Análise de Dados Genéticos

Marcus Nunes

Diamantina, Outubro de 2014

Este livro foi escrito com XƎLaTeX e `knitr`.

XIII Encontro Mineiro de Estatística

Coordenação Geral
Glaura da Conceição Franco - UFMG

Comissão Organizadora Local
Paulo César de Resende Andrade
Lucas Franco Ferreira
Marcelino Serretti Leonel
Emerson Cotta Bodevan
Josiane Magalhães Teixeira
Geovane da Conceição Máximo

Comissão Científica
Marcel de Toledo Vieira - UFJF
Lourdes Coral Contreras Montenegro - UFMG
Ricardo Luis dos Reis - UFVJM
Nilson Luiz Castelucio Brito - UNIMONTES
João Domingos Scalon - UFLA
Marcos Santos de Oliveira - UFSJ
Fernando Luiz Pereira de Oliveira - UFOP
Priscila Neves Faria - UFU
Eric Batista Ferreira - UNIFAL
Antonio Policarpo Souza Carneiro – UFV

Apoio Técnico
Gustavo Bastos e Godoi

Conteúdo

1 Introdução **7**

2 Genômica **9**
2.1 RNA-Seq 13
2.2 Pipeline 16

3 Análise de Dados **21**
3.1 Delineamento de Experimentos 21
3.1.1 Amostragem 22
3.1.2 Aleatorização 23
3.1.3 Replicação 24
3.1.4 Agrupamento em Blocos 25
3.2 Modelagem 27
3.2.1 Modelos Lineares Generalizados 27
3.2.2 edgeR 38
3.3 MA Plot 40
3.4 Comparações Múltiplas 41

4 Aplicação **49**
4.1 Análise Exploratória dos Dados 49
4.2 Testes de Expressão Gênica 62

Bibliografia **70**

1 Introdução

A área de biotecnologia cresce num ritmo vertiginoso no mundo inteiro. A quantidade de dados disponíveis para análise aumenta, e muito, a cada dia que passa. Estes dados são gerados para análise posterior não apenas nas áreas de pesquisa, mas principalmente na área médica. Por exemplo, Green and Guyer (2011) estimam que cada vez mais a ciência médica será capaz de utilizar este tipo de dado para prevenção e tratamento de doenças. Eles também afirmam que a partir de 2020 a medicina será capaz de utilizar estes recursos no dia-a-dia das clínicas e hospitais, devido à sua precisão e barateamento.

A tecnologia mais atual para sequenciamento genético chama RNA-Seq. Estamos interessados em introduzir os principais conceitos desta tecnologia, bem como conceitos biológicos e estatísticos relacionados a ela. Veremos desde os conceitos básicos de DNA, para entender com o que estamos trabalhando, até a análise dos dados tratados e resumidos.

Neste minicurso veremos uma introdução à análise de dados genéticos. Este trabalho não é, de maneira alguma, uma revisão extensiva da literatura da área. A ideia principal deste livro é situar os alunos em relação às atuais tecnologias e alguns dos métodos utilizados para a análise dos dados produzidos por elas. A evolução das ferramentas de análise não para nunca e métodos de análise atuais

podem se tornar obsoletos em pouco tempo.

Veremos alguns dos conceitos básicos do DNA e da genômica no Capítulo 2. Além disso, uma pequena introdução às tecnologias de sequenciamento é apresentada nesta parte do texto, bem como as ferramentas computacionais utilizadas para converter os resultados dos sequenciamentos obtidos diretamente das máquinas para dados que o R consegue analisar. O Capítulo 3 mostra algumas das ferramentas estatísticas utilizadas para a análise deste tipo de dados. São apresentados desde conceitos de delineamento de experimentos adaptados aos dados RNA-Seq até as ferramentas estatísticas mais utilizadas para analisá-los. No Capítulo 4 reuniremos alguns dos conceitos vistos durante o livro e analisaremos um experimento real de RNA-Seq.

2 Genômica

Os últimos 30 anos viram uma gigante transformação na genética e genômica. Podemos traçar suas origens (na verdade, as origens das ideias da seleção natural) até a Grécia e Roma antigas, com Empédocles e Lucrécio (400 a.C. e 55 a.C., respectivamente), e aos pensadores da época do Iluminismo, como Diderot(1749) e Maupertius (1756) (ver Zirkle (1941) para mais referências).

Entretanto, apenas com Gregor Mendel descobrimos que os genes possuem papel fundamental nos mecanismos da vida. Frei agostiniano e cientista, Mendel é hoje considerado o Pai da Genética. Entre 1856 e 1863, ele delineou e executou um experimento que o levou a descobrir a existência de genes dominantes e recessivos. Durante 7 anos, ele plantou e testou 29.000 ervilhas, mostrando que uma em cada quatro ervilhas possuíam apenas alelos recessivos, duas em cada quatro eram híbridas e uma possuía apenas alelos dominantes.

O maior avanço na genética, entretanto, ocorreu apenas no século XX. A primeira descrição da dupla hélice do DNA (ou ácido nucleico) apareceu em Watson and Crick (1953). O DNA é composto de quatro nucleotídeos diferentes, chamados adenina (A), timina (T), citosina (C) e guanina (G). A estrutura química dos quatro nucleotídeos do DNA está representada na Figura 2.1.

Cada nucleotídeo é chamado de base. Dizemos que adenina e

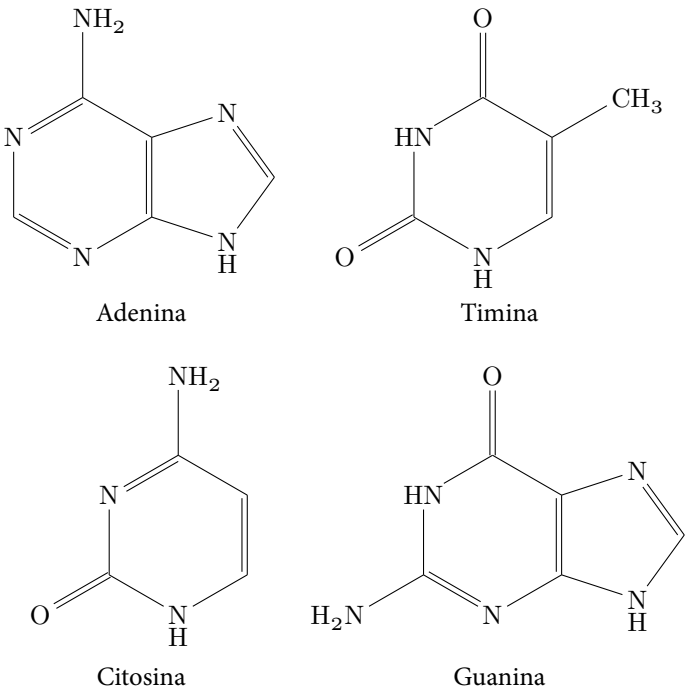

Figura 2.1: Estrutura química dos quatro nucleotídeos do DNA.

guanina são purinas, enquanto citosina e timina são chamadas de pirimidinas. A adenina liga-se apenas com a timina, enquanto a

citosina liga-se apenas com a guanina. Assim, embora a estrutura do DNA seja composta por uma dupla hélice, com duas fitas de material genético, basta que tenhamos uma destas fitas para, automaticamente, sabermos a composição da outra. A Figura 2.2 (em inglês) mostra estas peculiaridades em detalhes.

Na década de 1970, o método chamado "Sanger sequencing" (Sanger and Coulson (1975) Sanger et al. (1977)) tornou-se a primeira maneira padrão de sequenciar genomas. Este foi o método utilizado no Projeto Genoma Humano, por exemplo. Criado em 1993, com o objetivo de sequenciar completamente o DNA de um ser humano, levou 13 anos para ficar pronto, ao custo de US$2,7 bilhões (Green and Guyer (2011)). Atualmente, as máquinas de RNA-Seq, a tecnologia que nos interessa neste trabalho, são capazes de sequenciar o genoma de um ser humano em 8 horas, ao custo de US$1.000,00.

Apesar de todos os avanços na obtenção de dados genéticos, ainda não existe uma maneira definitiva de se testar a diferença de expressão gênica para experimentos de sequenciamento RNA-Seq. Embora a literatura da área já tenha métodos bastante conhecidos para realizar estes testes (por exemplo, Anders and Huber (2010) e Robinson et al. (2010), criadores dos pacotes `DESeq` e `edgeR` do `R`), eles apresentam deficiências. Por exemplo, uma das características dos experimentos de sequenciamento RNA-Seq é a geração de muitos zeros. Os métodos tradicionais de análise removem boa parte destes zeros durante sua execução.

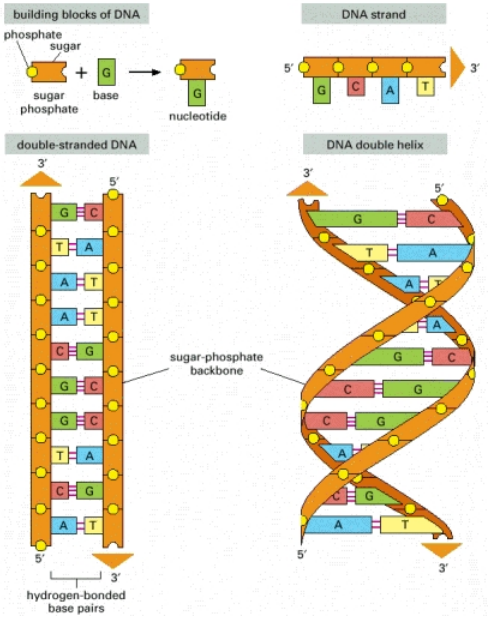

Figura 2.2: Estrutura de dupla hélice do DNA com seus quatro nucleotídeos: adenina (A), timina (T), guanina (G) e citosina (C). Cada posição nas fitas é chamada de base. Imagem retirada de Alberts et al. (2002).

2.1 RNA-Seq

A tecnologia mais recente disponível no mercado é chamada de RNA-Seq (ou MPSS - Massively Parallel Signature Sequencing). De acordo com Auer and Doerge (2010), Illumina's Genome Analyzer, Applied Biosystems' SOLiD e 454 Genome Sequencer FLX são as três soluções mais adotadas comercialmente. Em vez de detalharmos cada uma destas tecnologias neste trabalho, vamos nos focar na tecnologia criada pela Illumina. Detalhes sobre as outras duas tecnologias podem ser obtidos em Ansorge (2009).

Muitas das vezes em que realizamos experimentos a respeito de genomas, estamos interessados em comparar a expressão gênica de um ou mais grupos de seres vivos. Mas o que vem a ser expressão gênica? Alberts et al. (2002) define a expressão gênica como o processo em que a informação de um gene é utilizada na síntese de um produto gênico. Na grande maioria das vezes, isto significa transformar um ou mais aminoácidos em proteínas.

Em vez de medir a expressão gênica diretamente, a tecnologia criada pela Illumina isola o RNA mensageiro (mRNA) das células e replica-o aleatoriamente. Um processo chamado sonificação, fragmenta este mRNA aleatoriamente em milhões de pedaços, que são transcritos para DNA complementar (cDNA). Entretanto, apenas fragmentos de um certo tamanho, que varia entre 35 e 300 pares de base, dependendo da tecnologia, são selecionados. Pequenos adaptadores, que em geral são menores do que 20 pares de base, são anexados às extremidades destes fragmentos. Através de uma reação química chamada PCR, descrita em detalhes por Ochman et al. (1988) e Saiki et al. (1988), estes fragmentos são amplificados. Cada etapa de amplificação dobra a quantidade de cDNA dispo-

nível. Este cDNA amplificado é colocado num chip similar ao da Figura 2.3 e, enfim, sequenciado por uma máquina.

