Universidade do Minho
Escola de Engenharia

Catarina Isabel Ferreira Miranda Lemos

Seleção de genes diferencialmente expressos baseada em metodologia ROC (Receiver Operating Characteristic)

Abril de 2017

Universidade do Minho
Escola de Engenharia

Catarina Isabel Ferreira Miranda Lemos

Seleção de genes diferencialmente expressos baseada em metodologia ROC (Receiver Operating Characteristic)

Dissertação de Mestrado em Bioinformática

Trabalho efetuado sob a orientação de

Professora Doutora Ana Cristina Braga

Abril de 2017

Agradecimentos

"É na crise que nascem as invenções, os descobrimentos e as grandes estratégias. Quem supera a crise, supera-se a si mesmo".

Albert Einstein

A todos os que comigo caminharam, àqueles que comigo sonharam e a todos os que sempre acreditaram, o meu muito obrigada. Em particular,

- aos meus pais, pelo carinho e pelo amor de todos os dias. Obrigada pela força e pela vossa presença ao longo de toda esta jornada;
- ao André, irmão e amigo, pelos sorrisos e gargalhadas. Pelas palavras de coragem e de força, pela paciência e compreensão de todos os dias;
- à minha orientadora, Professora Doutora Ana Cristina Braga, por toda a ajuda e predisposição em enriquecer este trabalho. Obrigada pelo profissionalismo, pela motivação e dedicação em todos os momentos;
- à minha amiga Raquel por todo o apoio. A todos os meus amigos e amigas que ao longo deste ano viveram um pouco deste trabalho. Obrigada pela compreensão, pelos momentos de descontração e de felicidade;
- a todos vós, o meu muito obrigada. Apenas convosco foi possível aqui chegar... cada um de vós, como peça essencial para que este *puzzle* ficasse completo.

Resumo

A análise da expressão genética é essencial para uma identificação da função dos genes e para a identificação destes quando relacionados com doenças. Para a realização de um estudo em larga escala de mudanças na expressão genética é necessário encontrar um método que o faça com precisão e exatidão. Desta forma, foi aqui incluída, uma análise pela tecnologia de *microarrays*, uma ferramenta importante no diagnóstico de doenças.

A execução de um método que identificasse genes com regulação negativa e positiva e genes diferencialmente expressos simultaneamente, tornou-se, a principal motivação deste trabalho.

De entre as diferentes técnicas estatísticas, a metodologia ROC *(Receiver Operating Characteristic)* foi a escolhida para o efeito.

Quando se associa a metodologia ROC com a análise de dados de *microarrays* é possível ver que uma das principais aplicações é a identificação de grupos de genes associados ao desenvolvimento de qualquer patologia cancerígena. Para a análise deste último parâmetro é utilizado o *arrow plot* com a representação do OVL *(Overlapping Coefficient)* e da AUC *(Area Under the Curve)* para cada gene, numa experiência de *microarays* e comparar a sua eficácia com outros métodos existentes para o mesmo propósito.

Através da análise de um conjunto de dados de pacientes afetados pelo adenocarcinoma do pâncreas foi possível identificar os genes diferencialmente expressos, sendo este o principal objetivo do trabalho em questão.

Palavras-chave:*Microarrays*, genes, área abaixo da curva ROC, coeficiente de sobreposição, *arrow plot*.

Abstract

Genetic expression analysis is essential for the identification of gene function and when they are related with diseases. To perform a large-scale study of changes in gene expression it is necessary to find a method to do it with precision and accuracy. Thus, it was included here an analysis by microarray technology, an important tool in the diagnosis of diseases.

The execution of a method to identify genes with negative and positive regulation and differentially expressed genes simultaneously has become the main motivation of this work.

Among different statistical techniques, the receiver operating characteristic (ROC) was the chosen one.

When combining the ROC methodology with microarray data analysis it is possible to see that one of the main applications is the identification of gene groups associated with the development of any kind of cancer. For the analysis of this last parameter is used the arrow plot with the overlapping coefficient (OVL) and the area under the curve (AUC) representation for each gene of a microarray experience and compare its effectiveness with other existing methods for the same purpose.

Through the analysis of a set of affected patient data of pancreatic adenocarcinoma it was possible to identify differentially expressed genes, which is the main goal of this work.

Keywords: Microarrays, genes, area under the ROC curve, overlapping coefficient, arrow plot.

Índice

Lista de Tabelas

Lista de Figuras

Lista de Siglas e Acrónimos

ABCR *area between ROC curve and reference*
AD *average difference*
AUC *area under the curve*
cDNA *complementary DNA*
DE diferencialmente expresso
DNA *deoxyribonucleic acid*
FARMS *factor analysis for robust microarray summarization*
FC *fold change*
GAPDH *gliceral dehyde 3-phosphate dehydrogenase*
GCRMA *guanine cytosine robust multiarray analysis*
GEO *gene expression omnibus*
ibmT *intensity-based moderated statistics*
ID identificação
IM *ideal match*
MA *minus add*
MAS *Affymetrix microarray suite*
MBEI *model-based expression index*
MM *mismatch*
modT *moderated t-statistic*
mRNA *messenger RNA*
NCBI *national center for biotechnology information*
NPROC *not proper ROC curve*
NUSE *normalized unscaled standard error*
OVL *overlapping coefficient*
PCR reação de polimerização em cadeia
PLIER *probe logarithmic intensity error estimation*

PM *perfect match*
QC *quality control*
RLE *relative log expression*
RMA *robust multi-array average*
RNA *ribonucleic acid*
rRNA *ribossomic RNA*
RP *rank products*
SAM *significance analysis of microarrays*
TNRC *test for not proper ROC curves*
tRNA *transporter RNA*
WAD *weighted average difference*

Capítulo 1

Introdução

É de conhecimento geral que os procedimentos estudados pela bioestatística possibilitam a validação e a análise de dados da expressão genética.

A constante e rápida evolução da biologia celular e molecular nas últimas décadas, aliada à evolução da bioinformática, tem proporcionado uma evolução significativa na sequenciação genómica e na obtenção de informações sobre mecanismos de regulação, funções celulares e diferenças entre vários tipos de tecidos.

A tecnologia de *microarrays* torna-se, desta forma, uma ferramenta preponderante quando associada aos ramos da biologia acima mencionados. É uma técnica com grande influência na análise da composição genética de um determinado organismo mas também na identificação de genes que contêm uma variação nos seus níveis de expressão, quando estes se encontram sobre uma dada condição.

Com a junção desta técnica com métodos estatísticos, é possível selecionar os genes que são diferencialmente expressos (DE) em tecidos afetados com cancro e em tecidos sem esta patologia. A metodologia *Receiver Operating Characteristic* (ROC) foi a ferramenta utilizada para a seleção destes genes através do seu índice da AUC. Dado que o uso de apenas esta metodologia para a seleção deste tipo de genes se revelou pouco eficaz para retirar as conclusões esperadas, foi utilizada uma outra ferramenta, o gráfico *arrow plot*, construído com base em duas estimativas, área abaixo da curva ROC (AUC) e coeficiente de sobreposição entre duas densidades (OVL), determinadas com base numa abordagem não paramétrica.

Para que todo este processo se torne viável é essencial efetuar uma análise da qualidade dos dados a serem utilizados assim como a sua normalização.

Uma análise da bimodalidade e da multimodalidade, torna-se indispensável para a seleção dos genes mistos, que em conjunto com as medidas da AUC e OVL permite ainda selecionar genes com regulação positiva e negativa. Para se conseguir desenvolver todos estes pontos foi utilizado um *dataset* retirado da base de dados GEO (*Gene Expression Omnibus*) e posteriormente trabalhado no *Rstudio* com um algoritmo desenvolvido em R.

1.1 Motivação

Esta dissertação pretende comprovar a importância da seleção de genes DE no apoio ao desenvolvimento de fármacos e métodos de diagnóstico precoce para a patologia cancerígena. Sendo que este processo ainda se encontra pouco explorado para o cancro do pâncreas pretende-se ainda proceder à seleção de genes DE sobreposto ao mesmo com base na metodologia ROC. Dado que o cancro do pâncreas é o quinto tumor maligno mais frequente a nível mundial e que a taxa de mortalidade nos pacientes com esta patologia é de 98%, torna-se pertinente proceder à seleção de genes DE, quer para o desenvolvimento futuro de fármacos para o seu tratamento, quer no procedimento de diagnóstico da mesma. A principal motivação para o desenvolvimento desta dissertação é a procura de genes que geralmente não são identificados pelos métodos estatísticos usuais. Estes genes podem fornecer informações úteis acerca das funções das células que podem estar relacionadas com múltiplos tipos de cancro. Sendo este um tema alvo de enorme investimento na área da investigação, considera-se o mesmo pertinente para estudos clínicos que tenham como alvo a população, na tentativa de trabalhar genes DE e observar em que medida estarão relacionados com qualquer patologia e como é elaborada toda a análise envolvente.

1.2 Objetivos

Tendo em conta a importância da seleção de genes DE já descritos anteriormente, consideraram-se as seguintes hipóteses de trabalho:

1. A seleção de genes DE possui implicações relevantes para a perceção dos mecanismos moleculares e celulares;

2. A seleção de genes DE fornece informações para o desenvolvimento de novas terapias e de novos métodos de diagnóstico precoce do cancro do pâncreas;

3. A metodologia ROC é o método estatístico mais eficaz na seleção de genes DE em comparação com os métodos estatísticos mais usuais;

4. O gráfico *arrow plot* apresenta-se como sendo um complemento à metodologia ROC na análise dos genes DE.

Partindo das hipóteses formuladas, os objetivos centrais do trabalho a desenvolver passam pela representação de métodos que permitam a seleção de genes DE através da análise de dados de *microarrays* e pela exploração da metodologia ROC para posterior avaliação. Pode-se ainda destacar um objetivo secundário, o qual passará pela aplicação da ferramenta *arrow plot* a dados disponíveis publicamente, de modo a possibilitar a demonstração da metodologia em questão e posterior recolha de conclusões sobre genes com regulação positiva e negativa e genes com expressão diferencial nas amostras do *data set*.

1.3 Metodologia e Estrutura da dissertação

As limitações de precisão do diagnóstico como medida do desempenho da decisão requerem a introdução dos conceitos como a sensibilidade e a especificidade de um teste de diagnóstico. Essas medidas e os índices relacionados (verdadeiros positivos e falsos positivos) não fornecem uma descrição única do desempenho de diagnóstico, dado que dependem da seleção arbitrária de um limiar de decisão. A curva ROC sendo uma descrição empírica, simples e completa deste efeito de limiar de decisão, indica todas as combinações possíveis das frequências relativas dos vários tipos de decisões corretas e incorretas [23].

Tendo origem na teoria estatística, a metodologia ROC foi desenvolvida na década de 50 para avaliar a deteção de sinal em radar e na psicologia sensorial, sendo atualmente aplicada a uma grande variedade de testes de diagnóstico [12].

A metodologia ROC para a análise de dados de *microarrays* foi o principal método utilizado neste trabalho, sendo que, para o efeito, foi necessário desenvolver uma pesquisa em fontes de informação primárias e secundárias, alicerçadas em bibliografias, artigos científicos e bases de dados no âmbito da tecnologia de *microarrays* e da metodologia ROC. Para tal utilizaram-se as palavras chave: *microarrays*, genes DE, metodologia ROC, *arrow plot*.

Esta dissertação encontra-se estruturada em seis capítulos. Sendo este um trabalho alicerçado pela bioinformática, no capítulo 1 é feita uma breve introdução ao tema, passando ainda pela descrição dos objetivos propostos e metodologia adotada. No capítulo

2 é feita uma breve introdução à biologia molecular e celular, desenvolvendo os princípios inerentes a esta área de estudo. No capítulo 3 abordou-se a tecnologia de *microarrays* e a metodologia ROC, sendo descritos os métodos utilizados para a análise da qualidade dos dados e o consequente pré-processamento. É ainda abordada a seleção diferencial dos genes quando empregues por esta tecnologia e as características da curva ROC implícitas à mesma. No capítulo 4 descreveu-se os elementos característicos do gráfico *arrow plot*, sendo este constituído pela estimativa AUC e OVL, permitindo ao investigador selecionar genes com regulação positiva e negativa. O OVL é determinado pela abordagem não-paramétrica do mesmo. De forma a cumprir os objetivos inicialmente definidos, no capítulo 5 é apresentado um estudo do *dataset* do cancro do pâncreas [4], extraído da base de dados de informação de expressão genética GEO, ainda neste capítulo serão apresentados os respetivos resultados e consequente interpretação. Aqui, será também apresentada a distribuição dos genes de acordo com a classificação atribuída e uma análise do desenvolvimento do método proposto em comparação com outros métodos existentes. No último capítulo são apresentadas as considerações finais e as linhas orientadoras para trabalho futuro.

Capítulo 2

Introdução à Biologia Celular e Molecular

Em termos históricos (capítulo 2 [8]), sabe-se que, através das experiências de Avery, MacLeod e McCarthy e de Hershey e Chase em 1952, verificou-se que a molécula de DNA (*deoxyribonucleic acid*) seria a responsável pela herança genética de geração em geração. Isto aplica-se tanto para células eucariotas como procariotas. Com o modelo proposto por Watson e Crick em 1953, foi possível a compreensão de todo o mecanismo que envolve a hereditariedade. Neste modelo, a molécula de DNA é constituída por uma dupla hélice antiparalela, que se inicia por uma extremidade de carbono 5'e termina em carbono 3'. Esta hélice é formada por um conjunto de açucares e grupos fosfato na parte externa da molécula. Na parte interna da molécula, as duas cadeias laterais estão ligadas entre si por pontes de hidrogénio, responsáveis pelo emparelhamento de bases azotadas complementares. Este emparelhamento é feito sempre entre uma purina e uma pirimidina, tal que, a adenina (A) emparelha sempre com a timina (T) e a guanina (G) emparelha sempre com a citosina (C). Deste modo, é viável deduzir uma dada sequência de bases a partir da sua cadeia complementar [8].

O emparelhamento proposto por Watson e Crick sugeriu um mecanismo de replicação. Tendo em conta que a replicação do DNA se torna necessária cada vez que ocorre divisão celular, chegou-se à hipótese de uma possível separação das duas cadeias de DNA parental, em que cada uma origina uma nova cadeia que lhe seja complementar, resultando na formação de uma hélice dupla idêntica à parental, através da complementariedade de bases. Este processo é designado por **replicação semiconservativa** e foi comprovado por

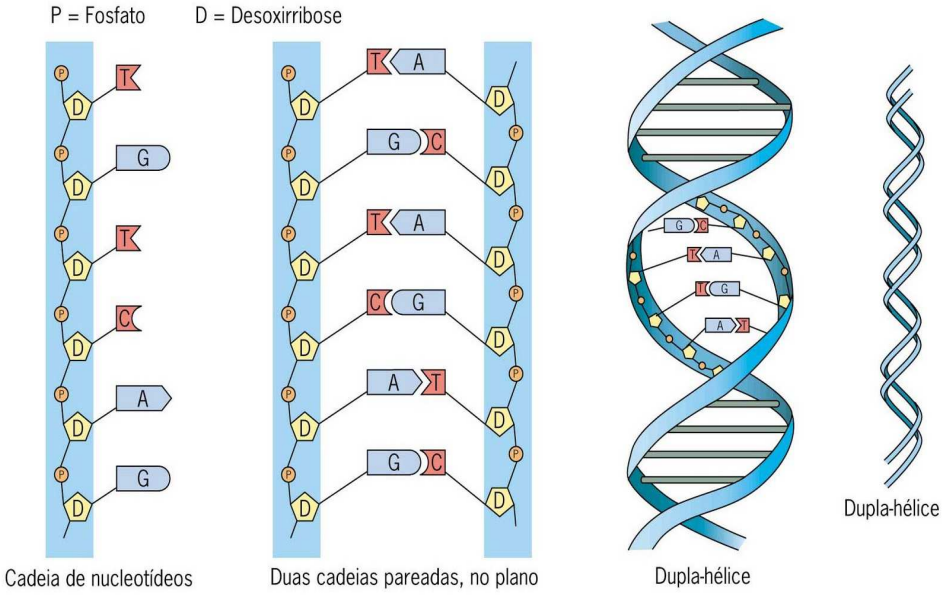

Figura 2.1: Ilustração da estrutura da molécula de DNA (Fonte: Criatividade e ciência).

