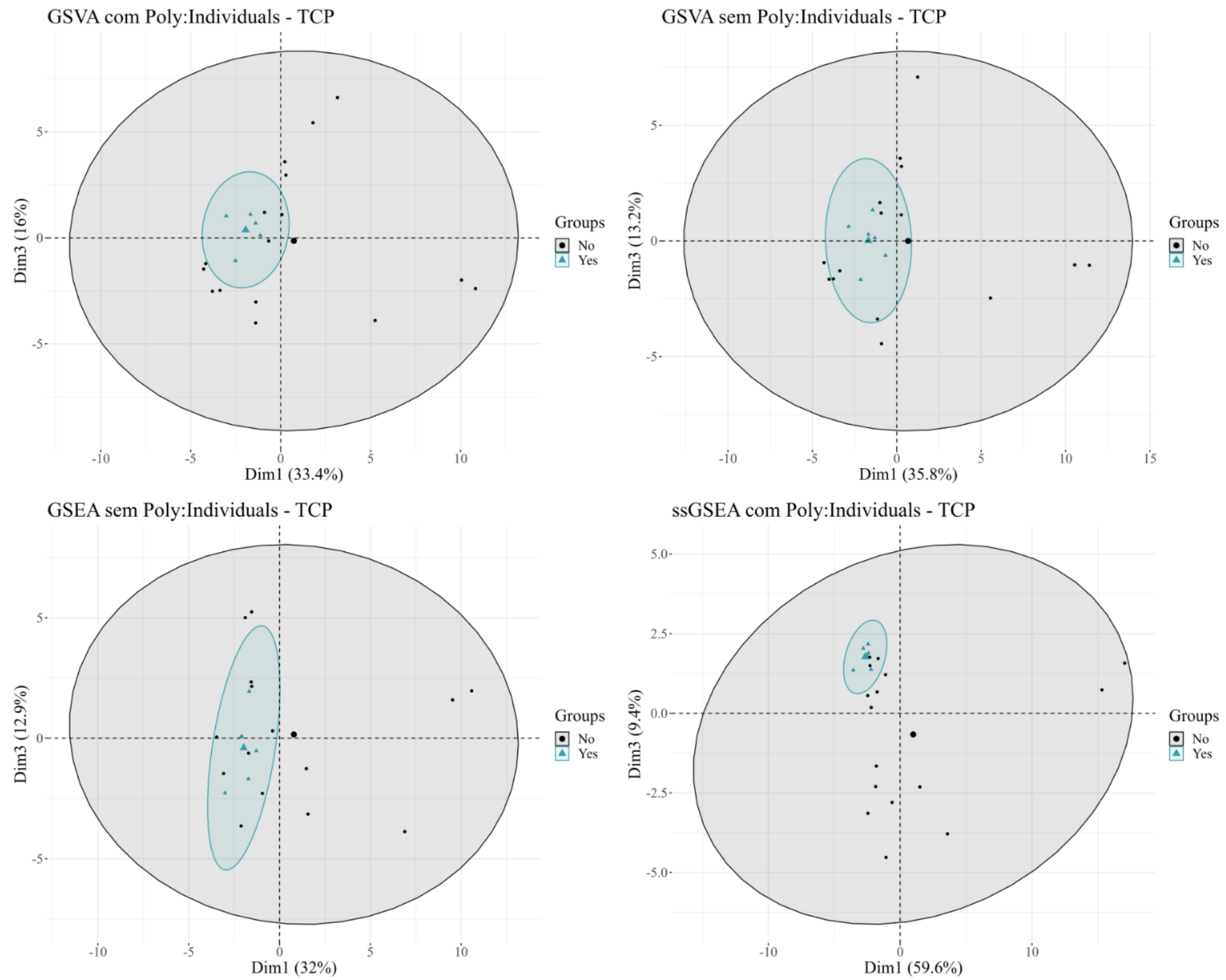


Figura Suplementar 25 - Análise de componentes principais (PCA), analisando a distribuição dos 21 biomateriais para o conjunto Hallmarks, com base na presença ou não de TCP na composição, sendo avaliados para os scores GSV (com POLY), GSV (sem POLY), GSEA e ssGSEA, respectivamente. Estudo feito para 1ª e 3ª dimensão.

