Sara
Leitão Roque

Metodologias estatísticas para análise de níveis de expressão genética.

Sara
Leitão Roque

Metodologias estatísticas para análise de níveis de expressão genética.

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Matemática e Aplicações, área de especialização Matemática Empresarial e Tecnológica, realizada sob a orientação científica da Doutora Adelaide de Fátima Baptista Valente Freitas, Professor Auxiliar do Departamento de Matemática da Universidade de Aveiro e co-orientação científica da Doutora Laura Cristina da Silva Carreto, Investigadora Auxiliar do Departamento de Biologia da Universidade de Aveiro.

à minha orientadora, Prof. Doutora Adelaide de Fátima Baptista Valente Freitas, pela dedicação, paciência e acima de tudo pela amizade

o júri / the jury

presidente / president

Doutora Isabel Maria Simões Pereira
Professora Auxiliar do Departamento de Matemática da Universidade de Aveiro

vogais / examiners committee

Doutora Lisete Maria Ribeiro de Sousa
Professora Auxiliar do Departamento de Estatística e Investigação Operacional da Universidade de Lisboa

Doutora Adelaide de Fátima Baptista Valente Freitas
Professora Auxiliar do Departamento de Matemática da Universidade de Aveiro (orientadora)

Doutora Laura Cristina da Silva Carreto
Investigadora Auxiliar do Departamento de Biologia da Universidade de Aveiro (co-orientadora)

agradecimentos / acknowledgements

À minha orientadora, Prof. Doutora Adelaide de Fátima Baptista Valente Freitas, pela excelente orientação e supervisão, pela dedicação, paciência e amizade sempre demonstradas ao longo do desenvolvimento deste trabalho.

À minha co-orientadora, Doutora Laura Cristina da Silva Carreto, pela disponibilidade sempre demonstrada em esclarecer todas as questões da área da biologia e pela revisão dos conceitos biológicos presentes nesta dissertação.

À minha família pelo apoio incondicional, pela constante motivação e pela criação de todas condições para a realização desta dissertação.

Ao Rui, meu companheiro, por todo o amor e pela paciência nos dias em que estive menos presente.

A todos os meus mais sinceros agradecimentos.

Palavras-chave

Nível de expressão genética, *Microarray* de DNA, SAM, Modelos lineares, Métodos de Bayes empíricos.

Resumo

A tecnologia de *microarrays* de DNA permite monitorizar a expressão de milhares de genes em simultâneo, constituindo um instrumento de grande apoio à investigação de grandes questões nas áreas da Biologia Molecular, Genética, Medicina, entre outras.

O uso de ferramentas estatísticas que permitam a detecção de genes diferencialmente expressos torna-se imprescindível no sentido de fornecer ao biólogo a identificação de diferenças entre as várias amostras comparadas durante a experiência de *microarrays*.

Nesta dissertação serão abordadas diferentes metodologias estatísticas com vista à detecção de genes que evidenciam diferenças significativas nos níveis de expressão sob duas condições distintas. Concretamente, estuda-se o procedimento estatístico de Análise de Significância de *Microarrays* (SAM) e vários métodos de Bayes empíricos. A metodologia SAM permite estabelecer a partir do valor observado de uma estatística de teste para cada gene, usando o método das permutações e controlando a taxa de falsas descobertas, quais os genes com níveis de expressão significativamente diferentes. Os métodos de Bayes empíricos assumem o ajustamento dos níveis de expressão genética a um dado modelo probabilístico teórico o qual, por sua vez, depende de uma distribuição *a priori* para o modelo dos parâmetros, sendo que os parâmetros da distribuição *a priori* são estimados com base nos dados observados. No presente trabalho serão abordadas quatro metodologias inseridas nos métodos de Bayes empíricos: um modelo linear e os modelos Gamma-Gamma, Log-Normal-Normal e Log-Normal-Normal com Variância Modificada.

Com o auxílio de *packages* do R obtidos do Bioconductor (nomeadamente, `limma` e `EBarrays`) e do package do R `samr`, aplicaram-se as metodologias referidas a duas bases de dados reais designadas por `ApoAI` e `Fermentation`. A `ApoAI` visa o estudo de ratos cujo gene ApoAI não está funcional e a forma como a deficiência deste gene afecta o desempenho dos outros genes no fígado. A base de dados `Fermentation` resulta de uma experiência de duas cores de *microarrays* de DNA recentemente realizada no Laboratório de *Microarrays* da Universidade de Aveiro. A análise destes dados visa comparar os níveis de expressão genética de cinco leveduras vínicas e duas leveduras não vínicas e identificar genes que permitam distinguir estirpes com uma boa resistência ao stress imposto pelo processo de fermentação.

Os resultados obtidos com cada uma das metodologias foram analisados e comparados obtendo-se uma lista de genes comuns identificados por todas as metodologias.

Key-words Gene expression level, DNA Microarray, SAM, linear models, empirical Bayes methods.

Abstract

The technology of DNA microarrays allows the monitoring of the expression levels of thousands of genes simultaneously in a single experiment. It has become a useful tool to support research in the fields of Molecular Biology, Genetics and Medicine, helping scientists to understand the patterns of gene activity in different cellular conditions. In the field of Statistics, the large amount of complex data emerging from DNA microarray technologies has created new challenges and stimulated the development of new methods.

In this dissertation, different statistical methodologies developed for the detection of differentially expressed genes in microarray experiments were studied and applied on two experimental datasets. These methodologies were, specifically, Significance Analysis of Microarrays (SAM) and various procedures based on empirical Bayes methods. The SAM procedure is a permutation-based statistical technique which considers gene specific statistical tests and measures the strength of the relationship between gene expression and condition types in order to decide whether there are statistically significant differences in gene expression levels, controlling the false discovery rate. Empirical Bayes procedures are bayesian methodologies in which the prior distribution for the model parameters is estimated from the data. Herein, four different theoretical models for the expression levels were included in the empirical Bayes approach: linear model, Log-Normal-Normal model, Gamma-Gamma model and, finally, the Log-Normal-Normal with modified variance model.

Using R packages (namely, `samr` and both `limma` and `EBarrays` from Bioconductor), those methodologies were applied on two real databases designated `ApoAI` and `Fermentation`. The `ApoAI` database has been largely studied in the specialized literature and it is aimed at identifying genes with altered expression in mice whose Apo AI gene is not functional. The `Fermentation` database was recently obtained at the National Facility for DNA Microarray at the University of Aveiro, and it comes from two colour DNA microarray experiment carried out to distinguish yeast strains with good resistance to stress imposed by the fermentation process.

The results generated with each methodology for each database were analyzed and compared to obtain a list of differentially expressed genes commonly identified by methodologies applied.

Sois belas, mas vazias. Não se pode morrer por vós. Minha rosa, sem dúvida um transeunte qualquer pensaria que se parece convosco. Ela sozinha é porém mais importante que vós todas, pois foi a ela que eu reguei. Foi a ela que pus a redoma. Foi a ela que abriguei com o para-vento. Foi dela que eu matei as larvas. Foi a ela que eu escutei queixar-se ou gabar-se, ou mesmo calar-se algumas vezes. É a minha rosa.

(ANTOINE DE SAINT-EXUPÉRY)

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Papel biológico dos genes

O genoma dos organismos eucarióticos encontra-se compartimentalizado no núcleo das células, dividido em estruturas condensadas de ADN de dupla hélice designadas por cromossomas [35]. Cada cromossoma contém informação relevante para a síntese de novas proteínas, as unidades de estrutura e catálise de reacções químicas das células, compostas por aminoácidos, estando esta informação codificada em unidades genómicas designadas por genes, numa sequência de quatro tipos diferentes de bases azotadas, ou nucleótidos, designados em abreviatura por A, C, T e G. Ver Figura 1.1.

O papel biológico da maioria dos genes é armazenar informação que especifica a composição química das proteínas ou os sinais regulatórios que irão conduzir à sua produção pela célula. Essa informação é codificada pela sequência de nucleótidos.

A estrutura primária de uma proteína é uma cadeia linear de aminoácidos[1], chamada polipeptídeo.

A primeira etapa adoptada pela célula para a formação de uma proteína é a chamada Transcrição, que consiste em transcrever a sequência de nucleótidos de apenas um dos filamentos do gene numa molécula unifilamentar complementar chamada ácido ribonucleico (ARN [2]). O transcrito de ARN representa uma "cópia funcional"da informação de um gene, um tipo de molécula "mensageira"chamada ARN mensageiro (ARNm [3]). O ARNm terá então a função de orientar a formação da proteína.

Terminado o processo de Transcrição é iniciado o processo de Tradução, sendo produzida uma cadeia de aminoácidos com base na sequência de nucleótidos do ARNm. A sequência de nucleótidos do ARNm é lida em grupos de três bases sucessivas que se denominam de codões. Dado que existem quatro nucleótidos diferentes, existem 64 codões

[1]Moléculas orgânicas formadas por átomos de Carbono, Hidrogênio, Oxigênio, e Nitrogênio unidos entre si.

[2]Em inglês *ribonucleic acid* (RNA). As principais diferenças entre o ARN e o ADN são poucas, no entanto faz com que o segundo seja mais estável que o primeiro. O ARN é formado por uma cadeia simples, sendo que a Timina é substituida pelo Uracilo.

[3]Em inglês *messenger RNA* (mRNA).

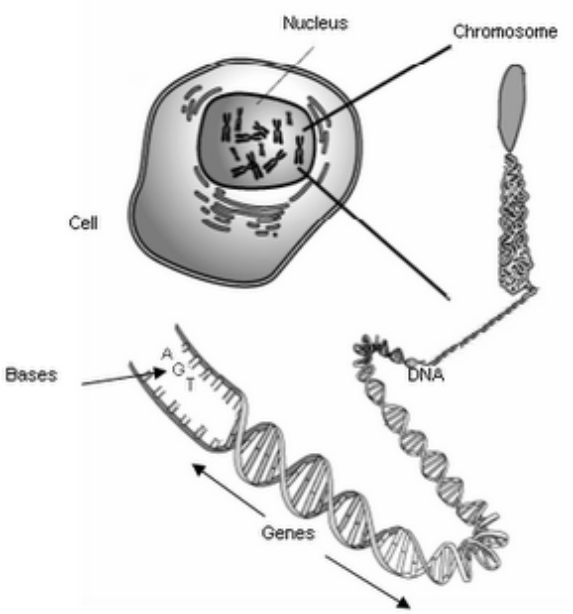

Figura 1.1: Ampliações sucessivas mostrando o material genético de uma célula. Na imagem pode ser vista uma célula contendo os cromossomas no seu núcleo. Cada cromossoma é uma longa sequência de ADN, que contém vários genes. A estrutura do ADN forma uma dupla hélice constituida por nucleótidos. Cada nucleótido contém uma das quatro bases nitrogenadas: Adenina (A), Guanina (G), Timina (T) ou Citosina (C). Imagem retirada de [33].

diferentes, cada um codificando um aminoácido ou sinal de finalização da Tradução.

A transcrição e posterior tradução de gene em proteína é designada por expressão desse gene. A activação ou repressão da expressão de grupos específicos de genes determinam a função e características da célula. A diferenciação celular pode ser vista como o processo de especialização das células vivas para realizar determinada função. Esta especialização conduz a alterações da função e do nível da estrutura celular.

O processo inverso também poderá ocorrer. As células já especializadas podem perder a sua função originando alterações celulares que conduzem a uma perda do controlo e autonomia do seu crescimento levando a uma proliferação celular anormal. Neste processo, em consequência de eventuais mudanças genéticas que regulam o crescimento e a diferenciação celular, as células reduzem ou perdem a capacidade de se diferenciar originando, por exemplo, as neoplasias (tumores).

A regulação genética dá à célula controle sobre a sua estrutura e função, sendo a base da diferenciação celular. As propriedades fisiológicas das células são largamente determinadas pelas proteínas activas e expressas nelas, pelo que a regulação permite a correcta adaptação às variações particulares de circunstância, tal como disponibilidade de nutrientes, invasão de agentes infecciosos, mudança de temperatura ou outras mudanças no estado de desenvolvimento da célula.

A expressão genética é a informação contida num gene que leva ao processamento de, por exemplo, uma proteína [35]. Assim, o estudo da expressão dos genes nas diferentes

células permite entender como as células funcionam normalmente e como são afectadas quando os genes não conduzem à correcta diferenciação celular.

1.2 *Microarrays* de ADN

O princípio usado pela técnica de *microarrays* baseia-se no facto de uma sequência de nucleótidos se colar ou hibridar à sua sequência complementar. Normalmente um único *microarray* contém milhares de pontos específicos e individualizados (*spots*) em que cada um representa um único gene e, eventualmente, colectivamente o genoma inteiro de um organismo, ver Figura 1.2 [37].

A tecnologia de *microarrays* de ADN permite a análise da expressão de milhares de genes em simultâneo, constituindo um instrumento de grande apoio à investigação de grandes questões nas áreas da Biologia Molecular, Genética, Medicina, entre outras.

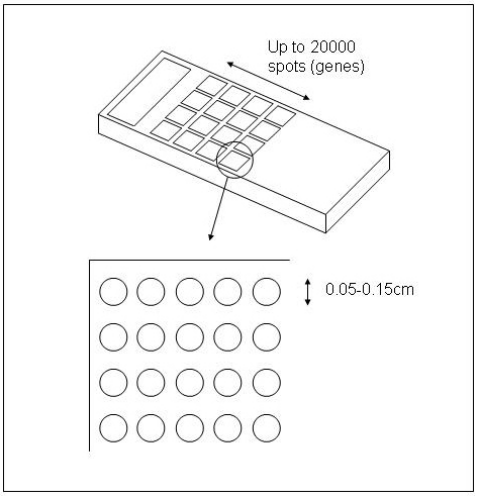

Figura 1.2: Ilustração de um *microarray* e *spots*. Cada *microarray* contém milhares de *spots*. Cada *spot* tem normalmente um diâmetro de 0.05 a 0.15 centímetros e representa apenas um gene. Os *spots* estão agrupados e ordenados por linhas e colunas. Imagem retirada de [38].

Uma experiência de estudos de expressão genética usando microarrays de ADN assume sete passos essenciais (Figura 1.3). Um maior detalhe sobre estes sete passos pode ser encontrado em [15, 20].

1. **Impressão dos *microarrays*.** A obtenção de *microarrays* de ADN é geralmente conseguida com auxílio de um robot especializado, chamado *Arrayer*, que fixa pequenas quantidades de ADN nos *spots* das lâminas de vidro. As quantidades de ADN fixadas na lâmina são muitas vezes chamadas sondas (*probes*), onde cada uma contém as sequências de um único gene.

2. **Planeamento experimental.** Num planeamento experimental é aconselhado incluir réplicas biológicas e técnicas das amostras de interesse, por forma a poder lidar com

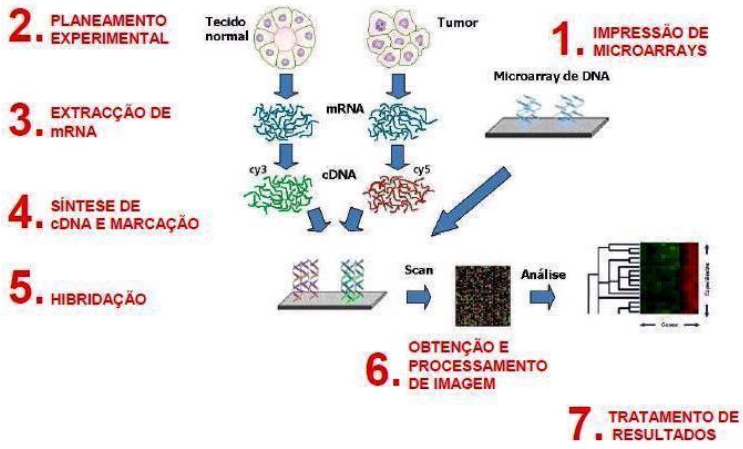

Figura 1.3: Esquema dos sete passos envolvidos numa experiência com *microarrays*. O primeiro passo constitui a Impressão dos *microarrays*, o segundo pelo planeamento experimental, no terceiro passo é feita a extracção de ARNm, no quarto a síntese de ADNc e a sua marcação, o quinto passo é constituido pela hibridação das amostras, no sexto são obtidas as imagens dos microarrays e por último feito tratamento de resultados. Imagem retirada de [15].

a variabilidade técnica e biológica. Entende-se por variabilidade técnica o grau de variabilidade obtido quando usadas técnicas diferentes (por exemplo, *dye swap*). A variabilidade biológica é devida às alterações de factores genéticos, idade, sexo, entre outros. A combinação de amostras para hibridação deve ser optimizada de modo a minimizar o número de *microarrays*, mas de forma a que o número de *microarrays* permita obter resultados conclusivos.

3. **Extracção de ARNm.** Para a maior parte dos genes, uma maior quantidade de ARNm traduz-se numa maior abundância celular da proteína que este codifica. Este passo é muito importante, já que o sucesso da posterior análise depende da qualidade da recolha.

4. **Síntese de ADNc e marcação.** A reacção de síntese de ADN complementar (ADN [4]) é realizada num recipiente apropriado (um tubo *eppendorf*). No recipiente adicionam-se:

   - Solução tampão apropriada
   - ARN total em estudo
   - Nucleótidos livres
   - nucleótido conjugado com fluoróforo (Cy3 ou Cy5)
   - Sequência poli-dT (que hibrida com poli-dA do ARNm)
   - Enzima Transcriptase Reversa (que sintetiza ADN a partir de sequência molde de ARN).

[4] ADN sintetizado a partir de ARN.

A síntese de ADNc é promovida a 42 ºC (temperatura óptima para a actividade da enzima). O ARNm é destruído e o ADNc é purificado recorrendo à filtração em microfiltros, sendo esta feita por centrifugação. O ADNc é retido enquanto os contaminantes da reacção são lavados através da membrana porosa.

A marcação do ADNc é feito com fluorescência. Adiciona-se um fluoróforo que emite fluorescência quando excitado com luz de energia apropriada, sendo que o fluoróforo Cy3 emite fluorescência com um máximo a 550 nm e o fluoróforo Cy5 com máximo a 650nm. Estes são bastante diferenciáveis pelos instrumentos ópticos, pelo que podem ser utilizados em conjunto. Uma vez que estes emitem luz de energia distinta a sua diferenciação é facilitada.

Moléculas de Cy5 ou Cy3 não ligadas ao ADNc são contaminantes indesejados e são eliminados por filtração em gel. A solução contendo o ADNc é passada através do gel com uma breve centrifugação. Como as moléculas de fluoróforo livre são menores, são retidas durante mais tempo nos poros do gel e não são recuperadas.

5. **Hibridação.** Para a preparação da hibridação as amostras marcadas são misturadas e diluídas em tampão apropriado de modo a que a interacção específica de ácidos nucleicos com sequências complementares, ou sondas, imobilizadas no *microarray*, seja estabilizada. A mistura de amostras (que permitirá a comparação de níveis de expressão) é colocada em contacto com o *microarray* recorrendo a suportes de metal que fazem o ajuste estanque entre o *microarray* e uma câmara de hibridação constituída por uma lâmina de vidro das dimensões do *microarray* delimitada por um vedante de borracha. A hibridação é feita num forno de incubação térmica entre 16 e 40 horas. O tempo de incubação é calculado experimentalmente. O objectivo será obter o equilíbrio químico entre a concentração de cadeias de ADNc livres e imobilizadas. Usualmente a incubação é feita entre 40 e 65 ºC. A temperatura é escolhida tendo em conta o tamanho das sondas e a composição de solução de hibridação.

   Feita a hibridação o suporte de metal é aberto e o slide é separado da lâmina com vedante. Recorrendo a uma série de lavagens é removido o excesso de amostra não hibridada.

   De modo a evitar a formação de manchas, o *microarray* é colocado no interior de um tubo e seco por centrifugação.

   O *microarray* é depois colocado no interior de um suporte adaptador para scanner.

6. **Obtenção e processamento de imagem** A imagem da hibridação é obtida através de um scanner de *microarrays* contendo lasers capazes de promover a emissão de fluorescência de Cy3 e Cy5. A medição da fluorescência é feita recorrendo a um *software* específico que detecta a posição e dimensões das manchas de sinal e ruído de fundo associados a cada sonda (Figura 1.4). O ruído de fundo (*background*) é corrigido antes de se proceder à normalização e análise dos resultados.

7. **Tratamento de resultados.** Obtidas as medidas de expressão genética e corrigido o *background*, metodologias estatísticas são aplicadas com vista à obtenção de resultados que conduzam à resposta de algumas questões biológicas.

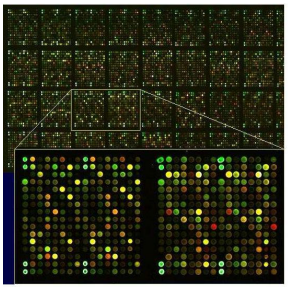

Figura 1.4: Imagem de um *microarray* obtida por scanner, como referido no Passo 6 (Obtenção e processamento de imagem). Na figura está a ampliação de uma parte do *microarray* sendo visível a intensidade do sinal de cada *spot*. Imagem adaptada de [15].

O nível de expressão de cada gene é calculado em termos do sinal de intensidade dado por:

$$\frac{\text{Intensidade de fluorescência do fluoróforo Cy5}}{\text{Intensidade de fluorescência do fluoróforo Cy3}},$$

que indica a abundância relativa nas duas amostras hibridadas. Se o quociente é inferior a uma unidade então o gene sob as condições da amostra que foi etiquetada com Cy3 é mais expresso; se superior à unidade então é mais expresso o gene sob as condições da amostra que foi etiquetada com Cy5.

1.3 Objectivos e estrutura da dissertação

Nesta dissertação pretende-se efectuar um estudo comparativo de três metodologias estatísticas que visam a detecção de genes diferencialmente expressos para duas classes em estudo. As metodologias estudadas são:

- A Análise de Significância de Microarrays (SAM [5]), proposta por Tusher, Tibsirani e Chu [27, 29] e implementada por Narasimhan e Tibsirani no pacote `samr` em linguagem R. A SAM calcula uma estatística de teste para cada gene de forma a medir a influência da variável resposta, ou classe (ex. tratamento e controlo), no nível de expressão desse gene. Para determinar se cada gene é ou não expresso faz uso de repetidas permutações dos dados.

- Os Métodos de Bayes Empíricos, em que os parâmetros são tratados como quantidades aleatórias, sendo-lhes associada uma distribuição *a priori*. O objectivo será estimar as probabilidades *a posteriori* de um dado gene ser diferencialmente expresso.

[5] Do inglês *Significance Analysis of Microarrays*

Várias metodologias têm sido propostas dependendo do modelo associado aos valores observados. Dentro dos Métodos de Bayes Empíricos, são abordadas duas metodologias:

Modelos Lineares. Esta metodologia foi sugerida e implementada no pacote `limma` do Bioconductor[6] por Gordon Smyth com contribuições de Matthew Ritchie, Natalie Thorne, James Wettenhall e Wei Shi [22, 19, 14, 13]. É assumido que os dados podem ser modelados por um modelo linear. Para este modelo são definidos os contrastes de interesse através de uma matriz de delineamento. São efectuados testes para verificar se estes contrastes são nulos, ou seja, se um dado gene não é diferencialmente expresso.

Métodos de Bayes Empíricos Paramétricos para *microarrays*. Esta metodologia foi desenvolvida por Michael A. Newton e Christina Kendziorski [18, 17, 16] e implementados no pacote `EBarrays` do Bioconductor com a colaboração de Ming Yuan, Ping Wang e Deepayan Sarkar [12]. Este pacote contém a implementação de três modelos:

O modelo Gama-Gama Considera uma distribuição Gama para os dados observados e uma distribuição Gama para a distribuição *a priori* do parâmetro de escala.

O modelo Lognormal-Normal Baseado numa distribuição Lognormal para as observações e uma distribuiçãao Normal para a distribuição *a priori* do parâmetro de escala.

O modelo Lognormal-Normal com Variância Modificada As condições são as mesmas do modelo anterior, mas em vez de considerar uma variância comum, considera-se uma variância para cada gene.

Cada um destes procedimentos foi aplicado em duas bases de dados, uma que se encontra amplamente estudada na literatura especializada (`ApoAI`) e uma outra recentemente obtida no Laboratório de *Microarrays* do Departamento de Biologia da Universidade de Aveiro (`Fermentation`).

- A `ApoAI`[7] pode ser obtida em http://bioinf.wehi.edu.au/marray/ibc2004/apoai.zip. Com esta base de dados pretende-se efectuar um estudo comparativo entre ratos cujo gene (ApoAI) não está funcional e ratos cujo gene (ApoAI) está funcional. Sabe-se que no primeiro caso os níveis de colesterol são muito baixos, pelo que o objectivo de estudo é avaliar de que forma a deficiência daquele gene afecta o desempenho dos outros genes do fígado (onde é recolhida a amostra), ou seja, que alterações dos níveis se detectam entre as duas classes de ratos.

[6] Projecto de *software* aberto baseado essencialmente na linguagem de programação do R, que permite a análise e compreensão de dados obtidos em estudos laboratoriais na Biologia Molecular. Disponível em www.bioconductor.org/

[7] Também conhecida por *apolipoproteinAI*.

A base de dados é composta por 16 *microarrays*, onde uma das classes é composta por 8 *microarrays* construídos a partir de ratos selvagens (*black six*) "normais"e a outra classe composta pelos restantes 8 *microarrays* construídos a partir de ratos cujo o gene ApoAI é deficiente. Para cada um dos 16 ratos foi extraído ARNm do fígado e etiquetado com Cy5 (vermelho). O ARN de cada rato foi hibridado em *microarrays* distintos com uma amostra de referência obtida de 8 ratos e rotulada com Cy3 (verde). Para cada rato foram selecionados 6382 genes. A questão essencial para esta base de dados será a determinação dos genes diferencialmente expressos.

- Na `Fermentation` foram recolhidas amostras de uma série de leveduras, tendo sido seleccionadas para estudo 7 leveduras, 5 vínicas e 2 não vínicas, das não vínicas, uma clínica e a outra laboratorial. Para cada levedura obteve-se dois *microarrays* em 6 tempos diferentes do seu desenvolvimento, um *microarray* com a amostra de referência rotulada com Cy3 e o mRNA da levedura etiquetada com Cy5, e outro com os fluoróforos trocados (*dye swap*), prefazendo um total de 12 *microarrays*. De cada uma das leveduras, obteve-se informação acerca da expressão de 6388 genes. No total a base de dados é composta por 84 microarrays. O objectivo é estabelecer comparações entre leveduras vínicas e não vínicas, leveduras vínicas e clínica e leveduras vínicas e laboratorial. Estas duas últimas comparações são importantes dada a diferente natureza das leveduras não vínicas. Deste modo fará sentido encontrar diferenças de comportamento dos genes considerando estas duas classes. Pretende-se com este estudo identificar genes que permitam distinguir estirpes com uma boa resistência ao stress imposto pelo processo de fermentação. A identificação de genes cuja expressão distinga as leveduras vínicas das leveduras não vínicas será particularmente interessante para desenvolvimento futuro de ferramentas moleculares de identificação de novas estirpes com bons padrões de fermentação alcoólica a partir de estirpes selvagens (isoladas do ambiente). A comparação de leveduras com fenótipo associado à manipulação laboratorial, infecção clínica e fermentação permitirá identificar genes com elevada variabilidade de expressão, bem como padrões de alteração de expressão genética potencialmente associados à patogénese.

A `ApoAI` conta com um grande historial no que respeita ao seu estudo. Inúmeros estudos já foram feitos com esta base de dados, entre os quais estão as referências [19, 24, 25]. Neste sentido para além da comparação dos resultados obtidos para esta base de dados com diferentes metodologias, este trabalho tem como principal contribuição a análise e comparação de resultados obtidos das diferentes metodologias, para a base de dados `Fermentation`, ainda muito pouco divulgada.

Esta dissertação é composta por esta introdução (Capítulo 1) e mais outros três capítulos.

O Capítulo 2 aborda a Expressão Genética e Testes Simultâneos, sendo apresentadas formas de representação de dados de *microarrays* e métodos de pré-processamento devem ser tidos em conta antes da aplicação de qualquer metodologia estatística. É feita uma abordagem à taxa de falsas descobertas (FDR[8]), sendo esta a principal ferramenta de con-

[8]Do inglês *false discovery rate*

trolo do erro em testes simultâneos aplicados na presente dissertação. É ainda estudada a taxa de falsas descobertas positiva (pFDR[9]), uma metodologia alternativa à FDR desenvolvida em [31], e suas principais propriedades. É também desenvolvida uma perspectiva bayesiana da mesma [32].

O Capítulo 3 aborda a Análise de Significância de Microarrays, sendo na primeira secção estudados os conceitos teóricos associados à metodologia em causa. O capítulo é concluído com um estudo experimental aplicando a metodologia às duas bases de dados `ApoAI` e `Fermentation`.

O Capítulo 4 aborda os Métodos de Bayes Empíricos. A primeira secção apresenta uma introdução ao conceito de Bayes empírico, passando pelo algoritmo EM. Na segunda secção são descritos os modelos lineares mais usados na literatura científica para a análise de dados de *microarrays* e alterações sobre este modelo que se espera fornecerem resultados mais estáveis. Os modelos lineares são aplicados às duas bases de dados. A terceira secção descreve a estrutura geral dos modelos desenvolvidos no pacote EBarrays. É feita uma descrição dos três modelos (Gama-Gama, Lognormal-Normal e Lognormal-Normal com Variância Modificada) e a aplicação dos modelos às bases de dados `ApoAI` e `Fermentation`.

Finaliza-se a presente dissertação com o Capítulo 5 apresentando as principais conclusões do trabalho realizado e sugestões para trabalho futuro.

Todos os resultados das análises realizadas foram obtidas recorrendo à linguagem R. Um breve resumo dos comandos e procedimentos utilizados pode ser consultado no Apêndice B.

[9]Em inglês *positive false discorey rate*

Capítulo 2

Expressão Genética e Testes Simultâneos

2.1 Dados de Microarrays

Numa experiência de *microarrays* de dois canais, quando se pretende comparar intensidades, é obtida uma medida relativa de expressão para cada *spot*. Essa medida relativa é normalmente dada pelo quociente

$$\frac{R}{G} = \frac{Intensidade\ do\ fluoróforo\ Cy5}{Intensidade\ do\ fluoróforo\ Cy3}. \tag{2.1}$$

Uma outra forma também muito comum de quantificar o nível de expressão de cada gene é calcular o logaritmo do quociente das intensidades (2.1). Uma das principais vantagens desta transformação é a redução da influência de eventuais valores atípicos, já que a distribuição dos dados torna-se mais simétrica.

A expressão das log-intensidades é geralmente designada por:

$$M = log_2(R/G) = log_2(R) - log_2(G),$$

onde R representa a intensidade do corante vermelho (*Red*), ou seja, a intensidade de Cy5 e G a intensidade do corante verde (*Green*), ou seja, a intensidade de Cy3. A atribuição da notação M deriva do facto da expressão se poder transformar numa subtracção de logaritmos (*minus*).

Uma outra medida bastante utilizada em conjunto numa análise dos valores M é dada por:

$$A = log_2(\sqrt{RG}) = \frac{1}{2}(log_2(R) + log_2(G)).$$

A letra A deriva da palavra em inglês *add* (soma).

2.2 Métodos de Pré-Processamento

2.2.1 Correcção de Background

Ao efectuar a leitura do *microarray* (através de *scan*) é obtida a intensidade específica (conhecida por intensidade de *foreground*) e a intensidade não específica (conhecida por intensidade de *background*) para cada gene. A correcção de *background* visa atenuar os efeitos causados pela imagem de fundo da lâmina que pode emitir alguma florescência por si só, pela falta de contribuição de sinal devido a moléculas que não se tenham hibridado com nenhuma molécula fluorescente ou devido a sinais inespecíficos decorrentes da eventual sujidade da lâmina ou hibridização inespecífica que contaminam o *background* [20]. Este não é o tema alvo desta dissertação, pelo que apenas será feita uma breve descrição de alguns dos métodos de correcção de *background* incluidos no pacote *limma*:

subtract Subtrai as intensidades de *background* às de *foreground*.

movingmin O *background* para cada *spot* é substituido pelo mínimo entre a estimativa de *background* para o *spot* e os seus oito vizinhos mais próximos.

half Todas as intensidades que sejam menores que 0.5, depois da subtracção do *background*, são corrigidos para 0.5.

minimum Todas as intensidades que, após a subtracção do *background*, sejam negativas são corrigidas para metade do mínimo das intensidades corrigidas positivas num determinado *microarray*.

edwards É feita a subtracção do *background* apenas quando a diferença entre as intensidades de *foreground* e *background* é superior a um determinado δ. Quando a diferença é inferior a esse δ, a subtracção é feita recorrendo a uma função monótona suave[1]. O valor de δ depende dos dados e pode ser consultado na função `backgroundCorrect` do *software* R.

normexp Este método é baseado na convolução das distribuições normal e exponencial. A intensidade corrigida passa a ser o valor esperado da intensidade verdadeira conhecendo a intensidade de *foreground* observada. Mais detalhes podem ser encontrados em [10].

O método cujo uso é mais comum é o *subtract*. Um dos problemas dos dois primeiros métodos acima descritos é o facto das intensidades de *background* poderem ser superiores às intensidades de *foreground*, levando à obtenção de resultados negativos. Quando obtidos os logaritmos destas intensidades, obter-se-ão valores omissos. De forma a evitá-los foram desenvolvidos os restantes métodos [11].

[1] O método foi proposto por Edwards D. em Edwards D., Non-linear normalization and background correction in one-channel cDNA microarray studies, Bioinformatics, Vol. 19, No. 7, pp. 825-833, 2003.

2.2.2 Normalização

As experiências com *microarrays* envolvem fontes de variação sistemática que podem afectar as medições dos níveis de expressão genética. Estas variações podem ter diversas causas tais como: diferenças na eficiência da incorporação dos fluoróforos, diferenças na quantidade de ARN inicial utilizado para marcação e hibridação, diferenças de ajuste dos parâmetros do scanner responsável pela leitura dos *microarrays*, falhas na impressão das sondas, imprecisão dos equipamentos utilizados, entre outros [20]. De forma a possibilitar a comparação dos *microarrays*, estas fontes de variação devem ser removidas. As técnicas de normalização são transformações dos dados que visam remover essas fontes de variação.

Existem dois tipos de normalizações (normalização dentro do *array*[2] e entre *arrays*[3]) e ambas estão implementadas no pacote *limma*:

Normalização dentro do Array

A normalização dentro do *array* é feita em cada *microarray* separadamente e pode envolver todos os genes do *microarray* ou apenas uma região deste. Esta normalização permite remover, por exemplo, o viés dos fluoróforos dentro de cada *array*. O método pode ser aplicado usando a função `normalizeWithinArrays()` que produz duas matrizes, uma com os valores M normalizados (M_{norm}) e outra com os valores A normalizados (A_{norm}), obtidos de forma semelhante a M_{norm}. Existem diversas metodologias para normalizar os dados dentro do *microarray*, entre elas as mais aplicadas são:

Normalização global É aplicada quando existe uma relação constante $Cy5 = k \cdot Cy3$. Os dados normalizados são obtidos através da expressão

$$M_{norm} = M - c,$$

onde c denota a mediana ou a média dos valores M do *microarray*.

Normalização dependente da intensidade de expressão Muitas vezes o viés está dependente da intensidade de expressão. Nestes casos será preferível usar métodos não-lineares, tal como a regressão Loess[4] (também conhecida por Regressão Lowess) para estimar a dependência dos valores M em relação às intensidades. Os valores normalizados são obtidos pela expressão

$$M_{norm}(A) = M - c(A),$$

onde $c(A)$ representa o valor ajustado pela regressão Loess e que depende de A.

O efeito da aplicação de um método de normalização pode ser observado num gráfico MA[5]. Nste gráfico, tal como o nome indica, são representados os valores M em função dos valores A. A Figura 2.1 ilustra gráficos MA obtidos antes e depois da normalização para a base de dados `ApoAI`.

[2]Em inglês conhecido por *Normalization Within-Arrays*

[3]Em inglês conhecido por *Normalization Between-Arrays*.

[4]Loess deriva da expressão em inglês *locally weighted polynomial regression*.

[5]Em inglês conhecido por *MA-plot*.

