Validação de procedimentos para medida de expressão gênica a partir de imagens de cDNA *Microarray*

Gustavo Henrique Esteves

São Paulo, Dezembro de 2002

"*Nenhum pássaro voa alto demais, se voa com suas próprias asas*"

William Blake.

Aos meus Pais, meu irmão e à Sibele.

AGRADECIMENTOS

Agradecer às pessoas que, de alguma forma, colaboraram com este trabalho não é tarefa fácil! Relembrar todos os nomes e rostos é difícil, mas sem dúvida nenhuma, todas as pessoas que cruzaram meu caminho colaboraram comigo de alguma forma. Talvez pela simples presença, ou ainda, pela imensa amizade, carinho e apoio que tenham me dado. Gostaria de agradecer a todas essas pessoas, mesmo que seus nomes não estejam aqui...

Entretanto, existem pessoas as quais eu não posso deixar de citar, pessoas que ajudaram a construir tudo o que sou hoje. Como o Antônio Carlos (popular Tonão) e o Fabrício Pires (vulgo Pirão), amigos desde os tempos da pré-escola passando pelos churrascos de fim de semana com o restante da turma: Régis, Daniel, Marcos, Toni etc..., e que estavam presentes até no dia de meu "trote" na faculdade. Com o tempo trouxeram consigo suas namoradas, Edineuza e Carina, (hoje já são esposas. – o rapaziada séria!). Com eles pude dividir boa parte da minha vida, se não a maior parte dela, e até hoje eles ainda têm o carinho de me emprestar um pouco de seu tempo e me dar apoio. Valeu!

Depois disso, veio a faculdade, juntamente com uma turma batalhadora, que "ralava" muito para superar as provas, trabalhos e dificuldades do tão temido curso de matemática. Deixo aqui meus agradecimentos aos companheiros Weber, Oscar, Sônia, Juninho, Patricia, Rita, Mônica, Fausto, entre vários outros que tiveram, junto comigo, o gostinho da vitória com a tão esperada formatura. Também, não posso deixar de citar a Dra. Eloiza Marino Silva, pela paciência e competência em me conduzir em um trabalho de iniciação científica, e que sem dúvida me ajudou a decidir sobre os passos mais importantes que eu daria na minha vida. Depois da faculdade veio o desemprego, e com ele mais duas pessoas fantásticas: as Dras. Paula Rahal e Eloiza Tajara, as biológas que estavam precisando de um matemático. Sou profundamente grato pela confiança depositada em mim e no meu trabalho. A Paula teve toda a paciência para me fazer relembrar o que era DNA, RNA e explicar cuidadosamente o que era um projeto genoma. Nessa época conheci todo o pessoal do laboratório de genética do IBILCE de Rio Preto: Fabrício, Janaína, Míriam, Nelson, Flávia, Alessandra, Sylvia, etc... Obrigado pelo convívio de todos vocês.

E então chegou a hora de pensar no Mestrado, e foi aí que eu caí no laboratório do Dr. Luiz Fernando Lima Reis em São Paulo. Esse é bioquímico, e neste momento eu já estava em um mundo em que vetores e *arrays* tinham dois significados completamente diferentes e eu não sabia mais se realmente era graduado em matemática. Mas, deixando os delírios a parte, o Luiz é uma pessoa admirável, sempre com os pés no presente e o olhar no futuro. Ele sempre me fez pensar de uma forma mais ampla. Obrigado Luiz, por acreditar no meu potencial e pelas dicas sempre bem colocadas. Mas felizmente o Luiz conseguiu uma colaboração com o grupo do Bioinfo do Instituto de Matemática e Estatística da USP e o Dr. Junior Barrera ficou como co-orientador do meu trabalho, e com ele voltei para a área de exatas, ao Junior também devo meus sinceros agradecimentos pelo paciente trablalho de leitura e correção das versões iniciais deste trabalho. Passei a utilizar *matlab* e a discutir os problemas de análise de imagens de *microarray*. No IME também conheci o Daniel Dantas, que me deu muitos toques sobre o *matlab*. Também conheci o Dr. Eduardo Jordão Neves, o Roberto Hirata e o Elier, a quem sou grato pelas enormes discussões sobre o meu projeto, e que sem dúvida me ajudaram a chegar até aqui.

Neste momento não posso deixar de agradecer a minha tia Maria de Lourdes e aos meus primos Fernando, Fabiana e Frederico, pelo carinho com que me receberam em casa, me tratanto como um membro da família. E, com certeza, eles foram minha segunda família. Serei eternamente grato por isso. Também quero registrar aqui meus agradecimentos ao Víctor, Thamíris, Tamara, Yéssica, Amanda, Armando e Khristian, por me fazerem mais feliz em muitos momentos no Jd. Grimaldi.

Com a vida e o trabalho em São Paulo também surgiram novos amigos. Estou me referindo a turma do LaBRI: Dr. Alex Fiorini de Carvalho (que também é meu co-orientador e a quem devo a qualidade dos dados obtidos nesse trabalho e a paciência em me ajudar nos experimentos de *microarray*), Adriana, Beatriz, Luciana, Chamberlein, Waleska, Abrantes, Suzana, Regina, Lara. São pessoas maravilhosas que me ajudaram a fazer PCR, seqüenciamento, e várias outras coisas que fazem parte do dia-a-dia de um laboratório de biologia molecular, se não fossem vocês... Sem me esquecer da Elaine e do Franco, que já não estão mais no laboratório, mas que nem por isso deixaram de ter participado dele. Todas essas pessoas me fizeram passar momentos muito especiais e divertidos nos últimos dois anos.

Eu não poderia deixar de agradecer a Sibele, pessoa muito especial por quem

aprendi a ter muito carinho, admiração e respeito, uma vez que isso é tudo o que ela sabe fazer pelas pessoas que a cercam. Sem deixar de dizer que ela sempre soube dividir as guloseimas mineiras feitas pela dona Marlene (e que guloseimas!) com todos os amigos do Labri naqueles cafés da tarde recheados de calor humano. Enfim, muito obrigado Sibele, pelo imenso carinho e atenção que tem dispendido desde que cheguei a São Paulo!

Nesse momento, não posso deixar de agradecer ao Instituto Ludwig de Pesquisa sobre o Câncer por ter me dado a oportunidade de trabalhar em um local tão conceituado e respeitado. Quero externar meus agradecimentos, também, à todos os funcionários deste instituto, que trabalham duro para que nós possamos trabalhar de maneira tranquila. Também não posso deixar de citar a Ana Maria e a Márcia, as regentes da Pós Graduação, que estão sempre prontas a nos atender e a Sueli pela atenção na revisão bibliográfica deste trabalho, muito obrigado.

Os meus agradecimentos a todos os meus avós, tios, primos e amigos que não foram citados aqui, mas que sem dúvida nenhuma foram peças importantíssimas na minha vida.

E finalmente os meus mais profundos agradecimentos ao meu irmão Ricardo e a duas pessoas mais do que especiais que constituem o alicerce de tudo o que eu sei e sou até hoje. Pessoas as quais Deus me deu a honra e o prazer de chamá-los de meus pais, seu Humberto e dona Maria Elena. Saibam vocês três que essa vitória também é vossa! Muito obrigado por tudo!!

Enfim, obrigado a CAPES pelo suporte financeiro.

LISTA DE FIGURAS

LISTA DE TABELAS

ÍNDICE

RESUMO

Esteves GH. Validação de procedimentos para medida de expressão gênica a partir de imagens de cDNA *Microarray*. São Paulo; 2002. [Dissertação de Mestrado - Fundação Antônio Prudente]

Introdução: A tecnologia de cDNA *microarray* tem sido intensamente utilizada na busca de diferenças em níveis de expressão gênica entre dois estados biológicos distintos. Essa metodologia consiste na deposição de milhares de fragmentos de cDNAs em lâminas de vidro. A área ocupada por cada cDNA é geralmente circular e recebe o nome de *spot*. Essas lâminas são hibridizadas com cDNAs marcados com nucleotídeos fluorescentes, e após excitadas por raio *laser*, as mesmas são digitalizadas, produzindo uma imagem de duas bandas, uma para cada corante. A imagem produzida é analisada computacionalmente com o objetivo de localizar cada *spot* da lâmina e quantificar os respectivos valores de intensidade. A comparação entre as intensidades de sinal em um dado *spot* deve refletir as diferenças na abundância do mRNA correspondente nas duas amostras estudadas. **Material e Métodos**: Neste trabalho investigamos a influência de vários parâmetros que influenciam na qualidade dos dados obtidos. Para fazer essa avaliação foram projetados experimentos bem controlados, onde fragmentos diferentes de cDNA com tamanhos variáveis foram fixados em posições específicas das lâminas, que foram hibridizadas com mRNAs sintéticos correspondentes a cada um desses fragmentos. Após a hibridização, as imagens foram adquiridas por *scanner laser* e quantificadas por diferentes metodologias de análise de imagens. Os valores de expressão obtidos pelas várias técnicas de localização e estimação de intensidade dos *spots* foram comparados com os resultados previstos nos experimentos controlados. **Resultados**: Foi observado que o tamanho dos fragmentos fixados interferem nos valores de intensidade observados. Além disso, a metodologia de análise de imagens baseada em segmentação por variação de intensidade se mostrou mais robusta e confiável além de ser totalmente automatizada e não necessitar de correção de *spots* mal localizados. Também foi mostrado a existência de uma dependência entre os valores de intensidade de sinal obtidos e os valores de razão esperados.

Descritores: 1. ANÁLISE DE EXPRESSÃO GÊNICA. 2. cDNA MICROARRAY/análise de imagens.

ABSTRACT

Esteves GH. Validação de procedimentos para medida de expressão gênica a partir de imagens de cDNA *Microarray* [Validation of procedures for measuring gene expression from images of cDNA Microarray]. São Paulo; 2002. [Dissertação de mestrado - Fundação Antônio Prudente]

Introduction: The cDNA microarray technology has been intensevely used in the search of differences in gene expression levels between two distinct biological states. This technology consists of the deposition of thousands of cDNA fragments in glass slides. The area where each cDNA is immobilized is generally circular and is called spot. These slides are hybridized with cDNAs labeled with fluorescent nucleotides, and after excitation by a laser bean, they are digitalized, producing one image of two bands, one for each dye. The generated image is computationally analyzed to locate each spot on the slide and quantify the respective values of intensity. The comparison between the signal intensities for a given spot should reflect the differences in the abundance of the correspondent mRNA in the two samples under study. **Material and Methods**: In this work, we investigated the influence of various parameters that may affect the quality of the data obtained. To make this evaluation we projected well controlled experiments, where different fragments of cDNA with variable sizes were fixed in specific positions of glass slides, which were hybridized with synthetic mRNAs corresponding to each one of these fragments. After hybridization, images were acquired by laser scanner and quantified by different metodologies of image analysis. The values of expression obtained by the several techniques of localization and estimation of intensities of the spots were compared with the predicted results in the controlled experiments. **Results**: We observed that the lenght of the immobilized fragments may affect the values of intensities observed. Moreover, the procedure of image analysis based on segmentation by intensity variation was the most powerfull and reliable method. Besides, it is totaly automated and doesn't require the correction of misidentified spots. Also, it was shown that there is a dependence between the signal intensities obtained and the expected ratio values.

Descriptors: 1. ANALYSIS OF GENE EXPRESSION. 2. cDNA MICROARRAY/image analysis.

1 INTRODUÇÃO

Todas as informações genéticas de um ser vivo são guardadas em uma biomolécula conhecida como ácido desoxirribonucléico (DNA) que se encontra no núcleo de todas as células de organismos eucariotos ou no citoplasma das células de organismos procariotos. De forma semelhante, o DNA abriga as informações necessárias para a manutenção da vida deste organismo. Esse processo é realizado a nível molecular onde pequenas porções do DNA, conhecidas como genes, são transcritas em moléculas de ácido ribonucléico (RNA) que é traduzido em proteínas. Essas moléculas protéicas promovem a manutenção de um ambiente favorável à sustentação da vida da célula e, conseqüentemente, de todo o organismo. Além disso, as proteínas também podem interagir com o DNA regulando a expressão de genes.

Da mesma forma que a expressão de genes encontrados nas moléculas de DNA é responsável pela manutenção dos processos vitais do organismo, genes expressos em momentos indesejados podem levar ao surgimento de diversas patologias. Assim, um dos principais desafios da Biologia Molecular é tentar entender esses perfis de expressão gênicos com o objetivo de caracterizar melhor o funcionamento do organismo como um todo.

1.1 OS PROJETOS GENOMA

A descoberta da estrutura do DNA em 1953 [49] abriu possibilidades para se desvendar o funcionamento dessa biomólecula, que desempenha papéis importantes para os processos vitais de todo organismo. Com isso foi possível entender como a herança genética é transmitida de geração para geração. Uma característica muito importante dessa molécula é a sua estrutura de dupla fita, que são complementares entre si e ligadas por pontes de hidrogênio. Essas novas informações tornaram possível o entendimento da transcrição do DNA em moléculas de RNA mensageiro (mRNA) fita simples que é, por sua vez, traduzido em proteínas.

A descoberta de enzimas de restrição juntamente com as enzimas de transcrição reversa (RT - *Reverse Transcriptase*) também tiveram grande importância para a Biologia Molecular. As enzimas de restrição são capazes de "cortar" a dupla fita de DNA em lugares específicos, o que possibilitou a clonagem de fragmentos genômicos

de DNA em bactérias, como a *E. Coli*. As RTs, são capazes de copiar uma fita complementar a partir de um mRNA molde, gerando assim um DNA complementar, ou cDNA, que também pode ser clonado em bactérias.

Todas essas descobertas levaram a biologia a desenvolver novas técnicas, cada vez mais avançadas, para manipular seqüências de ácidos nucléicos. Uma conseqüência marcante destes avanços foi a automação dos métodos de seqüenciamento de DNA que, juntamente com o advento da bioinformática, permitiram a geração de uma enorme quantidade de informação e dados, tornando viável a realização dos vários projetos genoma espalhados por todo o mundo. O principal objetivo desses projetos é a descoberta do código de toda seqüência de DNA de determinado organismo. Diversas bactérias e outros organismos já tiveram seu genoma completamente seqüenciado como *Mycoplasma genitalium*, *E. coli*, *S. cerevisiae*, *C. elegans* [11, 35]. As primeiras discussões sobre o seqüenciamento completo do genoma humano remonta da década 80, e o primeiro "rascunho" da seqüência de nossas $3x10^9$ bases foi publicado no início do ano de 2001 [43, 47]. Todos esses trabalhos resultam em diversos bancos de dados disponíveis em vários *sites* da *internet*, além de outros contendo ferramentas para busca de informações. A Tabela 1 lista alguns *sites* contendo os bancos mais populares e a ferramenta BLAST, muito utilizada para busca de informações em tais bancos.

Tabela 1: *Sites* encontrados na *internet*.
Tabela mostrando alguns *sites* da *internet*, onde podem ser encontrados dados sobre vários organismos, além de ferramentas para busca de informações.

Nome	***Site***
Genbank	*http://www.ncbi.nlm.nih.gov/Genbank/GenbankSearch.html*
SwissProt	*http://www.expasy.ch/sprot/*
Microbial Genomes	*http://www.tigr.org/tdb/mdb/mdbcomplete.html*
Human Genome Project	*http://www.sanger.ac.uk/HGP*
Blast	*http://www.ncbi.nlm.nih.gov/BLAST*

O Brasil também teve uma importante participação na busca do conhecimento da seqüência de DNA do homem com o Projeto Genoma Humano do Câncer, financiado pela FAPESP juntamente com o Instituto Ludwig de Pesquisa sobre o Câncer,

que identificou mais de um milhão de seqüências expressas em diferentes tumores como mama, pulmão, estômago, cabeça e pescoço, etc., que são alguns dos tumores que mais acometem a população brasileira. Estas seqüências, denominadas ORESTES (*Open Reading Frame Expressed Sequence Tags*) foram geradas utilizando-se uma metodologia inovadora que favorece a obtenção de dados referentes à porção central de um dado cDNA [31]. A conseqüência maior da utilização da técnica ORESTES foi a identificação de cerca de 300.000 seqüências que até então não estavam presentes em nenhum banco de dados visto que a grande maioria das seqüências obtidas por outras metodologias estão localizadas preferencialmente nas extremidades 3' e 5' das moléculas de cDNA.

Todos esses bancos de dados constituem uma enorme fonte de informações que certamente irão acelerar as pesquisas em vários ramos da Biologia Molecular, onde podemos destacar a caracterização de genes ainda não conhecidos, a busca de polimorfismos de um único nucleotídeo (SNPs), as localizações cromossômicas de todos os genes seqüenciados, etc. A maioria desses trabalhos tem sido feitos através de ferramentas computacionais, o que impulsionou o surgimento de um novo ramo da ciência, a Bioinformática ou Biologia Molecular Computacional.

Os projetos genoma, certamente alavancaram a identificação de genes até então desconhecidos, o que levou a um aumento significativo do número de genes bem caracterizados, o que favorece a análise comparativa da expressão gênica, uma vez que é sabido que todos os genes estão mergulhados em várias vias metabólicas relacionadas entre si.

1.2 ANÁLISE COMPARATIVA DA EXPRESSÃO GÊNICA

Com raras excessões, todas as células que constituem um certo organismo contém exatamente a mesma carga genética, ou seja, o mesmo DNA. O que diferencia dois grupos celulares morfologicamente diferentes (células de fígado e rim, por exemplo) são os genes expressos nesses dois tipos de células e os níveis de expressão desses genes. De maneira semelhante, podemos entender os processos patológicos que levam o organismo a desenvolver uma determinada doença. Além disso, o conhecimento dos níveis de expressão gênica leva a um melhor entendimento dos diferentes estágios de desenvolvimento de um tecido, a resposta a diferentes estímulos, etc.

Desta forma, vemos que a comparação dos níveis de expressão dos genes de diferentes tecidos podem levar ao entendimento dos diversos fenômenos encontrados em um organismo. Em especial é possível comparar tecidos sadios e doentes e buscar genes que podem ser possíveis alvos para tratamento e desenvolvimento de novas drogas. Hoje existem várias metodologias que podem ser utilizadas para a busca das diferenças entre os níveis de expressão.

A possibilidade de se transferir para filtros de nitrocelulose ou membranas de *nylon* fragmentos de RNA total (isto é, todos os tipos de RNA encontrados na célula) fracionados por eletroforese em géis de agarose que podem ser hibridizados contra cDNAs marcados com substâncias radioativas [41], possibilitou a busca de diferenças em níveis de expressão de genes conhecidos, uma vez que tais hibridizações podiam ser detectadas por radiografia e comparadas em seguida.

Com o passar dos anos esta metodologia serviu como alicerce para o surgimento de outras técnicas cujo objetivo principal era a comparação das populações de mRNAs de diferentes células ou tecidos. A hibridização subtrativa se utiliza da hibridização de mRNAs de uma certa população celular contra cDNAs construídos a partir de mRNAs de células diferentes (mas que preservem algum tipo de relação com o grupo anterior) onde os cDNAs de fita simples são recuperados e analisados, dando uma idéia das diferenças de expressão gênica entre os tipos celulares [25]. O DDRT-PCR (RT-PCR por amostragem diferencial) consiste da construção de cDNAs a partir de um oligonucleotídeo ancorado na cauda poli A dos mRNAs, posteriormente é feita uma reação em cadeia da *polimerase* (PCR) de baixo rigor utilizando-se o mesmo oligo inicial juntamente com outro arbitrário. A posterior comparação do padrão de corrida em um gel de seqüenciamento para duas populações celulares distintas nos dará as diferenças de expressão entre as amostras [27]. O RAP-PCR (RT-PCR por iniciação arbitrária) segue o mesmo princípio do DDRT-PCR, porém utilizando um oligonucleotídeo arbitrário na reação de transcrição reversa [50]. O RDA (Análise por representação diferencial) também segue o mesmo princípio do DDRT-PCR, porém, ele tem a vantagem de não amplificar os fragmentos que são comuns as duas populações estudadas [16]. A técnica SAGE permite a análise em série da expressão gênica com base na identificação de pequenos fragmentos, conhecidos como *tags* contendo, aproximadamente, de 9 a 10 pares de base (pb) [46].

A maioria dessas técnicas, entretanto, não permitem uma boa avaliação quan-

titativa dos níveis de expressão gênica das amostras em questão, além de avaliarem um número muito limitado de genes e serem relativamente demoradas, com exceção da técnica SAGE. Mais recentemente, a possibilidade de utilização de técnicas robotizadas permitiram a avaliação da expressão diferencial, em grande escala, de genes de duas populações celulares distintas, através da análise global de genes por microarranjos de cDNAs [6, 36, 38]. Essa metodologia é conhecida como DNA *microarray*, onde um grande número de cDNAs ou oligonucleotídeos podem ser fixados em pequenas superfícies de vidro ou membranas de *nylon* e os níveis de expressão dos genes correspondentes podem ser analisados em um único experimento, o que permite a análise da expressão funcional de um grande número de genes, mesmo aqueles cujas funções ainda não foram elucidadas.

Todos esses avanços da biologia molecular juntamente com a necessidade de uma técnica mais robusta e eficaz de análise de expressão gênica diferencial favoreceram o desenvolvimento de uma nova metodologia para esse fim. A tecnologia de cDNA *microarray*, também conhecida como *biochip* de DNA, foi inicialmente descrita por Schena *et al.* em 1995 [36] e consiste de lâminas de vidro, ou membranas de *nylon*, onde milhares de fragmentos de cDNA conhecidos são distribuídos em espaços mínimos para posterior hibridização contra cDNA marcado, preparado de alguma população celular de interesse. O cDNA fixado no substrato é conhecido como *cDNA sonda*, uma vez que suas seqüências são conhecidas, ao passo que o cDNA marcado por fluorescência é denominado *cDNA alvo*, de acordo com [40]. O cDNA sonda será freqüentemente chamado de *cDNA fixado*, ao passo que o cDNA alvo será chamado de *cDNA flutuante*. Em 1996 foi publicado um trabalho mais detalhado sobre os *microarrays* em lâminas de vidro [38]. Através da utilização de *microarrays*, é possível determinar a expressão gênica de dezenas de milhares de genes em um único experimento.

Atualmente existem várias revisões na literatura que tratam especifica e detalhadamente sobre esta nova tecnologia [7, 10, 23]. Vários protocolos para a construção de experimentos de cDNA *microarray* já estão bem definidos. Hedge *et al.* [14], por exemplo, descreve um procedimento completo a ser seguido para a realização desses tipos de experimentos, ele também fornece uma lista de vários outros protocolos disponíveis na *internet*. Também é importante mencionar que existem dois tipos de substratos que estão sendo mais freqüentemente utilizados para a fixação dos cDNAs, que são membranas de *nylon* ou lâminas de vidro, cada um com suas vantagens e desvantagens.

Entretanto, devido à maior utilização das lâminas de vidro na atualidade, esse trabalho está completamente voltado para esse tipo de *microarray*.

1.3 A METODOLOGIA DE cDNA *MICROARRAY*

Os experimentos de cDNA *microarray* consistem de uma etapa bioquímica e uma etapa computacional-estatística.

A etapa bioquímica consiste de duas fases que podem ser consideradas independentes, são elas:

- Fixação de cDNAs nas lâminas,
- Extração de RNA e hibridização.

Após a etapa bioquímica do processo, entramos na etapa computacional-estatística, que consiste das seguintes fases:

- Aquisição de imagens,
- Análise de imagens,
- Normalização de dados,
- Análise de dados.

A seguir descreveremos as fases envolvidas no processo de construção de experimentos dessa natureza. Apresentamos uma descrição suscinta de cada fase, que são listados em ordem de 'coisas a fazer'.

1.3.1 A fixação de fragmentos de cDNA nas lâminas de vidro

A fase inicial do trabalho consiste na seleção de clones de cDNA de interesse advindos de algum banco de clones. Esses bancos estão, geralmente, relacionados com algum projeto genoma específico. Os fragmentos selecionados são, então, amplificados por PCR e fixados de maneira automatizada através de um robô (ou *arrayer*) nas lâminas de vidro em posições específicas conhecidas como *spots*. As lâminas de vidro são geralmente revestidas por algum grupamento químico carregado positivamente, o que favorece a ligação do cDNA em sua superfície. Após a fixação, as lâminas podem ser guardadas para posterior utilização.

1.3.2 A extração de RNA e hibridização

A fase seguinte consiste da extração de RNA mensageiro ou total das populações celulares de interesse. A partir desses RNAs é produzido o cDNA alvo por transcrição reversa na presença de dCTP (ou dUTP) que contém grupamentos fluorescentes, *cy3* ou *cy5*, que são excitados e emitem luz em comprimentos de onda diferentes. A seguir os cDNAs alvo de duas amostras distintas são misturados e hibridizados contra as lâminas produzidas. Geralmente as duas amostras estudadas recebem o nome de *teste* e *referência*. A hibridização é feita em câmara úmida ou em estação de hibridização automatizada. Com a captação das imagens geradas pelas fluorescências, é possível compararmos as intensidades de sinal para cada *spot* da lâmina e buscar as diferenças de expressão entre os genes das amostras em estudo.

É neste momento do trabalho que a "mágica" dos *microarrays* acontece, aqui os mRNAs são copiados em fitas de cDNA na presença de nucleotídeos fluorescentes por uma enzima conhecida por RTase (do inglês, *Reverse Transcriptase*). Verifica-se que na síntese dessas novas fitas a proporção existente entre os diferentes mRNAs é mantida, é essa informação que é codificada nas imagens digitais na forma de intensidade dos *spots*.

Para tanto, é preciso que três coisas aconteçam. Primeiramente, a população de RNAs deve ser corretamente extraída e purificada de suas amostras de interesse. Segundo, os mRNAs presentes nesse grupo devem ser copiados em novas fitas com a efetiva incorporação dos nucleotídeos marcados. E, finalmente, esses cDNAs devem hibridizar contra suas respectivas seqüências nas lâminas. Existem vários trabalhos na literatura buscando aumentar os níveis de integridade de RNA, bem como a verificação da eficiência de incorporação de compostos fluorescentes em ácidos nucléicos visando a obtenção de resultados mais robustos e o uso mais eficiente dos experimentos de cDNA *microarray*, [21, 24, 30, 33, 51].

1.3.3 A aquisição de imagens

Uma vez que as lâminas foram hibridizadas, os sinais de intensidade devem ser adquiridos computacionalmente, para posterior quantificação e análise de dados. Esses sinais são geralmente armazenados como imagens de 16-*bit*, o que é feito por *scanners* apropriados para a digitalização desses dados. Atualmente existem dois tipos

de equipamentos diferentes para essa finalidade: os *scanners* CCD (*Charge Coupled Device*) e *laser*. Na tecnologia CCD as lâminas são excitadas com uma luz branca em toda a sua extensão e uma câmera fotografa a imagem decorrente da emissão de intensidade proveniente dos fluorocromos presentes nos alvos que foram utilizados para hibridização. Os *scanners* a *laser* fazem uma varredura na lâmina com um raio *laser* nos comprimentos de onda específicos digitalizando a imagem gerada. Vários equipamentos se encontram disponíveis no mercado, a Tabela 2 mostra uma listagem de alguns fabricantes desses equipamentos. Apesar disso, trabalhos recentes tem mostrado a busca de equipamentos ainda mais robustos para a captação de imagens de *microarray* [13].

Tabela 2: Alguns fabricantes de *scanners*.
Tabela mostrando alguns fabricantes de *scanners* utilizados para a captação das imagens geradas em experimentos de *microarray*.

Fabricante	**Site**
Axon	*http://www.axon.com*
GSI Lumonics	*http://www.gsilumonics.com*
Genomic Solutions	*http://www.genomicsolutions.com*
Packard BioSciences	*http://lifesciences.perkinelmer.com*
Molecular Dynamics	*http://www.mdyn.com*

Como já foi citado anteriormente, os fluorocromos *cy3* e *cy5* são excitados e emitem luz em comprimentos de onda diferentes, a Tabela 3 descreve as diferenças de excitação e emissão dos dois fluoróforos, lembrando que existem outros corantes que também podem ser utilizados para a mesma finalidade. Em geral os *scanners* fazem uma primeira leitura da lâmina utilizando o comprimento de onda específico para *cy3*, capturando uma imagem. Logo em seguida é feita uma segunda leitura utilizando o comprimento de onda específico para *cy5*, obtendo uma segunda imagem. É necessário enfatizar que cada lâmina de *microarray* gera duas imagens diferentes, uma para cada fluorocromo utilizado.

Geralmente uma pseudocoloração seguida de uma composição das duas imagens é utilizada para facilitar a visualização e verificação da qualidade do experimento em questão. O padrão de coloração mais freqüentemente utilizado também está descrito

na Tabela 3. A Figura 1 mostra um esquema de uma lâmina juntamente com uma imagem composta de um experimento de cDNA *microarray*. Na figura (A) temos um esquema de uma imagem de *microarray*, onde cada grupo de *spots*, como o delimitado pelo retângulo vermelho, constitui um bloco ou *subarray* [15], e na figura (B) temos uma imagem de um experimento de *microarray* em que os pontos verdes correspondem a genes mais expressos na amostra marcada com *cy3*, os vermelhos são genes mais expressos em *cy5*, os demais variam dentro de uma gradação de amarelo e laranja e são expressos nas duas amostras.

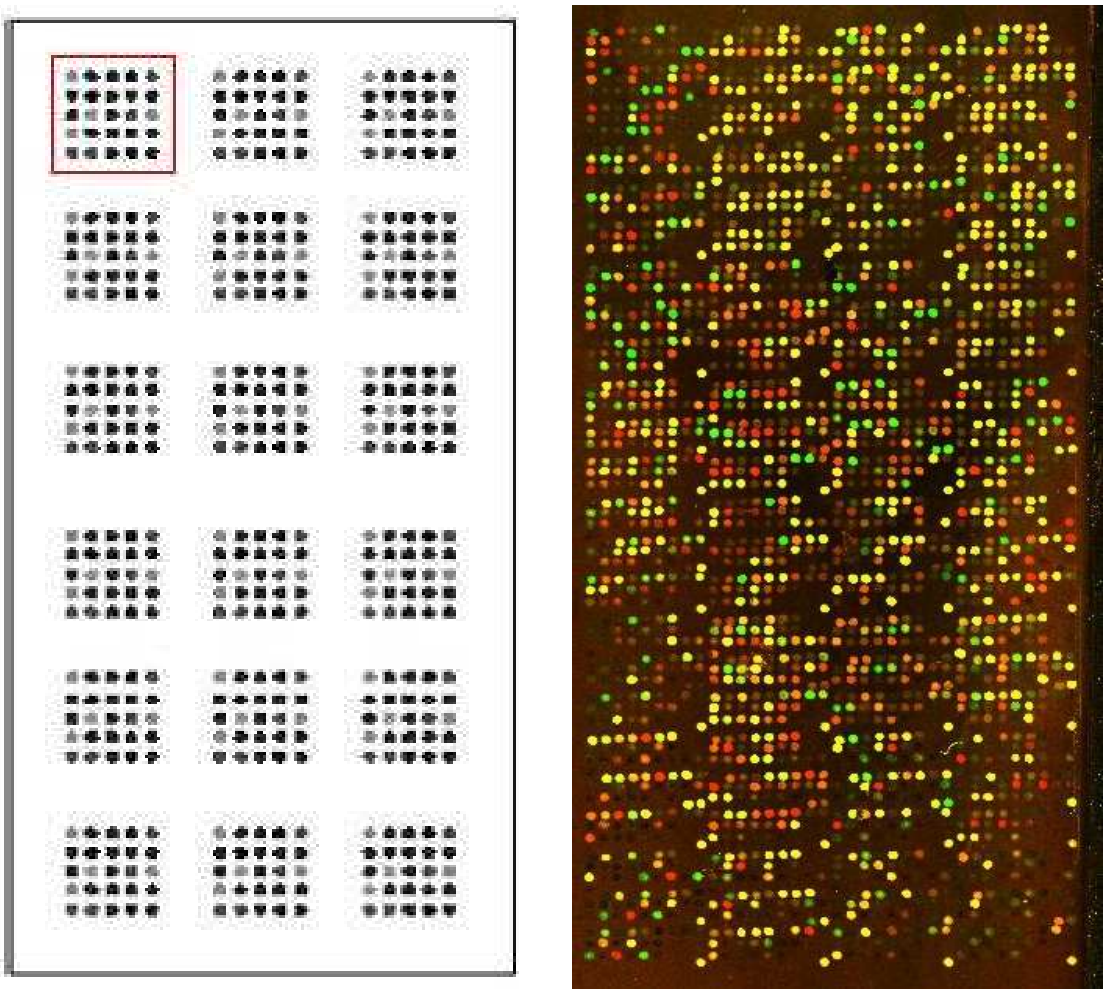
Tabela 3: As diferenças entre *cy3* e *cy5*.
Tabela mostrando as principais características entre *cy3* e *cy5*, os dois corantes utilizados com mais freqüência em experimentos de cDNA *microarray*.

Fluoróforo	Excitação (nm)	Emissão (nm)	Coloração
cy3	550	570	verde
cy5	649	670	vermelho

1.3.4 A análise de imagens

As imagens obtidas anteriormente constituem os dados 'crus' da análise de expressão gênica por experimentos de *microarray*. O trabalho seguinte consiste na quantificação dos sinais emitidos pelos diferentes *spots* que representam diferentes genes. O produto final desta etapa do trabalho é uma tabela de dados numéricos que deve ser posteriormente analisada por ferramentas computacionais-estatísticas.

Antes de detalhar um pouco mais o processo de análise de imagens decorrentes de experimentos de cDNA *microarray*, é necessário definir alguns termos que são freqüentemente usados na linguagem de processamento de imagens digitais. A região ocupada pelo *spot* é conhecida como *região de sinal* ou *foreground*. Os valores de intensidade medidos nessa região são decorrentes da emissão de fluorescência proveniente de moléculas de cDNA marcadas com *cy3* e/ou *cy5* que se anelam nas moléculas complementares fixadas na lâmina. Entretanto, algumas dessas móleculas fixadas não se anelam com nenhuma molécula fluorescente. A falta de contribuição dessas moléculas para o valor de sinal observado no *spot* é chamada de *ruído*. Em contrapartida, a



(A) (B)

Figura 1: A imagem de um experimento de *microarray*.

(A) Esquema geral de uma imagem de *microarray* [15], (B) Imagem de um experimento real de *microarray*.

imagem de fundo da lâmina (regiões onde não se encontram *spots*) é chamada de *background*. Eventualmente, podem existir sinais de intensidade mais intensos decorrentes de sujeira na lâmina ou hibridização inespecífica que contaminam o *background*. Esse tipo de sinal inespecífico é conhecido como *artefato*. A Figura 2 ilustra esses conceitos.

De acordo com Yang *et al.* [53] o processamento de imagens de *microarray* pode ser divido em três etapas, que são:

- Endereçamento ou gradeamento: processo de indexação de todos os *spots*.
- Segmentação do sinal (ou dos *spots*): é o que define os *pixels* que fazem parte das regiões de sinal (*foreground*) ou do *background* – região onde não se encontra DNA fixado.
- Quantificação da intensidade: etapa que estima, para cada *spot*, os valores de intensidade, *background* e, possivelmente, de qualidade.

Em geral, as bandas das imagens adquiridas pelo *scanner* são bem registradas, não exigindo cuidados adicionais de correção. Por outro lado, por vezes, as imagens adquiridas apresentam rotação, que deve ser corrigida, pois dificulta as etapas subseqüentes de processamento. Em geral, as bandas são compostas (e.g., por média ou máximo) em uma única imagem sobre a qual é efetuada a segmentação. O resultado da segmentação é usado como máscara nas duas bandas originais. A segmentação é o processo de particionar uma imagem em diferentes regiões, cada uma contendo certas características, o que no caso das imagens de *microarray* consiste em localizar cada um dos blocos que constituem a imagem. A seguir cada um desses blocos é também segmentado com o objetivo de localizar uma região ao redor de cada *spot*, essa região é, geralmente, retangular e é denominada *região de influência do spot*.

O processo de endereçamento consiste da indexação de cada bloco que constitui a imagem analisada, assim como de todos os *spots* para que eles sejam corretamente identificados de acordo com o arranjo definido no projeto do *biochip*. Os processos de estimação e segmentação dos *spots* são mais complexos e serão comentados em maiores detalhes a seguir.

A segmentação dos *spots* de uma imagem de *microarray* consiste em classificar cada *pixel* da imagem como pertencente ao *foreground* ou ao *background*. O resultado

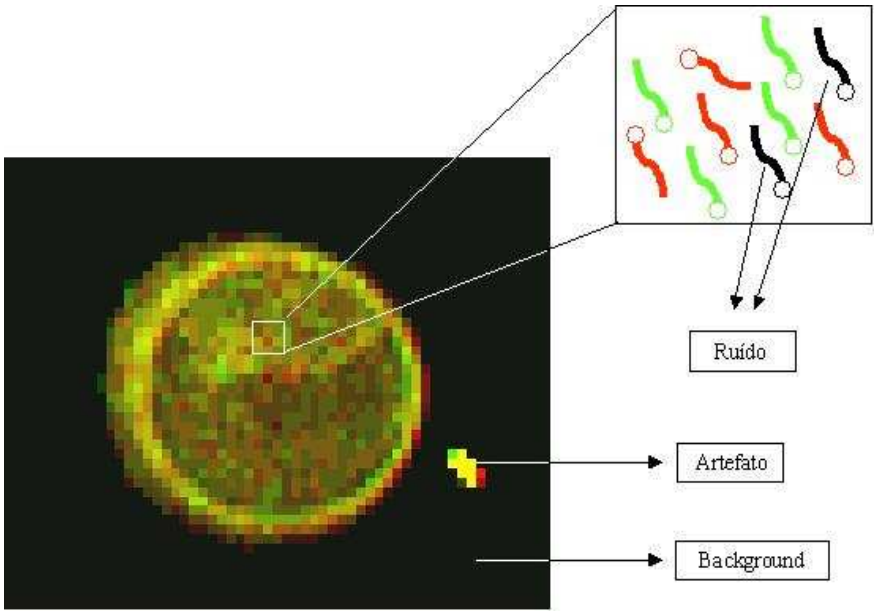


Figura 2: Termos utilizados na análise de imagens de *microarray*.
Esta figura ilustra os principais termos utilizados em análise de imagens de *microarray*. A imagem principal ilustra um *spot*. A região escura ao redor desse *spot* é conhecida como *background*, regiões de sinal de intensidade espúria dentro do *background* são conhecidas como *artefatos*. O esquema ilustrado no canto superior direito ilustra um *pixel* interno ao *spot*, onde são encontradas moléculas de cDNA marcadas com *cy3* (verdes), *cy5* (vermelhas) e moléculas fixadas na lâmina que não hibridizaram com nenhuma molécula marcada (pretas). Essas moléculas sem fluorescência constituem as regiões de *ruído*.

final dessa segmentação é uma 'máscara' que localiza todos os *spots* da lâmina. Os métodos de segmentação de sinal mais freqüentemente utilizados podem ser divididos em 4 grupos [53], que são:

- Segmentação de círculo fixo: assume que todos os *spots* de uma mesma lâmina tem o mesmo tamanho e forma circular, admite um valor de diâmetro fixo para todos os *spots* da lâmina. Esta metodologia tem a desvantagem de poder perder muitos *pixels* do *foreground* e incorporar muitos *pixels* do *background*. O erro está associado a variações de forma e tamanho do *spot*. O *ScanAlyze* é um exemplo de *software* que se utiliza deste método de segmentação [8].

