

UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO
OPERACIONAL



**APLICAÇÃO DA METODOLOGIA
ROC NA ANÁLISE DE DADOS DE
*MICROARRAYS***

Carina Soares da Silva Fortes

DOUTORAMENTO EM ESTATÍSTICA E
INVESTIGAÇÃO OPERACIONAL
(Especialidade de Probabilidades e Estatística)

2012

UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO
OPERACIONAL



**APLICAÇÃO DA METODOLOGIA
ROC NA ANÁLISE DE DADOS DE
*MICROARRAYS***

Carina Soares da Silva Fortes

Tese orientada pela Professora Doutora Maria Antónia Amaral Turkman e
pela Professora Doutora Lisete Sousa, especialmente elaborada para a
obtenção do grau de doutor em Estatística e Investigação Operacional
(Especialidade de Probabilidades e Estatística)

2012

Ao meu filho Diogo

*A ciência não pode resolver
o mistério final da natureza.
E isto porque, em última análise,
somos parte do mistério que tentamos resolver.*

Max Planck
Físico

Agradecimentos

Quero deixar o meu profundo agradecimento a todos aqueles que contribuiram direta ou indiretamente para a concretização deste trabalho. Em particular,

- às minhas orientadoras Professora Doutora Antónia Turkman e Professora Doutora Lisete Sousa, pela dedicação que sempre mostraram, orientando-me e conduzindo-me ao longo deste trabalho, mesmo nas alturas mais difíceis;
- à Fundação para a Ciência e Tecnologia, pelo suporte financeiro através da bolsa de doutoramento com a referência SFRH/ BD/ 45938/ 2008;
- ao Centro de Estatística e Aplicações da Universidade de Lisboa, pelo apoio da minha presença em eventos científicos que muito contribuiram para a aquisição de conhecimentos e divulgação do trabalho aqui desenvolvido;
- aos Biólogos Margarida Gama-Carvalho do IMM-FMUL, Lisete Fernandes da ESTeSL-IPL, Miguel Brito da ESTeSL-IPL e Ana Rita Grossou do IMM-FMUL, que contribuiram com bases de dados, partilha de conhecimento, descrição biológica dos genes selecionados, sugestões e esclarecimentos;
- à Júlia Teles, pelas valiosas sugestões e correções na elaboração do texto;
- aos meus amigos, que me apoiaram mesmo nas minhas longas ausências e pelo carinho em momentos mais difíceis;
- à minha sogra Judite Fortes, que em todos os momentos me apoiou e ajudou, particularmente com o Diogo;
- aos meus pais, porque sempre acreditaram em mim e pelo apoio incondicional em todos os momentos;
- ao meu marido Rui Fortes pela amizade e companheirismo, pela paciência nas minhas ausências, por ouvir sempre e por ter incutido o gosto pela programação;
- ao meu filho Diogo, a fonte da minha inspiração e alento, porque com ele tudo é mais fácil e por ele tudo vale a pena.

Resumo

Um objetivo muito comum na análise de dados de *microarrays* é determinar que genes são diferencialmente expressos sob dois (ou mais) tipos de tecido ou sob amostras submetidas a diferentes condições experimentais. Sabe-se que as amostras biológicas são heterogêneas devido a vários fatores, como por exemplo, antecedentes genéticos e subtipos moleculares, os quais são, na maior parte das vezes, do desconhecimento do investigador. Por exemplo, em experiências que envolvam a classificação de tumores é importante que se identifiquem subtipos do cancro em investigação. Distribuições bimodais ou multimodais geralmente refletem a presença de misturas de subclasses. Consequentemente, pode haver genes que sendo diferencialmente expressos (DE) quando se tem em conta os diferentes subgrupos, não são identificados pelos métodos usualmente utilizados para selecionar genes DE. Neste trabalho propõe-se uma nova representação gráfica que não só permite identificar genes com regulação positiva e regulação negativa, mas também genes DE em subgrupos. Esta ferramenta baseia-se em duas medidas, nomeadamente na área abaixo da curva (AUC) *receiver operating characteristic* (ROC) e no coeficiente de sobreposição entre duas densidades (OVL). Para a estimação do OVL desenvolveu-se um algoritmo que permite obter uma estimativa não-paramétrica desse coeficiente e que se baseia em determinar a área de sobreposição de duas densidades estimadas pelo método do núcleo. Foi também desenvolvido um algoritmo para identificar distribuições bimodais ou multimodais estimadas pelo método do núcleo, permitindo deste modo identificar os genes com as características acima descritas. A metodologia aqui proposta foi implementada em linguagem R. Compararam-se os resultados com os resultados obtidos através de métodos usualmente aplicados na seleção de genes DE, usando-se dados simulados e duas bases de dados disponíveis publicamente. Os resultados indicam que a nova ferramenta, *Arrow plot*, apresenta uma elevada performance na seleção de genes com diferentes tipos de expressão diferencial, sendo flexível e útil na análise de perfis da expressão de genes em dados de *microarrays*.

Palavras-chave: *Microarrays*, curvas ROC degeneradas, área abaixo da curva ROC, coeficiente de sobreposição, estimador do núcleo, *Arrow plot*.

Abstract

A common task in analyzing microarray data is to determine which genes are differentially expressed under two (or more) kinds of tissue samples or samples submitted under different experimental conditions. It is well known that biological samples are heterogeneous due to factors such as molecular subtypes or genetic background, which are often unknown to the investigator. For instance, in experiments which involve molecular classification of tumors it is important to identify significant subtypes of cancer. Bimodal or multimodal distributions often reflect the presence of subsamples mixtures. Consequently, truly differentially expressed genes on sample subgroups may be lost if usual statistical approaches are used. In this work it is proposed a new graphical tool which identifies genes with up and down regulation, as well as genes with differential expression which reveals hidden subclasses, that are usually missed if current statistical methods are used. This tool is based on two measures, namely the overlapping coefficient (OVL) between two densities and the area under (AUC) the receiver operating characteristic (ROC) curve. In order to estimate the OVL coefficient it is proposed an algorithm where a naive kernel density estimator is used to develop a nonparametric estimator of the OVL. It is also developed an algorithm to detect bimodal or multimodal kernel based estimated densities, essential to select genes with differential expression which reveals hidden subclasses. The methodology proposed here was implemented in the open-source R software. Results are compared with the ones obtained using some of the standard methods for detecting differentially expressed genes, both for simulated and public data sets. The results indicate that the *Arrow plot* represents a new flexible and useful tool for the analysis of gene expression profiles from microarrays.

Keywords: Microarrays, not proper ROC curves, area under ROC curve, overlapping coefficient, kernel estimator, Arrow plot.

Conteúdo

Lista de Figuras	xiii
Lista de Tabelas	xvi
Lista de Algoritmos	xvii
Lista de Siglas e Acrónimos	xix
Nota Introdutória	1
1 Breve Referência à Biologia Molecular	5
1.1 Introdução	5
1.2 O dogma central da Biologia Molecular	6
1.3 Biotecnologia	13
1.3.1 Sequenciação de DNA	14
1.3.2 Reação de Polimerização em Cadeia	14
1.3.3 Transcrição Reversa da Reação de Polimerização em Cadeia	16
1.3.4 <i>Microarrays</i>	16
1.4 Algumas considerações finais	23
2 Análise de dados de <i>microarrays</i> de um canal	25
2.1 Introdução	25
2.2 Pré-processamento	28
2.2.1 Filtragem	30
2.2.2 Valores omissos	30
2.2.3 Transformação dos dados	32
2.2.4 Avaliação da qualidade dos dados	33
2.2.5 Pré-processamento	44
2.2.6 Algumas considerações finais	52
2.3 Métodos para a seleção de genes DE	53
2.3.1 Introdução	53

2.3.2	Estado da Arte	54
2.3.3	Algumas considerações finais	63
3	Metodologia ROC na análise de dados de <i>microarrays</i>	65
3.1	Introdução	65
3.2	Definição e Propriedades da curva ROC	66
3.3	AUC e curvas ROC degeneradas	68
3.3.1	Métodos de estimação não-paramétricos da curva ROC e da AUC	72
3.4	AUC empírica <i>vs.</i> AUC núcleo — comparação do viés	77
3.5	Métodos baseados na metodologia ROC para a seleção de genes	81
3.5.1	Métodos de ordenação de genes propostos por Pepe <i>et al.</i> (2003)	81
3.5.2	SAMROC	82
3.5.3	Métodos propostos por Parodi <i>et al.</i> (2008)	84
3.6	Algumas considerações finais	86
4	Arrow Plot	87
4.1	Introdução	87
4.2	Coeficiente de sobreposição — OVL	89
4.2.1	Estimação não-paramétrica do OVL	90
4.3	<i>Arrow plot</i>	101
4.4	Considerações finais	108
5	Aplicações	111
5.1	Dados Simulados	112
5.1.1	Introdução	112
5.1.2	Seleção de genes DE e mistos	114
5.1.3	Comparação da performance com outros métodos	117
5.1.4	<i>Arrow plot</i> — AUC empírica <i>vs.</i> AUC núcleo	120
5.1.5	<i>Arrow plot</i> <i>vs.</i> <i>Volcano plot</i>	123
5.1.6	Considerações finais	123
5.2	Dados <i>Cancro da Bexiga</i> — Dyrskjot <i>et al.</i> (2004)	124
5.2.1	Introdução	124
5.2.2	Análise da qualidade dos dados	127
5.2.3	Pré-processamento	134
5.2.4	Seleção de genes DE e mistos	135
5.2.5	Comparação com outros métodos	138
5.2.6	<i>Arrow plot</i> — AUC empírica <i>vs.</i> AUC núcleo.	140
5.2.7	Considerações finais	143
5.3	Dados <i>Linfoma</i> — Alizadeh <i>et al.</i> (2000)	143

5.3.1	Introdução	143
5.3.2	Seleção de genes DE e mistos	145
5.3.3	Comparação com outros métodos	150
5.3.4	<i>Arrow plot</i> — AUC empírica <i>vs.</i> AUC núcleo	151
5.3.5	Considerações Finais	152
6	Conclusões	155
A	Código em R	161
A.1	Algoritmo 1	161
A.2	Algoritmo 2	164
A.3	Algoritmo 3	164
B	Figuras e Tabelas	167
B.1	Dados Cancro da Bexiga	167
B.2	Dados Linfoma	170
Referências		173

Listas de Figuras

1.1	Célula eucariótica	7
1.2	Estrutura do cromossoma	8
1.3	Estrutura do DNA	9
1.4	Relação entre os codões (no RNA) e os aminoácidos codificados	9
1.5	Fluxo de informação genética numa célula	10
1.6	Transformação da linguagem de DNA na linguagem proteica .	10
1.7	Molécula de RNA	11
1.8	Constituição de um gene	12
1.9	<i>Splicing</i>	13
1.10	Ciclo do PCR	15
1.11	<i>Microarrays</i> de cDNA	18
1.12	<i>Microarray</i> de um canal	19
1.13	Representação esquemática de um gene.	21
1.14	Representação de um <i>probeset</i>	21
1.15	Do <i>chip</i> à imagem	22
1.16	Parte da imagem de um <i>chip</i> com grelha sobreposta.	23
2.1	Esquema resumo de alguns ficheiros obtidos em <i>microarrays</i> da Affymetrix.	28
2.2	Fatores que influenciam a qualidade final dos dados	29
2.3	<i>Box plot</i> dos níveis de intensidade dos <i>arrays</i>	36
2.4	Densidades empíricas dos <i>arrays</i>	37
2.5	<i>Degradation plot</i>	38
2.6	<i>Simple Affy Plot</i>	41
2.7	Gráficos NUSE e RLE	42
2.8	Imagens dos <i>arrays</i>	43
2.9	Gráfico MA	47
3.1	Exemplo de uma Curva ROC.	68
3.2	Relação entre densidades e curvas ROC considerando variâncias iguais	71

3.3	Comparação das estimativas <i>bootstrap</i> da AUC estimada de forma empírica e da AUC estimada pelo método do núcleo com o valor exato	80
3.4	Curva ROC em função de PFP e PFN	83
4.1	Relação entre densidades e curvas ROC considerando variâncias diferentes, médias iguais e distribuições unimodais .	88
4.2	Área de sobreposição entre duas densidades — OVL	89
4.3	Representação gráfica dos pontos das densidades estimadas pelo método do núcleo pertencentes à região de interseção das duas densidades.	92
4.4	Comparação entre o valor exato do OVL e respetivas estimativas pelo método do núcleo.	99
4.5	Representação do OVL relativamente a genes com regulação positiva e negativa.	102
4.6	Representação do OVL de um gene misto.	103
5.1	<i>Box plot</i> — dados simulados.	115
5.2	<i>Arrow plot</i> — dados simulados. Distribuição dos genes de acordo com o seu verdadeiro estado	116
5.3	<i>Arrow plot</i> — dados simulados. Pontos de corte.	117
5.4	Curvas ROC empíricas. Comparação da performance na seleção de genes mistos.	119
5.5	Curvas ROC empíricas. Comparação da performance de métodos na seleção de genes DE e mistos.	121
5.6	Dados simulados — comparação do <i>Arrow plot</i> considerando a AUC estimada pelo método empírico e pelo método do núcleo.	122
5.7	<i>Volcano plot</i> — Dados simulados.	124
5.8	Representação dos vários tipos de amostras de tecido analisadas no estudo do <i>Cancro da Bexiga</i>	126
5.9	Densidades dos logaritmos dos níveis de intensidade PM em bruto — <i>Cancro da Bexiga</i>	127
5.10	<i>Box plot</i> dos logaritmos dos níveis de intensidade PM dos <i>arrays</i> relativos ao estudo do <i>Cancro da Bexiga</i>	128
5.11	<i>Degradation plot</i> — <i>Cancro da Bexiga</i>	129
5.12	Gráficos MA dos dados em bruto — <i>Cancro da Bexiga</i> — dados	130
5.13	Gráfico NUZE — <i>Cancro da Bexiga</i>	131
5.14	Gráfico RLE — <i>Cancro da Bexiga</i>	132
5.15	Gráfico QC dos dados em bruto — <i>Cancro da Bexiga</i>	133

Lista de Figuras

5.16	<i>Box plots</i> dos logaritmos dos níveis de expressão dos <i>arrays</i> após pré-processamento RMA e GCRMA— <i>Cancro da Bexiga</i> .	134
5.17	<i>Box plots</i> dos logaritmos dos níveis de expressão dos <i>arrays</i> após pré-processamento PLIER e FARMS— <i>Cancro da Bexiga</i> .	135
5.18	<i>Box plots</i> dos logaritmos dos níveis de expressão dos <i>arrays</i> após pré-processamento MAS5 e MBEI — <i>Cancro da Bexiga</i> .	136
5.19	Comparação das densidades dos logaritmos dos níveis de expressão dos <i>arrays</i> após RMA e FARMS — <i>Cancro da Bexiga</i> .	137
5.20	Gráfico MA após pré-processamento FARMS considerando a base de dados após transformação.	137
5.21	<i>Arrow plot</i> — Dados <i>Cancro da Bexiga</i> .	138
5.22	<i>Arrow plot</i> — Dados <i>Cancro da Bexiga</i> . Pontos de corte.	139
5.23	<i>Arrow plot</i> — AUC empírica <i>vs.</i> AUC núcleo.	141
5.24	<i>Box plots</i> dos logaritmos dos níveis de expressão dos dados <i>Linfoma</i> .	145
5.25	<i>Arrow plot</i> dos dados <i>Linfoma</i> .	146
5.26	<i>Arrow plot</i> dos dados <i>Linfoma</i> — selecção de genes DE e mistos.	147
5.27	Densidades e curvas ROC empíricas dos genes mistos dos dados <i>Linfoma</i> .	149
5.28	<i>Arrow plot</i> — AUC estimada pelo método empírico <i>vs.</i> pelo método do núcleo.	151
B.1	Gráfico QC após pré-processamento FARMS.	168
B.2	Gráfico QC após remoção do <i>array</i> C9.	169

Listas de Tabelas

2.1	Métodos de imputação valores omissos em <i>microarrays</i>	31
2.2	Exemplo de uma matriz de expressão.	53
2.3	Matriz de expressão	54
3.1	Estimativas <i>bootstrap</i> da AUC considerando os métodos empírico e do núcleo para $n = 500, 100$	79
3.2	Estimativas <i>bootstrap</i> da AUC considerando os métodos empírico e do núcleo para $n = 30, 15$	80
4.1	Lista de notações utilizadas no Algoritmo 1.	93
4.2	Lista de funções utilizadas no Algoritmo 1.	94
4.3	Estimativas da média MC, erro padrão MC e viés do OVL estimado pelo Algoritmo 1.	98
4.4	Erro padrão e viés relativo das estimativas <i>bootstrap</i> do OVL estimado pelo Algoritmo 1.	101
4.5	Lista de notações utilizadas nos Algoritmo 2 e 3.	104
4.6	Lista de funções utilizadas nos Algoritmos 2 e 3.	104
4.7	Lista de notações utilizadas no Algoritmo 3.	106
4.8	Lista de funções utilizadas no algoritmo 3.	107
5.1	Comparação da performance de métodos na seleção de genes mistos — AUC.	119
5.2	Comparação da performance de métodos na seleção de genes DE e mistos — AUC.	120
5.3	Lista de genes mistos nos dados <i>Cancro da Bexiga</i>	140
5.4	Comparação do número de genes selecionados considerando o <i>Arrow plot</i> com AUC estimada pelos métodos do núcleo e empírico, para os pontos de corte definidos anteriormente.	142
5.5	<i>Arrays</i> selecionados da base de dados original do estudo de Alizadeh <i>et al.</i> (2000).	144
5.6	Valores da AUC e OVL dos genes mistos nos dados <i>Linfoma</i>	148

Lista de Tabelas

5.7	Comparação do número de genes considerando o <i>Arrow plot</i> com AUC estimada pelos métodos do núcleo e empírico, considerando os pontos de corte anteriormente definidos.	152
B.1	Descrição biológica dos genes mistos selecionados a partir do <i>Arrow plot</i> nos dados <i>Linfoma</i>	170
B.2	Descrição biológica dos genes mistos selecionados a partir do <i>Arrow plot</i> nos dados <i>Linfoma</i> (cont.).	171

Listas de Algoritmos

1	Pseudo-código para a estimação não-paramétrica do OVL (Silva-Fortes <i>et al.</i> , 2012)	95
2	Pseudo-código para seleção de genes candidatos a genes mistos e identificação de distribuições bimodais (ou multimodais).	105
3	Pseudo-código para a construção do <i>Arrow plot</i> e identificação de genes com regulação positiva, regulação negativa e genes mistos.	107

Listas de Siglas e Acrónimos

A	adenina
ABCR	<i>area between ROC curve and reference</i>
AD	<i>average difference</i>
AGC	<i>array generation based gene centering</i>
AUC	<i>area under the curve</i>
B-CLL	leucemia linfocítica crônica de células B
BPCA	<i>bayesian principal component analysis</i>
C	citosina
CBN	células B normais
cDNA	<i>complementary DNA</i>
CIM	carcinomas invasivos do músculo
CIS	carcinoma <i>in situ</i>
CMVE	<i>collateral missing value estimation</i>
CST	carcinoma superficial de transição
dChip	<i>DNA-Chip analyzer</i>
DE	diferencialmente expresso
DFCM	<i>distribution free convolution model</i>
DFW	<i>distribution free weighted</i>
DLBCL	linfoma difuso de células B grandes
DNA	<i>deoxyribonucleic acid</i>
E	especificidade
EBAM	<i>empirical bayesian analysis of microarrays</i>
EQMI	erro quadrático médio integrado
FAR	<i>factor analysis regression</i>
FARMS	<i>factor analysis for robust microarray summarization</i>
FC	<i>fold change</i>
FDR	<i>false discovery rate</i>
FL	<i>follicular linfoma</i>
FWER	<i>family wise error rate</i>
G	guanina
GCRMA	<i>guanine cytosine robust multiarray analysis</i>
gFWER	<i>generalized family-wise error rate</i>
GMC	<i>gaussian mixture clustering</i>

Lista de Siglas e Acrónimos

ibmT	<i>intensity-based moderated t-statistic</i>
IM	<i>ideal mismatch</i>
iMSS	<i>integrative missing value estimation</i>
IP	imunoprecipitação
IQR	<i>interquartile range</i>
k-NN	<i>k-nearest neighbors</i>
LLSI	<i>local least square impute</i>
Linlmp	<i>linear model based imputation</i>
MAD	<i>median absolute deviance</i>
MAS	<i>Affymetrix microarray suite</i>
MBEI	<i>model-based expression index</i>
MM	<i>mismatch</i>
modT	<i>moderated t-statistic</i>
MPSS	<i>massively parallel signature sequencing technology</i>
mRNA	<i>messenger RNA</i>
NPROC	<i>not proper ROC curve</i>
NUSE	<i>normalized unscaled standard error</i>
OLSI	<i>ordinary least square input</i>
OVL	<i>overlapping coefficient</i>
pAUC	<i>partial area under the curve</i>
pb	pares de bases
PCR	reação de polimerização em cadeia
PDNN	<i>positional-dependent-nearest-neighbor</i>
PLIER	<i>probe logarithmic intensity error estimation</i>
PFN	proporção de falsos negativos
PFP	proporção de falsos positivos
PVN	proporção de verdadeiros negativos
PVP	proporção de verdadeiros positivos
PLM	<i>probe-level model</i>
POCS	<i>projection onto convex steps</i>
PM	<i>perfect match</i>
RLE	<i>relative log expression</i>
RMA	<i>robust multiarray analysis</i>
RNA	<i>ribonucleic acid</i>
ROC	<i>receiver operating characteristic</i>

RP	<i>rank products</i>
rRNA	<i>ribosomic RNA</i>
RT-PCR	<i>reverse transcription PCR</i>
S	sensibilidade
SAGE	<i>serial analysis of gene expression</i>
SAM	<i>significance analysis of microarrays</i>
SkNN	<i>sequential KNN</i>
sRMA	<i>small sample RMA</i>
SPA	<i>spliceosome associated proteins</i>
SPLOSH	<i>spacings loess histogram</i>
T	timina
TNRC	<i>test for not-proper ROC curves</i>
tRNA	<i>transfer RNA</i>
U	uracilo
VSN	<i>variance stabilization and normalization</i>
SVD	<i>singular value decomposition</i>
SVR	<i>support vector regression</i>
WAD	<i>weighted average difference</i>
WBL	<i>Weibull distribution based normalization</i>
WTSS	<i>Whole Transcriptome Shotgun Sequencing</i>

Nota Introdutória

A análise de dados da expressão genética é um dos grandes desafios que a Estatística e a Biologia têm enfrentado nos últimos anos. A mensuração da expressão genética a partir dos transcritos (mRNA) pode ser obtida através de várias tecnologias, tais como os *microarrays* ou o SAGE (*Serial Analysis of Gene Expression*). A análise destes dados pode fornecer informações importantes acerca das funções de uma célula, os mecanismos de regulação dos genes e as diferenças moleculares entre vários tipos de células ou tecidos.

A análise estatística pode ser realizada em diferentes níveis e de diferentes formas, variando de análises de baixo nível (*low-level analysis*), como análise da qualidade de imagem, normalização, correção de *background*, até análises de elevado nível (*high-level analysis*), como identificação de genes diferencialmente expressos, identificação de padrões, etc.

A maior parte das experiências associadas a *microarrays* têm como objetivo identificar genes cujos níveis de expressão variam em relação a alguma condição específica ou em resposta a um determinado estímulo. Os métodos mais comumente utilizados baseiam-se na comparação dos valores médios entre dois ou mais grupos, mesmo que existam subclasses escondidas num dos grupos (ou em ambos) com diferentes níveis de expressão.

É importante que se reconheça que diferentes tipos de questões podem ser colocadas numa experiência de *microarrays*. As abordagens estatísticas a utilizar terão que dar resposta a essas questões. Por exemplo, em experiências que envolvam diferentes tipos de tecido, como tecido com cancro e tecido sem cancro, incluem i) seleção de genes que são diferencialmente expressos nos diferentes tipos de tecido; ii) identificação de uma combinação de genes que providenciem a discriminação entre os diferentes tipos de tecido e iii) identificação de grupos de genes cujos níveis de expressão sejam correlacionados. Técnicas estatísticas como por exemplo análise de regressão e a análise discriminante são adequadas para ii), análise de clusters é

Nota Introdutória

apropriada para iii). Neste trabalho abordam-se métodos estatísticos com o objetivo de dar resposta a i).

A motivação particular para o desenvolvimento deste trabalho, foi a procura de genes que normalmente não são identificados pelos métodos estatísticos vulgarmente utilizados na identificação de genes com expressão diferencial, e que podem fornecer informações úteis acerca das funções das células, como por exemplo genes que possam ser biomarcadores de cancros e que possam ser utilizados no rastreio populacional. Estes genes serão designados neste trabalho por genes mistos.

Sendo este trabalho baseado na análise de dados provenientes de *microarrays*, no capítulo 1 faz-se uma breve introdução à biologia molecular e à tecnologia de *microarrays*, em particular de um canal da Affymetrix.

Os dados provenientes de *microarrays* são sujeitos a uma análise estatística preliminar até se obter a base de dados final para a seleção de genes diferencialmente expressos. Este processo, designado de *low-level analysis*, é complexo e a eleição das técnicas estatísticas usadas influenciam a qualidade final dos dados e a análise subsequente. Nesse sentido, no capítulo 2 descrevem-se os métodos usualmente utilizados para a análise da qualidade dos dados e de pré-processamento. Uma vez que o objetivo principal é desenvolver um método que permita selecionar genes diferencialmente expressos entre duas condições experimentais, neste capítulo descrevem-se alguns dos métodos mais utilizados para a seleção de genes DE entre duas condições experimentais.

Neste trabalho a metodologia *receiver operating characteristic* (ROC) é a principal ferramenta utilizada para a seleção de genes com as características acima descritas. Assim, o terceiro capítulo é dedicado à descrição das principais características da curva ROC e da sua aplicação na análise de dados de *microarrays*.

No capítulo 4 apresenta-se uma nova ferramenta na análise da expressão diferencial, o gráfico *Arrow plot*. Este gráfico é construído com base nas estimativas da área abaixo da curva ROC (AUC) e no coeficiente de sobreposição entre duas densidades (OVL). A abordagem não-paramétrica foi a eleita para o desenvolvimento do método proposto, pelo que se apresenta no capítulo 4 um algoritmo para a estimação não-paramétrica do OVL tendo por base densidades estimadas pelo método do núcleo. De modo a analisar o viés e o erro padrão do OVL estimado pelo algoritmo proposto, é realizada

uma análise de simulação usando métodos de Monte Carlo e *bootstrap*. A seleção dos genes mistos, para além da análise da AUC e OVL, pressupõe uma análise da bimodalidade das distribuições que representam os níveis de expressão dos genes nos dois grupos em estudo. Nesse sentido, neste capítulo é ainda desenvolvido um algoritmo para identificar distribuições bimodais ou multimodais estimadas pelo método do núcleo.

O método proposto para além dos genes mistos, permite selecionar genes com regulação positiva e negativa. No quinto capítulo apresenta-se uma análise da performance do método proposto em comparação com os métodos descritos nos capítulos 2 e 3. São utilizadas duas bases de dados disponíveis publicamente e uma base de dados simulada.

No capítulo 6 apresentam-se as principais conclusões e reflexões, e algumas linhas orientadoras para trabalho futuro.

Os apêndices são divididos em duas partes, apresentando-se no apêndice A o código em linguagem R relativo aos algoritmos propostos no trabalho e no apêndice B apresentam-se gráficos e tabelas relativos aos dados analisados no capítulo 5. Muito trabalho foi despendido na implementação em R em estudos de simulação e na análise de dados apresentada no capítulo 5 e, para não pesar o texto, foram colocados em CD os ficheiros R e ficheiros com os dados em condições de serem executados pelo leitor.

Capítulo 1

Breve Referência à Biologia Molecular

1.1 Introdução

O estudo do genoma humano tem proporcionado um vasto campo de investigação em várias ciências e conduzido ao aparecimento de novas disciplinas nos planos de estudo de alguns cursos, nomeadamente nos cursos de Estatística e Informática. O estudo dos genes e das suas funções tornou-se o Santo Graal na busca de respostas às questões que emergem da complexidade do funcionamento dos organismos vivos. A Genómica Funcional (Felix *et al.*, 2002) consiste no estudo da função dos genes e das suas relações com as doenças. Tradicionalmente os métodos existentes na Biologia Molecular estudavam um gene de cada vez (os humanos podem ter cerca de 30.000 a 40.000 genes e cerca de 1.000.000 de proteínas), o que tornava difícil compreender o mapa geral da vida. Nos últimos anos, uma nova tecnologia denominada de *microarrays*¹ tornou possível estudar milhares de genes simultaneamente. Esta tecnologia pode ser utilizada no estudo de grupos de genes envolvidos num determinado processo biológico ou numa doença em particular, identificando-se os vários genes cujos níveis de expressão se alteram em simultâneo sob certas circunstâncias.

Um *microarray* produz milhares de dados, tornando esta técnica da Biologia Molecular um atrativo para a exploração quer de métodos de análise estatística quer no desenvolvimento de capacidades e algoritmos

¹

Suporte sólido onde são depositadas gotas individuais de material genético dispostas de forma matricial que permite a sua análise em simultâneo, com o objetivo de alcançar um maior rendimento e velocidade.

computacionais para os suportar. O conhecimento dos processos biológicos associados a esta tecnologia é indispensável para o desenvolvimento de técnicas estatísticas para tratamento e análise de tal volume de dados. Nesse sentido, neste capítulo faz-se uma breve introdução à Biologia Molecular, enfatizando os processos biológicos e técnicos que estão relacionados com as experiências analisadas ao longo do trabalho. No entanto, pretende-se apenas apresentar definições necessárias para que os processos biológicos sejam entendidos e, portanto, esta caracterização é feita de forma simplificada.

1.2 O dogma central da Biologia Molecular

Porque diferem os seres vivos uns dos outros?

Uma explicação para esta diversidade reside no DNA². Esta molécula é o suporte da informação biológica e define as características de cada organismo.

A genética teve o seu ínicio em 1866, quando o monge Gregor Mendel realizou uma série de experiências que apontavam para a existência de elementos biológicos denominados de *genes*. Todos os seres humanos têm cerca de 50.000 a 100.000 genes diferentes no núcleo de cada célula do corpo. Os genes influenciam o funcionamento e o desenvolvimento dos órgãos e determinam a produção de proteínas. Mutações genéticas são responsáveis por uma série de doenças, como cancro, fibrose quística e esquizofrenia. Todas as células eucarióticas³ da mesma espécie (Figura 1.1) contêm o mesmo número de cromossomas (23 no caso dos humanos). Os cromossomas são estruturas constituídas por DNA e proteínas (Figura 1.2). O DNA é composto por nucleotídos e estes são constituídos por um grupo fosfato, uma pentose e por uma base azotada. Existem cinco tipos de bases azotadas, Adenina (A), Guanina (G), Citosina (C), Timina (T) e Uracilo (U), através das quais se forma o alfabeto genético. O comprimento do DNA humano é cerca de 3×10^9 pares de bases (pb) e cada indivíduo tem associado um código diferente. Apenas cerca de 0.1% varia de uma pessoa para outra em função da combinação dos genomas dos pais.

A estrutura do DNA (Figura 1.3) é descrita como uma dupla hélice que se assemelha a uma escada enrolada helicoidalmente, em que as bandas laterais são cadeias de nucleotídos ligados entre si por fosfatos e açúcares (cadeias polinucleotídicas), e os “degraus” centrais são pares de bases ligadas entre

²Do inglês *deoxyribonucleic acid*.

³

Células que possuem um núcleo com um envólucro nuclear, através do qual ocorrem trocas seletivas entre o núcleo e o citoplasma.

1.2. O dogma central da Biologia Molecular

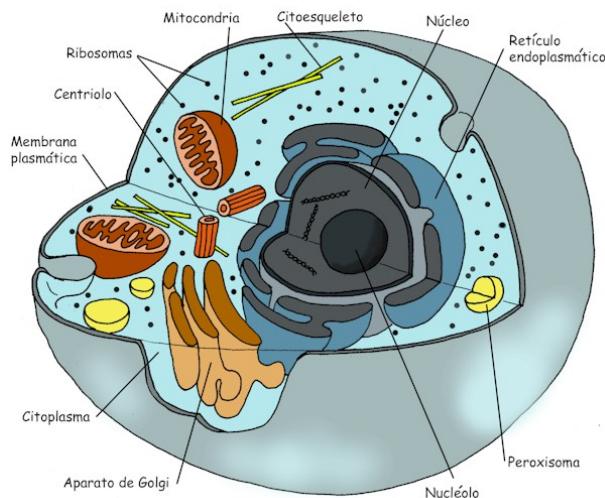


Figura 1.1: Célula eucariótica

(fonte: <http://www.ajudaalunos.com/cn/capi2.htm> em 3/02/2008).

si por pontes de hidrogénio. A adenina liga-se à timina por duas pontes de hidrogénio e a guanina liga-se à citosina por três pontes de hidrogénio. As bases que se emparelham dizem-se bases complementares. As duas cadeias polinucleotídicas da dupla hélice orientam-se em direções opostas. Cada uma inicia-se por uma extremidade de carbono 5' e termina no carbono 3', designando-se por cadeias antiparalelas.

Os genes, por sua vez, correspondem a regiões da molécula de DNA. A existência de células diferentes, tais como por exemplo as células da pele e as células do fígado, deve-se ao facto de que a uma dada altura um determinado grupo de genes é ativado configurando assim propriedades únicas a cada tipo de células.

As moléculas de DNA e as moléculas de proteínas são macromoléculas constituídas por, respetivamente, nucleótidos e aminoácidos. Cada aminoácido é codificado por sequências de três bases (codões), conhecendo-se cerca de 20 aminoácidos comuns a todos os organismos (Figura 1.4).

A ordenação dos aminoácidos numa molécula proteica confere-lhe características e funções biológicas muito específicas. A informação para a ordenação dos aminoácidos está contida no DNA sob a forma de um código que reside na sequência das bases (A, C, T e G). O processo resume-se, basicamente, na conversão da linguagem codificada do DNA (sequência de nucleótidos) para a linguagem das proteínas (sequência de aminoácidos)

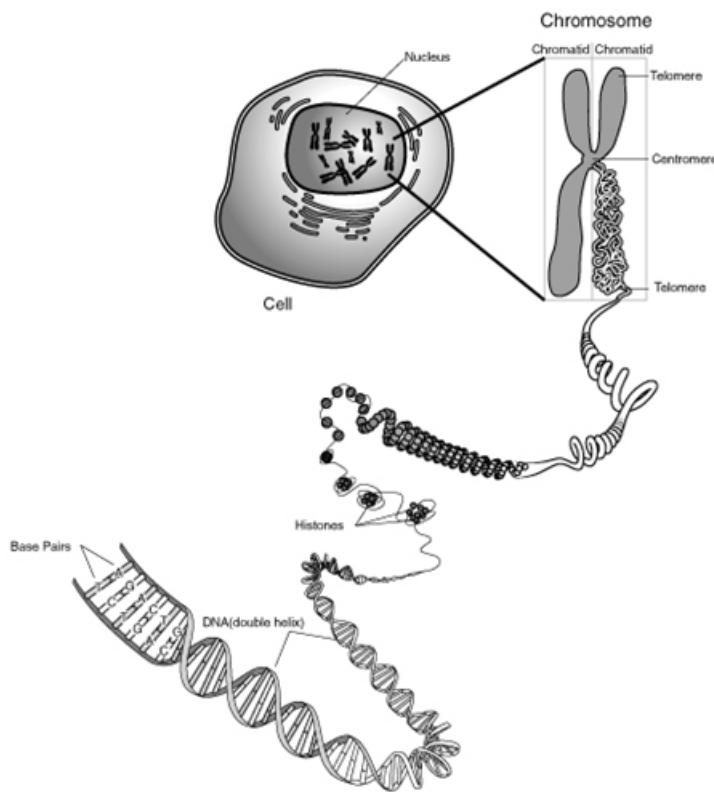


Figura 1.2: Estrutura do cromossoma
(fonte: <http://www.accessexcellence.org/> em 03/08/2008).

(Figura 1.5).

Este processo resulta de várias fases, nomeadamente na passagem da informação contida no DNA para uma sequência de nucleótidos que constituem uma molécula de RNA⁴, o mRNA⁵ (que serve de intermediário entre o espaço nuclear e os ribossomos citoplasmáticos e onde se processa a síntese proteica) abandona o núcleo, transportando a mensagem, ainda em código, para os ribossomas onde a mensagem é descodificada, ou seja traduzida para linguagem proteica (Figura 1.6).

O processo de sintetização das proteínas a partir do DNA ocorre em três

⁴Do inglês *ribonucleotide acid*.

⁵Do inglês *messenger RNA*.

1.2. O dogma central da Biologia Molecular

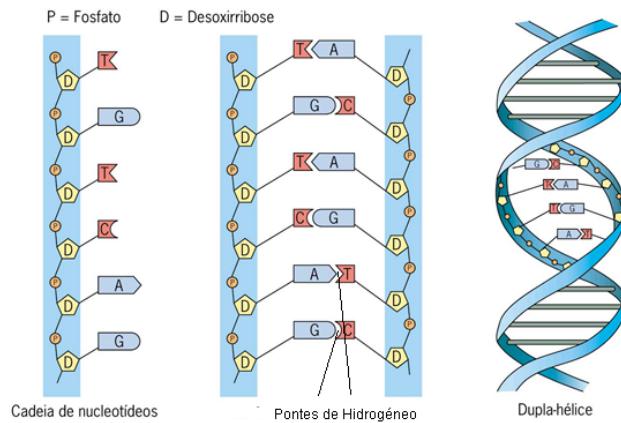


Figura 1.3: Estrutura do DNA
(fonte: <http://www.accessexcellence.org/> em 03/02/2008).

		2 ^a base			
		U	C	A	G
1 ^a base	U	UUU (Phe/F) Fenilalanina UUC (Phe/F) Fenilalanina UUA (Leu/L) Leucina UUG (Leu/L) Leucina, Start	UCU (Ser/S) Serina UCC (Ser/S) Serina UCA (Ser/S) Serina UCG (Ser/S) Serina	UAU (Tyr/Y) Tirosina UAC (Tyr/Y) Tirosina UAA "Ocre" (Stop) UAG "Âmbar" (Stop)	UGU (Cys/C) Cisteína UGC (Cys/C) Cisteína UGA "Opala" (Stop) UGG (Trp/W) Triptofano
	C	CUU (Leu/L) Leucina CUC (Leu/L) Leucina CUA (Leu/L) Leucina CUG (Leu/L) Leucina, Start	CCU (Pro/P) Prolina CCC (Pro/P) Prolina CCA (Pro/P) Prolina CCG (Pro/P) Prolina	CAU (His/H) Histidina CAC (His/H) Histidina CAA (Gln/Q) Glutamina CAG (Gln/Q) Glutamina	CGU (Arg/R) Arginina CGC (Arg/R) Arginina CGA (Arg/R) Arginina CGG (Arg/R) Arginina
	A	AUU (Ile/I) Isoleucina, Start AUC (Ile/I) Isoleucina AUA (Ile/I) Isoleucina AUG (Met/M) Metionina, Start	ACU (Thr/T) Treonina ACC (Thr/T) Treonina ACA (Thr/T) Treonina ACG (Thr/T) Treonina	AAU (Asn/N) Asparagina AAC (Asn/N) Asparagina AAA (Lys/K) Lisina AAG (Lys/K) Lisina	AGU (Ser/S) Serina AGC (Ser/S) Serina AGA (Arg/R) Arginina AGG (Arg/R) Arginina
	G	GUU (Val/V) Valina GUC (Val/V) Valina GUA (Val/V) Valina GUG (Val/V) Valina, Start	GCU (Ala/A) Alanina GCC (Ala/A) Alanina GCA (Ala/A) Alanina GCG (Ala/A) Alanina	GAU (Asp/D) Ácido aspártico GAC (Asp/D) Ácido aspártico GAA (Glu/E) Ácido glutâmico GAG (Glu/E) Ácido glutâmico	GGU (Gly/G) Glicina GGC (Gly/G) Glicina GGA (Gly/G) Glicina GGG (Gly/G) Glicina

Figura 1.4: Relação entre os codões (no RNA) e os aminoácidos codificados (fonte: www.wikipedia.org em 6/02/2008).

fases, transcrição da informação genética, migração dessa informação do núcleo para o citoplasma e tradução da mensagem em proteínas. Para que estas fases ocorram é necessário a participação de uma molécula designada



Figura 1.5: Fluxo de informação genética numa célula (“dogma central” da biologia molecular). Esquema baseado no modelo desenvolvido por Watson e Crick (1970).

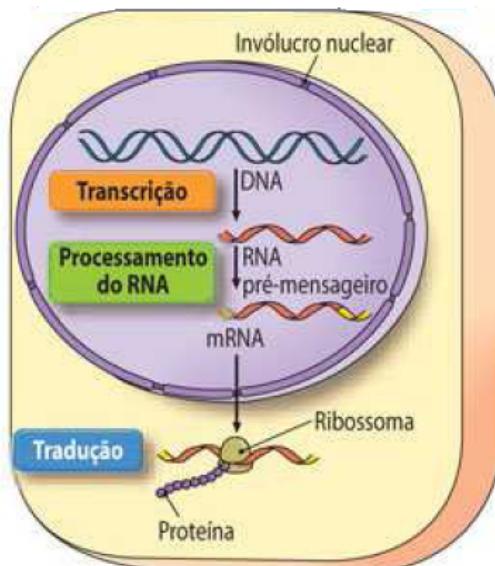


Figura 1.6: Transformação da linguagem de DNA na linguagem proteica (fonte: <http://biogeorecursos.files.wordpress.com> em 08/02/2008).

por RNA (Figura 1.7), que pode ter vários produtos, por exemplo: mRNA, tRNA⁶ e rRNA⁷. O RNA existe nas células em quantidades superiores às do DNA. Enquanto a quantidade de DNA é igual em todas as células, a quantidade de RNA celular é variável e relaciona-se com a maior ou menor actividade metabólica da célula. O RNA é formado por uma cadeia simples de nucleótidos, e não de dupla hélice como o DNA. Um filamento de RNA pode-se dobrar de tal modo que parte das suas próprias bases emparelham-se umas com as outras. Tal emparelhamento intramolecular de

⁶Do inglês *transfer RNA*.

⁷Do inglês *ribosomal RNA*.

1.2. O dogma central da Biologia Molecular

bases é um determinante importante da forma do RNA. Assim, formando pontes intracadeia o RNA é capaz de admitir uma variedade muito maior de formas moleculares do que a dupla hélice de DNA. Os nucleótidos das moléculas de RNA são constituídos pelas bases A, G e C e pela base - Uracilo (U) ($A \rightarrow U$ e $C \rightarrow G$). São cadeias mais simples e de dimensões muito inferiores às do DNA. Outra diferença muito importante é que enquanto o DNA encontra-se apenas no núcleo da célula o RNA também se encontra no citoplasma sob a forma de mRNA, tRNA e rRNA.

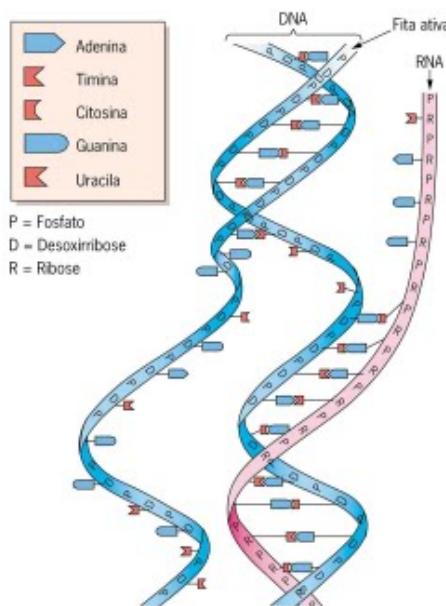


Figura 1.7: Molécula de RNA
(fonte: <http://biogeorecursos.files.wordpress.com> em 08/04/2008).

A *transcrição* realiza-se no núcleo, e corresponde à sintetização de RNA a partir de uma cadeia de DNA que contém a informação e que lhe serve de molde. Esta síntese faz-se na presença de uma enzima que se chama de RNA-polimerase. Esta enzima fixa-se sobre uma certa sequência de DNA, desliza ao longo dela provocando a sua abertura, iniciando-se a transcrição da informação. Ao dar-se a transcrição só uma das cadeias de DNA é utilizada como molde, e após a passagem da RNA-polimerase, a molécula de DNA reconstitui-se pelo restabelecimento das pontes de hidrogénio. Os produtos primários da transcrição genética sofrem ainda transformações profundas no núcleo, que variam com o tipo de RNA. Antes da célula de

RNA sair do núcleo, um dos processos a que se dá o nome de *splicing*, é elaborado por enzimas na remoção de sequências não codificantes.

Um gene corresponde a uma região do DNA constituída por sequências não codificáveis, os intrões, intercaladas entre sequências codificáveis, os exões (Figura 1.8).

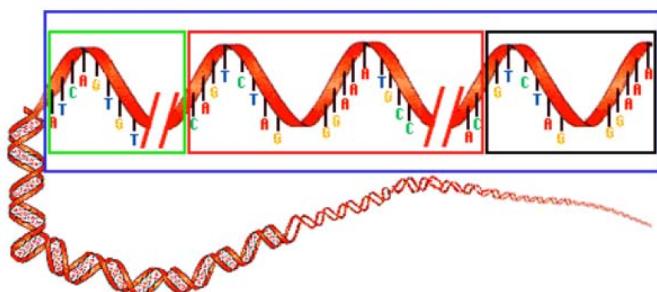


Figura 1.8: Constituição de um gene. Vermelho: região que codifica a seqüência das proteínas; preto: região não-codificante (um único gene normalmente contém mais do que uma); verde: seqüência reguladora (fonte: <http://morgan.rutgers.edu/MorganWebFrames/Level1/Page1/p1.html> em 03/02/2008).

A transcrição de um segmento de DNA forma um RNA pré-mensageiro. No processamento deste RNA, por ação de enzimas, são retirados os intrões, havendo posteriormente a união dos exões (Figura 1.9). Estas transformações conduzem à formação do RNA mensageiro (mRNA) maduro. A migração da informação genética consiste na exportação do mRNA funcional (ou maduro) do núcleo para o citoplasma.

A *tradução* corresponde à transformação da mensagem contida no mRNA em seqüências de aminoácidos. Esta fase comporta três etapas sucessivas: iniciação, alongamento e finalização (Alberts *et al.*, 1994) e encontram-se vários intervenientes: mRNA — contém a informação genética; ribossomos — que se podem traduzir por sistemas de leitura e cuja constituição é composta por proteínas e rRNA; aminoácidos e tRNA — pequenas moléculas de RNA que transportam os aminoácidos para junto dos ribossomos. No processo de iniciação verifica-se a ligação do mRNA e do tRNA. Na fase de alongamento verifica-se a tradução dos codões sucessivos e da ligação entre aminoácidos. Este processo repete-se ao longo do mRNA. Quando o ribossoma chega a

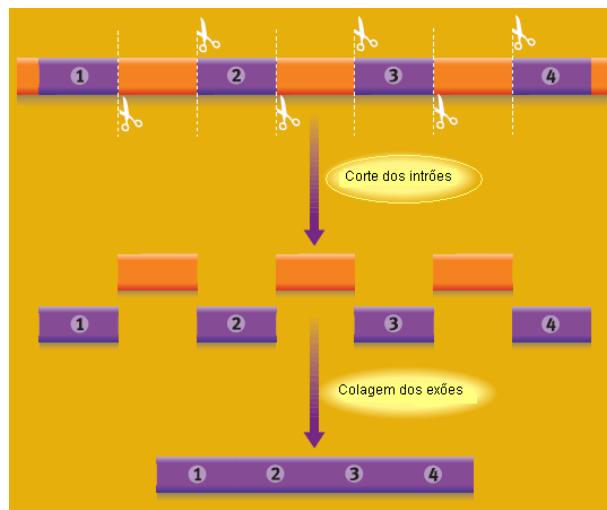


Figura 1.9: *Splicing*. Fatores específicos (representados por tesouras) reconhecem os sítios de corte do DNA, eliminando os intrôns (a laranja) e colando os exônes (a roxo) adjacentes, formando o mRNA (fonte: Penalva e Zario, 2001).

um codão de finalização, separa-se e fica livre para iniciar outro processo; esta fase denomina-se *finalização*.

Muitas das proteínas acabadas de sintetizar não possuem atividade biológica, sofrendo muitas alterações até atingirem a conformação definitiva. Estas moléculas condicionam todo o metabolismo celular. Umas podem ser funcionais nas células, como por exemplo as enzimas, outras são integradas em estruturas celulares como por exemplo no núcleo e outras ainda são exportadas para o meio extracelular, como por exemplo as hormonas proteicas.

1.3 Biotecnologia

A possibilidade de identificar a assinatura de cada tipo específico de patologia pela análise da expressão genética é atualmente uma realidade. A base da análise da expressão genética reside em comparar os níveis de expressão para um determinado conjunto de genes sob duas ou mais condições experimentais. A experiência começa pela obtenção de amostras de tecidos da região a ser avaliada, seguida da extração do mRNA, uma vez que o DNA é o mesmo em todas as células ao contrário do mRNA. A

quantidade de mRNA e as proteínas para as quais o mRNA é traduzido varia de célula para célula. Por exemplo, se se considerar dois tipos de células, A e B, e supondo que na célula A os genes 1 e 2 são transcritos em mRNA e depois traduzidos em proteínas. Na célula B, apenas o gene 2 é transcrito em mRNA e traduzido em proteínas mas num rácio superior ao da célula A. Pode dizer-se que o gene 1 apenas se expressa na célula A e o gene 2 expressa-se nas duas mas com níveis superiores na célula B. O estudo dos genes que têm expressão e os que não têm, em diferentes tipos de células ou sob diferentes condições experimentais e ambientais, permite a aprendizagem de como estes genes afetam o funcionamento das células.

A área da genética que procede à caracterização dos genomas baseia-se na deposição e imobilização de fragmentos de DNA em locais (*spots*), com coordenadas pré-definidas numa superfície sólida. Estas, poderão ser de nitrocelulose ou nylon, no caso dos *macroarrays*, ou de vidro com a dimensão de uma lâmina de microscópio, no caso dos *microarrays*, também designados por *chips* de DNA (Santos *et al.*, 2001). As moléculas de DNA, amplificadas a partir de bibliotecas genómicas, são aplicadas na respectiva superfície sólida através de instrumentação robotizada, *arrayers* de DNA. O material genético que é utilizado nestes suportes pode ser obtido a partir de vários processos laboratoriais, como por exemplo os que a seguir se descrevem.

1.3.1 Sequenciação de DNA

A sequenciação de DNA consiste em determinar a sequência exata das unidades estruturais que compõem o DNA - os nucleótidos. Uma molécula de DNA pode ser sequenciada pela geração de fragmentos através da interrupção controlada da replicação. A tecnologia mais avançada pode sequenciar simultaneamente 96 moléculas de DNA independentes em poucas horas. A tecnologia de sequenciação de DNA tem progredido rapidamente e atualmente vários genomas de microorganismos estão completamente sequenciados e muitos outros encontram-se em fase de sequenciação.