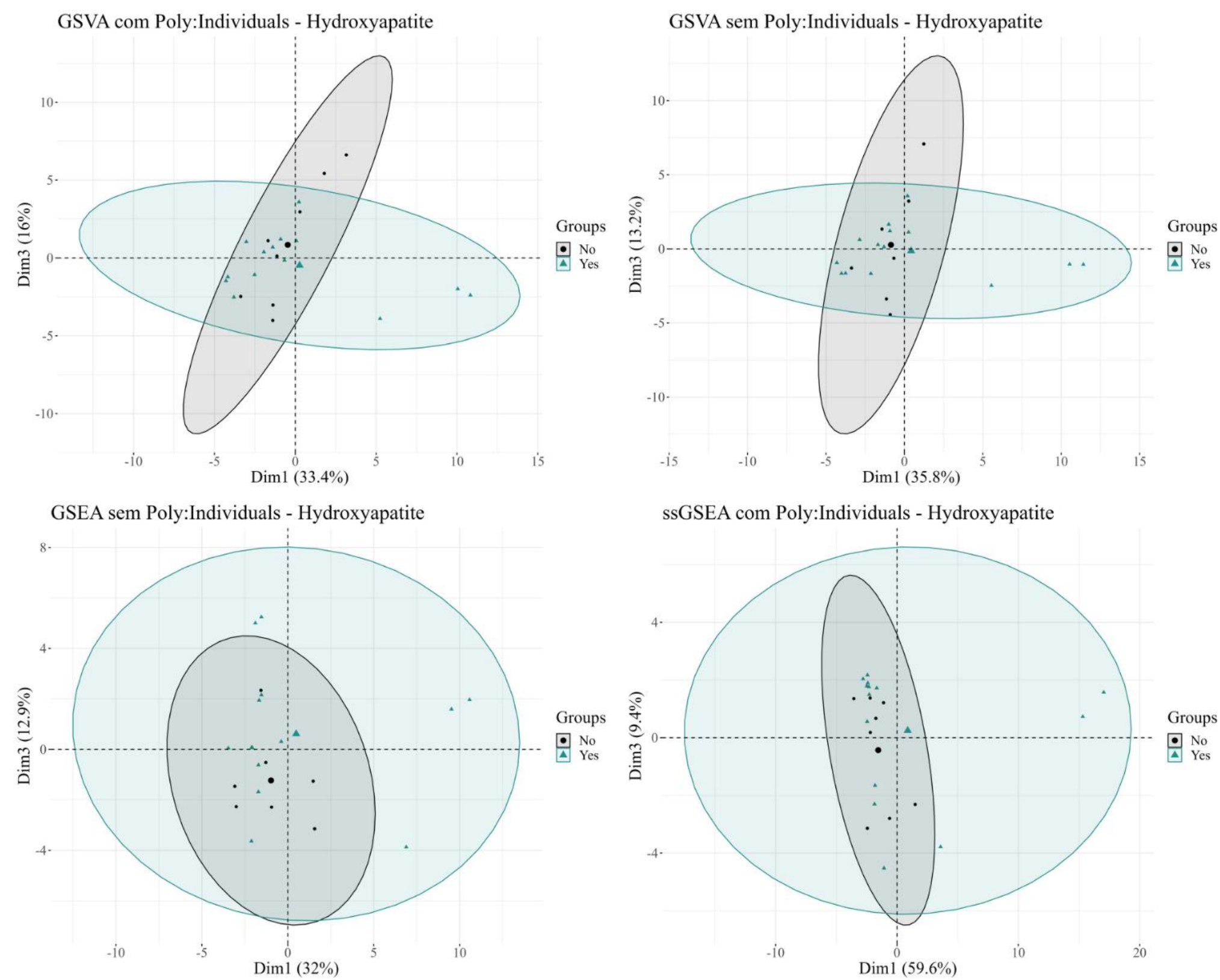


Figura Suplementar 26 - Análise de componentes principais (PCA), analisando a distribuição dos 21 biomateriais para o conjunto Hallmarks, com base na presença ou não de Hydroxyapatite na composição, sendo avaliados para os scores GSVA (com POLY), GSVA (sem POLY), GSEA e ssGSEA, respectivamente. Estudo feito para 1ª e 3ª dimensão.