Meselson e Stahl em 1958. Todo este processo culmina com a obtenção de duas moléculas idênticas à original [8].

2.1 Dogma Central da Biologia Molecular

Como é de conhecimento geral, o DNA contém um papel importante na manutenção da organização biológica. Nesta molécula estão todas as informações necessárias para a construção dos organismos em cada geração, e transmitir essa mesma informação para futuras gerações. Toda esta informação referida é mantida no genoma. Existem duas classes de DNA: composto por sequências de bases repetidas e composto por sequências de bases que apenas se repetem poucas vezes no genoma. Esta molécula, armazena toda a informação necessária para que seja possível obter todas as proteínas que estejam presentes em todas as reações químicas da célula, assim como a informação necessária para a síntese de ácidos nucleicos e ainda a informação relativa à regulação da expressão presente no genoma [17].

O RNA (*ribonucleic acid*) é um polímero de cadeia simples, presente em maiores quantidades que o DNA na célula, em que a base azotada, timina, é substituída pelo uracilo (U), que por sua vez, se emparelha da mesma forma com a adenina. O RNA é sintetizado

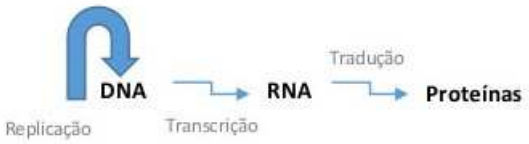

Figura 2.2: Dogma central da biologia.

através do DNA, na medida em que a sequência de bases do DNA é copiada em RNA, tendo por base a complementaridade com a cadeia de DNA que serve de molde. A enzima *RNA polimerase* promove esta síntese no sentido 5'- 3'[17].

As moléculas de RNA são transcritas a partir das moléculas de DNA e seguidamente transportados para diferentes compartimentos celulares. Os rRNA (*ribossomic RNA*), depois da sua conexão a determinadas proteínas, são associados nas subunidades ribossomais. O tRNA (*transporter RNA*) e o mRNA (*messenger RNA*), participam na síntese de proteínas por um processo de tradução. Este conjunto de processos, ilustrado na figura 2.2 denomina-se pelo **Dogma Central da Biologia** [7].

2.2 Análise da Expressão Genética

Os genes correspondem a uma pequena fração na molécula de DNA, esta fração contém informação genética para a síntese de uma determinada proteína, como exemplo pode ser tido em conta que existe um determinado gene que tem a informação para produzir a melanina, esta proteína irá determinar a cor da pele de um indivíduo. Quando ocorrem mutações genéticas, patologias como cancro ou esquizofrenia acabam por se manifestar. Estas mutações podem surgir quando se dão alterações na sequência de bases do DNA, tornando os genes não funcionais e que passam a denominar-se por pseudogenes [7].

Desta forma, é importante analisar cada tipo de patologia pela expressão genética, a base desta análise consiste em comparar os níveis de expressão de um determinado grupo de genes que estejam sujeitos a duas ou mais condições experimentais. O estudo dos genes que têm e não têm expressão, permite saber de que forma os genes afetam o funcionamento das células.

Todo este processo é iniciado pela extração de amostras de tecidos que correspondam à área de interesse. Posteriormente é extraído o mRNA e não o DNA dado que este é

igual para todas as células, enquanto que com o mRNA, tal já não acontece. Para se compreender como ocorre o processo de expressão, considere-se que numa mesma célula, o gene A e o gene B são transcritos em mRNA e posteriormente traduzidos em proteínas. Se numa outra célula apenas o mesmo o gene A for transcrito e traduzido, conclui-se, portanto, que o gene A expressa-se em ambas as células ao contrário do gene B que apenas se expressa na primeira célula [8].

Torna-se necessário utilizar uma tecnologia para a imobilização dos fragmentos de DNA em grandes quantidades, que pode ser desenvolvida a partir de alguns processos, tais como, por sequenciação de DNA, por uma PCR (reação de polimerização em cadeia), ou ainda por *microarrays* [8].

2.2.1 Sequenciação de DNA

Neste processo, o objetivo passa por determinar a sequência de nucleótidos que compõem uma molécula de DNA. Esta tecnologia pode sequenciar, em simultâneo, 96 moléculas de DNA.

Ao longo dos últimos anos, algumas plataformas de sequenciação de DNA foram ficando disponíveis reduzindo assim o custo desta operação. Hoje em dia dia já é possível encontrar vários genomas de microorganismos completamente sequenciados [17].

2.2.2 Reação de Polimerização em Cadeia

A técnica de PCR foi desenvolvida na década de 80 por Kary Mullis que mais tarde recebe o prémio Nobel da química por este mesmo motivo [50].

O seu uso começou por ter influência nos diagnósticos pré-natais, pela identificação de mutações e/ou polimorfismos. Hoje em dia facilita no processo de diagnóstico do cancro, medicina forense e saúde pública [24].

A técnica de PCR convencional permite a síntese de fragmentos de DNA com a utilização da enzima DNA polimerase que está associada à replicação de material genético. O procedimento geral inicia-se com a desnaturação do DNA molde, ou seja, a separação das duas cadeias desta molécula através do sobreaquecimento. Posteriormente, procede-se à ligação dos *primers* (segmentos de RNA que contêm 15 a 20 bases de nucleótidos complementares do DNA) para que a ligação às suas bases complementares ocorra e por fim, procede-se à extensão das cadeias de DNA pela ação da *Taq polimerase* para a obtenção de duas novas moléculas. O processo pode ser repetido 20 a 30 vezes [17].

O desenvolvimento de ferramentas para a amplificação de segmentos de DNA tem gerado algumas vantagens no que diz respeito à análise genética. Outra aplicação bastante útil deste processo é a clonagem de um dado fragmento de DNA permitindo, posteriormente, o estudo da expressão genética, a qual toma grande importância na medicina forense [50]. A técnica de PCR em tempo real, baseia-se nos mesmos princípios que a técnica convencional, no entanto, trouxe uma maior sensibilidade e exatidão e ainda um melhor controlo de qualidade associado a um menor risco de contaminação da amostra [50].

2.2.3 Introdução à Tecnologia de *Microarrays*

A tecnologia de *microarrays*, possibilita a avaliação simultânea dos níveis de expressão de milhares de genes presentes em diferentes tecidos ou amostras. É uma técnica que se baseia numa coleção de pontos microscópicos formados por sequências de DNA fixos numa superfície de vidro/silicone, estes pontos microscópicos são chamados de sondas (*probes*). Cada gene é representado por um *probeset*, cada um destes *prbesets* é constituído por 11 a 20 sondas. Em cada *array* podem existir de 12 a 20000 *probesets*. O uso de *microarrays* para medir concentrações relativas de sequências de ácidos nucleicos numa amostra, está cada vez mais a ser substituída por métodos de sequenciação [13].

Têm como principais aplicações a análise da expressão genética, a análise da ligação do fator de transcrição e a genotipagem.

A base desta tecnologia é o processo de hibridação, no qual é medida a quantidade de DNA ou RNA desconhecido, com base numa sequência complementar conhecida (sonda).

De entre os *microarrays* comercializados, existem dois mais utilizados, nomeadamente, *microarrays* de dois canais cDNA (*complementary DNA*) e *microarrays* de um canal (oligonucleótidos).

Os *microarrays* de dois canais são hibridizados com cDNA, preparados com duas amostras que se petendem analisar, células com patologia e células saudáveis, e etiquetadas com os fluorocromos, Cy5 (vermelho) e Cy3 (verde). As duas amostras são combinadas num único *array*. Posteriormente, os *arrays* são convertidos em imagens através de um *scanner* e medem-se as fluorescências para cada cor separadamente em cada *spot* do *array*. Se a quantidade de mRNA for abundante no tecido com patologia, o *spot* será vermelho, se a quantidade for abundante no tecido sem a patologia, o *spot* será verde, caso a quantidade seja igual nos dois tipos de tecido, o *spot* será amarelo e por fim, caso mRNA não esteja presente, o *spot* será preto [5].

Nos *microarrays* de um canal, milhares de sondas podem ser sintetizadas em cada

array. Cada uma é constituída por oligonucleótidos com sequências conhecidas e cada sonda é complementar do RNA que se pretende quantificar. Cada sonda é constituída por sequências designadas por PM (*Perfect Match*) (sequência idêntica à do alvo) e MM (*Mismatch*) (contêm uma mutação), estas sondas são sempre colocadas aos pares num determinado poço. A relação entre a intensidade do sinal das sondas PM e MM, indica se um gene está ou não ativo no tecido experimental. Este sinal produzido é proporcional à quantidade de RNA na amostra em questão. Estes *arrays* devem hibridar apenas com o mRNA do gene correspondente [6].

Para construir um array é necessário isolar-se o RNA que é de seguida reversamente transcrito em cDNA e hibridado no *array*. Após a lavagem dos *arrays*, emitem-se sinais luminosos e a quantidade de sinal emitido pelo *array* é armazenado num ficheiro **.DAT**. As intensidades das sondas são transformadas em valores, registados em ficheiros **.CEL** [6].

Esta técnica apresenta algumas limitações gerais, entre elas:

- os *arrays* permitem uma medição indireta da concentração do sinal numa dada posição; a concentração medida, assume-se como sendo proporcional à concentração da presumida espécie a hibridar no *array*;
- torna-se bastante difícil desenhar *arrays* em que múltiplas sequências de DNA e RNA não irão ligar à mesma sonda no *array*. Com isto, uma sequência que tenha sido desenhada no *array* para detetar o gene A pode também detetar o gene B, C ou D;
- se a solução a hibridar com o *array* possuir RNA ou DNA para os quais não existem sequências complementares no mesmo *array*, essas espécies não serão detetadas.

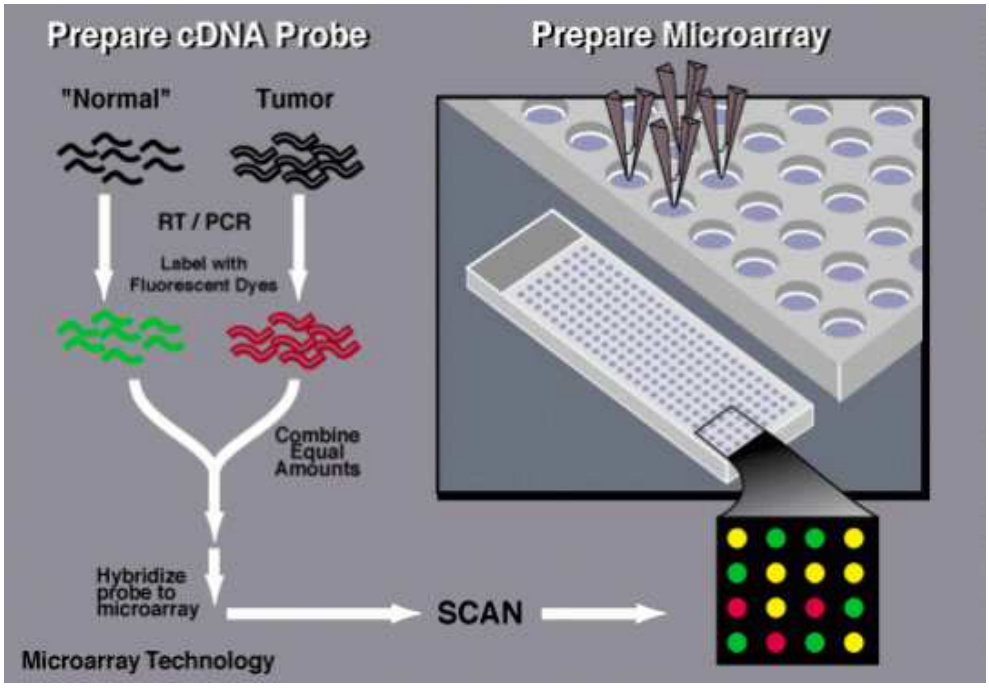

Figura 2.3: Tecnologia de *Microarrays* (Fonte: National Human Genome Research Institute)

Capítulo 3

Análise de dados de *Microarrays* e Metodologia ROC

3.1 *Microarrays*

Os *microarrays* de DNA são uma ferramenta muito utilizada para determinar a possibilidade de o material genético de um indivíduo conter uma mutação em algum gene.

Numa análise de dados de *microarrays* existem certos parâmetros que devem ser respeitados quando são avaliadas algumas medidas. Determinar o caminho correto para combinar os valores de PM e MM para obter o valor de expressão para uma dada sonda é um desses parâmetros. Para ser possível a sua avaliação devem ser seguidas as seguintes regras:

- Normalização: O objetivo passa por remover a variação não biológica sistemática entre *arrays*, baseado na premissa que a maioria dos genes não são diferencialmente expressos através dos *arrays* [29];

- Agregação de sondas: Minimiza os efeitos isolados. Neste caso não se verifica de imediato que as primeiras cinco sondas possam ser removidas como *outliers*. Estas cinco primeiras sondas vão corresponder a 1/3 do total das sondas de um *probeset* e por esta razão são aqui usadas como teste. São sondas que contêm repetições de oligonucleótidos de baixa complexidade, este facto não é tomado em conta pelos algoritmos de remoção de efeitos secundários [29];

- Testes estatísticos: Pretendem determinar quais os genes que denunciam uma ex-

pressão diferencial significativa através de dois ou mais grupos de réplicas [29].

Os *data sets* de *microarrays* incluem grandes quantidades de informação, que devem ser tratadas com metodologias eficazes. A avaliação da qualidade dos dados a analisar é considerada como o maior objetivo para os investigadores. Cada análise de dados deverá passar por uma análise de controlo de qualidade, pré-processamento de dados e identificação de variações genéticas para que, posteriormente, sejam convertidos em informação biológica [10].

3.1.1 Avaliação da Qualidade dos Dados

Atualmente, a utilização dos *arrays* de DNA em análises comparativas da expressão genética encontra-se bastante presente nas mais diversas áreas de investigação e diagnóstico [47]. Apesar da exploração de tais plataformas de *microarrays* ser de elevada importância, não deve ser descurada a credibilidade e a confiança das mesmas. Caso isto não se verifique, incorre-se no risco de utilização de dados falsos e, consequentemente, falsas interpretações. Para além destes fatores, a análise de dados de *microarrays*, está ainda dependente da sensibilidade e especificidade do processo e das amostras utilizadas, tipo de *arrays* usados e identificação das sondas. Uma filtragem de dados adequada e ainda uma normalização dos dados utilizados são também fatores essenciais para que se obtenha uma análise credível e viável. Para que isto seja possível, os *microarrays* de um canal são os mais adequados para uma análise de qualidade [10].

Uma forma para avaliarmos a qualidade dos dados é através de gráficos do tipo *boxplots*, ou seja, que permitam verificar de uma forma simples e eficaz a existência de algum *array* que se diferencie dos outros, dado que, o objetivo será obter distribuições semelhantes entre os vários *arrays*. Estas distribuições não uniformes são obtidas através dos níveis de expressão das sondas de um determinado *array*. Dado que este nível de expressão é dado pela diferença entre a média das sondas PM e a média das sondas MM, caso alguma destas sondas se apresente com uma expressão significativamente diferente , então o *array* em questão vai apresentar uma distribuição diferenciada dos restantes. Caso se verifique a existência de um *array* diferenciado, deve ser imediatamente questionada a sua remoção após o pré-processamento dos dados.

Outros métodos para a avaliação da qualidade dos dados, são o *degradation plot*, o *Simple Affy plot*, os gráficos RLE (*Relative Log Expression*) e NUSE (*Normalized Unscaled Standard Error*) e os gráficos *Pseudo Array Images* [10].