1.3.2 Reação de Polimerização em Cadeia

Podemos imaginar a dupla hélice de DNA como um fecho e com a presença de enzimas específicas estas separam-se por rotura das pontes de hidrogénio, dando origem a duas cadeias complementares, onde apenas uma servirá de molde à formação de duas moléculas de DNA idênticas à molécula original.

Estas novas moléculas resultam da condensação ordenada de nucleótidos que se encontram livres no núcleo da célula, processando-se no sentido 5' → 3'. Este processo é designado por *replicação*.

A replicação pode ser feita laboratorialmente através por exemplo da técnica de PCR⁸. A PCR é uma metodologia laboratorial que tem por objetivo produzir uma elevada quantidade de um segmento específico de DNA a partir de uma quantidade mínima (Griffiths *et al.*, 1999). De entre as várias aplicações salienta-se a clonagem de genes. O processo PCR consiste numa série de ciclos até à obtenção da quantidade desejada de clones. Um ciclo de PCR (Figura 1.10) consiste em três fases: (i) desnaturação da dupla hélice de DNA molde; (2) ligação dos *primers* (segmentos de RNA, com cerca de 1 a 60 nucleótidos complementares do DNA) às sequências dos pares de bases complementares da cadeia molde *in vitro*; (iii) procede-se à extensão. Normalmente são repetidos cerca de 20 a 30 ciclos, que demoram apenas algumas horas. Assim, duas novas cadeias são sintetizadas a partir da cadeia molde em cada ciclo completo de PCR, verificando-se um crescimento exponencial, pois ao fim de n ciclos há cerca de 2^n mais cópias do que no início.

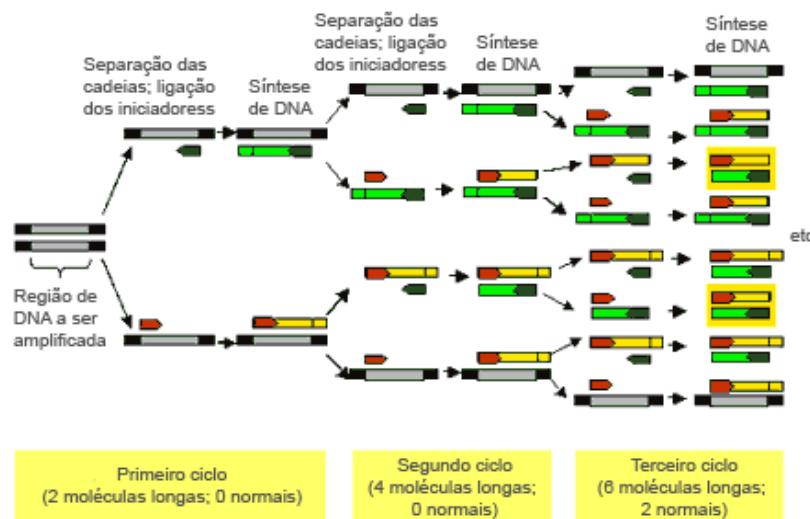


Figura 1.10: Ciclo do PCR (fonte: <http://www.e-escola.pt/> em 10/02/2008).

⁸Do inglês *polymerase chain reaction*.

1.3.3 Transcrição Reversa da Reação de Polimerização em Cadeia

A transcrição reversa (reação RT⁹- PCR) é um processo onde uma cadeia de RNA é reversamente transcrita em cDNA¹⁰. A técnica de RT-PCR combina a síntese de cDNA com a amplificação *in vitro* por PCR, possibilitando a obtenção de fragmentos de DNA, a partir dos respectivos mRNAs, e a análise de expressão genética. É uma técnica muito utilizada na deteção de vírus de RNA em amostras clínicas bem como no estudo de expressão genética em células provenientes de biopsias, uma vez que constituem amostras com pouco material biológico.

Dado o caráter exponencial da reação de PCR, a sensibilidade desta técnica é consideravelmente maior do que em outras técnicas de análise de expressão genética, requerendo quantidades menores de RNA para análise. Por outro lado, a contaminação do RNA por quantidades ínfimas de DNA genómico leva a falsos positivos ou à impossibilidade de distinção da amplificação obtida a partir de cDNA ou de DNA genómico. Adicionalmente, as duas reações que constituem o RT-PCR, são reações enzimáticas que dependem do emparelhamento de *primers*, sendo a sua eficiência fortemente dependente de condições ambientais como a temperatura, pH, força iônica e contaminação do próprio manipulador. Por isso, deve fazer-se um cuidadoso planeamento das reações, que deve passar pela seleção de um método apropriado de extração e purificação de RNA, escolha correta das enzimas a utilizar, *design* dos *primers*¹¹, temperatura de atuação da transcriptase reversa e programa de PCR (Griffiths *et al.*, 1999).

1.3.4 Microarrays

A utilização dos *arrays* de DNA em análises comparativas da expressão genética está definitivamente implantada nas mais diversas áreas de investigação e diagnóstico. Graças a esta tecnologia têm-se conseguido avanços importantes na identificação de genes específicos para determinados tecidos, com repercussões diretas no melhor entendimento dos processos que levam à diferenciação celular (Reinke *et al.*, 2000). Esta tecnologia tem contribuído também para a obtenção de alguns êxitos recentes na identificação de genes envolvidos na resposta a condições de *stress*, como

⁹Do inglês *reverse transcription*.

¹⁰Do inglês *complementary DNA*.

¹¹Curtas sequências sintéticas de nucleótidos, entre 20 e 30 bases.

sejam a presença de toxinas no desenvolvimento do ciclo celular (Nuwaysir *et al.*, 1999), ou fenómenos associados à replicação do DNA (Marton *et al.*, 1998). Finalmente, a área da saúde será porventura aquela onde a introdução dos *arrays* de DNA estará a causar maior impacto, nomeadamente, no desenvolvimento de novos fármacos (Heller *et al.*, 1999), na busca de polimorfismos associados a condições clínicas, por exemplo nos diabetes e em algumas condições cardíacas (Wang *et al.*, 1998), e na tipagem de alguns tipos de cancro (Scherf *et al.*, 2000; Ross *et al.*, 2000).

Os primeiros *arrays* surgiram na década de 80 e eram designados por *macroarrays*. Os *microarrays* de DNA foram inicialmente desenvolvidos por Schena *et al.* (1995) e têm sido largamente utilizados em várias áreas como fisiologia celular, farmacologia, fenotipagem molecular, sequenciação de DNA, etc.

A base fundamental dos *microarrays* é o processo de hibridação. Este princípio é explorado nesta técnica, no sentido em que se mede a quantidade de DNA (ou RNA) desconhecido (*target* ou alvo) com base numa sequência complementar conhecida (*probe* ou sonda). O nível de hibridação é usualmente quantificado a partir de uma etiqueta química usada para marcar o *target* numa determinada experiência. O *array* é lido num *scanner* e a imagem resultante é traduzida em intensidades. O objetivo é selecionar genes com diferentes níveis de expressão em duas (ou mais) amostras.

De entre os vários tipos de *microarrays* comercializados (Tan *et al.*, 2003), existem dois mais conhecidos, designados por *microarrays* de dois canais (*microarrays* de cDNA) e de um canal (*microarrays* de oligonucleótidos). Os *microarrays* de dois canais (Figura 1.11) consistem em milhares de sequências individuais de DNA impressas numa lâmina de vidro através de um robô. Duas amostras de mRNA (ou *targets*), correspondentes aos dois tipos de tecido que se pretendem comparar, por exemplo células com algum tipo de patologia (condição 1) e células sem essa patologia (condição 2), são reversamente transcritas em cDNA e etiquetadas com fluorocromos¹² diferentes (vermelho — Cy5 e verde — Cy3). Depois são misturadas em proporções iguais e hibridadas às sondas que se encontram no *array*. As duas amostras são combinadas num único *array*. Seguidamente os *arrays* são transformados em imagens através de um *scanner* e as medições das fluorescências são feitas separadamente para cada cor em cada *spot* do *array*. O rácio entre as fluorescências vermelho e verde de cada *spot* é um indicador

¹²Molécula que emite fluorescência quando excitada com um comprimento de onda adequado.

da abundância relativa de ácido nucleico das correspondentes sondas nas duas amostras (Dudoit *et al.*, 2002). Se a quantidade de mRNA na amostra da condição 1 for abundante o *spot* (ou poço) será vermelho, se a quantidade for abundante na amostra da condição 2 o *spot* será verde, se a quantidade for igual nas duas amostras o *spot* será amarelo e se não estiver presente será preto (Figura 1.11).

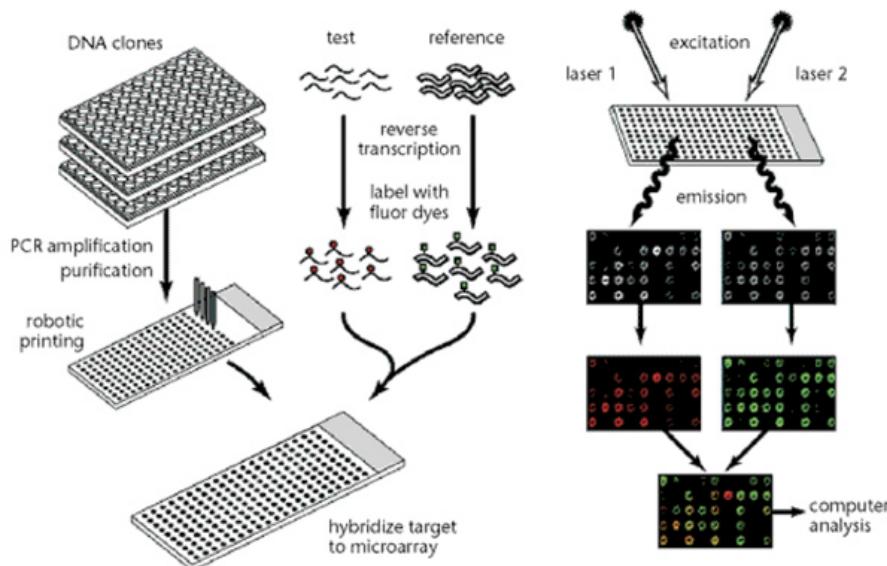


Figura 1.11: Passos na preparação de um *microarray* de dois canais (fonte: Goor, 2005).

Nos *microarrays* de um canal, mais especificamente da Affymetrix (Affymetrix, 1999), centenas de milhares de sondas oligonucleotídicas diferentes são sintetizadas em cada *array*. Cada oligonucleótido está localizado numa área específica do *array* denominada por *spot*, cada *spot* contém milhões de cópias de um determinado oligonucleótido (Figura 1.12). Na secção 1.3.4 será explicado em maior detalhe este tipo de *microarray*.

Atualmente, tem havido um grande esforço na padronização dos dados gerados a partir deste tipo de experiências, quer pela importância quer pela possibilidade de comparação entre diferentes experiências. Um exemplo desse esforço é o protocolo conhecido como MIAME¹³ (Brazma *et al.*, 2001),

¹³Do inglês *Minimum Information About a Microarray Experiment*.

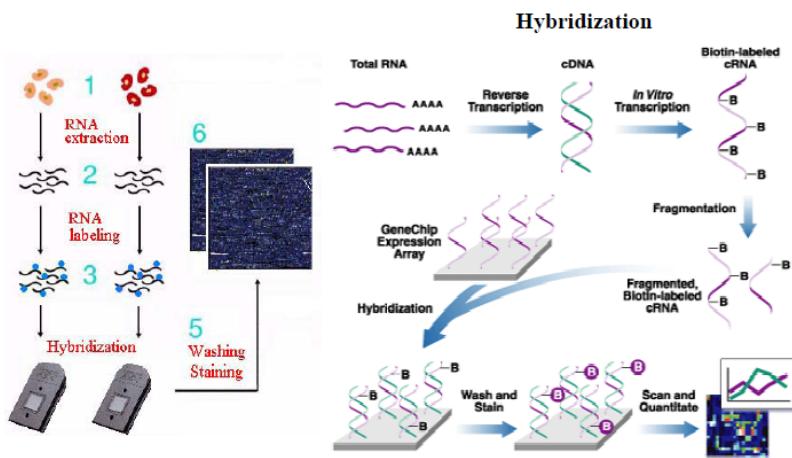


Figura 1.12: Passos na preparação de um *microarray* de um canal (fonte: Affymetrix, 1999).

outro exemplo é o BASE¹⁴ (Saal *et al.*, 2002), o qual foi desenhado para armazenar informações como: imagens, dados em bruto, descrições quer dos genes quer das amostras, normalizações e outros dados relevantes para a análise da expressão diferencial. Exemplos de outras bases de dados como o MAGE¹⁵ desenvolvido por Spellman *et al.* (2002), SMD¹⁶ (Sherlock *et al.*, 2001), OMIM¹⁷ *Gene Ontology*.

O projeto *BioConductor* (Gentleman *et al.*, 2004) é de livre acesso baseado em linguagem de programação R, e contém várias bibliotecas para análise de dados de *microarrays* e vários bancos de dados provenientes de *microarrays* de várias plataformas. Este é um projeto em permanente desenvolvimento, onde estatísticos e bioinformáticos contribuem com bibliotecas para a análise de dados genómicos. O projeto teve início em 2001 e é líder na análise de dados de *microarrays*.

Para além dos *microarrays*, existem técnicas alternativas para medir a expressão genética. Enquanto a tecnologia de *microarrays* baseia-se na hibridação, existem outras que se baseiam na sequenciação e fragmentação.

¹⁴Do inglês *BioArray Software Environment*.

¹⁵Do inglês *MicroArray Gene Expression*.

¹⁶Do inglês *Stanford Microarray Database*.

¹⁷Do inglês *Online Mendelian Inheritance in Man*.

Exemplos de técnicas baseadas em sequenciação são EST¹⁸, SAGE¹⁹, WTSS²⁰ e MPSS²¹ e baseadas em fragmentação são por exemplo o cDNA-AFLP²².

Os dados analisados neste trabalho provêm essencialmente de *microarrays* de um canal da Affymetrix e na aplicação e desenvolvimento de metodologias estatísticas em linguagem R.

Microarrays de um canal - Affymetrix

Os *arrays* de sondas oligonucleotídicas produzidas pela Affymetrix são conhecidos por Affymetrix GeneChip (Lockhart *et al.*, 1996). Neste trabalho os *GeneChips* são referidos por *arrays* ou *chips*.

Os *arrays* da Affymetrix são divididos em milhares de poços. Cada poço contém sondas, que por sua vez são oligonucleótidos constituídos por 25 pb, cujas sequências são conhecidas. As sondas são escolhidas de tal modo que cada uma é complementar do RNA alvo que se pretende quantificar. O RNA alvo é marcado com uma etiqueta química (biotina) cuja sequência complementar hibridará com a sonda. Os alvos que não hibridarem serão removidos. Cada gene é representado por dois conjuntos de entre 11 a 20 sondas oligonucleotídicas e cada sonda corresponderá a um fragmento do gene (Figura 1.13).

Ao conjunto de sondas que representa um gene dá-se o nome de *probeset* e podem existir entre 12.000 a 22.000 *probesets* no *array* (Figura 1.14). Por vezes existe mais do que um *probeset* para o mesmo gene, mas cada um usa diferentes partes da sequência.

Cada sonda é constituída por um par de sequências, designadas por *Perfect Match* (PM) e *Mismatch* (MM). As sondas PM têm a sequência de bases idêntica à do alvo complementar e as sondas MM contêm uma mutação, correspondente à base no meio da sequência (13^a posição) (Figura 1.14). A sequência MM tem por objetivo quantificar hibridações não específicas. O par PM e MM é designado por *probe-pair*.

As sondas PM e correspondente MM são sempre colocadas aos pares num determinado poço do *array*, no entanto os pares de um *probeset*

¹⁸Do inglês *Expressed Sequence TAG*.

¹⁹Do inglês *Serial Analysis of Gene Expression*.

²⁰Do inglês *Whole Transcriptome Shotgun Sequencing*

²¹Do inglês *Massively Parallel Signature Sequencing*.

²²Do inglês *Amplified Fragment Length Polymorphism on a cDNA template*.

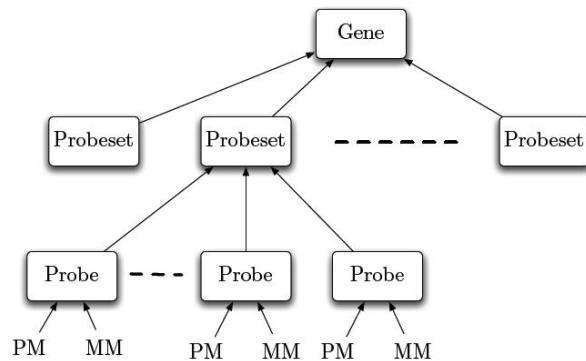


Figura 1.13: Representação esquemática de um gene.

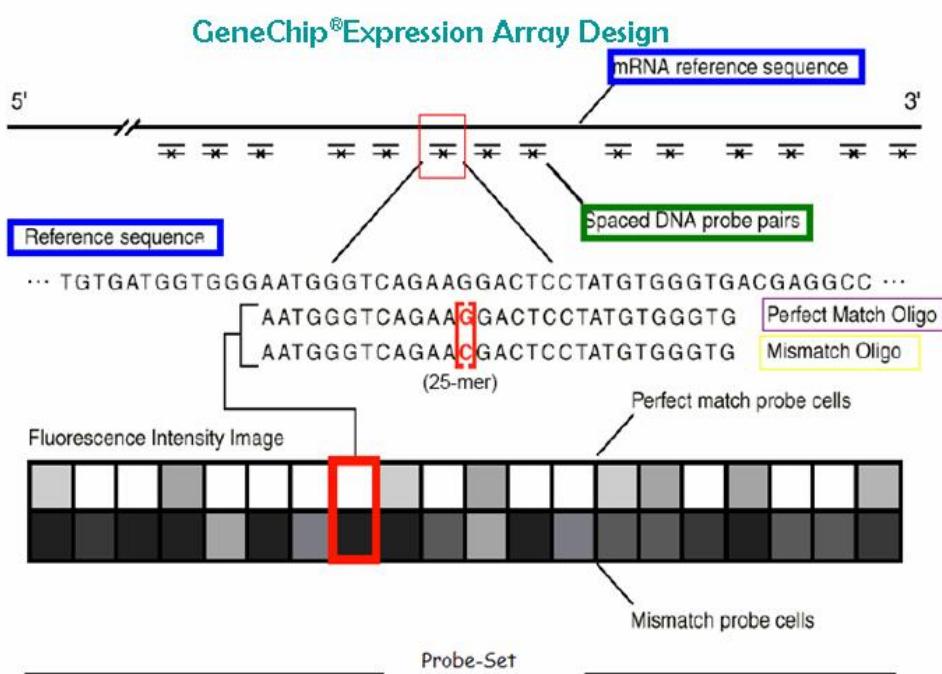


Figura 1.14: Representação de um *probeset* (fonte: Affymetrix, 1999).

s o distribu dos aleatoriamente no *array*. A rela o entre a intensidade do sinal das sondas PM e MM indica se um gene est  ou n o ativo na c ula ou tecido numa determinada situa o experimental. Este sinal

tende a ser proporcional à quantidade de RNA na amostra, até uma certa concentração de transcritos. Os diferentes *probes* de um mesmo *probeset* podem ter sinais muito diferentes entre si, no entanto quando comparados com vários *arrays* terão o mesmo comportamento. Estas diferenças de sinal entre os diferentes *probes* devem-se a vários fatores, nomeadamente na localização do *probe* no RNA, erros durante a construção do *array*, temperatura a que se dá a experiência, etc. Nos formatos mais recentes de *arrays* da Affymetrix, as sondas *mismatch* não são utilizadas (Dziuda, 2010).

Os *arrays* da Affymetrix podem ter entre 500.000 (HGU95Av2) a 1.300.000 (HGU133 plus 2.0) sondas. O processo de construção de um *array* é constituído por várias fases: primeiro isola-se o RNA total do tecido em estudo, seguidamente o RNA é reversamente transcrito em cDNA, etiquetado com biotina e hibridado no *array*. Um determinado número de controlos são também produzidos e hibridados nos *arrays*. Depois do processo de hibridação estar completo, são removidas as hibridações não específicas. Após lavagem dos *chips* são emitidos *lasers* com o objetivo de excitar as etiquetas químicas e sinais luminosos são emitidos. A quantidade de sinal emitido pelo *chip* é armazenado numa imagem (Figura 1.15), em ficheiros com extensão .DAT. A quantidade de luz emitida é proporcional à quantidade de moléculas-alvo ligadas a cada local.

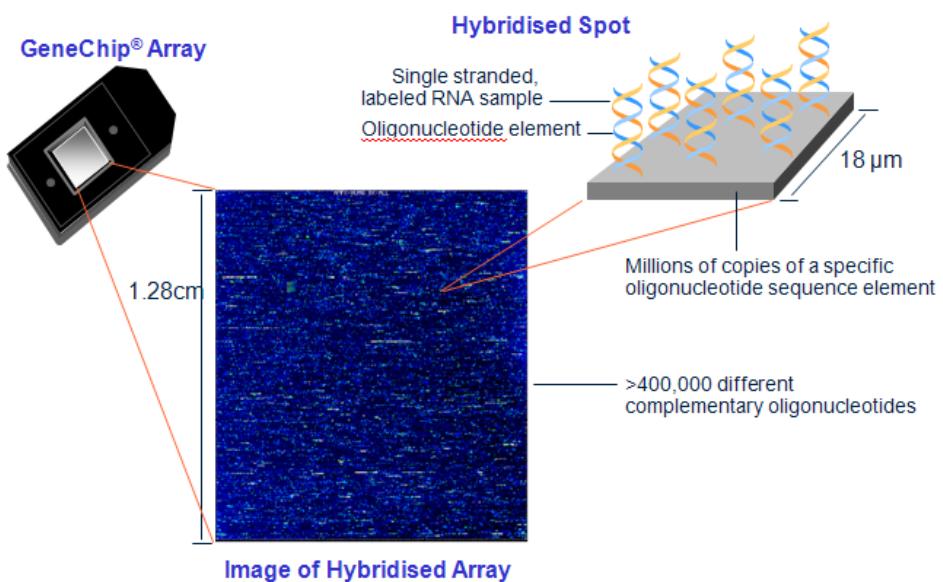


Figura 1.15: Do *chip* à imagem (fonte: Affymetrix, 1999).

A imagem é transformada em valores que representam as intensidades das sondas. Estas intensidades são obtidas através do seguinte processo: uma grelha quadriculada é colocada sobre a imagem e cada sonda é representada pelo quantil de ordem 75 dos píxeis correspondentes a cada quadrícula (Figura 1.16). Os valores das intensidades são registados em ficheiros com extensão .CEL, onde os genes são identificados através do código `número_at`. Nestes ficheiros outros códigos são utilizados, para além de `número_at` que reconhece múltiplos transcritos alternativos para o mesmo gene, um deles é o código `número_s`, que identifica genes diferentes que possuem sondas comuns, e o outro `número_x`, que identifica se o gene contém algumas sondas que são idênticas a outras sequências. Podem considerar-se quatro passos para a obtenção dos níveis de expressão (Stekel, 2003): a) identificação das posições dos transcritos no *microarray*, b) para cada transcrito, identificar os píxeis na imagem que fazem parte do transcrito, c) para cada transcrito identificar píxeis vizinhos que possam ser utilizados para o cálculo do *background* e d) calcular numericamente a informação da intensidade do transcrito, do *background* e do controlo da qualidade da imagem.

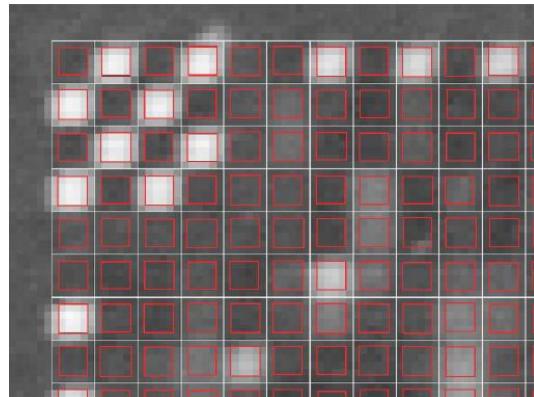


Figura 1.16: Parte da imagem de um *chip* com grelha sobreposta.

1.4 Algumas considerações finais

Este capítulo centrou-se na descrição dos processos biológicos e tecnológicos inerentes à produção dos dados em análise neste trabalho. Em particular descreveu-se os processos na célula, a estrutura base de todos os organismos vivos, assim como as suas principais partes integrantes. Foi descrita a unidade central de armazenamento da informação genética, o DNA, e a

forma como este é transcrito em mRNA, passa do núcleo para o citoplasma e é traduzido em proteínas. A compreensão do funcionamento destes processos tem sido possível graças ao desenvolvimento de novas ferramentas biomoleculares das quais se destacam os *microarrays*. Os *microarrays* são aqui apresentados como uma solução para a análise em larga escala dos níveis de expressão dos genes da célula. São vários os modos de operação dos *microarrays*, sendo os principais a monitorização da expressão genética e a deteção de mutações e polimorfismos. Sendo uma tecnologia relativamente recente, existem ainda bastantes condicionantes ao seu uso, nomeadamente na validação da qualidade dos resultados. Nos últimos anos, tem sido realizado um esforço coletivo para endereçar as questões existentes e para promover e generalizar o uso desta tecnologia.

Capítulo 2

Análise de dados de *microarrays* de um canal

2.1 Introdução

A produção de proteínas a partir de um gene é chamada de expressão do gene. A análise de dados de expressão genética é bastante útil e tem diversas aplicações na Biologia e na Medicina. Por exemplo, este tipo de análise pode fornecer informação importante acerca das funções de uma célula, uma vez que as mudanças na fisiologia de um organismo (fenótipo) são acompanhadas por mudanças nos padrões da expressão dos genes (genótipo) (Chan *et al.*, 2000). Existem duas abordagens para a avaliação da expressão genética: a análise do transcriptoma e a análise do proteoma. A primeira baseia-se na análise dos produtos da transcrição, mRNA, intermediários no processo de produção de uma proteína. A segunda, refere-se à análise direta das proteínas produzidas no processo de expressão. Uma proteína é produzida a partir de um mRNA, porém, a quantidade de mRNA presente numa célula, apesar de relacionada, nem sempre corresponde à quantidade de proteína produzida. Apesar de analisar-se diretamente o produto final da expressão de um gene, que são as proteínas, a análise do proteoma é muito mais complexa do que a análise do transcriptoma. Com os avanços da tecnologia relacionada com a análise das proteínas, as análises baseadas no proteoma têm proliferado rapidamente. Apesar disso, a análise da expressão genética tem sido feita, na sua grande maioria, pela análise do transcriptoma.

A mensuração da expressão genética a partir dos transcritos (mRNAs) de um organismo pode ser feita por meio de diversas técnicas que geram dados em larga escala, como por exemplo SAGE (*Serial Analysis of Gene Expression*) (Velculescu *et al.*, 1995), MPSS (*Massively Parallel Signature Sequencing*

Technology) (Brenner *et al.*, 2000), vários tipos de *microarrays* de DNA e RT-PCR. Stanton (2001) faz uma breve revisão de algumas destas técnicas.

Os cientistas que utilizam *microarrays* têm-se deparado com um desafio sem precedentes na análise de dados, muitos consideram extraordinária a capacidade de analisar quantidades astronómicas de dados (Leung *et al.*, 2001). O elevado volume de informação conduziu a um esforço interdisciplinar, e a uma necessidade de desenvolver novas ferramentas analíticas e computacionais. Várias metodologias e técnicas têm surgido. Porém, diferentes métodos para o mesmo objetivo têm conduzido a resultados diferentes, o que torna difícil a tarefa de eleger o método a aplicar. A dificuldade da escolha das técnicas deve-se ao elevado número de fatores que condicionam o tipo de dados, nomeadamente o tipo de *microarrays*, condições experimentais, desenho experimental, objetivo do estudo, etc, o que significa que muitas das situações são novas e a necessidade de criar metodologias que se adequem às exigências está sempre a verificar-se.

Assim, como o desenho experimental influencia a formulação do modelo subjacente à análise dos dados, a interpretação é algo que depende do número de réplicas¹ associadas à experiência. A redução do erro experimental requer que um número suficiente de réplicas seja utilizado (Lee *et al.*, 2000) e o não cumprimento deste princípio reduz o poder estatístico para detetar níveis de expressão diferencial.

As réplicas podem ser técnicas ou biológicas. Réplicas técnicas referem-se à replicação de hibridações da mesma amostra de mRNA, ou seja, que provêm da mesma fonte biológica não sendo réplicas verdadeiramente independentes umas das outras, porém, são úteis para validarem a exatidão das medições dos transcritos. No entanto, estas réplicas não dão informação acerca da variabilidade na população. Réplicas biológicas, referem-se a hibridações de amostras de mRNA provenientes de fontes biológicas independentes sob as mesmas condições, por exemplo amostras extraídas de diferentes indivíduos que receberam o mesmo tratamento (Ayroles e Gibson, 2006).

Um dos principais objetivos dos *microarrays* é medir a expressão de milhares de genes simultaneamente e identificar mudanças nas expressões entre diferentes estados biológicos. Assim, a partir de um elevado número de genes em análise é possível obter um número substancialmente mais reduzido de genes com a(s) característica(s) de interesse para posterior análise.

¹A cada réplica está associado um *array*.

O objetivo deste capítulo é dar uma ideia geral acerca de algumas das tendências na análise de dados em *microarrays*, em particular métodos para selecionar genes com expressão diferencial² (DE). A escolha dos métodos aqui expostos, justifica-se quer pelo interesse que têm no desenvolvimento das abordagens apresentadas no capítulo 4, quer pela necessidade de os comparar com o novo método que irá ser proposto neste trabalho.

Os métodos e experiências aqui abordados são direcionados a dados provenientes de *microarrays* de um canal da Affymetrix GeneChipTM.

Antes de poder aplicar-se métodos para selecionar genes DE é necessário proceder-se a uma intensa transformação dos dados em bruto e posterior análise. Este capítulo é organizado de modo a estabelecer um fio condutor entre as várias etapas necessárias até se chegar à lista final de genes selecionados como sendo genes DE.

Na primeira secção são abordados vários métodos de exploração e transformação dos dados, que na análise de *microarrays* corresponde à designada análise de baixo nível (*low-level analysis*). Serão descritos métodos de transformação dos dados, imputação de valores omissos, avaliação da qualidade dos dados e pré-processamento, que em *microarrays* da Affymetrix corresponde à correção de *background*, normalização, correção PM e sumariação. Esta análise é a que produz os níveis de expressão utilizados na designada análise de elevado nível (*high-level analysis*), nomeadamente na seleção de genes DE.

Na segunda secção apresentam-se vários métodos para a seleção de genes DE que são mais vulgarmente utilizados.

Os recursos computacionais de acesso livre serão privilegiados, em particular bibliotecas do R e *Bioconductor Project*.

A dificuldade que um estatístico sente quando se depara com a análise de dados de *microarrays*, para além da óbvia necessidade de entender os termos biológicos que descrevem a experiência em causa, é a de compreender como se encontram os dados disponíveis e como acceder à informação. Na Figura 2.1 apresenta-se de uma forma esquematizada alguns dos ficheiros (apenas os necessários para as análises realizadas neste trabalho) que a Affymetrix disponibiliza e a informação que se obtém em cada um deles.

²Genes que se expressam de forma diferenciada entre duas amostras ou mais.

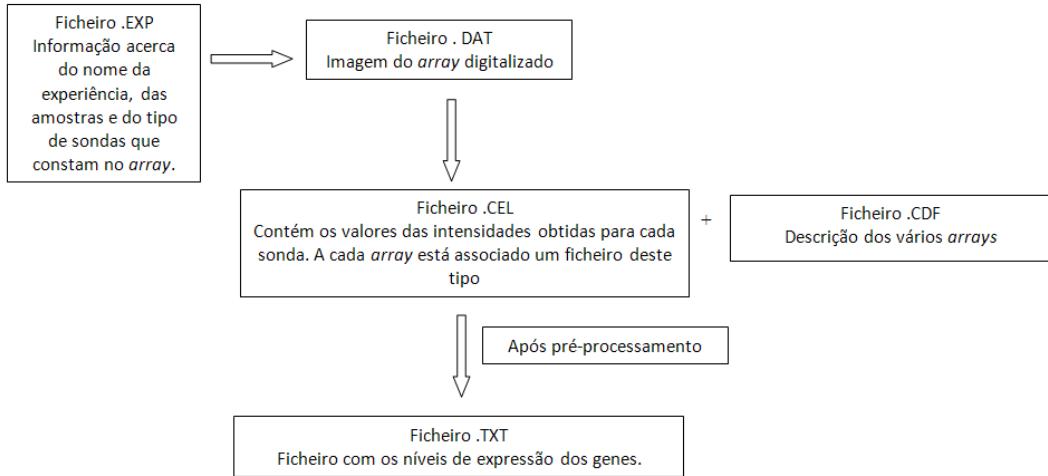


Figura 2.1: Esquema resumo de alguns ficheiros obtidos em *microarrays* da Affymetrix.

É importante relembrar que um gene é representado por um *probeset*, que é constituído por várias sondas (11 a 16) e estas encontram-se espalhadas no *array* (secção 1.3.4). Os dados que serão tratados nesta fase inicial da análise são relativos às sondas PM e MM que constituem os *probesets*. Assim, a notação utilizada é a seguinte:

PM_{ijg} — representa o nível de intensidade da j -ésima sonda PM do gene g no *array* i ;

MM_{ijg} — representa o nível de intensidade da j -ésima sonda MM do gene g no *array* i ;

$i = 1, \dots, n$, em que n representa o número total de *arrays* na experiência;

$j = 1, \dots, J$, em que J representa o número de sondas de um *probeset*;

$g = 1, \dots, G$, em que G representa o número total de genes na experiência.

2.2 Pré-processamento

A matriz de dados que serve de base para a análise de elevado nível é produto de um processo nada trivial e, nesta secção, pretende-se descrever as várias etapas conducentes à referida matriz de dados.

As fases na produção dos *microarrays* são várias³, e vão desde a aquisição do material genético à obtenção dos valores numéricos correspondentes às intensidades luminosas, intensidades essas que representam a expressão dos genes. Todas as fases estão sujeitas a fatores que contribuem para a existência de ruído e variabilidade sistemática (Figura 2.2).

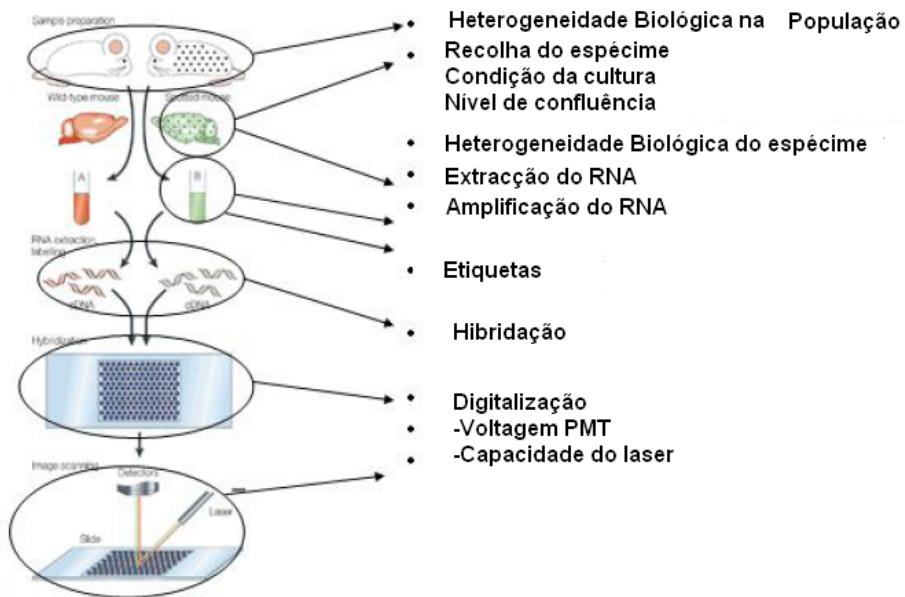


Figura 2.2: Fatores que contribuem para a variabilidade sistemática, influenciando a qualidade final dos dados. Figura adaptada de Geschwing e Greeg (2002).

A presença de valores estranhos é recorrente neste tipo de bases de dados, como por exemplo níveis de intensidade negativos ou de valores omissos e como tal, torna-se necessário o recurso a técnicas de correção.

As próximas secções são dedicadas a técnicas de refinamento da qualidade dos dados.

³Entenda-se produção dos níveis de intensidade e não o suporte propriamente dito.

2.2.1 Filtragem

A filtragem de genes é um processo opcional e consiste em remover sondas que introduzam viés ou acrescentem ruído aos resultados. Este processo pode ser executado tendo por base um limite, e sondas com valores abaixo do limite definido são removidas. Remover sondas com intensidades negativas é outro exemplo de filtragem. Contudo, a remoção conduz à presença de valores omissos cuja análise, se descreve na próxima secção.

2.2.2 Valores omissos

Os *microarrays* geram bases de dados com limitações técnicas, como já anteriormente foram referidas, que conduzem a sondas corrompidas. O *software* que produz as imagens dos *arrays*, marca estas sondas (*flagged spots*) e, caso não se proceda a nenhuma análise, são tratadas como valores omissos. Um dos problemas inerentes à existência de valores omissos na análise de dados de *microarrays*, é o facto de muitos dos métodos propostos para a análise destes dados não permitirem a existência de valores omissos, como por exemplo métodos clássicos de classificação supervisionada.

Na prática, há vários métodos para lidar com o problema dos valores omissos. O método mais simples consiste na eliminação dos genes com valores omissos (Statnikov *et al.*, 2005). No entanto, a sua eliminação pode conduzir à perda de genes importantes no estudo. Por vezes a lista de genes a eliminar pode ser muito grande comprometendo todo o estudo.

Outro método consiste em substituir todos os valores omissos por zero (Alizadeh *et al.*, 2000), o que na prática não resolve o problema na análise de dados de *microarrays*, porque valores dos níveis de intensidade próximos de zero podem ser confundidos com valores omissos.

Vários métodos de imputação têm sido propostos para substituir os valores omissos, desde os métodos estatísticos mais simples, como *Row Mean* (Bo *et al.*, 2004) ou *Row Average* (Kim *et al.*, 2005), ou métodos baseados no algoritmo EM (*Expectation-Maximization*), como *EM_gene* ou *EM_array* (Bo *et al.*, 2004). Celton *et al.* (2010) descrevem vários métodos para a imputação de valores omissos em dados de *microarrays* (Tabela 2.1). Um método bastante popular e pioneiro na análise de dados omissos nos *microarrays* é a abordagem *k*-NN⁴ (*k* vizinhos mais próximos), aplicado por Troyanskaya *et al.* (2001) em *microarrays* de dois canais. Não sendo

⁴Do inglês *k-Nearest Neighbours*.

o objetivo desde trabalho a análise dos vários métodos propostos para imputação de valores omissos, apresenta-se apenas um resumo na Tabela 2.1 e descreve-se sucintamente o método k -NN, uma vez que foi o eleito para a imputação de valores omissos nos dados de Alizadeh *et al.* (2001) (secção 5.3).

Tabela 2.1: Métodos de imputação de valores omissos em *microarrays*.

Tabela adaptada de Celton *et al.* (2010).

Método	Autores	Linguagem
<i>k</i> -NN	Troyanskaya <i>et al.</i> (2001)	C and R
BPCA (<i>Bayesian Principal Component Analysis</i>)	Oba <i>et al.</i> (2003)	JAVA
<i>Row Mean</i>	Bo <i>et al.</i> (2004)	JAVA
<i>EM_gene</i>	Bo <i>et al.</i> (2004)	JAVA
<i>EM_array</i>	Bo <i>et al.</i> (2004)	JAVA
<i>LSI_gene</i>	Bo <i>et al.</i> (2004)	JAVA
<i>LSI_combined</i>	Bo <i>et al.</i> (2004)	JAVA
<i>LSI_adaptative</i>	Bo <i>et al.</i> (2004)	JAVA
SkNN (<i>Sequential KNN</i>)	Kim <i>et al.</i> (2002)	R
LLSI (<i>Local Least Square Impute</i>)	Kim <i>et al.</i> (2005)	MATLAB
<i>Row Average</i>	Kim <i>et al.</i> (2005)	MATLAB
LinImp(<i>Linear Model Based Imputation</i>)	Scheel <i>et al.</i> (2005)	R
FAR (<i>Factor Analysis Regression</i>)	Feten <i>et al.</i> (2005)	-
OLSI (<i>Ordinary Least Square Impute</i>)	Nguyen <i>et al.</i> (2004)	-
SVR (<i>Support Vector Regression</i>)	Wang <i>et al.</i> (2006)	C++
GMC (<i>Gaussian Mixture Clustering</i>)	Ouyang <i>et al.</i>	MATLAB
SVD (<i>Singular Value Decomposition</i>)	Troyanskaya <i>et al.</i> (2001)	C
CMVE (<i>Collateral Missing Value Estimation</i>)	Sehgal <i>et al.</i> (2005)	MATLAB
<i>GO-Based Imputation</i>	Tuikkala <i>et al.</i> (2008)	-
iMISS (<i>Integrative Missing Value Estimation</i>)	Hu <i>et al.</i> (2006)	C++
POCS (<i>Projection Onto Convex Steps</i>)	Gan <i>et al.</i> (2006)	-
<i>Iterative k-NN</i>	Bras <i>et al.</i> (2007)	-

O método k -NN

O método k -NN permite selecionar genes com níveis de expressão semelhantes ao do gene de interesse, *i.e.*, o gene cujo nível de expressão se encontra omissos. Se se considerar um gene A, que apresenta um valor omissos no *array* 1, este método irá encontrar os k genes, que não apresentam valores omissos no *array* 1 e cujos níveis de expressão se encontram mais próximos do gene A nos *arrays* 2 a n (onde n representa o número total de *arrays*). Uma média ponderada dos valores no *array* 1 para os k genes mais próximos de A é usada como estimativa para o valor omissos correspondente ao gene A (2.1). Os pesos dos k genes na média ponderada são calculados em função da semelhança dos níveis de expressão dos genes em relação ao gene A (2.2).

$$\hat{x}_{A1} = \sum_{l=1}^k w_l \times \mathbf{x}_{l1}, \quad (2.1)$$

onde \mathbf{x}_{l1} ($l = 1, \dots, k$) representa o vetor dos níveis de expressão dos k genes mais próximos do gene A, com os pesos dados por:

$$w_l = \frac{\frac{1}{\psi_l}}{\sum_{i=1}^k \frac{1}{\psi_i}}, \quad (2.2)$$

onde ψ_l ($l = 1, \dots, k$) representa a distância euclidiana dos genes de referência ao gene A.

A métrica que Troyanskaya *et al.* (2001) adotam é a distância euclidiana, assumindo que os dados sofreram uma transformação logarítmica, reduzindo assim a presença de *outliers*, uma vez que a distância euclidiana é pouco robusta em relação à presença dos mesmos. A biblioteca `impute` do R permite usar esta metodologia para a imputação de valores omissos em *microarrays*. A biblioteca `arrayImpute`, para além desta metodologia permite aplicar outras, entre as quais algumas das descritas na Tabela 2.1.

2.2.3 Transformação dos dados

É prática comum proceder-se à transformação dos dados provenientes de *microarrays*, com o objetivo principal de uniformizar a variabilidade dos níveis de intensidade e simetrizar as respetivas distribuições.

Algumas técnicas de análise estatística propostas para a análise de dados de *microarrays* dependem da distribuição dos dados, particularmente a hipótese da normalidade (Dopazo *et al.*, 2001). Os níveis de intensidade geralmente têm uma distribuição normal após uma transformação logarítmica de base 2. Este tipo de transformação é muito, senão a mais comum nos dados de *microarrays* (Dopazo *et al.*, 2001; Murphy, 2002; Hariharan, 2003), pois é uma forma de tratar os níveis de expressão dos genes com regulação positiva ou negativa de igual forma já que passam a ter a mesma magnitude mas com sinais opostos. Outra razão (eventualmente) é o facto das variáveis que representam os níveis de intensidade serem positivas e, nesse sentido, tendem a ter distribuições enviesadas simplesmente porque são limitadas inferiormente e não superiormente, e o logaritmo torna a distribuição mais simétrica. Os grupos de genes com médias elevadas dos níveis de expressão, tendencialmente terão variâncias mais elevadas, e a estimativa do erro de medição do nível de expressão cresce com a média; nesse sentido os valores log-transformados terão uma variância aproximadamente constante para todos os genes. Alguns autores (Hariharan, 2003) criticam esta transformação referindo que é muito forte, aconselhando por exemplo outras transformações mais fracas como a raiz cúbica.

Em relação ainda às transformações de escala, há quem defenda que deve fazer-se uma análise em paralelo usando as escalas original e logarítmica, descartando os genes cuja significância desapareça sob uma escala ou outra.

Apesar da transformação dos dados ser apropriada para determinadas situações, deve ser aplicada de forma bastante criteriosa, e somente quando for necessário, pois cada transformação leva à perda de alguma informação contida nos dados (Sarle, 2002).

2.2.4 Avaliação da qualidade dos dados

Cada *microarray* produz níveis de intensidade de milhares de genes simultaneamente, sendo a análise de erros sistemáticos em bases de dados de grandes dimensões muito mais relevante do que a análise de erros aleatórios (Bolstad *et al.*, 2003). Existem muitas fontes de enviesamento neste tipo de experiências, desde a variabilidade biológica inerente às amostras, a temperatura aquando a hibridação, a quantidade de RNA utilizado, a digitalização dos *chips*, etc. Numa primeira abordagem podem analisar-se as imagens digitalizadas dos *chips*, onde é possível identificar artefactos que não sejam consistentes com a imagem do *chip*, como por exemplo riscos, bolhas, etc. Para que seja possível comparar os níveis de intensidade entre os

arrays, as distribuições das intensidades devem de ser semelhantes. Assim, antes de se proceder à seleção de genes, deve ser feita uma análise preliminar aos dados a fim de identificar fontes de enviesamento e, se for necessário, proceder a uma análise de pré-processamento.

Tendo por base a filosofia de George Fuechsel⁵, *garbage in garbage out*, a avaliação da qualidade dos dados é um passo muito importante em qualquer análise, em particular quando a quantidade de dados é tão elevada e os valores dependem de vários fatores extremamente sensíveis às condições da experiência. Assim, antes de se avançar para o pré-processamento propriamente dito, deve proceder-se a uma avaliação da qualidade dos dados. Como não existe consenso na escolha dos vários métodos que integram o pré-processamento, um procedimento, que é muitas vezes utilizado, é a aplicação de vários métodos e mediante a qualidade final dos dados, opta-se pelo método que produz melhores resultados.

A análise da qualidade dos dados em *microarrays* de um canal da Affymetrix pressupõe uma análise específica, não sendo extrapolada para outro tipo de plataforma. O *software* comercial GCOS da Affymetrix apresenta várias ferramentas para analisar a qualidade e no manual (GCOS, 2004; Affymetrix, 2001) são apresentados vários intervalos de referência para determinadas medidas, enfatizando a importância da consistência das medições dentro do conjunto dos *arrays* em análise, usando amostras e condições experimentais semelhantes. O utilizador é encorajado a utilizar várias ferramentas em conjunto.

Estas análises podem ser realizadas com recurso às bibliotecas disponíveis no Bioconductor, `simpleaffy`, `affy`, `affyPLM`, e as bibliotecas `affyQCReport` e `arrayQualityMetrics` produzem relatórios com as principais ferramentas propostas pela Affymetrix.

Box plot

A análise dos *box plots* em paralelo (Figura 2.3) dos vários *arrays* permite verificar se existe algum *array* que se diferencie dos restantes, revelando problemas que lhe possam estar associados. Usualmente o processo de normalização corrige estes problemas pois o objetivo é obter distribuições semelhantes entre os vários *arrays*. Normalmente em dados de *microarrays*

⁵Técnico de informática da IBM New York.

estes gráficos são construídos usando a transformação logarítmica dos níveis de intensidade pois permite uma melhor leitura. Como pode observar-se na Figura 2.3 A, estes gráficos não são legíveis usando os dados em bruto, no entanto, após uma transformação logarítmica (Figura 2.3 B) é possível verificar que por exemplo, o quarto *array* a contar da esquerda se diferencia dos restantes. Resta agora saber se após pré-processamento este efeito desaparece.

As Figuras 2.3 C–F foram obtidas após a aplicação de métodos de pré-processamento e, da sua análise, constata-se que o quarto *array* ainda se diferencia dos restantes, pelo que deve de ser questionada a sua remoção.

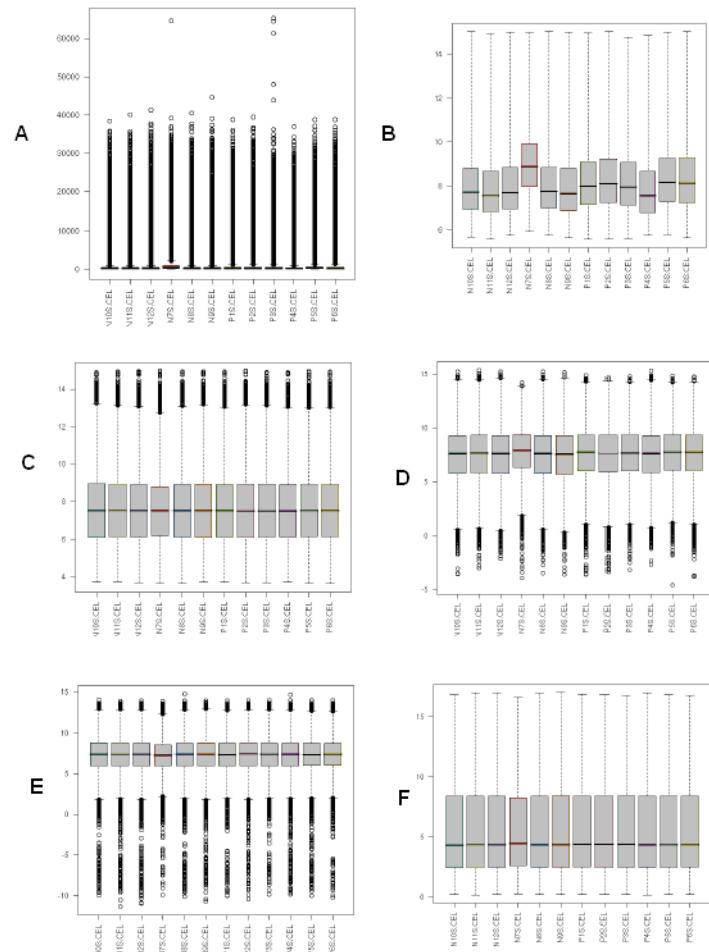


Figura 2.3: (A) — *Box plots* dos níveis de intensidade em bruto; (B) — transformação logarítmica de base 2: (C) — Pré-processamento usando o método RMA; (D) — Pré-processamento MAS5; (E) — Pré-processamento PLIER; (F) — Pré-processamento GCRMA (fonte: <http://arrayanalysis.org/main.html> em 20/05/2011).