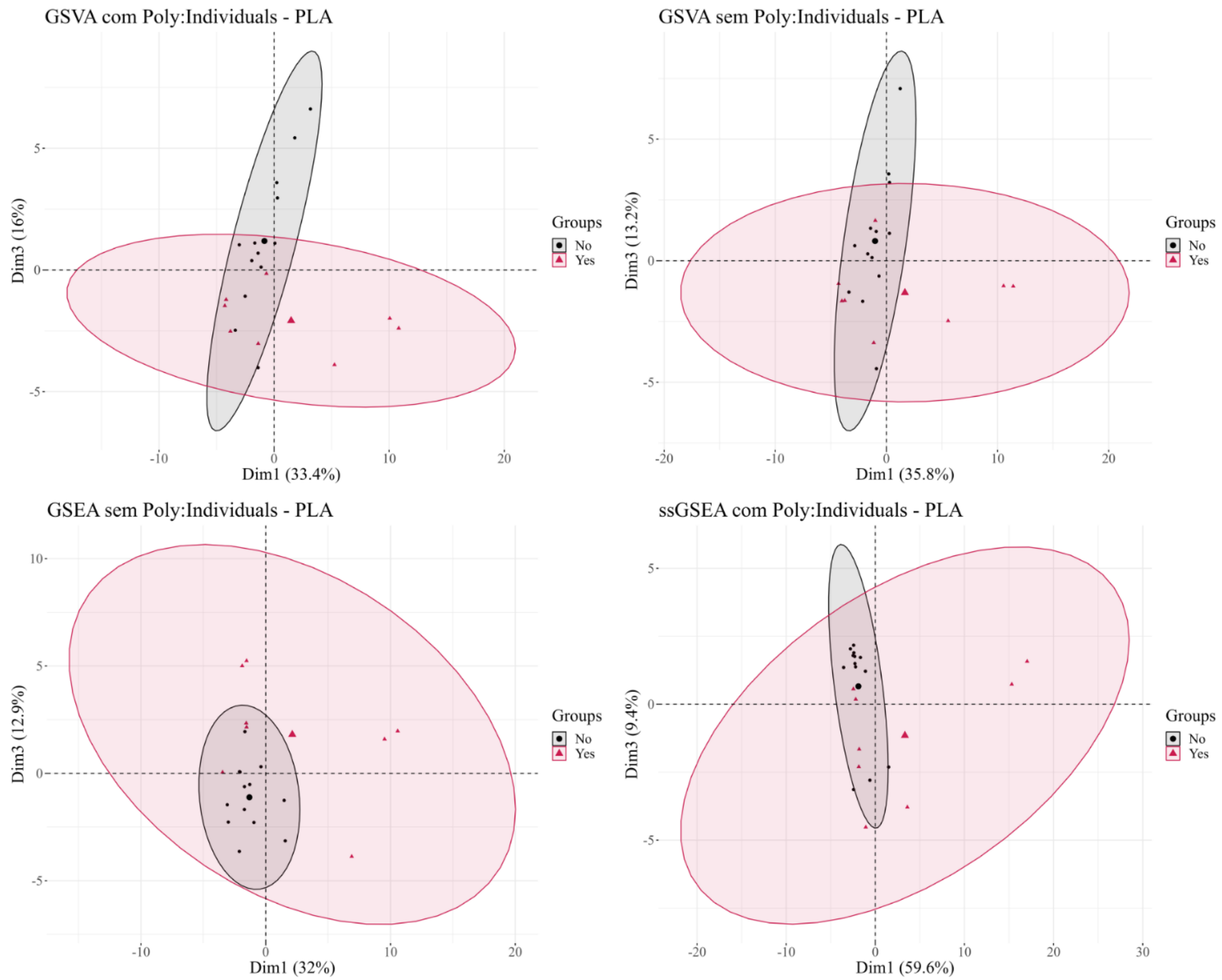


Figura Suplementar 27 - Análise de componentes principais (PCA), analisando a distribuição dos 21 biomateriais para o conjunto Hallmarks, com base na presença ou não de PLA na composição, sendo avaliados para os scores GSV (com POLY), GSV (sem POLY), GSEA e ssGSEA, respectivamente. Estudo feito para 1ª e 3ª dimensão.

