TESTES T DE STUDENT E BAYESIANOS APLICADOS A MICROARRANJOS:

IMPACTO DOS MÉTODOS DE TRANSFORMAÇÃO E DO TAMANHO DA AMOSTRA

Sandro Leonardo Martins Sperandei

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DA COORDENAÇÃO DOS PROGRAMAS DE PÓS-GRADUAÇÃO DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA BIOMÉDICA.

Aprovada por:

Prof. Flávio Fonseca Nobre, Ph.D.

Profa. Ana Beatriz Monteiro Fonseca, D.Sc.

Prof. Jurandir Nadal, D.Sc.

Prof. Ulisses Gazos Lopes , D.Sc.

RIO DE JANEIRO, RJ - BRASIL

JUNHO DE 2007

SPERANDEI, SANDRO LEONARDO MARTINS

Testes T de Student e Bayesianos Aplicados a Microarranjos de DNA: Impacto dos métodos de Transformação e do Tamanho da Amostra [Rio de Janeiro] 2007

IX, 95 p. 29,7 cm (COPPE/UFRJ, M.Sc., Engenharia Biomédica, 2007)

Dissertação – Universidade Federal do Rio de Janeiro, COPPE

1. Microarranjos; 2. Estatística

I. COPPE/UFRJ II. Título (série)

“Estas coisas disse ele em palavras. Mas muito no seu coração ficou por dizer. Porque ele próprio não podia falar do seu segredo mais profundo.”

O Profeta

Khalil Gibran

DEDICATÓRIA

Este trabalho é dedicado aos meus pais, responsáveis maiores por tudo de bom que consegui na minha vida. Obrigado por terem sempre acreditado em mim e estarem sempre presentes quando precisei!

Dedico também, como forma de reconhecimento, a três pessoas que foram fundamentais no meu desenvolvimento. À professora Maria Inês Ferreira, minha primeira tutora e orientadora, ao professor André Leta, exemplo profissional e acadêmico que sempre seguirei, e ao professor Marcelo Cabral, o maior irmão que já tive.

AGRADECIMENTOS

Em primeiro lugar, aos meus orientadores. Ao professor Flávio Fonseca Nobre, por sua infinita paciência com as demoras e confusões do seu orientando, e à professora Ana Beatriz Monteiro Fonseca, por ter me mostrado o mundo Bayesiano, ainda que ele continue muito misterioso.

Ao amigo Marcelo Ribeiro Alves, sem o qual esse trabalho jamais teria sido concluído. Não teria palavras para te agradecer, então, fica o meu muito obrigado! Com certeza, O mundo precisa de mais pessoas como você...

À amiga Alessandra Monteiro, companheira de batalha no laboratório, que resolveu seguir outro caminho. Tenho certeza que você terá sucesso, faça o que fizer!

Aos amigos professores Ricardo Sartorato e Rafaella Miranda, por continuarem sendo chatos, mesmo após todos esses anos, me forçando a estudar e me aprofundar cada vez mais.

Por último, a Samira Santana “Sperandei”, meu amor, por me aturar e servir sempre de inspiração. Você é o meu ponto de referência e o motivo pelo qual eu sigo em frente.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

TESTES T DE STUDENT E BAYESIANOS APLICADOS A MICROARRANJOS: IMPACTO DOS MÉTODOS DE TRANSFORMAÇÃO E DO TAMANHO DA AMOSTRA

Sandro Leonardo Martins Sperandei

Junho/2007

Orientadores: Flávio Fonseca Nobre

Ana Beatriz Monteiro Fonseca

Programa: Engenharia Biomédica

O objetivo deste trabalho foi comparar o desempenho do teste *t* de Student e do teste *t* Bayesiano aplicados a dados de microarranjos simulados, analisando também o impacto dos métodos de transformação *Shift, Lowess e Linlog* e do tamanho da amostra. Foram feitas simulações com diferentes tipos de ruído, contendo 3920 genes normoexpressos e 80 genes diferencialmente expressos, com 50 replicações para cada tipo de ruído. O teste *t* Bayesiano mostrou um desempenho superior ao teste de Student com número de amostras inferior a 20, e desempenho similar com mais de 20 amostras. O aumento no número de amostras melhorou o desempenho dos dois testes. Não houve um método de transformação que pudesse ser aplicado a todos os ruídos. O método *Lowess* teve aplicação mais geral e o método *Linlog* não mostrou eficácia. A presente metodologia deve ser utilizada na avaliação de outros métodos aplicados a microarranjos.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

STUDENT'S AND BAYESIANS T-TESTS APPLIED TO DNA MICROARRAY: IMPACT OF TRANSFORMATION METHODS AND SAMPLE SIZE

Sandro Leonardo Martins Sperandei

June/2007

Advisors: Flávio Fonseca Nobre

Ana Beatriz Monteiro Fonseca

Department: Biomedical Engineering

The purpose of this work was to compare the performance of Student t-test and Bayesian t-test applied to simulated microarray data, either analyzing the impact of *Shift, Lowess and Linlog* transformation methods and sample size. Simulations were generated with different noise types, each one with 3920 normo and 80 differentially expressed genes, with 50 replications. Bayesian t-test shown better performance compared to Student t-test with fewer replications and equal performance with more than 20 replications. Increasing sample size improved performance of both tests. There was no transformation method that could be used with all noise types. *Lowess* method showed a more general use and the *Linlog* method performs badly in all cases. The present methodology should be used to evaluate other methods applied to microarray.

ÍNDICE

CAPÍTULO 1

INTRODUÇÃO

1.1 Introdução

A informação necessária ao desenvolvimento, manutenção e reprodução de uma célula está armazenada nos genes. Desde os estudos de Mendel sobre a transmissão de características hereditárias entre gerações até os dias atuais, houve grandes avanços no sentido de compreender como as informações contidas nos genes atuam sobre o funcionamento do organismo e como os milhares de genes interagem entre si.

Nas décadas de 1940 e 1950 importantes mecanismos genéticos, que proporcionaram grandes avanços no conhecimento do armazenamento e transmissão da informação genética foram compreendidos. Na primeira década, a descoberta do ácido desoxirribonucleico (DNA) e a noção de que a informação genética consiste, basicamente, de instrução para a produção de proteínas [1]. Na década seguinte, a estrutura desse DNA foi revelada [68].

O DNA é composto pela combinação de oligonucleotídeos que apresentam apenas quatro diferentes bases. Toda a informação genética está contida na combinação destes quatro elementos em uma extensa molécula. Essa relativa simplicidade impediu, por muito tempo, que se aceitasse o DNA como a molécula que continha a complexa informação genética que regula todos os organismos vivos [1]. A formulação do Dogma Central da Biologia Molecular [17] esclareceu a forma como a

informação genética contida no DNA atua sobre os mecanismos celulares e é a base para muitas técnicas de estudo genômico.

O termo *genômica* descreve um campo do conhecimento que vai além da genética básica e foi utilizado pela primeira vez em 1987 [50]. O objetivo do estudo da genômica extrapola o da genética, focada na hereditariedade, englobando, entre outros, o seqüenciamento de ácidos nucléicos, identificação de genes e a análise funcional desses. Este último aspecto da genômica, conhecido como genômica funcional, tem recebido especial atenção, principalmente após a conclusão do seqüenciamento do genoma humano [67].

Cada célula do corpo humano possui a mesma informação contida no seu DNA. A diferenciação entre as células está na forma como cada uma utiliza essa informação em respostas aos estímulos a que é exposta. Diferentes técnicas vêm sendo utilizadas na análise do padrão de utilização da informação do DNA por um tipo de célula em resposta a um determinado estímulo, como o *Northern blot*, *dot blots* e outros [27]. Em todas elas, o princípio fundamental está na estreita relação entre a parte ativa do DNA (gene), o mensageiro que transmite a sua informação à célula (RNA) e o produto final (proteína). Até recentemente, essa análise funcional era feita com apenas um, ou com um pequeno número de genes a cada vez, limitando a possibilidade de observação da interação entre os genes.

A técnica de microarranjos de DNA é uma nova técnica de análise que permite a observação simultânea da expressão de milhares de genes [10]. Fundamentada no processo de hibridação competitiva [41], o microarranjo está voltado principalmente ao estudo do transcriptoma. Este representa o conjunto de RNAs de uma célula. Uma vez que esse RNA é o agente mensageiro que encaminha as informações contidas nos genes, a análise do RNA presente em uma célula nos permite inferir quais genes estão ativos e em que grau.

A partir do RNA extraído de células com e sem as características em questão, utilizando a transcrição reversa, amostras de cDNA são geradas e marcadas. Paralelamente, uma matriz de sondas contendo milhares de seqüências de DNA correspondentes a genes de interesse é preparada. A solução contendo as amostras marcadas de cDNA é, então, exposta às sondas para hibridação. Após a leitura por um *scanner*, é possível estabelecer o grau de atividade do gene correspondente a uma sonda pela quantidade de hibridação ocorrida e registrada pelo leitor.

As possibilidades desta informação são diversas. A oncologia é uma das áreas onde se espera um grande avanço com o uso de microarranjos de DNA. Um câncer é caracterizado pela reprodução anormal de células [36] e o diagnóstico correto é peça fundamental na determinação da terapia adequada, de modo a maximizar a eficácia do tratamento e minimizar os efeitos colaterais. Nessa área, GOLUB *et al.* [30] isolaram um grupo de 50 genes capazes de distinguir dois diferentes tipos de leucemia aguda. ALIZADEH *et al.* [2] utilizaram a técnica de microarranjos na definição de prognósticos em linfoma não-Hodkins. HWANG *et al.* [34], utilizando microarranjos, foram capazes de identificar um grupo de 45 genes altamente associados à ocorrência de câncer oral. Avanços similares vêm sendo alcançados em diferentes tipos de câncer.

Os microarranjos são aplicados também na análise de respostas a drogas e tratamentos [19, 40], no desenvolvimento de testes de diagnóstico [31, 70], e em campos tão diferentes como a psiquiatria [45] e a atividade física [3, 44],entre outros.

Em vista de todos esses avanços e resultados promissores, os microarranjos são, muitas vezes, vistos como uma ferramenta mágica que irá responder a todas as questões [54]. Apesar de já serem utilizados há mais de dez anos, diversos problemas ainda comprometem a confiabilidade dos resultados obtidos em experimentos de microarranjo. As normatizações de procedimentos, a automação do experimento e o

avanço das técnicas de normalização dos dados têm contribuído bastante para a creditação dos resultados.

Atualmente, a maior dificuldade no uso de microarranjos está na análise dos dados de expressão [66]. Do ponto de vista estatístico, esse é um problema incomum. São avaliados milhares de variáveis (genes) com apenas poucas amostras (replicações), dificultando o uso das técnicas estatísticas tradicionais. Além disso, erros de medida aumentam ainda mais a dificuldade na extração de informação confiável dos microarranjos.

Independentemente do cuidado na elaboração e execução de um experimento de microarranjo, pode-se identificar dois componentes de erro que podem estar presentes: sistemático e aleatório. Enquanto o primeiro componente pode ser definido como uma tendência de que as medidas sejam diferentes em uma direção em particular [6], podendo ser causado, por exemplo, devido à diferença na sensibilidade do leitor em relação ao marcador utilizado, o erro aleatório se relaciona a variações biológicas inerentes ao organismo e erros na execução do protocolo proposto [6]. A natureza do erro em experimentos de microarranjos, no entanto, permanece pouco compreendida [51].

Diferentes métodos de transformação de dados têm sido utilizados buscando minimizar os erros acima mencionados. Entretanto, o impacto dessas transformações no desempenho dos testes estatísticos aplicados tem sido pouco avaliado.

O teste da razão, que observa a média da razão entre as duas condições sob análise, foi a primeira abordagem utilizada na tentativa de detecção de genes diferencialmente expressos [20, 59, 60]. Esse método foi logo substituído por modelos probabilísticos [22], embora ainda seja utilizado. Diferentes abordagens como o teste *t* [22], modelos de mistura [48], testes não-paramétricos [62, 65] e testes Bayesianos [8, 37] vêm sendo aplicadas com resultados variados.

O teste *t* de Student vem sendo largamente aplicado [22, 46] e apresenta especial interesse por ser uma das primeiras abordagens probabilísticas e por ser relativamente simples [22, 46], apesar do frágil pressuposto de normalidade para a distribuição da amostra. Outra forma de análise que tem recebido bastante atenção é a utilização da inferência Bayesiana [8], sendo uma abordagem muito utilizada, com resultados bastante positivos.

A avaliação dos resultados apresentados por diferentes testes estatísticos torna-se ainda mais complexa pela escassez de informação sobre o resultado esperado.

Uma vez que é muito difícil determinar o resultado esperado a partir de um experimento real, torna-se impossível determinar se uma metodologia aplicada tem um desempenho melhor em encontrar respostas corretas [51]. Uma abordagem alternativa é a utilização de simulações.

A simulação de processos biológicos deve se aproximar do fenômeno em estudo e apresentar variabilidade realística, tornando-a um desafio [7]. Mesmo desenvolvida de forma cuidadosa, este processo apresentará sempre limitações, especialmente devido à forte relação com os modelos matemáticos teóricos que norteiam a simulação e a estrutura de erro apresentada [51]. Mas essa alternativa ainda se apresenta como boa na avaliação do desempenho de métodos de análise de microarranjos, pois a simulação possibilita a manipulação das variáveis envolvidas, possibilitanto o controle do resultado esperado.

1.2 Objetivos

1.2.1 Objetivo Geral

Comparar o desempenho do teste *t* de Student e do teste *t* Bayesiano aplicados a dados de microarranjos simulados.

1.2.2 Objetivos Específicos

- Comparar o número de verdadeiro-positivos e falso-positivos dos dois testes estatísticos na detecção de genes diferencialmente expressos em dados simulados de microarranjos;

- Avaliar o impacto de três diferentes métodos de transformação no desempenho dos testes estatísticos aplicados a dados simulados de microarranjo;

- Avaliar o impacto do número de replicações no desempenho dos testes estatísticos aplicados a dados simulados de microarranjo.

1.3 Estrutura do Trabalho

O Capítulo 2 apresenta a fundamentação teórica da técnica de microarranjo (2.1), dos métodos de transformação (2.2), dos testes estatísticos (2.3) e da simulação realizada (2.4). No tópico fundamentação da técnica de microarranjos, são apresentados os conceitos básicos de genética, esclarecendo a estrutura do DNA e sua função no organismo humano. Também são abordados os mecanismos de duplicação, transcrição e tradução, responsáveis pela transferência da informação genética entre as estruturas celulares e o modo de ação sobre o organismo. Além disso, são apresentados os fundamentos das principais técnicas de análise da expressão gênica que precederam o microarranjo, bem como as etapas na preparação do mesmo. Os métodos de transformação aplicados nesse trabalho estão descritos no tópico 2.2, abordando os métodos *Shift*, *Lowess* e *Linlog*. Serão apresentadas as bases teóricas e a forma esperada como cada um modifica o resultado apresentado.

Dois diferentes testes estatísticos são comparados nesse trabalho: o teste *t* de Student e uma forma de teste *t* Bayesiano empírico. Ambos são apresentados no tópico 2.3.

No Capítulo 3 encontra-se a metodologia utilizada na simulação e análise dos dados de microarranjo. Os resultados dessa análise estão no Capítulo 4. O Capítulo 5 se destina à discussão dos resultados encontrados e as conclusões do trabalho são apresentadas no Capítulo 6.

CAPÍTULO 2

FUNDAMENTAÇÃO TEÓRICA

2.1 Fundamentação da Técnica de Microarranjo

2.1.1 O DNA, o RNA e o Dogma Central da Biologia Molecular

O funcionamento do organismo humano depende da informação genética contida nas longas cadeias de DNA presentes no interior de cada célula. Essa informação consiste, essencialmente, de instruções para a produção de proteínas. A simplicidade química do DNA dificultou, em muito, a sua aceitação como material genético [1].

O DNA é composto por duas longas cadeias de nucleosídeos unidas por pontes de hidrogênio [1]. A descoberta dessa estrutura por WATSON *et al.* [68] permitiu um rápido avanço na compreensão da forma como a informação genética é mantida, utilizada e repassa às gerações subseqüentes.

Cada nucleotídeo presente no DNA, um nucleosídeo sem o revestimento de açúcar externo, é identificado por meio da sua base nitrogenada. Em todo o genoma, existem apenas quatro tipos diferentes de bases nitrogenadas: adenina (A), timina (T), guanina (G) e citosina (C). A grande gama de informação genética surge da combinação desses quatro tipos de nucleotídeos em seqüências de diferentes comprimentos. Os genes são seqüências específicas de nucleotídeos que contêm as informações necessárias para a produção de proteínas.

Uma vez que o DNA é composto de duas fitas opostas, existe um pareamento entre as bases de uma fita em relação àquelas na fita oposta. Essa relação é única, de forma que a base adenina deverá estar, invariavelmente, ligada a uma base timina na

fita oposta, e vice-versa. O mesmo se observa com as bases guanina e citosina. Essa relação entre as bases é crítica para os processos de transmissão da informação genética.

Antes da divisão celular, é necessário que as longas cadeias de DNA presentes no núcleo da célula sejam copiadas para serem repassadas à célula-filha. A esse processo dá-se o nome de replicação do DNA.

Durante a replicação, a dupla fita de DNA é separada e cada fita serve de molde para a criação de uma nova. Essa nova seqüência de DNA gerada contém os mesmos nucleotídeos da seqüência oposta, devido à complementaridade entre as bases (Figura 2.1). Cada fita dá origem, dessa forma, a uma cadeia de DNA complementar a si mesma e idêntica à fita oposta. Uma enzima, chamada DNA-polimerase, se liga ao DNA e segue de nucleotídeo em nucleotídeo, acrescentando um elemento à nova seqüência em acordo com o nucleotídeo lido na fita que serve de molde. Após a passagem dessa enzima, o DNA volta a se fechar.

A informação contida nos genes não age diretamente sobre a célula, mas através de uma molécula mensageira de ácido ribonucléico (mRNA) [53]. O mRNA é uma cadeia de nucleotídeos similar ao DNA, mas se apresenta como uma fita simples e não dupla e, ao contrário do DNA, o açúcar presente em seu nucleosídeo é uma ribose, ao invés de uma desoxirribose. Outra diferença importante é a presença da base nucleotídica uracil (U) em substituição à base timina.

A produção de uma molécula de mRNA é denominada de transcrição. Esse processo se dá de forma muito similar à replicação. A fita dupla de DNA se abre e a enzima RNA-polimerase se liga ao sítio específico do gene a ser expresso. Essa enzima se desloca, utilizando a seqüência de bases do gene como molde para o pareamento na formação da molécula de mRNA. Conforme o deslocamento segue, novos nucleotídeos são adicionados, até o fim da seqüência contida no gene. A

relação entre as bases também se mantém, exceção feita à base timina, que está ausente. O processo de transcrição é ilustrado na Figura 2.2.

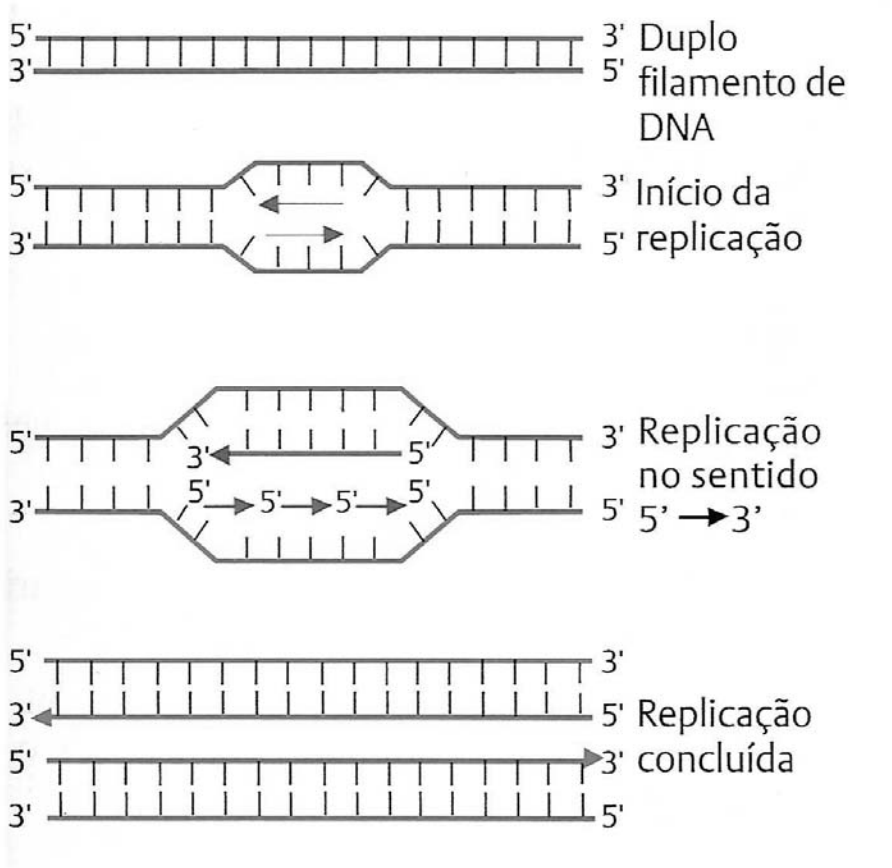


Figura 2.1 – Replicação do DNA. A dupla fita de DNA original se abre e cada fita individual serve de molde para uma nova [50].

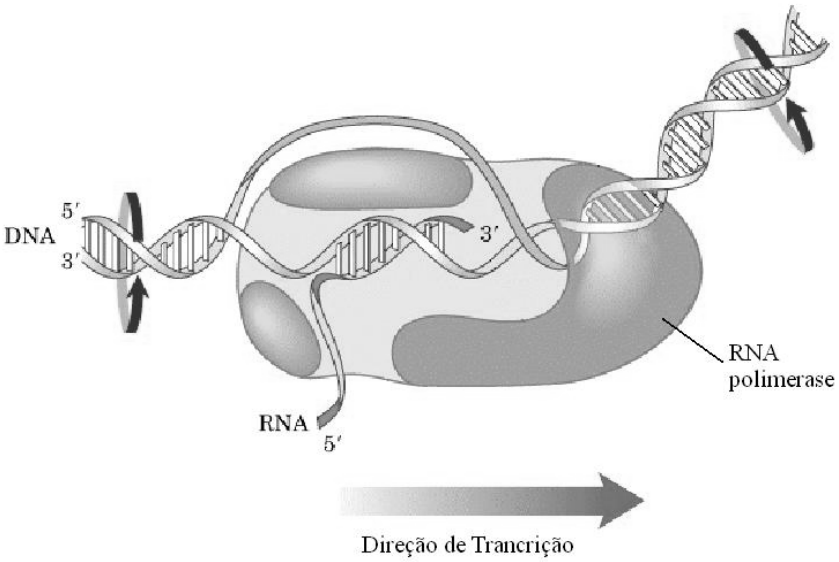


Figura 2.2 – Transcrição. A RNA polimerase separa as duas fitas de DNA e utiliza uma como molde para o mRNA. Adaptado de Lehninger *et al*. [42]

Essa molécula transporta a informação do gene até o ribossomo, presente no retículo endoplasmático. No ribossomo, o mRNA será decodificado em subunidades de três nucleotídeos, denominadas códons. Para cada códon existe um aminoácido correspondente que será adicionado à proteína em formação. A Tabela 2.1 apresenta a relação entre os códons e os aminoácidos. É através das proteínas que a informação genética contida no DNA irá atuar na regulação dos mecanismos celulares.