- Segmentação de círculo adaptativo: o diâmetro de cada *spot* é estimado separadamente. Esta metodologia também pode perder *pixels* do *foreground* e incorporar *pixels* do *background*. O erro tende a ser menor do que o de segmentação de círculo fixo, porque está associado apenas com variações de forma do *spot*. O *QuantArray* é um exemplo de utilização desta metodologia [32].

- Segmentação por histograma: separa o *foreground* do *background* baseado no histograma de intensidade da região de influência do *spot*. Esta é uma técnica de segmentação clássica que não restringe a alguma forma específica, porém ela tem um sério problema: incorpora *pixels* claros do *background* e perde *pixels* escuros do *foreground*, devido à sua exclusiva dependência da intensidade dos *pixels*. O *software QuantArray* [32], também pode utilizar uma implementação deste tipo.

- Segmentação por variação de intensidade: separa o *foreground* do *background* baseado no gradiente da imagem [12]. Algumas variantes mais robustas desta técnica usam também o conhecimento da região de influência do *spot* [1]. Essa técnica também não se restringe a formas específicas e é robusta a flutuações de intensidade dos *pixels* do *foreground* e *background*, pois depende de propriedades topológicas, além da variação de intensidade. Hirata Jr. *et al.* utiliza-se de ferramentas similares com aplicação específica para experimentos de *microarray* [15].

Cada *spot* de uma lâmina de *microarray* representa o valor de atividade real de algum gene em alguma característica biológica de interesse. Esses valores de atividade

gênica são codificados em forma de sinais de intensidade observados para cada *spot* da lâmina. A extração correta desses valores de intensidade é que possibilita a comparação entre diferentes estados biológicos. Entretanto, esses valores de intensidade estão sujeitos as variabilidades naturais de qualquer sistema biológico, além de fontes de incerteza intrínsecas à estratégia experimental. Porém, a estimação desses valores de intensidade não é uma tarefa muito simples, uma vez que a distribuição desses valores não é conhecida, sendo que existem diferentes técnicas de estimação sendo empregadas pelos *softwares* de análise de imagens.

Outro valor que deve ser extraído na etapa de análise de imagens é a intensidade do *background*, necessário para a correção dos valores de intensidade obtidos. A motivação para a utilização deste dado está no fato de que não é possível obter exatamente os valores de ruído encontrados nos *spots*, entretanto o efeito desse ruído pode ser minimizado através da correção dos valores de intensidade da região de sinal pelos valores de intensidade do *background*. Além disso, os valores de intensidade dos *spots* estão sujeitos a uma contribuição de outros fatores independentes da hibridização do cDNA alvo marcado contra a sonda fixada na lâmina. Esses fatores podem ser devidos a hibridização inespecífica ou ainda a componentes químicos presentes na superfície da lâmina. Existem duas técnicas distintas que são usualmente utilizadas com essa finalidade. Uma é o cálculo de um valor de *background* constante para todos os *spots*. Essa metodologia assume que a intensidade em regiões sem DNA fixado é uniforme em todo o substrato e é mais empregada em experimentos que utilizam membranas de *nylon*. A outra opção é o cálculo de *background* localmente, ou seja ao redor da máscara que define cada *spot*. Essa metodologia permite correções em experimentos cuja intensidade em regiões inespecíficas são mais variáveis, o que acontece com mais freqüência em experimentos que se utilizam de lâminas de vidro.

Muitos *softwares* também fornecem dados estatísticos de qualidade dos valores de intensidade obtidos, tais como, variabilidade dos valores de intensidade de *pixels* para cada *spot*, circularidade dos *spots*, etc.

1.3.5 A normalização dos dados

É extremamente importante que todos os dados a serem analisados sejam normalizados, isso deve ser feito para minimizar a variação introduzida por algumas fontes de variabilidade comumente encontradas em experimentos de cDNA *microarray*. Dentre essas fontes de imprecisão podemos destacar as diferenças na eficiência de incorporação dos diferentes corantes utilizados na marcação das amostras de interesse e erros devidos à imprecisão de equipamentos utilizados durante o manuseio das amostras estudadas. Alguns procedimentos de normalização dos efeitos dos fluoróforos em dados de *microarray* tem sido utilizados, onde podemos destacar os métodos de intensidades totais (ou energia total), as técnicas de regressão [34] e a autonormalização (ou *selfnormalization*) [54].

A normalização por energia total, aplica-se a experimentos onde os genes fixados na lâmina são escolhidos ao acaso, cuja maioria deles não apresentem diferenças de expressão entre as duas amostras estudadas ao passo que alguns genes tenham um aumento de expressão em uma das amostras e outros ganhem expressão na outra amostra. Dessa maneira a quantidade total de cDNA alvo hibridizado deve ser igual nos dois canais, conseqüentemente os valores de intensidade obtidos para todos os *spots* em *cy3* e *cy5* também deve ser aproximadamente igual. Isso permite o cálculo de um fator de correção capaz de corrigir os fatores citados no parágrafo anterior. Uma outra forma de normalização se baseia na dispersão dos valores de intensidade, onde a maioria dos *spots* devem se concentrar ao redor de uma reta com inclinação aproximada igual a 45^{o}, nas técnicas de regressão a inclinação dessa reta é corrigida usando métodos de regressão ajustando os valores de intensidade para atingir a inclinação esperada. Na autonormalização os experimentos são feitos em duplicatas com a inversão dos fluoróforos ou *swap*, onde a variação se anula quando é assumido que tal variação é aproximadamente a mesma nos dois experimentos.

Entretanto, as ferramentas de normalização ainda não estão padronizadas e bem definidas, cabendo ao pesquisador decidir qual é a metodologia que melhor se adequa às suas necessidades dentre as várias opções existentes atualmente [37, 54, 55].

1.3.6 A análise de dados

Os experimentos de *microarray* já são uma realidade para a Biologia Molecular atual, alguns bancos de dados exclusivamente referentes a essa nova metodologia de análise de expressão gênica já estão sendo descritos e podem ser acessados através da rede mundial de computadores, como o banco de dados de *microarray* da Universidade de Stanford [39] e o *GeneX* desenvolvido por um grupo de pesquisadores do NCGR (*National Center of Genome Resources*) [29], o que tem mostrado a força com que esses tipos de experimentos vem crescendo.

Os dados de *microarray* contém valores referentes aos níveis de atividade reais de milhares de genes em uma única lâmina tendo, em geral, um número bastante pequeno de amostras biológicas. Isso faz com que esses tipos de experimentos constituam um sistema dinâmico complexo sendo considerados como problemas computacionalmente intratáveis. Isso significa que a análise desses dados é uma tarefa muito complicada [5].

De uma forma geral, existem duas categorias de análise de dados: os algorítmos de aprendizagem não-supervisionada e supervisionada. Na análise não-supervisionada não se tem nenhum tipo de conhecimento prévio sobre os seus dados assim como não são obtidas medidas para o erro final cometido [5]. Alguns exemplos de ferramentas que se utilizam dessa metodologia são a análise de componentes principais (PCA) e os métodos de *clusterização*, como o *self-organizing maps* (SOM). A análise supervisionada caracteriza-se por fornecer medidas de erro e pela utilização de informações obtidas *a priori* de algum grupo de genes para guiar suas análises. Uma ferramenta bastante conhecida que emprega essa técnica de análise de dados é o *support vector machine* (SVM) [5, 34]. Um problema que acontece com ferramentas de análise supervisonada é que elas obtém melhores resultados quando temos um número relativamente pequeno de variáveis (genes, por exemplo) e muitos exemplos (pacientes, por exemplo), que é exatamente o contrário do que acontece em experimentos de *microarray*.

A Figura 3 apresenta um resumo geral de todo o procedimento de construção de experimentos de cDNA *microarray*, ilustrando as diferentes etapas envolvidas no trabalho.

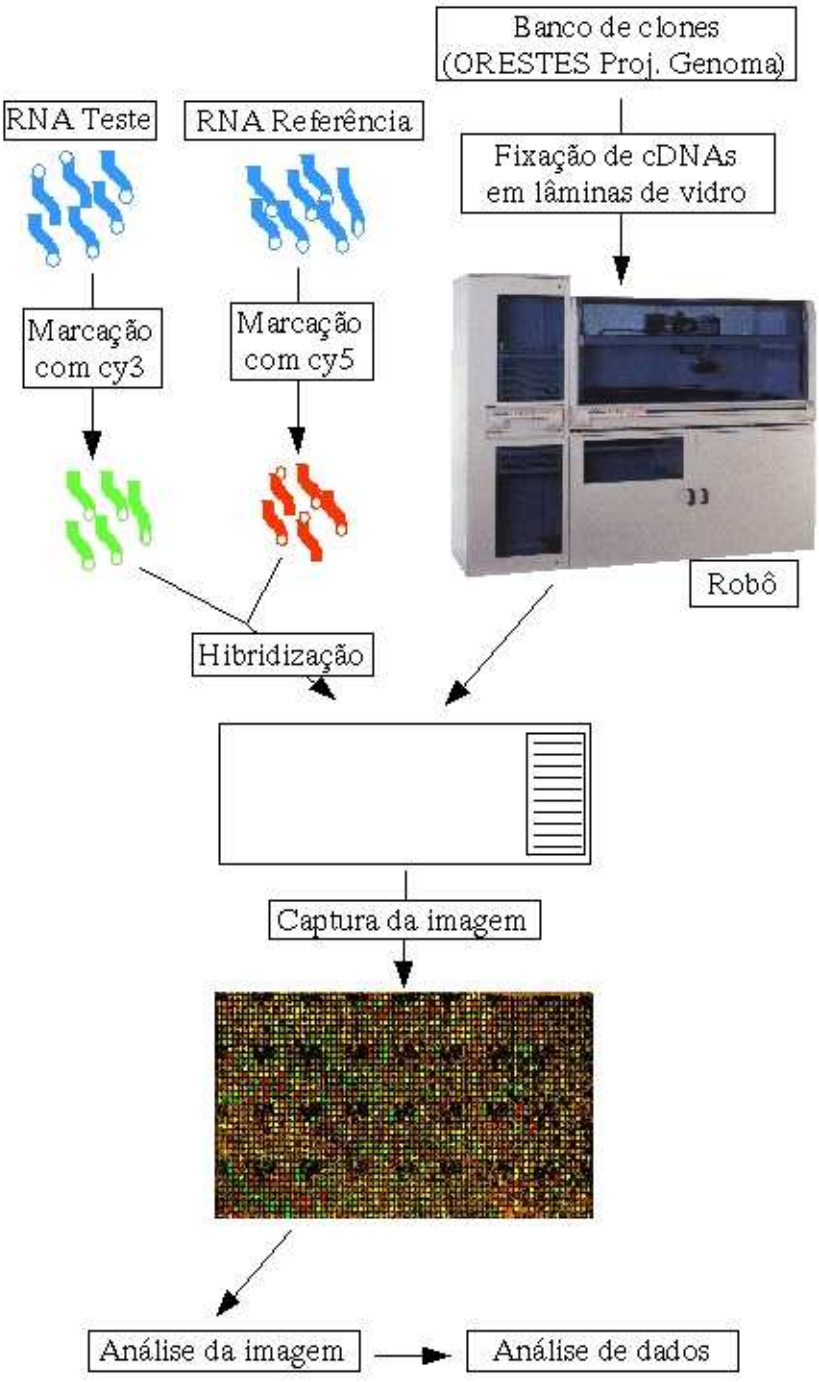


Figura 3: A metodologia de *microarray*.

Esquema mostrando todo o procedimento da metodologia de *microarray* [3]. Nota-se que todo o procedimento é constituído de três etapas diferentes: a extração de RNA das amostras de interesse, a produção e hibridização das lâminas e a quantificação e análise dos dados.

1.4 EXEMPLOS DA UTILIZAÇÃO DE cDNA *MICROARRAY*

Como foi citado anteriormente, experimentos de cDNA *microarray* tem sido extensivamente utilizados na busca de genes diferencialmente expressos entre dois tipos diferentes de tecido. Esta subsecção apresenta alguns trabalhos realizados através da utilização da tecnologia de cDNA *microarray*.

Diferenças nos níveis de expressão gênica entre uma linhagem celular melanótica que não apresentava potencial tumorigênico através da introdução de um cromossomo 6 humano normal (UACC-903(+6)) e uma outra sem a introdução do referido cromossomo (UACC-903) foram detectadas por meio de análises de *microarray* [6]. Um outro trabalho utilizando essa nova ferramenta de pesquisa mostrou que o linfoma difuso de células B grandes se divide em dois grupos molecularmente distintos, apesar desses dois grupos serem morfologicamente iguais. Esse fato se confirma quando vemos que 40% dos pacientes respondem bem ao tratamento atualmente empregado, ao passo que a outra parcela de pacientes simplesmente sucumbem a doença com o passar do tempo [2]. Lin *et al.*, por sua vez, detectou um novo gene (chamado de PART-1) que mostrou níveis de expressão aumentada em células LNCaP de câncer de próstata expostas a andrógenos [28]. Vários outros trabalhos têm confirmado o potencial dos *chips* de DNA nos estudos de expressão gênica [20, 44, 48].

1.5 FATORES POTENCIALMENTE PREJUDICIAIS PARA EXPERIMENTOS DE *MICROARRAY*

Embora os processos de fixação de fragmentos de cDNA nas lâminas de vidro e de hibridização já estejam bem padronizados. Vários cuidados devem ser tomados no momento da seleção de clones para a fixação e na identificação desses clones. Cerca de 1% a 5% dos clones existentes nos bancos de dados mais bem mantidos não contém a seqüência que deveriam conter [22], mostrando que alguns cuidados devem ser tomados para evitar a identificação errada de alguns genes no momento da análise de dados. Uma maneira de se resolver este problema é através do reseqüenciamento dos fragmentos selecionados para a utilização no experimento. A imprecisão dos equipamentos utilizados também constitui um grupo de fatores que podem contribuir

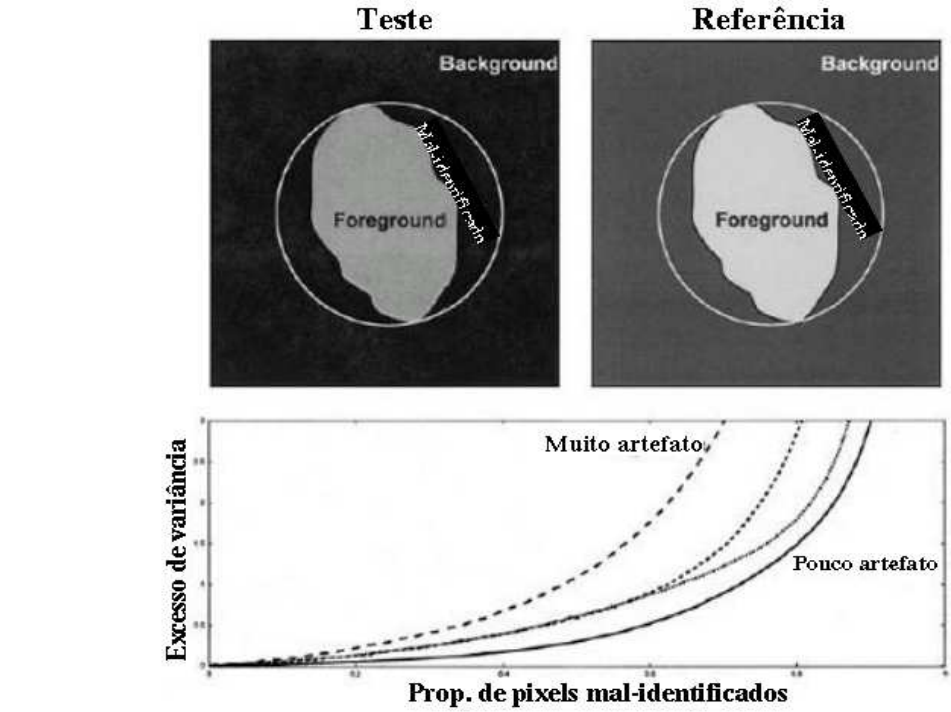
para a obtenção de diferenças que não são reais nos níveis de expressão gênica entre as populações celulares em estudo, mostrando que os equipamentos devem estar bem regulados.

Alguns cuidados devem ser tomados no momento da seleção dos fragmentos a serem utilizados nesses experimentos. Em especial, as seqüências devem estar o mais próximo possível da extremidade 3' do gene que ela representa. Isso vai garantir que o cDNA alvo produzido por transcrição reversa a partir de mRNAs e oligo dT seja capaz de encontrar seus respectivos pares no momento da hibridização. Além disso, a região 3' é mais gene específico, reduzindo hibridização cruzada. Também, *splicing* alternativo do último *exon* são menos freqüentes. No entanto, Iseli *et al.* mostrou a existência de heterogeneidade nas porções 3' de mRNAs humanos devido à utilização de sinais alternativos de poliadenilação [17], o que mostra a necessidade de um cuidado ainda maior na escolha do fragmento a ser fixado nas lâminas. Outro trabalho recente mostrou que fragmentos fixados com tamanhos superiores a 700pb parecem atingir um certo platô em valores de intensidade [42]. Isso nos sugere que o tamanho dos fragmentos utilizados para fixação também tem o potencial de influenciar os dados observados em experimentos de *microarray*.

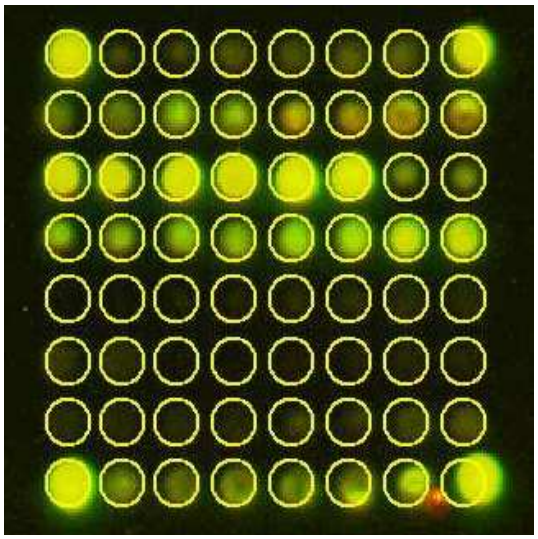
A manipulação da população de RNAs extraída das amostras de interesse também pode ser uma fonte de imprecisão para a metodologia de cDNA *microarray*, uma vez que a qualidade dos mRNAs extraídos é de fundamental importância para a qualidade dos sinais de intensidade dos *spots*. Sabe-se que existem tecidos cuja manutenção da integridade dos RNAs é difícil, como tecidos endoteliais e de estômago. Entretanto alguns trabalhos já apontam soluções para este tipo de problema [21].

Os procedimentos para localização e quantificação de *spots* adotados pelos *softwares* utilizados para a análise de imagens representam uma das principais fontes de erro para as posteriores análises, uma vez que a extração correta dos valores de intensidade de sinal e *background* é de fundamental importância para a confiabilidade dos dados obtidos. Alguns problemas são conhecidos no processamento de imagens de *microarray*, especialmente nas metodologias baseadas em segmentação por círculo fixo, uma vez que elas são mais susceptíveis a incorporar *pixels* do *background* nas regiões referentes aos *spots*, o que introduz um efeito de diluição ao sinal medido para esse ponto [19]. Outro problema muito freqüente com a segmentação por círculo fixo é a intensa necessidade de manipulação das máscaras produzidas, com a finalidade de

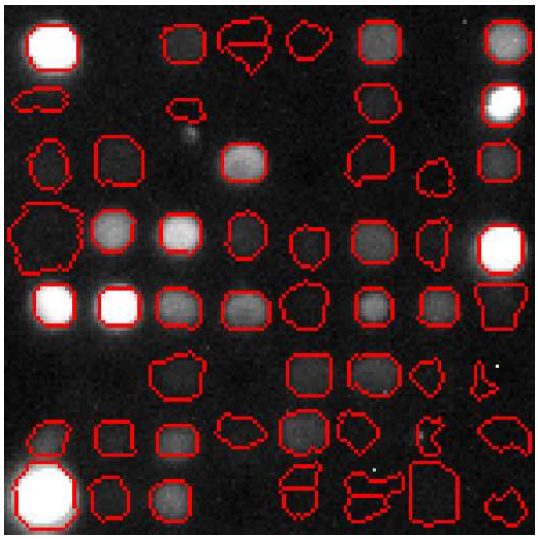
correção de *spots* mal localizados. Como essa correção é feita com base na acurácia visual do usuário, ela continua contendo erros, embora em menor proporção. Uma alternativa para contornar esse problema, e ajudar a minimizar a variabilidade dos dados obtidos, são os procedimentos baseados em segmentação por variação de intensidade, que além de proporcionar a identificação exata de cada *spot* não exige a interferência do usuário pelo fato de ser um processo totalmente automatizado [15]. A Figura 4 ilustra o problema que acontece quando existem *pixels* mal identificados como sinal, e mostra dois exemplos de segmentação: um por círculo fixo e outro por variação de intensidade.



(A)



(B)



(C)

Figura 4: Problemas na localização de *spots*.

(A) Esquema de *pixels* mal identificados [19], à medida que a proporção de *pixels* mal identificados aumenta, a variabilidade dos dados também aumenta. (B) Um fragmento de imagem segmentada por círculo fixo, alguns *spots* mal identificados. (C) Um exemplo de segmentação por variação de intensidade [15], uma solução para o problema.

2 JUSTIFICATIVA

A tecnologia de *microarray* é hoje uma ferramenta muito utilizada na busca de diferenças de expressão gênica entre diferentes tipos de tecidos e células. Essa nova tecnologia não recebeu tanto destaque por acaso, uma vez que ela possibilita uma análise muito próxima do que realmente acontece a nível molecular nas células de qualquer ser vivo, onde os genes se encontram mergulhados em várias vias metabólicas relacionadas e a expressão de um certo gene pode apontar a expressão de vários outros. Ao contrário das metodologias de análise de expressão gênica diferencial que o antecederam, o *microarray* incorpora as características dessas diferentes vias, uma vez que toda a população de mRNA presentes nas amostras de estudo é extraída e compete para hibridização contra o material fixado.

A grande virtude dessa nova tecnologia de análise de expressão gênica diferencial é a sua capacidade de medir numericamente os níveis de mRNAs gerados a partir de vários genes em um único experimento, ou seja, estabelecer valores numéricos para fenômenos biológicos. Esses valores numéricos são aproximações dos níveis reais de expressão dos diferentes genes. É imprescindível, portanto, que essas aproximações sejam boas o bastante para garantir uma análise de dados mais robusta e confiável. Entretanto, dados biológicos são sabidamente controlados por processos estocásticos, o que naturalmente dificulta a análise comparativa entre diferentes amostras de interesse. Somadas a essa variabilidade biológica natural é sabido que existem outras fontes de incertezas intrínsecas à metodologia de cDNA *microarray*.

Além dessas fontes de incertezas está a dificuldade de estimação dos valores de intensidade dos *spots* na etapa de análise de imagens, o que pode resultar em erros nos valores obtidos. Como foi visto no final da Seção 1, existem vários procedimentos que podem ser empregados para realizar essa tarefa, e os diversos *softwares* atualmente disponíveis utilizam várias implementações dessas diferentes estratégias. Assim, é extremamente importante avaliar essas diferentes metodologias de análise de imagens na busca das que sejam mais precisas, ou seja, que minimizem os erros cometidos.

Recentemente foi publicado um trabalho onde foram analisadas diferentes metodologias de análise de imagens decorrentes de experimentos de cDNA *microarray* [53]. As imagens de um conjunto de experimentos de cDNA *microarray* foram repro-

cessadas através de diversas metodologias distintas de análise de imagens, na busca daquelas que ofereceram menores variabilidades. Porém, essa estratégia de análise não permite avaliar o erro cometido pelas metodologias estudadas. Entretanto, esse tipo de avaliação é impossível de ser feito, uma vez que não se sabe antecipadamente qual é a proporção real entre os genes das duas amostras estudadas. Assim, fica evidente a necessidade do desenvolvimento de uma metodologia com base em experimentos controlados onde se saiba de antemão os valores de razão esperados para alguns genes específicos.

3 OBJETIVOS

3.1 OBJETIVO GERAL

O objetivo principal deste trabalho é o desenvolvimento de uma nova estratégia experimental, cuja finalidade principal é testar diferentes procedimentos de análise de imagens decorrentes de experimentos de cDNA *microarray*.

3.2 OBJETIVOS ESPECÍFICOS

1. Avaliar a segmentação de *spots* empregada por diferentes procedimentos.
2. Avaliar o efeito de diferentes tamanhos de cDNA fixados nos valores de intensidade de sinal obtidos.
3. Estudar a variação da razão entre as amostras de teste e referência com os valores de intensidade dos *spots*.
4. Comparar os erros cometidos entre as diferentes estratégias de quantificação de sinal em experimentos controlados onde se saiba a *priori* as razões esperadas.

4 MATERIAL E MÉTODOS

4.1 MATERIAL

4.1.1 Material utilizado nas reações de PCR

- Tampão de PCR 10X (*Invitrogen*, EUA).
- Cloreto de Magnésio ($MgCL_2$) 25 mM (*Invitrogen*, EUA).
- dNTPs 10mM (dATPs, dCTPs, dGTPs e dTTPs) (*Invitrogen*, EUA).
- ddH_2O: Água deionizada.
- Enzima *Taq Polimerase* 5U/μl (*Invitrogen*, EUA).

4.1.2 Soluções para eletroforese de DNA

- TAE 50X: 242g Tris base, 57,1ml ácido acético glacial, 100ml EDTA 0,5M pH8.
- Gel de agarose 1%: Pesar 0,25g de agarose e dissolver em 25ml de TAE 1X. Aquecer até que a agarose se dissolva. Adicionar 1μl de brometo de etídeo 10μg/ml.
- Solução tampão 6X para amostras de DNA: glicerol 30%, azul de bromofenol 0,25% e xilenocianol 0,25%.

4.1.3 Fragmentos de cDNA utilizados

Foram utilizados seis fragmentos de cDNAs distintos para este trabalho. Três deles (gene Q, LysA, TrpC) são utilizados com freqüência no laboratório como controles positivos (ou *LandMarker*) da reação de hibridização nos experimentos realizados. Outros dois são cDNAs de camundongo (Il-6 e Irf-1), gentilmente cedidos pelo Dr. Eduardo Abrantes, e um ORESTES (RC0-ST0280-021299-031-c05, que será chamado de ST0280), gerado durante o Projeto Genoma Humano do Câncer (FAPESP/Ludwig). A Tabela 4 mostra algumas características importantes desses cDNAs.

Tabela 4: Características dos cDNAs utilizados no projeto.
Nota-se a variação de tamanhos e proporções de nucleotídeos A, C, G e T. Essas proporções são dadas em porcentagem.

Genes	Tam. (pb)	Prop. A	Prop. C	Prop. G	Prop. T	Org. de orig.	nº
LysA	303	29,1	21,9	24,8	24,2	*B.subtilis*	1
TrpC	338	26,1	24,4	20,5	29	*B. subtilis*	2
Gene Q	637	27,7	23,6	28,3	20,4	fago λ	3
ST0280	659	30,7	16	18,3	35	ORESTES	4
Il-6	948	33,3	19,6	18,1	29	Murino	5
Irf-1	2069	25,4	25,8	26,7	22,1	Murino	6

4.1.4 Oligonucleotídeos utilizados nas reações de PCR

Os cDNAs do gene Q e ST0280 foram clonados em plasmídeo *pUC18*, os clones de Irf-1 e Il-6 estão em *BlueScript*, enquanto que TrpC e LysA em um vetor especial que já contém um sítio promotor p/ a enzima T3 e uma região poli T na outra extremidade, o que possibilita a produção do mRNA sintético. Os fragmentos de ST0280, Il-6 e Irf-1, necessitaram de um par de oligonucleotídeos adicionais cada com a introdução de um sítio promotor para a enzima SP6 no oligo senso e cauda poli T no antisenso. O gene Q já estava clonado com essas regiões inseridas e não precisou desses oligos especiais. A Tabela 5 lista todos os oligonucleotídeos utilizados nas reações de PCR para este projeto.

4.1.5 Soluções de lavagens de lâminas

- solução de lavagem 1: SSC 2X.
- solução de lavagem 2: SSC 0,1X e SDS 0,1%.
- solução de lavagem 3: SSC 0,1X.

4.2 MÉTODOS

4.2.1 Fixação dos cDNAs em lâminas de vidro

Os seis cDNAs citados anteriormente foram amplificados por PCR utilizando-se os oligonucleotídeos específicos (ver Tabela 5). A reação básica de PCR (volume final de 100μl) foi realizada, com os reagentes descritos no ítem 4.1.1, da seguinte maneira:

- 10μl de Tampão
- 2μl de dNTP *mix*
- 2μl de oligonucleotídeos senso 10μM
- 2μl de oligonucleotídeos antisenso 10μM
- 3μl de $MgCl_2$
- 0,5μl de *Taq DNA polimerase*
- 2μ de plasmídeo diluído 1/50 contendo o DNA a ser amplificado
- H_2O em quantidade suficiente para completar 100μl

O termociclador realizou 35 ciclos de 95°C/45seg, 55°C/45seg, 72°C/1min precedidos por um ciclo inicial de 95°C/5min e um ciclo de extensão final de 72°C/5min. Cerca de 5μl de cada produto amplificado foi adicionado em 1μl de tampão 6X (descrito na subseção 4.1.2) e fracionado em gel de agarose 1%, para imediata verificação do sucesso da reação.

Com a finalidade de retirar os oligonucleotídeos não incorporados na reação e evitar erros na quantificação dos cDNAs, esse produto de PCR foi purificado em colunas de *sephadex G50* (*Amersham Pharmacia*, EUA) montadas em um sistema de multifiltração em placas de 96 poços MAHV45 (*Millipore*, EUA), onde as amostras foram distribuídas no centro das 96 minicolunas de *Sephadex* seguido por uma centrifugação por cinco minutos a 910g. Com esse produto já purificado foi feita a leitura de absorbância a 260nm (A_{260}) em espectofotômetro de uma alíquota de 10μl dos produtos diluídos em 90μl de ddH_2O. A concentração dos DNAs foi baseada no seguinte fator de conversão: 1 A_{260} unidade de DNA dupla fita (dsDNA) equivale a 50μg/ml de concentração. Sabendo-se que em 1μg de dsDNA (com 1000pb) existem 9,1x10^{11}

Tabela 5: Oligonucleotídeos utilizados no trabalho.

Oligonucleotídeos para amplificação dos fragmentos utilizados no trabalho. Os trechos de seqüência em negrito representam a região que contém o sítio promotor para a enzima SP6, em itálico temos as caudas poli T.

Oligo	**Seqüência 5' – 3'**	**Classif.**
Puc senso	CGCCAGGGTTTTCCCAGTCACGAC	Universal
Puc antisenso	TTTCACACAGGAAACAGCTATGAC	Universal
TrpC senso	TTCTATTCAAACCACTCCC	Específico
TrpC antisenso	GCTCTCCGTCCTTTAATTC	Específico
LysA senso	TGGAAACAATAAGAGCAG	Específico
LysA antisenso	CGATATGGCAATGGACAC	Específico
IRF1 senso	CTTTCACAGTCTAAGCC	Específico
IRF1 antisenso	ACATATTTACACAGGTCC	Específico
Q-start plus	**TCATTTAGGTGACACTATAG**AGACTCGAAAGCGTA	Prod. mRNA
Q-end plus	*TTTTTTTTTTTTTTTTT*ACGTGTGACCGCATT	Prod. mRNA
IRF1 SP6	**TCATTTAGGTGACACTATAG**CTTTCACAGTCTAAGCC	Prod. mRNA
IRF1 DT	*TTTTTTTTTTTTTTTTT*ACATATTTACACAGGTCC	Prod. mRNA
IL6 SP6	**TCATTTAGGTGACACTATAG**CACCAAGAACGATAGTC	Prod. mRNA
IL6 DT	*TTTTTTTTTTTTTTTTT*GATTTTAGGTTATCATTTC	Prod. mRNA
ST0280 SP6	**TCATTTAGGTGACACTATAG**GGAACACTTACAAAGAGG	Prod. mRNA
ST0280 DT	*TTTTTTTTTTTTTTTTT*AGTAATGTCTTCTTAAGAAATG	Prod. mRNA

moléculas, foram calculados os números de moléculas/ml de todos os fragmentos de DNA utilizados.

Um número de aproximadamente 4,5x10^{12} moléculas de cada fragmento de DNA foi, então, precipitado em placas de 96 poços adicionando-se 1/10 do volume total do produto de acetato de sódio e volume igual de isopropanol. Essa reação foi incubada por uma hora a 4ºC, sendo centrifugada a 3200g por 60 minutos também a 4ºC. O sobrenadante foi descartado e foram adicionados 100μl de etanol 70% por poço. Após 5 minutos de incubação à temperatura ambiente, a placa foi centrifugada a 3200g/30min e então o sobrenadante descartado. O etanol restante foi evaporado em *SpeedVac* por 15 minutos. O DNA foi solubilizado em 6μl de DMSO 50% por poço.

A partir do produto concentrado de cada um dos seis cDNAs (genericamente chamados de g_i, onde esse i equivale a numeração da Tabela 4) foram feitas 4 diluições seriadas, também em DMSO 50%, que obedeceram a seguinte regra

$$d_n(g_i) = \frac{1}{2^{n-1}}, \quad n = 1, \cdots, 5 \quad \text{e} \quad i = 1, \cdots, 6. \tag{1}$$

onde $d_n(g_i)$ representa a *n-ésima* diluição do *i-ésimo* gene (cDNA). Para $n = 1$ temos a concentração inicial de cada cDNA, ao passo que para $n = 2$ até $n = 5$ temos as outras quatro diluições realizadas.

Todos os seis cDNAs e suas respectivas diluições foram distribuídas em posições específicas de seis placas de 384 poços para fixação em lâminas de vidro. A fixação foi feita pelo robô *Flexys* (*Genomic Solutions*, Inglaterra), utilizanado-se lâminas *CMT-GAPS*^{TM} *II* (*Corning*, EUA).

4.2.2 Construção dos mRNAs sintéticos

Os fragmentos dos genes Q, LysA e TrpC já estavam clonados em plasmídeos com cauda poli T em uma extremidade e sítio promotor p/ uma enzima *RNA polimerase*, sendo que o gene Q estava clonado com o sítio promotor da enzima SP6 e os outros dois com sítio p/ T3.

Os três cDNAs restantes não estavam clonados com nenhum sítio promotor e, então, foram amplificados em uma nova reação de PCR a partir de 2μl do PCR utilizado na etapa de fixação nas lâminas utilizando os oligonucleotídeos para construção de mensageiro descritos na Tabela 5. Nota-se que esses oligos também continham sítio promotor para a enzima SP6. Essa PCR foi semelhante aquela citada na subseção

4.2.1, apenas aumentando o número de ciclos para 40 e o tempo de extensão a 72°C para 1 min e 30seg no programa do termociclador. Esse produto de PCR foi purificado e concentrado utilizando o filtro *Microcon YM-100* (*Millipore*, EUA), seguindo instruções do fabricante. Alíquotas de 5μl do material desses três cDNAs foram fracionadas em gel de agarose, para verificação da amplificação. Todo o produto de Irf-1 foi fracionado em um gel de agarose para nova purificação a partir de uma fatia do gel com o objetivo de eliminar produtos contaminantes. Para isso foi utilizado o *kit Concert rapid gel extraction system* (*Invitrogen*, EUA) seguindo protocolo do fabricante.

A próxima etapa foi a transcrição *in vitro* para a produção de mRNA sintético. O produto inicial da reação deve conter DNA linearizado, para tanto os plasmídeos contendo os fragmentos de LysA e TrpC foram digeridos com a enzima de restrição *Not1* (*New England BioLabs*, Inglaterra). Para os demais cDNAs foram utilizados produtos iniciais derivados de PCR feitos com os oligonucleotídeos de produção de mRNA listados na Tabela 5.

A transcrição *in vitro* foi feita com o *kit Ribomax* (*Promega*, EUA) seguindo protocolo do fabricante. O produto da reação foi tratado com *RQ1 RNAse free DNAse* para degradação do DNA a 37°C por 15 minutos, seguido de uma purificação com PCI (fenol - 25 partes, clorofórmio - 24 partes e álcool isoamílico - 1 parte) centrifugando por 2 minutos a 20000g. Um outro tratamento com CI (clorofórmio - 24 partes e álcool isoamílico - 1 parte) foi feito para remoção de resíduos de fenol. Em seguida o produto da reação foi purificado em *Sephadex G50*, precipitado e lavado como descrito na seção anterior. Por fim, o RNA foi ressuspendido em H_2O DEPC (ddH_2O tratada com 0,1% de dietilpirocarbonato por 12 horas a 4°C e esterilizada em autoclave por 15 minutos a 121°C), quantificado em espectofômetro e armazenado a -70°C. O sucesso da reação foi verificado por gel de agarose.

4.2.3 Hibridizações

Os seis mRNAs produzidos foram misturados em quantidades específicas e conhecidas juntamente com oligo(dT) (*Invitrogen*, EUA). A partir dessa mistura, foram sintetizados os cDNAs utilizando a enzima de transcrição reversa *Superscript II* (*Invitrogen*, EUA) em presença de fluorocromos *cy3* ou *cy5* (*Amersham Pharmacia*, EUA). O cDNA marcado por fluorescência foi purificado em coluna *sephadex G50* (*Amersham*

Pharmacia, EUA), seguindo protocolo fornecido pelo fabricante. Reações marcadas com *cy3* e *cy5* para um mesmo experimento foram misturadas e, a essa reação foi adicionado tampão de hibridização. Esse produto foi hibridizado contra uma lâmina (ver subseção 4.2.1) a 42°C por um tempo aproximado de 20 horas. A hibridização foi feita em estação de hibridização *GeneTAC* (*Genomic Solutions*, Inglaterra). As lâminas foram previamente incubadas a 42°C por seis horas em solução de pré-hibridização, sendo lavadas em ddH_2O e centrifugadas a 30g por cinco minutos em tubo Falcon 50ml. Após a hibridização as lâminas foram lavadas com as soluções descritas na subseção 4.1.5 e digitalizadas com os *scanners GenePix* (*Molecular Dynamics*, EUA) ou *ScanArray* (*Packard BioScience*, EUA) que utilizam tecnologia *laser scanner*. Esses equipamentos permitem a digitalização de imagens cuja resolução dos *pixels* vão desde 5μm até 50μm, em todos os experimentos realizados aqui foi utilizada resolução de 10μm. Nas imagens obtidas neste trabalho os *spots* apresentaram diâmetro aproximado de 300μm com distâncias de centro à centro de aproximadamente 500μm (horizontal) e 450μm (vertical).

Como as quantidades de mRNAs sintéticos utilizadas no início das reações de transcrição reversa eram conhecidas, foi possível estabelecer antecipadamente as razões que eram esperadas entre os fragmentos utilizados para marcação com *cy3* e *cy5* e avaliar os erros cometidos ao final das análises.