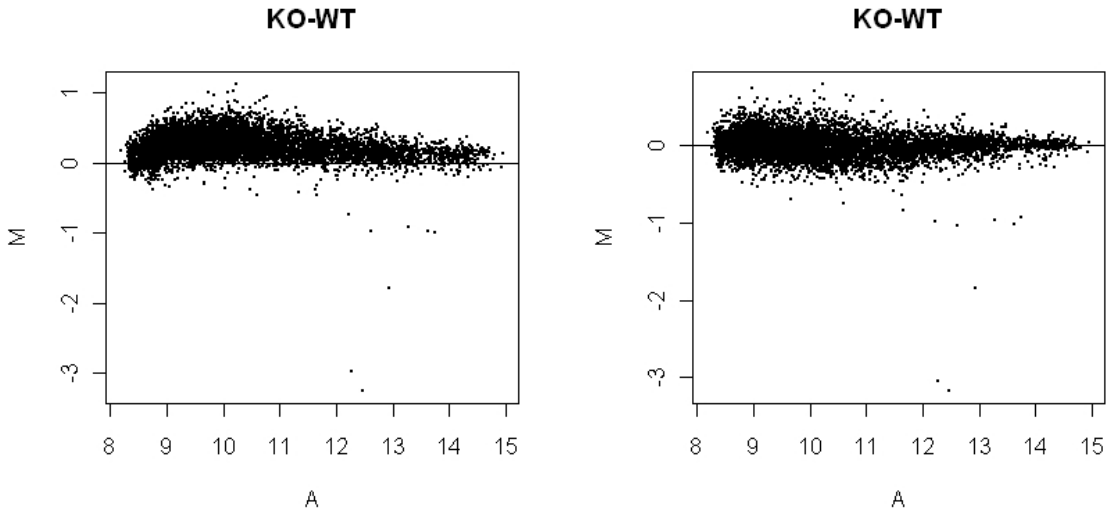

Figura 2.1: O gráfico da esquerda corresponde ao gráfico MA sem normalização e o da direita após a normalização baseada na regressão *Loess*. Verifica-se, neste último a diminuição da dependência dos valores M relativamente aos valores A. KO-WT denota os valores M que comparam as amostras vindas dos ratos cujo gene ApoAI é deficiente(KO) com os ratos cujo gene é normal(WT).

Maior detalhe sobre métodos de normalização dentro do *array* pode ser encontrado em [10].

Normalização entre arrays

A normalizaçao entre os *arrays* é utilizada para permitir a comparação de *microarrays*. É aplicada após a normalização dentro do *array* para correção dos efeitos de escala entre *arrays*. Esta normalização pode ser aplicada com a função `normalizeBetweenArrays()`. Existem diversos métodos subjacentes a este tipo de normalização [11]:

scale Os valores M são corrigidos de forma a que tenham o mesmo desvio mediano absoluto[6] ao longo dos *microarrays*. Este método deve ser usado apenas se não existirem razões biológicas para as diferenças entre os *microarrays*. Este é o método cujo uso é mais comum.

quantile Garante que as intensidades têm a mesma distribuição empírica para todos os *microarrays* e para todos os canais (R e G).

Aquantile Assegura que os valores A têm a mesma distribuição empírica. Os valores M não são alterados.

Gquantile Assegura que os valores G têm a mesma distribuição empírica. Os valores M não são alterados.

[6]Em inglês conhecido por *median absolute deviation* (MAD)

Rquantile Assegura que os valores R têm a mesma distribuição empírica. Os valores M não são alterados.

Tquantile Concretiza a normalização *quantile* separadamente para cada tipo de amostra. Esta é uma normalização bastante útil quando existem motivos para crer que as distribuições serão consideravelmente diferentes para diferentes amostras.

vns Do inglês *variance stabilizing normalization*. As variações são calibradas amostra-a-amostra através de deslocamento e escalonamento dos valores M. Os valores são transformados numa escala de tal forma que a variância é aproximadamente independente das intensidades. É assumido que a maioria dos genes não variam muito de amostra para amostra, ou seja, que existem poucos genes diferencialmente expressos.

2.3 A Taxa de Falsas Descobertas

Num teste de hipóteses o objectivo essencial será, com base em informação fornecida por uma dada amostra, decidir pela rejeição ou não rejeição da hipótese nula H_0 em estudo, quando esta está em confronto com uma hipótese alternativa H_1. A decisão é tomada usando uma estatística de teste T. Definida uma região de rejeição Γ, se $T \in \Gamma$, a hipótese nula é rejeitada, se $T \notin \Gamma$, a hipótese nula não é rejeitada. Um erro de tipo I ocorre quando se rejeita H_0, sendo ela verdadeira. Um erro de tipo II ocorre se H_0 não é rejeitada quando, na verdade a hipótese é falsa. Na Figura 2.2 apresenta-se esquematizado, em termos geométricos as probabilidades de ocorrência desses tipos de erro. Para estabelecer Γ escolhe-se um nível de significância α, para o qual o erro de tipo I é aceitável. A região corresponde a todos os valores da estatística de teste que levam à rejeição da hipótese nula. A probabilidade dessa região é igual ao nível de significância. Assim, a probalidade de erro tipo I é determinada pelo investigador.

Quando a hipótese H_0 é definida por uma intersecção de hipóteses nulas, $H_0 = \cap_{i=1}^{m} H_{0,i}$, tem-se um teste de hipóteses simultâneo, muitas vezes também denominado de teste de hipóteses múltiplas[7]. Nesse caso, se cada hipótese $H_{0,i}$ é testada isoladamente, estar-lhe-á associada um valor para a probabilidade de erro de tipo I e um valor para a probabilidade de erro de tipo II, conforme ilustra, a título de exemplo, a Figura 2.3, não sendo simples a escolha de uma medida única para avaliar o erro global associado ao teste de todas as hipóteses em simultâneo.

Considerando m hipóteses nulas a serem testadas em simultâneo tem-se, associado ao procedimento, um número aleatório de erros de tipo I e de tipo II conforme sumariado na Tabela 2.1.

A primeira medida sugerida para avaliar o erro global em testes de hipóteses simultâneos é conhecida pela abreviatura FWER (*Familywise error rate* [8]), que corresponde à probabilidade de cometer um ou mais erros de tipo I tendo em conta todas as m hipóteses.

[7] Miller, R.G. (1981). Simultaneous Statistical Inference 2nd Ed. Springer Verlag New York.

[8] Hochberg, Y., & Tamhane, A. C. (1987). Multiple comparison procedures. New York: Wiley.

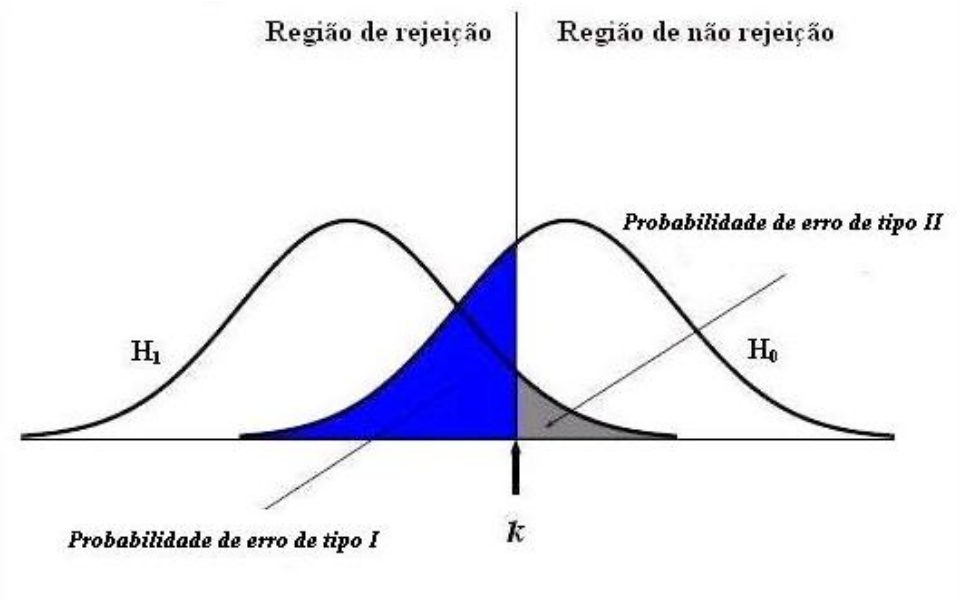

Figura 2.2: Probabilidade de erro de tipo I e probabilidade de erro de tipo II. A azul está a área correspondente à probabilidade de ocorrência de erro de tipo I e a cinza a probabilidade de ocorrência de erro de tipo II. A probabilidade do erro de tipo I corresponde ao nível de significância fixado à partida na realização do teste. Sob a hipótese nula, calcula-se o valor crítico k, que estabelece a fronteira da região de rejeição e corresponde ao quantil associado ao nível de significância definido. A probabilidade de erro de tipo II é definida sob a hipótese alternativa, correspondendo à probabilidade da região complementar. Imagem adaptada de [40].

Formalmente, da notação apresentada na tabela anterior, a FWER é dada por

$$FWER = Pr\left(V \geq 1\right).$$

Assim, em vez de controlar o erro de tipo I para cada teste, o procedimento ao testar H_0 é realizado de tal forma que $FWER \leq \alpha$, onde α representa o máximo de erro global pré-definido para o teste simultâneo. Um procedimento estatístico bem conhecido que permite controlar a medida FWER resulta da aplicação da correcção de Bonferroni, onde as m hipóteses são testadas individualmente ao nível $\frac{\alpha}{m}$, para um nível α escolhido para o teste simultâneo. Na realidade, a correcção de Bonferroni providencia um forte controlo do FWER pretendido para qualquer distribuição da estatística de teste sob a hipótese nula H_0. Considerando $H_{0,i}$ as hipóteses nulas em tese, para $i = 1, 2, ..., m$, Q_0 a distribuição da estatística de teste sob a hipótese nula $H_{0,i}$, m o número de hipóteses a testar, p_i a variável aleatória que representa o valor p associado ao procedimento para testar $H_{0,i}$ usando a distribuição Q_0 da estatística de teste e supondo, sem perda de generalidade, que as primeiras m_0 hipóteses hipóteses são verdadeiras, vem:

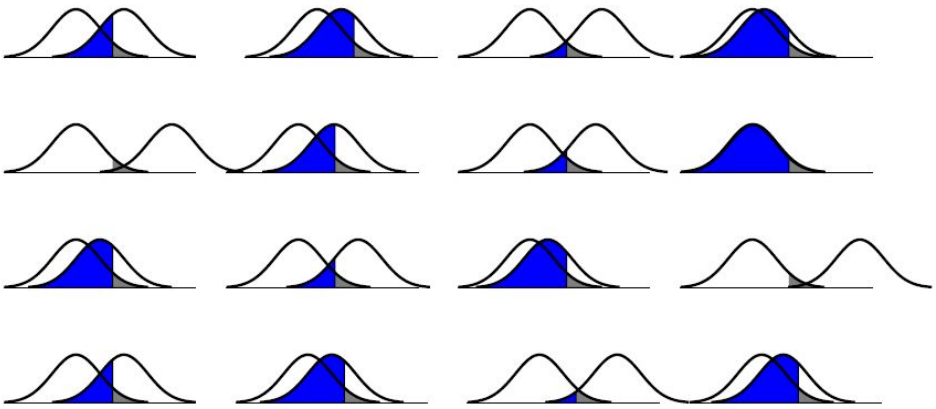

Figura 2.3: Problemática para a determinação de uma medida para o erro global. Visualização da variação da probabilidade de erro de tipo I (a azul) e de tipo II (a cinza) obtidos com a realização em separado de 16 testes para 16 médias de populações normais. Imagem retirada de [40].

Hipóteses	**Não Rejeitadas**	**Rejeitadas**	**Total**
Hipóteses Nulas Verdadeiras	U	V	m_0
Hipóteses Alternativas Verdadeiras	W	S	$m - m_0$
Total	$m - R$	R	m

Tabela 2.1: Resultados possíveis ao testar m hipóteses nulas em simultâneo. V e W representam a quantidade não observável de erros de tipo I e II, respectivamente, que podem ocorrer na realização do teste estatístico. U representa a quantidade de hipóteses nulas verdadeiras não rejeitadas e S a quantidade de hipóteses nulas em teste que são falsas e são rejeitadas. Todas as quantidades excepto R e m não são observáveis. Note-se que R, U, V, W e S constituem variáveis aleatórias, a primeira observável e as restantes não observáveis.

$$
\begin{aligned}
FWER &= Pr(\textit{cometer pelo menos um erro de tipo } I) \\
&= Pr_{Q_0}(\textit{rejeitar } H_{0,i},\ \textit{pelo menos para algum } i = 1, 2, ..., m_0) \\
&= Pr_{Q_0}(p_i \leq \frac{\alpha}{m},\ \textit{pelo menos para algum } i = 1, 2, ..., m_0) \\
&= Pr_{Q_0}\left(\bigcup_{i=1}^{m_0} \left\{p_i \leq \frac{\alpha}{m}\right\}\right) \\
&\leq \sum_{i=1}^{m_0} Pr_{Q_0}\left(p_i \leq \frac{\alpha}{m}\right) \quad \textit{pela desigualdade de Boole} \\
&\leq \sum_{i=1}^{m} Pr_{Q_0}\left(p_i \leq \frac{\alpha}{m}\right) \\
&= \sum_{i=1}^{m} \frac{\alpha}{m}
\end{aligned}
$$

Uma vez assumida a distribuição nula Q_0, p_i pode representar uma função de distribuição e, portanto, seguirá uma distribuição uniforme U[0,1]. Logo:

$$\begin{aligned} FWER &\leq \sum_{i=1}^{m} \frac{\alpha}{m} \\ &= m\frac{\alpha}{m} = \alpha. \end{aligned} \tag{2.2}$$

Outro exemplo de um procedimento estatístico que permite o controlo da FWER é a correcção de Sidák (também conhecida por correcção de Dunn-Sidák) [1]. Este procedimento assume que os testes individuais são independentes. Assim, definindo β o nível de significância de cada teste, a probabilidade de rejeitar pelo menos uma hipótese nula é dada por $1 - Pr(R = 0) = 1 - (1 - \beta)^m$, que se pretende igual ao nível de significância α; logo, $\beta = 1 - (1 - \alpha)^{1/m}$. A correcção de Bonferroni e a de Sidák estão ligada pela desigualdade [7]:

$$\alpha = 1 - (1 - \beta)^m \geq \frac{\alpha}{m}.$$

Assim, no caso dos testes serem dependentes a correcção de Sidák sobreestima o nível de significância β de cada teste. Se as estatísticas são independentes então a correcção de Sidák obtém uma estimativa menos pessimista que a fornecida pela correcção de Bonferroni sendo, deste modo, um procedimento não tão conservativo.

O principal problema com este tipo de procedimentos mais clássicos, que em muitas aplicações chega a ser inadequada, é a tendência a uma fraca sensibilidade. Muitas vezes, a falta de controlo da multiplicidade é tal que a protecção total resultante do controlo do FWER torna-se demasiado rigorosa. Considere-se a correcção de Bonferroni em que estão duas hipóteses em teste a um nível de significância de 0.05. Suponha-se que as estatísticas de teste são dependentes e que o primeiro teste não rejeita a hipótese nula (não diferencialidade dos genes). Definindo p_1 o valor p relativo ao primeiro teste então tem-se $p_1 > 0.05$. Como as estatísticas de teste são dependentes então, considerando p_2 o valor p relativo ao segundo teste ter-se-á $p_1 < p_2$ levando necessáriamente à não rejeição da hipótese nula relativa ao segundo teste. Voltando aos m testes, se as estatísticas para estes m testes forem dependentes, então a não rejeição de uma hipótese nula levará a uma maior probabilidade de não rejeição das restantes. Ou seja, a potência do teste (capacidade de detecção de genes diferencialmente expressos) é diminuida, tornando o teste mais conservativo [8].

Uma outra medida para definir a taxa de erro de tipo I em testes de hipóteses simultâneos é denotada por FDR (da abreviatura do inglês *False Discovery Rate*) e designada por Taxa de Falsas Descobertas. A FDR não é mais que a proporção esperada de hipóteses nulas verdadeiras rejeitadas no número total de rejeições. Vários estudos comparativos da FDR e FWER permitem concluir que, em geral, o primeiro sendo menos conservativo, torna-se mais potente no que respeita ao controlo de erros de tipo I [6, 5].

O conceito de FDR foi introduzido por Yoav Benjamini e Daniel Hochberg em 1995 [30]. Com a definição desta medida foi possível passar a ter em conta o valor esperado do

número de erros de tipo I e não apenas a probabilidade de cometer pelo menos um erro de tipo I. Além disso, aqueles autores provaram que o controlo da FDR implica o controlo da FWER e que a FDR admite procedimentos menos conservativos obtendo uma maior potência do teste. Formalmente, a FDR é definida como

$$FDR = E\left[\frac{V}{R \vee 1}\right] = E\left[\frac{V}{R} \mid R > 0\right] Pr\left(R > 0\right)$$

Saliente-se que o denominador $R \vee 1$ é assim construído apenas para que a fracção $\frac{V}{R}$ seja nula quando $R = 0$.

O procedimento de controlo da FDR construído por Yoav Benjamini e Daniel Hochberg tem por base a seguinte observação:

Considerando P $= (p_1, p_2, \ldots, p_m)$ o vector dos valores p dos m testes e $p_{(1)}, p_{(2)}, \ldots, p_{(m)}$ a respectiva sequência dos valores p ordenados de forma crescente, se se tomar

$$\hat{k} = \underset{1 \leq k \leq m}{\arg\max} \left\{ k : p_{(k)} \leq \frac{\alpha k}{m} \right\} \tag{2.3}$$

então, estabelecendo a rejeição de H_0 em termos da rejeição das hipóteses nulas $H_{0,i}$ $i = 1, 2, ..., \hat{k}$ significa que, nesta concretização, $R = \hat{k}$ e $V = m_0 \times \frac{\alpha \hat{k}}{m}$.

Esta observação levou os autores a demonstrar analiticamente que $FDR = E\left[\frac{V}{R}\right] = \frac{m_0}{m}\alpha \leq \alpha$, ficando garantido o controlo da medida de erro estabelecida. Importa salientar que $\hat{k} = 0$ quando nenhum p_i satisfaz a desigualdade $p_i \leq \frac{\alpha i}{m}$, e nenhuma das m hipóteses de teste é tida como significativa.

Relativamente ao método de Benjamini e Hochberg, Storey [32] coloca as seguintes questões:

- Até que ponto a expressão (2.3) fornece uma predição confiável para o valor observável de R numa dada concretização do teste simultâneo?
- Existe alguma forma de encontrar uma medida de erro para a variável aleatória R?

Storey observa que a garantia de obter um limite superior para a FDR é uma falsa sensação de segurança, quando na verdade o processo envolve estimação. Uma maior instabilidade do cálculo de $\hat{k}$ traduzirá num pior funcionamento do procedimento em termos práticos. Por outro lado, o valor esperado de $\frac{V}{R}$ é tal que $FDR \leq \alpha$, no entanto, não se sabe até que ponto o método é confiável caso a caso. Observa também que, usalmente, a potência de um teste de hipóteses simultâneo decresce à medida que o número de hipóteses a testar aumenta, mas se as hipóteses são independentes, então a potência não tem necessariamente que decrescer. Afirma ainda que, para um grande número de testes, maior será a informação contida nos valores p observados relativamente a m_0, devendo esta informação ser usada. Storey propõe assim um outro método usando essa informação, um método menos rigoroso que o de Benjamini e Hochberg, com uma maior potência e mantendo um forte controlo do

erro. Este procedimento faz uso do conceito de taxa de falsas descobertas positiva (pFDR[9]) definido por Storey [31], da seguinte forma:

$$pFDR = E\left[\frac{V}{R}\middle| R > 0\right]$$

De facto, para m grande muito dificilmente não existirão rejeições, pelo que fará sentido medir um erro de tipo I de um teste de hipóteses simultâneo pela proporção esperada de hipóteses erradamente rejeitadas no total das hipóteses rejeitadas, dado que existem rejeições das hipóteses $H_{0,i}$.

Em [32] é desenvolvida uma interpretação bayesiana para a FDR. Considerando m testes de hipóteses, baseados nas estatísticas de teste $T_1, T_2, ..., T_m$ para uma região de significância Γ, a pFDR é definida como

$$pFDR\left(\Gamma\right) = E\left[\frac{V\left(\Gamma\right)}{R\left(\Gamma\right)}\middle| R\left(\Gamma\right) > 0\right],$$

onde $V\left(\Gamma\right) = \#\left\{erros\ de\ tipo\ I\right\}$ e $R\left(\Gamma\right) = \#\left\{T_i : T_i \in \Gamma\right\}$. A cada hipótese nula $H_{0,i}$, Storey define uma variável aleatória indicatriz:

$$H_i = \begin{cases} 1, & se\ a\ hipótese\ H_{0,i}\ é\ falsa \\ 0, & se\ a\ hipótese\ H_{0,i}\ é\ verdadeira \end{cases}, \tag{2.4}$$

com $\pi_0 = \Pr\left(H_i = 0\right)$ e $\pi_1 = \Pr\left(H_i = 1\right)$ e H_i variáveis aleatorias *i.i.d* com distribuição *a priori* de Bernoulli de parâmetro π_1.

Storey prova que, assumindo (T_i, H_i) variáveis aleatória *i.i.d*, $T_i|H_i \sim \left(1 - H_i\right) F_0 + H_i F_1$, para alguma distribuição nula F_0 e uma distribuição alternativa F_1, e $H_i \sim Bernoulli\left(\pi_1\right)$, $i = 1, ..., m$, se tem

$$pFDR(\Gamma) = \Pr\left(H_i = 0 | T_i \in \Gamma\right), \quad m \geq 1\ e\ i = 1, 2, ..., m.$$

A título exemplificativo verifique-se a igualdade para $m = 2$ (para $m > 2$ o resultado segue da mesma forma tendo em conta a independência e distribuição idêntica das estatísticas de teste e das hipóteses H_i). Como,

$$E\left[\frac{V(\Gamma)}{R(\Gamma)}|R(\Gamma) > 0\right] = \int x\ dF_{\frac{V}{R}|R>0}(x) \tag{2.5}$$

[9]Em inglês conhecido como *positive false discovery rate*.

Assim, tomando m=2, vem

$$\begin{aligned} F_{\frac{V}{R}|R>0}(x) &= Pr\left(\frac{V}{R} \leq x | R>0\right) \\ &= \frac{Pr\left(\frac{V}{R} \leq x, R>0\right)}{Pr\left(R>0\right)} \\ &= \sum_{k=1}^{2} \frac{Pr\left(\frac{V}{R} \leq x, R=k\right)}{Pr(R>0)} \\ &= \sum_{k=1}^{2} \frac{Pr\left(\frac{V}{R} \leq x, R=k\right)}{Pr(R=k)} \frac{Pr(R=k)}{Pr(R>0)} \\ &= \sum_{k=1}^{2} Pr\left(\frac{V}{R} \leq x | R=k\right) Pr\left(R=k|R>0\right). \end{aligned}$$

Consequentemente,

$$\begin{aligned} E\left[\frac{V(\Gamma)}{R(\Gamma)}|R(\Gamma)>0\right] &= \int x \; d \sum_{k=1}^{2} Pr\left(\frac{V}{R} \leq x | R=k\right) Pr\left(R=k|R>0\right) \\ &= \sum_{k=1}^{2} \int x \; dF_{\frac{V}{R}|R=k}(x) Pr\left(R=k|R>0\right) \\ &= \sum_{k=1}^{2} Pr\left(R=k|R>0\right) \int x \; dF_{\frac{V}{R}|R=k}(x) \\ &= \sum_{k=1}^{2} Pr\left(R=k|R>0\right) E\left(\frac{V}{R}|R=k\right) \\ &= \sum_{k=1}^{2} Pr\left(R=k|R>0\right) \frac{E\left(V|R=k\right)}{k} \end{aligned}$$

Tomando em conta que

$$\begin{aligned} E\left(V|R=k\right) &= E\left(\sum_{i=1}^{2} I\left(T_i \in \Gamma \cap H_i = 0\right)|R(\Gamma)=k\right) \\ &= \sum_{i=1}^{2} Pr\left(T_i \in \Gamma \cap H_i = 0 | R(\Gamma)=k\right) \end{aligned}$$

Onde I(.) represenda a variável indicatriz, tomando o valor 1 se $T_i \in \Gamma$ e $H_i = 0$ e 0 caso contrário. Admitindo, sem perda de generalidade, que as primeiras k estatísticas de teste

pertencem à região crítica, vem:

$$\begin{aligned}\sum_{i=1}^{2} Pr\left(T_i \in \Gamma \cap H_i = 0 | R(\Gamma) = k\right) &= \sum_{i=1}^{k} Pr\left(H_i = 0 | T_1 \in \Gamma, T_2 \in \Gamma, ..., T_k \in \Gamma\right) \\ &= \sum_{i=1}^{k} Pr\left(H_i = 0 | T_i \in \Gamma\right) \\ &= kPr\left(H_i = 0 | T_i \in \Gamma\right) \qquad (2.6)\end{aligned}$$

Assim,

$$\begin{aligned}\sum_{k=1}^{2} Pr\left(R = k | R > 0\right) \frac{E\left(V | R = k\right)}{k} &= \sum_{k=1}^{2} \frac{1}{k} kP\left(H_i = 0 | T_i \in \Gamma\right) Pr\left(R = k | R > 0\right) \\ &= Pr\left(H_i = 0 | T_i \in \Gamma\right) \sum_{k=1}^{2} Pr\left(R = k | R > 0\right) \\ &= Pr\left(H_i = 0 | T_i \in \Gamma\right).\end{aligned}$$

Observe-se que $\Pr\left(H_i = 0 | T_i \in \Gamma\right)$ são idênticos para todos os $i = 1, ..., m$. É importante notar que esta reformulação torna a FDR independente de m, ou seja, a potência do teste de hipóteses múltiplo não depende do número de testes efectuados. Assim, pelo teorema de Bayes, e dada a condição de *i.i.d*, vem

$$\begin{aligned}pFDR &= \Pr\left(H_i = 0 | T_i \in \Gamma\right) \\ &= \frac{\Pr(H_i = 0)\Pr\left(T_i \in \Gamma | H_i = 0\right)}{\Pr\left(T_i \in \Gamma\right)} \\ &= \frac{\pi_0 \Pr\left(T_i \in \Gamma | H_i = 0\right)}{\pi_0 \Pr\left(T_i \in \Gamma | H_i = 0\right) + \pi_1 \Pr\left(T_i \in \Gamma | H_i = 1\right)} \\ &= \frac{\pi_0 \left\{Erros\ de\ tipo\ I\ em\ \Gamma\right\}}{\pi_0 \left\{Erros\ de\ tipo\ I\ em\ \Gamma\right\} + \pi_1 \left\{potência\ de\ \Gamma\right\}} \\ &= 1 - \left(1 + \frac{\pi_0 \left\{Erros\ de\ tipo\ I\ em\ \Gamma\right\}}{\pi_1 \left\{potência\ de\ \Gamma\right\}}\right)^{-1}. \qquad (2.7)\end{aligned}$$

o que mostra que a pFDR cresce com o aumento do erro de tipo I.

Note-se que o domínio das variáveis aleatórias valor p assume a forma $[0, \gamma]$, para algum $\gamma \geq 0$, onde γ representará o maior valor p (p_i) que leva à rejeição da hipótese nula em causa. Para entender o porquê desta afirmação, basta atender à definição de valor p. Fixado o nível de significância α, o objectivo de um teste será encontrar o menor nível de significância possível $\hat{\alpha}$ de tal forma que os dados observados conduzam à rejeição da hipótese nula. Por outras palavras, pretende-se

$$\hat{\alpha} = \inf_{\Gamma : T_i \in \Gamma} \left\{\Pr\left(T_i \in \Gamma | H_i = 0\right)\right\}. \qquad (2.8)$$

Este será o nível de significância atingido associado ao valor da estatística de teste, o que não é mais que o valor p associado ao teste. A Figura 2.4 permite um melhor entendimento do conceito anterior.

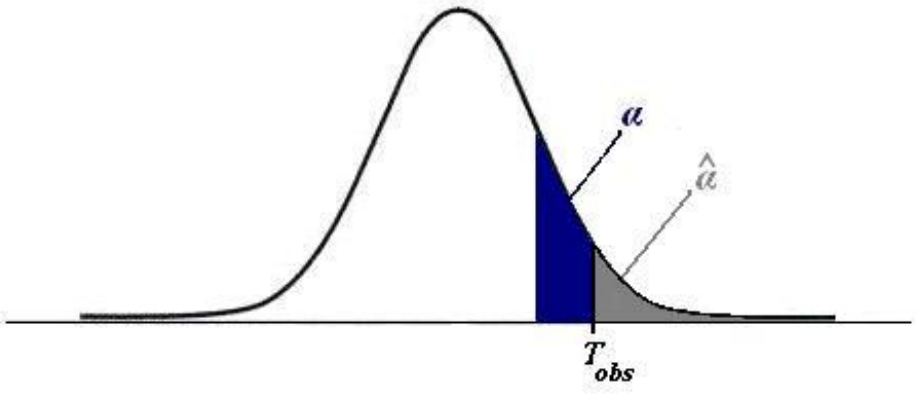

Figura 2.4: Valor p associado a um teste. O nível de significância definido pelo investigador é representado pela área a azul. Calculado o valor observado da estatística de teste sob a hipótese nula, o $\hat{\alpha}$ (valor p) é representado pela área mínima que evidencia a rejeição da hipótese nula (área a cinza).

Assim, dado dois valores p, p_1 e p_2, tal que $p_1 \leq p_2$, para as respectivas estatísticas observadas t_1 e t_2, se $t_2 \in \Gamma$ então também $t_1 \in \Gamma$, levando à rejeição de ambas as hipóteses nulas. Tal demonstra que os valores p que levam à rejeição das hipóteses nulas correspondentes pertencem a um intervalo da forma $[0, \gamma]$.

Tomando um valor fixo λ, π_0 pode ser estimado em termos de λ por:

$$\hat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda\}}{(1-\lambda)\, m}.$$

De facto, $\pi_0 m$ constitui o número esperado de hipóteses nulas verdadeiras e $\pi_0 m(1-\lambda)$ o número mínimo de hipóteses nulas verdadeiras que não são rejeitadas dado o nível de significância λ, que, para um λ bem escolhido, será uma aproximação da quantidade $\{\#p_i > \lambda\}$. Ou seja, $\pi_0 m(1-\lambda)$ representará uma aproximação do número de valores p maiores que o nível de significância λ e que não levam portanto à rejeição da hipótese nula. Em [32] é estabelecido um procedimento para encontrar o λ óptimo.

Uma vez estimado o λ óptimo, Storey propõe usá-lo para estimar γ tomando

$$\hat{\Pr}(p_i \leq \gamma) = \frac{\#\{p_i \leq \lambda\}}{m}$$

e estabelece $1-(1-\gamma)^m$ como um limite inferior para a probabilidade de existir pelo menos uma rejeição. Assim, de (2.7), em termos de valores p a pFDR pode ser reescrita como

$$pFDR\ (\gamma) = \frac{\pi_0 \Pr(T_i \in \Gamma_i | H_i = 0)}{\Pr(T_i \in \Gamma)} = \frac{\pi_0 \Pr(p_i \leq \gamma | H_i = 0)}{\Pr(p_i \leq \gamma)} = \frac{\pi_0 \gamma}{\Pr(p_i \leq \gamma)}.$$

Dado que a pFDR está condicionada à existência de pelo menos uma rejeição, uma estimativa pode ser dada por

$$\widehat{pFDR} = \frac{\widehat{\pi}_0(\lambda)\gamma}{\widehat{Pr}(p_i \leq \gamma)\{1-(1-\gamma)^m\}}.$$

Já a FDR, não estando condicionada à existência de rejeição, pode ser estimada por

$$\widehat{FDR} = \frac{\widehat{\pi}_0(\lambda)\gamma}{\widehat{Pr}(p_i \leq \gamma)}.$$

Storey prova ainda que para m grande estas estimativas de pFDR e de FDR são equivalentes, ou seja, em termos assimptóticos estas medidas são semelhantes para uma determinada região de rejeição.

A pFDR pode ser usada para definir o valor q, uma medida semelhante ao valor p, que fornece uma medida de erro para cada uma das estatísticas de teste, no que respeita ao pFDR. Para uma estatística observada $T = t$, Storey define

$$valor\ q\,(t) = \inf_{\Gamma: t\in\Gamma} \{pFDR\,(\Gamma)\} = \inf_{\Gamma: t\in\Gamma} \{Pr\,(H_i = 0 | T_i \in \Gamma)\}\,.$$

Assim, o valor q representa uma medida de força da estatística observada, no que respeita à pFDR, ou seja, é dado pelo menor valor de pFDR que pode ocorrer quando se rejeita uma estatística observada t para o conjunto das regiões de rejeição.

O exemplo seguinte, ilustrado na Figura 2.5 permite entender melhor o conceito [31].

Exemplo 2.3.1. *Suponha-se que se pretende efectuar m testes $H_{0,i} : \theta = 0$ vs $H_{0,1} : \theta = 2$ para m variáveis aleatórias com distribuição $N(\theta, 1)$, e estatísticas de teste $T_1, ..., T_m$. Especificamente (T_i, H_i) são i.i.d com $T_i|H_i \sim (1 - H_i)\,N(0,1) + H_i N(2,1)$.*

Assim, o valor $p(t_i) = \Pr(T \geq t_i | H_i = 0) = \Pr(N(0,1) \geq t_i)$ constitui toda a área a baixo da curva que está à direita do valor observado da estatística. O valor $q(t) = \inf_{\Gamma: t\in\Gamma} \{\Pr(H_i = 0 | T_i \in \Gamma)\}$ será calculado usando as áreas à direita da estatística observada e abaixo de ambas as curvas, $N(0,1)$ e $N(2,1)$, ou seja, usando as probabilidades da hipótese ser rejeitada, sendo ela a hipótese nula ou a hipótese alternativa.

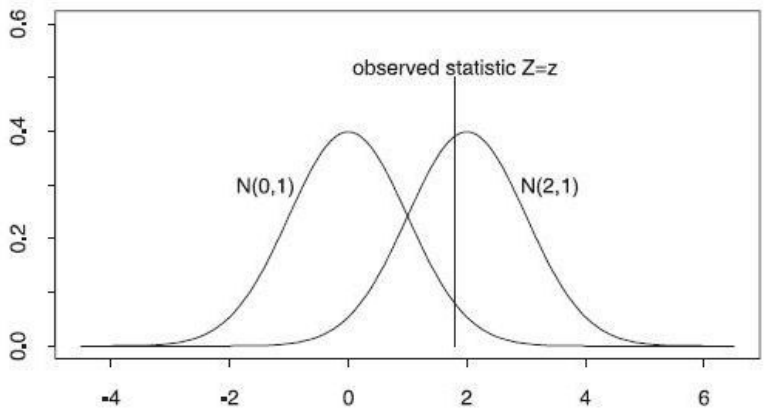

Figura 2.5: Gráfico das densidades $N(0,1)$ e $N(2,1)$. A linha vertical denota a estatística observada. O valor p é dado pela área debaixo de $N(0,1)$ à direita da estatística observada. O valor q é calculado usando as áreas debaixo de ambas as curvas à direita da estatística observada e ponderadas pelas probabilidades π_0 e π_1, respectivamente. Imagem retirada de [31].

Capítulo 3

Análise de Significância de *Microarrays*

Com vista à detecção de genes diferencialmente expressos efectuando o controlo da FDR, Tusher et al. (2001) [27] propôs um método de permutação, a SAM (Análise de Significância de Microarrays[1]), que posteriormente foi desenvolvida para um *software* livre por Storey e Tibshirani (2003) [4].

Com a introdução deste método foi possível ao utilizador decidir com que FDR trabalhar perante determinada base de dados e obter assim o número de genes diferencialmente expressos consoante a FDR estipulada.

O *software* é livre para pesquisas académicas sem fins comerciais, requerendo apenas o registo. O programa está implementado como uma extensão integrada do Microsoft Excel. Para uma correcta instalação, são pedidos alguns requisitos, tal como a última versão do *software* R. A instalação é então feita em simultâneo no Excel e no R, podendo o software SAM ser usado apenas recorrendo à biblioteca "`samr`"desenvolvida para o R. O software SAM pode ser obtido na página http://www-stat.stanford.edu/ tibs/SAM/Rdist/index.html.

A SAM permite a exploração de dados quando a variável resposta é:

Qualitativa, dada por duas classes (ex. dados vindos de indivíduos em tratamento e de indivíduos normais) ou multiclasses (ex. dados vindos de três ou mais espécies de cancro)

Quantitativa, por exemplo, na análise de tempos de sobrevivência censurados.

Para ambos os tipos de variável resposta, as amostras poderão ainda ser emparelhadas ou não emparelhadas.

Os valores omissos (*missing values*) que eventualmente existam nas bases de dados são substituídos pelo software SAM, sendo usado por defeito o método dos k-vizinhos mais próximos (*k-Nearest Neighbor*) com $k = 10$.

