UNIVERSIDADE ESTADUAL DE CAMPINAS

Faculdade de Ciências Médicas

Desenvolvimento de Ferramentas de Bioinformática para Análises de Expressão Gênica em Larga Escala

Cristiane de Souza Rocha

2011
UNICAMP

UNICAMP

UNIVERSIDADE ESTADUAL DE CAMPINAS

Faculdade de Ciências Médicas

Desenvolvimento de Ferramentas de Bioinformática para Análises de Expressão Gênica em Larga Escala

Cristiane de Souza Rocha

Tese de Doutorado apresentada à Pós-Graduação da Faculdade de Ciências Médicas da Universidade Estadual de Campinas, para a obtenção do título de Doutor em Fisiopatologia Médica, área de concentração em Neurociências.

Orientadora: Profa. Dra. Iscia Lopes Cendes

2011
UNICAMP

FICHA CATALOGRÁFICA ELABORADA POR
ROSANA EVANGELISTA PODEROSO – CRB8/6652
BIBLIOTECA DA FACULDADE DE CIÊNCIAS MÉDICAS
UNICAMP

R582d

Rocha, Cristiane de Souza, 1978-
Desenvolvimento de ferramentas de bioinformática para análises de expressão gênica em larga escala. / Cristiane de Souza Rocha. -- Campinas, SP : [s.n.], 2011.

Orientador : Iscia Lopes Cendes
Tese (Doutorado) - Universidade Estadual de Campinas, Faculdade de Ciências Médicas.

1. Tubulações. 2. Genética. 3. Software. I. Cendes, Iscia Lopes. II. Universidade Estadual de Campinas. Faculdade de Ciências Médicas. III. Título.

Informações para Biblioteca Digital

Título em inglês: Development of bioinformatics tools for large-scale gene expression analysis
Palavras-chave em inglês:
Pipelines
Genetics
Software
Área de concentração: Neurociências
Titulação: Doutor em Fisiopatologia Médica
Banca examinadora:
Iscia Teresinha Lopes Cendes [Orientador]
François Marie Artigurnave
Sara Teresinha Olalla Saad
Ana Lucia Brunialti Godard
Maricene Sabha
Data da defesa: 22-07-2011
Programa de Pós-Graduação: Faculdade de Ciências Médicas

Banca examinadora da tese de Doutorado

Cristiane de Souza Rocha

Orientador(a) : Prof(a). Dr(a). Iscia Teresinha Lopes Cendes

Membros:

1. Prof(a). Dr(a). Iscia Teresinha Lopes Cendes
2. Prof(a). Dr(a). François Marie Artiguenave
3. Prof(a). Dr(a). Sara Teresinha Olalla Saad
4. Prof(a). Dr(a). Ana Lucia Brunialti Godard
5. Prof(a). Dr(a). Maricene Sabha

Curso de pós-graduação em Fisiopatologia Médica da Faculdade de Ciências Médicas da Universidade Estadual de Campinas.

Data: 22/07/2011

DEDICATÓRIA

Dedico este trabalho a minha mãe, Eurides de Souza Rocha, pelas oportunidades que me proporcionou, por todo apoio nos momentos difíceis e pelo grande exemplo de garra e dedicação que sempre foi.

Aos meus amigos que sempre me apoiaram e incentivaram.

AGRADECIMENTOS

Agradeço a Deus pelas oportunidades colocadas em meu caminho.

À Profa. Dra. Iscia Lopes-Cendes, por ter me recebido em seu laboratório e pela oportunidade de desenvolver esta tese.

À minha mãe que, mesmo distante, esteve sempre presentes na minha vida. Obrigado por todos esses anos de dedicação, amor, carinho, respeito e, principalmente, confiança.

À Profa. Dra. Cláudia Vianna Maurer-Morelli por participar ativamente do desenvolvimento deste trabalho fornecendo os dados necessários e discussões para o seu aprimoramento.

A todos os colegas de laboratório.

A CAPES e a FAPESP pelas bolsas e financiamentos concedidos.

SUMÁRIO

LISTA DE FIGURAS

RESUMO

A construção de perfis de expressão usando *microarrays* tornou-se um método amplamente utilizado para o estudo dos padrões de expressão gênica. Estes estudos produzem uma grande quantidade de dados que faz com que a análise seja complexa e demorada. À medida que a qualidade dos *arrays* se torna mais confiável com a introdução dos *arrays* industriais, a quantidade de dados gerados aumenta e o ponto crítico dos experimentos, passa a ser garantir uma boa análise de bioinformática. Para facilitar esta análise desenvolvemos um *pipeline* que executa todos os passos de processamento de dados, tais como a correção de *background*, controle de qualidade, normalização, detecção de genes diferencialmente expressos e análises de clusterização. Além disso, nossa ferramenta permite a escolha de diversos testes estatísticos paramétricos e não paramétricos e também faz mineração de dados, buscando as informações relevantes dos genes e gerando *links* para diversos bancos de dados públicos. A função principal desta ferramenta é auxiliar pesquisadores que trabalham com *microarrays*, pois ela facilita a análise da grande quantidade de dados que este tipo de experimento gera, podendo fornecer uma análise personalizada com a escolha dos testes estatísticos e valores de corte de acordo com os parâmetros que o usuário julgar mais apropriado para as condições de seu experimento. A facilidade de uso e aplicações múltiplas implementadas apresentadas nesta ferramenta são inéditas e esperamos contribuir de maneira significativa para estimular e facilitar o uso dos estudos de expressão em larga escala.

ABSTRACT

Expression profile using microarrays has become a widely used method for studying gene expression patterns. These studies produce a large amount of data which makes the analysis complex and time consuming. As the arrays' quality becomes more reliable the critical point remains the bioinformatics tools used to process and analyze the data generated in these large experiments. Therefore, we aimed to develop a pipeline that runs all the steps of data processing such as background correction, quality control, normalization, detection of differentially expressed genes and clustering analysis. In addition, our tool allows the choice of several parametric and non-parametric tests to be used for group comparisons. Furthermore, it can be used for data mining, searching for relevant information of genes as well as creating links to various public available databases. The main goal of this tool is to provide researchers working with microarray data a user friendly tool which includes customized statistical analysis. This type of research tool with this many function is not previously available and we hope to contribute to make large scale expression studies easier and faster, specially for researchers that are starting to use this type of technology.

1 - INTRODUÇÃO

1.1- Expressão Gênica

De acordo com Lewin (2004) a natureza hereditária de cada organismo é definida por seu genoma, que consiste em uma longa cadeia de ácidos nucleicos que fornecem a informação necessária para a "construção" do organismo, através de uma complexa série de interações. No núcleo celular de um organismo eucarioto está presente o material genético que carrega estas informações que é a molécula de DNA (*Deoxyribonucleic Acid*). Todas as informações relativas à construção e ao funcionamento de um organismo estão embutidas nesta molécula que é formada por bilhões de bases nitrogenadas complementares em formato de dupla hélice (Watson, 2004) (Figura 1).

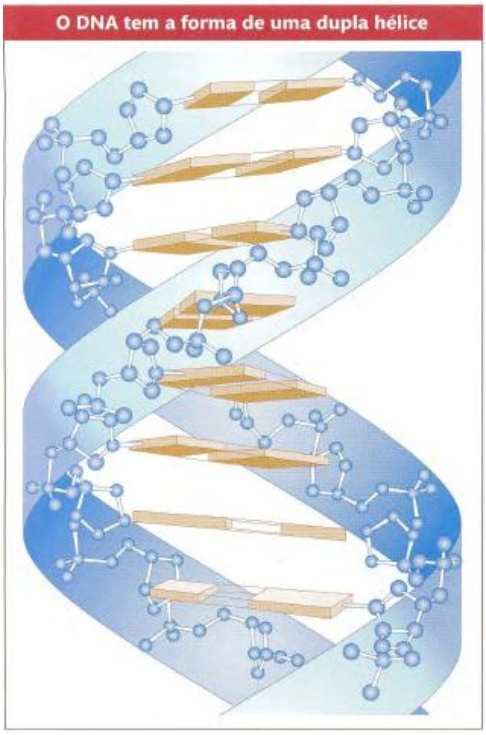


Figura 1 - Ilustração da dupla hélice formada pela molécula de DNA (imagem retirada de Lewis 2004)

O processo pelo qual a sequência de DNA é traduzida em proteína é chamado de dogma central da biologia molecular (Figura 2), em que a informação é perpetuada através da replicação do DNA, a informação contida em um gene é traduzida em estruturas presentes em determinado tipo celular através de quatro etapas: a transcrição, que converte a informação do DNA conhecido como gene em uma forma mais acessível, uma fita de RNA complementar, RNA mensageiro (RNAm); processamento do RNA, que envolve modificações do transcrito primário, permitindo a geração de uma molécula madura de RNAm, que servirá como molde e contém as informações para a síntese protéica; o transporte do RNAm para o citoplasma, onde se dará a síntese da proteína; e a tradução que converte a informação contida no RNAm em proteínas onde o RNAm liga-se a ribossomos proporcionando as informações necessárias para a síntese da proteína (Alberts, 2001).

A sequência de bases nitrogenadas é usada para produzir todas as proteínas de um organismo em um determinado momento e/ou local (Lewin, 2004). Ou seja, dependendo do tipo celular, diferentes grupos de genes estão ativos e produzindo determinadas proteínas, enquanto outros genes estão desligados.

Vários passos no processo de expressão gênica podem ser modulados, incluindo a transcrição do RNAm e a modificação pós-traducional de uma proteína. A regulação gênica dá à célula controle sobre sua estrutura e função e é a base para a diferenciação celular, morfogênese e para a versatilidade e adaptabilidade de qualquer organismo. A regulação gênica pode também ser responsável por mudanças evolutivas, ou surgimento de doenças, pois o

controle do momento, localização e quantidades de expressão gênica podem ter um efeito profundo nas funções do gene num organismo (Griffiths *et al.* 2006).

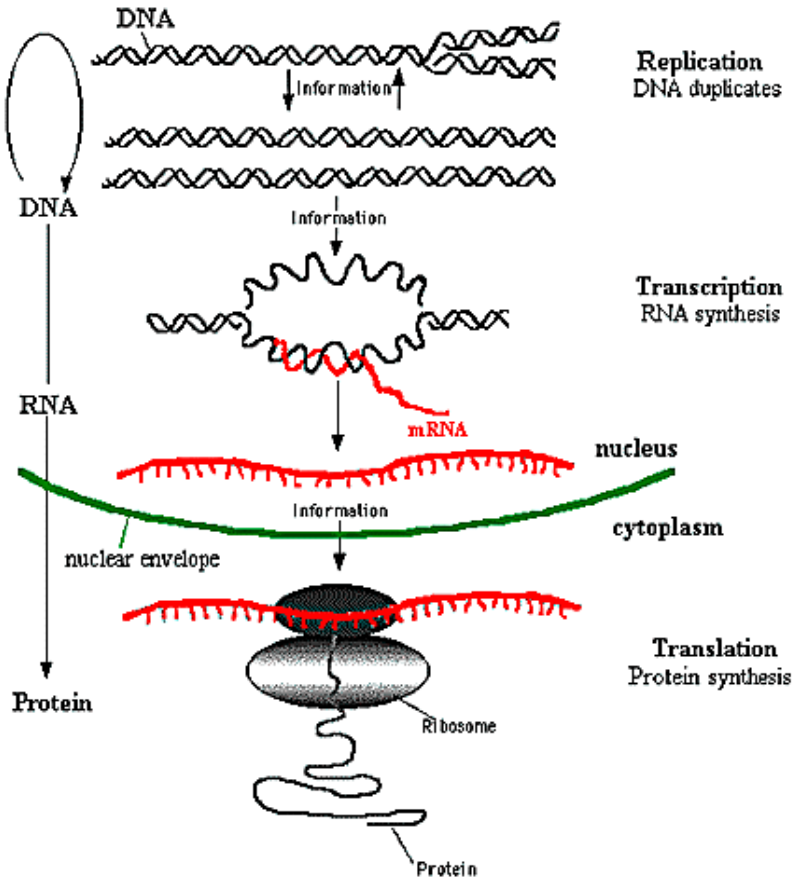


Figura 2 - Dogma Central da Biologia Molecular (imagem retirada de Access Excellence The National Health Museum)

No entanto existem moléculas de RNA que não são traduzidos em proteínas chamados RNAs não codificantes (ncRNAs), O genoma dos mamíferos codifica milhares de ncRNA (Pang *et al.*, 2007) como o RNA

transportador e o RNA ribossômico. E mais recentemente foi descoberto o microRNA.

MicroRNAs

MicroRNAs (miRNAs) são pequenos RNAs endógenos com cerca de 21 a 24 nucleotídeos de comprimento, que se ligam ao RNA mensageiro de forma sequência-específica tornando-o dupla fita, e assim regulam a expressão gênica pós-transcricionalmente (Bartel, 2004). Essa regulação pode ser realizada através da inibição da tradução ou da degradação do RNA mensageiro. Os miRNAs são diversificados em sequência e padrões de expressão, sugerindo que eles possam participar em uma ampla gama de vias regulatórias de expressão (Ambros, 2001).

Recentemente foi desenvolvido pela *Affymetrix*^{TM} um *microarray* comercial com sequências de microRNA para análise de sua expressão.

1.2- Bioinformática

Com a descoberta de Watson e Crick em 1953, de que o DNA é estruturado como dupla hélice e a posterior descoberta do código genético e do fluxo de informação biológica dos ácidos nucleicos para proteínas, definiu-se um novo ramo da biologia, a biologia molecular, mas ainda não era possível ler a informação guardada no código genético (Watson, 2004).

Na década de 40 foi inventado o computador digital, que foi chamado de digital por utilizar dígitos binários (zeros e uns) para armazenar informações e sua lógica também é digital baseada no fundamento de ligado ou desligado. O que de certa forma fez com que ele se tornasse o parceiro ideal para lidar com dados genéticos, pois a informação genética também é digital, só que com um alfabeto quaternário A, T, C, G (Adenina, Timina, Citosina, Guanina), as bases nitrogenadas que armazenam a "informação genética". Interessantemente, até certo ponto, os genes também tem um comportamento digital, pois podem ser "ligados" ou "desligados". A pesar dessas similaridades, somente após muitos anos surgiu a bioinformática, pois foi preciso esperar até a década de 80 para o aparecimento dos primeiros equipamentos que permitissem a "leitura", ou sequenciamento do código genético. Além disso a computação também precisou evoluir, com máquinas capazes de armazenar cada vez mais dados e processá-las com maior velocidade. Com esta evolução paralela, foi na década de 90 que a bioinformática ganhou destaque no mundo científico (Setubal, 2003).

Pode-se definir bioinformática como a aplicação da tecnologia da informação para o gerenciamento e manipulação de dados biológicos (Gibas & Jambeck, 2001). Isso envolve o desenvolvimento e o uso de ferramentas

computacionais visando à expansão e facilitação do uso de dados biológicos através da capacidade de adquirir, armazenar, organizar, analisar e visualizar tais dados. É uma disciplina em rápida evolução, nas últimas três décadas, o armazenamento de dados biológicos em bancos de dados públicos tem se tornado cada vez mais comum, e esses bancos de dados têm crescido exponencialmente (Figura 3). Atualmente há aproximadamente 126.551.501.141 bases em 135.440.924 registros de seqüência no GenBank (Benson, 2011).

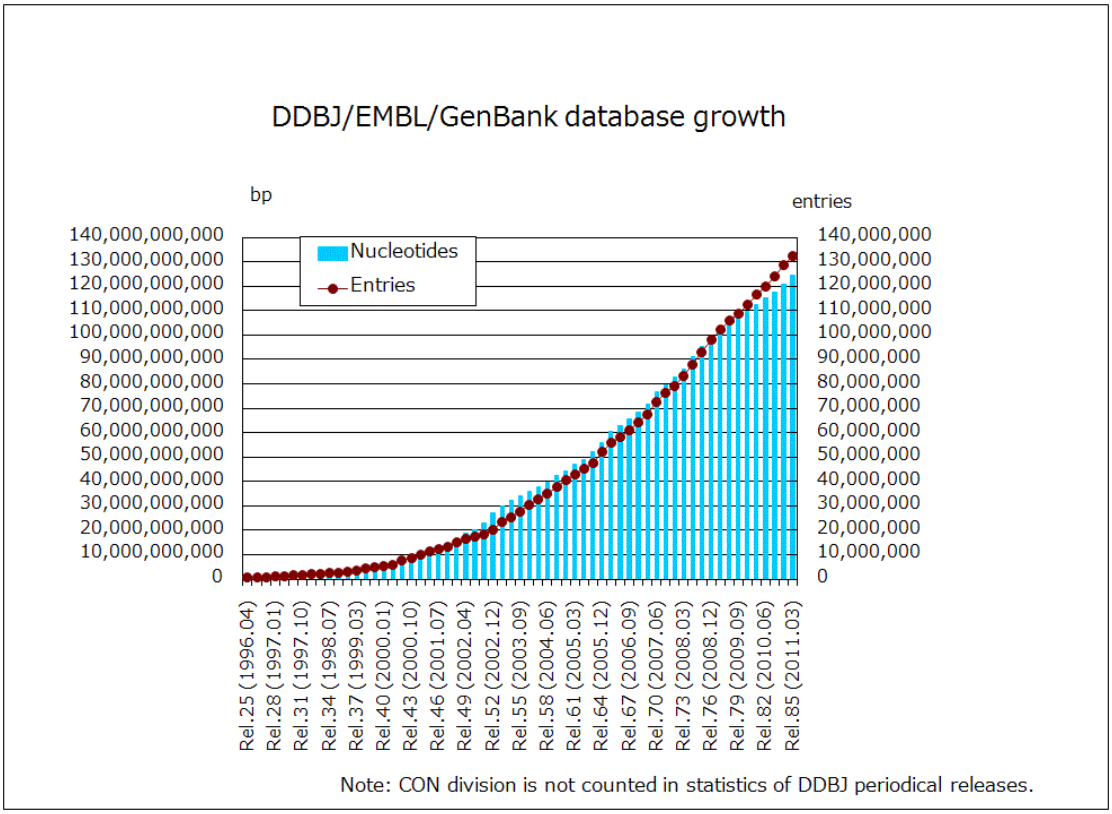


Figura 3 - Crescimento dos bancos de dados de sequências DDBJ (DNA Data Bank of Japan), EMBL (European Molecular Biology Laboratory) e GenBank (imagem retirada do site http://www.ddbj.nig.ac.jp/)

Com o auxílio da bioinformática é possível extrair informações relevantes a partir das sequências de DNA e de proteínas, obtidas pelo processo de seqüenciamento automático de nucleotídeos e de aminoácidos. A análise

computacional pode desvendar detalhes e revelar arranjos na organização de dados genômicos, ajudando a esclarecer a estrutura e a função dos genes e proteínas estudados (Baxevanis & Ouellette, 2005).

Atualmente, a utilização e desenvolvimento de modelos matemáticos e programas computacionais tem sido uma importante ferramenta na área da genética humana e médica, os quais têm auxiliado na determinação de parâmetros para o estudo de diversas doenças, cada vez mais complexas, tanto do ponto de vista clínico quanto genético. Muito tem sido desenvolvido nesta área e ferramentas computacionais são imprescindíveis para as análises.

1.3- Microarranjos de DNA – *Microarrays*

A conclusão do sequenciamento do genoma de cada vez mais organismos, fez com que o foco de pesquisa fosse transferido do sequenciamento, para a definição das funções biológicas de todos os genes codificados dentro do genoma de um organismo em particular. As metodologias de pesquisa biológica evoluíram do paradigma um gene em um experimento, para múltiplos genes em um experimento.

Um DNA *microarray*, consiste num arranjo pré-definido de moléculas de DNA (fragmentos de DNA genômico, cDNAs ou oligonucleotídeos) quimicamente ligadas a uma superfície sólida com compostos que conferem carga positiva. São utilizados na detecção e quantificação de ácidos nucléicos (RNAm na forma de cDNA ou DNA genômico) provenientes de amostras biológicas, as quais são hibridizadas com o DNA fixado no *array* -hibridização por complementaridade de bases (Coe & Antler, 2004).

A primeira forma de *array* foi o *Southern blot*, desenvolvido em 1975 pelo Dr. Edward Southern da universidade de Edimburgo. Nesta técnica, o DNA a ser analisado é digerido com enzimas de restrição e, em seguida, separado por eletroforese em gel de agarose, os fragmentos de DNA no gel são desnaturadas com solução alcalina e transferidos para um filtro de membrana de nitrocelulose ou nylon, preservando a distribuição dos fragmentos do DNA no gel, o filtro de nitrocelulose é incubado com uma sonda específica marcada e localização do fragmento de DNA que hibridiza com a sonda pode ser exibida por autoradiografia (Southern 1975).

Tecnologia de *arrays* já estava em uso no começo dos anos 1980 (Augenlicht 1984), mas não entrou em destaque até meados dos anos 1990,

quando foi desenvolvido o primeiro *array* miniaturizado em 1995 (Schena *et al.* 1995) onde os *microarrays* de cDNA surgiram como uma ferramenta de biologia molecular nova capaz de sondar o transcriptoma completo de uma célula (DeRisi *et al*. 1996; Lockhart, *et al*. 1996). Atualmente a tecnologia de *microarray* se tornou uma das ferramentas indispensáveis que muitos pesquisadores usam para monitorar níveis de expressão ao longo do genoma em um dado organismo.