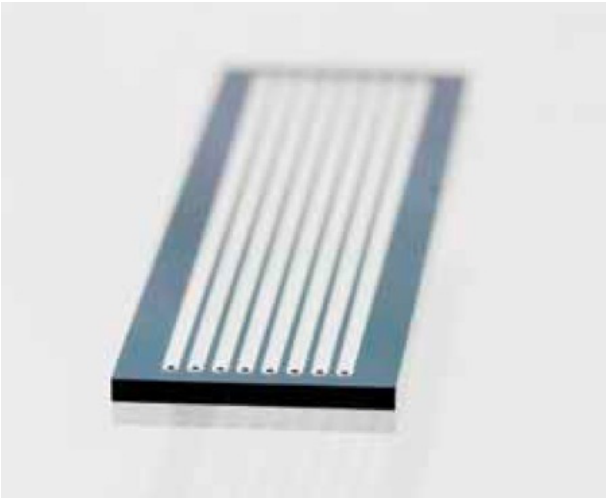

Figura 2.3: Chip da empresa Illumina, com suas 8 faixas que podem receber até 16 sujeitos cada uma.

Após o sequenciamento dos fragmentos de cDNA, estes são alinhados a um genoma de referência. Este genoma de referência tem regiões de interesse anotadas a ele, como genes e exons[1]. Há diversas maneiras de realizar este alinhamento. Uma destas maneiras é

[1]Região do DNA que codifica aminoácidos em proteínas.

através de um programa chamado Bowtie, descrito em Langmead et al. (2009). Note que alinhar um genoma sequenciado a um genoma de referência nada mais é do que encontrar uma ordem correta para os milhões de fragmentos obtidos durante o processo de sequenciamento. O Bowtie é capaz de fazer isto rapidamente, até mesmo num computador pessoal. Outros programas como BWA (Li and Durbin (2009)) e Maq (Li et al. (2008)) são boas alternativas para o Bowtie.

Após os dados obtidos terem sido alinhados, é necessário calcular a expressão gênica digital do experimento. A expressão gênica é definida como o número de fragmentos de cDNA que são alinhados a cada exon do genoma de referência. A Figura 2.4 exemplifica este cálculo para um gene qualquer.

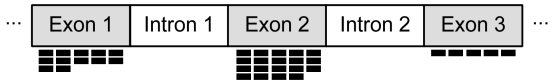

Figura 2.4: Exemplo de contagens alinhadas de dados RNA-Seq. A expressão gênica é definida como o número de fragmentos lidos (pequenos retângulos negros) que são mapeados aos exons (grandes retângulos cinzas) em cada gene. Neste exemplo, a expressão gênica digital é $12 + 19 + 5 = 36$.

Uma maneira de determinar estas contagens é através de um programa capaz de lidar com arquivos do tipo SAM. Uma alternativa popular é o pacote SAMtools, descrito em Li et al. (2009). Dentre as diversas funcionalidades deste programa, coverageBed é

aquela que precisamos para determinar a expressão gênica digital para experimentos de RNA-Seq.

Como os dados que obtemos vem de contagens, é fácil ver que a expressão gênica digital não é uma variável aleatória contínua. No passado, transformaríamos estes dados e usaríamos métodos já conhecidos de análise. Entretanto, com o desenvolvimento atual da Estatística e das novas ferramentas disponíveis, somos capazes de trabalhar com estes dados sem transformá-los.

2.2 Pipeline

Para que possamos testar a diferenciação na expressão gênica de dados provenientes de dados de experimentos de RNA-Seq, primeiro precisamos preparar o output das máquinas de sequenciamento. Há diversas maneiras de fazermos isto, mas elas são idênticas em seu cerne. De maneira geral, um pipeline para um experimento RNA-Seq consiste de cinco passos:

1. Preparação da amostra
2. Sequenciamento
3. Controle de qualidade
4. Alinhamento das leituras
5. Análise e descrição dos resultados

Reconhecemos a importância dos passos 1 e 2 na análise deste tipo de dados, mas devido a restrições diversas, eles estão fora do

escopo deste trabalho. Passaremos brevemente pelos passos 3 e 4, descrevendo-os superficialmente. Nos focaremos no passo 5 nos capítulos seguintes.

Pipeline é uma expressão inglesa que, no contexto de análise de dados genéticos, significa um conjunto de métodos utilizados para tratar os dados para análise. Existem diversas maneiras de prepararmos dados deste tipo. Para não nos estendermos em demasia, vamos mostrar como converter arquivos do tipo Sequence Read Archive ou Short Read Archive (SRA) nas contagens de genes que serão analisadas posteriormente pelo `R`.

O formato SRA foi definido por for Biotechnology Information (US). Em resumo, é um formato universal utilizado para armazenar dados de sequenciamento genético. Este formato é capaz de guardar as leituras das máquinas em um formato binário não-ambíguo. Atualmente, algumas agências de fomento e revistas científicas de acesso livre exigem que os pesquisadores disponibilizem os dados de sua pesquisa neste formato para que suas pesquisas sejam publicadas.

Entretanto, arquivos SRA são apenas recipientes binários para dados de experimentos RNA-Seq. Desta forma, precisamos extrair seu conteúdo de modo que ele possa ser convenientemente analisado. Felizmente, o NCBI disponibiliza via download gratuito, seu SRA Toolkit através do endereço `http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=toolkit_doc`. Há várias formas de extrair os dados destes arquivos. Cabe ao pesquisador escolher aquela que melhor se adapta às suas necessidades. Por exemplo, para extrair os dados para o formato FASTQ, a rotina a ser utilizadas é a `fastq-dump`.

O formato FASTQ, definido em Cock et al. (2010), é um arquivo texto usado para compartilhar os dados lios por máquinas de sequenciamento. Um das grandes vantagens deste formato é a gravação simultânea tanto da sequência genética, quanto da qualidade de cada base lida. Como é um formato desenvolvido pela empresa Solexa, Inc., é o formato utilizado nas máquinas Illumina. Abaixo vemos um exemplo de um destes arquivos.

```
@SRR014849.1 EIXKN4201CFU84 length=50
GGGGGGGGGGGGGGGCTTTTTTTGTTTGGAACCGAAAGGGTTTTGAAT
+SRR014849.1 EIXKN4201CFU84 length=50
3+&$#""""""""""7F@71,'";C?,B;?6B;:EA1EA 1EA5'9B:
```

```
@título e descrição opcional
linha com o que foi sequenciado
+repetição opcional do título
linha com as qualidades da sequência
```

Após o arquivo FASTQ ser obtido, os milhões de fragmentos presentes em cada arquivo devem ser alinhados a um genoma de referência. Minha escolha de alinhador é o programa Bowtie, descrito em detalhes por Langmead et al. (2009). Os genomas de referência mais recentes podem ser obtidos gratuitamente no UCSC[2] Genome Bioinformatics Site (`http://genome.ucsc.edu/`). O resultados do alinhamento feio pelo Bowtie é um arquivo Sequence Alignment/Map (SAM), que nada mais é do que um arquivo texto.

O programa BEDTools (Quinlan and Hall (2010)) é utilizado para converter os arquivos SAM obtidos até o momento em ar-

[2] University of California, Santa Cruz

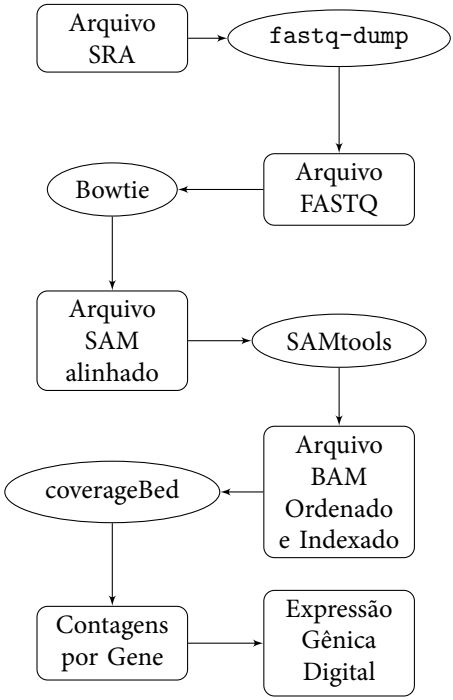

Figura 2.5: Fluxograma mostrando os passos necessários para transformar um arquivo SRA em contagens de genes que podem ser analisadas pelo R.

quivos BAM (Binary Alignment/Map). Este arquivo nada mais é do que a versão compactada do arquivo SAM. Após a obtenção do arquivo BAM, utilizamos o programa SAMTools (Li et al. (2009)) para finalmente converter os alinhamentos obtidos anteriormente em expressões gênicas digitais, que serão analisadas pelo R.

O procedimento completo está resumido em um fluxograma exibido na Figura 2.5.

3 Análise de Dados

Neste capítulo veremos diversas características dos experimentos de RNA-Seq. Vamos ver como planejar corretamente experimentos que utilizem esta tecnologia e como analisá-los posteriormente da melhor maneira possível.

3.1 Delineamento de Experimentos

Assim como todo estudo científico, experimentos que utilizarão dados de RNA-Seq devem ser planejados corretamente, de modo a obtermos o máximo de informação com o menor custo possível. Auer and Doerge (2010) fazem um excelente apanhado de técnicas que devem ser utilizadas para o delineamento de experimentos neste contexto. Faremos um breve resumo destas técnicas a partir de agora. O leitor familiarizado com a teoria de planejamento de experimentos encontrará diversas similaridades com as características clássicas de um experimento bem planejado.

Experimentos bem controlados e bem planejados são a chave para que possamos obter o máximo de informação a respeito do problema que desejamos atacar. É necessário saber escolher corretamente a população que interessa ao nosso estudo (amostragem), incluir a quantidade correta de sujeitos nele (replicação), verificar

se é possível reunir os sujeitos em grupos de similaridade (agrupamento em blocos) e alocar aleatoriamente os sujeitos nos tratamentos considerados (aleatorização).

Como visto acima, as quatro principais características da coleta e da análise de dados são amostragem, aleatorização, replicação e agrupamento em blocos. Embora estas características sejam simples de definir e entender, nem sempre a sua implementação é fácil de ser realizada. Veremos, a seguir, como definir cada uma delas e como elas se relacionam com o tipo de experimento que nos interessa aqui.

3.1.1 Amostragem

A amostragem de um experimento de RNA-Seq segue ideias similares a de qualquer outro experimento estatístico. Devemos definir claramente qual é nossa população de interesse, de modo a não extrairmos informações desnecessárias a partir do experimento. Após a população ser definida, é necessário definir como uma amostra representativa desta população será obtida. Há diversas maneiras disto ser feito, mas elas fogem do escopo deste texto.

Definir a população é, algumas vezes, uma tarefa óbvia. Por exemplo, se estivermos interessados em saber as intenções de voto para prefeito de Porto Alegre, devemos restringir as entrevistas a pessoas maiores de 16 anos, que possuam título de eleitor e que tenham domicílio eleitoral na cidade. Num cenário assim, de nada adianta entrevistarmos crianças ou americanos nascidos Nova Iorque, pois estes sujeitos não terão influência no resultado da eleição que estamos analisando.

Nos estudos que envolvem RNA-Seq, entretanto, surgem algu-

mas complicações no processo de amostragem. Não basta apenas selecionar os indivíduos certos para serem sequenciados. Também é necessário determinar de onde, em cada indivíduo, será retirada a amostra de tecido a ser sequenciada. Talvez não faça sentido retirar uma amostra do tecido do fígado de um sujeito se o experimento visa estudar pessoas que sofrem de câncer na faringe.

Não obstante, cada tecnologia de sequenciamento e cada técnica de purificação de RNA se comporta de maneira diferente de acordo com o tecido a ser sequenciado. Assim, o ideal é realizar estudos-piloto, com amostras pequenas, a fim de determinar qual combinação gera as amostras com melhor qualidade. Após isto ser definido, mantém-se constante, até o final do experimento, a melhor configuração de tecido e técnica de purificação.

3.1.2 Aleatorização

Com a população e a maneira de obter amostras dela definidas, o passo seguinte é a aleatorização da amostra. Geralmente, nos casos de diferenciação genômica, os pesquisadores estão interessado em fazer comparações entre tratamentos diferentes através de contrastes. Para que tenhamos um experimento bem realizado, é importante que os sujeitos sejam atribuídos de maneira aleatória entre os diversos tratamentos considerados. Devemos evitar ao máximo vícios que possam surgir devido à alocação errada de sujeitos em tratamentos.