A determinação dos genes diferencialmente expressos é feita tendo em conta o parâmetro *delta* (Δ) introduzido pelo utilizador, em finção do qual se calcula os falsos positivos que

[1]Em inglês conhecida como *Significance Analysis of Microarrays*

representam os genes tidos como diferencialmente expressos mas que na verdade não o são. É possível ainda escolher uma *fold change* para garantir que os genes significantes nos dois grupos, em média, diferem em uma proporção superior à *fold change* estipulada pelo utilizador.

Mais concretamente, considerando amostras não emparelhadas, dada uma base de dados com valores x_{ij}, $i = 1, 2, ..., m$ genes, $j = 1, 2, ..., n$ amostras, e a variavel resposta y_j para cada amostra, $j = 1, 2, ..., n$, tomando 0 se a amostra pertence a um grupo (ex. tratamento) e 1 se pertence ao outro grupo (ex. controlo), a SAM processa-se através dos seguintes passos [29, 28]:

1. Calcular a estatística:

$$d_i = \frac{r_i}{s_i + s_0},\ para\ cada\ gene\ i = 1, 2, ..., m \tag{3.1}$$

onde,

$$r_i = \bar{x}_{i2} - \bar{x}_{i1} \quad (pontuacão)$$

$$s_i = \sqrt{\frac{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\left\{\sum\limits_{j\in C_1}(x_{ij}-\bar{x}_{i1})^2+\sum\limits_{j\in C_2}(x_{ij}-\bar{x}_{i2})^2\right\}}{n_1+n_2-2}}\ (desvio\ padrão\ corrigido\ agrupado)$$

Para valores baixos do nível de expressão, a variabilidade dos d_i pode ser alta, devido aos baixos valores de s_i. De modo a comparar as diferenças para todos os genes, a distribuição dos valores d_i deve ser independente dos níveis de expressão e respectivo desvio padrão corrigido. O factor de intermutabilidade s_0 tem como objectivo tornar o coeficiente de variação de d_i aproximadamente constante, em função de s_i. Por defeito, o calculo de s_0 é proposto ser realizado do seguinte modo, denotando por s^α o percentil de ordem α dos valores $s_1, s_2, ..., s_m$, e

$$d_i^\alpha = \frac{r_i}{s_i + s^\alpha}, i = 1, 2, ..., m$$

- Calcular os 100 percentis dos valores s_i: $q_1 < q_2 < ... < q_{100}$.
- Para $\alpha \in \{0, 0.05, 0.1, ..., 1.0\}$
  (a) Calcular $v_j = \underset{\alpha=0,0.05,...,1}{mad}\left(d_i^\alpha : s_i \in [q_j, q_{j+1}[\right), j = 1, 2, ..., n$, onde mad é o desvio absoluto mediano em relação à mediana a dividir por 0.64.
  (b) Calcular o coeficiente de variação dos $v_j's$, $cv(\alpha)$.
- Escolher $\hat{\alpha} = \underset{\alpha=0,0.05,...,1}{\arg\min}\ \{cv(\alpha)\}$ e tomar $\hat{s}_0 = s^{\hat{\alpha}}$.

2. Ordenar os valores observados da estatística $d_{(1)} < d_{(2)} < ... < d_{(m)}$.

3. Tomar B conjuntos de permutações dos valores resposta. Para cada uma das permutações repetir os passos 1. e 2. até obter os valores d_i^{*b} da estatística e os correspondentes valores ordenados $d_{(1)}^{*b} < d_{(2)}^{*b} < ... < d_{(m)}^{*b}$.

4. Estimar as estatísticas de ordem usando o conjunto das B permutações,

$$\bar{d}_{(i)} = \frac{1}{B}\sum\nolimits_b d^{*b}_{(i)},\ i = 1, 2, ..., m.$$

5. Representar o gráfico de pontos $(d_{(i)}, \bar{d}_{(i)})$ e a recta $d_{(i)} = \bar{d}_{(i)}$.

6. Para um Δ representando a margem de variação para a qual os genes não serão considerados diferencialmente expressos, definir $t_1 = \underset{i=1,...,m}{\arg\max}\left(d_{(i)} - \bar{d}_{(i)} < -\Delta\right)$ e $t_2 = \underset{i=1,...,m}{\arg\min}\left(d_{(i)} - \bar{d}_{(i)} > \Delta\right)$
   Dos passos 5. e 6. é obtido um gráfico similar ao da Figura 3.1, que confronta os quantis observados $d_{(i)}$ com os esperados sob a hipótese nula.

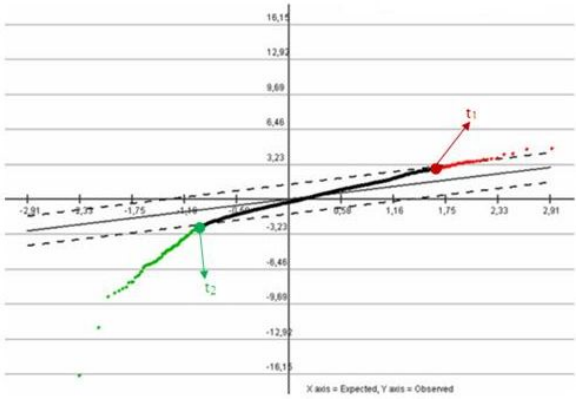

Figura 3.1: Imagem adaptada do output do SAM. No eixo dos yy estão representados os quantis observados $d_{(i)}$ e no eixo dos xx os quantis esperados sob a hipótese nula. A vermelho estão as observações com uma maior variação no sentido positivo que a definida pelo utilizador como não sendo significativa (ou seja, a intensidade R é significativamente superior à intensidade G), e a verde as observações com uma maior variação no sentido negativo (ou seja, a intensidade G é significativamente superior à intensidade R).

7. Estimar o número de rejeições sob a validação de todas as hipóteses nulas H_{0i} fazendo:

$$Falsos\ Positivos\,(\Delta) = \underset{b=1,...,B}{med}\left(\#\left\{i = 1, ..., m : d^{*b}_i\,(i) \geq t_1\ ou\ d^{*b}_i\,(i) \leq t_2\right\}\right)$$

   É de salientar que existem algumas variantes. Para além do cálculo da mediana dos genes classificados como significantes, para o conjunto das permutações, [29] faz referência ao uso do percentil 90 e [26, 27] à media dos genes diferencialmente expressos das permutações, sendo esta a utilizada na versão mais recente da SAM.

8. Estimar a probabilidade de H_0 ser verdadeira, π_0. A SAM usa a estimativa dada por Storey referida na Secção 2.3 com um $\lambda = 0.5$. A estimação é feita nos seguintes passos:

- Calcular o percentil 25 (q_{25}) e o percentil 75 (q_{75}) de todas as mB diferenças permutadas,
- Calcular $\hat{\pi}_0 = \#\{d_i \in (q_{25}, q_{75},)\}/0.5p$,
- Tomar $\hat{\pi}_0 = \min(\hat{\pi}_0, 1)$.

9. Estimar o número esperado de rejeições erradas por $\hat{\pi}_0 \times Falsos\ Positivos\,(\Delta)$.
10. Listar as hipóteses $H_{0,i}$ rejeitadas e quantificá-las (N).
11. Estimar a FDR por $\widehat{FDR}(\Delta) = \frac{\hat{\pi}_0 \times Falsos\ Positivos(\Delta)}{N}$.
12. ***Fold Change***. Se uma *fold change* t for especificada então, para que um gene seja significante ($H_{0,i}$ ser rejeitada) deverá ainda satisfazer $|\overline{x}_{i2}/\overline{x}_{i1}| \geq t$ ou $|\overline{x}_{i2}/\overline{x}_{i2}| \leq 1/t$.
13. **Valor q**. Se o utilizador pretender, também o valor q de um gene poderá ser indicado, o qual é estimado pela FDR correspondente ao menor $\hat{\Delta}$ para o qual esse gene é tido como significante. O valor q será a FDR correspondente a $\hat{\Delta}$, $valor\ q_i = \widehat{FDR_i}(\hat{\Delta}) = \widehat{FDR_i}(min(\Delta : |d_{(i)} - \overline{d}_{(i)}| \geq \Delta))$.
14. **A FDR local** para um gene é a FDR para genes que tenham níveis de expressão semelhantes. É estimada a FDR apenas para 0.5% dos genes mais próximos do gene em estudo. Se 1.0% do total de número de genes é menor que 50, então a percentagem é incrementada para que perfaça os 50 genes.

Os passos que se seguem, constituem apenas informação extra que o software SAM dará, caso o utilizador solicite essa informação.

Para respostas de tipo quantitativas, amostras emparelhadas, dados de sobrevivência censurados, multiclasse e apenas uma classe, as expressões para r_i e os respectivos desvios padrão s_i no passo 1. sofrem alterações no algoritmo. Informações detalhadas em [29].

Por defeito o SAM utiliza um teste t, no entanto, também o teste de Wilcoxon ou o de Mann-Whitney podem ser utilizados deste que definido pelo utilizador. Em [28] é referido que uma das desvantagens do SAM é a tendência para obter estimativas enviesadas do número de genes significantes, especialmente se o número for relativamente grande. Assim, para contornar este problema esse autor propõe uma alteração ao procedimento. A diferença entre o método proposto e o SAM tradicional reside na estatística escolhida. Em vez de usar a estatística (3.1), é usada uma estatística de teste linear *signed-rank*, baseada numa função *rank score*.

3.1 Análise Experimental

3.1.1 Base de dados `ApoAI`

Para proceder a uma análise de significância dos genes será necessário normalizar os dados antes de aplicar a SAM. O pacote `limma` do projecto **Bioconductor** (*http://bioconductor.org*)

permite realizar a normalização entre *microarrays* e dentro de cada *microarray* de dados de expressão genética. A normalização dos dados poderá ser feita através dos seguintes comandos:

```
library(limma)
load("ApoAI.RData")
MA <- normalizeWithinArrays(RG)
x<-MA$M
```

A variável resposta tem duas classes, uma relativa aos ratos "normais"(8 *microarrays*) e a outra relativa aos ratos com deficiência no gene ApoAI (8 *microarrays*), pelo que o vector da variável resposta pode ser obtido através do comando

```
y<-c(rep(1,8),rep(2,8))
```

onde 1 representa os ratos "normais"e 2 os ratos com deficiência no gene em estudo.

O pacote `samr` exige como dados de entrada uma lista com a base de dados, o vector das variáveis resposta e a identificação e nome de cada gene. Essa lista pode ser criada da seguinte forma:

```
data=list(x=x,y=y, geneid=MA$genes[,6], genenames=MA$genes[,5],
logged2=TRUE)
```

Estão criados os objectos no R necessários para aplicar a SAM à base de dados. Optou-se por fazer 1000 permutações. A base de dados, tal como descrita na Secção 1.3 é constituída duas amostras não emparelhadas (Ratos "normais"vs Ratos "com deficiência"). A aplicação da SAM à base de dados resulta do seguinte comando:

```
samr.obj<-samr(data, resp.type="Two class unpaired", nperms=1000)
```

Em [25] verificou-se que para um $\Delta = 0.61$ a SAM obtinha um bom desempenho (ou seja, para o número de genes significantes obtidos a FDR correspondente era muito baixa), pelo que se optou por esse mesmo Δ.

O gráfico obtido com a execução dos comandos

```
delta=.61
samr.plot(samr.obj,delta) [2]
```

permite obter o gráfico de quantis da Figura 3.2, onde são confrontados os valores das estatísticas observadas com os valores teóricos das estatísticas sob a hipotese nula (não diferencialidade). Deste modo será possível obter uma ideia dos valores que não se ajus-

[2]Por *defeito* o SAM considera que a *foldchange* é nula, ou seja, nunhum critério é aplicado.

tam, ou seja, dos genes que à partida serão diferencialmente expressos.

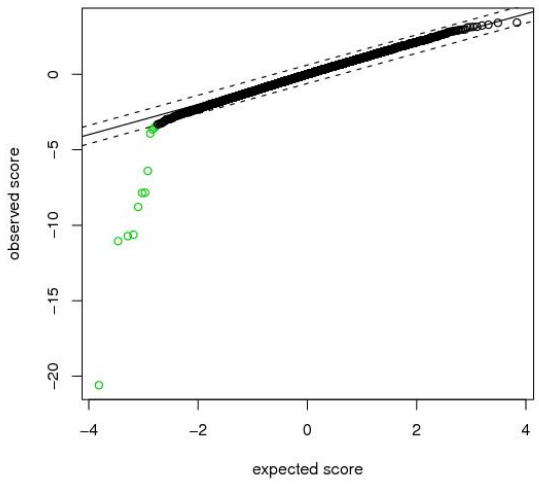

Figura 3.2: Gráfico $d_{(i)}$ *vs* $\bar{d}_{(i)}$ e identificação a verde dos genes significantes, quando se considera $\Delta = 0.61$.

O passo seguinte será obter uma estimativa da FDR e os $t_1 = \arg\max\left(d_{(i)} - \bar{d}_{(i)} < -\Delta\right)$ e $t_2 = \arg\min\left(d_{(i)} - \bar{d}_{(i)} > \Delta\right)$ para uma série de valores Δ. Com base nestas estimativas e na FDR que pretenda assumir o utilizador poderá escolher o Δ e obter o número de genes diferencialmente expressos associados a esse Δ. Essa informação pode ser obtida com o comando

```
delta.table <- samr.compute.delta.table(samr.obj);
```

Mais uma vez por defeito nenhum critério para a *fold change* é determinado. Também por omissão, o SAM considera 50 valores para o Δ, começando em $\Delta = 0$ e terminando no Δ que obtém todos os genes como não diferencialmente expressos.

A tabela dos genes significantes é obtida do seguinte modo:

```
siggenes.table<-samr.compute.siggenes.table(samr.obj,delta, data,
delta.table)
```

Do comando anterior não foram obtidos genes significantes positivos, ou seja, não foram encontrados genes acima de t_1. De facto, para os genes considerados significantes as estatísticas de teste tomaram sempre valores negativos. A informação relativa a estes genes encontra-se resumida na Tabela 3.1.

Em [24] usando o método das permutações para o cálculo dos valores p ajustados, são sugeridos 8 genes diferencialmente expressos. Todos os 8 genes estão incluidos na Tabela 3.1 obtida da SAM.

Note-se que uma das condições para a aplicação da estatística de teste t é a distribuição (aproximadamente) normal dos dados para cada um dos genes. A verificação desta condição

Linha	Nome	Valor d
2149	ApoAI,lipid-Img	-20,59
540	EST,HighlysimilartoA	-11,05
5356	CATECHOLO-METHYLTRAN	-10,63
4139	EST,WeaklysimilartoC	-10,72
1739	ApoCIII,lipid-Img	-8,79
2537	ESTs,Highlysimilarto	-7,85
1496	est	-7,87
4941	similartoyeaststerol	-6,40
947	EST,WeaklysimilartoF	-3,92
954	Caspase7,heart-Img	-3,65
5604		-3,56
4140	APXL2,5q-Img	-3,44

Tabela 3.1: Lista dos genes significantes obtidos da aplicação do SAM à base de dados `ApoAI`.

seria um processo moroso. Assim, perante esta situação e dado que apenas se tem oito observações para cada classe, optou-se por verificar os resultados obtidos com a estatística de teste de Wilcoxon (Mann-Whitney). Para a obtenção destes resultados aplicou-se o mesmo procedimento, indicando no comando `samr` o uso desta estatística. Foram obtidos 10 genes diferencialmente expressos que contemplam os genes sugeridos em [24]. Três destes genes foram detectados com o uso da estatística t ("Caspase7,heart-Img", "APXL2,5q-Img"e o gene da linha 5604) e não detectado o gene "NCAM-120,Brain-Img", linha 5343.

3.1.2 Base de dados `Fermentation`

O gráfico da Figura 3.3 representa as curvas de crescimentos das 7 leveduras (FF02, FF20, FRouge (FR), ICV, EC1118, J940047 (J047) e S288c) que constituem a base de dados `Fermentation`, já descrita na secção 1.3. Tendo em conta o gráfico, os tempos com maior interesse a estudar serão o Tempo 2 (T2), Tempo 3 (T3) e Tempo 5 (T5) já que representam fases de metabolismo distintas, sendo o T2 caracteristico da fase de multiplicação exponencial das células, T3 da transição entre exponencial e estacionária e T5 da fase estacionária (as células não se dividem). O Tempo 6 seria interessante estudar, se já todas as leveduras estivessem num estado de equilíbrio, o que de facto não se verifica, algumas continuam em crescimento. Neste sentido, dado que não é visível o mesmo comportamento para todas as leveduras, optou-se por não incluir este tempo na análise.

Como mencionado na secção 1.3, pretende-se determinar o número de genes diferencialmente expressos comparando leveduras vínicas e não vínicas, comparando as leveduras vínicas com a levedura clínica e comparando as leveduras vínicas com a levedura laboratorial.

Importa ainda referir que nos dados usados em estudo as hibridações *dye swap* já estão

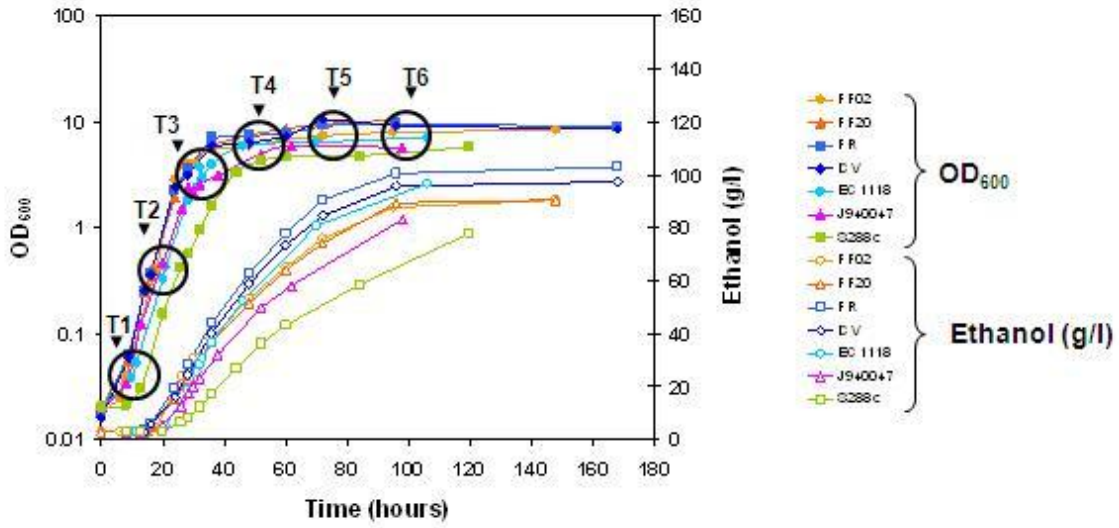

Figura 3.3: Gráfico representando as curvas de crescimento das 7 leveduras. Para a construção do gráfico, foram obtidas medidas da densidade das células (originando as 7 curvas mais acima, cujas medições estão representadas pelos pontos a cheio) e as respectivas concentrações de Etanol (originando as 7 curvas mais abaixo, cujas medições estão representadas pelos pontos "não a cheio". Para a construção da base de dados foram apenas consideradas 6 medidas da densidade das células (ou seja, medições em 6 tempos diferentes), para cada uma das 7 leveduras, e que estão representados no gráfico por T1, T2, T3, T4, T5 e T6. Gráfico obtido no Departamento de Biologia da Universidade de Aveiro.

---

invertidas, podendo deste modo serem comparadas directamente com os outros microarrays. De facto, quando se obtém os quocientes 2.1, é o mesmo que ter o quociente

$$\frac{Amostra\ A\ (rotulada\ com\ Cy5)}{Amostra\ B\ (rotulada\ com\ Cy3)}.$$

Quando se invertem os rótulos, deixa-se de ter o quociente 2.1, mas sim o seu inverso,

$$\frac{G}{R}$$

pelo que a inversão desta quantidade fará com que seja possível compará-la directamente com o quociente 2.1, facilitando a análise.

Os dados já se encontram pré-processados. Para a correcção de *background* foram eliminados os *spots* com má qualidade e calculada a média das réplicas das sondas em cada *microarray*. Para a normalização dos dados foi usada a normalização global, referida na secção 2.2.2 usando sondas para controlo como referência. Neste caso o uso de sondas para controlo assume um papel fundamental. Para todas as sondas foi usada uma referência comum vinda de células de crescimento exponencial que foram comparadas com alguns pontos da curva de crescimento em fase estacionária, deste modo, o uso de métodos baseados no comportamento global dos genes não seria aconselhável.

A primeira questão a colocar quando se pretende fazer uma SAM a uma base de dados da qual nada se sabe será qual o delta escolher. Interessará escolher um delta de forma a que se consiga obter o maior número de genes significantes com a menor FDR possível.

Pretendendo verificar quais os genes que são diferencialmente expressos quando comparando as leveduras vínicas com as não vínicas, com a clínica ou com a laboratorial para os tempos 2, 3 e 5, e na tentativa de averiguar o comportamento das bases de dados associadas às três análises em cada um dos tempo face à FDR, efectuou-se um estudo exploratório com uma análise gráfica da variável FDR estudada com o número de genes detectados como diferencialmente expressos, construindo-se o gráfico da Figura 3.4. Este gráfico possibilita a verificação da taxa de falsos positivos (taxa de genes considerados diferencialmente expressos quando na verdade não o são) que é obtida quando um determinado número de genes é detectado como diferencialmente expressos. Assim será possível obter uma ideia do número de genes não significantes que será obtido mediante a FDR que o investigador pretenda assumir e ajudá-lo numa escolha razoável da FDR a considerar. Para a construção deste gráfico, obteve-se o número de genes significantes e a respectiva FDR para uma tabela de valores de delta. Estes valores podem ser obtidos usando o comando `samr.compute.delta.table`.

Os gráficos da Figura 3.4 permitem concluir que em todos os tempos são detectados cerca de 200 genes diferencialmente expressos independentemente do confronto em causa (vínicas com a laboratorial, vínicas com as não vínicas ou as vínicas com a clínica), e com uma FDR muito baixa. Deste modo, existe uma grande "confiança"de que cerca de 200 genes têm um nível de expressão que difere, em média, quando são consideradas leveduras vínicas e não vínicas (laboratorial, clínica ou ambas). A partir destes 200 genes detectados como diferencialmente expressos existe uma grande distinção entre o comportamento das curvas. Para o gráfico do tempo 2, o facto da curva a azul crescer rapidamente após cerca de 200 genes no eixo dos xx indicará que as diferenças tidas como significativas entre a média de expressão genética das leveduras vínicas e a média de expressão genética da levedura laboratorial, podem não ser de facto diferenças significativas (a FDR começa a ser muito elevada, representando uma grande percentagem de detecção de falsos positivos, isto é, diferenças que na verdade não são). Já as curvas verde e vermelha apresentam um comportamento semelhante no que respeita ao número de genes detectados. Supondo que o limite máximo para a FDR aceite pelo biólogo é de 0.2, e considerando o comportamento das curvas apenas até à recta $Y = 0.2$, existem em média muitos genes com níveis de expressão diferentes entre as vínicas e as não vínicas e entre as vínicas e a clínica, e muitos genes com níveis de expressão semelhantes entre as vínicas e a laboratorial. Assim, aparentemente, a forma da curva verde deve-se às diferenças entre as vínicas e a clínica, e os níveis de expressão das levedura vínicas serão mais semelhantes com os da levedura laboratorial que com os da clínica.

No segundo gráfico as curvas têm um comportamento semelhante ao primeiro gráfico, no entanto o seu crescimento não é tão acentuado. As diferenças entre as leveduras vínicas e a laboratorial, para o tempo 3 aparentam ser maiores. Também a curva a vermelho parece evidenciar maiores diferenças quanto à expressão genética das leveduras vínicas e da clínica conduzindo a uma maior detecção de genes diferencialmente expressos com uma

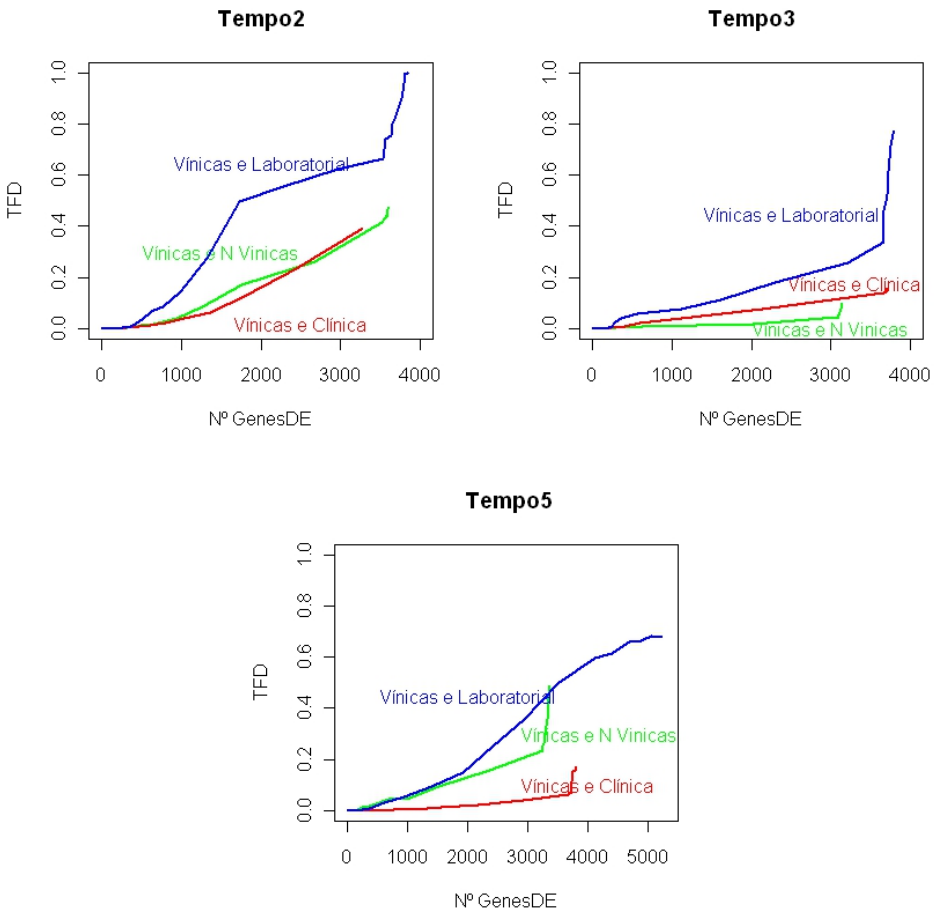

Figura 3.4: Análise gráfica da FDR para a base de dados `Fermentation`. As curvas representadas mostram o comportamento da base de dados face à FDR quando se colocam em confronto leveduras vínicas e não vínicas (curva a verde), leveduras vínicas e clínica (curva a vermelho) e leveduras vínicas e laboratorial (curva a azul), para os tempos 2, 3 e 5.

FDR bastante baixa, quando confrontadas as leveduras vínicas com as não vínicas.

O terceiro gráfico continua a evidenciar uma semelhança entre as leveduras vínicas e a laboratorial, mais acentuada que no tempo 3 e menos acentuada que no tempo 2. É neste gráfico onde as diferenças relativamente às médias dos níveis de expressão das leveduras vínicas e clínica aparentam ser maiores, de facto, serca de 3000 genes são detectados como diferencialmente expressos a uma FDR quase nula. Deste modo, o crescimento da curva verde, aparentemente, deve-se apenas ao rápido crescimento da curva a azul.

As Tabelas 3.2, 3.3 e 3.4, constituem parte das tabelas obtidas com a aplicação da SAM, onde constam os valores dos deltas, números de genes diferencialmente expressos obtidos e FDR.

Tendo em conta as Tabelas, 3.2, 3.3 e 3.4, o mais sensato será optar por um delta específico em cada um dos 9 casos (vínicas e não vínicas (tempo 1, tempo 2 e tempo 3),

Vínicas e Não Vínicas								
Tempo2			**Tempo3**			**Tempo5**		
delta	*genes DE*	*FDR*	*delta*	*genes DE*	*FDR*	*delta*	*genes DE*	*FDR*
1,001	494	0,0110	0,709	622	0,007	0,9070	248	0,0120
1,115	394	0,0070	0,785	395	0,004	1,0000	205	0,0090
1,236	310	0,0040	0,866	284	0,002	1,0970	174	0,0060
1,362	253	0,0000	0,950	214	0,000	1,1990	149	0,0000
1,495	218	0,0000	1,038	181	0,000	1,3060	130	0,0000
1,634	183	0,0000	1,130	156	0,000	1,4170	100	0,0000

Tabela 3.2: Excerto da tabela de deltas, genes diferenecialmente expressos e FDR obtida com a SAM confrontando as leveduras vínicas com as não vínicas. A tabela representa os resultados obtidos quando comparadas as leveduras vínicas com as não vínicas nos tempos 2, 3 e 5. Para cada tempo retirou-se um excerto da tabela fornecida com a SAM com os deltas e os correspondentes números de genes diferencialmente expressos detectados e FDR.

vínicas e clínica (tempo 1, tempo 2 e tempo 3) e vínicas e laboratorial (tempo 1, tempo 2 e tempo 3)). Pretendendo determinar o maior número de genes com a menor FDR possível, as escolhas da Tabela 3.5, parecem ser rasoáveis.

Para os deltas da Tabela 3.5, obtiveram-se os gráficos da Figura 3.5. O gráfico da Figura 3.5 permite verificar que de uma forma geral, para os deltas considerados, o número de genes diferencialmente expressos não varia muito para os três tipos de confrontos. Aparentemente a maior diferença é verificada quando confrontadas as leveduras vínicas e a levedura laboratorial, parecendo existir um aumento dos genes diferencialmente expressos negativos relativamente aos outros confrontos. No entanto, no total, o número de genes não parece variar muito. No tempo 5 quando feito o confronto entre as leveduras vínicas e a clínica também aparenta existir um aumento relativo do número de genes tidos como diferencialmente expressos. As Tabelas 3.6, 3.7 e 3.8 permitem retirar conclusões mais precisas acerca da dimensão destes genes (positivos e negativos).

Tal como na base de dados `ApoAI`, também na `Fermentation`, o número de observações para cada classe é pequeno e a validação do pressuposto da normalidade dos dados torna-se um processo demasiado moroso, face à quantidade de genes em estudo. Assim, optou-se também por efectuar a SAM usando a estatística de teste de wilcoxon e confrontar os resultados com os obtidos com a estatística de teste t. O procedimento aplicado é o mesmo que com a estatística de teste t, com excepção do comando `samr`, onde é específicado o uso do da estatística de teste de wilcoxon. As Tabelas 3.9,3.10 e 3.11 resumem o número de genes obtidos com o uso desta estatística de teste.

Como se terá oportunidade de verificar com as análises dos modelos abordados nas próximas secções desta dissertação, à excepção do estudo onde foram confrontadas as leveduras vínicas e as não vínicas, o número de genes detectados com o uso da estatística de teste de wilcoxon é muito elevado quando comparado com o número de genes diferencialmente expressos obtido com a aplicação das outras metodologias. Ao biólogo não

Vínicas e Clínica								
Tempo2			**Tempo3**			**Tempo5**		
delta	*genes DE*	*FDR*	*delta*	*genes DE*	*FDR*	*delta*	*genes DE*	*FDR*
0,720	464	0,0060	0,827	401	0,0090	1,165	1103	0,0050
0,784	384	0,0050	0,927	312	0,0050	1,299	764	0,0030
0,850	298	0,0030	1,032	223	0,0040	1,439	573	0,0010
0,920	249	0,0000	1,144	174	0,0000	1,586	424	0,0000
0,992	210	0,0000	1,261	148	0,0000	1,741	326	0,0000
1,066	196	0,0000	1,384	126	0,0000	1,903	242	0,0000

Tabela 3.3: Excerto da tabela de deltas, genes diferenecialmente expressos e FDR obtida com a SAM confrontando leveduras vínicas e a levedura clínica. A tabela representa os resultados obtidos quando comparadas as leveduras vínicas com a levedura clínica nos tempos 2, 3 e 5. Para cada tempo retirou-se um excerto da tabela fornecida com a SAM com os deltas e os correspondentes números de genes diferencialmente expressos detectados e FDR.

interessará tanto a obtenção de um grande conjunto de genes diferencialmente expressos, mas sim um pequeno conjunto de genes que mais evidenciem alterações no seu nível de expressão, para estudo futuro destes genes em laboratório. Dado que a grande maioria dos genes detectados com a estatística de teste t, foram detectados com a estatística de teste de wilcoxon, optou-se por prosseguir a restante análise com os resultados obtidos com a aplicação da estatística de teste t, já que o conjunto de genes é mais restrito, sendo à partida, o conjunto de genes que mais evidenciam alterações na sua expressão genética.

Ao biólogo interessará saber quais os genes diferencialmente expressos, mas principalmente os genes comuns às três análises, sendo estes os que à partida apresentarão as maiores diferenças de expressão genetica. Nesse sentido verificou-se quais os genes comuns para os três confrontos. Dada a quantidade de genes comuns entre as três análises não se fará uma lista com a identificação de cada gene, mas apenas uma tabela com o número de genes comuns para cada um dos tempos. A Tabela 3.12 representa essa informação.

Vínicas e Laboratorial								
Tempo2			**Tempo3**			**Tempo5**		
delta	*genes DE*	*FDR*	*delta*	*genes DE*	*FDR*	*delta*	*genes DE*	*FDR*
1,135	365	0,0070	0,531	276	0,022	0,672	421	0,012
1,252	316	0,0030	0,604	240	0,008	0,765	308	0,005
1,374	261	0,0020	0,681	197	0,005	0,863	245	0,003
1,501	229	0,0000	0,764	173	0,000	0,968	193	0,000
1,635	209	0,0000	0,861	151	0,000	1,078	155	0,000
1,774	190	0,0000	0,943	123	0,000	1,194	126	0,000

Tabela 3.4: Excerto da tabela de deltas, genes diferenecialmente expressos e FDR obtida com a SAM confrontando leveduras vínicas e a levedura laboratorial. A tabela representa os resultados obtidos quando comparadas as leveduras vínicas com a levedura laboratorial nos tempos 2, 3 e 5. Para cada tempo retirou-se um excerto da tabela fornecida com a SAM com os deltas e os correspondentes números de genes diferencialmente expressos detectados e FDR.

	Vínicas e Não Vínicas	**Vínicas e Clínica**	**Vínicas e Laboratorial**
Tempo	**Delta**	**Delta**	**Delta**
2	1,362	0,920	1,501
3	0,950	1,144	0,764
5	1,119	1,568	0,968

Tabela 3.5: Deltas escolhidos para a SAM. A primeira coluna é relativa aos tempos 2, 3 e 5 para o confronto entre leveduras vínicas e não vínicas, a segunda coluna é relativa aos três tempo para o confronto entre as leveduras vínicas e a clínica e a última é relativa aos mesmos três tempos para o confronto entre leveduras vínicas e a laboratorial.

	Vínicas e Não Vínicas		
	Tempo2	**Tempo3**	**Tempo5**
genes DE positivos	253	213	146
genes DE negativos	0	1	3

Tabela 3.6: Genes diferencialmente expressos obtidos para a base de dados `Fermentation` confrontando as leveduras vínicas com as não vínicas. A primeira coluna é realtiva ao tempo 2, a segunda ao tempo três e a última ao tempo 5. A tabela permite verificar que o número de genes diferencialmente expressos não variam muito ao longo do tempo. De qualquer forma, parecem existir menos diferenças com o passar do tempo, sendo no tempo 5 onde existem menos diferenças.