A Figura 2.3 mostra a relação entre a seqüência de bases presentes no DNA, as bases do mRNA e os aminoácidos das proteínas. Essa relação, conhecida como o Dogma Central da Biologia Molecular [17], foi fundamental no desenvolvimento de diferentes técnicas de análise genética, inclusive a de microarranjo.

2.1.2 O Microarranjo de DNA

De acordo com PASSARGE *et al.* [50], a genômica é a área de estudo interessada em aspectos que vão além da genética, como o seqüenciamento de mapas genômicos, identificação de genes e análise funcional, entre outros. Este último aspecto é objeto de um ramo fundamental, o da genômica funcional, que tem recebido especial atenção.

O foco da genômica funcional está na determinação da função que cada gene desempenha no funcionamento celular e na resposta aos estímulos sofridos por esta célula. O papel dos genes no desenvolvimento de doenças com origem genética é também um promissor ramo da genômica.

Diversas técnicas vêm sendo utilizadas há décadas na análise da expressão gênica em resposta a condições determinadas [25]. A despeito de suas diferenças e especificidades, todas apresentam em comum o uso da hibridação como princípio. Com base no Dogma Central da Biologia Molecular, duas seqüências de nucleotídeos só irão se ligar (hibridar) se houver complementaridade entre as suas bases.

O *Northern Blot* é o método mais utilizado para a determinação da abundância de mRNA em uma célula, devido à sua relativa simplicidade [53]. A técnica é uma adaptação do *Southern Blot* [61]. O mRNA extraído de células que apresentem uma determinada característica é fixado a uma membrana. A membrana é, então, exposta a uma solução contendo clones de DNA de um gene de interesse (sondas de DNA), possibilitando a hibridação de seqüências complementares. Após a lavagem da membrana, para retirar o excesso de solução não hibridada, o grau de hibridação pode ser observado pela marcação prévia das sondas de DNA, através de marcador químico ou radioativo.

Tabela 2.1. Relação entre os códons e os aminoácidos traduzidos. Em destaque, os aminoácidos que indicam o início (Met) e o fim da tradução.

Primeira Letra do Códon	Segunda Letra do Códon: U				Segunda Letra do Códon: C				Segunda Letra do Códon: A				Segunda Letra do Códon: G			
	U	C	A	G	U	C	A	G	U	C	A	G	U	C	A	G
U	Phe	Ser	Tyr	Cys	Leu	Pro	His	Arg	Ile	Thr	Asn	Ser	Val	Ala	Asp	Gly
C	Phe	Ser	Tyr	Cys	Leu	Pro	His	Arg	Ile	Thr	Asn	Ser	Val	Ala	Asp	Gly
A	Leu	Ser	Fim	Fim	Leu	Pro	Gln	Arg	Ile	Thr	Lys	Arg	Met	Ala	Glu	Gly
G	Leu	Ser	Fim	Trp	Leu	Pro	Gln	Arg	Val	Ala	Glu	Gly	Val	Ala	Glu	Gly

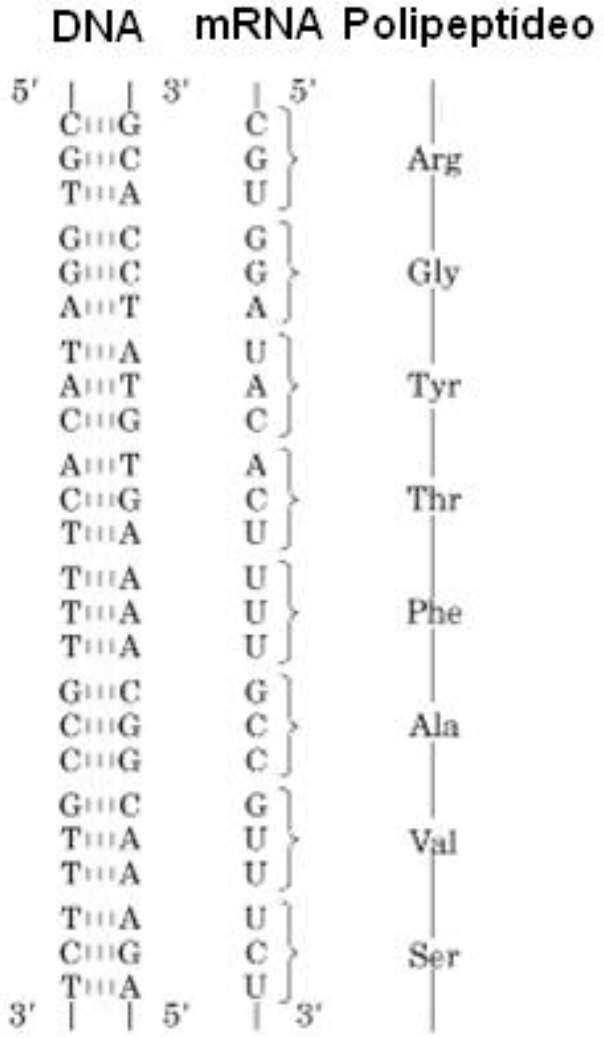


Figura 2.3 – Dogma Central da Biologia Molecular. Uma das fitas do DNA serve de molde ao mRNA durante a transcrição. Este, por sua vez, determina a seqüência de aminoácidos durante a tradução.

A técnica de Ensaio de Proteção à Ribonuclease (RPA) se dá de forma muito similar, mas tenta remover hibridações parciais e não-específicas. Outras técnicas como o *Dot Blot*, *Slot Blot* e *Fast Blot* foram desenvolvidas com objetivo de simplificar os procedimentos [53], mas possuem essencialmente os mesmos fundamentos e as mesmas limitações.

Um dos maiores problemas na utilização dessas técnicas está na necessidade de análise de um pequeno número de genes por vez. Mesmo que seja possível reutilizar a membrana contendo o mRNA ainda preso a ela, a análise de milhares de genes, ou mesmo de algumas centenas, torna-se uma tarefa inviável.

Em uma abordagem alternativa, a Reação em Cadeia da Polimerase (PCR) é uma técnica acelular, rápida e sensível para a amplificação de segmentos de DNA [50]. Utilizando seqüências de oligonucleotídeos complementares às seqüências adjacentes ao segmento investigado, a técnica força a fixação da polimerase nessa região e a ativação da seqüência-alvo. Numa reação em cadeia, cada seqüência expressa serva de molde para a produção de novas seqüências, ocasionando a amplificação do segmento de DNA. A técnica pode também ser utilizada a partir do mRNA presente no citoplasma, utilizando a enzima transcriptase reversa (RT-PCR). A principal limitação desta técnica, uma das principais utilizadas atualmente no estudo de expressão gênica, está na necessidade de conhecimento prévio do segmento de DNA a ser analisado. Assim, ela pode ser utilizada como forma de comprovar os achados de um microarranjo, por exemplo, mas não na descoberta de novas seqüências relacionadas a condições de interesse.

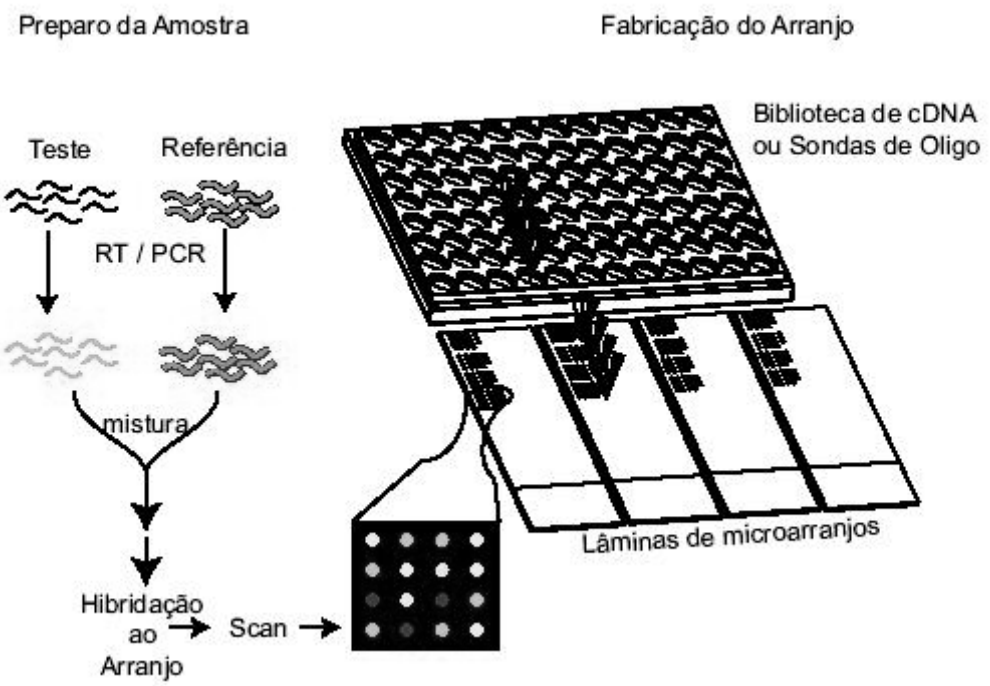
Os microarranjos de DNA possibilitam a análise simultânea da expressão de milhares de genes [58] e vêm sendo utilizados desde a década de 1990 [59]. Essa característica pode proporcionar uma grande quantidade de novos conhecimentos sobre os mecanismos de sistemas vivos [10]. Diferentemente das técnicas anteriores, no microarranjo, as sondas de DNA é que são imobilizadas em uma superfície e expostas à solução. É importante frisar que o microarranjo é uma técnica de “pré-seleção” de genes-candidatos, que devem ter sua relação com a condição de interesse confirmada por outros métodos, como a RT-PCR descrita acima.

2.1.3 Preparo de um Microarranjo de DNA

O microarranjo é desenvolvido a partir de dois passos que ocorrem paralelamente e um terceiro que combina esses dois. A Figura 2.4 mostra a preparação de uma lâmina de microarranjo. Inicialmente, uma amostra de RNA é extraída de uma célula de interesse, denominada condição controle, e, por meio da enzima transcriptase reversa, é convertida em uma amostra de cDNA, ou seja, uma fita simples de DNA com seqüência complementar ao RNA que lhe deu origem. Essa amostra de cDNA é marcada com um marcador específico. Os marcadores mais comuns são a cianina-3 (Cy3) e a cianina-5 (Cy5). Em geral, a cianina-3 é utilizada para a amostra da condição controle.

Uma segunda amostra é extraída de uma célula diferente, denominada condição experimental, e o processo de transcrição reversa é repetido. Essa amostra é individualizada utilizando um marcador (cianina-5) diferente daquele utilizado para a amostra controle. A seguir, as duas amostras, controle e experimental, são combinadas. O processo pode ser repetido tantas vezes quantos forem os protocolos experimentais a serem analisados.

Paralelamente, uma série de sondas de cDNA são depositadas em uma lâmina de vidro, seguindo um arranjo matricial. A lâmina é exposta à solução formada com as amostras de cDNA geradas a partir das células controle e experimental para que possa haver hibridação entre as amostras e as seqüências impressas na lâmina.



O microarranjo se apresenta como uma técnica de hibridação competitiva [10], onde a quantidade de hibridação ocorrida entre a amostra e uma determinada seqüência impressa na lâmina indica o grau de atividade do gene correspondente na célula no momento de extração do RNA.

2.1.4 Análise de Imagens de um Microarranjo de DNA

Após a exposição da lâmina à solução contendo os cDNAs marcados, a mesma é irradiada em microscópio confocal a laser, também chamado de *scanner*. Essa irradiação excita os marcadores fluorescentes e essa excitação é lida pelo *scanner* para cada um dos marcadores utilizados.

Como resultado, obtém-se uma imagem da lâmina para cada condição (ou canal) avaliada, em que cada ponto (*spot*) apresenta diferentes tons de cinza, de acordo com a quantidade de hibridação ocorrida, medida em quantidade de pixels em cada ponto, uma para cada marcador utilizado. Essas imagens são combinadas e

coloridas digitalmente, gerando-se as cores que são vistas: verde, vermelho e amarelo. Essa coloração visa facilitar a inspeção visual do resultado obtido (Figura 2.5).

Um importante ponto a ser considerado nessa etapa diz respeito à limitação técnica do *scanner*. Em geral, os equipamentos utilizados apresentam capacidade para gerar imagens de 16 bits (65,536 tonalidades de cinza). Dessa forma, caso ocorra um valor superior a esse limite, tem-se o que é denominado de "saturação". Nessa condição, apesar do volume de hibridação ter sido superior ao limite, o resultado apresentado será o próprio limite, com perda do restante do sinal. Em caso de grande número de *spots* saturados é recomendado que se faça novamente a passagem da lâmina pelo *scanner* em menor intensidade [18].

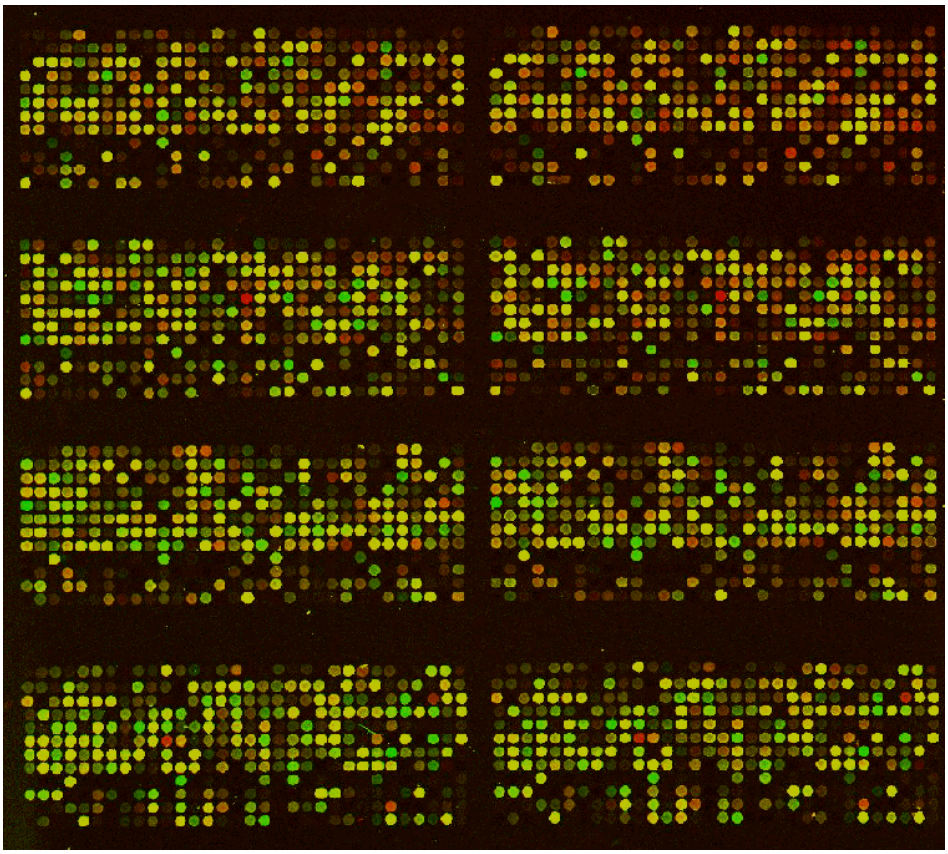


Figura 2.5 – Exemplo de uma imagem de microarranjo já colorida (adaptada de http://www.mgu.har.mrc.ac.uk/images/microarray.gif – Acesso em 03/06/07)

A intensidade dos pixels atribuídos a cada ponto em cada condição indica o grau de hibridação ocorrida entre a amostra de mRNA e a seqüência específica contida no ponto. Antes de se analisar os resultados obtidos, os dados necessitam passar por técnicas de normalização e transformação, reduzindo o efeito das fontes de variação inerentes ao experimento, conforme será visto a seguir.

2.2 Métodos de Transformação

Mesmo com a automação de boa parte do processo de construção de um experimento de microarranjo e a normatização crescente dos procedimentos, os resultados obtidos apresentam-se contaminados por erros de medida.

Erros de medida são variações entre as condições analisadas que não são originadas por diferenças reais entre as mesmas. Diversas são as fontes de variabilidade presentes em um experimento de microarranjo, como efeitos de saturação, flutuações na quantidade de sonda, hibridação não específica e outras, que dão origem a esses erros [41].

Basicamente, erros de medida podem ser classificados como sistemáticos e aleatórios. O erro sistemático se apresenta de maneira mais ou menos uniforme em todos os componentes da amostra analisada (lâminas). Aparece como uma tendência geral dos resultados serem diferentes em uma direção particular [6].

Em microarranjos, a principal fonte de erro sistemático está relacionada aos marcadores fluorescentes utilizados. A cianina-5 (Cy5) e a cianina-3 (Cy3) podem ter diferentes graus de incorporação às seqüências de mRNA da amostra, assim como o *scanner* de leitura poderá apresentar diferentes sensibilidades a cada um dos tipos de marcador. Essa fonte de erro foi descrita em experimentos onde a mesma amostra foi dividida, marcada com os dois corantes, e hibridada na mesma lâmina (experimentos

de marcação reversa). Ainda que não houvesse diferença entre as amostras, essa pôde ser observada devido ao erro sistemático relacionado ao marcador utilizado [22].

Por outro lado, o erro aleatório está ligado a variações biológicas das amostras ou mecânicas (instrumentais) inerentes ao experimento, ou a inconsistências no próprio protocolo experimental [15]. São fontes de erro aleatório as variações na quantidade de material depositado em cada ponto na construção do microarranjo, contaminação da lâmina por poeira ou outros tipos de sujeira e hibridação não específica.

O erro aleatório é mais difícil de ser retirado na fase de análise dos dados, devendo ser minimizado durante a elaboração do próprio experimento. No entanto, seu impacto nos resultados é menor, pois o resultado almejado é exatamente uma tendência no comportamento (sobre-expresso ou sub-expresso) de um ou mais genes na amostra analisada. Ao contrário, o erro sistemático pode influir profundamente nos resultados obtidos. Mas seu efeito é mais fácil de ser isolado e ele pode ser removido com maior ou menor grau de eficiência, dependendo do método de transformação utilizado.

2.2.1 Razão R/G

A primeira transformação aplicada aos resultados de um microarranjo é a relativização dos valores observados para cada canal (condição) em cada ponto (gene). O nome “Razão R/G” se deve ao uso do corante Cy5 (vermelho – *Red*) para a condição experimental e Cy3 (verde – *Green*) para a condição controle (Equação 2.1). Assim, a Razão R/G descreve o nível de expressão de um gene *j* na condição experimental (Y_{Rj}) em relação à expressão desse mesmo gene na condição controle (Y_{Gj}).

$$r_j = \frac{Y_{Rj}}{Y_{Gj}} \tag{2.1}$$

Essa transformação pode reduzir a variabilidade que surge da diversidade biológica, entre outras fontes, reduzindo o erro aleatório. Ao invés de observar e comparar valores absolutos que podem variar enormemente entre indivíduos, com a razão passa-se a observar a relação entre as condições analisadas.

2.2.2 Logaritmo

Apesar das vantagens descritas para a utilização da Razão R/G, o método também apresenta limitações. A mais importante é que a razão irá atribuir pesos diferentes a genes sobre-expressos e sub-expressos [52]. Genes sobre-expressos com um fator 2, por exemplo, irão apresentar uma razão de expressão igual a 2. Genes sub-expressos com fator 2, por outro lado, apresentarão razão igual a 0,5, mesmo que, biologicamente, tenham o mesmo significado.

O logaritmo da razão (Equação 2.2) proporciona a atribuição do mesmo valor a genes sobre- e sub-expressos pelo mesmo fator, ocasionando apenas inversão no sinal. Assim, $\log_2(2)= 1$ e $\log_2(½)= -1$. Logaritmos com diferentes bases podem ser utilizados, pois a base afeta apenas o valor obtido, não as propriedades desejadas [13].

$$Log_2(r_j) = Log_2\left(\frac{Y_{Rj}}{Y_{Gj}}\right) \tag{2.2}$$

Essa transformação possui a vantagem de proporcionar simetria à distribuição das razões (normalização), pressuposto básico de diversos testes estatísticos [23]. A transformação logarítmica tem uma vantagem adicional de contribuir para a estabilização da variância [14], outro pressuposto de muitos testes estatísticos paramétricos [64].

Os dois métodos descritos até aqui são geralmente aplicados aos dados de microarranjos. Porém, apesar de contribuir na análise dos dados, métodos adicionais de transformação devem ser aplicados. A escolha dos métodos seguintes depende da característica do erro assumido e eles podem ser aplicados individualmente ou em conjunto.

2.2.3 MA plot

Após a transformação dos dados resultantes de um experimento de microarranjo através da Razão R/G e do Logaritmo podem ainda restar erros de medida comprometedores para a análise dos dados.

Antes de selecionar algum método adicional de transformação dos dados, torna-se necessário observar a natureza dos dados. A melhor forma de visualização dos dados com respeito à análise da natureza do erro de medida é através de um gráfico da razão pela média da intensidade (MA *plot*). Nesta representação gráfica, a diferença é expressa em função da média entre os canais ou, no caso dos dados transformados, o logaritmo da razão entre os canais é expresso em função da média do logaritmo (Equação 2.3).

$$\begin{cases} Y_{Rj} - Y_{Gj} \Rightarrow Log_2\left(Y_{Rj}\right) - Log_2\left(Y_{Gj}\right) \Rightarrow Log_2\left(\dfrac{Y_{Rj}}{Y_{Gj}}\right) \\ \dfrac{Y_{Rj} + Y_{Gj}}{2} \Rightarrow \dfrac{Log_2\left(Y_{Rj}\right) + Log_2\left(Y_{Gj}\right)}{2} \Rightarrow \dfrac{1}{2} * Log_2\left(Y_{Rj} * Y_{Gj}\right) \end{cases} \tag{2.3}$$

O resultado dessa representação (Figura 2.6) é um diagrama de espalhamento similar a representar graficamente o canal R em função de G, mas com uma rotação de 45^0.

