Analise de dados de expressao genetica: Caso de Estudo - Pregnancy associated breast cancer

Introdução

O dataset escolhido para análise tem o id GDS4766 e é apelidado de "Pregnancy-associated breast cancer: laser capture microdissected epithelia and stroma". Este foi o dataset escolhido, uma vez que, inicialmente, observou-se a divisão das amostras em cancro da mama associado à gravidez e não associado à gravidez (PABC e não-PABC, respetivamente), bem como a divisão entre células tumorais e células normais. Isto levou-nos a concluir que poderia ser interessante analisar se seria possível distinguir as amostras entre os diferentes grupos. Para além disto, existe um número razoável de pacientes avaliado (33), que constituem as amostras. Por fim, o facto de o cancro da mama ser líder no número de casos em Portugal, com incidência progressivamente maior, com 6000 novos casos a sugirem anualmente, ajudou na escolha deste dataset.

Através da leitura do artigo correspondente ao dataset, foi possível observar que os dados foram retirados a partir de amostras de mulheres diagnosticadas com cancro da mama, separadas em dois grupos principais: grávidas e não grávidas.

No que toca a células tumorais provenientes de um tumor PABC, quando comparadas com células normais, os genes ligados à resposta imunitária parecem ser sobreexpressos, enquanto que os genes envolvidos na angiogénese e deposição extracelular matricial são subexpressos. Esta situação sugere que a alta severidade de um tumor PABC pode aumentar a predisposição para a formação de metástases, através de degradação da matriz extracelular.

Para além disto, é mencionado que os níveis elevados de estrogénio e progesterona durante a gestação poderão influenciar o desenvolvimento desta doença. Isto porque os recetores de estrogénio e progesterona são frequentemente negativos em aproximadamente 70% dos casos de tumor PABC, apesar de não ser bem claro que esta negatividade aconteça estritamente durante a gravidez.

As duas principais populações de células da mama são epiteliais, com aos vasos e lóbulos produtores de leite considerados como os 'targets' principais das hormonas da gravidez, e o estroma, que contribui para elementos de ligação de tecidos, como vascularização, células de imunidade, membrana de base e matriz extracelular. A maioria dos genes sobreexpressos nos tecidos epiteliais de tumores PABC lidam com a regulação da proliferação celular, metabolismo e severidade do tumor e incluem genes utilizados para validar a ocorrência de tumores. Quando comparados com os tecidos normais, há também um elevado número de genes fortemente associados ao ciclo celular, que são processos frequentemente regulados hormonalmente.

Para além disto, há genes codificantes de fatores proliferativos, sinalizadores e imunomodelares de morte celular que são hormonalmente regulados no estroma.

Deste modo, foi possível perceber que as amostras obtidas serão separadas em vários tipos de grupos, que constituirão os metadados do dataset obtido: cancro associado à gravidez vs cancro não associado à gravidez, células tumorais vs células normais, células com receptores de estrogénio positivos vs células com receptores de estrogénio negativo, e células do estroma vs células do epitélio. Isto revela que há padrões complexos de regulação hormonal neste tipo de cancro. Para além disto, os genes, que corresponderão aos dados "propriemente ditos", serão genes bastantes relacionados com a gravidez, resposta imunitária, angiogénese, desposição extracelular matricial, produção de receptores de estrogénio, proliferação e ciclo celular, metabolismo, entre outros.

Irá então ser feita análise de expressão diferencial, análise de enriquecimento, clustering e análise preditiva, por forma a corroborar algumas das hipóteses iniciais e obter informações adicionais.

Preparação dos Dados para Análise

De forma a obter o dataset escolhido sob a forma de um objecto expressionSet, foi necessário realizar vários passos. Primeiramente, utilizou-se os packages "GEOQuery", obtido no bioconductor, e "Biobase", de forma a baixar o dataset da base de dados GEO, através do seu identificador: GDS4766. Depois, baixou-se a anotação correspondente a este dataset. Apesar de o dataset não dizer qual a sua anotação, a página web GEO associada a este dataset explicitava o código da anotação: "hgu133plus2". Depois, converteu-se o dataset obtido para um objecto expressionSet, através de uma função que permitia ao mesmo tempo fazer transformação logarítmica de base 2 dos dados. Por fim, verificou-se a estrutura do dataset, sendo que este possui, de facto, 54675 atributos/variáveis e 33 amostras. Para além disto, os atributos dos metadatos foram os esperados:

Tabela 1: Atributos dos metadados do dataset escolhido e respectivo significado.

Nome do atributo	Significado
sample	Nome das Amostras
disease.state	Se amostra se encontra associada à gravidez ou não
cell.type	Se amostra provém do epitélio ou estroma
specimen	Se amostra provém de células tumorais ou normais
genotype	Se os receptores de estrogénio são positivos ou negativos
description	Descrição de cada uma das amostras

Com o dataset na forma de expressionSet, procedeu-se então à realização do pré-processamento destes dados, de forma a que os dados possuam dimensões e gamas de valores adequadas para a sua análise.

Como mencionado acima, a conversão do dataset inicial para um objecto expressionSet foi acompanhado pela transformação logarítmica dos dados. Esta transformação revela-se bastante importante, uma vez que os níveis de expressão de genes estão altamente enviesados a uma escala linear, ou seja, os genes menos expressados possuem valores entre 0 e 1 e os genes mais expressos com valores entre 1 e infinito positivo, sendo que 1 significa que não há mudança. Deste modo, a transformação logarítmica permite que os dados fiquem mais simétricos e, assim, seja possível obter resultados mais precisos e relevantes no que toca a testes paramétricos, como é o caso da análise de expressão diferencial, mencionada mais à frente.

De seguida, procedeu-se à contagem do número de valores omissos, de forma a que estes pudessem ser tratados. O número de valores omissos revelou-se nulo.

Depois, procedeu-se à filtragem de "Flat Patterns", ou seja, procurou-se remover os genes que possuem valores muito constantes ao longo das diferentes amostras, uma vez que não trazem informação relevante na distinção de grupos de amostras. Para a realização deste passo, foi utilizado o package "genefilter". Decidiu-se proceder então à remoção de genes cujo desvio padrão dos valores de expressão ao longo das amostras é menor do que duas vezes a mediana dos desvios padrão de todos os genes. Tal pode ser observado na figura xxxxxx. Com este passou-se a ter apenas 6378 genes, em vez dos anteriores 54675.

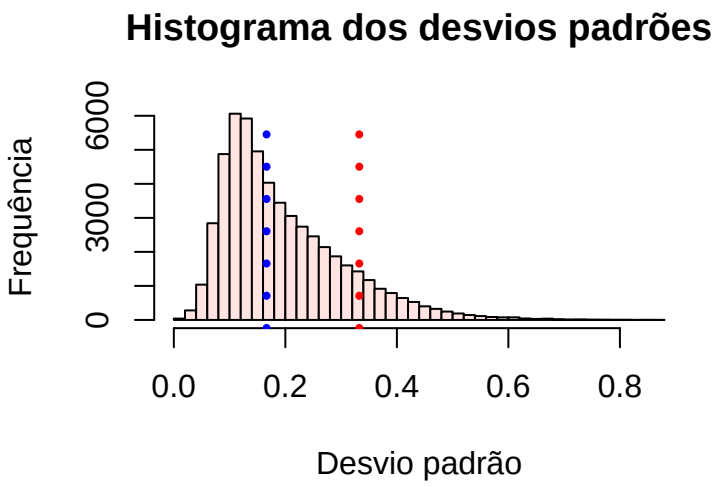


Figura 1: Histograma dos desvios padrões dos genes. A linha vertival azul corresponde à média dos desvios padrões. A linha vertical vermelha corresponde ao dobro da média dos desvios padrões. Os genes cujos desvios padrões se encontram à esquerda da linha vermelha foram retirados.

Por fim, realizou-se a normalização entre as diferentes amostras, de forma a que os dados fiquem normalizados, ou seja, média nula e desvio padrão 1.

Com a preparação dos dados finalizada, foi então possível proceder à análise dos mesmos.

Pipeline de Análise e Ferramentas Usadas

Uma questão básica na análise de dados da expressão genética é a identificação de genes com níveis significativamente diferentes de expressão de forma a distinguir duas condições experimentais. Isto é a realização da análise de expressão diferencial. Para isto, inicialmente, foram realizados testes estatísticos de hipóteses para todas as linhas do expressionSet de forma a distinguir as amostras relacionadas com a gravidez ou não. Os genes com maior evidência de expressão diferencial são aqueles que possuem menor p-value associado. Assim, os resultados obtidos pelos testes estatísticos foram ordenados de forma crescente pelo seu p-value, sendo selecionados apenas os 20 genes com menor p-value. Este processo de identificação dos genes diferencialmente expressos foi realizado da mesma forma para distinguir as amostras dos seguintes atributos dos metadados: “cell.type”, “specimen” e “genotype”. Para a realização desta tarefa, foi necessário utlizar o package, já referido anteriormente, “genefilter”.

Seguidamente foi realizada a análise de enriquecimento sobre os conjuntos de genes diferencialmente expressos. Esta análise passa por determinar quais os genes mais diferencialmente expressos, ao se escolher aqueles com p-value inferior a 0.05, e obter os respectivos IDs da base de dados EntrezGene. A partir destes IDs, foi então possível saber os termos GO associados. A análise de enriquecimento dará os termos GO que aparecem mais associados a estes genes, ou seja, as funções biológicas mais presentes no nosso dataset, o que enriquece o nosso dataset. Novamente, esta análise foi realizada para os mesmos atributos dos metadados usados na análise de expressão diferencial. A análise de enriquecimento foi feita através de testes estatísticos hipergeométricos, para os quais foi necessário criar parâmetros: o universo de genes utilizado para procurar os termos GO foram todos os genes presentes no dataset, sendo para isso fornecido os respectivos IDs da base de dados EntrezGene; os termos GO foram os associados à função molecular dos genes; termos GO com p-value maior que 0.025 não foram tidos em conta; e os termos GO obtidos teriam de ser os que se encontravam sobreexpressos. É de notar que, como os testes para determinar os genes mais diferencialmente expressos foram executados na análise de expressão diferencial,foi a partir destes resultados que se escolheu

os genes cujo p-value se revelava menor que 0.05. Por último, para que esta tarefa pudesse ser realizada, foi necessário utilizar o package "GOstats".

De forma a saber se é possível agrupar as amostras de acordo com cada atributo dos metadados através da distância entre todos os genes do dataset, realizou-se um cluster hierárquico. A métrica de distância utilizada para distinguir os genes foi a euclediana e o método de distância entre clusters foi a média das distâncias dos genes em cada cluster. Após a realização destes clusters, procurou-se saber se escolhendo os 20 melhores genes diferencialmente expressos para cada atributo dos metadados, obtidos através da análise de expressão diferencial, permitiriam um melhor agrupamento das amostras para cada um destes atributos. Também se realizou clusters hierárquicos para os 20 melhores genes de cada atributo dos metadados através da correlação de pearson, por forma a saber se era possível agrupar as amostras de acordo com a monotonia destes genes. É de notar que se utilizou a função my.plot.hc, fornecida nas aulas, para permitir perceber como as amostras estão agrupadas no que toca aos níveis de cada atributo dos metadados.

Outro ponto importante na análise deste tipo de dados é a classificação de amostras, em particular, a procura de modelos que melhor permitem distinguir as amostras no que toca a determinados atributos dos metadados. Neste caso, escolheu-se apenas procurar modelos que distingam as amostras relacionadas à gravidez das que não estão relacionadas com a gravidez e as amostras provenientes de um tumor das provenientes de células normais. Escolheu-se apenas estes dois atributos dos metadados, "disease.state" e "specimen", porque são os atributos que parecem ser mais relevantes de classificar quando se obtém uma amostra nova.