A primeira tecnologia utilizada foi o *microarray* de cDNA, que consiste em inúmeras sondas de fragmentos de cDNA amplificados em PCR e depositado em um padrão de matriz de pontos em uma superfície de vidro tratado (Figura 4). O alvo para essas sondas é uma solução de cDNA derivados do RNAm extraído de duas populações de células ou tecidos (ex.: tratamento e controle) marcado com fluoróforos diferentes Cy3 (Cianina 3) verde e Cy5 (Cianina 5) vermelho. As duas amostras são hibridizadas no mesmo *microarray,* que é chamada de hibridização competitiva, e escaneados com dois comprimentos de ondas diferentes relativos aos fluoróforos utilizados (Šášik *et al.* 2004), dependendo dos níveis de expressão das amostras se terá um padrão de coloração diferente para cada *spot*. A principal vantagem desta tecnologia é que ela elimina a necessidade de seqüenciamento de genomas inteiros uma vez que permite a avaliação da expressão gênica em organismos para os quais dados de seqüência são pouco disponíveis (Rathod *et al*. 2002). O que está mudando agora com as novas gerações de sequenciadores automáticos.

Os *microarrays* de cDNA se caracterizam por utilizarem longas sequências de DNA (sondas) fixadas nos *spots* dos *arrays*. Atualmente sondas de cDNA

para produção *microarrays* podem ser geradas a partir de bibliotecas de cDNA disponíveis comercialmente, podendo-se ter uma representação próxima do genoma completo do organismo a ser analisado (Coe & Antler, 2004). No entanto a densidade (quantidade de sondas diferentes) deste tipo de *array* em comparação ao *array* de oligonucleotídeo é bastante menor.

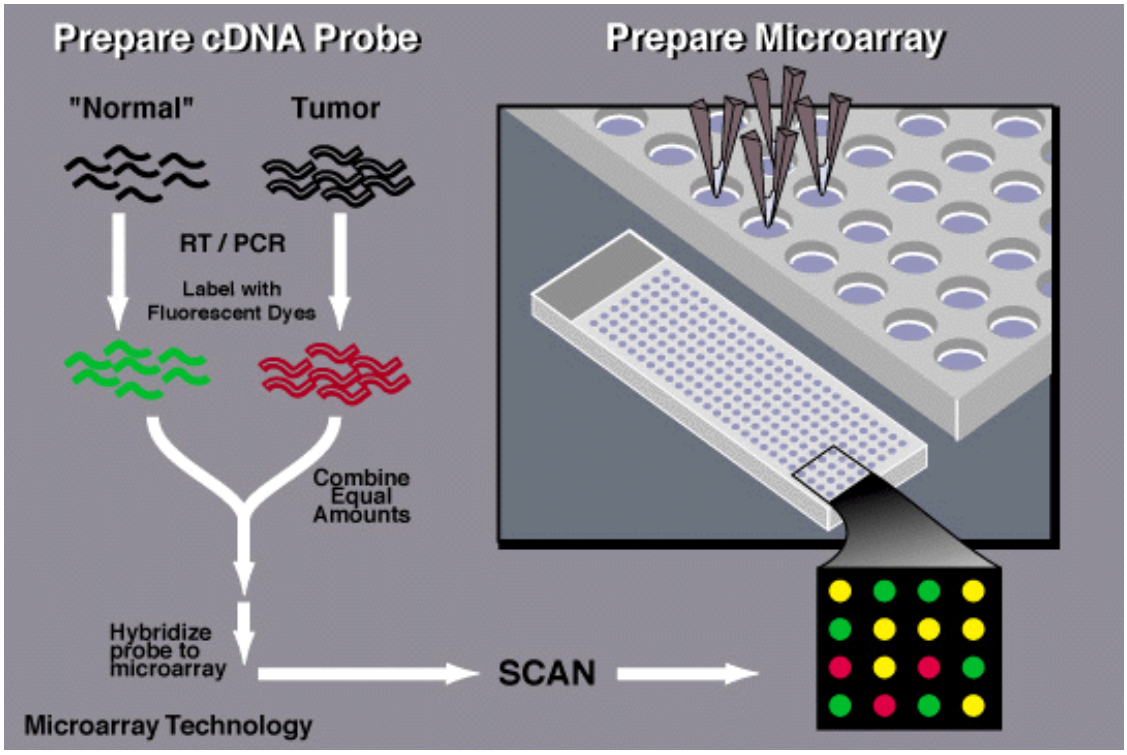


Figura 4 - Ilustra a preparação, hibridização e escaneamento do *microarray* de cDNA. (imagem retirada do site www.genome.gov)

A próxima tecnologia de *microarray* a surgir foi desenvolvida pela empresa Affymetrix™ com síntese *in-situ* utilizando a tecnologia de fotolitografia e química combinatória (Figura 5) Este processo consiste basicamente em bloquear a luz ultravioleta dos produtos químicos na superfície sólida usando uma máscara, pois a luz remove a proteção química que impede a ligação dos nucleotídeos fixados, inserindo novos nucleotídeos que se ligarão aos que estão desprotegidos (Pease *et al*. 1994). Permitindo assim a síntese específica

dos oligonucleotídeos, é uma técnica que permite a produção simultânea de milhares de sondas de forma relativamente rápida. (Coe & Antler, 2004)

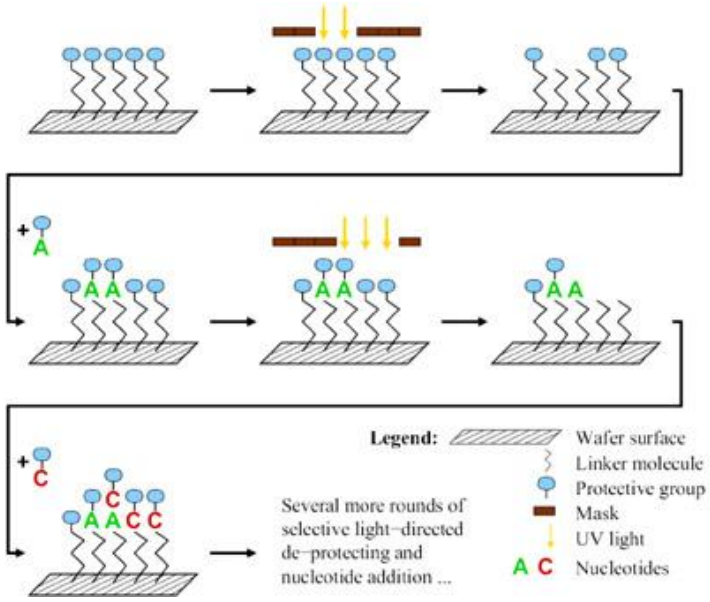


Figura 5 - Esquema da tecnologia de fotolitografia (imagem retirada do site http://www.rahmannlab.de/research/microarray-design)

Esta técnica se utiliza de uma série de sondas (*probes*) distintas (11-20) para cada gene alvo, essas sondas são denominadas coletivamente como *probe-set* (Figura 6). Cada sonda consiste em milhões de fitas simples de DNA de comprimento (25 pares de base) e sequência definidos, confinado a uma pequena área quadrada, que alinham com a sequência de um gene alvo, além disso, na plataforma Affymetrix para os chips modelo 3' IVT as sondas são fixadas aos pares (não no mesmo *spot*) sendo denominadas Perfect-Match (PM) e Miss-Match (MM) (Figura 6) (Irizarry *et al.* 2003a). PM é a sonda que alinha perfeitamente com a sequência do gene que se quer interrogar e MM é a sonda que só difere da PM pelo nucleotídeo do meio da sonda e é utilizada como correção da intensidade do sinal. É possível escolher entre computar as intensidades apenas com a intensidade do sinal das sondas PM ou pela

subtração da intensidade do sinal de PM pela intensidade do sinal MM (PM – MM) (Irizarry *et al.* 2003a). Para os outros modelos de chips a Affymetrix possui apenas as sondas PM.

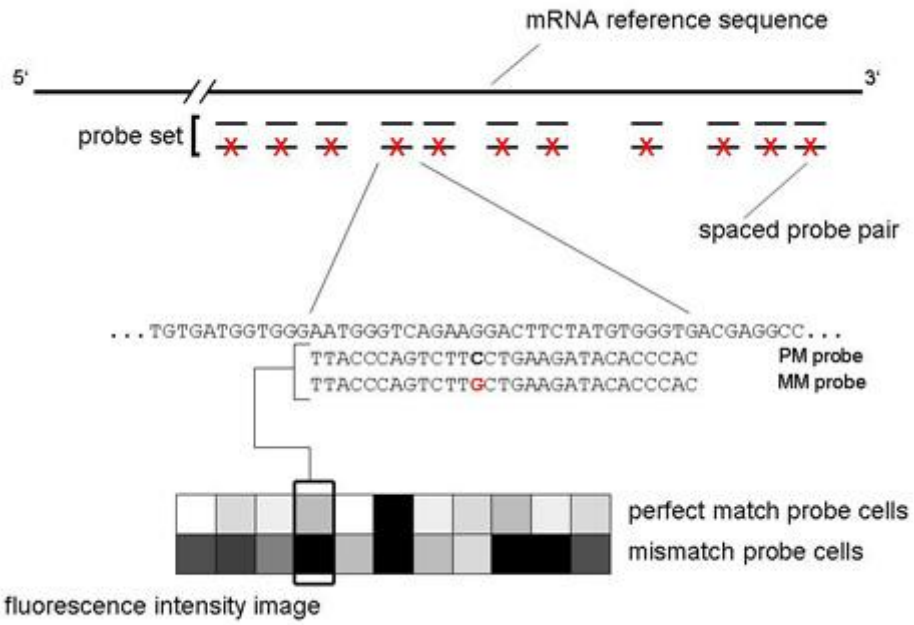


Figura 6 - Esquema de desenho de um probe-set (imagem retirada do site http://www.dkfz.de/gpcf/24.html)

Cada amostra deve ser hibridizada em um *microarray* diferente, não sendo mais possível a hibridização competitiva.

Após a comercialização da tecnologia de *microarray*s, muitos pesquisadores abandonaram a produção de seus próprios *arrays*, então a ênfase da pesquisa foi transferida do desenvolvimento de *arrays* para a sua análise como aquisição de imagem, quantificação, normalização e análises estatísticas (Šášik *et al.* 2004).

A tecnologia dos *microarray*s tem impulsionado de maneira importante a pesquisa de genômica funcional dos diferentes organismos, desde bactérias

até o homem, incluindo situações normais e patológicas (câncer, doenças autoimunes, doenças degenerativas entre outras).

A bioinformática desempenha múltiplos papéis em experimentos de *microarray*. De fato, é difícil imaginar a utilização de *microarrays* sem o envolvimento de computadores e bancos de dados, desde o desenvolvimento de chips para fins específicos, até a quantificação de sinais, e a detecção de grupos de genes com perfis de expressão ligados e/ou diferencialmente expressos, pois a quantidade de dados que envolve estes tipos de experimentos é gigantesca, principalmente com os *microarrays* industriais.

1.4- Testes estatísticos

Os estudos de m*icroarray* frequentemente têm o objetivo de identificar genes que são regulados diferencialmente em diferentes classes de amostras, ou encontrar genes marcadores que discriminam doentes de indivíduos saudáveis. Para este fim devem-se utilizar os testes estatísticos, mas deve-se tomar muito cuidado, pois o grande número de genes presentes em um único *array* faz com que o pesquisador leve em conta o problema de testes múltiplos (Dudoit *et al.* 2003)

O padrão de p-valor foi definido para testar hipóteses individuais e há um problema evidente na análise dos dados gerados através da expressão de genes por *microarrays*, uma vez que normalmente envolve o teste de várias dezenas de milhares de hipóteses simultaneamente (Shaffer, 1995). Mesmo se o p-valor estatístico atribuído a um dado gene indique que é extremamente improvável que a expressão diferencial deste gene seja devido a efeitos aleatórios, o número muito elevado de genes no *array* faz com que as expressões diferenciais de alguns genes representem falsos positivos (Dudoit *et al.* 2003) e muitas publicações de *microarray* apresentam p-valores, como se cada gene tivesse sido testado isoladamente. Uma das melhores maneira de especificar a confiança dos resultados de *microarray* é a taxa de falsa descoberta (*false discovery rate* - FDR). FDR de um conjunto de previsões é a porcentagem esperada de falsas previsões neste conjunto. A FDR é determinada a partir da distribuição do p-valor observada e, portanto, é adaptável à quantidade de sinal dos dados (Benjamini & Hochberg 1995).

Com relação à distribuição do sinal dos *microarrays*, a primeira coisa a notar é que a maioria dos genes é expressa em níveis muito baixos e alguns genes são expressos em número elevado de cópias. Estatísticos geralmente lidam com dados altamente distorcidos em uma escala logarítmica (Babu, 2002). Às vezes, a distribuição de intensidades aparece aproximadamente em forma de sino, após uma transformação, no entanto, dependendo de como o *background* é tratado e como os genes de baixa abundância são estimados, a distribuição de intensidades do *microarray* pode aparecer distorcida, mesmo em uma escala logarítmica. A melhor explicação para a forma típica da distribuição de sinal de *microarray* é que o sinal de cada gene é uma combinação de hibridização do gene mais algumas hibridizações não específicas, ou transcrições parciais na amostra, mais ruído (Figura 7).

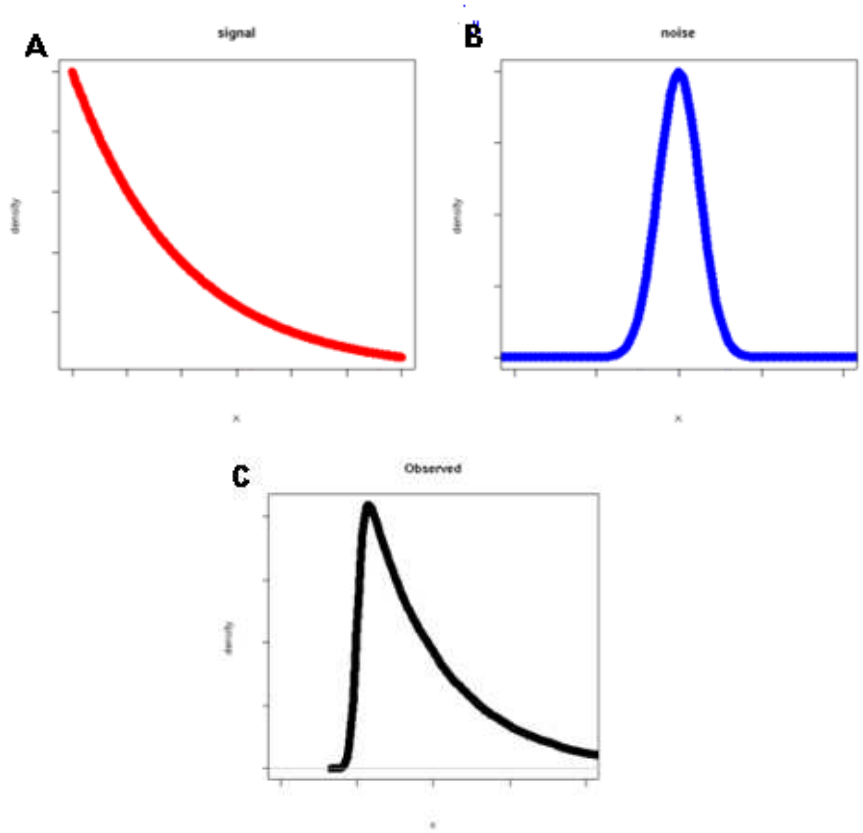


Figura 7 - Distribuição da intensidade do sinal de expressão (A), do ruido no chip (B) e o sinal observado na análise (C)

Dados de *microarrays* são freqüentemente utilizados como um guia para outros estudos mais precisos da expressão gênica como PCR em tempo real ou outros métodos. Então, o objetivo da análise estatística é heurística: proporcionar ao pesquisador uma lista ordenada de genes bons candidatos para acompanhamento.

A análise estatística de dados de *microarray* geralmente envolve três componentes: 1-Obtenção dos dados e eliminação de valores ilegítimos: correções de *background*, pré-ajuste dos dados para efeitos sistemáticos - normalização dos dados. 2- Análise estatística propriamente dita, a qual geralmente utiliza ferramentas metodológicas relativas a testes de significância, para detecção de genes diferencialmente expressos. 3- Análise de agrupamento, clusterização e classificação.

1.4.1. Obtenção de dados e correção pelo *background*

Após a hibridização o *microarray* é "lido" por um *scanner* que mede a intensidade do sinal das moléculas marcadas, a imagem digitalizada é transformada em uma planilha com os valores da média da intensidade dos pixels (menor ponto que forma uma imagem digital).

O processamento da imagem é diferente para cada tipo de *microarray* (Figura 8), no *microarray* de cDNA este processo pode ser dividido em quatro passos básicos: 1 – Localização dos *spots.* 2 – Segmentação da imagem, categorizando cada pixel como sinal dos *spots* ou sinal de *background*. 3 – Quantificação: determinar um valor para intensidade de sinal e para *background*. 4 – Verificação da qualidade de cada spot. Estes passos são

normalmente realizados com o auxílio de softwares especializados e pode envolver diferentes graus de interferência humana.

Processamento da imagem para GeneChips de oligonucleotídeos (ex. Affymetrix) geralmente é feito usando o software proprietário do fabricante. Existem duas grandes diferenças entre estes e os *microarrays* de cDNA: toda a superfície de um GeneChip é coberta com células em forma de quadrado contendo sondas e como as sondas são sintetizadas no chip em locais precisos a localização dos spots e segmentação da imagem deixam de ser um problema pois suas posições são conhecidas.

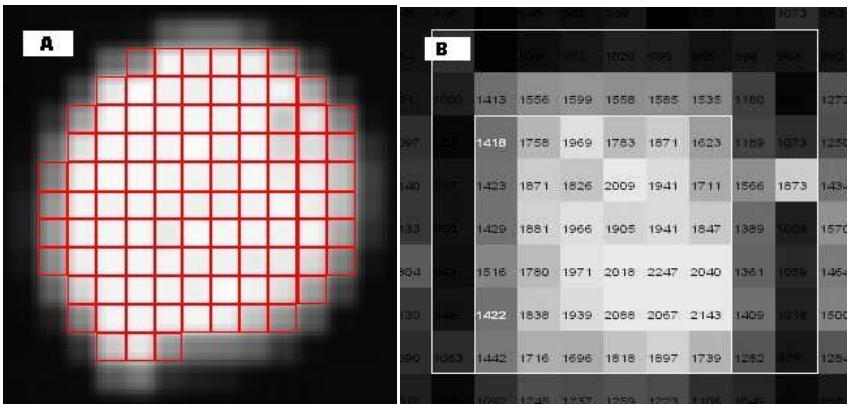


Figura 8 - Imagem de detecção de intensidade de sinal de *microarray* de cDNA(A) e de oligoarray(B)

A correção pelo *background* é definida como sendo a aplicação de metodologias com o objetivo de corrigir ou eliminar os ruídos de *background* (Bolstad 2004), ajustar as medidas de correção para uma escala adequada de forma que possa haver uma relação linear entre a intensidade do sinal e a expressão dos genes no *array*. Em experimentos de *microarrays* de cDNA esta correção é normalmente feita com os valores de intensidade que ficam em volta do *spot*, subtraindo a média destes valores da média dos valores do *spot* .

Já em *oligoarrays* como os da Affymetrix esta correção é feita utilizando os próprios valores de expressão medidos nos *spots*. Para o cálculo do *background*, é estabelecido um "piso" que será subtraído do valor de cada célula. Para este cálculo, o *array* é dividido em K zonas retangulares (o padrão é 16 zonas), as células são ranqueadas e 2% das que possuírem o menor valor de intensidade são escolhidas como *background* para esta zona. O desvio padrão destes 2% é calculado como uma estimativa da variação do *background* para cada zona. Para correção de ruído, um valor local de ruído é estimado baseado no desvio padrão dos 2% menores valores de *background* da zona (Affymetrix, 2002). Em seguida, um limite e um piso são estabelecidos por uma fração do valor do ruído local, de modo que nenhum valor é ajustado abaixo desse limiar. O valor do sinal é calculado com a combinação do *background* ajustado e os valores de PM e MM dos *probe sets*. O sinal é calculado da seguinte forma:

1. Intensidades das células são pré-processados para o *background* global.

2. Um valor de *mismatch* ideal é calculado e subtraído para ajustar a intensidade da PM.

3. As intensidades ajustadas PM são transformadas em logaritmo para estabilizar a variância.

4. Finalmente, o sinal é escalado com uma média aparada (Affymetrix, 2002) (Esse tipo de média corta os extremos reduzindo assim o efeito de *outliers* por remover uma proporção especificada da observação maior e menor).

A razão para a inclusão de uma sonda de MM é fornecer um valor que inclui a maior parte do *background* de hibridação cruzada e um sinal disperso

que afetam a sonda PM. Se o valor MM é inferior ao valor de PM, é uma estimativa fisicamente possível para o *background*, e pode ser usado diretamente (Affymetrix, 2002).

1.4.2. Normalização

Um dos passos mais importantes para a correção das medidas de expressão gênica é a normalização, que é o processo de remoção de variações indesejadas que afetam as medidas de intensidade da expressão, devido a pequenas diferenças na quantidade de RNA e flutuações geradas na própria técnica de *microarray* (Yang *et al.* 2002). O objetivo da normalização dos dados é minimizar os efeitos causados pelas variações de técnica e, como conseqüência, permitir que os dados sejam comparáveis, a fim de encontrar mudanças verdadeiramente biológicas. A figura 9 demonstra como a intensidade do sinal pode mudar depois da normalização. De forma geral, pode-se dividir os métodos de normalização em duas classes: os que utilizam todos os *arrays* e os que usam um *array* como referência e a partir deste os demais *arrays* são normalizados (Bolstad, 2004)

Várias formas de normalização foram propostas por diversos autores, como o uso de razões da intensidade de hibridização, métodos lineares, métodos não lineares ou globais, uso de genes controles como parâmetros ou o uso de genes invariantes (Quackenbush, 2002; Cullane, *et al*. 2002; Yang *et al.,*2002; Hill *et al.,*2001).