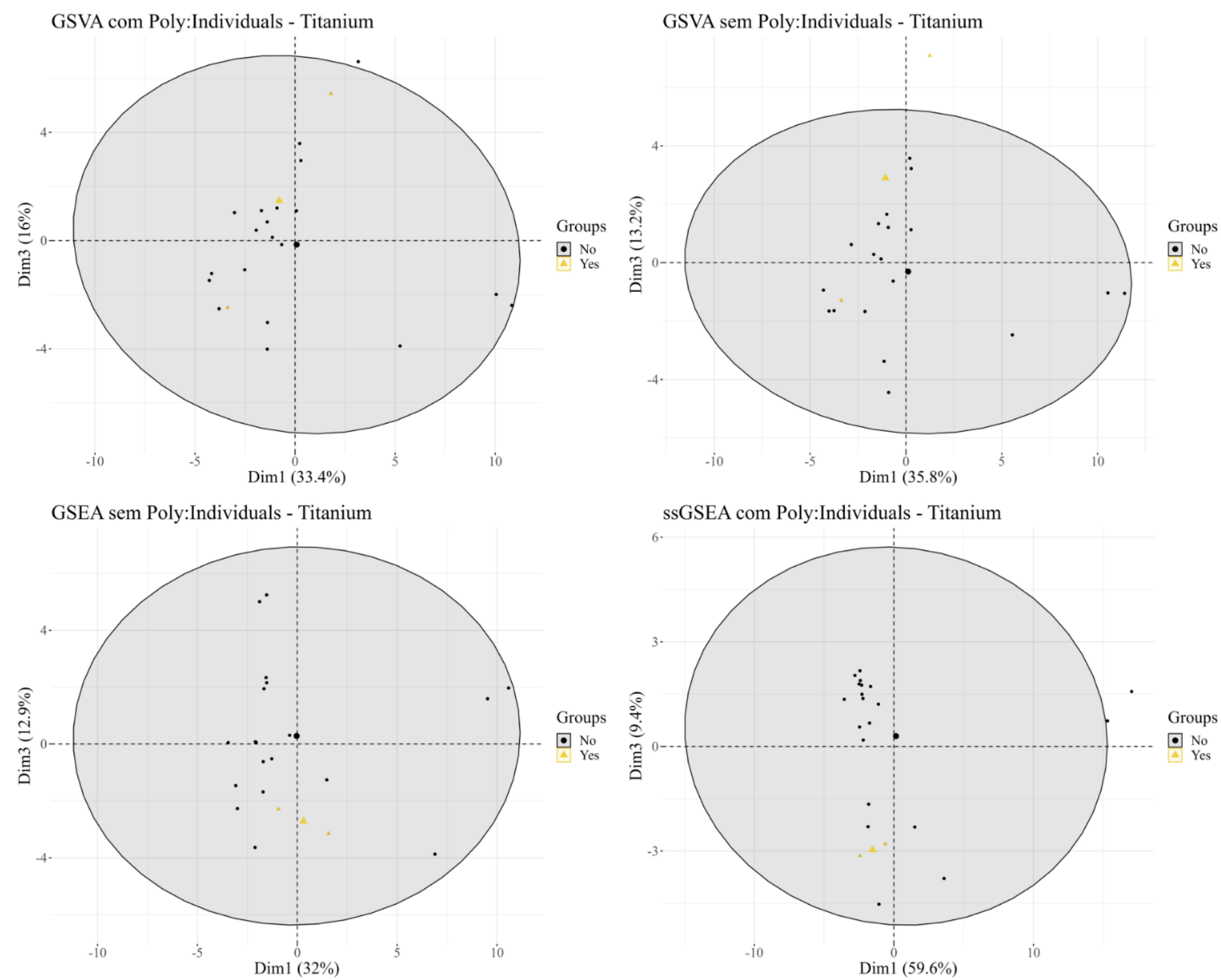


Figura Suplementar 28 - Análise de componentes principais (PCA), analisando a distribuição dos 21 biomateriais para o conjunto Hallmarks, com base na presença ou não de Titanium na composição, sendo avaliados para os scores GSVA (com POLY), GSVA (sem POLY), GSEA e ssGSEA, respectivamente. Estudo feito para 1ª e 3ª dimensão.