Densidade

A análise da representação empírica dos logaritmos dos níveis de intensidade PM dos *arrays* conduz-nos a informações semelhantes às obtidas pela análise dos *box plots*. Isto é, se as distribuições não forem semelhantes, revela a necessidade de se proceder à normalização. Na Figura 2.4 A, é possível, uma vez mais, identificar o mesmo *array* que da análise anterior se tinha diferenciado. Esta situação pode ser um indicador de ruído de *background* e, mesmo após pré-processamento RMA, este *array* ainda se diferencia dos restantes (ver secção 2.2.5). É possível ainda obter outras informações a partir deste gráfico, como por exemplo, se existir uma distribuição bimodal nos dados, é usualmente um indicador de que o *array* tem associado um problema na qualidade; se a distribuição estiver enviesada para a direita é sinónimo de problemas de *background*.

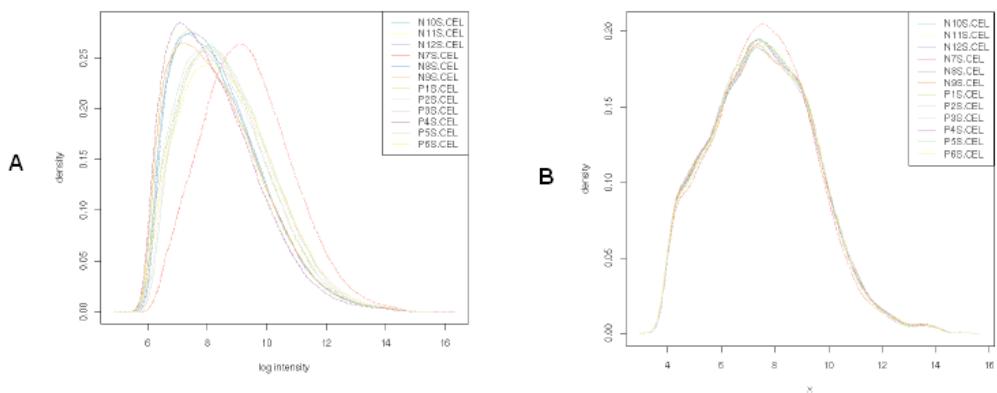


Figura 2.4: (A) — Densidades dos *arrays* apenas com transformação logarítmica dos dados; (B) — Pré-processamento RMA (fonte: <http://arrayanalysis.org/main.html> em 20/05/2011).

Degradation plot

Um gráfico que permite analisar a qualidade do material genético da amostra que hibrida com o *array* é o *degradation plot*. A degradação do mRNA ocorre quando a molécula começa a partir-se e torna-se impossível determinar o nível de expressão do gene. O processo de degradação da molécula tem início na extremidade 5' e avança para a extremidade 3'. Naturalmente as sondas que hibridam perto da extremidade 5' terão níveis de intensidade mais baixos. Para que os *arrays* sejam comparáveis, há que transformar a escala de modo que os erros padrão da média dos níveis de intensidade PM sejam aproximadamente 1. Um indicador de boa qualidade das amostras é a existência de um declive positivo dos níveis de intensidade médios das sondas PM dos *arrays*, geralmente a variar entre 0.5 e 1.7 dependendo do tipo de *array*. Se existir um declive superior a 2 pode ser um indicador de que a degradação é excessiva. Mas o mais importante é que todos os *arrays* possuam declives semelhantes como é o caso que se apresenta na Figura 2.5. O declive pode variar de experiência para experiência.

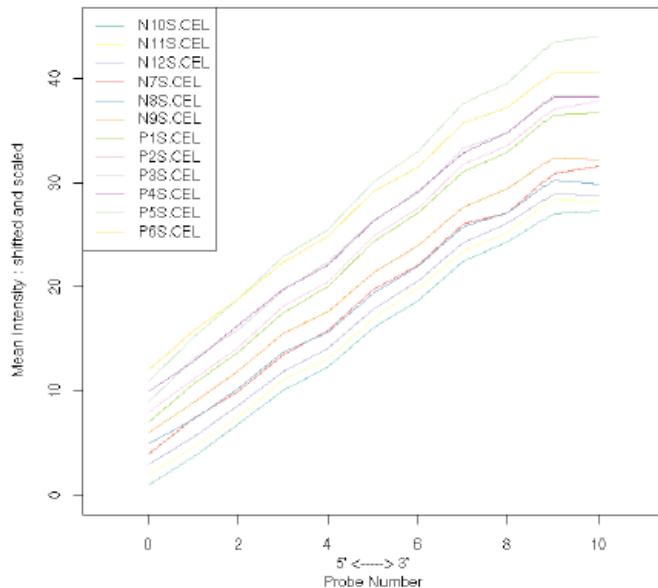


Figura 2.5: *Degradation plot*. Cada linha representa um *array* (fonte: <http://arrayanalysis.org/main.html> em 20/05/2011).

Simple Affy Plot

A Affymetrix (Affymetrix, 2004) recomenda que se examinem as seguintes medidas de avaliação da qualidade: médias de *background*, fator escala e percentagem de *present calls*.

A média de *background* indica o nível de ruído *background* que um *array* possui. Existem várias razões que justificam a variabilidade inerente aos diferentes níveis de ruído médio de *background* associado a cada *array*, como por exemplo, diferentes quantidades de mRNA presentes na hibridação, ou o processo de hibridação ter sido realizado de forma mais eficiente, produzindo *arrays* mais fluorescentes. É recomendado que estes valores sejam semelhantes em todos os *arrays*.

O fator escala refere-se ao tipo de escala aplicado aos *arrays* aquando do processo de normalização a partir do algoritmo MAS 5.0 da Affymetrix. Por omissão o MAS 5.0 transforma a escala das intensidades de cada *array* de modo que todos tenham a mesma média. A Affymetrix (Affymetrix, 2004) recomenda que o fator escala esteja compreendido entre -3 a 3 entre todos os *arrays*.

A percentagem de *present calls* é obtida através da diferença entre os pares de sondas PM e MM num *probeset* e representa a percentagem de *probesets* presentes num *array* cujos níveis de intensidade das sondas PM são superiores aos das sondas MM. Os *probesets* são designados de marginais ou ausentes quando os valores das sondas PM de um *probeset* não apresentarem valores significativamente superiores aos das sondas MM. A Affymetrix recomenda que estes valores também sejam semelhantes entre os *arrays*.

Na Figura 2.6 sumaria-se estas e outras medidas. Através da biblioteca *affyQCreport* do *Bioconductor Project* é possível obter-se este gráfico. É importante referir que esta análise de controlo da qualidade foi desenvolvida para *arrays* humanos HGU95A, o que significa que os valores especificados devem de ser interpretados caso a caso, uma vez que não foram testados para todos os tipos de *arrays*.

A análise da Figura 2.6 carece de alguma descrição uma vez que a sua interpretação não é imediata. Pode observar-se uns números a seguir à identificação dos *arrays* (.CEL), que representam a percentagem de *present calls* e a média do *background*. A superfície que se encontra a azul, representa o intervalo de variação do fator escala, e recomenda-se que os valores se

encontrem dentro do intervalo +/- 3 unidades da média de todos os *arrays*.

A Affymetrix desenhou *probesets* para hibridarem com certos transcritos, designados de *housekeeping genes*, como por exemplo o GAPDH (*Glyceral dehyde 3-phosphate dehydrogenase*) e o *beta-actin*, uma vez que estes se encontram expressos na maioria das células e são genes relativamente longos. Três *probesets* são obtidos em três regiões destes genes: terminação 5', centro (definido por M) e terminação 3'. Assim, é possível controlar se os transcritos não foram truncados e se foram etiquetados (quimicamente) de forma equitativa ao longo da sequência. É importante voltar a referir que, como a degradação do RNA tem início na extremidade 5', é comum que a intensidade do *probeset* associado a esta região tenha o nível de intensidade mais baixo. Na Figura 2.6 os valores GAPDH 3':5' são representados por círculos. Valores inferiores a 1.25 (recomendado) correspondem a círculos a azul e superiores a 1 a vermelho. Aqui também é importante que a análise seja feita caso a caso, podendo acontecer que todos possuam círculos a vermelho, não querendo dizer que todos os *arrays* devam ser excluídos da análise, mas considera-se que para esse tipo de *arrays* os valores devem ser ajustados. Os valores *beta-actin* (ou *b-actin*) 3':5' são representados por triângulos. Valores inferiores a 3 (recomendado) correspondem a triângulos azuis e valores superiores a 3 correspondem a triângulos vermelhos.

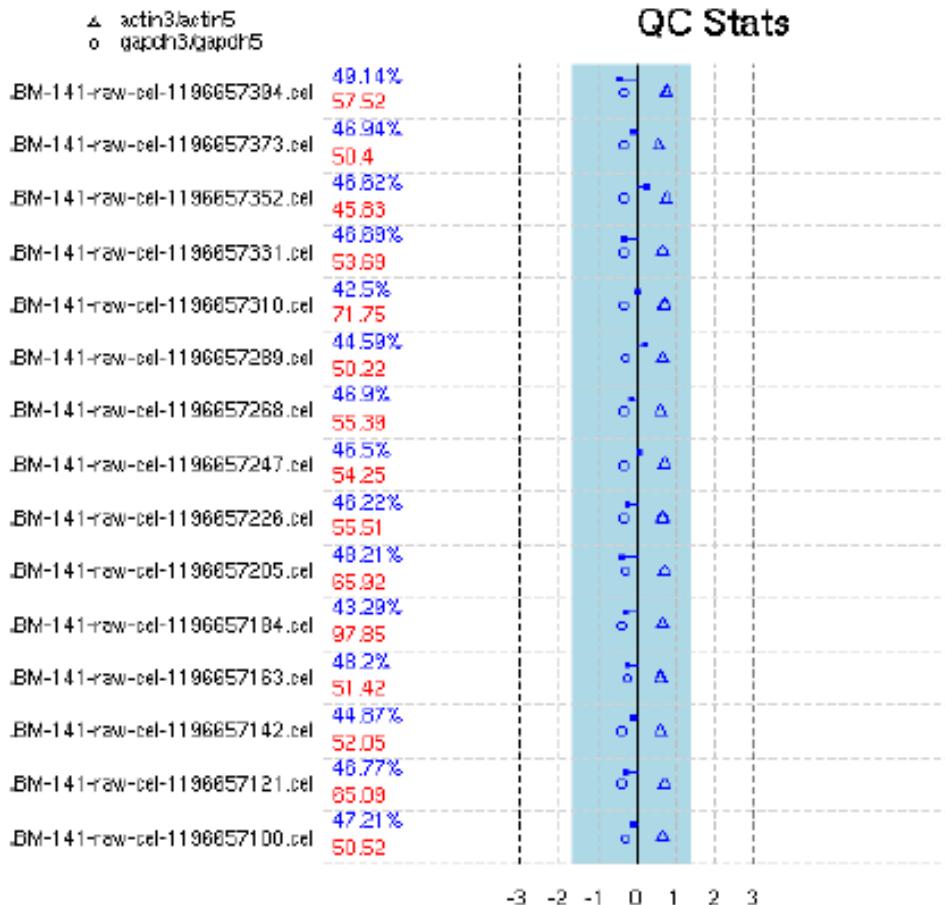


Figura 2.6: *Simple Affy Plot* (fonte: <http://arrayanalysis.org/main.html> em 20/05/2011).

Gráficos RLE e NUSE

O gráfico NUSE⁶ (Figura 2.7 A) descreve a distribuição das estimativas dos erros padrão dos níveis médios dos genes em cada *array*, ao se ajustar um modelo PLM (*Probe-Level Model*⁷). Tendo em conta que a variabilidade difere consideravelmente de gene para gene, a estimativa do

⁶Do inglês *Normalized Unscaled Standard Error*.

⁷O modelo PLM é dado por $Y_{ijk} = \mu_{ik} + \alpha_{jk} + \varepsilon_{ijk}$, onde μ_{ik} representa o nível de expressão na escala logarítmica para o gene k no *array* i , e α_{jk} representa o efeito de afinidade da sonda j no gene k .

erro é estandardizada de modo que a mediana do erro padrão em todos os *arrays* seja 1 para cada gene. *Arrays* de boa qualidade devem produzir *box plots* centrados em 1 e *box plots* que tenham maior dispersão interquartil ou se encontrem mais afastados da maioria dos *arrays*, são indicadores de *arrays* com má qualidade. Caso permaneçam com comportamento semelhante após pré-processamento deve considerar-se a sua remoção do estudo.

O gráfico RLE⁸ (Figura 2.7 B) é construído com base nos valores obtidos pela razão entre os níveis de expressão dos *probesets* e a mediana do nível de expressão do respetivo *probeset* de todos os *arrays*. É admitido que a maioria dos *probesets* não se altera ao longo dos *arrays*, por isso é esperado que os valores destas razões se encontrem em torno de zero na escala logarítmica. Assim, espera-se que os *box plots* construídos com base nestas razões para cada *array*, se encontrem centrados em zero e com amplitudes semelhantes.

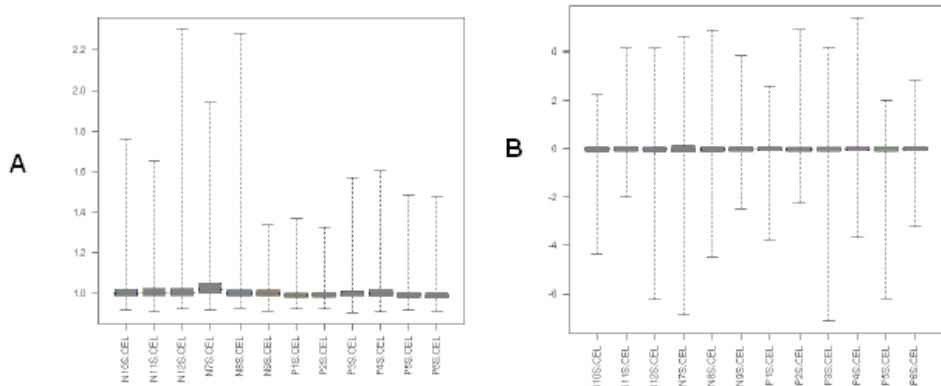


Figura 2.7: A — Gráfico NUSE; B — Gráfico RLE (fonte: <http://arrayanalysis.org/main.html> em 20/05/2011).

⁸Do inglês *Relative Log Expression*.

Gráficos *Pseudo Array Images*

Os gráficos *Pseudo Array Images* (Figura 2.8) são úteis para detetar possíveis artefactos nos *arrays* que colocam em causa a sua qualidade. A análise destas imagens tem como objetivo identificar regiões com marcas, por exemplo riscos, círculos, zonas com brilho muito elevado/baixo, etc. O que a Affymetrix regulamenta é que, desde que estas regiões não tenham uma área superior a 10% do total, as sondas que se encontram nestas áreas podem ser consideradas como valores omissos.

Existem várias funções associadas a algumas bibliotecas do *Bioconductor Project* que permitem a obtenção destas imagens. Na Figura 2.8 apresenta-se um exemplo de imagens que se obtêm através da função `spatialImagesfunction` da biblioteca `affyPLM`. No entanto esta opção consome muito tempo e espaço de memória. A função `image` da biblioteca `affy` é utilizada para obter imagens como a que está representada na Figura 2.8 A.

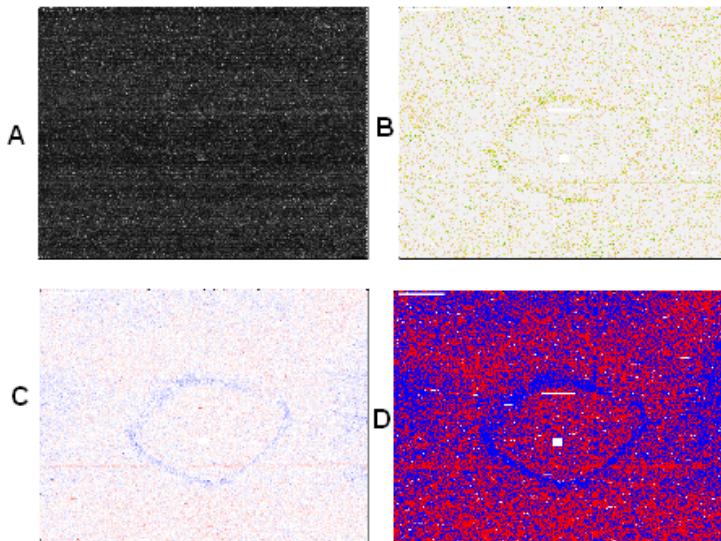


Figura 2.8: A — Imagem de um *array* para os dados em bruto; B — Imagem do *array* para os pesos PLM; C — Imagem de um *array* para os resíduos PLM; D — Imagem de um *array* para os sinais dos resíduos PLM (fonte: <http://arrayanalysis.org/main.html> em 20/05/2011).

2.2.5 Pré-processamento

Após a análise dos dados em bruto, usualmente é necessário proceder-se a uma análise de pré-processamento dos dados. No caso dos *microarrays* de um canal consiste na correção de *background*, normalização, correção PM e sumariação (Bolstad *et al.*, 2003), não necessariamente por esta ordem. Cada um destes passos é descrito sumariamente.

Correção de *Background*

Anteriormente já foi descrito que a molécula mRNA transcrita a partir de um gene é representada no *chip* por um *probeset*, que por sua vez é composto por um número de *probepairs* (pares de sondas), tipicamente entre 11 e 20. Este número de pares de sondas, deve-se ao facto de durante a transcrição só se poderem usar bases perto da extremidade 3' até a uma distância de 700pb. Cada *probepair* é composto por uma sonda PM e uma sonda MM. Cada sonda PM contém milhares de pequenas sequências oligonucleotídicas (25 mer) que coincidem com um segmento de um exão do gene de interesse (em alguns casos a sonda pode coincidir com um segmento de um intrão ou de uma outra região). A sonda MM contém sequências oligonucleotídicas idênticas à PM exceto o nucleótido que se encontra no centro da sequência, cujo objetivo é controlar hibridações não específicas ou cruzadas. Podem ocorrer três cenários: apenas a sequência PM hibrida (situação ideal), ambas as sondas PM e MM hibridam (esta sonda não trará nenhuma informação) ou PM e MM não hibridam (provavelmente o gene não está presente).

A correção de *background*, também referida por ajustamento de sinal (Bolstad *et al.*, 2003), tem como objetivos: corrigir o ruído e os efeitos de processamento dos *arrays*; ajustar as hibridações cruzadas ou não específicas (quando existem sequências não complementares que se ligam às sondas MM) e ajustar as estimativas das expressões por forma a serem linearmente relacionadas com as concentrações (Irizarry *et al.*, 2003).

A Affymetrix (1999) com o seu algoritmo MAS 4.0 apresenta um método de correção de *background* baseado nas diferenças PM-MM; no entanto, Irizarry *et al.* (2003a, 2003b) verificaram que estas diferenças conduziam a níveis de expressão com variâncias muito elevadas e propuseram o método RMA (*Robust Multi Array Analysis*). Este é um método de pré-processamento em que a correção de *background* utilizada é designada por *PMonly*, isto é, toma em consideração apenas as intensidades das sondas PM. Esta abordagem sacrifica alguma exatidão num ganho de precisão. Este método tem em consideração a transformação logarítmica das intensidades

e a biblioteca `affy` do projeto *Bioconductor* do R implementa este método. O método GCRMA (*Guanine Cytosine Robust Multi-Array Analysis*) proposto por Wu *et al.* (2004) é uma versão do RMA onde a correção de *background* integra a informação relativa às sequências das sondas MM; a biblioteca `gcrma` do projeto *BioConductor* do R implementa este método. Cope *et al.* (2006) sugerem uma alternativa ao RMA, RMA-noBG, onde não se faz correção de *background*. Em 2001 a Affymetrix apresenta o método de pré-processamento MAS 5.0, cuja correção de *background* tem em consideração ambas as sondas, PM e MM, e também a localização da sonda no *chip* (consultar Affymetrix (2002) para maior detalhe).

O método de pré-processamento PLIER (*Probe Logarithmic Intensity Error Estimation*) da Affymetrix (2005) oferece várias opções de correção de *background* de acordo com as características dos dados, nomeadamente: o já anteriormente descrito PM-MM (situações onde existe elevada variabilidade de amostra para amostra); PM-B (*perfectmatch* menos *background*) é útil para moderar a variabilidade de amostra para amostra e modera a sensibilidade a níveis baixos de intensidade; PMonly (utiliza apenas as sondas PM) é muito utilizado em experiências onde o *background* é admitido com variabilidade mínima ao longo da experiência e a opção PM+MM é utilizada em casos onde o *background* é considerado como irrelevante para o objetivo da experiência.

O método MBEI (*Model-Based Expression Index*) desenvolvido por Li *et al.* (2003) é implementado no software dChip⁹ (*DNA-Chip Analyzer*). Tem incluídas duas opções de correção *background*: (a) não ser feita qualquer correção de *background* (opção definida por omissão) e (b) ter em consideração as diferenças PM-MM.

Um método mais recente, proposto por Chen *et al.* (2006), DFCM (*Distribution Free Convolution Model*), utiliza os níveis de intensidade das sondas MM correspondentes às sondas PM com níveis de intensidade mais baixos para estimar o ruído associado ao *background*.

Normalização

Para além dos inevitáveis erros aleatórios, também existem erros sistemáticos devido aos protocolos experimentais, aos parâmetros de digitalização e às diferenças existentes nos *chips* por serem produzidos em diferentes “fornadas”.

⁹<http://biosun1.harvard.edu/complab/dchip/>.

Para que os *chips* sejam comparáveis é necessário que os níveis de expressão dos diferentes *arrays* tenham a mesma distribuição.

A *normalização* tem como objetivo principal a remoção de variações não biológicas que possam existir entre os *arrays* (Irizarry *et al.*, 2003a, 2003b; Bolstad *et al.*, 2003), para que seja possível compará-los entre si. Outro problema existente neste tipo de *microarrays* é a *saturação*, isto é, ambas as sondas PM e MM atingem o máximo da intensidade permitida pelo *scanner* (McLachlan *et al.*, 2004). Esta situação pode levar à não deteção de genes DE. Uma forma de controlar esta situação, é por exemplo, baixar a intensidade do *scanner* ou proceder à normalização.

Antes de se proceder à normalização, recomenda-se uma análise exploratória com intenção de detetar fontes de enviesamento. Por exemplo construir *box plots* para $\log_2(\text{PM})$, $\log_2(\text{MM})$, $\log_2(\text{PM/MM})$ ou $\text{PM} - \text{MM}$ para cada *array*. Alternativamente, também pode explorar-se os enviesamentos para cada combinação de pares de *arrays* com recurso aos gráficos MvA, mais conhecidos por gráficos MA (Figura 2.9). Para cada par de *arrays*, são calculados dois valores, M_i (2.3) e A_i (2.4), onde M_i representa a diferença dos logaritmos das intensidades (*fold change*) do gene i , e A_i representa a média dos logaritmos das intensidades do gene i (M de *minus* e A de *add*),

$$M_i = \log_2(x_{ij}) - \log_2(x_{ik}), \quad (2.3)$$

$$A_i = \frac{\log_2(x_{ij}) + \log_2(x_{ik})}{2}, \quad (2.4)$$

onde x_{ij} representa o nível de intensidade do gene i no *array* j e x_{ik} representa o nível de intensidade do gene i no *array* k .

A partir do gráfico representado na Figura (2.9) é possível identificar a tendência das intensidades através da curva *lowess* ajustada (a vermelho). Se não existirem diferenças sistemáticas entre os dois *arrays*, espera-se que os pontos se encontrem em torno de $M = 0$ para todos os valores de A. O desvio deste ajustamento revela a necessidade de se proceder à normalização.

Alternativamente, é possível obter gráficos MA em função da comparação dos *arrays* com um *array* de referência. A biblioteca *affy* do *Bioconductor*

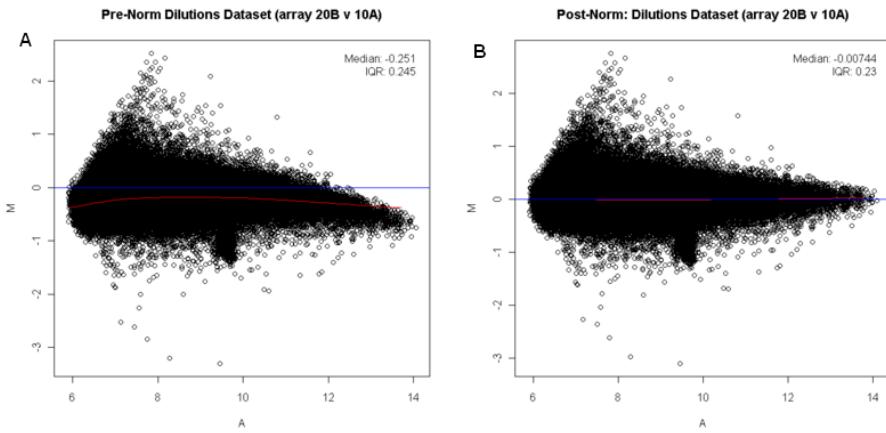


Figura 2.9: A — Gráfico MA antes da normalização. B — Gráfico MA após normalização. A mediana e amplitude interquartil (IQR^{10}) que surgem nos gráficos, são obtidas em função dos valores M_i (fonte: http://en.wikipedia.org/wiki/MA_plot em 17/04/2008).

Project tem a função `ma.plot`, cujo *array* de referência é obtido a partir da mediana de todos os *arrays* da experiência. Este gráfico não é nada mais do que o conhecido gráfico Bland-Altman (Bland e Altman, 1986), vulgarmente utilizado para analisar a concordância entre duas técnicas.

Quando se aplicam métodos de normalização, há que verificar os pressupostos, nomeadamente o número de genes com alterações nas expressões entre as diferentes condições experimentais, o qual deve ser pequeno, e o número de genes com aumentos e decréscimos das expressões, o qual deve ser equivalente, ou seja, o número de genes diferencialmente expressos deve ser pequeno (aproximadamente 1% a 5%) e o número de genes com regulação positiva deve ser semelhante ao número de genes com regulação negativa (Silva-Fortes *et al.*, 2007).

O processo de normalização envolve duas decisões: que genes e algoritmos usar. Os genes que servirão de base para a normalização podem ser escolhidos de acordo com quatro opções. Podem ser usados todos os genes existentes nos *arrays*, sendo esta opção baseada no facto de que a maioria dos genes tem o mesmo nível de expressão nas amostras que estão a ser alvo de comparação, considerando que menos de 5% dos genes são DE. Esta opção não pode ser tomada quando este pressuposto não se verifica, por

exemplo quando as amostras são altamente heterógenas. Outra opção é utilizar genes *housekeeping*, que são genes que estão sempre expressos, uma vez que codificam proteínas que são necessárias ao funcionamento da célula e supõe-se que a sua expressão não é afetada pelas condições experimentais. O problema associado a estes genes é que geralmente são expressos com níveis elevados, não sendo informativos em amplitudes de intensidade baixas. Os *Spike-in controls* são transcritos com níveis de concentração conhecidos (Schadt *et al.*, 2001) e são acrescentados em determinada quantidade a cada amostra. O método *invariant set*, proposto por Li e Wong (2003), é parte do método de pré-processamento MBEI seleciona os genes a usar na normalização após a análise de resultados. A ideia é detetar genes com níveis de expressão semelhantes em todos os *arrays* e basear a normalização neles. Uma forma de identificar estes genes é ordená-los de acordo com os níveis de expressão e usar os que têm a mesma ordem.

Existem vários métodos para a normalização dos níveis de intensidade das sondas. Os métodos existentes incluem normalizações lineares e extensões através de regressão não-linear. Bolstad *et al.* (2003) apresentam um método baseado nos gráficos MA, *cyclic loess*, onde a normalização é feita à custa dos resíduos $M_i - \hat{M}_i$ (onde \hat{M}_i é estimado em função do modelo *loess*). Caso existam mais de dois *arrays*, o método *orthonormal contrast-based* pode ser usado. *Quantile normalization* (Bolstad, 2001) consiste em normalizar os *arrays* de modo que tenham quantis iguais, garantindo que os *arrays* possuam as mesmas distribuições dos níveis de intensidade, no entanto não garante que os mesmos genes tenham os mesmos níveis de intensidade. *Global normalization (scaling ou constant)* é um método que tem por objetivo igualar os valores médios dos níveis de expressão, assumindo uma relação linear entre os *arrays*. No entanto é um método com algumas fragilidades (McLachlan, 2004), por exemplo admite que o valor médio dos níveis de expressão de todo o mRNA é constante. Rocke e Durbin (2001) introduzem o método VSN (*Variance Stabilization and Normalization*) que é construído tendo por base o facto da variância inherente aos dados de *microarrays* depender da intensidade dos sinais, sendo a transformação obtida após a variância ser aproximadamente constante. Uma vantagem da *vsn*-transformação em relação à *log*-transformação, é que a normalização pode ser aplicada a valores negativos após a subtração de *background*. Qspline (Workman *et al.*, 2002) é um método não linear que utiliza os quantis de cada *array* e ajusta-os a um sistema de *splines* cúbicas para normalizar os dados.

Outros métodos menos conhecidos como o *fastlo*, proposto por Ballman *et al.*

(2004), que é semelhante ao *cyclic loess* e ao *quantile normalization*, sendo no entanto computacionalmente mais rápido do que o *cyclic loess*. Zhang *et al.* (2003) propõem o modelo PDNN (*positional-dependent-nearest-neighbor*). Este modelo revela que os sinais das sondas dependem das sequências e tem em consideração dois diferentes modos de ligação das sondas — hibridação específica e hibridação não-específica — e atribui diferentes pesos a cada posição do nucleótido na sonda, de modo a refletir que diferentes partes da sonda podem contribuir de forma diferente para a estabilidade da hibridação. O método WBL (*Weibull distribution based normalization*), proposto por Autio *et al.* (2006), admite que os logaritmos dos níveis de intensidade das sondas podem ser ajustados pela distribuição de Weibull.

Se se estiver perante uma meta-análise (*arrays* provenientes de diferentes plataformas) o processo de normalização traz novos problemas, nomeadamente o número de genes em cada *array* poder ser diferente. O método designado por *equalized quantile normalization* (Kauraniemi *et al.*, 2004), pode ser entendido como uma extensão do *quantile normalization* anteriormente descrito. Outro método que tem em consideração *arrays* de várias plataformas é o método *array generation based gene centering* (AGC) proposto por Kilpinen *et al.* (2008). Este método admite que a média dos níveis de expressão de um gene em cada *array* de cada plataforma deve ser o mesmo. Se a média dos níveis de expressão de alguma das plataformas for substancialmente diferente da das outras, admite-se que é causado pela variação associada à plataforma. Este método tem como objetivo corrigir esta variação.

Quando a igualdade do número de genes com regulação positiva com o número de genes com regulação negativa não se verifica, a maior parte dos métodos de normalização falham, no entanto o método *global loess* (Freudenberg, 2005) é robusto relativamente a esta situação.

Todos os processos de normalização inevitavelmente alteram os dados e podem introduzir viés. Podem existir razões biológicas para que as distribuições sejam enviesadas, e o processo de normalização, de uma forma artificial, pode remover diferenças biológicas importantes. Não existe uma forma correta para a normalização, e por isso, é vantajoso que se apliquem vários e se comparem. Vários autores apresentam estudos com o propósito de comparar vários métodos de normalização e identificar os que melhor se ajustam aos dados como por exemplo Schadt *et al.* (2001), Bolstad (2002), Bolstad *et al.* (2003) e mais recentemente Autio *et al.* (2009).

Correção PM

Como já foi referido anteriormente, as sondas PM dos *arrays* da Affymetrix medem, simultaneamente, a abundância relativa do gene e uma certa quantidade correspondente a hidrilações não-específicas. Sondas MM são desenvolvidas para medir hidrilações não-específicas das suas correspondentes sondas PM. Parece óbvio que uma abordagem será subtrair os valores MM dos valores PM, mas geralmente cerca de 30% das sondas MM têm valores superiores às suas correspondentes sondas PM (Naef *et al.*, 2003). Chudin *et al.* (2001) demonstraram que as sondas MM possuem um sinal específico. Isto acontece porque elevadas quantidades de mRNA, que intencionalmente hibridam com as sondas PM, também hibridam com as sondas MM e demonstrou-se que a sensibilidade das sondas à hidrilação está relacionada com as bases que constituem a sonda. As sondas MM são mais sensíveis que as suas correspondentes sondas PM, se a base que se encontra no meio for uma purina (A ou G), e a relação é inversa se for constituída por uma pirimidina (C ou T). Quando os níveis de intensidade associados à sonda MM for superior à sonda PM, conduz a valores negativos do sinal. Muitos dos métodos de pré-processamento mais populares resolvem este problema simplesmente ignorando as sondas MM, sendo os valores das sondas PM corrigidos usando outros métodos.

O algoritmo MAS 5.0 da Affymetrix utiliza um método para evitar níveis de expressão negativos, onde as sondas MM são substituídas pelas estimativas *ideal mismatch* (IM) e que consiste em:

1. Se a intensidade da sonda MM for inferior à sua correspondente PM, então $IM=MM$.
2. Se a intensidade da sonda MM for superior à da sua correspondente PM, mas outras sondas MM correspondentes ao mesmo *probeset* não são, então a IM é estimada em função das restantes sondas PM e MM do mesmo *probeset*.
3. Se a maioria das sondas MM do mesmo *probeset* têm intensidades superiores às suas correspondentes sondas PM, então é atribuído a IM um valor ligeiramente inferior ao da sonda PM.

Finalmente, o valor ajustado do *probeset* é obtido subtraindo IM a PM.

Sumariação

Para se obter uma medida da expressão que represente a abundância de mRNA na amostra que hibridou com o *chip*, deve proceder-se à *sumariação* (este passo só se verifica nos *microarrays* de um canal). Mais especificamente, as intensidades das sondas PM e MM para cada *probeset* são combinadas de modo a representar os valores da expressão de um gene. Idealmente, os índices de expressão devem ser conjuntamente precisos (variância reduzida) e exactos (viés reduzido) (Zhou e Rocke, 2005).

A Affymetrix (1999) desenvolveu uma medida de sumariação designada por *Average Difference* (AD) integrada no método MAS 4.0. Neste método o nível de expressão do gene é estimado à custa da diferença das sondas PM e MM e, para eliminar algum viés, considera-se a média destas diferenças.

Irizarry *et al.* (2003) concluiram que esta medida, assim como outras com medida de escala linear, não é ótima. Em 2001 a Affymetrix (Affymetrix, 2001) desenvolveu uma nova medida de sumariação baseada na função Tukey Biweight, designada por MAS 5.0. Este método tem em consideração o facto de 1/3 das sondas MM terem níveis de intensidades inferiores às sondas PM. Também é conhecido por *Ideal Mismatch* (IM).

Li e Wong (2001) propõem o índice MBEI (*Model-Based Expression Indices*) disponível no *package* dChip (Li e Wong, 2003). Utilizam um modelo baseado nos dados originais (sem transformação de escala). Este método tem em consideração o facto de algumas sondas do mesmo gene possuirem uma maior afinidade de hibridação, fazendo com que tenham níveis de intensidade superiores. Bolstad *et al.* (2003) e Irizarry *et al.* (2003) propõem um método não-paramétrico designado por RMA (*Robust Multi-chip Average*). Utilizam a *median polish* sobre os dados numa escala logarítmica tendo em linha de conta a elevada correlação entre as sondas PM e MM. Este método considera apenas as sondas PM. A Affymetrix (2005) apresenta o PLIER (*Probe Logarithmic Intensity Error*). Este índice tem em consideração as diferentes afinidades de hibridação das sondas e é um método sensível a níveis de expressão baixos. Uma versão do RMA, o sRMA (*small-sample RMA*) proposto por Cope *et al.* (2005), é um exemplo da adaptação dos métodos a situações não previstas inicialmente. Neste caso, a correção de *background* e a normalização mantêm-se iguais ao método original, substituindo o método de sumariação *median polish* por um novo modelo, cuja implementação leva em consideração pesos associados às sondas das quais as posições se situem mais ou menos próximas das terminações 5' e 3' do transcrito.

Um método mais recente, FARMS (*Factor Analysis for Robust Microarray Summarization*), proposto por Hochreiter *et al.* (2006), utiliza dados na escala logarítmica e baseia-se num modelo linear. Tem em consideração apenas as sondas PM, justificando que as sondas PM têm ruído baixo para níveis de intensidade baixos, comparativamente com as diferenças PM-MM, enquanto para níveis de intensidade intermédios ou elevados, os valores do ruído das sondas PM e das diferenças PM-MM são semelhantes. O método DFW (*Distribution Free Weighted*), proposto por Chen *et al.* (2006), utiliza a informação acerca da variabilidade existente na sonda para estimar hibridações não específicas ou hibridações cruzadas. Este método não impõe qualquer condição às distribuições dos níveis de intensidade das sondas e também só consideram as sondas PM.

2.2.6 Algumas considerações finais

Dos vários métodos disponíveis, não existe um consenso de otimalidade. O projeto *BioConductor* do R tem uma variedade enorme de bibliotecas que implementam vários métodos de pré-processamento, oferecendo também a possibilidade de se realizarem várias combinações entre eles e decidir pelos que nos conduzam a distribuições com o melhor comportamento. É importante referir que as várias combinações entre métodos têm de ser feitas com precaução, uma vez que alguns métodos produzem valores negativos, nomeadamente métodos que envolvam as diferenças (PM-MM) (cerca de 30% das sondas têm as intensidades MM superiores às PM) logo não podem ser combinados com métodos que utilizam a escala logarítmica. Há uma função disponível na biblioteca **affy**, a função **expresso**, que aplica simultaneamente todos os passos de pré-processamento, deixando o utilizador escolher qual a combinação que pretende utilizar.

Vários autores realizaram comparações entre vários métodos de pré-processamento aqui descritos, como por exemplo Bolstad (2002), Lemon *et al.* (2002), Saviozzi e Calogero (2003), Freudenberg (2005), Shedden *et al.* (2005), Irizarry *et al.* (2006) e Jiang *et al.* (2008).

É preciso ter em consideração que os métodos utilizados têm uma profunda influência nos resultados. Segundo Quackenbush (2001) não existe consenso sobre o melhor método para revelar padrões de expressão nos dados e fica cada vez mais claro que não existe uma técnica superior. A aplicação de várias técnicas permite explorar aspetos diferentes dos dados.

Após a análise de pré-processamento vem a designada *high-level analysis*, que consiste, por exemplo, na seleção de genes DE, *screening*, análise de classificação supervisionada, etc. Os dados correspondentes aos níveis de expressão genética são geralmente representados numa matriz (Tabela 2.2), onde as linhas representam os genes e as colunas os *chips*. Esta matriz tem o nome de matriz de expressão. Cada posição da matriz contém um número que representa o nível de expressão de um gene em particular numa amostra. Devido ao elevado custo dos *microarrays*, uma matriz de expressão é, geralmente, constituída por um grande número de genes (milhares) e poucas amostras (dezenas ou menos).

Tabela 2.2: Exemplo de uma matriz de expressão.

Gene	Array 1	Array 2	...	Array n
Gene 1	1050	345	...	65
Gene 2	4578	98	...	345
Gene 3	1025	120	...	678
...
Gene g	5000	0	...	89

Na secção que se segue apresentam-se alguns dos métodos mais comumente utilizados para a seleção de genes DE.

2.3 Métodos para a seleção de genes DE

2.3.1 Introdução

Um dos objetivos da tecnologia de *microarrays* é medir a expressão de milhares de genes e identificar mudanças nas expressões entre diferentes estados biológicos. Assim, de um elevado número de genes em análise é possível obter um número substancialmente mais reduzido de genes com a(s) característica(s) de interesse para posterior análise.

As técnicas estatísticas que conduzem à seleção dos genes de interesse são empregues de acordo com as características das diferentes plataformas, das técnicas experimentais adotadas, das condições experimentais, etc. Nos

métodos baseados na hibridação, a experiência em causa pode referir-se a uma única amostra da condição de interesse ou à relação entre duas (ou mais) condições diferentes. Na primeira, pode dar-se o caso de ter interesse apenas em ler-se a quantidade de mRNA de uma amostra a partir da intensidade de um sinal de hibridação. Na segunda situação, pode dar-se o caso de se identificar os genes cujos níveis de expressão na condição A seja superior aos níveis de expressão na condição B. A estes genes dá-se o nome de genes com regulação positiva (*up-regulated*), ou então, identificar genes com níveis de expressão na condição A inferiores aos da condição B, a que se dá o nome de genes com regulação negativa (*down-regulated*). As condições experimentais podem ser relativas a diferentes tecidos, diferentes estádios de desenvolvimento do organismo ou das células de interesse, diferentes condições a que se submetem as amostras, etc.

Na próxima secção descrever-se-ão alguns dos métodos mais vulgarmente utilizados para a ordenação e seleção de genes DE. Sendo a metodologia ROC¹¹ um dos principais temas deste trabalho, métodos de seleção de genes DE especificamente baseados nesta metodologia serão abordados no capítulo 3. Todos os métodos serão abordados no sentido de se selecionar genes DE submetidos a duas condições mutuamente exclusivas (*e.g.*, amostra controlo *vs.* amostra experimental).

2.3.2 Estado da Arte

De uma forma geral, a notação utilizada para os dados na fase após pré-processamento é dada pela matriz de expressão representada na Tabela 2.3, onde sem perda de generalidade, x_{ij} representa o nível de expressão do i -ésimo gene ($i = 1, \dots, g$) no j -ésimo array ($j = 1, \dots, n$).

Tabela 2.3: Matriz de expressão para g genes e n arrays.

Genes	Array 1	...	Array n
1	x_{11}	...	x_{1n}
2	x_{21}	...	x_{2n}
...
g	x_{g1}	...	x_{gn}

¹¹Do inglês *Receiver Operating Characteristic*.

De futuro, para facilidade de notação, vai considerar-se que x_{ij} representa o nível de expressão do i -ésimo gene ($i = 1, \dots, g$) no array j da amostra controlo, $j = 1, \dots, n_0$; y_{ik} representa o nível de expressão do gene i no array k na amostra experimental, $k = 1, \dots, n_1$, e $n_0 + n_1 = n$ corresponde ao número total de arrays na experiência.

A análise de genes diferencialmente expressos baseia-se fundamentalmente em dois aspetos: ordenação e seleção. A ordenação requer a especificação de uma estatística (ou medida) que evidencie o nível de expressão de cada gene; a seleção requer a especificação de um procedimento que arbitre o que constitui ser DE. A escolha do método influencia a lista final de genes DE. Apesar da grande variedade de métodos existentes, os biólogos tendem a utilizar duas das abordagens mais antigas, o *fold-change* (FC) e a estatística- t , presumivelmente por serem simples de interpretar e fáceis de implementar. Nesta secção serão descritos métodos que se baseiam nestas duas abordagens.

Métodos baseados em *Fold Change* (FC)

Fold Change

Tusher *et al.* (2001) utilizam a definição clássica do FC para o gene i ,

$$\text{FC}_i = \frac{\bar{y}_{i\cdot}}{\bar{x}_{i\cdot}}, \quad (2.5)$$

onde $\bar{x}_{i\cdot} = \frac{1}{n_0} \sum_{j=1}^{n_0} x_{ij}$ e $\bar{y}_{i\cdot} = \frac{1}{n_1} \sum_{k=1}^{n_1} y_{ik}$.

Usualmente é utilizada a escala logarítmica (2.6), pois a sua interpretação é mais intuitiva. Se o valor for zero significa que o gene é não diferencialmente expresso, e se se fixar o *fold-change* na escala logarítmica por exemplo em 2, isto significa que para valores superiores a 1 o gene tem regulação positiva e para valores inferiores a -1 tem regulação negativa,

$$\log_2(\text{FC}_i) = \log_2 \left(\frac{\bar{y}_{i\cdot}}{\bar{x}_{i\cdot}} \right). \quad (2.6)$$

O *fold change* é conhecido por traduzir uma interpretação biológica dos dados. Os genes são ordenados de acordo com o valor absoluto de (2.5) ou (2.6), e de acordo com um determinado ponto de corte arbitrário, selecionam-se os primeiros genes da lista. Tradicionalmente utilizam-se pontos de corte superiores a 2 para selecionar os genes de interesse (Dalman *et al.*, 2012). De futuro, passa-se a dizer que um gene é significativo quanto maior for a importância do gene no estudo.

Average Difference (AD)

O *Average Difference* (AD) resulta da diferença das médias entre duas condições experimentais, *i.e.*, para cada gene i ,

$$\text{AD}_i = \bar{y}_{i\cdot} - \bar{x}_{i\cdot}. \quad (2.7)$$

Os genes são ordenados de acordo com o valor absoluto de AD e para valores elevados os genes correspondentes são considerados significativos. No entanto, alguns dos genes que ficam ordenados nas primeiras posições a partir da AD, podem exibir níveis de expressão baixos (Kadota *et al.*, 2008), pois esta medida não tem em consideração a variabilidade dos dados.

Weighted Average Difference (WAD)

Kadota *et al.* (2008) propõem o método WAD (*Weighted Average Difference*), que resulta da combinação do método AD e de um peso que pretende capturar a média relativa da intensidade do sinal,

$$\text{WAD}_i = \text{AD}_i \times w_i = \text{AD}_i \times \frac{(\bar{x}_{i\cdot} + \bar{y}_{i\cdot})/2 - \min}{\max - \min} \quad (2.8)$$

onde \max (ou \min) indica o valor máximo (ou mínimo) do vetor das médias $((\bar{x}_1 + \bar{y}_1)/2, \dots, (\bar{x}_g + \bar{y}_g)/2)$, quanto maior for o valor absoluto da estatística WAD, mais significativo é considerado o gene.

Rank Products (RP)

O método *Rank Products* (Breitling *et al.*, 2004; Breitling e Herzyk, 2005) é um método não-paramétrico que permite identificar genes com regulação positiva e/ou genes com regulação negativa entre duas condições experimentais. Originalmente desenvolvido para *microarrays* de dois canais, é também adaptado a dados provenientes de *microarrays* de um canal. Este método é o mais recomendado quando existem poucas réplicas (<10), situação muito comum devido aos elevados custos associados aos *microarrays*. No entanto, esta técnica admite variâncias iguais para todos os genes, podendo conduzir à seleção de muitos genes no caso desta condição ser violada. Para a aplicação desta técnica é muito importante que se realizem técnicas de normalização de modo a estabilizar a variância. Quando a normalização não for possível ou quando as variâncias forem diferentes, Breitling e Herzyk (2005) propõem uma variante do método RP, *average ranks*.

O RP baseia-se no pressuposto de que, sob a hipótese de não haver genes DE, um gene numa experiência que envolva g genes em $n_0 + n_1 = n$ réplicas, tem probabilidade $1/g^n$ de se encontrar na primeira posição de uma lista ordenada. Assim, é improvável que um determinado gene se encontre na primeira posição em todas as réplicas caso não seja diferencialmente expresso.

Para *microarrays* de um canal o cálculo da estatística de teste é feita de acordo com o seguinte algoritmo:

1. Para cada gene i calcule-se os *fold-change*¹² para todas as combinações entre as réplicas do grupo experimental e do grupo controlo :
 $y_{i1}/x_{i1}, y_{i2}/x_{i1}, y_{i3}/x_{i1}, \dots, y_{in_1}/x_{in_0} \Rightarrow n_1 \times n_0 = h$ combinações.
2. Para cada combinação, ordenem-se os genes segundo os *fold change*. Se o objetivo for selecionar genes com regulação positiva, ao gene com maior valor *fold-change* de cada lista atribua-se a ordem 1. Se o objetivo for selecionar genes com regulação negativa, ao menor valor *fold-change* de cada lista atribua-se a ordem 1.

¹²Observe-se que o *fold-change* aqui utilizado não é o método FC (que traduz uma razão de médias), o que se calcula é a razão do nível de expressão do gene i no *array* experimental e no *array* controlo.

3. Para cada gene i calcule-se o *rank product* de acordo com $RP_i = (\prod_{l=1}^h r_{il})^{1/h}$, onde r_{il} é a ordem do gene i da l -ésima combinação e $l = 1, \dots, h$.
4. De forma independente, realizem-se p permutações dos níveis de expressão dos genes em cada réplica, e para cada gene i repitam-se os passos 1 a 3, RP_{if}^* irá representar o *rank product* do gene i obtido na permutação f , $f = 1, \dots, p$.
5. Comparem-se os RP_{if}^* obtidos nas p permutações com o RP_i obtido com os dados originais, e conte-se quantas vezes os RP_{if}^* são inferiores a RP_i e seja este valor representado por c_i para cada gene i .
6. Calcule-se o valor esperado do *rank product* para cada gene i de acordo com $e_i = c_i/p$.
7. O valor- p associado a cada gene é dado por $\frac{e_i}{g \times p}$, onde g é o número total de genes.

Valores- p inferiores a um determinado ponto de corte conduzem à lista final de genes com regulação positiva (ou negativa). Este método tem poucos pressupostos, apenas que a variância seja igual para todos os genes.

A biblioteca `rankprod` do projeto *Bioconductor* (Hong e Wittner, 2007) implementa este método. A plataforma *on-line* `RankProdlt` através do sítio da internet <http://strep-microarray.sbs.surrey.ac.uk/RankProducts> oferece a possibilidade de se aplicar este método sem a necessidade de se editarem linhas de comando (Laing e Smith, 2010).

O método RP pode também ser estendido à meta-análise, isto é, oferece uma forma de ultrapassar a heterogeneidade existente entre as várias bases de dados obtidas de diferentes estudos (Hong e Breitling, 2007).

Métodos baseados na estatística- t

Estatística- t de Welch

A estatística- t de Welch (Welch, 1947) é outro método muito utilizado (2.9) quando se pretende comparar duas condições experimentais. A diferença em relação à estatística- t , é que esta não considera variâncias iguais para as

duas condições experimentais. A estatística-*t* de Welch é dada por:

$$\text{TW}_i = \frac{\bar{Y}_{i\cdot} - \bar{X}_{i\cdot}}{\sqrt{\frac{S_{X_{i\cdot}}^2}{n_0} + \frac{S_{Y_{i\cdot}}^2}{n_1}}}, \quad (2.9)$$

onde $\bar{X}_{i\cdot}$ e $\bar{Y}_{i\cdot}$ são as v.a. que representam as médias dos níveis de expressão nas condições controlo e experimental, para o gene i , $S_{X_{i\cdot}}^2$ e $S_{Y_{i\cdot}}^2$ são as variâncias dos níveis de expressão nos grupos controlo e experimental e n_0 e n_1 são o respetivo número de réplicas.

Esta estatística de teste nem sempre é a mais adequada para as experiências de *microarrays*. Por exemplo, no caso em que as diferenças das médias dos níveis de expressão são elevadas em comparação com as variâncias, pode admitir-se que o gene tem expressão diferencial. Por outro lado, mesmo que as diferenças das médias das expressões sejam elevadas, mas as variâncias também, o gene não será identificado como significativo.