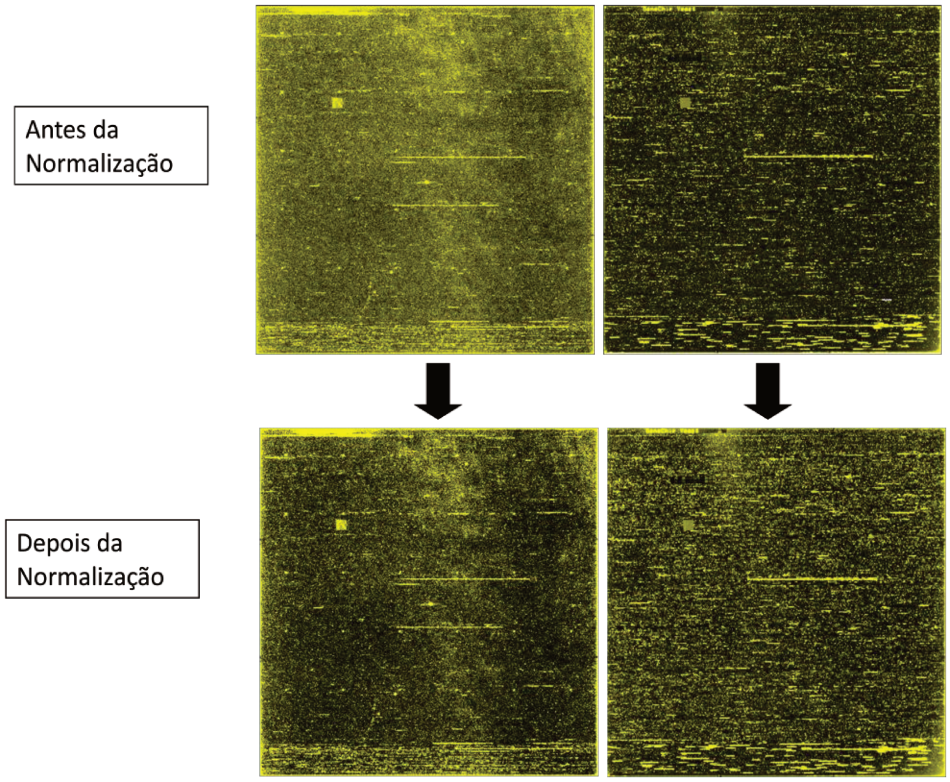


Figura 9 - Demonstração da correção de intensidade do sinal através da normalização.

1.4.3. Análise estatística

A configuração experimental mais simples e comum é comparar dois grupos: por exemplo, Controle vs. Paciente, mas a primeira coisa a se definir é o tipo do teste, se ele será paramétrico ou não-paramétrico. Os testes paramétricos são usados em amostras que possuem distribuição normal, os não-paramétricos para amostras que não se conhece a distribuição (casos com poucas amostras) ou a distribuição não é normal.

Teste de Hipóteses

Teste de Hipóteses é um método para verificar se os dados são compatíveis com alguma hipótese, podendo muitas vezes sugerir a não-validade de uma hipótese. Uma hipótese estatística é uma suposição ou

afirmação que pode ou não ser verdadeira, relativa a uma ou mais populações. A veracidade ou falsidade de uma hipótese estatística nunca é conhecida com certeza, a menos que, se examine toda a população, o que é impraticável na maioria das situações. O teste de hipóteses é um procedimento estatístico baseado na análise de uma amostra, através da teoria de probabilidades, usado para avaliar determinados parâmetros que são desconhecidos numa população (Fisher, 1925).

Os testes de hipóteses são sempre constituídos por duas hipóteses, a hipótese nula e a hipótese alternativa, a hipótese nula é a que espera a ausência do efeito que se quer verificar e a hipótese alternativa é a que o investigador quer verificar. E o nível de significância é a probabilidade de rejeitar a hipótese nula quando ela é efetivamente verdadeira (Cramer & Howitt, 2004)

O p-valor é muito utilizado para sintetizar o resultado de um teste de hipóteses e é definido como a probabilidade de se obter uma estatística de teste igual ou mais extrema quanto àquela observada em uma amostra, assumindo verdadeira a hipótese nula (Cramer & Howitt, 2004).

Um resultado para ser considerado estatisticamente significante é chamado de resultado positivo e um resultado que não é improvável sob a hipótese nula é chamado de um resultado negativo ou um resultado nulo.

Testes paramétricos:

Os testes paramétricos visam analisar a variabilidade dos resultados da variável dependente, em função da manipulação das variáveis independentes, de forma a que se possa refutar ou aceitar a hipótese nula, a qual postula que os resultados da investigação são devidos, não aos efeitos previstos pela

hipótese experimental, mas a diferenças aleatórias nos resultados, devidas a outras variáveis irrelevantes ou ao acaso (Geisser & Johnson, 2006).

Teste t de Student (t-test ou Student's t-test): teste paramétrico que utiliza duas amostras independentes. Este teste testa a diferença entre duas médias populacionais quando os desvios padrões populacionais são desconhecidos (o que ocorre na grande maioria dos casos).

Fórmula:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(s_1^2 / n_1 + s_2^2 / n_2)}}$$

Onde $\bar{x}_1$, $\bar{x}_2$ são as médias das amostras, s_1^2, s_2^2 são as variâncias das amostras e n_1, n_2 são os tamanhos das amostras (Dowdy *et al.* 2004)

ANOVA (ANalysis Of VAriance): teste de hipótese que objetiva comparar mais de duas médias. A análise de variância é um teste para comparar médias, que é realizado através das variâncias dentro e entre os conjuntos envolvidos. É uma extensão do Teste t para duas médias. Fazer múltiplos testes t de duas amostras resultaria em uma maior probabilidade de cometer erro de falso positivo. Por esta razão, ANOVAs são úteis na comparação três ou mais médias (Cui & Churchill, 2003).

Cálculo (Sewall, 1984):

1. Calcula a media de cada grupo:

$$\bar{a} = \frac{a_1 + a_2 + a_3 + \ldots + a_n}{n}$$

$$\bar{b} = \frac{b_1 + b_2 + b_3 + \ldots + b_n}{n}$$

$$\bar{c} = \frac{c_1 + c_2 + c_3 + \ldots + c_n}{n}$$

2. Calcula a media das médias:

$$\bar{x} = \frac{\bar{a} + \bar{b} + \bar{c} + \ldots}{m}$$

3. Calcula a variância das médias:

$$S^{*2} = \frac{s_a^2 + s_b^2 + s_c^2 + \ldots}{m}$$

4. Calcula a variância das amostras para cada grupo:

$$s_a^2 = \frac{(a_1 - \bar{a})^2 + (a_2 - \bar{a})^2 + \ldots (a_n - \bar{a})^2}{n-1}$$

$$s_b^2 = \frac{(b_1 - \bar{b})^2 + (b_2 - \bar{b})^2 + \ldots (b_n - \bar{b})^2}{n-1}$$

$$s_c^2 = \frac{(c_1 - \bar{c})^2 + (c_2 - \bar{c})^2 + \ldots (c_n - \bar{c})^2}{n-1}$$

5. Calcula a media da variância de todas as amostras:

$$S^2 = \frac{s_a^2 + s_b^2 + s_c^2 + \ldots}{m}$$

6. Calcula o valor do teste, estatística F:

$$F = \frac{nS^{*2}}{S^2}$$

Onde n é o número de componentes de cada grupo, m é o número de grupos S^{*2} é a variância das médias e S^2 é a média da variância de todas as amostras, a,b e c são as amostras

Testes não-paramétricos:

Teste não paramétrico é um teste que não faz suposições sobre a distribuição da amostra sob investigação. Testes não-paramétricos são geralmente muito simples de executar e, muitas vezes fazem uso de ranques (Corder *et al.*, 2009.).

***Rank Product*:** O *Rank Product* é um teste biologicamente motivados para a detecção de genes diferencialmente expressos em experimentos de *microarray* replicado. É um método não-paramétrico estatístico simples baseado em ranques dos *fold changes*. Além de seu uso em perfis de expressão, pode ser usado para combinar listas de ranqueamento em vários domínios de aplicação, incluindo a proteômica, metabolômica, meta-análise estatística, e seleção de recursos em geral (Breitling *et al.* 2004).

Fórmula:

$$RP(g) = (\Pi_{i=1}^{k} r_{g,i})^{1/k}$$

Onde dados n genes em k réplicas $r_{g,i}$ é o ranque do gene g na i-ésima réplica, o produto do ranque é calculado por média geométrica.

Teste dos sinais de Wilcoxon (*Wilcoxon's signed rank test*): Um teste não paramétrico ou de distribuição livre para testar a diferença entre duas populações utilizando amostras emparelhadas (Wilcoxon, 1945). O teste toma por base as diferenças absolutas dos pares de observações das duas amostras, ordenados de acordo com o seu valor onde cada posto (diferença) recebe o sinal da diferença original (Wild, 1988). Os seguintes passos devem ser seguidos em sua construção:

1. Calcular a diferença entre as observações para cada par.
2. Ignorar os sinais das diferenças e atribuir postos a elas.
3. Calcular a soma dos postos (S) de todas as diferenças negativas (ou positivas).

O p-valor deve ser obtido através de uma tabela especial.

Análise de correlação

No uso estatístico geral, correlação se refere à medida da relação entre duas variáveis. Neste sentido geral, existem vários coeficientes medindo o grau de correlação, adaptados à natureza dos dados.

Coeficiente de correlação de Pearson: o coeficiente de correlação de Pearson, é indicado para dados paramétricos, mede o grau da correlação e a direção dessa correlação, se positiva ou negativa, entre duas variáveis de escala métrica. Este coeficiente, normalmente representado por ρ assume apenas valores entre -1 e 1 (Edwards, 1976).

ρ = 1 Significa uma correlação perfeita positiva entre as duas variáveis.

ρ = − 1 Significa uma correlação negativa perfeita entre as duas variáveis - Isto é, se uma aumenta, a outra sempre diminui.

ρ = 0 Significa que as duas variáveis não dependem linearmente uma da outra.

Calcula-se o coeficiente de correlação de Pearson segundo a seguinte Fórmula:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y},$$

onde x e y são duas variáveis ou as médias amostrais de X e Y e s_x e s_y são os desvios-padrões da amostra de x e y.

Coeficiente de correlação de Spearman: Coeficiente de correlação de Spearman (Spearman, 1910) é uma medida de correlação não-paramétrica, ou seja, ele avalia uma função arbitrária que pode ser a descrição da relação entre duas variáveis, sem fazer nenhuma suposição sobre a distribuição de frequências das variáveis. O coeficiente de correlação de Spearman é definido como o coeficiente de correlação de Pearson entre as variáveis ranqueadas. Os valores n primários Xi, Yi são convertidos em ranques xi, yi, e ρ é calculado a partir destes ranques (Myers & Well, 2003)

o ρ é dado por:

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}.$$

1.4.4. Análise de agrupamento, clusterização e classificação

PCA – Análise de Componente principal ou Principal Component Analysis

A Análise de Componentes Principais, conhecida como *Principal Component Analysis* (PCA) é um procedimento matemático que utiliza uma transformação ortogonal ao converter um conjunto de observações de variáveis possivelmente correlacionadas em um conjunto de valores de variáveis não correlacionadas chamadas componentes principais (Jolliffe, 2002). O número de componentes principais é igual ou inferior ao número de variáveis originais. Esta transformação é definida de tal forma que o primeiro componente principal

tem uma variação tão alta quanto possível (isto é, representa o máximo da variabilidade dos dados quanto possível), e cada componente seguinte, por sua vez tem a maior variância possível, sob a restrição de que seja ortogonal (não correlacionados) aos componentes anteriores (Jolliffe, 2002). A técnica PCA essencialmente gira o conjunto de pontos ao redor de sua média tendo como objetivo o seu alinhamento com os primeiros componentes principais. Isto move a maior parte da variância possível, usando uma transformação linear, para as primeiras dimensões. Os valores nas dimensões restantes, portanto, tendem a ser altamente correlacionados e podem ser ignorados com perda mínima de informação. Trata-se de um extrator de característica não supervisionado, usado para identificar características entre classes que não têm uma definição prévia. É uma forma de identificar padrões em dados, e expressar os dados de forma a evidenciar as suas semelhanças e diferenças (Shlens, 2009). Uma vez que padrões de dados podem ser difíceis de encontrar em dados de elevada dimensão, o PCA é uma ferramenta poderosa para análise de dados. Com o uso do PCA pode-se observar o agrupamento das amostras de *microarray*, ou detectar algum problema nestes agrupamentos, pois é esperado que amostras de um mesmo grupo apareçam próximas umas das outras no gráfico.

Cluster Hierárquico

A análise de cluster ou clusterização é a atribuição de um conjunto de observações em subconjuntos (denominados clusters) de modo que as observações no mesmo cluster são similares em algum sentido. Clusterização é um método de aprendizagem não supervisionada e uma técnica comum para a análise dos dados estatísticos, utilizados em muitos campos, incluindo o

aprendizado de máquina, mineração de dados, reconhecimento de padrões, análise de imagens, recuperação de informação e bioinformática.

Algoritmos hierárquicos encontram grupos sucessivos usando clusters previamente estabelecidos. Esses algoritmos são geralmente ou aglomerativos ("*bottom-up*") ou divisivos ("*top-down*") (Hastie *et al*. 2009). Algoritmos aglomerativos começam com cada elemento como um grupo separado e os funde em conjuntos sucessivamente maiores. Algoritmos divisivos começam com o conjunto completo e avança para dividi-lo em conjuntos sucessivamente menores. Um passo importante na clusterização hierárquica é selecionar uma medida de distância, o que irá determinar como a semelhança de dois elementos é calculada. Isso vai influenciar na forma dos clusters, como alguns elementos podem estar perto um do outro de acordo com uma distância e mais longe de acordo com outra (Hastie *et al*. 2009). A medida de distância mais comumente usada é a distância euclidiana.

Os clusters hierárquicos ajudam bastante na visualização dos dados pois agrupa os genes que possuem um padrão de expressão semelhante, ele é geralmente feito com os genes selecionados como diferencialmente expressos para verificar perfis de expressão característicos nas amostras, (Eisen *et al.* 1998), mas pode ser feito com toda a amostra.

Self-Organizing Maps - SOM

Um mapa auto-organizável do inglês *Self-Organizing Map* – SOM representa o resultado de um algoritmo de quantização vetorial que coloca um número de vetores de referência em um espaço de dados de entrada de alta dimensão para agrupar o conjunto de dados de uma forma ordenada, é usado

para visualizar as relações métricas ordenadas das amostras (Kohonen 2001). Muitos campos da ciência adotaram o SOM como uma ferramenta padrão de análise: estatística, processamento de sinais, teoria de controle, análise financeira, física experimental, química e medicina (Kohonen 2001).

Diagrama de Venn

Diagramas de Venn são diagramas que mostram todas as relações lógicas hipoteticamente possíveis entre uma coleção finita de conjuntos, foram concebidos por volta de 1880 por John Venn (Sandifer 2004). Através de estudos relacionados à lógica, Jon Venn criou uma diagramação baseada em figuras no plano, esse método consiste basicamente em círculos ou outras figuras geométricas, que representam relações entre conjuntos numéricos (Venn, 1880) (Figura 10).

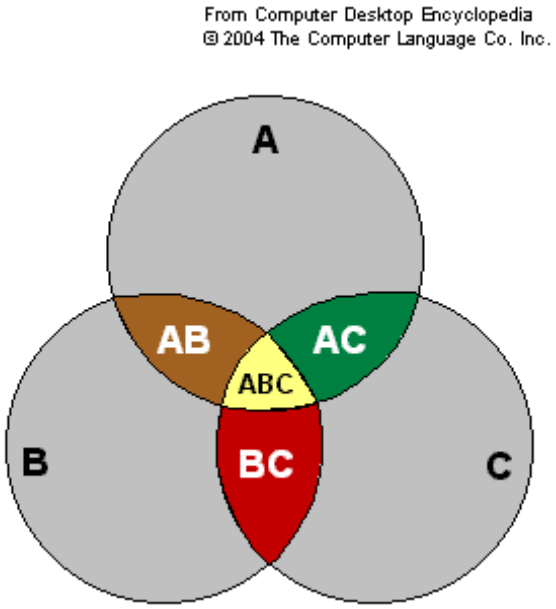


Figura 10 - Esquema de um diagrama de Venn mostrando como os conjuntos de dados se relacionam uns com os outros.

1.5- Linguagem de programação

A maior parte das ferramentas desenvolvidas nesta tese utilizam programação em *script* que são programas curtos com linguagem interpretada, que são linguagens de programação executadas do interior de programas e/ou de outras linguagens de programação; não se restringindo a esses ambientes. As linguagens de *script* servem para estender a funcionalidade de um programa e/ou controlá-lo, e sua utilização combinada com *softwares* de análises torna os processos mais rápidos e simples para o usuário, além de automatizar o armazenamento de informações em bancos de dados. Os *scripts* podem fazer a interface entre os programas fazendo com que a análise ocorra de forma mais simplificada.

Perl

No desenvolvimento de softwares de bioinformática, a linguagem Perl (*Practical Extraction and Report Language*) (Wall, 2009) é a mais utilizada, devido a seus mecanismos facilitados para tratamento de cadeias de caracteres. Perl é uma linguagem muito boa para manipulação de texto, embora as ciências biológicas envolvam uma boa dose de análise numérica agora, a maioria dos dados primários ainda é texto: nomes de clones, anotações, comentários, referências bibliográficas, sequências de DNA , etc., ainda estão em forma de texto e a conversão de formatos incompatíveis de dados é uma questão de gerenciamento de texto combinado com a criatividade do programador e Perl torna esta tarefa muito mais simples (Stein, 1996).

Aplicações web via CGI (*Common Gateway Interface*), também é uma característica da linguagem, que consiste em uma tecnologia que gera páginas dinâmicas, permitindo a um navegador passar parâmetros para um programa alojado num servidor web (Stein, 1998).

A linguagem suporta estruturas de dados arbitrariamente complexas. Ela também possui recursos vindos da programação funcional (as funções são vistas como outro valor qualquer para uma sub-rotina, por exemplo) e um modelo de programação orientada a objetos. Perl também possui variáveis com escopo léxico, que tornam mais fácil a escrita de código mais robusto e modularizado. Todas as versões de Perl possuem gerenciamento de memória automático e tipagem dinâmica (uma característica de determinadas linguagens de programação, que não exigem declarações de tipos de dados, pois são capazes de escolher que tipo utilizar dinamicamente para cada variável, podendo alterá-lo durante a compilação ou a execução do programa). Os tipos e necessidades de cada objeto de dados no programa são determinados automaticamente, a memória é alocada ou liberada de acordo com o necessário. A conversão entre tipos de variáveis é feita automaticamente em tempo de execução e conversões ilegais são erros fatais. Arquivos Perl são simples arquivos de texto ASCII (que é uma codificação de caracteres de oito bits baseada no alfabeto inglês) que contêm comandos na sintaxe Perl. Pode-se produzir tais arquivos com qualquer editor de texto que produza arquivos em ASCII puro. Para executar os comandos de um arquivo Perl é necessária a ação de um interpretador Perl (Perldoc – documentação online).

Além disso, esta linguagem possui vários módulos já desenvolvidos que facilitam o desenvolvimento de ferramentas biológicas como as que utilizam Blast e bancos de dados como o BioPerl (Stajich, *et al.* 2002).

Ambiente R

O ambiente R também é bastante utilizado para análises estatísticas sendo ideal para a manipulação de dados de *microarrays*. R é uma linguagem e ambiente para computação estatística e gráficos, fornece uma ampla variedade de métodos estatísticos (modelagem linear e não linear, testes estatísticos clássicos, análise de séries temporais, classificação, clusterização, etc.). Um dos pontos fortes de R é a capacidade de produzir gráficos com padrão de publicação, incluindo símbolos e fórmulas matemáticas, quando necessário, além de ser também uma linguagem de programação que permite o desenvolvimento de novas funções para lidar com formatos de dados personalizados(R Development Core Team, 2007).

Utilizando o ambiente R foi criado o *BioConductor que* é um projeto de desenvolvimento de software aberto para oferecer ferramentas para análise e compreensão de dados genômicos de alta densidade (Gentleman *et al.* 2004). *BioConductor* inclui suporte extensivo para análise de *arrays* de expressão como também para *exonarray, copy number*, SNP e outros tipos de experimentos (Dudoit *et al*., 2002).

1.6- Banco de Dados

A quantidade de dados gerada pela análise de *microarrays* é gigantesca e sem um bom sistema de banco de dados fica inviável o armazenamento e organização dos dados. Para os pesquisadores que se beneficiam com os dados guardados em um banco de dados, dois requisitos são necessários: Fácil acesso às informações (Eficácia) e métodos para extrair as informações necessárias para responder à uma pergunta biológica específica (Objetivo).

Um banco de dados pode ser definido como uma coleção compartilhada de dados logicamente relacionados, projetado para atender as necessidades de informação de múltiplos usuários em uma organização (Taylor, 2001). E um banco de dados precisa de um sistema gerenciador de banco de dados (SGBD) que é uma coleção de componentes de software para criar, gerenciar e consultar um banco de dados (Taylor, 2001). O MySQL é um SGBD, que utiliza a linguagem SQL (Linguagem de Consulta Estruturada, do inglês *Structured Query Language*) como interface. É atualmente um dos bancos de dados mais populares, com mais de 10 milhões de instalações pelo mundo, é robusto confiável e gratuito (www.mysql.com), por isso optamos por usá-lo como SGBD.