4.2.4 Estimação das intensidades

As imagens foram quantificadas utilizando quatro *softwares* distintos que empregam diferentes metodologias de localização e quantificação dos *spots*. Todas essas metodologias utilizam a imagem composta das duas bandas (*cy3* e *cy5*) para a segmentação dos *spots*. Uma vez definidos os *pixels* a serem computados, foi utilizada a média dos valores de intensidade de sinal e *background* para a estimação dos valores de interesse que são obtidos através da tabela de dados numéricos exportada pelo *software*. No total foram utilizados 9 procedimentos diferentes neste trabalho, que são:

1. ***Circfix*** – segmentação de círculo fixo. Assume que todos os *spots* são circulares e com mesmo tamanho. Para os experimentos construídos neste trabalho o diâmetro utilizado para os *spots* foi de 31 *pixels* (310μm). Essa metodologia

utiliza todos os *pixels* que estão dentro do raio citado anteriormente para o cálculo dos valores de sinal. Para os valores de *background* são usados todos os *pixels* que estão fora da área do *spot* mas que estejam dentro de um certo quadrado a partir do centro do *spot* com lado medindo 40 *pixels* (400μm).

2. ***Adap*** – segmentação adaptativa onde os *pixels* da região de sinal são selecionados com base em um teste estatístico que compara a distribuição dos *pixels* do *spot* com a distribuição dos *pixels* do *background*. Primeiramente é definido um diâmetro máximo para os *spots* (para os experimentos desenhados aqui foi utilizado 350μm). Para a seleção de *pixels* do *background* são definidos dois diâmetros diferentes (aqui foram utilizados 510μm e 675μm) que definem um 'anel' de onde os *pixels* são selecionados. A seguir é feito um teste estatístico de *Mann-Whitney*, cujo valor de $p < 0,0001$ é considerado significativo, entre os oito *pixels* medianos do *background* e cada grupo de oito *pixels* da região de sinal. Somente os *pixels* que apresentarem diferenças significativas em relação à distribuição dos *pixels* do *background* são considerados como pertencentes ao *spot*.

3. ***Circhist-50-50*** – nesta metodologia as áreas de sinal e *background* são definidas de maneira semelhante à metodologia adaptativa, sendo usado 310μm para o diâmetro do *spot* e 490μm e 675μm para a definição da região do *background*. Uma vez definidas essas regiões é calculado um histograma da distribuição dos valores de intensidade dos *pixels* do *foreground* e outro para a distribuição dos valores de intensidade do *background*. A seguir são definidos percentis máximos e mínimos para seleção dos *pixels* que serão quantificados. Aqui foram selecionados os *pixels* que estavam entre os percentis 45 e 95 da distribuição dos *pixels* da região de sinal e entre os percentis 5 e 55 da distribuição do *background*. Os *pixels* selecionados são independentes nos dois canais, ou seja, os *pixels* usados para *cy3* e *cy5* podem ser diferentes.

4. ***Circhist-100-20*** – é um procedimento exatamente igual ao da metodologia *circhist-50-50*, porém utilizando todos os *pixels* da região de sinal e os que estejam entre os percentis 1 e 20 da distribuição do *background*.

5. ***Circhist-30-10*** – é um método bastante semelhante à metodologia *circhist-50-50*, com a exceção de que nesta metodologia é calculado apenas um histograma para todos os *pixels* do quadrado centrado no *spot* com lados igual ao espaçamento entre *spots*. A seguir são selecionados 30% dos *pixels* mais intensos que estejam dentro de uma circunferência com diâmetro igual a metade do lado do quadrado menos um *pixel*. Para o *background* são selecionados 10% dos *pixels* menos intensos que estejam fora de uma circunferência com raio igual a metade do lado do quadrado mais um *pixel*. Nesta metodologia os *pixels* selecionados são sempre os mesmos para os dois canais da imagem.

6. ***Hist-15-15*** – método baseado na segmentação por histograma. Aqui é definido um retângulo, medindo 500μm por 450μm, ao redor do *spot*, onde é calculado o histograma de distribuição dos valores de intensidade de todos os *pixels* que estão dentro desse retângulo. A seguir é selecionada uma proporção dos *pixels* mais intensos para representar a região de sinal e uma outra proporção de *pixels* menos intensos para representar o *background*. Aqui foram utilizados os percentis de 80 a 95 para a região de sinal e de cinco a 20 para a região do *background*.

7. ***Segment-50-50*** – procedimento baseado em segmentação por variação de intensidade. A segmentação é feita através de operadores de morfologia matemática [15]. A seguir são calculados dois histogramas, um para as regiões de sinal e outro para o *background*, e selecionados percentis de maneira semelhante ao método *circhist-50-50*. Da mesma forma, os *pixels* selecionados são independentes nos dois canais.

8. ***Segment-100-20*** – metodologia igual à anterior com percentis de seleção iguais à da técnica *circhist-100-20*.

9. ***Segment-100-100*** – metodologia semelhante às duas anteriores, selecionando todos os *pixels* que constituem a região do *spot* e todos os da região de *background*.

O método *circfix* foi utilizado através do *software ScanAlyze*, [8]. Os métodos *adap*, *circhist-50-50*, *circhist-100-20* e *hist-15-15* foram aplicados através do *software QuantArray*, [32]. É importante mencionar aqui que todas as metodologias empregadas através do *software QuantArray* não selecionam nenhum *pixel* que esteja locali-

zado no exterior do retângulo definido na descrição da técnica *hist-15-15*. O método *circhist-30-10* foi empregado através do *software Spot*, [18, 19]. As estratégias de segmentação por variação de intensidade foram aplicadas através do *software* Bioinfo-USP [15]. Para maiores detalhes em relação a cada um destes programas consulte o Anexo A.

4.2.5 Normalização dos dados

Neste trabalho foram construídos dois tipos de experimentos diferentes. Em alguns experimentos foram utilizadas concentrações iguais de mRNAs para a síntese do cDNA utilizado para hibridização, ou seja, a razão esperada entre as amostras de teste (T) e referência (R) deve ser igual a um para todos os fragmentos utilizados. Em contrapartida, outros experimentos foram construídos a partir de concentrações maiores de mRNAs na amostra teste, onde eram esperadas razões entre T e R diferentes de um.

Desta maneira, as normalizações [34, 54] se deram de duas formas diferentes. Em experimentos que deveriam apresentar razão entre T e R igual a um, utilizamos

$$\sum_{i=1}^{n} T_{cy3_i} = k \sum_{i=1}^{n} R_{cy5_i} \tag{2}$$

onde k, que é a constante de normalização, é obtida de $\sum T_{cy3_i} / \sum R_{cy5_i}$, T_{cy3_i} e R_{cy5_i} são os valores de intensidade corrigidos pelo *background* para cada *spot* da lâmina nas amostras de teste e referência, marcados com *cy3* e *cy5*, respectivamente, e n é o número de *spots* viáveis, ou seja, sem *flag* (indicação de *spot* ruim) e com valor de intensidade de sinal positivo após correção pelo *background*. Esse tipo de correção é conhecida por normalização pela energia total.

Para experimentos onde as razões esperadas entre T e R são diferentes de um utilizamos a auto-normalização (*self-normalization*) [54], que é dada por

$$\log \frac{T}{R} = \frac{(\log T_{cy3} - \log R_{cy5}) + (\log T'_{cy5} - \log R'_{cy3})}{2} \tag{3}$$

onde T e R representam as amostras de teste e referência; T_{cy3} e R_{cy5} representam as intensidades corrigidas pelo *background* para *cy3* e *cy5* na primeira lâmina e T'_{cy5} e R'_{cy3} representam as intesidades corrigidas pelo *background* para *cy5* e *cy3* na lâmina de *swap*.

4.2.6 Estimação das expressões

As análises foram feitas através de medidas de razão obtidas entre os valores de intensidade nos dois canais. Inicialmente foi feita a correção dos valores de intensidade de sinal pelos valores de intensidade do *background* para cada *spot* das lâminas, sendo que os valores de razão foram calculados logo em seguida. Esses dados foram normalizados como descrito na seção anterior. Essa estratégia foi adotada para todos os procedimentos de análise de imagens, a fim de garantir que exatamente os mesmos cálculos fossem aplicados para todos os conjuntos de dados.

Para a avaliação das diferenças encontradas entre os valores de razão obtidos nos experimentos controlados e as razões esperadas de acordo com os experimentos construídos utilizamos a seguinte função de erro

$$E = \frac{\sum_{i=1}^{p} |r_i - r_e|}{p} \tag{4}$$

onde p é o número de *spots* de um dado cDNA fixado na lâmina, r_i e r_e são as razões observada para cada *spot* desse cDNA e esperada, respectivamente. As análises foram baseadas nas razões médias, desvios-padrão e erros obtidos para cada cDNA diferente fixado nas lâminas, onde elegemos o melhor conjunto de dados aquele que mais se aproximou da razão esperada de acordo com a construção do experimento juntamente com um menor desvio-padrão. Essas análises foram feitas através de *scripts* escritos em linguagem MatLab 6 (*The MathWorks*, EUA) [26].

5 RESULTADOS

5.1 A CONSTRUÇÃO DAS LÂMINAS

Todos os cDNAs do projeto, ver subseção 4.1.3, foram amplificados com sucesso para a fixação nas lâminas de vidro. A Figura 5 mostra uma foto de um gel de agarose onde foram fracionadas amostras dos produtos amplificados dos seis cDNAs. Após purificação e determinadas as A_{260nm}, o cDNA que apresentou menor concentração foi o Irf-1. Assim as concentrações de todos os outros produtos de PCR foram normalizadas pela concentração do Irf-1, que ficou em torno de 2,5 x 10^{13} moléculas/ml.

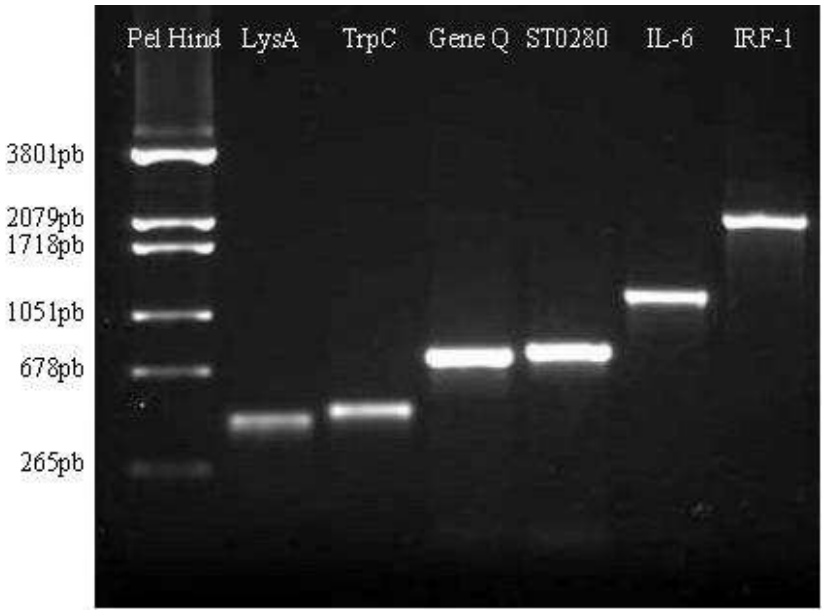


Figura 5: Amplificação dos fragmentos utilizados.
Foto do gel de agarose contendo amostras dos seis cDNAs amplificados para fixação em lâminas de vidro.

Para o presente trabalho foram produzidas 24 lâminas de vidro contendo *spots* dos seis cDNAs citados anteriomente juntamente com suas respectivas diluições indicadas na Equação (1). Essas lâminas contém 32 blocos (ou *subarrays*), dispostos em 4 colunas por 8 linhas, onde cada um destes blocos contém 64 *spots*, organizados em uma geometria de 8 linhas por 8 colunas, ver Figura 1 (A). Desta maneira, existe um número total de 2048 pontos contidos em cada uma das lâminas produzidas. A disposição de todos esses *spots* foi cuidadosamente planejada a fim de se aproveitar melhor todos os espaços disponíveis. Foram desenhados quatro blocos cujos *spots*

estão organizados de forma diferente e dispostos repetidamente dentro das lâminas, de maneira que são encontradas oito cópias de um *minichip* dentro de uma mesma lâmina. Para os cDNAs de LysA, gene Q, Il-6 e Irf-1 na concentração original (diluição 1) há um total de 96 pontos por lâmina, ao passo que os demais genes-diluições têm um total de 64 pontos por lâmina. As Figuras 6 e 7 esquematizam a fixação desses materiais nas lâminas, mostrando a disposição dos blocos dentro da lâmina e dos *spots* dentro de cada bloco diferente, respectivamente. Os pontos pretos da Figura 6 representam pontos fixados sempre com diluição um, ou seja, na concentração inicial, o esquema de cores representa as diluições como mostra a Tabela 6.

Tabela 6: Esquema de cores utilizado na Figura 6

Esta tabela é uma legenda para a coloração utilizada na Figura 6, pois foram utilizadas várias diluições nos cDNAs fixados nas lâminas de vidro.

Cor	**Diluição**
vermelho	1 (sem diluição)
verde	2 (diluição 1/2)
azul	3 (diluição 2/2)
amarelo	4 (diluição 3/2)
marrom	5 (diluição 4/2)

Os mRNAs sintetizados a partir das reações de transcrição *in vitro* foram quantificados em espectofotômetro e estocados a -70^{o}C. Para as hibridizações as concentrações desses mensageiros sintéticos foram novamente normalizadas.

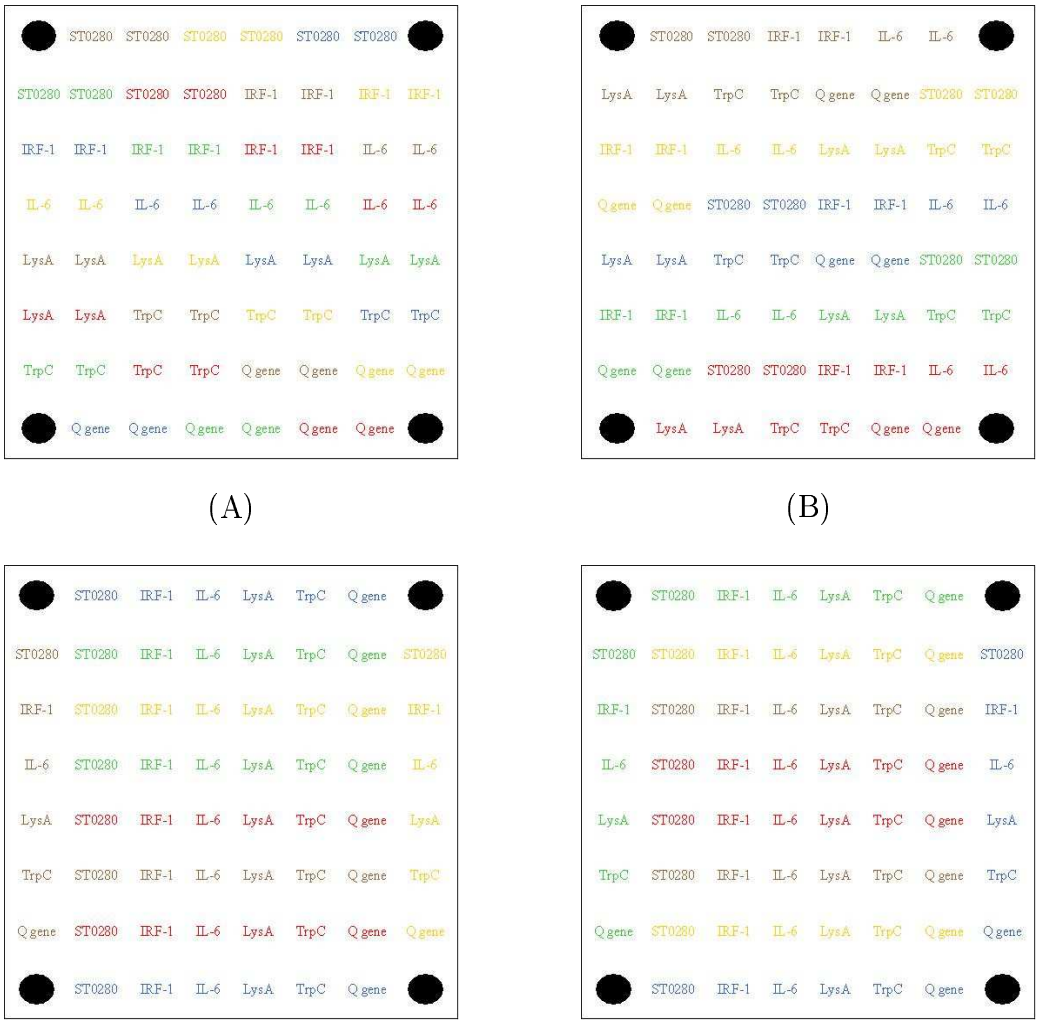


Figura 6: Os quatro blocos produzidos.

Essa figura ilustra um esquema dos diferentes blocos produzidos, ver Figura 7. (A) - Bloco 1. (B) - Bloco 2. (C) - Bloco 3. (D) - Bloco 4. As cores usadas aqui representam as diluições dos cDNAs fixados e estão indicadas na Tabela 6.

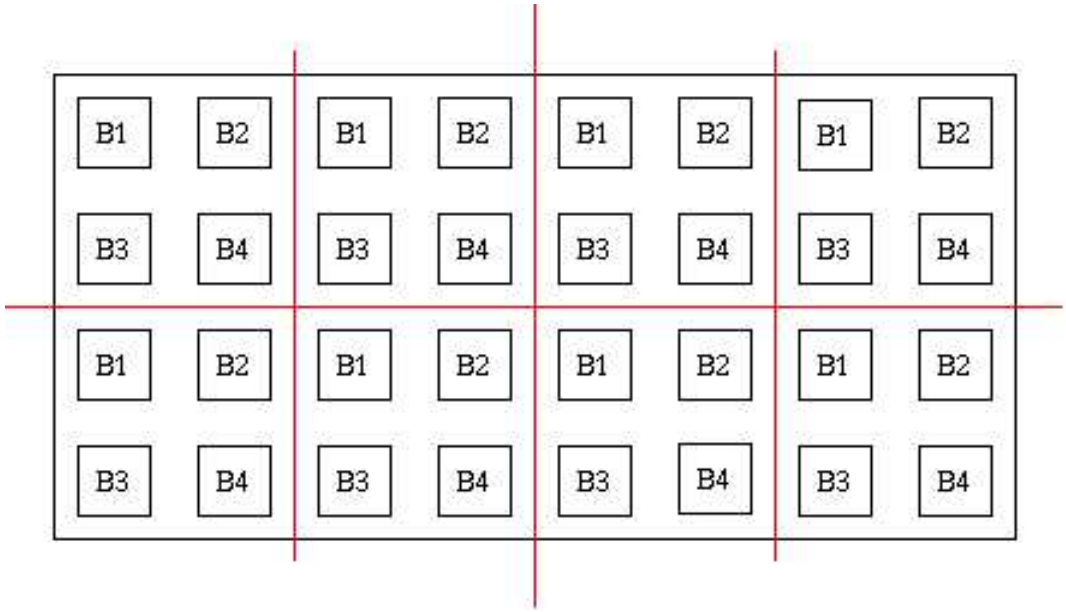


Figura 7: A fixação de quatro blocos diferentes.

Os fragmentos foram fixados em quatro blocos diferentes. Essa figura ilustra o esquema em que esses blocos foram organizados nas lâminas.

5.2 AS HIBRIDIZAÇÕES

Foram hibridizadas com sucesso 13 lâminas, sendo que outras duas foram utilizadas em testes da reação de transcrição *in vitro* na etapa de produção dos mRNAs. Essas 13 lâminas hibridizadas constituem um conjunto de oito experimentos diferentes que estão listados nas Tabelas 7 e 8. A Figura 8 mostra uma imagem composta dos dois canais de um experimento digitalizado no *scanner Genepix* onde a razão esperada entre teste e referência é um para todos os fragmentos utilizados, os retângulos e linhas vermelhas mostram o esquema ilustrado na Figura 7.

O protocolo de hibridização utilizado no laboratório já está bem definido, onde sabe-se qual é a quantidade de RNA total que deve ser utilizado na reação de marcação com *cy3* ou *cy5*. Entretanto, nos experimentos propostos nesse trabalho não se sabia qual era a quantidade de mRNA necessária para iniciar a reação de transcrição reversa, uma vez que foram utilizados mRNAs sintéticos construídos em laboratório por transcrições *in vitro*. Assim, era necessário saber qual a relação entre o cDNA alvo e o cDNA sonda que apresentaria melhores resultados.

Com o intuito de responder a essa questão foram desenhados três experimentos com concentrações iguais de mRNAs utilizados como amostras de teste e referência. Esse conjunto de experimentos obedeceu a uma grande variação entre a proporção de

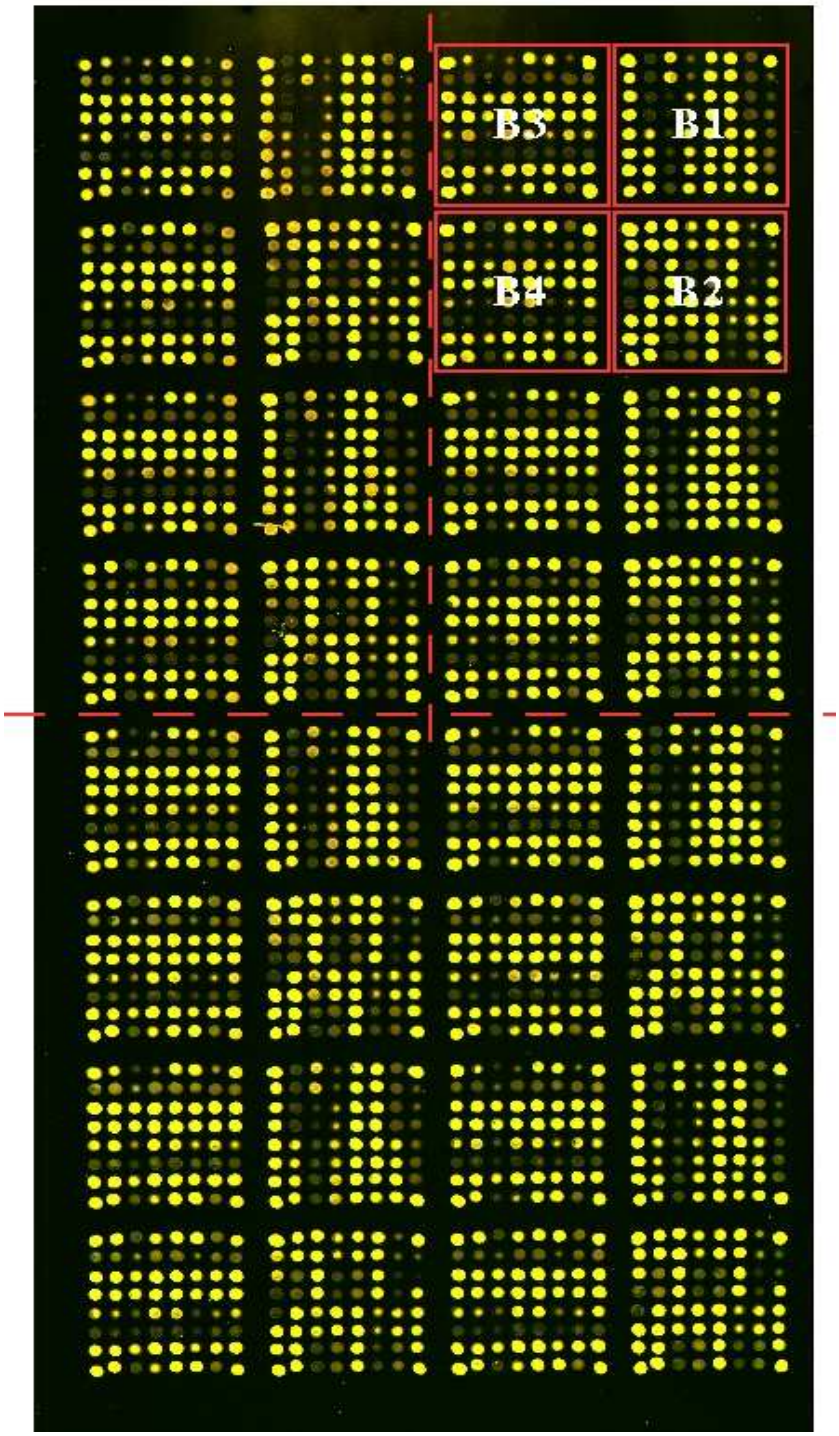


Figura 8: Imagem de um experimento com proporções iguais de amostras teste e referência.

Imagem composta de um experimento com proporções iguais de amostras teste e referência. Essa imagem foi digitalizada pelo *scanner Genepix*. As linhas tracejadas e os retângulos vermelhos indicam o esquema ilustrado na Figura 7.

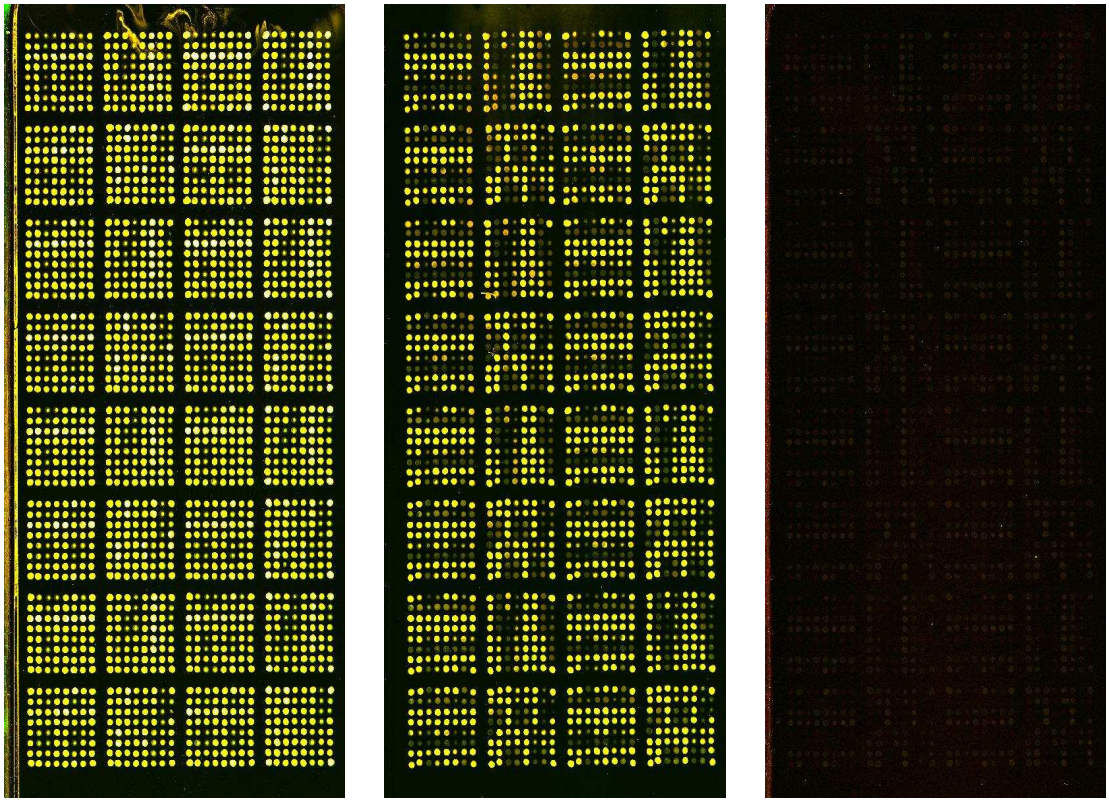
cDNA flutuante em relação ao cDNA fixado, como pode ser obervado na Tabela 7. Nota-se que o experimento 1 apresentou uma proporção de material flutuante/fixado de aproximadamente 500, ao passo que o experimento 2 tinha a proporção de 5 e o experimento 3 apresentava a proporção de 0,05. A Figura 9 mostra as imagens compostas desses três experimentos, onde nota-se a influência da proporção flutuante/fixado no sinal obtido. Os pontos brancos indicam *spots* saturados, ou seja, que atingiram o nível máximo de sinal de intensidade. Vê-se que o experimento 2, onde a proporção entre o cDNA flutuante e o cDNA fixado é aproximadamente equivalente apresentou *spots* bem definidos e sem saturação de sinal (Figura 9 (B)), o que nos levou a hibridizar os demais experimentos obedecendo a proporções de flutuante/fixado semelhantes à deste experimento.

Tabela 7: Experimentos para o ajuste da proporção flutuante/fixado. Esta tabela descreve três experimentos realizados com o objetivo de ajustar a proporção de material flutuante a ser utilizado. A variação dessa proporção pode ser notada na coluna Flut/Fix, que indica a proporção entre material flutuante e fixado utilizada.

Experim.	**Quant. lâm.**	***Scanner***	***Teste/Ref***	**Flut/Fix**
1	1	*Genepix*	1/1	500/1
2	1	*Genepix*	1/1	5/1
3	1	*Genepix*	1/1	1/20

O experimento 2, listado na Tabela 7, foi renomeado para *exp1/1* e também foi utilizado nas análises quantitativas, sendo que a razão esperada entre teste e referência era de um para todos os cDNAs utilizados. Entretanto experimentos baseados em razão entre teste e referência igual a um podem mascarar os dados obtidos, pois os *spots* que tiveram baixa intensidade devem naturalmente tender a essa razão por apresentarem sinais muito próximos do valor de intensidade do *background*. Para contornar esse problema, os experimentos exp3/1 e exp6/1 foram hibridizados com quantidades diferenciadas de cDNA marcado, apresentando razões esperadas de três e seis, respectivamente.

Além disso, os experimentos exp1/1-5/1, exp1/1-2/1 e exp1/1-10/1 tiveram dois grupos de proporções diferentes entre os mRNAs utilizados para a marcação com *cy3* e



(A) (B) (C)

Figura 9: Comparação entre as proporções de alvo/sonda.

Essa figura ilustra as diferenças observadas entre as proporções de material flutuante e fixado, com base nos experimentos 1, 2 e 3 descritos na Tabela 7. (A) - experimento 1 (500 vezes mais cDNA flutuante em relação ao fixado), (B) - experimento 2 (5 vezes mais cDNA flutuante em relação ao fixado), (C) - experimento 3 (20 vezes mais cDNA fixado em relação ao flutuante).

cy5, sendo que os mRNAs de TrpC, ST0280 e Irf-1 foram preparados com proporções iguais entre os dois fluorocromos, ao passo que os cDNAs de LysA, gene Q e Il-6 foram preparados respeitando-se proporções variáveis, onde o experimento exp1/1-5/1 apresentava cinco vezes mais material marcado em um dos canais, o experimento exp1/1-2/1 apresentava diferença de duas vezes e o exp1/1-10/1 uma diferença de dez vezes. Cabe ressaltar que todos esses experimentos foram feitos em duplicatas com a inversão dos fluoróforos ou *swap*. O *swap* de corantes, descrito na última coluna da Tabela 8, consiste de duas hibridizações com a inversão dos corantes entre as duas amostras estudadas e é utilizado com o objetivo de aplicar a auto-normalização, dada pela Equação (3).

Todos os experimentos utilizados para as análises quantitativas neste trabalho estão listados na Tabela 8.

Tabela 8: Experimentos realizados.

Esta tabela descreve os experimentos realizados neste trabalho para as análises quantitativas, juntamente com a quantidade de lâminas utilizadas, o *scanner* usado para captura das imagens, a proporção entre as amostras utilizadas como teste e referência e a realização de *swap* ou não.

Experim.	**Quant. lâm.**	***Scanner***	***Teste/Ref***	***Swap***
Exp1/1	1	*Genepix*	1/1	não
Exp3/1	2	*Genepix*	3/1	sim
Exp6/1	2	*Genepix*	6/1	sim
Exp1/1-5/1	2	*Genepix*	1/1 e 5/1	sim
Exp1/1-2/1	2	*ScanArray*	1/1 e 2/1	sim
Exp1/1-10/1	2	*ScanArray*	1/1 e 10/1	sim

5.3 A ANÁLISE DE DADOS

Uma vez capturadas as imagens, os dados foram quantificados. Como pode ser visto na subseção 1.3.4, existem diferentes metodologias que podem ser empregadas para a definição das áreas ocupadas por sinal (*spots*) e quantificação dos valores de intensidade de sinal e *background*. Neste trabalho foram utilizados diferentes procedimentos sendo que a comparação destas diferentes metodologias constituem um ponto importante dos nossos objetivos.

5.3.1 Avaliação da segmentação de *spots*

A definição correta das áreas onde estão localizados os *spots* são de fundamental importância para garantir que problemas do tipo descrito na Figura 4 (A) não aconteçam. A Figura 10 mostra alguns *spots* localizados pelas metodologias utilizadas nesse trabalho, a parte (A) mostra o procedimento de círculo fixo, em (B) tem-se a segmentação empregada pelas técnicas *adap*, *circhist-50-50* e *circhist-100-20*, em (C) vê-se a segmentação empregada pelo método *circhist-30-10* e em (D) a metodologia de segmentação por morfologia matemática, empregada em *segment-50-50*, *segment-100-20* e *segment-100-100*. As imagens visualizadas são referentes a dois *spots* do experimento exp1/1, um deles apresentava baixa intensidade e outro com alta intensidade. Na figura é possível notar que a metodologia de segmentação por variação de intensidade é mais precisa que todas as demais. Embora pareça que o *spot* de baixa intensidade não tenha sido corretamente localizado, quando o contraste da imagem é aumentado nota-se que existe sinal de intensidade em toda a região delimitada.

Os procedimentos de localização dos *spots* também podem explicar diferenças significativas entre os valores de intensidade obtidos. Tais diferenças podem ser explicadas através do critério de seleção de *pixels* que são usados para o cálculo da intensidade do *spot*, assim como do *background*. A Tabela 9 lista os valores médios de intensidade e *background*, em *cy3*, para todos os cDNAs sem diluição (diluição 1) para o experimento exp1/1.

Também, a Figura 11 mostra os *box plots* dos dados referentes a esse mesmo experimento, dando uma idéia da dispersão dos valores obtidos. Entretanto, aqui, apenas os valores do gene Q estão representados. Em (A) e (B) tem-se os valores de intensidade de sinal e *background*, respectivamente, do experimento exp1/1. No

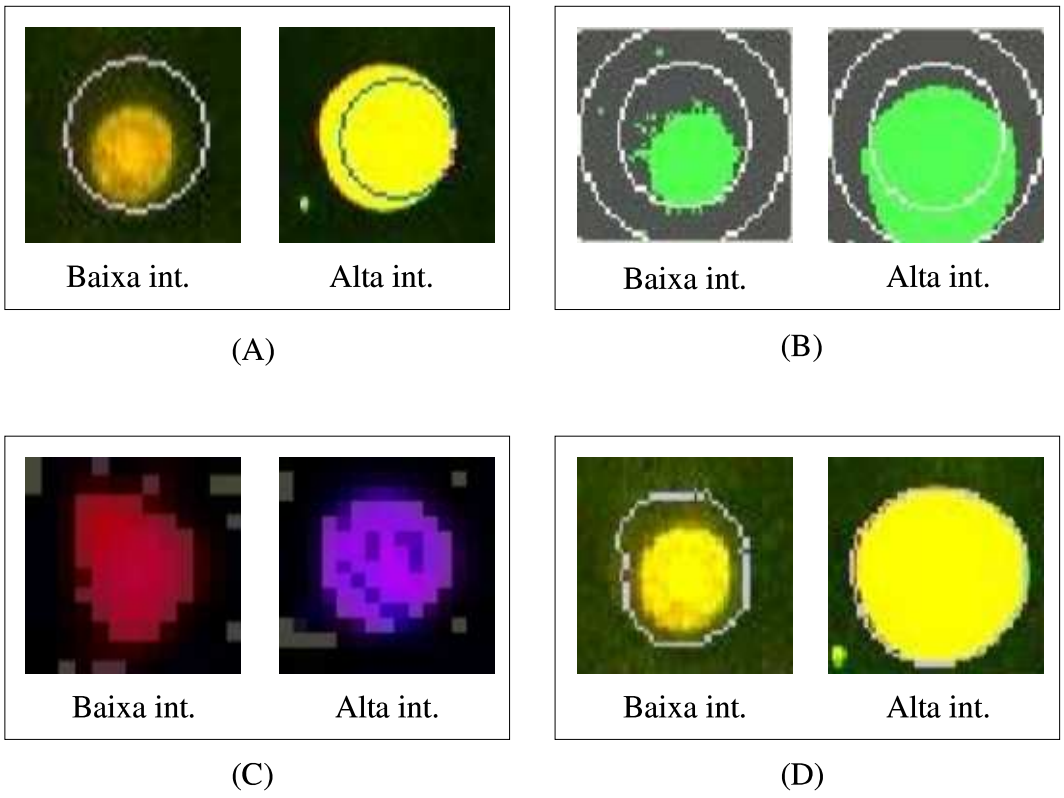


Figura 10: A capacidade de definição de *spots*.

As diferenças de definição de *spots* entre as metodologias distintas utilizadas neste trabalho. (A) metodologia de círculo fixo (empregada em *circfix*). (B) metodologia de segmentação empregada por *adap*, *circhist-50-50* e *circhist-100-20*. (C) segmentação empregada pela metodologia *circhist-30-10*. (D) segmentação por morfologia matemática (*segment-50-50*, *segment-100-20*, *segment-100-100*).

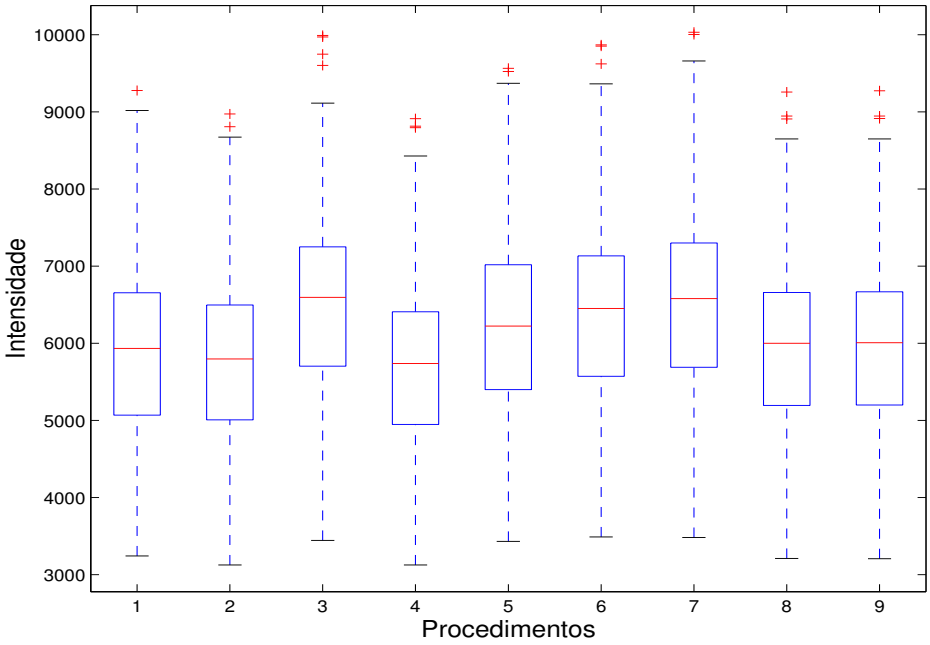
Tabela 9: Comparação entre os valores de intensidade de sinal e *background*.