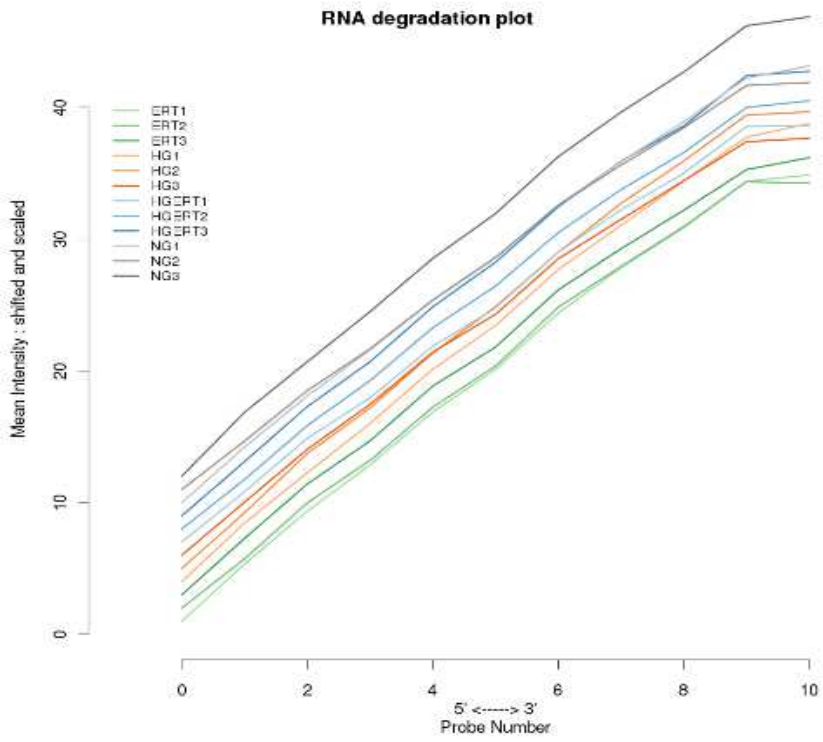

Figura 3.1: Exemplo de um *degradation plot*. (Fonte: Affymetrix QC)

Degradation plot

Este gráfico tem como principal objetivo verificar se os *arrays* são diferentes uns dos outros e se tal ocorrer, pode ser indicativo de potenciais problemas. Tenta quantificar a degradação de RNA de cada *array* e em como este se manifesta quando ocorrem mudanças na hibridação das sondas. Analisa ainda a qualidade da amostra, mais precisamente, no que diz respeito ao material genético a hibridar com o *array*. A análise da qualidade é feita através dos declives que as amostras apresentam. Quando se obtém um declive positivo dos níveis de intensidade das sondas PM dos *arrays*, entre 0,5 e 1,7 indica boa qualidade. Por outro lado, quando o declive se apresenta superior a 2, significa que existiu uma degradação excessiva de RNA. O ponto fulcral desta análise é que todos os *arrays* submetidos a este método tenham declives idênticos [27].

Simple Affy plot

Permitir que o utilizador tenha um rápido acesso à qualidade de um conjunto de *arrays* é o objetivo deste método. Para que se obtenha um ótimo controlo de qualidade, torna-se necessário avaliar médias de *background*, fatores escala e percentagens de *present calls*.

- Médias de *background*: Neste método é estimado o nível de ruído *background* presente num *array*. O nível de ruído vai depender de diferentes quantidades de mRNA na hibridação e ainda da forma mais ou menos eficiente de como é realizado este processo. No gráfico da figura 3.2 é representada pelos números que se fazem apresentar a seguir à respetiva identificação dos *arrays*;

- Fator escala: Ocorre quando o processo de normalização dos *arrays* é feito através do algoritmo MAS (*Affymetrix microarray suite*) 5.0. Este mesmo algoritmo, transforma a escala das intensidades de cada *array* do modo que todos venham a ter a mesma média de intensidades, devendo esta concentrar-se entre os valores de -3 e 3 em todos os *arrays*. Na figura 3.2, faz-se representar pela superfície a azul (intervalo da variação do fator escala);

- Percentagem de *present calls*: O seu valor é adquirido através da diferença entre os pares de sondas PM e MM. Reproduz a percentagem de *probesets* inseridos num *array* em que os níveis de intensidade das sondas PM são superiores aos das sondas MM. No gráfico da figura 3.2 fazem-se também representar pelos números presentes a seguir à identificação dos *arrays*.

Este gráfico *simple affy plot* obtém-se utilizando a biblioteca `affyQCreport` do *Bioconductor*. Neste exemplo, foram também desenhados *probesets* para hibridarem com o GAPDH (*Gliceral dehyde 3-phosphate dehydrogenase*) e o *beta-actin*. Os círculos presentes na figura 3.2 representam os valores de GAPDH. De acordo com a *Affymetrix* estes valores devem rondar a unidade, no entanto os *arrays* com mais potencial a virem a ser destacados irão apresentar valores acima de 1,25 . Os valores de *beta-actin* são representados por triângulos, em que, os valores recomendados devem ser inferiores a 3 [18].

Gráficos RLE e NUSE

Este tipo de gráficos é muito utilizado como método de controlo de qualidade dos *arrays* em estudo.

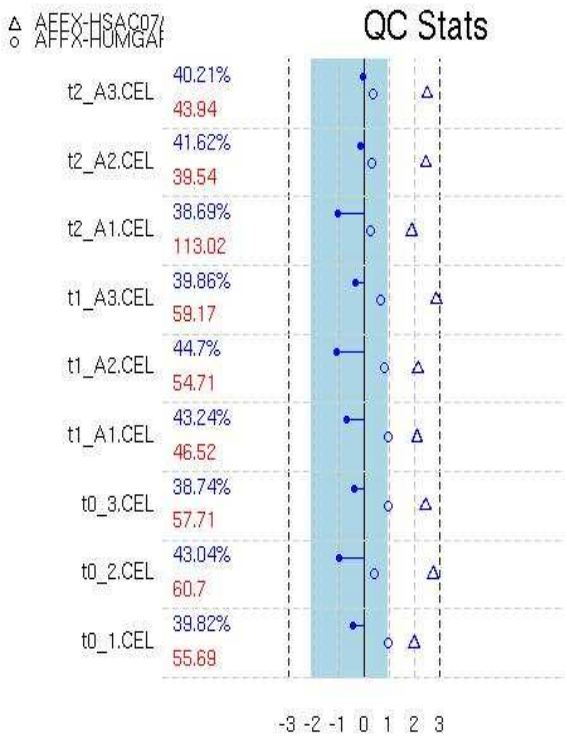

Figura 3.2: Exemplo de gráfico QC (Fonte: BCB Quality Control Report).

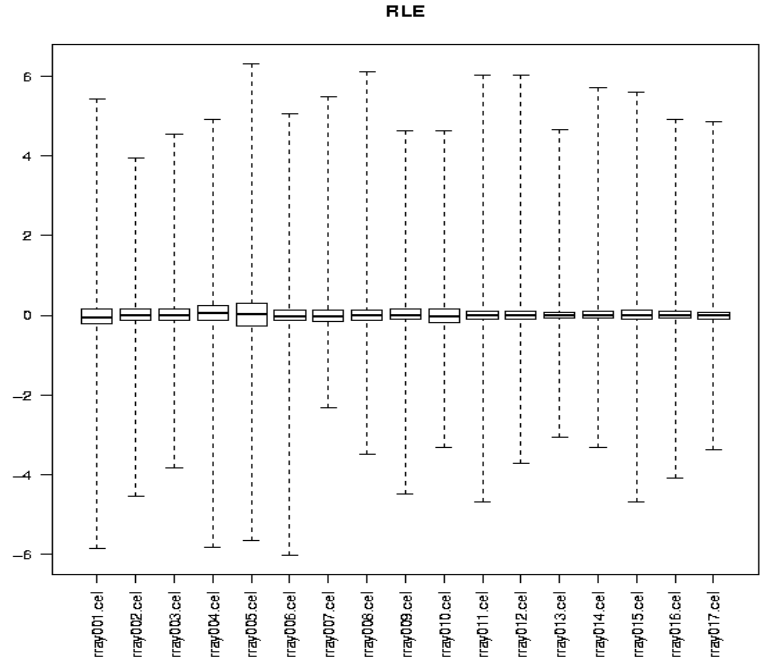

Figura 3.3: Exemplo de *box plot* RLE (Fonte: Affymetrix QC).

- RLE: Tem como base a razão dos valores de expressão dos *probe sets* e o nível de expressão médio do *probe set* de todos os *arrays*. Caso a qualidade de um dado *array* não se afaste muito da qualidade do *data set*, os valores de expressão irão rondar o zero e irão ainda conter distâncias interquartílicas similares a outros *arrays*. Este resultado é o esperado dado que os *probe sets* geralmente não se alteram ao longo dos *arrays* (figura 3.3).

- NUSE: Mede a precisão de uma estimativa dos valores de expressão dos genes em cada *array*. *Arrays* de boa qualidade irão resultar em *box plots* centrados em 1. Por outro lado, *arrays* de má qualidade irão resultar em largos valores de distribuição, ou seja, em grandes distâncias interquartílicas entre os *arrays* (figura 3.4) [45].

Gráficos *Pseudo Array Images*

Têm como principal objetivo a deteção de artefactos (pesos e resíduos do modelo de montagem) em *arrays* que colocam em causa a sua qualidade.

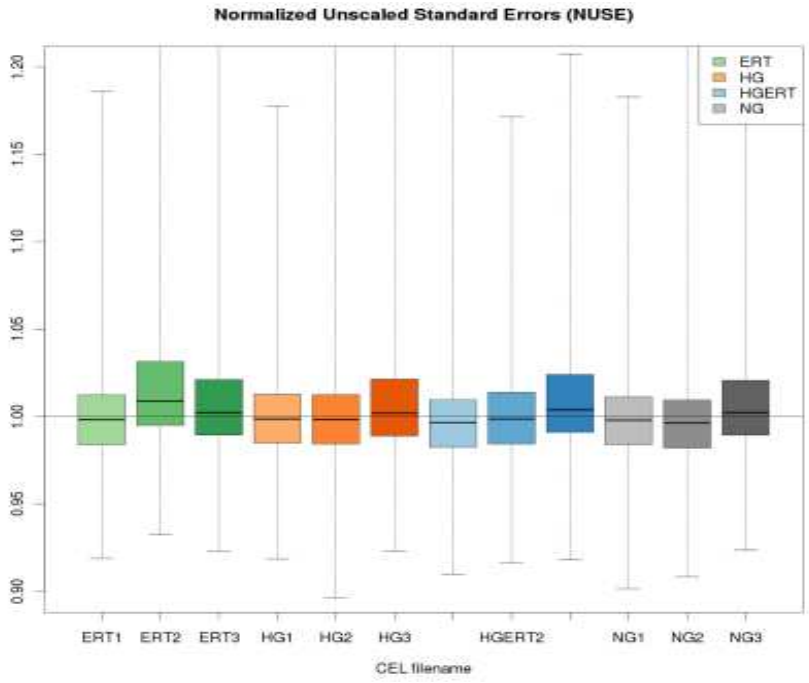

Figura 3.4: Exemplo de *box plot* NUSE (Fonte: Affymetrix QC).

Para se proceder à sua análise, deve-se ter em conta que desde que os artefactos encontrados nos *arrays* não ocupem mais do que 10% da área total, as sondas presentes nesse local podem ser consideradas como valores omissos. É possível obter este tipo de imagem utilizando a função `image` da biblioteca `affy` do *Bioconductor* [11].

3.1.2 Pré-processamento de dados

Após a avaliação da qualidade dos dados, procede-se ao pré-processamento, sendo este efetuado de acordo com os seguintes passos: correção de *background*, normalização, correção PM e sumariação.

Correção de *Background*

A correção de *background*, tal como o próprio nome indica, corrige o ruído dos *arrays*, ajusta as hibridações que não são específicas, nomeadamente, quando existem sequências não complementares que se ligam a sondas MM. Este passo é possível através de vários métodos:

- RMA (*Robust Multi Array Analysis*): Apenas tem em conta as sondas PM, o que num confronto direto entre precisão e exatidão, a exatidão consegue ser superior [41];

- GCRMA (*Guanine Cytosine Robust Multi Array Analysis*): Integra a informação relativa às sequências das sondas MM [41];

- MAS 5.0: Considera ambas as sondas PM e MM, bem como a localização da sonda no *chip* [41];

- PLIER (*Probe Logarithmic Intensity Error estimation*): Contém situações com elevada variabilidade de amostra para amostra, casos em que o *background* é admitido com a variabilidade mínima (apenas para sondas PM) e onde o *background* é irrelevante para o objetivo do estudo (para sondas PM e MM), bem como situações para moderar a sensibilidade da sonda quando estão presentes níveis baixos de intensidade [41];

- MBEI (*Model-Based Expression Index*): Considera casos em que não é feita nenhuma correção de *background*, no entanto, considera as diferenças entre sondas PM e MM [41].

Normalização

A normalização é responsável por remover variações não biológicas entre os *arrays* para que seja possível uma futura comparação dos mesmos. Este passo pode ser a solução para um processo de saturação, ou seja, o momento em que tanto as sondas PM como as MM atingem o seu pico de intensidade máxima e que o *scanner* consegue ler. Tal situação pode levar a uma falha na identificação de genes DE.

Antes mesmo da normalização, deve ser realizada a deteção de fontes de enviesamento, podendo tal processo ser efetuado através dos gráficos MA (*minus add*) [21]. Os gráficos MA permitem fazer a distinção entre as diferenças associadas à intensidade entre 2 *arrays* através da curva *lowess*. O esperado seria que todos os pontos se encontrem à volta de 0 e qualquer desvio mais acentuado revela a necessidade do processo de normalização.

Durante este processo é importante refletir sobre quais os genes e os algoritmos mais apropriados a usar. O ideal será identificar genes que tenham um nível de expressão semelhante em todos os *arrays*.

Todavia, existem métodos baseados nos níveis de intensidade das sondas:

- *Cyclic lowess*: Este método é baseado no gráfico MA. Aqui, o M representa a diferença entre os valores de expressão logarítmica e o A representa a média desses mesmos valores;

- *Orthonormal contrast*: Este processo é mais utilizado quando existem mais de 2 *arrays*;

- *Quantile normalization*: Tem como objetivo garantir que todos os *arrays* possuam a mesma intensidade, sendo graficamente representados por quantis idênticos. Tal não significa que genes iguais venham a ter intensidades semelhantes;

- *Global normalization*: Este método iguala os valores médios de expressão entre os *arrays*, indicando uma relação linear entre os mesmos.

É fulcral que se entenda que não existe um método de normalização ideal, logo o mais aconselhado é que se usem vários. Estes métodos não são infalíveis e quando não se verifica uma igualdade do número de genes com regulação positiva com os de regulação negativa, este processo pode não ser bem sucedido [41].

Correção PM

A correção PM remove sinais não específicos tais como ligações não específicas ou hibridações cruzadas nas próprias sondas PM. Por outro lado, as sondas MM medem as hibridações não específicas entre as sondas PM que lhes correspondem. Um passo ideal a dar seria subtrair os valores MM dos valores PM, no entanto, é importante ter em conta que algumas sondas MM têm um valor de intensidade superior aos das sondas PM correspondentes. É aqui que entra, uma vez mais, o algoritmo MAS 5.0, utilizando um método em que as sondas MM são substituídas pelas estimativas IM (*ideal match*): caso a intensidade da sonda MM seja inferior à sua correspondente PM então a estimativa IM vai ser igual aos valores da sonda MM; caso a maioria das sondas MM pertencentes ao mesmo *probeset* tenham intensidades superiores às suas correspondentes PM, então ao valor da estimativa IM é atribuído um valor inferior ao da sonda PM. Deste modo, o valor do *probeset* obtém-se a partir da diferença entre os valores de IM e MM [31].

Sumariação

A sumariação passa pela combinação de múltiplas intensidades de sondas MM e PM num único valor de expressão de um gene, isto, para cada *probeset*. Este passo apenas se verifica

nos *microarrays* de um canal [31].

Aqui são também utilizados alguns algoritmos já conhecidos:

- MBEI: tem em consideração que sondas do mesmo gene possuem uma maior afinidade de hibridação; isto reflete-se em níveis de intensidade superiores;

- RMA: considera apenas as sondas PM;

- PLIER: é extremamente sensível a níveis de expressão baixos;

- FARMS (*Factor Analysis for Robust Microarray Summarization*): considera que as sondas PM possuem ruído baixo para intensidades baixas, enquanto que para níveis de intensidade elevados o ruído vai ser diretamente proporcional.