2 - OBJETIVO

Objetivo principal:

Este trabalho foi desenvolvido com o objetivo de criar uma ferramenta de análise para estudos de expressão que utilizam *microarrays*.

Objetivos específicos:

- Controle de qualidade dos *arrays*;
- Correção do *background* dos *arrays;*
- Normalização de dados dos *arrays*;
- Implementação de diversos testes estatísticos para os dados dos *arrays*;
- Implementação de filtros e valores de corte definidos pelo usuário;
- Implementação de ferramenta de *data-mining;*
- Implementação de ferramentas de clusterização;
- Implementação de ferramenta de cruzamento de dados.

3 - Justificativa

A quantidade de dados gerados por um experimento de *microarrays* é bastante grande do ponto de vista biológico e estatístico e uma análise cuidadosa com a escolha de testes estatísticos apropriados e as devidas correções e filtragem de dados é imprescindível para que se obtenha resultados consistentes. Existem diversas ferramentas disponíveis gratuitamente para este fim, mas a grande maioria exige do usuário um conhecimento de programação e uso de linhas de comando, e muitas vezes os resultados das análises não são claros, exigindo do pesquisador um conhecimento avançado para extrair todas as informações necessárias ou utilizar todos os recursos que estas ferramentas fornecem, como gráficos, personalização de valores de corte, filtragem de dados, etc. Foi com este problema em mente que este trabalho foi desenvolvido. A ferramenta apresentada nesta tese utiliza várias dessas outras ferramentas já disponíveis e mais alguns scripts desenvolvidos originalmente, de uma forma simples e amigável para o usuário, onde ele terá acesso a uma grande quantidade de informação relativa a seu experimento, apenas selecionando o arquivo de dados brutos e escolhendo as análises que deseja fazer e que sejam aplicáveis a seu experimento.

4 – METODOLOGIA E RESULTADO

Como esta tese se trata do desenvolvimento de uma ferramenta computacional, para facilitar a compreensão do leitor, as sessões de metodologia e resultado serão apresentadas conjuntamente.

Este pipeline foi desenvolvido no sistema operacional Linux, em linguagem Perl, utilizando o módulo CGI para sua interface web e o módulo DBI para manipulação de banco de dados. Um banco de dados foi criado para armazenamento dos dados brutos de intensidade e outro com dados de anotação, utilizando linguagem SQL e MySQL como SGBD. O ambiente estatístico R foi usado para todas as análises estatísticas e tratamento de dados, utilizando diversas bibliotecas do *BioConductor*, que serão descritas mais detalhadamente abaixo. Todos os *sofwares* e ferramentas utilizadas para o desenvolvimento deste pipeline são gratuitos e estão disponíveis para download na internet.

O principal resultado deste trabalho é o desenvolvimento de uma *pipeline* de análise de *microarrays* de expressão (Figura 11) com interface web escrita em linguagem Perl e R. *Pipeline* consiste em uma cadeia de processos em que os elementos são organizados de forma que a saída de um processo é a entrada do outro, como em uma linha de produção.

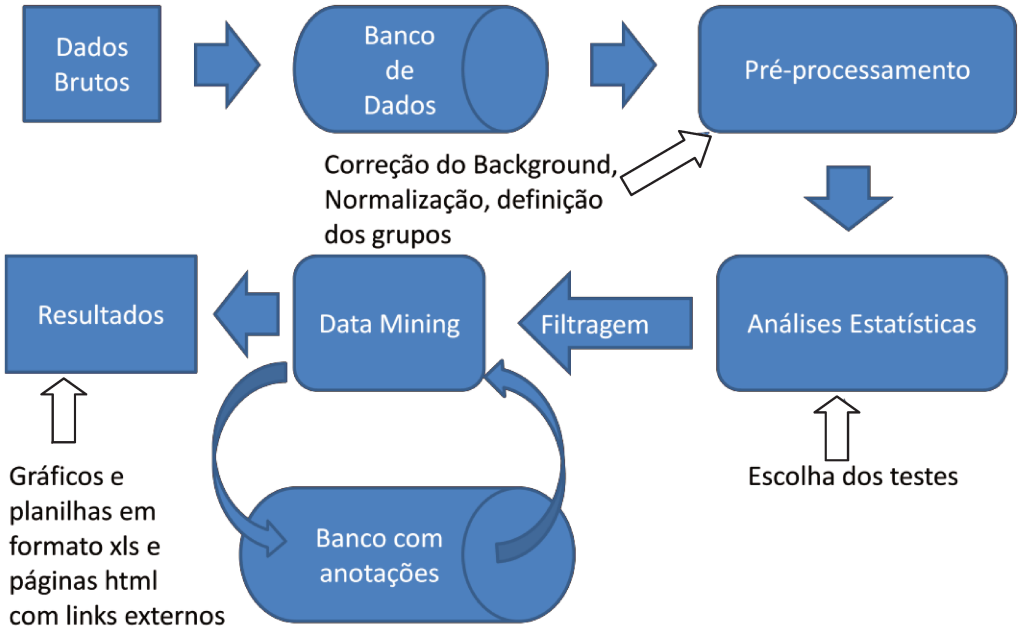


Figura 11 - Workflow de funcionamento do pipeline.

4.1- Aquisição dos dados

Os dados utilizados para o desenvolvimento desta ferramenta foram principalmente gerados pelo estudo "GENE EXPRESSION PROFILE IN MESIAL TEMPORAL LOBE EPILEPSY: CONTRIBUTION OF GENETIC AND SPORADIC FORMS AND INSIGHTS INTO MECHANISM" (apêndice 2) com material obtido de pacientes refratários com epilepsia de lobo temporal mesial que passaram por procedimento cirúrgico de hipocampectomia, e os dados de controle foram adquiridos através de autopsia.

O primeiro passo da metodologia de análise envolve a importação dos dados brutos com a leitura da intensidade do sinal do chip de *microarray,* arquivos *.CEL para chips de oligonucleotídeos e arquivo texto tabulado para chips de cDNA onde a primeira coluna possui os identificadores das sondas e as demais os valores de intensidades de sinal. O arquivo CEL armazena os resultados dos cálculos de intensidade sobre os valores de pixel do arquivo

DAT. Isso inclui um valor de intensidade, o desvio padrão da intensidade, o número de pixels usados para calcular o valor da intensidade, um sinalizador para indicar um *outlier*, calculado pelo algoritmo (*Affymetrix*, 2009). Utilizando o pacote *affy* (Gautier *et al.* 2004) dentro do ambiente R é feita a leitura da intensidade de chips Affymetrix, com a função *ReadAffy,* este pacote contém funções para análise exploratória de *arrays* de oligonucleotídeos. Se a função for chamada sem argumentos: `ReadAffy()` todos os arquivos CEL no diretório de trabalho serão lidos e armazenados. No entanto o uso de argumentos dá maior flexibilidade e permite uma maior organização dos dados.

O pacote inclui funções de plotagem dos dados em nível de sonda, útil para controle de qualidade, e neste passo nossa ferramenta gera gráficos de histograma, que é uma representação gráfica da distribuição de frequências de uma massa de medições, com os dados brutos (Figura 12).

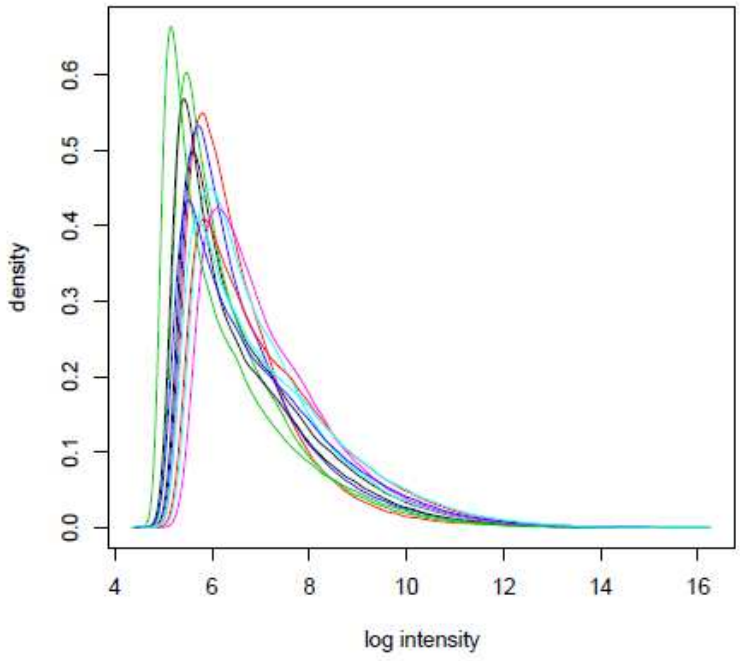


Figura 12 - Histograma dos dados brutos dos chips de epilepsia de lobo temporal mesial.

Como visto na Figura 7, a distribuição do sinal bruto de um *microarray* possui uma forma característica e o histograma é uma ferramenta boa para a identificação de saturação, que é vista como um pico adicional na maior intensidade de log no gráfico. Se houver saturação da sonda (valor de intensidade ≥ 46000), ela não será computada nas análises subseqüentes, pois é um indicador de problemas técnicos na hibridização.

O pacote fornece também a possibilidade de desenhar *Box Plot* (Figura 13), que é outra boa ferramenta de visualização para analisar a intensidade global de todas as sondas ao longo do *array*. O *Box Plot* é desenhado a partir do percentil 25 e 75 da distribuição de intensidades. A mediana ou percentil 50 é desenhada dentro da caixa (McGill, *et al.,* 1978). As linhas que se estende desde a caixa descrevem a propagação dos dados. *Arrays* de um mesmo grupo amostral devem ser semelhantes na distribuição, ou deverão ser descartados.

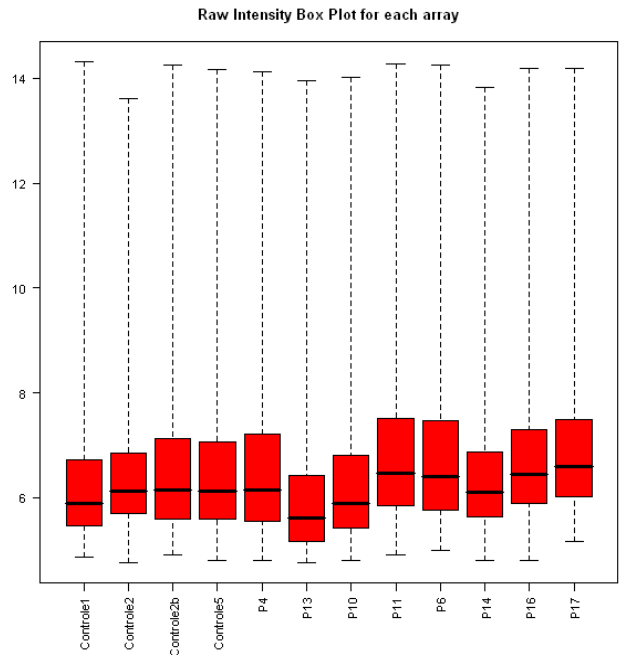


Figura 13 - Box Plot dos dados brutos dos chips de epilepsia de lobo temporal mesial.

4.2- Pré-processamento

O pré-processamento de dados de *microarray* inclui a correção de *background*, a normalização dos valores de expressão dentro de cada *microarray* e entre os *microarrays* (para ajustar os valores de expressão através dos *microarrays*). O objetivo da etapa de pré-processamento de dados é remover a variância técnica e erros sistemáticos, sem alterar a variação biológica dentro dos dados.

Usamos os pacotes *affy* e *gcrma* do *BioConductor* para este fim o que, permite o uso de métodos como o método padrão Affymetrix MAS5 ou métodos mais sofisticados como MBEI (*Model Based Expression Index*) (Li & Wong 2001), VSN (*Variance Stabilizing Normalization*) (Huber *et al.* 2002), RMA (*Robust Multiarray Average*) (Irizarry *et al*. 2003b) ou GCRMA (média robusta com correção de ligação não específica de acordo com o conteúdo GC de cada sequência das sondas)

Neste passo o usuário determina o método de correção de *background* e qual método usará para computar as intensidades (PM-only ou PM-MM Figura 14), normalização e define os grupos que serão usados na comparação (ex. quais arquivos são referentes ao controle e quais são referentes ao paciente).

Os métodos de correção de *background* disponíveis são: MAS 5 (Baseado no algoritmo padrão Affymetrix) e RMA (*Robust Multiarray Average*)

Ajuste da intensidade do Perfec Match:

PM-only – Sem ajuste. Usa os valores de intensidade de PM sem modificação.

PM-MM – Usa as sondas *mismatch* para ajustar o valor de intensidade subtraindo do valor de PM o valor de MM.

Figura 14 - Visualização da intensidade da imagem de um probe-set.

Os métodos de normalização são:

quantile: O objetivo do método de quantis é fazer a distribuição de intensidades de probe para cada *array* em um conjunto de *array*s. O método é motivado pela idéia de que um gráfico quantil-quantil mostra que a distribuição de dois vetores de dados é a mesma se o gráfico formar uma linha reta diagonal e não é a mesma se for diferente de uma linha diagonal (Bolstad *et al.* 2003) Com este método todos os *array* ficam com a mesma distribuição.

Loess: Esta abordagem é baseada na idéia do MA-plot, onde M é a diferença de valores de expressão em log e A é a média dos valores em log da expressão (Dudoit *et al.* 2002). No entanto, ao invés de ser aplicada a dois canais de cores no mesmo *array*, como é feito no caso de cDNA, é aplicada a intensidades das *probes* a partir de dois *array*s de cada vez. Um MA-plot de dados normalizados deve mostrar uma nuvem de pontos espalhados ao redor do eixo M=0.

Invariant Set: Escolhe um subconjunto de *probes* PM com diferença pequena dentro de um subconjunto classificado em dois *array*s, para servir como base para a montagem de uma curva de normalização (com o *array* de *baseline* sobre eixo Y e o *array* a ser normalizado no eixo X) (Cheng & Wong, 2001)

O objetivo da normalização dos dados é minimizar os efeitos causados pelas variações de técnica e, como conseqüência, permitir que os dados sejam comparáveis, a fim de encontrar mudanças verdadeiramente biológicas. É possível observar os efeitos da aplicação da normalização na figura 9, é perceptível a mudança no "brilho" da imagem antes e depois da normalização. Tendo selecionados os métodos desejados de normalização e correção de *background*, o usuário pode observar os resultados de tais tratamentos nos gráficos de histograma e box-plot gerados pelo programa com os dados normalizados (Figura 15) se comparado com as figuras 12 e 13. Em comparação com o histograma e o box-plot dos dados brutos fica claro o efeito da normalização que faz com que as intensidades dos sinais entre os *arrays* fique bem mais próximas.

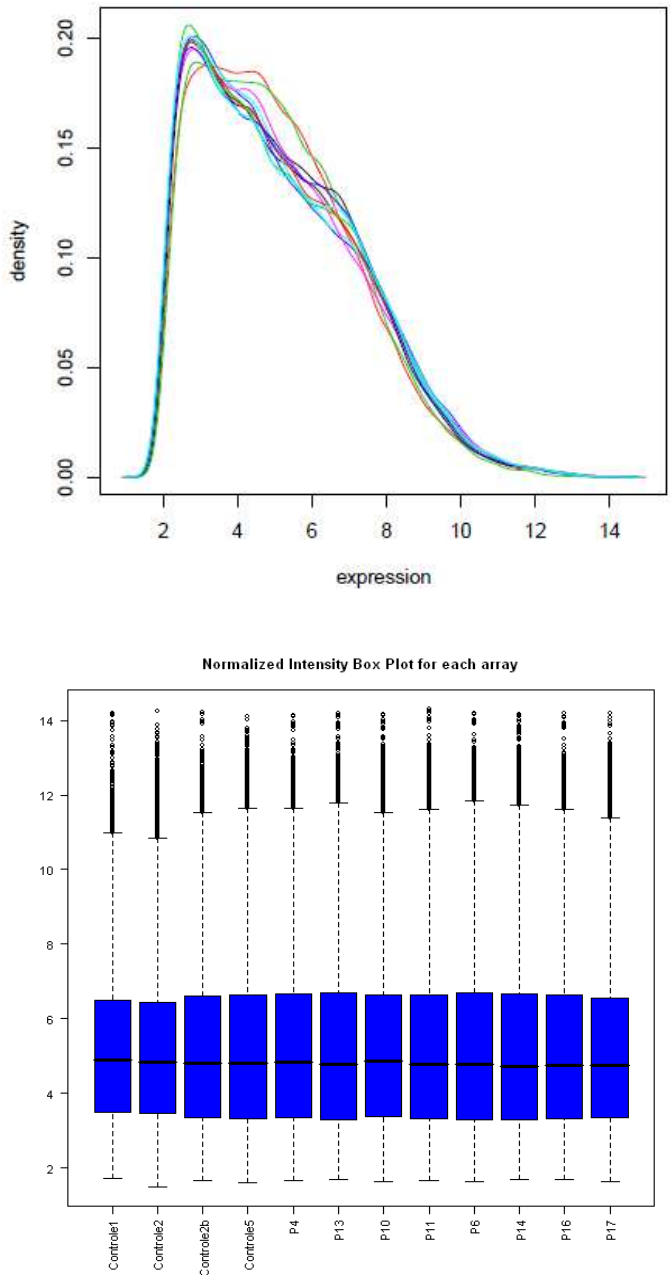


Figura 15 - Histograma e Box-plot dos dados normalizados dos chips de epilepsia de lobo temporal mesial.

4.3- Análises estatísticas

O usuário pode escolher o teste estatístico mais adequado ao seu experimento, dentre diversos testes paramétricos e não-paramétricos. Caso o número de amostras seja muito pequeno e o usuário selecione um teste paramétrico, um aviso irá aparecer na página, mas não impedirá a análise.

Os métodos para a detecção de genes diferencialmente expressos descrito a seguir podem ser aplicados aos valores de expressão normalizados para Affymetrix ou qualquer arquivo de texto tabulado contendo dados numéricos, não sendo necessário que sejam dados de *microarrays*, a única exigência é que o arquivo tenha na primeira coluna identificadores e nas demais colunas valores numéricos.

Genes diferencialmente expressos podem ser detectados usando testes estatísticos como o teste Wilcoxon, o teste t de Student, ANOVA, a estatística moderada do pacote limma (baseado em um abordagem bayesiana [Smyth, 2005]), SAM (Significance Analysis of Microarrays [Tusher *et al*. 2001]) e RankProduct (Breitling *et al.* 2004).

Experimentos de *microarray* geram problemas de multiplicidade grandes em que milhares de hipóteses são testadas simultaneamente dentro de um experimento, o pacote *multtest* do *BioConductor* fornece métodos adequados para ajustar p-valores de acordo com este problema de múltiplos testes. Métodos de ajuste disponíveis são, o procedimento introduzido por Benjamini e Hochberg (1995) para controle de taxa de falsa descoberta e Bonferoni ajuste de p-valores para forte controle de FWER (*family wise error rate*) (Abdi, 2007).

O gráfico de vulcão (Figura 16) é uma opção para quem possui dados em distribuição normal, pois ele utiliza o teste-t que é paramétrico, ele organiza os genes em dimensões de significância biológica e estatística. O primeiro eixo (horizontal) é o *fold change* entre os grupos em escala logarítmica, assim genes regulados positiva ou negativamente parecem simétricos. O segundo eixo (vertical) representa o p-valor para um teste-t em uma escala logarítmica

negativa assim, quanto menor o p-valor, mais alto no gráfico ele irá aparecer. O primeiro eixo indica o impacto biológico da mudança, o segundo indica os dados estatísticos, ou a confiabilidade da mudança (Cui & Churchill, 2003).

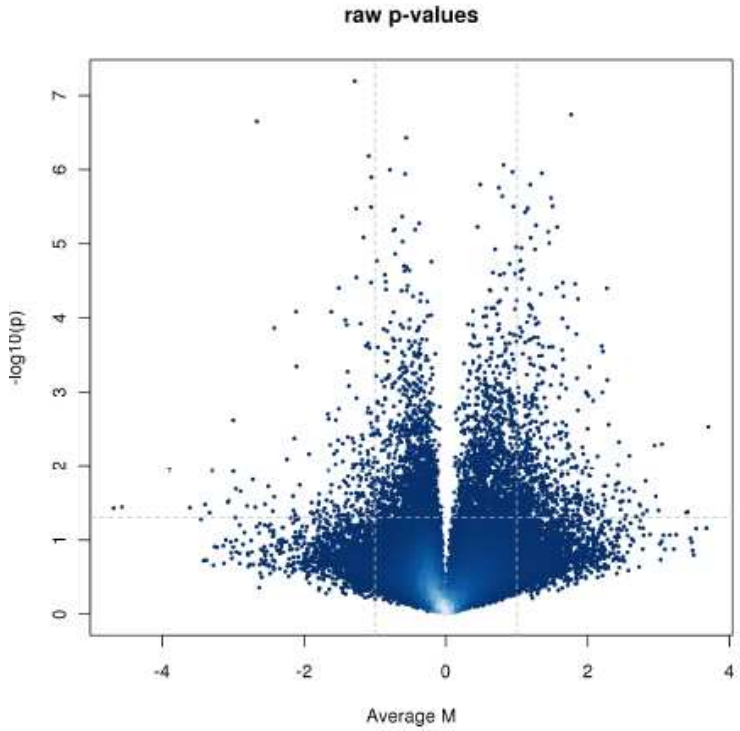


Figura 16 - Gráfico de vulcão dos dados de Epilepsia Mesial do Lobo Temporal.