Valores médios de sinal e *background* de todos os cDNAs (com diluição um), utilizados no trabalho. Esses dados são referentes ao experimento exp1/1.

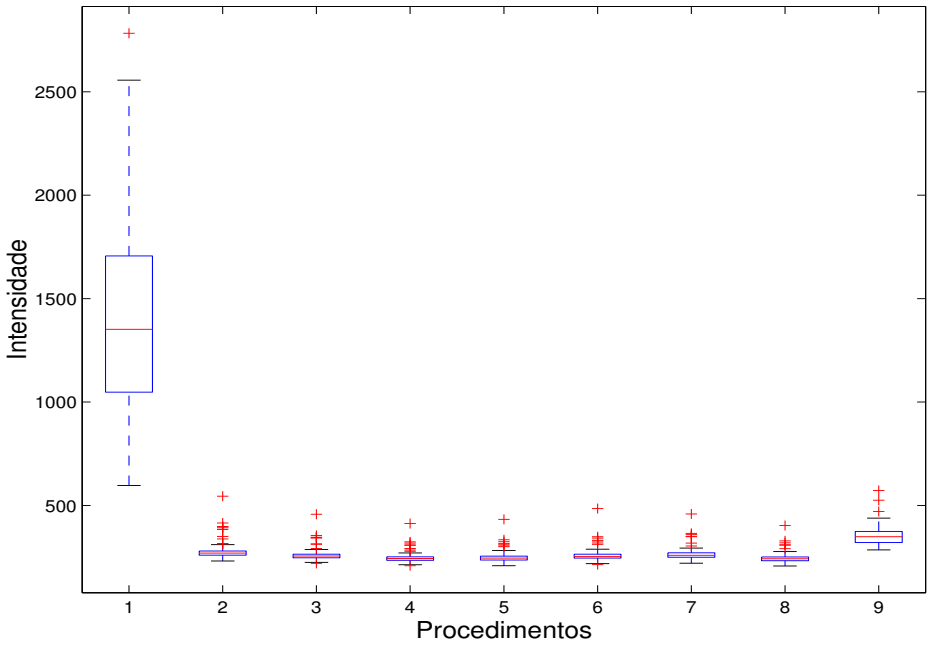
Proced.	**Tipo de Sinal**	**exp1/1**					
		LysA	**TrpC**	**Gene Q**	**ST0280**	**Il-6**	**Irf-1**
Circfix	Sinal	3827.68	922.20	5950.84	1311.48	11251.84	7878.76
	Back.	594.60	381.05	1411.53	451.05	1926.86	1727.77
Adap	Sinal	3653.88	983.77	5826.70	1335.65	10658.03	7659.67
	Back.	274.20	277.53	280.03	271.75	306.26	289.01
Circhist-50-50	Sinal	4372.11	1086.72	6563.72	1449.86	12321.87	8633.39
	Back.	249.67	255.11	262.00	253.46	256.19	256.11
Circhist-100-20	Sinal	3571.74	923.81	5728.85	1318.77	10778.16	7702.72
	Back.	236.26	241.78	247.82	238.72	242.24	242.39
Circhist-30-10	Sinal	4068.06	1150.36	6283.08	1393.74	11671.44	8268.02
	Back.	237.04	242.62	250.91	240.24	245.02	245.31
Hist-15-15	Sinal	4246.92	1030.07	6458.45	1420.85	12113.32	8456.73
	Back.	244.10	250.30	259.69	248.62	253.20	254.08
Segment-50-50	Sinal	4384.90	1084.47	6578.73	1445.44	12256.92	8582.32
	Back.	251.07	257.11	265.26	256.16	259.28	260.14
Segment-100-20	Sinal	3715.59	938.69	5981.99	1349.31	11041.72	7886.38
	Back.	233.73	240.22	247.05	237.05	240.98	241.68
Segment-100-100	Sinal	3715.18	938.61	5983.55	1349.36	11039.19	7882.61
	Back.	299.32	282.61	353.86	294.89	383.86	368.36

eixo das abscissas encontra-se os procedimentos de quantificação de sinal na mesma ordem encontrada na subseção 4.2.4, ou seja, 1 - *circfix*, 2 - *adap*, 3 - *circhist-50-50*, 4 - *circhist-100-20*, 5 - *circhist-30-10*, 6 - *hist-15-15*, 7 - *segment-50-50*, 8 - *segment-100-20*, 9 - *segment-100-100*.

A Figura 11 (A) mostra que não existem diferenças muito significativas nos valores de intensidade de sinal obtidos pelas diferentes metodologias utilizadas, dado que pode ser confirmado através da Tabela 9. Entretanto para os valores de intensidade do *background* a metodologia *circfix* quantifica valores maiores que os demais procedimentos, ver Tabela 9 e Figura 11 (B). Isso pode ser justificado pelo fato de que esse método de quantificação calcula a média de todos os *pixels* que estão na região do *background*, ao passo que as demais metodologias selecionam uma certa proporção de *pixels* com base no histograma da distribuição dos valores de intensidade que estejam na porção inferior da distribuição, ou seja, com baixa intensidade. Além disso, a Tabela 9 mostra que os valores de *background* calculados pelo procedimento *circfix* aumentam a medida que os valores de sinal também aumentam. Nota-se que os *spots* com intensidade de sinal muito fortes ocupam, visualmente, uma área maior que os que apresentam sinais de intensidade mais fracos. Através da Figura 10 (A) é possível notar que muitas vezes o *spot* fica maior que a mascára que delimita a sua região, o que faz com que *pixels* do *spot* sejam computados como sendo pertencentes ao *background*.



(A)



(B)

Figura 11: *Box plot* dos valores de intensidade do gene Q.

Esta figura ilustra os *box plots* dos dados citados na Tabela 9 apenas para o gene Q sem diluição, no eixo x temos as diferentes metodologias: 1 - *circfix*, 2 - *adap*, 3 - *circhist-50-50*, 4 - *circhist-100-20*, 5 - *circhist-30-10*, 6 - *hist-15-15*, 7 - *segment-50-50*, 8 - *segment-100-20*, 9 - *segment-100-100*. A figura mostra os valores de intensidade de (A) - sinal e (B) - *background* para o experimento exp1/1, respectivamente.

5.3.2 Análise das razões

O experimento exp1/1, que foi hibridizado com concentrações iguais de amostras de teste e referência, foi normalizado por energia total, descrita pela Equação (2). Como este experimento devia apresentar razão média de aproximadamente um para todos os *spots*, espera-se que os dados devam se concentrar em torno de uma reta com inclinação de aproximadamente 45º, quando são *plotados* os dados do canal um (*cy3*) no eixo x e do canal dois (*cy5*) no eixo y (gráfico de dispersão, ou *scatter plot*). Isso realmente aconteceu como pode ser observado pela Figura 12, referente ao experimento exp1/1.

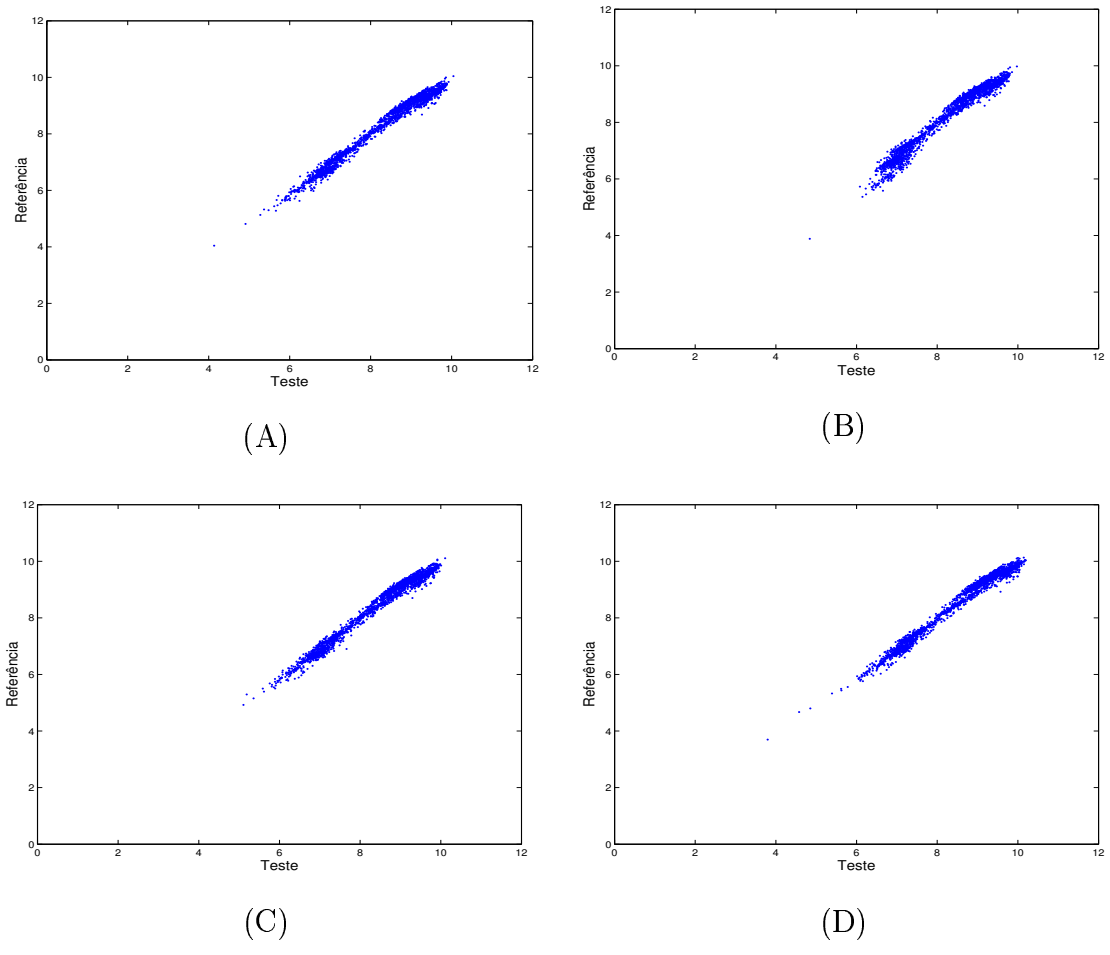


Figura 12: Gráficos de dispersão do experimento exp1/1.
Scatter plots referentes aos dados do experimento exp1/1. Esses dados foram quantificados através de 9 metodologias distintas, aqui estão representadas: (A) *segment-100-100*, (B) *adap*, (C) *circhist-30-10* e (D) *circhist-50-50*.

Uma outra forma de analisar esses dados é feita através dos gráficos de razão. Esses gráficos dispõem todos os *spots* no eixo das abscissas e seus respectivos valores de razão entre as amostras de teste e referência no eixo das ordenadas. A Figura 13 mostra esse tipo de gráfico para o experimento exp1/1. Note que o eixo y está em escala logarítmica nessa figura. De acordo com a parte (B) da figura nota-se que a metodologia *adap* apresenta maior variabilidade que as demais metodologias, sendo que todos os outros procedimentos apresentam perfis de dispersão muito parecidos entre si. Os padrões observados nas figuras 12 e 13 se conservam entre todas as metodologias utilizadas neste trabalho, por essa razão esses perfis foram ilustrados apenas para quatro metodologias.

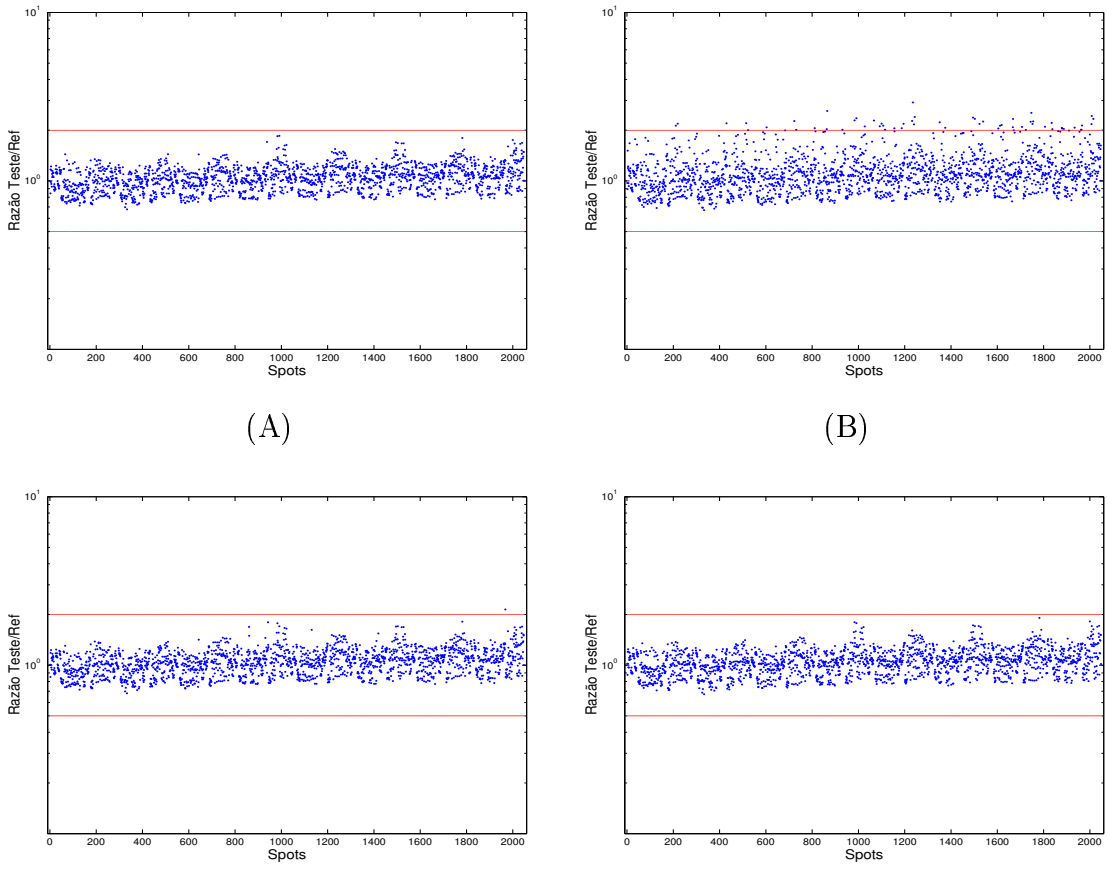


Figura 13: Gráficos de razão do experimento exp1/1.

Gráficos de razão do experimento exp1/1, cujas razões esperadas eram um para todos os fragmentos da lâmina. Estão representados os seguintes procedimentos: (A) *segment-100-100*, (B) *adap*, (C) *circhist-30-10* e (D) *circhist-50-50*.

Entretanto, em experimentos baseados somente na proporção teste/referência igual a um, a presença de vários *spots* que apresentam valores de intensidade muito próximos ao valor do *background* influencia para que esses *spots* apresentem valores de razão próximos de um, uma vez que os valores de intensidade do *background* em *cy3* e *cy5* são aproximadamente iguais. Visando contornar esse tipo de problema, foram desenhados outros experimentos onde foram variadas as concentrações dos mRNAs utilizados para a marcação do cDNA alvo. Os experimentos exp3/1 até exp1/1-10/1 foram construídos desta forma, as diferentes proporções de cDNA utilizados para hibridização podem ser verificadas na Tabela 8. As Figuras 14 e 15 mostram os *scatter plots* e gráficos de razões para o experimento exp3/1, onde esperamos razão igual a três, aproximadamente. As metodologias que não estão mostradas nas figuras apresentaram padrões de dispersão semelhantes às metodologias ilustradas, sendo que a metodologia *circfix* apresentou dispersão semelhante à da metodologia *segment-100-100*, e as demais apresentaram dados semelhantes aos dos procedimentos *circhist-30-10* e *segment-50-50*.

As Figuras 14 e 15 mostram menor variabilidade juntamente com uma maior exatidão nas metodologias *segment-100-100* e *circhist-50-50*. Nota-se, principalmente que a metodologia *adap* apresentou uma variabilidade muito alta para os fragmentos de TrpC e ST0280 (que são os fragmentos que apresentaram menores valores de intensidade de sinal), mostrando que este procedimento não é muito bom para *spots* que apresentam baixos valores de intensidade. É interessante dizer também que a técnica *segment-100-100* apresentou valores de razão mais precisos que as demais metodologias, note que os pontos azuis ficaram mais concentrados ao redor do grupo de pontos considerados bons, como mostra a Figura 15.

Os experimentos exp1/1-5/1, exp1/1-2/1 e exp1/1-10/1 foram construídos com dois grupos de proporções diferentes entre os cDNAs, onde os fragmentos de TrpC, ST0280 e Irf-1 mantiveram proporções iguais de material marcado como teste e referência sendo esperadas razões iguais a um, enquanto os fragmentos de LysA, gene Q e Il-6 tiveram proporções variáveis sendo esperadas razões diferentes de um. As Figuras 16 e 17 ilustram os *scatter plots* e gráficos de razão para o experimento exp1/1-5/1, que deveria apresentar razão cinco para os fragmentos cujas proporções foram variáveis. Novamente, os dados referentes às metodologias não apresentadas nas fi-

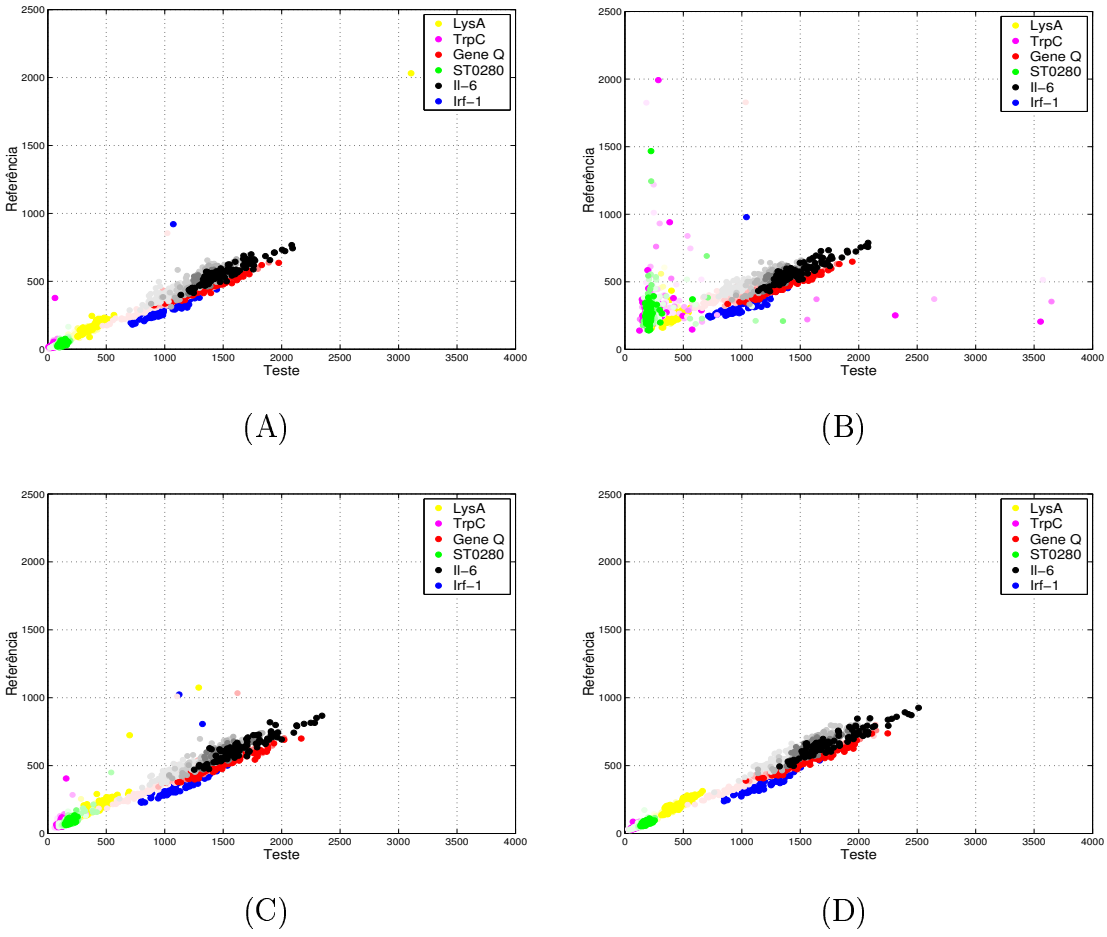


Figura 14: *Scatter plots* do experimento exp3/1.

Scatter plots do experimento exp3/1 que deveria apresentar razão três para todos os cDNAs. Os dados foram quantificados com 9 procedimentos diferentes. Aqui estão indicados: (A) *segment-100-100*, (B) *adap*, (C) *circhist-30-10* e (D) *circhist-50-50*. O esquema de cores utilizado ilustra os diferentes fragmentos de cDNA utilizados como indicado na legenda, o degradê de cada cor indica as diluições feitas para cada cDNA.

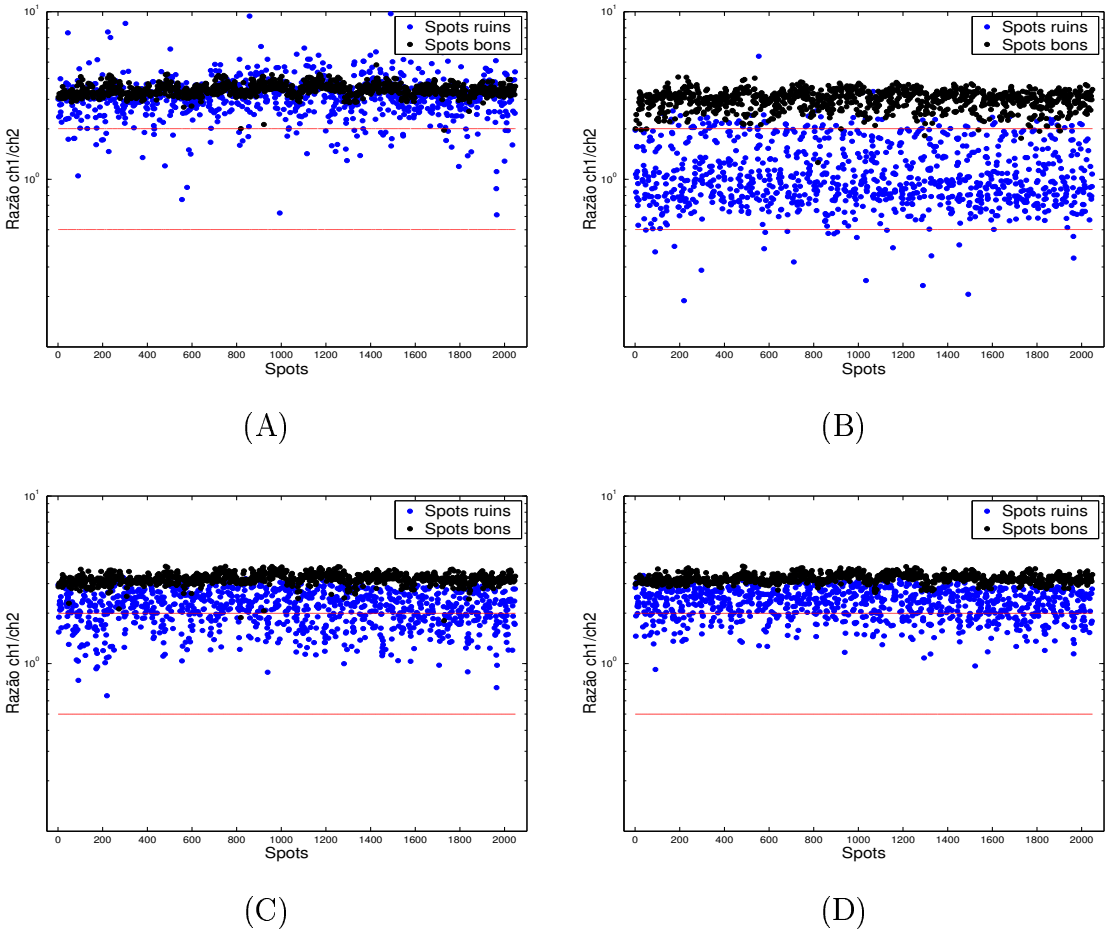


Figura 15: Gráficos de razão do experimento exp3/1.

Gráficos de razão do experimento exp3/1 que deveria apresentar razão três para todos os cDNAs. Os dados foram quantificados com 9 procedimentos diferentes. Aqui estão indicados: (A) *segment-100-100*, (B) *adap*, (C) *circhist-30-10* e (D) *circhist-50-50*. Os *spots* bons (indicados em preto na figura) são referentes aos fragmentos de gene Q, Il-6 e Irf-1, que apresentaram sinais de intensidade maiores que os demais cDNAs (que estão indicados na figura como *spots* ruins e coloridos em azul). Nota-se que os *spots* bons atingiram melhores resultados com menor variabilidade.

guras apresentaram perfis parecidos aos dados mostrados, de maneira semelhante ao experimento exp3/1, citado no parágrafo anterior.

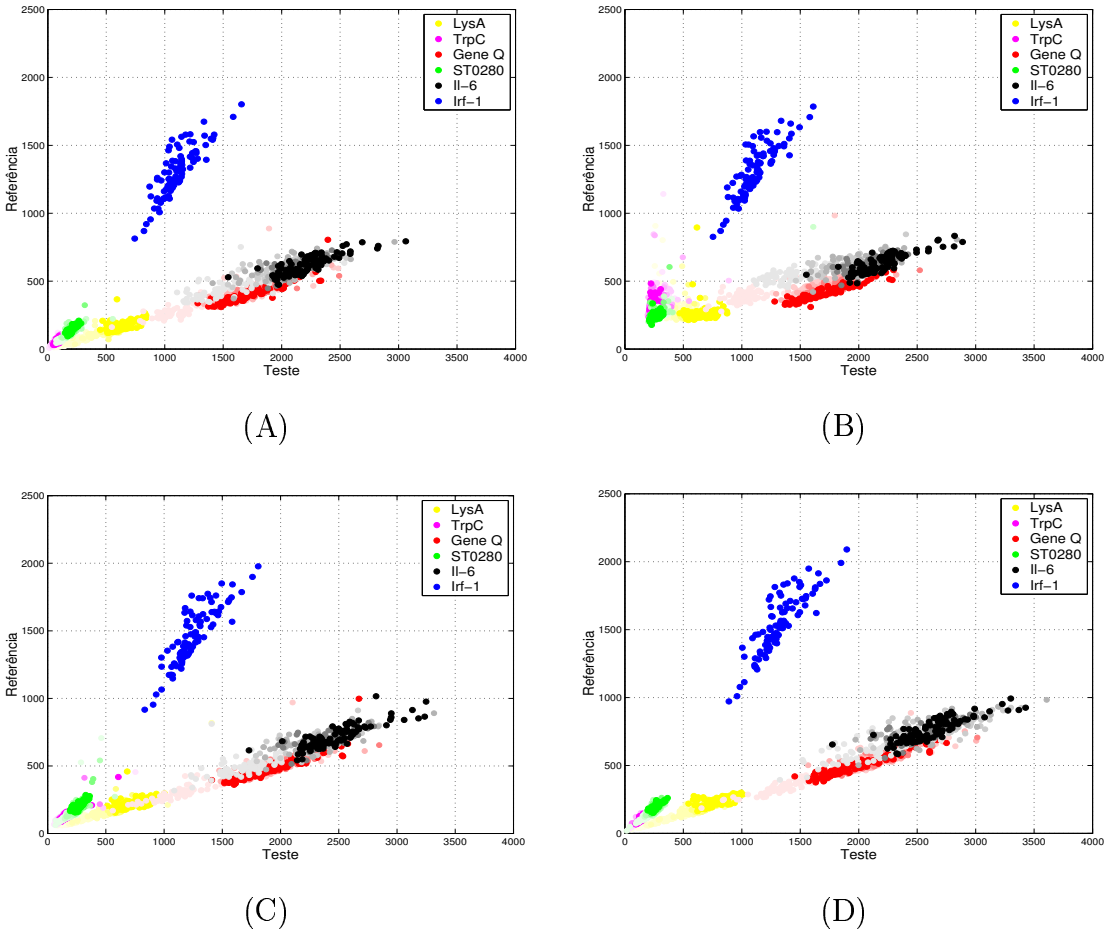


Figura 16: *Scatter plots* do experimento exp1/1-5/1.

Scatter plots do experimento exp1/1-5/1 que deveria apresentar razão um para os cDNAs de TrpC, ST0280 e Irf-1, e razão cinco para os outros fragmentos. Os dados foram quantificados com 9 procedimentos diferentes. Aqui estão indicados: (A) *segment-100-100*, (B) *adap*, (C) *circhist-30-10* e (D) *circhist-50-50*. O esquema de cores utilizado ilustra os diferentes fragmentos de cDNA utilizados como indicado na legenda, o degradê de cada cor indica as diluições feitas para cada cDNA.

Um fato que merece destaque são os maiores erros obtidos em todas as metodologias de análise de imagens para os fragmentos que apresentam valores de intensidade muito baixos. Na Figura 16 não é notada nenhuma diferença muito grande na variabilidade obtida pelas diferentes metodologias, ao contrário do que foi observado para a metodologia *adap* no experimento exp3/1 (Figura 14). Entretanto, a Figura 17 (B)

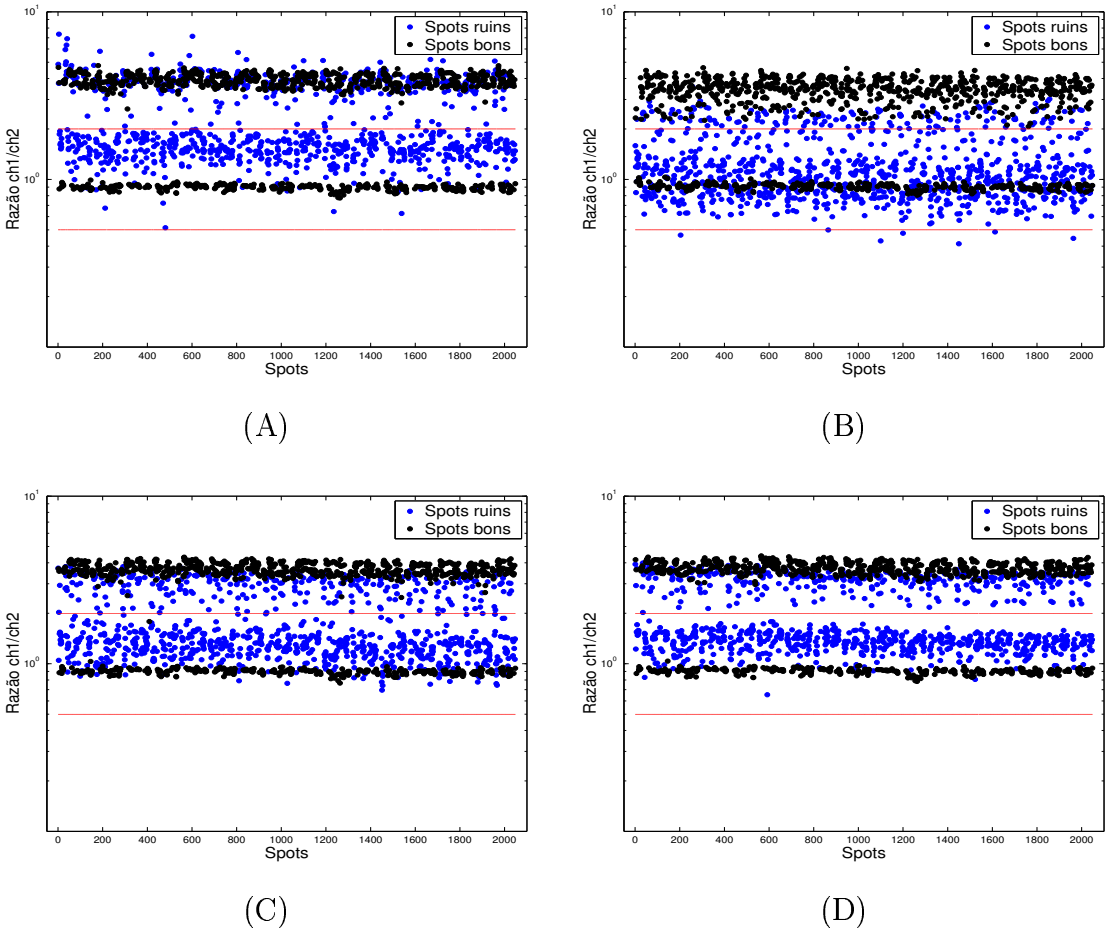


Figura 17: Gráficos de razão do experimento exp1/1-5/1.

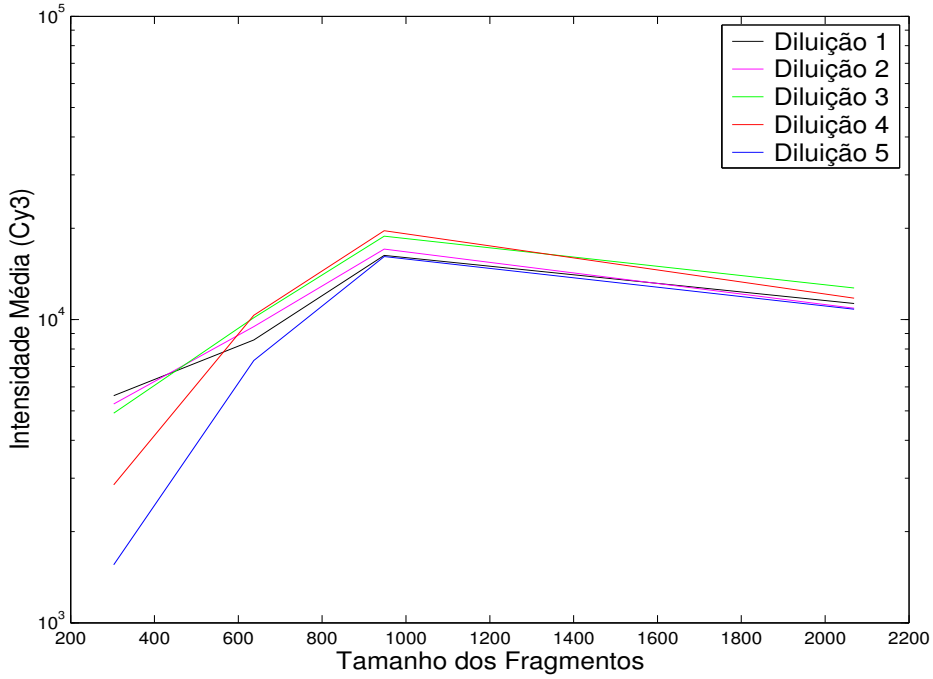
Gráficos de razão do experimento exp1/1-5/1, que deveria apresentar razão cinco para os cDNAs de LysA, gene Q e Il-6 e razão um para os demais fragmentos utilizados. Nesta figura são mostrados os dados referentes às metodologias: (A) *segment-100-100*, (B) *adap*, (C) *circhist30-10* e (D) *circhist-50-50*. Os *spots* bons (indicados em preto na figura) são referentes aos fragmentos do gene Q, Il-6 e Irf-1, que apresentaram sinais de intensidade maiores que os demais cDNAs (que estão indicados na figura como *spots* ruins e coloridos em azul). Nota-se que os *spots* bons atingiram melhores resultados com menor variabilidade.

mostra que o grupo de pontos coloridos em azul (*spots* ruins) apresentaram maior variabilidade que o grupo de pontos coloridos em preto (*spots* bons).

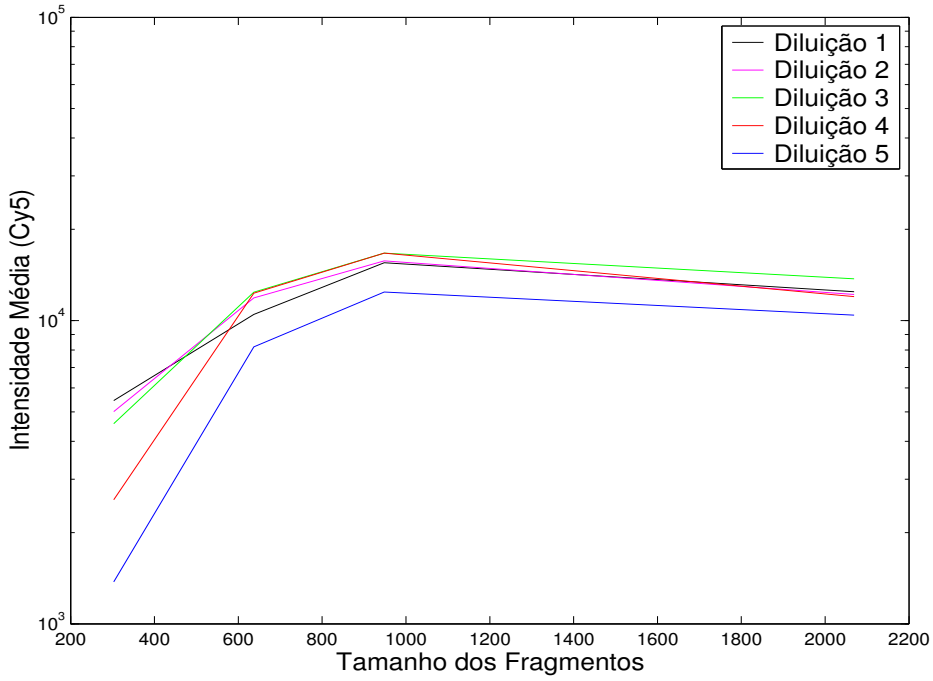
5.3.3 Efeito do tamanho do cDNA fixado

Para a construção das lâminas utilizadas neste projeto utilizamos fragmentos de cDNA de tamanhos variados que foram fixados em concentrações aproximadamente iguais. Assim, independentemente do procedimento utilizado para a quantificação dos dados, pôde-se também analisar o efeito do tamanho dos fragmentos fixados nos valores de intensidade observados. A Figura 18 mostra que esses valores de intensidade parecem decair de maneira mais acentuada em fragmentos menores que 650pb, aproximadamente. Essas figuras foram geradas a partir de dados do experimento exp1/1 quantificados através do procedimento segment-50-50, mas os perfis de queda de sinal observados aqui se conservam para as outras metodologias.

A Figura 18 mostra que a intensidade de sinal aumenta de maneira diretamente proporcional ao tamanho do cDNA fixado e atinge um platô em aproximadamente 650pb. Este dado indica que em insertos de até 650pb há uma dependência direta entre a intensidade de sinal e o tamanho do cDNA fixado. Porém, em insertos maiores esta dependência diminui sensivelmente, indicando que este deva ser um ponto importante de inflexão da curva; resultado que está de acordo com a literatura [42], que sugeria tal ponto de inflexão em 712pb, aproximadamente.



(A)



(B)

Figura 18: Gráfico de comparação dos diferentes tamanhos de fragmentos fixados.

Neste gráfico é possível notar que em fragmentos maiores que 650pb, aproximadamente, os valores de intensidade parecem atingir um certo platô. Esses dados são referentes ao experimento exp1/1 em (A) - *cy3* e (B) - *cy5*.