Para isto, começou-se por utilizar o package MLInterfaces, uma vez que é um package cujas funções disponibilizadas para realizar treino de modelos e previsão de amostras conseguem manipular datasets na forma de expressionSet. No entanto, com o uso deste package, foi possível notar que este não disponibiliza uma forma "automática" de realizar optimização de parâmetros de cada modelo, algo necessário para encontrar o melhor modelo possível que se adapte aos dados usados e preveja bem novos dados do mesmo tipo. Deste modo, recorreu-se ao package caret, uma vez que é capaz de fazer a optimização de parâmetros, retornando o modelo com o melhor parâmetro. Para treinar os modelos e fazer a sua validação, usou-se todos os dados do dataset, recorrendo à validação cruzada de três folds, com 10 repetições. Deste modo, todos os dados foram utilizados tanto para o treino como para o teste dos dados, mas numa só validação cada exemplo só foi utilizado ou para teste ou para treino. Procurou-se, sempre que possível, fazer a optimização de parâmetros dos modelos. Para cada modelo obtido, calculou-se as variáveis mais importantes para esse modelo, ou seja, as variáveis que mais influenciam a formação do modelo e consequente distinção entre os grupos de amostras, seleccionando-se as vinte melhores. Foi necessário tratar os dados de forma a que se pudesse utilizar o package caret. Este tratamento passou por obter apenas a matriz com os valores dos genes, obter a sua transposta e, por fim, passar a matriz para dataframe, uma vez que o package caret funciona com dados cujas amostras se situam nas linhas e atributos nas colunas dos dados, ao contrário do que acontece com os dados sob a forma de expressionSet. Os modelos escolhidos para fazer o seu treino foram tanto os modelos mais simples K-Nearest Neighbours e Naive-Bayes, como os modelos de árvores de decisão, Random Forests, Partial Least Squares (PLS) e Máquinas de Vectores de Suporte (SVMs).

Com o decorrer da classificação de amostras, foi possível perceber que, de facto, o elevado número de amostras torna o treino da maioria dos modelos num processo demorado. Para além disto, o elevado número de amostras pode implicar a existência de muito ruído e de maior complexidade dos modelos, algo que não é desejável por poder resultar num má previsão de novas amostras. Deste modo, procedeu-se à selecção dos conjuntos de atributos que seriam capazes de levar a uma previsão mais robusta, generalizada e fácil dos atributos dos metadados em causa: "disease.state", cujas classes são associada à gravidez e não associada à gravidez, e "specimen", cujas classes são células tumorais e células normais. O método de selecção de atributos usado foi o wrapper. Este tipo de método é realizado em conjunto com um modelo a optimizar, que neste caso foi Random Forests, de forma a optimizar o conjunto de atributos a seleccionar. Para realizar esta tarefa, foi utilizado, mais uma vez, o package caret.

Resultados e Discussão

Análise de Expressão diferencial Na análise de expressão diferencial foram identificados, através de testes estatísticos, os 20 genes com menor p-value, ou seja, com maior evidência de expressão diferencial para todos os atributos dos metadados. Os resultados obtidos são apresentados mas tabelas seguintes, estando estas organizadas da seguinte forma: nome do gene, nome da variável correspondente, p-value e função do gene. Esta última informação foi obtida através da base de dados Gene do NCBI. É de notar ainda que existem variáveis para as quais não são conhecidos os nomes dos genes correspondentes, nem a sua função.

Na tabela 2 encontram-se os vinte genes com maior evidência de expressão diferencial no que toca ao atributo dos metadados "disease.state", sendo possível constatar que os respectivos p-values são, de facto, bastante pequenos, o que permite aceitar que estes genes são diferencialmente expressos. Numa tentativa de perceber melhor os resultados, tentou-se procurar saber se alguns destes genes estariam, de alguma forma, relacionados com as células do estroma ou epiteliais, uma vez que, à partida, são os genes que melhor permitem distinguir estas duas classes.No entanto, houve alguma dificuldade em encontrar resultados que associem os genes a complicações ligadas à gravidez. O gene CXCL13 está associado a cancro da mama em idades inferiores aos 45 anos. O cancro da mama em idades mais precoces tem vindo a ocupar uma proporção cada vez maior do número de casos de cancro e está associado a um prognóstico mais desfavorável, que não pode ser totalmente explicado pelos fatores clínicos e moleculares. No entanto, o CXCL13 mostra grande relevância clínica no cancro de mama na juventude e representa um potencial alvo terapêutico. A expressão do gene PNMAL1 está ligada a dietas com maior grau de ingestão de carnes vermelhas processadas. Esta interação gene-dieta, onde a dieta pode modificar o efeito de variantes genéticas sobre o risco de doença, como acontece no cancro colorretal, que pode ter implicações importantes para a prevenção.

Tabela 2: Tabela com os 20 genes com maior evidência de expressão diferencial entre as duas condições do atributo dos metadados "disease.state".

Nome do Gene	Nome da Variável	p-Value	Função do Gene
PRKXP1	1559186_at	4.786574e-06	protein kinase, X-linked, pseudogene 1
PRKXP1	1559188_x_at	1.208066e-05	protein kinase, X-linked, pseudogene 1
Nome Desconhecido	238632_at	2.877108e-05	-
LY6K	223687_s_at	3.286119e-05	lymphocyte antigen 6 complex, locus K
Nome Desconhecido	230910_s_at	1.232951e-04	-
B4GALT6	206232_s_at	1.467213e-04	beta-1,4-galactosyltransferase 6
Nome Desconhecido	239678_at	1.702112e-04	-
Nome Desconhecido	241835_at	2.203584e-04	-
COL2A1	217404_s_at	2.783926e-04	collagen type II alpha 1 chain
CXCL13	205242_at	3.109222e-04	C-X-C motif chemokine ligand 13
COL2A1	213492_at	3.143965e-04	collagen type II alpha 1 chain
Nome Desconhecido	237023_at	3.877082e-04	-
Nome Desconhecido	1561489_at	4.152155e-04	-
ZNF334	238566_at	4.251556e-04	zinc finger protein 334
ELOVL2	220029_at	6.439369e-04	ELOVL fatty acid elongase 2
PNMAL1	218824_at	6.606305e-04	paraneoplastic Ma antigen family-like 1
BANK1	222915_s_at	6.678664e-04	B-cell scaffold protein with ankyrin repeats 1
GSTT1	203815_at	6.693770e-04	glutathione S-transferase theta 1
G0S2	213524_s_at	6.852797e-04	G0/G1 switch 2
Nome Desconhecido	242181_at	8.861605e-04	-

Na tabela 3 encontram-se os vinte genes com maior evidência de expressão diferencial no que toca ao atributo dos metadados "cell.type", sendo possível constatar que os respectivos p-values são, de facto, bastante pequenos, o que permite aceitar que estes genes são diferencialmente expressos. Numa tentativa de perceber melhor os resultados, tentou-se procurar saber se alguns destes genes estariam, de alguma forma, relacionados com as células do estroma ou epiteliais, uma vez que, à partida, são os genes que melhor permitem distinguir

estas duas classes. Ao procurar as suas funções, foi possível constatar que, de facto, alguns destes genes estão relacionados com estas células. O gene que demonstrou possuir menor p-value, o FGF7, codifica uma proteína que é um fator específico potente de crescimento de células epiteliais. Outros exemplos são os genes EMCN, que codifica um tipo de miosina formada em tecidos epiteliais, e o gene VGLL3, cuja expressão está associada a um fenótipo supressor de tumor no cancro epitelial do ovário. Foram também encontrados alguns genes neste conjunto relacionados com o estroma, sendo eles o gene FBN1, que codifica proteínas associadas à matriz extracelular, e o RECK, que codifica uma proteína extracelular cuja expressão é reprimida fortemente em muitos tumores. Por último, parece ser interessante salientar que o gene KIAA1462, apesar de não ter ainda uma função bem conhecida, parece ser um componente das junções entre células endoteliais [A coronary artery disease-associated gene product, JCAD/KIAA1462, is a novel component of endothelial cell-cell junctions.Akashi M, et al. Biochem Biophys Res Commun, 2011 Sep 23. PMID 21884682].

Tabela 3: Tabela com os 20 genes com maior evidência de expressão diferencial entre as duas condições do atributo dos metadados "cell.type".

Nome do Gene	Nome da Variável	p-Value	Função do Gene
FGF7	1555103_s_at	1.620709e-13	fibroblast growth factor 7
MGC24103	232568_at	7.672594e-12	-
ZFPM2	219778_at	4.897125e-11	zinc finger protein, FOG family member 2
Nome Desconhecido	233036_at	1.048204e-10	-
EMCN	219436_s_at	3.578784e-10	endomucin
LRRC32	203835_at	3.674869e-10	leucine rich repeat containing 32
MEDAG	227058_at	4.608431e-10	mesenteric estrogen dependent adipogenesis
KIAA1462	231841_s_at	4.833748e-10	-
ADGRF5	212951_at	7.975633e-10	adhesion G protein-coupled receptor F5
FBN1	235318_at	3.544551e-09	fibrillin 1
ROBO4	226028_at	3.544856e-09	roundabout guidance receptor 4
PREX2	228692_at	4.416276e-09	phosphatidylinositol-3,4,5-trisphosphate dependent Rac exchange factor 2
LUM	229554_at	4.441738e-09	lumican
ECSCR	227780_s_at	6.365953e-09	endothelial cell surface expressed chemotaxis and apoptosis regulator
CLEC1A	219761_at	6.896070e-09	C-type lectin domain family 1 member A
AKAP12	210517_s_at	7.406925e-09	A-kinase anchoring protein 12
KIAA1462	213316_at	1.887229e-08	-
VGLL3	227399_at	2.347628e-08	vestigial like family member 3
RECK	205407_at	2.610962e-08	reversion inducing cysteine rich protein with kazal motifs
ABCC9	208561_at	2.838895e-08	ATP binding cassette subfamily C member 9

Na tabela 4 encontram-se os vinte genes com maior evidência de expressão diferencial no que toca ao atributo dos metadados "specimen", sendo possível constatar que os respectivos p-values são, de facto, bastante pequenos, o que permite aceitar que estes genes são diferencialmente expressos. Mais uma vez, numa tentativa de perceber melhor os resultados, tentou-se procurar saber se alguns destes genes estariam, de alguma forma, relacionados com a proliferação de células tumorais, combate a tumores ou até se algum seria um gene oncogénico, uma vez que, à partida, são estes os genes que melhor permitem distinguir estas duas classes. O gene com p-value menor, CD300LG, é amplamente expresso em células hematopoiéticas, que

são células precursoras de um processo de formação, desenvolvimento e maturação dos elementos figurados no sangue. O segundo gene com menor p-value, o PPP1R14A, codifica uma proteína inibidora da fosfatase da miosina do músculo liso. A miosina é responsável pela contracção do músculo e o músculo liso reveste muitos dos tubos do corpo, como é o caso dos vasos sanguíneos, sendo que um tumor necessita sempre de estar bastante irrigado. O gene PAMR1 tem especial interesse, visto que é um supressor tumoral putativo que se encontra frequentemente inativado em tecidos de cancro da mama e está associado à regeneração do músculo liso.

Tabela 4: Tabela com os 20 genes com maior evidência de expressão diferencial entre as duas condições do atributo dos metadados "specimen".

Nome do Gene	Nome da Variável	p-Value	Função do Gene
CD300LG	1552509_a_at	3.076030e-14	CD300 molecule like family member g
PPP1R14A	227006_at	8.659842e-12	protein phosphatase 1 regulatory inhibitor subunit 14A
Nome Desconhecido	206742_at	4.396833e-11	-
MYH11	201496_x_at	4.636002e-11	myosin, heavy chain 11, smooth muscle
SAMD5	242626_at	6.747342e-11	sterile alpha motif domain containing 5
DLK1	209560_s_at	9.529554e-11	delta like non-canonical Notch ligand 1
CXCL11	211122_s_at	1.563660e-10	C-X-C motif chemokine ligand 11
Nome Desconhecido	227843_at	1.771849e-10	-
MYH11	207961_x_at	2.335500e-10	myosin, heavy chain 11, smooth muscle
Nome Desconhecido	1563295_at	3.186377e-10	-
COL10A1	217428_s_at	3.300154e-10	collagen type X alpha 1 chain
MYH11	201497_x_at	3.642100e-10	myosin, heavy chain 11, smooth muscle
Nome Desconhecido	236647_at	5.496594e-10	-
Nome Desconhecido	236414_at	9.490580e-10	-
SNX10	218404_at	1.165166e-09	sorting nexin 10
ASPM	232238_at	1.633276e-09	abnormal spindle microtubule assembly
PAMR1	213661_at	2.288774e-09	peptidase domain containing associated with muscle regeneration 1
C2orf40	223623_at	2.541102e-09	chromosome 2 open reading frame 40
ADAMDEC1	206134_at	2.855512e-09	ADAM like decysin 1
ACO1	237622_at	3.067827e-09	aconitase 1

Na tabela 5 encontram-se os vinte genes com maior evidência de expressão diferencial no que toca ao atributo dos metadados "genotype", sendo possível constatar que os respectivos p-values são, de facto, bastante pequenos, o que permite aceitar que estes genes são diferencialmente expressos. Numa tentativa de perceber melhor os resultados, tentou-se procurar saber se alguns destes genes estariam, de alguma forma, relacionados com receptores de estrogénio, uma vez que, à partida, são os genes que melhor permitem distinguir estas duas classes. No entanto, a grande maioria dos genes obtidos não estão muito diretamente relacionados com recetores de estrogénio. No entanto há três genes a destacar. O gene CEBPD leva a uma proteína que, ao ser estabilizada pelo recetor de atividade de estrogénio, inibe a expressão de SNAI2 e está associado a um bom prognóstico para o cancro da mama. É expressa nas células epiteliais normais e em cancros de baixo grau. Atenua também o crescimento cellular, mobilidade e capacidade invasiva do cancro. O CHST2 codifica uma proteína sulfotransferase associada a adenocarcinomas mucinosos de células claras, que são os principais tipos histológicos de cancro epitelial do ovário e estão associadas à sua resistência à quimioterapia. A anidrase carbónica XII, originada pelo gene CA12, é um novo alvo terapêutico para superar

a quimiorresistência em células cancerosas. Este produto génico é uma proteína de membrana do tipo I que é altamente expressa em tecidos normais, tais como rim , pâncreas e cólon e encontra-se sobre-expresso em 10% dos carcinomas de células renais. Estes três genes descritos acabaram por se revelar como estando relacionadas com o cancro da mama ou afetados pela gravidez. Para além destes genes, havia outros três ligados a células cancerosas ou carcinomas, demonstrando a relação entre este tipo de genes.