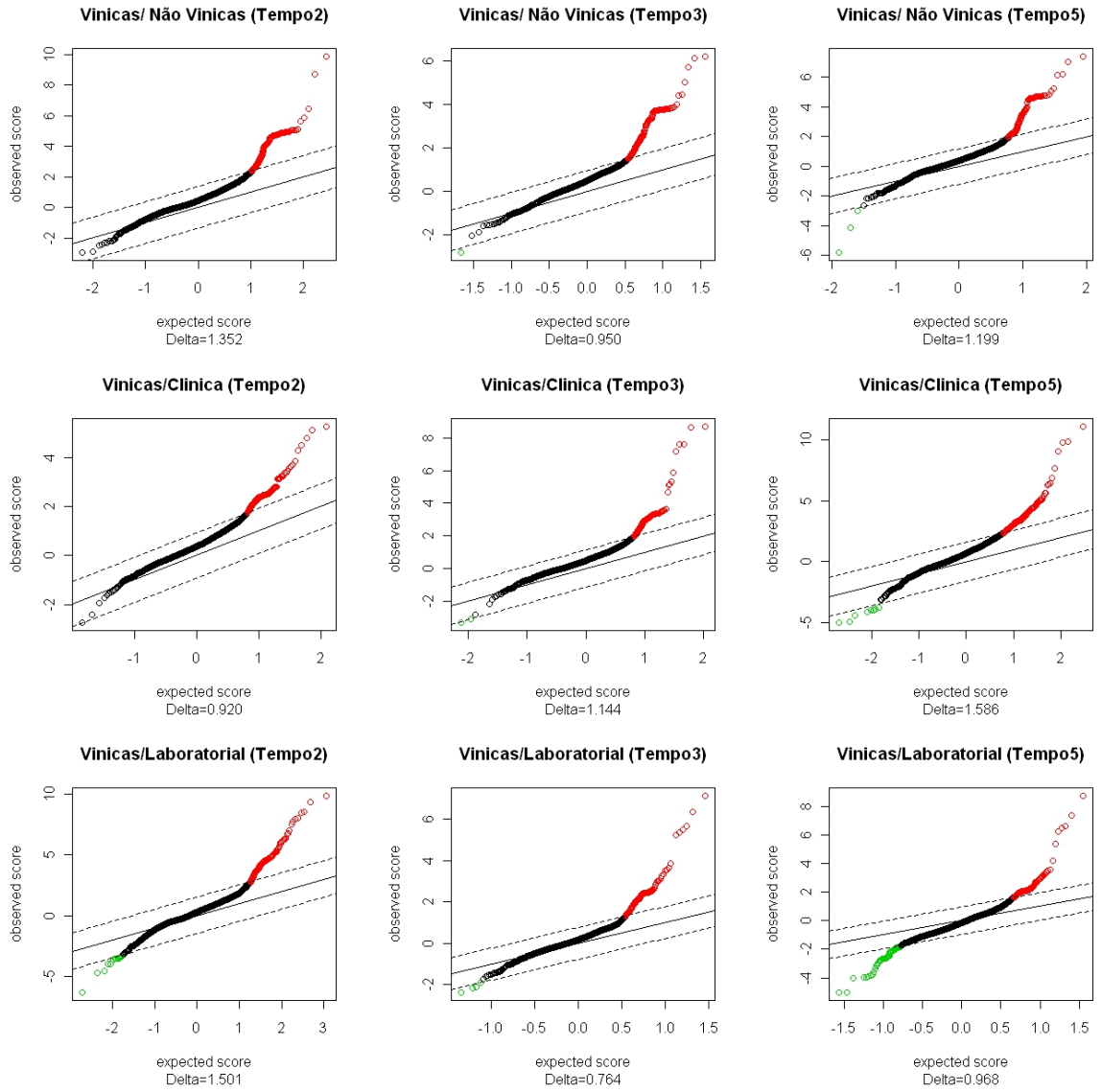

Figura 3.5: Gráficos dos genes diferencialmente expressos para a base de dados leveduras. Gráfico de quantis para as estatísticas de teste produzidas nos três tempos para as três análises. Os primeiros três gráficos representam os resultados produzidos confrontando as leveduras vínicas e não vínicas para os tempos 2, 3 e 5 respectivamente, os três seguintes para as leveduras vínicas e clínica e os últimos três para as vínicas e laboratorial.

	Vínicas e Clínica		
	Tempo2	**Tempo3**	**Tempo5**
genes DE positivos	249	172	416
genes DE negativos	0	2	8

Tabela 3.7: Genes diferencialmente expressos obtidos para a base de dados `Fermentation` confrontando as leveduras vínicas com a clínica. A primeira coluna é realtiva ao tempo 2, a segunda ao tempo três e a última ao tempo 5. O número de genes diferencialmente expressos são em menor quantidade no tempo 3 e em maior quantidade no tempo 5. Aparentemente as leveduras vínicas são mais semelhantes às leveduras clínicas no tempo 3. No tempo 5, parece que estas assumem comportamentos bastante diferenciados.

	Vínicas e Laboratorial		
	Tempo2	**Tempo3**	**Tempo5**
genes DE positivos	214	169	125
genes DE negativos	15	4	68

Tabela 3.8: Genes diferencialmente expressos obtidos para a base de dados `Fermentation` confrontando as leveduras vínicas com a laboratorial. A primeira coluna é realtiva ao tempo 2, a segunda ao tempo três e a última ao tempo 5. Como já referido, de todos os confrontos este é o confronto que obtém menos genes diferencialmente expressos. Aparentemente, a levedura laboratorial é a que mais se assemelha às leveduras vínicas. As maiores diferenças entre estas é visível no tempo 2, sendo no tempo 5 que se obtêm menores diferenças.

Vínicas e Não Vínicas		
T2	**T3**	**T5**
212	242	193

Tabela 3.9: Resultados obtidos com a estatística de teste de wilcoxon quando confrontadas as leveduras vínicas e as não vínicas. Quando usada a estatística de teste de wilcoxon e estando em confronto as leveduras vínicas e as não vínicas foram obtidos 212 genes no Tempo 2, 242 no Tempo 3 e 193 no Tempo 5 sendo que 198, 143 e 132 destes genes coincidiram com os genes obtidos com a estatística de teste t, respectivamente.

Vínicas e Clínica		
T2	**T3**	**T5**
1518	1231	1212

Tabela 3.10: Resultados obtidos com a estatística de teste de wilcoxon quando confrontadas as leveduras vínicas e a clínica. Quando usada a estatística de teste de wilcoxon e estando em confronto as leveduras vínicas e a clínica foram obtidos 1518 genes no Tempo 2, 1231 no Tempo 3 e 1212 no Tempo 5 sendo que 249, 172 e 416 destes genes coincidiram com os genes obtidos com a estatística de teste t, respectivamente.

Vínicas e Laboratorial		
T2	**T3**	**T5**
343	550	807

Tabela 3.11: Resultados obtidos com a estatística de teste de wilcoxon quando confrontadas as leveduras vínicas e a laboratorial. Quando usada a estatística de teste de wilcoxon e estando em confronto as leveduras vínicas e a laboratorial foram obtidos 343 genes no Tempo 2, 550 no Tempo 3 e 807 no Tempo 5 sendo que 205, 161 e 68 destes genes coincidiram com os genes obtidos com a estatística de teste t, respectivamente.

Genes DE Comuns		
T2	T3	T5
63	86	51

Tabela 3.12: Genes diferencialmente expressos comuns nas três análises. No tempo 2 existem 63 genes diferencialmente expressos coincidentes nas três análises, no tempo 3 existem 86 genes em comum e no tempo 5 50 genes.

Capítulo 4

Métodos de Bayes Empíricos

4.1 O Conceito Bayes Empírico

Na inferência clássica os parâmetros desconhecidos de um modelo são tratados como quantidades não aleatórias, ou seja, quantidades fixas. Já na inferência bayesiana os parâmetros assumem uma natureza aleatória, sendo-lhe associada uma distribuição (distribuição *a priori*). A informação de que se dispõe sobre os parâmetros, traduzida probabilisticamente através da distribuição *a priori*, pode ser actualizada utilizando a relação $p(\boldsymbol{y}\,|\boldsymbol{\theta})$, onde $\boldsymbol{Y}$ corresponde à matriz das observações aleatórias e $\boldsymbol{\theta}$ o vector de parâmetros desconhecidos. Após observar $\boldsymbol{Y} = \boldsymbol{y}$ a quantidade de informação obtida sobre os dados será maior, sendo possível obter uma aproximação do vector de parâmetros $\boldsymbol{\theta}$ mais próximo do verdadeiro valor. O teorema de Bayes constitui a regra de actualização utilizada para obter a designada distribuição *a posteriori* dada por

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y},\boldsymbol{\theta})}{p(\mathbf{y})} = \frac{p(\mathbf{y},\boldsymbol{\theta})}{\int p(\mathbf{y},\mathbf{u})d\mathbf{u}} = \frac{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int f(\mathbf{y}|\mathbf{u})\pi(\mathbf{u})d\mathbf{u}}. \tag{4.1}$$

Usualmente a distribuição *a priori* depende de outros parâmetros, chamados hiperparâmetros, que poderão ser ou não conhecidos.

O exemplo de seguida permite ilustrar a utilização do teorema de Bayes para estimar a probabilidade *a posteriori* de $\boldsymbol{\theta}$.

Exemplo 4.1.1. *[23]Seja Y uma variável aleatória com distribuição $N(\theta,\sigma^2)$. Logo $f(y|\theta) = \frac{1}{\sigma\sqrt{2\pi}}exp(-\frac{(y-\theta)^2}{2\sigma^2})$, com σ^2 conhecido e $y \in \Re$. Assuma-se uma distribuição a priori $N(\mu,\tau^2)$ para θ, onde $\eta = (\mu,\tau)$, $\mu \in \Re$, $\tau > 0$ constitui o vector hiperparâmetros. A distribuição a posteriori será dada pela regra (4.4). Efectuando os cálculos para o numerador obtém-se:*

$$\begin{aligned} f(y|\theta)\pi(\theta|\boldsymbol{\eta}) &= \frac{1}{\sigma\sqrt{2\pi}}exp\left\{-\frac{(y-\theta)^2}{2\sigma^2}\right\}\frac{1}{\tau\sqrt{2\pi}}exp\left\{-\frac{(\theta-\mu)^2}{2\tau^2}\right\} \\ &= \frac{1}{\sigma\tau(\sqrt{2\pi})^2}exp\left\{-(\frac{(y-\theta)^2}{2\sigma^2}+\frac{(\theta-\mu)^2}{2\tau^2})\right\} \\ &= \frac{1}{\sqrt{2\pi(\sigma^2+\tau^2)}}exp\left\{-\frac{(y+\mu)^2}{2(\tau^2+\sigma^2)}\right\}\left[\frac{1}{\frac{\sigma\tau}{\sqrt{\sigma^2+\tau^2}}\sqrt{2\pi}}exp\left\{-\frac{(\theta-\frac{\sigma^2\mu+\tau^2y}{\sigma^2+\tau^2})^2}{\frac{2\sigma^2\tau^2}{\sigma^2+\tau^2}}\right\}\right] \end{aligned} \tag{4.2}$$

Tendo em conta a expressão do numerador, após alguns cálculos algébricos o denominador de (4.4) vem dado por:

$$\begin{aligned} &\int f(y|u)\pi(u|\boldsymbol{\eta})du \\ &\quad= \frac{1}{\sqrt{2\pi(\sigma^2+\tau^2)}}exp\left\{-\frac{(y+\mu)^2}{2(\tau^2+\sigma^2)}\right\}\int\frac{1}{\frac{\sigma\tau}{\sqrt{\sigma^2+\tau^2}}\sqrt{2\pi}}exp\left\{-\frac{(u-\frac{\sigma^2\mu+\tau^2y}{\sigma^2+\tau^2})^2}{\frac{2\sigma^2\tau^2}{\sigma^2+\tau^2}}\right\}du \\ &\quad= \frac{1}{\sqrt{2\pi(\sigma^2+\tau^2)}}exp\left\{-\frac{(y+\mu)^2}{2(\tau^2+\sigma^2)}\right\} \end{aligned} \tag{4.3}$$

Consequentemente, a probabilidade a posteriori *é obtida como*

$$p(\theta|y) = \frac{\frac{1}{\sqrt{2\pi(\sigma^2+\tau^2)}}exp\left\{-\frac{(y+\mu)^2}{2(\tau^2+\sigma^2)}\right\}\left[\frac{1}{\frac{\sigma\tau}{\sqrt{\sigma^2+\tau^2}}\sqrt{2\pi}}exp\left\{-\frac{\left(\theta-\frac{\sigma^2\mu+\tau^2y}{\sigma^2+\tau^2}\right)^2}{\frac{2\sigma^2\tau^2}{\sigma^2+\tau^2}}\right\}\right]}{\frac{1}{\sqrt{2\pi(\sigma^2+\tau^2)}}exp\left\{-\frac{(y+\mu)^2}{2(\tau^2+\sigma^2)}\right\}}.$$

Pelo que $\theta|y \sim N(\theta|\frac{\sigma^2\mu+\tau^2y}{\sigma^2+\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2})$.

Denotando $B = \frac{\sigma^2}{\sigma^2+\tau^2}$ *tem-se que a média e variância da distribuição a posteriori são dadas por, respectivamente,* $\frac{\sigma^2\mu+\tau^2y}{\sigma^2+\tau^2} = B\mu + (1-B)y = y - B(y-\mu)$ *e* $\frac{\sigma^2\tau^2}{\sigma^2+\tau^2} = B\tau^2 = (1-B)\sigma^2$. *Como* $0 < B < 1$ *a média a posteriori é uma média ponderada da média a priori* (μ) *e das observações* (y) *cujos pesos são inversamente proporcionais às correspondentes variâncias* (σ^2, τ^2). *Assim, B é muitas vezes chamado de* "factor de redução" *uma vez que mede a distância da média a posteriori relativa aos valores observados. De facto, se* $\tau^2 \gg \sigma^2$, *isto é, existe pouca informação a priori, então B será pequeno e a média a posteriori será próxima do valor observado y, caso contrário (existe muita informação a priori) B será próximo de 1 e o valor da média a posteriori será próximo do valor da média a priori. No que respeita à variância, esta será sempre menor que qualquer uma das variâncias (das observações ou a priori).*

Concretamente, considere-se os valores observados representados por $\boldsymbol{y} = (\boldsymbol{y_1}, \boldsymbol{y_2}, ..., \boldsymbol{y_n})$ e seja $p(\boldsymbol{y}|\boldsymbol{\theta})$ a função verosimilhança de $\boldsymbol{y}$, onde $\boldsymbol{\theta}$ corresponde ao vector de parâmetros desconhecidos com distribuição *a priori* $\pi(\boldsymbol{\theta}|\boldsymbol{\eta})$ e $\boldsymbol{\eta}$ representa o vector dos hiperparâmetros que poderá ainda ser um vector de parâmetros desconhecidos. A estes hiperparâmetros, enquanto quantidades aleatórias, é associada a chamada distribuição *hiperpriori* $h(\boldsymbol{\eta})$. Este procedimento em cadeia, de atribuição de uma distribuição aos parâmetros desconhecidos, conduz a uma generalização do modelo inicial. É importante também referir que a distribuição *a priori* poderá ser paramétrica ou não paramétrica, dependendo do conhecimento ou desconhecimento da distribuição dos parâmetros. Este tipo de procedimento em cadeia leva à estruturação de um modelo hierárquico onde cada nova distribuição formada constitui um novo nível da hierarquia. O número de níveis a considerar dependerá do problema em estudo e dos dados observados. Esta hierarquia necessariamente terá de parar em determinado ponto havendo a necessidade de assumir que os restantes parâmetros *a priori* são conhecidos. É aqui que entra a abordagem de Bayes Empírica, que usa os dados observados para estimar estes parâmetros de modo a possibilitar o cálculo da distribuição *a posteriori* do vector de parâmetros desconhecido inicial.

Assim se o modelo estiver hierarquizado em apenas dois níveis, tem-se:

$$p(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\eta}) = \frac{p(\mathbf{y}, \boldsymbol{\theta}|\boldsymbol{\eta})}{p(\mathbf{y}|\boldsymbol{\eta})} = \frac{p(\mathbf{y}, \boldsymbol{\theta}|\boldsymbol{\eta})}{\int p(\mathbf{y}, \mathbf{u}|\boldsymbol{\eta}) d\mathbf{u}} = \frac{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\eta})}{\int f(\mathbf{y}|\mathbf{u})\pi(\mathbf{u}|\boldsymbol{\eta}) d\mathbf{u}}. \tag{4.4}$$

Se $\boldsymbol{\eta}$ é conhecido, então $\boldsymbol{\eta}$ pode ser eliminado e (4.4) pode ser reescrito como

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int f(\mathbf{y}|\mathbf{u})\pi(\mathbf{u}|\boldsymbol{\eta}) d\mathbf{u}}. \tag{4.5}$$

Se $\boldsymbol{\eta}$ não é conhecido e supondo que o modelo hierarárquico apenas tem dois níveis, a distribuição $h(\boldsymbol{\eta})$ será introduzida em (4.4) resultando na seguinte expressão:

$$p(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\eta}) = \frac{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\eta})}{\int f(\mathbf{y}|\mathbf{u})\pi(\mathbf{u}|\boldsymbol{\eta}) d\mathbf{u}} = \frac{\int f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\eta})h(\boldsymbol{\eta}) d\boldsymbol{\eta}}{\int\int f(\mathbf{y}|\mathbf{u})\pi(\mathbf{u}|\boldsymbol{\eta})h(\boldsymbol{\eta}) d\boldsymbol{\eta} d\mathbf{u}} \tag{4.6}$$

Este cálculo é muitas vezes complicado. Uma das formas de ultrapassar esta dificuldade é aplicar uma abordagem empírica em (4.6) obtendo um estimador para $\boldsymbol{\eta}$, $\widehat{\boldsymbol{\eta}}$. A inferência é então feita substituindo $\widehat{\boldsymbol{\eta}}$ em (4.6). Uma vez que $\boldsymbol{\eta}$ passará a ser conhecido, (4.4),(4.6) podem ser reescritos como em (4.5). Consequentemente

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \tag{4.7}$$

Ou seja, a probabilidade *a posteriori* é proporcional ao produto da função de verosimilhança com a distribuição *a priori*. De facto, qualquer constante ou função pode ser multiplicada pela distribuição verosimilhança sem que a distribuição *a priori* seja alterada.

Na abordagem de Bayes empírica diversos métodos têm sido propostos para a obtenção de uma estimativa dos parâmetros em causa. Se nada se sabe sobre a distribuição de $\boldsymbol{\theta}$, por

exemplo, se o modelo é não paramétrico, o método de Robbins tem sido considerado uma excelente alternativa para ultrapassar esta incerteza. Esse método, com base na estimativa de Bayes para θ_i, $\theta_i^B = \frac{(y_i+1)m_G(y_i+1)}{m_G(y_i)}$, propõe estimar as probabilidades marginais através das frequências empíricas, do seguinte modo [23]:

$$\widehat{\theta}_i^B = (y_i + 1)\frac{\#(y : y = y_i + 1)}{\#(y : y = y_i)}.$$

Assim, as estimativas dependem apenas dos dados observados, não sendo de todo necessário o envolvimento de uma distribuição para $\boldsymbol{\theta}$. As abordagens não paramétricas dado que não assumem nenhum modelo e apenas fazem uso da informação contida nos dados são em geral mais potentes. Como desvantagem, na prática oferecem muitas vezes grandes dificuldades na sua implementação [22]. Se a distribuição de $\boldsymbol{\theta}$ é conhecida e depende de um vector de hiperparâmetros desconhecido, este pode ser estimado usando métodos clássicos. Assim poderá ser usado o *Método dos Momentos* cuja ideia-base é utilizar os momentos da amostra para estimar os correspondentes momentos da população ou o *Método de Máxima Verosimilhança*, em geral mais complexo mas que origina estimadores de maior qualidade e cuja ideia-base será encontrar o valor do parâmetro que maximiza a função verosimilhança.

Muitas vezes a obtenção de estimadores de máxima verosimilhança não é fácil usando apenas o método iterativo padrão. Uma alternativa para obter esses estimadores nos casos mais complicados é o *algoritmo EM* do inglês (*Expectation-Maximization*).

4.1.1 Algoritmo EM

O algoritmo EM tornou-se uma ferramenta popular nos problemas de estimação envolvendo bases de dados com valores omissos (*missing values*) e distribuições de mistura. Trata-se de um algoritmo iterativo constituído, tal como o nome sugere, por dois passos em cada iteração: *Passo E* (de *Expectation*) e *Passo M* (de *Maximization*). Este pode ser aplicado em modelos paramétricos ou não paramétricos. Nesta dissertação apenas será descrito o algoritmo EM na estimação paramétrica, concretamente, na estimação do vector de hiperparâmetros $\boldsymbol{\eta}$ associado à distribuição *a priori* $\pi(\boldsymbol{\theta}|\boldsymbol{\eta})$.

Assim sendo, dada uma amostra, o objectivo do algoritmo é estimar o vector de hiperparâmetros $\boldsymbol{\eta}$ para o qual a função de verosimilhança $L(\boldsymbol{\eta}|\boldsymbol{y})$ atinge o seu valor máximo; ou seja, encontrar a estimativa de máxima verosimilhança de $\boldsymbol{\eta}$.

Considerando a função log-verosimilhança dada por

$$l(\boldsymbol{\eta}|\boldsymbol{y}) = log(L(\boldsymbol{\eta}|\boldsymbol{y})) \ ,$$

função do vector de hiperparâmetros $\boldsymbol{\eta}$ quando conhecidos os dados $\boldsymbol{y}$, e denotando por $\boldsymbol{\eta}^{(j)}$ a estimativa de $\boldsymbol{\eta}$ obtida na j-ésima iteração, os dois passos na $j+1$-ésima iteração do algoritmo EM podem ser descritos como [21]:

1. ***Passo E***: Calcular o valor esperado da função de log-verosimilhança $l(\boldsymbol{\eta}|\boldsymbol{y})$, tendo em conta a distribuição de θ e a estimativa actual $\boldsymbol{\eta}^{(j)}$. Ou seja, obter o valor de

$$Q(\boldsymbol{\eta}|\boldsymbol{\eta}^{(j)}) = E(l(\boldsymbol{\eta}|\boldsymbol{y}, \boldsymbol{\theta})|\boldsymbol{y}, \boldsymbol{\eta}^{(j)}).$$

2. ***Passo M***: Obter o vector de parâmetros $\boldsymbol{\eta}^{(j+1)}$ que maximiza a esperança obtida no *Passo E* relativamente a $\boldsymbol{\eta}$, resolvendo

$$\boldsymbol{\eta}^{(j+1)} = \arg\max_{\boldsymbol{\eta}} Q(\boldsymbol{\eta}|\boldsymbol{\eta}^{(j)}).$$

Vários estudos provaram que em cada iteração a verosimilhança marginal ou cresce ou mantém-se constante [23],

$$L(\boldsymbol{\eta}^{(j+1)}|\boldsymbol{y}) \geq L(\boldsymbol{\eta}^{(j)}|\boldsymbol{y}) \ \ para\ todo\ o\ j,$$

pelo que o algoritmo converge de forma monótona. Esta convergência é geralmente local (conduzindo apenas a um máximo local), pelo que não dispensa a consideração de várias aproximações iniciais. O algoritmo pára quando um critério de paragem for atingido. Geralmente considera-se como critério de paragem $\| \boldsymbol{\eta}^{(j+1)} - \boldsymbol{\eta}^{(j)} \| \leq \epsilon$, para um ϵ suficientemente pequeno.

4.2 Modelos Lineares

A maioria das experiências de *microarrays* com mais de duas amostras usam uma amostra de referência para facilitar a realização de todas as comparações possíveis. Uma lacuna desta abordagem é o gasto de uma quantidade de esforço na preparação da amostra de referência quando na verdade esta pode ter pouco ou nenhum interesse biológico. Os modelos lineares podem ser aplicados a qualquer experiência de *microarrays*, usando um modelo linear para cada gene. Estes permitem ao utilizador estabelecer quais os contrastes de interesse.

Para o modelo em causa assume-se que se tem n *microarrays* que produzem observações para cada gene. Denotando por $\boldsymbol{y}_i^T = (y_{i1}, y_{i2}, ..., y_{in})$ o vector de observações produzido pelos n *microarrays* para o i-ésimo gene, o modelo linear para as observações assume a seguinte estrutura:

$$\boldsymbol{y}_i = X\boldsymbol{\alpha}_i + \epsilon_i \ com\ \epsilon_i\ oerro\ aleatório.$$

Onde as condições do modelo são dadas por:

1. $E[\boldsymbol{Y_i}] = X\boldsymbol{\alpha}_i$,
   onde X representa a matriz de delineamento e $\boldsymbol{\alpha}_i$ o vector de coeficientes reflectindo diferenças de expressão em estudo.

2. $Var[\boldsymbol{Y_i}] = W_i\sigma_i^2$,
   onde W_i é uma matriz conhecida definida não negativa. O vector $\boldsymbol{y}_i$ poderá conter valores omissos e W_i poderá conter zeros na diagonal principal.

4.2.1 A Matriz de Delineamento

A condição 1. pode ser escrita em termos da matriz designada por matriz de delineamento (*design matrix*), a qual está associada ao desenho experimental das amostras de RNA hibridadas de todos os *microarrays* construídos. Cada linha da matriz corresponde a um microarray e cada coluna a um coeficiente, os quais devem ser independentes. Para ilustrar a construção de uma matriz de delineamento apresenta-se o seguinte exemplo.

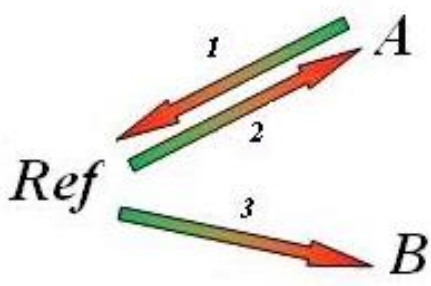

Figura 4.1: Experiência com três microarrays comparando as amostras A e B, usando uma referência comum *Ref*. A ponta verde identifica a amostra etiquetada com o fluóforo verde e a vermelha a amostra etiquetada com o fluóforo vermelho. Imagem retirada de [20].

Exemplo 4.2.1. *Considere-se uma experiência realizada com três* microarrays *onde são comparadas duas amostras (A e B) com uma referência comum (Ref), ver Figura 4.1, sendo que a amostra A é comparada com a referência usando a técnica de fluóforos trocados,* dye swap*, e que deve ser indicada invertendo o sinal do coeficiente associado a um dos* microarrays*.*

Cada microarray *é representado pela amostra que está no topo da seta menos a que está associada à parte inferior da seta. Ou seja, o* microarray 1 *será representado por* $(A-Ref)$*, o* microarray 2 *por* $(Ref-A)$ *ou* $-(A-Ref)$ *e o* microarray 3 *por* $(B-Ref)$*. Assim tem-se:*

$$\begin{bmatrix} microarray1 \\ microarray2 \\ microarray3 \end{bmatrix} = \begin{bmatrix} A-Ref \\ -(A-Ref) \\ B-Ref \end{bmatrix}.$$

Se se considerar por exemplo, $(A-Ref)$ *e* $(B-ref)$ *como coeficientes, então,*

$$\boldsymbol{\alpha}_i = \begin{bmatrix} A-Ref \\ B-Ref \end{bmatrix}$$

e

$$\begin{bmatrix} microarray1 \\ microarray2 \\ microarray3 \end{bmatrix} = \begin{bmatrix} A-Ref \\ -(A-Ref) \\ B-Ref \end{bmatrix} = X \begin{bmatrix} A-Ref \\ B-Ref \end{bmatrix}$$

pelo que,

$$X = \begin{array}{c} \\ microarray1 \\ microarray2 \\ microarray3 \end{array} \begin{array}{c} \begin{array}{cc} A\text{-}Ref & B\text{-}Ref \end{array} \\ \begin{bmatrix} 1 & 0 \\ -1 & 0 \\ 0 & 1 \end{bmatrix} \end{array}.$$

O modelo linear para o gene i vem então dado por:

$$E \begin{bmatrix} y_{i1} \\ y_{i2} \\ y_{i3} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} A - Ref \\ B - Ref \end{bmatrix}.$$

4.2.2 Matriz de Contrastes

Muitas vezes determinados contrastes são de grande interesse biológico. Como na matriz de delineamento todos os coeficientes devem ser independentes nem sempre é possível representar todos os contrastes de interesse. A matriz de contrastes possibilita a combinação dos coeficientes da matriz de delineamento por forma a estabelecer as comparações desejadas entre amostras de RNA.

Os contrastes de interesse para cada gene i são matematicamente definidos como [22]:

$$\beta_i = C^T \boldsymbol{\alpha}_i$$

onde C representa a matriz de contrastes.

Exemplo 4.2.2. *Considere-se o Exemplo 4.2.1. Note-se que as comparações efectuadas com a utilização de $\boldsymbol{\alpha}_i$ não contemplam a comparação pretendida (*B-A*). Esta comparação pode ser feita de duas formas, através da matriz contrastes ou especificando directamente essa comparação no vector de coeficientes.*

Especificando directamente a comparação dos microarrays A *e* B*, um modelo linear equivalente ao anterior seria:*

$$E \begin{bmatrix} y_{i1} \\ y_{i2} \\ y_{i3} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} A - Ref \\ B - A \end{bmatrix}$$

Utilizando a matriz de contrastes para efectuar a comparação entre A e B ter-se-ia

$$C = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

e o vector de coeficientes do Exemplo 4.2.1,

$$\boldsymbol{\alpha}_i = \begin{bmatrix} A - Ref \\ B - Ref \end{bmatrix}.$$

Na realidade,

$$\boldsymbol{\beta}_i = C^T \boldsymbol{\alpha}_i = \left[\begin{array}{c} B - A \end{array} \right]$$

O número de contrastes a introduzir na matriz de contrastes pode variar podendo inclusive ser em número superior ao número de coeficientes, nesse caso alguns deles terão de ser linearmente dependentes.

A questão a verificar será se para todo o gene i e para toda a medida de contraste j, se se tem $\beta_{ij} = 0$, ou seja, se não existem alterações dos níveis de expressão. As hipóteses a testar serão então

$$H_0 : \beta_{ij} = 0 \; vs \; H_1 : \beta_{ij} \neq 0.$$

Assumindo que o modelo linear está ajustado para cada um dos genes, $\widehat{\boldsymbol{\alpha}}_i$ representam os estimadores para os coeficientes, $var(\widehat{\boldsymbol{\alpha}}_i) = V_i s_i^2$ as respectivas matrizes de covariâncias estimadas, onde s_i^2 são os estimadores de σ_i^2 e V_i uma matriz definida positiva não dependente de s_i^2. Os estimadores de contraste são respresentados por $\widehat{\boldsymbol{\beta}}_i = C^T \; \widehat{\boldsymbol{\alpha}}_i$ tendo como matrizes de covariâncias estimadas $var(\widehat{\boldsymbol{\beta}}_i) = C^T V_i C s_i^2$. Assume-se que os estimadores de contraste são aproximadamente normais com média $\boldsymbol{\beta}_i$ e matriz de covariâncias $C^T V_i C \sigma_i^2$. Assume-se que as variâncias residuais seguem aproximadamente uma distribuição de qui-quadrado escalonada. A matriz de covariâncias não escalonada V_i (isto é, a matriz que multiplicada pela estimativa da variância do erro, produz uma estimativa da matriz de covariâncias para os coeficientes) pode depender de α_i. Segundo [22], nesse caso, assume-se que a matriz de covariâncias é avaliada em $\widehat{\boldsymbol{\alpha}}_i$, podendo essa variância ser ignorada para uma aproximação de primeira ordem.

Considere-se v_{ij} o j-ésimo elemento da diagonal de $C^T V_i C$ e que

$$\widehat{\boldsymbol{\beta}}_{ij} | \boldsymbol{\beta}_{ij}, \sigma_i^2 \sim N(\boldsymbol{\beta}_{ij}, v_{ij}\sigma_i^2) \tag{4.8}$$

e

$$s_i^2 | \sigma_i^2 \sim \frac{\sigma_i^2}{d_i} \chi_{d_i}^2 \tag{4.9}$$

onde d_i é o grau de liberdade residual para o modelo linear do gene i.

Assumindo que os estimadores $\widehat{\boldsymbol{\beta}}_i$ e s_i^2 são independentes. A estatistica t ordinária

$$t_{ij} = \frac{\widehat{\boldsymbol{\beta}}_{ij}}{s_i \sqrt{v_{ij}}} \tag{4.10}$$

quando $\beta_{ij} = 0$, segue uma distribuição t-Student com d_i graus de liberdade.

O pacote `limma` está implementado na linguagem R e disponível no site do projecto `Bioconductor`. Este faz uso dos modelos lineares para a análise do comportamento dos genes e detecção de genes diferencialmente expressos. De forma a produzir resultados mais estáveis, mesmo quando o número de *microarrays* é pequeno utiliza os métodos de Bayes empíricos.

4.2.3 Detecção de Genes Diferencialmente Expressos

Quando é delineada uma experiência com *microarrays* o seu número em geral é relativamente pequeno. Tal facto conduz a um número baixo de graus de liberdade associado às estimativas da variância para cada gene que consequentemente levará a uma grande instabilidade dos resultados produzidos pela estatística de teste (4.10). Para contornar este problema, [22] propõe a utilização de uma estatística de teste t moderada, onde são usados os desvios padrão residuais *a posteriori* em vez dos desvios padrão ordinais. Com o uso da abordagem de Bayes empírica as variâncias amostrais estimadas são reduzidas permitindo assim resultados mais estáveis.

O Modelo Hierárquico

Importa ainda relembrar, antes de prosseguir com a estruturação do modelo, que este assume a independencia dos estimadores $\widehat{\boldsymbol{\beta}}_i$ e S_i^2, que o objectivo será testar $H_0 : \beta_{ij} = 0$ *vs* $H_1 : \beta_{ij} \neq 0$ e obter uma estatistica de teste melhorada.

Como informação *a priori* [22] assume que

$$\frac{1}{\sigma_i^2} \sim \frac{1}{d_0 s_0^2}\chi^2_{d_0}$$

permitindo descrever como se espera que as variâncias variem ao longo dos genes. Para toda a medida de contraste j considera que a probabilidade $p(\beta_{ij} \neq 0) = p_j$ é conhecida e representa a proporção de genes diferencialmente expressos verdadeiros. Para estes genes assume a mesma distribuição sob a hipótese nula, ou seja,

$$\beta_{ij}|\sigma_i^2, \beta_{ij} \neq 0 \sim N(0, v_{0j}\sigma_i^2), \tag{4.11}$$

tal como em (4.8), com a excepção da variância, que é agora substituida pela variância não escalonada v_{0j}. A expressão (4.11) descreve a distribuição esperada para os "*log-fold changes*"dos genes diferencialmente expressos. A aplicação da abordagem de Bayes empírica ao método proposto modifica os erros padrão das "*log-fold changes*"estimadas permitindo uma inferência mais estável e um aumento da potência, particularmente para experiências envolvendo um número pequeno de *microarrays*.

Sob as condições acima descritas, [22] obtém a média *a posteriori* $\widetilde{s}_i^{-2}$ de σ_i^{-2} conhecido s_i^2, dada por

$$\widetilde{s}_i^2 = \frac{d_0 s_0^2 + d_i s_i^2}{d_0 + d_i}.$$

Deste modo, os valores *a posteriori* reduzem as variâncias através dos valores *a priori*, cujo grau de redução depende dos graus de liberdade d_0 e d_i.

A estatistica de teste (4.10), pode agora ser substituida por:

$$\widetilde{t}_{ij} = \frac{\widehat{\beta}_{ij}}{\widetilde{s}_i\sqrt{v_{ij}}}$$

Segundo [19] esta estatística de teste é designada de estatistica t moderada no sentido em que os erros padrão foram moderados ao longo dos genes, isto é, reajustados em direcção a um valor comum. Esta estatística produz valores p da mesma forma que a estatistica t ordinária, mas com um maior número de graus de liberdade (em vez de d_i graus de liberdade ter-se-ão $d_i + d_0$ graus de liberdade), conferindo uma maior confiança relativa aos erros padrão suavizados.

As distribuiçoes Marginais

Em [22] é ainda demonstrado que

$$p(\widetilde{t}_{ij}, s_i^2|\beta_{ij} = 0) = p(\widetilde{t}_{ij}|\beta_{ij} = 0)p(s_i^2|\beta_{ij} = 0) = p(\widetilde{t}_{ij}|\beta_{ij} = 0)p(s_i^2)$$

ou seja, $\widetilde{t}_{ij}$ e s_i^2 são independentes, sendo que

$$s_i^2 \sim s_0^2 F_{d_i,d_0}$$

$$\widetilde{t}_{ij}|\beta_{ij} = 0 \sim t_{d_0+d_i}.$$

e

$$\widetilde{t}_{ij}|\beta_{ij} \neq 0 \sim \left(1 + \frac{v_{0j}}{v_{ij}}\right)^{\frac{1}{2}} t_{d_0+d_i}.$$

Consequentemente, a distribuição marginal $\widetilde{t}$ ao longo de todos os genes é uma mistura da distribuição t-Student escalonada e da distribuição t-Student ordinária, cujos pesos são respectivamente p e $1 - p$. Ou seja,

$$\widetilde{t}_{ij} \sim (1 - p_j)t_{d_0+d_i} + p_j\sqrt{1 + \frac{v_{0j}}{v_{ij}}}t_{d_0+d_i}.$$

A Estatística B

Em [22] é desenvolvida uma outra estatística dada por $B_{ij} = log(O_{ij})$, que representa o logaritmo da chances *a posteriori* [1] do gene i ser diferencialmente expresso tendo em conta o contraste β_{ij}, onde

$$O_{ij} = \frac{p(\beta_{ij} \neq 0|\widetilde{t}_{ij}, s_i^2)}{p(\beta_{ij} = 0|\widetilde{t}_{ij}, s_i^2)}.$$

Mostra ainda, usando o Teorema de Bayes que

$$O_{ij} = \frac{p_j}{1 - p_j}\left(\frac{v_{ij}}{v_{ij} + v_{0j}}\right)^{\frac{1}{2}}\left(\frac{\widetilde{t}_{ij}^2 + d_0 + d_i}{\widetilde{t}_{ij}^2\frac{v_{ij}}{v_{ij}+v_{0j}} + d_0 + d_i}\right)^{\frac{1+d_0+d_i}{2}}.$$

Tendo em conta a definição de O_{ij}, uma estimativa para a probabilidade de que β_{ij} seja diferente de zero é então

[1] Do inglês *posterior odds* [22]

$$\widehat{p_j} = \frac{1}{G}\sum_{i=1}^{G}\frac{O_{ij}}{1+O_{ij}}$$

onde G representa o número total de genes e $\frac{O_{ij}}{1+O_{ij}}$ a probabilidade estimada de que o gene i seja diferencialmente expresso.