Filtragem dos dados: Depois de escolher o teste estatístico o usuário pode escolher um valor para o corte que será usado na determinação dos genes diferencialmente expressos, e terá também a liberdade de decidir se este corte será nos valores de p ou nos valores de correção para múltiplos testes do p. Pode também determinar uma diferença mínima entre os genes que aparecerão no arquivo final, por exemplo: apenas genes que tenham a diferença mínima de 50% entre eles.

Análise de Correlação: Como opção de análise também implementamos a correlação de Pearson ou de Spearman, em que o usuário

pode comparar dois grupos com o mesmo número de chips de forma pareada ou comparar as intensidades de expressão do grupo de chips contra um vetor numérico de tamanho n em que n seja o número de chips. Este tipo de análise pode ser usada, por exemplo, em experimentos em que se deseja observar a alteração de genes com dosagens diferentes de determinada droga, onde, no vetor numérico irão os valores das doses usadas e a correlação irá mostrar se o aumento da dose da droga regulou a expressão de forma positiva, negativa ou não causou alteração. Esta ferramenta possui também um filtro, em que o usuário poderá escolher o valor mínimo e/ou máximo de correlação que deseja observar, podendo também escolher apenas correlações positivas ou negativas.

4.4 – *Data-mining*

Para o *data-mining* foi desenvolvida a ferramenta *Annotation Search* (apêndice 1) onde um banco de dados SQL foi criado com os dados de anotação Affymetrix, com o campo Probe Set ID como chave primária. Com uma lista de identificadores (Probe Set IDs) pode-se facilmente pesquisar no banco de dados informações específicas para a anotação correspondente a estes identificadores. A fim de facilitar a mineração de dados, foram incluídos links para várias bases de dados públicas que podem conter informações biológicas relevantes, tais como: NetAffx, Unigene, Ensembl, SwissProt e OMIM. Como uma característica adicional, que pode ajudar em projetos de descoberta de genes, também foi implementado um filtro por cromossomo, que permite ao usuário diminuir a quantidade de dados a serem analisados, direcionando sua pesquisa a um cromossomo ou região cromossômica específica.

Para estudos de miRNA é gerado link para o banco mirBase dos miRNAs detectados como diferencialmente expressos (Figura 17). E um arquivo em formato FASTA com a sequência de todos os miRNAs detectados como diferencialmente expressos.

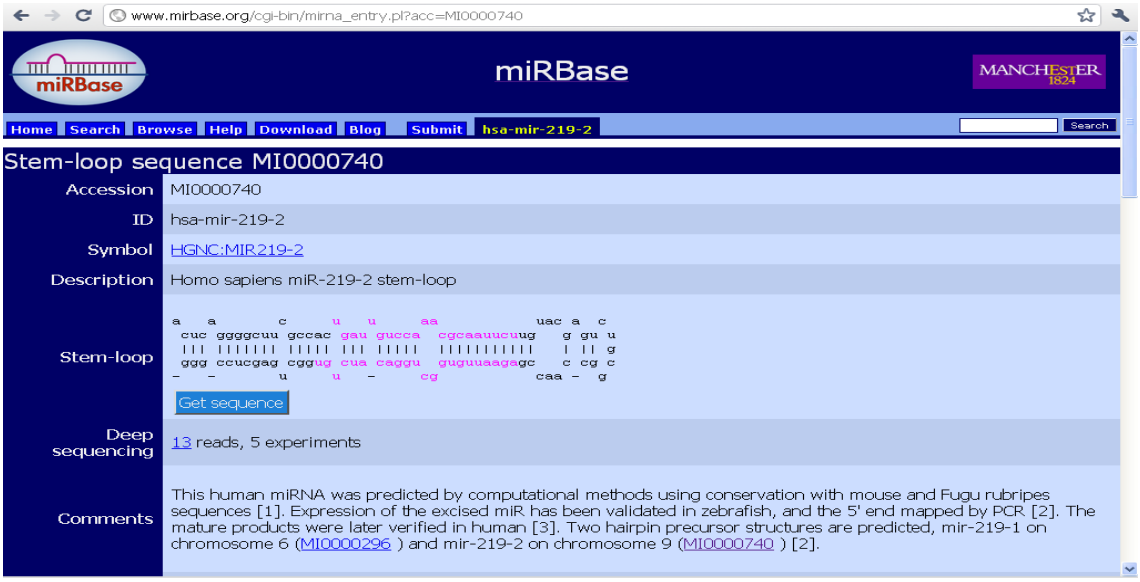


Figura 17 - Exemplo da página do miRBase que é exibida através do link da ferramenta de mineração de dados para microRNAs.

4.5 – Ferramentas de clusterização

PCA

A ferramenta faz automaticamente um gráfico de PCA com os dados normalizados, que mostra como as amostras estão agrupadas (Figura 18).

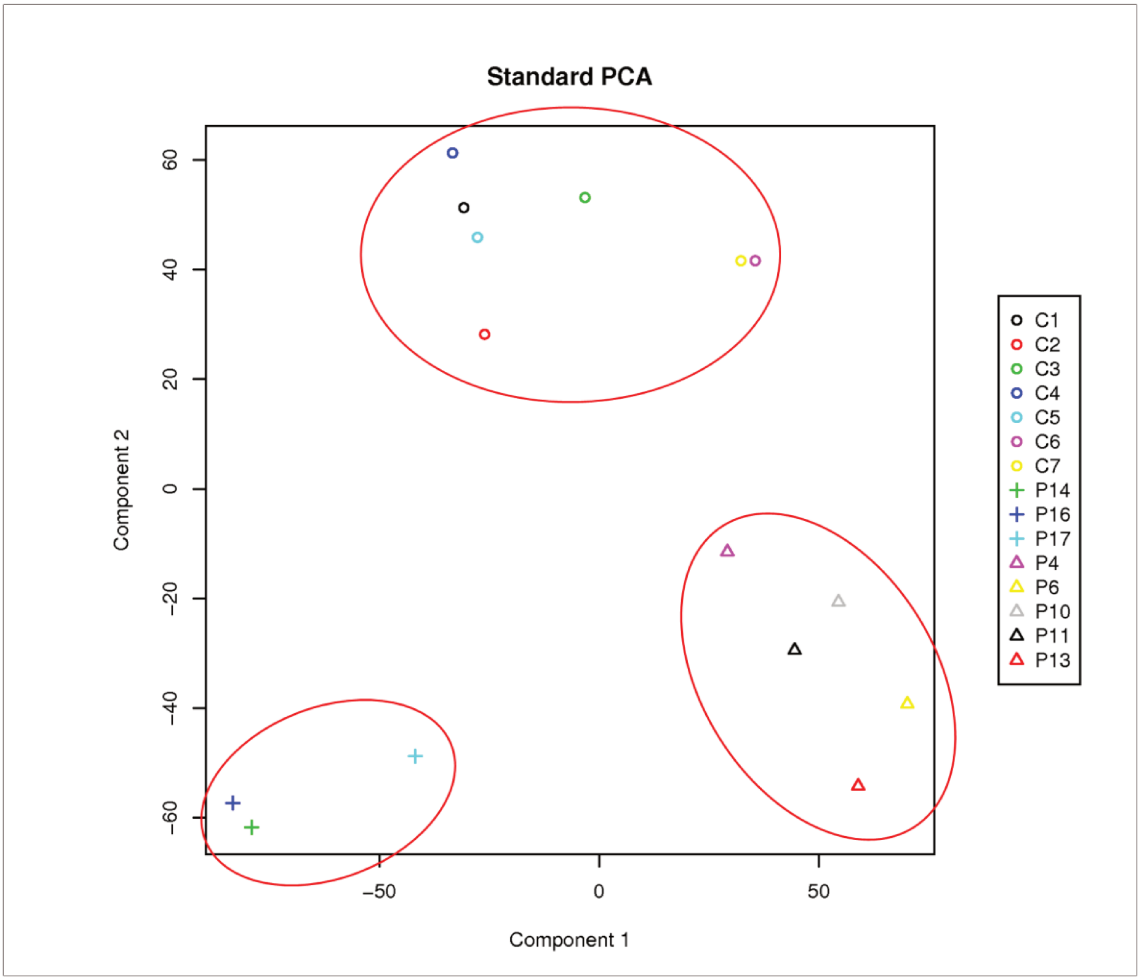


Figura 18 - Exemplo de PCA onde se vê claramente a formação de três grupos distintos entre as amostras. (Dados de epilepsia)

Cluster Hierárquico

Utilizando o pacote *stats* do R a ferramenta gera um cluster hierárquico com os dados selecionados pelo usuário, podendo ser com todos os genes da amostra ou apenas uma lista determinada pelo usuário, como a lista de genes detectados como diferencialmente expressos. O usuário determina o método que será usado para o cálculo da distância e um dendrograma também é gerado (Figura 19). O usuário poderá fazer download de um arquivo texto com a lista dos dados na mesma ordem da clusterização. É possível também escolher o padrão de cores que será usado para criar o cluster.

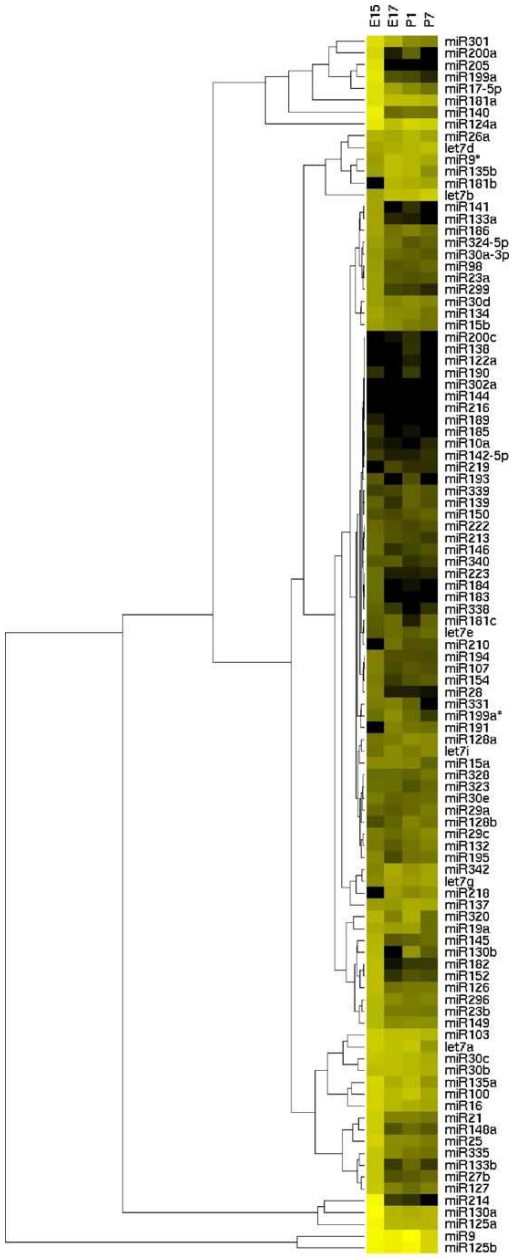


Figura 19 - Cluster hierárquico com os dados de miRNA do estudo MicroRNA Expression Profile in Murine Central Nervous System Development gerado com esta ferramenta.

Self-Organizing Maps

A ferramenta de *Self-Organizing Maps* recebe como entrada uma lista com valores de intensidade e o usuário deve determinar o número de clusters que deseja observar, sempre com uma relação múltipla, por exemplo: 2x2 ou

3x3 ou 3x2, etc (Figura 20). Um arquivo texto é gerado com a mesma lista de entrada acrescida de uma nova coluna que indica o número do cluster que aquele gene pertence.

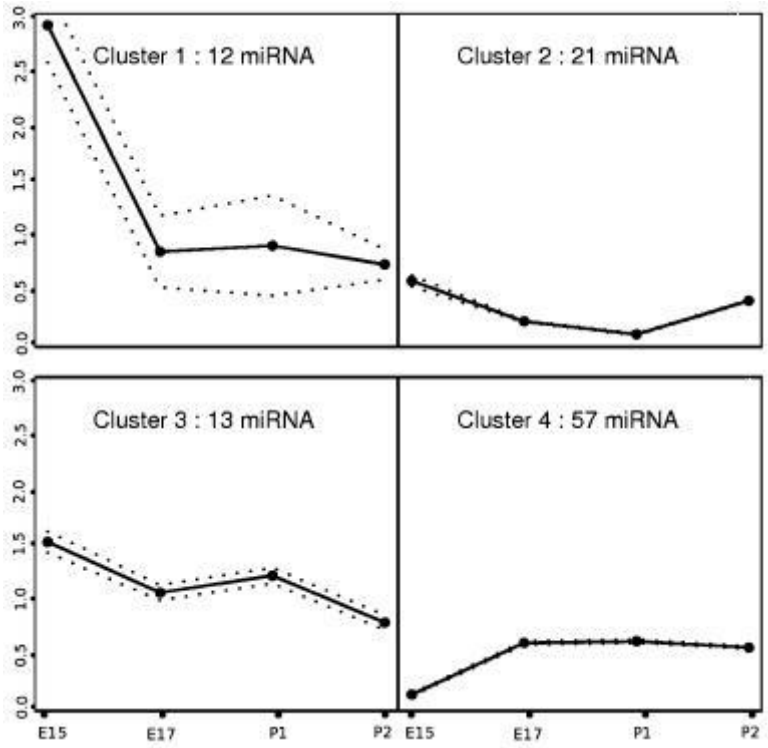


Figura 20 – SOM 2X2 gerado com os dados de miRNA do estudo "MicroRNA Expression Profile in Murine Central Nervous System Development" gerado com esta ferramenta.

4.6 – Ferramentas auxiliares

Desenvolvemos uma ferramenta para auxiliar pesquisadores que trabalham com os dados divididos em vários grupos, que permite que se selecionem dados comuns presentes em dois arquivos diferentes ou dados presentes em apenas um ou outro arquivo. Por exemplo, no estudo de epilepsia mesial de lobo temporal que desenvolvemos no laboratório (apêndice 2) tivemos três grupos: Controle, Paciente Familial e Paciente Esporádico. Com esta ferramenta conseguimos selecionar os genes que foram diferencialmente

expressos nas duas comparações: Controle X Familial e Controle X Esporádico, ou selecionar genes que foram diferencialmente expressos no grupo Controle X Familial e não foram diferencialmente expressos no grupo Controle X Esporádico**.**

Diagrama de Venn

Em conjunto com a ferramenta anterior é gerado um diagrama de Venn que mostra todas as correlações entre os conjuntos de dados mostrando o número de genes em comum (Figura 21).

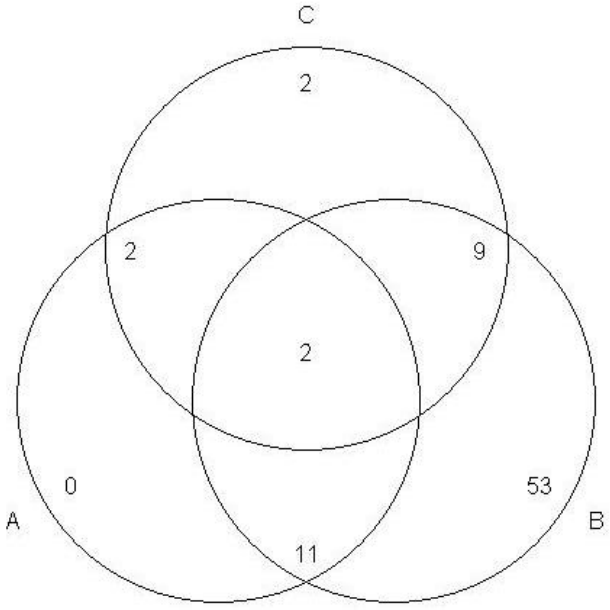


Figura 21 - Diagrama de Venn - (A) conjunto de dados 1; (B) conjunto de dados 2; (C) conjunto de dados 3

Banco de dados

Um banco de dados foi criado para armazenar os dados normalizados e resultados de análises para usuários cadastrados.

5 - Conclusão

A Biologia Molecular atual gera tal quantidade de informação que seria impraticável realizar análises sistêmicas eficientes sem ferramentas computacionais adequadas, além da tendência de disponibilidade pública de dados permitir aos cientistas uma grande economia de tempo e recursos. Muitos pesquisadores não possuem um profissional de bioinformática disponível para analisar dados de *microarrays*. Deste modo nós desenvolvemos uma ferramenta que auxilia pesquisadores que trabalham com *microarrays*, facilitando a análise da grande quantidade de dados que um experimento de *microarray* gera. Esta ferramenta possui uma interface amigável. Com o uso da ferramenta pesquisadores podem ter resultados das análises de forma clara e intuitiva, não havendo necessidades de aprender linguagens de programação ou uso de linha de comando. Nossa ferramenta foi aplicada com sucesso em dois estudos concluídos e está sendo aplicada em vários outros em andamento.

A ferramenta está disponível em: http://lgm.fcm.unicamp.br:9001/cgi-bin/pipeline/pipeline.cgi

6 - Referências Bibliográficas

Abdi, H. Bonferroni and Šidák corrections for multiple comparisons. In N.J. Salkind (ed.). *Encyclopedia of Measurement and Statistics*. Thousand Oaks, CA: Sage. 2007.

Alberts, B.; Bray, D.; Johnson, A.; Lewis, J.; Roberts, K.; Raff, M. & Walter, P. Fundamentos da biologia celular: uma introdução à biologia. Edição Universitária. Editora Artmed S.A., Porto Alegre. 780p. 2001.

Affymetrix, Affymetrix Developer Network, 2009. Documentação *online* disponível em: http://www.affymetrix.com/support/developer/powertools/changelog/gcos-agcc/cel.html

Affymetrix, Statistical Algorithms Description Document [Internet]. 2002. http://media.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf

Ambros, Victor. “microRNAs: tiny regulators with great potential.” Cell. 107.7: 823-6. 2001.

Augenlicht L.H., Wahrman M.Z., Halsey H., Anderson L., Taylor J., Lipkin M. Expression of cloned sequences in biopsies of human colonic tissue and in colonic carcinoma cells induced to differentiate in vitro. Cancer Research. 47 6017. 1987.

Babu, M. M., Introduction to microarray data analysis - in Computational Genomics (Ed: R. Grant), Horizon Press, U.K. 2002.

Baxevanis, A. D., and Ouellette, B. F. F. Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins. John Wiley, 2005.

Benjamini, Y., Hochberg, Y., Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological) Vol. 57, No. 1, 289-300. 1995.

Benson D.A., Karsch-Mizrachi I., Lipman D.J., Ostell J., Sayers E.W., National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA Nucleic Acids Res. Jan;39 (Database issue) 2011.

Bartel DP. MicroRNAs: genomics, biogenesis, mechanism and functions. Cell, 23:281-297. 2004.

Bolstad, B., Low Level Analysis of Hight-density Oligonucleotide Array Data: Background, Normalization and Summarization, Ph.D. Thesis, University of California, Berkeley, 2004.

Bolstad, B.M.; Irizarry, R. A.; Astrand, M.; Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics Vol. 19 no. 2 p185–193, 2003.

Breitling R, Armengaud P, Amtmann A, Herzyk P. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. FEBS Lett ;573:83-92. 2004

Cheng Li and Wing Hung Wong, Model-based analysis of oligonucleotides arrays: model validation, design issues and standard error application. Genome Biology 2001, 2(8):research0032.1-0032.11

Coe B., Antler C., Spot your genes – ana overvier of microarray, BioTeach. Disponível em: http://www.scq.ubc.ca/spot-your-genes-an-overview-of-the-microarray/ [internet] 2004.

Corder, G. W,, Dale I F. *Nonparametric statistics for non-statisticians: A step-by-step approach*. Wiley, 2009.

Cramer, D., & Howitt, D. The SAGE Dictionary of Statistics. SAGE Publications. 2004.

Cui, X. and Churchill, G.A., Statistical tests for differential expression in cDNA microarray experiments. Genome Biology, 4:210 , 2003

Cullane A.C., Pierre G, Considine E.C., Cotter T.G., Higgins D.G. Between-group analysys of microarray data. Bioinformatics. 18:1600-1608, 2002.

DeRisi J., Penland L., Brown P.O., Bittner M.L., Meltzer P.S., Ray M., Chen Y., Su Y.A. & Trent J.M. Use of a cDNA microarray to analyze gene expression patterns in human cancer. Nature 14 457–460, 1996.

Dowdy, S., Wearden, S., and Chilko, D. Statistics for Research, third edition Wiley: New York. 2004.

Dudoit S., Shaffer, J.P., Boldrick J.C., Multiple Hypothesis Testing in Microarray Experiments, *Statistical Science* Vol. 18, No. 1, pp. 71-103 2003

Dudoit S., Yang Y.H., Bolstad B., Using R for the analysis of DNA microarray data, R News 2 (1) 24–32. 2002

Edwards, A. L. An Introduction to Linear Regression and Correlation. San Francisco, CA: W. H. Freeman, pp. 33-46, 1976.

Eisen M.B., Spellman P.T., Brown P.O. Botstein D.: Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA, 95:14863-14868. 1998.

Fisher R. A. Statistical Methods for Research Workers, Edinburgh: Oliver and Boyd, p.43. 1925.

Gautier L, Cope L, Bolstad BM, Irizarry RA. affy--analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 2004; 20:307-315.

Geisser S. and Johnson W. M., *Modes of Parametric Statistical Inference*, John Wiley & Sons, 2006

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. BioConductor: open software development for computational biology and bioinformatics.Genome Biol . 5:R80. 2004

Gibas, C., Jambeck, P., Developing Bioinformatics Computer Skills, O'Reilly, 2001.

Hastie T., Tibshirani R. Friedman J. The Elements of Statistical Learning, 2nd ed., NY, 2009, Springer. pp. 520–528.