3.1.3 Replicação

De nada adianta um experimento ser bem amostrado e bem aleatorizado se não houverem replicações suficientes em sua execução. A quantidade de replicações a serem depende do experimento a ser realizado e da quantidade de verba disponível para realizar o experimento. De nada adianta determinar que 100 organismos serão sequenciados se não há dinheiro (ou até mesmo tempo) disponível para tal. São dois os principais tipos de replicação existentes em experimentos de RNA-Seq.

O primeiro deles, e que é mais similar àquilo que entendemos como replicação no sentido tradicional do termo, é chamado de replicação biológica. Temos replicações biológicas quando um mesmo organismo possui mais de um sujeito presente no experimento. Por exemplo, se estamos sequenciando um grupo de 10 humanos para estudar os componentes genéticos do Mal de Alzheimer, cada uma das pessoas deste grupo é uma replicação biológica.

O segundo tipo de replicação é chamado de replicação técnica. Devido à grande variabilidade presente nas tecnologias de sequenciamento atuais, devemos levar em conta que nem sempre teremos as mesmas leituras ao sequenciar um organismo. Por isso, é interessante sequenciar mais de uma vez o mesmo sujeito selecionado para o experimento, a fim de identificar os erros de medição dos equipamento utilizados. Portanto, a replicação técnica nada mais é do que o ressequenciamento de um sujeito presente no experimento.

Pensando novamente a respeito de um experimento hipotético sobre o Mal de Alzheimer, se tivermos 10 sujeitos no experimento, onde o genoma de cada um foi sequenciado 3 vezes, teremos 10 re-

plicações biológicas e 3 replicações técnicas por sujeito. Saber identificar as fontes de variação biológica e técnica de um experimento é importante para que os dados sejam analisados posteriormente.

Estudos observacionais sem replicações biológicas foram comuns durante um bom tempo (Marioni et al. (2008)). Entretanto, este tipo de estudo não possui a liberdade de um experimento controlado. Por exemplo, não é possível, em estudos observacionais, determinar quais sujeitos estarão nos grupos de tratamento e de controle. Por este motivo, não é possível estimar a variabilidade intragrupos nos casos em que não há replicação nos dados, prejudicando assim a sua análise e a generalização dos resultados obtidos.

3.1.4 Agrupamento em Blocos

A Figura 3.1 representa, esquematicamente, um chip da tecnologia Illumina. Como vimos anteriormente, estes chips possuem oito faixas. Graças a marcadores personalizados, cada faixa pode receber material genético de mais de um sujeito para ser sequenciado.

Em estatística, o agrupamento por blocos é um recurso utilizado para reduzir a variabilidade nos dados. Através dele, o pesquisador procura incluir um fator no modelo que reúna os sujeitos presentes no experimento em grupos. Estes grupos são criados a partir da similaridade entre eles. Por exemplo, ao testar quatro compostos diferentes de pneus, podemos considerar que os carros utilizados no experimento são blocos.

De modo a reduzir a variabilidade na análise de dados de experimentos RNA-Seq, podemos agrupar os genomas em blocos. Em geral, os chips e as faixas são as duas maneiras mais comuns de agrupar tais dados durante sua análise. Embora esperemos que

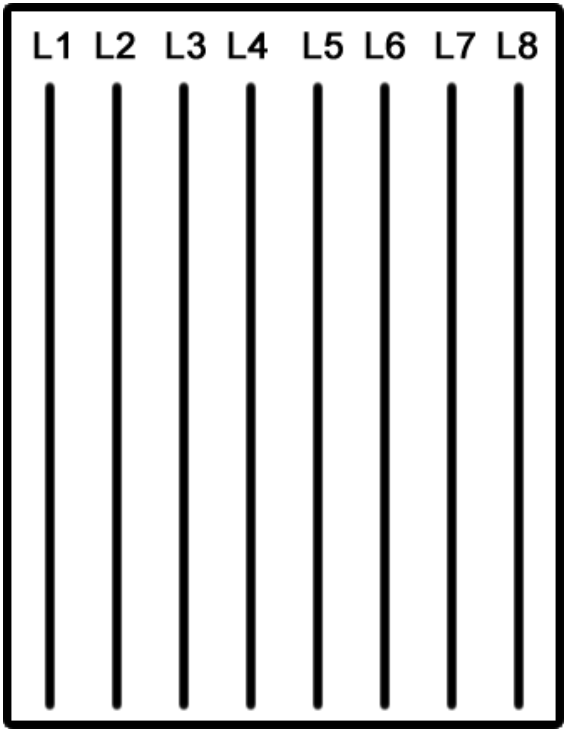

Figura 3.1: Representação esquemática de um chip Illumina e suas 8 faixas.

nem os chips e nem as suas faixas influenciem muito no resultado dos sequenciamentos, nem sempre podemos garantir este resultado na prática.

Devido à natureza deste tipo de experimento, em geral consideramos o delineamento por bloco incompleto equilibrado como a maneira mais adequada para analisar os resultados de sequenciamento obtidos.

3.2 Modelagem

A análise de dados de contagem não é simples. Por serem de natureza discreta, as ferramentas mais comuns de modelagem estatística não são capazes de lidar com este tipo de dado. Por isso, novos métodos tiveram que ser desenvolvidos a fim de que experimentos de RNA-Seq pudessem ser analisados corretamente.

Os principais métodos de análise deste tipo de dados são baseados em Modelos Lineares Generalizados (MLG). Veremos a seguir como utilizar o MLG para analisar dados discretos e, consequentemente, dados provenientes de experimentos RNA-Seq. Também veremos com alguns detalhes a teoria por trás de um dos métodos mais utilizados atualmente neste tipo de pesquisa e que está disponível no pacote `edgeR` do `R`.

3.2.1 Modelos Lineares Generalizados

O MLG é uma extensão do Modelo Linear baseado na distribuição normal. De acordo com McCullagh and Nelder (1989), há três componentes principais em um MLG:

- uma distribuição de probabilidade, da família exponencial, para o vetor resposta $\boldsymbol{Y}$
- um preditor linear para a esperança $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$, que especifica as variáveis explicativas do modelo
- uma função de ligação $g(\cdot)$ que relaciona $\boldsymbol{\eta}$ e $\boldsymbol{\mu}$ tal que $\boldsymbol{\eta} = g(\boldsymbol{\mu})$

Perceba que se $g(\cdot)$ for a função identidade, o método de mínimos quadrados torna-se um caso particular do MLG. Desta forma, o MLG se reduz à forma canônica e pode ser entendido como o método de mínimos quadrados tradicional.

O vetor com n observações independentes $(Y_1, \cdots, Y_n)$, representado por $\boldsymbol{Y}$, pertence à família exponencial. Sua função densidade de probabilidade pode ser expressa como

$$f(y|\boldsymbol{\theta}, \phi) = \exp\left\{\frac{y_i\theta_i - \kappa(\theta_i)}{\alpha_i(\phi)} + c(y_i|\phi)\right\}. \tag{3.1}$$

Na equação (3.1), $\kappa(\theta_i)$ é o cumulante, $\alpha(\phi)$ é o parâmetro de escala e $c(y_i|\phi)$ é o parâmetro de normalização, que garante que a integral da função densidade de probabilidade seja 1. Uma característica destes modelos é

$$\begin{aligned}\kappa'(\theta_i) &= \mathrm{E}(Y)\\ \kappa''(\theta_i) &= \mathrm{Var}(Y)\end{aligned}$$

Para a distribuição de Poisson,

$$
\begin{aligned}
f(Y_i|\lambda_i) &= \frac{\lambda_i}{Y_i!}\mathrm{e}^{-\lambda_i} \\
&= \frac{\exp\left[Y_i \log(\lambda_i) - \lambda_i\right]}{Y_i!} \\
&= \frac{1}{Y_i!}\exp\left(-\lambda_i\right)\exp\left[Y_i \log(\lambda_i)\right]. \qquad (3.2)
\end{aligned}
$$

Na equação (3.2), $h(Y_i) = 1/Y_i!$, $t(Y_i) = Y_i$, $c(\theta) = c(\lambda_i) = 1$, and $w(\theta) = w(\lambda_i) = \log(\lambda_i)$. McCullagh and Nelder (1989), ao definirem os Modelos Lineares Generalizados, usam esta família para desenvolver sua teoria.

As distribuições de probabilidade mais usadas para modelar dados de contagem são a distribuição de Poisson e a Binomial Negativa. Sabemos que se $X \sim \text{Poisson}(\lambda)$, então sua função massa de probabilidade é dada por

$$
f(x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}, \qquad y = 0, 1, \cdots, \qquad \lambda > 0.
$$

Além disso, podemos mostrar que

$$
\begin{aligned}
\mathrm{E}(X) &= \lambda \\
\mathrm{Var}(X) &= \lambda.
\end{aligned}
$$

Como podemos ver acima, a esperança e a variância de uma variável aleatória de Poisson são iguais. Nem sempre os dados provenientes de experimentos reais respeitam esta limitação. Ou seja,

por definição, a distribuição de Poisson acaba limitando sua aplicação na prática. Além disso, obrigatoriamente a sua variância vai aumentar proporcionalmente de acordo com a média, que é outro fator limitante para o tipo de análise que pretendemos fazer.

Quando a variância dos dados é maior do que sua média, dizemos que os dados são sobredispersos. Além disso, devido à natureza dos dados de experimentos RNA-Seq, muitos zeros são gerados. Como vimos anteriormente, estas contagens dizem respeito à expressão gênica. Assim, em alguns casos, não sabemos se tais zeros se referem a genes que não foram lidos durante o sequenciamento ou se os genes não possuem expressão alguma.

Graças a esta restrição na relação entre a média e a variância da distribuição de Poisson, precisamos procurar uma distribuição de probabilidade que seja mais versátil. Vários trabalhos, como Plackett (1981), McCullagh and Nelder (1989), Dobson and Barnett (2008), Madsen and Thyregod (2010) e Hilbe (2011), sugerem que dados de contagem que apresentam sobredispersão sejam modelados através de uma binomial negativa. Embora métodos capazes de lidar com sobredispersão sejam conhecidos há bastante tempo, muitos autores acabaram propondo modelos baseados na distribuição de Poisson para ajustar este tipo de dado. Como exemplo disso podemos citar Marioni et al. (2008), Auer and Doerge (2010), Auer (2010) e Blekhman et al. (2010).

A distribuição Binomial Negativa é uma maneira de modelar dados sobredispersos. Como a distribuição de Poisson, a Binomial Negativa é útil para modelar dados de contagem. Mas ao contrário do que ocorre com a Poisson, a variância da Binomial Negativa não é limitada por sua média. Assim, esta distribuição consegue lidar de maneira mais eficaz com dados sobredispersos.