A estimativa para p_j será provávelmente muito sensível à forma particular da distribuição *a priori* para β_{ij} e possívelmente também da dependência entre os genes. Uma forma de ultrapassar este problema será definir uma probabilidade esperada de que um gene seja diferencialmente expresso. Essa probabilidade fica ao cargo do utilizador e poderá assumir valores como 0.01, ou outros valores pequenos. No pacote limma, por defeito é assumida a probabilidade $p_j = 0.01$.

O seguinte exemplo permite ilustrar a importância da estatística B_{ij}.

Exemplo 4.2.3. *Suponhamos por exemplo que $B_{ij} = 2$. As chances de expressão diferencial para o gene i são dadas por $exp(B_{ij}) = exp(2) = 7.39$, o que significa que a probabilidade do gene i ser diferencialmente expresso é 7.39 vezes maior que a probabilidade de não ser. Neste sentido, a probabilidade de que este gene i seja diferencialmente expresso tendo em conta o contraste j é $\frac{O_{ij}}{1+O_{ij}} = \frac{7.39}{1+7.39} = 0.88$, ou seja, o probabilidade do gene i ser diferencialmente expresso é de 88%.*

Estimação dos Hiperparâmetros

As estatíticas B_{ij} e $\widetilde{t}_{ij}$ dependem no entanto dos hiperparâmetros d_0, s_0 e v_{0j}, que serão estimados com base nas observações. Smith, em [22] propõe estimar d_0 e s_0^2 a partir de s_i^2, estimando depois v_{0j} a partir da estatistica de teste $\widetilde{t}_{ij}$ assumindo que d_0 e p_j são conhecidos.

Para a estimação de d_0 e s_0^2 o autor propõe usar o método dos momentos onde em vez de considerar s_0^2 considera $log(s_0^2)$ sob as vantagens dos respectivos momentos serem finitos para qualquer valor dos graus de liberdade e a sua distribuição ser mais próxima da normal. Deste modo, os cálculos são facilitados e o método dos momentos funciona de forma mais eficiente. Como resultado as estimativas para d_0 e s_0^2, podem ser obtidas resolvendo o seguinte sistema de equações não linear.

$$\psi'\left(\frac{d_0}{2}\right) = m\acute{e}dia\left[(e_i-\overline{e})^2\, G/(G-1) - \psi'\left(d_i/2\right)\right]$$

e

$$s_0^2 = exp\left[\overline{e} + \psi\left(\frac{d_0}{2}\right) - log\left(\frac{d_0}{2}\right)\right]$$

onde G, representa o total de genes existentes, ψ e ψ' as funções digama (derivada logaritmica da função gama) e trigama (derivada da função digama), respectivamente.

O parâmetro v_{0j} é estimado igualando o valor p para cada observação da estatística de teste ($|\widetilde{t}_{ij}|$) ao seu valor nominal dada a posição do gene i. Assim, para cada gene em

particular e para cada ordem r será necessário resolver

$$2F(-|\widetilde{t}_{ij}|; v_{ij}, v_{0j}, d_{0j} + d_{ij}) = \frac{r - 0.5}{G} \tag{4.12}$$

cujo valor de v_{0j} que resolve (4.12) é

$$v_{0j} = v_{ij} \left(\frac{\widetilde{t}_{ij}^2}{q^2} - 1 \right)$$

com

$$q = F^{-1}(p; d_0 + d_i),$$

$$p = \frac{1}{p_j} \left(\frac{r - 0.5}{2G} - (1 - p_j) F(-|\widetilde{t}_j|; d_0 + d_i) \right)$$

sendo $0 < p < 1$, $q \leq |\widetilde{t}_{ij}|$, $F(\cdot; k)$ a função distribuição t-Student e r a posição do gene i quando $|\widetilde{t}_{ij}|$ estão ordenadas por ordem decrescente.

O estimador final é dado pela média dos estimadores decorrentes das $\frac{Gp_j}{2}$ maiores estatísticas ordinais.

A Estatística F

Tal como a forma quadrática das estatísticas t ordinárias seguem distribuições de *Fisher*, também apropriadas formas quadráticas das estatística t moderadas conduzem a distribuições de *Fisher*.

Suponha-se que se pretende testar se para um dado gene i todos os contrastes são nulos, isto é, $H_0 : \beta_i = 0$. A matriz de correlação para $\widehat{\beta}_i$ é dada por $R_i = U_i^{-1} C^T V_i C U_i^{-1}$, onde U_i é a matriz diagonal dos desvios padrão não escalonados $(v_{ij})^{\frac{1}{2}}$. Seja r o número de colunas de C, e Q_i uma matriz tal que $Q_i R_i Q_i = I_r$. Seja ainda $\boldsymbol{q}_i = Q_i^T \boldsymbol{t}_i$, onde $\boldsymbol{t}_i$ representa a estatística t observada para o gene i, então

$$F_i = \frac{\boldsymbol{q}_i^T \boldsymbol{q}_i}{r} = \frac{\boldsymbol{t}_i^T Q_i Q_i^T \boldsymbol{t_i}}{r} \sim F_{r, d_0 + d_i} \text{ [22].}$$

4.2.4 Análise Experimental

Base de Dados `ApoAI`

A estrutura da experiência com a base de dados pode ser esquematizada como na Figura 4.2. O contraste de interesse será (WT *vs* KO). Tomando este contraste como um

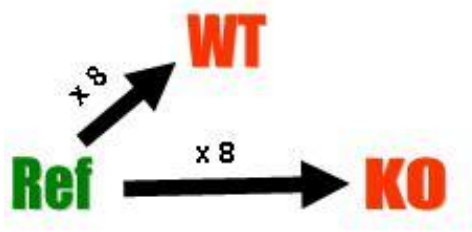

Figura 4.2: Delineamento da experiência relativamente à base de dados `ApoAI`. Uma referência comum (etiquetada a verde) foi hibridada com as sondas vindas dos ratos cujo gene ApoAI estava em condições normais de funcionamento ("`WT`"etiquetado a vermelho) e com as sondas vindas dos ratos cujo gene ApoAI estava deficiente ("`KO`"etiquetado a vermelho). Para cada hibridação foram consideradas 8 amostras, dando origem a oito *microarrays* distindos, prefazendo um total de 16 *microarrays*.

dos coeficientes, o vector de coeficientes poderá ser o seguinte:

$$\boldsymbol{\alpha}_i = \begin{bmatrix} (WT - Ref) \\ (KO - WT) \end{bmatrix}$$

Nestas condições, a matriz de delineamento poderá ser obtida recorrendo ao comando

```
design <- cbind("WT-Ref"=1,"KO-WT"=rep(0:1,c(8,8))).
```

e que define a seguinte matriz:

$$X = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}.$$

O ajustamento do modelo aos dados é feito da seguinte forma:

```
fit <- lmFit(MA,design=design)
```

onde MA, representa a matriz obtida com o comando `normalizeWithinArrays` referido em (2.2.2). Ajustado o modelo, procede-se ao cálculo das estatísticas de teste usando a abordagem de Bayes empírica. O cálculo das estatísticas de teste para cada gene é obtido com o comando:

```
fit <- eBayes(fit).
```

Um volcano plot, permite fornecer uma ideia resumida dos resultados produzidos, pelo que constitui uma ferramenta bastante útil. Este gráfico contrasta os $log_2(foldchange)$ com $-log_{10}(valor\ p)$. Assim será necessário especificar o uso desses valores produzidos pelo comando `topTable`, pedindo ainda para que sejam considerados todos os genes, que corresponde ao número de linhas da matriz M. O volcano plot da Figura 4.3 é produzido usando os comandos

```
y<- topTable(fit,coef="KO-WT", number=nrow(MA$M), adjust="fdr")
names(y)
plot(y[,9],-log(y[,12],12),xlab="log2(Fold-Change)",ylab="-log10(P.Value)"
,main="Volcano plot",cex=0.2,pch=19)
abline(v=c(-1,1),col="blue")
abline(h=-log(0.01,10),col="red")
```

Um gráfico quantil-quantil é outra ferramenta muito útil para resumir os resultados produzidos averiguando o ajustamento dos valores das estatísticas t_{ij} produzidos à distribuição t-Student. Para a contrução desse gráfico é considerada a média dos graus de liberdade dos genes, $\sum_{i=1}^{G} \frac{t_i}{G}$. Efectuando

```
qqt(fit$t[,2],df=fit$df.residual+fit$df.rior,main = "Gráfico de Quantis",
xlab = "quantis teóricos", ylab = "quantis observados")
abline(0,1)
title(sub = "Estatísticas de teste t vs t-student")
```

é possível obter o gráfico da Figura 4.4. Nesse gráfico aparentemente oito dos valores observados das estatísticas de teste afastam-se claramente da t-Student, sugerindo a não diferencialidade dos genes respeitantes a essas estatísticas.

Estando presentes testes simultâneos para aferir sobre a diferencialidade ou não diferencialidade dos genes, será necessário optar por uma medida de controlo do erro, como já discutido na secção 2.3 desta dissertação. Como medida de controlo dos falsos positivos, considerou-se a *false discovery rate*, mencionada na mesma secção. Outras medidas de controlo poderiam ser usadas, entre estas estão a correcção de bonferroni, o de método de holm [3] e o de hochberg [2]. No entanto, dado os bons resultados que a FDR tem

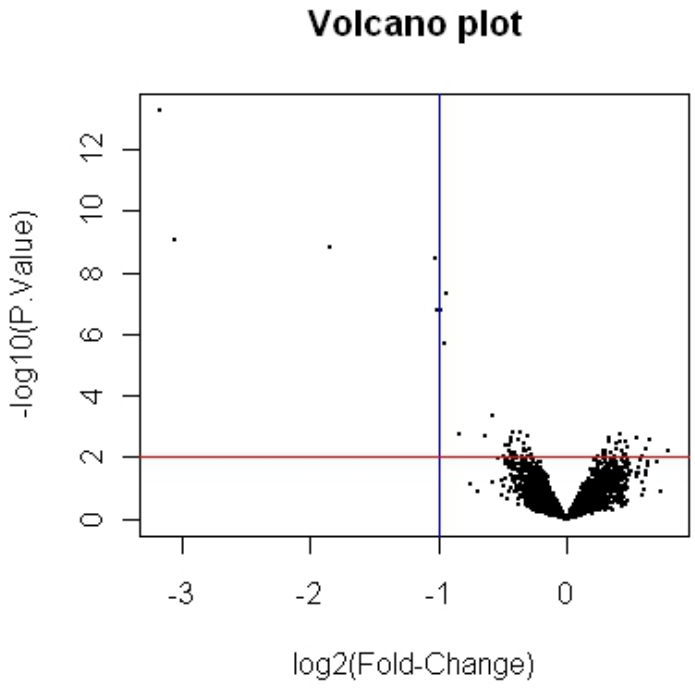

Figura 4.3: Volcano plot para a base de dados `ApoAI`. Os pontos a cima da linha vermelha, representam os genes cujo valor p é inferior a 0.01, ou seja, os genes que apresentam alterações dos níveis de expressão a um nível de significância de 1%. Os genes à esquerda da linha azul correspondem aos genes que expressam mais do que duas vezes o seu nível de expressão numa classe relativamente à outra classe.

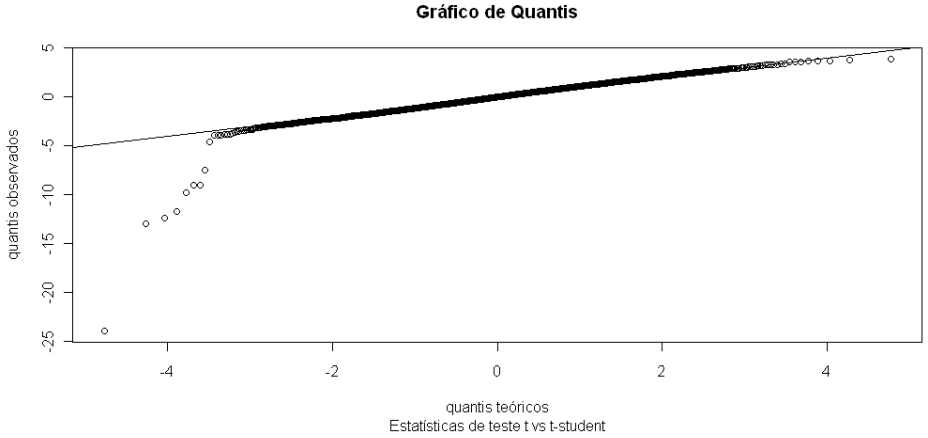

Figura 4.4: Gráfico de quantis para as estatísticas de teste t_{ij} produzidas pelo comando `eBayes` e distribuição t-Student com $\overline{d}$ graus de liberdade.

mostrado para estes *microarrays* em particular, optou-se por seguir com esta medida de controlo. A tabela dos genes diferencialmente expressos pode ser obtida usando o comando

```
topTable(fit,coef="KO-WT",adjust="fdr",p.value=0.01)
```

onde `coef` representa o contraste de interesse e `p.value` o nível de significância a que um gene é tido como diferencialmente expresso. Este comando permite obter informação descrita na Tabela 4.1. Da tabela é possível concluir que a um nível de significância de 0.01,

Linha	**Nome**	**Estatística t**	**Estatística B**	**Valor q**
2149	ApoAI,lipid-Img	-23,98	14,92663	3,05E-11
540	EST,HighlysimilartoA	-12,9621	10,81334	5,02E-07
5356	CATECHOLO-METHYLTRAN	-12,4408	10,44794	6,51E-07
4139	EST,WeaklysimilartoC	-11,7562	9,928528	1,21E-06
1739	ApoCIII,lipid-Img	-9,83535	8,19241	1,56E-05
2537	ESTs,Highlysimilarto	-9,01524	7,304899	4,22E-05
1496	est	-9,00232	7,290134	4,22E-05
4941	similartoyeaststerol	-7,44133	5,310665	0,000562

Tabela 4.1: Genes detectados como diferencialmente expressos através do ajustamento a um modelo linear usando a abordagem de Bayes empírica implementada no pacote `limma`. A primeira coluna representa a linha que identifica o gene na base de dados, a segunda coluna a designação do gene, a terceira coluna o valor da estatística de teste t moderada, a quarta coluna o valor da estatística B_{ij}, e por fim a última coluna representa o valor q. Tal como no valor p, estipulando um controlo do fdr ao nível de 0.01, todos os genes cujo valor q é inferior a este valor, são considerados diferencialmente expressos.

o modelo detecta oito genes diferencialmente expressos, o que vem confirmar as suspeitas obtidas aquando da análise do gráfico de quantis da Figura 4.4. A Figura 4.5 permite confirmar que os genes detectados como diferencialmente expressos tem de facto comportamentos bem distintos para as duas condições (ApoAI deficiente e normal). Quando comparando os resultados obtidos, com os obtidos através da aplicação da SAM (ver Tabela 3.1) verifica-se que o último conseguiu detectar mais quatro genes diferencialmente expressos. A Figura 4.6 mostra como são as distribuições das observações destes quatro genes para ambas as condições. Verifica-se uma certa aproximação das distribuições quando comparadas com as dos outros oito genes suspeitando-se assim de uma menor sensibilidade do modelo linear na detecção de genes diferencialmente expressos.

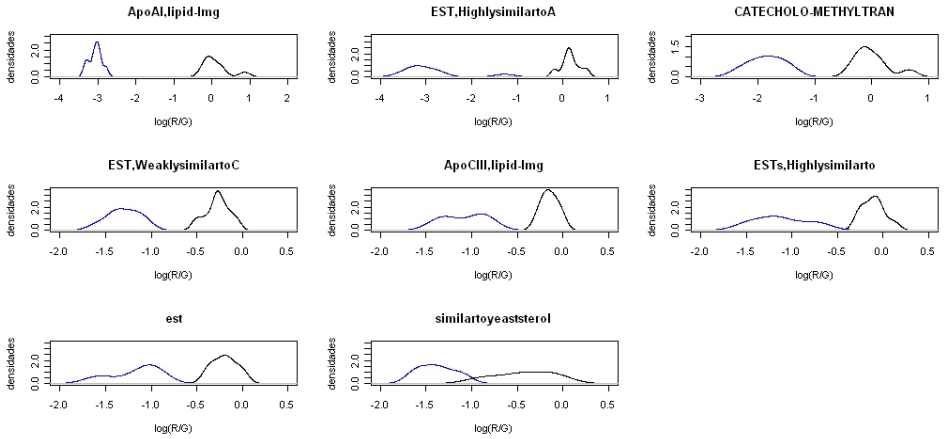

Figura 4.5: Densidades dos genes detectados como diferencialmente expressos. A curva a preto representa a densidade estimada para as observações com o gene ApoAI normal, e a curva a azul para as observações com o gene ApoAI deficiente. As densidades foram estimadas usando o comando `density` do pacote `stats` implementado em linguagem R.

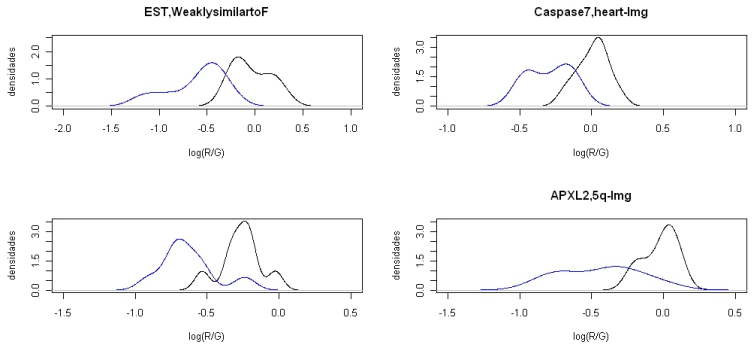

Figura 4.6: Densidades dos genes detectados como diferencialmente expressos pela SAM que não foram detectados usando o modelo linear. A curva a preto representa a densidade estimada para as observações com o gene ApoAI normal, e a curva a azul para as observações com o gene ApoAI deficiente. As densidades foram estimadas usando o comando `density` do pacote `stats` implementado em linguagem R.

Base de Dados `Fermentation`

Fixando um tempo específico, de entre os seis tempos considerados, o planeamento da experiência com a base de dados `Fermentation` pode ser esquematizada como na Figura 4.7. Para os restantes tempos o delineamento foi realizado de forma similar.

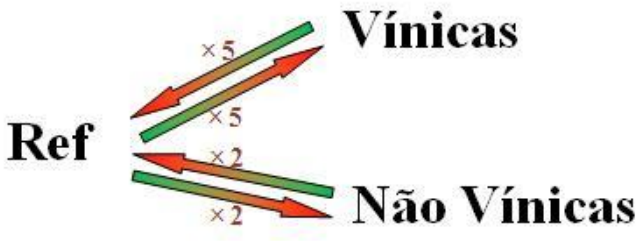

Figura 4.7: Delineamento da experiência relativamente à base de dados `Fermentation`. Uma referência comum (etiquetada a verde) foi hibridada com 5 sondas vindas de leveduras vínicas (etiquetadas a vermelho) e com 2 sondas vindas das leveduras não vínicas (etiquetadas a vermelho). O mesmo procedimento foi efectuado uma segunda vez, trocando os fluóforos (*dye swap*) e está representado pela seta em sentido contrário.

Relembrando que para estes dados as hibridações *dye swap* já estão invertidas, essas podem ser comparadas directamente com os outros *microarrays* (ver Secção 3.1.2). Por outro lado, pretendendo fazer comparações entre leveduras vínicas e leveduras não vínicas, leveduras vínicas e a levedura clínica e ainda comparações entre as leveduras vínicas e a laboratorial, as experiências de *microarrays* serão analizadas segundo três modelos lineares distintos, de acordo com os esquemas da Figura 4.8.

Deste modo, para o confronto entre vínicas e não vínicas, considerando o vector de coeficientes

$$\boldsymbol{\alpha_i} = \begin{bmatrix} \textit{Vínicas-Ref} \\ \textit{Vínicas-não Vínicas} \end{bmatrix}$$

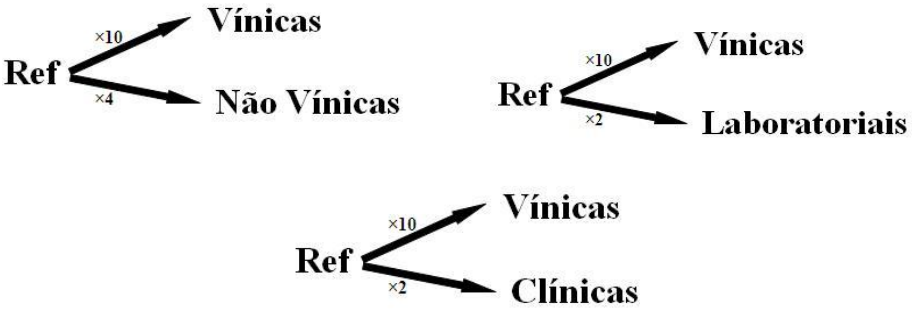

Figura 4.8: Esquematização das três abordagens consideradas para a análise dos dados da base de dados `Fermentation`. Na primeira imagem (superior esquerda), a levedura clínica e laboratorial são consideradas pertencentes a uma mesma classe (não vínicas) sendo efectuada uma comparação entre as leveduras vínicas e não vínicas. Deste modo considera-se a existência de 10 *microarrays* para a classe das leveduras vínicas e 4 *microarrays* para a classe das leveduras não vínicas (2 *microarrays* com observações de uma levedura clínica e 2 *microarrays* com observações de uma levedura laboratorial). Na segunda imagem (superior direita), é atribuida uma classe às leveduras laboratoriais, sendo feita a comparação dos níveis de expressão das leveduras vínicas com os níveis de expressão da levedura laboratorial, a classe das leveduras vínicas com 10 *microarrays* e a classe das leveduras laboratoriais com 2 *microarrays*. Na terceira imagem (inferior) é feita a comparação dos níveis de expressão das leveduras vínicas com a clínica, sendo que a classe das leveduras vínicas é constituida por 10 *microarrays* e a classe das leveduras clínicas constituida por 2 *microarrays*.

---

a matrix de delineamento será a seguinte:

$$X = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & -1 \\ 1 & -1 \\ 1 & -1 \\ 1 & -1 \end{bmatrix}$$

Para o confronto entre vínicas e clínica, considerando o vector de coeficientes

$$\boldsymbol{\alpha_i} = \begin{bmatrix} \textit{Vínicas-Ref} \\ \textit{Vínicas-Clínica} \end{bmatrix}$$

a matrix de delineamento será:

$$X = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & -1 \\ 1 & -1 \end{bmatrix} \tag{4.13}$$

E para o confronto entre vínicas e laboratorial, considerando o vector de coeficientes

$$\boldsymbol{\alpha_i} = \begin{bmatrix} \textit{Vínicas-Ref} \\ \textit{Vínicas-Laboratorial} \end{bmatrix}$$

a matrix de delineamento será (4.13).

Recorrendo aos comandos `lmfit` e `eBayes` é feito o ajustamento do modelo para as três análises. A Figura 4.9 permite visualizar a proporção de genes diferencialmente expressos e se a diferença das médias para esses genes são muito elevadas. De uma forma geral, os gráfico levam a crer que o número de genes diferencialmente expressos (genes situados acima da recta a vermelho) não variam muito para os três tempos considerados, nem para as três análises em estudo. A grande quantidade de observações encontradas acima da linha a vermelho à direita da linha azul $log_2(foldchange) = 1$ e à esquerda da linha azul $log_2(foldchange) = -1$ sugere a existência de uma grande quantidade de genes cujos níveis de expressão alteram em mais do que duas vezes quando pertencentes classes distintas.

A Figura 4.10 representa o ajustamento dos valores observados das estatísticas de teste à t-Student, sendo portanto uma outra medida para sumariar os resultados obtidos e fornecer uma ideia da proporção de genes diferencialmente expressos. A Figura 4.10 aparentemente confirma as suspeitas contruidas em torno dos volcano plots, sobre a proporção de genes não variar muito entre as três fases de crescimento (Tempo 2, 3 e 5) e entre o tipo de análise que é considerada (vínicas *vs* não vínicas, vínicas *vs* clínica ou vínicas *vs* laboratorial), isto é, o número de observações que não se ajusta à recta $Y = x$, por observação, não parece alterar muito para os 9 gráficos.

Determinou-se ainda o número exacto de genes diferencialmente expressos quando considerando um valor q igual a 0.05 e verificou-se o número de genes concordantes com os obtidos com a SAM para cada um dos nove casos. A Tabela 4.2 permite verificar que a grande maioria dos genes detectados com o Modelo Linear também são detectados com a SAM. Com efeito, à excepção do tempo 3 quando confrontando as vínicas com a laboratorial, o número de genes tidos como diferencialmente expressos e o número de genes coincidentes com os obtidos com a SAM estão muito próximos.

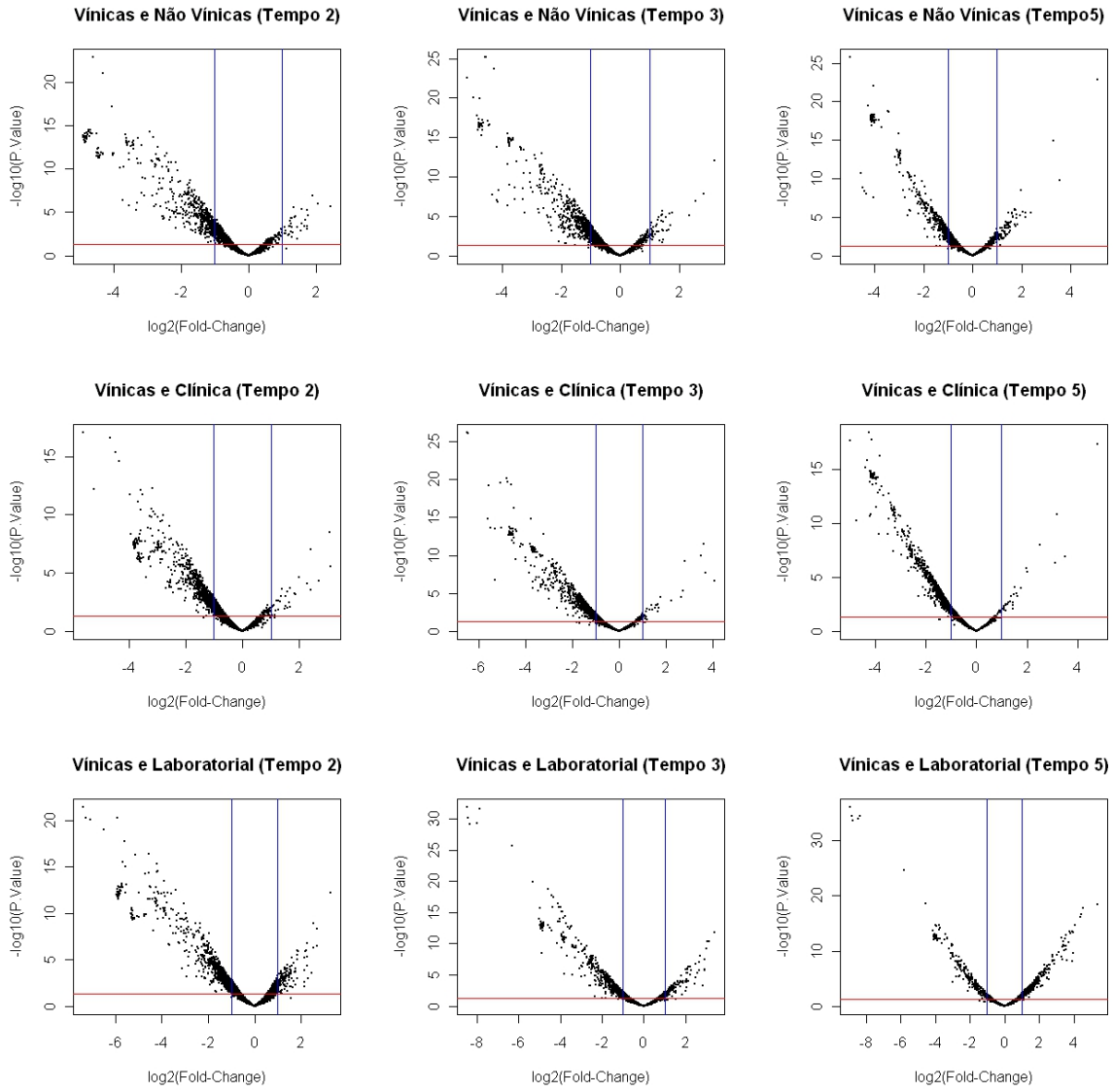

Figura 4.9: Volcano plot para a base de dados `Fermentation`. Os pontos acima da linha vermelha representam os genes cujo valor p é inferior a 0.05, ou seja, os genes que apresentam alterações dos níveis de expressão a um nível de significância de 5%. Os genes acima da linha vermelha e à esquerda da linha azul $log_2(foldchange) = -1$ e à direita da linha azul $log_2(foldchange) = 1$ correspondem aos genes diferencialmente expressos cuja expressão numa classe é superior ao dobro da expressão na outra classe.

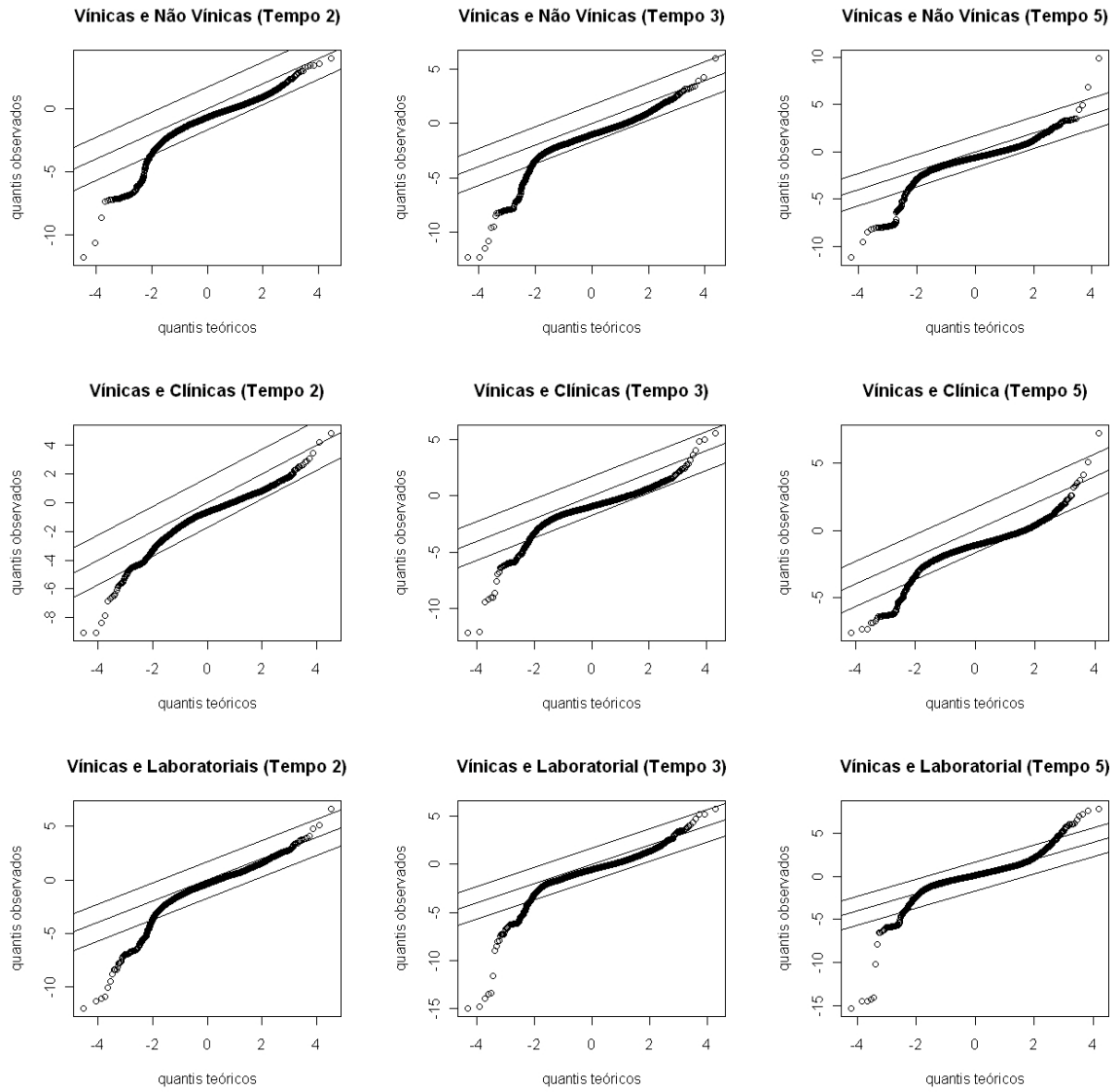

Figura 4.10: Gráficos de quantis para os 9 casos em estudo (três tempos para cada uma das análises) com a base de dados `Fermentation`. Para a obtenção da figura foi considerado um valor q de 0.05 e representado pelas linhas paralelas à recta $Y = x$. Os genes situados fora da banda limitada pelas duas linhas são tidos como genes significantes.

Vínicas e Não Vínicas		
Tempo	*Nº genes DE*	*Nº genes SAM*
2	185	183
3	164	162
5	119	118

Vínicas e Clínica		
Tempo	*Nº genes DE*	*Nº genes SAM*
2	110	108
3	157	146
5	165	110

Vínicas e Laboratorial		
Tempo	*Nº genes DE*	*Nº genes SAM*
2	193	187
3	160	154
5	145	94

Tabela 4.2: Número de genes diferencialmente expressos obtidos com o Modelo Linear para a base de dados `Fermentation` e número de genes concordantes com a SAM. A primeira coluna refere-se ao tempo de estudo, a segunda ao número de genes diferencialmente expressos que foram obtidos com o Modelo Linear e a terceira ao número de genes diferencialmente expressos comuns com os obtidos com a SAM. As primeiras quatro linha referem-se ao confronto das leveduras vínicas com as não vínicas, as quatro seguintes ao confronto das leveduras vínicas com a clínica e por fim os resultados obtidos quando confrontadas as leveduras vínicas com a laboratorial.

Como já referido será objectivo do biólogo verificar quais os genes detectados nas três análises em cada tempo (Tabela 4.3), mas mais que isso, saber quais os genes que mantêm acordo nas duas metodologias (SAM e modelo linear), pelo que se obteve o número de genes coincidentes com os genes comuns às três análises detectados pela SAM (ver Tabela 3.12). A Tabela 4.4 permite verificar que a grande maioria dos genes comuns às três análises obtidos com a SAM também são obtidos com o Modelo Linear. Assim, à partida tudo indica que estes serão os genes que deverão ter maior atenção por parte do biólogo.

Genes DE Comuns		
Tempo 2	*Tempo 3*	*Tempo 5*
83	99	68

Tabela 4.3: Genes diferencialmente expressos comuns nas três análise obtidos com o Modelo Linear. A primeira coluna é referente aos genes comuns nas três análises para o tempo 2, a segunda é referente aos genes comuns nas três análises para o tempo 3 e a terceira aos genes comuns nas três análises para o tempo 5.

Genes DE Comuns com a SAM		
Tempo 2	*Tempo 3*	*Tempo 5*
60	80	37

Tabela 4.4: Número de genes comuns às três análises obtidos com o Modelo Linear, coincidentes com os genes comuns às três análises obtidos com a SAM. A primeira coluna refere-se aos genes comuns para o tempo 2, a segunda para o tempo 3 e a última para o tempo 5.

4.3 EBarrays

O `EBarrays` é um pacote desenvolvido na linguagem R que implementa três modelos de mistura usando uma abordagem de Bayes empírica para a detecção de genes diferencialmente expressos. Assim sendo, estas metodologias têm como base o conceito de Bayes empírico já descrito na Secção 4.1 sendo que as diferenças entre elas residem essencialmente nas distribuições assumidas. Os modelos apresentados no pacote `EBarrays` são [16, 18]: Gama-Gama (GG), Lognormal-Normal (LNN) e Lognormal-Normal com variância modificada (LNNMV).