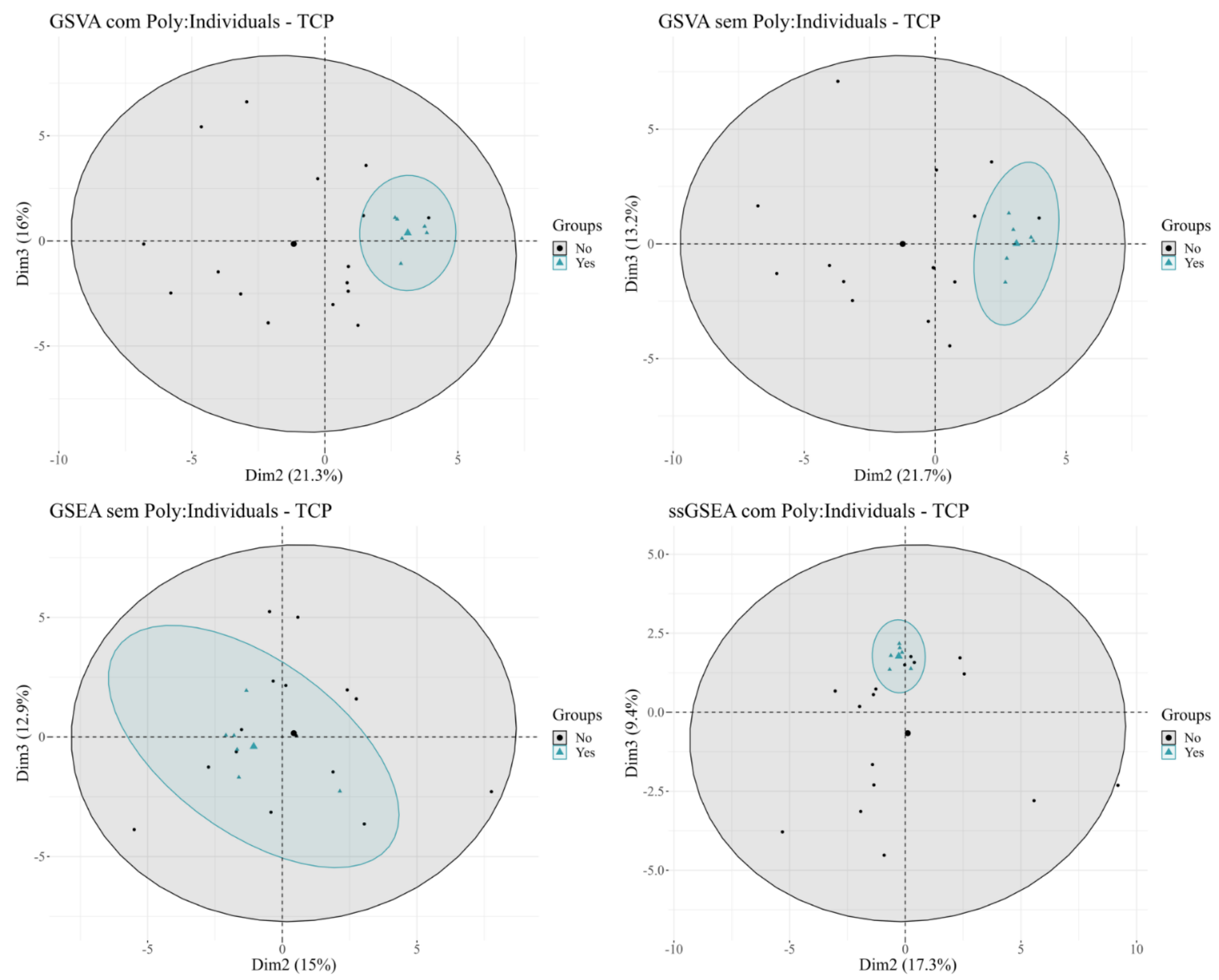


Figura Suplementar 29 - Análise de componentes principais (PCA), analisando a distribuição dos 21 biomateriais para o conjunto Hallmarks, com base na presença ou não de TCP na composição, sendo avaliados para os scores GSVA (com POLY), GSVA (sem POLY), GSEA e ssGSEA, respectivamente. Estudo feito para 2ª e 3ª dimensão.