Significance Analysis of Microarrays (SAM)

Chu *et al.* (2001) desenvolveram o popular método SAM (*Significance Analysis of Microarrays*), baseado numa versão modificada da estatística-*t* (2.10). A particularidade deste método é que tem em consideração as flutuações aleatórias que possam ocorrer nas amostras. Essas flutuações estão relacionadas com a especificidade de cada gene. Assim, cada gene terá um valor da estatística de teste (2.10) “personalizado”,

$$\text{SAM}_i(b) = \frac{\bar{Y}_{i\cdot} - \bar{X}_{i\cdot}}{S_{i\cdot} + b}, \quad (2.10)$$

onde $S_{i\cdot} = \sqrt{\frac{(n_0-1)S_{X_{i\cdot}}^2 + (n_1-1)S_{Y_{i\cdot}}^2}{n_0+n_1-2} \left(\frac{1}{n_0} + \frac{1}{n_1} \right)}$ e b é uma constante positiva.

Para assegurar que a variância de SAM_i seja independente dos níveis de expressão, acrescentou-se a constante b , também designada de constante *fudge*. Quando $b = 0$ obtém-se a estatística-*t*. A forma como Chu *et al.* (2001) propõem o cálculo desta constante não é imediata:

- Seja s^p o p -ésimo percentil do vetor dos desvios padrão s_i , $i = 1, \dots, g$, e $\text{sam}_i^p = \frac{\bar{y}_i - \bar{x}_i}{s_i + s^p}$;
- Calculem-se os 100 percentis do vetor dos desvios padrão s_i : $q_1 < q_2 < \dots < q_{100}$;
- Para $p \in \{0, 1, 2, \dots, 100\}$:
 - calcule-se $v_j = \text{mad}(\text{sam}_i^p : s_i \in [q_j; q_{j+1}])$, onde $j = 0, \dots, 99$ e mad é a mediana dos desvios absolutos da mediana dividido por 0.64;
 - calcule-se o coeficiente de variação dos v_j para cada p : $\text{cv}(p)$;
- Seja $\hat{p} = \text{argmin}[\text{cv}(p)]$ e \hat{b} será igual ao percentil \hat{p} dos valores de s_i .

Resumidamente, este método resulta de um processo composto por três componentes, a primeira relativa ao cálculo da estatística- t ajustada, a segunda da aproximação da distribuição baseada num teste de permutações e a terceira do controlo da FDR¹³. Para mais detalhes o manual dos autores Chu *et al.* (2001) é um bom documento de consulta.

Para além do *software SAM*, executável em Excel, que pode ser obtido, para fins não comerciais, através do sítio <http://www-stat-class.stanford.edu/tibs/clickwrap/sam.html>, é também possível utilizar as bibliotecas *samr*, *siggenes* e *st* do *Bioconductor*.

Estatística *B* e *Moderated t-statistic (modT)*

Lönnstedt e Speed (2002) propuseram uma estatística baseada numa abordagem bayesiana empírica para a análise de *microarrays* de dois canais. Lin *et al.* (2003) desenvolveram a metodologia de Lönnstedt e Speed (2002) para experiências de *microarrays* de um canal.

Lin *et al.* (2003) admitem que as variáveis aleatórias que representam o logaritmo dos níveis de expressão de cada gene em cada *array* são independentes e normalmente distribuídas:

¹³Do inglês *false discovery rate*.

2.3. Métodos para a seleção de genes DE

$$\begin{aligned} X_{ij} | \mu_i, \sigma_i &\sim N(\mu_i, \sigma_i^2), \\ Y_{ik} | \mu_i + \Delta_i, \sigma_i &\sim N(\mu_i + \Delta_i, \sigma_i^2). \end{aligned}$$

De forma equivalente à proposta de Lönnstedt e Speed (2002), Lin *et al.* (2003) consideraram a gama invertida como distribuição *a priori* para σ_i^2 . Para a distribuição *a priori* de Δ_i consideraram a distribuição normal e uma distribuição *a priori* não-informativa para $\mu_i | \tau_i$.

Seja $\tau_i = \frac{na}{2\sigma_i^2}$ (onde n representa o número total de réplicas $n = n_0 + n_1$), $\nu > 0$, $a > 0$ e $c > 0$ (a e c são parâmetros de escala) e considere-se:

$$\begin{aligned} \tau_i &\sim \Gamma(\nu, 1), \\ \mu_i | \tau_i &\sim 1, \\ \Delta_i | \tau_i &\begin{cases} = 0 & \text{se } I_i = 0 \\ \sim N(0, c \frac{na}{2\tau_i}) & \text{se } I_i = 1 \end{cases}. \end{aligned}$$

A variável I_i indica se os níveis de expressão do gene i são iguais nos grupos controlo e experimental ($\Delta_i = 0$) ou diferentes ($\Delta_i \neq 0$):

$$I_i = \begin{cases} 0 & \text{se } \Delta_i = 0 \\ 1 & \text{se } \Delta_i \neq 0 \end{cases},$$

para $i = 1, \dots, g$.

A estatística B representa os logaritmos das chances *a posteriori* e é dada por:

$$B_i = \log \left(\frac{p}{1-p} \right) \left(\frac{n}{n_0 n_1 c + n} \right)^{\frac{1}{2}} \left[\frac{na + (n-2)s_i^2 + \frac{n_0 n_1}{n} (\bar{Y}_{i.} - \bar{X}_{i.})^2}{na + (n-2)s_i^2 + \frac{n_0 n_1}{n_0 n_1 c + n} (\bar{Y}_{i.} - \bar{X}_{i.})^2} \right]^{\nu + \frac{n-1}{2}}, \quad (2.11)$$

onde p representa a proporção de genes diferencialmente expressos *a priori*.

Os parâmetros a e ν são estimados pelo método dos momentos. Os genes são ordenados por ordem decrescente da estatística B e os primeiros são

considerados genes potencialmente diferencialmente expressos.

Smyth (2005) propôs a *moderated t-statistic* (modT), uma versão da estatística-*t* que tem a vantagem em relação à estatística *B* de não depender do conhecimento da proporção de genes DE *a priori*.

A estatística modT é dada por:

$$t_i = \frac{\hat{\beta}_i}{\tilde{s}_i \sqrt{v_i}}, \quad (2.12)$$

e segue uma distribuição *t* de *Student* com $d_f = d_0 + d_i$ graus de liberdade.

Admite-se que a diferença dos logaritmos das médias dos níveis de expressão do gene i , $\hat{\beta}_i = \log(\bar{Y}_{i.}) - \log(\bar{X}_{i.})$, tem a seguinte distribuição de amostragem:

$$\hat{\beta}_i | \beta_i, \sigma_i^2 \sim N(\beta_i, v_i \sigma_i^2),$$

onde $v_i = \frac{1}{n_0} + \frac{1}{n_1}$.

Dada a variância σ_i^2 , a variância amostral do nível de expressão para cada gene i , segue uma distribuição qui-quadrado com d_i graus de liberdade:

$$s_i^2 | \sigma_i^2 \sim \frac{\sigma_i^2}{d_i} \chi_{d_i}^2,$$

Admite-se *a priori* que:

$$1/\sigma_i^2 \sim \frac{1}{d_0 s_{0i}^2} \chi_{d_0}^2,$$

onde d_0 e s_{0i} são hiperparâmetros.

A variância *a posteriori* e o número de graus de liberdade *a posteriori* são então dados por:

$$\begin{aligned} \tilde{s}_i^2 &= \frac{d_0 s_{0i}^2 + d_i s_i^2}{d_0 + d_i}, \\ d_f &= d_0 + d_i. \end{aligned}$$

A função `modt.stat` da biblioteca `st` em conjunto com a biblioteca `limma` do *Bioconductor* implementa este método.

Intensity-based moderated t-statistic (ibmT)

Sartor *et al.* (2006) propõem uma versão modificada de modT, *Intensity-based moderated t-statistic*. Utilizam um modelo bayesiano hierárquico normal completamente dependente dos dados, adotando uma filosofia empírica bayesiana para estimar os hiperparâmetros, e consequentemente não requer nenhuma especificação para os parâmetros livres. Este modelo tem como objetivo controlar a relação entre a variância e a intensidade dos sinais. Quando a relação é fraca ou quando não existe, este modelo é reduzido ao modT.

Este modelo é implementado através de código R disponível pelos autores no sítio <http://eh3.uc.edu/ibmt> em conjunto com a biblioteca `limma` do *Bioconductor*.

2.3.3 Algumas considerações finais

Com a tecnologia de *microarrays* têm-se desenvolvido vários métodos estatísticos para selecionar genes DE. O método FC é geralmente utilizado como primeira abordagem, mas pode conduzir a conclusões erradas devido não só à incerteza que é induzida pelo quociente entre dois níveis de intensidade, mas também porque implicitamente admite que a variância é constante para todos os genes. Existem várias variantes da estatística-*t*, mas o problema é que nestas experiências são aplicados vários testes de hipóteses para um número muito reduzido de réplicas, conduzindo a estatísticas muitos instáveis. Por exemplo, valores elevados da estatística de teste podem ocorrer devido à presença de variâncias muito pequenas, mesmo quando a diferença das médias dos níveis de expressão é muito pequena. As desvantagens do método FC e estatística-*t* têm sido apontadas por vários autores. Apesar da falha, estas estatísticas têm sido amplamente utilizadas na prática, incluindo a dupla filtragem considerando ambas as estatísticas, nomeadamente na construção do gráfico *Volcano plot* (Li, 2011).

Os métodos bayesianos aqui descritos têm a particularidade de serem apenas utilizados para estimar as variâncias *a posteriori* dos genes nos dois grupos em análise, que posteriormente são utilizadas numa abordagem clássica, em particular no cálculo de valores-*p*. Esta abordagem parece controversa, na medida em que é um híbrido entre duas correntes com pressupostos muito distintos.

Como já anteriormente foi referido, a tecnologia de *microarrays* permite a análise de milhares de genes simultaneamente e, tipicamente, para cada gene é aplicado um teste de hipóteses. Esta situação levanta o problema dos testes múltiplos. Existem várias formas de lidar com o problema de testes múltiplos.

No entanto a análise de testes múltiplos tem algumas desvantagens, e um exemplo é a dependência do número total de genes a ser testado. Um gene pode ter uma elevada expressão num determinado estudo clínico, desde que outros genes não sejam avaliados. Não obstante, e não sendo o objetivo deste trabalho na análise do problema de testes múltiplos, aconselha-se o uso de métodos simples de modo a reduzir o elevado número de potenciais candidatos num número de genes mais reduzido e qualificado para as subsequentes análises.

Capítulo 3

Metodologia ROC na análise de dados de *microarrays*

3.1 Introdução

A análise *receiver operating characteristic* (ROC) baseia-se na Teoria de Deteção de Sinais, desenvolvida durante a 2^a Guerra Mundial. Um operador de radar tinha de decidir se um ponto no ecrã representava um inimigo, um aliado ou simplesmente ruído (Metz e Pan, 1999). Os operadores eram denominados por *Radar Receiver Operators*. A Teoria de Deteção de Sinais tinha por objetivo avaliar a performance dos operadores.

O conceito de *Receiver Operating Characteristic Curves* foi introduzido na medicina por Lusted (1968) nos finais da década de 60. A principal aplicação da metodologia ROC na medicina é a análise da capacidade discriminativa de testes de diagnóstico (Silva, 2004; Silva-Fortes, 2011).

Uma abordagem muito comum para identificar genes relacionados com o cancro, baseia-se em comparar perfis dos níveis de expressão entre amostras normais e com cancro, e na seleção de genes que apresentem níveis de expressão elevados na amostra com cancro (Pepe *et al.*, 2003; Baker, 2003). Esta abordagem ignora a heterogeneidade existente nas amostras tumorais e não é adequada para encontrar genes com elevados níveis de expressão num determinado subgrupo de pacientes na população (Li *et al.*, 2007).

Várias técnicas estatísticas têm sido propostas para selecionar genes DE, entre as quais se encontram as curvas *Receiver Operating Characteristic* (ROC) (Silva-Fortes *et al.*, 2006). Numa experiência de *microarrays* a seleção de genes DE a partir da metodologia ROC pode ser realizada no

contexto da avaliação de quão afastadas se encontram as distribuições subjacentes às variáveis que representam os níveis de expressão dos genes em cada um dos grupos.

Atualmente, na pesquisa oncológica, a tecnologia de *microarrays* é amplamente utilizada. Uma das suas principais aplicações é a identificação de grupos de genes associados ao desenvolvimento de um determinado tipo de cancro. Diferentes estudos, baseados no perfil de expressão de genes em amostras tumorais, têm revelado a grande heterogeneidade transcripcional do cancro e têm permitido a classificação de novas subclasses clínicas e biológicas da doença (Perez-Diez *et al.*, 2007).

Na análise de dados de *microarrays* a metodologia ROC tem sido, particularmente, muito utilizada na comparação da performance de vários métodos, como por exemplo na comparação da performance de métodos na seleção de genes DE. No entanto, medidas baseadas na metodologia ROC, como por exemplo a AUC, podem ser utilizadas para selecionar genes. Este capítulo aborda a metodologia ROC na seleção de genes. No entanto, no capítulo 5, a metodologia ROC será também utilizada como uma ferramenta para avaliar a capacidade discriminativa dos métodos abordados neste trabalho.

Este capítulo tem como objetivo descrever as principais características e propriedades da curva ROC, métodos de estimação não-paramétricos da curva ROC e da área abaixo da curva ROC (AUC¹). Na secção 3.5 apresentam-se métodos baseados na metodologia ROC para a seleção de genes DE.

No contexto deste trabalho pretende-se aplicar a metodologia ROC na análise de dados de *microarrays*, de modo a selecionar genes diferencialmente expressos que dificilmente são identificados pelos métodos vulgarmente utilizados, alguns deles descritos na secção 2.3.

3.2 Definição e Propriedades da curva ROC

Considere-se que $X_1, X_2, X_3, \dots, X_{n_0}$ e Y_1, Y_2, \dots, Y_{n_1} são duas amostras aleatórias e independentes que representam os níveis de expressão de um

¹Do inglês *area under the curve*.

3.2. Definição e Propriedades da curva ROC

gene na população controlo e na população experimental, respetivamente. Seja F_X a função de distribuição associada a X_i e F_1 a função de distribuição associada a Y_j , e sem perda de generalidade para qualquer ponto de corte $c \in \mathbb{R}$, $F_X(c) > F_Y(c)$, sendo f_X e f_Y as correspondentes funções de densidade de probabilidade. A sensibilidade e a especificidade são dadas por:

$$\begin{aligned} S(c) &= 1 - F_Y(c) = q(c), \\ E(c) &= F_X(c) = p(c). \end{aligned}$$

Teoricamente a curva ROC φ é uma função definida da seguinte forma:

$$\varphi : [0, 1] \longrightarrow [0, 1]$$

$$\varphi(p) = 1 - F_Y(F_X^{-1}(1-p)), \quad (3.1)$$

onde F_X^{-1} é a função inversa de F_X definida por $F_X^{-1}(1-p) = \inf\{x \in W(F_X) : F_X(x) \geq 1-p\}$, e $W(F_X) = \{x \in \mathbb{R} : 0 < F_X(x) < 1\}$ é o suporte de F_X .

A curva ROC resulta da representação gráfica de todos os pares $(1-p(c), q(c))$ no plano unitário fazendo variar os pontos de corte no suporte da variável de decisão (Figura 3.1).

A curva ROC é uma função monótona crescente, e se para qualquer ponto de corte c se tiver $1 - F_X(c) = 1 - F_Y(c)$, o modelo de classificação é não informativo, *i.e.*, a curva ROC coincide com a diagonal principal (ou linha de referência) do plano unitário. Um modelo de classificação perfeito separa completamente os casos correspondentes à população experimental dos casos pertencentes à população controlo, *i.e.*, para um determinado ponto de corte c , ter-se-á $1 - F_Y(c) = 1$ e $1 - F_X(c) = 0$, que corresponde a uma curva que passa pelo vértice $(0,1)$ do plano unitário.

A curva ROC é invariante em relação a transformações monótonas da escala de X e Y , e se X for estocasticamente inferior a Y , isto é, $\forall c : F_X(c) \geq F_Y(c)$, a curva ROC estará definida acima da diagonal positiva do plano unitário, ou seja $\int_0^1 \varphi(p) dp \geq 0.5$.

No contexto da análise da expressão diferencial, dizer que um gene é diferencialmente expresso é equivalente a dizer que a distribuição dos níveis

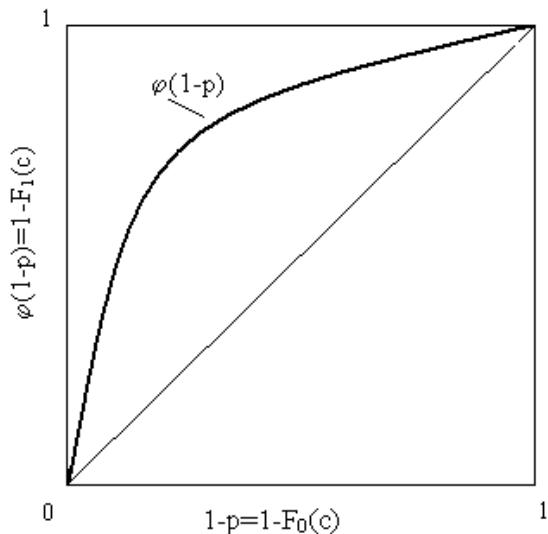


Figura 3.1: Exemplo de uma Curva ROC.

de expressão na população controlo é diferente da distribuição dos níveis de expressão na população experimental. Assim, a curva ROC pode ser usada para avaliar a separação das distribuições dos níveis de expressão entre os grupos controlo e experimental, desde que $\forall c : F_X(c) \geq F_Y(c)$.

3.3 AUC e curvas ROC degeneradas

A área abaixo da curva ROC é o índice mais comumente utilizado na metodologia ROC para avaliar o poder discriminativo de um modelo de classificação. Bamber (1975) demonstra que a AUC é obtida integrando a curva ROC no seu domínio, ou seja, $\int_0^1 \varphi(p) dp = P(X < Y)$. A AUC pode admitir valores entre 0.5 e 1, onde valores da AUC próximos de 1 indicam que o modelo de classificação tem um elevado poder discriminativo.

Wolf e Hogg (1971) recomendam este índice como uma medida que avalia diferenças entre duas distribuições. Para Pepe (2000) e Parodi *et al.* (2002) a AUC é uma medida não-paramétrica da distância entre as distribuições da variável de decisão nas duas classes. Esta caracterização da AUC conduziu

a uma reflexão, pois de acordo com a definição de *métrica* (ou distância), a AUC em rigor não pode ser considerada uma distância, uma vez que duas das propriedades não se verificam, tal como se demonstra a seguir.

Definição de métrica: Dado um conjunto S , uma métrica em S é uma função $d : S \times S \longrightarrow \mathbb{R}$ que satisfaz as seguintes propriedades:

1. $d(x, y) \geq 0, \forall x, y \in S;$
2. $d(x, y) = 0$ sse $x = y;$
3. $d(x, y) = d(y, x), \forall x, y \in S;$
4. $d(x, y) \leq d(x, z) + d(z, y), \forall x, y, z \in S.$

A AUC não satisfaz a propriedade 2, uma vez que o valor mínimo da AUC é igual a 0.5, correspondendo à sobreposição das distribuições das duas classes. Em relação à propriedade 3, a igualdade não se verifica, mas sim $d(x, y) = 1 - d(y, x)$. Conclui-se então, contrariando alguns autores (Pepe, 2003; Parodi *et al.*, 2008), que a AUC não é uma distância.

Como já foi referido, as curvas ROC são construídas com base na sensibilidade e especificidade de um sistema que classifique os dados em duas classes mutuamente exclusivas para todos os pontos de corte possíveis da variável de decisão. Estas probabilidades dependem da regra de classificação, sendo que na análise ROC tradicional um valor elevado da variável de decisão corresponde à presença do artefacto de interesse, *i.e.*, a variável que representa os níveis de expressão dos genes na população controlo, X , é estocasticamente inferior à variável que representa os níveis de expressão na população experimental, Y , ou seja, $F_X(c) > F_Y(c)$. No entanto, se se aplicar as curvas ROC para todos os genes de uma experiência de *microarrays* considerando a mesma regra de classificação, estas podem apresentar-se abaixo da diagonal positiva do plano unitário (Figura 3.2 E). Esta situação ocorre porque a regra de classificação não será a mesma para todos os genes numa experiência de *microarrays*, uma vez que alguns podem ter regulação positiva, $F_X(c) > F_Y(c)$, e outros regulação negativa, $F_X(c) < F_Y(c)$.

Todavia, não é apenas esta a situação que conduz a curvas ROC degeneradas². A presença de distribuições bimodais numa das condições experimentais e com médias semelhantes nos dois grupos, conduz a curvas ROC sigmoidais que cruzam a linha de referência do plano unitário (Figuras 3.2 C–D). No contexto deste trabalho, genes com os níveis de expressão que apresentem distribuições como as representadas nas Figuras 3.2 C–D serão designados de genes mistos. Esta designação é uma alternativa ao que é habitualmente a designação de um gene DE, onde particularmente, ou é um gene com regulação positiva ou com regulação negativa. Logo, os genes mistos, pelas características que apresentam, terão uma distribuição mista. A existência deste tipo de distribuições numa das condições (ou em ambas) pode indicar a presença de subclasses desconhecidas com diferentes níveis de expressão. Consequentemente, a identificação destas subclasses pode fornecer informações úteis sobre os mecanismos biológicos subjacentes a condições fisiológicas ou patológicas.

²Curvas que se apresentem abaixo ou que cruzem a diagonal principal do plano unitário designam-se de curvas ROC degeneradas (Figuras 3.2 C–E).

3.3. AUC e curvas ROC degeneradas

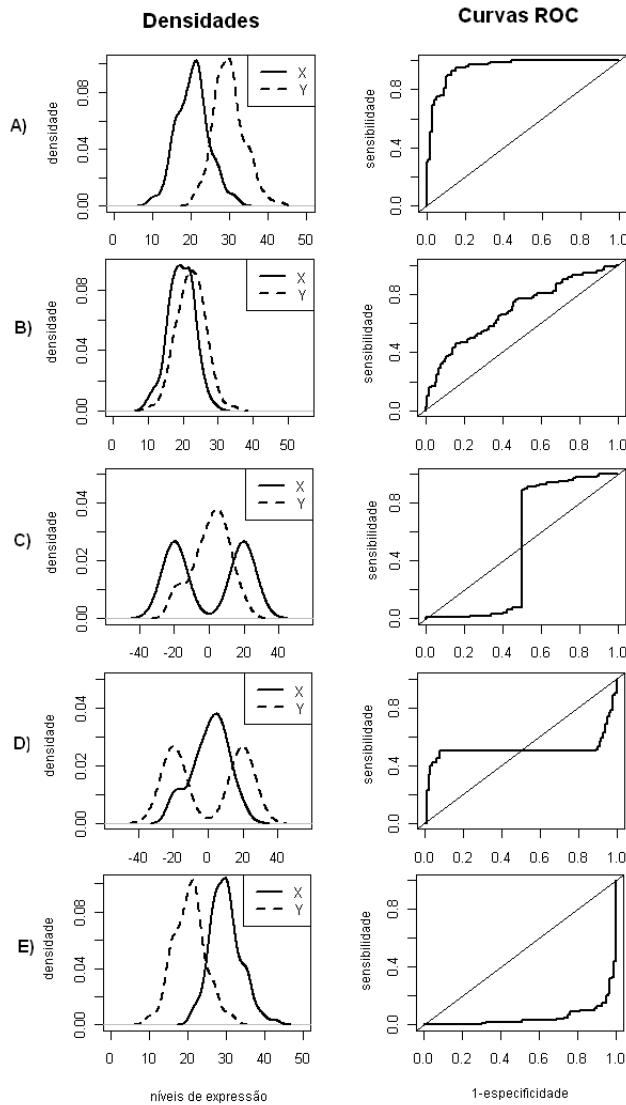


Figura 3.2: Densidades dos níveis de expressão de duas populações e respectivas curvas ROC empíricas. X representa níveis de expressão no grupo controlo e Y representa níveis de expressão no grupo experimental. Considerou-se a mesma regra de classificação para a estimação das curvas ROC. As densidades foram estimadas através do estimador de núcleo a partir de duas amostras, de dimensão 100, simuladas de duas normais. A) $X \sim N(20, 4)$, $Y \sim N(30, 4)$; B) $X \sim N(20, 4)$, $Y \sim N(22, 4)$; C) $X \sim 0.5N(-20, 2) + 0.5N(20, 2)$, $Y \sim N(0, 11)$; D) $X \sim N(0, 11)$, $Y \sim 0.5N(-20, 2) + 0.5N(20, 2)$; E) $X \sim N(30, 4)$, $Y \sim N(20, 4)$.

Sendo a AUC um índice que avalia o quão diferentes são as distribuições

dos grupos em análise, se o objetivo for selecionar genes DE, ao manter-se a mesma regra de classificação para todos os genes, os que apresentarem regulação negativa terão valores da AUC inferiores a 0.5, e os genes mistos apresentarão valores da AUC em torno de 0.5.

O modelo binormal próprio (Metz e Pan, 1999) e binormal contaminado (Dorfman *et al.*, 2000) são exemplos de métodos que obrigam as curvas ROC a apresentarem-se acima da diagonal principal e, consequentemente, os valores da AUC serão superiores ou iguais a 0.5. Mas ao utilizar-se esta abordagem perder-se-á a informação do tipo de regulação que os genes selecionados apresentam.

Assim, no contexto deste trabalho as curvas ROC degeneradas revelam-se úteis na deteção dos vários tipos de expressão diferencial. Nesse sentido, considera-se a mesma regra de classificação para todos os genes, *i.e.*, valores elevados dos níveis de expressão estão relacionados com a presença de genes com regulação positiva, e consequentemente os valores da AUC podem variar entre 0 e 1.

No entanto, a análise da AUC para selecionar genes mistos revela-se insuficiente, uma vez que os genes mistos (Figuras 3.2 C-D) não se distinguem dos genes não DE (Figura 3.2 B), pois os valores da AUC em ambas as situações são próximos de 0.5.

Uma vez que a análise da AUC mostra ser insuficiente para selecionar genes mistos, propõe-se a análise conjunta da AUC e do coeficiente de sobreposição entre duas distribuições (OVL³) (descrito com maior detalhe no capítulo seguinte) para selecionar simultaneamente genes DE e genes mistos.

3.3.1 Métodos de estimação não-paramétricos da curva ROC e da AUC

Considerou-se uma abordagem não-paramétrica para a estimação da AUC, uma vez que não necessita de pressupostos distribucionais, e numa experiência de *microarrays* usualmente o número de réplicas é pequeno e os dados não seguem uma distribuição normal (Hardin e Wilson, 2007). A abordagem não-paramétrica tem como desvantagem a perda de eficiência, que é contrabalançada com a redução do risco em interpretar resultados tendo por base especificações incorretas.

³Do inglês *coefficient of overlapping*.

Consideraram-se os métodos empírico e do núcleo para estimar a curva ROC e a AUC que a seguir se descrevem. Para aliviar a notação, vai considerar-se que os níveis de expressão são referentes apenas a um único gene.

Estimação empírica

Sejam X_1, \dots, X_{n_0} e Y_1, \dots, Y_{n_1} duas amostras aleatórias e independentes e que representam níveis de expressão de um determinado gene na amostra controlo e experimental, e as funções de distribuição empíricas de F_X e F_Y são dadas, respetivamente, por:

$$\hat{F}_X(t) = \frac{1}{n_0} \sum_{j=1}^{n_0} I[X_j \leq t], \quad (3.2)$$

$$\hat{F}_Y(t) = \frac{1}{n_1} \sum_{k=1}^{n_1} I[Y_k \leq t], \quad (3.3)$$

onde I é a função indicatriz.

Na prática para estimar a curva ROC empiricamente usam-se como pontos de corte os valores observados de X_j e Y_k ordenados por ordem crescente. Assim, para cada ponto de corte c , as estimativas empíricas da sensibilidade e especificidade são dadas por:

$$\hat{S}(c) = \frac{1}{n_1} \sum_{k=1}^{n_1} I[Y_k \geq c] = 1 - \hat{F}_Y(c) = \hat{q}(c), \quad (3.4)$$

$$\hat{E}(c) = \frac{1}{n_0} \sum_{j=1}^{n_0} I[X_j \leq c] = \hat{F}_X(c) = \hat{p}(c). \quad (3.5)$$

A estimativa empírica da curva é dada ROC:

$$\widehat{\varphi}(\hat{p}) = 1 - \widehat{F}_Y(\widehat{F}_X^{-1}(1 - \hat{p})). \quad (3.6)$$

Tecnicamente a curva ROC empírica é definida por todos os pares $(1 - \hat{E}, \hat{S})$ para todos os valores de c_i , $i = 1, \dots, (n_0 + n_1)$, que variam na amostra combinada dos valores dos níveis de expressão nos grupos controlo e experimental. Pontos adjacentes (sem empates na amostra combinada) são ligados por segmentos de reta verticais ou horizontais, que resultam numa curva em escada. As linhas diagonais que possam ocorrer correspondem a empates dos valores dos níveis de expressão nos dois grupos. Geralmente os pontos de corte c_i não são indicados no gráfico, no entanto cada ponto no gráfico corresponde ao intervalo do tipo $[c_i; c_{i+1}]$, cada segmento pode ser associado a um ponto de corte c_i . Se o valor observado corresponder a um array do grupo controlo, o segmento é horizontal; se o valor for de um array do grupo experimental, o segmento é vertical.

A AUC empírica corresponde à estatística de Mann-Whitney (McNeil e Hanley, 1984):

$$\widehat{\text{AUC}} = \frac{1}{n_0 n_1} \sum_{j=1}^{n_0} \sum_{k=1}^{n_1} \left(I[X_j < Y_k] + \frac{1}{2} I[X_j = Y_k] \right). \quad (3.7)$$

Na prática, o valor observado desta estatística coincide com o valor da AUC obtido pela regra do trapézio (Bamber, 1975):

$$\widehat{\text{AUC}}_{trap} = \frac{1}{2} \sum_{i=1}^r \left\{ (\hat{E}_i - \hat{E}_{i+1}) (\hat{S}_{i+1} - \hat{S}_i) + 2 (\hat{E}_i - \hat{E}_{i+1}) (\hat{S}_i) \right\}, \quad (3.8)$$

onde r representa o número total de pontos de corte.

Estimação pelo método do núcleo

Vários autores discutiram o refinamento da abordagem não-paramétrica de modo a originar curvas ROC suaves (Zou *et al.*, 1997; Goodard e Hinberg, 1990). De entre as várias metodologias não-paramétricas, um importante método para a estimação de funções densidade de probabilidade é a do estimador de núcleo. O desempenho do estimador de núcleo depende essencialmente da escolha do parâmetro de suavidade ou janela. Este parâmetro, geralmente denotado por h , determina o grau de suavização a ser feita.

Dada uma amostra aleatória X_1, \dots, X_n de uma distribuição univariada contínua com f.d.p. f desconhecida, um estimador de núcleo é dado por:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right), \forall x \in S, \quad (3.9)$$

onde K é a função núcleo, h a janela e S o suporte (Rosenblatt, 1956).

Existem vários tipos de estimadores do núcleo, sendo o representado na expressão (3.9) conhecido na literatura como estimador de núcleo fixo ou estimador de núcleo global, por considerar a janela h constante para todo o ponto x .

Um aspeto-chave do estimador de núcleo está associado ao facto do seu desempenho, em termos do erro, não depender diretamente da forma funcional do núcleo. Diversos trabalhos demonstram que a qualidade do estimador de núcleo depende essencialmente da escolha da janela h (Silverman, 1986).

Neste trabalho a função núcleo estará restrita ao núcleo gaussiano:
 $(2\pi)^{-\frac{1}{2}} \exp(-\frac{1}{2}u^2)$.

Em geral, a escolha da janela h está relacionada com a otimização de alguma medida de desempenho, isto é, a escolha de h está vinculada à minimização de alguma medida de exatidão, como por exemplo o erro quadrático médio integrado (EQMI). Ao longo dos anos, vários estudos foram feitos em busca de metodologias que estimassem “automaticamente” a janela ótima através da minimização de alguma medida. Esses métodos, designados como

métodos automáticos de seleção de janela, são baseados na ideia de que a quantidade ótima de suavização deve depender unicamente dos dados.

De acordo com Silverman (1986) a escolha ótima da janela h , quando a distribuição da população é normal e a função núcleo K é gaussiana, é dada por:

$$h = \left(\frac{4}{3}\right)^{\frac{1}{5}} s n^{-\frac{1}{5}}, \quad (3.10)$$

onde s é o desvio padrão empírico.

No entanto, se a população for multimodal⁴, esta janela pode suavizar em demasia o histograma. Assim, Silverman (1986) propõe para a janela h :

$$h = \left(\frac{4}{3}\right)^{\frac{1}{5}} \min\left(s, \frac{R}{1.34}\right) n^{-\frac{1}{5}}, \quad (3.11)$$

onde R é a amplitude interquartil.

Esta regra é atrativa em experiências de *microarrays*, uma vez que a escolha da janela ótima é feita de uma forma automática. A função **density** do R usa, por omissão, a função núcleo gaussiana e a escolha de h é feita de acordo com (3.11).

Sejam \tilde{f}_X e \tilde{f}_Y os estimadores do núcleo de f_X e f_Y , dados por:

$$\tilde{f}_X(t) = \frac{1}{n_0 h_0} \sum_{i=1}^{n_0} K\left(\frac{t - X_i}{h_0}\right), \quad (3.12)$$

$$\tilde{f}_Y(t) = \frac{1}{n_1 h_1} \sum_{j=1}^{n_1} K\left(\frac{t - Y_j}{h_1}\right). \quad (3.13)$$

Considerando um núcleo gaussiano, os estimadores da sensibilidade e da especificidade são dados por:

⁴Uma distribuição multimodal no contexto deste trabalho, é uma distribuição que apresenta vários picos não necessariamente com a mesma moda.

$$\tilde{S}(c) = 1 - \frac{1}{n_1} \sum_{j=1}^{n_1} \Phi \left(\frac{c - Y_i}{h_1} \right) = 1 - \tilde{F}_Y(c), \quad (3.14)$$

$$\tilde{E}(c) = \frac{1}{n_0} \sum_{i=1}^{n_0} \Phi \left(\frac{c - X_i}{h_0} \right) = \tilde{F}_X(c), \quad (3.15)$$

onde Φ representa a função de distribuição da normal padrão.

A curva ROC estimada pelo método de núcleo é uma curva suave, dada por:

$$\tilde{\varphi}(p) = 1 - \tilde{F}_Y(\tilde{F}_X^{-1}(1-p)), \quad 0 < p < 1. \quad (3.16)$$

Lloyd (1997) demonstrou que, considerando um núcleo gaussiano, o estimador pelo método do núcleo da AUC é dado por:

$$\widetilde{\text{AUC}} = \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} \Phi \left(\frac{Y_j - X_i}{\sqrt{h_0^2 + h_1^2}} \right). \quad (3.17)$$

Zou *et al.* (1997) demonstraram que, considerando o núcleo gaussiano, a janela ótima para h_0 e h_1 é obtida através de (3.11).

3.4 AUC empírica vs. AUC núcleo — comparação do viés

Lloyd e Yong (1999), demonstraram que a estimação da curva ROC pelo método do núcleo é assintoticamente superior em termos do erro quadrático médio (EQM) em relação à estimação da curva ROC pelo método empírico. Nesta secção pretende-se comparar os vieses da AUC quando estimada pelos métodos empírico e do núcleo, uma vez que no capítulo 5 são comparados os resultados usando as duas estimativas. Caso os vieses sejam significativamente diferentes considerando as duas abordagens, deve proceder-se à correção do viés.

Procedeu-se a um estudo de simulação utilizando o método de *bootstrap* de

modo a estimar o viés e o erro padrão das estimativas da AUC usando os métodos empírico e do núcleo. Na pasta “Capítulo 3” do CD, encontra-se o ficheiro `bootauc.R` com o código R que implementa a análise desenvolvida.

De modo a analisar situações onde os valores da AUC estão mais próximos de 1 ou de 0.5, simularam-se amostras das seguintes distribuições: $N(20, 4)$ (grupo experimental) e $N(10, 2)$ (grupo controlo); e $N(20, 4)$ (grupo experimental) e $N(18, 4)$ (grupo controlo). Simularam-se amostras de dimensões 15, 30, 100 e 500 respetivamente e consideraram-se $B=1000$ réplicas *bootstrap* para cada caso.

O valor exato da AUC quando as amostras provêm de populações normais e independentes é obtido a partir da expressão (Faraggi e Raiser, 2002):

$$\text{AUC} = \Phi \left(\frac{\mu_y - \mu_x}{\sqrt{\sigma_x^2 + \sigma_y^2}} \right), \quad (3.18)$$

onde Φ representa a função de distribuição normal padrão. No caso particular em que o par é $N(20, 4)$ e $N(10, 2)$ o valor exato da AUC é 0.987 e quando $N(20, 4)$ e $N(18, 4)$ o valor exato da AUC é 0.673.

A estimativa *bootstrap* da $\widehat{\text{AUC}}$ é dada por:

$$\widehat{\text{AUC}}_B = \frac{1}{1000} \sum_{i=1}^{1000} \widehat{\text{AUC}}_i^*, \quad (3.19)$$

onde $\widehat{\text{AUC}}_i^*$ é a estimativa da AUC (empírica ou do núcleo) em cada réplica *bootstrap*.

A estimativa *bootstrap* do erro padrão da $\widehat{\text{AUC}}$ é dada por:

$$\widehat{s}_e_B(\widehat{\text{AUC}}) = \sqrt{\frac{1}{999} \sum_{i=1}^{1000} (\widehat{\text{AUC}}_i^* - \widehat{\text{AUC}}_B)^2}. \quad (3.20)$$

3.4. AUC empírica vs. AUC núcleo — comparação do viés

A estimativa *bootstrap* do viés da $\widehat{\text{AUC}}$ é dado por:

$$\widehat{\text{viés}}_B(\widehat{\text{AUC}}) = \widehat{\text{AUC}}_B - \text{AUC}, \quad (3.21)$$

onde AUC corresponde ao valor exato.

Nas Tabelas 3.1 e 3.2 apresentam-se os valores das estimativas *bootstrap* da AUC obtidas pelos métodos empírico e do núcleo. O comportamento das estimativas *bootstrap* da AUC, quer da forma empírica e quer pelo método do núcleo, em relação ao valor exato, considerando as diferentes dimensões das amostras, está representado na Figura 3.3.

Tabela 3.1: Estimativas *bootstrap* da AUC considerando os métodos empírico e do núcleo para $n = 500, 100$.

		$n = 500$			$n = 100$		
		Média <i>Bootstrap</i>	Erro Padrão	Viés	Média <i>Bootstrap</i>	Erro Padrão	Viés
AUC=0.987	AUC empírica	0.985	0.004	-0.002	0.981	0.010	-0.006
	AUC núcleo	0.979	0.004	-0.009	0.969	0.009	-0.018
AUC=0.673	AUC empírica	0.698	0.017	0.026	0.629	0.041	-0.043
	AUC núcleo	0.694	0.016	0.022	0.624	0.039	-0.048

Como se pode observar, o viés *bootstrap* da AUC estimada pelos métodos empírico e do núcleo têm um comportamento aparentemente semelhante, observando-se uma maior diferenciação nos dois métodos na precisão. Verifica-se um aumento significativo do viés quando as distribuições associadas aos grupos experimental e controlo se encontram mais próximas, *i.e.*, quando o valor da AUC é mais próximo de 0.5 (linha de referência), à medida que a dimensão das amostras vai diminuindo. Esse aumento do viés não é tão evidente quando o valor da AUC é próximo da unidade. Embora se verifique a existência de viés, não existe a necessidade de se proceder à correção dos vieses para analisar resultados com base nos dois métodos de estimação.

Tabela 3.2: Estimativas *bootstrap* da AUC considerando os métodos empírico e do núcleo para $n = 30, 15$.

		$n = 30$			$n = 15$		
		Média <i>Bootstrap</i>	Erro Padrão	Viés	Média <i>Bootstrap</i>	Erro Padrão	Viés
AUC=0.987	AUC empírica	0.992	0.008	0.005	0.996	0.008	0.009
	AUC núcleo	0.997	0.012	-0.021	0.968	0.012	-0.019
AUC=0.673	AUC empírica	0.697	0.071	0.024	0.636	0.112	-0.037
	AUC núcleo	0.678	0.066	0.005	0.635	0.097	-0.038

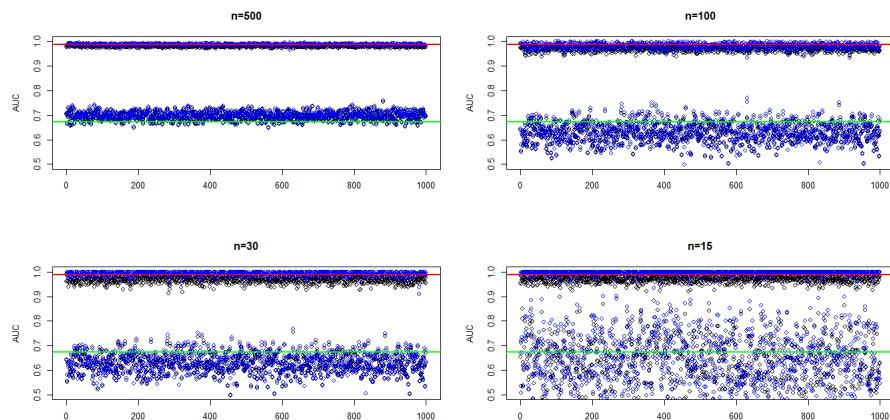


Figura 3.3: Comparação das estimativas *bootstrap* da AUC empírica e da AUC estimada pelo método do núcleo com o valor exato da AUC, para amostras de dimensões 500, 100, 30 e 15. Linhas a vermelho representam o valor exato da AUC=0.987 e linhas a verde representam o valor exato da AUC=0.673, pontos a azul representam as estimativas *bootstrap* da AUC empírica e pontos a preto representam as estimativas *bootstrap* da AUC estimada pelo método do núcleo.

De um modo geral, a exatidão e precisão das estimativas da AUC são superiores quando se considera o método empírico para estimar valores da AUC próximo da unidade, enquanto que o método do núcleo produz estimativas mais exatas e precisas quando os valores da AUC são próximas da linha de referência. A precisão é muito semelhante considerando os dois métodos de estimativa quando as dimensões das amostras são elevadas.

3.5 Métodos baseados na metodologia ROC para a seleção de genes

3.5.1 Métodos de ordenação de genes propostos por Pepe *et al.* (2003)

Pepe *et al.* (2003) utilizam medidas baseadas na metodologia ROC para obterem listas ordenadas de genes de modo a selecionarem genes com regulação positiva. O objetivo principal do estudo por eles desenvolvido foi o de identificar genes que possam ser considerados biomarcadores tumorais do cancro dos ovários.

Como habitualmente, sejam X_1, \dots, X_{n_0} e Y_1, \dots, Y_{n_1} duas amostras aleatórias e independentes, seja F_X a função de distribuição associada a X e que representa os níveis de expressão na população controlo e F_Y a função de distribuição associada a Y e que representa os níveis de expressão na população experimental.

Os genes são classificados como diferencialmente expressos de acordo com o quanto afastadas se encontram as distribuições associadas aos níveis de expressão dos grupos em análise. Numa lista ordenada por ordem descrescente de acordo com o valor de uma medida de interesse, os genes que se encontram nas primeiras posições corresponderão a genes com distribuições dos níveis de expressão completamente ou quase completamente separadas.

Assim, Pepe *et al.* (2003) propõem obter listas ordenadas de genes em função do valor da AUC (secção 3.3) ou, em alternativa, de acordo com uma das seguintes medidas de discriminação:

$$\varphi(t_0) = 1 - F_Y(F_X^{-1}(1 - t_0)), \quad (3.22)$$

$$\text{pAUC}(t_0) = \int_0^{t_0} \varphi(t) dt, \quad (3.23)$$

onde t_0 representa a proporção de falsos positivos (1-especificidade) definida pelo utilizador *a priori* e, pAUC é a área parcial abaixo da curva ROC.

A curva $\varphi(t_0)$ é interpretada, na prática, como a proporção de genes associados ao grupo experimental com níveis de expressão acima do quantil de probabilidade $(1-t_0)$ da distribuição dos níveis de expressão na população controlo. Se dois genes tiverem a mesma $\varphi(t_0)$, o que tiver o maior valor da área parcial abaixo da curva ROC, $pAUC(t_0)$, é o gene que terá as distribuições dos níveis de expressão dos dois grupos mais afastadas.

A escolha de t_0 varia de experiência para experiência, uma vez que depende de custos e consequências de erros cometidos. Valores muito baixos para t_0 são geralmente requeridos em experiências que envolvam a pesquisa na área oncológica. No entanto, segundo Pepe *et al.* (2003), para amostras de dimensões pequenas, a estimativa da $pAUC(t_0)$ e $\varphi(t_0)$ não é possível para valores muito baixos de t_0 .

Segundo Pepe *et al.* (2003) a ordem dos genes depende da variabilidade amostral e, porque o objetivo não é testar hipóteses nem estimar parâmetros, propõem o cálculo da probabilidade de um determinado gene g se encontrar entre os primeiros top k de uma lista ordenada ($P_g(k)$). Verdadeiros genes DE terão valores de $P_g(k)$ próximos de 1. Pepe *et al.* (2003) estimam estas probabilidades pelo método *bootstrap*.

3.5.2 SAMROC

Broberg (2002) propôs um método para ordenar genes baseado no método SAM e na metodologia ROC, designado de SAMROC, cujo objetivo principal é minimizar as proporções de falsos negativos e de falsos positivos na lista dos top k genes selecionados como DE. Este método é semelhante ao SAM (secção 2.3), no entanto a constante *fudge*, b , que é adicionada no denominador da expressão (2.10) é determinada de um modo diferente.

A constante *fudge* é obtida em função de um determinado nível de significância α e com base na metodologia ROC. O que se espera quando se obtém a lista final de genes selecionados como DE, é que destes existam poucos falsos positivos e poucos falsos negativos, na prática, o que se espera é que esta lista seja composta por muitos genes DE e poucos não DE.

3.5. Métodos baseados na metodologia ROC para a seleção de genes

Para cada combinação (α, b) calcula-se:

$$samroc_i(\alpha, b) = \frac{\bar{y}_i - \bar{x}_i}{s_i + b(\alpha)}, \quad (3.24)$$

e escolhe-se a combinação (α, b) que dá origem ao valor mínimo:

$$C(\alpha) = \sqrt{(1-E)(\alpha)^2 + (1-S)(\alpha)^2}. \quad (3.25)$$

Na prática, a escolha da constante b é feita com base na distância dos pontos $(1-E, 1-S)$ à origem (Figura 3.4), e escolhe-se o ponto que permite obter a proporção de falsos negativos e a proporção de falsos positivos mais próxima de zero.

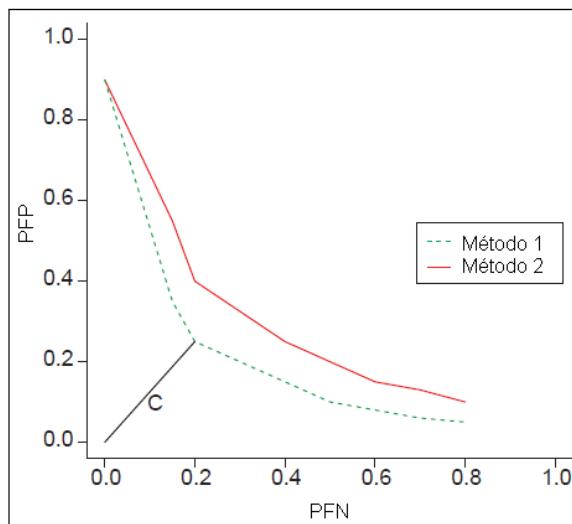


Figura 3.4: Gráfico obtido em função da proporção de falsos negativos *vs.* a proporção de falsos positivos. C corresponderá à distância mínima entre as curvas e a origem. Figura adaptada de Broberg (2003).

A proporção de falsos positivos ($1-E$) é obtida neste caso como sendo o quociente entre o número de falsos positivos de entre os genes selecionados como DE pelo número total de genes, a proporção de falsos negativos é obtida a partir do quociente entre o número de falsos negativos de entre os

genes selecionados como DE pelo número total de genes.

É possível obter-se os valores- p a partir de um teste de permutações considerando a estatística definida em (3.24) para a constante b que minimiza (3.25). Seja $samroc_i$ o valor da estatística de teste definida em (3.24) do i -ésimo gene e, $samroc_i^{*k}$ o valor da estatística de teste do i -ésimo gene na k -ésima permutação. Assim, o valor- p do gene i é dado por:

$$valor - p_i = \frac{\#\{samroc_i^{*k} : |samroc_i^{*k}| \geq |samroc_i|\}}{B \times g}, \quad (3.26)$$

e B é o número total de permutações.

Na biblioteca **SAGx** do R existe a função **samroc** que implementa este método.

3.5.3 Métodos propostos por Parodi *et al.* (2008)

Parodi *et al.* (2008) desenvolveram um método para selecionar genes, baseado na área entre a curva ROC e a diagonal de referência (ABCR⁵). Para separar curvas degeneradas das não informativas (que se situam perto da diagonal de referência) e das curvas ROC não degeneradas, Parodi *et al.* (2008) desenvolveram uma nova abordagem baseada na combinação de técnicas comuns de seleção de genes, por exemplo AUC ou estatística- t com um novo teste estatístico baseado numa variante da ABCR, o teste TNRC (*test for not-proper ROC curves*).

A proposta dos autores para o cálculo da ABCR é feita com base no cálculo da AUC segundo a regra trapezoidal (3.8). Para o cálculo da AUC, Parodi *et al.* (2008) não consideram empates na amostra global.

O algoritmo proposto por Parodi *et al.* (2008) permite selecionar genes DE, e destes, identificar os que têm curvas ROC degeneradas. Neste último caso, no contexto deste trabalho, estes genes são genes mistos.

Em primeiro lugar, obtém-se uma lista de genes considerados significativos a partir de um teste de permutações em função da estatística ABCR, e dessa lista, a partir de um teste de permutações em função da estatística TNRC, identificam-se os genes mistos.

⁵Do inglês *area between the ROC curve and the rising diagonal*.

3.5. Métodos baseados na metodologia ROC para a seleção de genes

A seleção dos genes de interesse é feita de acordo com os seguintes passos:

- Calcular a estatística ABCR para todos os genes de acordo com a expressão:

$$ABCR = \sum_{k=1}^m |\text{pAUC}_k - A_k|, \quad (3.27)$$

onde pAUC_k é a área parcial abaixo da curva ROC, A_k é a área parcial abaixo da diagonal de referência (curva ROC não informativa)

$$A_k = \frac{2k-1}{2m^2}, \quad (3.28)$$

onde m é a dimensão da amostra conjunta sem empates ($m \leq n$).

- Ordenar os genes de acordo com o valor da ABCR por ordem decrescente.
- Selecionar os top k da lista anterior considerando uma estimativa da FDR de 15% (voltar a repetir a análise para 10 % e 20%).
- A estimativa da FDR⁶ é feita com base num teste de permutações (a proposta é de 200 permutações).
- Para cada gene pertencente à lista dos top k , constrói-se a curva ROC e calcula-se a AUC a partir da regra do trapézio, onde cada gene é classificado como tendo regulação positiva ou regulação negativa de acordo com os valores da AUC serem próximos de 1 ou 0 respetivamente.
- Os top k genes são submetidos ao teste TNRC para verificar que genes têm curvas ROC são degeneradas:

$$\text{TNRC} = \sum_{k=1}^m |\text{AUC}_k - A_k| - |\text{AUC} - 0.5|, \quad (3.29)$$

⁶Do inglês *false discovery rate*.