4.3.1 O Modelo Hierárquico Geral

O objectivo das metodologias GG, LNN e LNNVM será ajustar uma distribuição de probabilidade ao conjunto dos dados proveniente da medição dos níveis de expressão genética dos diferentes genes em estudo. O número de padrões que se procura varia consoante o número de condições a que as células foram sujeitas. Se estas apenas forem sujeitas a duas condições, então apenas existem dois padrões distintos por gene, o gene ser diferencialmente expresso ou equitativamente expresso. Se três condições estiverem em estudo, então o número de padrões distintos por gene aumenta para cinco: o gene ser equitativamente expresso para todas as três condições (que equivale a um padrão), existir expressão diferencial apenas numa das condições (que equivale a três padrões), ou existir expressão diferencial em todas as condições (que equivale a um outro padrão).

Segundo [16], supondo que existem n amostras (*microarrays*) sendo $\boldsymbol{y}_i = (y_{i1}, y_{i2}, ..., y_{in})$ as medidas para o gene i nas n observações e considerando a existência de $m+1$ condições diferentes, então para cada padrão k as amostras experimentais $S = \{1, 2, ..., n\}$ são particionadas em $r(k)$ conjuntos disjuntos $\{S_{jk}, j = 1, 2, ..., r(k)\}$, onde as observações pertencentes ao grupo S_{jk} têm a mesma média. A hipótese nula independentemente do número de condições em estudo é a equidade nos níveis de expressão, neste caso todas as amostras terão a mesma média dos níveis de expressão e virão de uma distribuição conjunta $f_0(\boldsymbol{x}_i)$. Se a hipótese nula não se verifica então existirão diferenças nos níveis de expressão em pelo menos uma condição. Neste caso a distribuição conjunta do padrão k é dada por $f_k(\boldsymbol{x}_i)$, com $k \neq 0$ e

$$f_k(\boldsymbol{x}_i) = \prod_{j=1}^{r(k)} f(\boldsymbol{x}_{iS_{jk}}) \tag{4.14}$$

onde $f(\boldsymbol{x}_{iS_{jk}})$ é a distribuição dos dados associados ao subconjunto S_{jk}. À partida não se conhece a probablidade de um dado padrão k, sendo portanto necessária a introdução das quantidades desconhecidas p_k. A distribuição marginal dos dados é então obtida pela mistura das distribuições conjuntas com pesos p_k do seguinte modo:

$$\sum_{k=0}^{m} p_k f_k(\boldsymbol{x}_i)$$

Do teorema de Bayes vem que a probabilidade *a posteriori* é

$$p(k|\boldsymbol{x}_i) \propto p_k f_k(\boldsymbol{x}_i) \tag{4.15}$$

sendo as chances *a posteriori* definidas da seguinte forma:

$$O_{ik} = \frac{p(k|\boldsymbol{x}_i)}{1 - p(k|\boldsymbol{x}_i)}.$$

Assume-se que as medidas que partilham de uma mesma média comum μ_i são independentes e identicamente distribuidas com uma distribuição observada $f_{obs}(\cdot|\mu_i)$ e que μ_i provém de uma distribuição $\pi(\mu_i)$ que representa as flutuações na média dos níveis de expressão. Pelo teorema da probabilidade total, a distribuição dos dados associados ao subconjunto S_{jk} pode ser dada como

$$f(\boldsymbol{x}_{iS_{jk}}) = \int \left(\prod_{s \in S_{jk}} f_{obs}(x_{is}|\mu_i) \right) \pi(\mu_i)\, d\mu_i. \tag{4.16}$$

Cada probabilidade *a posteriori* determina a probabilidade de um gene ser diferencialmente expresso, o cálculo desta depende das probabilidades *a priori* e da distribuição dos dados. No pacote `EBarrays` encontra-se implementado um algoritmo de identificação de genes diferencialmente expressos com base em modelos de mistura onde são assumidos diferentes tipos de distribuições para π e f_{obs}. O algoritmo calcula as probabilidades *a posteriori* para identificar genes diferencialmente expressos em pelo menos uma condição, ordená-los sob determinada condição, classificá-los quanto à expressão diferencial e estimar a FDR.

A FDR é estimada recorrendo ao método de Yoav Benjamini e Daniel Hochberg, referido em 2.3. O ajustamento do modelo de Bayes empírico é feito recorrendo ao algoritmo EM. De seguida segue uma breve descrição de cada um dos modelos implementados no pacote `EBarrays`. Neste pacote são assumidos modelos específicos para o quociente das intensidades. Um maior detalhe sobre estes modelos pode ser consultado em [18, 17, 16].

Modelo Gama-Gama

No modelo Gama-Gama (GG) assume-se que a variável aleatória $\widetilde{X}$, que representa o quociente $\frac{R}{G}$, segue uma distribuição Gama com parâmetro de forma $\alpha > 0$ e média μ_i, pelo que o parâmetro de escala é dado por $\lambda_i = \frac{\alpha}{\mu_i}$. Mais especificamente, considera-se a seguinte função densidade de probabilidade para a componente observada:

$$f_{obs}(\widetilde{x}_{ij}|\mu_i) = \frac{\lambda_i^{\alpha} \widetilde{x}_{ij}^{\alpha-1} exp\{-\lambda_i \widetilde{x}_{ij}\}}{\Gamma(\alpha)}, para\ \widetilde{x}_{ij} > 0.$$

É de notar que se para um gene i, a v.a. $\widetilde{X}$ segue uma distribuição $Gama(\alpha, \lambda_i)$ então o seu coeficiente de variação é dado por

$$c_v = \frac{Var[\widetilde{X}]}{\sqrt{E[\widetilde{X}]}} = \frac{\sqrt{\frac{\alpha}{\lambda_i^2}}}{\frac{\alpha}{\lambda_i}} = \frac{1}{\sqrt{\alpha}},$$

independentemente de i, pelo que se o modelo Gama proposto se ajusta aos valores observados de $\widetilde{X} = \frac{R}{G}$, é de esperar que todos os genes tenham o mesmo valor para o coeficiente de variação.

Para a distribuição marginal $\pi(\mu_i)$ assume-se uma distribuição Gama inversa onde, fixando α, $\lambda_i = \frac{\alpha}{\mu_i}$ segue uma distribuição Gama com parâmetro de forma α_0 e parâmetro de escala ν.

Deste modo três parâmetros são envolvidos $\theta = (\alpha, \alpha_0, \nu)$. Estes serão posteriormente estimados usando o algoritmo EM, apresentado na Secção 4.1.1. Tendo em conta a fórmula (4.16) vem,

$$f(\boldsymbol{x_i}) = \frac{\nu^{\alpha_0}\Gamma(n\alpha + \alpha_0)}{\Gamma^n(\alpha)\Gamma(\alpha_0)} \frac{\left(\prod_{j=1}^n x_{ij}\right)^{\alpha-1}}{\left(\nu + \sum_{j=1}^n x_{ij}\right)^{n\alpha+\alpha_0}}$$

onde n corresponde ao número de observações da condição em causa.

Por (4.14) é então possível calcular a probabilidade *a posteriori* para qualquer padrão de expressão.

Particularizando, para o caso em que apenas duas condições estão em estudo, as chances *a posteriori* de que um gene i seja diferencialmente expresso vem dado por:

$$\begin{aligned} O_i &= \frac{p}{1-p}\frac{f_1(\boldsymbol{x}_i)}{f_0(\boldsymbol{x}_i)} \\ &= \frac{p}{1-p}K' \frac{\left(\sum_{j=1}^{n_1} x_{i,j} + \sum_{j=1}^{n_2} y_{i,j} + \nu\right)^{N\alpha+\alpha_0}}{\left(\sum_{j=1}^{n_1} x_{i,j} + \nu\right)^{n_1\alpha+\alpha_0}\left(\sum_{j=1}^{n_2} y_{i,j} + \nu\right)^{n_2\alpha+\alpha_0}} \end{aligned} \tag{4.17}$$

onde

$$K' = \frac{\nu^{\alpha_0}\Gamma(n_1\alpha + \alpha_0)\Gamma(n_2\alpha + \alpha_0)}{\Gamma(\alpha_0)\Gamma(N\alpha + \alpha_0)},$$

$N = n_1 + n_2$, n_1 o número de observações na condição 1 e n_2 o número de observações na condição 2.

Modelo Lognormal Normal

No modelo Lognormal Normal (LNN), μ_i representa a média dos logaritmos das intensidades de expressão $x_{ij} = log(\widetilde{x}_{ij})$ que se assume ajustadas a uma distribuição normal

com uma variância comum σ^2. Tal como no modelo Gama-Gama, também o coeficiente de variação se mantém constante ao longo dos genes. De facto, para este modelo vem

$$c_v = \frac{Var[X]}{\sqrt{E[X]}} = \frac{exp\left\{\mu + \frac{1}{2}\sigma^2\right\}\sqrt{exp\left\{\sigma^2\right\} - 1}}{exp\left\{\mu + \frac{1}{2}\sigma^2\right\}} = \sqrt{exp\left\{\sigma^2\right\} - 1}.$$

A distribuição *a priori* de μ_i é assumida normal com média μ_0 e variância τ_0^2.

Da expressão (4.16) decorre que a densidade $f(.)$ é também uma normal com o vector de médias $\boldsymbol{\mu_0} = (\mu_0, \mu_0, ..., \mu_0)^t$ e matrix de covariâncias permutável[2]

$$\boldsymbol{\Sigma}_n = \sigma^2 \begin{bmatrix} 1 & 0 & ... & 0 \\ 0 & 1 & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 1 \end{bmatrix} + \tau_0^2 \begin{bmatrix} 1 & 1 & ... & 1 \\ 1 & 1 & ... & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & ... & 1 \end{bmatrix}$$

Mais precisamente $f(.)$ toma a forma

$$f(\boldsymbol{x_i}) = \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{|\boldsymbol{\Sigma_n}|}} exp\left\{-\frac{1}{2}\left(\boldsymbol{x}_i - \boldsymbol{\mu_0}\right)^t \boldsymbol{\Sigma_n^{-1}}\left(\boldsymbol{x}_i - \boldsymbol{\mu_0}\right)\right\}.$$

Deste modo três parâmetros são envolvidos: $\theta = (\mu_0, \sigma^2, \tau_0^2)$. Estes poderão ser estimados usando o algoritmo EM.

No caso particular de apenas duas condições, as chances de que um gene i seja diferencialmente expresso vem dado por:

$$O_i = \frac{p}{1-p}\frac{f_1(\boldsymbol{x}_i)}{f_0(\boldsymbol{x}_i)} = \frac{p}{1-p}\sqrt{\left(\frac{|\boldsymbol{\Sigma_n}|}{|\boldsymbol{\Sigma_*}|}\right) exp\left\{-\frac{1}{2}\boldsymbol{\delta}_i^T\left(\boldsymbol{\Sigma_*^{-1}} - \boldsymbol{\Sigma_n^{-1}}\right)\boldsymbol{\delta}_i\right\}}$$

onde $\boldsymbol{\delta}_i = (\boldsymbol{x}_i, \boldsymbol{y}_i)^T - \boldsymbol{\mu}_0$ representa os dados centrados, com $\boldsymbol{x}_i$ a representar os logaritmos das medidas numa condição e $\boldsymbol{y}_i$ os logaritmos das medidas na outra condição. $\boldsymbol{\Sigma_*}$ representa a matriz $n \times n$ diagonal por blocos, com a matriz $\boldsymbol{\Sigma_{n_1}}$ situada no bloco esquerdo superior e $\boldsymbol{\Sigma_{n_2}}$ no bloco direito inferior.

Modelo Lognormal-Normal com variância modificada

O modelo Lognormal-Normal com variância modificada (LNNMV) assume os mesmos pressupostos que o modelo Lognormal-Normal com a excepção de que, em vez de ser considerada uma variância comum ao longo dos genes, ter-se-á em conta que esta poderá modificar de gene para gene. Assim, em vez de σ^2 ter-se-á σ_i^2. Para σ_i^2 é obtida a sua estimativa *a posteriori* pelo que para este modelo apenas dois parâmetros $\theta = (\mu_0, \tau_0^2)$ são envolvidos. Para entender melhor como é feita a estimativa para σ_i^2, destaca-se o caso em que apenas existem duas condições.

[2]Do inglês *exchangeable covariance matrix*

Considerando μ_{i1} e μ_{i2}, as médias das condições 1 e 2 para um gene i, conhecidas, assume-se que σ_i^2 tem distribuição de qui-quadrado inversa escalonada com ν_0 graus de liberdade e parâmetro de escala σ_0. Pelo teorema de Bayes tem-se que a distribuição *a posteriori* para σ_i^2 é também uma distribuição de qui-quadrado inversa escalonada com $n_1 + n_2 + \nu_0$ graus de liberdade e parâmetro de escala

$$\sqrt{\frac{\nu_0\sigma_0^2 + \sum_{j=1}^{n_1}(x_{ij} - \mu_{i1})^2 + \sum_{j=1}^{n_2}(y_{ij} - \mu_{i2})^2}{n_1 + n_2 + \nu_0}},$$

onde x_{ij} representa o logaritmo das observações na condição 1, y_{ij} representa o logaritmo das observações na condição 2 e n_1 e n_2 o número de observações da condição 1 e 2, respectivamente. Se se tomar as variâncias aleatórias combinadas

$$\widetilde{\sigma}_i^2 = \frac{\sum_{j=1}^{n_1}(x_{ij} - \overline{x}_i)^2 + \sum_{j=1}^{n_2}(y_{ij} - \overline{y}_i)^2}{n_1 + n_2 - 2}$$

como uma amostra aleatória obtida da distribuição *a priori* de σ_i^2 então, pelo método dos momentos obtêm-se as estimativas para ν_0 e σ_0. Dadas as estimativas $\widehat{\nu}_0$ e $\widehat{\sigma}_0$ a estimativa para σ_i^2 resulta então da média *a posteriori*, dada por

$$\widehat{\sigma}_i^2 = \frac{\widehat{\nu}_0\widehat{\sigma}_0 + \sum_{j=1}^{n_1}(x_{ij} - \overline{x}_i)^2 + \sum_{j=1}^{n_2}(y_{ij} - \overline{y}_i)^2}{n_1 + n_2 + \widehat{\nu}_0 - 2}.$$

4.3.2 Análise Experimental

Base de dados `ApoAI`

Cada um dos três modelos serão agora aplicados à base de dados `ApoAI` com vista a analisar e identificar os genes diferencialmente expressos, usando o pacote EBarrays.

Relembrando que para esta base de dados o nível de expressão genética foi medido sobre duas condições, ambas com 8 *microarrays* cada, então apenas dois padrões são possíveis: diferencialmente expresso *versus* não diferencialmente expresso. Estes padrões podem ser denotados respectivamente pelos vectores padrão:

(1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1), representando a hipótese nula $\mu_{i1} = \mu_{i2}$.

(1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2), representando a hipótese alternativa $\mu_{i1} \neq \mu_{i2}$.

Recorrendo ao pacote `EBarrays`, definem-se aqueles padrões através do comando `ebPatterns`.

```
padroes<-ebPatterns(c("1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1",
"1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2")).
```

A base de dados a entrar no pacote `EBarrays` deve corresponder a uma matriz onde as linhas representam os genes e as colunas os *microarrays*. Os dados devem estar numa escala não logaritmica pelo que, para obter os dados tal como exigido pelo pacote efectuou-se o seguinte procedimento.

```
x<-RG$R/RG$G.
```

O método implementado no `EBarrays` não prevê a existência de valores omissos. Além disso, genes contendo pelo menos uma observação negativa são excluidos da análise. Esta última situação não é no entanto o caso das bases de dados analisadas nesta dissertação.

Dada a existência de valores omissos para a `ApoAI`, foi aplicada uma metodologia de imputação de valores. Optou-se pelo algoritmo k-nn (k vizinho mais próximo[3]) com $k = 10$. O algoritmo pode ser aplicado usando a função `pamr.knnimpute` do pacote `pamr`. Feita a imputação dos valores omissos é então possível determinar a distribuição dos dados sob cada padrão. Essa distribuição é obtida através do comando

```
fit<-emfit(dados, family="GG", hypotheses=padroes)
```

onde `dados` representa a matriz `x` dos quocientes com os valores omissos imputados, `family` o modelo a aplicar (neste caso é ilustrado para o modelo Gama-Gama, nos outros casos ter-se-ia `family`="LNN"ou `family`="LNNMV") e `hypotheses` as hipóteses a testar que, neste caso concreto, é não ser diferencialmente expresso *vs* ser diferencialmente expresso representadas pelo vector dos padrões `padroes`.

O passo seguinte será calcular as probabilidades *a posteriori*. O cálculo destas probabilidades é obtido fazendo

```
posteriori<-postprob(fit,dados)$pattern
```

onde `fit` representa o modelo ajustado no passo anterior. A seguir define-se um limite para as probabilidades *a posteriori* de forma a que a FDR seja controlada a um determinado nível desejado. Esse limite é determinado pelo valor obtido através do comando

```
threshold<-crit.fun(posteriori[,1],0.001)
```

onde o primeiro parâmetro é dado pelo valor da subtracção da probabilidade *a posteriori* dos genes serem diferencialmente expressos à unidade. Neste caso, dada a existência de apenas dois padrões, estas probabilidades são o mesmo que as probabilidades de não ser diferencialmente expresso e, portanto, as probabilidades *a posteriori* do primeiro padrão que é representado pelo vector `posteriori[,1]`. O segundo parâmetro refere-se ao nível de significância a que se pretende controlar a FDR.

[3]Do inglês *k-nearest neighbors*.

Os genes diferencialmente expressos serão identificados por possuirem probabilidades *a posteriori* superiores ao limite `threshold`. Uma forma de obter a quantidade de genes considerados diferencialmente expressos será fazer

```
sum(posteriori[,2]>threshold)
```

onde `posteriori[,2]` representa o vector das probabilidades *a posteriori* de ser diferencialmente expresso e `threshold` o limite calculado atrás. Para visualizar concretantente quais os genes seleccionados pelo modelo basta então percorrer todo o vector `posteriori[,2]` e verificar quais os índices em que as probabilidades ultrapassaram o `threshold`. As Tabelas 4.5 e 4.6 sumariam o número total de genes obtidos em cada um dos três modelos, e qualifica os genes seleccionados em comum pelos modelos dois a dois e para os três modelos simultâneamente.

GG	LNN	LNNVM
65	11	7

Tabela 4.5: Número de genes diferencialmente expressos obtidos com os três modelos usando o pacote EBarrays. Modelo Gama-Gama (GG), modelo Lognormal-Normal (LNN) e Lognormal-Normal com Variância Modificada(LNNVM).

Da Tabela 4.6 conclui-se que o modelo LNN concorda com todos os genes seleccionados pelo modelo LNNVM, e com apenas oito dos 65 genes seleccionados pelo modelo GG. Já o modelo LNNVM concorda apenas com seis dos genes seleccionados pelo GG. Em comum, os três modelos concordam na selecção de seis genes.

Comparando com os resultados obtidos através do modelo linear e da SAM (ver Tabelas 4.1 e 3.1), verifica-se que ambos concordam com os sete genes obtidos confrontando os resultados dos modelos LNN e LNNVM, e consequentemente com os seis genes seleccionados em comum pelos três modelos.

Um dos pressupostos para o ajustamento dos modelos GG e LNN aos dados é a verificação de um coeficiente de variação constante. Assim será importante verificar a relação entre a média dos quocientes e o coeficiente de variação. O comando `checkCCV` permite efectuar gráficamente essa comparação. O ajustamento é feito usando regressão polinomial, sendo que, no caso ideal o ajustamento aos dados deverá ser feito por uma recta horizontal, ou seja, os coeficientes que acompanham os termos cujo grau é superior a um deverão ser zero. O gráfico obtido para esta base está representado na Figura 4.11. Verifica-se que o ajustamento não é o desejado pois não se observa a aproximação a uma recta. Deste modo levantam-se incertezas quanto à validade dos modelos LNN e GG. Apesar disso, ambos os modelos aplicados conduzem a resultados bastante satisfatórios, coincidindo com os resultados dos outros métodos.

Uma outra forma de verificar o ajustamento do modelo paramétrico será através do gráfico de quantis para a Gama ou Normal consoante o modelo em causa. Note-se no

GG e LNN		GG e LNNVM	
Linha	**Nome**	**Linha**	**Nome**
540	EST,HighlysimilartoA	540	EST,HighlysimilartoA
799	Cy3RT	1496	est
1496	est	1739	ApoCIII,lipid-Img
1739	ApoCIII,lipid-Img	2149	ApoAI,lipid-Img
2149	ApoAI,lipid-Img	2537	ESTs,Highlysimilarto
2537	ESTs,Highlysimilarto	5356	CATECHOLO-METHYLTRAN
5356	CATECHOLO-METHYLTRAN		
5986	Cy3RT		

LNN e LNNVM		GG, LNN e LNNVM	
Linha	**Nome**	**Linha**	**Nome**
540	EST,HighlysimilartoA	540	EST,HighlysimilartoA
1496	est	1496	est
1739	ApoCIII,lipid-Img	1739	ApoCIII,lipid-Img
2149	ApoAI,lipid-Img	2149	ApoAI,lipid-Img
2537	ESTs,Highlysimilarto	2537	ESTs,Highlysimilarto
4139	EST,WeaklysimilartoC	5356	CATECHOLO-METHYLTRAN
5356	CATECHOLO-METHYLTRAN		

Tabela 4.6: Comparações dos genes seleccionados em comum pelos modelos dois a dois e pelos três modelos simultâneamente.

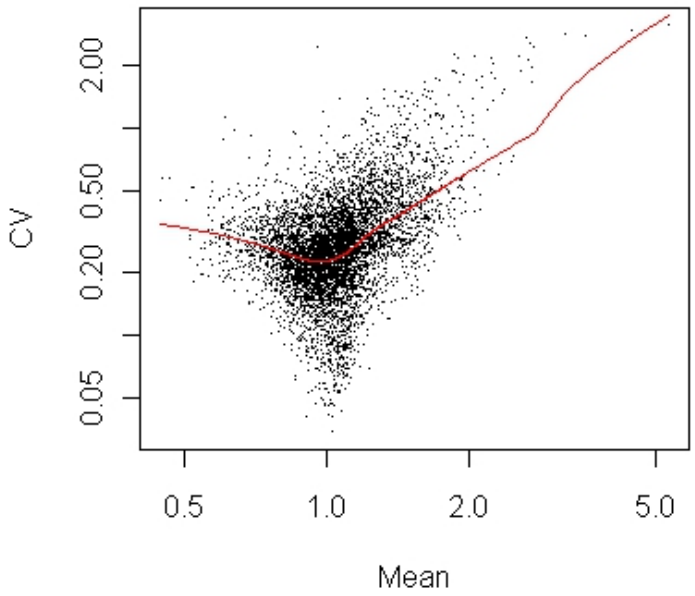

Figura 4.11: Gráfico verificando a relação entre os quocientes das intensidades e o coeficiente de variação para a base de dados `ApoAI`. À medida que a média das intensidades aumenta verifica-se um aumento do coeficiente de variação maior que o esperado. O caso ideal seria obter um coeficiente de variação constante para todos os valores da média das intensidades.

entanto, que cada gene terá uma distribuição cujos parâmetros variam de gene para gene. Assim, uma forma sumária de concentrar a representação dos gráficos de cada gene será efectuar agrupamentos em 9 subconjuntos de dados, construindo nove gráficos. Para cada gráfico é seleccionada uma média, sendo representados os genes cujas médias estejam situadas em torno dessa média. Este gráfico pode ser construido usando a função `checkModel` cujos parâmetros de entrada são a base de dados, o modelo ajustado, o modelo aplicado (GG, LNN ou LNNVM) e o número de genes a imprimir em cada gráfico. Os gráficos para os três modelos sobre a `ApoAI` estão representados na Figura 4.12.

Por observação da Figura 4.12 é de crer que o ajustamento é bastante razoável para qualquer um dos modelos. Quando comparando os três modelos, o modelo GG mostra um maior desajuste, o que poderá implicar uma detecção de um maior número de genes diferencialmente expressos por este modelo, confirmando assim os resultados da Tabela 4.5.

Para os modelos GG e LNN a distribuição marginal dos dados não depende das variâncias aleatórias dos genes, seguindo uma distribuição bem definida e portanto possível de representar. Estas distribuições marginais podem ser obtidas usando o comando `plotMarginal` tendo como parâmetros de entrada o modelo ajustado e a base de dados. Na Figura 4.13 podem ser encontrados os gráficos das marginais para os dois modelos referidos.

O modelo LNN assume que as variâncias amostrais seguem uma distribuição Qui-Quadrado inversa escalonada e pode ser avaliado gráficamente através da execução do comando `checkVarsQQ`. Este comando permite construir um gráfico de quantis contrastando os quantis das variâncias amostrais com os quantis da distribuição assumida. Para este gráfico é construido um histograma para as variâncias aleatórias e a curva de densidade da Qui-quadrado inversa escalonada. A Figura 4.14 permite averiguar gráficamente que, de uma forma geral, as variâncias amostrais não se afastam muito da distribuição assumida, sugerindo um bom ajustamento do modelo aos dados.

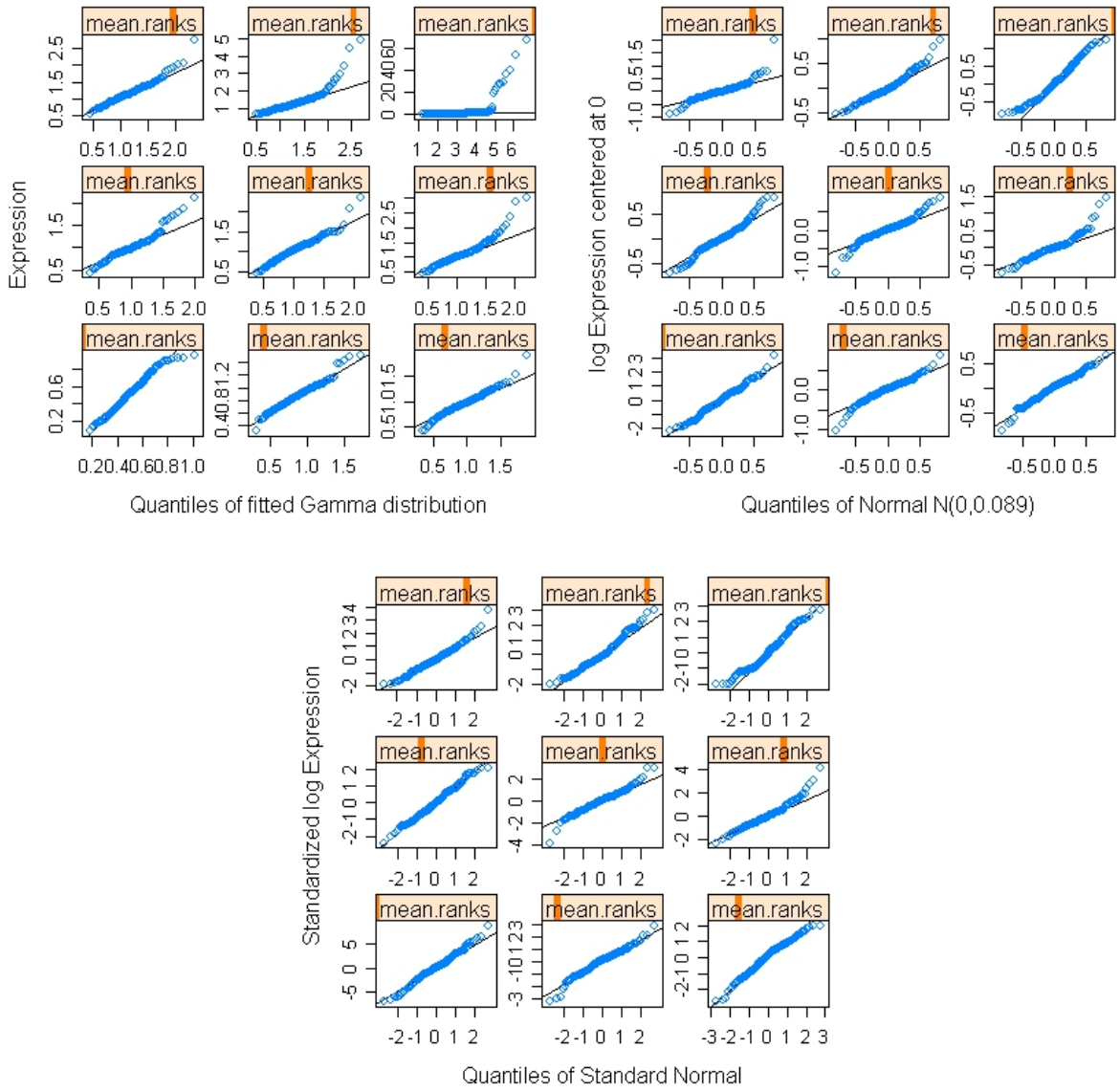

Figura 4.12: Gráficos de quantis sob a hipótese nula para os três modelos do EBarrays. Modelo Gama-Gama (cima, esquerda), modelo Lognormal-Normal (cima, direita) e Lognormal-Normal com Variância Modificada (baixo). Note-se que os gráficos apenas assumem os genes não diferencialmente expressos. Assim, prevendo a existência de genes diferencialmente expressos é esperada a observação de casos que violem a distribuição assumida.

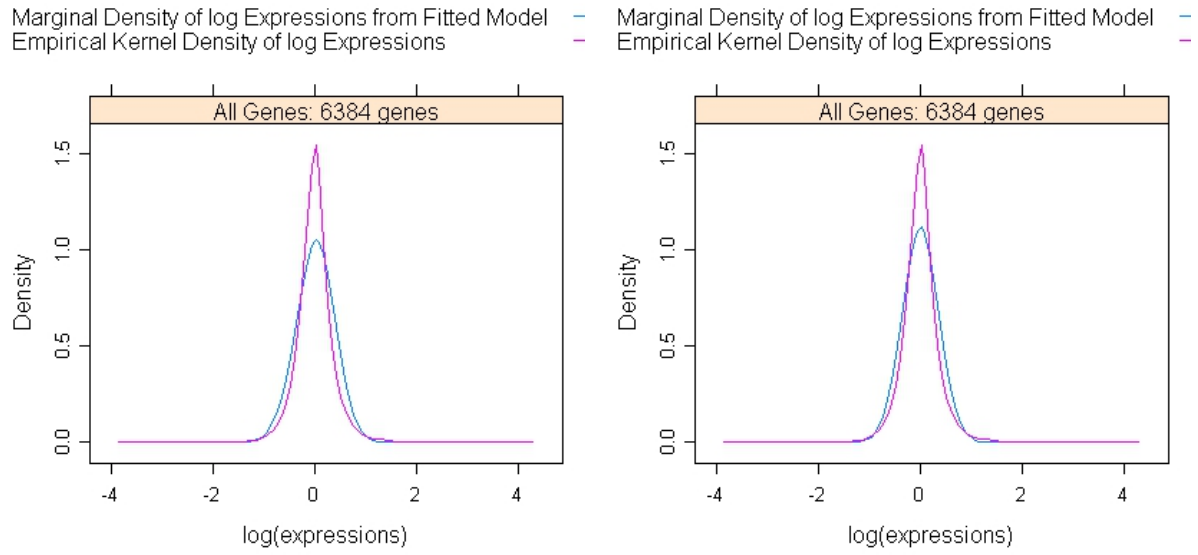

Figura 4.13: Gráfico das marginais para os modelos Gama-Gama (esquerda) e Lognormal-Normal (direita). Para cada gráfico, a curva a vermelho representa a marginal empírica usando o kernel gaussiano e a curva a azul a marginal do modelo ajustado. Os dois modelos mostram ajustamentos bastante semelhantes, no entanto, aparentemente o modelo Lognormal-Normal acompanha um pouco melhor a marginal empírica.

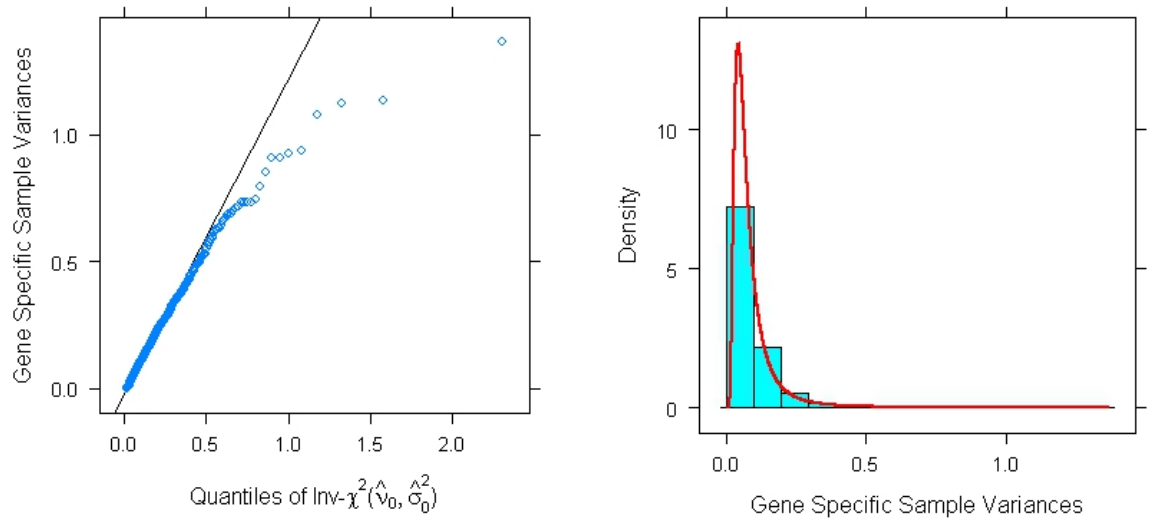

Figura 4.14: Gráfico de quantis para as variâncias amostrais (esquerda) e Histograma das variâncias amostrais e densidade da Qui-quadrado inversa escalonada (direita). Observa-se que as variâncias amostrais, de uma forma geral, não se afastam da distribuição Qui-quadrado inversa escalonada. Os parâmetros para a distribuição são estimados usando a base de dados.

Base de dados `Fermentation`

Relembrando, os objectivos desta base de dados são a detecção de genes diferencialmente expressos quando confrontadas as leveduras vínicas com as não vínicas, com a clínica ou com a laboratorial. Constitui ainda objectivo de estudo o confronto dos resultados obtidos e verificar que genes são comuns às três abordagens. Assim, para as três abordagens existem sempre duas condições, pelo que existirão apenas dois padrões: não diferencialidade e diferencialidade. Esses padrões são denotados da seguinte forma:

(1 1 1 1 1 1 1 1 1 1 1 1 1 1), representando a hipótese nula $\mu_{i1} = \mu_{i2}$.

(1 1 1 1 1 1 1 1 1 1 2 2 2 2), representando a hipótese alternativa $\mu_{i1} \neq \mu_{i2}$.

Como os valores M da base de dados devem estar na escala não logaritmica efectuou-se a exponencial dos valores. As metodologias associadas aos modelos de mistura GG, LNN e LNNVM não prevêem a existência de valores omissos, pelo que foi aplicado o algoritmo $k - nn$ com $k = 10$ para efectuar a imputação dos valores em falta, à semelhança do que se fez com a base de dados `ApoAI`.

Os gráficos da Figura 4.15 mostram que, de uma forma geral, o pressuposto de que os coeficientes de variação se mantenham constantes ao longo das médias das intensidades, não se verifica. O Tempo 3 é o que, de alguma forma, verifica um coeficiente de variação próximo de um valor constante, sendo este o caso em que os modelos GG e LNN parecem ajustar-se melhor. O Tempo 2 é o que aparentemente obtém um pior ajustamento.

As Figuras 4.16 e 4.17 ilustram gráficos construidos para verificar o pressuposto do modelo LNN das variâncias amostrais seguirem uma distribuição Qui-quadrado inversa escalonada. Observa-se que, em geral, o pressuposto não é violado. De facto, nos gráficos de quantis a generalidade dos pontos encontram-se sobre a recta $y = x$ e, nos histogramas as curvas de densidade parecem ajustar-se bem ao histograma das variâncias aleatórias. Deste modo, é de crer que o modelo LNNVM estabeleça um ajustamento aceitável dos dados.

As Figuras 4.18, 4.19 e 4.20 mostram o ajustamento dos dados à Normal e à Gama para os Tempos 2, 3 e 5, respectivamente, através de um gráfico de quantis. Para o Tempo 2 a quantidade de pontos que não se ajustam à recta $y = x$ aparenta ser muito semelhante para os três modelos, não sendo evidenciado o modelo que detectará mais genes diferencialmente expressos em cada uma das três análises (Vínicas e não vínicas, vínicas e clínica e vínicas e laboratorial). Eventualmente as maiores diferenças na detecção de genes diferencialmente expressos pelos três modelos estarão quando comparadas as leveduras vínicas com a laboratorial. Nesse caso, parece existir um maior desajuste para os modelos LNN e LNNVM e, consequentemente, uma maior quantidade de genes diferencialmente expressos. No Tempo 3, também os gráficos são bastante semelhantes entre si, não parecendo possível obter para já grandes conclusões. Eventualmente existirão mais genes detectados pelo modelo LNN quando comparadas as leveduras vínicas com a laboratorial. Da mesma forma, conclusões

prévias são difíceis de retirar observando apenas o gráfico do Tempo 5.