Tabela 5: Tabela com os 20 genes com maior evidência de expressão diferencial entre as duas condições do atributo dos metadados "genotype".

Nome do Gene	Nome da Variável	p-Value	Função do Gene
SLC7A8	217248_s_at	1.056480e-05	solute carrier family 7 member 8
PSAT1	223062_s_at	1.299513e-05	phosphoserine aminotransferase 1
S100B	209686_at	2.982062e-05	S100 calcium binding protein B
CYP27C1	1568868_at	4.993839e-05	cytochrome P450 family 27 subfamily C member 1
Nome Desconhecido	243605_at	5.989080e-05	-
MLPH	218211_s_at	1.453842e-04	melanophilin
BCL11A	219497_s_at	1.605149e-04	B-cell CLL/lymphoma 11A
CEBPD	213006_at	2.164142e-04	CCAAT/enhancer binding protein delta
BCL11A	222891_s_at	2.740405e-04	B-cell CLL/lymphoma 11A
BCL11A	219498_s_at	2.756799e-04	B-cell CLL/lymphoma 11A
CHST2	203921_at	2.878188e-04	carbohydrate sulfotransferase 2
Nome Desconhecido	208451_s_at	3.211029e-04	-
PRR15	226961_at	3.213069e-04	proline rich 15
Nome Desconhecido	1559078_at	3.328573e-04	-
ZG16B	228058_at	3.358580e-04	zymogen granule protein 16B
CA12	215867_x_at	4.316243e-04	carbonic anhydrase 12
Nome Desconhecido	239536_at	4.541480e-04	-
RNF150	236038_at	4.743470e-04	ring finger protein 150
TMC5	240304_s_at	5.042911e-04	transmembrane channel like 5
CA3	204865_at	5.349679e-04	carbonic anhydrase 3

Análise de Enriquecimento Para cada atributo mencionado dos metadados, foi realizada a análise de enriquecimento de forma a encontrar os termos GO que aparecem associados aos genes mais diferencialmente expressos. As tabelas seguintes apresentam os resultados obtidos sendo indicado para cada ID de GO, o p-value, a contagem do número de vezes que cada termo GO aparece e os próprios termos.

Na tabela 6, é possível observar os dez termos GO mais associados ao atributo dos metadados "disease.state". O termo mais presente é "antigen binding", que é uma função que consiste na capacidade de induzir uma resposta imune específica e reagir com os produtos dessa resposta. Para além deste termo, há ainda termos associados ao complexo MHC classe II, que é um complexo presente em várias células apresentadoras de antigénios, como é o caso de células dendríticas, que são células do sistema imunitário dos mamíferos que interagem com células T que, por sua vez, são células que podem reconhecer as moléculas MHC classe II e gerar respostas imunitárias que induzam a morte das células tumorais. Existem ainda outros termos também relacionados com a apresentação de antigénios. Deste modo, pode-se concluir que os genes que permitem diferenciar entre um cancro da mama associado à gravidez ou não associado à gravidez têm, na sua maioria, funções moleculares associadas à resposta imunitária, algo que seria de esperar.

Tabela 6: Tabela com os 10 termos GO mais associados aos genes diferencialmente expressos no que toca à distinção entre amostras associadas à gravidez das não associadas.

IDs de GO	p-values	Contagens de termos GO	Termos
GO:0003823	3.119040e-06	11	antigen binding
GO:0032395	3.327343e-06	5	MHC class II receptor activity
GO:0023026	1.967449e-03	3	MHC class II protein complex binding
GO:0023023	4.624995e-03	3	MHC protein complex binding
GO:0001602	6.520142e-03	2	pancreatic polypeptide receptor activity
GO:0004833	6.520142e-03	2	tryptophan 2,3-dioxygenase activity
GO:0045519	6.520142e-03	2	interleukin-23 receptor binding
GO:0042605	8.700028e-03	3	peptide antigen binding
GO:0015103	1.504406e-02	5	inorganic anion transmembrane transporter activity
GO:0001601	1.851398e-02	2	peptide YY receptor activity

Na tabela 7, é possível observar os dez termos GO mais associados ao atributo dos metadados "cell.type". Os termos GO estão relacionados com a junção das células umas às outras, como é o caso do termo "collagen binding", que consiste numa função onde há interacção com o colagénio, que é um grupo de proteínas fibrosas que são o principal componente do tecido conjuntivo dos mamíferos, ao qual pertence as células epiteliais, e do termo "integrin binding", uma vez que a integrina é um receptore transmembranar associado à junção de células. Pode-se também ainda referir o termo "extracellular matrix structural constituent" que, tal como o nome indica, refere-se a proteínas que façam parte da matriz extracelular das células. Deste modo, pode-se concluir que os genes que permitem diferenciar entre células epiteliais e células do estroma têm, na sua maioria, funções moleculares associadas à junção das células e contituição das suas membranas, algo que seria de esperar.

Tabela 7: Tabela com os 10 termos GO mais associados aos genes diferencialmente expressos no que toca à distinção entre as amostras provenientes do estroma ou do epitélio.

IDs de GO	p-values	Contagens de termos GO	Termos
GO:0005518	2.931631e-05	21	collagen binding
GO:0005178	5.965939e-04	23	integrin binding
GO:0034987	1.743214e-03	7	immunoglobulin receptor binding
GO:0005201	2.249051e-03	25	extracellular matrix structural constituent
GO:0017124	2.703665e-03	18	SH3 domain binding
GO:0004714	3.154349e-03	20	transmembrane receptor protein tyrosine kinase activity
GO:0008329	4.326570e-03	6	signaling pattern recognition receptor activity
GO:0038187	4.326570e-03	6	pattern recognition receptor activity
GO:0004713	5.922451e-03	26	protein tyrosine kinase activity
GO:0001228	8.347281e-03	40	transcriptional activator activity, RNA polymerase II transcription regulatory region sequence-specific binding

Na tabela 8, é possível observar os dez termos GO mais associados ao atributo dos metadados "specimen". Estes termos GO estão bastante relacionados com a ligação entre proteínas, regulação de actividade enzimática, e ligação da cadeia dupla de DNA. Este tipo de actividades tem de estar bastante presente em células tumorais para que estas possam proliferar rapidamente. Um dos termos que talvez possa causar mais atenção é "chemokine activity", uma vez que as quimiocinas são citocinas, moléculas capazes de induzir respostas imunitárias.

Tabela 8: Tabela com os 10 termos GO mais associados aos genes diferencialmente expressos no que toca à distinção entre as amostras relacionadas com o tumor das células normais.

IDs de GO	p-values	Contagens de termos GO	Termos
GO:0005515	0.001200044	788	protein binding
GO:0030234	0.003142029	81	enzyme regulator activity
GO:0004857	0.005951155	39	enzyme inhibitor activity
GO:0043565	0.006356016	86	sequence-specific DNA binding
GO:0008146	0.008037550	9	sulfotransferase activity
GO:0016782	0.008037550	9	transferase activity, transferring sulfur-containing groups
GO:0019887	0.008283282	20	protein kinase regulator activity
GO:0008009	0.008447016	13	chemokine activity
GO:0005102	0.008708768	162	receptor binding
GO:0003690	0.010142694	67	double-stranded DNA binding

Na tabela 9, é possível observar os dez termos GO mais associados ao atributo dos metadados "genotype". É possível constatar que estes termos GO estão bastante relacionados com proteínas e actividade transmembranar, o que é algo de esperar, uma vez que se trata dos genes que melhor permitem distinguir entre as amostras que contêm receptores de estrogénio e as que não têm.

Tabela 9: Tabela com os 10 termos GO mais associados aos genes diferencialmente expressos no que toca à distinção entre as amostras com receptores de estrogénio positivos dos negativos.

IDs de GO	p-values	Contagens de termos GO	Termos
GO:0008509	1.141562e-05	19	anion transmembrane transporter activity
GO:0008514	1.201442e-05	15	organic anion transmembrane transporter activity
GO:0015171	1.319697e-04	9	amino acid transmembrane transporter activity
GO:0005342	1.481790e-04	12	organic acid transmembrane transporter activity
GO:0046943	1.481790e-04	12	carboxylic acid transmembrane transporter activity
GO:0002162	2.562307e-04	4	dystroglycan binding
GO:0005506	1.428050e-03	13	iron ion binding
GO:0015291	1.794573e-03	14	secondary active transmembrane transporter activity
GO:0015179	1.898009e-03	6	L-amino acid transmembrane transporter activity
GO:0015464	2.032862e-03	3	acetylcholine receptor activity

Clustering No que toca a esta análise de dados, obteve-se três plots para cada atributo dos metadados analisados, que consistem num cluster das amostras usando todos os genes e dois clusters das amostras usando apenas os vinte melhores genes do atributo dos metadados correspondente, sendo um obtido através do uso da distância euclediana e o outro através da correlação de Pearson. Todos os clusters encontram-se coloridos de acordo com a classe do atributo em questão a que cada amostra pertence.

Nas figuras 2, 3 e 4, é possível observar os clusters correspondentes ao atributos dos metadados "disease.state". O cluster da figura 3, que corresponde ao cluster hierárquico da distância euclediana de todos

os genes, permite concluir que o uso das distâncias entre todos os genes revela não ser útil para agrupar de forma razoável as amostras nos grupos cancro associado à gravidez e cancro não-associado à gravidez. No entanto, quando se usa apenas os genes obtidos na expressão diferencial, o agrupamento das amostras melhora consideravelmente. Isto deve-de ao facto de os genes usados serem os que melhor ajudam na diferenciação entre os dois grupos de amostras mencionados. Deste modo, usando apenas os vinte melhores genes obtidos na análise de expressão diferencial, é possível agrupar as amostras nos dois grupos mencionados através da distância entre os genes. No que toca ao agrupamento das amostras através da monotonia (correlação) dos vinte melhores genes, ilustrado na figura 4, é possível perceber que este agrupamento também permite separar as amostras em dois grupos um pouco distantes. No caso deste último cluster, é possível observar uma amostra pertencente ao grupo do cancro associado à gravidez bem mais perto das amostras pertencentes ao outro grupo. Isto pode dever-se ao facto de, segundo o artigo correspondente a este dataset, uma das amostras, considerada como cancro associado à gravidez, provir de uma mulher que deu à luz oito semanas antes da recolha da amostra e não durante o período de gestação, tal como aconteceu com as outras.

Figura 2: Cluster das 33 amostras, usando apenas os 20 genes que melhor distinguem as amostras associadas à gravidez das não associadas. A vermelho são as amostras associados à gravidez e a preto as não-associadas à gravidez.

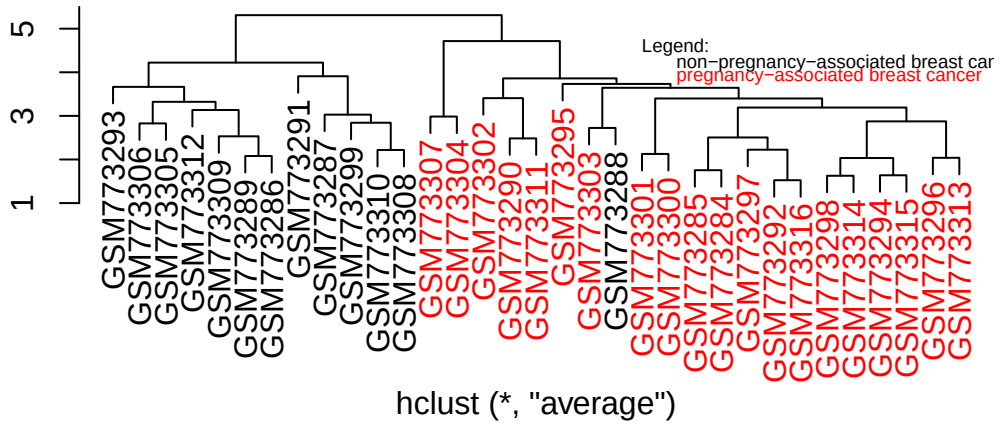


Figura 3: Cluster das 33 amostras, usando todos os genes. A vermelho são as amostras associados à gravidez e a preto as não-associadas à gravidez.