São calculados valores- p com base num teste de permutações, e os genes com valores- p inferiores a um determinado nível de significância são declarados como tendo associadas curvas ROC degeneradas, que no contexto desta tese correspondem a genes mistos.

3.6 Algumas considerações finais

Muitos autores consideram a AUC como uma medida de distância entre duas distribuições, demonstrou-se que esta designação para a AUC não é correta.

As curvas ROC degeneradas revelam-se uma ferramenta muito útil na seleção de genes diferencialmente expressos, em particular a análise da AUC associada a este tipo de curvas, uma vez que a partir desta medida é possível identificar os vários tipos de genes diferencialmente expressos. No entanto, a AUC por si só, revela-se insuficiente na análise de genes mistos, pois estes não se distinguem dos genes não DE, uma vez que em ambas as situações a AUC terá valores próximos de 0.5.

A partir do estudo de simulação verificou-se que os vieses da AUC obtidos de forma empírica e pelo método do núcleo revelaram-se com diferenças aparentemente pouco significativas, podendo-se comparar resultados sem a necessidade de se proceder à respetiva correção do viés. No entanto, após uma reflexão sobre a análise da precisão e exatidão das estimativas da AUC obtidas pelos dois métodos no contexto da seleção de genes DE, verifica-se que quando o número de réplicas é pequeno, as estimativas da AUC pelo método empírico serão mais exatas e precisas para selecionar genes DE e, as estimativas da AUC pelo método do núcleo serão mais exatas e precisas para selecionar genes mistos.

Capítulo 4

Arrow Plot: uma nova ferramenta para a análise de genes DE

4.1 Introdução

A análise da AUC, considerando a mesma regra de classificação para todos os genes numa experiência de *microarrays*, permite-nos identificar vários tipos de genes diferencialmente expressos, nomeadamente genes com regulação positiva e negativa. No entanto, a análise da AUC por si só não permite identificar genes mistos, já que estes não se irão distinguir dos genes não DE pois, os valores da AUC em ambas as situações serão próximos de 0.5. Assim, propõe-se a análise conjunta da AUC e do coeficiente de sobreposição entre duas distribuições (OVL) que representam os níveis de expressão nos dois grupos em análise. Apresenta-se uma nova ferramenta na análise da expressão diferencial, o gráfico *Arrow plot*.

Neste trabalho propõe-se a construção do *Arrow plot* usando estimativas não-paramétricas da AUC e do OVL. Propõe-se um método de estimação não-paramétrico do OVL baseado no cálculo da área sobreposta de duas densidades estimadas pelo método do núcleo, e para o qual se desenvolveu um algoritmo que se descreve detalhadamente na secção 4.2.

Os níveis de expressão dos genes mistos não se distinguirão dos genes não DE que apresentem médias semelhantes, variâncias com grande discrepância entre os dois grupos e distribuições unimodais (Figura 4.1), uma vez que também terão valores da AUC próximos de 0.5 e valores baixos do OVL, *i.e.*, à semelhança dos genes mistos, estes genes apresentam-se na mesma

zona no *Arrow plot*. Para que seja possível diferenciar os genes mistos dos genes não DE com as características descritas anteriormente, desenvolveu-se um algoritmo que se baseia na identificação de distribuições bimodais ou multimodais, estimadas pelo método do núcleo. Assim, os genes que possuam bimodalidade (ou multimodalidade) em pelo menos uma das distribuições dos níveis de expressão de entre os genes com AUC em torno de 0.5 e OVL baixo, serão classificados de genes mistos.

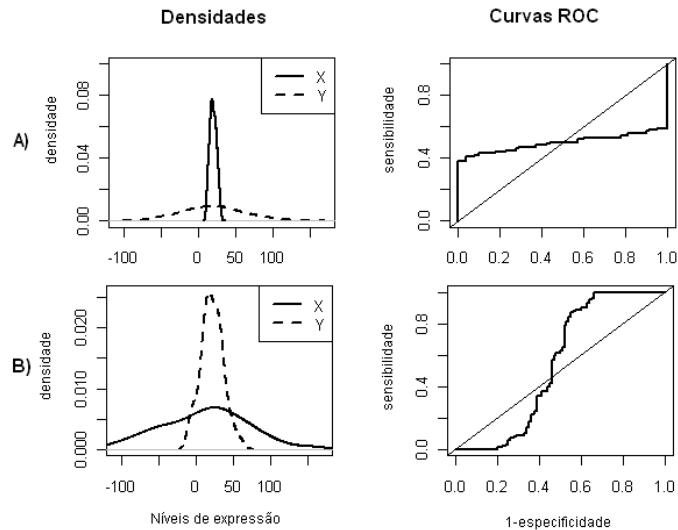


Figura 4.1: Densidades dos níveis de expressão de duas populações e respetivas curvas ROC empíricas. X representa os níveis de expressão no grupo controlo e Y representa os níveis de expressão no grupo experimental. Considerou-se a mesma regra de classificação para a estimação das curvas ROC. As densidades foram estimadas através do estimador de núcleo a partir de duas amostras de dimensão 100 simuladas de duas normais. A) $X \sim N(20, 15)$, $Y \sim N(20, 60)$; B) $X \sim N(20, 40)$, $Y \sim N(20, 5)$.

Finalmente para a construção do *Arrow plot* apresenta-se o algoritmo necessário para a sua construção e respetiva identificação de genes DE e genes mistos.

4.2 Coeficiente de sobreposição — OVL

O coeficiente de sobreposição de duas densidades (OVL) é definido como a área em comum entre duas funções de densidade de probabilidade (Figura 4.2). Este coeficiente é utilizado como medida de concordância entre duas distribuições (Weitzman, 1970; Inman e Bradley, 1989).

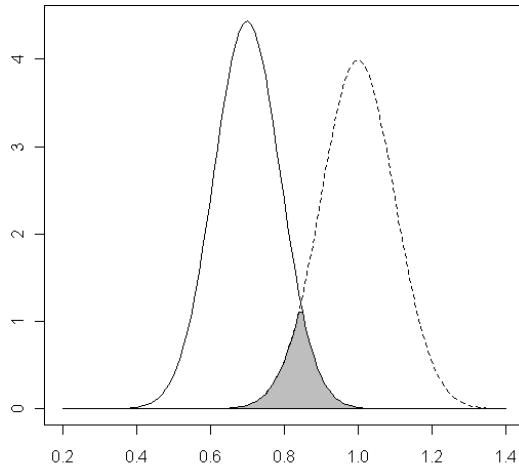


Figura 4.2: Área de sobreposição de duas densidades — OVL (área a cinzento). As distribuições representadas são normais, onde a densidade a tracejado tem distribuição $N(1,0.1)$ e a densidade a traço contínuo tem distribuição $N(0.7, 0.09)$.

A expressão do OVL pode ser representada sob várias formas. Weitzman (1970) propõe a expressão (4.1).

$$\text{OVL} = \int_c \min[f_X(c), f_Y(c)]dc, \quad (4.1)$$

onde f_X e f_Y são as f.d.p. das variáveis aleatórias X e Y , respectivamente. Os resultados são diretamente aplicáveis a distribuições discretas substituindo o integral por um somatório.

Por outro lado, a partir da conhecida expressão (4.2) é possível obter-se uma representação alternativa para o OVL (4.3) tomando $u = f_X(c)$ e $v = f_Y(c)$.

$$\min(u, v) = \frac{1}{2}(u + v) - \frac{1}{2}|u - v|. \quad (4.2)$$

$$\text{OVL}(X, Y) = 1 - \frac{1}{2} \int_{-\infty}^{+\infty} |f_X(c) - f_Y(c)| dc. \quad (4.3)$$

O OVL é invariante em relação a transformações de escala estritamente crescentes e diferenciáveis das variáveis X e Y . Esta propriedade revela-se muito importante no contexto da análise de dados de *microarrays*, uma vez que na maioria das vezes as variáveis são sujeitas a transformações, nomeadamente o logaritmo de base 2.

4.2.1 Estimação não-paramétrica do OVL

Sejam X_1, \dots, X_{n_0} e Y_1, \dots, Y_{n_1} duas amostras aleatórias e independentes e sejam \tilde{f}_0 e \tilde{f}_1 os estimadores do núcleo de f_X e f_Y dados em (3.12) e (3.13) respectivamente. A partir da expressão (4.3), o estimador do OVL pelo método do núcleo é dado por:

$$\widetilde{\text{OVL}} = 1 - \frac{1}{2} \int_{-\infty}^{+\infty} |\tilde{f}_X(t) - \tilde{f}_Y(t)| dt. \quad (4.4)$$

Schmid e Schmidt (2006) aplicam uma transformação adequada a X_i e Y_j de modo que o suporte fique restrito ao intervalo unitário, e com base em (4.4) o integral definido no intervalo unitário é aproximado pela regra do trapézio sendo o estimador do núcleo do OVL dado por:

$$\widetilde{\text{OVL}_2} = 1 - \frac{1}{2} \left(\frac{1}{q} \sum_{i=1}^q \frac{1}{2} \left(\left| \tilde{f}_X \left(\frac{i}{q} \right) - \tilde{f}_Y \left(\frac{i}{q} \right) \right| + \left| \tilde{f}_X \left(\frac{i-1}{q} \right) - \tilde{f}_Y \left(\frac{i-1}{q} \right) \right| \right) \right), \quad (4.5)$$

onde q representa o número de subintervalos equidistantes no intervalo unitário.

Algoritmo — Estimação não-paramétrica do OVL

Para estimar o OVL em função de duas densidades estimadas pelo método do núcleo, desenvolveu-se um algoritmo, onde no Algoritmo 1 apresenta-se o pseudo-código que o implementa e nas Tabelas 4.1 e 4.2 as notações e funções utilizadas. No apêndice A.1 encontra-se o código em linguagem R.

À semelhança da estimação da AUC pelo método do núcleo (secção 3.3), vai considerar-se um núcleo gaussiano e os parâmetros de suavização vão ser de acordo com os definidos em (3.11).

Ao estimar-se uma função densidade de probabilidade pelo método do núcleo, na prática o que se obtém são os pontos onde a densidade é estimada (Figura 4.3), uma vez que de acordo com (3.9), a cada ponto da amostra é centrada uma densidade normal, e cada valor de \tilde{f} é obtido somando os valores das densidades normais para esse ponto e que posteriormente são ligados de modo a dar uma ideia de continuidade.

O algoritmo proposto para a estimação não-paramétrica do OVL, tem como objetivo determinar os pontos das densidades estimadas pelo método do núcleo que delimitam a região de interseção (Figura 4.3: pontos a vermelho, verde e azul) das duas densidades (Algoritmo 1: linhas 1–21). Os pontos onde as densidades se intersetam designam-se de pontos de salto (Figura 4.3: pontos a azul). Quando não existe um ponto que pertença simultaneamente às duas densidades, este é estimado por interpolação linear (Algoritmo 1: linhas 22–46). Os pontos das densidades que delimitam a zona de interseção e os pontos de salto são combinados numa única lista e ordenados por ordem crescente das abscissas (Algoritmo 1: linha 49). Finalmente aplica-se a regra do trapézio considerando a grelha de pontos da lista final (Algoritmo 1: linha 50).

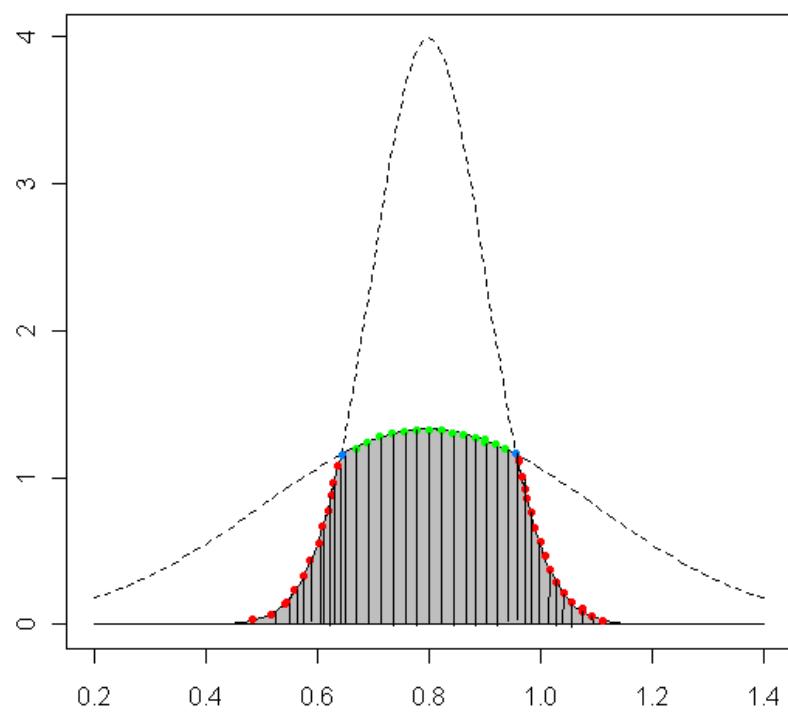


Figura 4.3: Representação gráfica dos pontos das densidades estimados pelo método do núcleo pertencentes à região de interseção das duas densidades (pontos verdes e vermelhos), pontos azuis correspondem à região de interseção e correspondem aos pontos de salto entre densidades.

Tabela 4.1: Lista de notações utilizadas no Algoritmo 1. Os símbolos estão listados pela ordem em que surgem no algoritmo.

Símbolo	Definição
G^A, G^B	Listas de pontos (abcissas e ordenadas) relativos às densidades estimadas pelo método do núcleo nas amostras A e B.
A, B	Pares de coordenadas obtidas das listas G^A e G^B que correspondem aos pontos da área de interseção das duas densidades.
$(G_x^A[i], G_y^A[i])$	Indexa um par de coordenadas da lista G^A , onde $G_x^A[i]$ corresponde à abcissa e $G_y^A[i]$ correspondente à ordenada (o mesmo para G^B e G).
$\#(.)$	Dimensão de uma lista.
G	Lista de pontos ordenada em função das abcissas, e que resulta da união de A com B.
P	União da lista G com novos pares de coordenadas correspondentes aos pontos de salto entre densidades.
$G[i]$	Indexa um par de coordenadas da lista G.
x_{new}	Nova abcissa.
y_{new}	Nova ordenada.
F	Lista final de pontos para estimar OVL.
OVL	Área de sobreposição de duas densidades estimadas pelo método do núcleo.

Tabela 4.2: Lista de funções utilizadas no Algoritmo 1. As funções estão listadas pela ordem em que surgem no algoritmo.

Função	Definição
<code>xMatch(abcissa, lista)</code>	Se numa lista existir mais de uma abcissa igual, devolve o par de coordenadas cuja ordenada corresponde ao mínimo.
<code>ordinate(abcissa, ordenada)</code>	Devolve a ordenada de um par de coordenadas.
<code>xPrev(abcissa, lista)</code>	Devolve o par de coordenadas de uma lista imediatamente anterior de uma dada abcissa.
<code>xNext(abcissa, lista)</code>	Devolve o par de coordenadas de uma lista imediatamente a seguir de uma dada abcissa.
<code>Union(lista, lista)</code>	Concatena listas.
<code>order(lista)</code>	Ordena uma lista por ordem crescente das abcissas.
<code>abscissa(abcissa, ordenada)</code>	Devolve a abcissa de uma par de coordenadas.
<code>Trapez(lista)</code>	Regra do trapézio.

```

input :  $G^A, G^B$ 
output: Estimação não-paramétrica do OVL

1  $i \leftarrow 1, A \leftarrow \text{vazio};$ 
2 while  $i <= \sharp(G^A)$  do
3    $x_1 \leftarrow G_x^A[i];$ 
4    $y_1 \leftarrow G_y^A[i];$ 
5   if { [xMatch( $x_1, G^B$ ) $\neq$  vazio  $\wedge$   $y_1 \leq \text{ordinate}(\text{xMatch}(x_1, G^B))]$   $\vee$ 
6   [ xMatch( $x_1, G^B$ ) $=$ vazio  $\wedge$  xPrev( $x_1, G^B$ ) $\neq$  vazio  $\wedge$  xNext( $x_1, G^B$ ) $\neq$ vazio  $\wedge$ 
7    $y_1 \leq \text{ordinate}(\text{xPrev}(x_1, G^B)) \wedge y_1 \leq \text{ordinate}(\text{xNext}(x_1, G^B))]$  } then
8     |  $A \leftarrow (G_x^A[i], G_y^A[i]);$ 
9   end
10   $i \leftarrow i + 1;$ 
11 end
12  $i \leftarrow 1;$ 
13  $B \leftarrow \text{vazio};$ 
14 while  $i <= \sharp(G^B)$  do
15    $x_2 \leftarrow G_x^B[i];$ 
16    $y_2 \leftarrow G_y^B[i];$ 
17   if { [xMatch( $x_2, G^A$ ) $\neq$  vazio  $\wedge$   $y_2 \leq \text{ordinate}(\text{xMatch}(x_2, G^A))]$   $\vee$ 
18   [ xMatch( $x_2, G^A$ ) $=$ vazio  $\wedge$  xPrev( $x_2, G^A$ ) $\neq$  vazio  $\wedge$  xNext( $x_2, G^A$ ) $\neq$ vazio  $\wedge$ 
19    $y_2 \leq \text{ordinate}(\text{xPrev}(x_2, G^A)) \wedge y_2 \leq \text{ordinate}(\text{xNext}(x_2, G_y^A))]$  } then
20     |  $B \leftarrow (G_x^B[i], G_y^B[i]);$ 
21   end
22    $i \leftarrow i + 1;$ 
23 end
24  $G \leftarrow \text{order}(\text{Union}(A, B));$ 
25  $i \leftarrow 1;$ 
26  $P \leftarrow \text{vazio};$ 
27 while  $i \leq \sharp(G) - 1$  do
28   if ( $G[i] \in A \wedge G[i + 1] \in B$ ) then
29     |  $x_1 \leftarrow G_x[i];$ 
30     |  $y_1 \leftarrow G_y[i];$ 
31     |  $x_4 \leftarrow G_x[i + 1];$ 
32     |  $y_4 \leftarrow G_y[i + 1];$ 
33     | if  $G[i] \in A$  then  $x_2 \leftarrow \text{abscissa}(\text{xNext}(x_1, G^A));$ 
34     | else  $x_2 \leftarrow \text{abscissa}(\text{xNext}(x_1, G^B));$ 
35     | if  $G[i] \in A$  then  $y_2 \leftarrow \text{ordinate}(\text{xNext}(x_1, G^A));$ 
36     | else  $y_2 \leftarrow \text{ordinate}(\text{xNext}(x_1, G^B));$ 
37     | if  $G[i + 1] \in A$  then  $x_3 \leftarrow \text{abscissa}(\text{xPrev}(x_4, G^A));$ 
38     | else  $x_3 \leftarrow \text{abscissa}(\text{xPrev}(x_4, G^B));$ 
39     | if  $G[i + 1] \in A$  then  $y_3 \leftarrow \text{ordinate}(\text{xPrev}(x_4, G^A));$ 
40     | else  $y_3 \leftarrow \text{ordinate}(\text{xPrev}(x_4, G^B));$ 
41     |  $D \leftarrow (x_1 - x_2) \times (y_3 - y_4) - (y_1 - y_2) \times (x_3 - x_4);$ 
42     | if  $D \neq 0$  then
43       | |  $x_{new} \leftarrow (1/D) \times [(x_3 - x_4)(x_{12} - y_1 \times x_2) - (x_1 - x_2)(x_3 \times y_4 - y_3 \times x_4)];$ 
44     | |  $y_{new} \leftarrow (1/D) \times [(y_3 - y_4)(x_1 \times y_2 - y_1 \times x_2) - (y_1 - y_2)(x_3 \times y_4 - y_3 \times x_4)];$ 
45     | | else  $P \leftarrow (P, (x_1, y_1), (x_{new}, y_{new}));$ 
46   else
47     | |  $P \leftarrow (P, (G_x[i], G_y[i]));$ 
48   end
49    $i \leftarrow i + 1;$ 
50 end
51  $F \leftarrow \text{order}(\text{Union}(P, (G_x[\sharp(G)], G_y[\sharp(G)])));$ 
52  $OVL \leftarrow \text{Trapez}(F)$ 

```

Algoritmo 1: Pseudo-código para a estimação não-paramétrica do OVL (Silva-Fortes *et al.*, 2012)

A implementação do Algoritmo 1 em linguagem R numa base de dados com 10000 genes demora aproximadamente 60 minutos num Pentium 533 MHz.

Procedeu-se a um estudo de simulação, para analisar o erro padrão e viés do OVL obtido pelo Algoritmo 1, onde foram aplicados métodos de Monte Carlo (MC) e *bootstrap*. No CD na pasta “Capítulo 4” encontra-se o ficheiro `bootstrap.R` com a análise desenvolvida.

De modo a analisar o comportamento do Algoritmo 1 em situações em que as distribuições são mais ou menos afastadas, consideraram-se os seguintes pares de distribuições: $N(20, 4)$ e $N(10, 4)$; $N(20, 4)$ e $N(15, 4)$; e $N(20, 4)$ e $N(19, 4)$. Para cada par foram simuladas amostras de dimensões 100 e 500, perfazendo um total de 6 combinações.

No caso particular em que as distribuições são normais, com valores médios μ_1 e μ_2 e variâncias iguais, σ^2 , as densidades cruzam-se num único ponto, $x = \frac{\mu_1 + \mu_2}{2}$, e o valor exato do OVL é dado por:

$$\text{OVL} = 2\Phi\left(-\frac{|\delta|}{2}\right), \quad (4.6)$$

onde $\delta^2 = \frac{(\mu_2 - \mu_1)^2}{\sigma^2}$.

Demonstração:

Seja $X \sim N(\mu_1, \sigma)$ e $Y \sim N(\mu_2, \sigma)$, o ponto onde as duas densidade se cruzam, qualquer que seja μ_1 e $\mu_2 \in \mathbb{R}$ é dado por:

$$\begin{aligned} f_X(x) &= f_Y(x) \\ \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu_1)^2}{\sigma^2}} &= \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu_2)^2}{\sigma^2}} \\ x &= \frac{\mu_1 + \mu_2}{2} \end{aligned}$$

Para $\mu_1 \leq \mu_2$, O OVL é dado por:

$$\begin{aligned}
 \text{OVL} &= P\left(Y \leq \frac{\mu_1 + \mu_2}{2}\right) + P\left(X \geq \frac{\mu_1 + \mu_2}{2}\right) \\
 &= P\left(Y \leq \frac{\mu_1 + \mu_2}{2}\right) + 1 - P\left(X \leq \frac{\mu_1 + \mu_2}{2}\right) \\
 &= 2\Phi\left(\frac{\mu_1 - \mu_2}{2\sigma}\right)
 \end{aligned}$$

Para $\mu_1 \geq \mu_2$ o OVL é dado por:

$$\begin{aligned}
 \text{OVL} &= P\left(X \leq \frac{\mu_1 + \mu_2}{2}\right) + P\left(Y \geq \frac{\mu_1 + \mu_2}{2}\right) \\
 &= P\left(X \leq \frac{\mu_1 + \mu_2}{2}\right) + 1 - P\left(Y \leq \frac{\mu_1 + \mu_2}{2}\right) \\
 &= 2\Phi\left(\frac{-(\mu_1 - \mu_2)}{2\sigma}\right)
 \end{aligned}$$

▽

Para cada uma das combinações atrás descritas, foram consideradas 1000 réplicas MC e obtiveram-se as seguintes estimativas:

- estimativa MC do $\widetilde{\text{OVL}}$:

$$\bar{x}_{MC}(\widetilde{\text{OVL}}) = \frac{\sum_{i=1}^{1000} \widetilde{\text{OVL}}_i}{1000},$$

onde $\widetilde{\text{OVL}}_i$ é a estimativa do OVL obtida a partir do Algoritmo 1 na réplica i ;

- estimativa do erro padrão MC do $\widetilde{\text{OVL}}$:

$$\widehat{s}_{eMC}(\widetilde{\text{OVL}}) = \sqrt{\frac{\frac{1}{999} \sum_{i=1}^{1000} (\widetilde{\text{OVL}}_i - \bar{x}_{MC})^2}{\sqrt{1000}}}$$

- estimativa do viés relativo do \widetilde{OVL} :

$$\widehat{VR}_{MC}(\widetilde{OVL}) = \frac{\bar{x}_{MC} - OVL}{OVL},$$

onde OVL é o valor exato obtido a partir de (4.6).

Na Tabela 4.3 apresentam-se as estimativas MC da média, do erro padrão e do viés relativo do \widetilde{OVL} .

Tabela 4.3: Estimativas da média MC, erro padrão MC e viés do OVL estimado pelo Algoritmo 1.

OVL exato	$n_1 = n_2 = 100$			$n_1 = n_2 = 500$		
	Média MC	Erro padrão MC	Viés relativo	Média MC	Erro padrão MC	Viés relativo
$X_1 \sim N(20, 4)$ $X_2 \sim N(10, 4)$ $OVL=0.2112$	0.2342	0.0012	0.1088	0.2257	5.58E-04	0.0687
$X_1 \sim N(20, 4)$ $X_2 \sim N(15, 4)$ $OVL= 0.532$	0.5512	0.0017	0.0362	0.5449	8.3E-04	0.0244
$X_1 \sim N(20, 4)$ $X_2 \sim N(19, 4)$ $OVL=0.901$	0.8687	0.0014	-0.0359	0.8953	8.1E-04	-0.0063

Na Figura 4.4 estão representadas as estimativas do OVL pelo Algoritmo 1, nas 1000 réplicas, para cada uma das 6 combinações atrás descritas e, respetivos valores exatos.

Pela análise da Tabela 4.3 e da Figura 4.4, pode concluir-se que o viés decresce com o aumento da dimensão da amostra e por outro lado o viés é mais baixo quando as distribuições se encontram mais próximas. Pode observar-se também, que o algoritmo tende a sobreestimar o OVL quando as distribuições são mais afastadas e tende a subestimar quando são mais próximas.

Para estimar o erro padrão e viés do OVL estimado pelo Algoritmo 1 pelo método *bootstrap*, consideraram-se $B=200$ réplicas *bootstrap* para cada

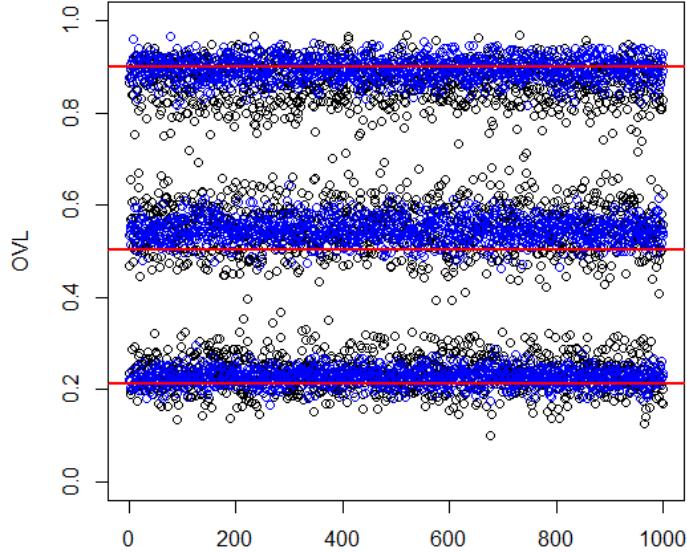


Figura 4.4: Comparação entre o valor exato do OVL e respetivas estimativas pelo Algoritmo 1 considerando 1000 réplicas MC para cada par de distribuições. Do topo para a base no gráfico: $N(20, 4)$ e $N(10, 4)$; $N(20, 4)$ e $N(15, 4)$ e $N(20, 4)$ e $N(19, 4)$. As linhas a vermelho representam o valor exato do OVL, pontos a azul representam as estimativas do OVL considerando amostras de dimensão 500, pontos a preto representam as estimativas do OVL considerando amostras de dimensão 100.

réplica MC i , $i = 1, \dots, 1000$, do estudo de simulação anterior.

Para cada uma das combinações consideradas no estudo de simulação anterior, obtiveram-se as seguintes estimativas:

- para cada réplica i , obteve-se a estimativa *bootstrap* do erro padrão de $\widetilde{\text{OVL}}$:

$$\widehat{s}_e_{B_i}(\widetilde{\text{OVL}}) = \sqrt{\frac{1}{199} \sum_{j=1}^{200} \left(\widetilde{\text{OVL}}_{ij}^* - \widetilde{\text{OVL}}_{B_i}^* \right)^2},$$

onde \widetilde{OVL}_{ij}^* é a estimativa do OVL pelo Algoritmo 1 na j -ésima réplica *bootstrap* da i -ésima réplica e, o $\widetilde{OVL}_{B_i}^*$ é a estimativa *bootstrap* do \widetilde{OVL} na i -ésima réplica:

$$\widetilde{OVL}_{B_i}^* = \frac{1}{200} \sum_{j=1}^{200} \widetilde{OVL}_{ij}^*;$$

- a média da estimativa *bootstrap* dos erros padrão do \widetilde{OVL} nas 1000 réplicas é dada por:

$$\bar{x}_{\widehat{s}e_B}(\widetilde{OVL}) = \frac{\sum_{i=1}^{1000} \widehat{s}e_{B_i}}{1000};$$

- o erro padrão da estimativa *bootstrap* do erro padrão do \widetilde{OVL} é dado por:

$$\widehat{s}e_{\widehat{s}e_B}(\widetilde{OVL}) = \sqrt{\frac{1}{999} \sum_{i=1}^{1000} (\widehat{s}e_{B_i} - \bar{x}_{\widehat{s}e_B})^2};$$

- o viés relativo da estimativa *bootstrap* do \widetilde{OVL} é dado por:

$$\widehat{VR}_B(\widetilde{OVL}) = \frac{\bar{x}_{\widehat{s}e_B} - \widehat{s}d_{MC}}{\widehat{s}d_{MC}},$$

onde $\widehat{s}d_{MC}$ representa o desvio padrão da estimativa MC do \widetilde{OVL}

$$\widehat{s}d_{MC} = \sqrt{\frac{1}{999} \sum_{i=1}^{1000} (\widetilde{OVL}_i - \bar{x}_{MC})^2}.$$

Na Tabela 4.4 apresentam-se os valores obtidos das respetivas estimativas. Clemons e Bradley (2000) também propõem a estimação do OVL usando funções de densidade estimadas pelo método do núcleo, e Schmid e Schmidt (2006) propõem 5 estimadores não-paramétricos para o OVL. Comparando os resultados aqui obtidos com os destes autores, conclui-se que o Algoritmo 1 permite obter estimativas com viés mais reduzido, em particular quando as distribuições são mais próximas. Conclui-se que o Algoritmo 1 é mais eficiente.

Tabela 4.4: Erro padrão e viés relativo das estimativas *bootstrap* do OVL estimado pelo Algoritmo 1.

	$n_1 = n_2 = 100$			$n_1 = n_2 = 500$		
OVL exato	$\bar{x}_{\widehat{s}_B}$	Erro padrão	Viés relativo	$\bar{x}_{\widehat{s}_B}$	Erro padrão	Viés relativo
$X_1 \sim N(20, 4)$ $X_2 \sim N(10, 4)$ $OVL=0.2112$	0.0378	1.21E-05	-0.0043	0.0174	0.0011	-0.0097
$X_1 \sim N(20, 4)$ $X_2 \sim N(15, 4)$ $OVL=0.532$	0.0527	0.0034	-0.0035	0.0248	0.002	-0.0501
$X_1 \sim N(20, 4)$ $X_2 \sim N(19, 4)$ $OVL=0.901$	0.1674	3.77	2.822	0.0303	0.0881	0.2611

4.3 Arrow plot

O *Arrow plot* resulta da representação gráfica num plano unitário das estimativas da AUC e do OVL para todos os genes de uma experiência de *microarrays* (Silva-Fortes *et al.*, 2012). A partir deste gráfico é possível selecionar genes mistos, genes com regulação positiva e genes com regulação negativa. Espera-se que estes genes tenham valores baixos para as estimativas do OVL. Particularmente, quando os genes têm regulação negativa ou positiva, espera-se que as distribuições dos níveis de expressão para os dois grupos se encontrem suficientemente afastadas (Figura 4.5).

Relativamente aos genes mistos, apesar de terem médias dos níveis de expressão dos dois grupos semelhantes, também é esperado que tenham valores baixos para as estimativas do OVL (Figura 4.6).

Quanto aos valores da AUC, para os genes mistos, é esperado que se obtenham estimativas em torno de 0.5. Como já foi referido anteriormente, genes não DE com variâncias dos níveis de expressão nos dois grupos significativamente diferentes, valores médios semelhantes e distribuições unimodais, apresentar-se-ão na mesma zona do gráfico que os genes mistos. Para ultrapassar esta situação, desenvolveu-se um algoritmo (Algoritmo 2), de modo que de entre os genes com valores baixos para as estimativas do OVL e estimativas à volta de 0.5 para a AUC, seja possível identificar os genes mistos.

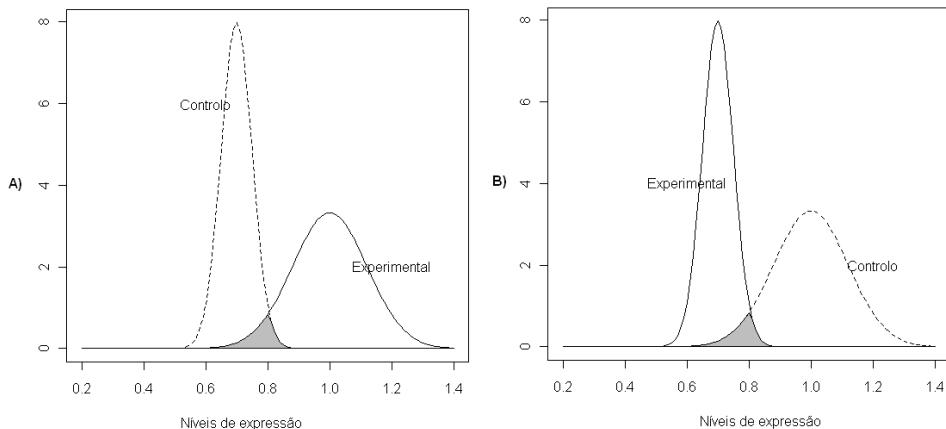


Figura 4.5: A) Representação do OVL de um gene com regulação positiva (hipotética). B) Representação do OVL de um gene com regulação negativa (hipotética). O OVL corresponde às áreas a cinzento.

O algoritmo baseia-se na identificação dos genes de entre os descritos acima, que possuam bimodalidade (ou multimodalidade) numa das distribuições, ou em ambas as distribuições, estimadas pelo método do núcleo.

No Algoritmo 2 apresenta-se o pseudo-código para a identificação de bimodalidade e seleção de genes mistos, estando no apêndice A.2 o código em linguagem R. Nas Tabelas 4.5 e 4.6 estão descritos os símbolos e funções utilizadas.

A seleção dos genes mistos consiste em dois passos principais. Num primeiro passo, selecionam-se todos os genes com valores da AUC em torno de 0.5 (*e.g.*, $0.4 < \text{AUC} < 0.6$) e valores baixos do OVL (Algoritmo 2: linhas 1–6), estes genes são designados de candidatos a mistos. De entre os genes selecionados neste passo (Algoritmo 2: linha 9), procede-se à seleção dos genes mistos que dependem da identificação de bimodalidade (ou multimodalidade) nas distribuições dos grupos em análise. A análise da bimodalidade baseia-se no comportamento das ordenadas dos pontos onde as densidades são estimadas pelo método do núcleo nos dois grupos. Os pares de coordenadas em ambos os grupos são ordenados por ordem crescente das abscissas (Algoritmo 2: linha 10). Em cada grupo verifica-se se o valor de uma ordenada é igual ou inferior ao valor da ordenada imediatamente a seguir (do mesmo grupo),

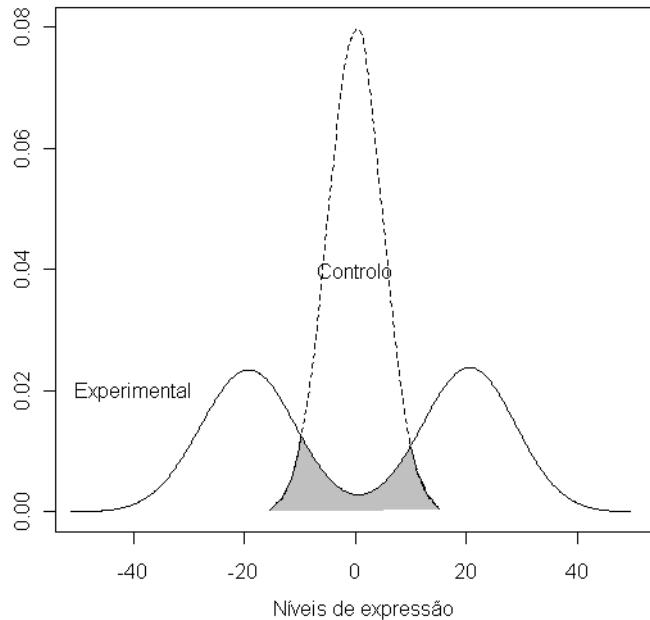


Figura 4.6: A) Representação do OVL de um gene misto (hipotético). O OVL corresponde à área a cinzento.

e em caso afirmativo é atribuída a etiqueta um a esse ponto, e zero caso contrário (Algoritmo 2: linhas 11–16 e 24–29). Isto permite-nos analisar a variação das densidades no intervalo observado. Considerando apenas as ordens dos pontos onde a função é decrescente, *i.e.*, os pontos cujas etiquetas são zero (Algoritmo 2: linhas 17 e 30), calculam-se as diferenças das ordens adjacentes, e se existir alguma diferença cujo valor absoluto é superior a um, conclui-se que existe bimodalidade ou (multimodalidade) (Algoritmo 2: linhas 19–22 e 31–35). Para que um gene seja considerado misto, é suficiente que se identifique bimodalidade num dos grupos (Algoritmo 2: linhas 37–40).

Para o biólogo é importante identificar em que grupo se verifica a bimodalidade, e nesse sentido, de entre os genes classificados como mistos, atribuem-se diferentes cores aos respetivos pontos do *Arrow plot* (Algoritmo 2: linhas 41–59, e Algoritmo 3: linhas 22–26).

Tabela 4.5: Lista de notações utilizadas nos Algoritmos 2 e 3. Os símbolos estão listados pela ordem em que surgem no algoritmo.

Símbolo	Definição
X	Matriz do tipo $p \times n$ correspondente à amostra A onde as colunas representam os <i>arrays</i> e as linhas os genes.
Y	Matriz do tipo $p \times m$ correspondente à amostra B onde as colunas representam os <i>arrays</i> e as linhas os genes.
$X[i], Y[i]$	Indexam um gene (linha da matriz) nas amostras A e B.
k_1, k_2	Pontos de corte definidos pelo utilizador ¹ . $k_1 \in]0.5; 0.6]$ e $k_2 \in [0.4, 0.5[, w \in]0; 0.6]$.
S^A, S^B	Listas de pontos das densidades estimadas pelo método do núcleo das amostras A e B.
$S^A[j]$	Indexa um gene de uma subamostra S da amostra A.
$S_y^{A[j]}[i]$	Indexa a ordenada do gene j da subamostra S obtida da amostra A.
$Bim_X[j]$	Indexa um gene apresentando bimodalidade na densidade estimada pelo método do núcleo na amostra A.
$Bim_Y[j]$	Indexa um gene apresentando bimodalidade na densidade estimada pelo método do núcleo na amostra B.
$Bim[j]$	Indexa um gene apresentando bimodalidade em pelo menos uma das amostras.
MIX	Lista de genes candidatos a genes mistos.
BIM	Lista com identificação da bimodalidade dos genes da lista MIX.

Tabela 4.6: Lista de funções utilizadas nos Algoritmos 2 e 3. As funções estão listadas pela ordem em que surgem no algoritmo.

Função	Definição
<code>AUC(lista,lista)</code>	Área abaixo da curva ROC estimada pelo método do trapézio ou método do núcleo.
<code>OVL(lista,lista)</code>	Área de sobreposição de duas densidades estimada pelo Algoritmo 1.
<code>kernel(lista)</code>	Método de estimação do núcleo.
<code>rank(lista)</code>	Devolve as ordens de uma lista.

```

input :  $X, Y$ 
output: MIX e BIM

1 K $\leftarrow$  vazio;  $i \leftarrow 1$ ;
2 for  $i \leftarrow p$  do
3   if AUC( $X[i], Y[i]$ ) $< k_1 \wedge$  AUC( $X[i], Y[i]$ ) $> k_2 \wedge$  OVL(kernel( $X[i]$ ), kernel( $Y[i]$ )) $< w$ 
4     then
5       | K $\leftarrow i$ ;
6   end
7  $z_1 \leftarrow$  vazio;  $z_2 \leftarrow$  vazio;  $Bim_X \leftarrow$  vazio;  $Bim_Y \leftarrow$  vazio;  $Bim \leftarrow$  vazio;
8 MIX $\leftarrow$  vazio;  $j \leftarrow 1$ ;
9 for  $j \in K$  do
10    $S^A \leftarrow$  order(kernel( $X[j]$ ));  $S^B \leftarrow$  order(kernel( $Y[j]$ ));
11   while  $i \leq \#(S^A[j])$  do
12     if  $S_y^{A[j]}[i] \leq S_y^{A[j]}[i + 1]$  then
13       |  $z_1^{[j]}[i] \leftarrow 1$ ;
14     else
15       |  $z_1^{[j]}[i] \leftarrow 0$ ;
16     end
17     for  $z_1^{[j]} \neq 1$  do
18       |  $r_1^{[j]}[i] \leftarrow$  rank( $z_1^{[j]}[i]$ ) - rank( $z_1^{[j]}[i + 1]$ );
19       if  $\exists |r_1^{[j]}| > 1$  then
20         | |  $Bim_X[j] \leftarrow$  VERDADEIRO ;
21     end
22   end
23 end
24 while  $i \leq \#(S^B[j])$  do
25   if  $S_y^{B[j]}[i] \leq S_y^{B[j]}[i + 1]$  then
26     |  $z_2^{[j]}[i] \leftarrow 1$ ;
27   else
28     |  $z_2^{[j]}[i] \leftarrow 0$ ;
29   end
30   for  $z_2^{[j]} \neq 1$  do
31     |  $r_2^{[j]}[i] \leftarrow$  rank( $z_2^{[j]}[i]$ ) - rank( $z_2^{[j]}[i + 1]$ );
32     if  $\exists |r_2^{[j]}| > 1$  then
33       | |  $Bim_Y[j] \leftarrow$  VERDADEIRO ;
34   end
35 end
36 end
37 if  $Bim_X[j] =$  VERDADEIRO  $\vee Bim_Y[j] =$  VERDADEIRO then
38   | | MIX $\leftarrow j$ ;
39 end
40 end
41 BIM $\leftarrow$  vazio; group1 $\leftarrow$  vazio; group2 $\leftarrow$  vazio; both $\leftarrow$  vazio;
42 for  $i \in MIX$  do
43   if  $Bim_X[i] =$  VERDADEIRO then
44     | | group1[i]  $\leftarrow 1$ ;
45   else
46     | | group1[i]  $\leftarrow 0$ ;
47   end
48   if  $Bim_Y[i] =$  VERDADEIRO then
49     | | group2[i]  $\leftarrow 1$ ;
50   else
51     | | group2[i]  $\leftarrow 0$ ;
52   end
53   if  $Bim_X[i] =$  VERDADEIRO  $\wedge Bim_Y[i] =$  VERDADEIRO then
54     | | both[i]  $\leftarrow 1$ ;
55   else
56     | | both[i]  $\leftarrow 0$ ;
57   end
58 end
59 BIM $\leftarrow$ (group1,group2,both)

```

Algoritmo 2: Pseudo-código para seleção de genes candidatos a genes mistos e identificação de distribuições bimodais (ou multimodais).

Para a construção do *Arrow plot* e identificação de genes DE e genes mistos, apresenta-se no Algoritmo 3 o pseudo-código e nas Tabelas 4.7 e 4.8 os símbolos e funções utilizadas. No apêndice A.3 encontra-se o código em linguagem R.

Após a estimação das AUC e OVL de todos os genes, estes representam-se num diagrama de dispersão, onde no eixo das abcissas representam-se as estimativas do OVL e no eixo das ordenadas as estimativas da AUC (Algoritmo 3: linhas 1–7 e 19). Espera-se que os genes com regulação positiva apresentem valores da AUC próximos de um, os quais são representados a vermelho (Algoritmo 3: 12–13 e 20). Os genes com regulação negativa, espera-se que tenham valores da AUC próximos de zero e que são representados a azul (Algoritmo 3: linhas 15–16 e 21). Os genes mistos serão representados com cores diferentes em função da presença da bimodalidade, no grupo controlo (cor azul claro), grupo experimental (cor laranja) ou em ambos (cor verde) (Algoritmo 3: linhas 22–25).

Tabela 4.7: Lista de notações utilizadas no Algoritmo 3. Os símbolos estão listados pela ordem em que surgem no algoritmo.

Símbolo	Definição
auc	Lista de todos os genes com valores da AUC.
ovl	Lista de todos os genes com valores do OVL estimados estimados pelos Algoritmo 1.
UP	Lista de genes com regulação positiva.
DOWN	Lista de genes com regulação negativa .
$c1, c2$	Pontos de corte definidos pelo utilizador. $c1 > 0.5$ e $c2 < 0.5$.
$BIM[i, j]$	Indexa o gene i e a coluna j ($j=1,2,3$) 1=grupo A, 2=grupo B, 3=Grupos A e B.

Tabela 4.8: Lista de funções utilizadas no algoritmo 3. As funções estão listadas pela ordem em que surgem no algoritmo.

Função	Definição
plot(lista,lista)	Diagrama de dispersão, onde o 1º argumento corresponde à lista de valores representados no eixo das abscissas e o segundo argumento no eixo das ordenadas.
points(lista,cor)	Função que atribui uma cor à lista de pontos definida no 1º argumento.

```

input : X, Y, MIX, BIM
output: Arrow plot

1 auc← vazio;
2 ovl← vazio;
3 i← vazio;
4 for i ← p do
5   | auc[i]← AUC(X[i],Y[i]);
6   | ovl← OVL(kernel(X[i]),kernel(Y[i]));
7 end
8 UP← vazio;
9 DOWN← vazio;
10 i ← 1;
11 for i ← p do
12   | if AUC(X[i],Y[i])> c1  $\wedge$  OVL(kernel(X[i]),kernel(Y[i]))< w then
13   |   | UP← i;
14   |
15   | if AUC(X[i],Y[i])< c2  $\wedge$  OVL(kernel(X[i]),kernel(Y[i]))< w then
16   |   | DOWN← i;
17   |
18 end
19 plot(ovl,auc);
20 points(UP,vermelho);
21 points(DOWN,azul);
22 for i ∈ MIX do
23   | points(BIM[i,1]=1,azul claro);
24   | points(BIM[i,2]=1,laranja);
25   | points(BIM[i,3]=1,verde);
26 end

```

Algoritmo 3: Pseudo-código para a construção do *Arrow plot* e identificação de genes com regulação positiva, regulação negativa e genes mistos.

4.4 Considerações finais

Neste capítulo apresentou-se um algoritmo para a estimativa não-paramétrica do OVL. Este algoritmo permite estimar o OVL a partir de densidades estimadas pelo método do núcleo com um número de interseções ilimitado e não impõe restrições quanto ao suporte das variáveis. Para determinar a curva de interseção entre as duas densidades, numa base ponto-por-ponto, não é necessário que os pontos correspondentes às estimativas das densidades pelo método do núcleo tenham as mesmas abscissas. O algoritmo permite obter estimativas não enviesadas do OVL. O estudo de simulação revelou que este algoritmo estima o OVL com viés reduzido, com tendência a subestimar quando as densidades são mais afastadas e a sobreestimar quando as densidades se encontram mais próximas.

É também proposto um algoritmo para identificar a presença de bimodalidade (ou multimodalidade) em distribuições estimadas pelo método do núcleo. É importante referir que este algoritmo foi desenvolvido em particular para genes que possuam distribuições dos níveis de expressão com AUC em torno de 0.5 e valores do OVL baixos. Nestas circunstâncias espera-se que as distribuições dos níveis de expressão em ambos os grupos tenham médias semelhantes e, ou são bimodais em pelo menos um dos grupos (genes mistos), ou as distribuições dos níveis de expressão são unimodais com variâncias significativamente diferentes (genes não DE).

Apresentou-se um novo gráfico, *Arrow plot*, que permite selecionar para além de genes com regulação positiva e negativa, genes com um comportamento biológico de interesse (genes mistos). Este gráfico baseia-se nas estimativas do OVL e AUC, onde neste trabalho foram usados métodos não-paramétricos. A análise do gráfico é bastante intuitiva, uma vez que ambas as estimativas variam entre 0 e 1. Uma propriedade importante que AUC e OVL possuem, é o facto de serem invariantes a transformações de escala das variáveis X e Y , que no caso particular dos dados de *microarrays*, estes são sujeitos na grande maioria das vezes a transformações. O *Arrow plot* permite-nos obter uma fotografia global do comportamento dos genes e com base na sua análise podemos escolher pontos de corte para a AUC e OVL, apesar da escolha destes valores ser arbitrária.

No capítulo que se segue apresenta-se a aplicação do *Arrow plot* em duas bases de dados disponíveis publicamente e a dados simulados, com objetivo

4.4. Considerações finais

de avaliar a sua performance em comparação com outros métodos descritos nos capítulos 2 e 3.

Capítulo 5

Aplicações

Neste capítulo pretende-se avaliar a performance do novo método proposto e realizar uma análise comparativa com os métodos descritos nos capítulos 2 e 3. A análise é realizada em dados simulados e em duas bases de dados disponíveis publicamente.

Os critérios para a seleção das duas bases de dados tiveram a haver essencialmente com as condições de acesso e o desenho experimental, *i.e.*, dados de acesso livre, com duas condições experimentais e, com a particularidade de em uma ou nas duas condições experimentais houvesse misturas de tecidos.

A base de dados *Cancro da Bexiga* (secção 5.2) é usada com o objetivo de se realizar um estudo desde a análise de baixo nível até à aplicação do método proposto.

Os dados *Linfoma* (secção 5.3) apesar de serem provenientes de *microarrays* de cDNA (de dois canais), têm um desenho experimental semelhante aos *arrays* de um canal, *i.e.*, as amostras controlo e experimental são hibridadas em *arrays* independentes. O objetivo do estudo desta base de dados é o de comparar os resultados obtidos no estudo de Parodi *et al.* (2008) (secção 3.5.3), com os resultados obtidos com o novo método aqui proposto.

5.1 Dados Simulados

5.1.1 Introdução

Com o objetivo de analisar a performance do método aqui proposto, procedeu-se à criação de uma base de dados simulada que traduzisse o comportamento de dados provenientes de *microarrays*. A simulação permite controlar determinados fatores, permitindo assim avaliar a performance de um método de uma forma pouco dispendiosa, mais simples e rápida.