Griffiths, A. J. F.; Miller, J. H.; Suzuki, D. T.; Lewontin, R. C.; Gelbart, W. M.; Wessler,S. R. Introdução a Genética. 8 ed. Rio de Janeiro: Editora Guanabara Koogan, 2006.

Hill A.A., Brown E.L, Whitley M.Z, Tucker-Kellogg G., Hunter C.P., Slonim D.K. Evaluation of normalization procedures for oligonucleotide array data based on spiked cDNA controls. Genome Biol. 2:RESEARCH005, 2001

Huber, W., Heydebreck, A., Sultmann,H., Poustka,A., Vingron,M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics, 18 Suppl 1:96–104, 2002

Irizarry R.A., Bolstad B.M., Collin F., Cope L.M., Hobbs B., Speed T.P. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res. Feb 15;31(4):e15 2003 a.

Irizarry,R.A, Hobbs,B., Collin, F., D Beazer-Barclay,Y.D. Antonellis, K.J., Scherf,U,. Speed, T.P. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics, 4(2):249–264, Apr 2003 b.

Jolliffe, I.T. Principal Component Analysis, Series: Springer Series in Statistics, 2nd ed., Springer, NY, XXIX, 487 p. 28, 2002.

Kohonen, T. Self-organizing maps. Third, extended edition New York, NY: Springer. 2001.

Lewin, B., Genes VIII. Pearson Education, Upper Saddle River, NJ, USA. 2004

Li, C., Wong, W.H. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc Natl Acad Sci U S A, 98(1):31–36, 2001.

Lockhart D.J., Dong H., Byrne M.C., Follettie M.T., Gallo M.V., Chee M.S., Mittmann M, Wang C., Kobayashi M., Horton H. *et al*. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nature Biotechnology 1996 14 1675–1680

McGill R., Tukey W. J., Larsen, W. A., Variations of Box Plots. The American Statistician 32 (1): 12–16. 1978

Myers, J. L..; Well, A. D. Research Design and Statistical Analysis (second edition ed.). Lawrence Erlbaum. 2003.

Pang, K.C., Stephen,S., Dinger, M.E., Engstrom, P.G., Lenhard, B. e Mattick,J.S. NAdb 2.0-an expanded database of mammalian non-coding RNAs.Nucleic Acids Res. 35 (Database issue):D178-D182. 2007.

Pease A.C., Solas D., Sullivan E.J., Cronin M.T., Holmes C.P., Fodor S.P.A. Light-generated oligonucleotide arrays for rapid DNA-sequence analysis. *Proc Natl Acad Sci.* 91:5022–502, 1994

Perldoc – Perl Programming Documentation – disponível em: http://perldoc.perl.org/perl.html

Quackenbush J. Microarray data normalization and transformation. Nat. Genet. 32:496-501, 2002.

R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing [Internet]. 2007 ;Available from: http://www.r-project.org

Rathod P.K., Ganesan K., Hayward R.E., Bozdech Z. & DeRisi J.L. DNA microarrays for malaria. Trends in Parasitology. 2002. 18, 39–45.

Šášik, R., Woelk, C.H., Corbeil, J. Microarray truths and consequences. Journal of Molecular Endocrinology (2004) 33, 1–9

Sandifer, E. (2004). "How Euler Did It" The Mathematical Association of America: MAA [internet] Disponível em http://www.maa.org/editorial/euler/How%20Euler%20Did%20It%2003%20Venn%20Diagrams.pdf

Schena M., Shalon D., Davis R.W., Brown P.O. "Quantitative monitoring of gene expression patterns with a complementary DNA microarray". Science 270, (5235): 467–470. 1995

Setubal, J.C. A origem e o sentido da bioinformática. Com Ciência [internet]. 2003 http://www.comciencia.br/reportagens/bioinformatica/bio10.shtml

Shaffer, J. P. Multiple hypothesis testing. Annual Review of Psychology 46.1: 561-584. 1995.

Shlens, J. A Tutorial on Principal Component Analysis. Center for Neural Science, New York University. 2009. Disponível em http://www.snl.salk.edu/~shlens/pca.pdf

Spearman, C. Correlation Calculated from Faulty Data. British Journal of Psychology, 1904-1920, 3: 271–295. 1910.

Stein, L. D. Official guide to programming with CGI. pm: [the standard for building Web scripts]. New York: Wiley. 1998

Stein, L. How Perl saved human genome – BioPerl. The Perl Journal 1996. Disponível em: http://www.bioperl.org/wiki/How_Perl_saved_human_genome

Storey, J.D. and R. Tibshirani, Statistical significance for genomewide studies. Proc Natl Acad Sci U S A, 100(16): p. 9440-5. 2003.

Southern, E. M. "Detection of specific sequences among DNA fragments separated by gel electrophoresis". Journal of Molecular Biology, 98 (3): 503–517. 1975

Sewall, W. *Evolution and the genetics of populations*. Chicago: University of Chicago Press. 1984.

Smyth,G.M., Limma: linear models for microarray data. In R C Gentleman, V J Carey, S Dudoit, R A Irizarry, andWHuber, editors, Bioinformatics and Computational Biology Solutions using R and BioConductor , page Chapter 23. Springer, New York, 2005.

Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JGR, Korf I, Lapp H, Lehvaslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka ED, Wilkinson M, Birney E. The Bioperl Toolkit: Perl modules for the life sciences. Genome Research. Oct;12(10):1161-8. 2002.

Taylor, A.G., SQL Para Dummies, Tradução da quarta edição, Editora Campus. Rio de Janeiro. 2001.

Tusher, V.G., R. Tibshirani, and G. Chu, Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA, 2001. 98(9): p.5116-21.

Wall, L., The Perl Programming Language, version 5.10.1, 2009. http://www.cpan.org/src/ [internet]

Watson, J. DNA: The secret of life. Arrow Books. 2004

Wilcoxon, F. Individual comparisons by Ranking Methods. Biometrics Bulletin 180–83. 1945.

Wild, C. J.. The Wilcoxon Rank-Sum Test. Behaviour 2.4 : 1-10. 1988.

Yang Y.H., Dout S., Luu P., Lin D.M, Peng V., Ngai J., Speed T.P. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res. 30:e15, 2002.

7 - APÊNDICE

Apêndice 1: Manuscrito da ferramenta de anotação Annotation Search

Annotation Search: a tool for data mining with gene expression microarray systems.

Cristiane S. Rocha, B.Sc.; Claudia V. Maurer-Morelli, Ph.D and Iscia Lopes–Cendes, MD, PhD.

Department of Medical Genetics, Faculty of Medical Sciences, University of Campinas -UNICAMP, Campinas, SP, Brazil

Correspondence to: Iscia Lopes-Cendes, MD, PhD
Department of Medical Genetics,
Faculty of Medical Sciences – UNICAMP
Tessália Vieira de Camargo, 126
Cidade Universitária "Zeferino Vaz"
Campinas SP, Brazil, CEP 13083-887
Tel: +55 19 3521 8909
Fax: +55 19 3289 1818
e-mail: icendes@unicamp.br

Abstract

Annotation Search (AS) is a web tool developed to facilitate assembly of a number o different biological information from data obtained directly using Affymetrix™ GeneChip expression system. AS was written in perl language using the CGI module for web application and the DBI module to database manipulation A SQL database was created with the Affymetrix™ annotation data, with the Probe Set ID field as primary key. Starting with a list of identifiers (Probe Set IDs) AS can easily search a specific database for the annotation corresponding to these identifiers. In order to facilitate data mining we included links to several databases which may contain relevant biological information, such as: NetAffx, UniGene, Ensembl, SwissProt and OMIM. As an additional feature, which may help in gene discovery projects, we also implemented a filter ´by chromosome´ which allows the user to decrease the amount of data to analyze be analyzed at once by directing his/her search to a specific chromosome or chromosomal region.

Keywords: multiple data base search, filter by chromosome, gene discovery

Introduction

Expression studies using microarrays produce a great amount of data which can lead to very complex and time consuming analyses [1]. As the quality of the arrays become more reliable (ref.) the critical point today is to guarantee a good bioinformatics analysis of the data generated; therefore, the usefulness of gene expression data depends on how much information is available for each identified gene. In other words, the identities of the genes associated with each spot on a microarray must be accessible as the analysis is performed, descriptions and classifications of each gene on the array must be readily available. Moreover, it is necessary to organize the data and relate it to other databases, such as proteins, metabolic pathways and putative biological functions. Therefore, our main goal was to develop a tool that fulfill this requirements as it analysis data of gene expression generated by the Affymetrix™ GeneChip microarray system.

Annotation Search is a web tool developed to facilitate for biologists to get annotation data from Affymetrix™ GeneChip. Starting with a list of identifiers (Probe Set IDs) this tool can search on a database the annotation corresponding to these identifiers. To make it easier the data mining process, it was included links to others databases such as: NetAffx, UniGene, Ensembl, SwissProt and OMIM. In addition we implemented a filter by chromosome, which allows the user to significantly diminish the amount of data to analyze directing his search to a specific chromosome or interest.

Methods

This tool was developed to be web based. A SQL database was created with the Affymetrix™ annotation data, with the Probe Set ID field as primary key using MySQL[3] as DBMS. The web tool was written in perl[4] language using the CGI

module for web application and the DBI module to database manipulation. The database has the annotation of all 54.675 Probe Sets presents in the HG-U133 Plus 2 Affymetrix™ GeneChip and can be fed with the others Affymetrix™ chips as requested. External links was implemented for: NetAffx, UniGene, Ensembl, SwissProt, OMIM and NCBI. The input file must have one Probe Set ID by row, and must be .txt.

Results

This tool is freely available and was developed to help researchers of biological science to gain access to annotation information of data from GeneChip da Affymetrix™. With the filter by chromosome, the user can diminish the amount of data to analyze and focalise the research. In our laboratory we use this tool to search annotation for epilepsy data, the analysis identified 4751 genes differently expressed, but the researcher want to study the only chromosome 18 at the moment, with the filter option that number decrease to 60.

Conclusion

The main function of the tool Annotation Search is to help researchers that work with GeneChip Affymetrix™, because it facilitates the analysis of the great amount of data that an microarray experiment generates. With the option of filter by chromosome, the user can diminish the volume of data significantly to analyze, generating more specific searches for genes candidates in the projects of positional cloning. With external links, a great number of useful functional information can be had easily about the Probe Set what helps in the hierarchization of the data of gene expression gotten in the experiments.

Availability and requirements:

Software home page: http://lgm.fcm.unicamp.br:9001/cgi-bin/affy/affy_annotation.cgi

Operating system(s): Platform independent

Programming language: Perl (5.8.8) and MySQL

Other requirements: none

License: Freely available

Acknowledgments

This study was supported by FAPESP. CSR is a recipient of a scholarship from CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), Brazil.

References

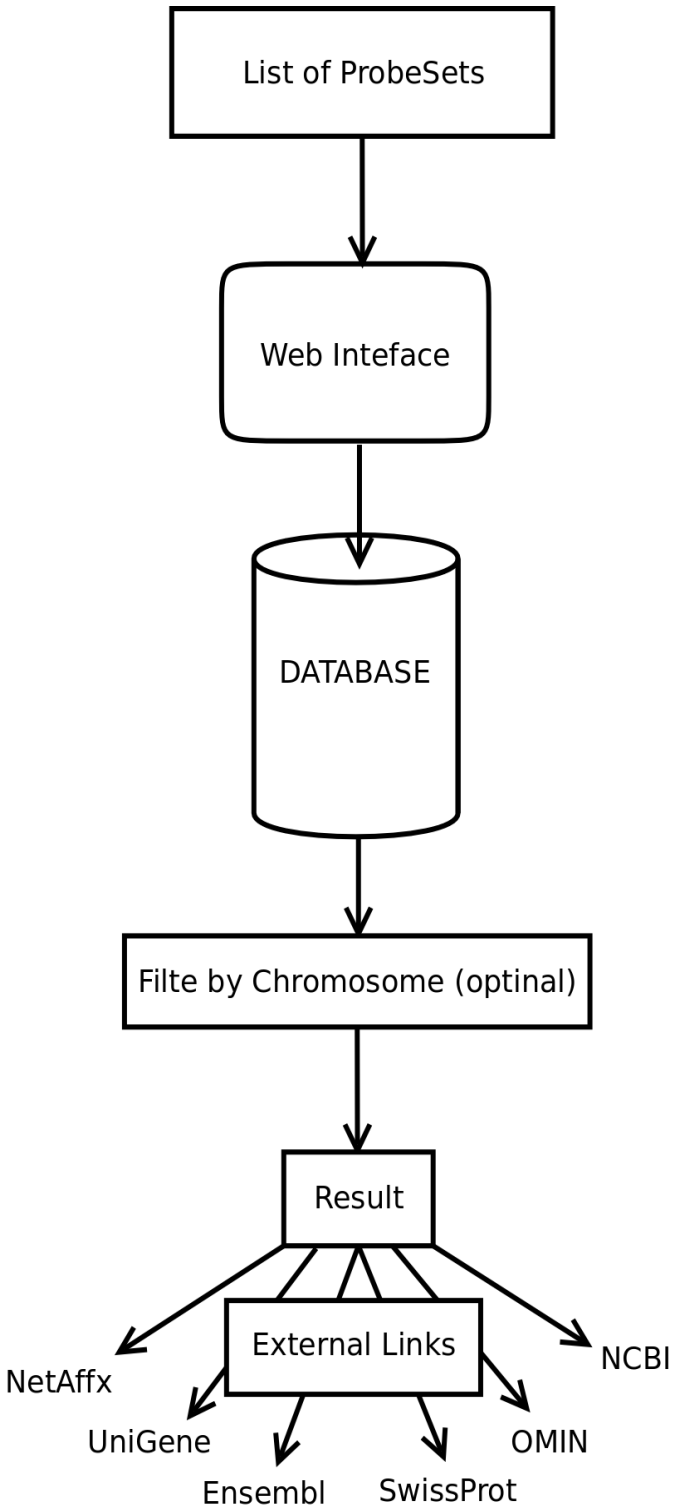
1 - Genetics Encyclopedia gene chip on Answers.com. Genetics Copyright © 2003 by The Gale Group, Inc. Published by The Gale Group, Inc.

2 - Duggan, D. J., Bittner, M., Chen, Y., Meltzer, P., Trent, J. M. (1999). Expression profiling using cDNA microarrays. *Science* **283**(5398), 83-87. 1/1999.

3 - Perl.com: The Source for Perl, http://www.perl.com/

4 - MySQL The world's most popular open source database, http://www.mysql.com/

Fig. 1. Diagram of Annotation Search, showing the workflow of data.



Apêndice 2: Manuscrito do estudo que forneceu os dados para a criação da maior parte da ferramenta apresentada nesta tese.

GENE EXPRESSION PROFILE IN MESIAL TEMPORAL LOBE EPILEPSY INDICATES DIFFERENT MOLECULAR MECHANISMS FOR SPORADIC AND FAMILIAL FORMS

Running Title: Gene expression in MTLE

Claudia V Maurer-Morelli[1], Cristiane de Souza Rocha[1], Jaira F de Vasconcellos[1], Fernanda CR Pinto[1], Romenia R Domingues[1], Clarissa L Yasuda[2], Helder Tedeschi[2], Evandro De Oliveira[2], Fernando Cendes[2], Iscia Lopes–Cendes[1*].

1. Departments of Medical Genetics and 2. Neurology; Faculty of Medical Sciences; University of Campinas (UNICAMP), Campinas, SP, Brazil

Running Title: Gene expression in MTLE

Key words: Familial mesial temporal lobe epilepsy, hippocampal sclerosis, microarray technology.

***Corresponding author:** Iscia Lopes-Cendes, M.D., Ph.D.
Department of Medical Genetics
Faculty of Medical Sciences
University of Campinas (UNICAMP)
Tessália Vieira de Camargo, 126
Cidade Universitária "Zeferino Vaz"
Campinas SP, Brazil, CEP 13083-887
TEL: +55 19 3521 8909
FAX: +55 19 3289 1818
e-mail: icendes@unicamp.br

Abstract

Patients with mesial temporal lobe epilepsy (MTLE) and hippocampal atrophy (HA) determined by magnetic resonance image, who are pharmacoresistant may undergo unilateral surgical removal of hippocampal formation as an alternative for seizure control. Surgical specimens of these patients offer a unique opportunity to investigate the molecular mechanisms underlying hippocampal sclerosis (HS), the pathological hallmark of MTLE. In this study we have accessed patients with a positive familial history of MTLE (n=4), as well as patients with a negative familial history of MTLE (sporadic form, n=4) in order to determine gene expression profiles in hippocampal tissue obtained from epilepsy surgery. In addition, we compared expression profiles obtained from patients with hipocampal tissue from autopsy controls (n=3). We used microarray technology, followed by validation of target genes using quantitative Real Time PCR. We found a total of 582 genes differentially expressed when comparing samples from MTLE patients with a positive and negative family. In this paper we discussed about the main altered function and cellular pathways in both forms using Gene Ontology, Ingenuity Pathways Analysis software and DAVID analysis. This is the first study using a high throughput microarray platform in MTLE with HA to explore the differential gene expression between positive and negative familial history of MTLE. Our results clearly show that, although similar in the clinical features, both forms of MTLE have distinct molecular signatures.

Introduction

Classical histopathological studies and, more recently, different high-resolution neuroimaging modalities identify hippocampal sclerosis (HS) as the most prominent pathological substrate in patients with intractable mesial temporal lobe epilepsy (MTLE) (Gloor 1991; Berkovic et al., 1991; Cendes et al., 1993). However, the precise pathogenesis of HS and its relationship with the molecular and cellular mechanisms underlying MTLE is not completely understood. Surgical specimens of patients with MTLE who are resistant to clinical treatment may undergo unilateral surgical removal of hippocampal formation as an alternative for seizure control offering a unique opportunity to investigate the molecular mechanisms underlying HS.

Microarray technology makes possible to study thousands of genes at the same time, providing insight into the entire set of molecular events taking place in the examined tissue. The use of microarray technology to investigate global gene expression changes in epilepsy has grown in animal models for epilepsy as much as for human brain tissue (Majores et al., 2004; Jamali et al., 2006; Wang et al., 2010) but none of these studies in human brain have considered the potential differences exiting between samples obtained from patients with differences in their genetic predisposition to seizures.

Therefore, we have designed the present study to specifically address the question of whether differences in genetic predisposition to seizures, assessed by the presence of family history, has an impact on gene expression profile obtained from hippocampal tissue of patients with pharmacoresistant MTLE.

Methods

Patients and Controls

hippocampi from patients used in this study were obtained surgically from individuals with medically refractory MTLE. Patients had a complete clinical, EEG and imaging investigation as part of the clinical protocol for epilepsy surgery in our University Hospital (Yasuda 2006 and 2010). Briefly, patients had detailed neurological examination, series of electroencephalography (EEG), magnetic resonance imaging (MRI), neuropsychological and psychological assessments. Seizures were lateralized according to the medical history, interictal EEGs and comprehensive neurological examination (Yasuda 2006 and 2010). None of the patients included in the present study had any generalized or complex partial seizures documented in the 48h before or during epilepsy surgery.

Control specimens (with no history of epilepsy or any neurological condition) were obtained from autopsies performed with no more than six to eight hours after death. In the microarray experiments we included samples from patients with a positive family history (FH) of MTLE (n=4), a negative-FH of MTLE (n=4) and control specimens (n=3). In addition, we performed quantitative Real Time PCR (qRT-PCR) validation, using samples from patients with a positive–FH of MTLE (n=10), a negative-FH of MTLE (n=10) and control specimens (n=9), including samples previously analyzed in the microarray experiments. All specimens were obtained after informed consent which was approved by the Research Ethics Committee of our institution.

Tissue processing

Fresh surgical specimens were sectioned, immediately frozen in liquid nitrogen and store frozen at -80°C until use. For microarray and qRT-PCR analysis, total RNA was isolated by Trizol™ reagent (Invitrogen, Carlsbad, USA) according to manufacturer´s instructions and its concentration, purity and quality were determined by spectrophotometry at 260/280nm and by electrophoresis on agarose gels.

Gene Expression Profiling Assays

Transcriptional profiling of total RNA [5 µg] from surgical specimens of patients with positive-FH and negative-FH was performed using the Affymetrix (Santa Clara, CA) oligonucleotide microarray Human Genome U133 Plus 2.0, according to a standard One-Cycle Target Labeling protocol array (Affymetrix) for cDNA synthesis, in vitro transcription, production of biotin-labeled cRNA, hybridization of cRNA with the plate, and scanning of image output files using the GeneChip Scanner 3000 (Affymetrix). The quality of hybridized chips was assessed using Affymetrix guidelines on the basis of average background, scaling factor, number of genes called present, 3' to 5' ratios for endogenous control genes. All the high quality arrays were included for low and high level bioinformatics analysis.

Data processing was perfomed in R environment (http://www.r-project.org) using the computer packages Affy and RankProd from Bioconductor (Breitling et al., 2004; Gautier et al., 2004; Gentleman et al, 2004). Genes were considered differentially expressed when statistical significance reached a $p<0.01$. The cut-off for minimal difference between groups was 20%. In addition, we took into account the false

discovery rate (FDR) due to multiple comparisons and corrected the p-value ($p<0.05$) using the Benjamini-Hochberg test (Benjamini et al., 1995).

We analyzed overrepresented gene ontology categories modulated in both patients groups (those with positive-FH and negative-FH) using the Database for Annotation, Visualization and Integrated Discovery (DAVID, http://david.abcc.ncifcrf.gov/home.jsp National Institute of Allergy and Infectious Diseases, National Institutes of Health). DAVID produces a list of overrepresented categories applying a score to each category by using "-log" of EASE score (a modified Fisher Exact p-value) in order to show the significantly enriched gene ontology categories (Dennis et al., 2003; Huang et al., 2009). The related gene ontology categories were combined in a cluster, and considered statistically significant when enrichment scores were ≥ 2 and $p<0.05$. The increased confidence in gene clusters overrepresentation analysis is observed by an increase in the enrichment scores. The gene interactions and correlation networks were identified with the Ingenuity Pathways Analysis software version 7.6 (Ingenuity Systems, Mountain View, CA). The fold change threshold was >2 or ≤ -2 and the signaling pathways were considered statistically significant with a $p<0.05$.