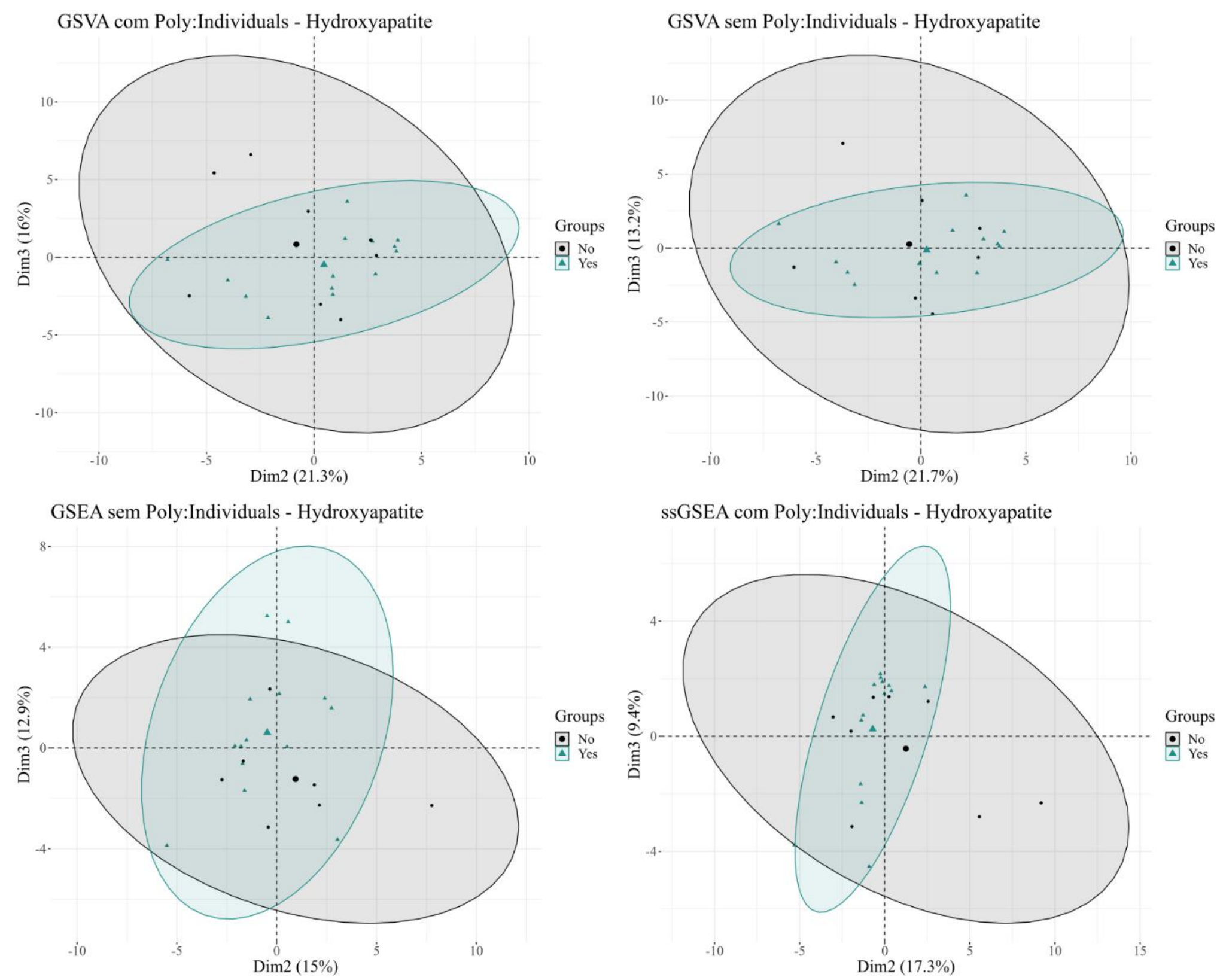


Figura Suplementar 30 - Análise de componentes principais (PCA), analisando a distribuição dos 21 biomateriais para o conjunto Hallmarks, com base na presença ou não de Hydroxyapatite na composição, sendo avaliados para os scores GSVA (com POLY), GSVA (sem POLY), GSEA e ssGSEA, respectivamente. Estudo feito para 2ª e 3ª dimensão.