Em um MA *plot* é possível observar, com maior facilidade, a presença e magnitude dos erros. Espera-se que a maioria absoluta dos genes em um

microarranjo não apresente diferença significativa entre as condições. Desta forma, a maior parte dos pontos no gráfico deverá possuir valor igual a zero para o logaritmo da razão entre os canais. Uma reta de regressão linear a partir destes pontos teria intercepto e coeficiente de inclinação iguais a zero. Pequenas variações a partir desta reta são, em geral, associadas ao erro aleatório (Figura 2.6).

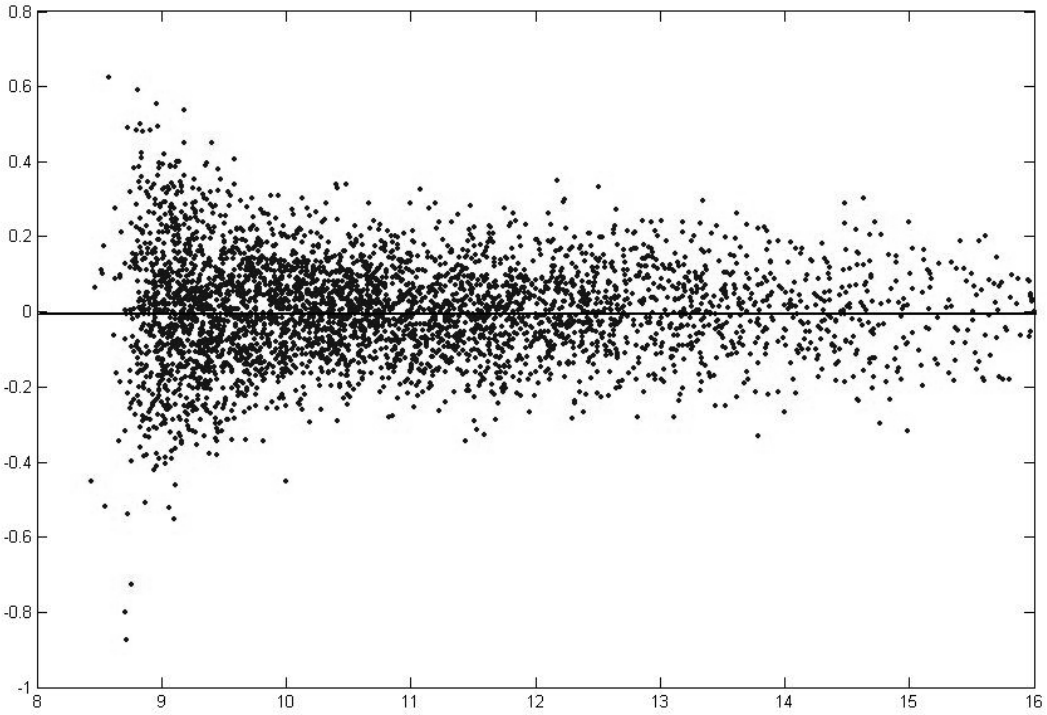


Figura 2.6 – Exemplo de MA *plot*. Observar a reta representando a média das razões igual a zero e o erro aleatório dos dados.

O erro sistemático mais simples se mostra como um desvio significativo do intercepto da reta de regressão. Esse desvio será ocasionado pela tendência de dominância de um canal em relação ao outro (Figura 2.7).

Quando o coeficiente de inclinação dessa reta de regressão se mostra significativo, existe um tipo mais complexo de erro sistemático, onde existe relação entre o tamanho do erro sistemático e a intensidade do sinal verificado. Em microarranjos essa forma de erro sistemático é comum, mas raramente é linear,

tornando necessária a utilização de métodos de transformação não-lineares (Figura 2.8).

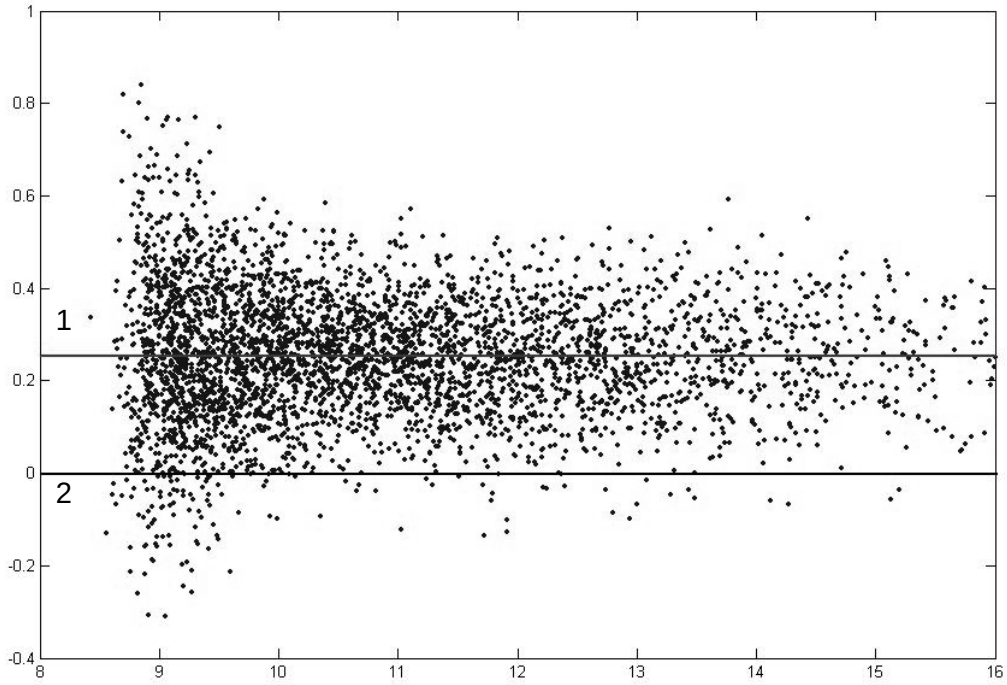


Figura 2.7 – MA plot indicando um erro sistemático aditivo. Note como a linha que indica a média das razões (1) se afasta do zero (2).

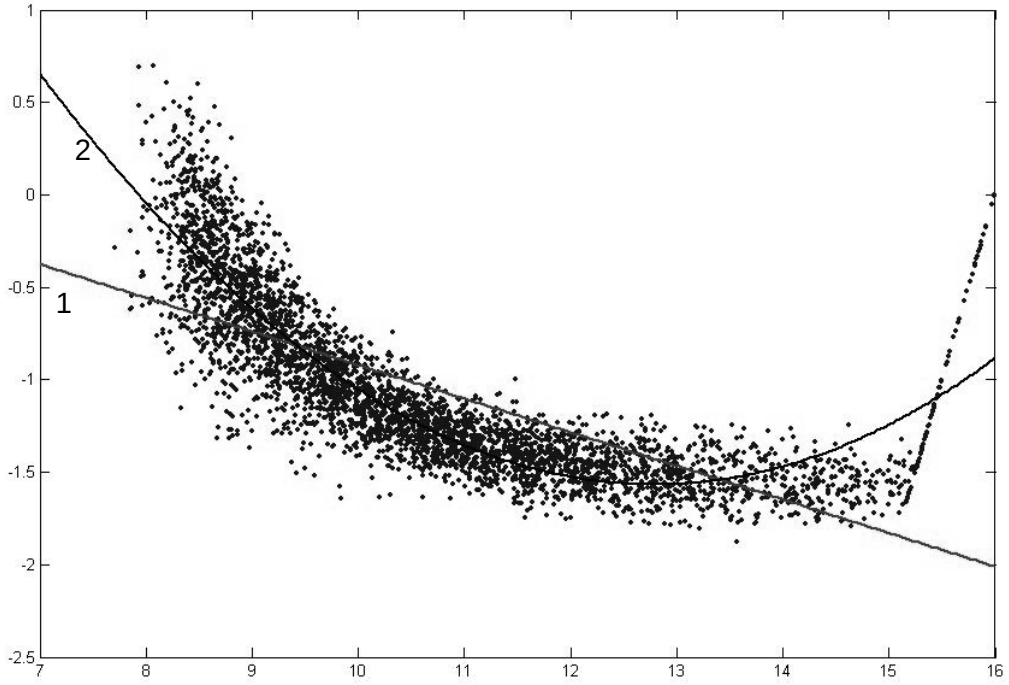


Figura 2.8 – MA plot indicando um erro sistemático com dependência entre a razão e a média. A relação não é linear, o que pode ser visto pelo fraco ajuste da reta (1) em relação ao ajusto quadrático (2).

2.2.4 Método Shift *de Transformação*

O método *shift* [39], também conhecido como método de transformação global [71], é o mais simples e um dos primeiros métodos de transformação aplicados a procedimentos de microarranjos [22, 71]. O objetivo é estimar uma constante a ser aplicada ao sinal bruto de cada canal do microarranjo de maneira a reduzir o componente aditivo do erro sistemático [39].

Neste método, o valor transformado (Z_{ij}) é o logarítmo do sinal medido (Y_{ij}) adicionado a uma constante C (Equação 2.4).

$$\begin{cases} Z_{Ri} = Log_2\left(Y_{Ri} + C\right) \\ Z_{Gi} = Log_2\left(Y_{Gi} - C\right) \end{cases} \tag{2.4}$$

KERR *et al.* [39] recomendam a estimação da constante C que irá reduzir o desvio absoluto médio (MAD). O uso desta estatística ao invés do método dos mínimos quadrados objetiva reduzir a interferência dos genes diferencialmente expressos, cujas razões se afastam muito da média. Numa abordagem mais simples, a constante C pode ser estimada, simplesmente, pela média ou mediana das razões [22].

A utilização de uma mesma constante a ser adicionada a um canal e subtraída do canal oposto é preferível, uma vez que diferentes valores para cada canal torna a definição da constante a ser utilizada muito difícil, pois apresenta um grande número de soluções possíveis [57].

O método *shift* apresenta a limitação de desconsiderar possíveis dependências do erro em relação à intensidade do sinal e variações espaciais, comprometendo seu resultado [22, 38, 71]. KERR *et al.* [39], por outro lado, afirmam uma preferência em relação a este método por ser robusto e com interpretação simples, embora assumam que não há garantias de que o método vá funcionar bem em qualquer conjunto de dados de microarranjos.

2.2.5 Método Lowess *de Transformação*

O método *Lowess* de transformação foi desenvolvido como um método de suavização de diagramas de espalhamento ainda na década de 1970 [16]. Após a sua aplicação inicial em microarranjos [71], este método vem sendo amplamente utilizado na transformação desse tipo de dado [9, 18].

O *Lowess* é aplicado diretamente sobre o MA *plot*. Inicialmente, uma fração *f* dos genes é selecionada em cada lâmina. Para essa fração, é ajustado um polinômio de ordem *d*, sendo atribuídos pesos diferentes para os genes de acordo com a distância de cada gene dentro da fração selecionada em relação ao gene central

analisado, utilizando uma função de pesos *W*. Em seguida, os pesos são recalculados a partir da diferença entre o valor inicial e o ajustado. O processo se repete por um número *t* de interações[16]. Dessa forma, é definido o fator de calibração (constante C) a ser subtraído ao sinal para a correção do erro sistemático [52].

Uma possível limitação do método é que o polinômio é obtido a partir do método dos mínimos quadrados. Os genes diferencialmente expressos podem influenciar a estimação por este método [39]. No entanto, seu efeito é, em geral, pequeno, pois apenas uma pequena proporção dos genes serão expressos de forma significativamente diferente entre os canais [9, 71]. Além disso, o processo de interações reduz o impacto desses genes através da atribuição de pesos reduzidos [16].

Uma dificuldade na utilização do *Lowess* é a determinação do tamanho da fração dos genes utilizada na determinação do espaço reconhecido como local (*span*) [39]. Se este parâmetro, identificado por *f*, for muito pequeno, haverá uma super-adequação aos dados, de modo que genes diferencialmente expressos terão seu valor alterado. Por outro lado, ao utilizar uma fração muito grande, não será alcançado o efeito desejado sobre o erro sistemático [9, 18, 39]. Formas complexas já foram sugeridas para estimar o parâmetro *f* [23], porém um valor entre 0,2 e 0,5 é normalmente utilizado [16, 71].

Os demais parâmetros a serem selecionados (*d, t* e *W*) possuem valores de referência que podem ser utilizados com sucesso em quase todas as situações [16].

2.2.6 Método Linlog *de Transformação*

O método *Linlog* é um dos métodos menos utilizados até o presente. Ele foi proposto por CUI *et al.* [18], como uma simplificação da transformação arco-seno hiperbólico (arsinh) de HUBER *et al.* [33].

Novamente, a homocedasticidade é pressuposto de diversos métodos estatísticos, embora essa característica nem sempre esteja presente em resultados de microarranjos. Freqüentemente, a variância diminui com o aumento da intensidade [56].

O principal objetivo do método *Linlog* é a estabilização da variância por todo o espectro de intensidade de um microarranjo, sem necessariamente corrigir curvaturas que possam existir no MA *plot* [18].

Em um microarranjo, erros aditivos originados de diferenças de background têm maior impacto em genes com menor intensidade de sinal. Por outro lado, erros multiplicativos, relacionados especialmente ao efeito do corante utilizado, afetam principalmente os genes com maior intensidade de sinal [18].

O método *Linlog* transforma os genes com baixo nível de expressão a partir de um modelo linear, enquanto os genes de alta expressão sofrem uma transformação logarítmica, transformando os erros multiplicativos em aditivos.

$$\begin{cases} Z_{ij} = Log_2\left(d_i\right) - \dfrac{1}{Ln2} + \dfrac{Y_{ij}}{\left(d_i Ln2\right)} & Y_{ij} < d_i \\ Z_{ij} = Log_2\left(Y_{ij}\right) & Y_{ij} \geq d_i \end{cases} \qquad (2.5)$$

O que precisa ser determinado neste método é o ponto de divisão entre os genes que serão considerados de baixa expressão e aqueles que serão considerados como de alta expressão. É o parâmetro d_i da Equação 2.5, acima. CUI *et al.* [18] citam que este ponto pode ser estimado pela minimização do desvio absoluto médio na amplitude interquartílica dos logaritmos das razões em relação à mediana, mas também pode ser fixado, em termos práticos, entre 25 e 30% dos genes.

Diferentes métodos de transformação vêm sendo aplicados a procedimentos de microarranjos, utilizando diversas informações como a localização do ponto na

lâmina ou grupos de genes-controle. Aqui, restringiu-se à descrição destes três métodos, por apresentarem a essência dos demais métodos.

2.3 Testes Estatísticos

Um teste estatístico é um método que pretende inferir as características de uma população a partir de informações extraídas de uma amostra. Os testes paramétricos apresentam, em geral, a mesma dinâmica: cálculo de uma estatística, baseada em valores de diferenças e dispersões, comparação desta estatística com uma população padrão e determinação da probabilidade da estatística nessa população padrão.

A grande limitação dos testes estatísticos paramétricos está na sua forte dependência de pressupostos em relação à sua população padrão. O comprometimento destes pressupostos compromete também a validade do resultado obtido. Em microarranjos, que apresentam quase sempre um baixo número de replicações (pequeno tamanho amostral), a garantia desses pressupostos é muito fraca. No entanto, a utilização de testes não-paramétricos, mais independentes em relação aos pressupostos teóricos, apresentam comprometimento do poder estatístico, especialmente com baixo número de replicações.

A seguir, serão descritos dois dos principais testes estatísticos utilizados na análise de microarranjos.

2.3.1 Teste t *de Student*

A base teórica que fundamenta o teste *t* de Student é a mesma da distribuição normal de probabilidade, e já estava disponível na última metade do século XIX [26]. O teste surgiu, no entanto, no início do século XX, a partir das observações e experimentações práticas de STUDENT [63].

O teste *t* é um teste para comparação de médias, onde o objetivo é observar a probabilidade de ocorrência do resultado encontrado para a média calculada a partir da hipótese nula que é a de igualdade entre as médias populacionais das amostras comparadas.

Após o cálculo da média para o logaritmo das razões entre os canais, essa média é subtraída da média teórica esperada sob a hipótese nula. No caso dos microarranjos, a hipótese nula é a inexistência de diferença no nível de expressão de um gene nas duas diferentes condições analisadas. O valor esperado para a razão é, portanto, igual a 1. Como se está trabalhando com o logaritmo da razão, esse valor será, na realidade, igual a zero.

O valor resultante dessa diferença, que será a própria média dos logaritmos das razões, é, então, padronizado através da divisão pelo erro padrão da média (Equação 2.6). O resultado é a estatística *t*.

$$t = \frac{\bar{r}}{\sqrt{\frac{s_r^2}{n}}} \tag{2.6}$$

Onde $\bar{r}$ é a média dos logs das razões, s_r é o desvio-padrão dos logs das razões e n é o número de replicações (lâminas).

O método leva em consideração não só a diferença (razão) entre os canais, mas também a dispersão dos dados. Assim, pequenas diferenças médias podem ser significativas na presença de uma pequena dispersão, enquanto mesmo grandes diferenças podem não ser significativas, caso a dispersão seja muito grande.

A estatística *t* é utilizada para calcular a probabilidade utilizando uma distribuição *t* padronizada com média zero, erro-padrão 1 e n-1 graus de liberdade. Quanto maior o valor da estatística *t*, menor a probabilidade de aquela diferença

pertencer à distribuição sob a hipótese nula e, por isso, estar relacionada ao acaso. Para valores de probabilidade iguais ou menores que um valor definido *a priori*, a diferença é considerada estatisticamente significativa.

Os genes diferencialmente expressos (genes DE) são aqueles que apresentam diferença estatisticamente significativa entre as duas condições.

A dependência dos graus de liberdade (n-1) é o principal avanço oferecido pelo trabalho de Gosset (que assumiu o pseudônimo de Student) [63], posteriormente fundamentado matematicamente por FISHER [29]. Através de observações práticas e de simulações de Monte Carlo, Gosset observou que a teoria da distribuição normal, incluindo o teorema do limite central, era robusta apenas para amostras de tamanho grande (n>30). Em pequenas amostras, a curva de distribuição de probabilidades se afastava da distribuição normal, mostrando-se tão mais achatada quanto menor fosse o tamanho amostral (Figura 2.9).

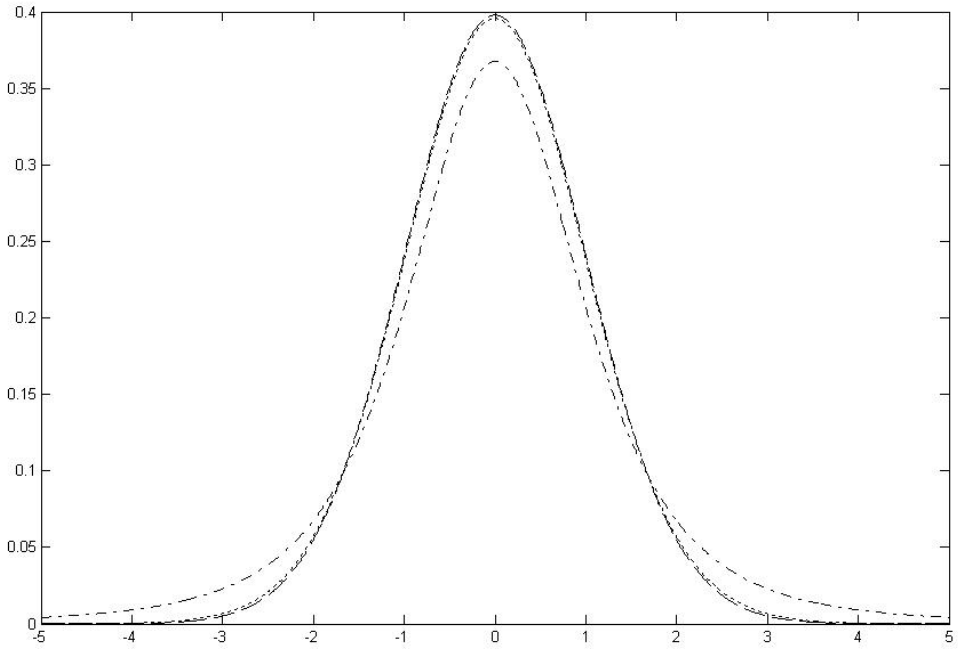


Figura 2.9 – Distribuição *t* de Student. A distribuição com três graus de liberdade (traço-ponto) é mais achatada que aquela com 30 graus de liberdade (pontilhado). Existe uma diferença muito pequena entre esta última e aquela com 100 graus de liberdade (traço).

É importante observar que esse pequeno número de replicações, comum em microarranjos, possui dois possíveis efeitos que podem impactar no resultado de um teste *t* de Student. Em primeiro lugar, uma distribuição de probabilidade mais achatada irá acarretar na necessidade de uma estatística *t* mais elevada de forma a garantir um valor de probabilidade baixo o suficiente para alcançar uma significância estatística (Figura 2.9).

O segundo efeito está relacionado ao cálculo do erro-padrão da média (Equação 2.6). Para uma mesma média e desvio-padrão das razões, quanto maior o número de replicações, maior o valor da estatística *t*.

O pressuposto de normalidade da população de origem da amostra pode ser prescindido sem maiores prejuízos, podendo ser substituído pela exigência de uma população aproximadamente simétrica [14]. Mais ainda, a transformação logarítmica da razão, citada anteriormente, tende a normalizar a distribuição das razões [14, 28].

2.3.2 Teste t *Bayesiano*

A inferência Bayesiana é uma abordagem alternativa aos testes de hipótese. Seu nome se deve a Thomas Bayes, cujo trabalho, publicado postumamente em 1763, serviu de base para o desenvolvimento da inferência Bayesiana [35].

A probabilidade calculada a partir do Teorema de Bayes (Equação 2.7) recebe o nome de probabilidade condicional. Nesta abordagem do cálculo de probabilidades, é possível adicionar à informação obtida através da coleta de dados, o conhecimento do pesquisador sobre o fenômeno estudado.

$$\text{Prob}(P \mid D) = \frac{\text{Prob}(D \mid P) * \text{Prob}(P)}{\text{Prob}(D)} \tag{2.7}$$

No Teorema de Bayes, a probabilidade de que os dados (D) venham de uma determinada distribuição (P) é dado pelo produto da probabilidade que esses dados apresentam nessa distribuição (Prob(D|P)) e da probabilidade subjetiva que o pesquisador tem sobre a mesma distribuição (Prob(P)), ou probabilidade *a priori*, normalizados pela própria probabilidade total dos dados (Prob(D)). O resultado é a probabilidade posterior, ou *a posteriori*, de que os dados tenham sido originados na distribuição em questão.

A possibilidade de acrescentar informação subjetiva à informação obtida com os dados permite o tratamento estatístico de situações com baixa replicação ou mesmo eventos que não permitem replicação alguma [35].