Simular dados que representem níveis de expressão de genes de uma experiência de *microarrays* não é uma tarefa trivial. Os sistemas biológicos são complexos e existem genes que têm algum tipo de interação. As possibilidades em simular este tipo de dados são intermináveis, mas podem não ser todas razoáveis, pois alguma combinação de padrões pode não ser realista.

Existem várias abordagens para criar dados desta natureza. Uma opção é por exemplo adicionar genes com níveis de concentração conhecidos a uma base de dados já conhecida. Por exemplo os dados *spike-in* HG-U95 e HG-U133 da Affymetrix (McGee e Chen, 2006) acrescentam genes à amostra que naturalmente não se encontram. Por exemplo, genes de plantas podem ser adicionados a uma amostra de genes de ratos.

Aqui pretende-se simular dados que representem o comportamento mais comum de dados provenientes de *microarrays* e, que reflitam particularmente três padrões: genes com regulação positiva, genes com regulação negativa e genes mistos. Numa experiência de *microarrays* cujo objetivo seja analisar a expressão diferencial, tipicamente é esperado que cerca de 1% a 5% do número total de genes seja DE (Silva-Fortes *et al.*, 2007), e o número de genes com regulação positiva seja semelhante ao número de genes com regulação negativa. Assim, de um total de 10000 genes, simularam-se 225 genes¹ com regulação positiva, 225 genes com regulação negativa e 50 genes mistos, sendo os restantes não DE.

Muitos estudos têm admitido que os dados correspondentes aos níveis de

¹Por conveniência, sempre que há referência a simulação de genes entenda-se que o que se simula são os níveis de expressão dos genes.

expressão provenientes de *microarrays* têm distribuições normais. Contudo, existe pouca literatura que confirme a normalidade, sendo esta a exceção e não a regra (Hardin e Wilson, 2007). A variabilidade biológica torna complicada a identificação da fonte da não-normalidade. Dados não-normais podem ser fruto por exemplo de misturas de dados com distribuição normal. Outra dificuldade neste tipo de experiências, é o facto de amostras de dimensões reduzidas não terem o poder de justificar a normalidade dos dados.

É conhecido que as distribuições dos níveis de expressão são altamente enviesadas (geralmente enviesadas à direita), com muitos valores extremos. Assim, para imitar os dados provenientes de *microarrays*, os dados correspondendo ao grupos controlo e experimental foram simulados de distribuições log-normais e posteriormente foram sujeitos a uma transformação logarítmica para tornar os dados mais simétricos, *i.e.*, $\log N(x; \mu, \sigma)$ representa a distribuição log-normal com parâmetro de localização $\mu \in \mathbb{R}$ e de escala $\sigma > 0$.

Considere-se X e Y duas variáveis aleatórias independentes, onde X é a variável aleatória que representa os níveis de expressão no grupo controlo e $X \sim \log N(\mu_x, \sigma_x)$, e seja Y a variável aleatória que representa os níveis de expressão no grupo experimental, onde $Y \sim \log N(\mu_y, \sigma_y)$. Simularam-se 30 *arrays* para cada grupo e um total de 10000 genes. As simulações foram feitas de forma independente, embora se tenha conhecimento de que os níveis de expressão dos genes estão longe de serem independentes.

Foram consideradas três características para simular esta base de dados: a magnitude da diferença entre os valores médios, a magnitude da diferença entre as variâncias dos grupos controlo e experimental, e a existência ou não de bimodalidade nas distribuições num dos grupos. Consideraram-se várias combinações destes parâmetros.

Para assegurar que os genes sejam verdadeiros não DE, as médias dos grupos controlo e experimental foram simuladas usando o mesmo vetor de médias. De modo a providenciar vários padrões de distribuições, consideraram-se diferentes variâncias. O efeito de alterar as variâncias parece não afetar os níveis de expressão destes genes, no sentido de se manterem sempre não DE, uma vez que em todos os *arrays* simulados foram gerados do mesmo vetor de médias. No entanto, alguns destes genes não DE, que apresentem variâncias significativamente diferentes, irão confundir-se com os genes mistos quando representados no *Arrow plot*, mas

que após uma análise de bimodalidade espera-se que sejam diferenciados dos genes mistos. A diferença das médias dos *arrays* controlo e experimental variou entre -0.9 e 0.9 e a diferença das variâncias variou no intervalo 0 a 6.25.

Para os genes com regulação positiva e negativa, considerou-se para a diferença das médias entre os grupos controlo e experimental, os intervalos 3.5 a 12.5 e -12.5 a -3.5, respetivamente. A diferença das variâncias variou entre 0 e 6.25.

Para os genes mistos considerou-se uma mistura de distribuições log-normais num dos grupos, de modo a obter distribuições bimodais. Se X for uma variável aleatória cuja distribuição é uma mistura de log-normais, definida por $\alpha(x; \mu_0, \sigma_0) + (\alpha - 1)(x; \mu_1, \sigma_1)$, $x > 0$. O parâmetro $\alpha \in (0, 1)$ especifica a contribuição de cada componente na mistura. Ao simular-se a mistura num dos grupos, considerou-se $\mu_0 = 3.5$ fixo e μ_1 foi gradualmente incrementado assumindo valores entre 7 a 17, mantendo-se $\sigma_0 = \sigma_1 = 1.2$ fixo. Considerou-se $\alpha = 0.5$. Para o outro grupo considerou-se uma densidade log-normal com parâmetro de localização igual a $\alpha \times \mu_0 + (1 - \alpha) \times \mu_1$. Finalmente aplicou-se uma transformação logarítmica aos níveis de expressão dos 10000 genes em ambos os grupos, de modo a tornar as distribuições mais simétricas.

No CD na pasta “Capítulo 5” encontra-se o ficheiro `simulação.R` com o código R que implementa a simulação dos dados e a análise que a seguir se descreve.

5.1.2 Seleção de genes DE e mistos

Na Figura 5.1 estão representados os *box plots* dos níveis de expressão simulados dos *arrays* nas amostras controlo e experimental. Como se pode observar, os *arrays* apresentam um comportamento semelhante quanto à distribuição e dispersão dos níveis de expressão, garantindo assim o pressuposto dos níveis de expressão dos *arrays* possuírem distribuições semelhantes.

Na Figura 5.2 apresenta-se o *Arrow plot*, cuja AUC foi estimada pelo método do núcleo. Com o objetivo de analisar se os genes se encontram na região esperada do *Arrow plot*, adicionaram-se etiquetas aos genes de acordo com a legenda da Figura 5.2. Como se pode observar, os genes distribuem-se de acordo com o esperado, ou seja, os genes com regulação positiva

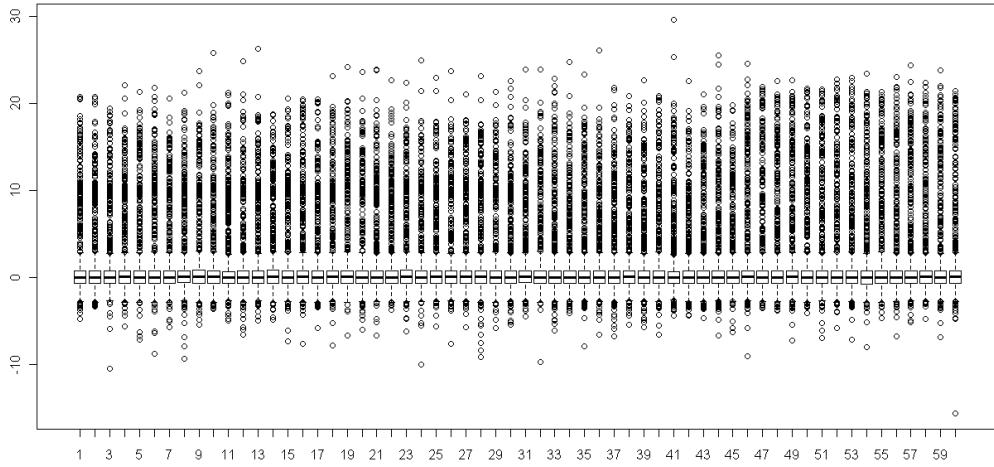


Figura 5.1: *Box plot* — dados simulados. *Box plots* 1 a 30 correspondem aos arrays da amostra controlo e os restantes à amostra experimental.

apresentam valores elevados da AUC e baixos para o OVL, os genes mistos apresentam valores da AUC em torno de 0.5 e valores baixos para o OVL, e os genes com regulação negativa, apresentam valores baixos para o OVL e AUC.

A seleção dos pontos de corte para a AUC e OVL é arbitrária, no entanto, a partir do *Arrow plot* é possível definir pontos de corte de modo a selecionar os genes de interesse. Na Figura 5.3 estão representados os pontos de corte escolhidos para a AUC e OVL e, selecionaram-se 225 genes com regulação positiva, 224 genes com regulação negativa e 35 genes candidatos a mistos². Dos genes candidatos a mistos, há 3 genes que foram simulados como não DE (pontos a laranja na Figura 5.3). Procedeu-se a uma análise da bimodalidade a partir do Algoritmo 2 e, dos 35 genes candidatos a mistos removeu-se um gene que foi simulado como não DE. Verifica-se que apesar de dois genes terem sido simulados como não DE, apresentam no entanto bimodalidade numa das densidades (de acordo com o algoritmo para identificar bimodalidade — Algoritmo 2), e como tal foram selecionados como genes mistos, facto que se comprovou pela análise das respetivas

²Os genes candidatos a genes mistos têm o valor da OVL pequeno e da AUC em torno de 0.5, sendo posteriormente classificados como mistos se apresentarem bimodalidade em pelo menos uma das distribuições dos níveis de expressão das amostras em análise.

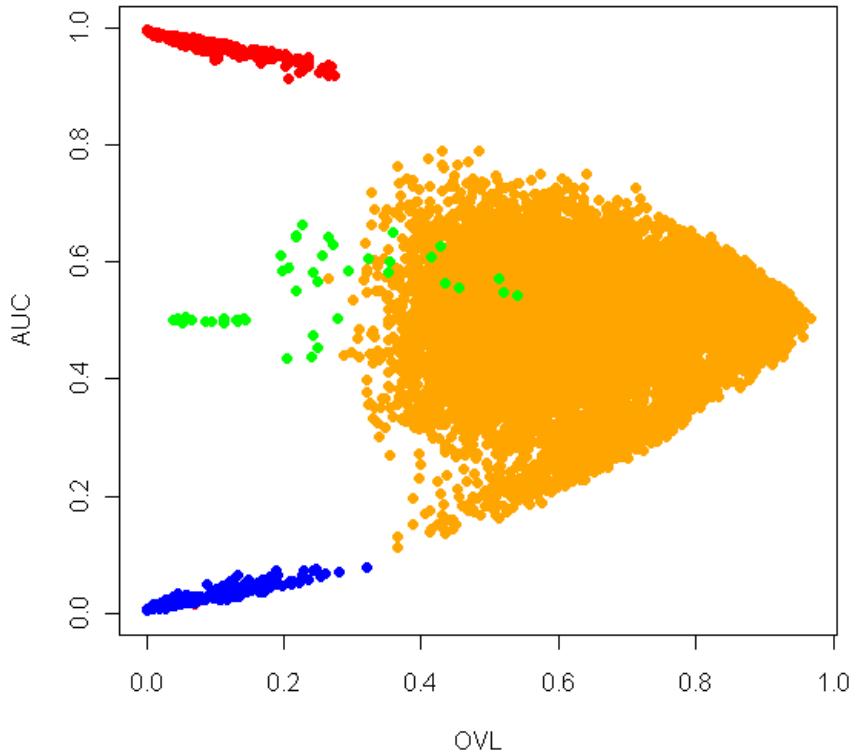


Figura 5.2: Arrow plot — dados simulados. Distribuição dos genes de acordo com o seu verdadeiro estado: pontos a laranja correspondem aos genes não DE; pontos a azul correspondem aos genes com regulação negativa; pontos a vermelho correspondem aos genes com regulação positiva; pontos verdes correspondem aos genes mistos (fonte: Silva-Fortes *et al.*, 2012).

densidades. Todos os genes simulados como mistos e que posteriormente foram candidatos a mistos foram selecionados pela análise da bimodalidade.

Conclui-se que de um total de 10000 genes simulados e, considerando os pontos de corte acima definidos, selecionaram-se 224 genes com regulação positiva, 225 genes com regulação negativa e 34 genes mistos. Todos os genes selecionados como tendo regulação positiva e negativa são verdadeiros genes com regulação positiva e negativa. Em relação aos genes mistos, 32 são verdadeiros genes mistos e 2 são verdadeiros não DE. Contudo, foram classificados como mistos de acordo com o algoritmo que permite identificar

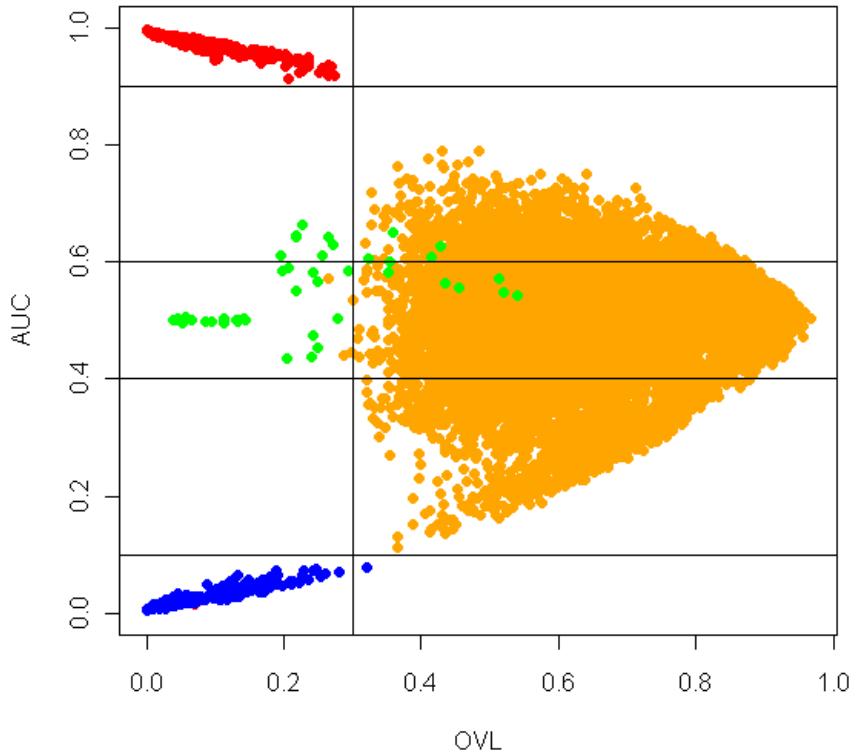


Figura 5.3: Arrow plot - dados simulados. Pontos de corte: OVL < 0.3; AUC: genes com regulação positiva: AUC > 0.9; genes com regulação negativa AUC < 0.1; genes mistos: $0.4 < \text{AUC} < 0.6$ (fonte: Silva-Fortes *et al.*, 2012).

a bimodalidade em pelo menos um dos grupos, e pela análise das densidades, foram corretamente identificados. Assim, o método proposto apresentou uma eficiência global de 99.8% na seleção de genes DE e mistos.

5.1.3 Comparação da performance com outros métodos

Com o objetivo de comparar a performance do método aqui proposto com os descritos nos capítulos 2 e 3, vai usar-se a metodologia ROC para avaliar a capacidade discriminativa, *i.e.*, se os métodos têm boa capacidade em classificar os genes de entre duas classes, genes DE ou genes não DE. Para o

efecto vão comparar-se curvas ROC e os valores das AUC de cada método, estimados pelo método empírico.

Os métodos a comparar são: *Fold Change* (FC), *Average Difference* (AD), *Weighted Average Difference* (WAD), *Rank Products* (RP), estatística-*t* de Welch (Welch-*t*), *Significance Analysis of Microarrays* (SAM), *Moderated t-statistic* (modT), *intensity-based moderated t-statistic* (ibmT), área abaixo da curva ROC estimada pelo método empírico (empAUC) e estimada pelo método do núcleo (kerAUC), SAMROC e o coeficiente de sobreposição entre duas densidades (OVL).

A capacidade discriminativa vai ser avaliada considerando dois cenários: (i) considerando que os grupos a discriminar são genes não DE e genes mistos; (ii) considerando que os grupos a discriminar são genes não DE e genes DE, onde neste caso genes DE são genes com regulação positiva, regulação negativa e genes mistos.

A construção das curvas ROC em relação aos métodos FC, AD, WAD, RP, Welch-*t*, SAM, SAMROC, ibmT, modT vai ser em função dos valores absolutos das respetivas estatísticas e, valores elevados estão relacionados com o artefacto de interesse, onde no cenário (i) são os genes mistos e, no cenário (ii) são os genes DE.

Pretende-se avaliar a capacidade discriminativa da empAUC e kerAUC, nas situações em que vulgarmente são utilizadas, *i.e.*, vai considerar-se que os valores admissíveis variam entre 0.5 e 1, ou seja, não se vão admitir curvas ROC degeneradas. Assim, valores elevados das AUC (próximos da unidade) estão associados à presença do artefacto de interesse onde no cenário (i) são os genes mistos e, no cenário (ii) são os genes DE.

Em relação ao método proposto, a construção da curva ROC vai contemplar apenas a análise do OVL. Isto porque, a capacidade discriminativa é avaliada independentemente do tipo de expressão diferencial, e valores baixos do OVL estão associados a genes DE independentemente do tipo. Assim, valores baixos do OVL estão associados à presença do artefacto de interesse, apenas genes mistos no caso (i) e genes DE no caso (ii).

Comparação da performance na seleção de genes mistos

Com o objetivo de avaliar a performance dos métodos em discriminar entre genes não DE e genes mistos, apresenta-se na Figura 5.4 as curvas ROC e na Tabela 5.1 apresentam-se os valores das respectivas AUC.

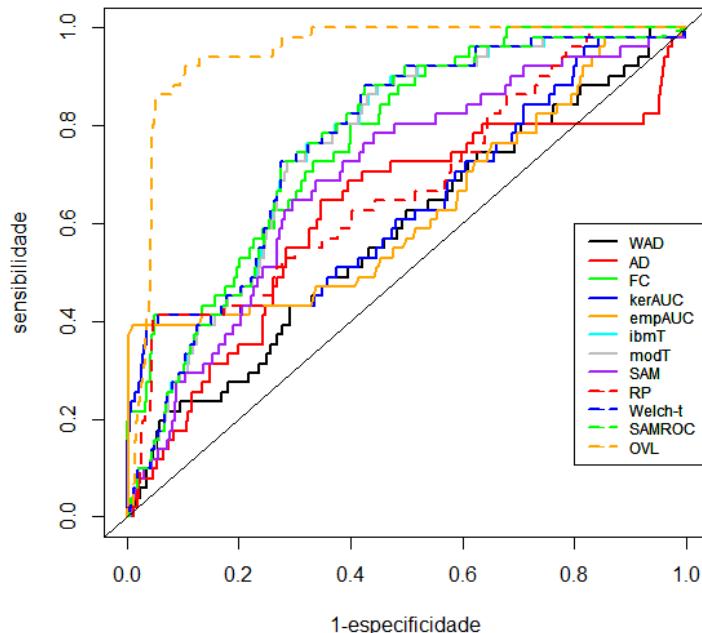


Figura 5.4: Curvas ROC empíricas. Comparação da performance na seleção de genes mistos (fonte: Silva-Fortes *et al.*, 2012).

Tabela 5.1: Comparação da performance de métodos na seleção de genes mistos — AUC. Métodos ordenados por ordem descrescente da AUC. AUC estimadas pelo método empírico.

OVL	FC	SAMROC	<i>t</i> -Welch	ibmT	modT
0.946	0.779	0.761	0.760	0.756	0.755
SAM	RP	kerAUC	empAUC	AD	WAD
0.693	0.673	0.641	0.629	0.614	0.579

Conclui-se que o método proposto é o que apresenta a melhor performance

na seleção de genes mistos com um valor da AUC perto da unidade, revelando ser um método com elevada performance na seleção de genes mistos. Os métodos FC, SAMROC, t -Welch, ibmT e modT, têm uma performance semelhante apresentando valores compreendidos entre 0.75 e 0.78 para a AUC. Os métodos SAM, RP, kerAUC, empAUC, AD e WAD foram os que apresentaram a pior performance na seleção de genes mistos.

Comparação da performance na seleção de genes DE e mistos

Com o objetivo de avaliar a performance dos métodos em discriminar entre genes não DE e genes DE e mistos. Considere-se neste caso, para efeitos de dicotomização, as classes genes não DE e genes DE, onde neste caso, considerou-se genes DE os que apresentam regulação positiva, regulação negativa e genes mistos. Apresenta-se na Figura 5.5 as curvas ROC e na Tabela 5.2 os respetivos valores das AUC.

Tabela 5.2: Comparação da performance de métodos na seleção de genes DE e mistos — AUC. Métodos ordenados por ordem descendente da AUC. AUC estimadas pelo método empírico.

OVL	RP	WAD	FC	AD	empAUC
0.998	0.969	0.959	0.953	0.939	0.937
kerAUC	SAM	ibmT	modT	t -Welch	SAMROC
0.936	0.930	0.924	0.924	0.924	0.921

Em relação à seleção de genes DE e mistos, o método proposto é o que apresenta a melhor performance, com um valor próximo da unidade para a AUC. Os restantes métodos também apresentaram valores elevados da AUC, tendo os métodos RP, WAD e FC valores semelhantes da AUC seguido dos métodos AD, empAUC, kerAUC, SAM, ibmT, modT, t -Welch e SAMROC, que apresentaram performances semelhantes, no entanto com valores ligeiramente mais baixos da AUC.

5.1.4 Arrow plot — AUC empírica vs. AUC núcleo

Compararam-se os *Arrow plots* considerando as AUC estimadas pelo método do núcleo e pelo método empírico (Figura 5.6). Não se procedeu à correção do viés, uma vez que, pela análise realizada na secção 3.3, os vieses obtidos pelos dois métodos de estimação, não apresentaram diferenças aparentemente significativas. Pelo que se observa, e como seria de esperar, as AUC estimadas

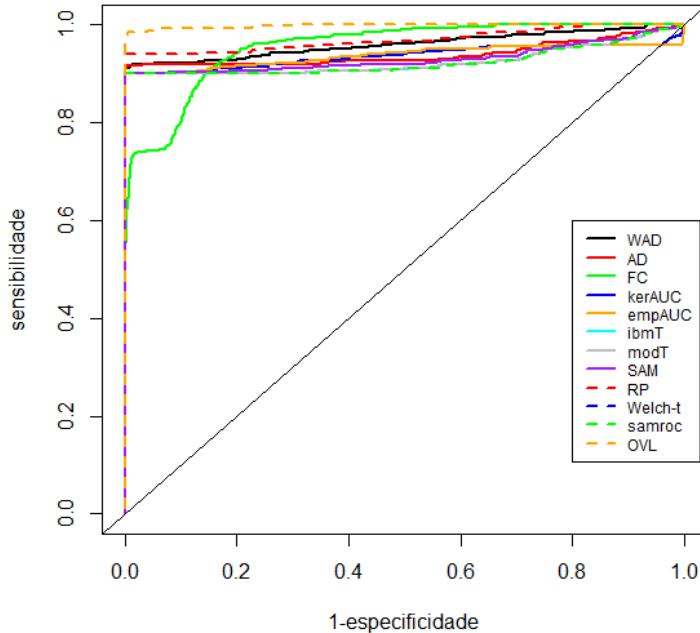


Figura 5.5: Curvas ROC empíricas. Comparação da performance de métodos na seleção de genes DE (regulação positiva, regulação negativa e mistos) (fonte: Silva-Fortes *et al.*, 2012).

pelo método empírico são mais otimistas, *i.e.*, têm tendência para sobreestimar a AUC. No entanto, as diferenças mais evidentes estão associadas à seleção de genes com regulação positiva e regulação negativa, enquanto para a seleção de genes mistos as diferenças não são significativas. Se se considerarem os pontos de corte anteriormente definidos para a AUC e OVL, o número de genes DE e mistos selecionado é o mesmo considerando os dois métodos de estimação para a AUC.

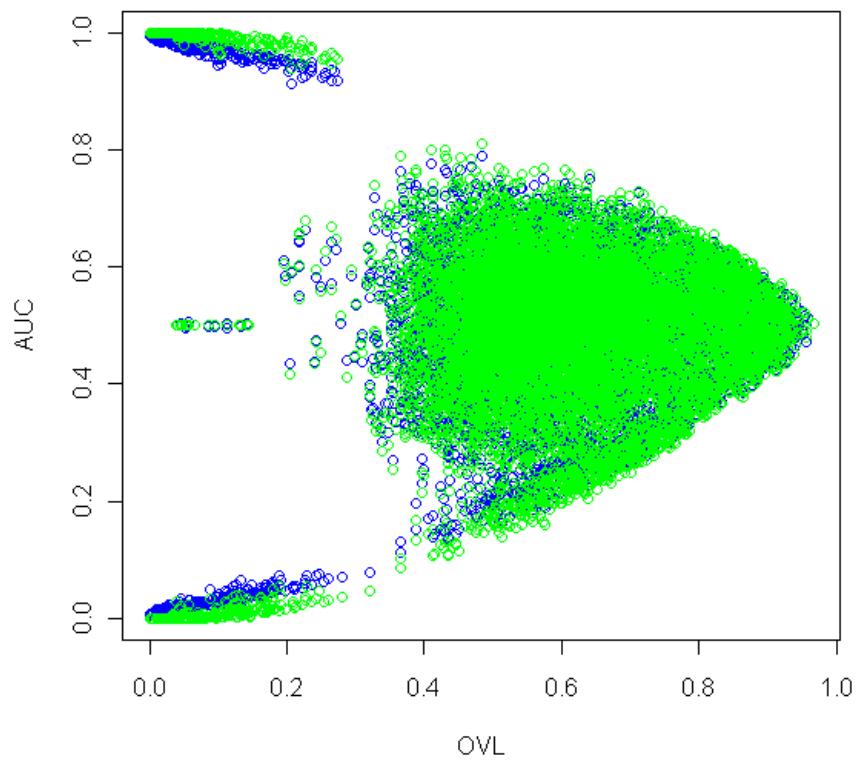


Figura 5.6: Comparação do *Arrow plot* considerando a AUC estimada pelo método empírico e pelo método do núcleo. Pontos a azul correspondem à AUC estimada pelo método do núcleo e pontos a verde correspondem à AUC estimada pelo método empírico.

5.1.5 Arrow plot vs. Volcano plot

O *Volcano plot* é um gráfico muito utilizado na análise de dados de *microarrays*. Permite identificar de uma forma rápida genes com diferentes níveis de expressão entre duas condições experimentais (Cui e Churchill, 2003). No eixo das ordenadas representam-se os valores negativos dos logaritmos na base decimal dos valores-*p* obtidos a partir de uma estatística de teste de interesse ($-\log_{10}(\text{valor-}p)$), representando a significância estatística; no eixo das abscissas representam-se os correspondentes valores do logaritmo de base 2 do FC ($\log_2(\text{FC})$), representando a significância biológica. Um gene é considerado significativo se tiver significância estatística e biológica de acordo com os pontos de corte selecionados arbitrariamente.

Na Figura 5.7 está representado o *Volcano plot* para os dados simulados, onde os valores-*p* foram obtidos a partir da estatística SAMROC. Graficamente é possível identificar apenas os genes com regulação positiva e os genes com regulação negativa. De entre os genes selecionados de acordo com os pontos de corte especificados na Figura 5.7, nenhum gene simulado como misto foi selecionado. Os genes simulados como mistos (a verde na Figura 5.7) encontram-se misturados com os genes não DE.

5.1.6 Considerações finais

A partir dos dados simulados foi possível analisar a performance do método proposto face à seleção de genes mistos e à seleção de genes DE e mistos conjuntamente. Conclui-se que o método proposto tem uma elevada performance considerando os dois cenários, e no caso particular da seleção de genes mistos os métodos usualmente utilizados têm uma fraca performance, sendo o método FC o único que selecionou genes mistos e num número não significativo.

O número de genes mistos selecionados considerando a AUC estimada pelo método do núcleo e pelo método empírico é semelhante, sendo a diferença mais evidente na seleção de genes com regulação positiva e regulação negativa, onde a AUC empírica permite selecionar um maior número de genes com estas características.

O *Volcano plot*, sendo um gráfico muito utilizado para a seleção de genes que apresentem níveis de expressão diferentes entre duas condições

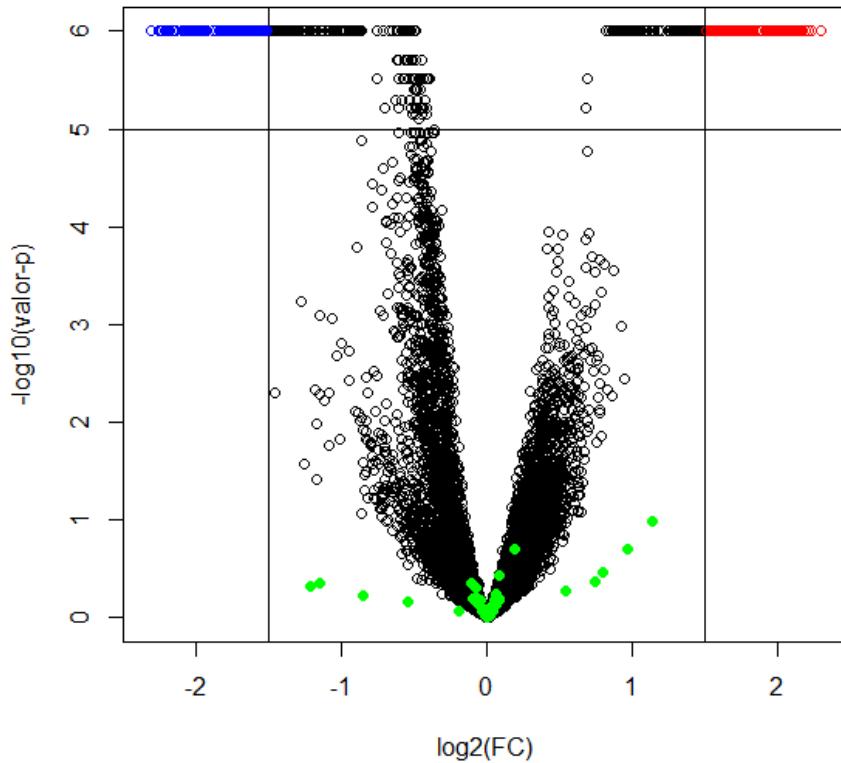


Figura 5.7: *Volcano plot* — Dados simulados. Os valores- p foram obtidos a partir da estatística SAMROC. Genes com regulação positiva correspondem aos pontos a vermelho e genes com regulação negativa correspondem aos pontos a azul. Pontos verdes representam os verdadeiros genes mistos.

experimentais, revelou-se desadequado na análise de genes mistos.

5.2 Dados *Cancro da Bexiga* — Dyrskjot *et al.* (2004)

5.2.1 Introdução

A presença de lesões de carcinomas *in situ* (CIS) na bexiga está associada a um alto risco de progressão da doença para um estado invasivo muscular.

5.2. Dados Cancro da Bexiga — Dyrskjot et al. (2004)

Dyrskjot *et al.* (2004) desenvolveram um estudo com o objetivo de identificar padrões nos níveis de expressão em células de carcinomas superficiais de transição (CST) com CIS, células CST sem CIS e em carcinomas invasivos do músculo (CIM).

Os dados são constituídos por 15 casos no grupo sem tumor (*array* C1 a C15), onde 10 casos dizem respeito a amostras sem qualquer tipo de cancro e 5 casos são provenientes de amostras normais adjacentes a CIS; 45 casos fazem parte do grupo com carcinoma, onde 15 casos são provenientes de amostras de CST sem CIS, 13 casos de amostras provenientes de CST com CIS, 13 casos provenientes de amostras com CIM e 4 casos são provenientes de amostras com CIS. Foi analisado um total de 22283 genes. Na Figura 5.8 estão ilustrados os vários tipos de tecido analisados.

Os dados estão publicados no sitio da internet <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE3167>, e são provenientes de *microarrays* da Affymetrix U133A.

No CD na pasta “Capítulo 5” encontra-se o ficheiro **bexiga.R** com o código em R que implementa a análise que a seguir se descreve.

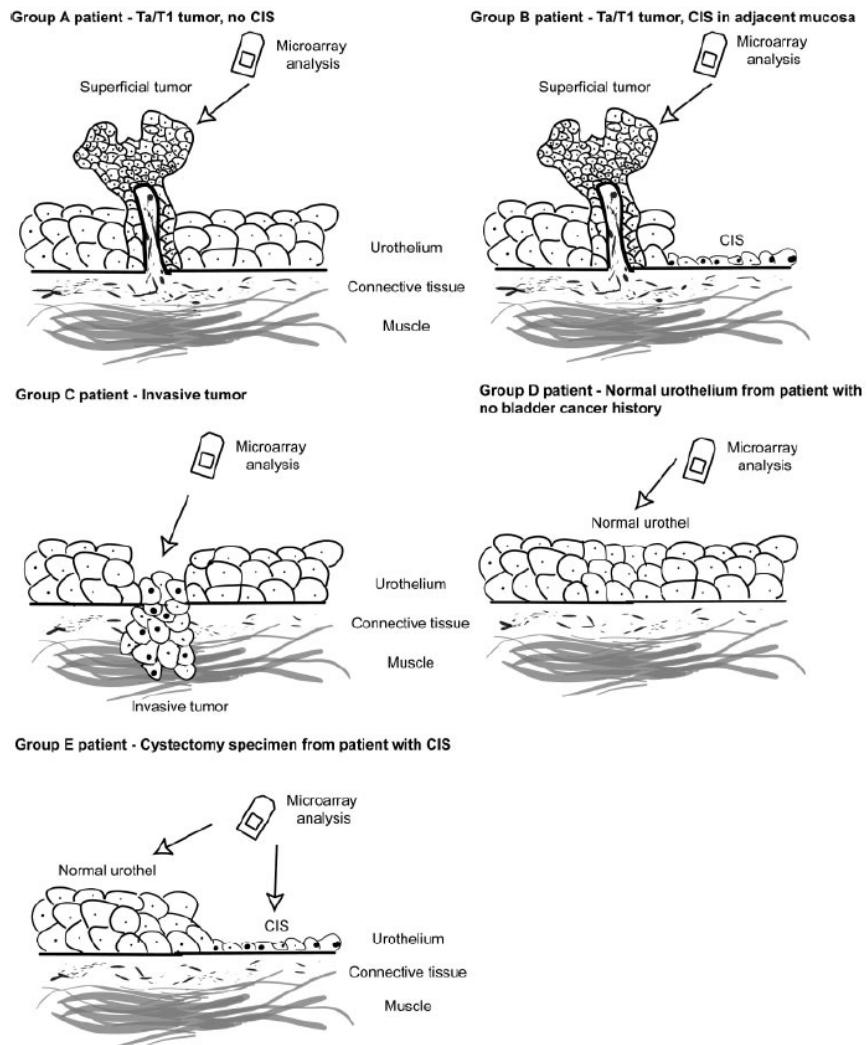


Figura 5.8: Representação dos vários tipos de amostras de tecido analisadas no estudo do *Cancro da Bexiga*. Figura adaptada de Dyrskjot *et al.* (2004).

5.2.2 Análise da qualidade dos dados

De acordo com o capítulo 2, o primeiro passo a realizar quando se tem os dados em bruto é uma análise da qualidade dos *arrays*. Iniciou-se esta análise verificando se existiam valores omissos a partir da aplicação da função `table(is.na(.))` do R, tendo-se concluído que não. De modo a verificar se se encontravam artefactos que revelassem a necessidade de se remover algum *array*, procedeu-se à análise das imagens dos 60 *arrays*, tendo-se optado por colocá-las no CD na pasta “Capítulo 5” (`image1cancro.png`, `image2cancro.png`, `image3cancro.png` e `image4cancro.png`). Pela análise das imagens não há nenhum *array* que apresente algum tipo de artefacto que justifique a sua remoção.

A partir da análise das densidades (Figura 5.9) e dos *box plots* (Figura 5.10) dos logaritmos dos níveis de intensidade PM dos 60 *arrays*, verifica-se que estes devem de ser submetidos a um processo de normalização, destacando-se o *array* C9 apresentando um comportamento mais diferenciado dos restantes.

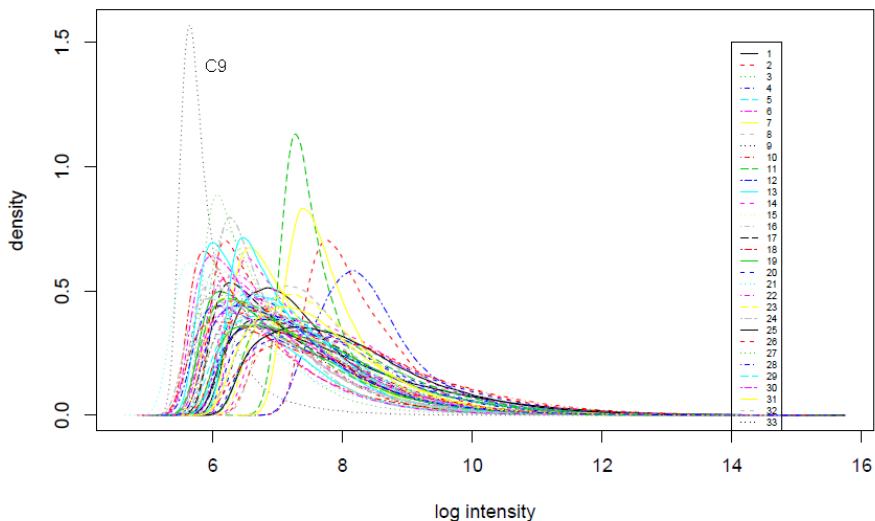


Figura 5.9: Densidades dos logaritmos dos níveis de intensidade PM em bruto.

Pela análise do *degradation plot* (Figura 5.11) pode concluir-se que o material genético é de boa qualidade, uma vez que as linhas representativas dos *arrays* são paralelas e com declive positivo.

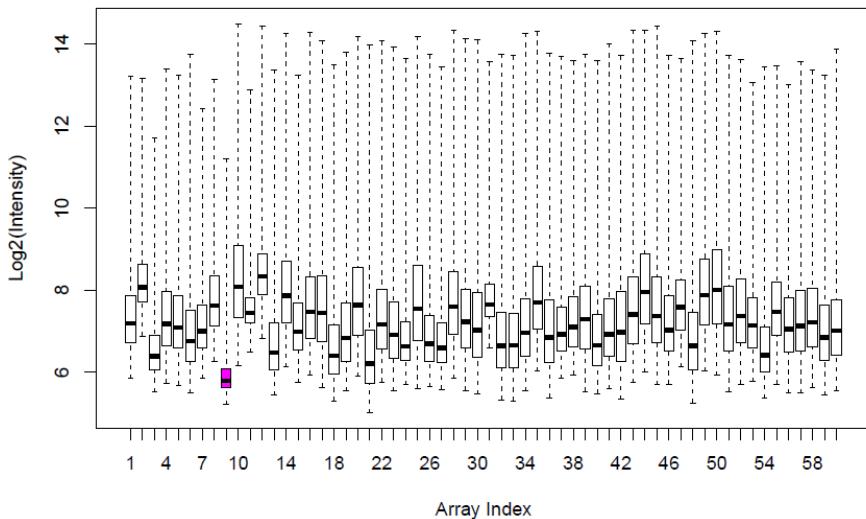


Figura 5.10: Box plot dos logaritmos dos níveis de intensidade PM dos arrays relativos ao estudo do *Cancro da Bexiga*. Array C9 a rosa.

Na Figura 5.12 estão representados os gráficos MA dos arrays C1 a C6 (não se justificando a necessidade de apresentar os gráficos MA para todos os arrays), verificando-se um desvio do ajustamento da curva *lowess*, que revela a necessidade de se proceder à normalização dos arrays.

Os gráficos NUSE e RLE representados nas Figuras 5.13 e 5.14 respetivamente, revelam que o array C9 encontra-se mais afastado dos restantes, indicando má qualidade.

Da análise do gráfico QC (Figura 5.15), destacam-se os arrays C9 e E45, revelando a necessidade de se proceder à sua remoção.

Assim, procedeu-se em primeiro lugar à remoção do array C9, uma vez que foi o que mais se destacou pela análise anterior. No apêndice B.2 apresenta-se o gráfico QC dos dados em bruto após remoção do array C9, e pelo que se pode observar não há mais nenhum array que revele necessidade de ser removido.

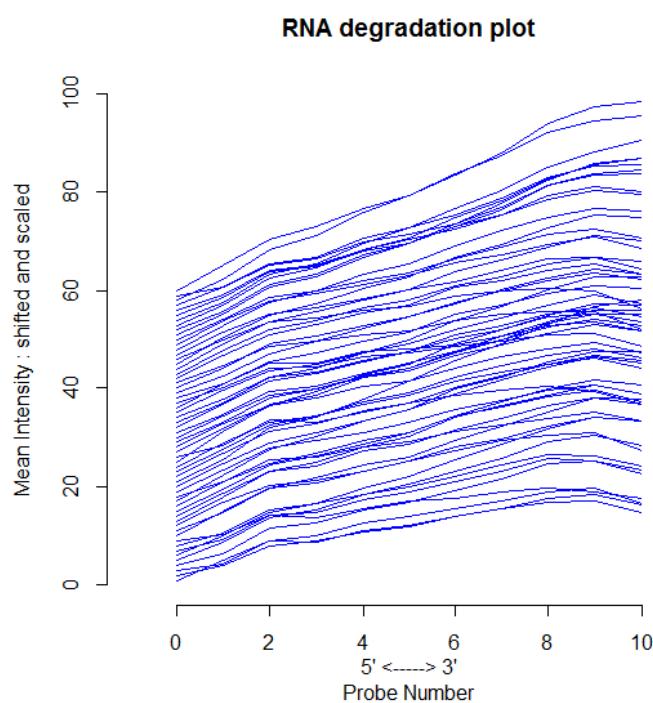


Figura 5.11: *Degradation plot.*

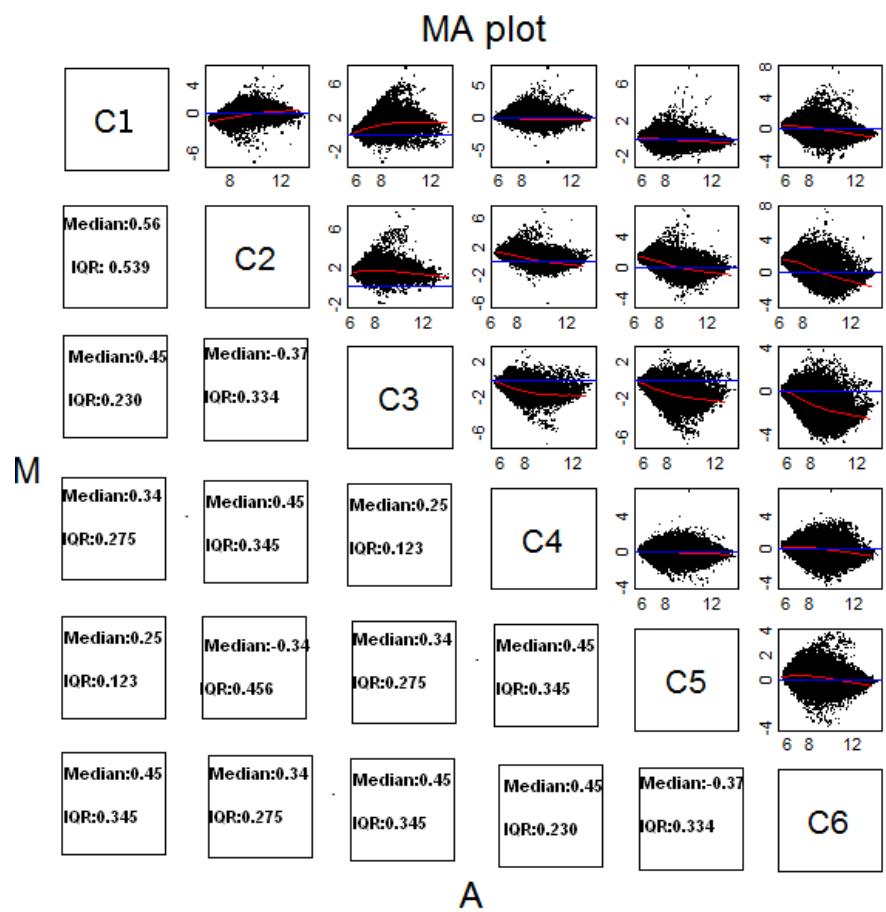


Figura 5.12: Gráficos MA dos arrays C1 a C6 — dados em bruto.

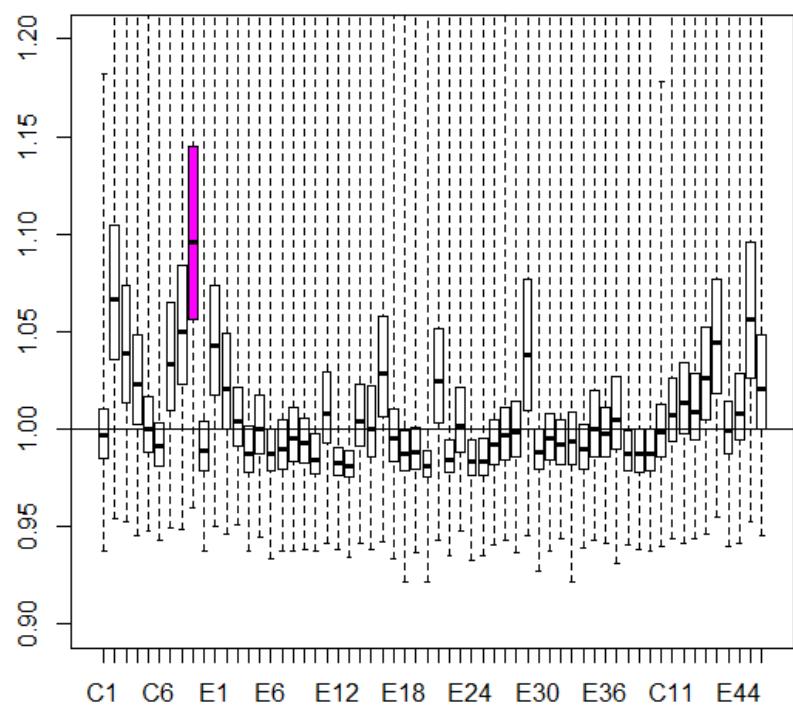


Figura 5.13: Gráfico NUSE. *Array C9 a rosa.*

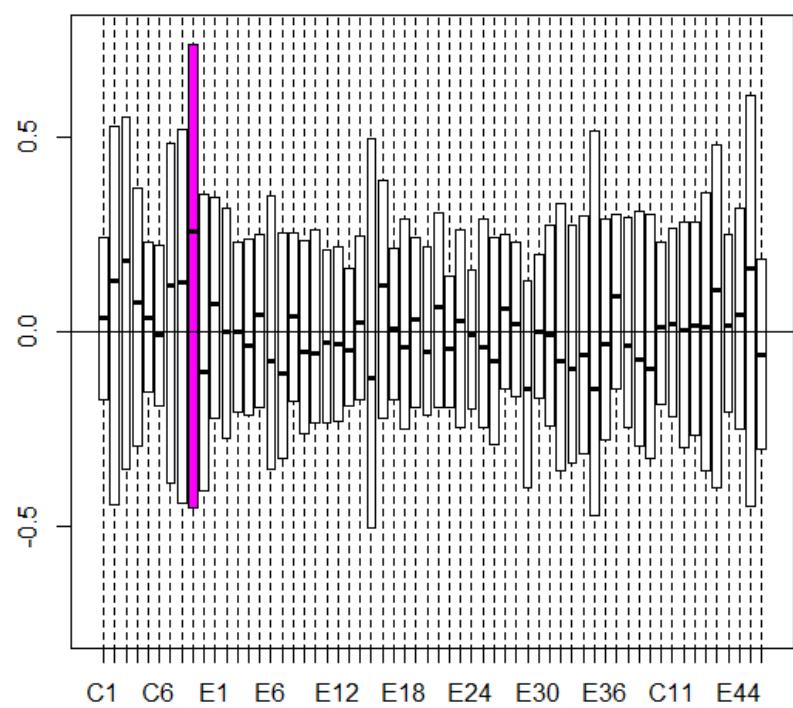


Figura 5.14: Gráfico RLE. *Array C9 a rosa.*

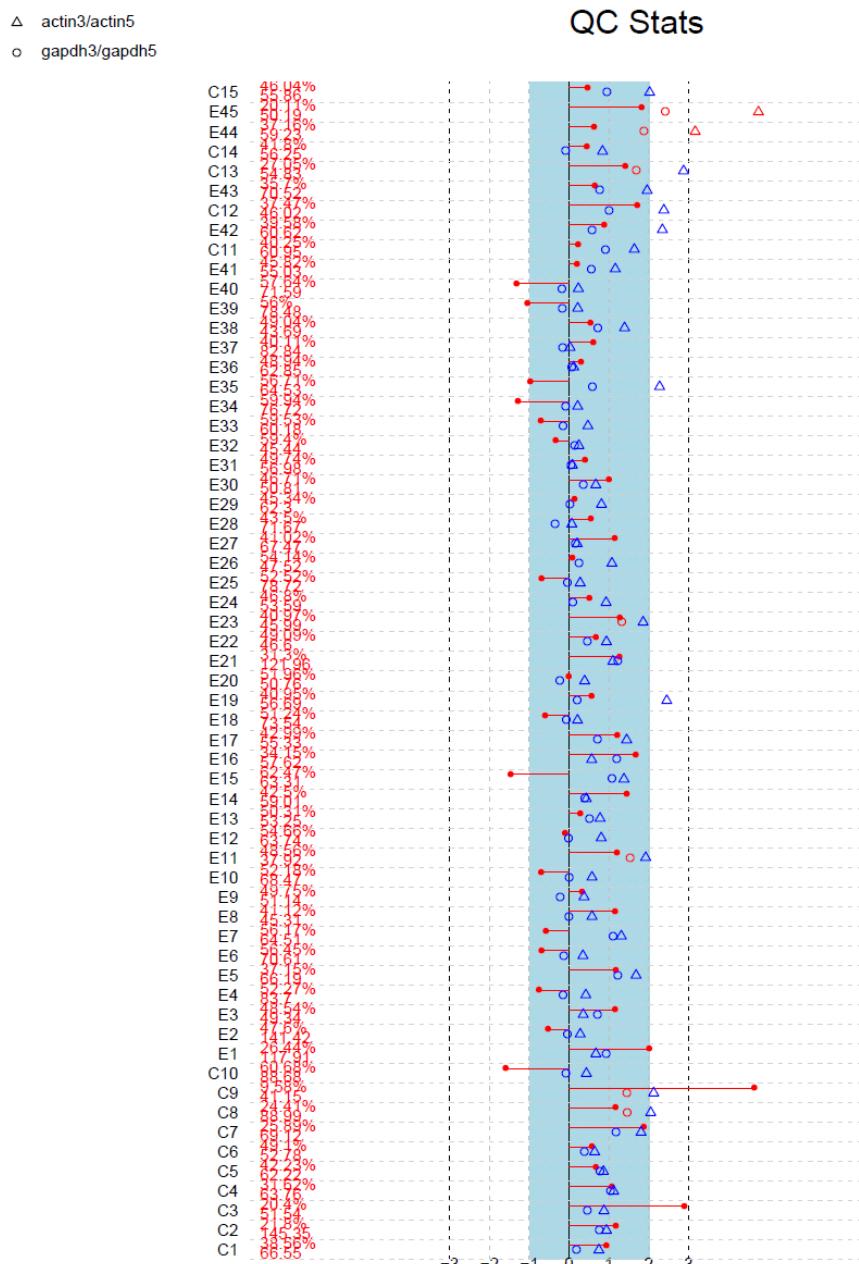


Figura 5.15: Gráfico QC dos dados em bruto.