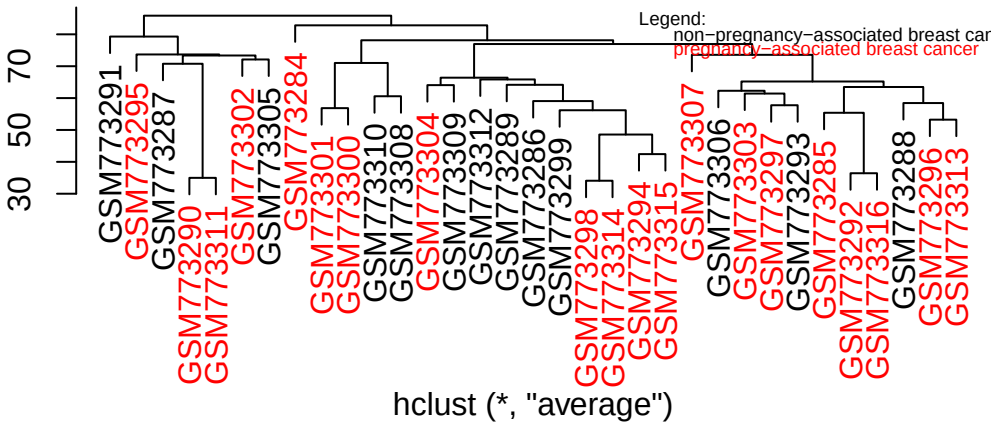
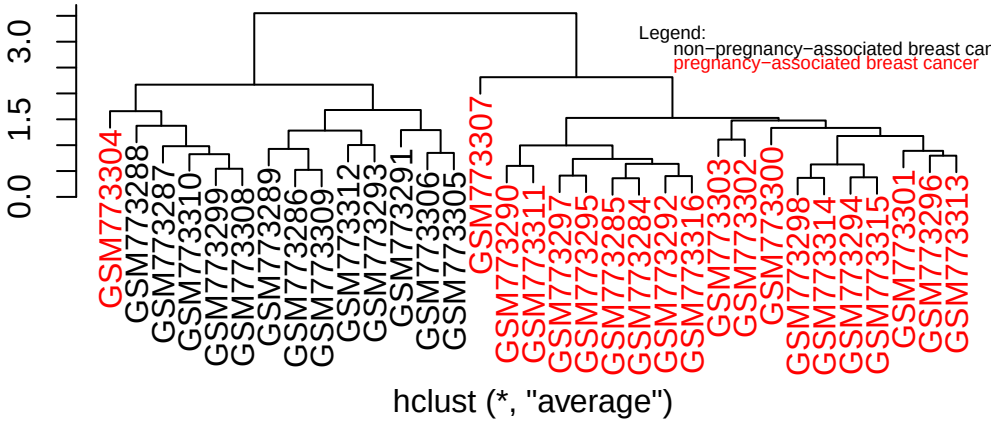


Figura 4: Cluster das 33 amostras, usando apenas os 20 genes que melhor distinguem as amostras associadas à gravidez das não associadas, usando correlação de pearson. A vermelho são as amostras associados à gravidez e a preto as não-associadas à gravidez.



Nas figuras 5, 6 e 7, é possível observar os clusters correspondentes ao atributos dos metadados “cell.type”. O cluster da figura 6, que corresponde ao cluster hierárquico da distância eucledіana de todos os genes, permite concluir que o uso das distâncias entre todos os genes revela não ser útil para agrupar de forma razoável as amostras nos grupos células do estroma e células epiteliais. No entanto, quando se usa apenas os genes obtidos na expressão diferencial, o agrupamento das amostras melhora consideravelmente. Isto deve-de ao facto de os genes usados serem os que melhor ajudam na diferenciação entre os dois grupos de amostras mencionados. Deste modo, usando apenas os vinte melhores genes obtidos na análise de expressão diferencial, é possível agrupar as amostras nos dois grupos mencionados através da distância entre os genes. No que toca ao agrupamento das amostras através da monotonia (correlação) dos vinte melhores genes, ilustrado na figura 7, é possível perceber que não é possível agrupar as amostras em células do estroma e células epiteliais através da correlação dos genes.

Figura 5: Cluster das 33 amostras, usando apenas os 20 genes que melhor distinguem as amostras provenientes do estroma ou epitelial. A vermelho são as amostras provenientes do estroma e a preto as do

epitélio.

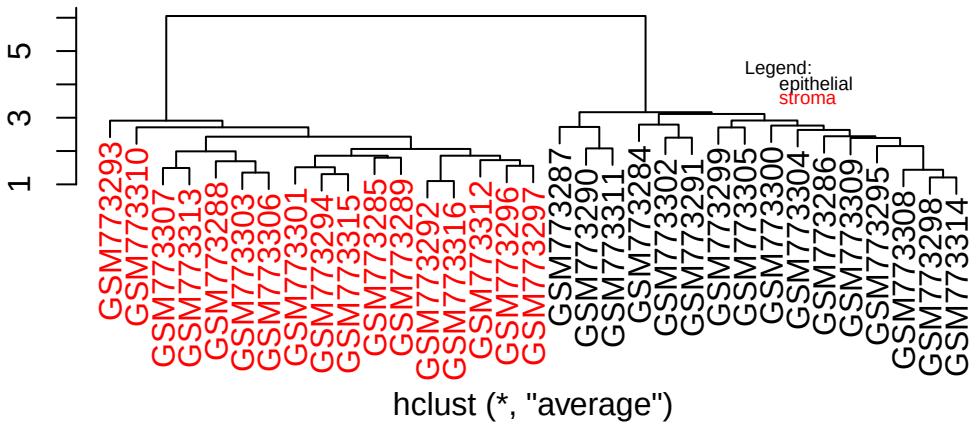


Figura 6: Cluster das 33 amostras, usando todos os genes. A vermelho são as amostras provenientes do estroma e a preto as do epitélio.

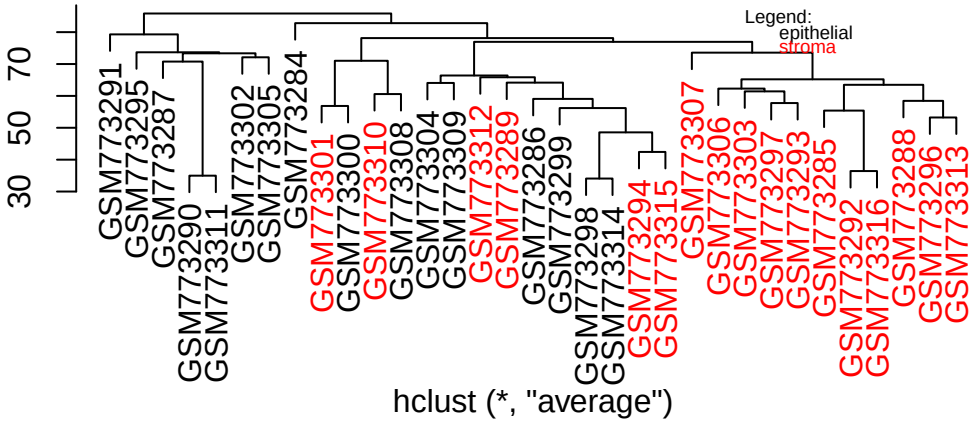
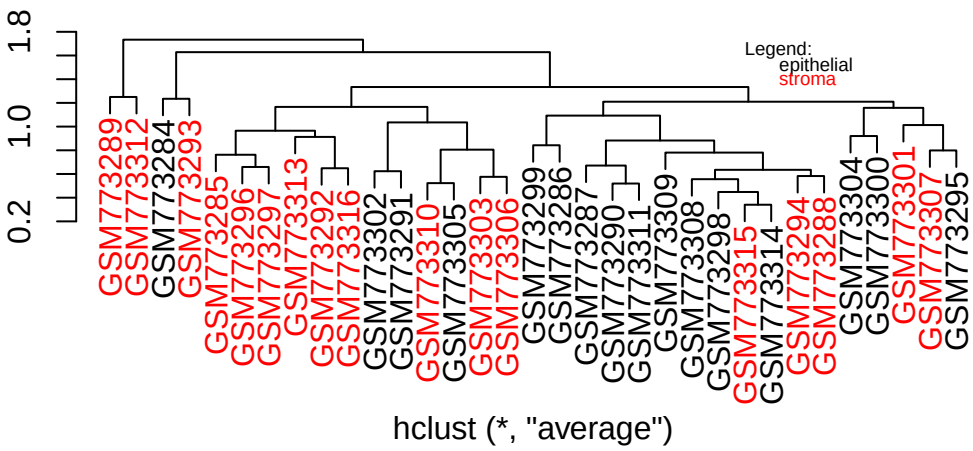


Figura 7: Cluster das 33 amostras, usando apenas os 20 genes que melhor distinguem as amostras associadas à gravidez das não associadas, usando correlação de pearson. A vermelho são as amostras provenientes do estroma e a preto as do epitélio



Nas figuras 8, 9 e 10, é possível observar os clusters correspondentes ao atributos dos metadados "specimen". O cluster da figura 9, que corresponde ao cluster hierárquico da distância eucledianа de todos os genes, permite concluir que o uso das distâncias entre todos os genes revela não ser útil para agrupar de forma razoável as amostras nos grupos células tumorais e células normais. No entanto, quando se usa apenas os genes obtidos na expressão diferencial, o agrupamento das amostras melhora consideravelmente. Isto deve-de ao facto de os genes usados serem os que melhor ajudam na diferenciação entre os dois grupos de amostras mencionados. Deste modo, usando apenas os vinte melhores genes obtidos na análise de expressão diferencial, é possível agrupar as amostras nos dois grupos mencionados através da distância entre os genes. No que toca ao agrupamento das amostras através da monotonia (correlação) dos vinte melhores genes, ilustrado na figura 10, é possível perceber que este agrupamento também permite separar as amostras em dois grupos um pouco distantes.

Figura 8: Cluster das 33 amostras, usando apenas os 20 genes que melhor distinguem as amostras provenientes do tumor ou de células normais. A vermelho são as amostras provennientes do tumor e a preto das células normais.

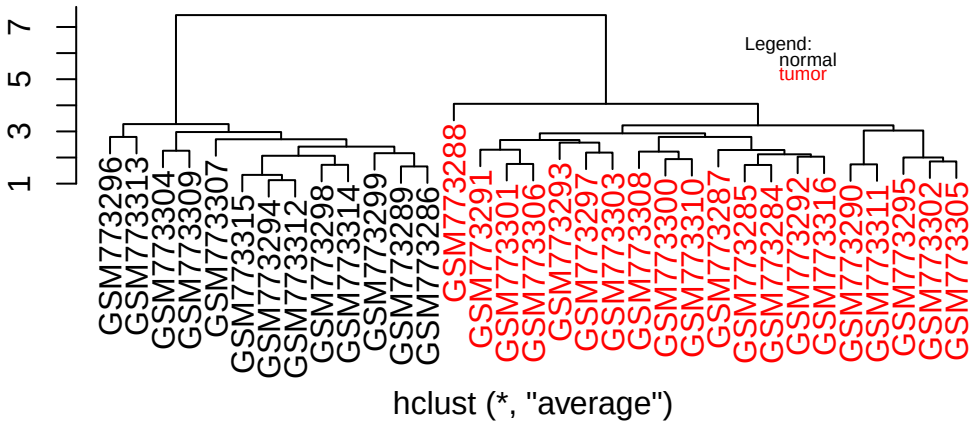


Figura 9: Cluster das 33 amostras, usando todos os genes. A vermelho são as amostras provenientes do tumor e a preto das células normais.