Antes de passar aos resultados de forma concreta, importa ainda verificar se as distribuições teóricas se ajustam aos dados de forma satisfatória. Os gráficos das Figuras 4.21, 4.22 e 4.23 representam os ajustamentos das distribuições teóricas dos modelos GG e LNN para as três análises para os Tempos 2, 3 e 5, respectivamente. Para o Tempo 2 e Tempo 3 ambos os modelos parecem ajustar-se de forma semelhante às distribuições empíricas. Eventualmente, para o Tempo 3 o ajustamento seja melhor quando considerado o modelo LNN. No Tempo 5 algumas diferenças evidenciam um maior ajustamento por parte do modelo LNN.

A Tabela 4.7 mostra o número de genes diferencialmente expressos obtidos com os três modelos para cada um dos 9 casos.

	Vínicas e Não Vínicas		
	Tempo 2	*Tempo 3*	*Tempo 5*
GG	286	292	196
LNN	299	265	157
LNNVM	291	326	137
	Vínicas e Clínica		
	Tempo 2	*Tempo 3*	*Tempo 5*
GG	144	207	192
LNN	201	223	191
LNNVM	219	209	261
	Vínicas e Laboratorial		
	Tempo 2	*Tempo 3*	*Tempo 5*
GG	278	237	174
LNN	359	227	183
LNNVM	413	263	181

Tabela 4.7: Número de genes diferencialmente expressos obtidos da aplicação dos três modelos para a `Fermentation`. Para cada base de dados (vínicas e não vínicas, vínicas e clínica e vínicas e laboratorial) em cada linha está o número de genes diferencialmente expressos obtidos para os modelos GG, LNN e LNNVM, respectivamente. Cada coluna representa o número de genes diferencialmente expressos obtidos para o Tempo 2, 3 e 5, respectivamente.

A Tabela 4.8 apresenta o número de genes concordantes entre os três modelos, verificando-se que, para cada um dos casos, o número de genes concordantes entre os modelos é bastante elevado quando comparado com a totalidade dos genes detectados pelos modelos.

Será também importante verificar o número de genes que coincidem nas três análises, isto é, à partida se um gene é detectado quando confrontando as vínicas com as não

	VNV	VC	VL
Tempo 2	186	98	243
Tempo 3	181	152	185
Tempo 5	121	160	142

Tabela 4.8: Número de genes concordantes nos três modelos para cada um dos casos de estudo da `Fermentation`. A primeira coluna contém o número de genes coincidentes nos três modelos para a base de dados contendo as leveduras vínicas e não vínicas nos Tempos 2, 3 e 5, respectivamente. A segunda coluna o número de genes coincidentes nos três modelos para os três tempos quando consideradas as leveduras vínicas e a clínica. E a terceira coluna o número de genes coincidentes nos três modelos para os três tempos quando consideradas as leveduras vínicas e a laboratorial.

vínicas, a laboratorial ou a clínica, então dará mais relevância à conclusão que esse gene assume diferenças significativas para as duas condições em estudo. Nas três análises da `Fermentation` e para os três modelos determinou-se o número de genes concordantes, tendo-se observado 44 genes no Tempo 2, 86 no Tempo 3 e 62 no Tempo 5. Sugere-se que estes genes tenham especial atenção por parte do biólogo, já que para as condições em causa as diferenças de expressão destes genes são tais que, independentemente da análise aplicada os três modelos detectam estes genes.

Será também de todo o interesse confrontar os resultados obtidos nas metodologias anteriores (SAM e Modelo Linear) e verificar quais destes genes coincidem entre elas. Verificou-se existirem 24 genes comuns no Tempo 2 nas três análises e nos cinco modelos, 74 genes no Tempo 3 e 37 no Tempo 5. A lista destes genes pode ser consultada no Apêndice A.

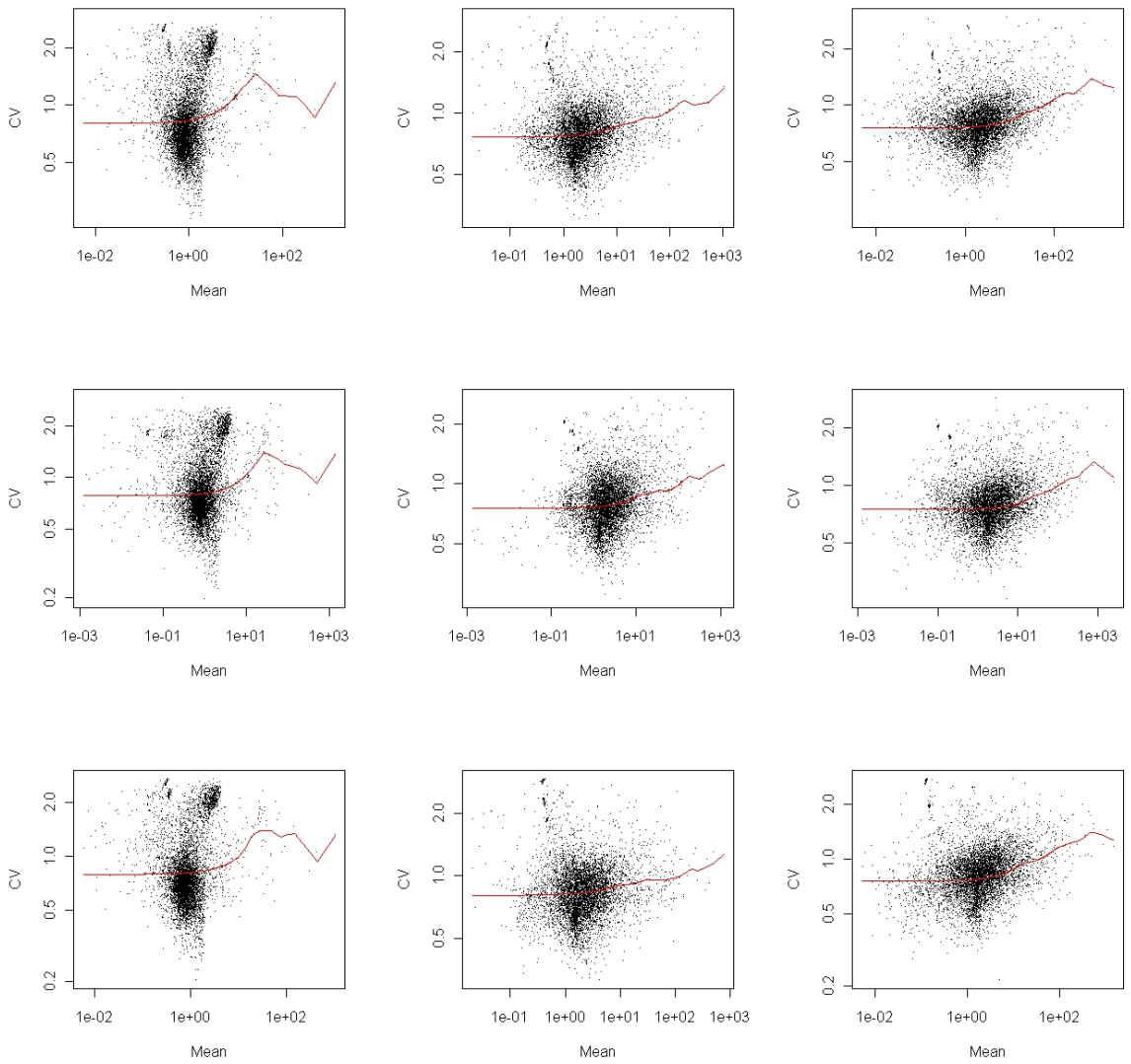

Figura 4.15: Gráfico das médias das intensidades versus coeficiente de variação para a `Fermentation`. A primeira linha representa os gráficos para as comparações entre leveduras vínicas e as não vínicas, a segunda para as vínicas e a clínica e a terceira para as vínicas e laboratorial. A primeira coluna é referente ao Tempo 2, a segunda ao Tempo 3 e a terceira ao Tempo 5.

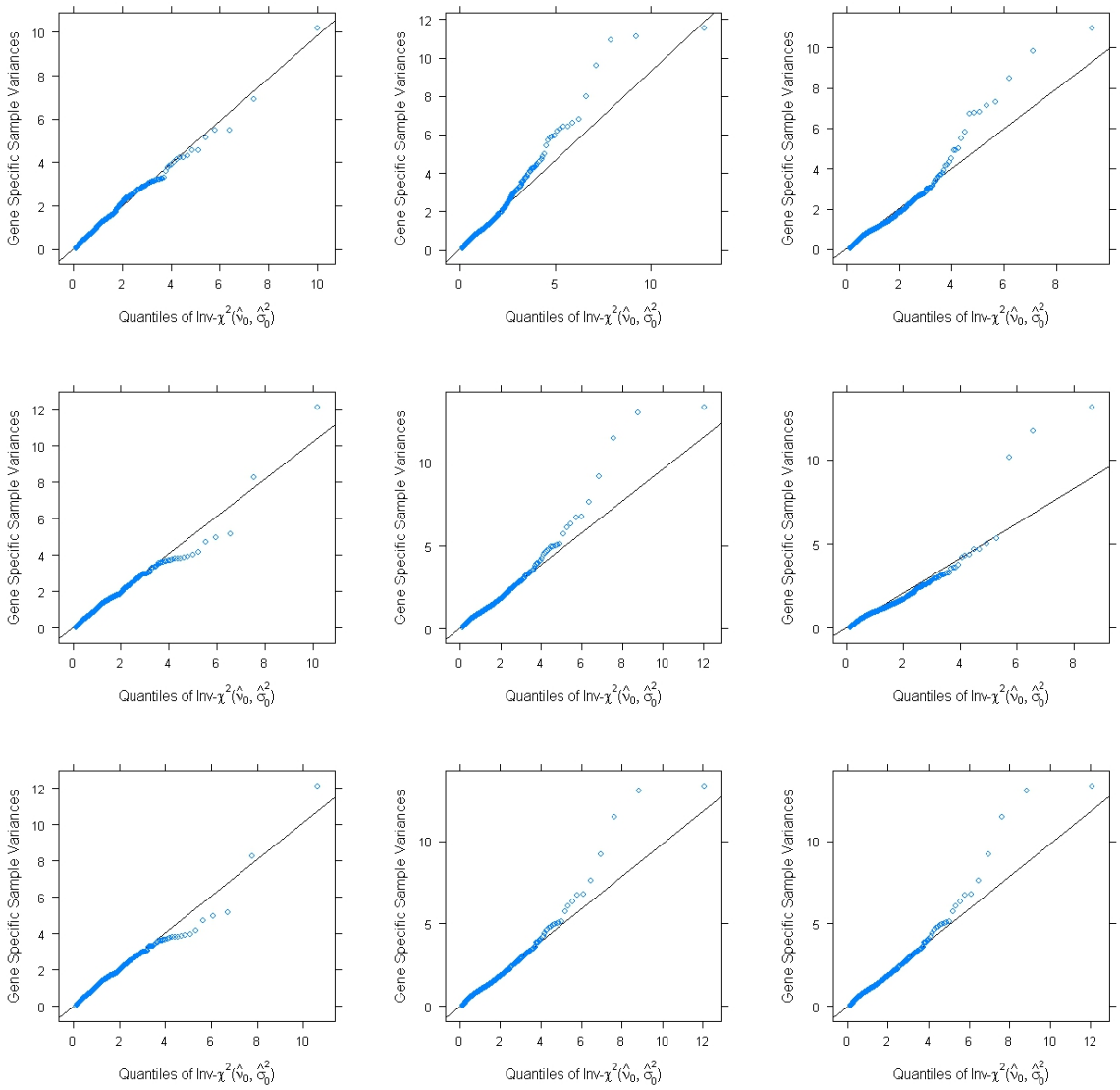

Figura 4.16: Gráfico de quantis para as variâncias aleatórias e Qui-quadrado inversa escalonada. Na primeira linha estão representados os gráficos dos Tempos 2, 3 e 5, respectivamente, para as leveduras vínicas e não vínicas. Na segunda linha os três tempos para as leveduras vínicas e a levedura clínica. Na terceira linha os gráficos para os mesmos três tempos mas para as leveduras vínicas e a levedura laboratorial.

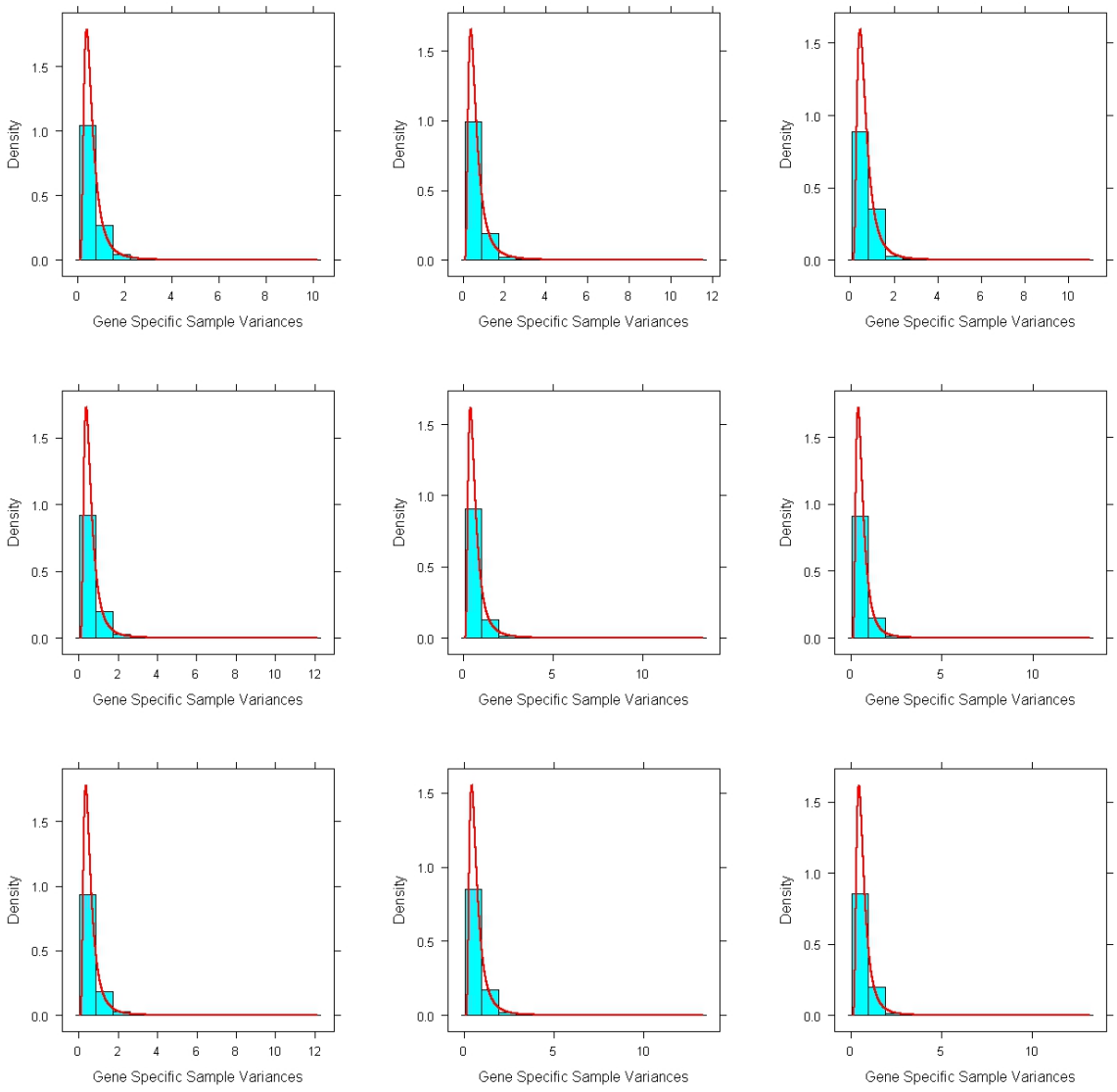

Figura 4.17: Histogramas para as variâncias aleatórias e sobreposição da curva de densidade da Qui-quadrado inversa escalonada. Na primeira linha estão representados os histogramas dos Tempos 2, 3 e 5, respectivamente, para as leveduras vínicas e não vínicas. Na segunda linha os três tempos para as leveduras vínicas e a levedura clínica. Na terceira linha os historamas para os mesmos três tempos mas para as leveduras vínicas e a levedura laboratorial.

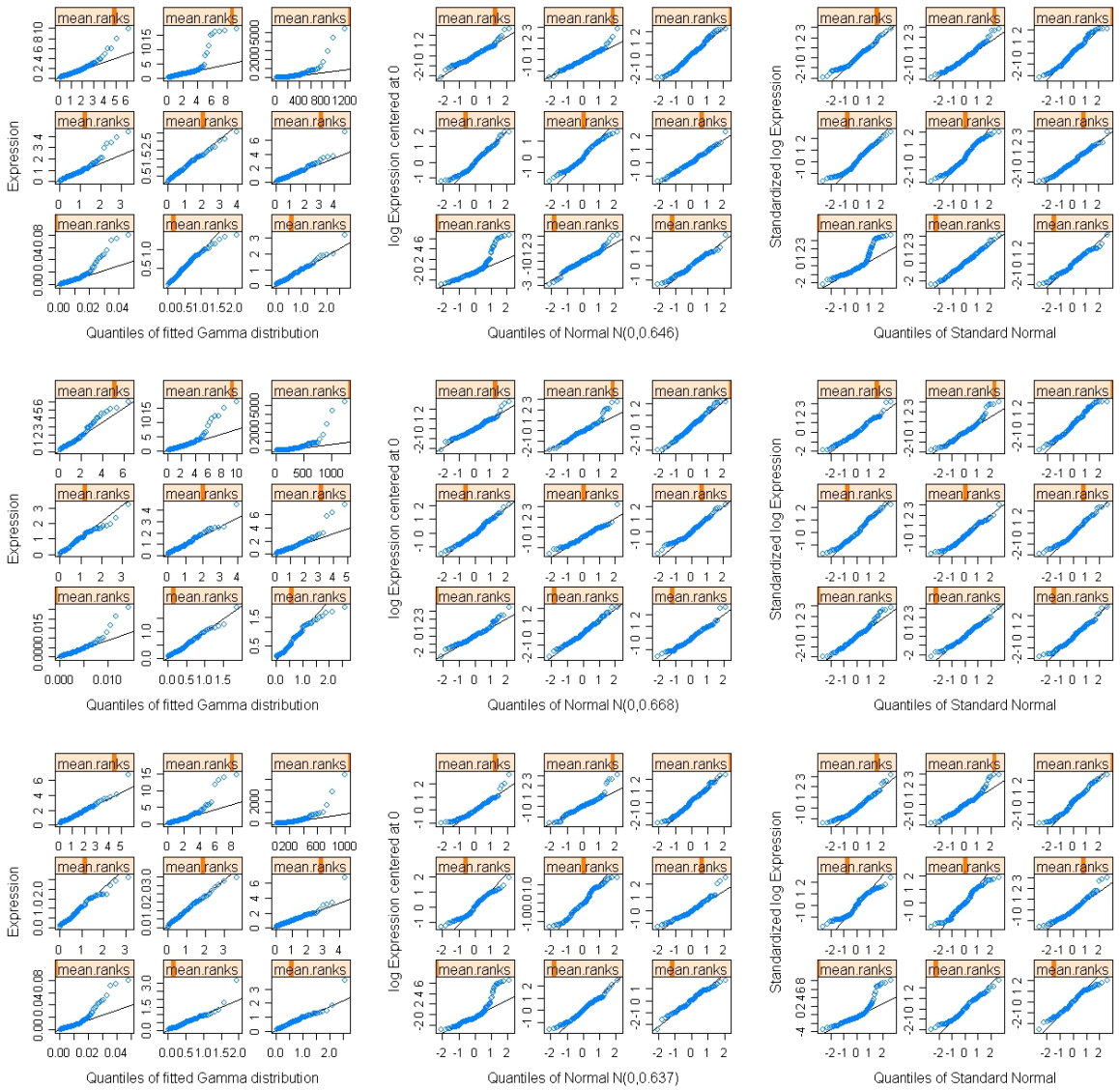

Figura 4.18: Gráficos quantis Tempo 2 referente aos modelos GG (gráficos da coluna à esquerda), LNN (gráficos da coluna central) e LNNVM (gráficos da coluna à direita), quando se confrontam as leveduras vínicas e não vínicas (gráficos da linha superior), leveduras vínicas e a levedura clínica (gráficos da linha central) e leveduras vínicas e a levedura laboratorial (gráficos da linha inferior).

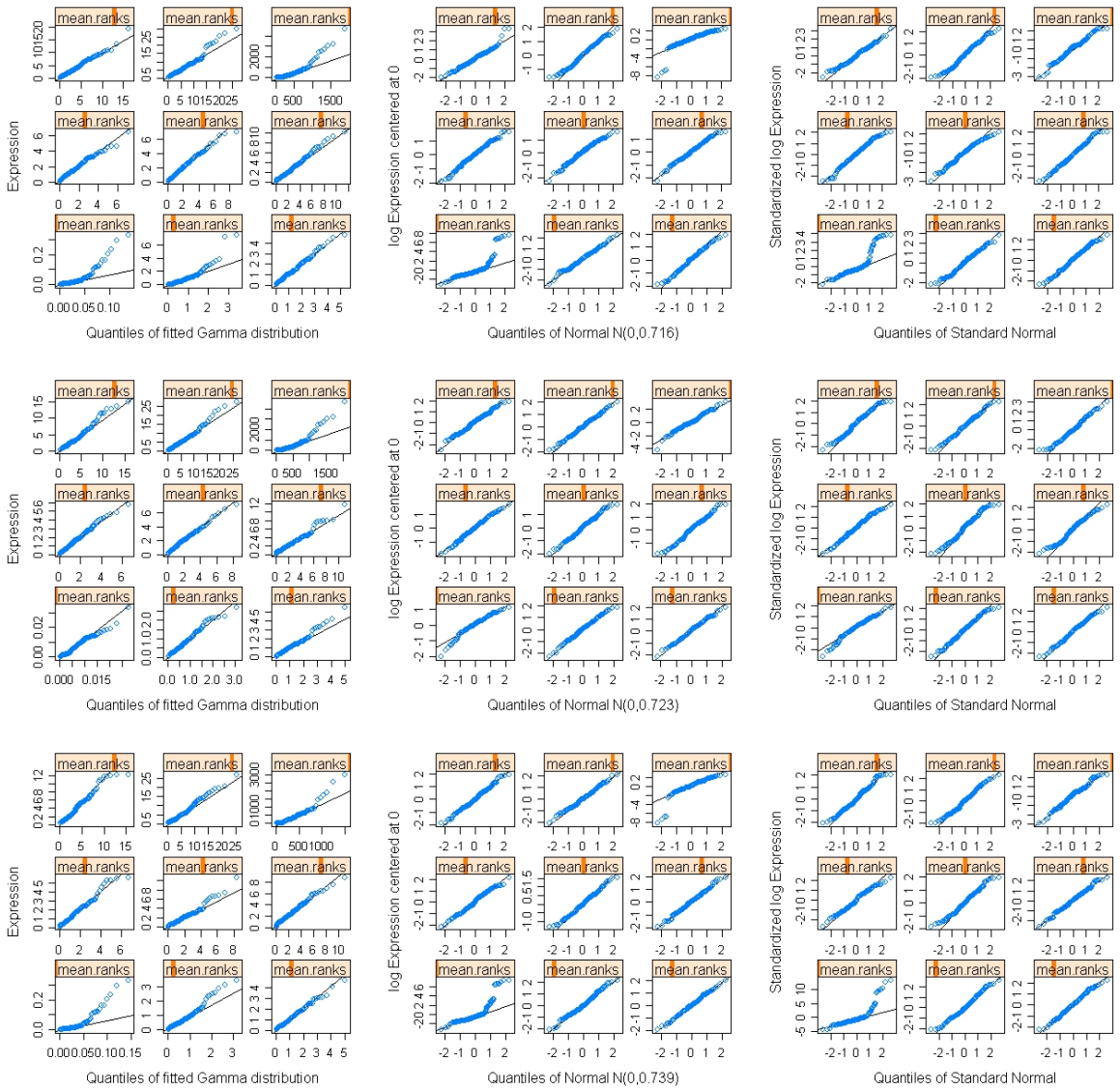

Figura 4.19: Gráfico quantis Tempo 3 referente aos modelos GG (gráficos da coluna à esquerda), LNN (gráficos da coluna central) e LNNVM (gráficos da coluna à direita), quando se confrontam as leveduras vínicas e não vínicas (gráficos da linha superior), leveduras vínicas e a levedura clínica (gráficos da linha central) e leveduras vínicas e a levedura laboratorial (gráficos da linha inferior).

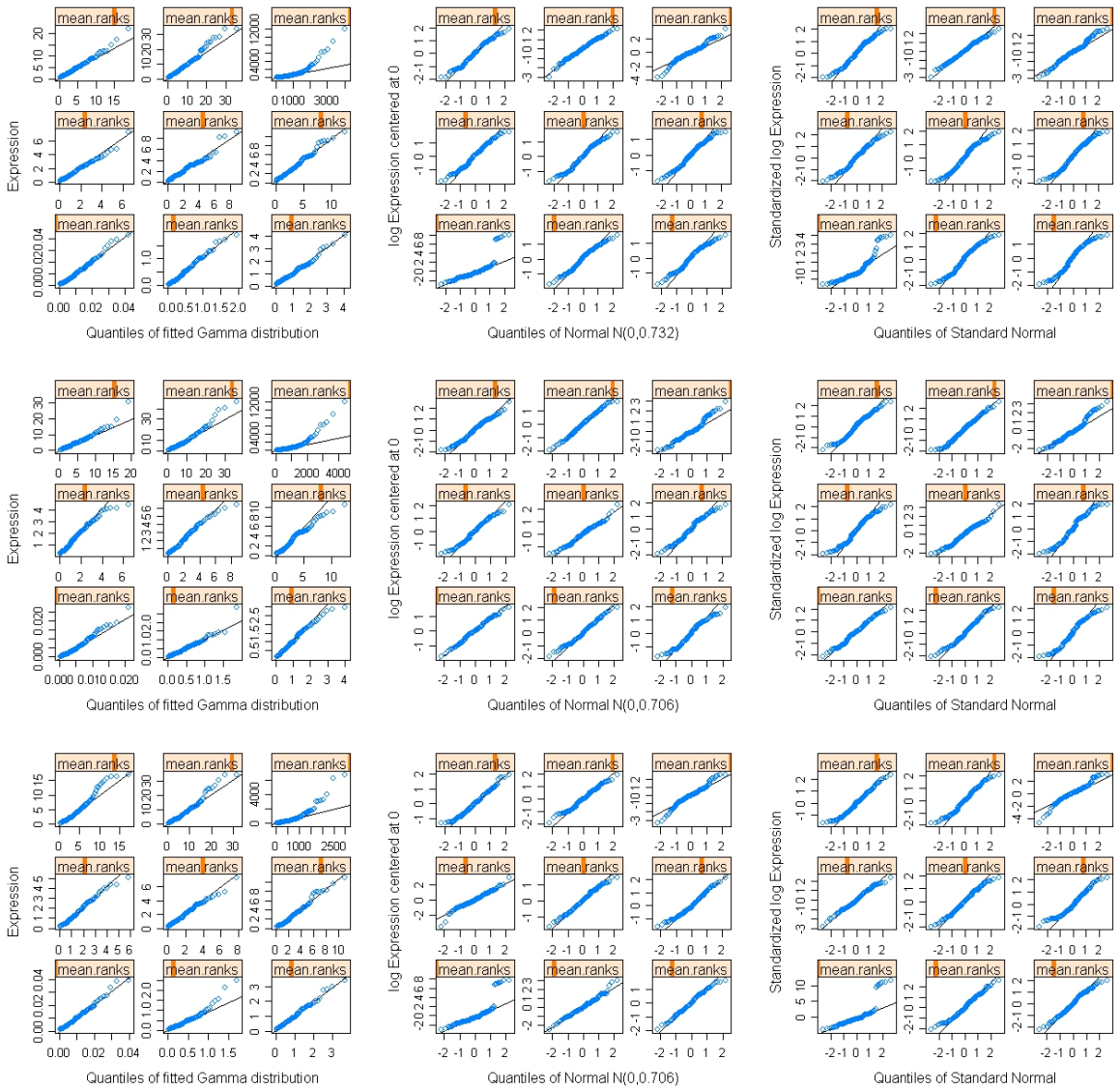

Figura 4.20: Gráfico quantis Tempo 5 referente aos modelos GG (gráficos da coluna à esquerda), LNN (gráficos da coluna central) e LNNVM (gráficos da coluna à direita), quando se confrontam as leveduras vínicas e não vínicas (gráficos da linha superior), leveduras vínicas e a levedura clínica (gráficos da linha central) e leveduras vínicas e a levedura laboratorial (gráficos da linha inferior).

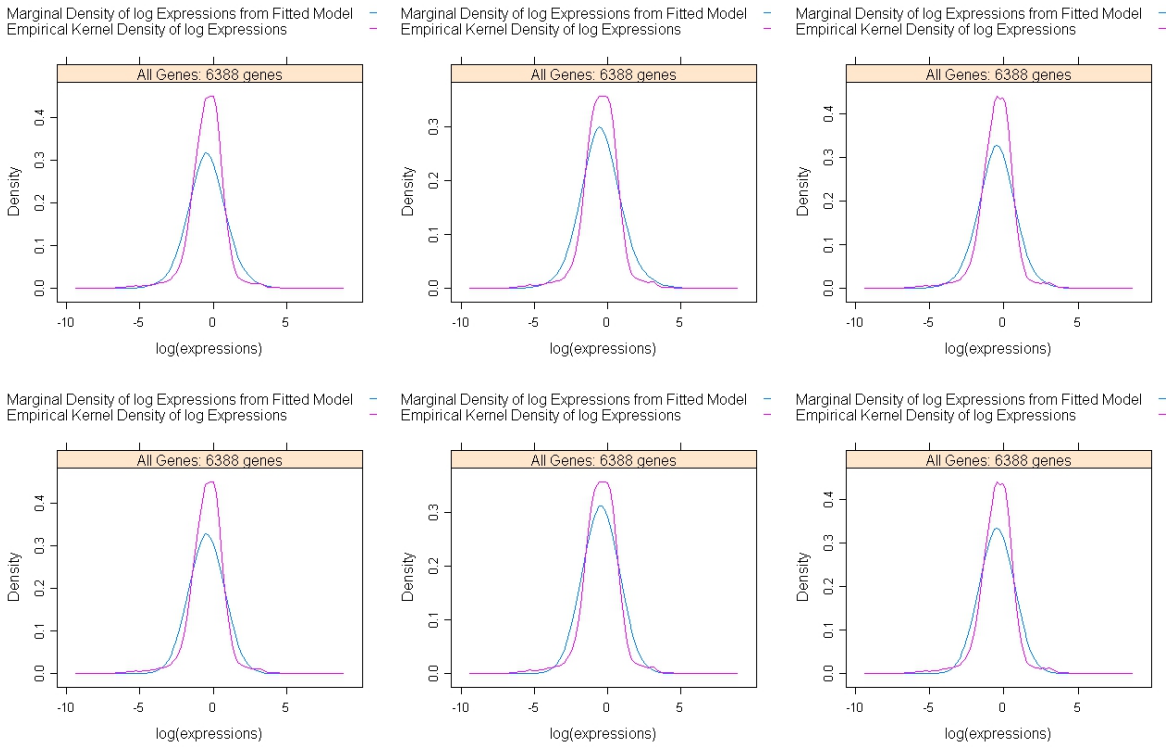

Figura 4.21: Marginais para o Tempo 2. A primeira linha é referente ao ajustamento do modelo GG para a base de dado quando consideradas as leveduras vínicas e as não vínicas, as leveduras vínicas e a clínica e as leveduras vínicas e a laboratorial, respectivamente. A segunda linha é respectiva ao ajustamento do modelo LNN para as bases de dados quando consideradas as leveduras vínicas e as não vínicas, as leveduras vínicas e a clínica e as leveduras vínicas e a laboratorial, respectivamente.

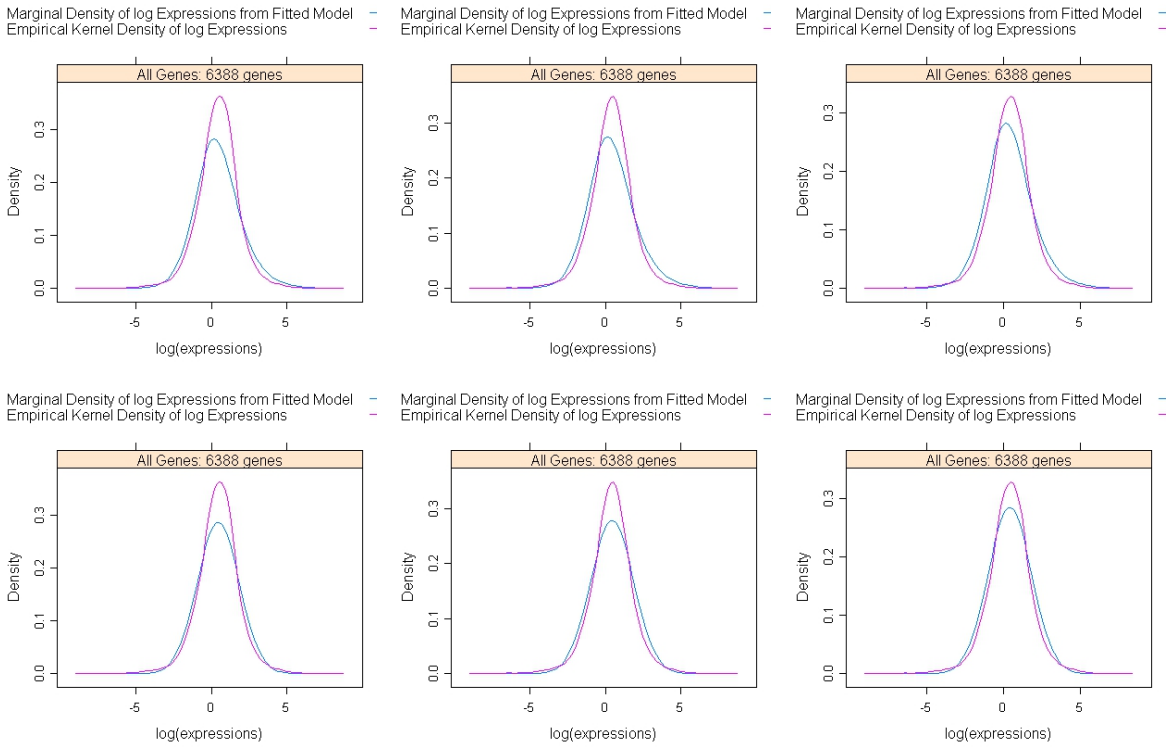

Figura 4.22: Marginais para o Tempo 3. A primeira linha é referente ao ajustamento do modelo GG para as bases de dados quando consideradas as leveduras vínicas e as não vínicas, as leveduras vínicas e a clínica e as leveduras vínicas e a laboratorial, respectivamente. A segunda linha é respectiva ao ajustamento do modelo LNN para as bases de dados quando consideradas as leveduras vínicas e as não vínicas, as leveduras vínicas e a clínica e as leveduras vínicas e a laboratorial, respectivamente.