3.1.3 Métodos para a seleção de genes DE

O principal objetivo deste estudo, passa por selecionar genes DE, sendo que para o efeito é necessário compreender o conceito inerente aos mesmos. De todos os genes que são propostos a uma análise é possível obter um grupo mais reduzido de genes com as características de interesse definidas para uma análise posterior [32]. No entanto, com esta afirmação surge uma questão: se o genoma é todo igual para todas as células somáticas, como é possível estas diferenciarem-se umas das outras? A resposta a esta questão surge através dos três postulados da expressão diferencial: [51]

1. Todos os núcleos celulares contêm o genoma completo no "ovo" fertilizado. Em termos moleculares, o DNA de todas as células diferenciadas são idênticos;

2. Os genes inutilizados em células diferenciadas não são mutados nem destruídos, estes retêm o potencial de virem a ser expressos;

3. Apenas uma pequena percentagem do genoma é expresso em cada célula e apenas uma porção de RNA sintetizado na célula é específico para aquele determinado tipo de célula em questão.

Inúmeros métodos foram propostos para prever a expressão diferencial. Uma forma de o fazer passa por calcular a estatística para cada gene e, posteriormente, ordenar os valores dados para cada gene de acordo com os valores calculados. É importante ter em conta que

métodos diferentes podem dar ordenações diferentes. Um valor elevado pode ser sinónimo de expressão diferencial [32].

Os métodos mais utilizados para a seleção destes genes são os métodos baseados em FC (*Fold Change*) e os métodos baseados em *estatística-t*.

Métodos baseados em FC

***Fold Change*-** Este método realiza uma interpretação biológica dos dados através de uma determinação de pontos de corte arbitrários que, geralmente, são valores superiores a 2. Deste modo, para valores acima de 1, o gene tem regulação positiva, para valores abaixo de 1, o gene tem uma regulação negativa. Se o valor for zero, os genes não são diferencialmente expressos [19]. É geralmente o mais utilizado, no entanto, pode conduzir a erros dado que considera uma variabilidade constante em todos os genes da experiência em questão.

***Average Difference* (AD)-** De acordo com este método, os genes são ordenados tendo em conta o valor de AD (*average difference*), em que valores elevados têm, normalmente, um grande significado em termos de expressão diferencial. Contudo, podem haver alguns genes classificados nas primeiras posições que têm baixos valores de expressão, tal acontece dado que esta medida não tem em consideração a variabilidade dos dados [33].

***Weighted Average Difference* (WAD)-** É o produto da combinação do método AD com um peso significativo da média da intensidade do sinal. Quanto mais elevado o valor de WAD (*Weighted Average Difference*), mais relevante é considerado o gene [33].

***Rank Products* (RP)-** É um método que torna possível a identificação de genes com regulação positiva e negativa. É mais aconselhado quando existe um número reduzido de *arrays*. Ao contrário do método AD, este considera uma igual variância entre os genes.

Métodos baseados na estatística-*t*

***Significance Analysis of Microarrays* (SAM)-** Esta medida tem em conta as flutuações aleatórias que vão ocorrendo nas amostras, relacionadas com a especificidade de cada gene. Este método é composto por três fases, sendo a primeira referente ao cálculo de uma estatística-*t*, a segunda é baseada num teste de permutações e a terceira baseada no controlo de "falsos positivos" [49].

***Moderated t-statistic* (modT) e *Intensity-based moderated t-statistic* (ibmT)-** São dois métodos baseados na estatística-*t*, tendo ambos em consideração a variância entre os genes.

O modT (*moderated t-statistic*) tem utilidade tanto quando a experiência trata a análise de *microarrays* de dois canais como de um canal.

O ibmT (*intensity-based moderated t-statistic*) é uma versão modificada de modT. Quando a relação entre a variância dos genes e a intensidade dos sinais é fraca ou até inexistente, aplica-se o método modT em vez do ibmT.

3.2 Metodologia ROC

A metodologia ROC tornou-se, nas últimas décadas, uma ferramenta bastante utilizada na medicina, na medida em que, contribui para determinar a exatidão de um dado teste de diagnóstico, através de um valor de corte. Este, representa a concordância entre testes propostos e uma dada referência para identificar uma condição alvo [25]. As curvas ROC podem também ser utilizadas para determinar os níveis de fator de crescimento dos biomarcadores (proteínas, antigénios, enzimas, hormonas...) que melhor diferenciam os casos de cancro dos casos de controlo (amostras normais) [52].

Um objetivo muito comum desta metodologia é a identificação dos genes relacionados com um determinado tipo de cancro, comparando os níveis de expressão de cada um entre amostras normais e amostras com cancro [42]. Os que despertarão mais interesse para o decorrer deste projeto, serão os genes que apresentarem um elevado nível de expressão numa determinada amostra de cancro [44].

Na análise de dados de *microarrays*, esta ferramenta é igualmente utilizada, para que se torne possível comparar a performance de métodos que selecionam genes DE.

A curva ROC aponta a relação entre a proporção de verdadeiros positivos (sensibilidade) e a proporção de falsos positivos (especificidade), resultantes de um valor de corte. A sensibilidade é a probabilidade de um sujeito com cancro ser corretamente classificado, enquanto que a especificidade é a probabilidade de um sujeito sem cancro ser classificado como não tendo a doença em questão [46]. A curva ROC tradicional surge quando se faz variar um valor de corte num eixo de decisão contínuo, com distribuições de probabilidade para os casos negativos (sem cancro) e positivos (com cancro), resultando diferentes pares de sensibilidade e 1-especificidade (pontos da curva) (figura 3.5).

No entanto, tudo isto depende de uma regra de classificação, pois, se uma análise

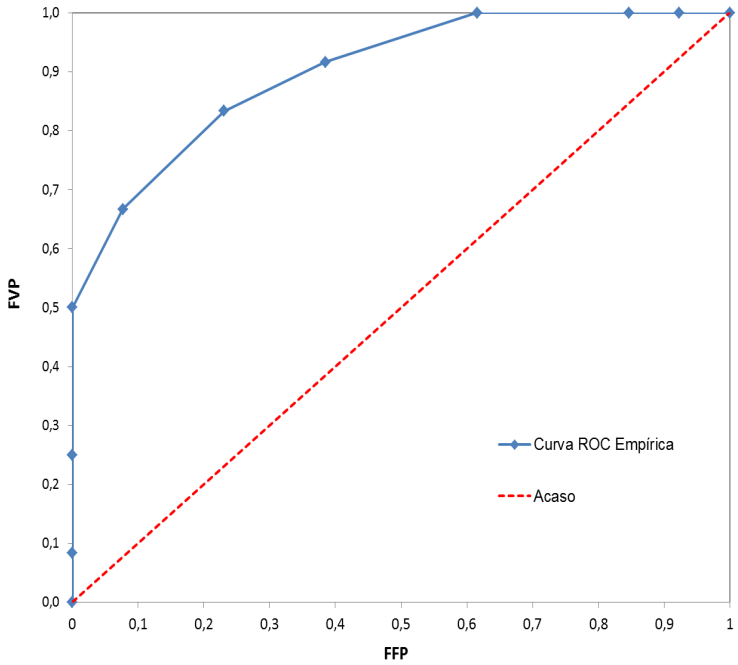

Figura 3.5: Exemplo de Curva ROC.

ROC for aplicada à seleção de genes DE, mantendo a mesma regra de classificação para todos os genes, as curvas NPROC (*not proper ROC*), ou degeneradas (figura 3.6), poderão ser produzidas, dado que genes com uma regulação positiva têm regras de classificação opostas aos de regulação negativa [43]. No entanto, esta não é a única razão pela qual surgem as curvas NPROC. Quando ocorrem casos em que existem distribuições bimodais ou multimodais num dos grupos experimentais, e esses mesmos grupos apresentam médias semelhantes, então, são produzidas curvas ROC sigmoidais, ou seja, curvas que cruzam a linha de referência do plano unitário (figura 3.6 C e D). A ocorrência deste tipo de distribuições num dos grupos, pode ser indicativo da existência de subclasses de genes desconhecidas com diferentes níveis de expressão. A identificação destas mesmas subclasses, pode fornecer informações bastante satisfatórias sobre mecanismos biológicos que estejam subjacentes a condições fisiopatológicas. Assim, pode-se afirmar que as curvas NPROC podem fornecer informações essenciais sobre vários tipos de expressão e permitir ainda a distinção de genes DE de genes não DE mas que inicialmente apresentam características como sendo DE [47].

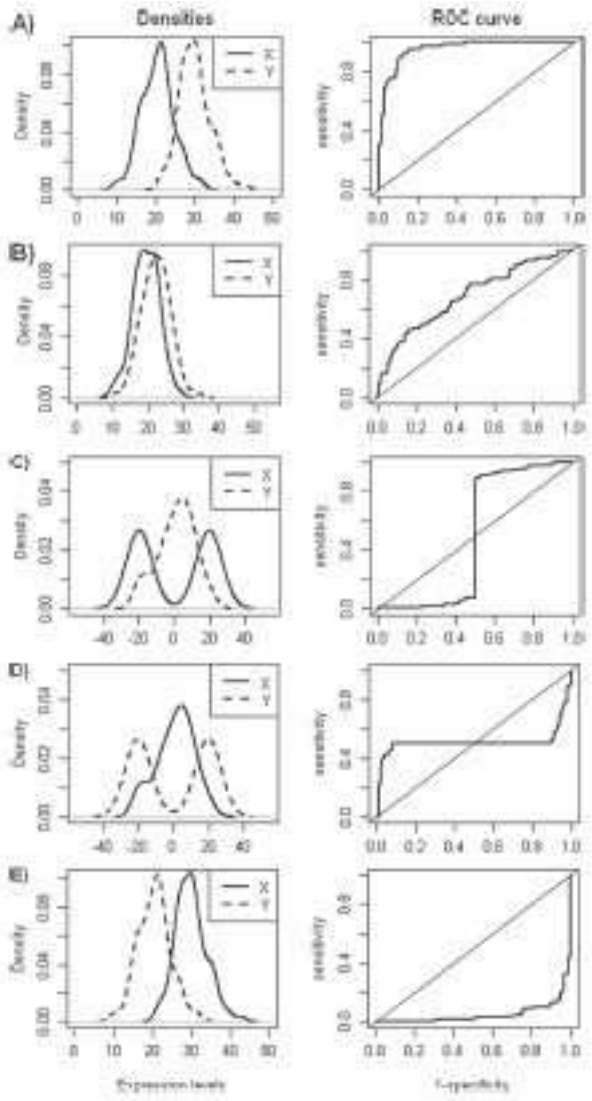

Figura 3.6: Exemplos de Curvas ROC degeneradas (Fonte: Silva Fortes [48]).

3.2.1 Seleção de genes DE baseados na metodologia ROC

Existem alguns métodos para a seleção de genes DE baseados na metodologia ROC. De entre os quais se destacam:

- SAMROC: Este método minimiza o número de falsos negativos e falsos positivos, que haviam sido classificados como DE. Tem por base o método SAM e a metodologia ROC;

- TNRC (*Test for not proper ROC curves*): Este método permite a seleção de genes DE e, a partir destes, fazer nova seleção para identificar os que têm curvas ROC degeneradas;

- ABCR (*Area between the ROC curve and the rising diagonal*):Permite identificar diferentes tipos de genes DE [43].

3.2.2 Índice de precisão da curva - AUC

A AUC é a medida de discriminação mais utilizada na metodologia ROC. É uma medida não-paramétrica da distância entre as distribuições nas duas classe a avaliar, sendo considerado o método standard para ter acesso à exatidão de um modelo de distribuição, evitando a subjetividade no processo de seleção do valor de corte [36]. Quando se pretende avaliar a regra de classificação de um modelo, a AUC deve apresentar valores entre 0,5 e 1 enquanto que para as NPROC deve apresentar valores abaixo de 0,5. Com isto, pode-se afirmar que genes com regulação positiva, apresentarão uma AUC próxima de 1, genes com regulação negativa, terão uma AUC abaixo de 0,5 e os genes DE (com maior interesse para o estudo) terão uma AUC em torno de 0,5. Um valor de corte que minimize a diferença entre a sensibilidade e especificidade tem uma melhor performance. Este valor, graficamente, representa a interseção de uma curva ROC com a linha perpendicular à diagonal do ponto de discriminação [36].

Uma vez que, apenas com a medida discriminativa AUC, não é possível diferenciar os genes mistos dos genes que não são DE, será necessário envolver outra medida para selecionar genes DE e genes mistos, nomeadamente o OVL (coeficiente de sobreposição de duas densidades) [47], [48].

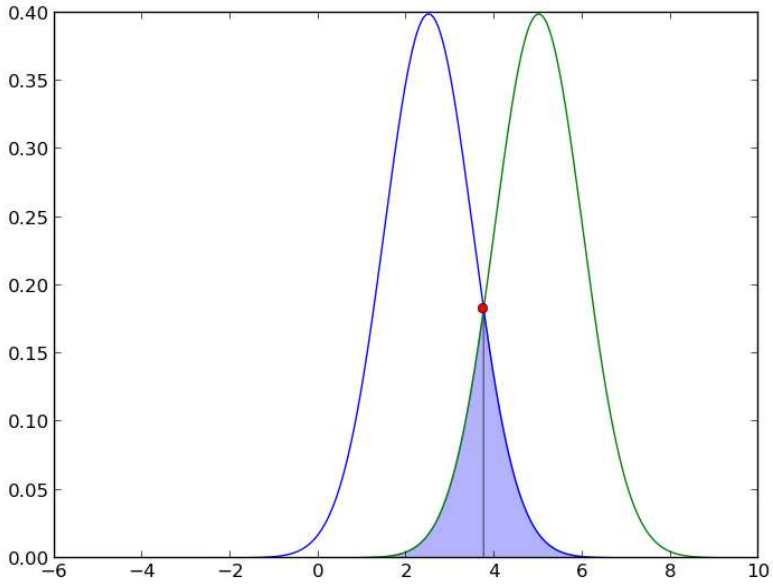

Figura 3.7: Exemplo de ilustração de OVL (Fonte: OVL)

.

3.2.3 Coeficiente de Sobreposição - OVL

O coeficiente de sobreposição de duas distribuições, o OVL, foi uma medida proposta por Fortes [47], para que conjuntamente com a AUC permita diferenciar os casos com AUC próximos de 0,5, sendo aqui utilizado como medida de concordância e similaridade entre duas distribuições [34].

É ainda bastante útil na comparação de modelos de probabilidade e é definido com sendo a área intersetada entre duas funções de densidade e probabilidade (área de sobreposição entre duas densidades) [14].

Capítulo 4

Arrow Plot

Tal como referido anteriormente a análise de dados de *microarrays* para a identificação de genes DE, tem vindo a ganhar importância nos últimos anos. Determinar quais os genes que apresentam características de expressão diferencial quando estão inseridos em diferentes tecidos ou em amostras submetidas a diferentes condições experimentais é o grande objetivo deste estudo. No entanto, podem existir genes DE que não sejam identificados como tal quando se utiliza os métodos mais comuns para este procedimento. Desta forma, recorreu-se a esta nova ferramenta denominada de *arrow plot*, desenvolvida e estudada por Fortes [48]. Esta técnica identifica não só genes com regulação positiva e negativa como também, todos os genes DE presentes em diferentes grupos.

Esta recente tecnologia, permite analisar simultaneamente as medidas não-paramétricas de AUC e OVL, correspondentes ao eixo das abcissas e ordenadas, respetivamente, num gráfico de eixos coordenados. Trata-se de uma técnica eficaz para a seleção de genes DE e mistos (genes com interesse biológico) e consiste em dois processos. Numa primeira fase, são selecionados os genes que tenham uma AUC perto de 0,5 e um valor de OVL baixo, na segunda fase destaca-se a importância da análise da bimodalidade e multimodalidade.