5.3.4 Dispersão das razões com as intensidades de sinal

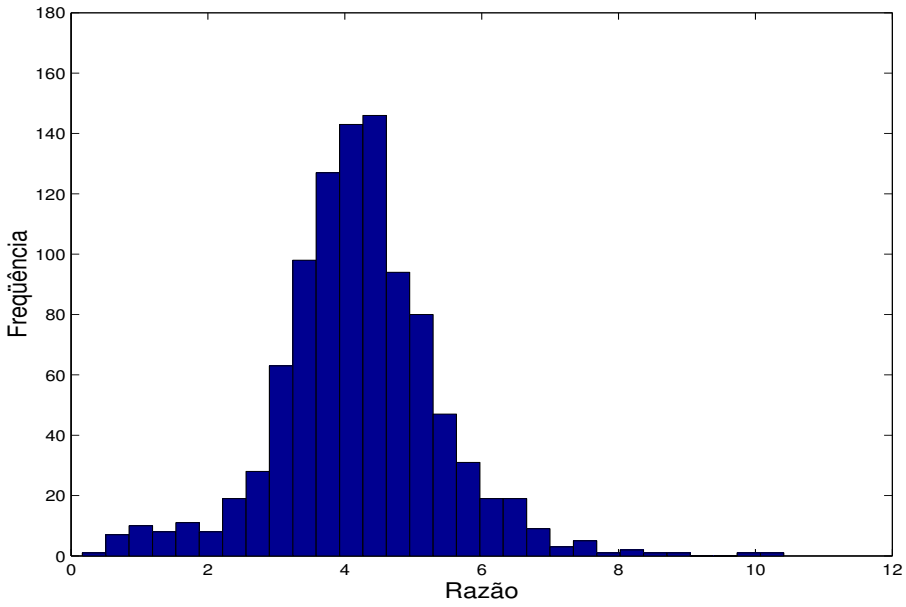
Analisando os experimentos onde a razão esperada difere de um notamos que a variabilidade dos dados é muito maior do que a que foi observada nos experimentos que foram construídos com a razão unitária. Mas, como podemos ver nas Figuras 15 e 17, existe um conjunto de pontos (pontos pretos) cuja razão fica muito próxima de 3 e 5, respectivamente. Entretanto, existe um outro grupo de dados, referenciados como *spots* ruins e coloridos em azul no gráfico, que se afastam muito da razão esperada. Isso pode ser explicado pelo fato de que os cDNAs de LysA, TrpC e ST0280 apresentaram sinais de intensidade mais baixos que os demais, o que pode ser verificado nas Figuras 14 e 16. Os fragmentos de *LysA* e *TrpC* tinham tamanho próximo de 300pb, e foi mostrado que esses fragmentos apresentam baixa intensidade (Figura 18), já o mRNA transcrito de ST0280 apesar de possuir 650pb (semelhante ao do gene Q) sempre apresentou intensidades de sinal próximas dos mRNAs menores. Isto pode ser devido à problemas na qualidade deste mRNA transcrito *in vitro*.

Observando mais detalhadamente esses *spots* com baixos valores de intensidade, nota-se que os valores de sinal dos *pixels* que constituem suas regiões apresentam uma variabilidade maior que os *spots* que apresentam valores de intensidade altos. A Figura 19 mostra os histogramas dos valores de razão calculados *pixel* à *pixel* para um *spot* de intensidade alta (A) e baixa (B), onde nota-se que o histograma do *spot* com baixa intensidade é mais "gordo" que o do *spot* que apresenta alta intensidade, mostrando que os *spots* que acendem com intensidade baixa apresentam uma maior dispersão em relação aos *spots* de intensidade maior. Essa maior dispersão atrapalha a estimação dos *pixels* a serem utilizados para o cálculo dos valores de intensidade.

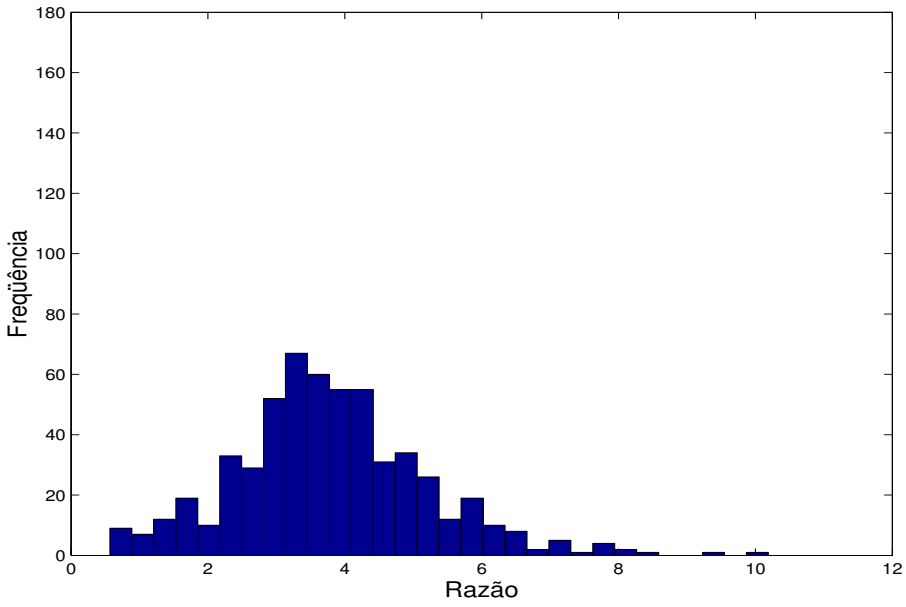
Uma conseqüência importante desse fato é mostrada a seguir. A Figura 20 mostra que a exatidão da razão obtida decai proporcionalmente aos valores de intensidade observados para quase todas as metodologias empregadas, com exceção das metodologias *circfix* e *segment-100-100* que apresentam uma exatidão maior, pagando o preço de apresentar um número maior de *outliers*, como mostra a figura 21.

Essas duas figuras (20 e 21) mostram *scatter plots* tridimensionais referentes ao experimento exp3/1, relacionando os valores de intensidade para *cy3* e *cy5* com os valores de razão calculados. Note que o gráfico está em escala logarítmica e que esse experimento deveria apresentar razão igual a três para todos os cDNAs utilizados.

Os pontos azuis e pretos são definidos de maneira semelhante à definição dada nas Figuras 15 e 17.



(A)



(B)

Figura 19: Histogramas de razão *pixel* à *pixel*.

Este gráfico ilustra os histogramas dos valores de razão *pixel* à *pixel* de dois *spots* diferentes, com intensidade de sinal (A) - alta e (B) - baixa. Esses *spots* são referentes ao experimento exp3/1, que deveriam apresentar razão 3.

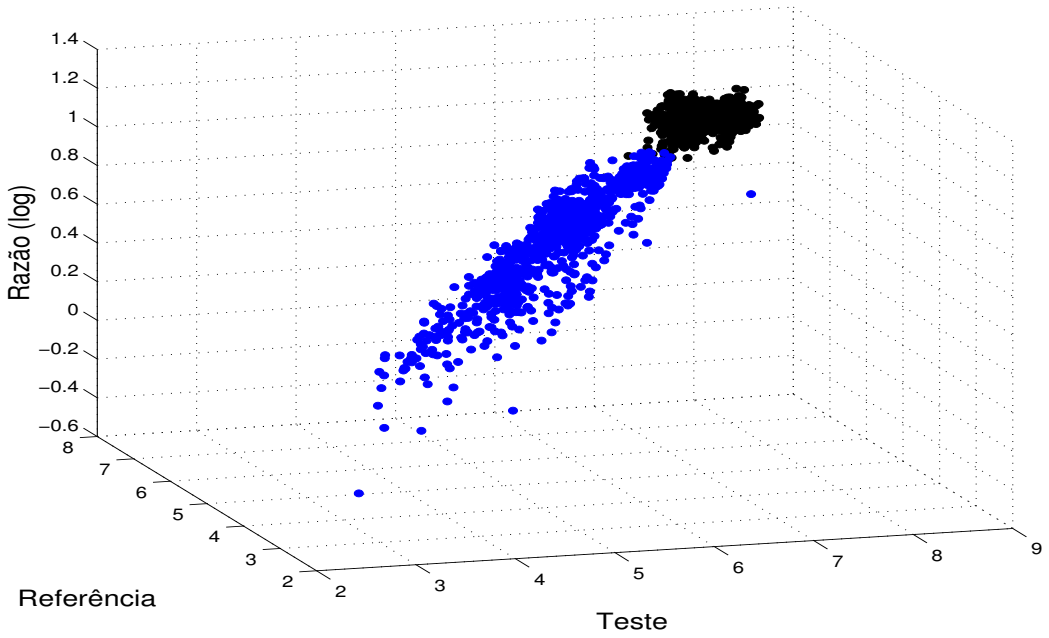


Figura 20: Variação da razão com a intensidade (metodologia *segment-50-50*).

Gráfico relacionando os valores de intensidade em *cy3* e *cy5* com os valores de razão observados. Esses dados são referentes ao experimento exp3/1 (com razão esperada igual a três) quantificados através do procedimento *segment-50-50*. Observando este gráfico é possível notar que existe uma forte dependência entre os valores de intensidade de sinal dos *spots* e os valores de razão calculados. À medida que os valores de intensidade de sinal decaem, a exatidão da razão medida também decai.

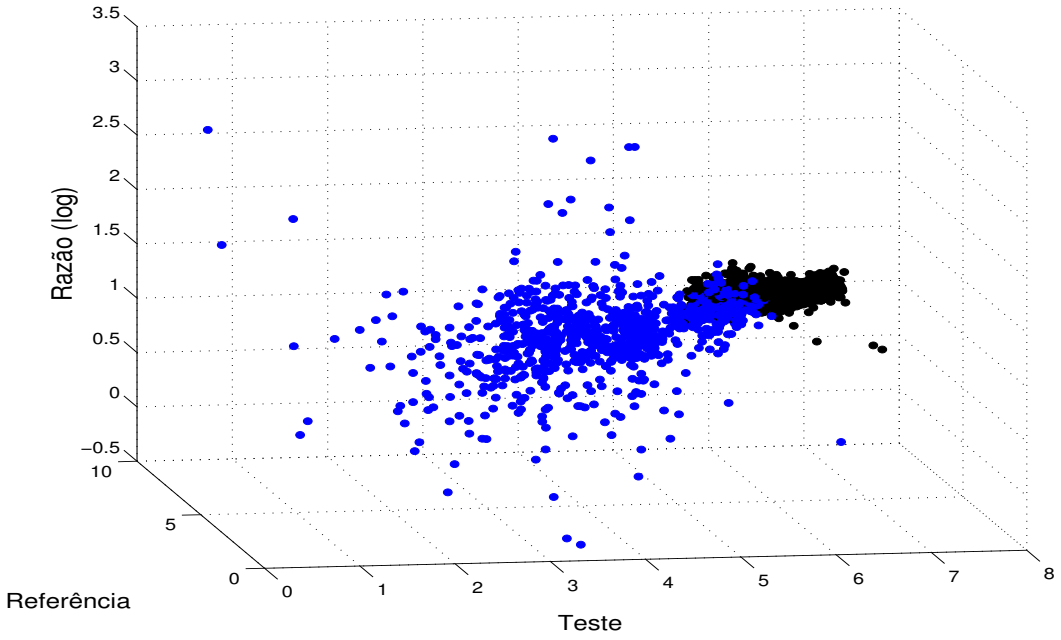


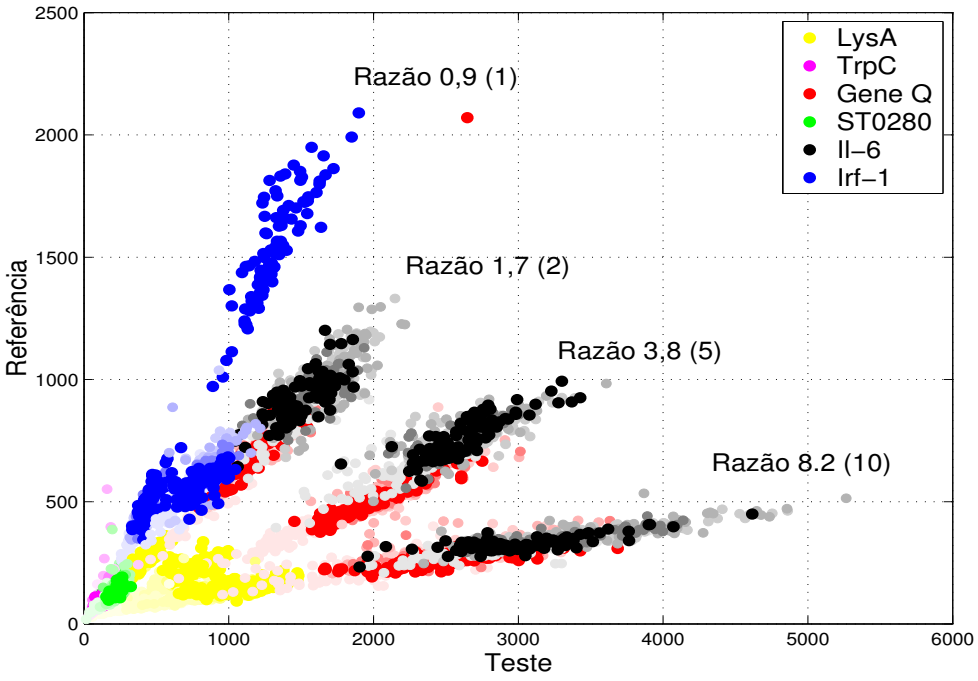
Figura 21: Variação da razão com a intensidade (metodologia *segment-100-100*).

Gráfico relacionando os valores de intensidade em *cy3* e *cy5* com os valores de razão observados. Esses dados são referentes ao experimento exp3/1 (com razão esperada igual a três) quantificados através do procedimento *segment-100-100*. Observando este gráfico é possível notar que aquela forte dependência observada através da metodologia *segment-50-50* (Figura 20) tende a diminuir. Entretanto aqui nota-se vários *outliers* dos valores de razão obtidos.

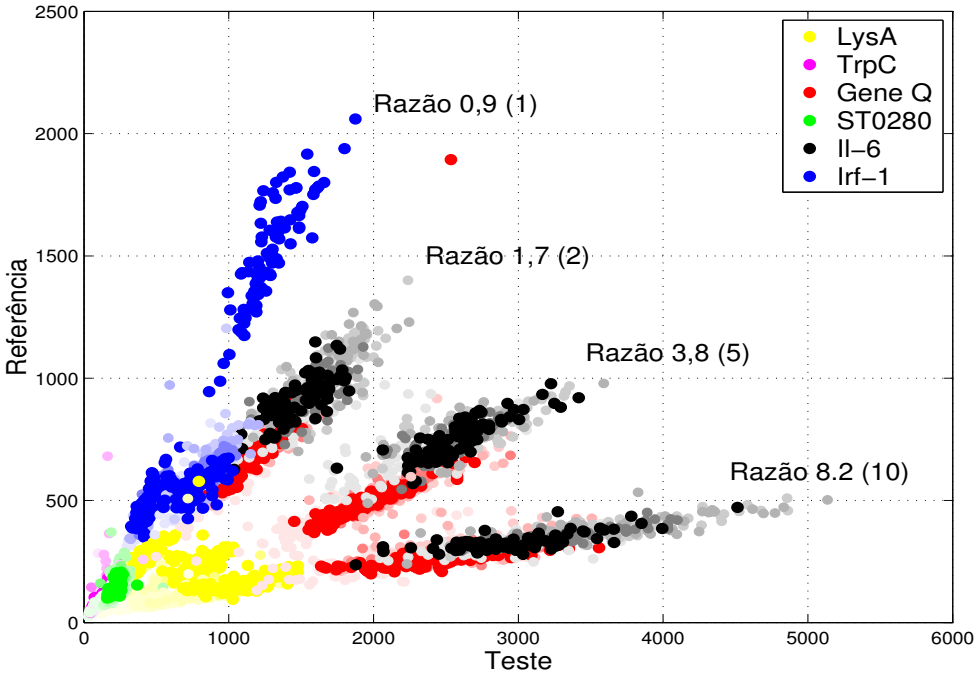
5.3.5 Análise dos erros cometidos

Os experimentos exp1/1-5/1, exp1/1-2/1 e exp1/1-10/1, nos quais se pretendia avaliar as diferenças entre as metodologias de localização e quantificação de *spots*, tiveram grande variação das razões esperadas entre as amostras de teste e referência (note que as razões variaram de 1 até 10). Assim, ao se olhar de uma forma mais crítica para esses dados, temos uma idéia da precisão da metodologia de cDNA *microarray* na detecção de diferenças em níveis de expressão de RNA mensageiros entre duas populações celulares distintas. A Figura 22 mescla os dados desses três experimentos com o intuito de analisar a exatidão da tecnologia de *arrays*, onde se pode notar que as diferenças entre as proporções de *cy3* e *cy5* foram detectadas satisfatoriamente (para as metodologias *circhist-50-50* em (A) e *segment-50-50* em (B)). Nestes gráficos estão destacados, aproximadamente, a razão obtida após a auto-normalização e a razão esperada entre parênteses. O degradê de cores para cada cDNA distinto se refere às diferentes diluições feitas para o material fixado.

Nos experimentos construídos neste trabalho foram utilizadas cinco diluições dos cDNAs que foram fixados nas lâminas. Nos fragmentos que apresentaram valores de intensidade altos (Il-6 e Irf-1) não foi notada muita diferença nos valores médios de sinal entre essas diferentes diluições. Entretanto para os fragmentos que apresentaram intensidade de sinal baixa (LysA, TrpC e ST0280) notamos que tais diluições tem um efeito mais drástico, como pode ser observado na Figura 18. Por outro lado, foi mostrado aqui que existe uma dependência entre os valores de intensidade dos *spots* e os valores de razão calculados, ver Figuras 20 e 21. Assim, nota-se que a análise dos fragmentos que apresentam baixos valores de intensidade vão apresentar maiores erros que a análise dos mesmos fragmentos na concentração inicial. Esse fato também é esperado para as diferentes diluições encontradas nas lâminas, principalmente para os cDNAs de tamanho pequeno (LysA e TrpC). Desta forma, as análises de erro apresentadas aqui foram feitas apenas para os cDNAs que não apresentavam diluição (diluição 1), para evitar superestimação dos erros e desvios-padrão obtidos. A seguir são apresentadas algumas tabelas e figuras que resumem, numericamente, os resultados obtidos neste trabalho para os cDNAs que foram fixados na concentração inicial (diluição um), com excessão da Tabela 10 e da Figura 23 que apresentam os dados referentes ao experimento exp1/1 para os cDNAs que apresentavam diluição cinco. Nesses gráficos são apresentados os valores de erro encontrados para todos os



(A)



(B)

Figura 22: Validação da metodologia de cDNA *microarray*.
Dados dos experimentos exp1/1-5/1, exp1/1-2/1 e exp1/1-10/1. Quantificados com (A) *circhist-50-50* e (B) *segment-50-50*. Observa-se que os valores de razão obtidos para fragmentos de intensidade alta ficam muito próximos dos valores esperados (indicados entre parênteses). O degradê observado para cada cDNA se refere às diluições utilizados no cDNA da sonda.

cDNAs, que estão indicados no eixo x na seguinte ordem: 1 - LysA, 2 - TrpC, 3 - Gene Q, 4 - ST0280, 5 - Il-6, 6 - Irf-1. Quando comparamos a Tabela 10 e a Figura 23 com a Tabela 11 e a Figura 24, é possível notar as diferenças entre os erros cometidos para os cDNAs com diluição um e com diluição cinco, especialmente para os fragmentos LysA, TrpC e ST0280, que apresentaram valores de intensidade menores que os demais.

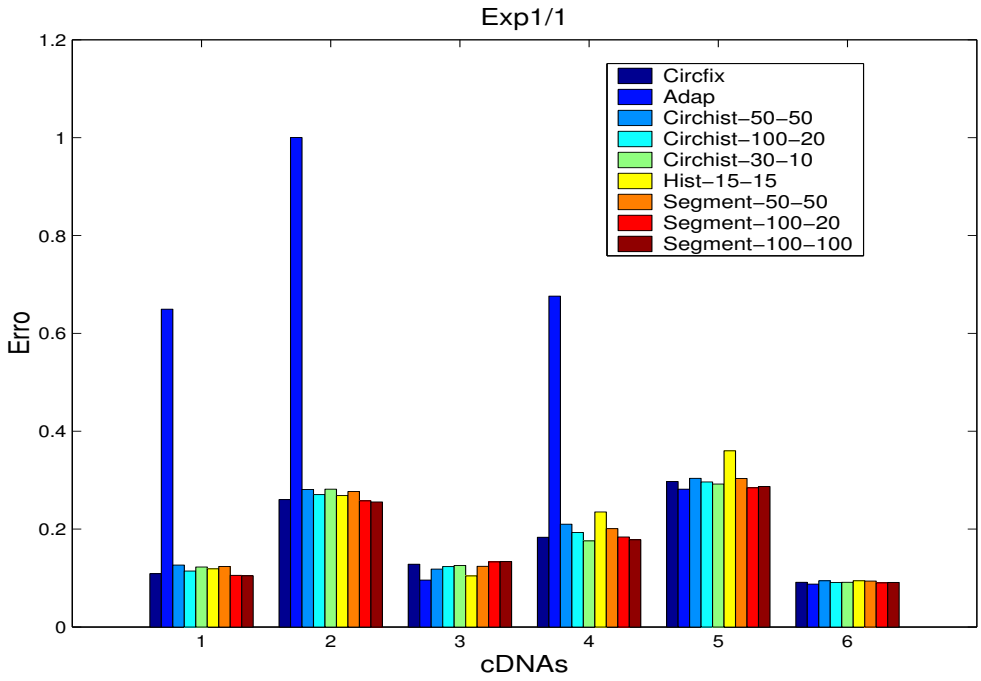


Figura 23: Erros cometidos no experimento exp1/1 (com diluição 5). Erros cometidos pelas diferentes metodologias empregadas para a quantificação dos experimentos. Nesta figura são mostrados os erros cometidos para *spots* de todos os cDNAs utilizados com diluição cinco no experimento exp1/1. Os cDNAs estão indicados no eixo x na seguinte ordem: 1 - LysA, 2 - TrpC, 3 - Gene Q, 4 - ST0280, 5 - Il-6, 6 - Irf-1.

O experimento exp1/1 deveria apresentar razão entre as amostras de teste e referência igual a um. Quando os dados referentes a esse experimento foram analisados, foi notado que não existiram diferenças muito significativas entre os erros cometidos por quase todas as metodologias empregadas referentes à todos os cDNAs utilizados, ver Tabelas 10 e 11 e Figuras 23 e 24. Entretanto nota-se que a metodologia *adap* apresenta erros bastante superiores para os fragmentos que apresentaram valores de intensidade muito baixos (LysA, TrpC e ST0280). Entranto experimentos baseados

Tabela 10: Dados obtidos para o experimento exp1/1 (diluição cinco).

Médias, desvios padrão e erros obtidos para todos os fragmentos da sonda que apresentam diluição cinco no experimento exp1/1. Neste experimento é esperada razão um para todos os cDNAs.

Experimento exp1/1, com diluição cinco																		
Soft	**LysA**			**TrpC**			**Gene Q**			**ST0280**			**Il6**			**Irf1**		
	μ	**E**	**DP**	μ	**E**	**DP**	μ	**E**	**DP**	μ	**E**	**DP**	μ	**E**	**DP**	μ	**E**	**DP**
Circfix	1.09	0.11	0.12	1.25	0.26	0.18	0.88	0.13	0.08	1.16	0.18	0.19	1.30	0.30	0.16	1.03	0.09	0.13
Adap	1.65	0.65	0.29	2.00	1.00	0.30	0.93	0.10	0.08	1.67	0.68	0.30	1.28	0.28	0.16	1.04	0.09	0.12
Circhist-50-50	1.12	0.13	0.12	1.28	0.28	0.17	0.90	0.12	0.08	1.20	0.21	0.19	1.30	0.30	0.17	1.03	0.09	0.13
Circhist-100-20	1.11	0.11	0.12	1.27	0.27	0.16	0.89	0.12	0.08	1.18	0.19	0.18	1.30	0.30	0.16	1.03	0.09	0.13
Circhist-30-10	1.11	0.12	0.13	1.28	0.28	0.18	0.88	0.13	0.08	1.12	0.18	0.19	1.29	0.29	0.16	1.03	0.09	0.13
Hist-15-15	1.11	0.12	0.12	1.27	0.27	0.16	0.92	0.10	0.08	1.23	0.23	0.19	1.36	0.36	0.17	1.04	0.09	0.13
Segment-50-50	1.12	0.12	0.12	1.28	0.28	0.16	0.89	0.12	0.09	1.19	0.20	0.20	1.30	0.30	0.17	1.03	0.09	0.13
Segment-100-20	1.09	0.11	0.12	1.26	0.26	0.16	0.88	0.13	0.08	1.17	0.18	0.18	1.28	0.28	0.16	1.03	0.09	0.13
Segment-100-100	1.09	0.10	0.12	1.26	0.26	0.17	0.88	0.13	0.08	1.15	0.18	0.19	1.29	0.29	0.16	1.03	0.09	0.13

Tabela 11: Dados obtidos para o experimento exp1/1.

Médias, desvios padrão e erros obtidos para todos os fragmentos com diluição um da sonda no experimento exp1/1. Neste experimento é esperada razão um para todos os cDNAs.

Experimento exp1/1, com diluição um																		
Soft	**LysA**			**TrpC**			**Gene Q**			**ST0280**			**Il6**			**Irf1**		
	μ	**E**	**DP**	μ	**E**	**DP**	μ	**E**	**DP**	μ	**E**	**DP**	μ	**E**	**DP**	μ	**E**	**DP**
Circfix	1.02	0.05	0.06	1.22	0.22	0.07	0.80	0.20	0.05	1.02	0.06	0.07	1.07	0.07	0.05	0.89	0.12	0.07
Adap	1.00	0.04	0.06	1.27	0.27	0.18	0.79	0.21	0.05	1.02	0.06	0.07	1.05	0.07	0.06	0.88	0.12	0.07
Circhist-50-50	1.02	0.05	0.05	1.22	0.22	0.07	0.81	0.19	0.04	1.02	0.06	0.07	1.05	0.06	0.05	0.89	0.11	0.06
Circhist-100-20	1.01	0.04	0.06	1.21	0.21	0.07	0.80	0.20	0.04	1.02	0.06	0.07	1.06	0.07	0.05	0.89	0.12	0.07
Circhist-30-10	1.01	0.05	0.06	1.20	0.21	0.10	0.81	0.19	0.04	1.03	0.06	0.07	1.06	0.07	0.05	0.89	0.11	0.07
Hist-15-15	1.02	0.04	0.06	1.22	0.22	0.08	0.81	0.19	0.04	1.02	0.06	0.06	1.04	0.05	0.05	0.89	0.12	0.06
Segment-50-50	1.02	0.05	0.05	1.23	0.23	0.07	0.81	0.19	0.04	1.02	0.06	0.06	1.05	0.06	0.05	0.90	0.11	0.06
Segment-100-20	1.01	0.04	0.06	1.21	0.21	0.07	0.81	0.19	0.04	1.02	0.06	0.07	1.07	0.08	0.05	0.90	0.11	0.07
Segment-100-100	1.01	0.05	0.06	1.21	0.21	0.07	0.81	0.19	0.04	1.02	0.06	0.07	1.07	0.08	0.05	0.90	0.11	0.07

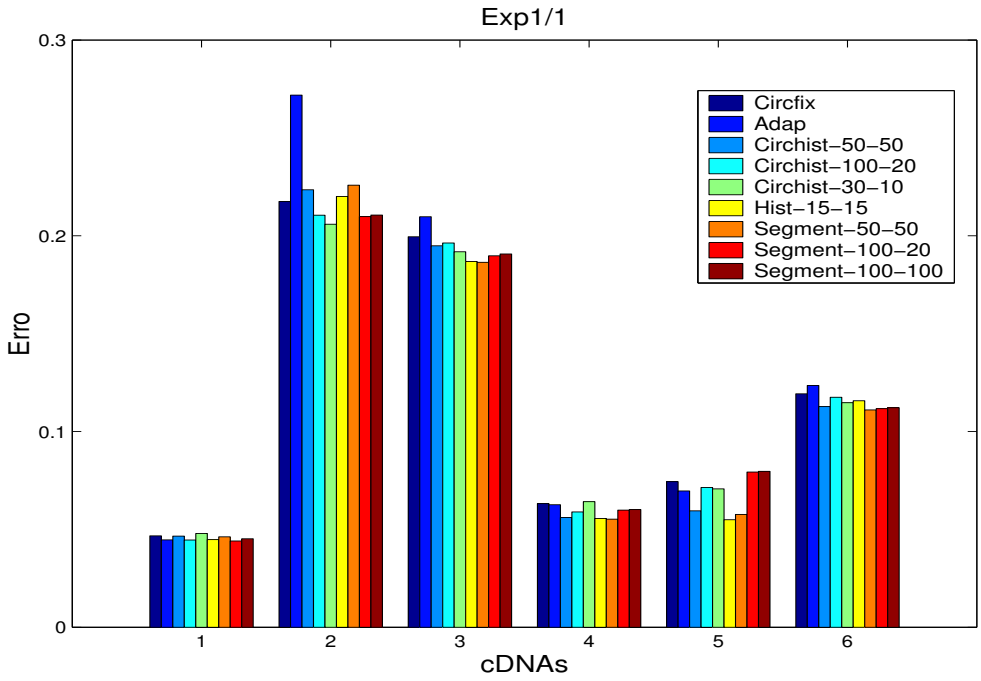


Figura 24: Erros cometidos no experimento exp1/1.

Esta figura ilustra os erros cometidos pelas diferentes metodologias de quantificação para o experimento exp1/1. Os cDNAs estão indicados no eixo x na seguinte ordem: 1 - LysA, 2 - TrpC, 3 - Gene Q, 4 - ST0280, 5 - Il-6, 6 - Irf-1.

somente em razão entre teste e referência igual a um podem mascarar resultados importantes. Assim, foram desenhados outros experimentos cujas razões esperadas eram diferentes de um.

Inicialmente foi construído um experimento no qual a proporção dos mRNAs utilizados como amostras de teste e referência foi de 3/1 (exp3/1). A Figura 25 (A) mostra os erros cometidos neste experimento onde nota-se que, para o fragmento LysA as metodologias *circhist-50-50* e *segment-50-50* apresentaram erros menores, *adap* apresentou o maior erro e as demais metodologias ficaram aproximadamente equivalentes. Para o fragmento de TrpC a metodologia *circhist-50-50* apresentou menor erro que todas as demais, sendo que *circfix* e *adap* apresentaram os maiores erros. Os fragmentos do gene Q, Il-6 e Irf-1 em geral apresentaram um pequeno aumento do erro cometido pelas metodologias *circfix* e *segment-100-100*, se mantendo estáveis nas demais metodologias. O fragmento de ST0280 apresentou os maiores erros nas metodologias *adap* e *segment-100-100*. Entretanto, observando a tabela 12 é notado que nos fragmentos do gene Q, Il-6 e Irf-1, que apresentam os sinais de intensidade mais fortes, nota-se que todas as metodologias calcularam corretamente

Tabela 12: Dados obtidos para o experimento exp3/1.

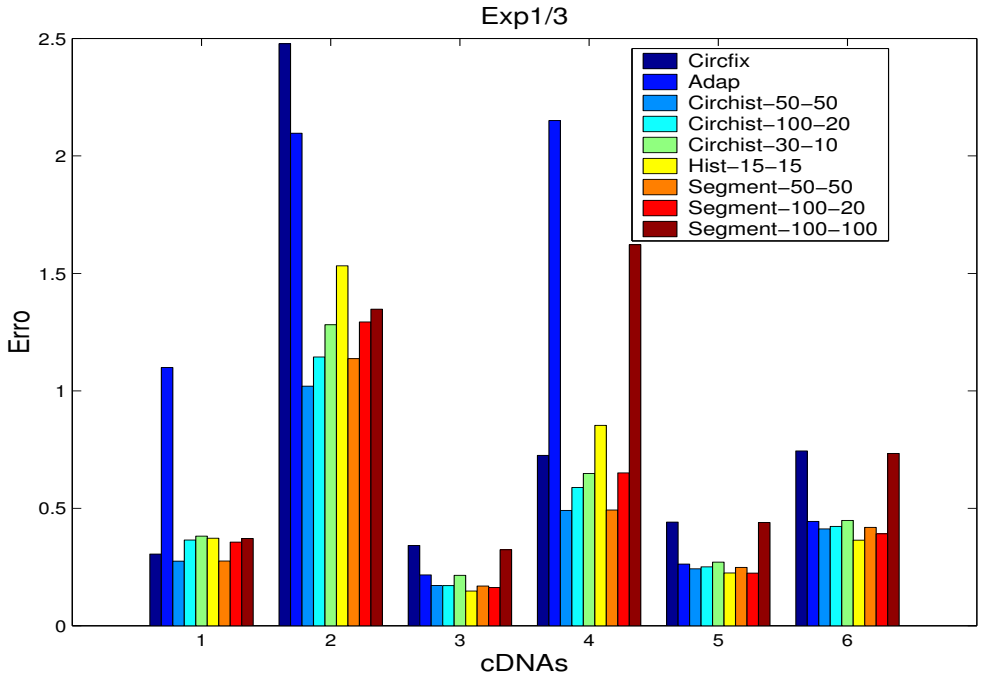
Médias, desvios padrão e erros obtidos para todos os fragmentos com diluição um da sonda no experimento exp3/1. Neste experimento é esperada razão três para todos os cDNAs.

Experimento exp3/1, com diluição um																		
Soft	**LysA**			**TrpC**			**Gene Q**			**ST0280**			**Il6**			**Irf1**		
	μ	**E**	**DP**	μ	**E**	**DP**	μ	**E**	**DP**	μ	**E**	**DP**	μ	**E**	**DP**	μ	**E**	**DP**
Circfix	3.16	0.31	0.38	4.93	2.48	8.37	3.34	0.34	0.18	3.61	0.73	0.75	3.44	0.44	0.24	3.72	0.74	0.31
Adap	1.90	1.10	0.38	0.99	2.10	0.72	3.19	0.22	0.17	0.85	2.15	0.22	3.24	0.26	0.21	3.41	0.44	0.29
Circhist-50-50	2.75	0.27	0.25	1.98	1.02	0.29	3.15	0.17	0.15	2.53	0.49	0.29	3.22	0.24	0.19	3.41	0.41	0.20
Circhist-100-20	2.64	0.36	0.27	1.86	1.14	0.31	3.14	0.17	0.14	2.42	0.59	0.31	3.22	0.25	0.20	3.39	0.42	0.26
Circhist-30-10	2.65	0.38	0.38	1.72	1.28	0.41	3.19	0.21	0.17	2.36	0.65	0.37	3.26	0.27	0.19	3.41	0.45	0.26
Hist-15-15	2.64	0.37	0.27	1.47	1.53	0.22	3.12	0.15	0.13	2.15	0.85	0.26	3.19	0.22	0.19	3.36	0.36	0.19
Segment-50-50	2.75	0.28	0.25	1.86	1.14	0.36	3.15	0.17	0.14	2.52	0.49	0.30	3.23	0.25	0.19	3.41	0.42	0.20
Segment-100-20	2.65	0.36	0.27	1.71	1.29	0.33	3.14	0.16	0.13	2.35	0.65	0.30	3.20	0.22	0.18	3.36	0.39	0.24
Segment-100-100	3.23	0.37	0.47	3.41	1.35	1.97	3.32	0.32	0.18	4.44	1.62	2.91	3.44	0.44	0.23	3.71	0.73	0.31

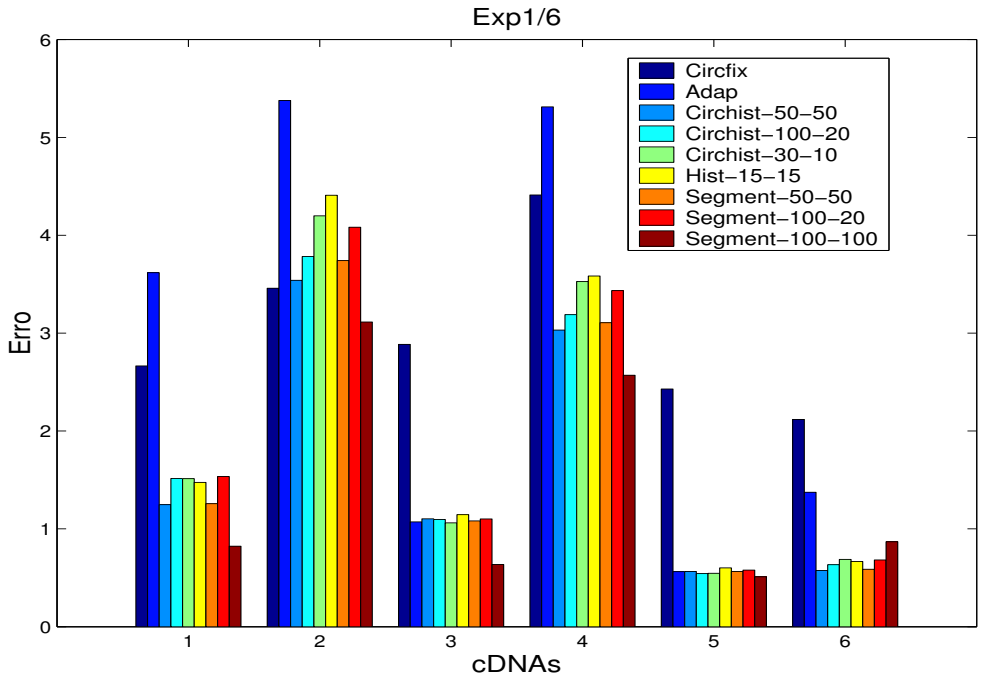
Tabela 13: Dados obtidos para o experimento exp6/1.

Médias, desvios padrão e erros obtidos para todos os fragmentos com diluição um da sonda no experimento exp6/1. Neste experimento é esperada razão seis para todos os cDNAs.