A informação acrescentada aos dados na inferência Bayesiana se dá pela atribuição de uma distribuição de probabilidade ao parâmetro utilizado. Essa distribuição é denominada de distribuição *a priori* e, ao acrescentar informação aos dados, aumenta o poder estatístico destes. Na prática, a distribuição *a priori* tem o mesmo efeito que aumentar o tamanho da amostra utilizada. A informação acrescentada pode ter origem em conhecimento prévio do pesquisador sobre o problema ou nos próprios dados (Bayesiano empírico).

Um grande problema na utilização da inferência Bayesiana pura aplicada a experimentos de microarranjos está relacionada, novamente, ao baixo conhecimento não só sobre a atividade de um gene específico em uma condição determinada, mas também ao pouco conhecimento sobre a relação na atividade de vários genes analisados simultaneamente.

Uma possível solução para a estimação da distribuição *a priori* a ser utilizada em microarranjos é o método Bayesiano Empírico [37]. Neste método, os parâmetros da distribuição *a priori* são determinados a partir dos próprios dados gerados pelo experimento. Por isso, esse método não é considerado propriamente um método

Bayesiano [55]. Exemplos de aplicação em dados de microarranjos desse método foram apresentados por BALDI *et al.* [8], TUSHER *et al*. [66], EFRON *et al*. [24] e outros. Atualmente, o número de métodos Bayesianos desenvolvidos para microarranjos é muito grande e continua crescendo. Na prática de microarranjos, os métodos Bayesianos têm sido utilizados melhorar a estimação da dispersão dos dados com poucas replicações.

O método proposto por BALDI *et al*. [8], por exemplo, utiliza a média dos desvios-padrões dos genes com intensidade de sinal similar para estimar o desvio-padrão da distribuição *a priori* de cada gene, que é combinada no denominador da estatística *t*.

$$t = \frac{\bar{r}}{\sqrt{\left(\frac{\nu_0 \sigma_0^{\,2} + (n-1)s^2}{\nu_0 + n - 2}\right)\frac{1}{n}}} \tag{2.8}$$

A média das razões ($\bar{r}$) é estimada da forma usual, assim como a variância das razões (s^2). O desvio-padrão da *priori* (σ_0) tem grau de confiança (ν_0) determinado pelo pesquisador e varia em função do número de amostras utilizadas, segundo a equação

$$K = \nu_0 + n \tag{2.9}$$

Os autores descrevem que o valor de K pode ser estabelecido como sendo igual a duas ou três vezes o número de amostras e utilizam a distribuição Normal como modelo a ser adotado para a distribuição *a priori*.

A estatística *t* observada aqui será utilizada como aquela calculada pelo teste de Student, sendo genes DE aqueles com diferença significativa.

CAPÍTULO 3

MATERIAIS E MÉTODOS

3.1 Simulação

3.1.1 Conjunto Sem Distorção

Os dados utilizados na comparação do desempenho dos testes estatísticos foram simulados através do modelo de CUI *et al.* [18].

$$Y_{ij} = a_i + b_i X_{ij} e^{\eta_j + \zeta_{ij}} + \varepsilon_j + \delta_{ij} \quad (3.1)$$

Onde o sinal medido no gene *j* na condição *i* (Y_{ij}) é a combinação linear do ruído de fundo da lâmina, ou *background* (a_i), do sinal real de hibridação ocorrida (X_{ij}) e dos erros aditivos (ε_j – comum às duas condições; δ_{ij}– específico de cada condição) e multiplicativos (η_j – comum; ζ_{ij} – específico). O valor do sinal real X_{ij} foi extraído de uma distribuição Lognormal (7±1,9) [18, 32]. Para maior controle dos resultados, o sinal real foi igual nas duas condições para um mesmo gene (X_{1j}=X_{2j}) [18]. Os erros aditivos e multiplicativos foram extraídos de distribuições Normais com média igual a zero. Os desvios das distribuições de erro e os valores de *background* e do coeficiente do modelo são apresentados na Tabela 3.1. Estes valores foram replicados de CUI *et al.* [18].

Cada conjunto de simulação foi composto de 50 lâminas, com 4000 genes em cada. Foram selecionados aleatoriamente 80 genes em cada lâmina para representarem genes diferencialmente expressos (DE), ou seja, aqueles que apresentam nível de expressão maior ou menor em uma das condições analiadas. Os mesmos genes foram utilizados em todas as lâminas de cada conjunto a cada

simulação. Quatro fatores de expressão (1,5, 2, 3 e 4) foram aplicados de maneira uniforme aos genes DE, multiplicando o sinal real (X_{ij}) em uma das duas condições.

Tabela 3.1. Valor do desvio-padrão das distribuições dos erros, *background* e coeficiente do modelo de simulação utilizado.

Parâmetro	Símbolo	σ
Erro Aditivo Comum	ε_j	20
Erro Aditivo Específico	δ_{ij}	50
Erro Multiplicativo Comum	η_j	0,2
Erro Multiplicativo Específico	ζ_{ij}	0,1
Background	a_i	300
Coeficiente	b_i	1,0

O valor obtido para cada gene em cada condição (Y_{ij}) foi transformado pelo logaritmo e a razão entre as condições foi tomada como o valor para o gene na lâmina, sendo utilizado para os testes estatísticos.

3.1.2 Conjuntos Com Distorção

Com o objetivo de avaliar o impacto dos métodos de normalização no desempenho dos testes estatísticos, foram simulados três diferentes conjuntos simulando distorções nos dados que podem ser observadas em experimentos de microarranjos: diferença de *background*, diferença de inclinação e distorção

heterogênea. As distorções desejadas foram obtidas através da manipulação dos parâmetros do modelo de simulação, conforme descrito na Tabela 3.2.

A distorção heterogênea foi simulada de forma que metade dos genes [2] apresentasse as características da diferença de *background* e o restante apresentasse as características da diferença de inclinação. Os conjuntos com distorção foram simulados da mesma forma que os sem distorção, exceto pelas alterações dos parâmetros descritas acima.

O processo de simulação com e sem distorção foi repetido 10 vezes, sendo utilizados os valores médios nas comparações.

3.2 Métodos de Transformação

Três diferentes métodos de transformação foram aplicados aos conjuntos simulados com e sem distorção: *Shift*, *Lowess* e *Linlog*. Os métodos de transformação foram aplicados utilizando o pacote MAANOVA (disponível em http://www.jax.org/staff/churchill/labsite/software/anova/index.html, acessado em 03/10/2006), tendo sido utilizados os valores padrões dos parâmetros. De forma simplificada, para o método *shift*, foram utilizados os valores de –200 e 200 como limites para a constante C (ver Equação 2.4). No método *Lowess*, os três parâmetros foram 0,2 para a fração de genes f e três iterações (t). O fator de corte para o método *Linlog* foi de 0,3, indicando que os 30% dos genes com menor valor de expressão serão linearmente transformados.

Tabela 3.2. Valor do desvio-padrão das distribuições dos erros, *background* e coeficiente utilizados nas simulações com distorção.

		Distorção	
Parâmetro	**Símbolo**	***Background***	**Inclinação**
Erro Aditivo Comum (valor para σ)	ε_j	20	20
Erro Aditivo Específico (condição 1-valor para σ)	δ_{1j}	70	50
Erro Aditivo Específico (condição 2-valor para σ)	δ_{2j}	50	50
Erro Multiplicativo Comum (valor para σ)	η_j	0,2	0,2
Background (condição 1)	a_1	450	300
Background (condição 2)	a_2	150	300
Coeficiente (condição 1)	b_1	1,0	0,5
Coeficiente (condição 2)	b_2	1,0	1,5

3.3 Testes Estatísticos

O teste *t* de Student (Equação 2.6) e o teste *t* Bayesiano (Equação 2.8) foram aplicados aos conjuntos com e sem distorção simulados para comparação do seu desempenho na detecção de genes DE.

O resultado dos testes foi avaliado em função do número de genes DE apontados (verdadeiro-positivos, VP) e do número de genes normo-expressos apontados como DE (falso-positivos, FP).

Para o teste *t* Bayesiano, o parâmetro K foi estabelecido como duas vezes o número de lâminas utilizadas em cada comparação. Este valor foi descrito por BALDI *et al.* [8] como satisfatório.

Devido ao grande número de variáveis (genes) analisadas simultaneamente, fez-se necessária a realização de correção para testes múltiplos [12], utilizando-se a correção de etapa única de Sidák [22],

$$\tilde{p} = 1-\left(1-\hat{p}\right)^{g} \tag{3.2}$$

onde $\tilde{p}$ é o valor observado de *p* ($\hat{p}$) ajustado pelo número de testes realizados (*g*), no caso, 4000 ou um teste para cada gene analisado.

3.4 Número de Replicações

Para observar o efeito do número de replicações no desempenho dos testes, com e sem distorção e transformação, foram analisados os resultados das comparações entre as duas condições utilizando três, cinco, dez, 15, 20, 25, 30, 35, 40, 45 e 50 lâminas.

CAPÍTULO 4

RESULTADOS

As simulações utilizando os parâmetros descritos nas Tabelas 3.1 e 3.2 apresentaram os efeitos esperados para os conjuntos com e sem distorção. A Figura 4.1 mostra um exemplo de cada conjunto simulado antes e após a aplicação dos métodos de transformação.

O impacto das transformações sobre o desempenho dos dois testes foi similar, seja nos casos em que houve melhoria ou diminuição na capacidade de identificação dos genes DE (Tabelas 4.1 a 4.4). Os desvios-padrões apresentados em todos os resultados foram extremamente baixos em relação ao número de genes apontados, indicando uma grande estabilidade nos resultados. Dada a essa estabilidade, os valores das Tabelas são apresentados apenas em médias, sem os desvios-padrões, facilitando a visualização. Os resultados completos para cada teste e método de transformação são apresentados no Apêndice. O teste Bayesiano apresentou um desempenho melhor do que o teste de Student, independentemente do método de transformação utilizado.

O método *Lowess* foi capaz de aumentar o número de genes DE corretamente identificados tanto no conjunto com distorção de inclinação quanto naquele com distorção de *background*. O método *Shift* mostrou ser vantajoso apenas no conjunto com distorção de *background*. O método *Linlog* não ocasionou melhoria no desempenho dos testes estatísticos em nenhum dos conjuntos.

Com qualquer um dos métodos de transformação utilizados, o aumento no número de lâminas melhorou o desempenho dos dois testes. As figuras 4.2 e 4.3

mostram a relação entre o número de amostras e o percentual de genes DE encontrados em cada um dos testes estatísticos utilizados.

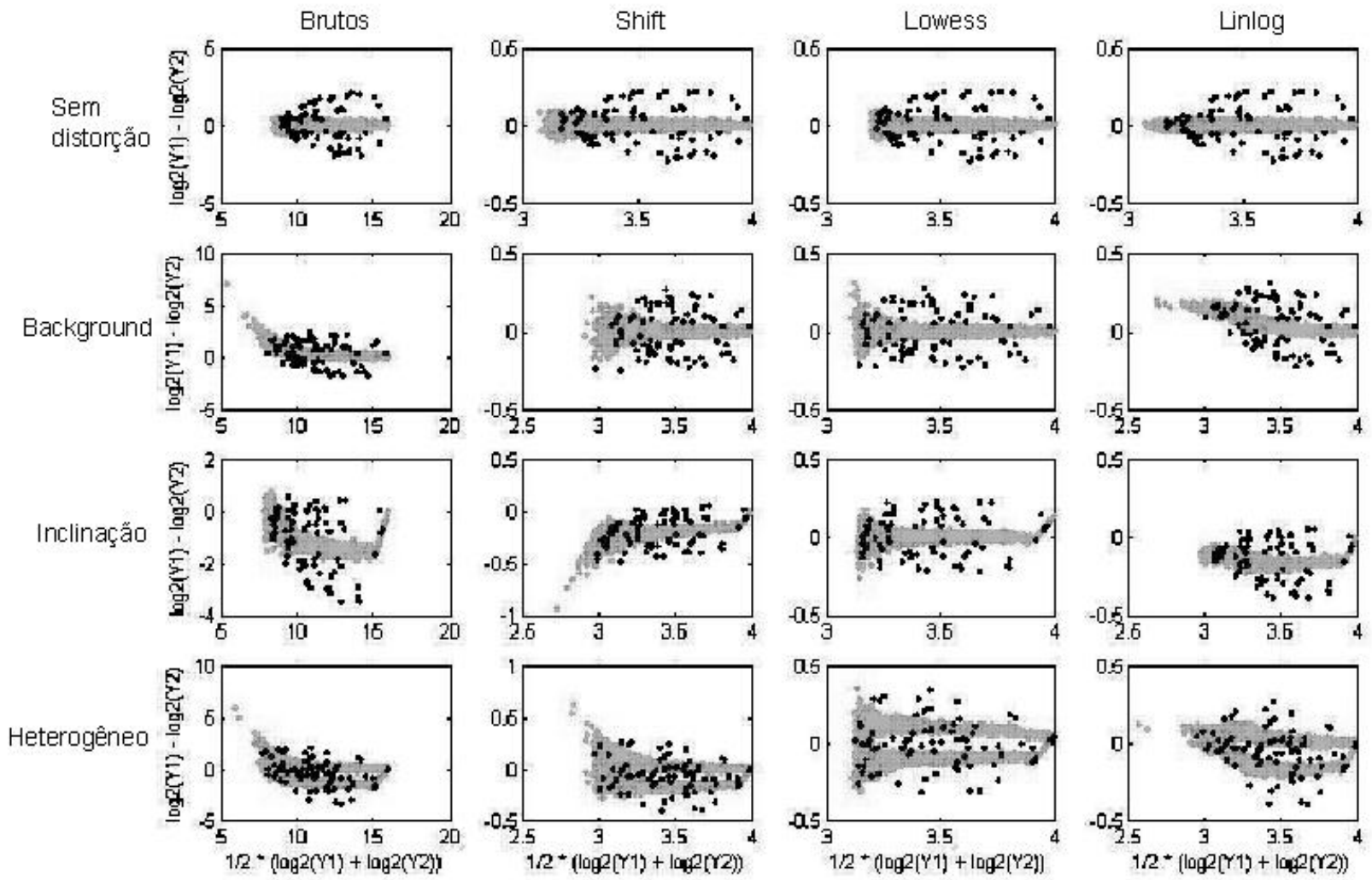


Figura 4.1 – Exemplos de resultados com e sem distorções, antes e após as transformações. Em cinza, genes normo-expressos; Em preto, genes diferencialmente expressos.

Tabela 4.1 – Número médio de genes DE indicados pelos dois testes no conjunto “Sem distorção”. Verdadeiro-positivos (VP) e falso-positivos (FP).

		3		5		10		15		20		25		30		35		40		45		50	
		FP	VP	FP	VP	FP	VP	FP	VP	FP	VP	FP	VP	FP	VP	FP	VP	FP	VP	FP	VP	FP	VP
Student	Bruto	0	0	0	1	0	23	0	66	0	78	0	80	0	80	0	80	0	80	0	80	0	80
	Shift	0	0	0	4	0	33	0	69	0	78	0	80	0	80	0	80	0	80	0	80	0	80
	Lowess	0	0	0	4	0	33	0	69	0	78	0	80	0	80	0	80	0	80	0	80	0	80
	Linlog	0	0	0	3	0	20	0	63	0	78	0	80	0	80	1	80	0	80	1	80	0	80
Bayesiano	Bruto	0	17	0	45	0	77	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0	80
	Shift	0	21	0	47	0	76	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0	80
	Lowess	0	21	0	47	0	76	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0	80
	Linlog	0	21	0	43	0	78	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0	80

Tabela 4.2 – Número de genes DE indiVPados pelos dois testes no VPonjunto VPom distorção de *Background*. Verdadeiro-positivos (C) e falso-positivos (FP).

		3		5		10		15		20		25		30		35		40		45		50	
		FP	VP	FP	VP	FP	VP	FP	VP	FP	VP	FP	VP	FP	VP	FP	VP	FP	VP	FP	VP	FP	VP
Student	Bruto	0	0	0	4	5	35	93	46	646	52	1882	56	3087	58	3698	60	3878	62	3915	63	3920	64
	Shift	0	0	0	3	0	34	0	66	0	76	0	78	0	80	0	80	0	80	0	80	0	80
	Lowess	0	0	0	3	0	34	0	66	0	76	0	78	0	80	0	80	0	80	0	80	0	80
	Linlog	0	0	1	4	14	35	190	45	1091	49	2564	53	3543	57	3853	59	3913	60	3920	61	3920	62
Bayesiano	Bruto	0	7	6	31	135	54	791	59	1943	62	2983	63	3583	64	3822	66	3897	67	3914	68	3918	68
	Shift	0	0	0	3	0	34	0	66	0	76	0	78	0	80	0	80	0	80	0	80	0	80
	Lowess	0	20	0	44	0	74	0	79	0	80	0	80	0	80	0	80	0	80	0	80	0	80
	Linlog	1	5	9	30	231	53	1251	58	2600	60	3458	62	3802	63	3894	63	3915	65	3919	66	3920	66

Tabela 4.3 – Número de genes DE indicados pelos dois testes no conjunto com distorção de Inclinação. Verdadeiro-positivos (VP) e falso-positivos (FP).

		3		5		10		15		20		25		30		35		40		45		50	
		FP	VP	FP	VP	FP	VP	FP	VP	FP	VP	FP	VP	FP	VP	FP	VP	FP	VP	FP	VP	FP	VP
Student	Bruto	1	0	23	0	652	19	2873	55	3849	67	3919	69	3920	70	3920	70	3920	70	3920	70	3920	70
	Shift	6	0	397	9	3628	47	3920	58	3920	60	3920	62	3920	65	3920	68	3920	69	3920	70	3920	70
	Lowess	0	0	0	1	0	28	0	58	0	70	0	75	0	78	0	79	0	79	0	80	0	80
	Linlog	3	0	100	3	3008	40	3918	61	3920	61	3920	61	3920	61	3920	62	3920	63	3920	64	3920	65
Bayesiano	Bruto	24	3	252	20	2995	42	3883	48	3920	51	3920	54	3920	58	3920	60	3920	63	3920	66	3920	68
	Shift	1576	30	3649	48	3920	61	3920	64	3920	67	3920	69	3920	70	3920	70	3920	70	3920	71	3920	71
	Lowess	0	13	0	39	0	67	0	75	0	78	0	79	0	80	0	80	0	80	0	80	0	80
	Linlog	181	11	2126	34	3920	54	3920	60	3920	60	3920	60	3920	61	3920	62	3920	62	3920	62	3920	63

Tabela 4.4 – Número de genes DE indicados pelos dois testes no conjunto com distorção Heterogênea. Verdadeiro-positivos (VP) e falso-positivos (FP).

		3		5		10		15		20		25		30		35		40		45		50	
		FP	VP	FP	VP	FP	VP	FP	VP	FP	VP	FP	VP	FP	VP	FP	VP	FP	VP	FP	VP	FP	VP
Student	Bruto	1	0	12	3	326	28	1489	51	2257	60	2904	64	3511	65	3817	67	3902	67	3918	68	3920	68
	Shift	1	0	31	5	242	28	865	48	1603	58	2203	61	2773	65	3318	66	3675	68	3841	69	3901	70
	Lowess	1	0	4	4	64	26	133	44	241	58	386	67	565	69	800	70	1064	70	1347	70	1634	70
	Linlog	1	0	52	4	1297	38	2031	53	2424	56	3155	58	3682	60	3880	62	3916	63	3920	64	3920	65
Bayesiano	Bruto	55	5	512	24	1852	47	1980	58	2074	62	2320	65	2724	66	3147	68	3496	68	3723	69	3842	69
	Shift	64	6	403	20	1479	45	1897	56	1966	61	1998	64	2050	65	2159	67	2341	68	2580	69	2849	70
	Lowess	0	3	7	16	38	47	109	57	245	61	433	64	679	66	976	68	1298	70	1650	70	1984	70
	Linlog	66	6	833	31	2059	52	2513	60	3171	61	3623	62	3837	63	3903	64	3917	64	3920	65	3920	66

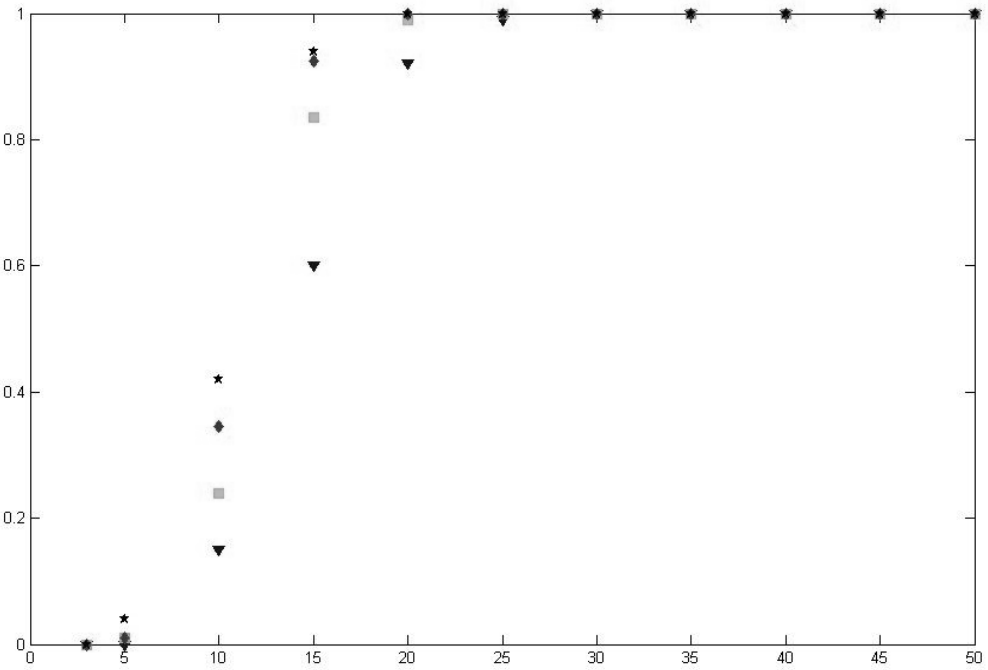


Figura 4.2 – Relação entre o percentual de genes DE encontrados, de acordo com o fator de expressão, e o número de amostras utilizados no teste *t* de Student. Fator 1,5: triângulo; Fator 2: quadrado; Fator 3: losango; Fator 4: estrela.