A identificação da bimodalidade e multimodalidade nos genes em análise, tornou-se relevante neste tipo de investigação, dado que, a distinção dos grupos de pacientes que obtêm expressões genéticas elevadas ou reduzidas é mais fácil quando comparada com genes que apresentam expressões unimodais. Genes com expressão bimodal, têm papéis importantes na diferenciação e sinalização celular, assim como na própria progressão da doença. A manutenção e regulação da expressão bimodal pode levar à proliferação celular descontrolada e ainda à formação de tumores malignos [30]. Deste modo, é necessária a

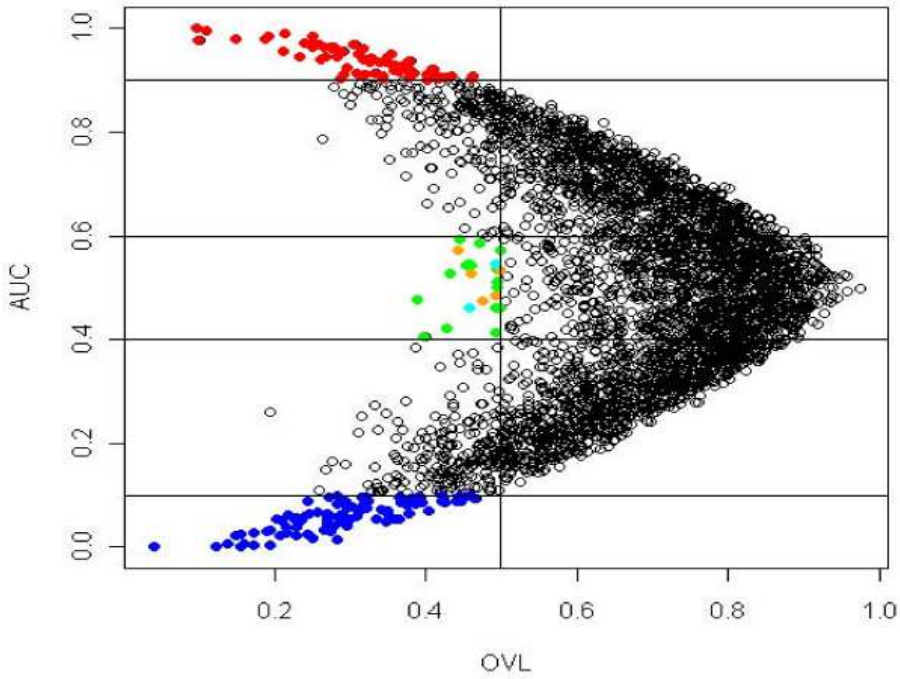

Figura 4.1: Exemplo de *arrow plot* (Fonte: Silva Fortes [48]).

remoção dos genes que não apresentam a expressão pretendida. Para que um determinado gene seja considerado misto, basta que apresente bimodalidade em apenas um dos seus grupos [48].

Tal como referido por Silva Fortes [48], para que seja viável a exploração da tecnologia *arrow plot* na seleção proposta, foram considerados os seguintes valores de corte: genes regulados positivamente apresentariam uma AUC superior a 0,9 e um valor de OVL abaixo de 0,5; genes regulados negativamente apresentariam uma AUC abaixo de 0,1 e um valor de OVL abaixo de 0,5; para a seleção dos genes diferencialmente expressos, estes apresentariam uma AUC entre 0,4 e 0,6 e um valor de OVL abaixo de 0,5. Os valores de corte são escolhidos arbitrariamente, no entanto, apenas um investigador com vasta experiência se deve ocupar desta tarefa, havendo riscos para a interpretação dos resultados [47].

4.1 *Arrow Plot vs Volcano Plot*

O *volcano plot* é outro tipo de gráfico utilizado na seleção de genes DE sendo útil para identificar rapidamente alterações em conjuntos de dados, sendo muito comum em investi-

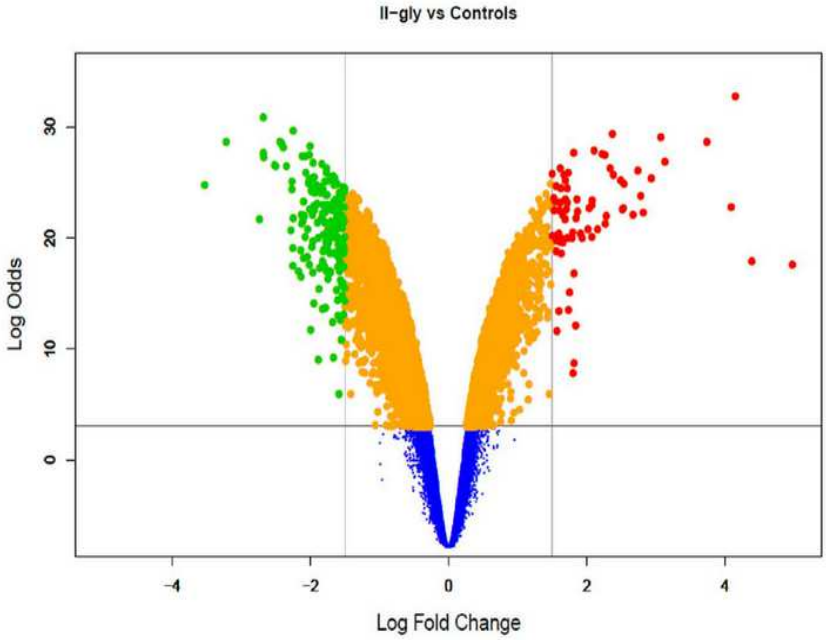

Figura 4.2: Exemplo de *volcano plot* (Fonte: Volcano plot)

gações que envolvam temas como a genómica, proteómica e metabolómica. Aqui é obtida uma lista de milhares de dados replicados em que uma condição imediata que se põe é identificar rapidamente alterações significativas.

Apresenta-se como tendo uma função muito semelhante à do *arrow plot*, exibe tanto o método *fold change* como a *estatística-t*. O método *fold change* é exibido no eixo das abcissas, representado a significância biológica pelos valores do logaritmo de base 2 do *fold change*. Por sua vez a *estatística-t* faz-se representar no eixo das ordenadas pelos valores negativos dos logaritmos dos p-*values* obtidos através da *estatística-t*, representando a significância estatística. Todavia, alguns genes verdadeiramente expressos podem escapar a uma análise posterior, sendo esta a principal razão pela qual foi substituído pelo *arrow plot* [35].

Capítulo 5

Dados de expressão de tumor pancreático e amostras normais

5.1 O cancro do pâncreas

O pâncreas é uma glândula de aproximadamente 15 centímetros e pertence simultaneamente ao sistema digestivo e ao sistema endócrino. Localiza-se atrás do estômago e possui funções exócrinas, com secreção do suco pancreático que contém enzimas digestivas que ajudam o trânsito das proteínas, açucares e gorduras para o intestino, e funções endócrinas, através das quais produz hormonas importantes como a insulina, o glucagon (hormonas que regulam os níveis de açúcar no sangue) e a somatostatina que regulam a forma como o organismo utiliza os açucares ingeridos. Estas entram na corrente sanguínea e ajudam o organismo a utilizar ou a armazenar a energia obtida a partir dos alimentos. As hormonas endócrinas e as enzimas exócrinas são responsáveis pela constituição do tecido pancreático, possibilitando a regeneração celular do órgão [20].

Quando as células do pâncreas perdem o mecanismo de controlo e sofrem alterações no seu genoma formam, consequentemente, um tumor, o qual em caso de malignidade é, frequentemente, denominado por cancro. A origem deste pode ser explicada através de alterações em três grandes grupos de genes: aceleradores, travões e desestabilizadores. Este último grupo corresponde aos "genes envolvidos na estabilidade genética", nomeadamente aqueles que restauram os erros genéticos que ocorrem nas mitoses (divisões celulares) e que permitem regenerar as células que todos nós produzimos diariamente. As alterações nos

genes envolvidos na reparação do ADN são determinantes no aparecimento do cancro (carcinogénese). Havendo perda de atividade destes genes, maior será o número de alterações genéticas, não corrigidas, em oncogénese e genes supressores tumorais [51].

O carcinoma do pâncreas possui diferentes tipos, sendo o mais comum o adenocarcinoma pancreático. É de difícil controlo e não é detetado precocemente, sendo mais frequente na população masculina e possuindo maior incidência nos doentes com pancreatite crónica e em alguns tipos de diabetes. Está associado ao consumo de tabaco, álcool e a uma alimentação rica em carne gorda. O seu diagnóstico é complexo, dado que os sintomas surgem tardiamente e podem ser confundidos com outras patologias. Evolui de forma silenciosa e está associado a sintomas como a icterícia, dores lombares e na parte superior do abdómen, perda de apetite e de peso, fraqueza, náuseas e vómitos. Este tipo de carcinoma possui cinco estadios diferentes, de acordo com a magnitude do tumor [51]:

- Estadio 0: Cancro apenas no revestimento do pâncreas (*carcinoma in situ*);
- Estadio 1: Cancro apenas no pâncreas;
- Estadio 2: O cancro invade os tecidos próximos ou gânglios linfáticos próximos do órgão;
- Estadio 3: O cancro atinge os vasos sanguíneos próximos do pâncreas e poderá estar disseminado para os gânglios linfáticos próximos;
- Estadio 4: O cancro pode alcançar qualquer tamanho e invade órgãos distantes (fígado, pulmão, cavidade abdominal...).

5.2 Descrição da experiência

Para ilustrar a aplicação das metodologias apresentadas foram usados *microarrays* para a identificação de diferenças de expressão dos genes entre amostras de tumor pancreático e amostras normais. O *data set* contém 36 amostras tumorais, representadas pelos *arrays* com a letra "E"e 16 amostras normais, representadas pelos *arrays* com a letra "C", num total de 52 ficheiros. Das 36 amostras tumorais, 16 possuem dados de expressão tumoral e de normalidade, enquanto 20 possuem apenas dados de cancro [4].

Para efetuar o procedimento foi utilizado um código em R disponível em [47] e adaptado e corrigido para este novo conjunto de dados.

5.3 Procedimentos

Os dados pioneiros foram publicados a 7 de Novembro de 2003, tendo sido realizada uma última atualização a 31 de Agosto de 2016. Os mesmos encontram-se na GEO (*Gene Expression Omnibus*), uma base de dados de informação para a expressão genética do NCBI (*National Center for Biotechnology Information*), e foram fabricados pela *Affymetrix*, uma companhia produtora de *microarrays* de DNA, muito utilizados na área biomédica.

Tal como refere *Talloen et al.* [37], os *arrays* da *Affymetrix* consistem em sondas desenhadas para interrogar quantos dos genes transcritos (mRNA) da sequência complementar para a sequência de DNA estão presentes numa amostra, possibilitando também a oportunidade de saber se os genes foram ou não detetados em cada *array*. Isto ocorre porque cada alvo de transcritos é sondado por um par de oligonucleótidos, em que uma sonda PM mede a concentração de mRNA e a sonda MM tem em conta a medida de *background*.

A diferença entre as sondas PM e MM é usada para determinar se o transcrito se encontra ou não presente, sendo cada um dos transcritos representado por 11 a 20 pares de sondas diferentes. As intensidades das mesmas são sumariadas para cada conjunto de sondas, de modo a providenciar um nível de expressão para o respetivo transcrito.

O caso de estudo apresentado analisa exatamente estes *microarrays* de oligonucleótidos.

5.4 Análise da qualidade dos dados

Tal como referido no capítulo 3, é essencial uma avaliação da qualidade dos dados, na medida em que se torna crucial perceber se existem *arrays* com grandes diferenças comparativamente aos restantes. Desta forma, iniciou-se este procedimento com a análise das imagens de cada *array*, nas quais não foram encontrados nenhuns *arrays* que justifiquem a sua remoção, figura 5.1.

Com a análise das densidades dos logaritmos dos níveis de intensidade PM (figura 5.2) e respetivos *box plots* de todos os 52 *arrays* foi possível verificar que os dados em questão deveriam ser submetidos a um processo de normalização. Através do gráfico da figura 5.3 destaca-se principalmente o *array* E28, este *array* contém a intensidade das sondas PM aumentadas, logo o seu nível de expressão vai estar elevado em relação aos outros *arrays* em análise.

Através do gráfico QC (*Quality Control*) (figura 5.4) foi possível verificar que os *arrays* C7 e C15 se destacaram em relação aos outros. Através da análise dos fatores escala foi

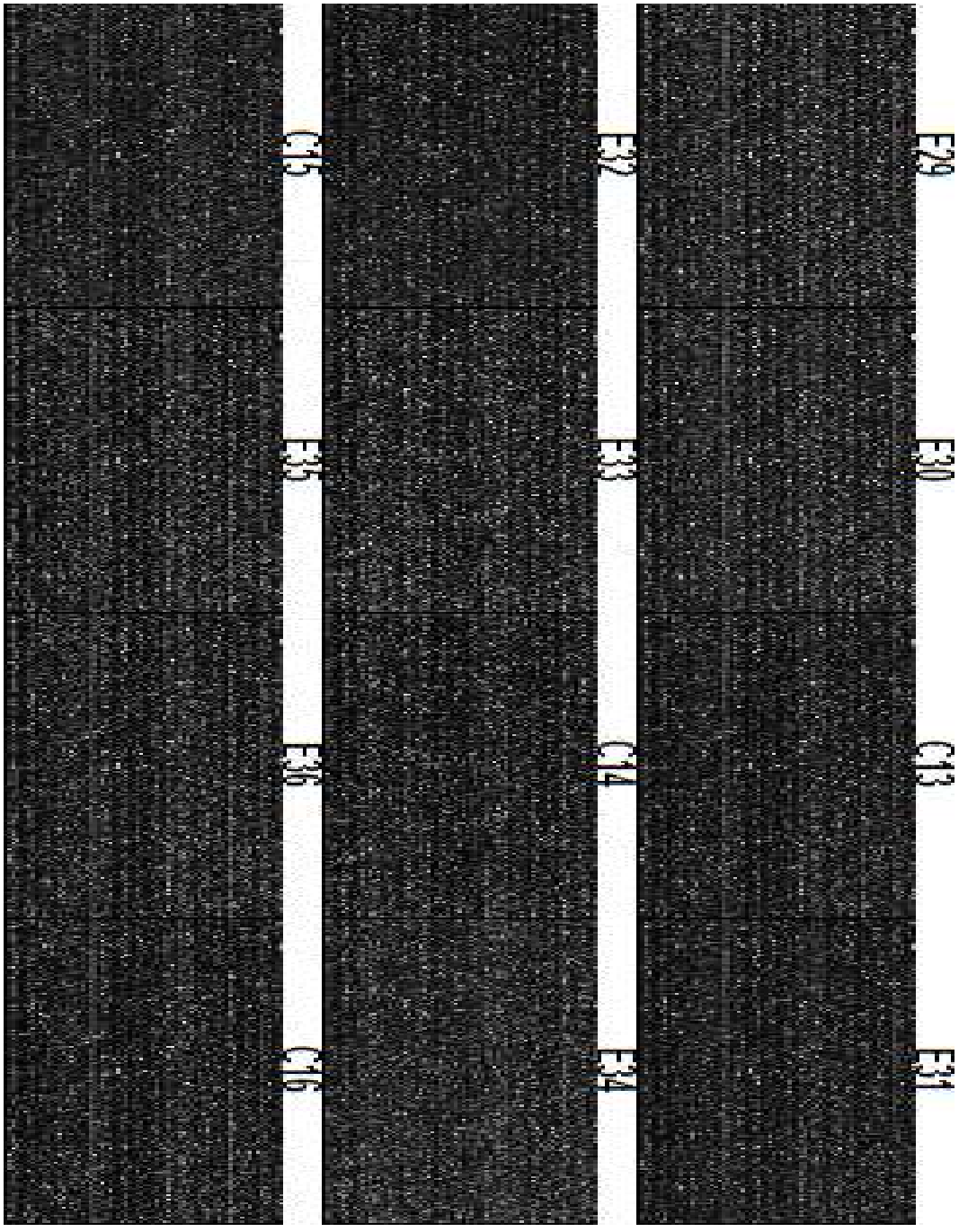

Figura 5.1: Imagens dos *arrays* do *data set*.