Experimento exp6/1, com diluição um																		
Soft	**LysA**			**TrpC**			**Gene Q**			**ST0280**			**Il6**			**Irf1**		
	μ	**E**	**DP**	μ	**E**	**DP**	μ	**E**	**DP**	μ	**E**	**DP**	μ	**E**	**DP**	μ	**E**	**DP**
Circfix	3.33	2.67	0.57	2.78	3.46	1.90	3.11	2.89	0.27	4.08	4.41	8.74	3.57	2.43	0.40	3.88	2.12	0.58
Adap	2.38	3.62	0.61	0.62	5.38	0.24	4.93	1.07	0.38	0.69	5.31	0.29	5.59	0.56	0.52	4.65	1.37	0.64
Circhist-50-50	4.75	1.25	0.59	2.46	3.54	0.85	4.90	1.10	0.36	2.97	3.03	0.65	5.56	0.56	0.52	5.67	0.57	0.61
Circhist-100-20	4.49	1.51	0.66	2.22	3.78	0.82	4.90	1.10	0.36	2.81	3.19	0.63	5.61	0.54	0.51	5.62	0.63	0.68
Circhist-30-10	4.49	1.51	0.74	1.80	4.20	0.52	4.94	1.06	0.38	2.47	3.53	0.66	5.60	0.55	0.51	5.56	0.69	0.71
Hist-15-15	4.52	1.48	0.59	1.59	4.41	0.29	4.85	1.15	0.36	2.42	3.58	0.49	5.51	0.60	0.52	5.49	0.67	0.64
Segment-50-50	4.74	1.26	0.60	2.26	3.74	0.71	4.92	1.08	0.37	2.89	3.11	0.62	5.59	0.56	0.52	5.68	0.59	0.63
Segment-100-20	4.47	1.53	0.62	1.92	4.08	0.52	4.90	1.10	0.36	2.57	3.43	0.56	5.54	0.58	0.51	5.53	0.68	0.69
Segment-100-100	5.98	0.82	1.09	4.11	3.11	2.94	5.37	0.64	0.42	5.20	2.57	3.09	6.14	0.51	0.66	6.63	0.87	0.95



(A)



(B)

Figura 25: Erros cometidos nos experimentos exp3/1 e exp6/1.
Erros cometidos pelas diferentes metodologias empregadas para a quantificação dos experimentos (A) - exp3/1 e (B) - exp6/1. Os cDNAs estão indicados no eixo x na seguinte ordem: 1 - LysA, 2 - TrpC, 3 - Gene Q, 4 - ST0280, 5 - Il-6, 6 - Irf-1.

os valores de razão esperados, apresentando erros e DPs equivalentes em todos os procedimentos empregados. Porém, isso não aconteceu nos fragmentos com sinais de intensidade mais baixos. No fragmento de ST0280 a metodologia *segment-100-100* apresentou razão média de 4,44, sendo que a mediana é 3,58 (valor não apresentado na tabela), mostrando que alguns *outliers* atrapalharam a média e o erro cometido. Esse fato realmente se confirma, sendo que dois *spots* desse cDNA apresentaram razão 19 e 14,65, respectivamente (dados não mostrados). Note que a razão média obtida pela metodologia *circfix* foi de 3,34. O mesmo fato pode ser observado nos dados obtidos para o fragmento de TrpC, onde a metodologia *segment-100-100* apresentou razão média de 3,41 e *circfix* apresentou razão média de 4,93 e mediana de 3,68, sendo que este mesmo fragmento teve um *outlier* com razão de 68,62. Entretanto, os demais procedimentos de quantificação de dados apresentaram razões médias de 0,99 a 1,98, embora tenham apresentado valores de erro menores. Nos dados referentes ao cDNA LysA também observamos que as técnicas *circfix* e *segment-100-100* apresentam valores de razão médias melhores que as demais, embora apresentem erros piores que algumas delas, o que pode ser justificado pela presença de alguns valores extremos.

Analisando os dados do experimento exp6/1, onde é esperada razão seis, Figura 25 (B) e Tabela 13, é possível notar que a metodologia *segment-100-100* apresentou erros menores para os fragmentos de LysA, TrpC, gene Q, ST0280 e Il-6. Observando os fragmentos de Il-6 e Irf-1, que são os cDNAs que apresentam maiores valores de intensidade de sinal em nosso conjunto de dados, nota-se que a maioria das metodologias apresentam dados bastante consistentes, com exceção da metodologia *circfix* que apresentou razão média bastante inferior ao que era esperado com baixo desvio padrão (sugerindo que não existe nenhum *outlier* muito extremo). Por outro lado, a metodologia *segment-100-100* foi mais precisa. Este dado indica que a segmentação por variação de intensidade foi peça fundamental para a exatidão dos dados obtidos, uma vez que a principal diferença entre as metodologias *circfix* e *segment-100-100* consiste no processo de localização dos *spots*. Quando olhamos para os fragmentos que apresentaram sinais de intensidade inferiores ao dos cDNAs citados anteriormente, nota-se que existe uma certa dependência entre os valores de intensidade e os erros cometidos nas razões calculadas. Em especial para os fragmentos que apresentaram sinais de intensidade muito baixos (ST0280 e TrpC) nota-se que todas as metodologias apresentaram valores de razão abaixo do esperado, embora nota-se uma maior

exatidão nos dados gerados pelo procedimento *segment-100-100*. Essa maior exatidão dos dados obtidos pelo procedimento *segment-100-100* é claramente notada nos dados referentes aos cDNAs de LysA e gene Q, onde tal procedimento apresentou valores de razão média iguais a 5,98 e 5,37 com baixo DP, respectivamente, ao passo que as demais metodologias detectaram uma razão média de, no máximo, 4,9.

Para os experimentos exp1/1-5/1, exp1/1-2/1 e exp1/1-10/1, eram esperados dois grupos de razões diferentes. Os cDNAs de TrpC, ST0280 e Irf-1 deveriam apresentar razão igual a um nos três experimentos citados. Observando os gráficos da Figura 26 vê-se que os fragmentos que apresentaram valores de intensidade baixos mostraram algumas diferenças entre os procedimentos diferentes utilizados, onde as metodologias *circfix* e *segment-100-100* apresentaram erros maiores que as demais. Isso pode ser explicado pelo fato de que os três experimentos citados nesse parágrafo apresentaram muitos artefatos pequenos no *background*. Como essas duas técnicas computam a média de todos os *pixels* que compõem o *background* da região de influência de cada *spot* e os cDNAs de TrpC e ST0280 apresentaram valores de intensidade de sinal baixos, os *spots* desses fragmentos tendem a incorporar maiores erros. Esse fato pode ser confirmado através dos dados obtidos para o fragmento do cDNA Irf-1, que apresentou valores de intensidade maiores que TrpC e ST0280, onde nota-se que as diferenças drásticas citadas anteriormente desaparecem.

Ainda, nesses três últimos experimentos os fragmentos de LysA, gene Q e Il-6 deveriam apresentar razão entre teste e referência de aproximadamente cinco (exp1/1-5/1), dois (exp1/1-2/1) e dez (exp1/1-10/1). Nota-se que para esses três fragmentos, os erros cometidos pelas metodologias *circfix* e *segment-100-100* sempre apresentaram erros menores que as demais, em especial no cDNA LysA. Outro ponto importante a destacar é o fato que o procedimento *adap* oferece erros muito altos para fragmentos que apresentam baixo valor de intensidade (LysA) ou em experimentos com uma grande diferença entre as amostras de teste e referência (exp1/1-10/1, ver Figura 26 (C)).

Tabela 14: Dados obtidos para o experimento exp1/1-5/1.

Médias, desvios padrão e erros obtidos para todos os fragmentos com diluição um da sonda no experimento exp1/1-5/1. Neste experimento é esperada razão um para os cDNAs de TrpC, ST0280 e Irf-1 e razão cinco para os demais.

Experimento exp1/1-5/1, com diluição um																		
Soft	**LysA**			**TrpC**			**Gene Q**			**ST0280**			**Il6**			**Irf1**		
	μ	**E**	**DP**	μ	**E**	**DP**	μ	**E**	**DP**	μ	**E**	**DP**	μ	**E**	**DP**	μ	**E**	**DP**
Circfix	3.76	1.24	0.37	1.53	0.53	0.23	4.30	0.70	0.21	1.50	0.50	0.17	3.80	1.21	0.32	0.91	0.09	0.04
Adap	2.39	2.61	0.36	0.79	0.21	0.12	4.09	0.93	0.25	1.08	0.11	0.11	3.54	1.46	0.18	0.90	0.10	0.03
Circhist-50-50	3.33	1.67	0.26	1.36	0.36	0.11	3.98	1.02	0.15	1.40	0.40	0.09	3.52	1.48	0.17	0.90	0.10	0.03
Circhist-100-20	3.20	1.80	0.25	1.31	0.31	0.11	3.92	1.08	0.13	1.37	0.37	0.09	3.48	1.52	0.16	0.90	0.10	0.03
Circhist-30-10	3.21	1.79	0.28	1.18	0.19	0.17	3.96	1.04	0.19	1.31	0.31	0.11	3.51	1.49	0.17	0.90	0.10	0.03
Hist-15-15	3.15	1.85	0.23	1.17	0.17	0.12	3.91	1.09	0.14	1.33	0.33	0.11	3.49	1.51	0.16	0.90	0.10	0.03
Segment-50-50	3.32	1.68	0.24	1.30	0.30	0.12	3.97	1.03	0.14	1.37	0.37	0.10	3.52	1.48	0.17	0.90	0.10	0.03
Segment-100-20	3.16	1.84	0.26	1.24	0.25	0.14	3.88	1.12	0.16	1.33	0.33	0.09	3.42	1.58	0.15	0.89	0.11	0.03
Segment-100-100	3.88	1.13	0.41	1.70	0.72	0.87	4.24	0.77	0.28	1.49	0.49	0.17	3.67	1.33	0.18	0.90	0.10	0.03

Tabela 15: Dados obtidos para o experimento exp1/1-2/1.

Médias, desvios padrão e erros obtidos para todos os fragmentos com diluição um da sonda no experimento exp1/1-2/1. Neste experimento é esperada razão um para os cDNAs de TrpC, ST0280 e Irf-1 e razão dois para os demais.

Experimento exp1/1-2/1, com diluição um																		
Soft	**LysA**			**TrpC**			**Gene Q**			**ST0280**			**Il6**			**Irf1**		
	μ	**E**	**DP**	μ	**E**	**DP**	μ	**E**	**DP**	μ	**E**	**DP**	μ	**E**	**DP**	μ	**E**	**DP**
Circfix	1.72	0.28	0.17	1.22	0.25	0.37	1.87	0.14	0.09	1.20	0.20	0.08	1.81	0.19	0.09	0.98	0.05	0.07
Adap	1.23	0.77	0.17	0.87	0.20	0.21	1.75	0.25	0.08	0.97	0.14	0.21	1.69	0.31	0.09	0.95	0.08	0.09
Circhist-50-50	1.62	0.38	0.10	1.08	0.09	0.06	1.76	0.24	0.07	1.13	0.13	0.05	1.71	0.29	0.08	0.97	0.04	0.05
Circhist-100-20	1.59	0.41	0.10	1.06	0.07	0.06	1.77	0.23	0.07	1.11	0.11	0.05	1.72	0.28	0.08	0.97	0.05	0.05
Circhist-30-10	1.53	0.47	0.18	1.01	0.06	0.08	1.78	0.22	0.07	1.08	0.10	0.08	1.72	0.28	0.09	0.97	0.05	0.06
Hist-15-15	1.60	0.40	0.09	1.01	0.03	0.04	1.76	0.24	0.07	1.07	0.07	0.04	1.71	0.29	0.08	0.97	0.05	0.05
Segment-50-50	1.61	0.39	0.10	1.07	0.08	0.06	1.76	0.24	0.07	1.12	0.12	0.05	1.71	0.29	0.08	0.97	0.05	0.05
Segment-100-20	1.58	0.42	0.12	1.04	0.06	0.06	1.76	0.24	0.06	1.10	0.10	0.05	1.71	0.29	0.08	0.97	0.05	0.06
Segment-100-100	1.79	0.27	0.22	1.22	0.23	0.24	1.84	0.16	0.07	1.22	0.22	0.12	1.78	0.22	0.10	0.98	0.05	0.08

Tabela 16: Dados obtidos para o experimento exp1/1-10/1.

Médias, desvios padrão e erros obtidos para todos os fragmentos com diluição um da sonda no experimento exp1/1-10/1. Neste experimento é esperada razão um para os cDNAs de TrpC, ST0280 e Irf-1 e razão dez para os demais.

Experimento exp1/1-10/1, com diluição um																		
Soft	**LysA**			**TrpC**			**Gene Q**			**ST0280**			**Il6**			**Irf1**		
	μ	**E**	**DP**	μ	**E**	**DP**	μ	**E**	**DP**	μ	**E**	**DP**	μ	**E**	**DP**	μ	**E**	**DP**
Circfix	7.82	2.32	1.83	1.39	0.40	0.31	10.55	1.23	1.52	1.78	0.78	0.34	8.87	1.20	1.17	1.23	0.23	0.08
Adap	1.66	8.34	0.47	0.76	0.26	0.17	4.38	5.62	1.00	0.70	0.30	0.17	5.33	4.67	0.80	1.28	0.28	0.10
Circhist-50-50	6.43	3.57	0.81	1.38	0.38	0.15	8.52	1.48	0.70	1.61	0.61	0.14	7.52	2.48	0.41	1.24	0.24	0.05
Circhist-100-20	5.64	4.36	0.98	1.23	0.23	0.16	8.10	1.90	0.86	1.48	0.48	0.18	7.41	2.59	0.61	1.23	0.22	0.07
Circhist-30-10	5.10	4.90	1.18	1.07	0.17	0.21	7.58	2.42	1.21	1.27	0.29	0.25	6.97	3.03	0.94	1.19	0.19	0.09
Hist-15-15	5.61	4.39	0.74	1.13	0.15	0.13	8.02	1.98	0.73	1.40	0.40	0.12	7.23	2.77	0.48	1.23	0.23	0.06
Segment-50-50	6.15	3.85	0.87	1.34	0.34	0.14	8.44	1.56	0.70	1.51	0.51	0.17	7.39	2.61	0.46	1.24	0.24	0.06
Segment-100-20	5.41	4.59	1.01	1.18	0.19	0.15	7.86	2.14	0.86	1.36	0.37	0.22	7.05	2.95	0.74	1.22	0.22	0.07
Segment-100-100	8.68	2.55	3.36	1.74	0.74	0.72	10.94	1.45	2.29	1.77	0.77	0.83	8.66	1.36	0.91	1.26	0.26	0.07

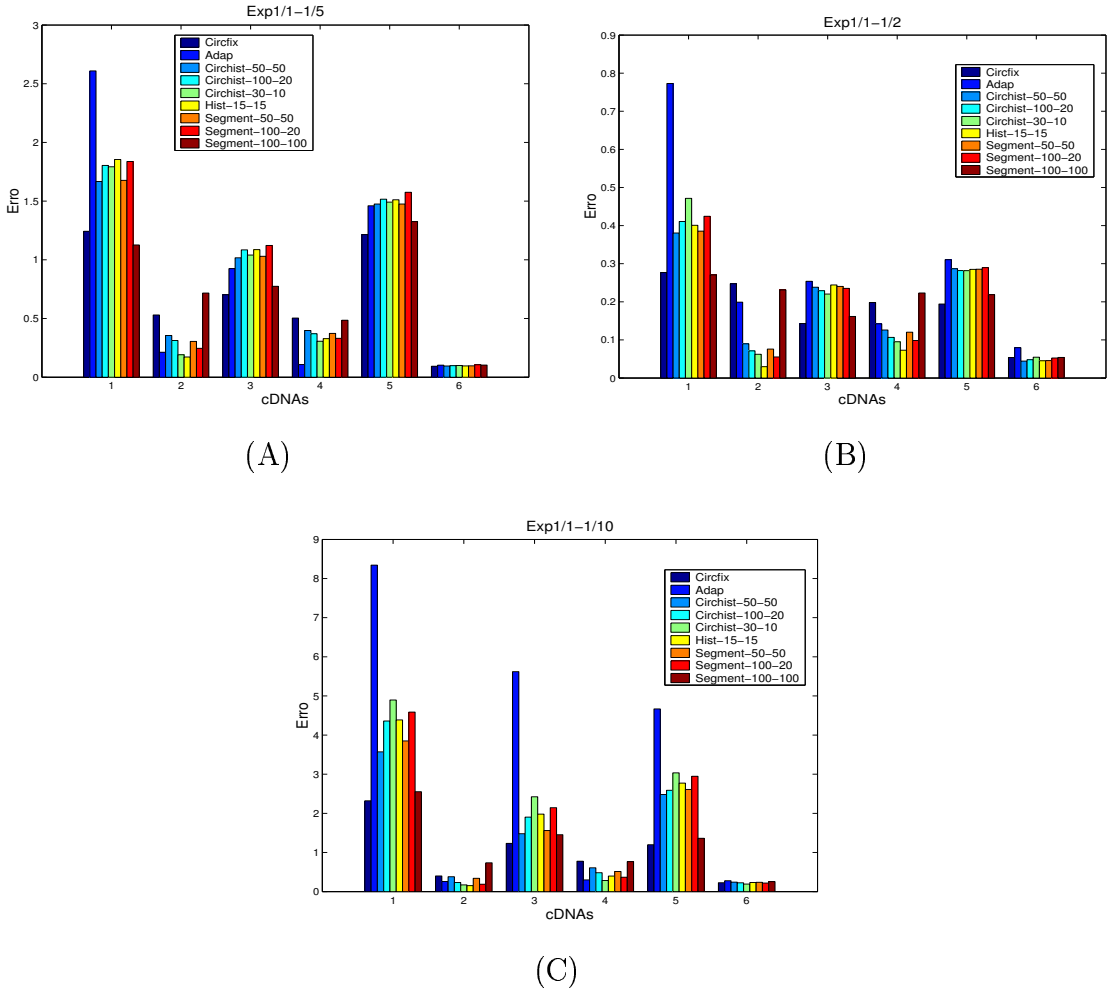


Figura 26: Erros cometidos nos experimentos exp1/1-5/1, exp1/1-2/1 e exp1/1-10/1.

Essa figura mostra os erros cometidos pelas diferentes metodologias nos experimentos (A) - exp1/1-5/1, (B) - exp1/1-2/1 e (C) - exp1/1-10/1. Os cDNAs estão indicados no eixo x na seguinte ordem: 1 - LysA, 2 - TrpC, 3 - Gene Q, 4 - ST0280, 5 - Il-6, 6 - Irf-1.

6 DISCUSSÃO

A tecnologia de cDNA *microarray* é uma ferramenta bastante utilizada atualmente. Dentre as suas principais utilidades, destaca-se a busca de diferenças de expressão de vários genes em um único experimento. Entretanto, os dados gerados através dessa metodologia estão sujeitos a várias fontes de incertezas que podem levar a erros nas análises finais destes dados, como mostra Finkelstein *et al* [9]. Além disso, a dificuldade de estimação dos valores de intensidade reais no momento da análise de imagens também pode influenciar nos resultados finais. Uma vez que existem vários procedimentos sendo atualmente empregados por diferentes *softwares* de análise de imagens de *microarray*, a comparação desses diferentes procedimentos constitui um importante passo na busca daquela metodologia que minimize os erros cometidos e favoreça a aquisição de dados mais robustos e confiáveis. Recentemente foi publicado um trabalho onde foram analisadas diferentes metodologias de processamento de imagens provenientes de experimentos de cDNA *microarray* [53]. Este trabalho buscou por metodologias que minimizam a variabilidade dos dados obtidos. Entretanto, pouco se sabe sobre a exatidão destas diferentes metodologias de processamento de imagens, o que mostra a necessidade do desenvolvimento de uma estratégia experimental onde seja possível ter uma idéia do erro cometido por tais metodologias.

No trabalho apresentado aqui foi desenvolvida uma estratégia experimental com o objetivo principal de validar diferentes metodologias de localização e quantificação de *spots* em imagens provenientes de experimentos de cDNA *microarray*. Esta metodologia se baseia na construção de experimentos onde são fixados fragmentos de cDNA conhecidos em lâminas de vidro que são hibridizadas contra cDNAs marcados por fluorêscencia construídos a partir de mRNAs que continham exatamente a mesma seqüência nucleotídica dos fragmentos fixados como sonda. Esses mRNAs foram construídos através de reações de transcrição *in vitro* utilizando como molde os mesmos fragmentos de cDNA fixados. Esse cDNA molde foi previamente amplificado por PCR utilizando oligonucleotídeos contendo a seqüência promotora para a enzima T3 (ou SP6) *RNA polimerase* e outro oligo contendo uma seqüência poli A. Após essa etapa, quantidades específicas e conhecidas de mRNAs eram misturadas e seguiam o protocolo de hibridização normalmente utilizado nos experimentos de cDNA *micro-*

array desenvolvidos rotineiramente no laboratório. Desta forma, através de experimentos controlados, foi possível determinar antecipadamente qual a razão esperada entre os diversos experimentos construídos e compará-las com os resultados obtidos. Existem fortes evidências bioquímicas que garantem a semelhança entre esses experimentos controlados e os experimentos biológicos normalmente realizados através da utilização da tecnologia de cDNA *microarray*. O sucesso da hibridização foi garantido, uma vez que as lâminas construídas eram hibridizadas contra cDNAs marcados por fluorescência que continham exatamente as mesmas seqüências nucleotídicas que estavam fixadas nas lâminas. Além disso, as lâminas produzidas foram hibridizadas seguindo-se os protocolos de hibridização comumente utilizados no laboratório.

Utilizando esta nova estratégia experimental, foram construídos oito experimentos, sendo que três deles foram utilizados para ajustar a quantidade inicial de mRNA a ser utilizada nas reações de transcrição reversa para marcação do cDNA flutuante (cDNA alvo). Seis desses experimentos foram quantificados através de 9 procedimentos diferentes de análise de imagens, sendo um deles baseado em segmentação de círculo fixo (*circfix*) [8], outros três baseados em segmentação de círculo fixo sendo que é permitido uma certa mobilidade aos *spots* (*circhist-50-50*, *circhist-100-20* e *adap*) [32], um outro procedimento define as regiões de sinal baseando-se na distribuição do histograma de maneira automatizada com base nos espaçamentos entre os *spots* [18, 19] e outros três baseados em segmentação por morfologia matemática [15]. Além dessas metodologias, também foi utilizada uma outra baseada em segmentação por histograma.

Desta forma, foram utilizadas cinco metodologias diferentes de localização de *spots* em experimentos de cDNA *microarray*, onde foi possível avaliar a capacidade de localização de *spots* nessas diferentes metodologias estudadas, com excessão do procedimento de segmentação por histograma, uma vez que ele não delimita uma máscara para os *spots*. A metodologia de segmentação por círculo fixo (*circfix*) é muito susceptível a problemas de localização errada de *pixels* como pertencentes a região de sinal, como indicado na Figura 10 (A). Além disso, essa metodologia é totalmente baseada em computação gráfica, forçando o usuário a corrigir muitos *spots* que ficam fora da sua posição estimada, o que faz com que o processamento de cada experimento seja extremamente trabalhoso e cansativo para o usuário, além de não ser reprodutível. A metodologia de segmentação por círculo fixo que permite uma

certa mobilidade à máscara que define cada *spot*, embora seja mais robusta que a metodologia de círculo fixo tradicional, também não define exatamente todos os *spots* como mostra a Figura 10 (B), levando à problemas semelhantes ao da metodologia anterior. A metodologia baseada em segmentação que delimita os *spots* com base nas medidas da imagem (*circhist-30-10*) é um processo totalmente automatizado e requer do usuário apenas as imagens em *cy3* e *cy5* juntamente com uma descrição da geometria da lâmina. A partir daí, o programa processa as imagens e gera um arquivo texto contendo os dados referentes aos *spots* e um outro arquivo (em formato TIFF) com a imagem da segmentação realizada. Entretanto, esta metodologia é muito susceptível a erros de segmentação de blocos. Para corrigir erros desse tipo foi necessário reprocessar as imagens alterando parâmetros no campo *Layout adjustment options*, ver subseção A.3 do Anexo A. Além disso, a descoberta dos parâmetros ideais não é muito trivial, o que faz com que isso se torne uma tarefa não muito simples para o usuário. Ressaltamos, ainda, que esses erros acontecem freqüentemente. Ao contrário do que foi citado até aqui, a metodologia de segmentação por morfologia matemática é altamente robusta e eficaz. Todas as imagens dos experimentos produzidos neste trabalho foram corretamente processadas. Eventualmente alguma linha gerada no processo de gradeamento dos blocos ou *spots* devem ser ajustadas, o que sempre foi observado em conjunto com algum problema nas imagens. Quanto a localização das áreas de sinal, a metodologia foi precisa para todos os experimentos produzidos aqui, mesmo para os pontos que apresentavam baixa intensidade, ver Figura 10 (D). A metodologia baseada em histograma utilizada neste trabalho não fornece nenhuma imagem da região delimitada pelos *spots*, e por isso não foi discutida aqui.

A metodologia de processamento de imagens de *microarray* também pode explicar algumas diferenças nos valores de intensidade calculados para o *background*. Em especial, os valores de intensidade do *background* estimados através da metodologia *circfix* são maiores que os valores obtidos pelas outras metodologias, como mostra a Tabela 9 e a Figura 11 (B). Esse fato não é devido à presença de artefatos no *background*, já que o procedimento *segment-100-100* também utiliza todos os *pixels* que constituem a região de influência dos *spots* e não oferece valores de *background* similares aos obtidos na metodologia *circfix*. A Figura 10 (A) mostra que existem casos em que *spots* que apresentam valores de sinal mais intensos ocupam uma área visualmente maior que os *spots* menos intensos, levando a metodologia *circfix* a computar

pixels do *foreground* no *background*.

A melhor forma de compararmos as diferentes metodologias de localização e quantificação de dados decorrentes de experimentos de *microarray* é observando as razões entre teste e referência em experimentos controlados onde sabemos de antemão os resultados esperados. Os experimentos produzidos aqui tem essa característica e, por isso, nos permitiram fazer esse tipo de análise.

No experimento exp1/1 era esperada razão um para todos os fragmentos utilizados neste trabalho, a Figura 12, mostra que não existe muita diferença entre os diferentes procedimentos de análise de imagens com relação a disperão dos *spots*, sendo que a metodologia *adap* apresentou um padrão relativamente mais "gordo". Isso sugere uma maior variabilidade nos dados obtidos por tal técnica, o que é confirmado nos gráficos de razão mostrados na Figura 13. Entretanto, análises baseadas apenas em razão um podem esconder muitas características importantes. Uma justificativa muito simples para isso está no fato de que muitos fragmentos que apresentaram valores de intensidade de sinal baixos (por terem tamanho pequeno – LysA e TrpC, ou problemas relacionados à qualidade do mRNA sintetizado – TrpC e ST0280), sendo esses valores próximos aos de *background*. Uma vez que os valores de *background* não variam muito entre os dois fluorocromos utilizados, os resultados das razões tederiam para um. Para contornar esse problema foram desenhados experimentos que apresentavam proporções diferentes entre os fragmentos de cDNA utilizados como amostras de teste e referência.

A maior variabilidade obtida para o procedimento *adap* se repetiu em todos os outros experimentos construídos, que deveriam apresentar razões diferentes de um. Entretanto, com base nesses novos experimentos foi notado que essa maior variabilidade aconteceu principalmente para os fragmentos dos cDNAs de LysA, TrpC e ST0280, que apresentaram valores de intensidade de sinal abaixo dos demais cDNAs, como mostra as Figuras 14, 15, 16 e 17. Embora essa variabilidade tenha sido maior para os dados estimados a partir da metodologia *adap*, ela também aconteceu para todos os outros procedimentos.

Com relação ao tamanho dos fragmentos fixados, este trabalho mostra que a intensidade de sinal aumenta de maneira proporcional ao tamanho dos cDNAs fixados nas lâminas de vidro. Dado semelhante foi mostrado por Stillman & Tonkinson [42], os quais demonstraram que a dinâmica da hibridização é influenciada pelo tamanho

do cDNA imobilizado na lâmina. Neste estudo ficou demonstrado que a taxa de hibridização era diretamente dependente ao tamanho do cDNA imobilizado, havendo um ponto de inflexão da curva em cDNAs com 712pb de tamanho. Para cDNAs maiores que 712pb a influência do tamanho na taxa de hibridização diminuía.

É interessante salientar que no trabalho citado foram utilizados cDNAs alvo com tamanho fixo de 712pb, variando-se apenas os tamanhos dos fragmentos fixados (116, 454, 712, 1233 e 2057 pares de base), o que pode favorecer os dados obtidos, uma vez que o material alvo tem tamanho único de 712pb. No trabalho apresentado aqui foram utilizados fragmentos de 300, 650, 960 e 2000 pares de bases, aproximadamente, onde mostramos que o ponto de inflexão parece estar em 650pb. Nota-se que existe uma lacuna entre 450 e 650bp, quando são considerados os dois trabalhos realizados. Isso pode indicar a possibilidade de que o ponto de inflexão real da curva esteja nesse intervalo. Novos experimentos são necessários para uma avaliação mais detalhada desta característica.

Além da maior variabilidade observada nos valores de razão obtidos para os cDNAs que apresentaram valores de intensidade de sinal muito baixos, também foi observada a existência de uma certa dependência entre os valores de intensidade de sinal dos *spots* e os valores de razão calculados, como é mostrado nas Figuras 20 e 21, onde observa-se que a medida que os valores de intensidade caem os valores de razão também caem. Esse fato pode ser explicado pela maior variabilidade dos valores de intensidade dos *pixels* que constituem os *spots* de baixa intensidade. Além disso, nota-se que as metodologias que não selecionam todos os *pixels* da região de sinal e do *background*, apresentam uma dependência maior que as metodologias que selecionam todos os *pixels* que constituem as regiões citadas.

Com relação à exatidão das metodologias estudadas, foi observado que as metodologias que computam todos os *pixels* da região de sinal para a extração do valor de intensidade dos *spots* e todos os *pixels* da região de *background* para o cálculo dos valores de intensidade dessa região oferecem dados mais corretos que as outras metodologias, especialmente para os fragmentos que apresentaram valores de intensidade de sinal muito baixos, ver Tabelas 13, 14, 15 e 16 e Figuras 25 (B) e 26. Nota-se que as metodologias *circfix-50-50*, *circfix-100-20*, *circfix-30-10*, *segment-50-50* e *segment-100-20* apresentam resultados inferiores aos observados para as duas metodologias citadas. Entretanto, os artefatos que aparecem na região de *background* interferem so-

bremaneira no cálculo das razões para as metodologias *circfix* e *segment-100-100* como foi observado, principalmente, nos dados referentes ao experimento exp3/1, como pode ser visto na Tabela 12 e na Figura 25 (A). A melhor forma de se evitar esse tipo de problema seria a detecção exata destes artefatos, separando *background* de artefato, o que é relativamente difícil já que eles não apresentam uma forma homôgenea. Entretanto, é possível calcular o *background* selecionando-se uma proporção grande dos *pixels* da região do *background*, evitando-se a seleção daqueles que estejam nos percentis superiores da distribuição (selecionar os *pixels* compreendidos entre os percentis 1 e 90, por exemplo). Isso pode ser feito através do *software* desenvolvido pelo grupo do Bioinfo-USP, que permite a seleção de qualquer percentil de *pixels* tanto para a medida do *foreground* como do *background*. Esses dados mostram a necessidade de filtrar os *pixels* pertencentes as regiões de artefatos encontrados no *background*, fato que já foi mencionado em outros trabalhos [4, 45, 52].

Além disso, a metodologia baseada em variação de intensidade constitui um processo totalmente automatizado, evitando a necessidade de intensa manipulação de *spots* por parte do usuário, o que evita a introdução de erros humanos durante esse processo, além de economizar um bom tempo que seria dispendido com a correção de *spots* mal localizados.

É importante lembrar que seria ideal filtrar o verdadeiro ruído das imagens decorrentes de experimentos de cDNA *microarray*, que são as regiões do interior dos *spots* que não contribuem nos valores de intensidade de sinal. Entretanto a correlação entre os valores de intensidade dos *pixels* que constituem a área de um *spot* pode ser muito útil para filtrar esse ruído. Nenhum *software* utilizado na atualidade utiliza um procedimento semelhante a esse, entretanto a implementação de um algorítmo capaz de filtrar os *pixels* que não estejam dentro de uma boa correlação linear não é uma tarefa muito complicada e provavelmente faria com que a etapa de quantificação dos valores de intensidade do *background* seja irrelevante.

7 CONCLUSÕES

Uma conclusão importante que pode ser extraída do presente trabalho está na utilidade da estratégia experimental desenvolvida aqui. Nota-se que essa nova estratégia apresenta um importante impacto para a avaliação e validação de diferentes procedimentos de análise de imagens de *microarray*, sendo que ela oferece o pontencial de se avaliar os erros cometidos, além da variabilidade dos dados.

Os dados obtidos aqui também sugerem que cuidados especiais devem ser tomados no momento da seleção dos clones para fixação nas lâminas, uma vez que os valores de intensidade medidos são diretamente proporcionais ao tamanho do fragmento fixado, dado que já havia sido observado anteriormente [42]. Esse fato ainda pode ser reforçado pela dependência observada entre os valores de intensidade e os valores de razão calculados, onde *spots* que que apresentam valores de intensidade muito baixos tendem a apresentar erros maiores que os *spots* que apresentam sinais de intensidade maiores.

Esses erros podem ser minimizados através da utilização de *softwares* de análise de imagens baseados em segmentação por variação de intensidade e que selecionam todos os *pixels* da região de sinal e o maior número possível dos *pixels* pertencentes à região de *background*, evitando a seleção daqueles que estejam contaminados por artefatos, uma vez que tais procedimentos resultam em dados mais robustos e precisos para os valores de razão obtidos. Isso é especialmente notado nos *spots* que apresentam valores de intensidade de sinal muito baixos.

Outra contribuição importante do presente trabalho está na validação dos experimentos de *microarray*, sendo que os experimentos realizados neste trabalho mostraram que, os dados decorrentes de tais experimentos são confiáveis, reprodutíveis e apresentam erros aceitáveis, como mostra a figura 22. Entretanto, um ponto importante deve ser destacado, que é a perda de precisão observada para fragmentos de cDNA que apresentavam valores de intensidade de sinal muito baixos.

8 PERSPECTIVAS

Neste trabalho foi desenvolvida uma nova estratégia experimental, onde as razões esperadas entre alguns fragmentos de cDNA são conhecidas. Embora, existam evidências bioquímicas, que assegurem a semelhança entre esse tipo de experimento e os experimentos "reais" de *microarray*, seria ideal uma validação estatística (por teste de hipótese) para garantir a similaridade das distribuições dos *pixels* de um *spot* nas imagens construídas a partir de mRNA sintético e real.

Os dados observados neste trabalho mostram a existência de uma dependência entre os valores de intensidade dos *spots* e os valores de razão calculados, além de indicar que a presença de artefatos na região do *background* tem o potencial de influenciar os valores de razão observados. Assim, nota-se a necessidade de se desenvolver um método para a detecção automática de qualidade dos *spots* com base no padrão de distribuição dos valores de intensidade de sinal e do *background*. Uma outra possibilidade é a determinação de alguns limites inferiores para os valores de intensidade de sinal corrigidos pelo *background*, uma vez que foi mostrado aqui que os dados decorrentes de *spots* com sinal de intensidade muito baixos apresentam maior imprecisão nos resultados obtidos. Além disso, seria interessante a utilização de algumas medidas de incerteza, tais como intervalo de confiança, relacionando os valores de intensidades e razões calculadas.

Essa nova estratégia experimental, desenvolvida neste trabalho, também pode ser empregada para vários outros procedimentos de análise de imagens de *microarray*, sendo uma ferramenta de validação dos *softwares* utilizados para o processamento de imagens.

Também seria muito interessante desenvolver uma estratégia de votação hierárquica que atribuiria um voto para o melhor procedimento para cada um dos seis cDNAs utilizados em todos os experimentos. A soma desses votos ajudará na escolha do melhor procedimento de análise de imagens. Entretanto, o critério para atribuição desses votos deve levar em conta tanto o erro quanto o desvio padrão cometidos.

9 REFERÊNCIAS BIBLIOGRÁFICAS

[1] Adams R, Bischof L. Seeded region growing. **IEEE Trans Pattern Analysis Machine Intelligence** 1994; 16:641–7.

[2] Alizadeh AA, Eisen MB, Davis RE, et al. Distinct types of diffuse large B-cell lymphoma indentified by gene expression profiling. **Nature** 2000; 403:503–11.

[3] Bowtell DD. Options available – from start to finish – for obtaining expression data by microarray. **Nat Genet** 1999; 21 (1 Suppl):25–32.

[4] Brown CS, Goodwin PC, Sorger PK. Image metrics in the estatistical analysis of DNA microarray data. **Proc Natl Acad Sci USA** 2001; 98:8944–9.

[5] Burke HB. Discovering patterns in microarrays data. **Mol Diagn** 2000; 5:349–57.

[6] DeRisi J, Penland L, Brown PO, et al. Use of a cDNA microarray to analyse gene expression patterns in human cancer. **Nat Genet** 1996; 14:457–60.

[7] Duggan DJ, Bittner M, Chen Y, Meltzer P, Trent JM. Expression profiling using cDNA microarrays. **Nat Genet** 1999; 21 (1 Suppl):10–4.

[8] Eisen M. **ScanAlyze**. [Programa de computador]. Berkeley, 1999.

[9] Finkelstein D, Ewing R, Gollub J, Sterky F, Cherry JM, Somerville S. Microarray data quality analysis: lessons from AFGC project. **Plant Mol Biol** 2002; 48:119–31.

[10] Friend SH, Stoughton RB. The magic of microarrays. **Sci Am** 2002; 286:44–53.

[11] Goffeau A. Four years of post-genomic life with 6000 yeast genes. **FEBS Lett** 2000; 480:37–41.

[12] Gonzalez RC, Woods RE. **Digital image processing**. Addison-Wesley Publishing Company, Boston, 1992.

[13] Graves DJ, Su HJ, Addya S, Surrey S, Fortina P. Four-laser scanning confocal system for microarray analysis. **Biotechniques** 2002; 32:346–54.

[14] Hedge P, Qi R, Abernathy K, et al. A concise guide to cDNA microarray analysis. **Biotechniques** 2000; 29:548–62.

[15] Hirata Jr R, Barrera J, Hashimoto RF, Dantas DO, Esteves GH. Segmentation of microarray images by mathematical morphology. **Real-Time Imaging** 2002; 8:491–505.

[16] Hubank M, Schatz DG. Identifying differences in mRNA expression by representational difference analysis of cDNA. **Nucleic Acids Res** 1994; 22:5640–8.

[17] Iseli C, Stevenson BJ, de Souza SJ, et al. Long-range heterogeneity at the 3' ends of human mRNAs. **Genome Res** 2002; 12:1068–74.

[18] Jain AN, Tokuyasu TA. **Spot 2.0**. [Programa de Computador]. San Francisco, 2002.

[19] Jain AN, Tokuyasu TA, Snijders AM, Segraves R, Albertson DG, Pinkel D. Fully automatic quantification of microarray image data. **Genome Res** 2002; 12:325–32.

[20] Jin W, Riley RM, Wolfinger RD, White KP, Passador-Gurgel G, Gibson G. The contributions of sex, genotype and age to transcriptional variance in drosophila melanogaster. **Nat Genet** 2001; 29:389–95.

[21] Khodarev NN, Yu J, Nodzenski E, et al. Method of RNA purification from endothelial cells for DNA array experiments. **Biotechniques** 2002; 32:316–20.

[22] K̄night J. When the chips are down. **Nature** 2001; 410:860–1.

[23] Kurian KM, Watson CJ, Wyllie AH. DNA chip technology. **J Pathol** 1999; 187:267–71.

[24] Lage JM, Hamann S, Gribanov O, Leamon JH, Pejovic T, Lizardi PM. Microgel assessment of nucleic acid integrity and labeling quality in microarray experiments. **Biotechniques** 2002; 32:312–4.

[25] Lee SW, Tomasetto C, Sager R. Positive selection of candidate tumor-suppressor genes by subtractive hybridization. **Proc Natl Acad Sci USA** 1991; 88:2825–9.