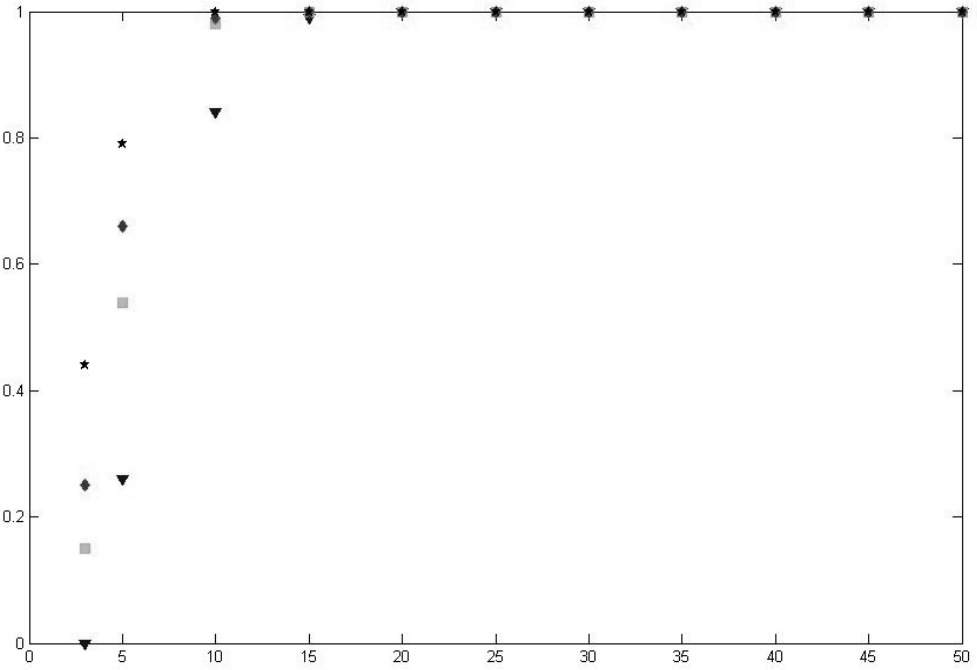


Figura 4.3 – Relação entre o percentual de genes DE encontrados, de acordo com o fator de expressão, e o número de amostras utilizados no teste *t* Bayesiano. Fator 1,5: triângulo; Fator 2: quadrado; Fator 3: losango; Fator 4: estrela.

CAPÍTULO 5

DISCUSSÃO

A tecnologia de microarranjos vem recebendo crescente atenção, demonstrada pelo número cada vez maior de publicações sobre o assunto, aparecendo como uma das principais ferramentas de análise genética. Entretanto, para que possam ser exploradas as possibilidades da técnica, é necessário que as ferramentas de análise sejam confiáveis.

Das diferentes técnicas utilizadas na detecção de genes diferencialmente expressos em microarranjos, duas formas de teste *t* merecem destaque. A primeira, o teste de Student, pela sua simplicidade e larga utilização. A segunda, o teste Bayesiano de Baldi e Long [8], pelos aparentes excelentes resultados descritos.

Um número grande de pesquisadores vem abordando diferentes métodos de seleção de genes em microarranjos. No entanto, a confiabilidade dos resultados de um experimento de microarranjo depende da qualidade dos dados gerados. Diversas fontes de variabilidade podem estar presentes em um experimento desse tipo, originadas no procedimento experimental ou durante o tratamento dos dados. De especial interesse são as fontes de erro sistemático, que podem induzir a resultados enganosos.

A utilização de dados simulados oferece uma saída para a falta de conhecimento sobre o resultado esperado de um microarranjo, permitindo maior controle dos resultados.

Uma importante limitação nessa abordagem está na capacidade do modelo utilizado na simulação de reproduzir as características dos dados reais gerados em um

microarranjo. O modelo utilizado aqui, apresentado por CUI *et al.* [18] demonstrou um comportamento bastante similar a dados gerados em experimentos reais, o que foi demonstrado pelos próprios autores[18]. Também observa-se esse comportamento ao compararmos os dados simulados com dados reais analisados em nosso laboratório (dados não-publicados). Dessa forma, acredita-se que os dados simulados reproduzem suficientemente a realidade de experimentos de microarranjo.

Os dois tipos de testes *t* aqui comparados objetivam a detecção de tendências nos resultados de um mesmo gene ao longo das replicações. A principal diferença entre eles está na determinação da dispersão dos dados (denominador das Equações 2.6 e 2.8). A utilização de uma distribuição *a priori* pelo teste Bayesiano acrescenta informação ao teste, aumentando seu poder estatístico [35]. Neste trabalho, os dois testes analisados mostraram grande sensibilidade quando combinados a um método de transformação adequado. Porém, o teste Bayesiano se mostrou sempre igual ou superior ao teste de Student na detecção de genes DE, independentemente do método de transformação ou do número de lâminas utilizados.

O número excepcionalmente baixo de genes falso-positivos em qualquer dos dois testes, na maioria dos casos, pode estar relacionado às características do modelo de simulação utilizado, apesar do que já foi discutido acima.

Os três tipos de transformação comparados representam três diferentes abordagens à transformação de dados de microarranjos. O método *Shift* objetiva a redução do erro aditivo, o *Lowess* tem por alvo o erro dependente da intensidade, essencialmente multiplicativo, e o método *Linlog* atua sobre os dois tipos de erro, objetivando a estabilização da variância. No exemplo da Figura 4.1 pode ser observado que os três métodos de transformação foram capazes de reduzir a variabilidade dos dados, sem alterar a forma da distribuição dos mesmos. No entanto,

a redução da variabilidade não se traduziu, necessariamente, em aumento no desempenho dos testes.

Os resultados demonstraram que o método *Shift* foi capaz de reduzir o erro sistemático ocasionado pela diferença de *background*, reduzindo a curvatura apresentada pelos dados, e estão em conformidade com os resultados de CUI *et al.* [18]. Estes mesmos autores descrevem um bom desempenho também no caso de diferenças de inclinação, o que não foi confirmado no trabalho atual. Também Kerr *et al.* [39] preferem o método *Shift* ao *Lowess* como forma de transformação. Ainda que o método tenha sido capaz de reduzir a curvatura dos dados, ele não contribuiu para a melhoria do desempenho dos testes estatísticos.

O método *Lowess*, por outro lado, foi capaz de reduzir a curvatura dos dados do conjunto com diferença de inclinação, conforme esperado, e também do conjunto com diferença de *background*, contribuindo para o aumento do desempenho dos testes estatísticos nos dois casos. Os resultados concordam com CUI *et al.* [18].

Embora este método venha se tornando um dos mais utilizados na análise de microarranjos, ele apresenta limitações, como demonstrado no conjunto heterogêneo simulado aqui. Nesse caso, o *Lowess* ocasionou o maior aumento no número de genes FP (piora de desempenho). No entanto, o aumento ocasionado pelos outros métodos também foi muito elevado, de forma que nenhum dos três seria recomendado para esse tipo de distorção.

Apesar de descrito por CUI *et al.* [18] como o método que proporciona o maior aumento no poder estatístico entre os três apresentados aqui, a transformação *Linlog* não foi capaz de melhorar a performance dos testes estatísticos em nenhum dos conjuntos simulados. A diferença nos resultados pode estar relacionada à utilização, por aqueles autores, do método da Razão na detecção de genes DE, ao invés de métodos estatísticos. Pela sua característica de estabilização da variância das razões

em uma única lâmina, o método pode levar a um maior número de genes DE apontados pela análise simples da razão entre os canais. Mas com o uso de replicações e testes estatísticos, esse resultado não foi observado. O método da Razão apresenta, porém, a importante limitação de não considerar a dispersão dos resultados quando mais de uma lâmina são analisadas no mesmo experimento.

Nenhum dos métodos de transformação utilizados neste trabalho foi capaz de melhorar o desempenho dos testes estatísticos no conjunto com distorção heterogênea. Ao contrário, a transformação dos dados gerou uma tendência a aumento no número de genes FP, sem aumento nos genes VP.

Resumindo, comparando-se os resultados dos métodos *Lowess* e *Shift*, pode-se afirmar que o primeiro é mais geral que o segundo, podendo ser utilizado em dois dos três tipos de distorção simulados neste trabalho. Essa conclusão está de acordo com YANG *et al.* [71] e PARK *et al.* [49]. Dada a dificuldade de estabelecer se a curvatura apresentada pelo gráfico das razões é originada de diferenças de inclinação ou *background*, recomenda-se, nesse caso, o uso do método *Lowess*. O terceiro tipo de distorção simulado (heterogêneo) necessita de métodos mais detalhados para a redução do erro sistemático presente.

Outro grande problema em microarranjos se relaciona ao baixo poder estatístico geralmente observado, devido ao grande número de genes avaliados e o pequeno número de replicações utilizado. Enquanto um experimento super-dimensionado, que utiliza uma amostra maior do que aquela necessária para a identificação dos efeitos esperados, ocasiona desperdício de material e tempo de pesquisa, um experimento sub-dimensionado apresenta baixo poder estatístico e pode comprometer a validade científica dos resultados encontrados, levando a conclusões equivocadas [37]. Em microarranjos, WEI e BUMGARNER [69] afirmam ser esse o caso mais comum.

O número mínimo de replicações depende, principalmente, do nível de significância estatística estabelecido, do poder estatístico desejado, do tamanho da variação esperada entre as condições sob análise e da variabilidade entre as amostras de uma mesma condição [47].

O surgimento, nos últimos anos, de pesquisas sobre a confiabilidade dos resultados de microarranjos [4, 5, 11] começam a trazer mais informações sobre a variabilidade nesse tipo de experimento, mas a determinação mais precisa do erro de medida ainda está distante de ser alcançada. Também a diferença esperada entre os níveis de expressão de diferentes genes em diferentes condições é difícil de determinar, dado o pouco conhecimento sobre a interação entre os genes e a resposta do seu nível de atividade sobre o fenótipo humano. A partir de que nível de sobre-expressão ou sub-expressão de um determinado gene pode-se observar variações fenotípicas ainda é uma questão a ser respondida.

Estas duas limitações, por si, dificultam, ou mesmo inviabilizam, a utilização de procedimentos padrões de determinação *a priori* do tamanho amostral ideal em microarranjos. Atualmente, as já citadas limitações de custo e tempo é que determinam o número de arranjos do experimento.

Iniciou-se a análise aqui a partir de três amostras (lâminas), número comum em diversas publicações, até 50 amostras, número excepcionalmente alto para microarranjos. Para a observação do impacto do número de amostras no desempenho dos testes foram considerados apenas os resultados dos dados simulados sem distorção e sem a aplicação de nenhum método de transformação (Tabela 4.1).

O número de genes verdadeiro-positivos encontrados pelos dois testes cresceu de acordo com o número de arranjos utilizados, conforme esperado. Para o teste de *Student*, foram necessários 20 arranjos para que mais de 95% dos genes diferencialmente expressos fossem identificados. No caso do teste Bayesiano, apenas

dez arranjos foram suficientes. Essa diferença parece estar associada ao uso de informação *a priori* no teste Bayesiano.

Valores de tamanho de amostra superiores a esses ocasionaram super dimensionamento da amostra, causando aumento nos custos sem melhoria no desempenho dos testes. Um importante achado é o fato de que, apesar do aumento nos custos do experimento, não foi observada redução do desempenho dos testes por aumento no número de genes FP.

Por outro lado, a maior diferença entre os testes foi observada nos casos dos menores tamanhos amostrais. Enquanto, com três amostras, o teste de *Student* não foi capaz de localizar nenhum gene DE, o teste Bayesiano identificou corretamente 17 genes. Com apenas cinco amostras, o teste Bayesiano foi capaz de localizar mais de 50% dos genes DE presentes, enquanto o teste de *Student* identificou apenas 1 gene DE. Essas diferenças são muito importantes devido ao custo financeiro e de tempo de um experimento de microarranjo. Um experimento desse tipo realizado com três a cinco amostras, valores normalmente utilizados, e analisado com o teste de *Student* não será capaz, de acordo com os resultados apresentados, de identificar um número significativo de genes DE que compense o custo do experimento.

Métodos mais recentes de análise de microarranjos foram desenvolvidos de forma a aumentar o poder estatístico mesmo em amostras muito pequenas. DRUMMOND *et al*. [21] propuseram uma forma de transformação dos dados, baseada na família Box-Cox, com seleção dos *outliers* como diferencialmente expressos e relataram uma sensibilidade próxima a 1,0, com taxas de falso-positivos inferiores a 0,15. Numa abordagem similar, também selecionando *outliers* em uma distribuição Normal, LOGUINOV *et al*. [43] reportaram resultados excelentes utilizando amostras unitárias, em dados simulados e reais. A avaliação destes métodos utilizando o

modelo de simulação aqui descrito seria importante na comparação com os resultados apresentados.

Com relação ao fator de expressão entre as condições, os resultados indicam uma grande dificuldade na detecção de genes DE com fator de expressão 1,5. No teste *t* de Student, por exemplo, enquanto mais de 80% dos genes com fator de expressão igual ou superior a 2 já haviam sido detectados com 15 amostras, menos de 60% dos genes com fator 1,5 foram encontrados.

Fatores de expressão mais baixos serão ainda mais difíceis de serem detectados, necessitando de um número maior de amostras. Existe pouco conhecimento, conforme já destacado, sobre o fator de expressão necessário para que alterações genotípicas possam causar efeito no fenótipo do organismo analisado. No entanto, esses resultados indicam que, se fatores de expressão muito baixos são o interesse do pesquisador de microarranjos, é necessário que um número alto de replicações seja utilizado.

CAPÍTULO 6

CONCLUSÕES E RECOMENDAÇÕES

Os dois testes estatísticos analisados mostraram bom desempenho. O teste t Bayesiano apresentou desempenho igual ou superior ao teste *t* de Student em todos os números de arranjos analisados, sendo o mais indicado na análise de microarranjos.

Com relação ao número de amostras, foram necessárias dez amostras para o teste Bayesiano alcançar 95% dos genes DE simulados, e o dobro foi necessário para o teste *t* de Student. Métodos de análise desenvolvidos para aplicação em pequenas amostras podem ser avaliados utilizando a o modelo de simulação utilizado.

Métodos de transformação podem melhorar o desempenho dos testes estatísticos, mas é necessário cuidado na seleção do método, de acordo com o tipo de distorção presente. Nenhum método de transformação analisado foi capaz de melhorar o resultado dos testes estatísticos no conjunto com distorção heterogênea, aqui simulado.

A metodologia de simulação aplicada aqui apresenta a importante vantagem do controle dos resultados esperados e a possibilidade de gerar um grande número de replicações sem custo adicional. Dessa forma, recomenda-se que a mesma seja utilizada na avaliação de outros métodos de transformação e análise de dados de microarranjos, podendo ainda ser adaptada de maneira a ser aplicada a métodos multivariados.

REFERÊNCIAS

1. ALBERTS,B.; JOHNSON,A.; LEWIS,J.; RAFF,M.; ROBERTS,K.; WALTER,P. **Biologia Molecular Da Célula**. 4a ed. Porto Alegre: ArtMed, 2004.
2. ALIZADEH,A.A.; EISEN,M.B.; DAVIS,R.E. et. al. "Distinct Types of Diffuse Large B-Cell Lymphoma Identified by Gene Expression Profiling", **Nature**, v. 403, n.6769, pp.503-511, 2000.
3. AOI,W.; ICHIISHI,E.; SAKAMOTO,N. et. al. "Effect of Exercise on Hepatic Gene Expression in Rats: a Microarray Analysis", **Life Sci**, v. 75, n.26, pp.3117-3128, 2004.
4. ASYALI,M.H. e ALCI,M. "Reliability Analysis of Microarray Data Using Fuzzy C-Means and Normal Mixture Modeling Based Classification Methods", **Bioinformatics**, v. 21, n.5, pp.644-649, 2005.
5. ASYALI,M.H.; SHOUKRI,M.M.; DEMIRKAYA,O. et. al. "Assessment of Reliability of Microarray Data and Estimation of Signal Thresholds Using Mixture Modeling", **Nucleic Acids Res**, v. 32, n.8, pp.2323-2335, 2004.
6. ATKINSON,G. e NEVILL,A.M. "Statistical Methods for Assessing Measurement Error (Reliability) in Variables Relevant to Sports Medicine", **Sports Med**, v. 26, n.4, pp.217-238, 1998.
7. BALAGURUNATHAN,Y.; DOUGHERTY,E.R.; CHEN,Y. et. al. "Simulation of CDNA Microarrays Via a Parameterized Random Signal Model", **J Biomed Opt**, v. 7, n.3, pp.507-523, 2002.
8. BALDI,P. e LONG,A.D. "A Bayesian Framework for the Analysis of Microarray Expression Data: Regularized t -Test and Statistical Inferences of Gene Changes", **Bioinformatics**, v. 17, n.6, pp.509-519, 2001.

9. BERGER,J.A.; HAUTANIEMI,S.; JARVINEN,A.K. et. al. "Optimized LOWESS Normalization Parameter Selection for DNA Microarray Data", **BMC Bioinformatics**, v. 5, n.1, pp.194, 2004.
10. BERRAR,D.P.; DUBITZKY,W.; GRANZOW,M. **A Practical Approach to Microarray Data Analysis**. Boston, MA: Kluwer Academic Publishers, 2003.
11. BILKE,S.; BRESLIN,T.; SIGVARDSSON,M. "Probabilistic Estimation of Microarray Data Reliability and Underlying Gene Expression", **BMC Bioinformatics**, v. 4, pp.40, 2003.
12. BLAND,J.M. e ALTMAN,D.G. "Multiple Significance Tests: the Bonferroni Method", **BMJ**, v. 310, n.6973, pp.170, 1995.
13. BLAND,J.M. e ALTMAN,D.G. "Statistics Notes. Logarithms", **BMJ**, v. 312, n.7032, pp.700, 1996.
14. BLAND,J.M. e ALTMAN,D.G. "Transforming Data", **BMJ**, v. 312, n.7033, pp.770, 1996.
15. BLAND,J.M. e ALTMAN,D.G. "Measuring Agreement in Method Comparison Studies", **Stat Methods Med Res**, v. 8, n.2, pp.135-160, 1999.
16. CLEVELAND,W.S. "Robust Locally Weighted Regression and Smoothing Scatterplots", **J Am Stat Assoc**, v. 74, pp.829-836, 1979.
17. CRICK,F.H. "On Protein Synthesis", **Symp Soc Exp Biol**, v. 12, pp.138-163, 1958.
18. CUI,X.; KERR,M.K.; CHURCHILL,G.A. "Transformations for CDNA Microarray Data", **Statistical Applications in Genetics and Molecular Biology**, v. 2, n.1, pp.1-20, 2003.
19. DEBOUCK,C. e GOODFELLOW,P.N. "DNA Microarrays in Drug Discovery and Development", **Nat Genet**, v. 21, n.1 Suppl, pp.48-50, 1999.
20. DERISI,J.; PENLAND,L.; BROWN,P.O. et. al. "Use of a CDNA Microarray to Analyse Gene Expression Patterns in Human Cancer", **Nat Genet**, v. 14, n.4, pp.457-460, 1996.

21. DRUMMOND,R.D.; PINHEIRO,A.; ROCHA,C.S. et. al. "ISER: Selection of Differentially Expressed Genes From DNA Array Data by Non-Linear Data Transformations and Local Fitting", **Bioinformatics**, v. 21, n.24, pp.4427-4429, 2005.
22. DUDOIT,S.; YANG,Y.H.; CALLOW,M.J. et. al. "Statistical Methods for Identifying Differentially Expressed Genes in Replicated CDNA Microarray Experiments", **Statistica Sinica**, v. 12, pp.111-139, 2002.
23. DURBIN,B. e ROCKE,D.M. "Estimation of Transformation Parameters for Microarray Data", **Bioinformatics**, v. 19, n.11, pp.1360-1367, 2003.
24. EFRON,B. e TIBSHIRANI,R. "Empirical Bayes Methods and False Discovery Rates for Microarrays", **Genet Epidemiol**, v. 23, n.1, pp.70-86, 2002.
25. EKINS,R.P. "Ligand Assays: From Electrophoresis to Miniaturized Microarrays", **Clin Chem**, v. 44, n.9, pp.2015-2030, 1998.
26. FEINSTEIN,A.R. "Clinical Biostatistics. LVI. The t Test and the Basic Ethos of Parametric Statistical Inference (Conclusion)", **Clin Pharmacol Ther**, v. 30, n.1, pp.133-146, 1981.
27. FELIX,J.M.; DRUMMOND,R.D.; NOGUEIRA,F.T. et. al. "Genoma Funcional", **Biotecnologia Ciência & Desenvolvimento**, n.24, pp.60-67, 2002.
28. FINNEY,D.J. "On the Distribution of a Variate Whose Logarithm Is Normally Distributed", **Supplement to the Journal of the Royal Statistical Society**, v. 7, n.2, pp.155-161, 1941.
29. FISHER,R.A. "Application of Student's Distribution", **Metron**, v. 5, pp.90-104, 1925.
30. GOLUB,T.R.; SLONIM,D.K.; TAMAYO,P. et. al. "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring", **Science**, v. 286, n.5439, pp.531-537, 1999.
31. GORDON,G.J.; JENSEN,R.V.; HSIAO,L.L. et. al. "Translation of Microarray Data into Clinically Relevant Cancer Diagnostic Tests Using Gene Expression

Ratios in Lung Cancer and Mesothelioma", **Cancer Res**, v. 62, n.17, pp.4963-4967, 2002.

32. HOYLE,D.C.; RATTRAY,M.; JUPP,R. et. al. "Making Sense of Microarray Data Distributions", **Bioinformatics**, v. 18, n.4, pp.576-584, 2002.
33. HUBER,W.; VON HEYDEBRECK,A.; SULTMANN,H. et. al. "Variance Stabilization Applied to Microarray Data Calibration and to the Quantification of Differential Expression", **Bioinformatics**, v. 18 Suppl 1, pp.S96-104, 2002.
34. HWANG,D.; ALEVIZOS,I.; SCHMITT,W.A. et. al. "Genomic Dissection for Characterization of Cancerous Oral Epithelium Tissues Using Transcription Profiling", **Oral Oncol**, v. 39, n.3, pp.259-268, 2003.
35. IVERSEN,G.R. **Bayesian Statistical Inference**. 3a ed. London: Sage, 1989.
36. JOOS,L.; ERYUKSEL,E.; BRUTSCHE,M.H. "Functional Genomics and Gene Microarrays--the Use in Research and Clinical Medicine", **Swiss Med Wkly**, v. 133, n.3-4, pp.31-38, 2003.
37. KENDZIORSKI,C.M.; NEWTON,M.A.; LAN,H. et. al. "On Parametric Empirical Bayes Methods for Comparing Multiple Groups Using Replicated Gene Expression Profiles", **Stat Med**, v. 22, n.24, pp.3899-3914, 2003.
38. KEPLER,T.B.; CROSBY,L.; MORGAN,K.T. "Normalization and Analysis of DNA Microarray Data by Self-Consistency and Local Regression", **Genome Biol**, v. 3, n.7, pp.RESEARCH0037, 2002.
39. KERR,M.K.; AFSHARI,C.A.; BENNETT,L. et. al. "Statistical Analysis of a Gene Expression Microarray Experiment With Replication", **Statistica Sinica**, v. 12, pp.203-217, 2002.
40. KIM,I.J.; KANG,H.C.; PARK,J.G. "Evaluation of Microarray Analysis for Predicting Treatment Responsiveness in Patients With Chronic Hepatitis C Viral Infection", **Gastroenterology**, v. 129, n.5, pp.1803-1804, 2005.
41. KNUDSEN,S. **A Biologist's Guide to Analysis of DNA Microarray Data**. New York: Wiley-Interscience, 2002.