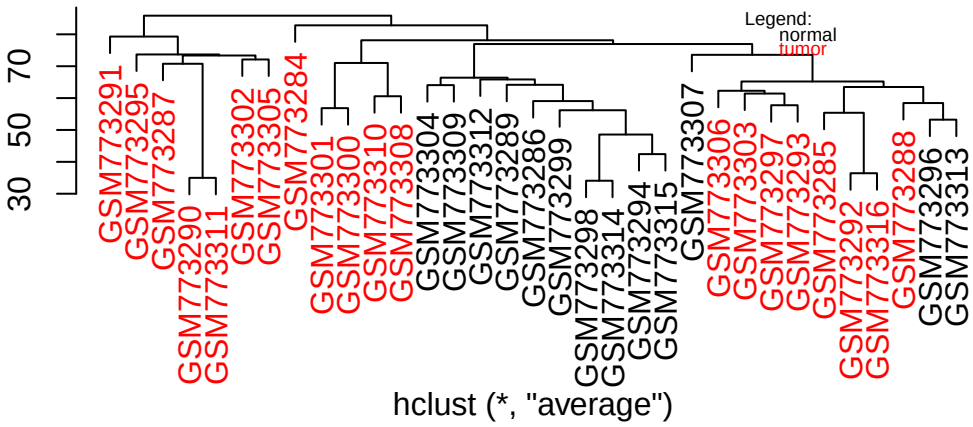
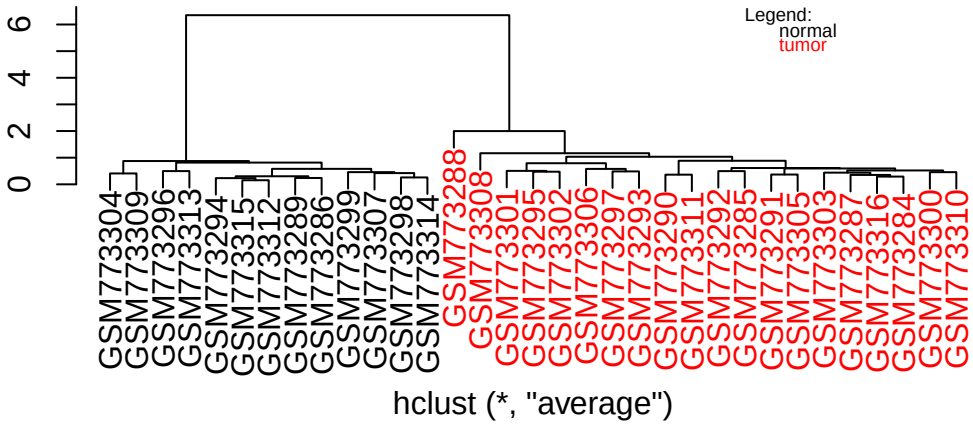


Figura 10: Cluster das 33 amostras, usando apenas os 20 genes que melhor distinguem as amostras associadas à gravidez das não associadas, usando correlação de pearson. A vermelho são as amostras provenientes do tumor e a preto das células normais



Nas figuras 11, 12 e 13, é possível observar os clusters correspondentes ao atributos dos metadados "genotype". O cluster da figura 12, que corresponde ao cluster hierárquico da distância euclediana de todos os genes, permite concluir que o uso das distâncias entre todos os genes revela não ser útil para agrupar de forma razoável as amostras nos grupos receptores de estrogénio positivos e negativos. APesar de quando se usa apenas os genes obtidos na expressão diferencial o agrupamento das amostras melhorar, não é bom o suficiente para se ver uma separação clara entre os dois grupos mencionados. Com isto, pode-se concluir que, apesar de os genes poderem estar até relacionados com receptores de estrogénios, os genes do dataset não serem os ideais para permitir um agrupamento claro destes dois grupos. No que toca ao agrupamento das amostras através da monotonia (correlação) dos vinte melhores genes, ilustrado na figura 13, é possível perceber que este agrupamento também não permite separar as amostras nos dois grupos.

Figura 11: Cluster das 33 amostras, usando apenas os 20 genes que melhor distinguem as amostras relacionadas com receptores de estrogénio positivo das com receptores de esttrogénio negativo. A vermelho são as amostras com receptores positivos os negativos.

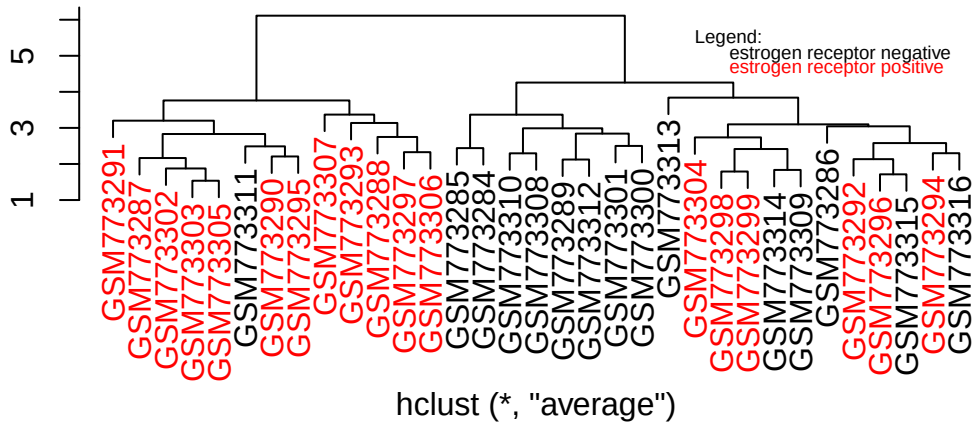


Figura 12: Cluster das 33 amostras, usando todos os genes. A vermelho são as amostras com receptores positivos e a preto os negativos.

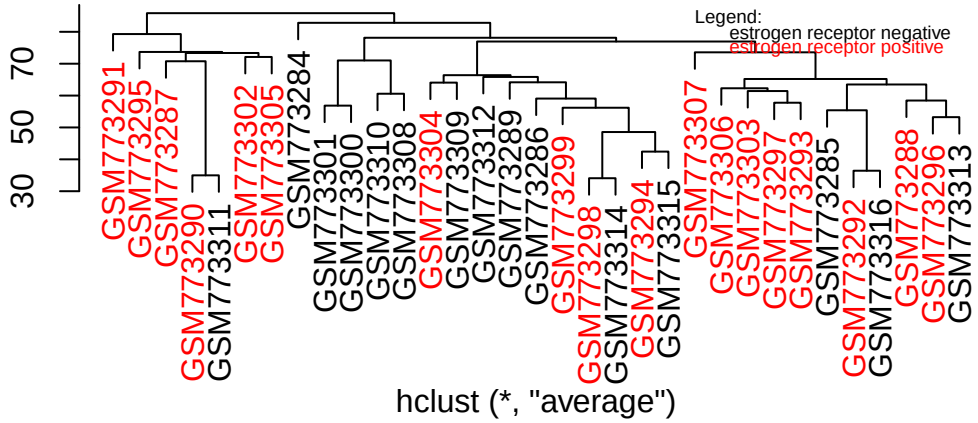
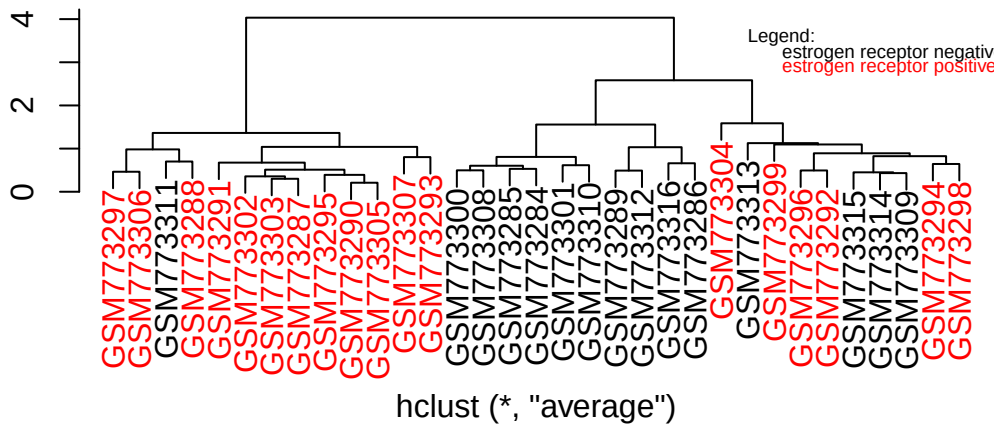


Figura 13: Cluster das 33 amostras, usando apenas os 20 genes que melhor distinguem as amostras associadas à gravidez das não associadas, usando correlação de pearson. A vermelho são as amostras com receptores positivos e a preto os negativos.



Análise Preditiva Usando o package MLInterfaces, realizou-se apenas o treino de modelos para prever o atributo dos metadados "disease.state": se associada à gravidez ou não. Os modelos treinados foram K-Nearest Neighbours, Naive Bayes, árvore de decisão e random forests. Estes dados não serão mostrados no relatório por não serem relevantes na discussão dos resultados obtidos, mas o código para a sua execução está disponível no ficheiro Rmd em anexo. Deste modo, apenas serão mostrados os resultados referentes ao package caret, tanto para prever o atributos dos metadados "disease.state" como para o atributo "specimen", que são compostos pela matriz de confusão do melhor modelo obtido para cada tipo de modelo, com respectiva especificação dos parâmetros, a accuracy média do modelo e as vintes variáveis mais importantes na previsão do modelo.

O primeiro modelo realizado para "disease.state" foi o modelo Knn, obtendo-se 13 como o melhor valor para o parâmetro k do modelo. A accuracy média obtida é 0.597. Na tabela 9, observa-se a matriz de confusão correspondente, sendo possível concluir que este modelo é mau a prever as amostras que pertencem ao grupo cancro não-associado à gravidez. No que toca às variáveis mais importantes do modelo, que podem ser observadas na tabela 10, nota-se que a grande maioria destes genes fazem parte dos vinte melhores genes obtidos na análise de expressão diferencial para o atributo "disease.state", o que seria algo de esperar, uma vez que revelaram ser os genes que mais ajudam na diferenciação entre os dois grupos de amostras. É importante também salientar que os genes PLIN5X, LOC101927967, LOC100288860, FCRL1, LOC105374809 e TAP2 são os genes presentes nas variáveis mais importantes para este modelo mas que não fazem parte dos 20 melhores genes da expressão diferencial para o atributo "disease.state".

Tabela 9: Matriz de Confusão para o melhor modelo knn, cujo parâmetro k é 13, para a previsão do atributo dos metadados "disease.state".

		Valores Reais	
		Non-Pregnancy-Associated Breast Cancer	Pregnancy-Associated Breast Cancer
Valores Previstos	Non-Pregnancy-Associated Breast Cancer	0.9	1.8
	Pregnancy-Associated Breast Cancer	38.5	58.8

Tabela 10: Tabela dos 20 genes mais importantes no melhor modelo knn, cujo parâmetro k é 13, para a previsão do atributo dos metadados "disease.state".

Nome do Gene	Nome da Variável	Importância
PRKXP1	1559186_at	100.00
PRKXP1	1559188_x_at	99.33
LY6K	223687_s_at	96.00
CXCL13	205242_at	94.00
PLIN5	241368_at	92.00
Nome Desconhecido	230910_s_at	91.33
LOC101927967	1557778_at	90.67
Nome Desconhecido	239678_at	90.67
G0S2	213524_s_at	90.00
GSTT1	203815_at	89.33
PNMAL1	218824_at	89.33
LOC100288860	239127_at	88.67
Nome Desconhecido	1561489_at	88.67
COL2A1	217404_s_at	88.00
ELOVL2	220029_at	88.00
ZNF334	238566_at	88.00
FCRL1	235982_at	87.33
LOC105374809	1556173_a_at	86.67
TAP2	208428_at	86.00
Nome Desconhecido	241835_at	86.00

O modelo naive-Bayes para a classe "disease.state", cujo parâmetro useKernel é falso, possui accuracy média de 0.6567. Na tabela 11, observa-se a matriz de confusão correspondente, sendo possível concluir que este modelo prevê de forma um pouco pior as amostras que pertencem ao grupo cancro não-associado à gravidez do que as amostras que pertencem ao grupo cancro associado à gravidez. No que toca às variáveis mais importantes do modelo, que podem ser observadas na tabela 12, nota-se que foram as mesmas obtidas para o modelo Knn para "disease.state", o que era algo de esperar, uma vez que se tratam de modelos onde ambos pretendem prever a classe "disease.state".

Tabela 11: Matriz de Confusão para o melhor modelo naiveBayes, cujo parâmetro usekernel é falso, para a previsão da classe "disease.state".

		Valores Reais	
		Non-Pregnancy-Associated Breast Cancer	Pregnancy-Associated Breast Cancer
Valores Previstos	Non-Pregnancy-Associated Breast Cancer	14.5	9.4
	Pregnancy-Associated Breast Cancer	24.8	51.2

Tabela 12: Tabela dos 20 genes mais importantes no melhor modelo naiveBayes, cujo parâmetro usekernel é falso, para a previsão do atributo dos metadados "disease.state".