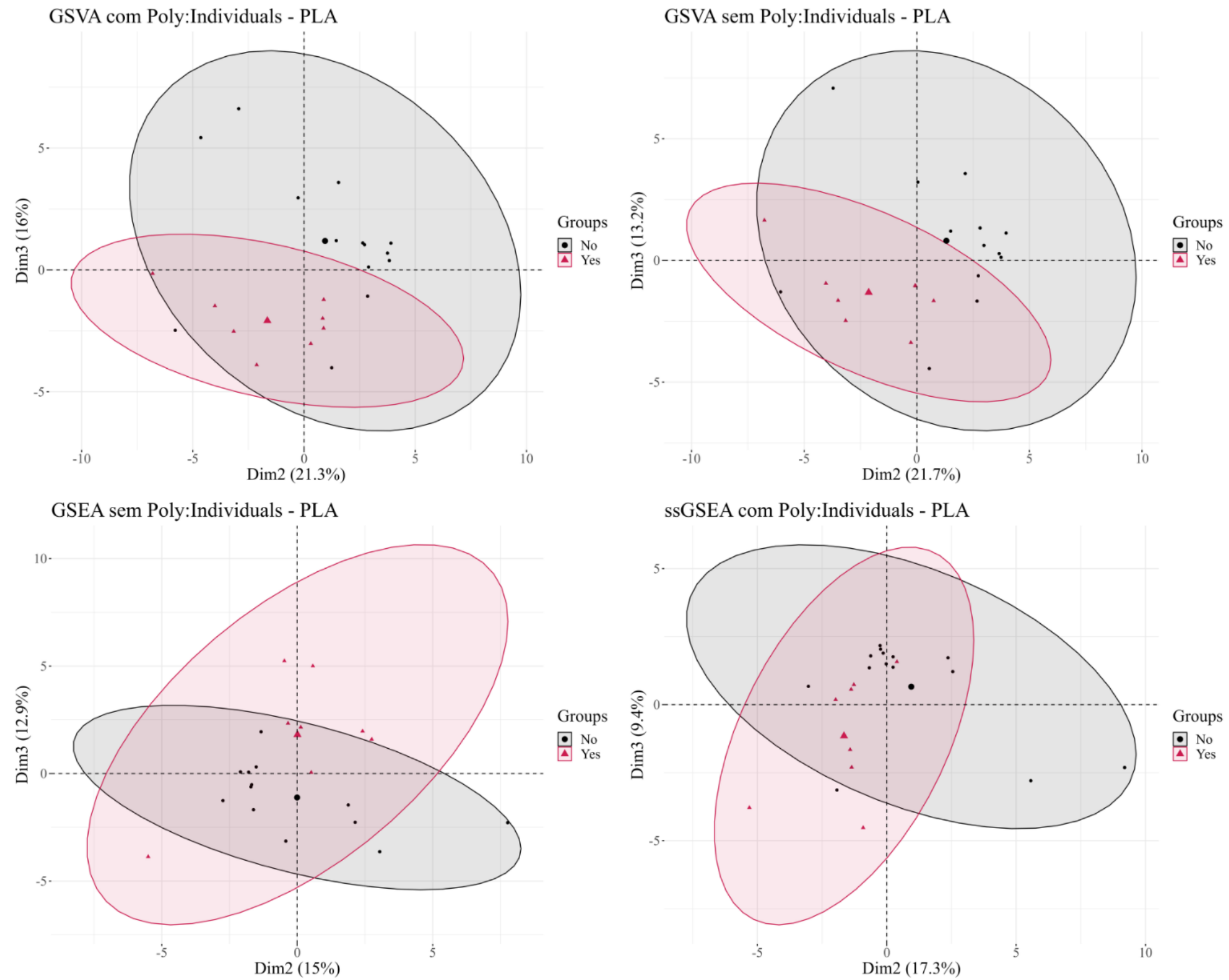


Figura Suplementar 31 - Análise de componentes principais (PCA), analisando a distribuição dos 21 biomateriais para o conjunto Hallmarks, com base na presença ou não de PLA na composição, sendo avaliados para os scores GSVa (com POLY), GSVa (sem POLY), GSEA e ssGSEA, respectivamente. Estudo feito para 2ª e 3ª dimensão.