42. LEHNINGER,A.L.; NELSON,D.L.; COX,M.M. **Lehninger Principles of Biochemistry**. 4a ed. New York: W.H. Freeman, 2005.
43. LOGUINOV,A.V.; MIAN,I.S.; VULPE,C.D. "Exploratory Differential Gene Expression Analysis in Microarray Experiments With No or Limited Replication", **Genome Biol**, v. 5, n.3, pp.R18, 2004.
44. MAEDA,S.; IEMITSU,M.; MIYAUCHI,T. et. al. "Aortic Stiffness and Aerobic Exercise: Mechanistic Insight From Microarray Analyses", **Med Sci Sports Exerc**, v. 37, n.10, pp.1710-1716, 2005.
45. MURPHY,G.M., JR. "Application of Microarray Technology in Psychotropic Drug Trials", **J Psychopharmacol**, v. 20, n.4 Suppl, pp.72-78, 2006.
46. PAN,W. "A Comparative Review of Statistical Methods for Discovering Differentially Expressed Genes in Replicated Microarray Experiments", **Bioinformatics**, v. 18, n.4, pp.546-554, 2002.
47. PAN,W.; LIN,J.; LE,C.T. "How Many Replicates of Arrays Are Required to Detect Gene Expression Changes in Microarray Experiments? A Mixture Model Approach", **Genome Biol**, v. 3, n.5, pp.research0022, 2002.
48. PAN,W.; LIN,J.; LE,C.T. "A Mixture Model Approach to Detecting Differentially Expressed Genes With Microarray Data", **Funct Integr Genomics**, v. 3, n.3, pp.117-124, 2003.
49. PARK,T.; YI,S.G.; KANG,S.H. et. al. "Evaluation of Normalization Methods for Microarray Data", **BMC Bioinformatics**, v. 4, pp.33, 2003.
50. PASSARGE,E. **Genética - Texto e Atlas**. 2a ed. Porto Alegre: Artmed, 2004.
51. QIN,L.X. e KERR,K.F. "Empirical Evaluation of Data Transformations and Ranking Statistics for Microarray Analysis", **Nucleic Acids Res**, v. 32, n.18, pp.5471-5479, 2004.
52. QUACKENBUSH,J. "Microarray Data Normalization and Transformation", **Nat Genet**, v. 32 Suppl, pp.496-501, 2002.

53. REUE,K. "MRNA Quantitation Techniques: Considerations for Experimental Design and Application", **J Nutr**, v. 128, n.11, pp.2038-2044, 1998.
54. RIVA,A.; CARPENTIER,A.S.; TORRESANI,B. et. al. "Comments on Selected Fundamental Aspects of Microarray Analysis", **Comput Biol Chem**, v. 29, n.5, pp.319-336, 2005.
55. ROBERT,C.P. **The Bayesian Choice**. 2a ed. New York: Springer, 2001.
56. ROCKE,D.M. e DURBIN,B. "A Model for Measurement Error for Gene Expression Arrays", **J Comput Biol**, v. 8, n.6, pp.557-569, 2001.
57. SAPIR, M. e CHURCHILL, G. A., 2000, "Estimating the Posterior Probability of Differential Gene Expression From Microarray Data" Acesso em 20/7/2006.
58. SCHENA,M.; HELLER,R.A.; THERIAULT,T.P. et. al. "Microarrays: Biotechnology's Discovery Platform for Functional Genomics", **Trends Biotechnol**, v. 16, n.7, pp.301-306, 1998.
59. SCHENA,M.; SHALON,D.; DAVIS,R.W. et. al. "Quantitative Monitoring of Gene Expression Patterns With a Complementary DNA Microarray", **Science**, v. 270, n.5235, pp.467-470, 1995.
60. SCHENA,M.; SHALON,D.; HELLER,R. et. al. "Parallel Human Genome Analysis: Microarray-Based Expression Monitoring of 1000 Genes", **Proc Natl Acad Sci U S A**, v. 93, n.20, pp.10614-10619, 1996.
61. SOUTHERN,E.M. "Detection of Specific Sequences Among DNA Fragments Separated by Gel Electrophoresis", **J Mol Biol**, v. 98, n.3, pp.503-517, 1975.
62. STOLOVITZKY,G. "Gene Selection in Microarray Data: the Elephant, the Blind Men and Our Algorithms", **Curr Opin Struct Biol**, v. 13, n.3, pp.370-376, 2003.
63. STUDENT "The Probable Error of a Mean", **Biometrika**, v. 6, n.1, pp.1-25, 1908.
64. TRIOLLA,M.F. **Introdução à Estatística**. 9a ed. Rio de Janeiro: LTC, 2005.

65. TROYANSKAYA,O.G.; GARBER,M.E.; BROWN,P.O. et. al. "Nonparametric Methods for Identifying Differentially Expressed Genes in Microarray Data", **Bioinformatics**, v. 18, n.11, pp.1454-1461, 2002.

66. TUSHER,V.G.; TIBSHIRANI,R.; CHU,G. "Significance Analysis of Microarrays Applied to the Ionizing Radiation Response", **Proc Natl Acad Sci U S A**, v. 98, n.9, pp.5116-5121, 2001.

67. VENTER,J.C.; ADAMS,M.D.; MYERS,E.W. et. al. "The Sequence of the Human Genome", **Science**, v. 291, n.5507, pp.1304-1351, 2001.

68. WATSON,J.D. e CRICK,F.H. "Molecular Structure of Nucleic Acids; a Structure for Deoxyribose Nucleic Acid", **Nature**, v. 171, n.4356, pp.737-738, 1953.

69. WEI,C.; LI,J.; BUMGARNER,R.E. "Sample Size for Detecting Differentially Expressed Genes in Microarray Experiments", **BMC Genomics**, v. 5, n.1, pp.87, 2004.

70. WU,L.; WILLIAMS,P.M.; KOCH,W.H. "Clinical Applications of Microarray-Based Diagnostic Tests", **Biotechniques**, v. 39, n.4, pp.577-582, 2005.

71. YANG,Y.H.; DUDOIT,S.; LUU,P. et. al. "Normalization for CDNA Microarray Data: a Robust Composite Method Addressing Single and Multiple Slide Systematic Variation", **Nucleic Acids Res**, v. 30, n.4, pp.e15, 2002.

APÊNDICE

TABELAS DE RESULTADOS

Tabela A1. Número de genes VP (C) e FP (FP) apontdos pelo teste t de Student no conjunto sem distorção e sem transformação, em relação ao tamanho da amostra

	3		5		10		15		20		25		30		35		40		45		50	
n	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**
1	0	0	2	0	18	0	66	0	79	0	80	0	80	1	80	3	80	1	80	1	80	1
2	0	0	2	0	22	0	63	0	77	0	80	0	80	0	80	0	80	0	80	1	80	0
3	0	0	0	0	23	0	65	0	79	0	80	0	80	0	80	0	80	0	80	0	80	0
4	0	0	1	0	19	0	67	0	78	0	80	0	80	0	80	0	80	0	80	0	80	0
5	0	0	1	0	24	0	67	1	79	0	79	0	80	0	80	0	80	0	80	0	80	0
6	0	0	2	0	22	0	64	0	77	0	80	0	80	0	80	0	80	0	80	0	80	0
7	0	0	2	0	26	0	71	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0
8	0	0	0	0	31	0	65	0	79	0	80	0	80	0	80	1	80	1	80	0	80	0
9	0	0	2	0	22	0	62	0	75	0	79	0	80	0	80	0	80	0	80	0	80	0
10	0	0	0	0	24	0	70	0	79	0	80	0	80	1	80	0	80	1	80	1	80	2
Média	**0**	**0**	**1**	**0**	**23**	**0**	**66**	**0**	**78**	**0**	**80**	**0**	**80**	**0**	**80**	**0**	**80**	**0**	**80**	**0**	**80**	**0**
Desvio-Padrão	**0,0**	**0,0**	**0,9**	**0,0**	**3,4**	**0,0**	**2,7**	**0,3**	**1,4**	**0,0**	**0,4**	**0,0**	**0,0**	**0,4**	**0,0**	**0,9**	**0,0**	**0,5**	**0,0**	**0,5**	**0,0**	**0,6**

Tabela A2. Número de genes VP (VP) e FP (FP) apontdos pelo teste t Bayesiano no conjunto sem distorção e sem transformação, em relação ao tamanho da amostra

	3		5		10		15		20		25		30		35		40		45		50	
n	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**
1	21	0	42	0	77	0	80	0	80	0	80	0	80	0	80	1	80	0	80	0	80	0
2	12	0	45	0	76	0	79	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0
3	18	0	43	0	76	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0
4	12	0	39	0	78	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0
5	25	0	51	0	75	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0
6	16	0	45	0	76	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0
7	16	0	44	0	76	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0
8	15	0	52	0	77	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0
9	16	0	40	0	74	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0
10	19	0	52	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0	80	1	80	2
Média	**17**	**0**	**45**	**0**	**77**	**0**	**80**	**0**	**80**	**0**	**80**	**0**	**80**	**0**	**80**	**0**	**80**	**0**	**80**	**0**	**80**	**0**
Desvio-Padrão	**3,8**	**0,0**	**4,6**	**0,0**	**1,6**	**0,0**	**0,3**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**	**0,3**	**0,0**	**0,0**	**0,0**	**0,3**	**0,0**	**0,6**

Tabela A3. Número de genes VP (VP) e FP (FP) apontdos pelo teste t de Student no conjunto sem distorção após transformação *Shift*, em relação ao tamanho da amostra

	3		5		10		15		20		25		30		35		40		45		50	
n	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**
1	0	0	3	0	29	0	70	0	79	0	80	0	80	1	80	2	80	1	80	1	80	1
2	0	0	1	0	32	0	68	0	77	0	80	0	80	0	80	0	80	0	80	1	80	0
3	0	0	6	0	33	0	68	0	78	0	80	0	80	0	80	0	80	0	80	0	80	0
4	1	0	1	0	33	0	68	0	78	0	79	0	80	0	80	0	80	0	80	0	80	0
5	0	0	7	0	33	0	69	1	77	0	79	0	79	0	80	0	80	0	80	0	80	0
6	0	0	4	0	35	0	69	0	78	0	80	0	80	0	80	0	80	0	80	0	80	0
7	0	0	4	0	31	0	73	0	78	0	80	0	80	0	80	0	80	0	80	0	80	0
8	1	0	4	0	37	0	71	0	77	0	80	0	80	0	80	1	80	1	80	0	80	0
9	0	0	4	0	33	0	66	0	75	0	79	0	80	0	80	0	80	0	80	0	80	0
10	0	0	7	0	34	0	72	0	79	0	80	0	80	1	80	0	80	0	80	1	80	2
Média	**0**	**0**	**4**	**0**	**33**	**0**	**69**	**0**	**78**	**0**	**80**	**0**	**80**	**0**	**80**	**0**	**80**	**0**	**80**	**0**	**80**	**0**
Desvio-Padrão	**0,4**	**0,0**	**2,0**	**0,0**	**2,0**	**0,0**	**2,0**	**0,3**	**1,1**	**0,0**	**0,5**	**0,0**	**0,3**	**0,4**	**0,0**	**0,6**	**0,0**	**0,4**	**0,0**	**0,5**	**0,0**	**0,6**

Tabela A4. Número de genes VP (VP) e FP (FP) apontdos pelo teste t Bayesiano no conjunto sem distorção após transformação *Shift*, em relação ao tamanho da amostra

	3		5		10		15		20		25		30		35		40		45		50	
n	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**
1	23	0	48	0	76	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0
2	16	0	50	0	76	0	79	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0
3	22	0	44	0	76	0	79	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0
4	15	0	40	0	78	0	79	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0
5	28	0	52	0	74	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0
6	22	0	47	0	75	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0
7	19	0	44	0	75	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0
8	23	0	56	0	76	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0
9	19	0	44	0	72	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0
10	22	0	49	0	79	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0	80	1
Média	**21**	**0**	**47**	**0**	**76**	**0**	**80**	**0**	**80**	**0**	**80**	**0**	**80**	**0**	**80**	**0**	**80**	**0**	**80**	**0**	**80**	**0**
Desvio-Padrão	**3,6**	**0,0**	**4,4**	**0,0**	**1,8**	**0,0**	**0,5**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**	**0,3**

Tabela A5. Número de genes VP (VP) e FP (FP) apontdos pelo teste t de Student no conjunto sem distorção após transformação *Lowess*, em relação ao tamanho da amostra

	3		5		10		15		20		25		30		35		40		45		50	
n	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**
1	0	0	3	0	28	0	69	0	80	0	80	0	80	1	80	2	80	1	80	1	80	1
2	0	0	2	0	33	0	68	0	77	0	80	0	80	0	80	0	80	0	80	0	80	0
3	0	0	7	0	32	0	68	0	78	0	80	0	80	0	80	0	80	0	80	0	80	0
4	1	0	1	0	32	0	70	0	78	0	79	0	80	0	80	0	80	0	80	0	80	0
5	0	0	7	0	33	0	69	1	77	0	79	0	79	0	80	0	80	0	80	0	80	0
6	0	0	3	0	34	0	69	0	78	0	80	0	80	0	80	0	80	0	80	0	80	0
7	0	0	4	0	31	0	73	0	79	0	80	0	80	0	80	0	80	0	80	0	80	0
8	1	0	5	0	38	0	71	0	77	0	80	0	80	0	80	1	80	1	80	0	80	0
9	0	0	4	0	32	0	66	0	75	0	80	0	80	0	80	0	80	0	80	0	80	0
10	0	0	7	0	34	0	71	0	78	0	80	0	80	1	80	0	80	1	80	1	80	2
Média	**0**	**0**	**4**	**0**	**33**	**0**	**69**	**0**	**78**	**0**	**80**	**0**	**80**	**0**	**80**	**0**	**80**	**0**	**80**	**0**	**80**	**0**
Desvio-Padrão	**0,4**	**0,0**	**2,1**	**0,0**	**2,4**	**0,0**	**1,9**	**0,3**	**1,3**	**0,0**	**0,4**	**0,0**	**0,3**	**0,4**	**0,0**	**0,6**	**0,0**	**0,5**	**0,0**	**0,4**	**0,0**	**0,6**

Tabela A6. Número de genes VP (VP) e FP (FP) apontdos pelo teste t Bayesiano no conjunto sem distorção após transformação *Lowess*, em relação ao tamanho da amostra

	3		5		10		15		20		25		30		35		40		45		50	
n	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**
1	24	0	48	0	76	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0
2	17	0	50	0	76	0	79	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0
3	23	0	43	0	76	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0
4	15	0	40	0	77	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0
5	27	0	52	0	74	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0
6	21	0	48	0	75	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0
7	19	0	44	0	75	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0
8	25	0	56	0	76	0	79	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0
9	18	0	44	0	73	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0
10	23	0	49	0	79	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0	80	1
Média	**21**	**0**	**47**	**0**	**76**	**0**	**80**	**0**	**80**	**0**	**80**	**0**	**80**	**0**	**80**	**0**	**80**	**0**	**80**	**0**	**80**	**0**
Desvio-Padrão	**3,7**	**0,0**	**4,5**	**0,0**	**1,6**	**0,0**	**0,4**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**	**0,3**

Tabela A7. Número de genes VP (VP) e FP (FP) apontdos pelo teste t de Student no conjunto sem distorção após transformação *Linlog*, em relação ao tamanho da amostra

	3		5		10		15		20		25		30		35		40		45		50	
n	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**
1	0	1	3	0	18	0	63	0	79	0	80	0	80	1	80	3	80	1	80	1	80	1
2	0	0	1	0	19	0	61	0	77	0	80	0	80	0	80	0	80	0	80	0	80	0
3	0	0	4	0	20	0	58	0	76	0	80	0	80	0	80	0	80	0	80	0	80	0
4	0	0	1	0	15	0	63	0	79	0	80	0	80	0	80	0	80	0	80	0	80	0
5	0	1	6	1	20	0	63	0	79	0	79	0	80	0	80	0	80	0	80	1	80	0
6	0	0	3	0	20	0	63	0	78	0	80	0	80	0	80	0	80	0	80	0	80	0
7	0	0	3	0	22	0	68	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0
8	1	0	5	0	26	0	63	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0
9	0	0	4	0	17	0	57	0	76	0	80	0	80	0	80	0	80	0	80	0	80	0
10	0	0	3	1	21	0	66	0	79	0	80	0	80	1	80	2	80	2	80	3	80	2
Média	**0**	**0**	**3**	**0**	**20**	**0**	**63**	**0**	**78**	**0**	**80**	**0**	**80**	**0**	**80**	**1**	**80**	**0**	**80**	**1**	**80**	**0**
Desvio-Padrão	**0,3**	**0,4**	**1,5**	**0,4**	**2,8**	**0,0**	**3,1**	**0,0**	**1,4**	**0,0**	**0,3**	**0,0**	**0,0**	**0,4**	**0,0**	**1,0**	**0,0**	**0,6**	**0,0**	**0,9**	**0,0**	**0,6**

Tabela A8. Número de genes VP (VP) e FP (FP) apontdos pelo teste t Bayesiano no conjunto sem distorção após transformação *Linlog*, em relação ao tamanho da amostra

	3		5		10		15		20		25		30		35		40		45		50	
n	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**
1	25	0	39	0	78	0	80	0	80	0	80	1	80	1	80	2	80	0	80	0	80	0
2	15	0	39	0	77	0	79	0	80	1	80	0	80	0	80	0	80	0	80	0	80	0
3	23	0	39	0	77	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0
4	15	0	38	0	78	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0
5	29	0	45	0	78	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0
6	21	0	42	0	77	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0
7	21	0	40	0	76	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0
8	20	0	51	0	79	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0
9	20	0	40	0	76	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0
10	21	0	53	0	80	0	80	0	80	0	80	0	80	0	80	0	80	1	80	2	80	3
Média	**21**	**0**	**43**	**0**	**78**	**0**	**80**	**0**	**80**	**0**	**80**	**0**	**80**	**0**	**80**	**0**	**80**	**0**	**80**	**0**	**80**	**0**
Desvio-Padrão	**4,0**	**0,0**	**5,1**	**0,0**	**1,2**	**0,0**	**0,3**	**0,0**	**0,0**	**0,3**	**0,0**	**0,3**	**0,0**	**0,3**	**0,0**	**0,6**	**0,0**	**0,3**	**0,0**	**0,6**	**0,0**	**0,9**

Tabela A9. Número de genes VP (VP) e FP (FP) apontdos pelo teste t de Student no conjunto com diferença de *background* sem transformação, em relação ao tamanho da amostra

	3		5		10		15		20		25		30		35		40		45		50	
n	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**
1	0	0	2	0	33	2	46	85	51	636	55	1879	58	3091	59	3699	62	3879	62	3917	62	3920
2	0	0	1	0	33	5	45	108	52	588	54	1863	58	3087	61	3679	62	3867	63	3914	63	3920
3	0	0	6	0	39	5	45	85	53	614	54	1887	56	3105	60	3694	61	3879	62	3913	63	3920
4	0	0	0	0	36	9	46	92	53	684	57	1872	60	3087	61	3703	60	3872	62	3910	63	3919
5	0	0	3	0	33	4	47	79	51	662	56	1865	59	3082	60	3712	63	3881	64	3912	64	3919
6	0	0	4	1	32	10	45	99	49	674	54	1902	56	3073	58	3691	62	3874	63	3917	65	3919
7	0	0	8	0	38	4	47	98	52	670	55	1906	58	3101	60	3726	63	3889	63	3917	65	3920
8	0	0	2	0	39	1	46	107	52	642	57	1875	56	3073	61	3681	63	3871	64	3916	64	3920
9	0	0	4	0	35	4	46	75	51	632	57	1890	58	3118	60	3701	61	3886	62	3916	63	3920
10	0	0	7	2	35	7	46	100	52	657	59	1880	60	3055	62	3691	63	3878	65	3915	64	3918
Média	**0**	**0**	**4**	**0**	**35**	**5**	**46**	**93**	**52**	**646**	**56**	**1882**	**58**	**3087**	**60**	**3698**	**62**	**3878**	**63**	**3915**	**64**	**3920**
Desvio-Padrão	**0,0**	**0,0**	**2,5**	**0,6**	**2,5**	**2,7**	**0,7**	**10,9**	**1,1**	**28,1**	**1,6**	**13,7**	**1,4**	**17,1**	**1,1**	**13,3**	**1,0**	**6,5**	**1,0**	**2,3**	**0,9**	**0,7**

Tabela A10. Número de genes VP (VP) e FP (FP) apontdos pelo teste t Bayesiano no conjunto com diferença de *background* sem transformação, em relação ao tamanho da amostra

	3		5		10		15		20		25		30		35		40		45		50	
n	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**
1	7	1	33	8	53	125	59	763	61	1917	61	2946	62	3589	63	3819	64	3895	65	3914	68	3917
2	7	0	25	4	51	131	57	781	64	1920	64	2978	66	3566	67	3818	68	3893	68	3911	67	3917
3	6	0	29	7	54	145	57	790	61	1901	62	2966	63	3576	65	3820	66	3897	67	3914	70	3918
4	6	0	32	5	56	130	61	797	61	1973	62	2991	65	3594	65	3825	67	3898	67	3917	67	3920
5	5	0	28	5	56	139	60	783	62	1943	64	2985	66	3583	66	3827	66	3900	69	3913	69	3920
6	6	0	29	6	53	136	59	772	62	1959	61	2994	61	3583	64	3828	66	3894	67	3915	66	3919
7	7	0	34	10	55	147	61	842	63	1980	63	3008	62	3590	66	3821	69	3899	69	3910	70	3916
8	9	1	31	7	57	140	58	800	63	1945	65	2987	66	3571	67	3818	69	3899	67	3915	69	3918
9	8	0	36	8	52	123	59	791	62	1931	62	2961	63	3602	65	3817	67	3894	68	3918	69	3919
10	6	0	29	2	55	134	61	788	62	1965	65	3014	66	3576	68	3826	69	3903	69	3915	69	3918
Média	**7**	**0**	**31**	**6**	**54**	**135**	**59**	**791**	**62**	**1943**	**63**	**2983**	**64**	**3583**	**66**	**3822**	**67**	**3897**	**68**	**3914**	**68**	**3918**
Desvio-Padrão	**1,1**	**0,4**	**3,1**	**2,2**	**1,8**	**7,6**	**1,5**	**20,1**	**0,9**	**24,7**	**1,4**	**19,9**	**1,9**	**10,5**	**1,4**	**4,0**	**1,6**	**3,0**	**1,2**	**2,3**	**1,3**	**1,2**