Nome do Gene	Nome da Variável	Importância
PRKXP1	1559186_at	100.00
PRKXP1	1559188_x_at	99.33
LY6K	223687_s_at	96.00
CXCL13	205242_at	94.00
PLIN5	241368_at	92.00
Nome Desconhecido	230910_s_at	91.33

Nome do Gene	Nome da Variável	Importância
LOC101927967	1557778_at	90.67
Nome Desconhecido	239678_at	90.67
G0S2	213524_s_at	90.00
GSTT1	203815_at	89.33
PNMAL1	218824_at	89.33
LOC100288860	239127_at	88.67
Nome Desconhecido	1561489_at	88.67
COL2A1	217404_s_at	88.00
ELOVL2	220029_at	88.00
ZNF334	238566_at	88.00
FCRL1	235982_at	87.33
LOC105374809	1556173_a_at	86.67
TAP2	208428_at	86.00
Nome Desconhecido	241835_at	86.00

O modelo de árvore de decisão para o atributo dos metadados "disease.state", cujo melhor valor para o parâmetro de complexidade se revelou ser 0.6923077, possui accuracy média de 0.6242. Na tabela 13, observa-se a matriz de confusão correspondente, sendo possível concluir que este modelo prevê de forma um pouco pior as amostras que pertencem ao grupo cancro associado à gravidez, quando comparado com os modelos anteriores, apesar de as amostras que pertencem ao grupo cancro não-associado à gravidez serem melhor previstas quando comparadas com os mesmos modelos. No que toca às variáveis mais importantes do modelo, que podem ser observadas na tabela 14, nota-se que, mais uma vez, foram as mesmas obtidas para os modelos Knn e naive-Bayes para "disease.state".

Tabela 13: Matriz de Confusão para o melhor modelo de árvore de decisão, cujo parâmetro de complexidade é 0.6923077, para a previsão do atributo dos metadados "disease.state".

		Valores Reais	
		Non-Pregnancy-Associated Breast Cancer	Pregnancy-Associated Breast Cancer
Valores Previstos	Non-Pregnancy-Associated Breast Cancer	21.2	19.4
	Pregnancy-Associated Breast Cancer	18.2	41.2

Tabela 14: Tabela dos 20 genes mais importantes no melhor modelo de árvore de decisão, cujo parâmetro de complexidade é 0.6923077, para a previsão do atributo dos metadados "disease.state".

Nome do Gene	Nome da Variável	Importância
PRKXP1	1559186_at	100.00
PRKXP1	1559188_x_at	99.33
LY6K	223687_s_at	96.00
CXCL13	205242_at	94.00
PLIN5	241368_at	92.00
Nome Desconhecido	230910_s_at	91.33
LOC101927967	1557778_at	90.67
Nome Desconhecido	239678_at	90.67
G0S2	213524_s_at	90.00
GSTT1	203815_at	89.33
PNMAL1	218824_at	89.33

Nome do Gene	Nome da Variável	Importância
LOC100288860	239127_at	88.67
Nome Desconhecido	1561489_at	88.67
COL2A1	217404_s_at	88.00
ELOVL2	220029_at	88.00
ZNF334	238566_at	88.00
FCRL1	235982_at	87.33
LOC105374809	1556173_a_at	86.67
TAP2	208428_at	86.00
Nome Desconhecido	241835_at	86.00

O modelo de Random Forests para o atributo dos metadados "disease.state", cujo melhor valor para o parâmetro mtry é 6378, possui accuracy média de 0.6333. Na tabela 15, observa-se a matriz de confusão correspondente, sendo possível concluir que o uso de várias árvores de decisão para prever as amostras aumenta ligeiramente a accuracy média em relação ao modelo da árvore de decisão. Isto porque, apesar de a capacidade de previsão de amostras de cancro não-associado à gravidez serem pior previstas, a previsão das amostras de cancro associado à gravidez melhora bastante. No entanto, não é o suficiente para prever melhor, ou seja, ter uma melhor accuracy, que o modelo naive-Bayes. No que toca às variáveis mais importantes do modelo, que podem ser observadas na tabela 16, já existem algumas diferenças em relação às variáveis obtidas no modelos anteriores, apesar de a maioria continuar a pertencer aos melhores genes obtidos na análise de expressão diferencial do atributo "disease.state". É importante também salientar que os genes PLIN5X, LOC100288860, FCRL1, HLA-DRB4, CD36, DOCK9-AS2, DNAJC28 e CYP1B1 são os genes presentes nas variáveis mais importantes para este modelo mas que não fazem parte dos 20 melhores genes da expressão diferencial para o atributo "disease.state".

Tabela 15: Matriz de Confusão para o melhor modelo random Forests, cujo parâmetro mtry é 6378, para a previsão do atributo dos metadados "disease.state".

		Valores Reais	
		Non-Pregnancy-Associated Breast Cancer	Pregnancy-Associated Breast Cancer
Valores Previstos	Non-Pregnancy-Associated Breast Cancer	11.5	8.8
	Pregnancy-Associated Breast Cancer	27.9	51.8

Tabela 16: Tabela dos 20 genes mais importantes no melhor modelo random Forests, cujo parâmetro mtry é 6378, para a previsão do atributo dos metadados "disease.state".

Nome do Gene	Nome da Variável	Importância
GSTT1	203815_at	100.00
CXCL13	205242_at	87.68
HLA-DRB4	209728_at	81.63
PRKXP1	1559188_x_at	58.91
PLIN5	241368_at	41.78
B4GALT6	206232_s_at	39.19
Nome Desconhecido	241835_at	29.73
LOC100288860	239127_at	18.22
PRKXP1	1559186_at	17.10
LY6K	223687_s_at	16.47
Nome Desconhecido	238632_at	15.27

Nome do Gene	Nome da Variável	Importância
FCRL1	235982_at	14.43
CD3G	206804_at	14.41
DOCK9-AS2	231062_at	14.24
Nome Desconhecido	237023_at	13.98
DNAJC28	220372_at	13.46
G0S2	213524_s_at	13.11
CYP1B1	202437_s_at	11.68
ELOVL2	213712_at	11.62
COL2A1	213492_at	11.18

O modelo de Partial Least Squares (PLS) para o atributo dos metadados "disease.state", cujo melhor valor para o parâmetro de número de componentes é 4, possui accuracy média de 0.7424. Na tabela 17, observa-se a matriz de confusão correspondente, sendo possível concluir que, no geral, este modelo consegue prever melhor tanto o grupo de cancro não-associado à gravidez como o grupo de cancro associado à gravidez que os modelos já mencionados. No que toca às variáveis mais importantes do modelo, que podem ser observadas na tabela 18, existem bastantes diferenças em relação às variáveis obtidas em relação aos vinte melhores genes obtidos na análise de expressão diferencial do atributo "disease.state", sendo 15, em 20, o número de genes que não pertence aos genes obtidos pela análise de expressão diferencial. No entanto, é importante ter em mente que estes genes que não pertencem aos 20 melhores obtidos na expressão diferencial podem ser genes que também são dos melhores mas que não figuram os 20 melhores. O mesmo pode acontecer para os genes obtidos nos modelos anteriores e que não fazem parte dos vinte melhores genes obtidos na análise de expressão diferencial para o atributo "disease.state".

Tabela 17: Matriz de Confusão para o melhor modelo Partial Least Squares, cujo parâmetro de número de componentes é 4, para a previsão do atributo dos metadados "disease.state".

		Valores Reais	
		Non-Pregnancy-Associated Breast Cancer	Pregnancy-Associated Breast Cancer
Valores Previstos	Non-Pregnancy-Associated Breast Cancer	23.0	9.4
	Pregnancy-Associated Breast Cancer	16.4	51.2

Tabela 18: Tabela dos 20 genes mais importantes no melhor modelo Partial Least Squares, cujo parâmetro de número de componentes é 4, para a previsão do atributo dos metadados "disease.state".

Nome do Gene	Nome da Variável	Importância
HLA-DRB4	209728_at	100.00
CXCL13	205242_at	97.33
MAGEA6	214612_x_at	89.13
MSMB	210297_s_at	81.27
Nome Desconhecido	209942_x_at	79.82
Nome Desconhecido	238632_at	76.53
IGSF1	207695_s_at	76.26
TNFSF11	210643_at	76.16
CPB1	205509_at	75.84
MSMB	207430_s_at	73.79
HLA-DQA1	203290_at	73.10
COL2A1	213492_at	69.55

Nome do Gene	Nome da Variável	Importância
HOTAIR	239153_at	69.20
CSN1S1	208350_at	66.69
MAGEA12	210467_x_at	66.36
FCRL1	235982_at	65.94
Nome Desconhecido	214603_at	63.47
PRKXP1	1559186_at	63.31
MS4A1	210356_x_at	63.29
PLEKHS1	1554190_s_at	63.05

O modelo Máquinas de Vectores de Suporte (SVMs) possui uma accuracy média obtida é 0.7061. Na tabela 19, observa-se a matriz de confusão correspondente, sendo os seus valores semelhantes aos obtidos com o modelo de árvore de decisão, já mencionado anteriormente, no que toca à previsão das amostras de cancro não-associado à gravidez. No entanto, a previsão do outro grupo de amostras é melhor. No que toca às variáveis mais importantes do modelo, que podem ser observadas na tabela 20, nota-se que, mais uma vez, foram as mesmas obtidas para os modelos Knn, naive-Bayes e árvore de decisão para "disease.state".

Tabela 19: Matriz de Confusão para o melhor modelo Máquinas de Vectores de Suporte para a previsão do atributo dos metadados "disease.state".

		Valores Reais	
		Non-Pregnancy-Associated Breast Cancer	Pregnancy-Associated Breast Cancer
Valores Previstos	Non-Pregnancy-Associated Breast Cancer	21.2	11.2
	Pregnancy-Associated Breast Cancer	18.2	49.4

Tabela 20: Tabela dos 20 genes mais importantes no melhor modelo Máquinas de Vectores de Suporte para a previsão do atributo dos metadados "disease.state".

Nome do Gene	Nome da Variável	Importância
PRKXP1	1559186_at	100.00
PRKXP1	1559188_x_at	99.33
LY6K	223687_s_at	96.00
CXCL13	205242_at	94.00
PLIN5	241368_at	92.00
Nome Desconhecido	230910_s_at	91.33
LOC101927967	1557778_at	90.67
Nome Desconhecido	239678_at	90.67
G0S2	213524_s_at	90.00
GSTT1	203815_at	89.33
PNMAL1	218824_at	89.33
LOC100288860	239127_at	88.67
Nome Desconhecido	1561489_at	88.67
COL2A1	217404_s_at	88.00
ELOVL2	220029_at	88.00
ZNF334	238566_at	88.00
FCRL1	235982_at	87.33
LOC105374809	1556173_a_at	86.67
TAP2	208428_at	86.00

Nome do Gene	Nome da Variável	Importância
Nome Desconhecido	241835_at	86.00

Com base em toda esta informação, o modelo Partial Least Squares (PLS), com 4 componentes, parece ser o modelo que melhor se adequa a prever o atributo dos metadados "disease.state", ou seja, a distinguir se as amostras pertencem a um cancro não-associado à gravidez ou de um cancro associado à gravidez. Isto porque é o modelo que apresenta melhor accuracy média de entre todos os modelos treinados e testados.

O primeiro modelo realizado para "specimen" foi o modelo Knn, obtendo-se 5 como o melhor valor para o parâmetro k do modelo. A accuracy média obtida é 0.9424. Na tabela 21, observa-se a matriz de confusão correspondente, sendo possível ver que, de facto, o modelo prevê com bastante qualidade as amostras de células tumorais e amostras de células normais. No que toca às variáveis mais importantes do modelo, que podem ser observadas na tabela 22, nota-se que a grande maioria destes genes fazem parte dos vinte melhores genes obtidos na análise de expressão diferencial para o atributo "specimen", o que seria algo de esperar, uma vez que revelaram ser os genes que mais ajudam na diferenciação entre os dois grupos de amostras. É importante também salientar que os genes OAS2, SCN4B, FPR3, SULT1C2 e SEMA5A são os genes presentes nas variáveis mais importantes para este modelo mas que não fazem parte dos 20 melhores genes da expressão diferencial para o atributo "specimen".

Tabela 21: Matriz de Confusão para o melhor modelo knn, cujo parâmetro k é 5, para a previsão do atributo dos metadados "specimen".