Seja Y uma variável aleatória com distribuição Binomial Negativa. Assim, Y modela o número de sucessos y numa sequência de tentativas de Bernoulli independentes, até que o r-ésimo fracasso ocorra. A função massa de probabilidade de Y é dada por

$$f(y|r,p) = \binom{r+y-1}{y} p^y (1-p)^r, \qquad y = 0, 1, \cdots, \qquad 0 \leq p \leq 1, \tag{3.3}$$

onde y é o número de sucessos até observarmos r fracassos durante a realização do experimento e p é a probabilidade de sucesso para cada tentativa de Bernoulli. Podemos reparametrizar (3.3) definindo dois novos pariametros $\mu = \mathrm{E}(Y)$ e $\phi = 1/r$, chamados de média e parâmetro de dispersão, respectivamente. Esta reparametrização pode ser justificada utilizando a equação (3.3) reescrita na sua forma exponencial, dada por (3.1):

$$f(y|p,r) = \exp\left\{ y\log(1-p) + r\log(p) + \log\binom{y+r-1}{r-1} \right\},$$

onde $y\log(1-p)$ e $(\theta_i) = r\log(p)$ são a função de ligação e o cumulante, respectivamente. Podemos definir

$$\begin{aligned} \theta = \log(1-p) &\rightarrow p = 1 - \exp(\theta) \\ \kappa(\theta) = -r\log(p) &\rightarrow \kappa(\theta) = -r\log(1-\exp(\theta)). \end{aligned}$$

Diferenciando $\kappa(\theta)$ com respeito a θ, temos o valor esperado de Y, dado por

$$\begin{aligned}\kappa'(\theta) &= \frac{\partial \kappa}{\partial p}\frac{\partial p}{\partial \theta} \\ &= -\frac{r}{p}(-(1-p)) \\ &= \frac{r(1-p)}{p}\end{aligned} \tag{3.4}$$

e sua variância, dada por

$$\begin{aligned}\kappa''(\theta) &= \frac{\partial^2 \kappa}{\partial p^2}\left(\frac{\partial p}{\partial \theta}\right)^2 + \frac{\partial \kappa}{\partial p}\frac{\partial^2 p}{\partial \theta^2} \\ &= \frac{r}{p^2}(1-p)^2 + \frac{-r}{p}(1-(1+p)) \\ &= \frac{r(1-p)}{p^2}.\end{aligned} \tag{3.5}$$

De modo a ter um modelo mais prático para se trabalhar, podemos reparametrizar p e r utilizando (3.4):

$$\begin{aligned} & \tfrac{r(1-p)}{p} && = \mu \\ \Leftrightarrow \quad & \tfrac{1-p}{p} && = \frac{\mu}{r} \\ \Leftrightarrow \quad & \tfrac{1}{p} - 1 && = \frac{\mu}{r} \\ \Leftrightarrow \quad & \tfrac{1}{p} && = 1 + \frac{\mu}{r}\end{aligned}$$

$$
\begin{aligned}
\Leftrightarrow \quad & \tfrac{1}{p} &&= \frac{r+\mu}{r} \\
\Leftrightarrow \quad & p &&= \frac{r}{r+\mu} \\
\Leftrightarrow \quad & p &&= \frac{r}{r\left(1+\frac{\mu}{r}\right)} \\
\Leftrightarrow \quad & p &&= \frac{1}{1+\phi\mu},
\end{aligned}
$$

onde $\phi = 1/r$. Podemos substituir os novos parâmetros μ e ϕ em (3.3) para obter

$$
\begin{aligned}
f(y|\mu,\phi) &= \frac{\Gamma(y+\phi^{-1})}{\Gamma(\phi^{-1})\Gamma(y+1)}\left(\frac{1}{\phi\mu+1}\right)^{\phi^{-1}}\left(1-\frac{1}{\phi\mu+1}\right)^{y} \\
&= \frac{\Gamma(y+\phi^{-1})}{\Gamma(\phi^{-1})\Gamma(y+1)}\left(\frac{1}{\phi\mu+1}\right)^{\phi^{-1}}\left(\frac{\mu}{\phi^{-1}+\mu}\right)^{y} \qquad (3.6)
\end{aligned}
$$

De (3.4) e (3.5), podemos mostrar que sua esperança e variância são dadas por

$$
\mathrm{E}(Y) = \frac{r(1-p)}{p} = \mu \qquad (3.7)
$$

$$
\mathrm{Var}(Y) = \frac{r(1-p)}{p^2} = \mu + \phi\mu^2, \qquad (3.8)
$$

respectivamente, de acordo com a parametrização dada em (3.6). Portanto, em vez de utilizarmos os parâmetros r e p, a função massa de probabilidade da distribuição Binomial Negativa pode ser definida através da sua esperança e variância, dadas por (3.7) e (3.8),

respectivamente. Como $\text{Var}(Y) = \mu + \phi\mu^2$, dizemos que $\phi \in (0, \infty)$ é o parâmetro de dispersão. Perceba que se $\phi \to 0, Y \to X$, onde $X \sim \text{Poisson}(\mu)$. Portanto, a distribuição de Poisson é um caso especial da distribuição Binomial Negativa.

Se $y_i, i = 1, 2, \cdots, n$ é uma amostra i.i.d. $f(y|\mu, \phi)$, definida em (3.6), podemos escrever $\mathcal{L}(\mu_i; y, \phi)$, sua função de log-verossimilhança, como

$$
\begin{aligned}
L(y_i|\mu, \phi) &= \log(\prod_{i=1}^{n} f(y_i|\mu, \phi)) \\
&= \log\left\{\prod_{i=1}^{n} \frac{\Gamma(\phi^{-1})}{\Gamma(\phi^{-1})\Gamma(y_i+1)} \left(\frac{1}{1+\phi\mu}\right)^{\phi^{-1}} \left(\frac{\phi\mu}{1+\phi\mu}\right)^{y_i}\right\} \\
&= \sum_{i=1}^{n} \left\{y_i \log\left(\frac{1}{1+\mu\phi}\right) - \phi^{-1}\log(1+\phi\mu)\right. \\
&\quad \left. + \ + \log\Gamma(y_i+\phi^{-1})\log\Gamma(y_i+1) + \log\Gamma(\phi^{-1})\right\} \\
&= \sum_{i=1}^{n} \{y_i \log(\mu\phi) - (y_i+\phi^{-1})\log(1+\phi\mu) \\
&\quad + \log\Gamma(y_i+\phi^{-1}) - \log\Gamma(y_i+1) - \log\Gamma(\phi^{-1})\}
\end{aligned}
$$

A maneira mais eficaz de estimar os parâmetros dos MLGs é através de uma algoritmo chamado IRLS (Iteratively Reweighted Least Squares - Mínimos Quadrados Reponderados Iterativamente). A ideia básica deste procedimento é dada abaixo:

1. Inicializar $\boldsymbol{\beta}$, o vetor com os parâmetros do modelo, e o preditor linear $\boldsymbol{\eta}$

2. Calcular os pesos $W^{-1} = \boldsymbol{V} g'(\boldsymbol{\beta})$, onde $\boldsymbol{V}$ é a variância dada por $\kappa''(\boldsymbol{\theta})$, onde $\boldsymbol{\theta} = [\boldsymbol{\beta}, \phi]$
3. Calcular $\boldsymbol{Z} = \boldsymbol{\eta} + (\boldsymbol{Y} - \boldsymbol{\beta}) g'(\boldsymbol{\beta})$
4. Regressar $\boldsymbol{Z}$ nos preditores $x_1, x_2, \cdots, x_n$ com pesos W para obter atualizações para $\boldsymbol{\beta}$
5. Calcular $\boldsymbol{\eta}$ baseado nas estimações da regressão
6. Calcular $\boldsymbol{\beta}$ como $g^{-1}(\boldsymbol{\eta})$
7. Calcular a função de log-verossimilhança
8. Iterar até a convergência

Variações deste método podem ser encontradas em todas as principais implementações do algoritmo IRLS (Hilbe (2011)), embora possam existir pequenas variações entre elas.

Utilizamos o Teste de Razão de Verossimilhanças (TRV) para realizar inferências a respeito da diferença na expressão gênica. Este método tem sido usado há anos para comparar modelos encaixados. Um destes modelos, chamado alternativo, terá mais parâmetros, enquanto o outro modelo, chamado nulo, terá menos parâmetros. Outro uso para os Testes de Razão de Verossimilhanças é testar, dentro dois diferentes modelos, com distribuições diferentes (*e.g.*, Poisson e Binomial Negativa), possuem um ajuste melhor aos dados. O TRV também é útil para decidir se devemos adicionar ou remover preditores do modelo ajustado.

A intuição por trás dos Testes de Razão de Verossimilhanças se baseiam no fato de que, para uma certa realização de um certo processo, possui uma chance de ser melhor modelada por um modelo

ou outro. Quando tomamos a razão das log-verossimilhanças destes modelos, calculamos quão mais provável os dados estão de serem melhor ajustados por um modelo do que sob o outro.

Suponhamos que $\boldsymbol{\theta}$ seja o vetor de parâmetros do modelo que estamos ajustando aos dados. Definimos o conjunto ω como o espaço nulo e o conjunto Ω, complementar de ω, como o conjunto alternativo. Portanto, $\omega \cup \Omega$ é o espaço completo, com todos os parâmetros possíveis. As hipóteses sendo testadas são

$$\begin{aligned} H_0 &: \boldsymbol{\theta} \in \omega \\ H_A &: \boldsymbol{\theta} \in \Omega \end{aligned}$$

A justificativa por trás dos TRV pode ser entendida através de um exemplo dado em Casella and Berger (2001). Seja $f(x|\theta)$ a função massa de probabilidade de uma variável aleatória discreta. Seja a estatística do teste de razão de verossimilhança $\Lambda(\boldsymbol{X})$ definida por

$$\Lambda(\boldsymbol{X}) = \frac{\sup_{\theta \in \omega} L_0(\theta|\boldsymbol{X})}{\sup_{\theta \in \Omega} L_A(\theta|\boldsymbol{X})}, \tag{3.9}$$

onde L_0 e L_A são as verossimilhanças sob as hipóteses nula e alternativa, respectivamente.

Pela definição da máxima verossimilhança, o numerador de (3.9) é a probabilidade máxima da amostra observada, sob a hipótese nula. O denominador de (3.9)é a probabilidade máxima da amostra observada sob o espaço de todos os parâmetros. Em outras palavras, a verossimilhança no numerados é calculada sobre os parâmetros presentes em H_0, enquanto a verossimilhança no denominador é calculada sobre todos os parâmetros presentes.

Se a razão destes dois máximos é pequena, então há pontos no espaço da hipótese alternativa para os quais a amostra observada é muito mais provável do que qualquer ponto no espaço nulo. Nesta situação, o critério do TRV afirma que H_0 deve ser rejeitada em favor de H_A.

No contexto dos MLGs, o principal interesse é testar as hipóteses

$$\begin{array}{ll} H_0: & \beta_q = \beta_{q+1} = \cdots = \beta_{p-1} \\ H_A: & \text{nem todos os } \beta_k \text{ em } H_0 \text{ são iguais} \end{array}$$

onde, por conveniência, o modelo é ajustado de tal forma que os últimos $l - h$ coeficientes do modelo de regressão são aqueles que serão testados. O modelo em H_0 é definido como o modelo reduzido e o modelo em H_A é definido como modelo completo.

A expressão (3.9) pode ser reescrita como

$$\Lambda(\boldsymbol{X}) = -2 \log \left(\frac{L_0(\theta|\boldsymbol{X})}{L_A(\theta|\boldsymbol{X})} \right),$$

ou, ainda,

$$\Lambda(\boldsymbol{X}) = -2 \left(\mathcal{L}_0(y|\hat{\boldsymbol{\theta}}) - \mathcal{L}_A(y|\hat{\boldsymbol{\theta}}) \right),$$

onde $\mathcal{L}_0(y|\hat{\boldsymbol{\theta}})$ e $\mathcal{L}_A(y|\hat{\boldsymbol{\theta}})$ são os máximos das log-verossimilhanças sob os espaços nulo e alternativo, respectivamente. Para amostras grandes,

$$\Lambda(\boldsymbol{X}) \xrightarrow{D} \chi^2_p$$

onde s é a diferença no número de parâmetros entre os modelos nulo e alternativo.

3.2.2 edgeR

Atualmente, um dos métodos mais utilizados para realizar a análise de diferenciação genômica é a máxima verossimilhança condicional ajustada pelos quantis. Este método está implementado no pacote `edgeR` do `R`, disponível no site do projeto Bioconductor - `http://www.bioconductor.org/`. Este pacote executa a análise de diferenciação para qualquer característica genética, como genes, exons, transcrições ou intervalos genéticos gerais. Normalmente, genes são utilizados neste tipo de análise. Este método é descrito com mais detalhes em Robinson et al. (2010).