Quantitative real-time PCR validation

In order to validate data obtained in the microarray experiments, we performed an independent qRT-PCR gene expression analysis using additional surgical specimens from MTLE patients with positive-FH (n=10) and negative-FH (n=10), as well as autopsy controls (n=9). Total RNA from each patient and control was isolated as described above. Complementary DNA (cDNA) was obtained using SuperScript III™ reverse transcriptase and random primers, both from Invitrogen, Carlsbad, USA. Gene

expression assays were carried out in 7500 Real-time PCR system (Applied Biosystems Foster City, CA) with the TaqMan® system (Applied Biosystems, Foster City, CA). Validation of probes and primers, PCR conditions and the relative genes expression were carried out by the 2(-Delta Delta C(T)) method as described by Livak and Schmittgen (2001). Briefly, this method uses de CT variation from target and control genes (ΔCT = CT target gene - CT control gene), and the ΔCT from target and calibrator [-ΔΔCT = -(ΔCT target - ΔCT calibrator)] to calculate the amount of target. Therefore, the amount of target is given by $2^{-\Delta\Delta CT}$. QRT-PCR reactions were performed in triplicates, and on eight genes along with *ß-ACTIN* and *HPRT* (TaqMan™-Applied Biosystems, Foster City, CA) as endogenous control. Statistical analysis was performed by Kruskall- Wallis test with $p \leq 0.05$ for a significant difference between groups.

Results

Clinical characterization of MTLE Patients

Presurgical evaluation of all patients indicated MTLE, including MRI evidence of HS. The predominant type of seizure pattern was complex partial seizures uncontrolled by the use of optimal doses of antiepileptic medications for at least two years. The mean pre-operative seizure frequency for patients with positive-FH was 7.25 seizures/month and for patients with negative-FH was 9.25 seizures/month. The most relevant clinical information for both groups of patients included in the microarray study is shown in Table 1.

General features of gene expression for positive-FH and negative-FH MTLE

At a global level, transcriptional changes were differentially distributed (up and down-regulation) when comparing MTLE patients with a positive-FH with patients with a negative-FH. Figure 1 shows cluster analysis for all conditions studied. One remarkable feature was the tendency for overall low gene expression seen in samples from patients with negative-FH (Figure 1). In the comparison with autopsy controls samples from MTLE patients with a positive-FH showed a total of 170 differentially expressed genes; whereas, samples from patients with a negative-FH had 341 differentially expressed genes. Interestingly, when comparing samples from both groups of patients (with a positive-FH and a negative-FH) we identified a significant difference in gene expression profiling, with a list of 582 differentially expressed genes. In Figure 2, we represent the overall number of genes that were differentially expressed in the two by two comparisons among patients with positive-FH, negative-FH and controls. A complete list of the differentially expressed genes can be found in the supplemental material section which is available "on line".

Gene Ontology processes and pathways modulated by MTLE differentially expressed genes

The enrichment analysis from the differently expressed genes in positive-FH and negative-FH MTLE were performed to identify gene ontology processes and pathways that would occur more often than the expected in a random distribution.

The analyses of overrepresented gene ontology categories modulated in both in positive-FH and negative-FH MTLE was performed with the Database for Annotation, Visualization and Integrated Discovery (DAVID), and the cut-off criteria were enrichment score ≥ 2 and $p<0.05$. The top enriched gene ontology categories, which include biological processes and metabolic functions, in positive-FH compared to

negative-FH MTLE gene expression profiling included “cell fraction”, “cytoplasmic vesicle”, “actin filament organization”, “synapse”, “protein localization”, “neuron differentiation”, “protein kinase binding”, “cell adhesion”, “cytoskeleton”, “GTPase activity”, and “regulation of apoptosis”, as demonstrated in Figure 3. The most highly enriched clusters of the gene ontology categories in negative-FH MTLE compared to controls included “neurological system process”, “dendrite and cell soma”, “axonogenesis”, “synaptic vesicle membrane”, “neurotransmitter secretion”, “cell-cell adhesion”, “synaptic vesicle”, and “learning or memory” (Supplementary Figures S1 and S2).

Furthermore, Ingenuity Pathways Analysis we identified signaling pathways differently modulated in positive-FH and negative-FH MTLE. The fold change threshold was >2 or ≤-2 and the signaling pathways were considered statistically significant with a $p<0.05$. Among the most activated signaling pathways was the Glutamate Receptor Signaling in negative-FH MTLE, and the Pathogenesis of Multiple Sclerosis in positive-FH MTLE, which is significantly associated with inflammatory response processes. The most highly ranked signaling pathways are shown in Figure 4.

Validation of Candidate Genes by qRT-PCR

JAK1, ARHGDIA, IL6ST, GRIN1, GABRA3, SCN2B and BDNF genes, which were differentially expressed in microarray analysis, were chosen for validation by qRT-PCR. The experiments were run in triplicates using three endogenous controls to normalize the analysis. Changes on gene expression observed by microarray analysis were confirmed in these selected genes by *HPRT* endogenous control (Figure 5). Another candidate gene, NRCAM, was validated only using *β-ACTIN* as endogenous

control (data not shown). Our analysis also demonstrated that *GAPDH* and *ß-ACTIN* were more subjected to instability in RNAm expression among the analyzed samples.

Discussion

Microarray technology has been successfully employed to understand molecular basis of many diseases including neurological diseases (Glanzer et al., 2004; Mirnics and Pevsner, 2004; Jamali et al., 2006). In this way, microarray analyses performed on tissue obtained from humans during resective surgery for intractable human temporal lobe epilepsy and from animal models of epilepsy have given clues about the molecular mechanisms underlying the epileptogenesis as well as the chronicle phases. Many genes, differentially expressed, have been categorized into functional groups such as inflammation, synaptic transmission, cell death and survival, differentiation and cellular proliferation and cell adhesion (Gorter et al., 2006, Jamali et al. 2006). Nevertheless, none of these studies involving humans specimens have attempted for differences between patients with positive-FH of epilepsy and those with negative-FH of epilepsy (Arion et al., 2006; Jamali et al., 2006; Fernández-Medarde et al., 2007; van Gassen et al., 2007). So, in the present study, we performed microarray investigation in order to elucidate the molecular mechanisms underlying the pathology of HA in MTLE, and mainly to investigate positive-FH and negative-FH forms of MTLE.

Recently, we identified a locus for familial MTLE using linkage studies corroborating previous studies from our group, in which a genetic predisposing to HA was observed in these families (Kobayashi et al., 2001; Maurer-Morelli et al., 2007). Since we had access to positive and negative-FH of epilepsy in our service, our group

has conducted clinical, imaging and molecular investigating in both forms of MTLE (Maurer-Morelli et al., 2007; Yasuda et al., 2010; Conz et al., 2011).

It is important to note that Yasuda and colleagues (2010) recently described clinical, neuropsychological, and MRI abnormalities in surgical MTLE patients. In this study, voxel-based morphometry (VBM) was performed on preoperative MRIs and the possible clinical and neuropsychological differences between the two groups investigated. VBM and t tests were used to compare the patients' groups with normal controls. The results showed that patients with negative-FH are more likely to develop more severe neuropsychological deficits, HA, and widespread extrahippocampal damage, compared with patients with positive-FH. On the other hand, positive-FH patients present less severe neuropsychological deficits and more restricted structural abnormalities. These results demonstrate that different mechanisms might underlie pathological findings in MTLE with HA; patients with positive-FH are under a stronger genetic influence and less subjected to environmental factors, and negative-FH patients are under much more environment factors that are the cause of the HA (Yasuda et al., 2010). Another recently paper designed to evaluate progression of HA in patients with positive-FH of MTLE by longitudinal MRIs demonstrated that these patients have progressive hippocampal volume reduction independently of seizure frequency; however, such progression of HA seems to be slower than in negative-FH of MTLE (Conz et al., 2011).

Considering clinical and histological aspects, Andrade-Valença and colleagues (2008) demonstrated that even clinical and hippocampal histological features of intractable patients with MTLE and HA in both, positive and negative-FH, were similar positive-FH patients had less pronounced mossy fibers sprouting demonstrating that this

group of patients respond differently to plastic changes induced by cellular losses in epilepsy.

Our study comes from to corroborate these clinical, neuroimaging and histopathogical findings shedding light on the molecular events underlying the differences between both forms of MTLE with HA. Analyzing our list of transcripts it is possible to find genes related to crucial cellular functions such as basic cellular metabolism, cell growth and proliferation, morphology, cell cycle, neurotransmission, immune response, cell death and maintenance in both forms of MTLE compared to control specimens. However, when we compared both groups of epilepsy it became clear the difference between groups, where negative-FH presents much more down regulated differentiation genes such as *BDNF* and its receptor *NTKR3, MAPK1, JAK*; anti-apoptosis genes as *ARGHDIA*; signaling genes as *GABRA3*, *GRIN1A*, and so on (Figures 1 and 2, and list of genes in Supplemental data).

The heterogeneity in the transcription profiles of positive and negative-FH of MTLE samples were identified using a hierarchical cluster analysis, reflecting the global differences between samples (Figure 1). Comparison of the differentially expressed transcripts in MTLE samples using Venn diagrams revealed that a few transcripts (9 up-regulated and 6 down-regulated) are differentially expressed between positive and negative-FH of MTLE compared to controls.

In up-regulated Venn diagram is notable that MALAT1 gene was the only one differential expressed in all conditions analyzed, C vs (+)FH, C vs (-)FH and (+)FH vs (-)FH (Figure 2). Despites, MALAT1 is overexpressed in many healthy tissues, including brain, it was overexpressed in epilepsy condition not only when compared to control hippocampus, but also was overexpressed more than 20 times in negative-FH

patients when compared with positive-FH. MALAT1 gene is highly abundant in neurons and has been suggested that it plays a role in synapse functions, as synapse formation and/or maintenance (Bernard et al., 2010). In epilepsy, this up-regulation may indicates

Hierarchical cluster analysis clearly showed a positive-FH patient, who has a particular gene expression profile, often similar to negative-FH of MTLE (Figure1). We found that this patient (identified as P4) had bilateral HA and more diffuse lesion observed by MRI. Because this patient has a familial antecedent of epilepsy, the molecular finding might be due by both genetic background and environmental influence. In fact, this patient has more diffuse lesions in extratemporal areas. This could explain her particular molecular signature when compared to the other patients belonging to positive-FH group. Because we used a nonparametric test, we could consider this patient in our sample without concerns.

An independently validation of the microarray results was performed by qRT-PCR. *HPRT*, *GAPDH* and *ß-ACTIN* endogenous controls were used to normalize the experiments. *HPRT* demonstrated to be the most stable housekeeping gene for human brain and validated the results obtained within the microarray analysis (Figure 5). However, it is important to note that some of the gene expression analysis by qRT-PCR did not achieve significant differences, probably due to the number of patients analyzed in each group. Moreover, we could not validate candidate genes using *GAPDH,* and only one candidate gene (NRCAM) was validated within *ß-ACTIN*, overall we observed that RNAm expression was more subjected to instability when samples were normalized with both these endogenous controls. These findings are in accordance to recently published data from Wierschke et al. (2010), which demonstrated that both *GAPDH* and *ß-ACTIN* are among a group of unstable housekeeping genes to be used in human

epileptogenic brain tissue. Interestingly, Pernot et al. (2010) using a kainite-induced mouse epilepsy model demonstrated that *GAPDH* was also an unstable housekeeping gene, but *β-ACTIN* was a stable endogenous control, and therefore suitable to be used to an efficient qRT-PCR normalization in this epilepsy mouse model.

Many articles have brought huge lists of genes based on microarray analysis; however, in order to avoid those series of individual differently expressed genes, we performed clustering analysis of functionally related genes based on gene ontology categories.

Recent studies of complex biological networks have shown that their organizations are not random; rather, they follow modular principles (Barabási & Oltvai 2004). Scientists defined a module as a group of genes cooperating to achieve a particular physiological function (Hartwell et al., 1999). A comprehensive interpretation of this huge amount of information is possible due to the development of algorithms that interpret the expression profile data into differently expressed genes. Further analysis can also classify the profiling data into enriched clusters of genes or significant signaling pathways, exploiting the potential of the data to elucidate biological features and generate a functional profiling. The main purpose in clustering analysis is to create groups (clusters) that share common features, with the main advantage of reducing complexity, which makes possible to represent more clearly the biological processes, cellular components and/or molecular functions represented in the analyzed gene expression profile. In our study, to perform the enrichment analysis for gene ontology categories in positive and negative-FH MTLE gene expression profiles we used the Database for Annotation, Visualization and Integrated Discovery (DAVID) program, and to investigate the significant signaling pathways in these gene expression profiles we used Ingenuity Pathways Analysis software (Figures 3 and 4, respectively) .

Interesting, using DAVID program to compare control vs positive-FH and control vs negative-FH , we achieved many similar functional group differential expressed in both forms of MTLE, as those linked to structural functional and communication (Figures 1and 2, Supplemental data). Furthermore, when we compared positive-FH vs negative-FH, besides those previous mentioned, we found that main functional classes altered were those linked to cell, membrane and insoluble fraction, which by definitions are fraction cells component, generated by disruptive biological procedures.

Ingenuity Pathway analysis showed that glutamate pathway is the most affected in negative-FH patients. In fact, subunits of ionotropic N-methyl-D-aspartate (NMDA), such as *GRIN1* and *GRIN2A,* and the subunit of metabotropic AMPA, *GRIA* were down-regulated in surgical specimens of these patients. Glutamate is the most abundant excitatory neurotransmitter in the central nervous system and has been related to the initiation and spread of seizure activity. Garrido-Sanabria et al., 2008 described mGlu2/3 impairment in animal model of pilocarpine and discussed a possible contribution of this finding to abnormal presynaptic plasticity with amplification of glutamate release and hyperexcitability in temporal lobe epilepsy. In our negative-FH patients samples we found impairment not only in metabotropic receptors but also in ionotropic receptor subunits. In addition, there are many other genes involving in glutamate turn over, which are down regulation as *SLC14A1, SLC17A7, GLUL* and *GLS*. Besides, is important to point out that genes related to calcium influx, *CACNA1C* and *CANB2* are up-regulated and possible contributing for glutamate release. This scenario corroborates to amplification of glutamate release and may contribute to hyperexcitability and damage to neurons.

Another analysis performed in the Ingenuity Pathway showed that Pathogenesis of Multiple Sclerosis is the main differential expressed pathway in positive-FH patients. In this pathway the main genes involved were *CCL3* and *CCL4,* which are mainly expressed in glial cells, as microglia and astrocyte cells (Biber et al., 2006). The precise biological role of these chemokines in the brain is still unknown. In addition, Xu and colleagues (2009) using pilocarpine animal model of epilepsy demonstrated the CCL3 and its receptor were down-regulated in both RNAm and protein levels suggesting that this gene/receptor might play a role in decreasing neuroprotective mechanisms. In fact, *CCL3* gene ontology function is described as cellular calcium ion homeostasis, which is a possible lesion mechanism in this group of positive-FH patients of MTLE.

Database generated by our study provides a complete view of cellular picture of happenings in pathogenesis of HA. It is important to note that hippocampus is a singular structure with much different type of cells and function on generate excitatory and inhibitory synapses. For this time our option was to performed this experimental design, however to better understand the lesion on each region affected in the hippocampus e its relation to pathology of epilepsy Ammon´s corn we are underway to dissect each region.

Conclusion

This is the first study using microarray in MTLE with HA to explore the differential gene expression between positive and negative-FH of MTLE. Corroborating the new evidences for different lesion seems by MRI, our results clearly show that both forms of MTLE have distinct molecular signatures.

.

Conflict of Interest

The authors have no conflict of interest to declare.

Acknowledgments

This study was supported by FAPESP grants 2005/56578-4, 2008/54789-6 and 2010/17440-5. We are grateful to our patients and their families for their helpful cooperation.

References

Abou-Khalil B, Andermann E, Andermann F, Olivier A and Quesney LF Temporal lobe epilepsy after prolonged febrile convulsions: excellent outcome after surgical treatment. 1993. Epilepsia 34: 878-883.

Arion D, Sabatini M, Unger T, Pastor J, Alonso-Nanclares L, Ballesteros-Yanez I, et al. Correlation of transcriptome profile with electrical activity in temporal lobe epilepsy. Neurobiol Dis. 2006; 22:374-87.

Andrade-Valença LP, Valença MM, Velasco TR, Carlotti CG Jr, Assirati JA, Galvis-Alonso OY, Neder L, Cendes F, Leite JP. Mesial temporal lobe epilepsy: clinical and neuropathologic findings of familial and sporadic forms. Epilepsia 2008; 49:1046-1054.

Barabási AL, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet. 2004; 5(2):101-13.

Becker AJ, Wiestler OD, Blumcke I. Functional genomics in experimental and human temporal lobe epilepsy: powerful new tools to identify molecular disease mechanisms of hippocampal damage. Prog Brain Res 2002;135:161-173.

Benjamini, Yoav; Hochberg, Yosef Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Statist. Soc.1995; 1:289–300.

Berkovic SF, Andermann F, Olivier A, Ethier R, Melanson D, Robitaille Y, et al. Hippocampal sclerosis in temporal lobe epilepsy demonstrated by magnetic resonance imaging. Ann Neurol. 1991; 29:175-182.

Bernard D, Prasanth KV, Tripathi V, Colasse S, Nakamura T, Xuan Z, et al. A long nuclear-retained non-coding RNA regulates synaptogenesis by modulating gene expression. EMBO J. 2010; 29(18):3082-3093.

Biber K, de Jong EK, vanWeering HR, Boddeke HW. Chemokines and their receptors in central nervous system disease. Curr Drug Targets 2006; 7:29–46.

Breitling R, Armengaud P, Amtmann A, Herzyk P. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. FEBS Lett 2004;573:83-92.

Cendes F, Andermann F, Gloor P, Evans A, Jones-Gotman M, Watson C, MRI volumetric measurement of amygdala and hippocampus in temporal lobe epilepsy. Neurology. 1993; 43: 719-725.

Conz L, Morita ME, Coan AC, Kobayashi E, Yasuda CL, Pereira AR, Lopes-Cendes I, Cendes F. Longitudinal MRI volumetric evaluation in patients with familial mesial temporal lobe epilepsy. Front Neurol. 2011; 14:2-5.

Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol 2003;4:P3.

Fernández-Medarde A, Porteros A, de las Rivas J, Núñez A, Fuster JJ, Santos E. Laser microdissection and microarray analysis of the hippocampus of Ras-GRF1 knockout mice reveals gene expression changes affecting signal transduction pathways related to memory and learning. Neuroscience. 2007; 146:272-285.

Garrido-Sanabria ER, Otalora LF, Arshadmansab MF, Herrera B, Francisco S, Ermolinsky BS. Impaired expression and function of group II metabotropic glutamate

receptors in pilocarpine-treated chronically epileptic rats. Brain Res. 2008;1240:165-176.

Gautier L, Cope L, Bolstad BM, Irizarry RA. affy--analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 2004; 20:307-315.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics.Genome Biol 2004;5:R80.

Gorter JA, van Vliet EA, Aronica E, Breit T, Rauwerda H, Lopes da Silva FH et al. Potential new antiepileptogenic targets indicated by microarray analysis in a rat model for temporal lobe epilepsy. J Neurosci. 2006; 25:11083-110.

Gloor P. Mesial temporal sclerosis: historical background and overview from a modern perspective. In: Luders H, (Ed.) Epilepsy surgery. New York: Raven press, 1991: 689-703.

Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. Nature. 1999;402(6761 Suppl):C47-52.

Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 2009;4:44-57.

Jamali S, Bartolomei F, Robaglia-Schlupp A, Massacrier A, Peragut JC, Regis J, et al. Large-scale expression study of human mesial temporal lobe epilepsy: evidence for dysregulation of the neurotransmission and complement systems in the entorhinal cortex. Brain. 2006; 129:625-641.

Ji P, Diederichs S, Wang W, Böing S, Metzger R, Schneider PM, et al. MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. Oncogene. 2003; 22:8031-8041.

Kobayashi E, Lopes-Cendes I, Guerreiro CA, Sousa SC, Guerreiro MM, Cendes F. Seizure outcome and hipocampal atrophy in familial mesial temporal lobe epilepsy. Neurology 2001; 56:166-172.

Kobayashi E, Li LM, Lopes-Cendes I, Cendes F. MRI evidence of hippocampal sclerosis in asymptomatic first degree relatives of patients with familial mesial temporal lobe epilepsy. Arch Neurology 2002; 59:1891-1894.

Kobayashi E, D'Agostino MD, Lopes-Cendes I, Andermann E, Dubeau F, Guerreiro CA et al. Outcome of surgical treatment in familial mesial temporal lobe epilepsy. Epilepsia.2003;44:1080-1084.

Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods. 2001;25:402-408

Majores M, Eils J, Wiestler OD, Becker AJ. Molecular profiling of temporal lobe epilepsy: comparison of data from human tissue samples and animal models. Epilepsy Res. 2004; 60:173-178.

Matzilevich DA, Rall JM, Moore AN, Grill RJ, Dash PK: High-density microarray analysis of hippocampal gene expression following experimental brain injury. J Neurosci Res 2002; 67:646-663.