		Valores Reais	
		Normal	Tumor
Valores Previstos	Normal	35.8	2.1
	Tumor	3.6	58.5

Tabela 22: Tabela dos 20 genes mais importantes no melhor modelo Máquinas de Vectores de Suporte para a previsão do atributo dos metadados "specimen".

Nome do Gene	Nome da Variável	Importância
CD300LG	1552509_a_at	100.00
MYH11	201497_x_at	100.00
MYH11	207961_x_at	100.00
SNX10	218404_at	100.00
PPP1R14A	227006_at	100.00
OAS2	204972_at	99.40
SAMD5	242626_at	99.40
MYH11	201496_x_at	99.80
ADAMDEC1	206134_at	99.80
SAMD5	228653_at	99.80
SCN4B	236359_at	99.80
Nome Desconhecido	236414_at	99.80
Nome Desconhecido	236647_at	99.80
ACO1	237622_at	99.80
FPR3	214560_at	98.19
SULT1C2	205342_s_at	97.59
Nome Desconhecido	206742_at	97.59
COL10A1	217428_s_at	97.59
Nome Desconhecido	1563295_at	96.99
SEMA5A	213169_at	96.99

Nome do Gene	Nome da Variável	Importância

O modelo naive-Bayes para o atributo dos metadados "specimen", cujo parâmetro useKernel é falso, possui accuracy média de 0.9394. Na tabela 23, observa-se a matriz de confusão correspondente, sendo possível concluir que este modelo consegue prever, aparentemente, na totalidade as amostras provenientes de um tumor. No entanto, nem todas as amostras de células normais foram correctamente classificadas, originando falsos positivos no que toca à existência de tumor, tendo de facto sido observado uma pior previsão deste grupo de amostras em relação ao modelo Knn, sendo esta a razão pela qual a accuracy deste modelo é menor que a do modelo Knn. No que toca às variáveis mais importantes do modelo, que podem ser observadas na tabela 24, nota-se que foram as mesmas obtidas para o modelo Knn para "specimen", o que era algo de esperar, uma vez que se tratam de modelos onde ambos pretendem prever o atributo "specimen".

Tabela 23: Matriz de Confusão para o melhor modelo naiveBayes, cujo parâmetro usekernel é falso, para a previsão do atributo dos metadados "specimen".

		Valores Reais	
		Normal	Tumor
Valores Previstos	Normal	33.3	0.0
	Tumor	6.1	60.6

Tabela 24: Tabela dos 20 genes mais importantes no melhor modelo naiveBayes, cujo parâmetro usekernel é falso, para a previsão do atributo dos metadados "specimen".

Nome do Gene	Nome da Variável	Importância
CD300LG	1552509_a_at	100.00
MYH11	201497_x_at	100.00
MYH11	207961_x_at	100.00
SNX10	218404_at	100.00
PPP1R14A	227006_at	100.00
OAS2	204972_at	99.40
SAMD5	242626_at	99.40
MYH11	201496_x_at	99.80
ADAMDEC1	206134_at	99.80
SAMD5	228653_at	99.80
SCN4B	236359_at	99.80
Nome Desconhecido	236414_at	99.80
Nome Desconhecido	236647_at	99.80
ACO1	237622_at	99.80
FPR3	214560_at	98.19
SULT1C2	205342_s_at	97.59
Nome Desconhecido	206742_at	97.59
COL10A1	217428_s_at	97.59
Nome Desconhecido	1563295_at	96.99
SEMA5A	213169_at	96.99

O modelo de árvore de decisão para o atributo dos metadados "specimen", cujo melhor valor para o parâmetro de complexidade se revelou ser 0.75, possui accuracy média de 0.9636. Na tabela 25, observa-se a matriz de confusão correspondente, sendo possível ver que, de facto, no geral, a previsão das amostras é melhor que os outros modelos até agora mencionados. No que toca às variáveis mais importantes do modelo,

que podem ser observadas na tabela 26, nota-se que alguns dos genes não são os mesmos observados para os modelos anteriores. Para além disto, é possível ver que cinco variáveis possuem importância 100, sendo eles CD300LG, MYH11, SNX10 e PPP1R14A, todos eles pertencentes ao grupo de vinte melhores genes na análise de expressão diferencial para o atributo "specimen". Todas as restantes quinze possuem importância 0, o que pode indicar que apenas os cinco genes mencionados foram importantes na construção da árvore de decisão.

Tabela 25: Matriz de Confusão para o melhor modelo de árvore de decisão, cujo parâmetro de complexidade é 0.75, para a previsão do atributo dos metadados "specimen".

		Valores Reais	
		Normal	Tumor
Valores Previstos	Normal	37.3	1.5
	Tumor	2.1	59.1

Tabela 26: Tabela dos 20 genes mais importantes no melhor modelo de árvore de decisão, cujo parâmetro de complexidade é 0.75, para a previsão do atributo dos metadados "specimen".

Nome do Gene	Nome da Variável	Importância
CD300LG	1552509_a_at	100
MYH11	201497_x_at	100
MYH11	207961_x_at	100
SNX10	218404_at	100
PPP1R14A	227006_at	100
RFC2	1053_at	0
CCL5	1405_i_at	0
CYP2A6	1494_f_at	0
MAPK1	1552263_at	0
ADAM32	1552266_at	0
SPATA17	1552269_at	0
SLC46A1	1552278_a_at	0
CILP2	1552289_a_at	0
GIMAP1	1552316_a_at	0
GIMAP1	1552318_at	0
FAM122C	1552322_at	0
CFAP53	1552325_at	0
CFAP53	1552326_a_at	0
CENPBD1	1552330_at	0
MCMDC2	1552359_at	0

O modelo de Random Forests para o atributo dos metadados "specimen", cujo melhor valor para o parâmetro mtry é 6378, possui accuracy média de 0.9818. Na tabela 27, observa-se a matriz de confusão correspondente, sendo possível concluir que, de facto, o uso de várias árvores de decisão para prever as amostras aumenta a accuracy média em relação ao modelo da árvore de decisão. No que toca às variáveis mais importantes do modelo, que podem ser observadas na tabela 28, a sua grande maioria continua a fazer parte dos vinte melhores genes na análise de expressão diferencial para "specimen". É importante também salientar que os genes OAS2, MYBL1, SCN4B, SEMA5A e CASC5 são os genes presentes nas variáveis mais importantes para este modelo mas que não fazem parte dos 20 melhores genes da expressão diferencial para o atributo "specimen".

Tabela 27: Matriz de Confusão para o melhor modelo random Forests, cujo parâmetro mtry é 6378,

para a previsão do atributo dos metadados “specimen”.

		Valores Reais	
		Normal	Tumor
Valores Previstos	Normal	37.9	0.3
	Tumor	1.5	60.3

Tabela 28: Tabela dos 20 genes mais importantes no melhor modelo random Forests, cujo parâmetro mtry é 6378, para a previsão do atributo dos metadados “specimen”.

Nome do Gene	Nome da Variável	Importância
MYH11	207961_x_at	100
CD300LG	1552509_a_at	96.24
MYH11	201497_x_at	88.40
OAS2	204972_at	84.24
PPP1R14A	227006_at	75.52
ACO1	237622_at	57.92
SNX10	218404_at	57.11
SAMD5	228653_at	51.80
MYBL1	213906_at	46.94
SCN4B	236359_at	46.50
Nome Desconhecido	236414_at	42.34
SAMD5	242626_at	38.55
ADAMDEC1	206134_at	36.63
MYH11	201496_x_at	35.71
DLK1	209560_s_at	34.46
SEMA5A	229427_at	33.21
Nome Desconhecido	206742_at	27.47
CASC5	1552682_a_at	25.48
Nome Desconhecido	236647_at	24.12
SEMA5A	213169_at	24.04

O modelo de Partial Least Squares (PLS) para o atributo dos metadados “specimen”, cujo melhor valor para o parâmetro de número de componentes é 5, possui accuracy média de 0.9879. Na tabela 29, observa-se a matriz de confusão correspondente, sendo possível ver que, no geral, este modelo é o modelo que, até agora, consegue fazer melhor a previsão das classes. Isto porque, aparentemente, as amostras de células tumorais foram sempre correctamente classificadas e as amostras de células normais foram muito pouco incorrectamente classificadas. No que toca às variáveis mais importantes do modelo, que podem ser observadas na tabela 30, existem bastantes diferenças em relação às variáveis obtidas em relação aos vinte melhores genes obtidos na análise de expressão diferencial do atributo “disease.state”, sendo 11, em 20, o número de genes que não pertence aos genes obtidos pela análise de expressão diferencial. No entanto, mais uma vez, é importante ter em mente que estes genes que não pertencem aos 20 melhores obtidos na expressão diferencial podem ser genes que também são dos melhores mas que não figuram os 20 melhores. O mesmo pode acontecer para os genes obtidos nos modelos anteriores e que não fazem parte dos vinte melhores genes obtidos na análise de expressão diferencial para o atributo “specimen”.

Tabela 29: Matriz de Confusão para o melhor modelo Partial Least Squares, cujo parâmetro de número de componentes é 5, para a previsão do atributo dos metadados “specimen”.

		Valores Reais	
		Normal	Tumor
Valores Previstos	Normal	38.2	0.0
	Tumor	1.2	60.6

Tabela 30: Tabela dos 20 genes mais importantes no melhor modelo Partial Least Squares, cujo parâmetro de número de componentes é 5, para a previsão do atributo dos metadados "specimen".

Nome do Gene	Nome da Variável	Importância
CXCL11	211122_s_at	100
DLK1	209560_s_at	94.80
C2orf40	223623_at	94.59
TAC1	206552_s_at	92.84
ADAMDEC1	206134_at	83.80
Nome Desconhecido	206742_at	83.65
CD300LG	1552509_a_at	83.37
CAPN6	202965_s_at	82.89
COL10A1	217428_s_at	78.22
PIGR	204213_at	72.66
LRRC15	213909_at	72.36
EDN3	208399_s_at	71.77
NEK2	211080_s_at	70.33
WIF1	204712_at	68.42
Nome Desconhecido	227843_at	67.92
Nome Desconhecido	207016_s_at	67.67
BMPR1B	210523_at	67.35
BMPR1B	242579_at	67.13
SCARA5	229839_at	66.79
CDKN2B	236313_at	66.06

O modelo Máquinas de Vectores de Suporte (SVMs) possui uma accuracy média obtida é 0.9667. Na tabela 31, observa-se a matriz de confusão correspondente, sendo os seus valores semelhantes aos obtidos com o modelo de árvore de decisão, já mencionado anteriormente, no que toca à previsão das amostras de células normais. No entanto, a previsão do outro grupo de amostras é ligeiramente melhor. No que toca às variáveis mais importantes do modelo, que podem ser observadas na tabela 32, nota-se que, mais uma vez, foram as mesmas obtidas para os modelos Knn e naive-Bayes para "specimen".

Tabela 31: Matriz de Confusão para o melhor modelo Máquinas de Vectores de Suporte para a previsão do atributo dos metadados "specimen".

		Valores Reais	
		Normal	Tumor
Valores Previstos	Normal	37.3	1.2
	Tumor	2.1	59.4

Tabela 32: Tabela dos 20 genes mais importantes no melhor modelo Máquinas de Vectores de Suporte para a previsão do atributo dos metadados "specimen".

Nome do Gene	Nome da Variável	Importância
CD300LG	1552509_a_at	100.00
MYH11	201497_x_at	100.00
MYH11	207961_x_at	100.00
SNX10	218404_at	100.00
PPP1R14A	227006_at	100.00
OAS2	204972_at	99.40
SAMD5	242626_at	99.40
MYH11	201496_x_at	99.80
ADAMDEC1	206134_at	99.80
SAMD5	228653_at	99.80
SCN4B	236359_at	99.80
Nome Desconhecido	236414_at	99.80
Nome Desconhecido	236647_at	99.80
ACO1	237622_at	99.80
FPR3	214560_at	98.19
SULT1C2	205342_s_at	97.59
Nome Desconhecido	206742_at	97.59
COL10A1	217428_s_at	97.59
Nome Desconhecido	1563295_at	96.99
SEMA5A	213169_at	96.99

Com base em toda esta informação, o modelo Partial Least Squares (PLS), com 5 componentes, parece ser o modelo que melhor se adequa, mais uma vez, a prever o atributo "specimen", ou seja, a distinguir se as amostras pertencem a células tumorais ou a células normais. Isto porque é o modelo que apresenta melhor accuracy média de entre todos os modelos treinados e testados.

Selecção de atributos Nesta análise do dataset, obteve-se um plot para cada atributo dos metadados usado na análise preditiva ("disease.state" e "specimen") que demonstra a evolução da accuracy ao longo dos vários conjuntos de variáveis seleccionados aleatoriamente pelo método escolhido.