De maneira similar aos mais recentes métodos estatísticos, o pacote `edgeR` assume que as contagens dos genes seguem a distribuição Binomial Negativa. Este método utiliza um estimador de máxima verossimilhança condicional ajustada pelos quantis (CML)para o parâmetro de dispersão da distribuição Binomial Negativa. Para fazer isto, os autores assumiram que se todas as amostras i no experimento possuem o mesmo tamanho (*i.e.*, $m_i = m$), a soma $Z = Y_1 + Y_2 + \cdots + Y_k \sim \mathrm{NB}(km\lambda, \phi k^{-1})$ é verdadeira. Condicionando a verossimilhança em Z e tomando seu logaritmo natural, temos

$$\mathcal{L}(z|, \phi) = \left[\sum_{i=1}^{k} \log \Gamma(y_i + \phi^{-1})\right] + \log \Gamma(n\phi^{-1}) - \log \Gamma(z + k\phi^{-1}) - k \log \Gamma(\phi^{-1}). \quad (3.10)$$

Com a equação (3.10) é possível construir um método de estimação para o parâmetro ϕ. Entretanto, a hipótese do mesmo tamanho para as bibliotecas i no experimento não é sempre verdadeira. Neste caso, eles também propuseram um método para igualá-las. Este método é capaz de gerar "pseudo-dados" ajustados pelos quantis, permitindo assim a estimação do parâmetro ϕ.

Seja $m^* = \left(\prod_{i=1}^{k} m_i\right)^{\frac{1}{k}}$ a média geométrica dos tamanhos das bibliotecas. Os dados observados são ajustados como se eles tivessem sido todos amostrados a partir de uma distribuição NB$(m^*\lambda, \phi)$, utilizando o seguinte algoritmo:

1. Encontre ϕ, o estimador CML que maximiza (3.10)
2. Dada a estimativa de ϕ, estime λ
3. Assumindo que $y_i \sim$ NB$(m_i\lambda, \phi)$, calcule os percentis observados

$$p_i = P(Y < y_i | m_i\lambda, \phi) + \frac{1}{2}P(Y = y_i | m_i\lambda, \phi),$$
$$i = 1, 2, \cdots, k$$

4. Utilizando a interpolação linear das funções dos quantis, gere pseudo-dados de uma distribuição NB$(m^*\lambda, \phi)$, com quantis p_i
5. Calcule ϕ utilizando a CML nos pseudo-dados
6. Repita os passos 2 a 5 até ϕ convergir

É possível, através do ajustamento por quantis, definir um teste exato. Para um teste com dois grupos A e B, definimos Z_{tA} e Z_{tB} como as somas das pseudo-contagens destes grupos, sobre o número de amostras k_A e k_B. Sob a hipótese nula,

$$Z_{tl} \sim \text{NB}(n_l m^* \lambda_t, \phi n_l^{-1}), \qquad l \in \{A, B\}.$$

Assim, é possível construir um teste exato similar ao teste exato de Fisher para tabelas de contingência.

Condicionando na soma das pseudo-contagens totais, $Z_{tA} + Z_{tB}$ também é uma variável aleatória Binomial Negativa. Assim, o cálculo da probabilidade de observarmos classes que são tão ou mais extremas do que aquelas que foram observadas é trivial. Portanto, o p-valor para o teste exato bilateral é definido como a soma total das probabilidades das classes que não são mais prováveis do que aquelas observadas.

3.3 MA Plot

O MA Plot é uma aplicação do gráfico de Bland-Altman Bland and Altman (1986) em estudos genéticos. Durante muito tempo foi utilizado para representar visualmente as microarrays de duas cores transformadas nas escalas M (logaritmos) e A (média).

Entretanto, o uso original do gráfico de Bland-Altman envolvia analisar o quão concordantes estavam duas replicações de um mesmo experimento. Embora seja possível, com um gráfico de dispersão, verificar se duas replicações de um experimento estão correlacionadas, não é possível, por exemplo, detectar diferenças sistemáticas entre elas.

Se estamos interessados na certa característica R de um experimento com duas replicações R_1 e R_2, então as coordenadas cartesianas (x, y) do MA Plot são dadas por

$$R(x, y) = \left(\frac{R_1 + R_2}{2}, R_1 - R_2 \right).$$

Se o valor médio das diferenças é significantemente diferente de zero, então há a presença de um vício nas medições realizadas. É muito comum este gráfico possuir limites de concordância a 95% entre as medições, a fim de facilitar a visualização de suas características e a detecção de outliers.

Na Figura 3.2 vemos um exemplo do gráfico de Bland-Altman em ação. O conjunto de dados utilizado é o mesmo utilizado em Bland and Altman (1986). O vemos na figura são as duas medições de débito expiratório máximo instantâneo de 17 pacientes. Note que, aparentemente, não há nenhum tipo de relação entre a diferença e a média das medições analisadas. As três linhas horizontais representadas na figura dizem respeito às quantidades $\overline{d} = R_1 - R_2$ e $\overline{d} \pm 1,96\sigma$, onde σ é o desvio-padrão de $R_1 - R_2$.

3.4 Comparações Múltiplas

Comparações múltiplas é como chamamos o fato de realizarmos duas ou mais inferências simultâneas. No caso de testarmos apenas uma hipótese, definimos uma região de rejeição para controlar a taxa de falsos positivos, conhecidos como Erros do Tipo I, enquanto atingimos o mínimo possível para a taxa de falsos negativos, chamados de Erros do Tipo II.

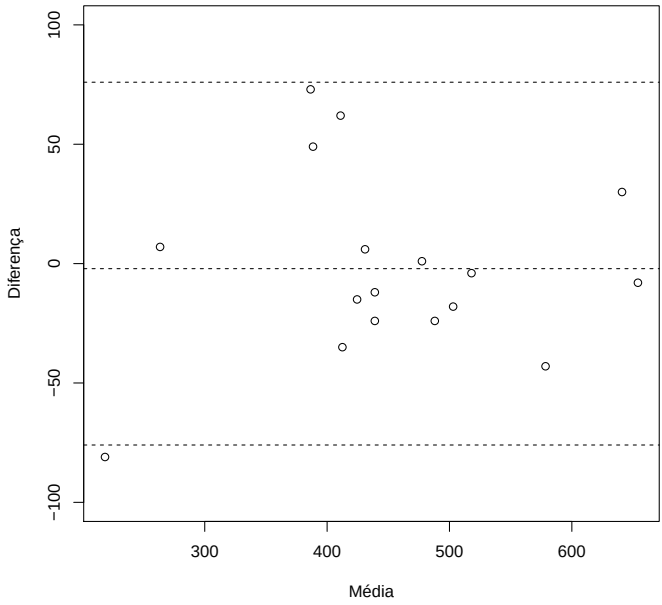

Figura 3.2: Exemplo do gráfico de Bland-Altman.

Para ilustrar como este problema pode ocorrer, imagine que desejamos testar a eficácia de uma nova droga. Para isto, dividimos um certo número de sujeitos em dois grupos. Em destes grupos

será o controle e, o outro, tratamento. Suponha ainda que estamos interessados em testar mais de uma característica entre estes grupos. Neste caso, dado o nível α de significância do teste, há uma chance de $100 \times \alpha\%$ de rejeitarmos a hipótese nula, mesmo que ela seja verdadeira.

Assim, conforme o número de testes aumenta, torna-se cada vez mais provável que os grupos controle e tratamento diferenciem-se em pelo menos uma característica apenas devido à chance.

Desta forma, ao realizarmos múltiplos testes, devemos ajustar o nível de significância α para controlar a taxa de Erro Tipo I para a família de testes. Quando realizamos apenas um teste, geralmente utilizamos $\alpha = \alpha_0$ tal que a probabilidade de um Erro do Tipo I seja menor do que α_0.

O objetivo de métodos de comparação múltipla é encontrar uma maneira de balancear as características opostas de sensibilidade e especificidade. A literatura estatística descreve diversas maneiras de ajustarmos o nível α dos testes de hipóteses. Este tem sido um objeto de interesse de pesquisa desde os anos 1950 (Benjamini and Hochberg (1995)). As origens destes trabalhos podem ser traçadas até os trabalhos de Scheffé e Tukey, com seus limites para intervalos de confiança dois a dois.

Definimos a taxa de erro do tipo I por família (FWER - family-wise type I error rate) como a probabilidade de pelo menos uma rejeição falsa entre m testes de hipóteses. Entre os métodos de comparações múltiplas para FWER, podemos citar o método de Bonferroni como o exemplo mais conhecido. Embora seja ingênuo e muito conservador, foi muito utilizado desde a sua introdução em 1959 (Kutner et al. (2004)). Ao aplicarmos a correção de Bonferroni, o teste é rejeitado se seu p-valor é inferior a α/m. Entretanto,

se m é muito grande, esta correção é muito conservadora. Assim, embora seja um método de fácil aplicação, é pouco utilizado atualmente.

Tornou-se muito comum, em muitas aplicações do mundo real, termos que realizar milhares (ou até mesmo milhões) de testes de hipóteses simultâneos. O poder do teste é crítico nestas aplicações, pois os efeitos mais interessantes estão no limite de detecção. Por este motivo, quando as tecnologias de sequenciamento genético tornaram-se mais populares, os métodos tradicionais de comparações múltiplas começaram a parecer muito conservadoras. No caso de dados de RNA-Seq, com dezenas de milhares de comparações múltiplas, o controle da FWER torna-se crítico.

Para corrigir este excesso de conservadorismo, Benjamini and Hochberg (1995) propuseram um novo método de controlar a taxa de falsos positivos em comparações múltiplas. À época, seu método pareceu mais apropriado do que os que existiam até o momento. Sua justificativa é devida à ideia principal por trás da FWER. Este método controla a probabilidade de observarmos pelo menos um falso positivo, *i.e.*, quase nenhum falso positivo pode ocorrer se a FWER é pequena.

Em Benjamini and Hochberg (1995), o foco do método foi colocado na False Discovery Rate (FDR). Este método, quando comparado com os métodos FWER (Bonferroni, por exemplo), apresenta um menor controle estrito sobre as descobertas falsas. Isto é importante justamente para identificar os poucos efeitos que realmente dentre aqueles muitos que são testados.

A ideia por trás da FDR é a de que um conjunto de predições possui um percentual esperando de falsas predições, *i.e.*, há um certo número de testes que rejeitará a hipótese nula incorretamente. A

FDR é a proporção esperada de hipóteses nulas incorretamente rejeitadas em um dado conjunto de testes.

Para uma série de testes de hipóteses independentes, a FDR é dada por

$$\text{FDR} = \text{E}\left(\frac{V}{V+S}\right),$$

onde V é o número de falsos positivos e S é o número de verdadeiros positivos. O que desejamos é manter o valor da FDR abaixo de α. Entretanto, este número não é calculado facilmente. Geralmente,

$$\text{FDR} = \frac{\text{E}(V)}{S}$$

é usado como um valor aproximado da FDR (Benjamini and Hochberg (1995)).

A Tabela 3.1, reproduzida de Verhoeven et al. (2005), mostra uma maneira de identificar os possíveis resultados de uma série de testes de hipóteses.

Podemos entender a Tabela 3.1 utilizando a seguinte legenda:

- m é o número de hipóteses testadas
- m_0 é o número de hipóteses nulas verdadeiras
- $m - m_0$ é o número de hipóteses alternativas verdadeiras
- V é o número de falsos positivos (ou "false discoveries")
- S é o número de positivos verdadeiros (ou "true discoveries")

Tabela 3.1: Como identificar os possíveis resultados de uma série de testes de hipóteses.

Truth	Decision		Total
	Not Significant	Significant	
Null hypothesis	U	V	m_0
Alternative hypothesis	T	S	$m - m_0$
Total	$m - r$	r	m

- T é o número de falsos negativos
- U é o número de negativos verdadeiros
- r é o número de hipóteses nulas rejeitadas (ou "discoveries")
- em m testes de hipóteses nos quais m_0 deles são hipóteses nulas verdadeiras, $m - m_0$ é uma variável aleatória observável e S, T, U e V são variáveis aleatórias que não podem ser observadas

Para testar a diferenciação na expressão gênica através das m hipóteses nas quais estamos interessados, o procedimento padrão é combinar os p-valores de cada teste num único vetor de p-valores. Após este vetor ser compilado, duas etapas são realizadas:

1. Ordenar os m p-valores calculados do menor para o maior, denominando-os como $p_{(1)}, p_{(2)}, \cdots, p_{(m)}$
2. Encontrar o maior k tal que $p_{(k)} \leq \frac{k}{m}\alpha$

Assumindo que os testes de hipóteses são independentes, este método controla a FDR desejada. Entretanto, se os testes são positivamente correlacionados, este pode ser um método conservador. Se os testes são correlacionados negativamente, devemos encontrar o maior k tal que $p_{(k)} \leq \frac{k}{m(\log(m)+\gamma)}\alpha$, onde $\gamma = 0.5772\cdots$ é a constante de Euler-Mascheroni[1] (Benjamini and Hochberg (1995)).