Tabela A11. Número de genes VP (VP) e FP (FP) apontdos pelo teste t de Student no conjunto com diferença de *background* após transformação *Shift*, em relação ao tamanho da amostra

	3		5		10		15		20		25		30		35		40		45		50	
n	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**
1	0	0	6	0	31	0	63	0	75	0	77	0	80	0	80	0	80	0	80	0	80	0
2	0	0	3	0	36	0	69	0	73	0	77	0	79	0	80	0	80	0	80	0	80	0
3	0	0	3	0	35	0	63	0	77	0	79	0	80	0	80	0	80	0	80	0	80	0
4	0	0	1	0	30	0	65	0	78	0	80	0	80	0	80	0	80	0	80	0	80	0
5	0	0	4	0	33	0	69	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0
6	0	0	7	0	32	0	65	0	72	0	75	0	80	0	80	0	80	0	80	0	80	0
7	0	0	4	0	36	0	66	0	77	0	79	0	79	0	79	0	80	0	80	0	80	0
8	0	0	2	0	39	0	69	0	77	0	80	0	80	0	80	0	80	0	80	0	80	0
9	0	0	1	0	37	0	65	0	76	0	79	0	80	0	80	0	80	0	80	0	80	0
10	0	1	3	0	30	0	66	0	75	0	78	0	80	0	80	0	80	0	80	0	80	0
Média	**0**	**0**	**3**	**0**	**34**	**0**	**66**	**0**	**76**	**0**	**78**	**0**	**80**	**0**	**80**	**0**	**80**	**0**	**80**	**0**	**80**	**0**
Desvio-Padrão	**0,0**	**0,3**	**1,9**	**0,0**	**3,0**	**0,0**	**2,2**	**0,0**	**2,2**	**0,0**	**1,6**	**0,0**	**0,4**	**0,0**	**0,3**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**

Tabela A12. Número de genes VP (VP) e FP (FP) apontdos pelo teste t Bayesiano no conjunto com diferença de *background* após transformação *Shift*, em relação ao tamanho da amostra

	3		5		10		15		20		25		30		35		40		45		50	
n	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**
1	0	0	6	0	31	0	63	0	75	0	77	0	80	0	80	0	80	0	80	0	80	0
2	0	0	3	0	36	0	69	0	73	0	77	0	79	0	80	0	80	0	80	0	80	0
3	0	0	3	0	35	0	63	0	77	0	79	0	80	0	80	0	80	0	80	0	80	0
4	0	0	1	0	30	0	65	0	78	0	80	0	80	0	80	0	80	0	80	0	80	0
5	0	0	4	0	33	0	69	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0
6	0	0	7	0	32	0	65	0	72	0	75	0	80	0	80	0	80	0	80	0	80	0
7	0	0	4	0	36	0	66	0	77	0	79	0	79	0	79	0	80	0	80	0	80	0
8	0	0	2	0	39	0	69	0	77	0	80	0	80	0	80	0	80	0	80	0	80	0
9	0	0	1	0	37	0	65	0	76	0	79	0	80	0	80	0	80	0	80	0	80	0
10	0	1	3	0	30	0	66	0	75	0	78	0	80	0	80	0	80	0	80	0	80	0
Média	**0**	**0**	**3**	**0**	**34**	**0**	**66**	**0**	**76**	**0**	**78**	**0**	**80**	**0**	**80**	**0**	**80**	**0**	**80**	**0**	**80**	**0**
Desvio-Padrão	**0,0**	**0,3**	**1,9**	**0,0**	**3,0**	**0,0**	**2,2**	**0,0**	**2,2**	**0,0**	**1,6**	**0,0**	**0,4**	**0,0**	**0,3**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**

Tabela A13. Número de genes VP (VP) e FP (FP) apontdos pelo teste t de Student no conjunto com diferença de *background* após transformação *Lowess*, em relação ao tamanho da amostra

	3		5		10		15		20		25		30		35		40		45		50	
n	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**
1	0	1	6	0	31	0	63	0	76	0	77	0	80	0	80	0	80	0	80	0	80	0
2	0	0	2	0	35	0	69	0	74	0	77	0	79	0	80	0	80	0	80	0	80	0
3	0	0	4	0	35	0	63	0	78	0	79	0	80	0	80	0	80	0	80	0	80	0
4	0	0	1	0	31	0	65	0	78	0	80	0	80	0	80	0	80	0	80	0	80	0
5	0	0	4	0	33	0	69	0	79	0	80	0	80	0	80	0	80	0	80	0	80	0
6	0	0	7	0	32	0	65	0	72	0	75	0	80	0	80	0	80	0	80	0	80	0
7	1	0	3	0	35	0	64	0	75	0	79	0	79	0	79	0	80	0	80	0	80	0
8	0	0	2	0	39	0	69	0	77	0	80	0	80	0	80	0	80	0	80	0	80	0
9	0	0	2	0	38	0	64	0	76	0	79	0	80	0	80	0	80	0	80	0	80	0
10	0	0	3	0	31	0	67	0	76	0	78	0	80	0	80	0	80	0	80	0	80	0
Média	**0**	**0**	**3**	**0**	**34**	**0**	**66**	**0**	**76**	**0**	**78**	**0**	**80**	**0**	**80**	**0**	**80**	**0**	**80**	**0**	**80**	**0**
Desvio-Padrão	**0,3**	**0,3**	**1,8**	**0,0**	**2,8**	**0,0**	**2,4**	**0,0**	**2,0**	**0,0**	**1,6**	**0,0**	**0,4**	**0,0**	**0,3**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**

Tabela A14. Número de genes VP (VP) e FP (FP) apontdos pelo teste t Bayesiano no conjunto com diferença de *background* após transformação *Lowess*, em relação ao tamanho da amostra

	3		5		10		15		20		25		30		35		40		45		50	
n	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**
1	18	0	37	0	75	0	80	0	79	0	80	0	80	0	80	0	80	0	80	0	80	0
2	20	0	44	0	75	0	77	0	79	0	80	0	80	0	80	0	80	0	80	0	80	0
3	19	0	48	0	73	0	79	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0
4	18	0	41	0	73	0	79	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0
5	17	0	40	0	68	0	79	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0
6	24	0	43	0	71	0	77	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0
7	17	0	48	0	74	0	79	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0
8	18	0	50	0	79	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0
9	21	0	48	0	75	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0
10	26	0	44	0	72	0	76	0	80	0	80	0	80	0	80	1	80	1	80	0	80	1
Média	**20**	**0**	**44**	**0**	**74**	**0**	**79**	**0**	**80**	**0**	**80**	**0**	**80**	**0**	**80**	**0**	**80**	**0**	**80**	**0**	**80**	**0**
Desvio-Padrão	**2,9**	**0,0**	**4,0**	**0,0**	**2,8**	**0,0**	**1,4**	**0,0**	**0,4**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**	**0,3**	**0,0**	**0,3**	**0,0**	**0,0**	**0,0**	**0,3**

Tabela A15. Número de genes VP (VP) e FP (FP) apontdos pelo teste t de Student no conjunto com diferença de *background* após transformação *Linlog*, em relação ao tamanho da amostra

	3		5		10		15		20		25		30		35		40		45		50	
n	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**
1	0	0	7	0	32	10	46	178	48	1064	51	2554	57	3531	59	3852	59	3918	60	3920	60	3920
2	0	0	1	0	33	22	45	194	50	1033	52	2534	55	3555	59	3844	60	3909	60	3920	61	3920
3	0	0	3	0	38	12	45	186	49	1102	53	2599	54	3556	57	3852	60	3915	61	3920	61	3920
4	0	0	2	2	38	17	44	180	49	1114	55	2576	59	3538	58	3849	60	3911	60	3920	61	3920
5	0	0	4	1	32	12	45	196	50	1121	54	2546	58	3532	59	3853	59	3911	62	3919	63	3920
6	0	0	5	1	31	15	42	193	46	1076	52	2573	56	3547	57	3852	59	3912	61	3919	62	3920
7	0	0	3	2	40	12	44	198	50	1123	54	2590	56	3567	59	3860	61	3912	62	3919	63	3920
8	0	0	2	2	39	9	45	191	49	1070	53	2570	55	3536	59	3854	61	3913	63	3919	64	3920
9	0	0	5	2	36	9	45	175	50	1087	52	2551	58	3545	60	3865	60	3915	60	3919	60	3920
10	0	0	5	2	35	18	45	211	51	1123	56	2549	60	3519	59	3850	61	3912	62	3920	63	3920
Média	**0**	**0**	**4**	**1**	**35**	**14**	**45**	**190**	**49**	**1091**	**53**	**2564**	**57**	**3543**	**59**	**3853**	**60**	**3913**	**61**	**3920**	**62**	**3920**
Desvio-Padrão	**0,0**	**0,0**	**1,7**	**0,9**	**3,1**	**4,1**	**1,0**	**10,3**	**1,3**	**28,9**	**1,5**	**19,7**	**1,8**	**13,5**	**0,9**	**5,5**	**0,8**	**2,4**	**1,0**	**0,5**	**1,3**	**0,0**

Tabela A16. Número de genes VP (VP) e FP (FP) apontdos pelo teste t Bayesiano no conjunto com diferença de *backgrond* após transformação *Linlog*, em relação ao tamanho da amostra

	3		5		10		15		20		25		30		35		40		45		50	
n	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**
1	3	1	29	9	52	217	57	1230	60	2596	61	3456	61	3802	62	3893	63	3917	64	3917	65	3920
2	5	1	28	8	51	219	57	1206	60	2583	62	3469	63	3787	65	3887	67	3914	66	3918	66	3920
3	4	3	28	11	53	231	56	1241	61	2575	62	3480	62	3811	63	3894	64	3916	65	3920	67	3920
4	8	0	31	6	55	247	59	1259	60	2603	61	3442	63	3807	63	3889	63	3914	64	3920	64	3920
5	4	1	28	6	55	243	60	1240	60	2594	64	3462	63	3806	65	3898	65	3914	68	3920	67	3920
6	2	2	29	6	51	236	59	1270	61	2609	60	3443	61	3795	62	3892	64	3915	64	3919	65	3920
7	6	0	32	14	54	237	57	1285	60	2659	63	3499	63	3808	64	3900	67	3913	68	3918	69	3920
8	5	1	32	5	55	235	58	1248	59	2591	63	3432	65	3789	63	3902	64	3917	65	3919	67	3919
9	6	1	33	18	51	214	55	1241	60	2584	61	3442	61	3813	62	3896	64	3916	66	3919	66	3920
10	5	4	29	6	54	233	58	1288	60	2604	63	3457	64	3803	65	3892	66	3913	66	3920	67	3920
Média	**5**	**1**	**30**	**9**	**53**	**231**	**58**	**1251**	**60**	**2600**	**62**	**3458**	**63**	**3802**	**63**	**3894**	**65**	**3915**	**66**	**3919**	**66**	**3920**
Desvio-Padrão	**1,6**	**1,2**	**1,8**	**4,0**	**1,6**	**10,5**	**1,4**	**24,0**	**0,5**	**22,1**	**1,2**	**19,3**	**1,3**	**8,5**	**1,2**	**4,5**	**1,4**	**1,4**	**1,4**	**1,0**	**1,3**	**0,3**

Tabela A17. Número de genes VP (VP) e FP (FP) apontdos pelo teste t de Student no conjunto com diferença de inclinação sem transformação, em relação ao tamanho da amostra

	3		5		10		15		20		25		30		35		40		45		50	
n	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**
1	0	2	0	20	22	628	56	2892	69	3839	70	3920	70	3920	70	3920	70	3920	70	3920	70	3920
2	0	0	1	17	21	658	62	2875	68	3850	69	3919	70	3920	70	3920	70	3920	70	3920	70	3920
3	0	1	1	24	29	663	55	2851	67	3861	67	3919	70	3920	70	3920	70	3920	70	3920	70	3920
4	0	0	0	22	22	650	56	2834	65	3861	70	3920	70	3920	70	3920	70	3920	70	3920	70	3920
5	0	1	0	25	14	647	51	2857	66	3828	70	3917	70	3920	70	3920	70	3920	70	3920	70	3920
6	0	2	1	17	16	645	54	2868	64	3854	69	3920	70	3920	70	3920	70	3920	70	3920	70	3920
7	0	0	1	19	20	656	55	2902	66	3852	70	3919	70	3920	70	3920	70	3920	70	3920	70	3920
8	0	2	0	25	15	664	52	2914	65	3853	70	3919	70	3920	70	3920	70	3920	70	3920	70	3920
9	0	0	0	27	19	626	53	2894	68	3842	69	3920	69	3920	70	3920	70	3920	70	3920	70	3920
10	0	0	0	35	16	683	57	2841	67	3846	69	3919	70	3920	70	3920	70	3920	70	3920	70	3920
Média	**0**	**1**	**0**	**23**	**19**	**652**	**55**	**2873**	**67**	**3849**	**69**	**3919**	**70**	**3920**	**70**	**3920**	**70**	**3920**	**70**	**3920**	**70**	**3920**
Desvio-Padrão	**0,0**	**0,9**	**0,5**	**5,2**	**4,2**	**16,1**	**2,9**	**25,8**	**1,5**	**9,7**	**0,9**	**0,9**	**0,3**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**

Tabela A18. Número de genes VP (VP) e FP (FP) apontdos pelo teste t Bayesiano no conjunto com diferença de inclinação sem transformação, em relação ao tamanho da amostra

	3		5		10		15		20		25		30		35		40		45		50	
n	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**
1	4	26	21	290	42	3013	48	3880	53	3920	55	3920	58	3920	60	3920	62	3920	65	3920	67	3920
2	1	22	18	242	44	2975	49	3882	51	3919	56	3920	59	3920	60	3920	60	3920	61	3920	66	3920
3	3	19	24	257	41	3015	48	3886	51	3920	53	3920	56	3920	58	3920	63	3920	68	3920	69	3920
4	4	31	17	264	42	2980	47	3886	50	3920	52	3920	60	3920	60	3920	62	3920	66	3920	70	3920
5	1	26	23	209	42	2951	49	3879	52	3918	54	3920	57	3920	60	3920	65	3920	68	3920	69	3920
6	3	28	17	235	41	2952	48	3878	51	3920	54	3920	56	3920	61	3920	62	3920	65	3920	68	3920
7	10	19	25	258	41	3013	49	3890	51	3920	53	3920	59	3920	61	3920	62	3920	64	3920	69	3920
8	2	30	22	268	42	3032	49	3885	51	3920	55	3920	58	3920	59	3920	62	3920	64	3920	66	3920
9	3	23	18	249	40	2993	47	3888	50	3919	54	3920	59	3920	61	3920	65	3920	67	3920	68	3920
10	2	18	19	244	42	3023	48	3878	51	3920	54	3920	56	3920	61	3920	64	3920	67	3920	68	3920
Média	**3**	**24**	**20**	**252**	**42**	**2995**	**48**	**3883**	**51**	**3920**	**54**	**3920**	**58**	**3920**	**60**	**3920**	**63**	**3920**	**66**	**3920**	**68**	**3920**
Desvio-Padrão	**2,5**	**4,5**	**2,8**	**20,5**	**1,0**	**27,6**	**0,7**	**4,1**	**0,8**	**0,7**	**1,1**	**0,0**	**1,4**	**0,0**	**0,9**	**0,0**	**1,5**	**0,0**	**2,1**	**0,0**	**1,3**	**0,0**

Tabela A19. Número de genes VP (VP) e FP (FP) apontdos pelo teste t de Student no conjunto com diferença de inclinação após transformação *Shift*, em relação ao tamanho da amostra

	3		5		10		15		20		25		30		35		40		45		50	
n	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**
1	0	2	10	412	45	3650	59	3919	60	3920	61	3920	65	3920	67	3920	67	3920	69	3920	70	3920
2	0	8	7	372	45	3649	58	3920	60	3920	61	3920	64	3920	68	3920	70	3920	70	3920	70	3920
3	1	5	12	404	46	3613	57	3920	61	3920	61	3920	66	3920	68	3920	70	3920	70	3920	70	3920
4	0	5	7	377	49	3614	59	3920	60	3920	61	3920	63	3920	66	3920	69	3920	69	3920	70	3920
5	0	5	13	393	45	3635	59	3919	61	3920	61	3920	63	3920	69	3920	70	3920	70	3920	70	3920
6	0	6	7	429	48	3614	56	3920	61	3920	61	3920	63	3920	68	3920	70	3920	70	3920	70	3920
7	0	9	12	406	50	3630	59	3920	60	3920	61	3920	63	3920	69	3920	70	3920	70	3920	70	3920
8	0	3	8	382	50	3630	59	3920	59	3920	62	3920	65	3920	68	3920	69	3920	70	3920	70	3920
9	0	7	8	392	44	3627	59	3920	61	3920	64	3920	67	3920	68	3920	70	3920	70	3920	70	3920
10	0	7	9	404	45	3617	58	3920	61	3920	62	3920	66	3920	67	3920	69	3920	69	3920	70	3920
Média	**0**	**6**	**9**	**397**	**47**	**3628**	**58**	**3920**	**60**	**3920**	**62**	**3920**	**65**	**3920**	**68**	**3920**	**69**	**3920**	**70**	**3920**	**70**	**3920**
Desvio-Padrão	**0,3**	**2,1**	**2,2**	**16,5**	**2,2**	**13,1**	**1,0**	**0,4**	**0,7**	**0,0**	**0,9**	**0,0**	**1,4**	**0,0**	**0,9**	**0,0**	**0,9**	**0,0**	**0,5**	**0,0**	**0,0**	**0,0**

Tabela A20. Número de genes VP (VP) e FP (FP) apontdos pelo teste t Bayesiano no conjunto com diferença de inclinação após transformação *Shift*, em relação ao tamanho da amostra

	3		5		10		15		20		25		30		35		40		45		50	
n	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**
1	29	1664	47	3671	59	3920	62	3920	67	3920	69	3920	70	3920	70	3920	70	3920	70	3920	70	3920
2	31	1595	44	3670	62	3920	62	3920	67	3920	69	3920	69	3920	70	3920	70	3920	70	3920	70	3920
3	31	1668	47	3664	60	3920	65	3920	66	3920	69	3920	70	3920	70	3920	70	3920	70	3920	70	3920
4	29	1573	49	3629	62	3920	66	3920	68	3920	68	3920	70	3920	70	3920	71	3920	71	3920	71	3920
5	30	1516	48	3636	61	3920	65	3920	67	3920	70	3920	70	3920	70	3920	70	3920	71	3920	71	3920
6	26	1551	46	3654	61	3920	65	3920	68	3920	69	3920	70	3920	70	3920	70	3920	70	3920	72	3920
7	35	1671	53	3654	62	3920	63	3920	67	3920	69	3920	71	3920	71	3920	71	3920	71	3920	71	3920
8	32	1455	49	3625	60	3920	64	3920	67	3920	69	3920	71	3920	71	3920	71	3920	70	3920	70	3920
9	28	1508	50	3638	61	3920	63	3920	67	3920	69	3920	71	3920	70	3920	70	3920	70	3920	71	3920
10	28	1555	47	3648	61	3920	66	3920	69	3920	70	3920	70	3920	70	3920	70	3920	72	3920	72	3920
Média	**30**	**1576**	**48**	**3649**	**61**	**3920**	**64**	**3920**	**67**	**3920**	**69**	**3920**	**70**	**3920**	**70**	**3920**	**70**	**3920**	**71**	**3920**	**71**	**3920**
Desvio-Padrão	**2,4**	**70,4**	**2,3**	**15,7**	**0,9**	**0,0**	**1,4**	**0,0**	**0,8**	**0,0**	**0,5**	**0,0**	**0,6**	**0,0**	**0,4**	**0,0**	**0,5**	**0,0**	**0,7**	**0,0**	**0,7**	**0,0**

Tabela A21. Número de genes VP (VP) e FP (FP) apontdos pelo teste t de Student no conjunto com diferença de inclinação após transformação *Lowess*, em relação ao tamanho da amostra

	3		5		10		15		20		25		30		35		40		45		50	
n	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**
1	1	0	1	0	30	0	61	0	69	0	74	0	76	0	79	0	79	0	80	0	80	0
2	0	0	2	0	31	0	58	0	68	0	75	0	77	0	79	0	79	0	80	0	80	0
3	0	0	0	0	28	0	60	0	72	0	76	0	80	0	79	0	79	0	80	0	80	0
4	0	0	1	0	29	0	58	0	72	0	75	0	79	0	80	0	79	0	80	0	80	0
5	0	0	2	0	21	0	59	0	73	0	75	0	76	0	78	0	80	0	80	0	80	0
6	0	1	3	1	30	0	61	0	68	0	77	0	79	0	79	0	79	0	80	0	80	0
7	0	0	1	0	34	0	57	0	72	0	76	0	77	0	79	0	80	0	80	0	80	0
8	0	0	1	0	23	0	54	0	66	0	72	0	80	0	80	0	80	0	80	0	80	0
9	0	1	1	0	28	0	52	0	68	0	75	0	79	0	80	0	80	0	80	0	80	0
10	0	0	1	0	24	0	55	0	67	0	72	0	76	0	78	0	79	0	79	0	79	0
Média	**0**	**0**	**1**	**0**	**28**	**0**	**58**	**0**	**70**	**0**	**75**	**0**	**78**	**0**	**79**	**0**	**79**	**0**	**80**	**0**	**80**	**0**
Desvio-Padrão	**0,3**	**0,4**	**0,8**	**0,3**	**3,8**	**0,0**	**2,9**	**0,0**	**2,4**	**0,0**	**1,6**	**0,0**	**1,6**	**0,0**	**0,7**	**0,0**	**0,5**	**0,0**	**0,3**	**0,0**	**0,3**	**0,0**

Tabela A22. Número de genes VP (VP) e FP (FP) apontdos pelo teste t Bayesiano no conjunto com diferença de inclinação após transformação *Lowess*, em relação ao tamanho da amostra