No que toca ao atributo dos metadados "disease.state", foi possível observar que o melhor conjunto de atributos para a previsão de modelos é composto por oito variáveis. Tal informação pode ser observada na figura 14, bem como na figura 15, que se trata do mesmo plot presente na figura 14, mas com um menor limite no eixo dos xx, por forma a observar melhor a que número de variáveis corresponde o conjunto com maior Accuracy.

Figura 14: Plot da evolução da Accuracy ao longo dos conjuntos de variáveis seleccionados, na selecção de atributos para "disease.state". O conjunto de variáveis com maior accuracy possui a bola azul preenchida.

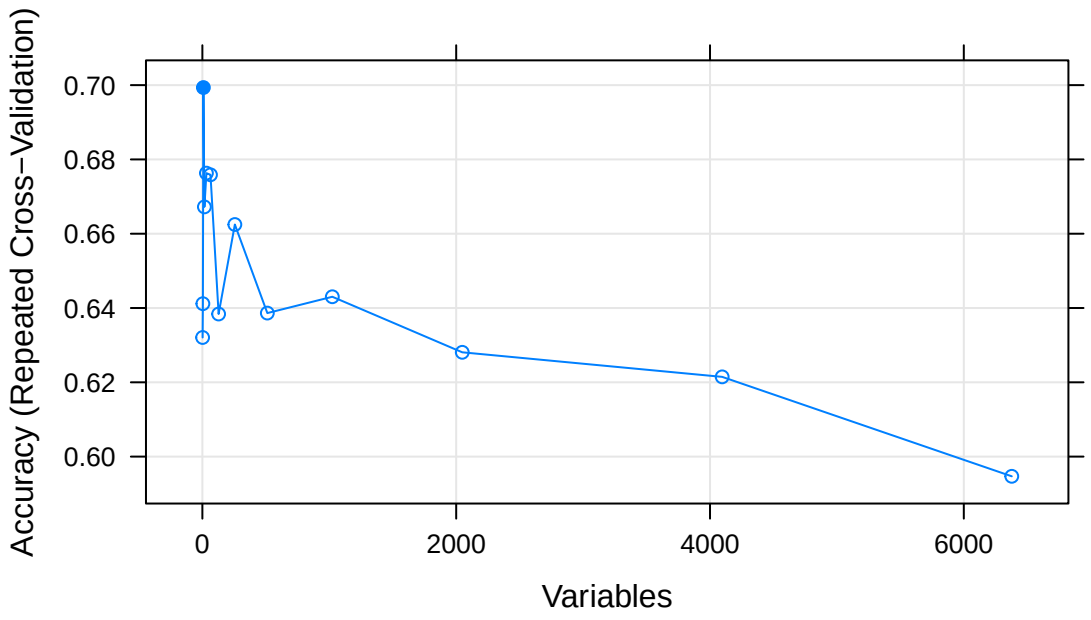
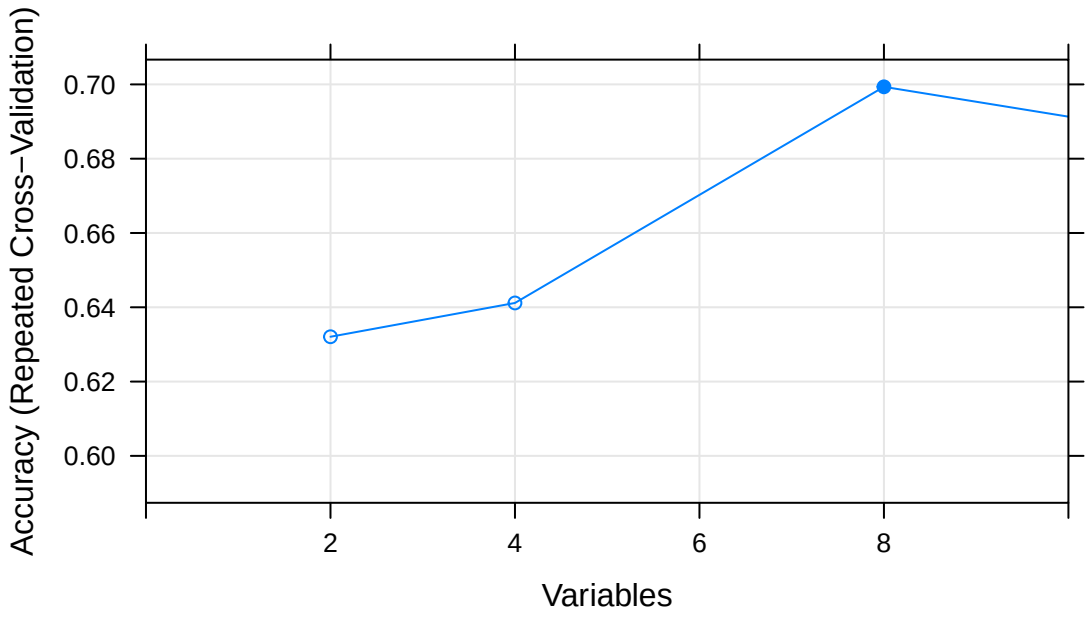


Figura 15: Reprodução do plot da figura xxxx, mas com os limites do eixo das abcissas menores. O conjunto de variáveis com maior accuracy possui a bola azul preenchida.



Ao se obter apenas oito variáveis, procurou-se saber quais os nomes dos genes destas variáveis, uma vez que poderia revelar alguma informação interessante. Deste modo, foi possível observar que todas as oito variáveis obtidas são genes que apareceram nas várias variáveis mais importantes para cada modelo relativo ao atributo dos metadados "disease.state". Cinco destas variáveis, de nome PRKXP1, 238632_at, B4GALT6, CXCL13 e GSTT1, fazem parte dos vinte melhores genes na análise de expressão diferencial para "disease.state".

Para além disto, surgiu a curiosidade de observar se, de facto, o uso de apenas estas variáveis iria provocar um aumento na accuracy dos modelos treinados. Após a realização desta tarefa, cujo código

e posteriores detalhes dos resultados pode ser consultado em anexo, constatou-se, de facto, um aumento considerável das accuracies destes modelos, para além de que o processo foi bem menos demorado.

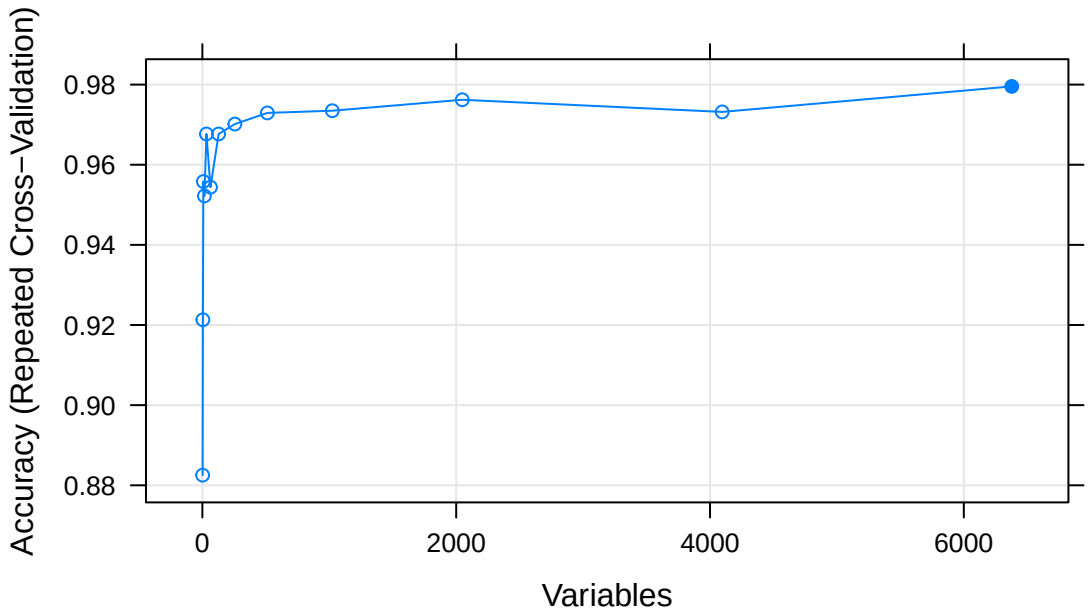
Através das accuracies observadas, o modelo Random Forests, com o parâmetro mty igual a dois, revelou ser o modelo que melhor se adequava a estimar se as amostras estão associadas à gravidez ou não, com os novos atributos seleccionados, com uma accuracy de 0.9545. No entanto, poderá ser preciso ter em atenção que a função escolhida para fazer a selecção de atributos foi random forests. O segundo melhor modelo obtido foi o modelo de Partial Least Squares (PLS), com o parâmetro de número de componentes igual a 3 e uma accuracy de 0.9455.

Tabela 33: Tabela com o conjunto de 8 variáveis seleccionadas para o atributo "disease.state".

Nome do Gene	Nome da Variável
LOC100288860	239127_at
GSTT1	203815_at
CXCL13	205242_at
B4GALT6	206232_s_at
HLA-DRB4	209728_at
PRKXP1	1559188_x_at
PLIN5	241368_at
Nome Desconhecido	238632_at

Por último, no que toca ao atributo dos metadados "specimen", foi possível observar que o melhor conjunto de atributos para a previsão de modelos é composto por todos os genes do dataset. Tal informação pode ser observada na figura 16.

Figura 16: Plot da evolução da Accuracy ao longo dos conjuntos de variáveis seleccionados, na selecção de atributos para o atributo "specimen". O conjunto de variáveis com maior accuracy possui a bola azul preenchida.



Conclusão

Primeiramente, a análise de expressão diferencial realizada sobre cada atributo dos metadados mostrou de facto a existência de genes diferencialmente expressos, dado que foram obtidos p-values bastante baixos em todos os casos. Após verificada a função dos genes dos atributos "cell.type" e "specimen" conclui-se que estes estão associados de forma evidente às condições de cada atributo, ou seja, são genes que permitem uma melhor distinção de cada condição analisada. No entanto para os atributos "disease.state" e "genotype", não foi tão evidente a associação da função dos genes com as suas condições. Mas é de salientar que os genes diferencialmente expressos destes atributos encontram-se bastante relacionados com as células tumorais.

No que diz respeito à análise de enriquecimento, novamente realizada para cada atributo dos metadados, verificou-se que os termos GO mais associados aos genes diferencialmente expressos estão dentro do que seria expectável, para todos os casos, permitindo um enriquecimento do nosso dataset. Por exemplo, os genes que permitem diferenciar células normais de células tumorais têm funções moleculares muito relacionadas com a ligação entre proteínas, regulação de atividade enzimática, e ligação da cadeia dupla de DNA.

Na realização do "clustering" das amostras de cada atributo dos metadados, foi sempre evidente que, usando apenas os genes obtidos na análise da expressão diferencial, o agrupamento das amostras através da distância euclediana melhora consideravelmente relativamente ao uso de todos os genes. Isto acontece pois os genes mais diferencialmente expressos ajudam na diferenciação entre os grupos de amostras de cada atributo. Para todos os atributos dos metadados, com exceção do "genotype" foi possível agrupar as amostras através da distância entre os genes. No que toca ao agrupamento das amostras através da correlação dos genes mais diferencialmente expressos este só agrupou as amostras em dois grupos distintos para os atributos "disease.state" e "specimen". Para o atributo "genotype" não foi possível uma separação clara entre os dois grupos de amostras (recetores de estrogénio positivos ou negativos), tanto através da distância entre os genes como através da sua correlação, isto permite concluir que apesar dos genes estarem relacionados com recetores de estrogénio, não permitem a separação das amostras dos dois grupos.

Para a anáilse preditiva dos atributos dos metadados "disease.state" e "specimen", o modelo PLS revelou ser, inicialmente, o melhor modelo para prever estes atributos, embora com parâmetros diferentes: número de componentes 4 no que toca a prever o atributo "disease.state" e 5 para prever o atributo "specimen".

Por fim, na selecção de atributos, foi possível observar que havia um menor conjunto de genes, neste caso oito, que levavam a um aumento da accuracy na precisão do atributo "disease.state", pelo qual se repetiu o treino de modelos só com estes genes. Isto levou a que se verificasse que, de facto, todos os modelos aumentaram a sua capacidade de precisão, especialmente o Random Forests, que se revelou o melhor. No que toca ao atributo dos metadados "specimen", o melhor conjunto de atributos para a previsão de modelos revelou ser todos os genes do dataset.

Trabalho Realizado Por: Andreia Campos Nº PG30384 José Dias Nº PG26550 Sara Cardoso Nº PG30386