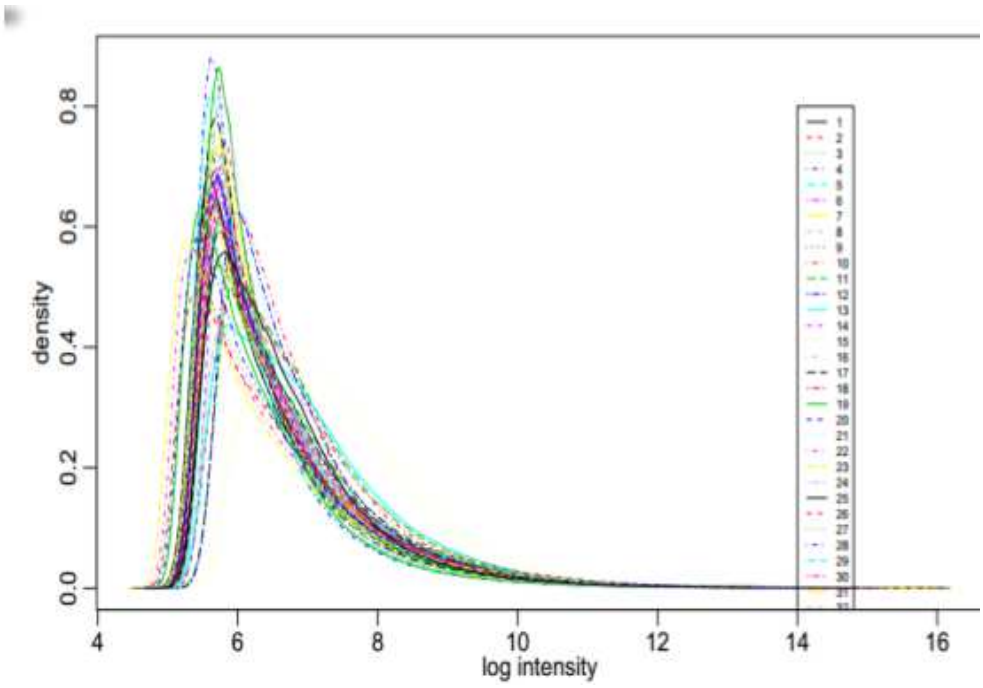

Figura 5.2: Gráfico de densidades dos logaritmos dos níveis de intensidade PM.

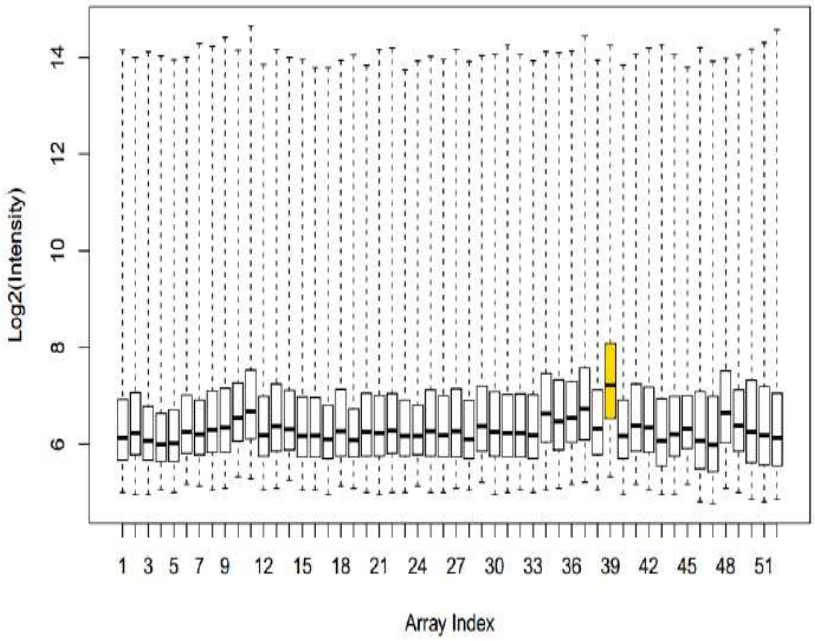

Figura 5.3: *Box plot* dos níveis de intensidade PM dos *arrays* do caso de estudo.

possível observar que os *arrays* anteriormente referidos encontram-se muito próximos do limite deste fator, devendo por isso, proceder-se à sua remoção.

Nos gráficos MA, foram analisados os *arrays* desde o E1 ao E5 incluindo ainda o C1, isto porque não é necessária a apresentação deste tipo de gráfico para a totalidade dos *arrays* (figura 5.5). Aqui verificou-se uma necessidade de normalização dos dados tendo em conta o desvio do ajustamento à curva *lowess*.

Pela análise do *degradation plot*, (figura 5.6) é possível concluir que os ficheiros em estudo, em específico o material genético emanado dos mesmos, é de boa qualidade dado que todas as linhas (que representam os *arrays*) se apresentam paralelas umas às outras e com um declive positivo.

Os gráficos RLE (figura 5.7) e NUSE (figura 5.8) apresentam o *array* C15 como o que possui uma maior diferença comparativamente aos outros. No entanto, no gráfico NUSE, também os *arrays* C7 e C13 apresentaram essa mesma diferença.

Desta forma, procedeu-se à remoção dos *arrays* C7, C13 e C15, sendo estes os que mais se destacaram nos vários métodos de avaliação de qualidade de dados apresentados.

5.5 Pré-processamento

Tendo em consideração a análise supra, foi possível concluir que os *arrays* em estudo devem ser submetidos a uma análise de pré-processamento. Os métodos que permitem realizar esta análise foram já descritos no capítulo 3, sendo eles o RMA, MAS5, PLIER, FARMS, MBEI e GCRMA.

Com a aplicação dos mesmos verificou-se que os métodos RMA e FARMS são os que produzem densidades mais semelhantes.

Para comprovar que os níveis de expressão se tornaram mais semelhantes após a remoção dos *arrays* acima mencionados, foi feita uma nova análise de pré-processamento através do método FARMS, (figura 5.15)

Através da análise do gráfico da figura 5.15 é possível verificar uma maior concordância entre os *arrays* após a normalização, deste modo, torna-se viável a continuação do estudo do conjunto de dados com o objetivo de identificar os genes DE aqui presentes.

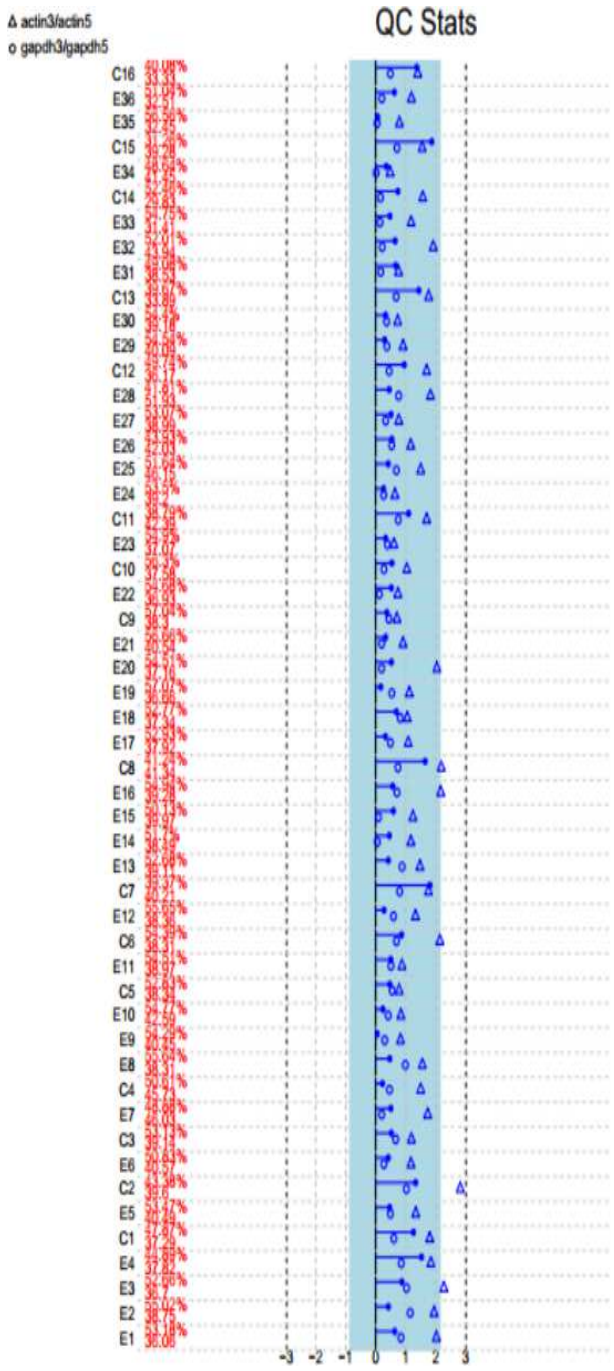

Figura 5.4: Gráfico QC dos dados.

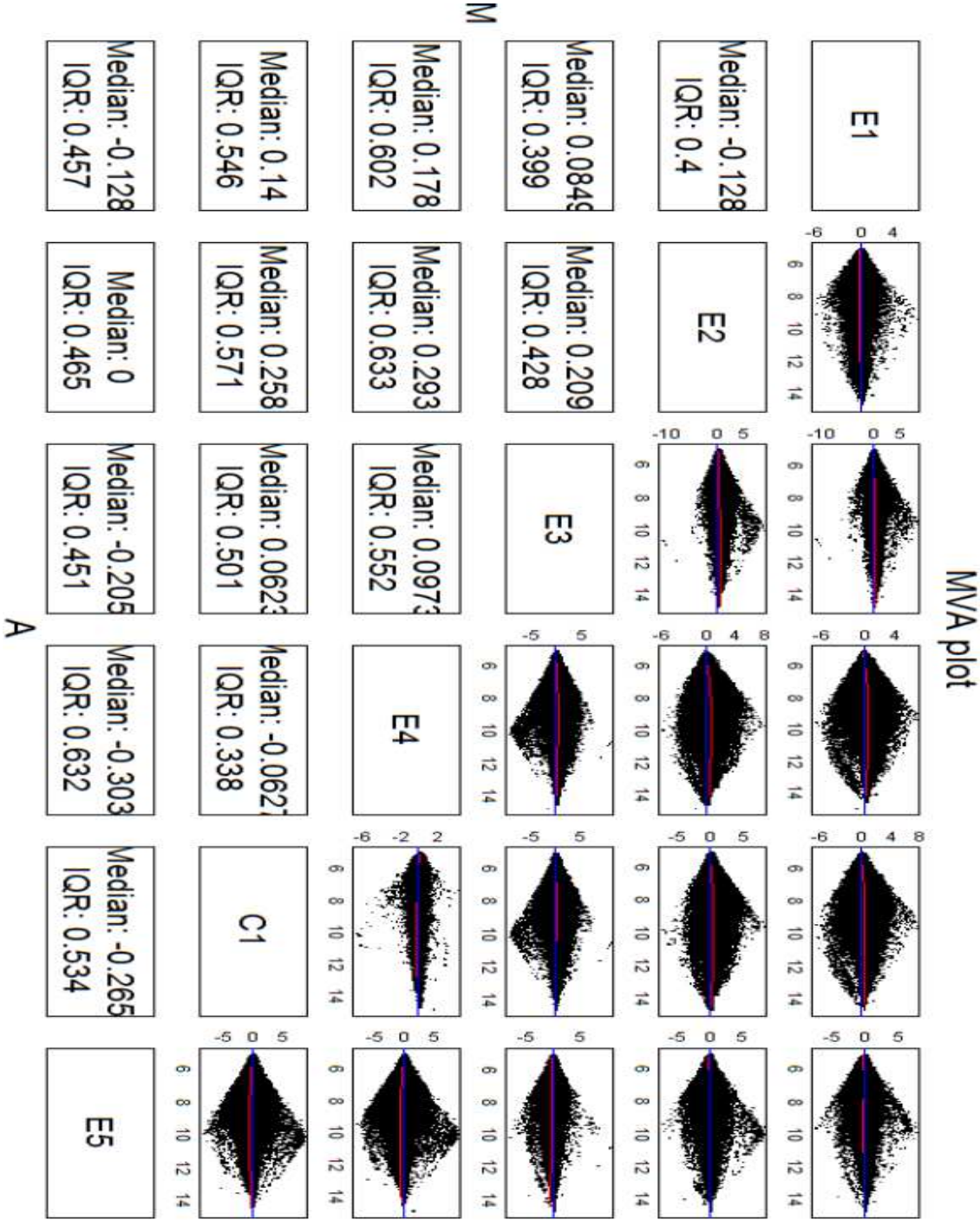

Figura 5.5: Gráficos MA dos *arrays* representados.

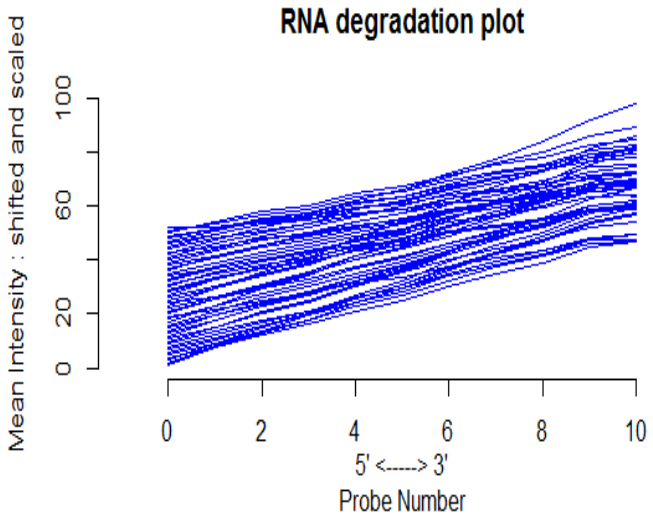

Figura 5.6: *Degradation plot.*

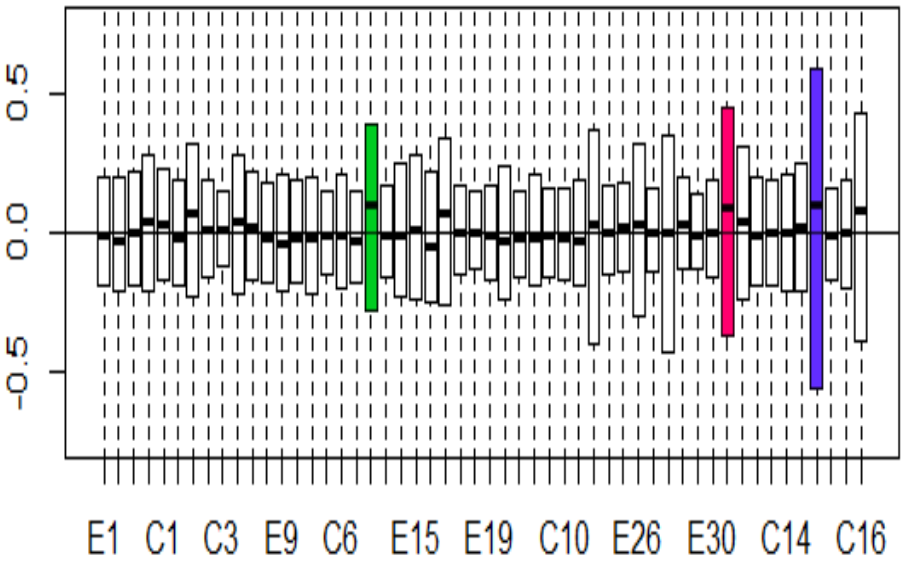

Figura 5.7: Gráfico RLE.

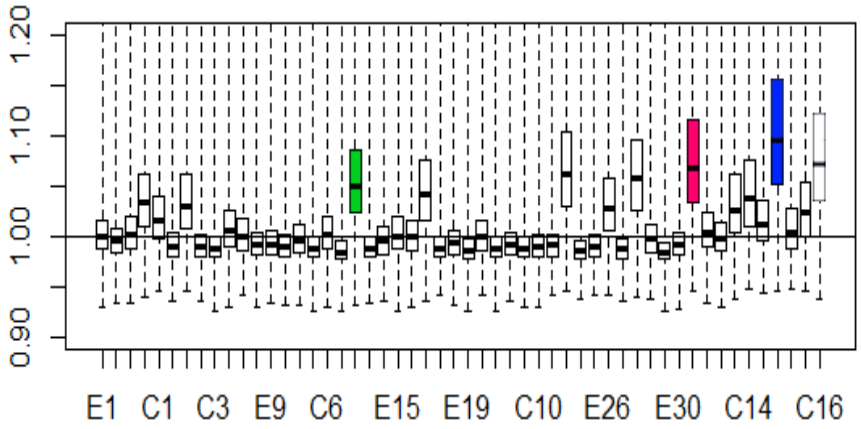

Figura 5.8: Gráfico NUSE.