[26] Liang P, Pardee AB. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. **Science** 1992; 257:967–70.

[27] Lin B, White JT, Ferguson C, et al. Part-1: A novel human prostate-specific, androgen-regulated gene that maps to chromossome 5q12. **Cancer Res** 2000; 60:858–63.

[28] Mangalam H, Stewart J, Zhou J, et al. Genex: an open source gene expression database and integrated tool set. **IBM Syst J** 2001; 40:552–69.

[29] Matsumoto EY. **Matlab 6 - Fundamentos de programação**. São Paulo: Editora Érica; 2001.

[30] Nagai T, Chapman Jr WH. Analysis of microliter volumes of dye-labeled nucleic acids. **Biotechniques** 2002; 32:356–64.

[31] Neto ED, Correa RG, Verjovski-Almeida S, et al. Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. **Proc Natl Acad Sci USA** 2000; 97:3491–6.

[32] Pachard BioScience. **QuantArray Microarray Analysis Software**. [Programa de Computador]. Boston, 2001.

[33] Pradet-Balade B, Boulmé F, Müllner EW, Garcia-Sanz JA. Reliability of mRNA profiling: verification for samples with different complexities. **Biotechniques** 2001; 30:1352–7.

[34] Quackenbush J. Computational analysis of microarray data. **Nat Rev Genet** 2001; 2:418–27.

[35] Sansom CE, Smith CA. Computer applications in biomolecular sciences. part 2: bioinformatics and genome projects. **Biochem Educ** 2000; 28:127–31.

[36] Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. **Science** 1995; 270:467–70.

[37] Schuchhardt J, Beule D, Malik A, et al. Normalization strategies for cDNA microarrays. **Nucleic Acids Res** 2000; 28:e47.

[38] Shalon D, Smith SJ, Brown PO. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. **Genome Res** 1996; 6:639–45.

[39] Sherlock G, Hernandez-Boussard T, Kasarskis A, et al. The stanford microarray database. **Nucleic Acids Res** 2001; 29:152–5.

[40] Southern E, Mir K, Shchepinov M. Molecular interactions on microarrays. **Nat Genet** 1999; 21(1 Suppl):5–9.

[41] Southern EM. Detection of specific sequences among DNA fragments separated by gel electrophoresis. **J Mol Biol** 1975; 98:503–17.

[42] Stillman BA, Tonkinson JL. Expression microarray hybridization kinetics depend on length of the immobilized DNA but are independent of immobilization substrate. **Anal Biochem** 2001; 295:149–57.

[43] The International Human Genome Mapping Consortium. A physical map of the human genome. **Nature** 2001; 409:934–941.

[44] Triche TJ, Schofield D, Buckley J. DNA microarrays in pediatric cancer. **Cancer J** 2001; 7:2–15.

[45] Tseng GC, Oh MK, Rohlin L, Liao JC, Wong WH. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. **Nucleic Acids Res** 2001; 29:2549–57.

[46] Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. **Science** 1995; 270:484–7.

[47] Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. **Science** 2001; 291:1304–51.

[48] Walker J, Rigley K. Gene expression profiling in human peripheral blood mononuclear cells using high-density filter-based cDNA microarrays. **J Immunol Methods** 2000; 239:167–79.

[49] Watson J, Crick F. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. **Nature** 1953; 171:737–8.

[50] Welsh J, Chada K, Dalal S, Cheng R, Ralph D, McClelland M. Arbitrarily primed PCR fingerprinting of RNA. **Nucleic Acids Res** 1992; 20:4965–70.

[51] Wildsmith SE, Archer GE, Winkley AJ, Lane PW, Bugelski PJ. Maximization of signal derived from cDNA microarrays. **Biotechniques** 2001; 30:202–8.

[52] Yang YH, Buckley MJ, Speed TP. Analysis of cDNA microarray images. **Brief Bioinform** 2001; 2:341–9.

[53] Yang YH, Buckley MJ, Dudoit S, Speed TP. Comparison of methods for image analysis on cDNA microarray data. **J Comput Graphical Stat** 2002; 11:108–36.

[54] Yang YH, 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** 2002; 30:e15.

[55] Zien A, Aigner T, Zimmer R, Lengauer T. Centralization: a new method for the normalization of gene expression data. **Bioinformatics** 2001; 17 (1 Suppl):S323–31.

A *SOFTWARES* UTILIZADOS

Uma das principais etapas no desenvolvimento de experimentos de cDNA *microarray* é a análise das imagens geradas. Nessa etapa destaca-se principalmente a localização dos *spots* e a extração das intensidades obtidas nos dois canais como já foi citado no decorrer do texto. Neste apêndice vamos citar as principais características dos *softwares* utilizados neste trabalho.

A.1 *SCANALYZE*

O *ScanAlyze* foi desenvolvido por Michael B. Eisen na Universidade de Stanford. Ele opera em imagens provenientes de hibridizações obtidas de um ou dois fluorocromos e tem o objetivo final de fornecer uma tabela numérica para análises posteriores. Este *software* se baseia totalmente em computação gráfica e oferece um ambiente muito iterativo, exigindo intensa interferência do usuário. O procedimento completo de análise de um experimento de *microarray* é descrito a seguir:

- Carregar as imagens,
- Ajustar o ganho para que a maioria dos *spots* fique visível,
- Criar todos os *grids* necessários (ou utilizar *grids* previamente criados),
- Ajustar a localização dos *grids* até que todos os blocos estejam corretamente localizados,
- Fazer o refinamento da localização,
- Corrigir *spots* eventualmente mal localizados,
- Marcar *spots* problemáticos (*flag*),
- Salvar os *grids*,
- Exportar a tabela de dados.

A Figura 1 mostra uma imagem sendo analisada com o *software ScanAlyze*, dando uma visão geral do seu funcionamento. As imagens são carregadas através dos botões *load*. Atualmente o *ScanAlyze* pode trabalhar com dois tipos de arquivos de

imagem: com extensão *.scn* utilizado na Universidade de Stanford e arquivos TIFF de 8 e 16 *bits*. A composição das duas imagens é pseudocolorida e mostrada na tela através do botão *redraw*. Também é interessante notar a existência de vários outros campos que controlam o brilho (*gain*) e a normalização entre as duas imagens (*norm*). Todos esses botões e campos citados aqui, podem ser visualizados na Figura 1 (A).

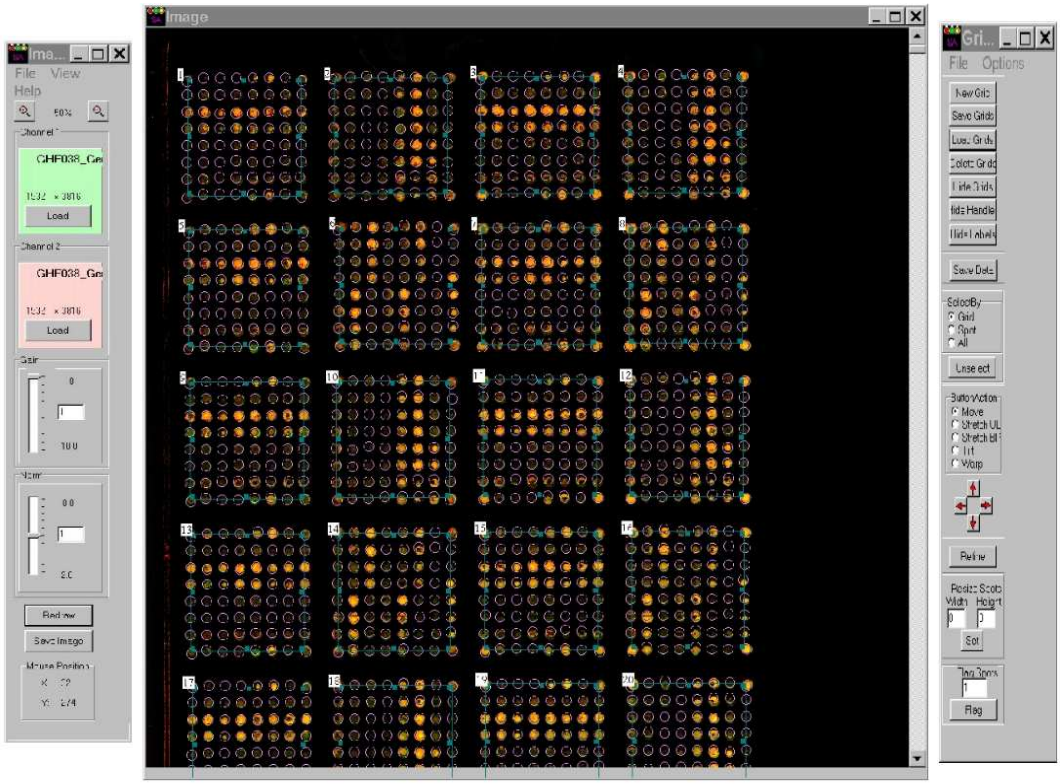


Figura 1: Visão geral do *ScanAlyze*.
Nesta figura temos os três componentes principais da *interface* do *ScanAlyze*, que são: (A) - *Image control form*, (B) - *Image* e (C) - *Grid control form*.

A etapa seguinte, e talvez mais importante, é a construção da grade que localiza todos os *spots* da lâmina. Todos os procedimentos referentes a essa fase do trabalho são controlados através da janela *Grid Control Form* (ver Figura 1 (C)). Sempre que uma lâmina de um novo lote de experimentos é analisada pela primeira vez um novo *grid* deve ser criado. Isso é feito clicando-se no botão *new grid* e preenchendo-se um formulário como indicado pela Figura 2. A descrição de cada parâmetro necessário nessa etapa se encontra na Tabela 1.

Os campos encontrados em *First Grid Position* (*Left* e *Top*) definem o centro do *spot* que está localizado no canto superior esquerdo do primeiro bloco da imagem.

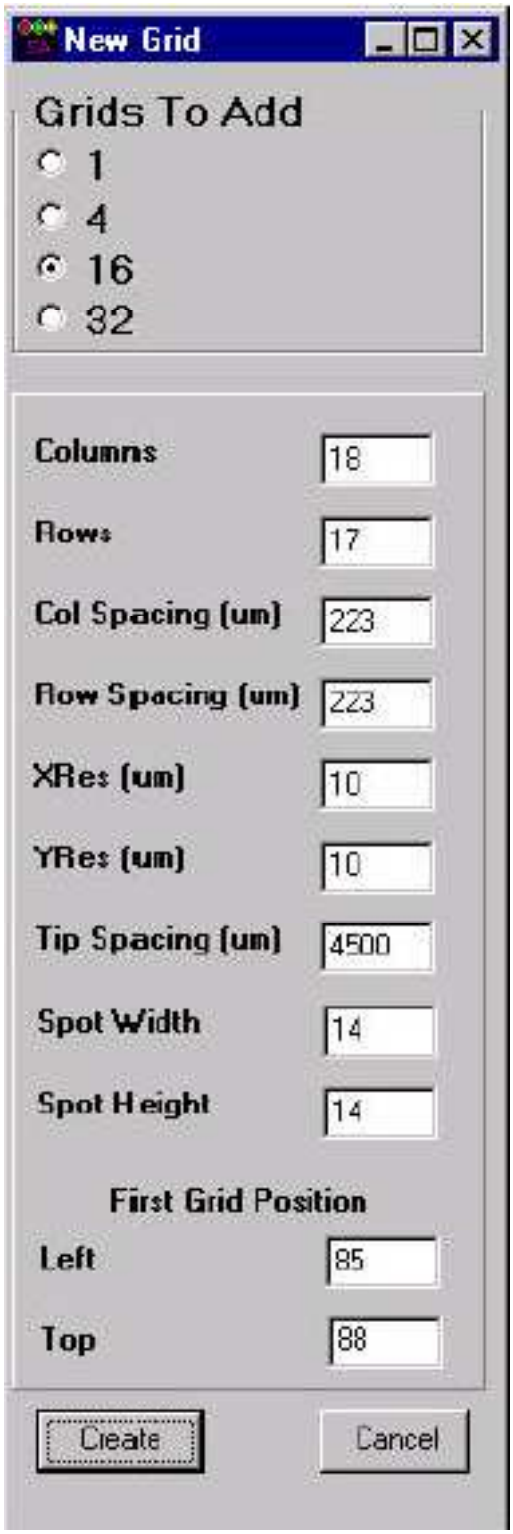


Figura 2: A criação de novos *grids* no *ScanAlyze*.

Janela ilustrando a etapa de criação de novos *grids* no *ScanAlyze*. Os campos a serem preenchidos aqui são referentes às informações da geometria da lâmina, bem como algumas dicas de espaçamento entre blocos e *spots*. A descrição de cada ítem está na Tabela 1.

Tabela 1: Os parâmetros necessários para a criação de um novo *grid*. Descrição dos parâmetros necessários para a criação de um novo *grid*, de acordo com a Figura 2.

Parâmetro	Descrição
Grids to add	Número de blocos
Columns	colunas de *spots* em cada bloco
Rows	linhas de *spots* em cada bloco
Col Spacing	espaçamento hor. entre *spots* (μm)
Row Spacing	espaçamento vert. entre *spots* (μm)
XRes	tamanho hor. de cada *pixel* (μm)
Yres	tamanho vert. de cada *pixel* (μm)
Tip Spacing	espaçamento entre blocos (μm)
Spot Width	tamanho hor. do *spot* (*pixels*)
Spot Height	tamanho vert. do *spot* (*pixels*)

Como nem sempre é possível determinar corretamente o valor de todos esses parâmetros, o usuário deve corrigir manualmente posições que estiverem erradamente localizadas. A localização e inclinação entre outros problemas podem ser ajustados para todos os *grids* diretamente via *mouse* ou atavés das setas direcionais que se encontram na janela *Grid Control Form* bastando deixar o campo *Select By* selecionado na posição *grid*, essas correções também podem ser feitas para todos os *grids* de uma só vez selecionando a posição *all*. Uma vez que todos os blocos estejam próximos da sua posição ideal o usuário pode (e deve) utilizar o botão *refine*, que se encontra logo abaixo das setas direcionais, para otimizar os parâmetros automaticamente.

Entretanto, muitos *spots* ainda podem ficar fora da posição inferida pelo *grid*, o que pode ser corrigido selecionando-se o botão *spot* no campo *Select By* e ajustando a sua posição com o auxílio do *mouse* ou das setas direcionais. Também é possível ajustar o tamanho dos *spots* através do campo *Resize Spots* e marcar *spots* que estejam dentro de áreas problemáticas do *array* no campo *Flag Spots*, onde qualquer valor diferente de zero não deve ser considerado em análises posteriores. O *ScanAlyze* também oferece a opção de marcação automática de *spots* com qualidade ruim. O

principal problema aqui é que essas correções devem ser feitas manualmente, *spot* por *spot*, o que torna esse procedimento extremamente trabalhoso e cansativo.

Uma vez que todos os *spots* foram corretamente localizados, um clique no botão *Save Data* faz com o *ScanAlyze* faça uma série de cálculos internos e grave um arquivo texto delimitado por tabulações com os dados obtidos.

Primeiramente o *software* separa a imagem em *pixels* que estão dentro de alguma região identificada como *spot* dos que não foram indentificados. A Figura 3 ilustra essa divisão de *pixels* para um *spot* em particular. Qualquer *pixel* que esteja dentro ou sobre o círculo que define o *spot* é contado como pertencente à área ocupada por sinal (também conhecida como *foreground*) e qualquer outro que não esteja dentro deste círculo mas que esteja dentro de um quadrado centrado no *spot* com lado igual a 2 vezes o raio do *background* (um parâmetro, em *pixels*, definido pelo usuário cujo *default* é 20) é definido como *pixel* do *background*. É importante notar que se as medidas desse quadrado ultrapassarem os limites dos *spots* vizinhos ao que está sendo analisado, os *pixels* pertencentes a tais *spots* não serão computados.

A Tabela 2 mostra os valores dados pelo *ScanAlyze* na tabela exportada, vários desses dados são úteis para avaliar a qualidade dos *spots*.

A.2 *QUANTARRAY*

O *software QuantArray*, da *Packard BioScience*, oferece uma série de vantagens em relação ao *ScanAlyze*, o que o torna mais robusto e estável do que este último. Ele pode ser utilizado em dois níveis, como um programa *stand-alone* operado manualmente ou de maneira semi automatizada usando um protocolo previamente criado ou em conjunto com o Sistema *ScanArray* de análise de *microarray*. A Figura 4 mostra uma visão geral desse *software*.

Basicamente, o *QuantArray* guia o usuário por uma série de passos até a quantificação da intensidade de fluorescência nos vários *spots* existentes na lâmina. Os parâmetros importantes desses passos, tais como a geometria da lâmina, são descritos por um protocolo que também inclui preferências do usuário para vários parâmetros de quantificação. Cabe salientar que existe um assistente de criação para esse protocolo e que ele é necessário antes do início da análise de algum experimento.

Uma vez criado o protocolo que será usado (também é possível utilizar protocolos

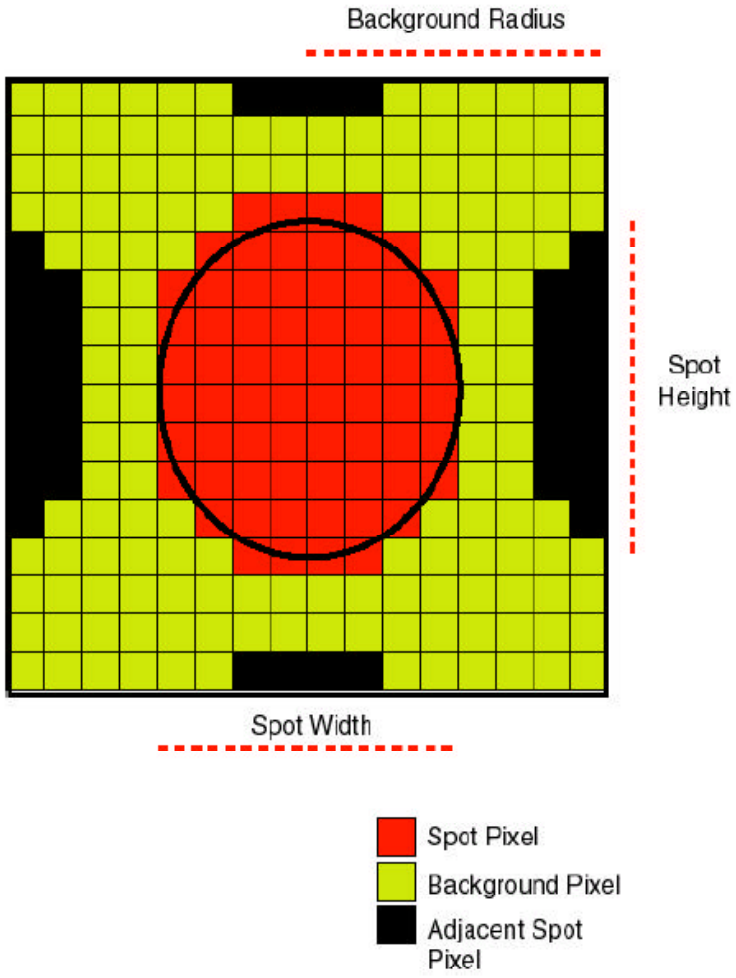


Figura 3: A segmentação empregada pelo *ScanAlyze*.

Seleção de *pixels* que constituem a área do *spot* e do *background* empregado pelo *ScanAlyze*.

Tabela 2: Os valores exportados pelo *ScanAlyze*.

Um clique no botão *Save Data*, fará com que o *ScanAlyze* faça uma série de cálculos internos, para a construção do arquivo de dados numéricos. Esta tabela descreve quais são os valores exportados neste arquivo pelo *software*.

Valor exportado	Descrição
CH1I e *CH2I*	Intensidade média dos *pixels* do *spot* (canais 1 e 2)
SPIX	número de *pixels* no *spot*
CH1B e *CH2B*	mediana dos *pixels* do *background* local (canais 1 e 2)
CH1BA e *CH2BA*	média dos *pixels* do *background*
BGPIX	número de *pixels* do *background*
MRAT	razão *pixel* a *pixel* calculada entre os dois canais
REGR	inclinação da regressão linear simples entre os dois canais
LFRAT	inclinação da regressão de mínimos quadrados
CORR	correlação entre os *pixels* dos dois canais
CH1GTB1 e *CH2GTB1*	fração de *pixels* maiores que o *background* no *spot*
CH1GTB2 e *CH2GTB2*	fração de *pixels* maiores que 1,5 vezes o *background*
CH1KSD e *CH2KSD*	Teste estatístico entre *foreground* e *background*
CH1KSP e *CH2KSP*	as probabilidades do teste anterior
CH1EDGEA e *CH2EDGEA*	magnitude média dos vetores *Sobel* de cada *spot*
SPOT	indexador único para cada *spot* da lâmina
GRID, ROW, COL	*grid* onde o *spot* está localizado e a posição dentro do *grid*

já criados por outros usuários) todo o processo de análise de imagens com o *QuantArray* é realizado em várias etapas independentes entre si (campo *Analysis Step* na Figura 4), que serão descritos a seguir.

Register Images Para a correta quantificação dos dados de um experimento de *microarray*, as duas imagens estudadas (*cy3* e *cy5*) devem ser sobrepostas, isso garante que os *spots* que tenham hidridizado somente em um ou outro canal sejam localizados. Com esta opção selecionada é possível mover uma imagem sobre a outra através das setas direcionais do teclado ou clicando-se nas setas situadas logo abaixo do campo *Analysis Step*, Figura 4, com o objetivo de ajustar a sobreposição das imagens.

Specify Location Uma vez que as imagens estejam ajustadas, é necessário selecionar

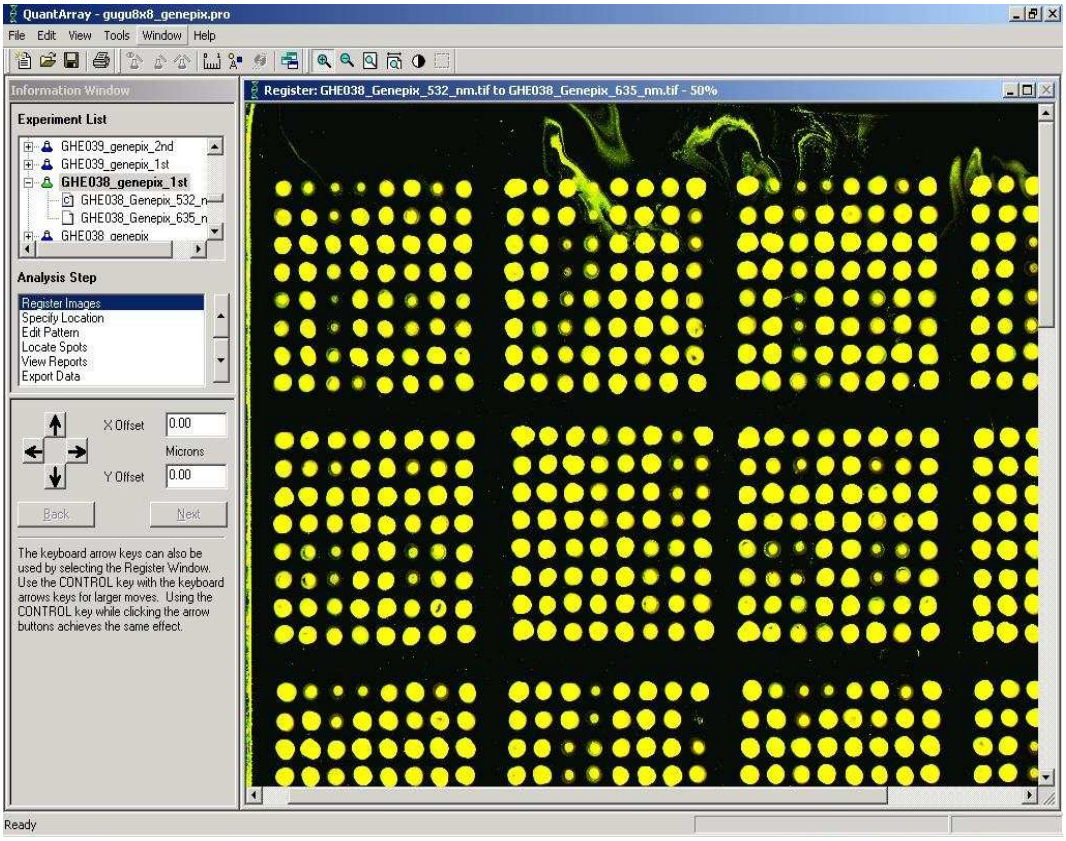


Figura 4: Visão geral do *QuantArray*.

Esta figura mostra a interface gráfica do *QuantArray*, onde vemos a imagem composta do experimento que está sendo analisado juntamente com outros campos de controle do programa.

esta opção para indicar a localização do *spot* superior esquerdo da lâmina. Isto pode ser feito clicando com o botão esquerdo do *mouse* diretamente na região desejada ou arrastando-se o quadrinho que aparece na tela até tal posição.

Edit Pattern Quando este item é selecionado, um *grid* é colocado sobre a imagem, o qual espera-se que cubra corretamente a maioria dos *spots*, a Figura 5 ilustra esta etapa da análise. Como isto nem sempre acontece, é possível mover linhas, colunas ou todo um bloco de *spots* (campo *Selection Type* na Figura 5), o campo *rotate* corrige eventuais erros de inclinação. O *grid* utilizado aqui é construído com base nas informações sobre a geometria da lâmina encontradas no protocolo citado no início desta seção.

Locate Spots A etapa seguinte é a localização dos pontos ocupados por sinal. No *QuantArray* isso significa determinar o centro de cada *spot* juntamente com uma área subjacente, que geralmente é retangular. É dentro desta área que o

software irá determinar os *pixels* que irão constituir o *foreground* e o *background*, ver Figura 6. Uma vez selecionado o campo *Locate Spots* basta clicar em *Start Locate*. Eventualmente alguns *spots* podem ficar mal localizados e devem ser corrigidos com o *mouse*.

View Reports Quando a etapa de localização dos *spots* está completa é possível visualizar, individualmente, a morfologia dos diversos *spots* existentes na lâmina, ver Figura 7. Isso pode ajudar a verificar se os parâmetros de definição do *foreground* e *background* estão corretos. Também é possível visualizar os dados em vários outros tipos de gráficos, que também são úteis para a criação de filtros de qualidade para os *spots* da lâmina analisada, entretanto esses ítens não são muito relevantes para o presente trabalho.

Export Data Finalmente podemos exportar a tabela de dados para posterior análise. Isso é feito selecionando-se o item *Export Data* no campo *Selection Type* e então clicando-se no botão *Export*. Uma caixa de diálogo será aberta para que seja especificado um nome de arquivo que será salvo como um arquivo texto delimitado por tabulações.

O arquivo criado pelo *QuantArray* pode ser dividido em três partes. A primeira descreve diversos parâmetros do protocolo, imagens e fluorocromos utilizados entre outras coisas, a segunda parte descreve filtros utilizados ou não e finalmente a terceira parte fornece os valores de intensidade de sinal e *background* para cada *spot*; este programa também fornece valores de desvio-padrão, diâmetro, área, circularidade, etc, referentes tanto aos valores de intensidade como os de *background* para todos os pontos da lâmina.

Uma das principais características do *QuantArray* são os três métodos distintos de quantificação que ele oferece, são eles: o método do histograma, círculo fixo e adaptativo. Cada um desses métodos tem suas características peculiares que serão descritas a seguir.

O primeiro método, como o próprio nome diz, calcula um histograma para todos os *pixels* que se encontram dentro do retângulo definido na etapa de localização dos *spots*, enfatizamos que aqui não é feita nenhuma restrição quanto à máscara do *spot*. Uma vez que este histograma foi calculado, quatro parâmetros são usados para quantificar os valores de intensidade e *background*. Estes parâmetros estão listados na

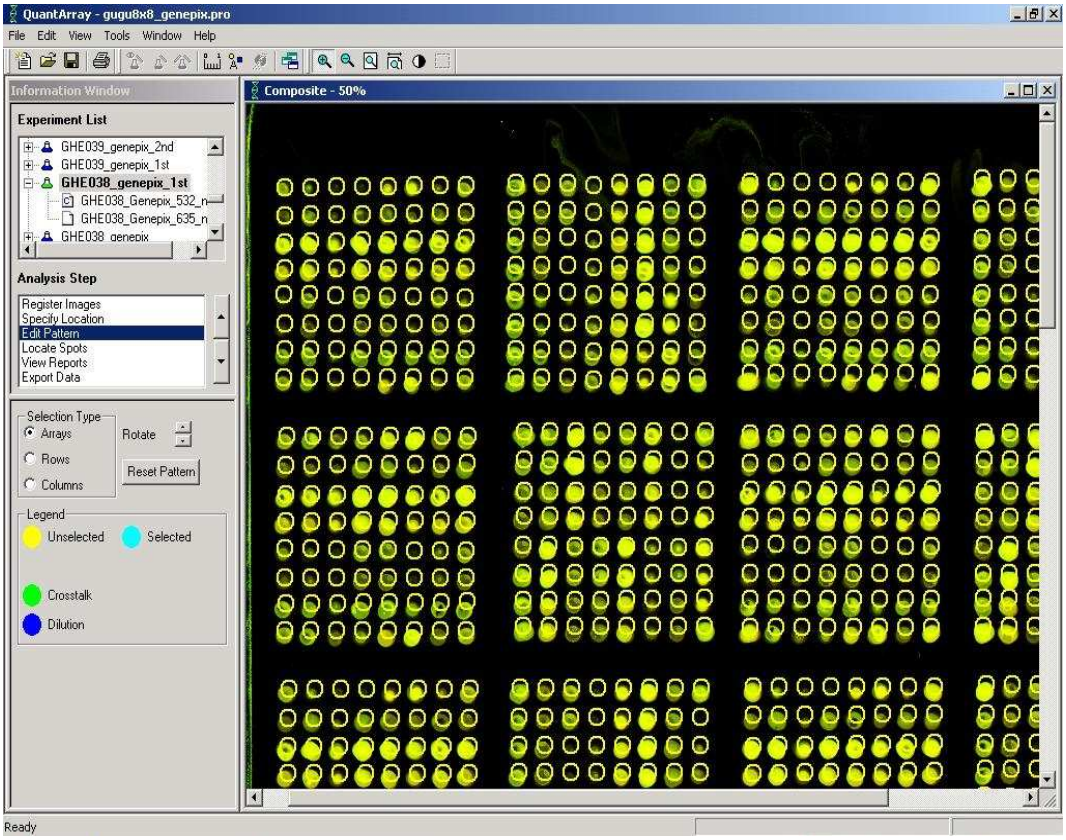


Figura 5: A correção do *grid* original no *QuantArray*.

Etapa de correção do *grid* original (*Edit Pattern*), onde a máscara dos *spots* deve ficar o mais próximo possível do ideal.

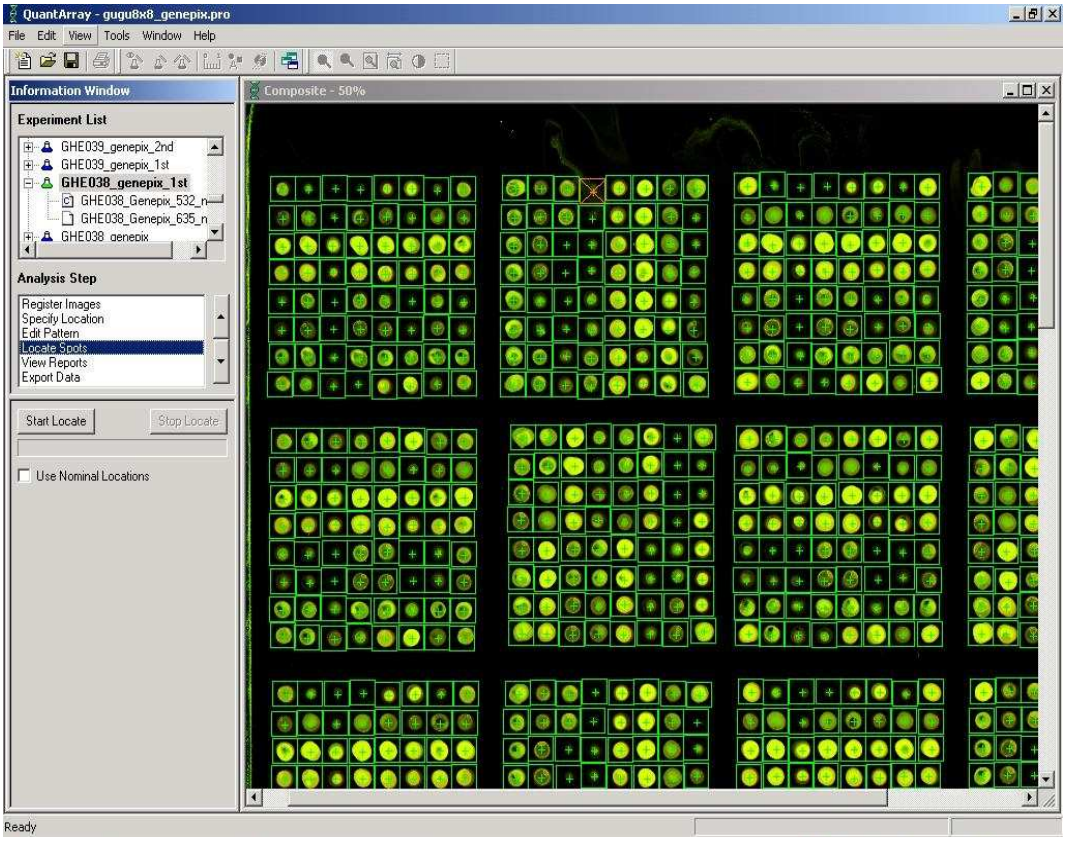


Figura 6: A localização de *spots* no *QuantArray*.

Uma vez que a máscara inicial foi corretamente ajustada vem a etapa de localização de *spots*, o que é feito com um clique no botão *Start locate*.

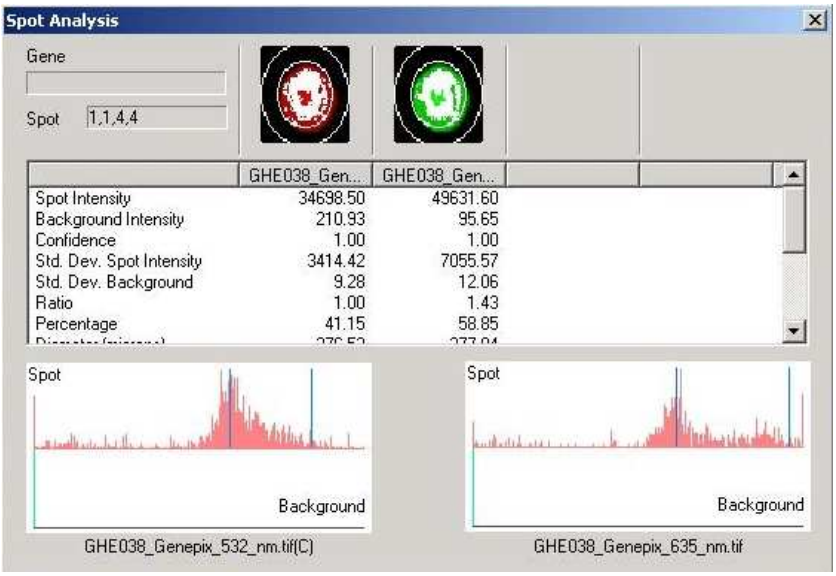


Figura 7: Visualização da morfologia de um *spot*.

Depois que os *spots* foram localizados, é possível avaliar a morfologia dos mesmos. Esta figura ilustra este procedimento, onde é possível ver a máscara do *spot*, alguns dados estatísticos junto com um histograma da distribuição dos *pixels* nos dois canais.

Tabela 3, onde os valores padrão estão listados na última coluna. O método do círculo fixo é semelhante ao utilizado pelo *software ScanAlyze*, entretanto ele tem uma ligeira modificação na seleção de *pixels* que é indicada na Figura 8. Diferente do *ScanAlyze*, na metodologia empregada pelo *QuantArray* se calcula um histograma para os *pixels* do sinal e outro histograma para os do *background*. A partir dos histogramas podemos selecionar diferentes percentis para a quantificação dos valores de sinal e *background*. Assim, o método do círculo fixo utiliza quatro parâmetros idênticos aos utilizados pelo histograma além de outros três que estão listados na Tabela 4, valores padrão se encontram na última coluna.

Tabela 3: Parâmetros utilizados pelo método do histograma.
Esta tabela descreve os parâmetros necessários para o método do histograma empregado pelo *QuantArray*.

Parâmetro	Descrição	Padrão
Signal Low	Perc. mín. para cálculo da intensidade do *foreground*	80
Signal High	Perc. máx. para cálculo da intensidade do *foreground*	95
Background Low	Perc. mín. para cálculo da intensidade do *background*	5
Background High	Perc. máx. para cálculo da intensidade do *background*	20

Tabela 4: Parâmetros utilizados pelo método do círculo fixo.
Esta tabela descreve os parâmetros necessários para o método do círculo fixo empregado pelo *QuantArray*.

Parâmetro	Descrição	Padrão
Signal Low	Perc. mín. para cálculo da intensidade do *foreground*	45
Signal High	Perc. máx. para cálculo da intensidade do *foreground*	95
Background Low	Perc. mín. para cálculo da intensidade do *background*	5
Background High	Perc. máx. para cálculo da intensidade do *background*	20
Spot Diameter	Diâmetro do *spot*	x[1]
Back. Inner Diameter	Diâm. da circ. interior que define o *background*	y[1]
Back. Outer Diameter	Diâm. da circ. exterior que define o *background*	z[1]

[1]Parâmetros dados em *microns* (μm) que dependem das características das imagens analisadas. Esses parâmetros são diferentes entre si e podem variar de experimento para experimento.

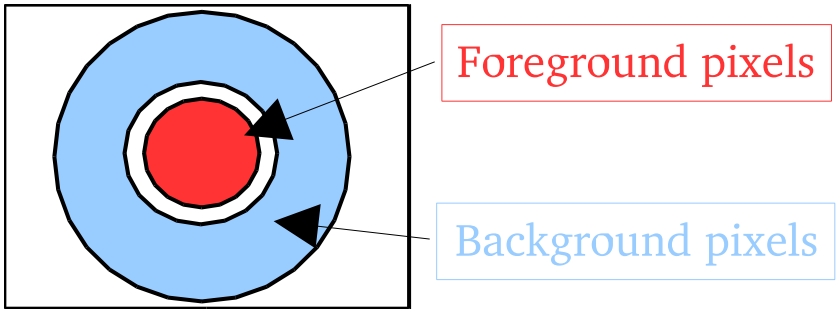


Figura 8: A segmentação empregada pelo *QuantArray*.
Esta figura ilustra o esquema de segmentação empregado pelo *QuantArray*. Esse tipo de seleção de *pixels* é empregado pelo método do círculo fixo e adaptativo.