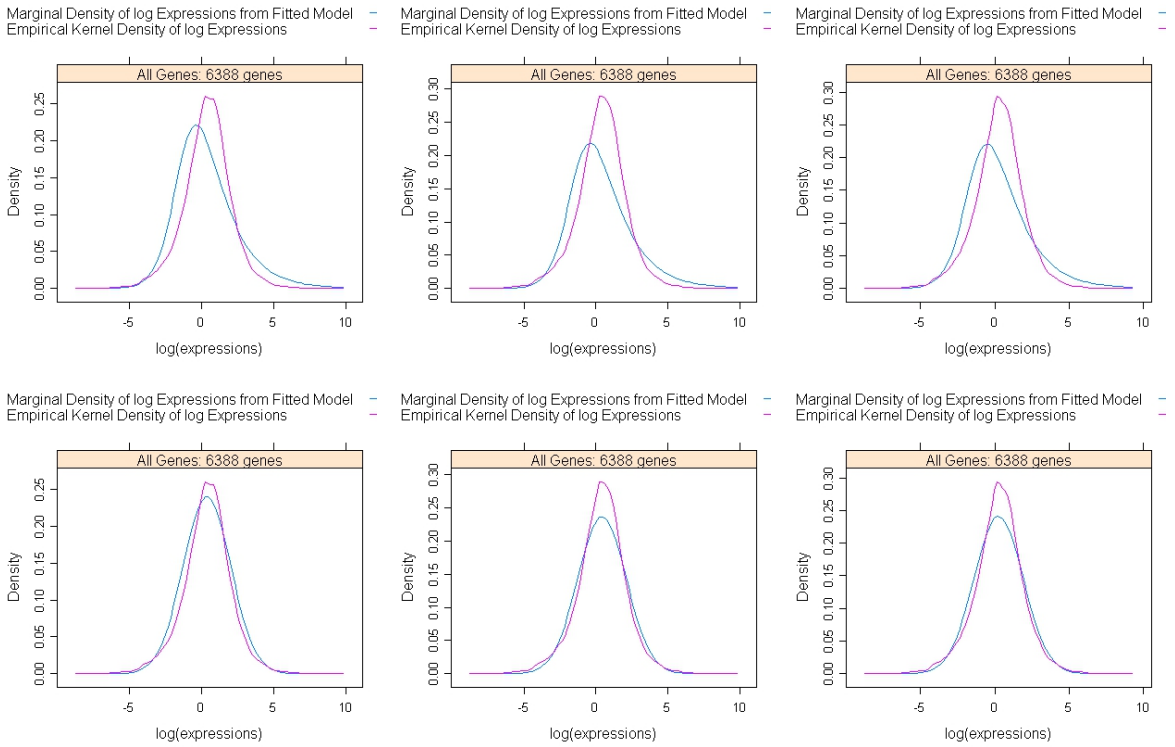

Figura 4.23: Marginais para o Tempo 5. A primeira linha é referente ao ajustamento do modelo GG para as bases de dados quando consideradas as leveduras vínicas e as não vínicas, as leveduras vínicas e a clínica e as leveduras vínicas e a laboratorial, respectivamente. A segunda linha é respectiva ao ajustamento do modelo LNN para as bases de dados quando consideradas as leveduras vínicas e as não vínicas, as leveduras vínicas e a clínica e as leveduras vínicas e a laboratorial, respectivamente.

Capítulo 5

Conclusões e trabalho futuro

No presente trabalho foram abordadas diversas metodologias estatísticas para a identificação de genes diferencialmente expressos, nomeadamente a SAM e métodos que recorrem ao conceito de Bayes empírico (modelo linear, modelo Gama-Gama (GG), Lognormal-Normal (LNN) e Lognormal-Normal com Variância Modificada (LNNVM)). Essas metodologias foram aplicadas a duas bases de dados (`ApoAI` e `Fermentation`) com vista à detecção de genes diferencialmente expressos.

O estudo efectuado com a primeira base de dados, `ApoAI`, possibilitou mais do que a detecção dos genes diferencialmente expressos. Dada a existência de uma vasta literatura para esta base de dados, onde são sugeridos os genes que evidenciam diferenças nos seus níveis de expressão, foi possível avaliar a capacidade de detecção de genes diferencialmente expressos dessas metodologias. Quando confrontando os genes obtidos com o ajustamento dos dados a um modelo linear e os genes obtidos com a aplicação da SAM, verificou-se que este último detectou mais quatro. Obteve-se as densidades dos genes identificados como diferencialmente expressos pela SAM. Para os quatro genes não detectados com o modelo linear, as densidades das observações obtidas dos ratos com o gene ApoAI normal e das observações obtidas dos ratos com o gene ApoAI deficiente mostraram uma maior aproximação quando comparadas com as densidades dos outros genes detectados pelo modelo linear. O facto deste modelo não detectar os quatro genes referidos leva a crer que a SAM terá, à partida, uma maior sensibilidade às diferenças dos níveis de expressão. No que respeita à aplicação dos modelos contidos no pacote `EBarrays`, apenas os modelos GG e LNNVM obtiveram uma quantidade de genes bastante superior ao esperado, no entanto, após o cruzamento da informação e verificação de quais os genes em que ambos os modelos estavam de acordo, verificou-se que esses genes coincidiam com os já obtidos com as outras metodologias e com os referidos na literatura.

Para a base de dados `Fermentation`, foi muito importante a aplicação de várias metodologias, no sentido de verificar os genes diferencialmente expressos que tinham acordo por parte de todas as metodologias, obtendo-se uma maior confiança no que respeita à classificação desses genes como diferencialmente expressos. Além disso, foram verificados quais dos genes mantinham acordo nas três análises efectuadas, isto é, quando comparadas as leveduras vínicas com as não vínicas, quando comparadas as leveduras vínicas com a clínica

ou quando comparadas as leveduras vínicas com a laboratorial. Foi possível ainda, através do cruzamento de toda esta informação, obter um conjunto de genes considerados diferencialmente expressos, independentemente da metodologia aplicada e da análise efectuada. Verificou-se que, de uma forma geral, o número genes comuns nas três análises, detectados com o ajustamento de um modelo linear é semelhante ao número de genes detectados pela aplicação da SAM e dos modelos do pacote `EBarrays`, no entanto, quando feito o cruzamento dos resultados obtidos para todas as metodologias, o número de genes considerados diferencialmente expressos diminuiu consideravelmente. A determinação deste conjunto constitui assim, um ponto de partida para o biólogo, de forma a estabelecer que genes deverão ser estudados em laboratório para retirar conclusões mais precisas sobre o seu comportamento nos processos de fermentação. A lista destes genes pode ser consultada no Apêndice A.

Em ambas as bases de dados o cruzamento da informação foi feita usando os genes diferencialmente expressos obtidos com o uso da estatística t na SAM e os obtidos nas restantes metodologias.

Como trabalho futuro fica em aberto a aplicação de um estudo mais aprofundado verificando quais os genes comuns entre os detectados com a estatística de teste de wilcoxon e os detectados com a estatística t. Este estudo poderia, eventualmente, alterar a dimensão do conjunto de genes detectados independentemente da análise e da metodologia aplicada, obtendo possivelmente resultados mais conclusivos.

Para a base de dados `Fermentation` também alguns trabalhos ficam em aberto. As análises foram efectuadas considerando sempre em separado apenas duas classes, isto é, a comparação entre vínicas e não vínicas, vínicas e clínica e entre vínicas e laboratorial. Uma outra abordagem poderia ser efectuada, onde apenas seriam consideradas duas abordagens, a comparação entre vínicas e não vínicas ou a comparação entre as vínicas, clínica e laboratorial tomando para o efeito três classes, a classe das vínicas, a classe das clínicas e a classe das laboratoriais. Este tipo de abordagem trará alterações em todas as metodologias aplicadas. Na SAM ter-se-á de efectuar uma análise multiclasse. Nos modelos lineares a matriz de delineamento terá outra dimensão e o vector de parâmetros contemplará ambos os contrastes $v\acute{i}nicas - cl\acute{i}nicas$ e $v\acute{i}nicas - laboratorial$. Esta abordagem poderá ainda fazer uso da matriz de contrastes onde um novo contraste de interesse $v\acute{i}nicas - \frac{cl\acute{i}nicas+laboratorial}{2}$ (comparação entre os níveis de expressão das vínicas e a média dos níveis de expressão da clínica e laboratorial) poderá ser definido e posteriormente comparados os seus resultados com os obtidos com o contraste $v\acute{i}nicas - n\tilde{a}o\ v\acute{i}nicas$. Nos modelos definidos no pacote `EBarrays`, se consideradas três classes, obter-se-ão mais padrões que os dois considerados (diferencialidade e não diferencialidade de expressão). Os resultados obtidos quando considerados os diversos padrões poderão ser comparados com as diferenças encontradas pela abordagem adoptada nesta dissertação.

Bibliografia

[1] Ewens, Warren J. e Grant, Gregory R. (2005) *Statistical Methods in Bioinformatics, An Introduction*, Springer, second edition.

[2] Hochberg, Y. (1988), A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75, 800:803.

[3] Holm, S. (1979), A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6, 65:70.

[4] John D. Storey, e Robert Tibshirani. (2003), Statistical significance for genomewide studies, Edited by Philip P. Green, University of Washington School of Medicine, Seattle, WA

[5] Qiong Yang, et al. (2005), Power and type I error rate of false discovery rate approaches in *genome-wide association studies*,BMC Genet; 6(Suppl 1): S134. 2005 December 30. doi: 10.1186/1471-2156-6-S1-S134.

[6] Shaffer J.P. (1995) Multiple hypothesis testing, *Annual Review of Psychology* 46:561-584

[7] Abdi, H (2007). "Bonferroni and Sidák corrections for multiple comparisons". in N.J. Salkind (ed.). *Encyclopedia of Measurement and Statistics*. Thousand Oaks, CA: Sage. http://www.utdallas.edu/ herve/Abdi-Bonferroni2007-pretty.pdf.

[8] Bland, Martin (2000). Multiple significance tests In: *An Introduction to Medical Statistics*, Oxford University Press. Secção disponível em http://www-users.york.ac.uk/ mb55/intro/bonf.htm

[9] Speed, Terry. Multiple testing in large-scale gene expression experiments, disponível em
http://www.stat.berkeley.edu/users/terry/Classes/s246.2004/Week13/2004L22Stat246.pdf

[10] Marcos, Ana Luísa Romão de São, Avaliação de metodologias de pré-processamento de dados de microarrays, Universidade de Aveiro, Departamento de Matemática, 2009, Dissertação, Matemática e Aplicações, área de especialização Matemática Empresarial e Tecnológica, Universidade de Aveiro, 2009.

[11] Grosso, Ana Rita Fialho, Statistical Methodologies for the Analysis of DNA Microarray Data, Universidade de Lisboa, Faculdade de Ciências da Universidade de Lisboa, Departamento de Estatística e Investigacão Operacional, 2006, Dissertação, Mestrado em Bioinformática.

[12] Documentação do pacote limma. Disponível em http://bioconductor.org/packages/2.4/bioc/html/limma.html

[13] Documentação do pacote EBarrays. Disponível em http://bioconductor.org/packages/2.2/bioc/html/EBarrays.html

[14] Gordon K. Smyth, Matthew Ritchie, Natalie Thorne, James Wettenhall e Wei Shi, limma: Linear Models for Microarray Data User's Guide, Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, Melbourne, Australia, 22 October 2009 http://bioconductor.org/packages/2.4/bioc/html/limma.html

[15] Portfolio realizado pelo grupo de Bioinformática da Universidade de Aveiro no âmbito do projecto Novas metodologias Estatísticas para Análise de dados de Microarrays de ADN. http://bioinformatics.ua.pt/resources/pub/pfma.pdf.

[16] Newton, M.A. e Kendziorski, C.M. (2003) Parametric Empirical Bayes Methods fpr Microarrays in *The analysis of gene expression data: methods and software*. Eds. G.Parmigiani, E.S. Garrett, R. Irizarry and S.L. Zeger, New york: Springer Verlag.

[17] Newton, M.A., Noueiry, A., Sarkar, D., e Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture model. *Biostatistics* 5: 155-176.

[18] Kendziorski, C.M., Newton, M.A., Lan, H. e Gould, M.N. (2003). On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine* 22:3899-3914.

[19] Smith, 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, pages 397-420.

[20] Filho, D. F. e Leandro, R. A. (2009). Análise de Microarray usando o R e o Bioconductor, Universidade de São Paulo.

[21] Borman, Sean, The Expectation Maximization Algorithm A short tutorial, July 18 2004.

[22] Smith, Gordon K. (2004), Linear Models e Empirical bayes Methods for Assessing Differential Expression in Microarray Experiments, *Statistical Applications in Genetics and Moloecular Biology* 3, No.1, Article 3.

[23] Bradley P. Carlin and Thomas A. Louis,(1996) *Bayes and Empirical Bayes for Data Analysis*, Second edition, New York: Chapman & Hall

[24] Dudoit S., Yang Y., Callow M. and Speed T. (2002), Statistical Methods for Identifying Differentially Expressed Genes in Replicated cDNA Microarray Experiments, *Statistica Sinica* 12, 111-139.

[25] Dean N. and Raftery A. (2005), Normal uniform mixture differencial gene detection for cDNA microarrays, *BMC Bioinformatics*, 6:173

[26] Schwender H., Krause A., and Ickstadt K., Identifying Interesting Genes with siggenes

[27] Tusher V., Tibshirani R. and Chu G. (2001), Significance analysis of microarrays applied to the ionizing radiation response, *Bradley Efron*, Stanford University, Stanford, CA, February 6, vol. 98, no. 9, 5116−5121

[28] Wiel, M.A. van (2004), Significance Analysis of Microarrays using Rank Scores, *Kwantitatieve Methoden* 71, 25-37.

[29] G. Chu, B. Narasimhan, R.Tibshirani and V.Tusher, SAM, Significance Analysis of Microarrays, *Users guide and tecnical document*

[30] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser.* B 57, 289-300.

[31] Storey John D. (2003), The Positive False Discovery Rate: A Bayesian Interpretation and the q-Value. *The Annals os Statistics*, 2003 Vol. 31, No. 6, 2013-2035, Institute of Mathematical Statistics.

[32] Storey John D. (2002), A direct approach to false discovery rates. *J.R Statist. Soc. B*, 64, Part 3, pp. 479-498.

[33] http:3.bp.blogspot.com/_NupbWI2XxeA/SNSBGP−gVeI /AAAAAAAACf4/−P3HCZcZECTU/s320/gene+cromossomo+dna.gif.

[34] http://www.bing.com/images/search?q=DNA&FORM=BIFD#focal= 0c68f92bfe813c8685−d3f98cd103564b&furl=http%3A%2F% 2Fwww.shoppingblog.com%2Fpics%2Fdna_strand.jpg

[35] al, A. J. (cop.1986). Introduction to genetic analysis. New York : W. H. Freeman and Company.

[36] Gonçalves, F. S. (2008). Obtido em 24 de 10 de 2009, de InfoEscola-Navegando e aprendendo: http://www.infoescola.com/genetica/transcricao/

[37] Silva, L. M. (2009) Selecção de variáveis em *microarrays* de adn. *Tese de Mestrado.* Departamento de Matemática Aplicada, Faculdade de Ciências da Universidade do Porto, Porto, Portugal.

[38] Ozan Gundogdu, A. E. (2009) London School of Hygiene & Tropical Medicine University of London. Genome Resource Facility: http://www.lshtm.ac.uk/itd/grf/contactus.htm

[39] Rikshospitalet. (s.d.) http://www.rr−research.no/lothe/?k=lothe/methods&aid=2823

[40] Genovese, C. R. (s.d.). A Tutorial on False Discovery Control. Department of Statistics, Carnegie Mellon University.

Apêndice A

Genes Comuns

" GENES COMUNS A NAS TRÊS ANÁLISES PARA TODAS AS METODOLOGIAS NO TEMPO 2"

Linha	Gene
2885	YIL014CA
2887	YIL015CA
333	YBR069C
2984	YIL111W
4320	YLR410WB
2319	YGR038CB
2450	YGR161WB
592	YCL019W
1236	YDR210WD
4937	YNL054WB
5416	YOL106W
1116	YDR098CB
6316	YPR137CB
1402	YDR365WB
1234	YDR210WB
4130	YLR227WB
248	YBL101WB
5831	YOR343CB
3351	YJR027W
1048	YDR034CD
5675	YOR192CB
1291	YDR261WB
6176	YPR002CA
1980	YFR026C

" GENES COMUNS A NAS TRÊS ANÁLISES PARA TODAS AS METODOLOGIAS NO TEMPO 3"

Linha	Gene
2885	YIL014CA
2887	YIL015CA
2885	YIL014CA
1980	YFR026C
5474	YOL163W
5475	YOL164W
5412	YOL103WA
5622	YOR142WA
1230	YDR210CC
4160	YLR256WA
3352	YJR028W
1288	YDR261CC
2447	YGR161CC
4426	YML040W
1821	YER137CA
4587	YMR051C
1401	YDR365WA
6147	YPL257WA
1347	YDR316WA
1115	YDR098CA
1235	YDR210WC
1845	YER159CA
274	YBR012WA
3350	YJR026W
4936	YNL054WA
4431	YML045WA
2863	YHR214CC
2318	YGR038CA
4057	YLR157CA
110	YAR010C
4129	YLR227WA
145	YBL005WA
1822	YER138C
109	YAR009C
5413	YOL103WB

2448 YGR161CD
5623 YOR142WB
5170 YNL284CB
4425 YML039W
4058 YLR157CB
6148 YPL257WB
2307 YGR027WB
1846 YER160C
275 YBR012WB
3931 YLR035CA
4580 YMR045C
4586 YMR050C
146 YBL005WB
5169 YNL284CA
2306 YGR027WA
4430 YML045W
1348 YDR316WB
2862 YHR214CB
1231 YDR210CD
4581 YMR046C
3351 YJR027W
4320 YLR410WB
5675 YOR192CB
2319 YGR038CB
2450 YGR161WB
1402 YDR365WB
4937 YNL054WB
1116 YDR098CB
592 YCL019W
3093 YIR042C
6182 YPR007C
248 YBL101WB
5831 YOR343CB
1236 YDR210WD
1234 YDR210WB
1291 YDR261WB
1048 YDR034CD
4130 YLR227WB
1289 YDR261CD

" GENES COMUNS A NAS TRÊS ANÁLISES PARA TODAS AS METODOLOGIAS NO TEMPO 5"

Linha	Gene
2885	YIL014CA
1980	YFR026C
5474	YOL163W
2887	YIL015CA
2885	YIL014CA
3837	YLL010C
6315	YPR137CA
6341	YPR158WA
5412	YOL103WA
5622	YOR142WA
3352	YJR028W
6147	YPL257WA
6338	YPR158CC
1230	YDR210CC
1288	YDR261CC
1401	YDR365WA
1845	YER159CA
4426	YML040W
4160	YLR256WA
3350	YJR026W
1347	YDR316WA
4057	YLR157CA
4587	YMR051C
1235	YDR210WC
1115	YDR098CA
274	YBR012WA
4936	YNL054WA
2863	YHR214CC
1821	YER137CA
2447	YGR161CC
2318	YGR038CA
110	YAR010C
145	YBL005WA
4129	YLR227WA
4431	YML045WA
1822	YER138C
3175	YJL078C

A utilização do nome sistemático do gene na base de dados *Saccharomyces Genome Database* (www.yeastgenome.org) permite acesso a várias informações funcionais e estruturais acerca deste gene, nomeadamente a sua localização no genoma, sequência nucleotídica e função da proteína que codifica.

Apêndice B

Comandos em R

```
BASE DE DADOS APOAI

############################################################
#####
##### SAM
#####
############################################################


## MUDAR O DIRECTORIO PARA A PASTA QUE CONTÉM A BASE DE DADOS
## Fazer o load dos pacotes samr e limma e da base de dados

library(samr)
library(limma)
load("ApoAI.RData")


## Normalizar os dados antes de usa-los, neste caso usando o método
## normalização global

MA <- normalizeWithinArrays(RG)

## Obter a matriz dos valores M a utilizar na análise

x<-MA$M


## Obter o vector das variáveis de resposta

y<-c(rep(1,8),rep(2,8)) # 1 representa os ratinhos de controlo e
```

```
                              # 2 os ratinhos doentes


## Criar a lista de dados a entrar no SAM

data=list(x=x,y=y, geneid=MA$genes[,6], genenames=MA$genes[,5],
logged2=TRUE)


## Obter a análise de significância usando a estatística t e a
## estatistica de wilcoxon

      samr.obj<-samr(data,  resp.type="Two class unpaired", nperms=1000)

  samr.objW<-samr(data,  resp.type="Two class unpaired",
nperms=1000,testStatistic="wilcoxon")

## Obter a tabela de deltas para ambas as estatisticas

delta.table <- samr.compute.delta.table(samr.obj);
delta.tableW <- samr.compute.delta.table(samr.objW);

## Escolher o melhor delta, tendo em conta a FDR
## para ambas as estatísticas

delta=.64
deltaW=1.833437e-01

## Fazer o gráfico das estatísticas observadas
## versus estatísticas estimadas

      samr.plot(samr.obj,delta)

## Obter informação acerca dos genes diferencialmente expressos para o
## delta escolhido para ambas as estatísticas

      siggenes.table<-samr.compute.siggenes.table(samr.obj,delta, data,
delta.table)

  siggenes.tableW<-samr.compute.siggenes.table(samr.objW,deltaW,
data, delta.tableW)

########################################################
```

```
#######
####### Empirical Bayes
#######
#########################################################

#######
#######     AJUSTAMENTO DO MODELO LINEAR
#######


## Construir a matriz de design

design <- cbind("WT-Ref"=1,"KO-WT"=rep(0:1,c(8,8)))
design

##Ajustar o modelo linear

fit <- lmFit(MA,design=design)
colnames(fit)
names(fit)

## Usar o Empirial Bayes

fit <- eBayes(fit)
names(fit)
summary(fit)

## Ver a tabela dos genes diferencialmente expressos

table<-topTable(fit,coef="KO-WT",adjust="fdr",p.value=0.01)

## Contrução do vulcano plot
    y<- topTable(fit,coef="KO-WT", number=nrow(MA$M), adjust="fdr")
    names(y)
    plot(y[,9],-log(y[,12],12),xlab="log2(Fold-Change)"
,ylab="-log10(P.Value)",main="Volcano plot",cex=0.2,pch=19)
    abline(v=c(-1,1),col="blue")
    abline(h=-log(0.01,10),col="red")

## Obter o gráfico de quantis das estatísticas t (t-student)

qqt(fit$t[,2],df=fit$df.residual+fit$df.prior,
    main = "Gráfico de Quantis", xlab = "quantis teóricos",
```

```
    ylab = "quantis observados")

    abline(0,1)
title(sub = "Estatísticas de teste t vs t-student")


## Obter o gráfico MA dos coeficientes do modelo ajustado

plotMA(fit, 2)


## construção das densidades teoricas e empíricas para os genes DE
## -> Exemplo para um gene

#Gene "EST,WeaklysimilartoF"

plot(density(MA$M[947,1:8]), main="EST,WeaklysimilartoF", xlab="log(R/G)",
    ylab="densidades",xlim=c(-2,1),ylim=c(0,2.5))
par(new=TRUE)
plot(density(MA$M[947,9:16]), col="blue", main="", xlab="", ylab="",
    xlim=c(-2,1),ylim=c(0,2.5))


############################################################
###
### EBarrays
###
############################################################

## Fazer o load do pacote EBarrays

library("EBarrays")


## declaração dos padrões possíveis, neste caso expressão diferencial
## ou não diferencial

padroes<-ebPatterns(c("1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1",
                      "1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2"))
padroes

# Obter os valores não logaritmizados da base de dados
```

```
x<-RG$R/RG$G

## Obter o vector de variáveis de resposta para entrar no
## k-nn para repor valores omissos

y<-c(rep(1,8),rep(2,8)) # 1 representa os ratos de controlo e
                        # 2 os ratos doentes

# Criar a lista total de dados para entrar no k-nn

data <- list(x=x,y=factor(y))

## Imputar valores omissos

# fazer o load do pacote pamr que contém a metodologia knn
library(pamr)

## Obter a nova base, sem valores omissos

data<-pamr.knnimpute(data ,k = 10, rowmax = 0.5, colmax = 0.8
, maxp = 96000)

## Obter a matrix dos valores R/G

dados<-data$x

## Obter os graficos de dispersão para verificação dos
## pressupostos sobre o CV e variâncias

print(checkCCV(dados))
print(checkVarsQQ(dados, groupid=c(1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2)))
print(checkVarsMar(dados, groupid=c(1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2)))



#####
##### Modelo Gama-Gama
#####


## Ajustar o modelo

fit<-emfit(dados, family="GG", hypotheses=padroes)
```

```
fit

## Calcular a probabilidade a posteriori para cada hipotese,
## para cada gene.

posteriori<-postprob(fit,dados)$pattern

## Encontrar um threshold para a probabilidade a posteriori de forma a
## controlar a FDR

threshold<-crit.fun(posteriori[,1],0.001)

## Obter o número de genes diferencialmente expressos

sum(posteriori[,2]>threshold)

## Obter a lista de genes DE

for(i in 1:6384){
  if(posteriori[i,2]>threshold)
  print(i)
}

## Gráfico de quantis para o modelo

print(checkModel(dados,fit, model="gamma"))

## Gráfico das marginais empírica e teórica

print(plotMarginal(fit,dados))


#####
##### Modelo LogNormal-Normal
#####


## Ajustar o modelo

fit.LNN<-emfit(dados, family="LNN", hypotheses=padroes, num.iter=10)
fit.LNN

## Calcular a probabilidade a posteriori para cada hipotese,
```

```
## para cada gene.

posteriori.LNN<-postprob(fit.LNN,dados)$pattern

## Encontrar um threshold para a probabilidade a posteriori de forma
## a controlar a FDR

threshold.LNN<-crit.fun(posteriori.LNN[,1],0.001)

## Obter o número de genes DE

sum(posteriori.LNN[,2]>threshold.LNN)

## Obter a lista ge genes DE

for(i in 1:6384){
  if(posteriori.LNN[i,2]>threshold.LNN)
  print(i)
}

## Obter o gráfico de quantis para o modelo

print(checkModel(dados,fit.LNN, model="lognormal"))

## Obter as marginais empírica e teórica

print(plotMarginal(fit.LNN,dados))


#####
##### Modelo LogNormal-Normal com variancia modificada
#####


##Ajustar o modelo

fit.LNNMV<-emfit(dados, family="LNNMV", hypotheses=padroes,
groupid=c(1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2), num.iter=10)
fit.LNNMV

## Calcular a probabilidade a posteriori para cada hipotese,
## para cada gene.
```

```
 posteriori.LNNMV<-postprob(fit.LNNMV,dados,groupid=c(1,1,1,1,1,1,1,1,
2,2,2,2,2,2,2,2))$pattern

## Encontrar um threshold para a probabilidade a posteriori de forma
## a controlar o FDR

threshold.LNNMV<-crit.fun(posteriori.LNNMV[,1],0.001)

## Obter o número de genes DE

sum(posteriori.LNNMV[,2]>threshold.LNNMV)

## Obter a lista de genes DE

for(i in 1:6384){
  if(posteriori.LNNMV[i,2]>threshold.LNNMV)
  print(i)
}

## Obter o gráfico de quantis para o modelo

print(checkModel(dados,fit.LNNMV,
            model="lnnmv",groupid=c(1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2)))


## BASE DE DADOS FERMENTATION

################################################
####                                        ####                    SAM
################################################

## Ler informação do ficheiro txt

dados<-read.delim("Fermentation_All data_Normalized.txt",
header = TRUE,sep = "\t", quote = "\"", dec = ".", fill = TRUE)

## Isolar a matriz de dados dos níveis de expressão do resto da informação

niveis_expressao<-dados[,9:92];

##
## Análise para a comparação Vínicas e não Vínicas - Tempo 2
##
```

```
## Obter a matriz com os níveis de expressão apenas para o tempo 2

x_T2<-niveis_expressao[,c(3,4,15,16,27,28,39,40,51,52,63,64,75,76)]
x_T2<-matrix(unlist(x_T2),nrow=6388,ncol=14)

## Obter o vector das variáveis resposta

y_T2<-c(rep(1,10),rep(2,4)) # 1 representa as leveduras vínicas e
                            # 2 as não vínicas

## Obter a lista de dados a entrar no SAM

data=list(x=x_T2,y=y_T2, geneid=dados[,1], genenames=dados[,2],
logged2=TRUE)

## Aplicar a SAM aos dados do tempo 2 com as estatísticas t e wilcoxon

samr.obj_T2<-samr(data,  resp.type="Two class unpaired", nperms=100)
samr.obj_T2W<-samr(data,  resp.type="Two class unpaired",
nperms=100,testStatistic="wilcoxon")

## Obter a tabela de deltas para ambas as estatísticas

delta.table_T2<- samr.compute.delta.table(samr.obj_T2);
delta.table_T2W<- samr.compute.delta.table(samr.obj_T2W);

## Escolher os deltas tendo em conta a FDR para ambas as estatísticas

delta=1.362
deltaW=0.6212865181

## Obter o gráfico das estatísticas observadas versus
## estatísticas estimadas

samr.plot(samr.obj_T2,delta)
title(main="Vinicas/ Não Vinicas (Tempo2)", sub="Delta=1.352")

## Obter as informações sobre os genes diferencialmente expressos
## para ambas as estatísticas

siggenes.table_T2<-samr.compute.siggenes.table(samr.obj_T2,delta,
data,delta.table_T2)
```

```
siggenes.table_T2W<-samr.compute.siggenes.table(samr.obj_T2,deltaW,
data,delta.table_T2W)

#Construir os gráficos Ngenes vs FDR
## Apenas exemplificado como construir a curva para a comparação e tempo
## em causa

plot(delta.table_T2[,4],delta.table_T2[,5],col="white", main=" Tempo2 ",
xlab="N° GenesDE", ylab="TFD", xlim=c(0,4000),ylim=c(0,1))

lines(delta.table_T2[,4],delta.table_T2[,5], col="green", lwd=2)
text(1500,0.3,"Vínicas e N Vinicas", col="green")

#########################################################
#######
####### Empirical Bayes
#######
#########################################################

#######
#######     AJUSTAMENTO DO MODELO LINEAR
#######

#####
##### Vinicas/N Vinicas Tempo2
#####


#Juntar todas as colunas respectivas ao tempo2

x_T2<-niveis_expressao[,c(3,4,15,16,27,28,39,40,51,52,63,64,75,76)]


# Construir a matriz design

design<-cbind(rep(1,14),c(rep(0,10),rep(-1,4)))


##Ajustar o modelo linear

fit <- lmFit(x_T2,design=design)
```

```
## Usar o Empirial Bayes

fit <- eBayes(fit)
names(fit)
summary(fit)

## ver a tabela dos genes diferencialmente expressos

 table<-topTable(fit,number=6000,coef=2,genelist=dados[,2],adjust="fdr"
,p.value=0.05)

## Contrução do vulcano plot
    y<- topTable(fit,coef=2, number=nrow(x_T2), adjust="fdr")
    names(y)
    plot(y[,1],-log(y[,3],3),xlab="log2(Fold-Change)",
ylab="-log10(P.Value)",main="Vínicas e Não Vínicas (Tempo 2)",cex=0.2
,pch=19)
    abline(v=c(-1,1),col="blue")
    abline(h=-log(0.05,10),col="red")

## Obter o gráfico de quantis das estatísticas t (t-student)
#par(mfrow=c(2,2))

qqt(fit$t[,2],df=fit$df.residual+fit$df.prior,
main = "Vínicas e Não Vínicas (Tempo 2)", xlab = "quantis teóricos",
ylab = "quantis observados")
a<-qt(0.05,mean(fit$df.residual+fit$df.prior))
abline(0,1)
abline(-a,1)
abline(a,1)

############################################################
###
### EBarray
###
############################################################

## Efectuar o load do pacote EBarrays

library("EBarrays")


###
```

```
### Vinicas e Não Vinicas
### Tempo2
###

## Obter a matriz referente ao dados do tempo 2

x_T2<-niveis_expressao[,c(3,4,15,16,27,28,39,40,51,52,63,64,75,76)]

## Obter os padrões, neste caso diferencialidade e não diferencialidade

padroes<-ebPatterns(c("1,1,1,1,1,1,1,1,1,1,1,1,1,1",
"1,1,1,1,1,1,1,1,1,1,2,2,2,2"))

## Obter os dados não logaritmizados

x<-exp(x_T2)
x<-matrix(unlist(x),nrow=6388,ncol=14)

## Vector para Vinicas e N Vinicas

y<-c(rep(1,10),rep(2,4)) # 1 representa vinicas e 2 não vinicas


# Criar lista de dados a entrar no k-nn

data <- list(x=x,y=factor(y))



## Imputar valores omissos

library(pamr)
data<-pamr.knnimpute(data ,k = 10, rowmax = 0.5, colmax = 0.8)
dados<-data$x

#####
##### Modelo Gama-Gama
#####


##Ajustar o modelo

fit<-emfit(dados, family="GG", hypotheses=padroes, num.iter=10)
```

```
fit

## Calcular a probabilidade a posteriori para cada hipotese,
## para cada gene.

posteriori<-postprob(fit,dados)$pattern

## Encontrar um threshold para a probabilidade a posteriori de forma a
## controlar a FDR

threshold<-crit.fun(posteriori[,1],0.05)

## Obter os genes DE

sum(posteriori[,2]>threshold)

sigGGVNV2<-matrix(nrow=286,ncol=1)
j=1;
for(i in 1:6388){
  if(posteriori[i,2]>threshold){
  sigGGVNV2[j,1]<-i;
  j=j+1;}}

## Obter o gráfico de quantis para o modelo

print(checkModel(dados,fit, model="gamma"))

## Obter as marginais empírica e teórica

print(plotMarginal(fit,dados))

#####
##### Modelo LogNormal-Normal
#####

## Ajustar o modelo

fit.LNN<-emfit(dados, family="LNN", hypotheses=padroes, num.iter=10)
fit.LNN

## Calcular a probabilidade a posteriori para cada hipotese,
## para cada gene.
```

```
posteriori.LNN<-postprob(fit.LNN,dados)$pattern

## Encontrar um threshold para a probabilidade a posteriori de forma a
## controlar a FDR

threshold.LNN<-crit.fun(posteriori.LNN[,1],0.05)

## Obter os genes DE

sum(posteriori.LNN[,2]>threshold.LNN)

sigLNNVNV2<-matrix(nrow=299,ncol=1)
j=1;
for(i in 1:6388){
  if(posteriori[i,2]>threshold){
  sigLNNVNV2[j,1]<-i;
  j=j+1;}}

## Obter o gráfico de quantis para o modelo

print(checkModel(dados,fit.LNN, model="lognormal"))

## Obter as marginais teórica e empírica

print(plotMarginal(fit.LNN,dados))

#####
##### Modelo LogNormal-Normal com variancia modificada
#####


##Ajustar o modelo

fit.LNNMV<-emfit(dados, family="LNNMV", hypotheses=padroes,
groupid=c(1,1,1,1,1,1,1,1,1,1,2,2,2,2), num.iter=10)
fit.LNNMV

## Calcular a probabilidade a posteriori para cada hipotese,
## para cada gene.

 posteriori.LNNMV<-postprob(fit.LNNMV,dados,groupid=c(1,1,1,1,1,1,1,1
,1,1,2,2,2,2))$pattern
```

```
## Encontrar um threshold para a probabilidade a posteriori de forma
## a controlar a FDR

threshold.LNNMV<-crit.fun(posteriori.LNNMV[,1],0.05)

## Obter os genes DE

sum(posteriori.LNNMV[,2]>threshold.LNNMV)

sigLNNVMVNV2<-matrix(nrow=291,ncol=1)
j=1;
for(i in 1:6388){
  if(posteriori[i,2]>threshold){
  sigLNNVMVNV2[j,1]<-i;
  j=j+1;}}

## Obter o gráfico de quantis para o modelo

print(checkModel(dados,fit.LNNMV,
        model="lnnmv",groupid=c(1,1,1,1,1,1,1,1,1,1,2,2,2,2)))

# Gráficos para verificar os pressupostos do CV e variância

print(checkCCV(dados))
print(checkVarsQQ(dados, groupid=c(1,1,1,1,1,1,1,1,1,1,2,2,2,2)))
print(checkVarsMar(dados, groupid=c(1,1,1,1,1,1,1,1,1,1,2,2,2,2)))

#### Concordâncias dos modelos GG e LNN

sum(posteriori[,2]>threshold & posteriori.LNN[,2]>threshold.LNN)

for(i in 1:6384){
  if(posteriori[i,2]>threshold & posteriori.LNN[i,2]>threshold.LNN)
  print(i)
}

#### Concordâncias dos modelos GG e LNNMV

sum(posteriori[,2]>threshold & posteriori.LNNMV[,2]>threshold.LNNMV)

for(i in 1:6384){
  if(posteriori[i,2]>threshold & posteriori.LNNMV[i,2]>threshold.LNNMV)
  print(i)
```

```
}

#### Concordâncias dos modelos LNN e LNNMV

sum(posteriori.LNN[,2]>threshold.LNN &
            posteriori.LNNMV[,2]>threshold.LNNMV)

for(i in 1:6384){
  if(posteriori.LNN[i,2]>threshold.LNN &
            posteriori.LNNMV[i,2]>threshold.LNNMV)
  print(i)
}

#### Concordâncias dos modelos GG, LNN e LNNMV

sum(posteriori.LNN[,2]>threshold.LNN &
        posteriori.LNNMV[,2]>threshold.LNNMV &
        posteriori[,2]>threshold )

for(i in 1:6384){
  if(posteriori.LNN[i,2]>threshold.LNN &
            posteriori.LNNMV[i,2]>threshold.LNNMV &
            posteriori[i,2]>threshold )
  print(i)}
```