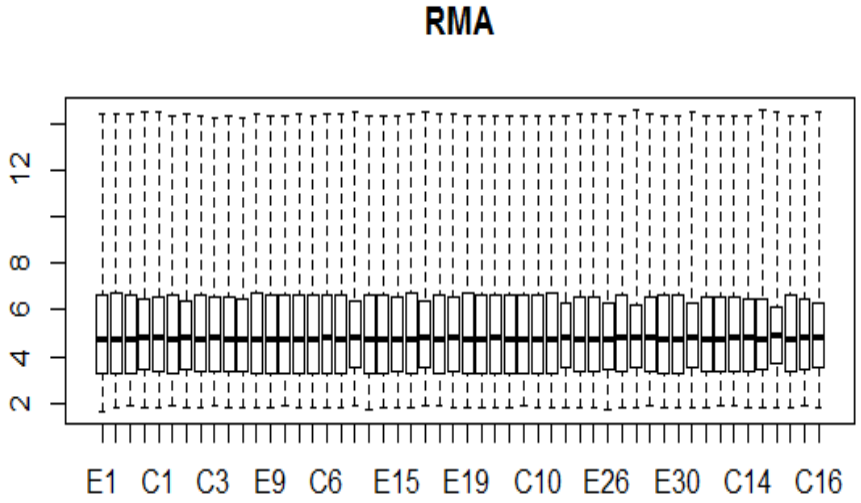

Figura 5.9: *Box plot* dos níveis de expressão dos *arrays* após pré-processamento RMA.

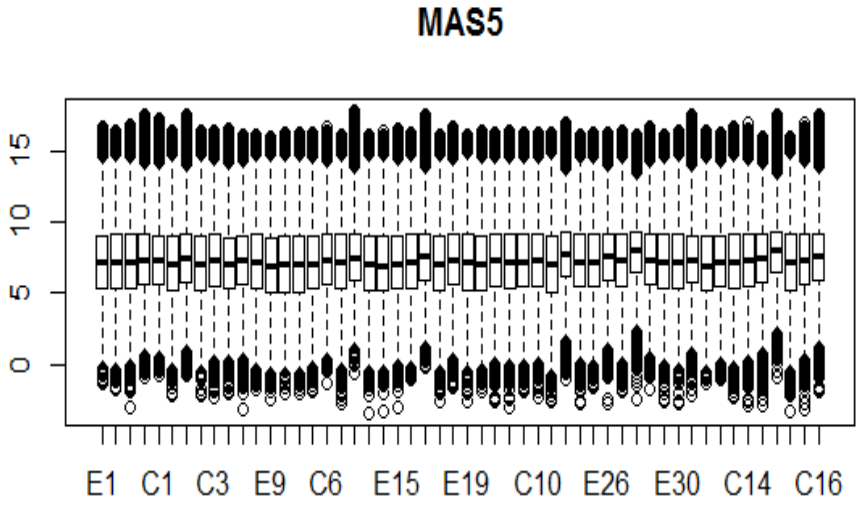

Figura 5.10: Box plot dos níveis de expressão dos *arrays* após pré-processamento MAS5.

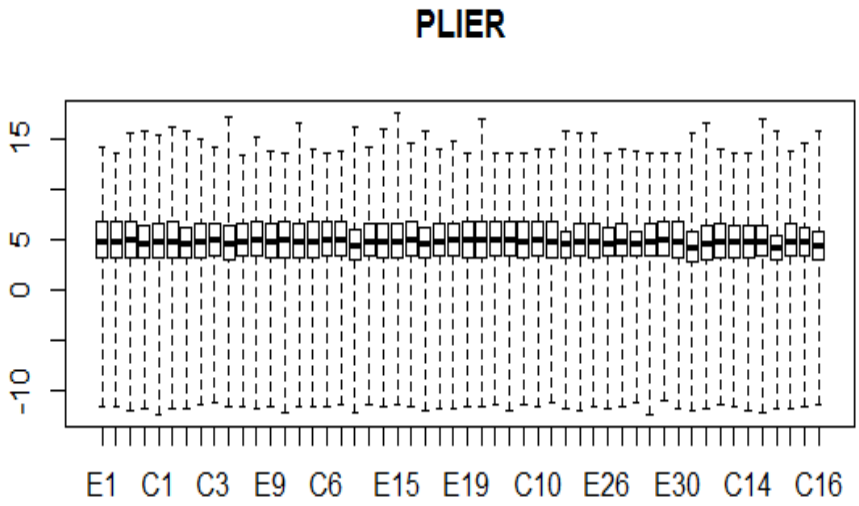

Figura 5.11: Box plot dos níveis de expressão dos *arrays* após pré-processamento PLIER.

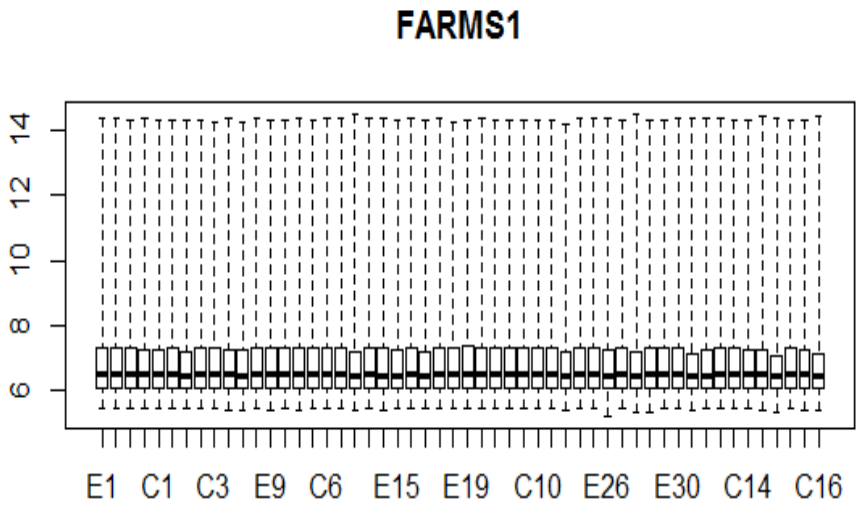

Figura 5.12: Box plot dos níveis de expressão dos *arrays* após pré-processamento FARMS.

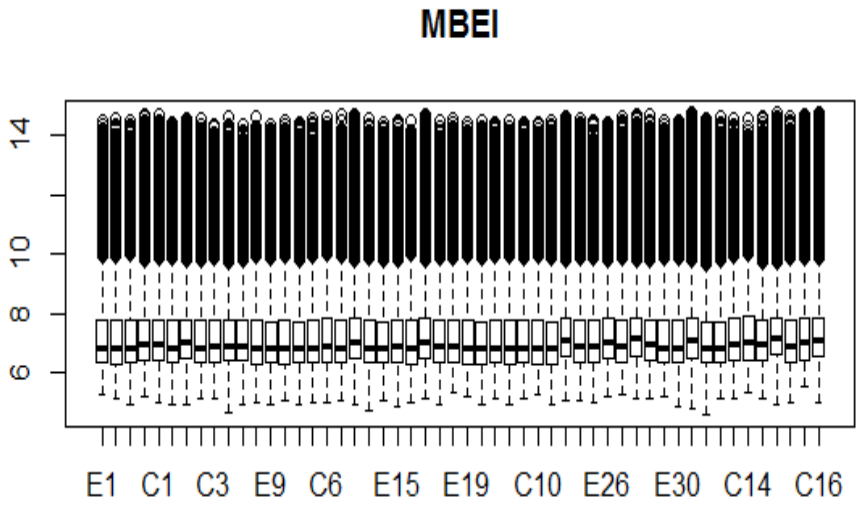

Figura 5.13: Box plot dos níveis de expressão dos *arrays* após pré-processamento MBEI.

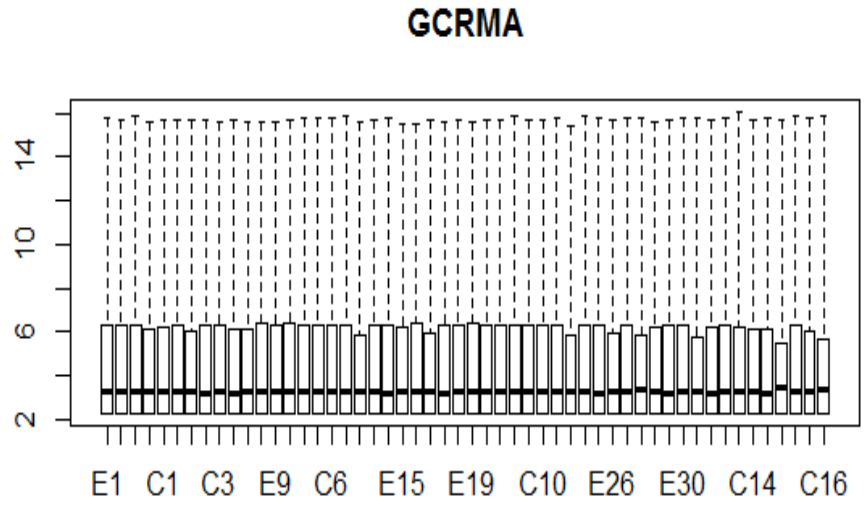

Figura 5.14: Box plot dos níveis de expressão dos *arrays* após pré-processamento GCRMA.

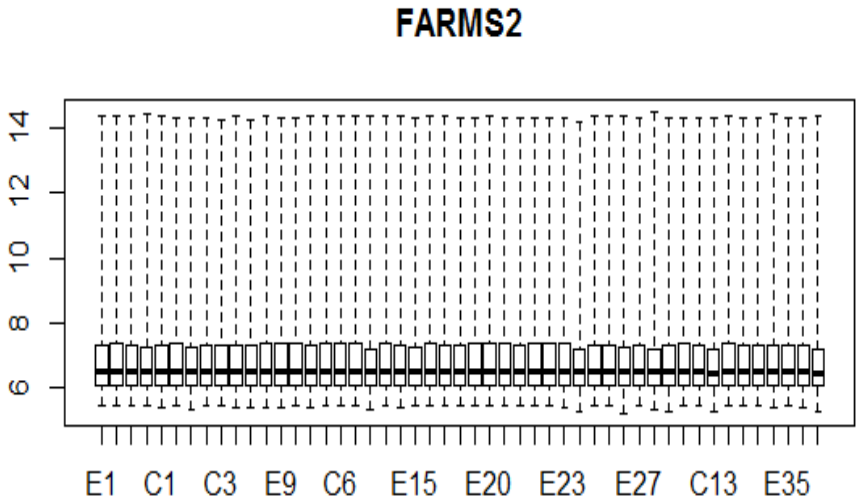

Figura 5.15: *Box plot* dos níveis de expressão após a remoção dos *arrays* C8, E31 e C15 com o método de pré-processamento FARMS.

5.6 Seleção de genes DE

Para se proceder à análise dos 54675 genes através do *arrow plot*, a AUC foi estimada a partir do método de *kernel*, ou seja, pelo método do núcleo, este é referido como sendo um método mais robusto para a seleção de genes. Neste caso o *kernel* utilizado foi o que é definido por defeito no *R*. É importante ter em conta que as AUC estimadas pelo método empírico têm tendência a serem mais otimistas, no entanto, comparando o métodos do núcleo com o método empírico, a maior diferença é notada na seleção de genes com regulação positiva e negativa em que o número destes é significativamente maior no caso da AUC estimada pelo método empírico, o que para o objetivo deste estudo não é relevante, pois apesar de serem genes DE, não apresentam a característica de genes mistos para o pré-diagnóstico do cancro do pâncreas [47].

O ponto de corte definido para o OVL foi de menor ou igual a 0,4. Para a AUC, no que diz respeito aos genes com regulação positiva e de regulação negativa definiu-se um valor maior que 0,9 e menor que 0,1 respetivamente. Para selecionar os genes de interesse para a investigação corrente selecionou-se um valor de AUC entre 0,4 e 0,6 [47].

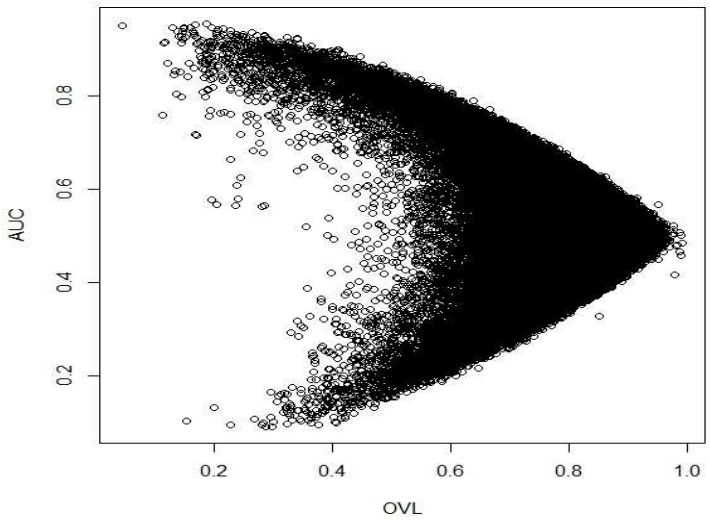

Figura 5.16: *Arrow plot* dos dados do cancro do pâncreas.

Tendo em conta os pontos de corte acima referidos, foram selecionados 10 genes candidatos a mistos, 170 genes com regulação positiva e 7 de regulação negativa tal como é

possível observar na figura 5.17.

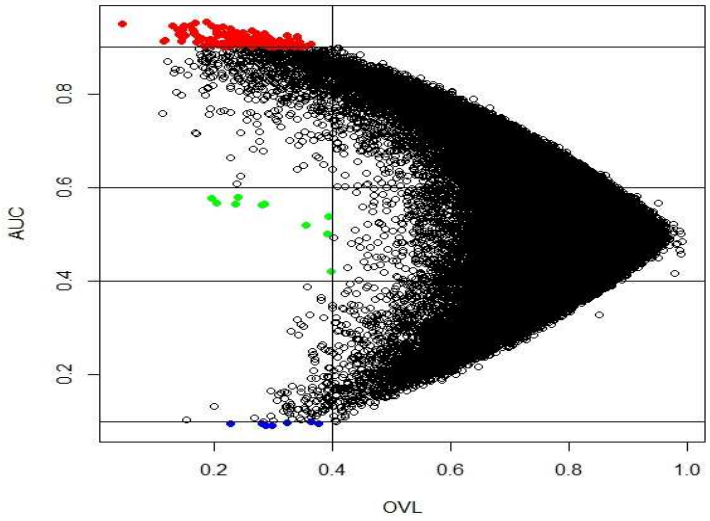

Figura 5.17: *Arrow plot* dos dados do cancro do pâncreas. Pontos a vermelho representam os genes com regulação positiva, os pontos a azul representam os genes de regulação negativa, os pontos a verde representam os genes candidatos a possuírem interesse biológico.

Após a análise da bimodalidade dos genes candidatos a mistos, apenas 3 revelaram não pertencer a este grupo (figura 5.18).

No gráfico da figura 5.18, os pontos a vermelho representam os genes com regulação positiva, os pontos a azul representam os genes com regulação negativa, os pontos a verde representam os genes mistos com bimodalidade em ambos os grupos (controlo e experimental), pontos a laranja representam genes mistos com bimodalidade no grupo experimental, pontos a azul claro representam genes mistos com bimodalidade apenas no grupo de controlo e, por fim, pontos a negro ilustram os genes que após a análise da bimodalidade foram excluídos do grupo de genes mistos.