[1] $\gamma = \lim_{n\to\infty}\left(\sum_{k=1}^{n}\frac{1}{k} - \log(n)\right)$

4 Aplicação

Neste capítulo veremos uma aplicação da teoria desenvolvida nos capítulos anteriores. O conjunto de dados analisado aqui foi disponibilizado por Blekhman et al. (2010). Estas amostras foram obtidas a partir dos fígados de machos e fêmeas de três espécies de primatas: humanos (*Homo sapiens*), chimpanzés (*Pan troglodytes*) e macacos-rhesus (*Macaca mulatta*).

Veremos como fazer uma análise exploratória dos dados a serem analisados, de modo a ter uma ideia geral da sua qualidade e do que podemos esperar de nossa análise. Em seguida, utilizaremos o pacote `edgeR` para procurar diferenças na expressão gênica entre grupos de nosso interesse.

4.1 Análise Exploratória dos Dados

Precisamos visualizar e tratar os dados antes de realizarmos os testes de diferenciação da expressão gênica que nos interessam. Neste exemplo vamos assumir que os dados já foram sequenciados e pré-processados através dos programas vistos no Capítulo 2. Realizaremos aqui apenas a parte da análise que depende de pacotes do R.

A análise exploratória dos dados começa ao percebermos que

haviam 3 replicações biológicas de cada tratamento neste experimento. Cada replicação biológica foi dividida em 2 faixas. Foram coletadas 20689 características de cada uma das 36 faixas, totalizando 71 milhões de leituras de 35 pares de base mapeáveis. As leituras não-processadas podem ser baixadas do Gene Expression Omnibus (GEO) através do conjunto de dados GSE17274. Estas contagens já processadas também podem ser baixadas através do link `http://marcusnunes.me/mgest-2014/GSE17274.txt`.

```
> dados <- read.table(file = "GSE17274.txt",
+     header = TRUE)
> colnames(dados)
```

```
##  [1] "EnsemblGeneID" "R1L1.HSM1"     "R1L2.PTF1"
##  [4] "R1L3.RMM1"     "R1L4.HSF1"     "R1L6.PTM1"
##  [7] "R1L7.RMF1"     "R2L2.RMF2"     "R2L3.HSM2"
## [10] "R2L4.PTF2"     "R2L6.RMM2"     "R2L7.HSF2"
## [13] "R2L8.PTM2"     "R3L1.RMM3"     "R3L2.HSF2"
## [16] "R3L3.PTM1"     "R3L4.RMF3"     "R3L6.HSM3"
## [19] "R3L7.PTF3"     "R3L8.RMM1"     "R4L1.HSM3"
## [22] "R4L2.HSF1"     "R4L3.RMM3"     "R4L4.PTF1"
## [25] "R4L6.PTM2"     "R4L7.RMF3"     "R4L8.HSM2"
## [28] "R5L1.RMF1"     "R5L2.HSM1"     "R5L3.PTF3"
## [31] "R5L4.RMM2"     "R5L8.RMF2"     "R6L2.PTM3"
## [34] "R6L4.PTM3"     "R6L6.PTF2"     "R8L1.HSF3"
## [37] "R8L2.HSF3"
```

A primeira coluna, chamada `EnsembleGeneID`, identifica cada um dos genes dos organismos consideramos neste experimento.

Perceba que cada coluna, a partir da segunda, possui um código associado a ela. Este código nos dá informações a respeito do chip onde o experimento foi realizado e do sujeito analisado. Ele deve ser interpretado da seguinte maneira:

- Os dois primeiros caracteres, com a letra R e um número, referem-se ao *run* do experimento. Ou seja, identifica em qual chip a amostra de material genético estava. `R2` significa, portanto, que a amostra foi colocada no segundo chip utilizado neste experimento.

- Os dois caracteres seguintes referem-se à faixa (*lane*) do chip onde o material genético foi colocado. Conforme vimos no Capítulo 3, até oito faixas podem ser utilizadas em cada experimento. Desta forma, `L6` informa que o material a ser analisado foi colocado na faixa 6.

- O ponto marca a separação entre as identificações do chip e do sujeito.

- As duas letras a seguir indicam a espécie do sujeito. Como vimos anteriormente, este estudo analisou humanos (*Homo sapiens*), chimpanzés (*Pan troglodytes*) e macacos-rhesus (*Macaca mulatta*). As iniciais do gênero e da espécie são utilizadas para identificar a espécie de cada sujeito. Portanto, `RM` indica um macaco-rhesus.

- Os caracteres finais indicam se o sujeito é macho (M) ou fêmea (F). Assim, `M2` indica que este sujeito é o segundo macho desta espécie.

Seguindo esta convenção, o código `R2L6.RMM2` nos informa que esta amostra foi obtida no chip 2, faixa 6 e é do macaco-rhesus macho número 2.

Como vimos no Capítulo 3, a distribuição dos dados de RNA-Seq segue a distribuição Binomial Negativa, o que implica numa assimetria alta. Desta forma, a maioria dos gráficos para sua análise são mais interpretáveis na escala logarítmica. Como não podemos calcular $\log(0)$, e estes dados geram muitas contagens nulas, vamos adicionar $0,25$ aos dados. Assim, $\log_2(0,25) = \log_2(1/4) = -2$. Portanto, todos os zeros do conjunto de dados estarão localizados em -2. Como vamos fazer vários gráficos dos logaritmos destas contagens, é mais prático criar um novo data frame com estes valores, que vamos chamar de `log.dados`.

```
> log.dados <- log2(dados[, 2:37] + 0.25)
```

Através de gráficos, podemos ter uma ideia da qualidade dos dados a serem analisados. Começamos fazendo histogramas de uma faixa do chip para cada espécie considerada no experimento.

```
> par(mfrow = c(2, 3))
> for (i in 1:6) {
+     hist(log.dados[, i], xlab = "log(Contagem)",
+      ylab = "Frequência", main = colnames(log.dados)[i])
+ }
```

Perceba, na Figura 4.1, como as distribuições das contagens são altamente similares através das espécies e como há um predomínio das contagens iguais a zero.

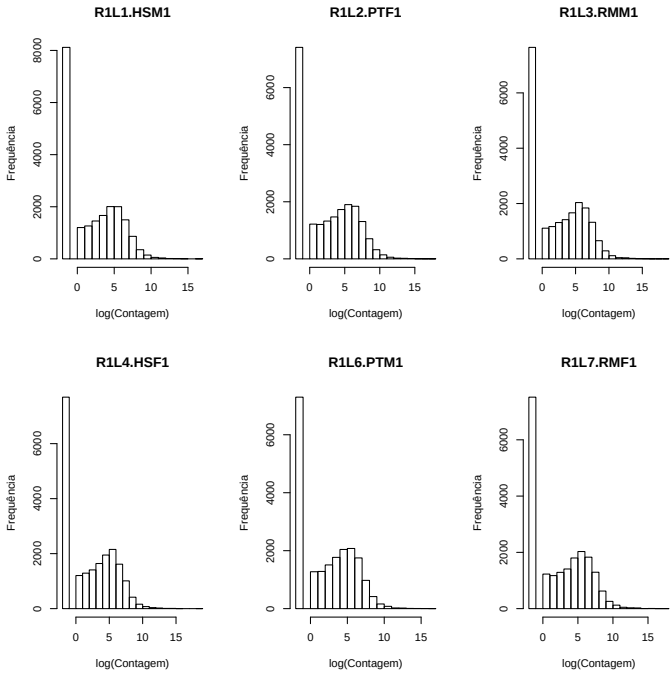

Figura 4.1: Histogramas dos logaritmos das contagens de seis sujeitos do experimento.

A seguir, faremos gráficos de dispersão para avaliar as replicações técnicas do experimento. Vamos comparar as contagens de

duas replicações do experimento para cada espécie e sexo analisada. Novamente, faremos apenas os 6 primeiros gráficos. Note que estamos aplicando uma transparência na cor dos pontos do gráfico para que possamos ter uma ideia da densidade destes pontos.

```
> par(mfrow = c(2, 3))
> repl.bio <- unique(substr(colnames(log.dados)[1:36],
+     6, 9))
> for (i in repl.bio[1:6]) {
+     col.n <- grep(i, colnames(log.dados))
+     plot(log.dados[, col.n[1]], log.dados[,
+         col.n[2]], main = i, xlab = "Repl 1",
+         ylab = "Repl 2", col = rgb(0, 0,
+             0, 0.005), pch = 16)
+ }
```

Note que há uma concentração grande de pontos próximos à origem dos gráficos exibidos na Figura 4.2. À direita da origem, a densidade destes pontos diminui, indicando que densidade de observações diminui, tornando a aumentar logo depois. Perceba também que a dispersão dos dados também é maior próximo à origem e vai diminuindo conforme as contagens aumentam.

Conforme vimos no Capítulo 3, podemos utilizar MA plots para fazer comparações entre duas amostras aleatórias. Note que não é necessário utilizar os logaritmos dos dados neste caso, pois a função `maPlot` calcula estes valores automaticamente.

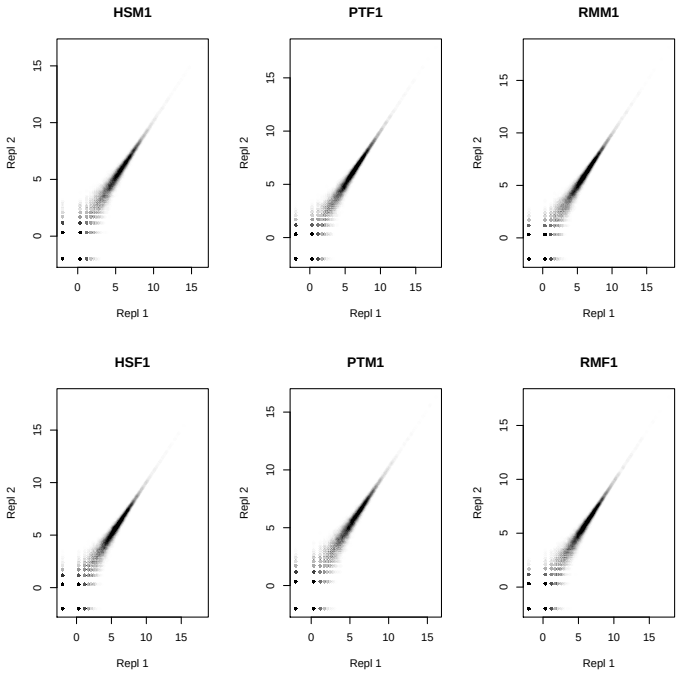

Figura 4.2: Gráfico de dispersão comparando as replicações técnicas de alguns sujeitos do experimento.

```
> library(edgeR)

## Loading required package: limma
## Loading required package: methods
```

```
> par(mfrow = c(2, 3))
> for (i in repl.bio[1:6]) {
+     col.n <- grep(i, colnames(dados))
+     tmp <- maPlot(dados[, col.n[1]], dados[,
+         col.n[2]], main = i)
+     abline(h = 0, col = "white")
+ }
```

Note que a Figura 4.3 exibe as relações entre a concentração e o fold-change através dos genes. Complementando as informações da Figura 4.2, podemos perceber a alta variabilidade dos genes com baixa contagem, enquanto os genes de alta contagem apresentam baixa variabilidade.