5.2.3 Pré-processamento

Da análise anterior concluiu-se que os *arrays* devem de ser submetidos a uma análise de pré-processamento. Para o efeito foram aplicados os métodos de pré-processamento RMA, GC-RMA, FARMS, MAS5, PLIER e MBEI. Nas Figuras 5.16, 5.17 e 5.18 estão representados os logaritmos dos níveis de expressão dos 60 *arrays* após pré-processamento. Pela sua análise, os métodos RMA e FARMS revelam ter o melhor comportamento. Com o objetivo de verificar qual dos dois métodos se vai considerar na análise subsequente, apresenta-se na Figura 5.19 as densidades dos logaritmos dos níveis de expressão dos *arrays* após pré-processamento RMA e FARMS, e, da sua análise verifica-se que o método FARMS é o que produz densidades mais semelhantes.

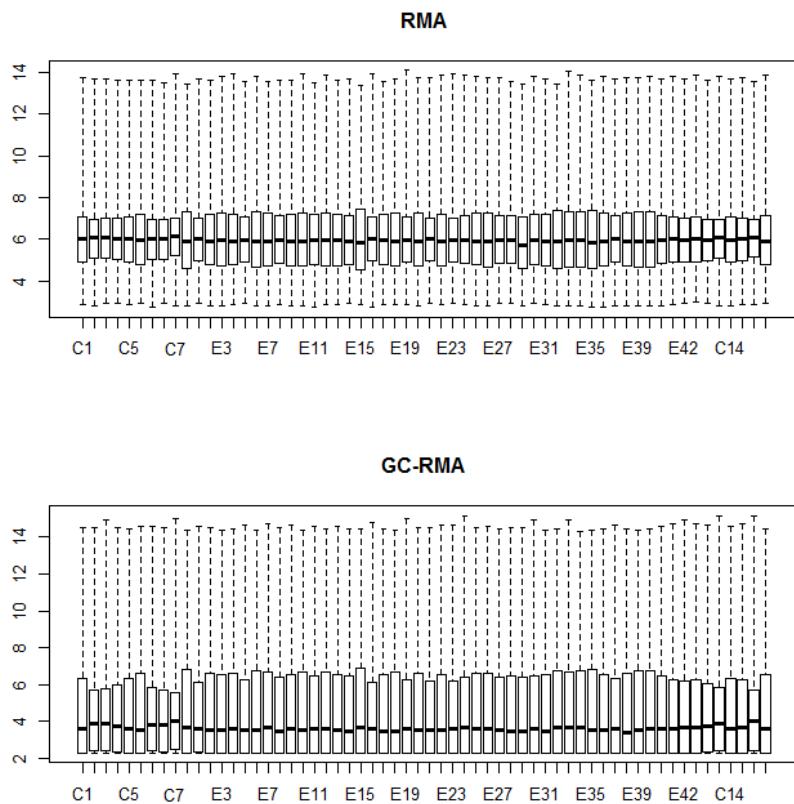


Figura 5.16: *Box plots* dos logaritmos dos níveis de expressão dos *arrays* após pré-processamento RMA e GCRMA.

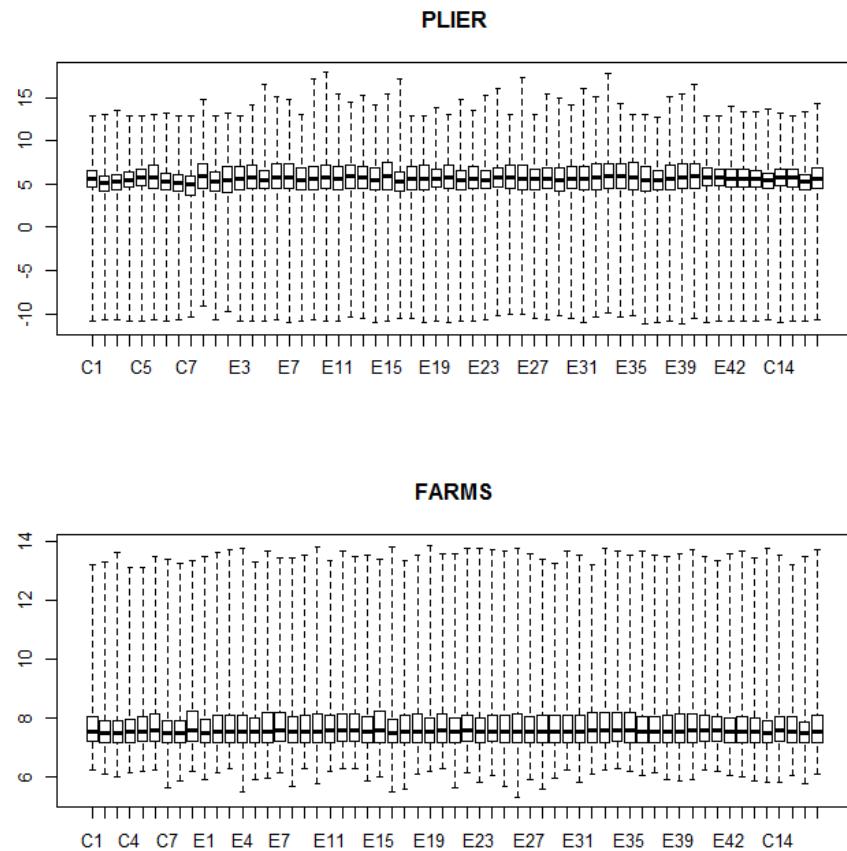


Figura 5.17: *Box plots* dos logaritmos dos níveis de expressão dos *arrays* após pré-processamento PLIER e FARMS.

Na Figura 5.20 apresenta-se o gráfico MA após pré-processamento FARMS e verifica-se um ajustamento em torno de $M=0$.

Em resumo, da análise da qualidade dos dados e da análise pré-processamento, removeu-se o *array* C9 da base de dados e os dados foram submetidos ao método de pré-processamento FARMS.

5.2.4 Seleção de genes DE e mistos

Na Figura 5.21 apresenta-se o *Arrow plot* para os 22283 genes e 59 *arrays*, onde a AUC foi estimada pelo método do núcleo.

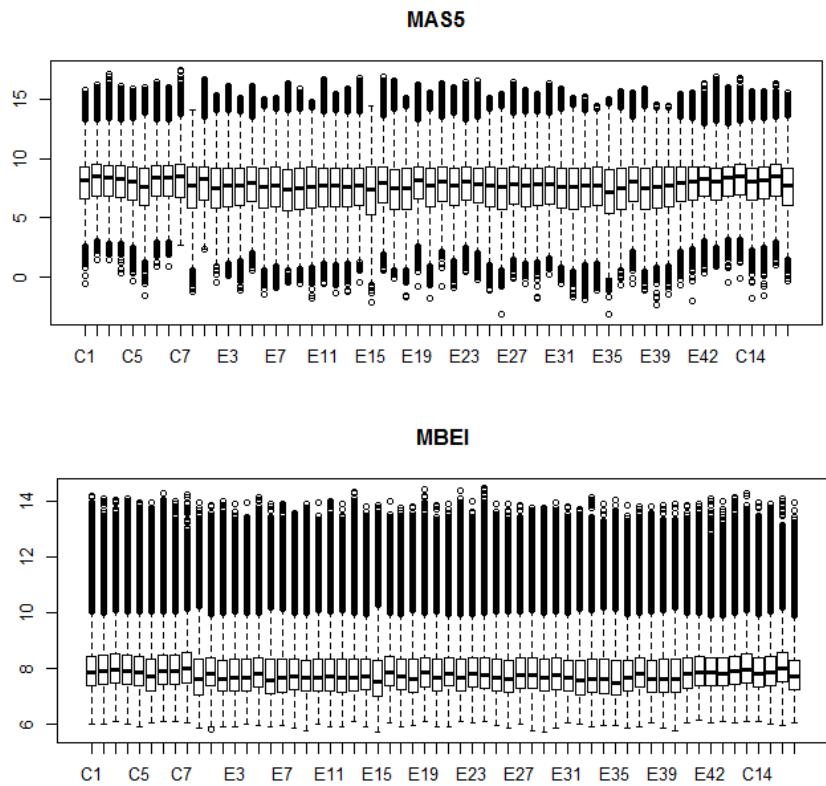


Figura 5.18: *Box plots* dos logaritmos dos níveis de expressão dos *arrays* após pré-processamento MAS5 e MBEI.

Pela análise da Figura 5.21 selecionaram-se os seguintes pontos de corte para o OVL e AUC: $OVL \leq 0.4$; para selecionar genes com regulação positiva considerou-se $AUC > 0.9$; para selecionar genes com regulação negativa considerou-se $AUC < 0.1$ e para selecionar genes mistos considerou-se $0.4 < AUC < 0.6$ (Figura 5.22). Após uma análise da bimodalidade dos genes candidatos a mistos, apenas um não revelou bimodalidade em ambos os grupos.

De acordo com os pontos de corte acima definidos, selecionaram-se 10 genes mistos, 20 genes com regulação positiva e 52 genes com regulação negativa. Na Tabela 5.3 apresentam-se os genes mistos e os valores da AUC e OVL. Procedeu-se a uma comparação dos resultados aqui obtidos com os resultados obtidos no estudo de Dyrskjot *et al.* (2004). No entanto, há que salientar que

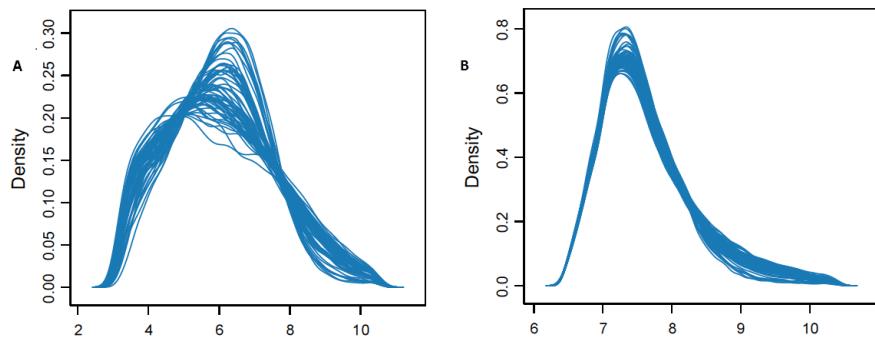


Figura 5.19: Comparação das densidades dos logaritmos dos níveis de expressão dos *arrays* após RMA (A) e FARDS (B). Escala logarítmica.

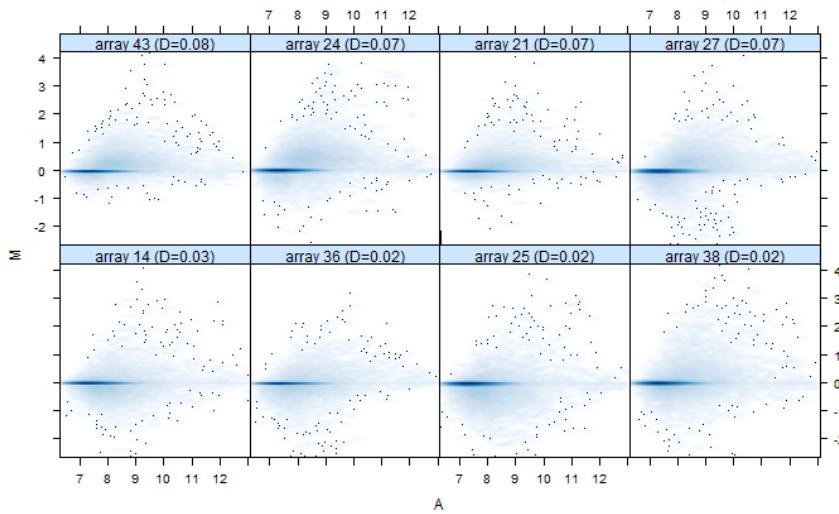


Figura 5.20: Gráfico MA apóis pré-processamento FARDS considerando a base de dados sem o *array* C9.

os dados foram sujeitos a um processo de filtragem, foram pré-processados com o método RMA e não houve indicação de remoção de algum *array*. Aplicaram um teste de permutações com base na estatística-*t* e selecionaram os primeiros 50 genes com regulação positiva e 50 genes com regulação negativa. Comparando os resultados com os obtidos por Dyrskjot *et al.* (2004), nenhum dos genes mistos foram referidos no estudo, *i.e.*, não foram selecionados nem como tendo regulação positiva nem negativa. Os genes

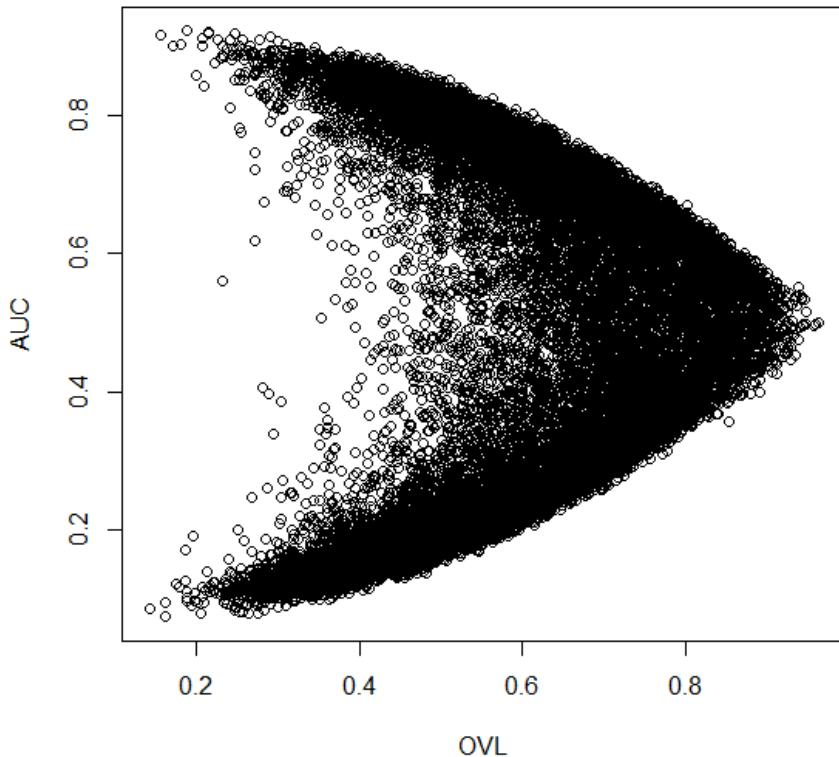


Figura 5.21: *Arrow plot* — Dados *Cancro da Bexiga*. AUC estimada pelo método do núcleo.

selecionados com regulação positiva e negativa pelo *Arrow plot* também foram selecionados no estudo de Dyrskjot *et al.* (2004).

Conclui-se que os genes mistos foram selecionados apenas pelo *Arrow plot*.

5.2.5 Comparação com outros métodos

Ordenaram-se por ordem decrescente os valores absolutos das estatísticas kerAUC, empAUC, WAD, FC, AD, ibmT, modT, samT, SAMROC e *t*-Welch e selecionaram-se os 82 primeiros genes da lista (sendo este o número de genes selecionado a partir do *Arrow plot*). Cruzou-se a lista de genes mistos e DE selecionados pelo novo método com os primeiros 82

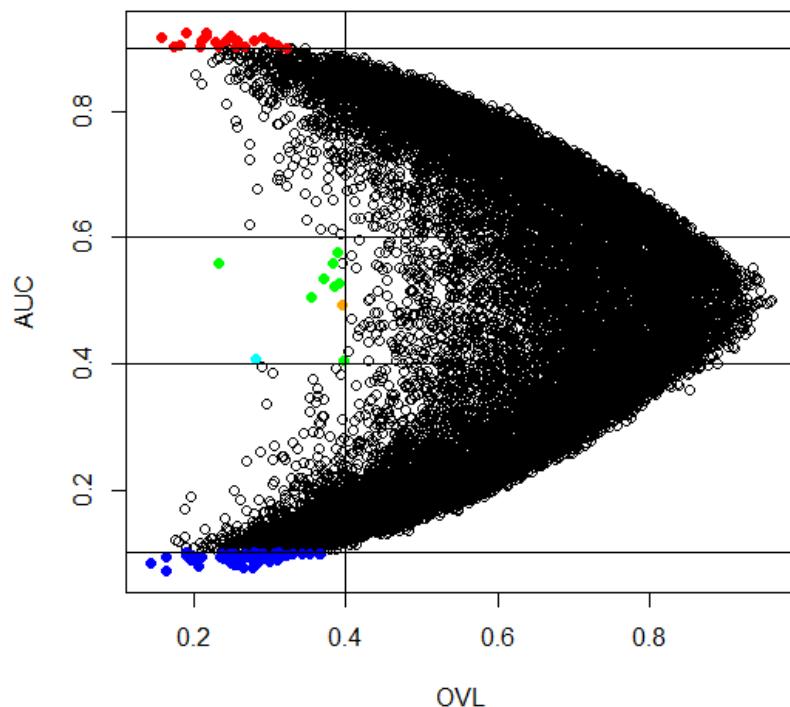


Figura 5.22: Arrow plot — Dados *Cancro da Bexiga*. Pontos a vermelho: genes com regulação positiva; pontos a azul: genes com regulação negativa; pontos a verde: genes mistos com bimodalidade em ambos os grupos; pontos a azul claro: genes mistos com bimodalidade na amostra sem cancro; pontos a laranja: genes mistos com bimodalidade no grupo com cancro.

Tabela 5.3: Lista de genes mistos nos dados *Cancro da Bexiga*. Valores dos pontos de corte para a seleção de genes mistos: $0.4 < \text{AUC} < 0.6$ e $\text{OVL} < 0.4$. C—bimodalidade no grupo sem cancro, E—bimodalidade no grupo com cancro, B—bimodalidade em ambos os grupos. Os genes estão ordenados por ordem crescente do OVL.

ID do Gene	Nome do gene	AUC	OVL	Grupo
214595_at	<i>KCNG1</i>	0.5599368	0.233	B
207818_s_at	<i>HTR7</i>	0.4066269	0.282	C
213076_at	<i>D38169</i>	0.5060773	0.354	B
218638_s_at	<i>SPON2</i>	0.5343976	0.371	B
209942_x_at	<i>MAGEA3</i>	0.5581244	0.383	B
218888_s_at	<i>FLJ10430</i>	0.5231348	0.385	B
205927_s_at	<i>CTSE</i>	0.5772683	0.390	B
216074_x_at	Semelhante a <i>KIAA0869</i>	0.5259593	0.391	B
218918_at	<i>HMIC</i>	0.4933609	0.396	E
210674_s_at	<i>PCDH</i>	0.4054118	0.397	B

genes obtidos para cada método e, apenas o FC selecionou 2 genes como DE da lista de genes mistos selecionados pelo *Arrow plot*, e dos 72 genes DE selecionados pelo método proposto, obtiveram-se os seguintes resultados para cada método: kerAUC — 72; empAUC — 57; modT — 52; ibmT e samT — 47; *t*-Welch — 42; SAMROC — 31; AD — 7; WAD e FC — 6.

5.2.6 Arrow plot — AUC empírica vs. AUC núcleo.

Na Figura 5.23 estão representados os *Arrow plots* sobrepostos, onde se considerou a AUC estimada pelo método do núcleo e pelo método empírico. Considerando a análise realizada na secção 3.3 decidiu-se não se proceder a uma correção do viés dos valores das AUC estimadas pelos métodos empírico e do núcleo.

Pode observar-se que as diferenças significativas são em relação ao número de genes selecionados com regulação positiva e regulação negativa (Tabela 5.4). O método empírico revela-se muito mais otimista selecionando mais genes, no entanto no contexto da seleção de genes DE, espera-se que este número seja muito reduzido, na ordem de 1% a 5%. Esta discrepância tão grande não era

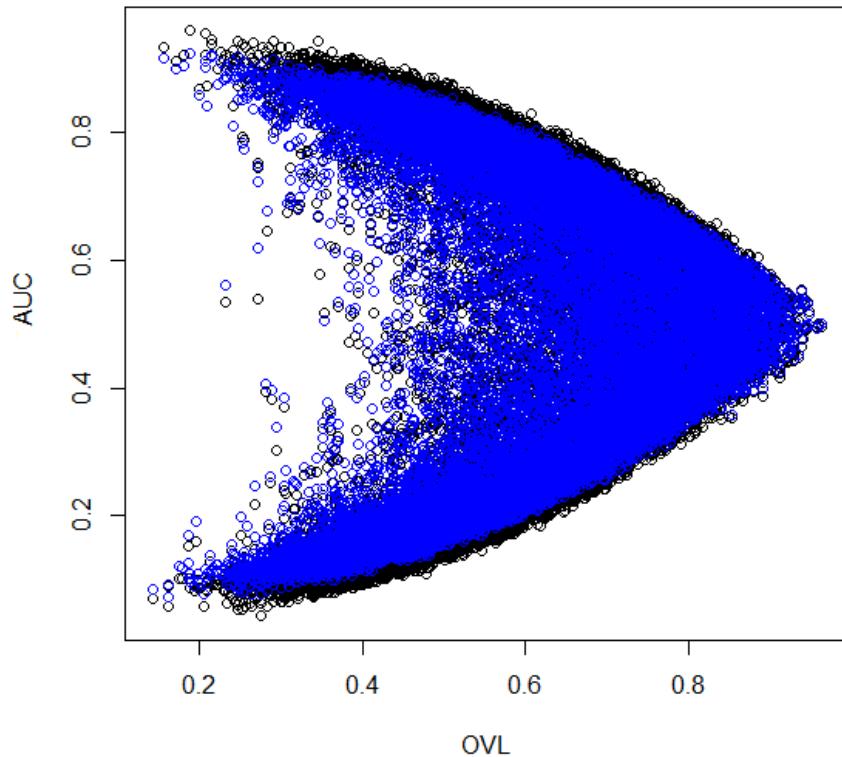


Figura 5.23: *Arrow plot* — AUC empírica *vs.* AUC núcleo. Pontos a azul representam genes onde a AUC foi estimada pelo método do núcleo e pontos a preto onde a AUC foi estimada pelo método empírico.

de esperar após o estudo de simulação anterior. Este problema necessita de uma investigação mais profunda e é com certeza um tema de trabalho futuro. Métodos *bootstrap* para correção de viés podem ter alguma relevância.

Tabela 5.4: Comparação do número de genes selecionados considerando o *Arrow plot* com AUC estimada pelos métodos do núcleo e empírico, para os pontos de corte definidos anteriormente.

Número de genes	AUC núcleo	AUC empírica
Mistos	10	11
Regulação positiva	20	136
Regulação negativa	52	468

5.2.7 Considerações finais

A partir da análise do *Arrow plot* com AUC estimada pelo método do núcleo e com os pontos de corte definidos, selecionou-se 10 genes mistos que não foram selecionados no estudo de Dyrskjot *et al.* (2004). Apenas o método FC selecionou 2 genes como sendo DE da lista de genes mistos obtida pelo *Arrow plot*. Em relação ao número de genes DE com regulação positiva e regulação negativa, a estatística kerAUC (AUC estimada pelo método do núcleo) selecionou os mesmos genes (como era expectável), e os métodos com o menor número de genes DE selecionados em comum foram o AD, WAD e FC.

O número de genes mistos considerando a AUC empírica e estimada pelo método do núcleo diferiu apenas em um gene, onde as diferenças significativas foram relacionadas com o número de genes com regulação positiva e negativa, onde o *Arrow plot* considerando a AUC empírica permite selecionar um maior número de genes (7 ou 9 vezes mais) com essas características. Uma vez que se espera que o número de genes DE a selecionar seja reduzido, o *Arrow plot* considerando a AUC estimada pelo método do núcleo, parece ir mais ao encontro do que é expetável em termos biológicos.

5.3 Dados *Linfoma* — Alizadeh *et al.* (2000)

5.3.1 Introdução

Alizadeh *et al.* (2000) desenvolveram um estudo com o objetivo de avaliar a heterogeneidade em linfomas não-Hodgkin. Pacientes a quem foram diagnosticados o mesmo tumor têm percursos clínicos diferentes e respostas diferentes aos vários tratamentos. Alizadeh *et al.* (2000) defendem que a classificação de tumores com base na análise da expressão genética permite identificar previamente importantes subtipos de cancro não detetáveis clinicamente.

Para o efeito construíram *microarrays* de cDNA mistos designados de *Lymphochips*, onde utilizam genes expressos preferencialmente em células linfáticas e genes com um papel importante (ou com suspeita de um papel importante) em processos imunológicos ou oncológicos. Utilizaram estes *microarrays* para caracterizar padrões de expressão genética de tumores

Tabela 5.5: *Arrays* selecionados da base de dados original do estudo de Alizadeh *et al.* (2000).

<i>Arrays</i> Controlo	<i>Arrays</i> Experimental
B_i $i=1, \dots, 14$	FL9, FL9CD19, FL12, FL10CD19, FL10, FL11, FL11CD19 FL6, FL5, CLL60, CLL68, CLL9, CLL14, CLL51 CLL65, CLL71Richter, CLL71, CLL13, CLL39, CLL52

linfáticos nas três situações mais prevalentes em adultos: DLBCL³ (linfoma difuso de células B grandes), FL⁴ (linfoma folicular) e B-CLL⁵ (leucemia linfocítica crónica de células B).

A base de dados aqui analisada é uma parte da base de dados original, coincidindo com a utilizada no estudo de Parodi *et al.* (2008). Inclui níveis de expressão de 4026 genes, onde a amostra controlo é constituída por: 14 casos com células B normais (CBN), onde 6 casos as células foram altamente estimuladas e 9 casos foram ligeiramente estimuladas ou não foram estimuladas. O grupo experimental é constituído por 20 casos onde 9 correspondem ao tipo FL e 11 correspondem ao tipo B-CLL. Como se pode verificar em ambos os grupos verifica-se a existência de subpopulações. Os dados de Alizadeh *et al.* (2000) encontram-se disponíveis no sítio da internet <http://llmp.nih.gov/lymphoma/data/figure1/>. Na Tabela 5.5 apresentam-se os *arrays* selecionados da base de dados original.

Apesar dos níveis de expressão serem provenientes de *microarrays* de cDNA, as amostras controlo e experimental foram hibridadas separadamente em *microarrays* distintos, tal como nos desenhos experimentais dos *arrays* de um canal. O objetivo é que as amostras controlo e experimental hibridem com o material genético dos *Lymphochips*. Os dados refletem a abundância relativa de cada gene em cada condição experimental com a referência. As razões dos níveis de expressão foram sujeitos a uma transformação logarítmica de base 2. Os valores omissos foram estimados de acordo com método k-NN (secção 2.2.2).

Na Figura 5.24 apresentam-se os *box plots* dos logaritmos dos níveis de expressão dos *arrays* descritos na Tabela 5.5, e pela sua análise podemos

³Do inglês *diffuse large B-cell lymphoma*.

⁴Do inglês *follicular lymphomas*.

⁵Do inglês *chronic lymphocytic leukemia*.

concluir que as distribuições são aproximadamente simétricas e semelhantes entre si. Decidiu-se não se proceder a transformações para que as condições sejam as mesmas do estudo de Parodi *et al.* (2008).

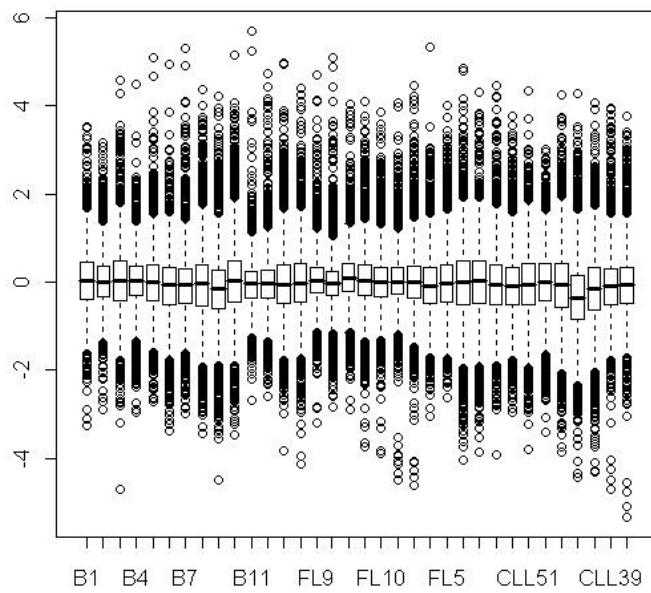


Figura 5.24: *Box plots* dos logaritmos dos níveis de expressão dos dados *Linfoma*.

No CD na pasta “Capítulo 5” encontra-se o ficheiro `linfoma.R` com o código R que implementa a análise que a seguir se descreve.

5.3.2 Seleção de genes DE e mistos

Para a construção do *Arrow plot* (Figura 5.24) estimaram-se os valores da AUC pelo método empírico, uma vez que nos métodos propostos por Parodi *et al.* (2008) as curvas ROC são estimadas por este método.

Para a seleção de genes DE e genes mistos consideraram-se valores do $OVL < 0.5$, e em relação à AUC consideraram-se os seguintes pontos de

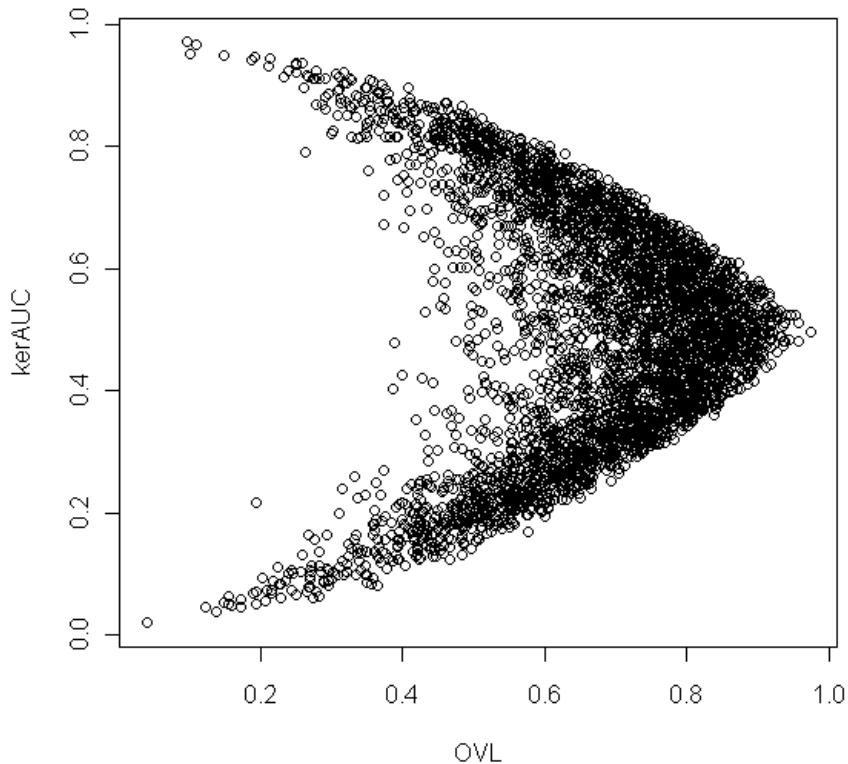


Figura 5.25: *Arrow plot* dos dados *Linfoma* (fonte: Silva-Fortes *et al.*, 2012).

corte: para selecionar genes com regulação positiva - $AUC > 0.9$, para selecionar genes com regulação negativa - $AUC < 0.1$ e para a seleção de genes mistos considerou-se $0.4 < AUC < 0.6$. A seleção dos pontos de corte para a AUC e OVL foi feita com base na análise do *Arrow plot* (Figura 5.25). Na Figura 5.26 apresenta-se o *Arrow plot* com a seleção de genes DE e mistos de acordo com os pontos de corte definidos anteriormente.

Um total de 68 genes apresentou regulação positiva e 90 genes regulação negativa. Vinte genes foram selecionados como candidatos a mistos e após uma análise da bimodalidade, foram todos selecionados como mistos, onde 2 apresentaram bimodalidade na amostra controlo, 5 na amostra experimental e 13 genes apresentaram bimodalidade nos dois grupos. Na Tabela 5.6 apresenta-se a lista de genes mistos e respetivos valores da AUC e OVL.

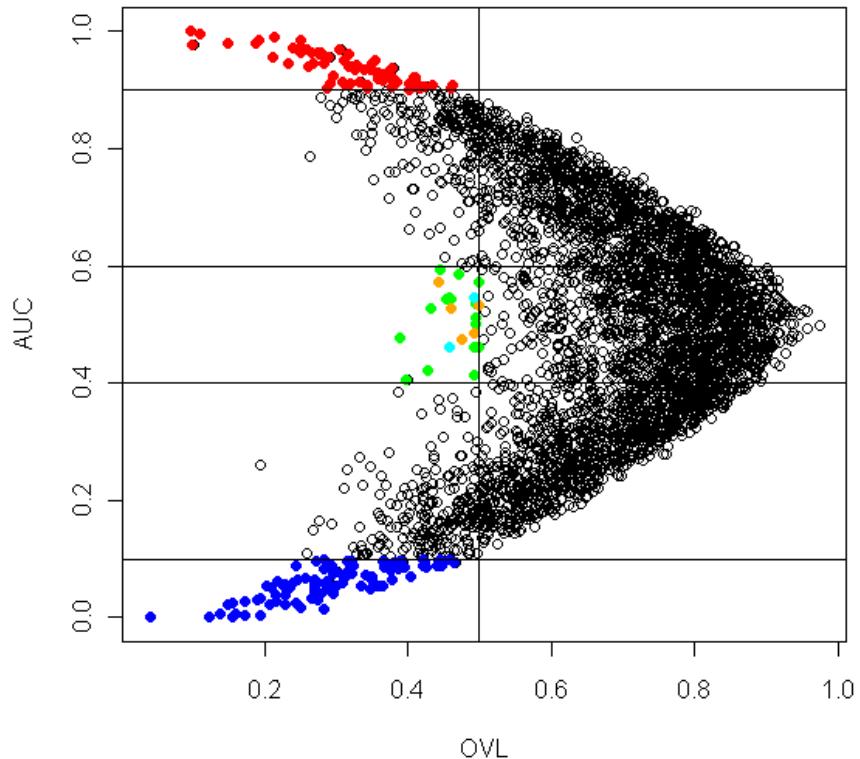


Figura 5.26: Arrow plot dos dados *Linfoma*. Pontos a vermelho representam os genes com regulação positiva, pontos a azul representam os genes com regulação negativa. Os genes mistos com bimodalidade no grupo experimental estão representados a laranja, os pontos a azul claro representam os genes mistos com bimodalidade na amostra controlo e os pontos a verde representam os genes mistos com bimodalidade em ambos os grupos.

Na Figura 5.27 apresentam-se as densidades dos grupos controlo e experimental estimadas pelo método do núcleo e respetivas curvas ROC empíricas dos 20 genes mistos.

Tal como esperado, todos os genes mistos apresentam curvas ROC sigmoidais e pela análise das densidades verifica-se a existência de bimodalidade em pelo menos um dos grupos.

Tabela 5.6: Lista de genes mistos dos dados *Linfoma*. Valores dos pontos de corte para a seleção de genes mistos: $0.4 < \text{AUC} < 0.6$ e $\text{OVL} < 0.5$. C — bimodalidade no grupo controlo, E — bimodalidade no grupo experimental, B — bimodalidade em ambos os grupos. Os genes estão ordenados por ordem crescente do OVL.

ID do Gene	Nome do gene	OVL	AUC	Grupo
GENE3323X	<i>BCL7A</i>	0.389	0.477	B
GENE3473X	desconhecido	0.399	0.407	B
GENE1877X	desconhecido	0.428	0.421	B
GENE3388X	<i>Cadeia J da imunoglobulina</i>	0.432	0.529	B
GENE1141X	<i>MAPKKK5</i>	0.443	0.571	E
GENE3521X	semelhante a <i>KIAA0050</i>	0.446	0.593	B
GENE3407X	<i>Histona desacetilase 3</i>	0.453	0.543	B
GENE75X	<i>VRK2 cinase</i>	0.457	0.546	C
GENE2519X	desconhecido	0.461	0.529	E
GENE3343X	<i>LR11</i>	0.461	0.543	B
GENE1817X	<i>BL34</i>	0.472	0.586	B
GENE3389X	<i>Cadeia J da imunoglobulina</i>	0.476	0.475	E
GENE3909X	<i>Bicurina placentária</i>	0.492	0.463	C
GENE2887X	<i>LBR</i>	0.492	0.486	E
GENE3547X	<i>Cadeia leve kappa da imunoglobulina</i>	0.493	0.413	B
GENE1004X	<i>BNIP3</i>	0.494	0.511	B
GENE2547X	<i>CLK3 cinase</i>	0.495	0.500	B
GENE2778X	<i>DNA Ligase III</i>	0.496	0.536	B
GENE3322X	<i>BCL7A</i>	0.498	0.532	E
GENE463X	<i>PARP</i>	0.499	0.461	B

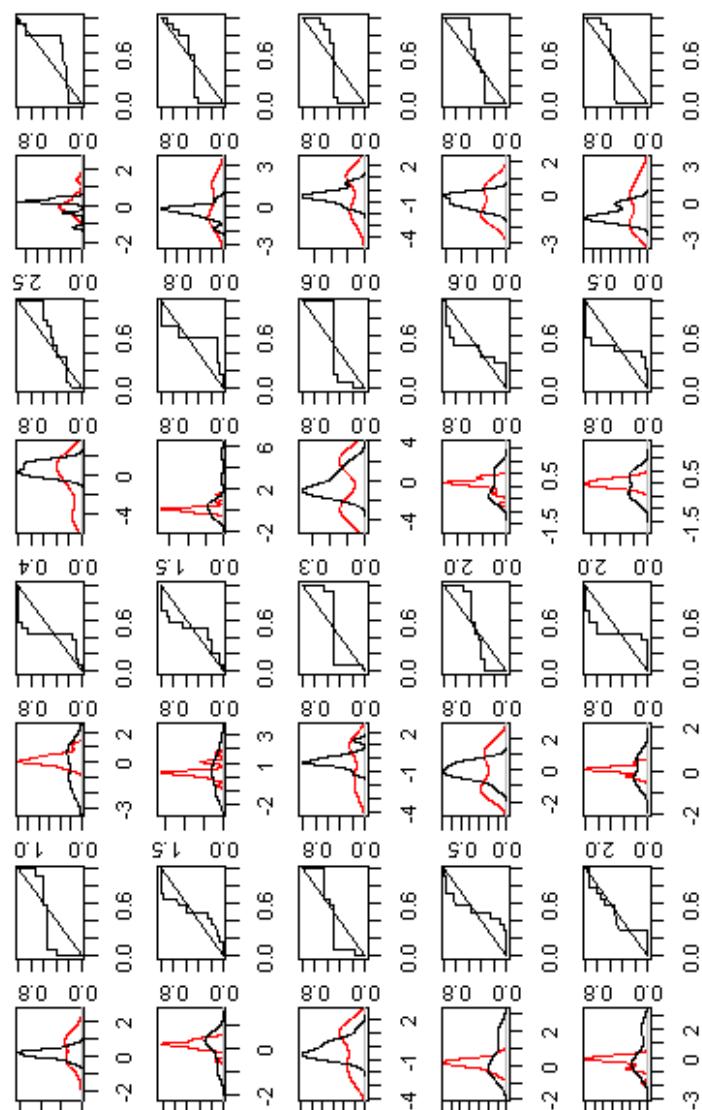


Figura 5.27: Densidades estimadas pelo método do núcleo e curvas ROC empíricas dos genes mistos dos dados *Linfoma*. As densidades a vermelho representam o grupo experimental e a preto as densidades no grupo controlo. Da esquerda para a direita e de baixo para cima, cada par de gráficos é referente a um gene, cujas identificações são: GENE1141X, GENE3521X, GENE3547X, GENE3473X, GENE2547X, GENE2519X, GENE1877X, GENE3343X, GENE3322X, GENE3323X, GENE3389X, GENE3388X, GENE3909X, GENE2887X, GENE2778X, GENE463X, GENE1004X, GENE3407X, GENE75X, GENE1817X.

Dos 20 genes mistos 3 têm funções regulatórias desconhecidas, os restantes 17 genes estão relacionados com codificação de proteínas. Os genes GENE3323X (*BCL7A*) e GENE3388X (*Cadeia J da imunoglobulina*) estão presentes noutras clones da mesma base de dados, e foram também selecionados como genes mistos, GENE 3322X e GENE3389X respetivamente. Nas Tabelas B.1 e B.2 do apêndice B.2 descrevem-se as características biológicas dos 20 genes mistos.

5.3.3 Comparação com outros métodos

Compararam-se os resultados obtidos a partir da análise do *Arrow plot*, com os resultados obtidos por Parodi *et al.* (2008), onde utilizaram as estatísticas ABCR e TNRC (secção 3.5.3). De um total de 1607 genes DE, 16 corresponderam a curvas ROC degeneradas de acordo com os valores mais elevados de TNRC, ou seja, Parodi *et al.* (2008) selecionaram 16 genes mistos. Oito genes desta lista, foram classificados como mistos de acordo com a nova abordagem aqui proposta. Os restantes 8 genes da lista de Parodi *et al.* (2008) apresentaram valores da AUC e OVL ligeiramente superiores aos pontos de corte aqui definidos. No entanto, se se escolhessem pontos de corte para a AUC e OVL de modo a selecionar esses 8 genes, selecionar-se-iam mais 85 genes mistos de acordo com o método aqui proposto.

Ordenaram-se os valores absolutos das estatísticas kerAUC, empAUC WAD, FC, AD, ibmT, modT, samT, SAMROC e *t*-Welch por ordem decrescente e selecionaram-se os 178 primeiros genes da lista de cada método, sendo este o número total de genes selecionado a partir do *Arrow plot* e para os pontos de corte definidos anteriormente, onde 20 genes são mistos e os restantes 158 são DE com regulação positiva ou negativa. Cruzou-se a lista de genes mistos e DE selecionados pelo novo método com 178 primeiros genes da lista obtida a partir de cada método, e apenas o FC selecionou 1 gene em comum com a lista de genes mistos obtida pelo *Arrow plot*, e de um total de 158 genes DEs selecionados pelo método proposto, obtiveram-se os seguintes resultados: kerAUC — 100; empAUC — 100; modT e ibmT— 98; samT — 94; *t*-Welch — 99; SAMROC — 90; AD — 59; WAD e FC — 51.

5.3.4 Arrow plot — AUC empírica vs. AUC núcleo

Na Figura 5.28 apresenta-se os *Arrow plots* sobrepostos, considerando a AUC estimada pelos métodos empírico e do núcleo.

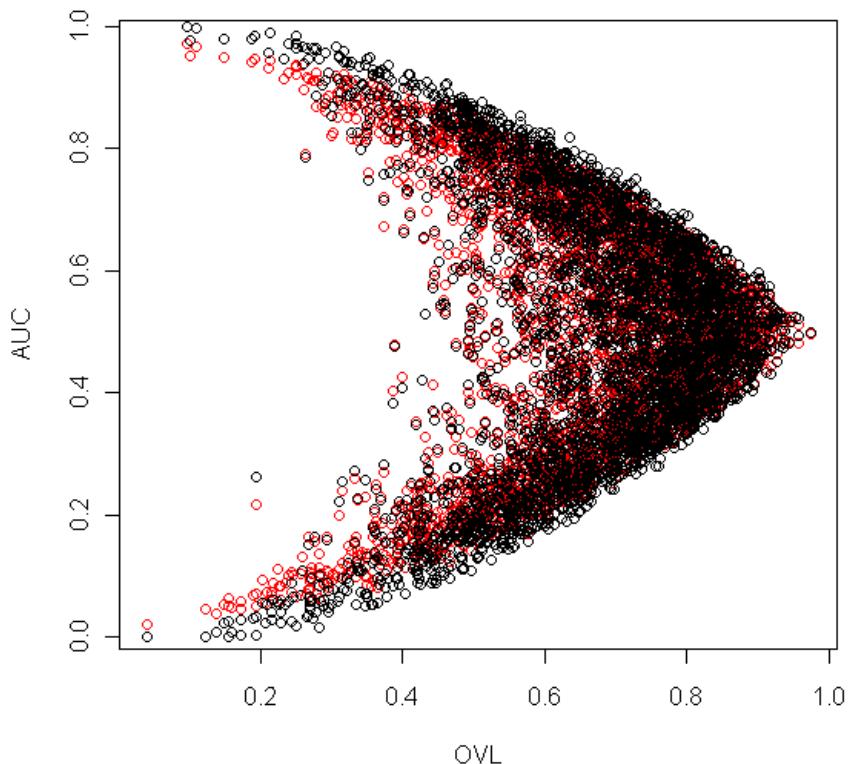


Figura 5.28: Comparação do *Arrow plot* com AUC estimada pelo método empírico e pelo método do núcleo. Pontos a vermelho correspondem à AUC estimada pelo método do núcleo e pontos a preto correspondem à AUC estimada pelo método empírico.

Cruzou-se a lista de genes mistos e DE selecionados pelo novo método considerando a AUC estimada pelo método do núcleo e considerando os mesmos pontos de corte (total de 99 genes), com os primeiros 99 genes da lista ordenada por ordem decrescente dos valores das estatísticas de cada um dos métodos descritos anteriormente. Concluiu-se que apenas o FC selecionou 1 gene da lista de genes mistos selecionada pelo *Arrow*

Tabela 5.7: Comparação do número de genes considerando o *Arrow plot* com AUC estimada pelos métodos do núcleo e empírico, considerando os pontos de corte anteriormente definidos.

Número de genes	AUC núcleo	AUC empírica
Mistos	21	20
Regulação positiva	31	68
Regulação negativa	47	90

plot. De um total de 78 genes DE selecionados pelo método aqui proposto obtiveram-se os seguintes resultados: kerAUC — 78; empAUC — 77; ibmT— 73; modT — 72; *t*-Welch — 71; samT — 64; SAMROC — 60; AD — 42; FC — 39; WAD — 37; estes valores correspondem ao número de genes coincidentes entre o *Arrow plot* e a lista de genes de cada um dos métodos.

5.3.5 Considerações Finais

Foram selecionados 20 genes mistos pelo método proposto e comparando os resultados com os obtidos por Parodi *et al.* (2008), apenas 8 genes foram selecionados em comum. Caso se considerassem pontos de corte para o OVL e AUC de modo a selecionar os outros 8 genes da lista de 16 genes selecionados por Parodi *et al.* (2008), ter-se-ia de selecionar mais 85 genes mistos. O método proposto por Parodi *et al.* (2008) não admite variâncias diferentes em ambos os grupos e apenas seleciona genes que revelem distribuições bimodais.

Comparando os resultados com os obtidos pelos outros métodos aqui analisados, apenas o FC selecionou um gene da lista de genes mistos obtida pelo *Arrow plot*. Quanto ao número de genes DE, os métodos AD, WAD e FC foram os que apresentaram o menor número de genes selecionados em comum com o *Arrow plot*.

Comparando os resultados obtidos com o *Arrow plot* considerando a AUC estimada pelo método do núcleo e AUC empírica e, os mesmos pontos de corte, o número de genes mistos é equivalente, sendo as diferenças significativas ao nível do número de genes com regulação positiva e negativa, onde o *Arrow plot* considerando a AUC empírica seleciona o dobro do número de genes em comparação com o *Arrow plot* com a AUC estimada pelo método

do núcleo.

Capítulo 6

Conclusões

A tecnologia de *microarrays* permite monitorizar em simultâneo milhares de genes num único *chip*, tornando possível aos investigadores perceberem as suas interações. Embora a base química dos *microarrays* não seja recente, o estudo simultâneo de centenas de genes transformou os *microarrays* numa ferramenta de análise global muito importante com aplicações na Biologia e na Medicina. A tecnologia de *microarrays* permite identificar genes associados ao problema sob investigação, sendo considerada uma ferramenta de diagnóstico ou prognóstico, onde a identificação do gene não é um ponto crítico, mas sim a informação quantitativa associada ao perfil ou assinatura de expressão.

Atualmente, na pesquisa oncológica, abordagens baseadas em *microarrays* têm sido amplamente usadas, onde uma das principais aplicações é a identificação de grupos de genes associados ao desenvolvimento de um tipo de cancro em particular. Diferentes estudos baseados no perfil de expressão de genes em amostras tumorais têm revelado a grande heterogeneidade transcripcional do cancro e têm permitido a classificação de novas subclasses clínicas e biológicas importantes da doença. Dessa forma, as vantagens desta tecnologia têm sido demonstradas em estudos realizados numa extensa variedade de cancros.

A análise de dados provenientes de *microarrays* representam um grande desafio. A complexidade advém não só da elevada quantidade de dados que é gerada, mas também pela interação de diferentes áreas, como a biologia, a

estatística e a informática.

Neste momento ainda não existem processos perfeitos de gestão e manuseamento de grandes quantidades de dados, pois o atual grande desafio não aparenta ser o modelo de produção de *arrays*, mas sim a manipulação e análise das matrizes de dados. Outro dos grandes problemas associados aos *microarrays* é a ausência de um protocolo estandardizado para manuseamento de dados, embora este problema caminhe a passos largos para a sua resolução com a criação de grupos de trabalho, como o MGED (*Microarray Gene Expression Data Society*), que procuram criar um processo de uniformização de manipulação e armazenamento de dados de expressão genética.

O principal objetivo deste trabalho foi o de apresentar um método que permita selecionar genes que revelem a presença de subclasses, e deste modo, obter mais informação acerca dos mecanismos biológicos dos genes subjacentes a condições físicas ou patológicas.

Sendo a seleção de genes um passo na análise de dados de *microarrays* que depende de um prévio e complexo tratamento de dados e, que por sua vez produz transformações nos dados que influenciam os resultados finais, dedicou-se parte deste trabalho na pesquisa dos métodos mais vulgarmente utilizados. Conclui-se que não existe um critério de otimalidade entre os métodos apresentados, devendo o utilizador comparar vários métodos e optar pelo que produzir os melhores resultados.

Com o objetivo de avaliar a performance do método proposto em comparação com os usualmente utilizados para a seleção de genes DE em *microarrays* de um canal, apresentaram-se alguns dos métodos mais vulgarmente utilizados baseados no *Fold Change*, na estatística-*t* e na metodologia ROC.

A estatística-*t*, as suas variantes e o método FC, apesar das críticas, são dos métodos mais utilizados na seleção de genes DE. Autores consideram ainda que aplicando ambas as estatísticas em simultâneo para selecionar genes (dupla filtragem), conseguem retirar das duas as suas melhores características, reduzindo a possibilidade de se produzirem resultados errados.