No método adaptativo, as máscaras iniciais dos *spots* e *background* são construídas da mesma maneira que o método do círculo fixo. A diferença é que este método é refinado por um teste estatístico (teste de *Mann Whitney*). Este teste compara uma amostra de oito *pixels* que estão na região do *spot* com outros oito que estão na região do *background*, sendo o processo repetido até que o *QuantArray* encontre diferença estatisticamente significativa entre os *pixels* do *spot* e do *background*. Esse método usa três parâmetros de diâmetro que definem as áreas do sinal e *background* semelhantes ao do método do círculo fixo acrescido de um quarto parâmetro que é o *p-Value* – valor que controla a confiança estatística do teste realizado. O *QuantArray* utiliza como padrão $p - Value < 0,0001$.

Além das três diferentes metodologias de quantificação de sinal utilizada pelo *QuantArray*, ainda podemos selecionar o tipo de medida que pode ser usada para quantificar o valor de intensidade dos *pixels* selecionados:

Total Intensities Soma das intensidades de todos os *pixels* do *spot* e *background*.

Mean Intensity Intensidade média dos *pixels*.

Mode Intensity Intensidade mais freqüente entre todos os *pixels*, ou seja, a moda.

Median Intensity Intensidade mediana dos *pixels*.

A.3 *SPOT*

O *software Spot* foi desenvolvido na Universidade da Califórnia e está disponível apenas para uso acadêmico, consulte [19] para maiores detalhes. A Figura 9 mostra a *interface* gráfica do *Spot*. O seu algorítmo é baseado em cinco passos principais, que são eles:

- Cálculo dos espaçamentos entre *spots* e *subarrays*.
- Localização e cálculo da posição dos *subarrays*.
- Localização e cálculo da posição de *spots* individuais.
- Identificação de *pixels* do *foreground* e *background* para cada *spot*.
- Cálculo dos dados e criação do arquivo texto.

Basicamente é necessário informar apenas as imagens do experimento (em formato TIFF) no campo *De Novo + Hint parameters and options*, aqui também é possível utilizar uma outra imagem que tenha sido corada com algum marcador de DNA (DAPI, por exemplo) apenas para a construção dos *grids*. Cabe ressaltar que quando não se utiliza esta terceira imagem, deve-se deixar esse campo preenchido com a palavra *blank*, veja Figura 9. Também é necessário especificar um nome de arquivo no campo *Output file prefix* para a saída do *software* e a geometria da lâmina utilizada no campo *De Novo parameters and options*, essa geometria é definida pelo número de blocos (linhas X colunas) e o número de *spots* dentro de cada bloco (linhas X colunas). Ainda existem outros parâmetros que são opcionais, tais como dicas de espaçamento ou tamanho do diâmetro dos *spots*. Para o processamento das imagens basta clicar no botão *Run* no canto superior direito da janela.

O cálculo dos espaçamentos é feito através do somatório das intensidades de sinal nas direções x e y das imagens, onde o padrão dos picos é usado para determinar tanto os espaçamentos entre *spots* como entre blocos, a Figura 10 ilustra esse padrão de picos em ambas as direções. A seguir é aplicada uma função de custo que calcula as diferenças entre as regiões do centro dos *spots* e regiões entre eles. Um refinamento final é feito mudando os posicionamentos dos *grids* dos blocos a fim de se maximizar o valor dessa função de custo. Os posicionamentos dos *spots* também são ajustados individualmente de maneira semelhante.

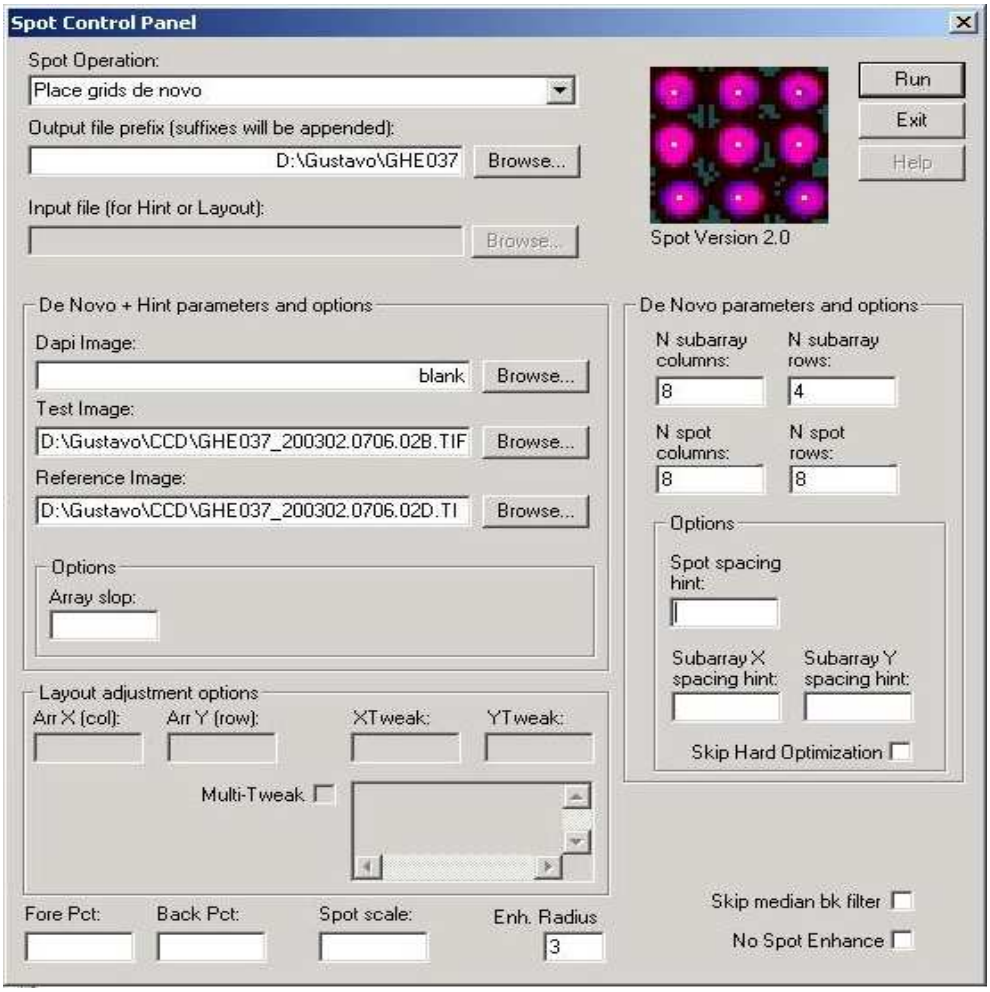


Figura 9: *Interface* gráfica do *software Spot*.

Esta figura ilustra a imagem principal do *software Spot*. Cabe salientar que este programa não mostra a imagem que está sendo analisada.

Os *pixels* a serem utilizados como *foreground* ou *background* são definidos com base no cálculo do histograma local da distribuição dos sinais de intensidade. Esse histograma é computado sobre uma área retangular com centro no *spot* com lado igual ao espaçamento entre *spots*. Os *pixels* do *foreground* são definidos como sendo um certo percentil com sinal mais intenso (30% é o padrão) e que esteja dentro de um certo raio a partir do centro do *spot* (raio padrão: metade da medida lateral do retângulo menos um *pixel*). Um outro percentil de *pixels* com intensidade mais baixa é definido para a quantificação do *background* (padrão: 10%) e que esteja fora de um outro raio, cujo padrão é a metade da medida do retângulo mais um *pixel*. Os valores de intensidade dos *pixels* do *background* são pós processados, onde os valores extremos (*outliers*) são substituídos pela mediana, o que evita a quantificação de alguns sinais espúrios. Pelo visto aqui, a seleção de *pixels* feita pelo *Spot* é bastante

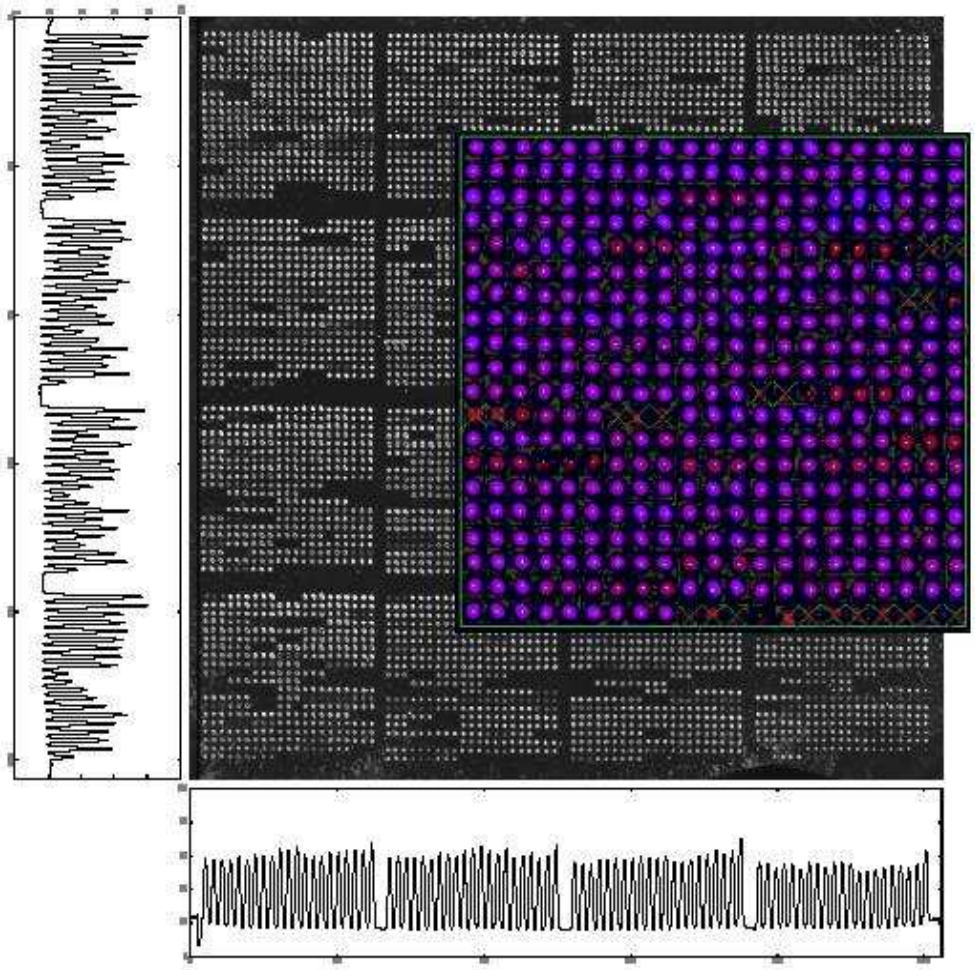


Figura 10: A segmentação empregada pelo *Spot*.

Este programa calcula a soma de intensidades nos eixos x e y com o objetivo de localizar os blocos e *spots* com base nos picos de intensidade observados. Aqui vemos um exemplo do padrão de picos obtidos após o somatório dos sinais de intensidade nas direções x e y. Em destaque temos o exemplo de segmentação de um dos blocos da imagem.

similar à metodologia do círculo fixo empregada no *QuantArray*, com a diferença de calcular apenas um histograma para a região delimitada pelo retângulo ao redor do *spot*. A imagem ampliada na Figura 10 ilustra este esquema de segmentação, onde as regiões avermelhadas representam os *spots* corretamente identificados. O *software* produz uma imagem em formato TIFF semelhante a essa onde é possível avaliar a qualidade da segmentação obtida, a Figura 11 tem um exemplo dessa imagem. Todos esses parâmetros podem ser ajustados pelo usuário nos campos *Fore Pct*, *Back Pct*, *Spot Scale* e *Enh. Radius*.

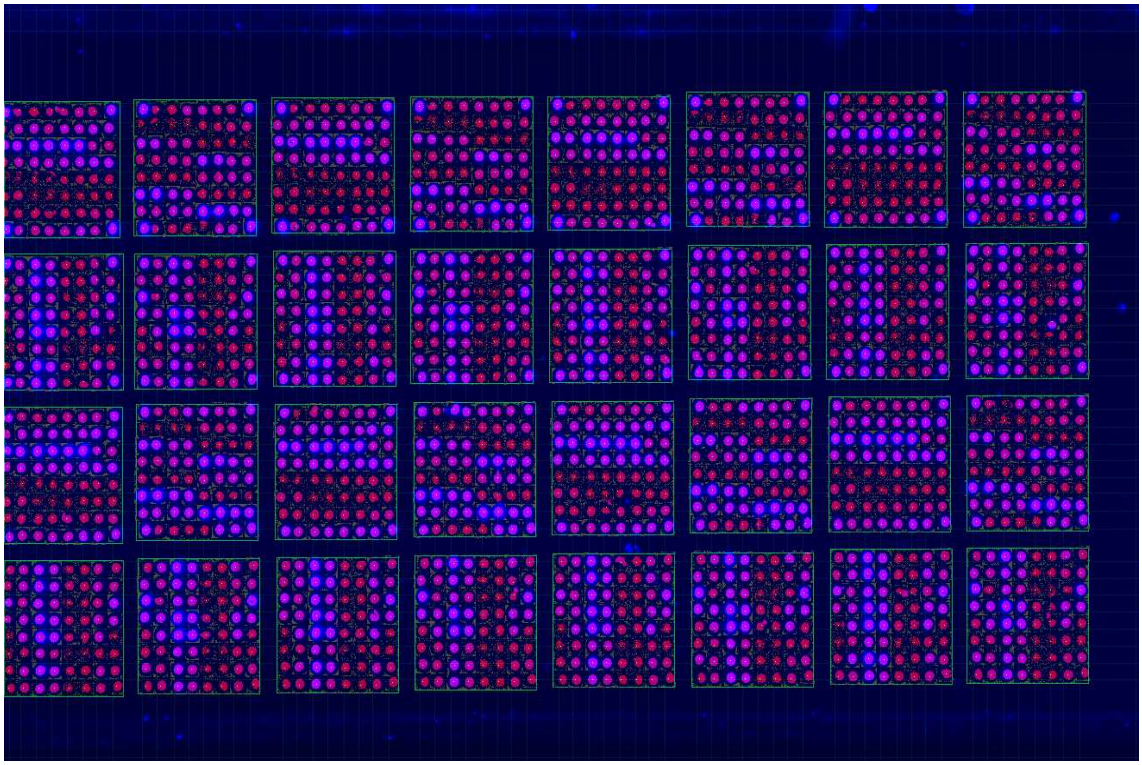


Figura 11: Exemplo de segmentação do *Spot*.

Figura da segmentação de um experimento de *microarray* através do *software Spot*.

Um problema muito freqüente na metodologia empregada pelo *software Spot* acontece na segmentação dos blocos, onde muitas vezes os blocos da imagem não são corretamente localizados. Um exemplo do tipo de erro cometido está indicado pelas setas brancas na Figura 12.

Depois de determinados os *pixels* que serão utilizados como *foreground* e *background*, a etapa final é o cálculo dos valores de intensidade nos dois canais diferentes juntamente com outros valores de qualidade e razão entre eles. A Tabela 5 descreve os principais campos exportados pelo *Spot* em um arquivo texto.

A.4 BIOINFO-USP

Este é um novo *software* que está sendo desenvolvido pelo Laboratório de Bioinformática do IME-USP (Bioinfo), cuja principal característica é a segmentação por variação de intensidade baseada em Morfologia Matemática, que torna a análise das imagens mais confiável e totalmente automatizada. Cabe salientar que ele ainda está em fase de aprimoramento, embora já exista uma versão bastante robusta definida.

Tabela 5: A tabela de dados gerada pelo *Spot*.

Valores exportados pelo *Spot* na tabela de dados. É possível ver que este *sotware* também fornece vários dados estatísticos que podem ser usados para controle de qualidade dos *spots* da lâmina.

Valor exportado	**Descrição**
SpotNum	Indexador único para cada *spot*
Flag	Indicador de qualidade: 0 se o *spot* é bom
nfore	Número de *pixels* do *foreground*
nback	Número de *pixels* do *background*
TestFore	Média do *foreground* do canal 1
TestForeSD	Desvio padrão do *foreground* do canal 1
TestBack	Média do *background* do canal 1
TestBackSD	Desvio padrão do *background* do canal 1
RefFore	Média do *foreground* do canal 2
RefForeSD	Desvio padrão do *foreground* do canal 2
RefBack	Média do *background* do canal 2
RefBackSD	Desvio padrão do *background* do canal 2
RawRat	Razão entre canal 1 e canal 2 corrigidos pelo *background*
SpotCorr	Correlação entre os *pixels* dos canais 1 e 2
TestZstat	Teste estatístico entre sinal e *back.* do canal 1
RefZstat	Teste estatístico entre sinal e *back.* do canal 2
Slope	Regressão linear entre os canais 1 e 2 (*pixels*)
MeanRat	Razão média entre os canais 1 e 2 (*pixel* a *pixel*)
MedianRat	Razão mediana entre os canais 1 e 2 (*pixel* a *pixel*)

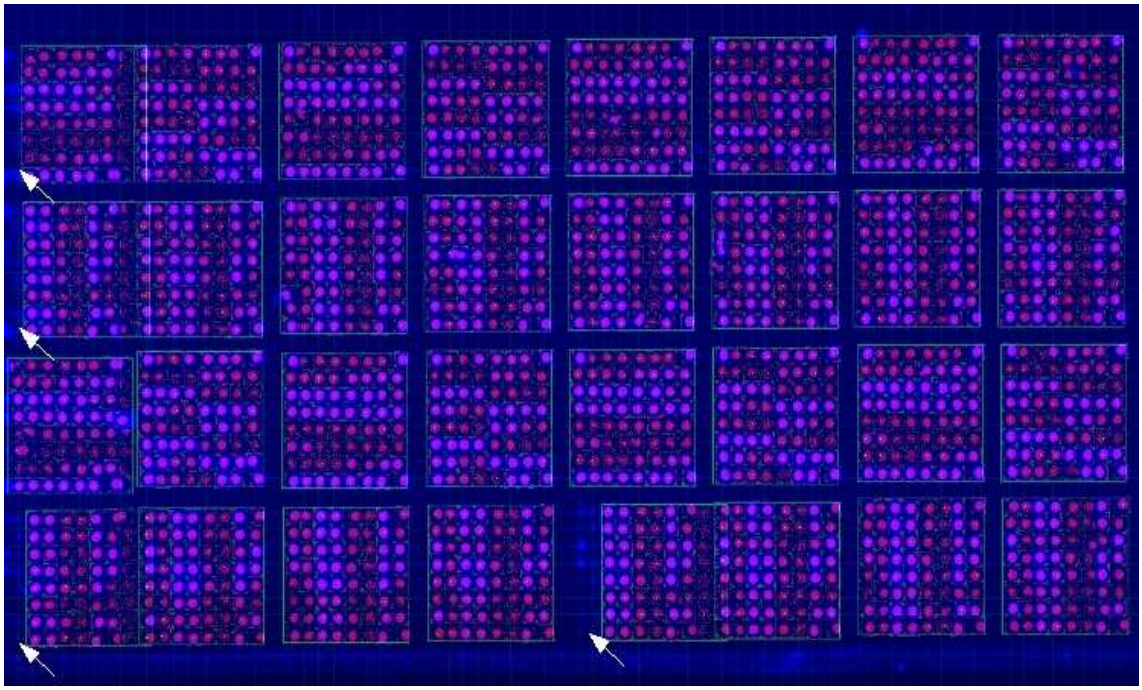


Figura 12: Erro de segmentação do *software Spot*.

Figura da segmentação de um experimento de *microarray* através do *software Spot*. Esse é um exemplo de segmentação defeituosa, como pode ser observado pelas setas brancas.

Também enfatizamos que este programa ainda não tem um nome comercial definido, o que nos motivou a chamá-lo de Bioinfo no presente trabalho. A Figura 13 mostra a *interface* gráfica principal deste *software*.

O procedimento básico de segmentação de imagens de *microarray* é descrito pela Figura 14. Nesta figura notamos que primeiro ocorre um gradeamento dos diferentes blocos que compõem a lâmina, a seguir para cada um desses blocos é feito um gradeamento semelhante para a localização dos *spots*, que são posteriormente segmentados por operadores morfológicos.

O gradeamento dos blocos feito pelo Bioinfo utiliza a projeção dos perfis verticais e horizontais dos sinais de intensidade da imagem. Esses perfis nada mais são do que o somatório das intensidades nas direções x e y, como no *Spot*, com a vantagem de que aqui são aplicados três filtros (fechamento morfológico, operador *basin* e abertura morfológica) para filtrar artefatos da imagem. Para o gradeamento dos *spots* é utilizada uma idéia semelhante, empregando filtros mais simples. Uma função de custo é aplicada com o objetivo de corrigir eventuais discrepâncias entre as linhas obtidas e as características da imagem analisada em ambos gradeamentos. As Figu-

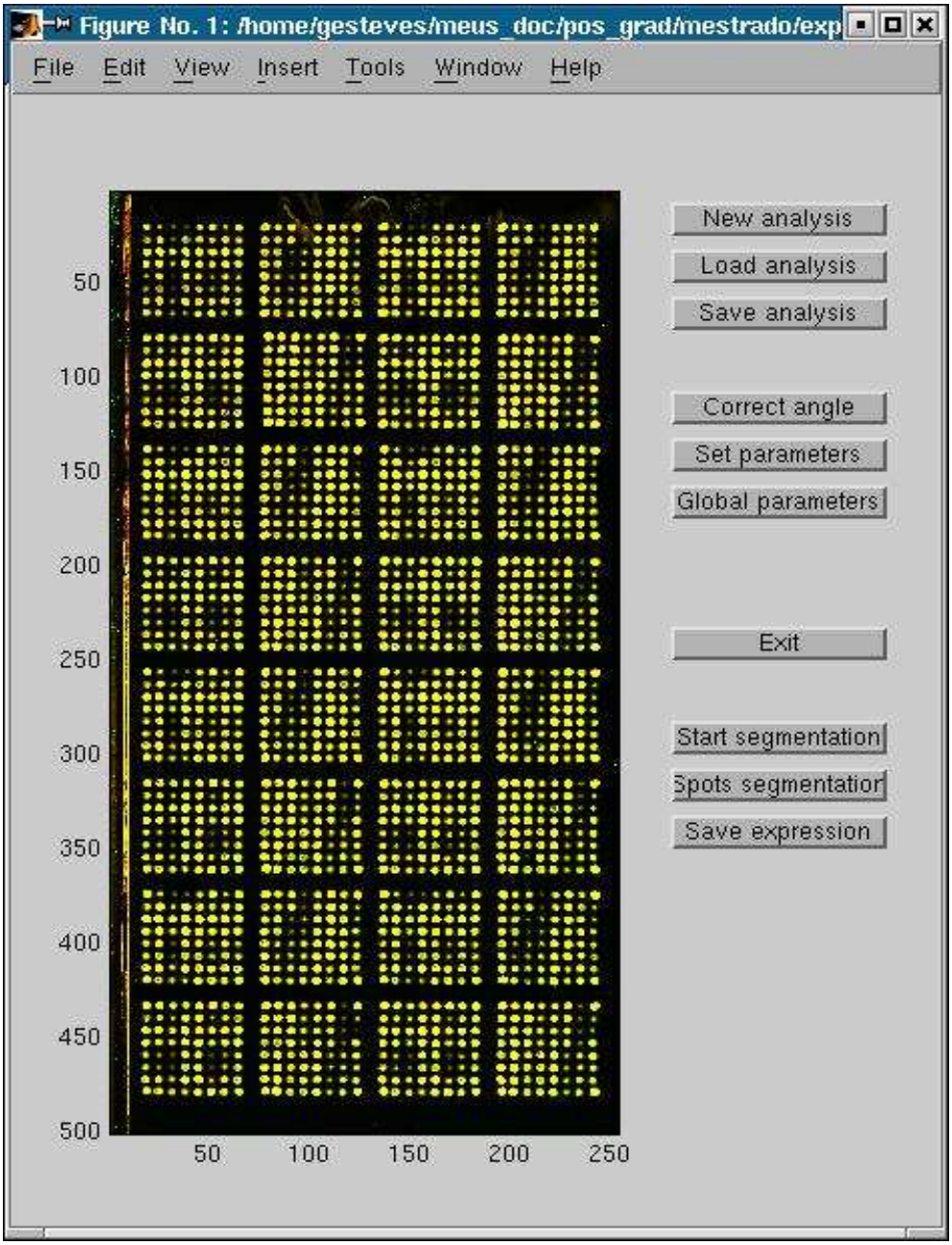


Figura 13: *Interface* principal do *software* Bioinfo.

Esta figura mostra a tela principal do Bioinfo. Aqui vemos que é possível salvar um experimento que está sendo conduzido e continuar um experimento já começado.

ras 15 e 16 mostram exemplos do processo de segmentação para os blocos e *spots*, respectivamente.

O último passo de processamento das imagens é a detecção dos *spots*. Esse é o ponto forte desse *software*, uma vez que cada *spot* também é segmentado por um operador morfológico, o operador *top-hat*. A Figura 16 também mostra um exemplo de um bloco cujos *spots* foram corretamente localizados. Essa metodologia evita a

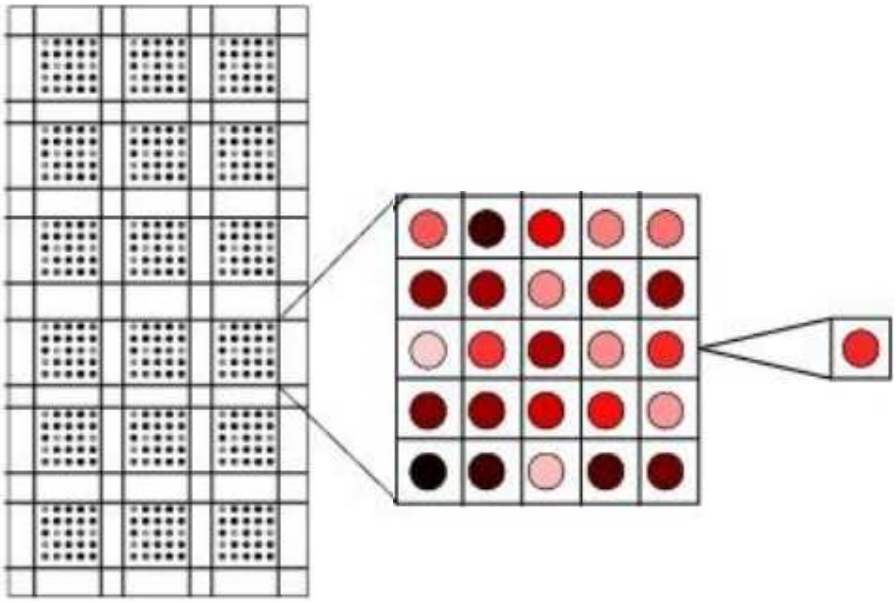


Figura 14: O processo de segmentação de imagens de *microarray* por morfologia matemática.

O processo todo é dividido basicamente em três etapas. Primeiramente é feito o gradeamento dos blocos da lâmina seguido do gradeamento dos *spots* em cada um dos blocos, por fim se dá a localização de cada *spot* através de ferramentas de morfologia matemática.

necessidade de intensa manipulação das máscaras dos *spots* por parte do usuário. Maiores detalhes de todo o procedimento de segmentação de imagens de *microarray* podem ser vistos em [15].

Para verificar a segmentação de *spots* dentro de cada bloco basta clicar com o botão direito do *mouse* e depois em *process block*. Ainda, na nova janela com a imagem individual do bloco também é possível analisar cada *spot* individualmente seguindo um procedimento similar. Um dos pontos interessantes a notar são os *scatter plots* de intensidades dos *pixels* de cada *spot*, onde podemos ter alguma informação sobre a qualidade dos mesmos. *Spots* bons (intensidade de sinal homogênea) apresentam menor dispersão dos dados ao passo que os ruins (intensidade de sinal não muito homogênea) apresentam maior dispersão, ver Figura 17.

Para fazer o gradeamento dos blocos é necessário ajustar os parâmetros globais da imagem (botão *Global Parameters*, Figura 13), onde a única informação que o usuário deve fornecer é o número de blocos (linha x coluna) e o número de *spots* dentro de cada bloco (linha x coluna). As distâncias entre blocos e *spots* podem ser ajustas interativamente na imagem. A seguir basta clicar em *Start Segmentation* para

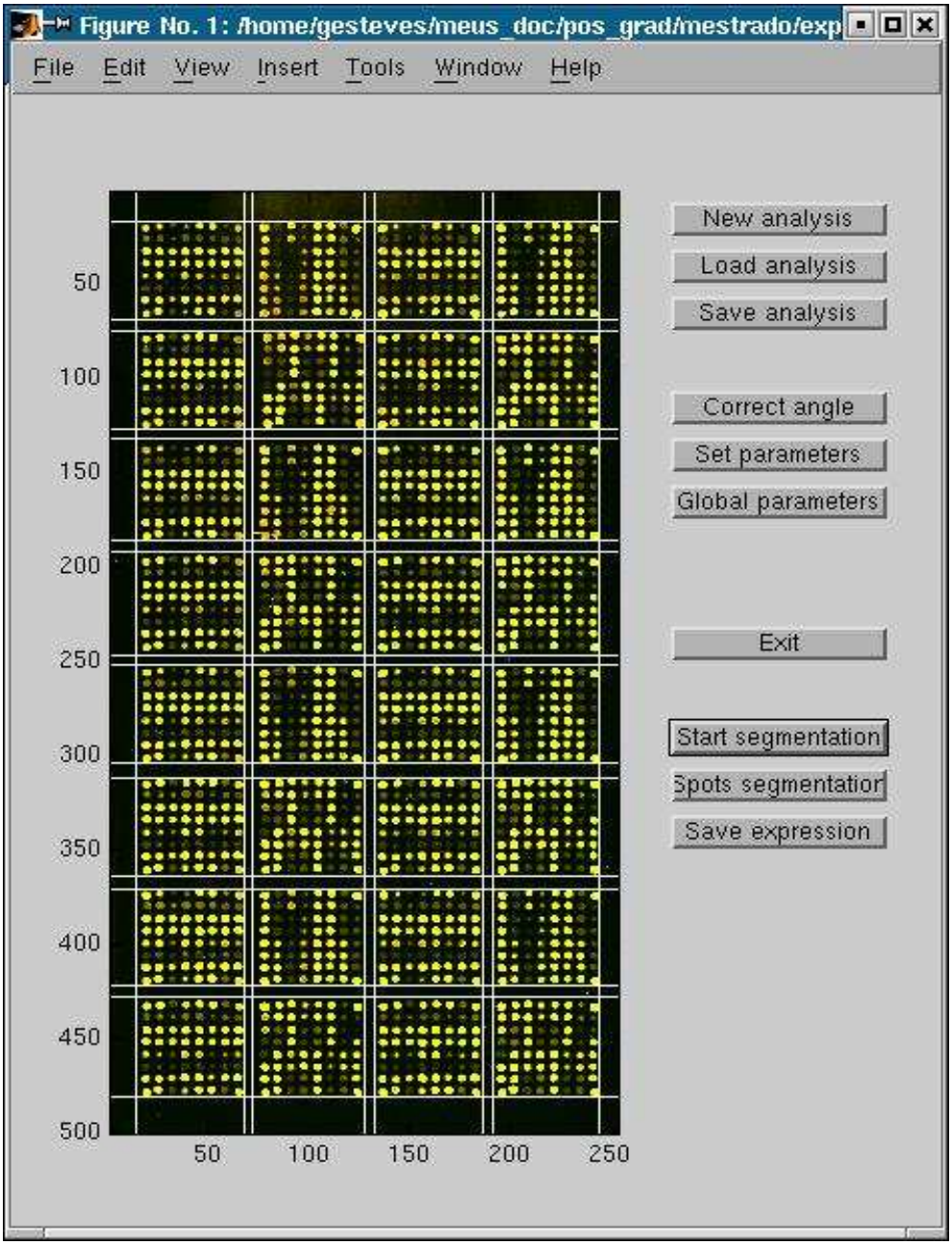


Figura 15: O gradeamento dos blocos no *software* Bioinfo.

Esta figura exemplifica o processo de gradeamento dos blocos de um experimento de *microarray* feita pelo Bioinfo. Note que todos os blocos são corretamente localizados.

a segmentação dos blocos e em *Spots Segmentation* para o gradeamento e segmentação dos *spots*. A etapa final é a exportação da tabela de dados, que é feita através do botão *Save expression*. O esquema de seleção de *pixels* utilizados para o cálculo dos valores de intensidade é bastante similar ao método de círculo fixo do *QuantArray* com a diferença de que aqui a única restrição para os *pixels* do *background* é estar fora da área seguimentada para o *spot*. A tabela de dados contém os mesmos valores

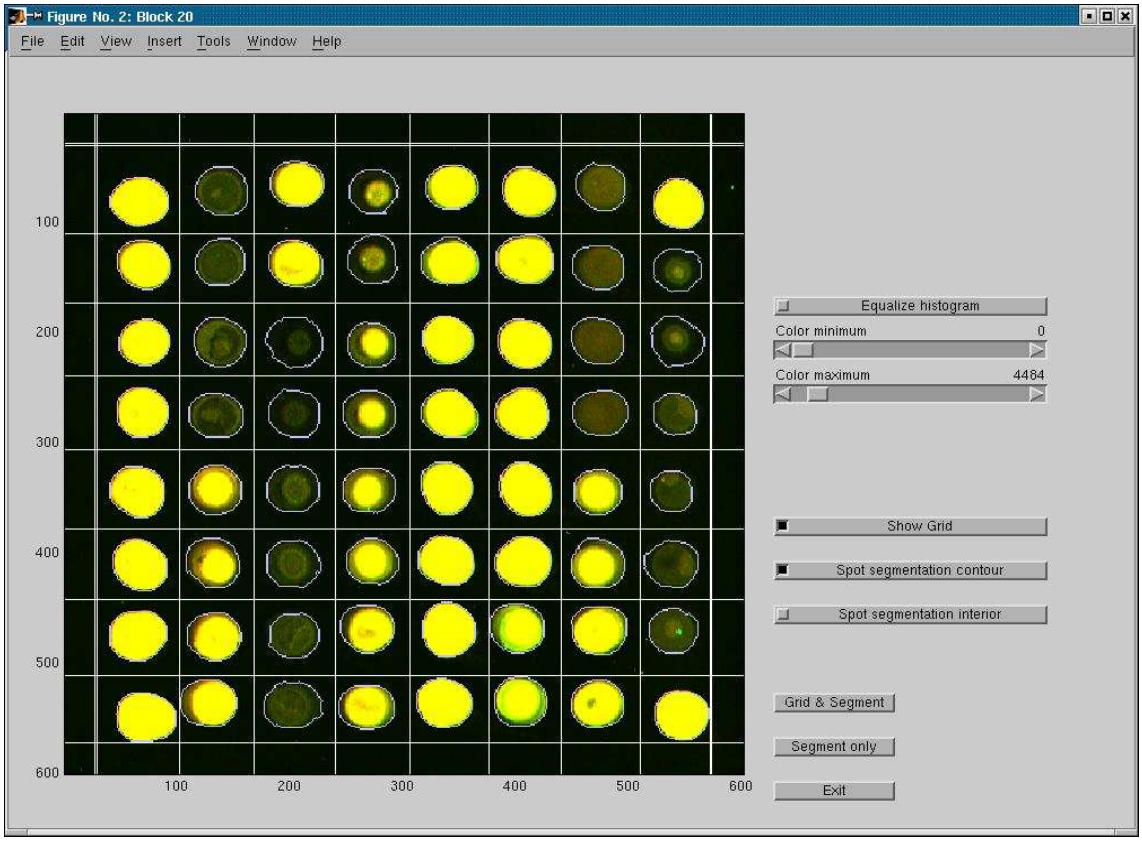
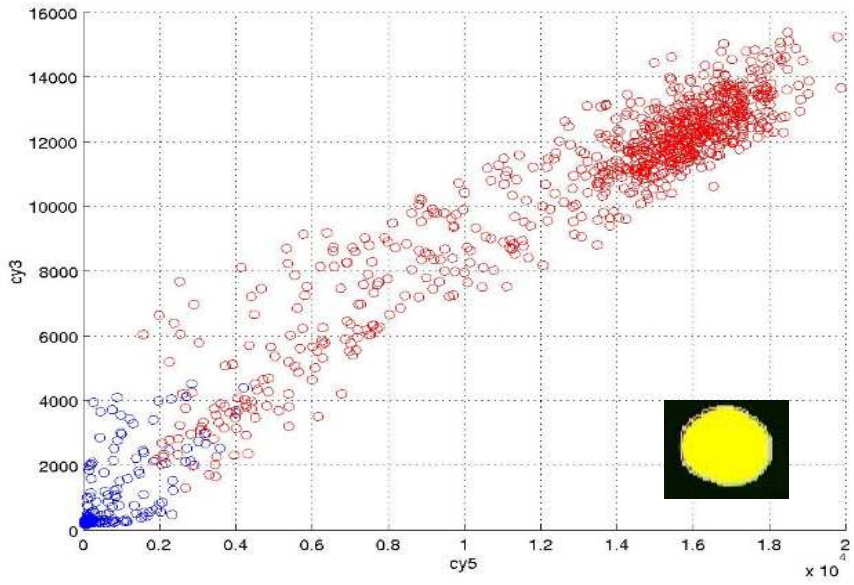


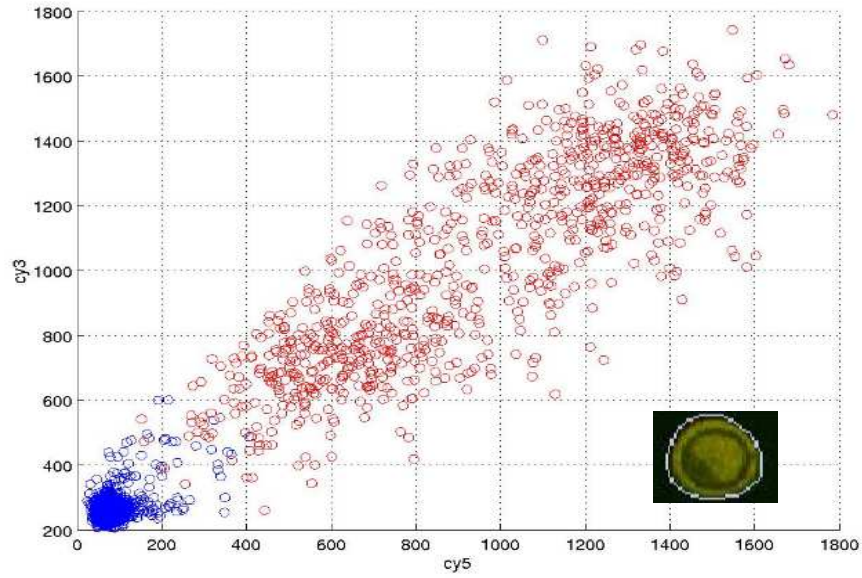
Figura 16: O gradeamento e segmentação dos *spots*.

Esta figura exemplifica o gradeamento de um bloco individual e a segmentação de *spots* desse bloco. Vê-se que todos os *spots* são corretamente localizados.

dados pelo *ScanAlyze*.



(A)



(B)

Figura 17: *Scatter plots* de *spots* individuais.

Exemplo de *scatter plots* individuais dado pelo Bioinfo, onde os pontos vermelhos representam *pixels* do *foreground* e os azuis do *background*. (A) - *Spot* bom, que apresenta boa correlação entre os dois canais. (B) - *Spot* ruim, nota-se uma pior correlação. Nos detalhes temos as imagens dos *spots*.