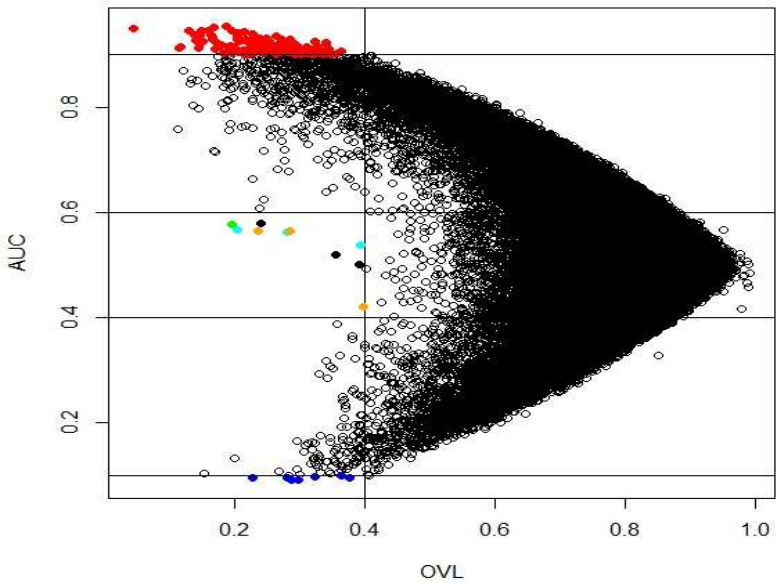

Figura 5.18: *Arrow plot* dos dados do cancro do pâncreas. Pontos a vermelho representam os genes com regulação positiva, os pontos a azul representam os genes de regulação negativa, os pontos a verde representam os genes de interesse biológico com bimodalidade em ambos os grupos, pontos a laranja representam genes de interesse com bimodalidade no grupo experimental, pontos a azul claro representam genes de interesse com bimodalidade apenas no grupo de controlo e pontos a negro representam os genes que após a análise da bimodalidade foram excluídos do grupo de genes com interesse biológico.

5.6.1 Comparação com outros métodos de seleção de genes DE

Considerando os genes DE selecionados pelo *arrow plot*, quando comparado com outros métodos de seleção, obtiveram-se os resultados apresentados na tabela 6. Deste modo, é possível verificar que o método utilizado nesta investigação (*arrow plot*) é, verdadeiramente, o melhor método para a seleção de genes DE, dado que foi o que conseguiu selecionar um maior número destes genes. Isto permite não só visualizar genes com regulação positiva e negativa mas também genes com interesse biológico para o investigador, sendo assim possível enriquecer o estudo em questão.

Tabela 5.1: Comparação dos genes DE selecionados com o *arrow plot* com outros métodos de seleção.

Método proposto	Número de genes DE selecionados
Arrow plot	170
empAUC	93
kerAUC	100
FC	25
WAD	18
AD	24
ibmT	84
modT	84
SAM	84
SAMROC	85

5.6.2 Análise de genes DE mistos

Dos 10 genes candidatos a mistos nesta investigação referidos anteriormente, apenas 7 (genes a verde, laranja e azul claro) revelaram possuir verdadeiramente interesse biológico para o investigador. Dada a (ID) identificação do gene, o próximo passo passou por verificar qual o nome e a função de cada um destes genes. Assim, foi concebida a tabela 5.2.

Através de uma análise da tabela 5.2 constatou-se que 5 dos 7 genes mistos selecionados apresentam-se como sendo **XIST** (transcritos específicos inativos para o cromossoma X) o que significa que são genes que pertencem ao RNA e presentes no cromossoma X dos seres humanos. Foi demonstrado que o **XIST** (gene) interage com o BRCA1, um gene supressor de tumor humano e normalmente expresso nas células de cancro da mama e outros tecidos [3].

Outro gene com interesse biológico é o **MIR1182** (microRNA 1182) e está envolvido na regulação pós-transcricional da expressão genética, afetando tanto a estabilidade como a tradução dos mRNAs [2].

O último gene identificado foi o **C18orf61**, um gene de RNA não codificante, o que significa que não é traduzido numa proteína, sendo possível afirmar que a síntese proteica não vai ocorrer. Com a ausência da síntese proteica, o organismo tem muito mais dificul-

Tabela 5.2: Identificação dos genes com interesse biológico.

ID do Gene	Nome do Gene	AUC	OVL	Bimodalidade
214218_s_at	XIST	0,578	0,195	Ambos os grupos
224588_at	XIST	0,565	0,284	Grupo de controlo
224589_at	XIST	0,563	0,280	Grupo experimental
224590_at	XIST	0,565	0,235	Grupo de controlo
226448_at	MIR1182	0,539	0,393	Grupo experimental
227671_at	XIST	0,568	0,203	Grupo experimental
241943_at	C18orf61	0,421	0,398	Grupo de controlo

dades para reagir positivamente à quimioterapia e outros tratamentos utilizados em casos de cancro [1].

Capítulo 6

Conclusões

Este estudo permitiu confirmar que, tal como já foi mencionado, através da tecnologia de *microarrays* é possível obter um número de genes com as características de interesse pretendidas para uma análise posterior de entre um extensivo número de genes. A aplicação da metodologia ROC posteriormente à tecnologia de *microarrays* permitiu verificar o nível de expressão genética no conjunto de dados propostos. Dada a insuficiência desta metodologia para a seleção de genes, partiu-se para a exploração de outros métodos.

Dos métodos utilizados, o método do *arrow plot* evidenciou-se como sendo o método com a melhor *performance* em comparação com os outros métodos de seleção de genes DE, dado que foi o único a obter um número de genes DE superior a 100 .

Após a análise pelo *arrow plot*, verificou-se que o número total de genes DE com regulação positiva foi de 153, tal significa que em 153 genes verificou-se um nível de expressão genética muito mais elevada na condição experimental (cancro) do que na condição normal, concluindo-se que os genes regulados positivamente são indicadores da condição de cancro do pâncreas nas amostras abrangidas na análise. O oposto acontece nos 7 genes regulados negativamente, significando que não possuem qualquer indicação de doença nas amostras analisadas. Dos 10 genes candidatos a mistos, após a análise da bimodalidade, apenas 7 revelaram características de interesse. São estes genes considerados de interesse biológico e vulgarmente conhecidos como biomarcadores, neste trabalho designados por mistos. Estes biomarcadores permitem conhecer o estado de uma condição patológica ou a resposta desta a um fármaco, no entanto, o seu papel como ferramenta de ajuda ao diagnóstico é ainda controversa.

Dado o estudo por concluído, é ainda possível responder aos objetivos que foram traça-

dos no capítulo 1. O *arrow plot* em associação com a metodologia ROC, constitui o método mais eficaz na seleção e análise de genes DE quando comparado com outros métodos estatísticos mais utilizados. Sendo assim, e de acordo com a informação acima descrita, é possível afirmar que futuramente podem vir a ser desenvolvidas novas terapias, tratamentos e métodos de diagnóstico para certas patologias utilizando o método proposto.

Como trabalho futuro, propõem-se que seja explorada de que forma os genes mistos selecionados influenciam diretamente a doença em estudo e ainda abranger esta técnica de análise a outros tipos de patologias.

Bibliografia

[1] Gene c18orf61. https://www.ncbi.nlm.nih.gov/gene/?term=c18orf61. Consultado em janeiro 2017.

[2] Gene mir1182. https://www.ncbi.nlm.nih.gov/gene/100302132. Consultado em janeiro 2017.

[3] Gene xist. https://www.ncbi.nlm.nih.gov/gene/7503. Consultado em janeiro 2017.

[4] Gse16515. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gse16515. Consultado em julho 2016.

[5] Introdução à tecnologia de microarray. http://www.lce.esalq.usp.br. consultado em janeiro 2016.

[6] one.channel.arrays.analysis.pdf. http://www.bioinformatica.unito.it/downloads/. Consultado em janeiro 2016.

[7] Alexandre Quintas, Ana Ponces Freire, M. J. H. *Bioquímica: Organização Molecular da Vida*. LIDEL, 2008.

[8] Azevedo, C. *Biologia Celular e Molecular*, 4ª edição ed. LIDEL, 2005.

[9] Baran-Gale, J., Fannin, E. E., Kurtz, C. L., and Sethupathy, P. Beta cell 5' -shifted isomirs are candidate regulatory hubs in type 2 diabetes. *PLoS ONE 8*, 9 (09 2013), e73240.

[10] Behzadi P., Behzadi E., R. R. Microarray data analysis. *Albanian Medical Journal*, 4 (2014), 84–90.

[11] Bolstad, B. affyPLM: Model Based QC Assessment of Affymetrix GeneChips, October 2016.

[12] Braga, A. C. *Curvas ROC: Aspectos Funcionais e Aplicações*. PhD thesis, Universidade do Minho, 2000.

[13] Bumgarner, R. DNA microarrays: Types, Applications and their future. *Curr Protoc Mol Biol. 6137*, 206 (2013), 1–17.

[14] Clemons, T. E., and Bradley, E. L. Nonparametric measure of the overlapping coefficient. *Computational Statistics and Data Analysis 34*, 1 (2000), 51–61.

[15] Cleves, M. A. Comparative assessment of three common algorithms for estimating the variance of the area under the nonparametric receiver operating characteristic curve. *Stata Journal 2*, 3 (2002), 280–289.

[16] Colombo, J. REVISÃO A tecnologia de microarray no estudo do câncer de cabeça e pescoço. 64–72.

[17] Cooper, G. M., and Hausman, R. E. *The Cell: A Molecular Approach*. Sinauer Associates, 2011.

[18] Craig Parman, C. H. affyQCReport: A Package to Generate QC Reports for Affymetrix Array Data, October 2016.

[19] Dalman, M. R., Deeter, A., Nimishakavi, G., and Duan, Z.-H. Fold change and p-value cutoffs significantly alter microarray interpretations. *BMC Bioinformatics 13*, 2 (2012), 1–4.

[20] Damjanov, I. *Pathology for the health professions*, 3rd edition ed. Elsevier Saunders, 2006.

[21] Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., and Huber, W. BioMart and Bioconductor: A powerful link between biological databases and microarray data analysis. *Bioinformatics 21*, 16 (2005), 3439–3440.

[22] Dyrskjø T., L., Kruhø ffer, M., Thykjaer, T., Marcussen, N., Jensen, J. L., Mø ller, K., and Ø rntoft, T. F. Gene Expression in the Urinary Bladder : A Common Carcinoma in Situ Gene Expression Signature Exists Disregarding Histopathological Classification Gene Expression in the Urinary Bladder : A Common Carcinoma in Situ Gene Expression Signature Exists Disrega. *Cancer research 64* (2004), 4040–4048.

[23] E., M. C. Basic principles of roc analysis. *Seminars in nuclear medicine 8*, 4 (1978), 283–298.

[24] Erlich, H. A. Polymerase chain reaction. *Journal of Clinical Immunology 9*, 6 (1989), 437–447.

[25] Florkowski, C. M. Sensitivity, specificity, receiver-operating characteristic (ROC) curves and likelihood ratios: communicating the performance of diagnostic tests. *The Clinical biochemist. Reviews / Australian Association of Clinical Biochemists 29 Suppl 1*, August (2008), S83–S87.

[26] Gbarton. ROC Curves Analysis. 1–20.

[27] Gentleman, R. Some Quality Methods for Affymetrix Microarrays, January 2007.

[28] Hanley, J. A., and McNeil, B. J. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology 148*, 3 (1983), 839–843. PMID: 6878708.

[29] Hariharan, R. The analysis of microarray data. *Pharmacogenomics 4* (2003), 477–497.

[30] Hellwig, B., Hengstler, J. G., Schmidt, M., Gehrmann, M. C., Schormann, W., and Rahnenführer, J. Comparison of scores for bimodality of gene expression distributions and genome-wide evaluation of the prognostic relevance of high-scoring genes. *BMC Bioinformatics 11*, 1 (2010), 276.

[31] Hochreiter, S., Clevert, D.-A., and Obermayer, K. A new summarization method for affymetrix probe level data. *Bioinformatics 22*, 8 (2006), 943.

[32] Jeffery, I. B., Higgins, D. G., and Culhane, A. C. Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC bioinformatics 7* (2006), 359.

[33] Kadota, K., Nakai, Y., and Shimizu, K. A weighted average difference method for detecting differentially expressed genes from microarray data. *Algorithms for molecular biology : AMB 3* (2008), 8.

[34] Larson, B. Meet the Overlapping Coefficient: A Measure for Elevator Speeches, June 2014.

[35] Li, W. Application of Volcano Plots in Analyses of mRNA Differential Expressions with Microarrays. *Journal of Bioinformatics and Computational Biology 10*, 6 (2012), 1–25.

[36] Lobo, J. M., Jiménez-Valverde, A., and Real, R. Auc: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography 17*, 2 (2008), 145–151.

[37] McGee, M. A review of: "gene expression studies using affymetrix microarrays, by h. göhlmann and w. talloen". *Journal of Biopharmaceutical Statistics 20*, 2 (2010), 488–490.

[38] Motulsky, H. *Intuitive Biostatistics: A Nonmathematical Guide to Statistical Thinking*. 1995.

[39] Mukherjee, S., and Roberts, S. A theoretical analysis of gene selection. *IEEE Computational Systems Bioinformatics* (2004), 125–135.

[40] Park, S. H., Goo, J. M., and Jo, C.-H. Receiver Operating Characteristic (ROC) Curve: Practical Review for Radiologists. *Korean Journal of Radiology 5*, 1 (2004), 11.

[41] Park, T., Yi, S.-G., Kang, S.-H., Lee, S., Lee, Y.-S., and Simon, R. Evaluation of normalization methods for microarray data. *BMC Bioinformatics 4*, 1 (2003), 33.

[42] Parodi, S., Izzotti, a., and Muselli, M. Re: The central role of receiver operating characteristic (ROC) curves in evaluating tests for the early detection of cancer. *J Natl Cancer Inst 97*, 7 (2003), 511–515.

[43] Parodi, S., Pistoia, V., and Muselli, M. Not proper ROC curves as new tool for the analysis of differentially expressed genes in microarray experiments. *BMC bioinformatics 9* (2008), 410.

[44] Pepe, M. S., Longton, G., Anderson, G. L., and Schummer, M. Selecting Differentially Expressed Genes from Microarray Experiments. 133–142.

[45] Rosikiewicz, M., and Robinson-Rechavi, M. Iqrray, a new method for affymetrix microarray quality control, and the homologous organ conservation score, a new benchmark method for quality control metrics. *Bioinformatics 30*, 10 (2014), 1392–1399.

[46] Schisterman, E. F., Faraggi, D., Reiser, B., and Trevisan, M. Statistical inference for the area under the receiver operating characteristic curve in the presence of random measurement error. *American Journal of Epidemiology 154*, 2 (2001), 174–179.

[47] Silva-Fortes, C. *Aplicação da Metodologia ROC na Análise de Dados Microarrays*. PhD thesis, Faculdade de Ciências de Lisboa, 2012.

[48] Silva-Fortes, C., Amaral Turkman, M. A., and Sousa, L. Arrow plot: a new graphical tool for selecting up and down regulated genes and genes differentially expressed on sample subgroups. *BMC bioinformatics 13* (2012), 147.

[49] Tusher V. G., Tibshirani R, C. G. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America* (2001).

[50] Valones, M. A. A., Guimarães, R. L., Brandão, L. A. C., de Souza, P. R. E., de Albuquerque Tavares Carvalho, A., and Crovela, S. Principles and applications of polymerase chain reaction in medical diagnostic fields: a review. *Brazilian Journal of Microbiology 40*, 1 (2009), 1–11.

[51] Weinburg, R. A. *The biology of cancer*, 2nd edition ed. Taylor and Francis, 2013.

[52] Weiss, H. L., Niwas, S., Grizzle, W. E., and Piyathilake, C. Receiver operating characteristic (roc) to determine cut-off points of biomarkers in lung cancer patients. *Disease Markers 19*, 6 (2004), 273–278.

Apêndices

Tabela A1: Identificação dos *arrays*

Número	ID *array*
1	E1
2	E2
3	E3
4	E4
5	C1
6	E5
7	C2
8	E6
9	C3
10	E7
11	C4
12	E8
13	E9
14	E10
15	C5
16	E11
17	C6
18	E12
19	C7
20	E13
21	E14
22	E15
23	E16
24	C8
25	E17
26	E18
27	E19
28	E20
29	E21
30	C9

Continua na próxima página

Tabela A1 – continuação da página anterior

Número	ID *array*
31	E22
32	C10
33	E23
34	C11
35	E24
36	E25
37	E26
38	E27
39	E28
40	C12
41	E29
42	E30
43	C13
44	E31
45	E32
46	E33
47	C14
48	E34
49	C15
50	E35
51	E36
52	C16