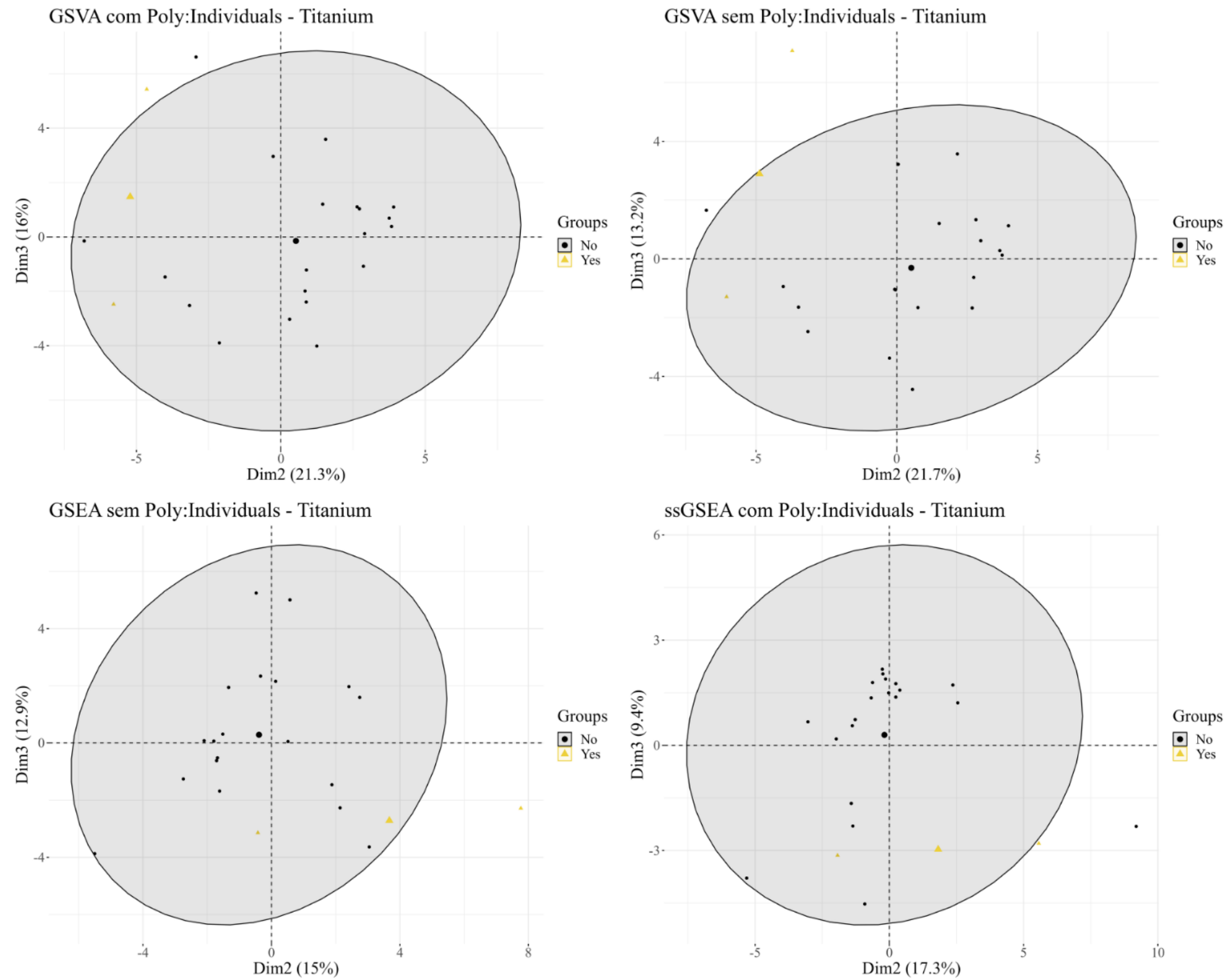


Figura Suplementar 32 - Análise de componentes principais (PCA), analisando a distribuição dos 21 biomateriais para o conjunto Hallmarks, com base na presença ou não de Titanium na composição, sendo avaliados para os scores GSVA (com POLY), GSVA (sem POLY), GSEA e ssGSEA, respectivamente. Estudo feito para 2ª e 3ª dimensão.