Em muitas experiências de *microarrays* o primeiro objetivo é ordenar os genes segundo o valor de uma estatística em vez de calcular os correspondentes valores-*p*. Isto porque apenas um número muito reduzido de genes pode ser

posteriormente monitorizado, mesmo que se selezionem genes considerados significativos.

Neste trabalho, a abordagem para a seleção de genes DE, foi com base na ordenação e seleção dos primeiros genes de uma lista ordenada a partir de valores obtidos de uma estatística. No entanto, se se pretende fazer inferência e, quando o objetivo é selecionar genes de entre um número consideravelmente elevado, a questão dos testes múltiplos é largamente abordada nas inúmeras publicações acerca deste tema. A seleção de genes a partir dos valores- p , quer sejam corrigidos ou não, parece ser a abordagem de eleição para selecionar genes. Mas há questões que muitos investigadores parecem não colocar, como por exemplo, será que a análise dos resultados de uma experiência de *microarrays* deve ser vista do ponto de vista de inferência ou observacional?

Foi proposto um método de seleção de genes a partir da análise exploratória de um gráfico, designadamente *Arrow plot*. Este gráfico permite selecionar para além de genes com regulação positiva e negativa, genes com um comportamento biológico de interesse, que neste trabalho foram designados de genes mistos. Este gráfico baseia-se na representação das estimativas do OVL e da AUC de todos os genes pertencentes à experiência. A análise em simultâneo do OVL e AUC é bastante intuitiva, uma vez que ambos admitem valores no intervalo [0,1], admitindo no caso da AUC a existência de curvas ROC degeneradas. Outra característica muito útil, particularmente nestas experiências, é o facto de ambas as estatísticas serem invariantes em relação a transformações de escala estritamente crescentes e diferenciáveis das variáveis que representam os níveis de expressão dos genes em ambos os grupos. Estimadores que partilhem esta propriedade de invariância não dependem das observações diretamente, mas das suas ordens. Esta propriedade é muito útil, uma vez que na análise de dados de *microarrays* é muito comum procederem-se a transformações das variáveis para remover enviesamentos, como por exemplo o logaritmo de base 2.

Neste trabalho, a estimação da AUC e do OVL foi realizada numa abordagem não-paramétrica, que em dados de *microarrays* revela-se mais adequada, uma vez que não exige pressupostos distribucionais e geralmente o número de *arrays* é consideravelmente pequeno.

Para a estimação do OVL foi proposto um algoritmo que se baseia na determinação dos pontos que delimitam a área de interseção das densidades dos níveis de expressão dos genes em ambas os grupos estimadas pelo método

do núcleo. De acordo com o estudo de simulação, a estimativa *bootstrap* do viés do OVL calculado pelo algoritmo proposto apresentou valores mais reduzidos em comparação com outros métodos não-paramétricos para a estimação do OVL, particularmente quando as distribuições são mais próximas. Este algoritmo não impõe restrições no número de interseções das densidades e para determinar a curva de interseção entre as duas densidades, numa base ponto-por-ponto, não é necessário que as estimativas das densidades pelo método do núcleo tenham as mesmas abscissas nos dois grupos em análise.

Compararam-se resultados na seleção de genes considerando a AUC estimada pelo método empírico e pelo método do núcleo. Verificou-se que quando as distribuições são normais, o comportamento do viés é semelhante considerando os dois métodos de estimação. No entanto, o número de genes selecionado com regulação positiva e regulação negativa é diferente quando se consideram os mesmos pontos de corte no *Arrow plot*, nomeadamente quando se considera a AUC empírica o número de genes selecionado é substancialmente maior. O *Arrow plot* com a AUC estimada pelo método empírico revela-se mais otimista, no sentido de permitir selecionar um número mais elevado de genes, num entanto presumivelmente menos realista, uma vez que biologicamente espera-se selecionar um número muito reduzido de genes DE. Relativamente ao número de genes mistos o número de genes selecionado considerando os dois métodos de estimação é semelhante. No entanto a discrepância encontrada nos resultados considerando os dois métodos de estimação não era esperada. Pretende-se como trabalho futuro realizar um estudo mais detalhado com recurso a métodos *bootstrap* considerando outras distribuições para além da normal e considerando mais do que um ponto de interseção entre as densidades.

Houve ainda a necessidade de se acrescentar ao *Arrow plot*, a identificação de genes que tenham as distribuições dos níveis de expressão bimodais ou multimodais, uma vez que, genes com valores da AUC à roda de 0.5 e valores baixos do OVL, não implica necessariamente que sejam genes mistos. Genes não DE com níveis de expressão médios semelhantes nos dois grupos, variâncias significativamente diferentes e distribuições unimodais também se apresentam na mesma região do *Arrow plot*. Assim, desenvolveu-se um algoritmo para identificar distribuições bimodais ou multimodais, estimadas pelo método do núcleo, para os genes que se encontram nessa zona particular do gráfico. Verificou-se que este algoritmo classifica as distribuições associadas aos genes de interesse, com uma exatidão de 100%. No entanto como trabalho futuro, pretende-se avaliar a performance deste

algoritmo, para os genes que se encontrem nas restantes zonas do gráfico.

O OVL revelou-se o método com a melhor performance para selecionar em simultâneo genes com regulação positiva e negativa e genes mistos, em comparação com os métodos descritos neste trabalho.

Um gráfico muito utilizado na análise de dados de *microarrays* é o *Volcano plot*, no entanto na análise de genes mistos revelou-se desadequado.

Uma das limitações que pode ser associada ao *Arrow plot* é o facto da seleção dos genes de interesse obrigar o utilizador a definir pontos de corte, e nesse sentido, deve ser um utilizador experiente, uma vez que à partida deve ter uma ideia do número de genes que pretende selecionar. No entanto, esta situação é semelhante quando o investigador tem uma lista de genes ordenada segundo o valor de uma estatística de teste e, tem de selecionar os primeiros genes. Aqui também tem de decidir a partir de que ponto vai considerar a lista final.

Como trabalho futuro pretende-se investigar a viabilidade de usar a área parcial abaixo da curva ROC na seleção de genes mistos. Neste caso, o utilizador pode à partida definir qual a percentagem de falsos positivos que admite na sua investigação.

Pretende-se ainda estender a aplicação deste gráfico na análise de dados em experiências *RNA-Seq* também designado por *Whole Transcriptome Shotgun Sequencing—WTSS*. É uma tecnologia que faz parte da *Next Generation Sequencing*, com aplicação na análise do transcriptoma, cujo objetivo é sequenciar o cDNA de modo a obter informação acerca do RNA. Esta tecnologia visa substituir os *microarrays*. A aplicação particular do *Arrow plot* será com o objetivo de identificar misturas de diferentes tipos de células, que se traduzem pela presença de distribuições bimodais ou multimodais.

Apêndice A

Código em R

A.1 Algoritmo 1

```
# Seja (x,y) as coordenadas de um ponto obtidas a partir da densidade
# estimada pelo método do núcleo
# L1 lista de pontos no grupo 1 e L2 lista de pontos no grupo 2

#####
# mínimo das ordenadas quando existe mais de um ponto com
#abcissas iguais na mesma lista

m_y<-function(x,L){
  z<-which(L[,1]==x)
  return(min(L[z,2]))}

#####

# obtem o ponto imediatamente anterior a uma dada abscissa x na mesma lista

xant<-function(x,L){
  z<-which(L[,1]< x)
  r<-max(L[z,1])
  y<-m_y(r,L)
  a<-ifelse(is.null(r)|r=="-Inf",NA,r)
  y1<-ifelse(y=="Inf",NA,y)
  return(c(a,y1))}

#####
```

Código em R

```
#obtem o ponto imediatamente a seguir a uma dada abcissa x na mesma lista

xseg<-function(x,L){
  z<-which(L[,1]>x)
  if(length(z)==0){r<-NA y<-NA }
  else
  {r<-min(L[z,1])
  y<-m_y(r,L)}
  return(c(r,y))}

#####
#obtem os pontos com o mesmo x na mesma lista

xigu<-
function(x,L){
  z<-which(L[,1]==x)
  y<-m_y(x,L)
  y1<-ifelse(y=="Inf",NA,y)
  return(c(x,y1))
}

#####
# Determina os pontos de uma densidade que devem constar na lista final
## tem de se correr para as duas listas: (L1,L2) e depois (L2,L1)

pontos<- function(L1,L2) {
  x<-L1[,1]
  y<-L1[,2]
  L2igu<-sapply(x,function(x)xigu(x,L2))
  t<-L1[which(L2igu[2,]!="NA" & L2igu[2,]>=y),]
  L2ant<-sapply(x,function(x)xant(x,L2))
  L2seg<-sapply(x,function(x)xseg(x,L2))
  k<-L1[which(is.na(L2igu[2,]) & L2ant[2,]!="NA" & L2seg[2,]!="NA"
    & y<=L2ant[2,] & y<=L2seg[2,]),]
  return(rbind(t,k))}

#####
## junta os pontos obtidos nas duas listas anteriores e ordena
## por ordem crescente das abcissas

junta_pontos <- function(L1,L2) {
  L3<-rbind(L1,L2)
  return(L3[order(L3[,1]),])}

#####
## lista final para o cálculo da área
## acrescenta os pontos de salto entre densidades
## A lista L1 tem de corresponder à lista 1 e L2 a 2
## Tem de se alterar o nome das densidades para 1 e 2
```

```

## Lista total é a que corresponde aos pontos de interesse

p<-data.frame()
final<-function(L,L1,L2){
  while (i <= nrow(L)-1 ){
    if (L[i,3]!=L[(i+1),3]) {

      x1<-L[i,1]
      y1<-L[i,2]
      x4<-L[(i+1),1]
      y4<-L[(i+1),2]

      if (L[i,3]==1) {x2<-xseg(L[i,1],L1)[1]} else
        x2<-xseg(L[i,1],L2)[1]
      if (L[i,3]==1) {y2<-xseg(L[i,1],L1)[2]} else
        y2<-xseg(L[i,1],L2)[2]
      if (L[(i+1),3]==1) {x3<-xant(L[(i+1),1],L1)[1]} else
        x3<-xant(L[(i+1),1],L2)[1]
      if (L[(i+1),3]==1){y3<-xant(L[(i+1),1],L1)[2]} else
        y3<-xant(L[(i+1),1],L2)[2]

      D<-(x1-x2)*(y3-y4)-(y1-y2)*(x3-x4)

      if (D!=0){
        x_int<-(1/D)*((x3-x4)*(x1*y2-y1*x2)-(x1-x2)*(x3*y4-y3*x4))
        y_int<-(1/D)*((y3-y4)*(x1*y2-y1*x2)-(y1-y2)*(x3*y4-y3*x4))
        r<-c(x_int,y_int)
      }

      p<-rbind(p,c(L[i,1],L[i,2]))
      p<-rbind(p,c(x_int,y_int))

      } else
      p<-rbind(p,c(L[i,1],L[i,2]))
    i<-i+1
  }

  total<-rbind(p,c(nrow(L),1),L[nrow(L),2]))
  return(total)
}

## Cálculo do OVL pela regra do trapézio

library(bitops)
library(caTools)
OVL<-trapz(z[,1],z[,2])

```

A.2 Algoritmo 2

```
bimodality<-function(L1,L2){  
  L1ord<-L1[order(L1$x),]  
  L2ord<-L2[order(L2$x),]  
  
  x1<-L1ord[,1]  
  x2<-L2ord[,1]  
  
  L1ord.seg<-sapply(x1,function(x1)xseg(x1,L1ord))  
  L1ord.igu<-sapply(x1,function(x1)xigu(x1,L1ord))  
  
  z1<-ifelse(L1ord.igu[2,]<=L1ord.seg[2,],1,0)  
  z1<-na.omit(z1)  
  r1<-diff(which(z1!=1))  
  
  bim1<-ifelse(length(subset(r1,r1>1))!=0,TRUE,FALSE)  
  L2ord.seg<-sapply(x2,function(x2)xseg(x2,L2ord))  
  L2ord.igu<-sapply(x2,function(x2)xigu(x2,L2ord))  
  
  z2<-ifelse(L2ord.igu[2,]<=L2ord.seg[2,],1,0)  
  z2<-na.omit(z2)  
  r2<-diff(which(z2!=1))  
  
  bim2<-ifelse(length(subset(r2,r2>1))!=0,TRUE,FALSE)  
  
  group1<-ifelse(bim1==TRUE,1,0)  
  group2<-ifelse(bim2==TRUE,1,0)  
  both<-ifelse(bim1==TRUE & bim2==TRUE, 1,0)  
  
  return(c(group1,group2,both))  
}  
  
#####  
BIM<-NULL  
for (i in 1:z) {  
  BIM<-rbind(BIM,bimodality(G1,G2)) # z número de genes candidatos a mistos  
  # i.e. genes que satisfazem a condição  
  # 0.4<AUC<0.6 and OVL<0.4  
}  
#####
```

A.3 Algoritmo 3

```
# Estimação da AUC pelo método empírico  
library(ROC)  
AUC<-NULL
```

```

for (i in 1:n){
AUC[i]<-AUC(rocdemo.sca(rec.info,data[i,],dxrule.sca, caseLabel="",
markerLabel=""))
}

# Estimação da AUC pelo método do núcleo

library(pROC)
kerAUC<-NULL
for (i in 1:n){
roc.i<-roc(controls=log.control[i,], cases=log.cases[i,],direction=<,auc=TRUE)
kerAUC[i]<-AucROC$auc[1]
}

##### arrow plot

plot(OVL,AUC,ylab="AUC",xlab="OVL", main="Arrow plot") #equivalente para kerAUC

# pontos de corte

abline(h=0.1)
abline(h=0.9)
abline(v=0.4)
abline(h=0.6)
abline(h=0.4)

OVL.AUC<-data.frame(AUC,OVL)
rownames(OVL.AUC)<-rownames(data)

OVL.AUC2<-OVL.AUC[AUC>0.9 & OVL<0.4,]    # genes com regulação positiva
OVL.AUC3<-OVL.AUC[AUC<0.1 & OVL<0.4,]    # genes com regulação negativa

points(OVL.AUC2$OVL,OVL.AUC2$AUC,col="red",pch=19) # genes com regulação positiva
points(OVL.AUC3$OVL,OVL.AUC3$AUC,col="blue",pch=19) # genes com regulação negativa
points(OVL.AUC$OVL[v1],OVL.AUC$AUC[v1],
col="green",pch=19)#genes especiais com bimodalidade
#em ambos os grupos, v1 é um vetor
#com os genes com estas características

points(OVL.AUC$OVL[v2],OVL.AUC$AUC[v2],
col="cyan",pch=19) #genes especiais com bimodalidade
#no grupo experimental

points(OVL.AUC$OVL[v3],OVL.AUC$AUC[v3],
col="orange",pch=19)#genes especiais com bimodalidade
#no grupo controlo

```

Código em R

Apêndice B

Figuras e Tabelas

B.1 Dados Cancro da Bexiga

B. Figuras e Tabelas

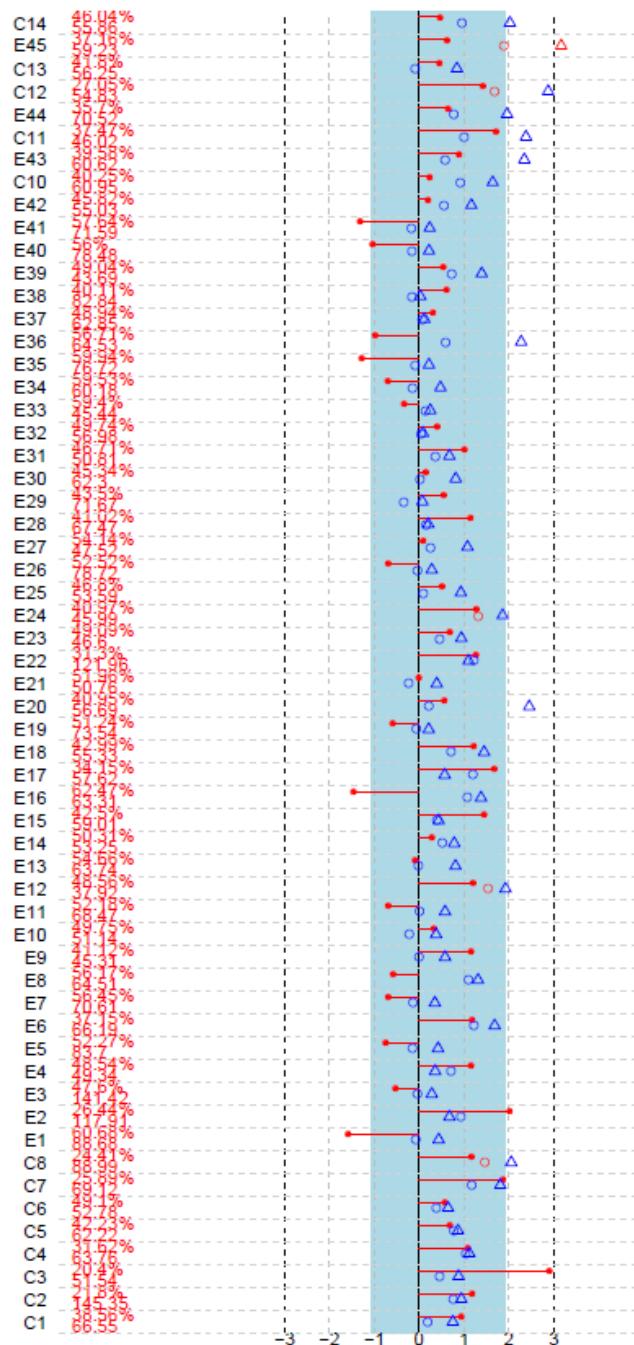


Figura B.1: Gráfico QC apóis pré-processamento FARMS

B.1. Dados Cancro da Bexiga

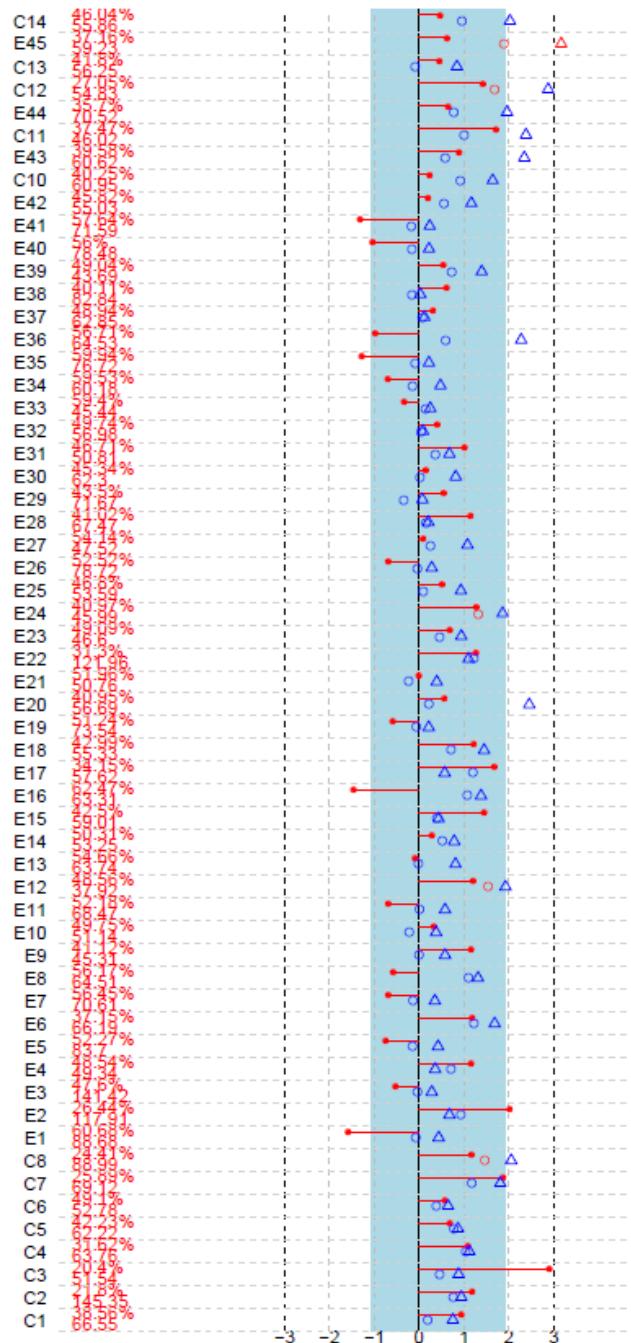


Figura B.2: Gráfico QC apóis remoção do array C9.

B.2 Dados Linfoma

Tabela B.1: Descrição biológica dos genes mistos selecionados a partir do *Arrow plot* nos dados *Linfoma*.

Nome do gene	Descrição
<i>BCL7A</i>	A região cromossómica 12q24.1 é um local recorrente de quebra em Linfomas das células B não-Hodgkin, na região do terminal amina do BCL7A, definindo assim de patogénese um novo mecanismo neste tipo de linfomas.
<i>Cadeia J da imunoglobulina</i>	A cadeia J encontra-se na IgM pentamérica, e também é importante nas imunoglobulinas segregadas.
<i>MAPKKK5</i>	As cinases da JUN N terminal (JNKs) são MAPKs que estimulam a atividade transcripcional do JUN em resposta a fatores de crescimento, citocinas proinflamatórias, e alguns fatores de stresse ambiental.
semelhante a <i>KIAA0050</i>	A ACAP1 e a ACAP2 são recrutadas pelo fator de crescimento derivado das plaquetas (PDGF), induzindo alterações na membrana dos fibroblastos NIH 3T3 de rato, e a sua sobreexpressão inibem as alterações de membrana.
<i>Histona desacetilase 3</i>	A acetilação e a desacetilação das proteínas histónicas alteram a estrutura dos cromossomos, afetando o acesso dos fatores de transcrição ao DNA.
<i>VRK2 Cinase</i>	Genes envolvidos na regulação da divisão celular.
<i>LR11</i>	Gene geneticamente associado à manifestação tardia da doença de Alzheimer, parecendo estar envolvido no processo neurodegenerativo.
<i>BL34</i>	Sequência específica de um gene prematuro das células B humanas, que é induzível em resposta a vários sinais de ativação das células B.
<i>Bicunina placentária</i>	É um potente inibidor das serina peptidases envolvidas na coagulação sanguínea e na fibrinólise tais como a plasmina, calicreina e fator XIa.
<i>LBR</i>	Codifica o receptor da lâmina B, uma proteína interna da membrana nuclear que se liga à lâmina B. Uma vez que este receptor pode ser uma redutase do esterol, a perda da maior parte da sua expressão pode levar a alterações no metabolismo dos esteróis que provocarão anomalias de desenvolvimento.

Tabela B.2: Descrição biológica dos genes mistos selecionados a partir do *Arrow plot* nos dados *Linfoma* (cont.).

Nome do gene	Descrição
<i>Cadeia leve kappa da imunoglobulina</i>	Pequena subunidade polipeptídica do anticorpo. Os anticorpos são produzidos pelos linfócitos B, sendo que cada um expressa um só tipo de cadeia leve (kappa ou lambda).
<i>BNIP3</i>	Ativa a expressão do gene que codifica a NIP3, induzindo a apoptose celular em condições de privação persistente de oxigénio. Esta via parece ter um papel importante na morte celular resultante de isquémia cerebral e miocárdica.
<i>CLK3 Cinase</i>	Família de enzimas que catalisam a fosforilação de proteínas em resíduos de serina e treonina.
<i>DNA Ligase III</i>	Forma as ligações fosfodiester após quebra das cadeias de DNA durante o processo de recombinação meiótica em células germinais e em consequência de dano do DNA em células somáticas.
<i>PARP</i>	Necessária para a reparação celular. Inibidores desta enzima potenciam os efeitos letais de agentes nocivos.

B. Figuras e Tabelas

Referências

Affymetrix. (1999). *Gene Chip Analysis Suite User Guide*. Affymetrix, Santa Clara, C.A., versão 4.

Affymetrix. (2001). *Microarray suite user guide*. Affymetrix, Santa Clara, C.A., versão 5.

Affymetrix. (2002). *Statistical Algorithms Description Document*. Inc. Santa Clara, C. A.

Affymetrix. (2004). *Expression Analysis Technical Manual*. Affymetrix, Santa Clara, C.A.

Affymetrix. (2005). *Guide to probe logarithmic intensity error (PLIER) estimation*. Affymetrix, Santa Clara, C.A.

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K. e Watson, J. D. (1994). *Molecular Biology of The Cell*. Garland Publishing, Inc.

Alizadeh, A. A., Elsen, M. B., Davis, E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Marti, G. E., Moore, T., Hudson Jr, J., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G. Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O. e Staudt, L. M. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**:503–511.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D. e Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Cell Biology*, **96**:6745–6750.

Referências

- Autio, R., Kilpinen, S., Saarela, M., Hautaniemi, S., Kallioniemi, O. e Astola, J. (2006). The Weibull distribution based normalization method for Affymetrix gene expression microarray data. In Proceedings of the 2006 IEEE International Workshop on Genomic Signal Processing and Statistics (Gensips 2006), 2830 May 2006 College Station, Texas, USA: 9–10.
- Autio, R., Kilpinen, S., Saarela, M., Kallioniemi, O., Hautaniemi, S. e Astola J. (2009). Comparison of Affymetrix data normalization methods using 6,926 experiments across five array generations. *BMC Bioinformatics*, 10:S24.
- Ayroles, J. F. e Gibson, G. (2006). *Analysis of variance of microarray data in DNA microarrays*, Part B. Elsevier Inc.
- Baldi, P. e Long, A.D. (2001). A Bayesian framework for the analysis of microarry expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, 17:509–519.
- Ballman, K. V., Grill, D., Oberg, A. e Therneau, T. (2004). Faster cyclic loess: normalizing DNA arrays via linear models. *Tecnhical Report*, nº 68, Mayo Foundation.
- Baker, S. G. (2003). The central role of receiver operating characteristic (ROC) curves in evaluating tests for the early detection of cancer. *Journal of the National Cancer Institute*, 95(7):511–515.
- Bamber, D.C. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic curve graph. *Journal of Mathematical Psychology*, 12:387–415.
- Bland, J.M. e Altman, D.G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, 327(8476):307-10.
- Bo, T.H., Dysvik, B. e Jonassen, I. (2004). LSimpute: accurate estimation of missing values in microarray data with least squares methods. *Nucleic Acids Research*, 32(3):e34.
- Bolstad, B. (2001). Probe level quantile normalization of high density oligonucleotide array data. <http://bmbolstad.com/stuff/qnorm.pdf>.

- Bolstad, B. (2002). Comparing the effects of background, normalization and summarization on gene expression estimates. <http://bmbolstad.com/stuff/components.pdf>.
- Bolstad, B., Irizarry, R.A., Astrand, M. e Speed, T.P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**(2):185–193.
- Bradley, E. L. Jr. (1985). Overlapping Coefficient. In *Encyclopedia of Statistical Sciences*, **6**:546–547. Chapman and Hall: New York.
- Bras, L.P. e Menezes, J.C. (2007). Improving cluster-based missing value estimation of DNA microarray data. *Biomolecular Engineering*, **24**(2):273–282.
- Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C.A., Causton, H. C., Gaasterland, T., Glenisson, P., Holstege, F. C.P., Kim, I. F., Markowitz, V., Matese, J. C., Parkinson, H., Robinson, Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J. e Vingron, M. (2001). Minimum information about a microarray experiment (MIAME)toward standards for microarray data. *Nature Genetics*, **29**:365–371.
- Breitling, R., Armengaud, P., Amtmann, A. e Herzyk, P. (2004). Rank Products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Letters*, **573**: 83–92.
- Breitling, R. e Herzyk, P. (2005). Rank-based methods as a non-parametric alternative of the t-statistic for the analysis of biological microarray data. *Journal of Bioinformatics and Computational Biology*, 11:32.
- Brenner, S., M. Johnson, J. Bridgham, G. Golda, D. H. Lloyd, D. Johnson, S. Luo, S. McCurdy, M. Foy, M. Ewan, R. Roth, D. George, S. Eletr, G. Albrecht, E. Vermaas, S. R. Williams, K. Moon, T. Burcham, M. Pallas, R. B. DuBridge, J. Kirchner, K. Fearon, J. i Mao, K. Corcoran (2000). Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature Biotechnology*, **18**(6):630-634.
- Broberg, P. (2003). Statistical methods for ranking differentially expressed genes. *Genome Biology*, **4**:R41.

Referências

- Celton, M., Malpertuy, A., Lelandais, G. e Brevern, A. G. (2010). Comparative analysis of missing value imputation methods to improve clustering and interpretation of microarray experiments. *BMC Genomics*, 11:15.
- Chan, V., Hontzeas, N. e Park, V. (2000). Gene expression. Preprint.
- Chen, Z., McGee, M., Liu, Q. e Scheuermann, R. H. (2006). A distribution free summarization method for Affymetrix GeneChip arrays. *Bioinformatics*, **23**(3):321–327.
- Chu, G. Narasimhan, B., Tibshirani, R. e Tusher,V. (2001). SAM: Significance Analysis of Microarrays. *User Guide and Technical Document*.
- Chudin,E., Walker, R., Kosaka, A., Wu, S. X., Rabert, D., Chang, T. K. e Kreder, D. E. (2001). Assessment of the relationship between signal intensities and transcript concentration for affymetrix genechip arrays. *Genome Biology*, **3**(1).
- Clemons, T. E. e Bradley Jr., L. (2000) A nonparametric measure of the overlapping coefficient. *Computational Statistics and Data Analysis*, **31**(1): 51–61.
- Cope, L. M., Irizarry, R. A., Jaffee, H. (2003). A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics*, **20**:323–331.
- Cope, L. M., Hartman, S. M., Gohlmann, H. W. H., Tiesman, J. P. e Irizarry, R. A. (2005). Analysis of Affymetrix GeneChip Data Using Amplified RNA. *Johns Hopkins University, Dept. of Biostatistics Working Papers*, **84**.
- Crick, F. H. C.(1970). The Central Dogma of Molecular Biology. *Nature*, **227**:561–563.
- Cui, X. e Churchill, G. A. (2003). Statistical tests for differential expression in cDNA microarray experiments. *Genome Biology*, **4**:210.
- Cui, X., Hwang, J. T., Qiu, J., Blades, N. J. e Churchill, G. A. (2005). Improved Statistical Tests for Differential Gene Expression by Shrinking variance components estimates. *Biostatistics*, **6**(1):59–75.

- Dalman, M. R., Deeter, A., Nimishakavi, G. e Duan, Z.-H. (2012). Fold change and p-value cutoffs significantly alter microarray interpretations. *BMC Bioinformatics*, **13**(Suppl2):S11.
- Dopazo, J., Zanders, E., Dragoni, I., Amphlett, G. e Falciani, F. (2001). Methods and approaches in the analysis of gene expression data. *Journal of Immunological Methods*, **250**: 93-112.
- Dorfman, D. D., Berbaum, K.S., Brandser, E. A. (2000). A contaminated binormal model for ROC data: Part I. Some Interesting examples of binormal degeneracy. *Academic Radiology*, **7**(6):420–426.
- Dudoit, S., Yang, Y. H., Callow, M. J., e Speed, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarrays experiments. *Statistics Sinica*, **12**, 111–139.
- Dziuda, D. (2010). *Data mining for genomics and proteomics: analysis of gene and protein*. Wiley-Interscience.
- Efron, B. e Tibshirani. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.
- Faraggi, D. e Reiser, B. (2002). Estimation of the area under the ROC curve. *Statistics in Medicine*, **31**: 3093–3106.
- Felix, J. M., Drummond, R. D., Nogueira, F. T. S., Junior, V. E. R., Jorge, R. A., Arruda, P. e Menossi, M. (2002). Genoma Funcional. *Biotecnologia Ciência e Desenvolvimento*, **24**, 60–67.
- Feten, G., Almoy, T. e Aastveit, A.H. (2005). Prediction of missing values in microarray and use of mixed models to evaluate the predictors. *Statistical Applications in Genetic and Molecular Biology*, **4**:10.
- Freudenberg, J.M. (2005). Comparison of background correction and normalization procedures for high-density oligonucleotide microarrays. *Technical Report 3, Leipzig Bioinformatics Working Paper*.
- Gan, X., Liew, A.W. e Yan, H. (2006). Microarray missing data imputation based on a set theoretic framework and biological knowledge. *Nucleic Acids Research*, **34**(5):1608–1619.

Referências

GCOS (2004). *GeneChip Expression Analysis – Data Analysis Fundamentals*. Affymetrix, Inc., Santa Clara, C.A.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. e Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, **5**, 10:R80.

Goor, T. A. (2005). A History of DNA Microarrays. *Pharmaceutical Discovery*.<http://www.pharmadd.com/>.

Griffiths, A. J. F., Miller, J. H., Suzuki, D. T., Lewontin, R. C. e Gelbart, W. M. (1999). *An Introduction to genetic analysis*. 7ª Edição, Freeman.

Hardin, J. e Wilson, J. (2007). Oligonucleotide microarray data are not normally distributed. *Bioinformatics*, **10**:1–6

Hariharan, R. (2003). The analysis of microarray data. *Pharmacogenomics*, **4**(4):477-497.

Heller R. A., Allard J., Zuo F., Lock C., Wilson S., Klonowsky P., Gmuender H., Van Hart H., e R. Booth (1999). *Gene chips and microarrays: applications in disease profiles, drug target discovery, drug action and toxicity*. In *DNA Microarrays: A practical approach*. Edited by Mark Schena, Oxford University press Inc. New York, 187–206.

Hill, J. M., Oxley, M. E. e Bauer, K. W. (2003). Receiver operating characteristic curves and fusion of multiple classifiers. *ISIF*, 815–822.

Hochreiter, S., Clevert, D-A. e Obermayer, K. (2006). A new summarization method for Affymetrix probe level data. *Bioinformatics*, **22**:943-949.

Hong, F. e Breitling, R. (2007). A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics*, **24**(3):374–382.

Hong, F. e Wittner, B. (2010). *Bioconductor RankProd Package Vignette*. Software Manual from Bioconductor.

- Hu, J., Li, H., Waterman, M.S. e Zhou, X.J. (2006). Integrative missing value estimation for microarray data. *BMC Bioinformatics*, **7**:449.
- Inman, H. F. e Bradley E. L. (1989). The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities. *Communications in Statistics – Theory Methods*, **18**(10):3851–3874.
- Irizarry, R. A., Collin, H. B., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U. e Speed, T. P. (2003a). Exploration, normalization, and summaries of high density oligonucleotide array probe level. *Biostatistics*, **4**(2):249–64.
- Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B. e Speed, T. P. (2003b). Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research*, **31**(4).
- Irizarry, R. A., Wu, Z. e Jaffee, H. A. (2006). Comparison of Affymetrix GeneChip expression measures. *Bioinformatics*, **22**(7):789–794.
- Jiang N., Leach, L. J., Hu, X., Potokina, E., Jia, T., Druka, A., Waugh, R., Kearsey, M. e Luo, Z. W. (2008). Methods for evaluating gene expression from Affymetrix microarray datasets. *BMC Bioinformatics*, **9**:284.
- Jovaag, K. A. (1998). *Explanation of Median Polish Algorithm*. Technical Report, Iowa State University.
- Kadota, K., Nakai, Y. e Shimizu, K. (2008). A weighted average difference method for detecting differentially expressed genes from microarray data. *Algorithms for Molecular Biology*, **3**:8.
- Kauraniemi, P., Hautaniemi, S., Autio, R., Astola, J., Monni, O., Elkahloun, A. e Kallioniemi, A. (2004). Effects of Herceptin treatment on global gene expression patterns in HER2-amplified and nonamplified breast cancer cell lines. *Oncogene*, **23**:1010–1013.
- Kilpinen, S., Autio, R., Ojala, K., Iljin, K., Bucher, E., Sara,H., Pisto, T., Saarela, M., Skotheim, R. I., Björkman, M., Mpindi, J-P., Haapa-Paananen, S., Vainio, P., Edgren, H., Wolf, M., Astola, J., Nees, M., Hautaniemi, S. e Kallioniemi, O. (2008). Systematic bioinformatic analysis of expression levels of 17,330 human genes across 9,783 samples from 175 types of healthy and pathological tissues. *Genome Biology*, **9**:R139.

Referências

- Laing, E. e Smith, C. P. (2010). RankProdIt: A web-interactive Rank Products analysis tool. *BMC Research Notes*, **3**:221.
- Lemon, W. J., Palatini, J. J. T., Krahe, R. e Wright, F. A. (2002). Theoretical and experimental comparisons of gene expression indexes for oligonucleotide arrays. *Bioinformatics*, **18**: 1470–1476.
- Lee, M. L., Kuo, F. C., Whitmore, G. A. e Sklar, J. (2000). Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proceedings of the National Academy of Sciences*, **97**:9834–9839.
- Leung, Y. F., Lam, D. S. C. e Pang, C. P. (2001). The miracle of microarray data analysis. *Genome Biology*, **2**(9), reports 4021.1–4021.2.
- Li, W. (2011). Application of volcano plots in analysis of mRNA differential expression with microarrays. *Techical Report*: arXiv:1103.3434v1.
- Li, C. e Wong, W. H. (2001). Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceeding of the National Academy of Sciences*, **98**(1):31–36.
- Li, C. e Wong, W. H. (2003). DNA-Chip Analyzer (dChip). In *The analysis of gene expression data: methods and software*. Edited by G Parmigiani, ES Garrett, R Irizarry and SL Zeger. Springer, New York. pp. 120–141.
- Li, L., Chaudhuri, A., Chant, J. e Tang, Z. (2007). PADGE: analysis of heterogeneous patterns of differential gene expression. *Physiologic Genomics*, **32**:154–159.
- Lin, C.-H. e Patton, J. G. (1995). Regulation of alternative 3' splice site selection by constitutive splicing factors. *RNA*, **1**: 234–245.
- Lin, Y., Reynolds, P. e Feingold, E. (2003). An empirical bayesian method for differential expression studies using one-channel microarray data. *Statistical Applications in Genetics and Molecular Biology*, **2**(1): Article 8.
- Lloyd, C. J. (1997). The use of smoothed ROC curves to summarize and compare diagnostic systems. *Journal of American Statistical Association*, **93**:1356–1364.

- Lloyd, C. J. e Yong, Z.(1999). Kernel estimators of the ROC curve are better than empirical. *Statistics and Probability Letters*, **3**:221–228.
- Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H. e Brown, E. L. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays, *Nature Biotechnology*. **14**:1675-1680.
- Lönnstedt, I. e Speed, T. P. (2002). Replicated microarray data. *Statistical Sinica*, **12**:31–36.
- Lusted, L. (1968). *Introduction to Medical Decision Making*. C.C. Thomas, Springfield.
- Kim, H., Golub, G.H. e Park, H. (2005). Missing value estimation for DNA microarray gene expression data: local least squares imputation. *Bioinformatics*, **21**(2):187–198.
- Kim, K.Y., Kim, B.J. e Yi, G.S. (2002). Reuse of imputed data in microarray analysis increases imputation efficiency. *BMC Bioinformatics*, **5**:160.
- Marton M. J., DeRisi J. L., Bennett H. A., Iyer V. R., Meyer M. R., Roberts C. J., Stoughton R., Burchard J., Slade D., Dai H., Bassett D. E. Jr, Hartwell L. H., Brown P. O., e S. H. Friend [Rosetta/Stanford] (1998). Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nature Medicine*, **4**:1293–301.
- McLachlan, G. J., Do, K-A. e Ambroise, C. (2004). *Analyzing Microarray Gene Expression Data*. Wiley.
- McNeil, B. J. e Hanley, J. A. (1984). Statistical Approaches to the Analysis of ROC curves. *Medical Decision Making*, **4**(2):136–149.
- Metz, C.E. e Pan, X. (1999) Proper binormal ROC curves: Theory and maximum-likelihood estimation. *Journal of Mathematical Psychology*, **43**:1–33.
- Murphy, D. (2002). Gene expression studies using microarrays: Principles, problems, and prospects. *Advances in Physiology Education*, **26**(4):256-270.
- Naef, F., Lim, D. A. e Magnasco, M. (2002). Dna hybridization to

Referências

- mismatched templates: A chip study. *Physical Review E*, **65**(4).
- Nguyen, D.V., Wang, N., Carroll, R.J.(2004). Evaluation of Missing Value Estimation for Microarray Data. *Journal of Data Science*, **2**:347–370.
- Nuwaysir E. F., Bittner M., Trent J., Barrett J. C., e C. A. Afshari (1999). Microarray and Toxicology: The advent of Toxigenomics . *Molecular Carcinogenesis*, **24**:153–159.
- Oba, S., Sato, M.A., Takemasa, I., Monden, M., Matsubara, K. e Ishii, S. (2003). A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, **19**(16):2088–2096.
- Ouyang, M., Welsh, W.J., Georgopoulos, P.(2004). Gaussian mixture clustering and imputation of microarray data. *Bioinformatics*, **20**(6):917–923.
- Parodi, S., Muselli, M., Fontana, V. e Bonassi, S. (2003). ROC curves are a suitable and flexible tool for the analysis of gene expression profiles. *Cytogenet Genome Research*, **101**(1):90–91.
- Parodi, S., Pistoia, V. e Musseli, M. (2008). Not proper ROC curves as new tool for the analysis of differentially expressed genes in microarray experiments. *BMC Bioinformatics*, **9**:410.
- Penalva, L. O. F. e Zario, D. A. R. (2001). A leitura do DNA: Como é processada a informação dos genes. *Ciência Hoje*, **29**(171):34–39.
- Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. Oxford (UK): Oxford University Press.
- Pepe, M. S., Longton, G. M., Anderson, G. L. e Schummer, M. (2003). Selecting differentially expressed genes from microarray experiments. *UW Biostatistics Working Paper Series*, nº 184.
- Perez-Diez, A., Morgun, A. e Shulzhenko, N. (2007). Microarrays for cancer diagnosis and classification. *Advances in Experimental Medicine and Biology*, **593**:74–85.
- Pollard, K. S., Ge, Y., Taylor, S. e Dudoit, S. (2009). multtest: resampling-based multiple hypothesis testing. R package version 1.22.0.

- Pounds, S. B. (2005). Estimation and control of multiple testing error rates for microarray studies. *Briefings in Bioinformatics*, **7**(1):25–36.
- Pounds, S. e Cheng, C.(2004) Improving false discovery rate estimation. *Bioinformatics*, **20**:1737-45.
- Quackenbush, J. (2001). Computational analysis of microarray data. *Nature Reviews Genetics*, **2**(6), 418427.
- Reinke V., Smith H. E., Nance J., Wang J., Van Doren C., Begley R., Jones S. J., Davis E. B., Scherer S., Ward S., e S. K. Kim (2000.) A global profile of germline gene expression in *C. elegans*. *Molecular Cell*, **6**:605–616.
- Rocke, D. M. e Durbin, B. (2001) A model for measurement error for gene expression arrays. *Journal of Computational Biology*, **8**(6):557-569.
- Rosenblatt, M. (1956). Remarks on some nonparametric estimate of a density function.*Annals of Mathematical Statistics*, **27**:832-837.
- Ross D. T., Scherf U., Eisen M. B., Perou C. M., Rees C., Spellman P., Iyer V., Jeffrey S. S., Van De Rijn M., Waltham M., Pergamenschikov A., Lee J. C., Lashkari D., Shalon D., Myers T. G., Weinstein J. N., Botstein D., e P. O. Brown (2000). Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics*. **24**:227–35.
- Saal, L. H., Troein, C., Vallon-Christersson, J., Gruvberger, S., Borg, A. e Peterson, C. (2002). BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data. *Genome Biology*, **3**(8).
- Santos, J. M, Silva, R. M., Domingues, P., Amado, F. e Santos, M. S. (2001). Genómica funcional em Aveiro. *Boletim de biotecnologia*, **68**:13–18.
- Sarle, W. S. (2002). Artificial neural network FAQ - should I normalize/standardize/rescale the data? <ftp://ftp.sas.com/pub/neural/FAQ.html>.
- Sartor, M. A., Tomlinson, C. R., Wesselkamper, S. C., Sivaganesan, S., Leikauf, G. D. e Medvedovic, M. (2006). Intensity-based hierarchical Bayes method improves testing for differentially expressed genes in microarray

Referências

- experiments. *BMC Bioinformatics*, **7**:538.
- Saviozzi, S. e Calogero, R. A. (2003). Microarray probe expression measures, data normalization and statistical validation. *Comparative and Functional Genomics*, **4**:442–446.
- Schadt, E. E., Li, C., Ellis, B. e Wong, W. H. (2001). Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *Journal of Cellular Biochemistry Supplement*, **37**:120–125.
- Schena, M., Shalon, D., Davis R. W. e Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with complementary DNA microarray. *Science*, **270**: 467–470.
- Scherf U., Ross D. T., Waltham M., Smith L. H., Lee J. K., Tanabe L., Kohn K. W., Reinhold W. C., Myers T. G., Andrews D. T., Scudiero D. A., Eisen M. B., Sausville E. A., Pommier Y., Botstein D., Brown P. O. e Weinstein J. N. (2000). A gene expression database for the molecular pharmacology of cancer. *Nature Genetics*, **24**:236–44.
- Schmid, F. e Schmidt, A. (2006). Nonparametric estimation of the coefficient of overlapping-theory and empirical application. *Computational Statistics and data Analysis*, **50**:1583–1596.
- Shedden, K., Chen, W., Kuick, R., Ghosh, D., Macdonald, J., Cho, K. R., Giordano, T. J., Gruber, S. B., Fearon, E. R., Taylor, J. M. G. e Hanash, S. (2005). Comparison of seven methods for producing Affymetrix expression scores based on False Discovery Rates in disease profiling data. *BMC Bioinformatics*, **6**:26.
- Scheel, I., Aldrin, M., Glad, I. K., Sorum, R., Lyng, H. e Frigessi, A. (2005). The influence of missing value imputation on detection of differentially expressed genes from microarray data. *Bioinformatics*, **21**(23):4272–4279.
- Silva, C. (2004). As curvas ROC como instrumento na análise estatística de testes de diagnóstico. Tese de mestrado. DEIO-FCUL.
- Silva-Fortes, C., Amaral Turkman, M. A. e Sousa, L. (2006). Aplicação das curvas ROC na análise de dados de microarrays. *Livro de Actas Estatística e Qualidade na Saúde, EQS2006*, 15–21.

- Silva-Fortes, C., Barreto-Hernandez, E., Sousa, L., Amaral Turkman, M. A. e Gama-Carvalho, M. (2007). Análise por microarrays de complexos ribonucleoproteicos: normalização e selecção de genes com associação diferencial. *Actas do XV Congresso Anual de Estatística*, 537–547.
- Silva-Fortes, C. (2011). *Testes de Diagnóstico e Curvas ROC*. Em Bioestatística e Qualidade na Saúde. Cunha, G., Eiras, M. e Teixeira, N. (Eds). LIDEL.
- Silva-Fortes, C., Amaral Turkman, M. A. e Sousa, L. (2012). Arrow plot: a new graphical tool for selecting up and down-regulated genes and genes differentially expressed on sample subgroups. *BMC Bioinformatics*, 13:147.
- Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, **3**(1):Article 3.
- Smyth, G. K. (2005). *Limma: linear models for microarray data*. In: *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, W. Huber (eds.), Springer, New York, pp. 397–420.
- Stanton, L. W. (2001). Methods to profile gene expression. *Trends in Cardiovascular Medicine*, **11**(2):49-54.
- Statnikov, A., Aliferis, C. F., Tsamardinos, I., Hardin, D., Levy, S. (2005). A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, **21**(5):631–643.
- Stekel, D. (2003). *Microarray Bioinformatics*. Cambridge University Press.
- Storey, J. D. (2001). *A direct approach to false discovery rates*. Technical Report. Stanford, CA: Stanford University.
- Storey, J. D.(2002). A direct approach to false discovery rates. *Journal of Royal Statistics Society B*, **64**:479-498.
- Storey, J. D. (2003a). A direct approach to false discovery rates. *Journal of Royal Statistical Society -B*, **24**:479–498.
- Storey, J. D. (2003b). The positive false discovery rate: a Bayesian

Referências

- interpretation and the q-value. *Annals of Statistics*, **31**:201335.
- Storey, J. D. e Tibshirani, R. (2003). Statistical significance for genome-wide studies. *Proceeding of the National Academy of Sciences*, **100**:9440–9445.
- Tan, P. K., Downey, T. J., Spitznagel, E. L., Xu, P., Fu, D., Dimitrov, D. S., Lempicki, R. A., Raaka, B. M. e Cam, M. C. (2003). Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acid Research*, **31**:5676–5684.
- Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**(6):520–525.
- Tuikkala, J., Elo, L. L., Nevalainen, O. S., Aittokallio, T. (2008) Missing value imputation improves clustering and interpretation of gene expression microarray data. *BMC Bioinformatics*, **9**:202.
- Tusher, V. G., Tibshirani, R. e Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *PNAS*, **98**(9):5116–5121.
- Velculescu, V. E., Zhang, L., Vogelstein, B. e Kinzler, K. W. (1995). Serial analysis of gene expression. *Science*, **270**:484-487.
- Wang, D. G., Fan J.B., e E. S. Lander (1998). Large-scale identification, mapping, and genotyping of singlenucleotide polymorphisms in the human genome. *Science*, **280**:1077–82.
- Wang, X., Li, A., Jiang, Z. e Feng, H.(2006) Missing value estimation for DNA microarray gene expression data by Support Vector Regression imputation and orthogonal coding scheme. *BMC Bioinformatics*, **7**:32.
- Watson, J. D. e Crick, F. H. C. (1953). Molecular Stucture of Nucleic Acids: A structure for Deoxyribose Nucleic Acid. *Nature*, **171**:737–738.
- Weitzman, M. S. (1970). *Measure of the overlap of income distribution of white and negro families in the United States*. Technical Report No. 22, U.S. Department of Commerce, Bureau of the Census, Washington, D.C.
- Welch, B. L. (1947). The generalization of Student's problem when several

different population variances are involved. *Biometrika*, **34**:28–35.

Workman, C., Jensen, L. J., Jarmer, H., Berka, R., Gautier, L., Nielsen, H. B., Saxild, H.-H., Nielsen, C., Brunak, S. e Knudsen, S. (2002). A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biology*, **3**(9):research0048.

Wu, J. Y. e Maniatis, T. (1993). Specific interactions between proteins implicated in splice site selection and regulated alternative splicing. *Cell*, **75**:1061–1070.

Wu Z., Irizarry, R. A., Gentleman, R., Murillo, F. M. e Spencer, F.(2004). A model-based background adjustment for oligonucleotide expression arrays. *Journal of the American Statistical Association*, **99**:909–917.

Zhang, L., Miles, M. F. e Aldape, K. D. (2003). A model of molecular interactions of short oligonucleotide microarrays. *Nature Biotechnology*, **21**: 818–821.

Zhou, L. e Rocke, D. M. (2005). An expression index for Affymetrix GeneChips based on the generalized logarithm. *Bioinformatics*, **21**: 3983–3989.

Zou, K. H., Hall, W. J. e Shapiro, D. E. (1997). Smooth non-parametric receiver-operating characteristic (ROC) curves for continuous diagnostic tests. *Statistics in Medicine*, **16**:2143-2156.

Referências