Além disso, note que traçamos uma reta horizontal passando por 0. Através dela, percebemos que os valores M estão sistematicamente acima ou abaixo desta linha de referência para valores altos de A. Uma hipótese para este fato seria uma diferença no somatório da contagem total das faixas dos chips. Podemos verificar esta hipótese através de uma normalização destas contagens, dividindo-as pela soma das contagens das faixas do chip.

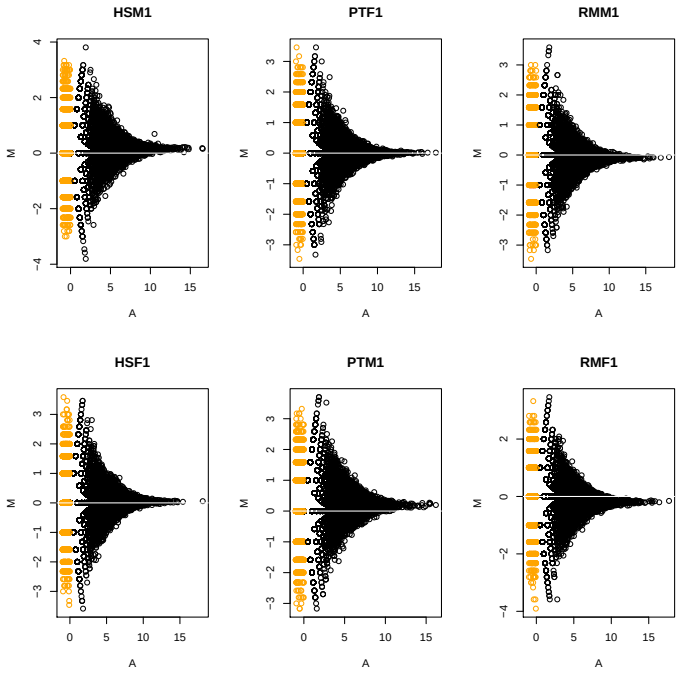

Figura 4.3: MA plots dos primeiros sujeitos, de ambos os sexos, das três espécies.

```
> lib.n <- apply(dados[, 2:37], 2, sum)
> dados.lib <- t(dados[, 2:37])/lib.n
> dados.lib <- t(dados.lib)
> par(mfrow = c(2, 3))
> for (i in repl.bio[1:6]) {
+     colno <- grep(i, colnames(dados.lib))
+     tmp <- maPlot(dados.lib[, colno[1]],
+         dados.lib[, colno[2]], main = i)
+     abline(h = 0, col = "white")
+ }
```

A Figura 4.4 sugere que, de fato, a diferença percebida anteriormente pode ser explicada pela diferença no total de contagens de cada faixa. Assim, vamos somar as duas replicações de cada sujeito antes de proceder com a normalização dos dados. Vamos criar um dendrograma para checar se as duas faixas com as expressões gênicas de cada sujeito são similares. Este dendrograma está exibido na Figura 4.5.

```
> dist <- as.dist(1 - cor(log.dados))
> plot(hclust(dist), main = "Dendrograma",
+     ylab = "Altura")
```

Se as faixas com as replicações dos sujeitos estiverem próximas entre si, podemos assumir que elas são suficientemente similares. Note que é exatamente isto que vemos na Figura 4.5. Portanto, podemos considerar a soma destas replicações em nossa análise.

Ao somar as colunas com as replicações dos sujeitos, vamos aproveitar e eliminar os genes com menos de 10 contagens somadas por

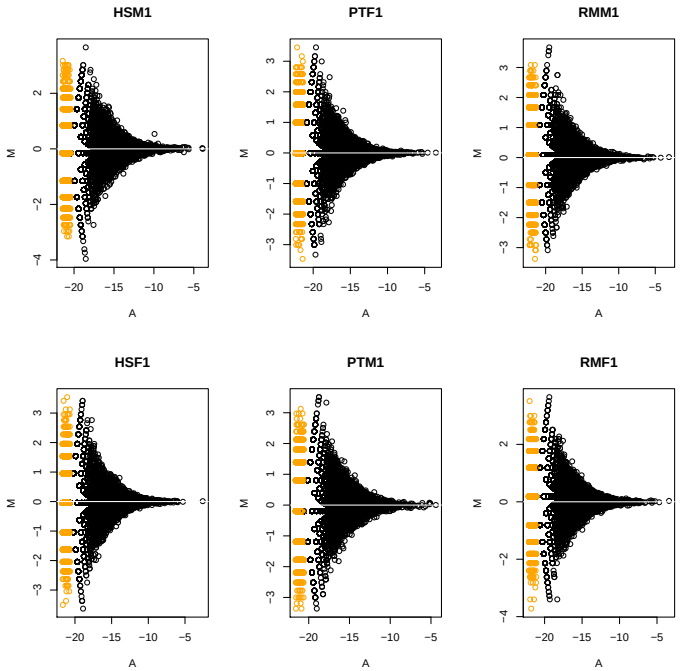

Figura 4.4: MA plots dos primeiros sujeitos, de ambos os sexos, das três espécies, com as contagens normalizadas em respeito à soma das contagens das faixas do chip.

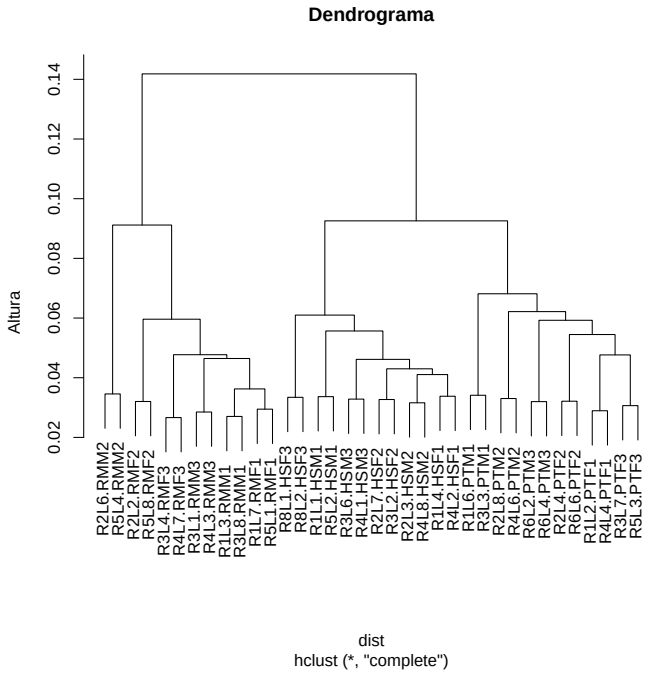

Figura 4.5: Dendrograma das expressões gênicas dos sujeitos do experimento.

todas as faixas. Fazemos isto devido a limitações no método que utilizamos para testar a diferença na expressão gênica. Remover

estes genes com baixas contagens reduzirá o tempo computacional da estimação, aumentará o poder do teste e melhorará da estimação do FDR para os outros genes.

```
> dados.10 <- apply(dados[, 2:37], 1, sum) >
+     9
> dados.total <- as.matrix(dados[dados.10,
+     2:19])
> colnames(dados.total) = repl.bio
> rownames(dados.total) = dados[dados.10, 1]
> for (i in repl.bio) {
+     col.n <- grep(i, colnames(dados))
+     dados.total[, i] <- apply(dados[dados.10,
+         col.n], 1, sum)
+ }
```

A seguir, vamos criar fatores com os tratamentos de casa amostra, para que possamos, posteriormente, realizar os testes de diferenciação da expressão gênica.

```
> tratamentos <- substr(colnames(dados.total),
+     1, 3)
> tratamentos <- factor(tratamentos)
> tratamentos.n <- unique(tratamentos)
```

4.2 Testes de Expressão Gênica

Utilizaremos o pacote edgeR na análise que realizaremos a seguir. Antes de iniciá-la, precisamos transformar a matriz que obtivemos anteriormente em um objeto do tipo DGEList. Fazemos isto porque o pacote que utilizamos exige que os dados de contagem e as informações sobre os tratamentos do experimento constem dentro deste objeto. Além disso, o parâmetro de dispersão ϕ também é estimado com um valor comum para todos os genes através da função estimateCommonDisp. Como efeito colateral, a média geométrica dos tamanhos das bibliotecas é calculada e um conjunto de contagens normalizado é criado.

```
> d <- DGEList(counts = dados.total, group = tratamentos,
+     genes = rownames(dados.total))
> d <- estimateCommonDisp(d)
> d$common.dispersion
```

```
## [1] 0.1749
```

Nós poderíamos fazer um dendrograma similar ao da Figura 4.5 para os dados normalizados, mas o pacote edgeR possui a função plotMDS, mais adequada para exibir graficamente o que estamos desejando verificar. Esta função utiliza uma fórmula para calcular distância mais adequada para dados discretos.

```
> plotMDS(d)
```

Note como os sujeitos da mesma espécies tendem a ficar próximos entre si na Figura 4.6. As distâncias neste gráfico podem ser

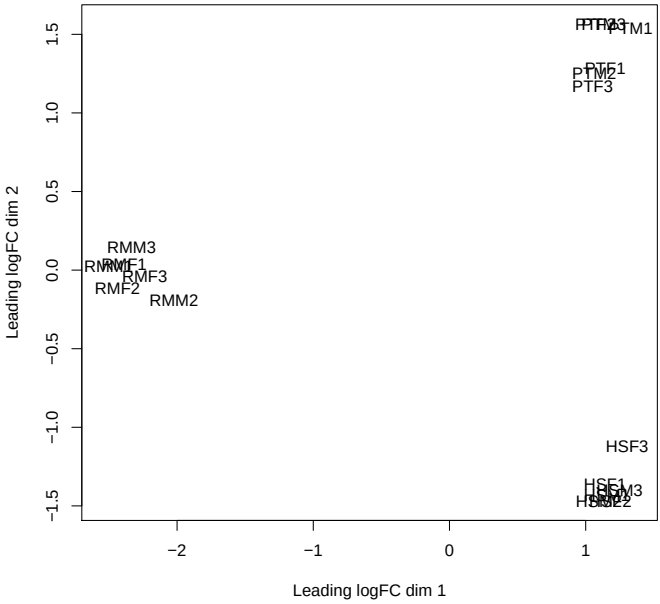

Figura 4.6: Gráfico das relações da amostra baseado em escalonamento multi-dimensional.

interpretadas como mudanças $\log_2$-fold.

Vamos, enfim, proceder à análise da diferenciação gênica entre

grupos. De acordo com Robinson et al. (2010), isto é feito através de um teste exato de Fisher modificado, que leva em conta a dispersão nos dados. Esta dispersão pode ser definida como 0, se estivermos interessados em ajustar um modelo baseado na distribuição de Poisson. Entretanto, parece ser mais indicado que utilizemos o valor da dispersão calculada anteriormente na hora de estimar as estatísticas deste teste.