	3		5		10		15		20		25		30		35		40		45		50	
n	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**
1	15	0	41	0	69	0	73	0	76	0	79	0	80	0	80	0	80	0	80	0	80	0
2	10	0	41	0	69	0	75	0	78	0	79	0	80	0	80	0	80	0	80	0	80	0
3	17	0	43	0	71	0	76	0	79	0	80	0	80	0	80	0	80	0	80	0	80	0
4	11	0	33	0	67	0	77	0	80	0	80	0	80	0	80	0	80	0	80	0	80	0
5	12	0	39	0	66	0	77	0	77	0	78	0	80	0	80	0	80	0	80	0	80	0
6	8	0	36	0	68	0	77	0	79	0	79	0	79	0	80	0	80	0	80	0	80	0
7	18	0	44	0	65	0	75	0	78	0	80	0	80	0	80	0	80	0	80	0	80	0
8	11	0	40	0	65	0	74	0	79	0	80	0	80	0	80	1	80	0	80	0	80	0
9	15	0	41	0	64	0	76	0	79	0	80	0	80	0	80	0	80	0	80	0	80	0
10	12	0	36	0	66	0	74	0	75	0	78	0	79	0	79	0	80	0	80	0	80	0
Média	**13**	**0**	**39**	**0**	**67**	**0**	**75**	**0**	**78**	**0**	**79**	**0**	**80**	**0**	**80**	**0**	**80**	**0**	**80**	**0**	**80**	**0**
Desvio-Padrão	**3,0**	**0,0**	**3,3**	**0,0**	**2,1**	**0,0**	**1,4**	**0,0**	**1,5**	**0,0**	**0,8**	**0,0**	**0,4**	**0,0**	**0,3**	**0,3**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**

Tabela A23. Número de genes VP (VP) e FP (FP) apontdos pelo teste t de Student no conjunto com diferença de inclinação após transformação *Linlog*, em relação ao tamanho da amostra

	3		5		10		15		20		25		30		35		40		45		50	
n	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**
1	0	3	3	110	44	2984	60	3919	61	3920	61	3920	61	3920	61	3920	61	3920	62	3920	63	3920
2	0	3	4	100	39	2988	60	3916	61	3920	61	3920	60	3920	60	3920	61	3920	63	3920	65	3920
3	0	1	0	99	45	3039	61	3919	60	3920	61	3920	60	3920	62	3920	62	3920	65	3920	64	3920
4	0	5	2	84	43	2992	60	3918	60	3920	60	3920	61	3920	61	3920	63	3920	62	3920	64	3920
5	0	5	4	88	33	2970	60	3919	61	3920	63	3920	63	3920	64	3920	65	3920	65	3920	66	3920
6	0	0	4	99	37	2977	60	3918	60	3920	60	3920	61	3920	63	3920	61	3920	63	3920	64	3920
7	0	1	2	102	45	3026	61	3917	60	3920	61	3920	60	3920	61	3920	62	3920	63	3920	65	3920
8	0	2	4	104	38	3061	60	3918	60	3920	60	3920	61	3920	61	3920	62	3920	62	3920	62	3920
9	1	3	2	106	39	3048	63	3919	62	3920	62	3920	63	3920	64	3920	64	3920	65	3920	66	3920
10	0	2	2	111	39	2993	60	3919	61	3920	63	3920	63	3920	65	3920	66	3920	66	3920	66	3920
Média	**0**	**3**	**3**	**100**	**40**	**3008**	**61**	**3918**	**61**	**3920**	**61**	**3920**	**61**	**3920**	**62**	**3920**	**63**	**3920**	**64**	**3920**	**65**	**3920**
Desvio-Padrão	**0,3**	**1,6**	**1,3**	**8,2**	**3,7**	**30,9**	**0,9**	**1,0**	**0,7**	**0,0**	**1,1**	**0,0**	**1,2**	**0,0**	**1,6**	**0,0**	**1,7**	**0,0**	**1,4**	**0,0**	**1,3**	**0,0**

Tabela A24. Número de genes VP (VP) e FP (FP) apontdos pelo teste t Bayesiano no conjunto com diferença de inclinação após transformação *Linlog*, em relação ao tamanho da amostra

	3		5		10		15		20		25		30		35		40		45		50	
n	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**
1	13	196	34	2137	57	3920	59	3920	60	3920	61	3920	61	3920	61	3920	61	3920	61	3920	62	3920
2	9	181	34	2065	57	3920	60	3920	60	3920	60	3920	60	3920	60	3920	60	3920	60	3920	62	3920
3	10	191	39	2158	54	3920	60	3920	60	3920	60	3920	60	3920	61	3920	62	3920	63	3920	62	3920
4	14	178	28	2065	53	3920	59	3920	60	3920	60	3920	61	3920	61	3920	62	3920	61	3920	62	3920
5	8	147	36	2028	52	3920	60	3920	60	3920	60	3920	62	3920	64	3920	64	3920	64	3920	66	3920
6	9	174	31	2006	53	3920	60	3920	60	3920	60	3920	61	3920	62	3920	62	3920	64	3920	64	3920
7	15	186	40	2238	53	3920	60	3920	60	3920	60	3920	60	3920	60	3920	60	3920	61	3920	62	3920
8	11	191	33	2236	54	3920	60	3920	60	3920	60	3920	60	3920	61	3920	62	3920	62	3920	62	3920
9	11	176	34	2209	57	3919	60	3920	62	3920	62	3920	62	3920	63	3920	63	3920	63	3920	64	3920
10	9	186	33	2120	52	3920	60	3920	60	3920	61	3920	63	3920	63	3920	65	3920	65	3920	65	3920
Média	**11**	**181**	**34**	**2126**	**54**	**3920**	**60**	**3920**	**60**	**3920**	**60**	**3920**	**61**	**3920**	**62**	**3920**	**62**	**3920**	**62**	**3920**	**63**	**3920**
Desvio-Padrão	**2,3**	**13,1**	**3,3**	**80,1**	**1,9**	**0,3**	**0,4**	**0,0**	**0,6**	**0,0**	**0,7**	**0,0**	**1,0**	**0,0**	**1,3**	**0,0**	**1,5**	**0,0**	**1,6**	**0,0**	**1,4**	**0,0**

Tabela A25. Número de genes VP (VP) e FP (FP) apontdos pelo teste t de Student no conjunto com distorção heterogênea sem transformação, em relação ao tamanho da amostra

	3		5		10		15		20		25		30		35		40		45		50	
n	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**
1	0	0	2	15	21	332	53	1485	63	2257	68	2921	70	3539	70	3823	70	3908	70	3918	70	3920
2	0	2	6	9	28	339	54	1504	63	2270	66	2905	69	3509	69	3818	70	3901	70	3920	70	3920
3	0	0	1	14	29	350	48	1472	57	2240	59	2916	63	3497	65	3810	66	3905	67	3918	68	3920
4	0	0	3	8	27	312	50	1502	62	2280	65	2920	66	3526	68	3812	68	3900	68	3919	70	3920
5	0	0	3	15	26	313	51	1501	58	2251	60	2869	60	3489	61	3819	61	3908	62	3920	63	3920
6	0	0	4	15	27	316	53	1462	59	2258	65	2928	64	3525	66	3823	67	3905	67	3918	69	3920
7	0	1	2	13	32	341	51	1464	60	2270	62	2929	63	3536	66	3834	67	3905	67	3919	67	3920
8	0	1	4	9	32	328	53	1517	62	2237	65	2881	66	3520	67	3819	67	3898	67	3917	67	3920
9	0	1	2	16	28	302	50	1494	61	2225	64	2873	66	3478	69	3804	67	3895	70	3915	69	3920
10	0	2	7	9	26	325	50	1488	58	2278	65	2898	66	3490	67	3806	68	3894	68	3918	68	3920
Média	**0**	**1**	**3**	**12**	**28**	**326**	**51**	**1489**	**60**	**2257**	**64**	**2904**	**65**	**3511**	**67**	**3817**	**67**	**3902**	**68**	**3918**	**68**	**3920**
Desvio-Padrão	**0,0**	**0,8**	**1,8**	**3,0**	**3,0**	**14,3**	**1,8**	**17,3**	**2,1**	**17,5**	**2,6**	**21,5**	**2,8**	**20,3**	**2,4**	**8,6**	**2,4**	**4,8**	**2,2**	**1,4**	**2,0**	**0,0**

Tabela A26. Número de genes VP (VP) e FP (FP) apontdos pelo teste t Bayesiano no conjunto com distorção heterogênea sem transformação, em relação ao tamanho da amostra

	3		5		10		15		20		25		30		35		40		45		50	
n	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**
1	9	51	26	514	49	1834	60	1976	65	2069	68	2308	69	2717	69	3109	70	3465	70	3698	70	3828
2	7	74	21	514	47	1837	60	1975	64	2069	67	2312	68	2712	70	3152	70	3504	71	3723	73	3852
3	4	50	22	479	48	1868	60	1990	59	2086	63	2337	65	2741	67	3145	67	3481	67	3739	69	3849
4	4	42	28	488	47	1840	63	1976	64	2074	66	2320	68	2729	69	3153	70	3511	70	3724	70	3833
5	3	58	20	505	47	1847	53	1989	56	2078	59	2298	62	2719	62	3156	64	3495	64	3718	64	3855
6	2	47	23	512	45	1861	56	1979	64	2077	67	2335	67	2724	68	3164	69	3518	70	3723	70	3840
7	5	48	24	485	45	1847	54	1978	62	2069	66	2345	66	2747	67	3166	67	3537	67	3730	67	3844
8	5	60	24	556	49	1854	56	1985	62	2079	65	2325	65	2749	66	3155	67	3483	69	3730	69	3834
9	4	70	23	540	47	1869	56	1972	60	2057	64	2288	65	2677	69	3123	71	3469	72	3714	72	3829
10	7	52	25	530	47	1863	59	1981	63	2084	65	2336	68	2725	69	3145	69	3494	69	3730	69	3851
Média	**5**	**55**	**24**	**512**	**47**	**1852**	**58**	**1980**	**62**	**2074**	**65**	**2320**	**66**	**2724**	**68**	**3147**	**68**	**3496**	**69**	**3723**	**69**	**3842**
Desvio-Padrão	**2,0**	**9,8**	**2,2**	**23,4**	**1,3**	**12,2**	**3,0**	**5,8**	**2,7**	**8,1**	**2,4**	**17,7**	**2,0**	**19,7**	**2,2**	**17,0**	**2,0**	**21,3**	**2,2**	**10,7**	**2,4**	**9,6**

Tabela A27. Número de genes VP (VP) e FP (FP) apontdos pelo teste t de Student no conjunto com distorção heterogênea após a transformação *Shift*, em relação ao tamanho da amostra

	3		5		10		15		20		25		30		35		40		45		50	
n	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**
1	0	0	4	32	20	254	46	884	59	1585	62	2218	64	2767	66	3340	67	3688	67	3836	68	3905
2	0	0	6	20	30	242	51	860	63	1587	65	2208	68	2770	68	3315	69	3670	72	3838	72	3898
3	0	0	2	30	29	234	52	811	57	1562	59	2179	67	2755	69	3311	69	3657	71	3833	72	3894
4	0	1	4	26	27	228	48	858	60	1627	64	2228	67	2827	69	3343	71	3676	70	3844	71	3902
5	0	3	3	34	30	252	44	867	56	1603	57	2183	59	2723	59	3284	60	3659	63	3843	65	3907
6	0	1	4	37	32	236	49	851	57	1615	61	2207	65	2810	66	3347	68	3708	69	3837	71	3902
7	0	2	3	27	30	260	47	828	53	1581	57	2216	63	2819	66	3349	68	3693	67	3846	69	3908
8	0	1	9	25	32	230	49	885	56	1612	63	2168	65	2737	66	3272	66	3664	69	3848	70	3908
9	0	3	6	36	24	230	50	895	58	1597	64	2184	66	2736	68	3291	71	3644	72	3829	73	3889
10	1	2	6	39	26	258	44	915	57	1658	61	2238	62	2790	67	3329	70	3686	71	3851	72	3901
Média	**0**	**1**	**5**	**31**	**28**	**242**	**48**	**865**	**58**	**1603**	**61**	**2203**	**65**	**2773**	**66**	**3318**	**68**	**3675**	**69**	**3841**	**70**	**3901**
Desvio-Padrão	**0,3**	**1,1**	**2,0**	**5,8**	**3,6**	**11,9**	**2,6**	**29,5**	**2,5**	**25,7**	**2,7**	**22,0**	**2,6**	**35,0**	**2,7**	**26,6**	**3,0**	**18,4**	**2,7**	**6,7**	**2,3**	**5,9**

Tabela A28. Número de genes VP (VP) e FP (FP) apontdos pelo teste t Bayesiano no conjunto com distorção heterogênea após a transformação *Shift*, em relação ao tamanho da amostra

	3		5		10		15		20		25		30		35		40		45		50	
n	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**
1	10	57	23	414	47	1498	59	1895	63	1954	64	1995	64	2048	64	2146	66	2306	67	2555	68	2804
2	6	66	17	400	42	1486	56	1888	65	1963	67	1997	68	2051	69	2119	72	2297	73	2513	73	2815
3	5	68	19	376	48	1436	58	1892	60	1961	62	1990	67	2034	69	2152	69	2344	70	2561	71	2845
4	6	47	24	386	48	1490	60	1895	62	1962	67	2002	68	2083	70	2187	70	2359	70	2603	70	2856
5	4	65	16	394	47	1474	55	1910	57	1981	58	1999	61	2036	60	2135	61	2299	63	2512	65	2830
6	3	54	20	404	43	1462	56	1897	61	1973	62	2008	64	2057	66	2190	69	2390	71	2629	71	2912
7	5	71	22	366	44	1466	53	1890	59	1963	62	1997	64	2056	66	2187	67	2370	67	2671	69	2923
8	5	75	22	424	46	1482	53	1901	59	1964	64	1994	66	2044	67	2134	69	2301	69	2554	69	2834
9	7	83	18	434	42	1473	55	1901	65	1959	67	1992	67	2022	68	2146	71	2290	71	2531	72	2767
10	6	58	21	434	44	1527	57	1903	63	1979	62	2003	63	2069	67	2196	68	2453	69	2666	70	2905
Média	**6**	**64**	**20**	**403**	**45**	**1479**	**56**	**1897**	**61**	**1966**	**64**	**1998**	**65**	**2050**	**67**	**2159**	**68**	**2341**	**69**	**2580**	**70**	**2849**
Desvio-Padrão	**1,8**	**10,1**	**2,5**	**22,3**	**2,3**	**22,8**	**2,2**	**6,3**	**2,5**	**8,4**	**2,8**	**5,2**	**2,2**	**16,8**	**2,8**	**26,6**	**2,9**	**50,2**	**2,6**	**56,4**	**2,1**	**48,1**

Tabela A29. Número de genes VP (VP) e FP (FP) apontdos pelo teste t de Student no conjunto com distorção heterogênea após a transformação *Lowess*, em relação ao tamanho da amostra

	3		5		10		15		20		25		30		35		40		45		50	
n	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**
1	0	1	4	5	21	65	43	147	60	247	70	387	72	566	72	789	72	1063	72	1331	72	1653
2	0	0	2	2	27	48	49	137	60	221	63	385	63	572	66	788	66	1077	66	1357	66	1625
3	0	1	2	4	28	60	42	144	60	234	69	368	69	553	69	776	69	1024	69	1309	69	1603
4	0	1	6	7	26	74	45	128	54	256	67	401	68	583	71	800	71	1050	71	1345	71	1648
5	0	3	5	4	25	61	42	131	56	241	69	406	71	536	71	802	71	1079	72	1341	73	1617
6	0	1	1	4	28	62	42	128	53	243	66	398	67	584	69	829	70	1049	70	1359	70	1629
7	0	0	3	5	28	68	47	130	55	247	66	378	72	556	72	798	72	1062	72	1353	72	1646
8	0	3	5	3	24	71	44	141	62	241	69	408	69	566	70	787	70	1051	70	1356	70	1661
9	0	0	4	3	24	68	44	118	59	243	64	370	65	575	65	793	66	1086	66	1355	66	1654
10	0	0	4	5	27	63	43	123	56	239	66	359	70	560	72	841	72	1101	72	1368	72	1608
Média	**0**	**1**	**4**	**4**	**26**	**64**	**44**	**133**	**58**	**241**	**67**	**386**	**69**	**565**	**70**	**800**	**70**	**1064**	**70**	**1347**	**70**	**1634**
Desvio-Padrão	**0,0**	**1,1**	**1,5**	**1,3**	**2,2**	**6,8**	**2,2**	**8,9**	**2,9**	**8,7**	**2,2**	**16,2**	**2,8**	**13,9**	**2,4**	**18,9**	**2,2**	**21,1**	**2,2**	**16,1**	**2,3**	**19,6**

Tabela A30. Número de genes VP (VP) e FP (FP) apontdos pelo teste t Bayesiano no conjunto com distorção heterogênea após a transformação *Lowess*, em relação ao tamanho da amostra

	3		5		10		15		20		25		30		35		40		45		50	
n	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**
1	4	0	17	8	53	45	60	109	65	246	66	438	66	682	71	955	71	1267	72	1653	72	2008
2	3	0	19	5	43	37	55	104	56	237	59	423	60	653	63	956	65	1287	66	1647	66	1971
3	2	0	18	6	50	33	60	116	63	241	63	414	65	649	68	948	68	1284	69	1612	69	1961
4	6	-1	18	4	50	36	60	118	62	255	66	431	68	663	69	965	71	1286	71	1667	71	2028
5	3	0	15	7	43	35	61	110	64	260	67	441	68	676	72	988	73	1295	73	1645	73	1995
6	2	0	17	8	45	38	55	113	58	250	65	449	67	699	68	998	70	1284	70	1655	70	1939
7	2	0	17	8	46	35	55	88	61	236	64	422	65	663	67	955	70	1310	71	1644	71	1991
8	4	0	12	9	45	39	55	121	59	248	61	447	66	699	68	980	69	1317	69	1675	70	2004
9	2	1	16	8	48	45	57	94	61	227	61	446	63	708	66	1013	66	1325	66	1661	66	1987
10	3	0	14	6	47	32	53	116	61	251	64	421	67	696	72	998	72	1329	72	1642	72	1955
Média	**3**	**0**	**16**	**7**	**47**	**38**	**57**	**109**	**61**	**245**	**64**	**433**	**66**	**679**	**68**	**976**	**70**	**1298**	**70**	**1650**	**70**	**1984**
Desvio-Padrão	**1,2**	**0,4**	**2,0**	**1,5**	**3,1**	**4,2**	**2,7**	**10,2**	**2,6**	**9,4**	**2,5**	**12,0**	**2,3**	**20,1**	**2,7**	**21,6**	**2,4**	**19,5**	**2,3**	**16,3**	**2,3**	**25,8**

Tabela A31. Número de genes VP (VP) e FP (FP) apontdos pelo teste t de Student no conjunto com distorção heterogênea após a transformação *Linlog*, em relação ao tamanho da amostra

	3		5		10		15		20		25		30		35		40		45		50	
n	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**
1	0	0	4	54	37	1287	56	2027	61	2425	64	3158	66	3687	67	3878	67	3917	67	3919	67	3920
2	0	1	3	43	41	1309	57	2028	57	2424	60	3171	60	3691	61	3880	64	3918	65	3920	68	3920
3	0	1	0	43	36	1328	51	2025	53	2424	55	3157	56	3676	58	3883	60	3916	62	3920	64	3920
4	0	3	4	47	36	1294	53	2049	56	2429	61	3152	64	3686	66	3877	65	3916	67	3920	66	3920
5	0	2	4	55	33	1317	50	2042	50	2415	53	3125	55	3668	56	3881	57	3919	59	3920	60	3920
6	0	1	2	50	37	1257	54	2030	57	2421	59	3177	60	3678	62	3880	64	3915	66	3919	65	3920
7	0	1	3	50	37	1275	54	2026	55	2453	59	3181	57	3704	59	3893	60	3917	63	3919	64	3920
8	0	1	5	55	46	1315	54	2029	58	2404	58	3169	59	3688	62	3869	63	3913	64	3920	66	3920
9	0	2	5	65	36	1294	53	2017	54	2405	55	3117	58	3668	61	3876	62	3913	63	3919	63	3920
10	0	2	6	59	37	1290	52	2036	54	2444	58	3138	61	3677	64	3878	65	3917	66	3920	67	3920
Média	**0**	**1**	**4**	**52**	**38**	**1297**	**53**	**2031**	**56**	**2424**	**58**	**3155**	**60**	**3682**	**62**	**3880**	**63**	**3916**	**64**	**3920**	**65**	**3920**
Desvio-Padrão	**0,0**	**0,8**	**1,6**	**6,6**	**3,4**	**20,2**	**2,0**	**8,7**	**2,9**	**14,5**	**3,1**	**20,7**	**3,3**	**10,5**	**3,3**	**5,7**	**2,8**	**1,9**	**2,4**	**0,5**	**2,2**	**0,0**

Tabela A32. Número de genes VP (VP) e FP (FP) apontdos pelo teste t Bayesiano no conjunto com distorção heterogênea após a transformação *Linlog*, em relação ao tamanho da amostra

	3		5		10		15		20		25		30		35		40		45		50	
n	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**	**VP**	**FP**
1	9	57	35	835	55	2049	65	2498	65	3179	67	3611	67	3832	67	3900	67	3918	67	3920	67	3920
2	7	81	26	861	53	2061	61	2514	62	3187	62	3637	63	3847	63	3901	64	3918	65	3920	67	3920
3	5	71	30	791	52	2052	59	2516	58	3191	60	3643	61	3835	62	3906	62	3918	64	3919	64	3919
4	6	48	35	835	54	2064	63	2499	63	3177	64	3621	65	3827	66	3906	68	3917	67	3920	69	3920
5	5	72	29	803	48	2069	54	2520	57	3147	57	3611	57	3854	57	3907	59	3920	60	3920	61	3920
6	4	67	28	823	51	2063	60	2526	63	3153	64	3608	65	3839	66	3902	66	3918	66	3920	67	3920
7	6	65	31	801	48	2059	56	2515	57	3185	61	3633	60	3841	61	3903	61	3916	63	3920	65	3920
8	5	73	34	883	52	2066	59	2536	61	3190	62	3627	63	3827	63	3899	63	3916	64	3920	66	3920
9	6	77	32	842	53	2044	58	2466	60	3140	62	3612	63	3832	64	3902	64	3915	67	3918	68	3920
10	7	51	28	858	50	2064	61	2537	64	3165	64	3629	65	3837	66	3903	67	3917	66	3919	67	3920
Média	**6**	**66**	**31**	**833**	**52**	**2059**	**60**	**2513**	**61**	**3171**	**62**	**3623**	**63**	**3837**	**64**	**3903**	**64**	**3917**	**65**	**3920**	**66**	**3920**
Desvio-Padrão	**1**	**10**	**3**	**27,9**	**2,2**	**7,7**	**3,0**	**19,9**	**2,8**	**17,9**	**2,6**	**11,8**	**2,8**	**8,1**	**2,9**	**2,5**	**2,8**	**1,3**	**2,1**	**0,7**	**2,2**	**0,3**