Para testar a diferença de expressão gênica entre os grupos de machos e fêmeas humanos, fazemos

```
> HSM.vs.HSF <- exactTest(d, pair = c(1, 2))
> HSM.vs.HSF

## An object of class "DGEExact"
## $table
##                    logFC  logCPM PValue
## ENSG00000000003 -0.06484  6.9082 0.8985
## ENSG00000000005  1.79476 -0.5125 1.0000
## ENSG00000000419 -0.35879  4.1865 0.4994
## ENSG00000000457  0.18363  5.0917 0.7215
## ENSG00000000460  0.02462  1.4939 1.0000
## 14976 more rows ...
##
## $comparison
## [1] "HSF" "HSM"
##
## $genes
## [1] "ENSG00000000003" "ENSG00000000005"
## [3] "ENSG00000000419" "ENSG00000000457"
```

```
## [5] "ENSG00000000460"
## 14976 more rows ...
```

Temos, dentro do objeto `HSM.vs.HSF`, o resultado do teste comparando os grupos que nos interessam. Perceba que `table` é o componente que mais nos interessa aqui, pois é dentro dele em que estão os resultados dos testes.

```
> head(HSM.vs.HSF$table)
```

```
##                    logFC  logCPM  PValue
## ENSG00000000003 -0.06484  6.9082 0.89854
## ENSG00000000005  1.79476 -0.5125 1.00000
## ENSG00000000419 -0.35879  4.1865 0.49944
## ENSG00000000457  0.18363  5.0917 0.72150
## ENSG00000000460  0.02462  1.4939 1.00000
## ENSG00000000938 -1.01120  4.4394 0.05387
```

De maneira análoga, para testar a diferença de expressão gênica entre os grupos de machos humanos e chimpanzés, fazemos

```
> HSM.vs.PTM <- exactTest(d, pair = c(2, 4))
> head(HSM.vs.HSF$table)
```

```
##                    logFC  logCPM  PValue
## ENSG00000000003 -0.06484  6.9082 0.89854
## ENSG00000000005  1.79476 -0.5125 1.00000
## ENSG00000000419 -0.35879  4.1865 0.49944
## ENSG00000000457  0.18363  5.0917 0.72150
```

```
## ENSG00000000460  0.02462  1.4939 1.00000
## ENSG00000000938 -1.01120  4.4394 0.05387
```

Através destes outputs, podemos fazer algumas análises a respeito dos genes. Por exemplo, no caso em que comparamos as grupos de machos e fêmeas humanos, não encontramos nenhum gene que indicasse alguma diferença na expressão entre estes grupos. O mesmo ocorre quando interpretamos os resultados da comparação entre machos humanos e chimpanzés: nenhum dos seis primeiros genes apresenta diferença de expressão entre os dois grupos.

Antes que façamos uma análise mais apurada dos resultados obtidos, note que utilizamos um valor comum para a dispersão dos genes. Tal suposição não é factível. Não possuímos nenhum indício biológico que sustente esta característica. Por isso, faremos novamente o ajuste feito anteriormente, mas desta vez utilizando valores de dispersão diferentes entre os genes.

```
> d10 <- estimateTagwiseDisp(d, prior.df = 10)
> TW.HSM.vs.HSF <- exactTest(d10, pair = c(1,
+     2))
> head(TW.HSM.vs.HSF$table)
```

```
##                    logFC  logCPM PValue
## ENSG00000000003 -0.06455  6.9082 0.8861
## ENSG00000000005  1.78073 -0.5125 1.0000
## ENSG00000000419 -0.35694  4.1865 0.3604
## ENSG00000000457  0.18184  5.0917 0.6731
## ENSG00000000460  0.02430  1.4939 1.0000
## ENSG00000000938 -1.01006  4.4394 0.1192
```

```
> TW.HSM.vs.PTM <- exactTest(d10, pair = c(2,
+     4))
> head(TW.HSM.vs.PTM$table)
```

```
##                   logFC  logCPM   PValue
## ENSG00000000003  0.9215  6.9082 0.039618
## ENSG00000000005  0.8848 -0.5125 1.000000
## ENSG00000000419  0.3491  4.1865 0.362794
## ENSG00000000457  1.2639  5.0917 0.003564
## ENSG00000000460 -0.3789  1.4939 0.536722
## ENSG00000000938  0.7353  4.4394 0.258405
```

A opção `prior.df` está relacionada com a suavização no cálculo da estimativa do parâmetro ϕ. Como aqui utilizamos `prior.df=10`, isto significa que as estimativas da dispersão de 10 genes foram utilizados para calcular $\hat{\phi}$. Se tivéssemos utilizado `prior.df=100`, então estimativas da dispersão de 100 genes seriam utilizados para calcular $\hat{\phi}$. Note que quanto maior `prior.df` for, mais suavizadas as estimativas serão e mais próximas elas estarão do valor da dispersão comum.

Com os resultados obtidos acima, podemos listar os genes que possuem a maior diferença entre os grupos para os dois casos que analisamos.

```
> topTags(TW.HSM.vs.HSF, adjust.method = "BH")
```

```
## Comparison of groups:  HSM-HSF
##                  genes  logFC logCPM    PValue
## 6515  ENSG00000138131  5.089 5.3944 2.756e-07
```

```
## 9245  ENSG00000163017 -3.058 3.8267 3.206e-05
## 11722 ENSG00000178297  2.189 3.1773 4.298e-05
## 5185  ENSG00000128285 -2.304 1.9543 7.879e-05
## 11344 ENSG00000175084 -2.163 4.9850 9.342e-05
## 6819  ENSG00000140403 -3.190 4.1706 9.848e-05
## 4457  ENSG00000120694 -2.669 6.3453 1.234e-04
## 12473 ENSG00000185031 -4.979 0.9932 2.070e-04
## 750   ENSG00000060566  1.866 5.7418 4.106e-04
## 8849  ENSG00000160181 -4.296 0.5679 6.616e-04
##            FDR
## 6515  0.004129
## 9245  0.214623
## 11722 0.214623
## 5185  0.245900
## 11344 0.245900
## 6819  0.245900
## 4457  0.264174
## 12473 0.387569
## 750   0.683441
## 8849  0.991113
```

Perceba que pedimos para os p-valores dos genes com maior diferença entre os grupos sejam reportados em ordem crescente. Assim, o topo da lista mostra o gene que apresenta a maior diferença entre os grupos, e esta diferença vai diminuindo conforme vamos descendo na tabela.

Praticamente não encontramos diferenças significativas quando comparamos o grupo de homens com o grupo de mulheres. Apenas o gene ENSG00000138131 apresenta diferença significativa na

expressão gênica entre estes dois grupos. Entretanto, isto não ocorre se considerarmos humanos e chimpanzés machos:

```
> topTags(TW.HSM.vs.PTM, adjust.method = "BH")

## Comparison of groups:  PTM-HSM
##                 genes  logFC logCPM    PValue
## 14119 ENSG00000208587 14.776  7.931 9.663e-64
## 14118 ENSG00000208570 14.097  8.107 2.310e-54
## 8570  ENSG00000157399 -6.777  5.252 1.262e-32
## 5911  ENSG00000134391 -9.228  6.204 2.453e-29
## 399   ENSG00000025423 -6.002  8.026 4.218e-28
## 9325  ENSG00000163444  5.757  4.229 6.610e-26
## 6712  ENSG00000139540 -8.208  6.585 1.885e-24
## 1901  ENSG00000099834 -6.730  6.475 2.650e-23
## 14854 ENSG00000219802  9.927  3.164 1.353e-22
## 14936 ENSG00000220688  4.526  5.146 5.968e-20
##             FDR
## 14119 1.448e-59
## 14118 1.731e-50
## 8570  6.303e-29
## 5911  9.186e-26
## 399   1.264e-24
## 9325  1.650e-22
## 6712  4.034e-21
## 1901  4.963e-20
## 14854 2.252e-19
## 14936 8.940e-17
```

O que percebemos nestes resultados é que, nos casos da comparação entre genes de humanos machos e fêmeas, não encontramos nenhum gene que apresentasse diferença na sua expressão entre os grupos considerados.

Para efeitos de reproducibilidade, seguem as informações a respeito do sistema onde a análise foi realizada.

```
> sessionInfo()

## R version 3.1.1 (2014-07-10)
## Platform: x86_64-apple-darwin13.1.0 (64-bit)
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_U
##
## attached base packages:
## [1] methods   stats     graphics  grDevices
## [5] utils     datasets  base
##
## other attached packages:
## [1] edgeR_3.6.8  limma_3.20.9 knitr_1.6
##
## loaded via a namespace (and not attached):
## [1] evaluate_0.5.5 formatR_1.0    highr_0.3
## [4] stringr_0.6.2  tools_3.1.1
```

Bibliografia

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002). *Molecular Biology of the Cell, fourth edition*. Garland Science.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, 11(10).

Ansorge, W. J. (2009). Next-generation dna sequencing techniques. *New Biotechnology*, 25(4):195–203.

Auer, P. L. (2010). *Statistical design and analysis of next-generation sequencing data*. PhD thesis, Purdue University.

Auer, P. L. and Doerge, R. W. (2010). Statistical design and analysis of rna sequencing data. *Genetics*, 185:405–416.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Bland, J. M. and Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, 327(8476):307–10.

Blekhman, R., Marioni, J. C., Zumbo, P., Stephens, M., and Gilad, Y. (2010). Sex-specific and lineage-specific alternative splicing in primates.

Casella, G. and Berger, R. L. (2001). *Statistical Inference*. Duxbury Press, 2 edition.

Cock, P. J. a., Fields, C. J., Goto, N., Heuer, M. L., and Rice, P. M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic acids research*, 38(6):1767–71.

Dobson, A. and Barnett, A. (2008). *An introduction to generalized linear models*. Texts in statistical science. CRC Press.

for Biotechnology Information (US), B. M. N. C. (2010-). Sra handbook [internet]. URL `http://www.ncbi.nlm.nih.gov/books/NBK47537/`. Accessed on February 28th, 2013.

Green, E. D. and Guyer, M. S. (2011). Charting a course for genomic medicine from base pairs to bedside. *Nature*, 470(7333):204–13.

Hilbe, J. M. (2011). *Negative Binomial Regression*. Cambridge University Press, New York.

Kutner, M., Nachtsheim, C., Neter, J., and Li, W. (2004). *Applied Linear Statistical Models - Fifth Edition*. McGraw-Hill/Irwin, New York.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, 10(3):R25.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25(14):1754–60.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16):2078–9.

Li, H., Lovci, M. T., Kwon, Y.-S., Rosenfeld, M. G., Fu, X.-D., and Yeo, G. W. (2008). Determination of tag density required for digital transcriptome analysis: application to an androgen-sensitive prostate cancer model. *Proceedings of the National Academy of Sciences of the United States of America*, 105(51):20179–84.

Madsen, H. and Thyregod, P. (2010). *An Introduction to General and Generalized Linear Models*. Chapman & Hall.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., and Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*, 18(9):1509–17.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall / CRC, London.

Ochman, H., Gerber, A. S., and Hartl, D. L. (1988). Genetic Applications of an Inverse Polymerase Chain Reaction. *Science*, 120(3):621–623.

Plackett, R. L. (1981). *The Analysis of Categorical Data*. Macmillan Publishing Co., Inc., New York.

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, 26(6):841–2.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–40.

Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Horn, G. T., Mullis, K. B., and Erlich, H. A. (1988). Primer-Directed Enzymatic Amplification of DNA with a Thermostable DNA Polymerase. *Science*, 239(4839):487–491.

Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, J. C., III, C. A. H., Slocombe, P. M., and Smith, M. (1977). Nucleotide sequence of bacteriophage ϕX174 DNA. *Nature*, 265:687–695.

Sanger, F. and Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, 94:441–8.

Verhoeven, K. J. F., Simonsen, K. L., and Mcintyre, L. M. (2005). Oikos 108: 643 /647, 2005 implementing false discovery rate control: increasing your power.

Watson, J. D. and Crick, F. H. (1953). Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, 171(4356):737–738.

Zirkle, C. (1941). Natural Selection before the "Origin of Species". *Proceedings of the American Philosophical Society*, 84(1):71–123.

Índice

agrupamento por blocos, 25
aleatorização, 23
amostragem, 22
análise exploratória dos dados, 49
aplicação, 49

BAM, 20
BEDTools, 18
Binomial Negativa, 30
Bowtie, 18

cDNA, 13
CML, 38
Comparações múltiplas, 41
coverageBed, 15

delineamento de experimentos, 21
DNA, 9

edgeR, 38
Erros do Tipo I, 41
Erros do Tipo II, 41
exons, 14
expressão gênica, 13, 15

FASTQ, 18
FDR, 44

GEO, 50
Gregor Mendel, 9

MA Plot, 40
Modelos Lineares Generalizados, 27
mRNA, 13

nucleotídeos, 9

pipeline, 16
Poisson, 28
Projeto Genoma Humano, 11
pseudo-dados, 39

qualidade dos dados, 52

Replicação, 24
replicação biológica, 24
replicação técnica, 24
RNA-Seq, 13

SAM, 15
SAMtools, 15
Sanger sequencing, 11
SRA, 17
SRA Toolkit, 17

Teste de Razão de Verossimilhanças, 35