Maurer-Morelli CV, Secolin R, Domingues RR, Marchesini RB, Santos NF, Kobayashi E Cendes F, Lopes-Cendes I. Identification of the Locus for Familial Mesial Temporal Lobe Epilepsy with Hippocampal Atrophy and Search for Candidate Genes. In: 59th Annual Meeting - American Academy of Neurology, 2007, Boston - MA. Neurology, 2007; 68: A338-A338.

Pernot F, Dorandeu F, Beaup C, Peinnequin A. Selection of reference genes for real-time quantitative reverse transcription-polymerase chain reaction in hippocampal structure in a murine model of temporal lobe epilepsy with focal seizures. J Neurosci Res. 2010; 88(5):1000-8.

Pitkänen A & Lukasiuk K. Mechanisms of epileptogenesis and potential treatment targets. Lancet Neurol 2011; 10: 173–186.

Tang FR, Chia SC, Chen PM, Gao H, Lee WL, Yeo TS, et al., Metabotropic glutamate receptor 2/3 in the hippocampus of patients with mesial temporal lobe epilepsy, and of rats and mice after pilocarpine-induced status epilepticus. Epilepsy Res. 2004; 59:167-180.

van Gassen KL, de Wit M, Koerkamp MJ, Rensen MG, van Rijen PC, Holstege FC, Lindhout D, de Graan PN. Possible role of the innate immunity in temporal lobe epilepsy. Epilepsia 2008; 49: 1055-1065.

Wang YY, Smith P, Murphy M, Cook M. Global expression profiling in epileptogenesis: does it add to the confusion? Brain Pathol. 2010; 20:1-16.

Wierschke S, Gigout S, Horn P, Lehmann TN, Dehnicke C, Bräuer AU, Deisz RA. Evaluating reference genes to normalize gene expression in human epileptogenic brain tissues. Biochem Biophys Res Commun. 2010; 403(3-4):385-90.

Wieser HG, Ortega M, Friedman A, Yonekawa Y. Long-term seizure outcomes following amygdalohippocampectomy. J Neurosurg 2003; 98:751-63.

Xu JH, Long L, Tang YC, Zhang JT, Hut HT, Tang FR. CCR3, CCR2A and macrophage inflammatory protein (MIP)-1a, monocyte chemotactic protein-1 (MCP-1) in the mouse hippocampus during and after pilocarpine-induced status epilepticus (PISE) Neuropathol Appl Neurobiol. 2009; 35 :496-514.

Yasuda CL, Tedeschi H, Oliveira EL, Ribas GC, Costa AL, Cardoso TA, et al. Comparison of short-term outcome between surgical and clinical treatment in temporal lobe epilepsy: a prospective study. Seizure. 2006;15:35-40.

Yasuda CL, Morita ME, Alessio A, Pereira AR, Balthazar ML, Saúde AV, et al. Relationship between environmental factors and gray matter atrophy in refractory MTLE. Neurology. 2010;74:1062-1068.

Figure Legends

Figure 1 – Hierarchical cluster analysis depicting the expression of positive-FH and negative-FH MTLE versus control group, and positive-FH versus negative-FH MTLE. Colorgram reflects the global differences between samples. Columns represent different samples and rows represent different genes. Gene expression is shown with a pseudocolor scale with red color denoting high expression level and green color denoting low expression level. The distances among high-dimensional expression profiles are represented as a dendrogram that arranges the clustered genes in terms of similarity to one another. FH, familial history; MTLE, mesial temporal lobe epilepsy.

Figure 2 – Overlap in gene expression between MTLE and control groups. (**A**) Venn diagram shows the number of differently expressed genes up-regulated in positive-FH MTLE, negative-FH MTLE and control groups. (**B**) Venn diagram indicates the number of differently expressed genes down-regulated in positive-FH MTLE, negative-FH MTLE and control groups. FH, familial history; MTLE, mesial temporal lobe epilepsy.

Figure 3 – Enrichment analysis of positive-FH *versus* negative-FH MTLE. The analysis was performed using the DAVID tool. The Bar graphs depict the enriched Gene Ontology process categories and -log of the P value as well as the EASE score (Escore) for a cluster of related GO categories. The P value depicts the significance of enrichment, the smaller is the P value the more significant is the enrichment. FH, familial history; MTLE, mesial temporal lobe epilepsy.

Figure 4 – Activated signaling pathways in (A) positive-FH and (B) negative-FH MTLE. The statistical threshold (orange line without squares) represents the cut-off to the significance in the log scale (y axis, left). The ratio (orange line with squares) of the number of significant genes from a group of data correlated with a signaling pathway divided by the total number of genes of the signaling pathway is also demonstrated (y axis, right). The analyses were performed with Ingenuity Pathways Analysis software. FH, familial history; MTLE, mesial temporal lobe epilepsy.

Figure 5 – Validation of a selected subset of positive-FH and negative-FH MTLE genes by quantitative RT-PCR. Validation was conducted using positive-FH and negative-FH MTLE patients and control samples (up to n = 9 per group). The gene symbol is listed for each analysis and *HPRT* gene was used as endogenous control. The median RNAm expression value in the control group has been set to 1.0 and the relative expression of all the other samples where calculated in comparison to this sample. FH, familial history; MTLE, mesial temporal lobe epilepsy.

Supplementary Figure 1 – Enrichment analysis of negative-FH MTLE *versus* health control group. The analysis was performed using the DAVID tool. The Bar graphs depict the enriched Gene Ontology process categories and -log of the P value as well as the EASE score (Escore) for a cluster of related GO categories. The P value depicts the significance of enrichment, the smaller is the P value the more significant is the enrichment. FH, familial history; MTLE, mesial temporal lobe epilepsy.

Supplementary Figure 2 – Enrichment analysis of positive-FH MTLE *versus* health control group. The analysis was performed using the DAVID tool. The Bar graphs

depict the enriched Gene Ontology process categories and -log of the P value as well as the EASE score (Escore) for a cluster of related GO categories. The P value depicts the significance of enrichment, the smaller is the P value the more significant is the enrichment. FH, familial history; MTLE, mesial temporal lobe epilepsy.

Table 1 - Clinical characteristics of patients with mesial temporal lobe epilepsy submitted to microarray analysis.

Individual	Gender	FH of epilepsy	Lateralization	Seizures/Month	Antiepileptic medication	Duration of epilepsy (years)
P14A	M	positive	RHA	12	Valproate	34
P16A	M	positive	BHA	4	Carbamazepine	48
P17A	M	positive	LHA	10	Carbamazepine/ Clobazam	31
P04	F	positive	BHA	3	Carbamazepine/ Clobazam	37
P06	M	negative	HA	10	Carbamazepine/ Clobazam	57
P10	F	negative	HA	15	Carbamazepine/ Clobazam	36
P11	F	negative	RHA	4	Carbamazepine/ Clobazam	36
P13	M	negative	HA	8	Clobazam/ Lamotrigine	8

FH – familial history; LHA- left hippocampal atrophy; RHA- right hippocampal atrophy; BHA- bilateral hippocampal atrophy.

Figure 1

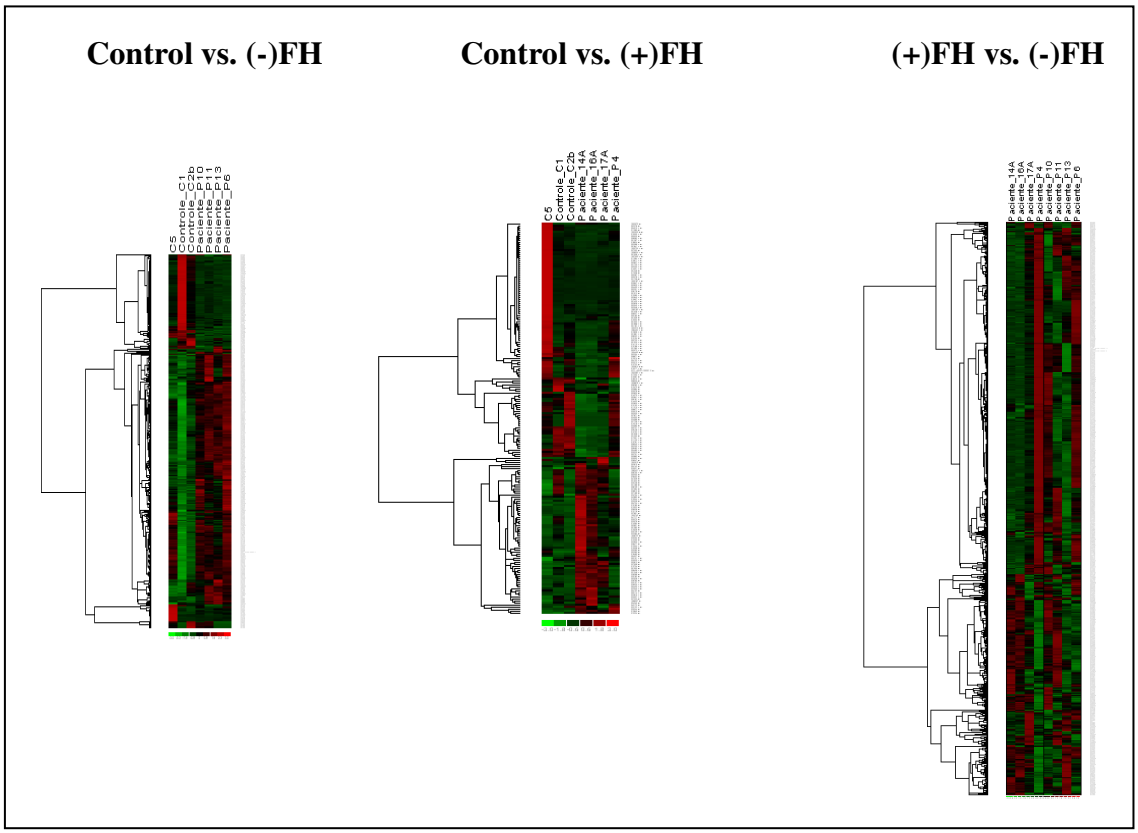


Figure 2

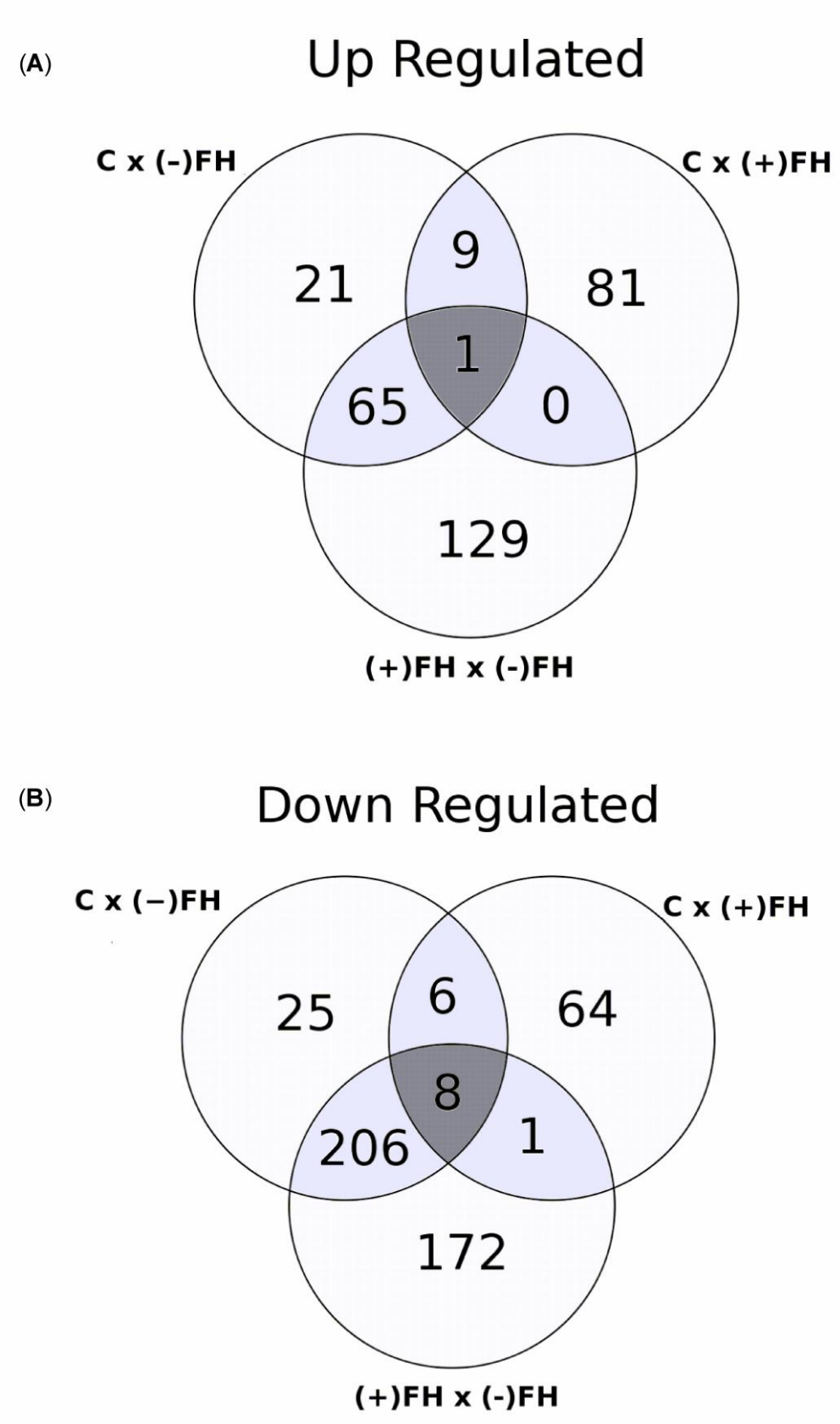

(A)
Up Regulated
C x (-)FH
C x (+)FH
21
9
81
1
65
0
129
(+)FH x (-)FH
(B)
Down Regulated
C x (−)FH
C x (+)FH
25
6
64
8
206
1
172
(+)FH x (-)FH

Figure 3

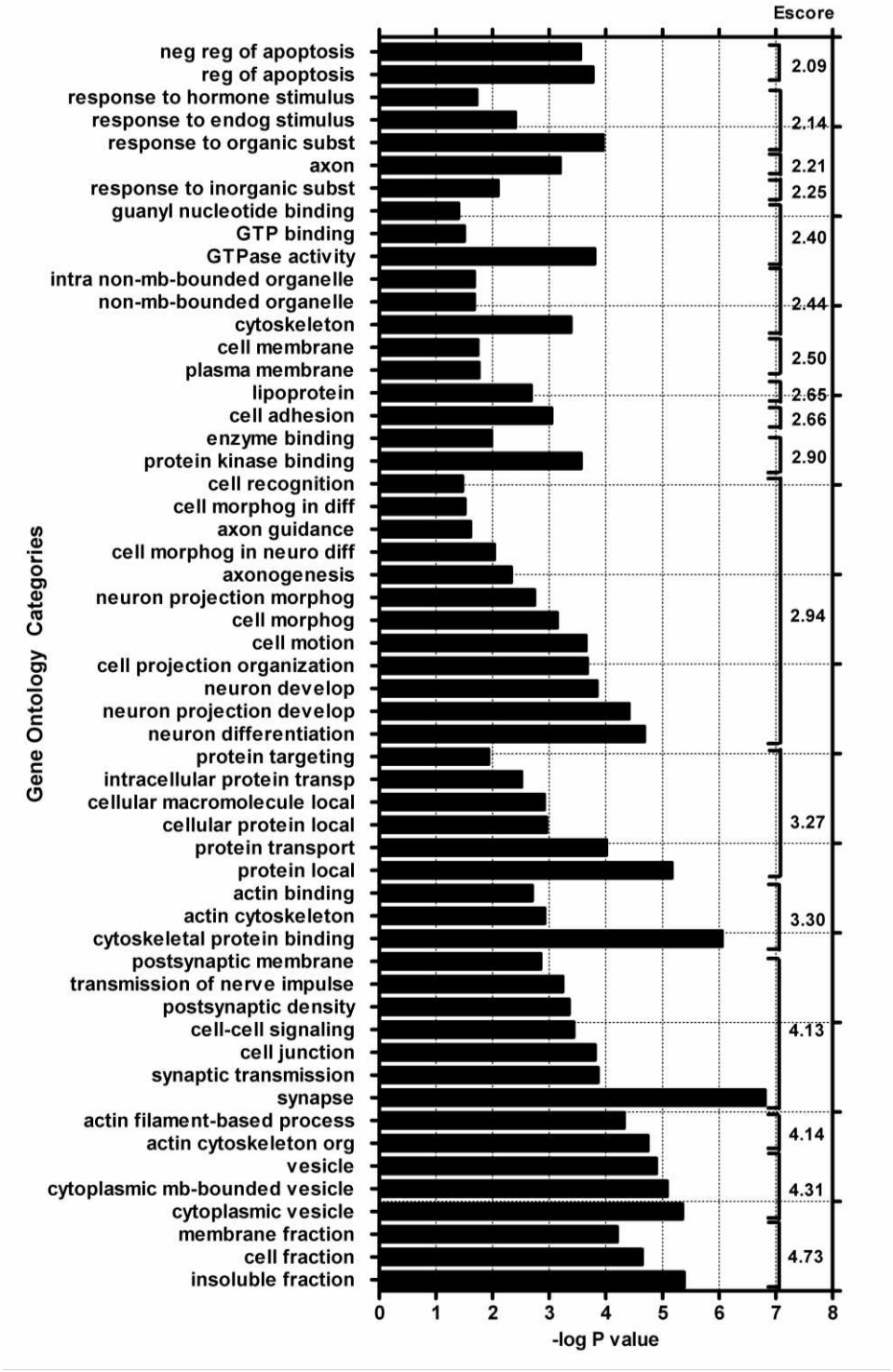
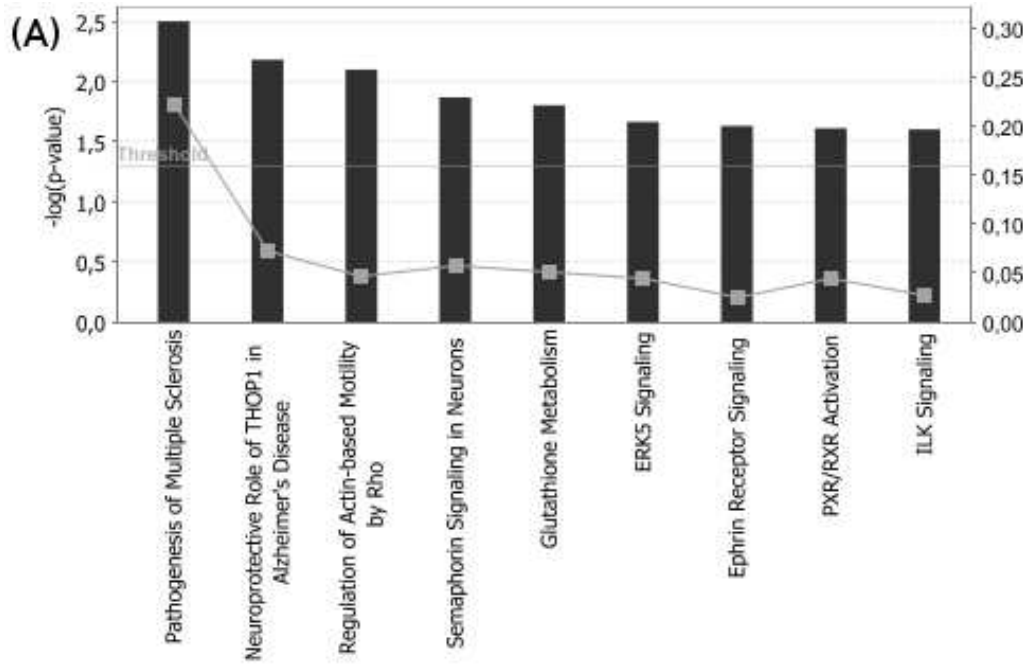

Escore
neg reg of apoptosis
reg of apoptosis
response to hormone stimulus
response to endog stimulus
response to organic subst
axon
response to inorganic subst
guanyl nucleotide binding
GTP binding
GTPase activity
intra non-mb-bounded organelle
non-mb-bounded organelle
cytoskeleton
cell membrane
plasma membrane
lipoprotein
cell adhesion
enzyme binding
protein kinase binding
cell recognition
cell morphog in diff
axon guidance
cell morphog in neuro diff
axonogenesis
neuron projection morphog
cell morphog
cell motion
cell projection organization
neuron develop
neuron projection develop
neuron differentiation
protein targeting
intracellular protein transp
cellular macromolecule local
cellular protein local
protein transport
protein local
actin binding
actin cytoskeleton
cytoskeletal protein binding
postsynaptic membrane
transmission of nerve impulse
postsynaptic density
cell-cell signaling
cell junction
synaptic transmission
synapse
actin filament-based process
actin cytoskeleton org
vesicle
cytoplasmic mb-bounded vesicle
cytoplasmic vesicle
membrane fraction
cell fraction
insoluble fraction
2.09
2.14
2.21
2.25
2.40
2.44
2.50
2.65
2.66
2.90
2.94
3.27
3.30
4.13
4.14
4.31
4.73
Gene Ontology Categories
0
1
2
3
4
5
6
7
8
-log P value

Figure 4


(A)
-log(p-value)
2,5
2,0
1,5
1,0
0,5
0,0
0,30
0,25
0,20
0,15
0,10
0,05
0,00
Threshold
Pathogenesis of Multiple Sclerosis
Neuroprotective Role of THOP1 in Alzheimer's Disease
Regulation of Actin-based Motility by Rho
Semaphorin Signaling in Neurons
Glutathione Metabolism
ERK5 Signaling
Ephrin Receptor Signaling
PXR/RXR Activation
ILK Signaling

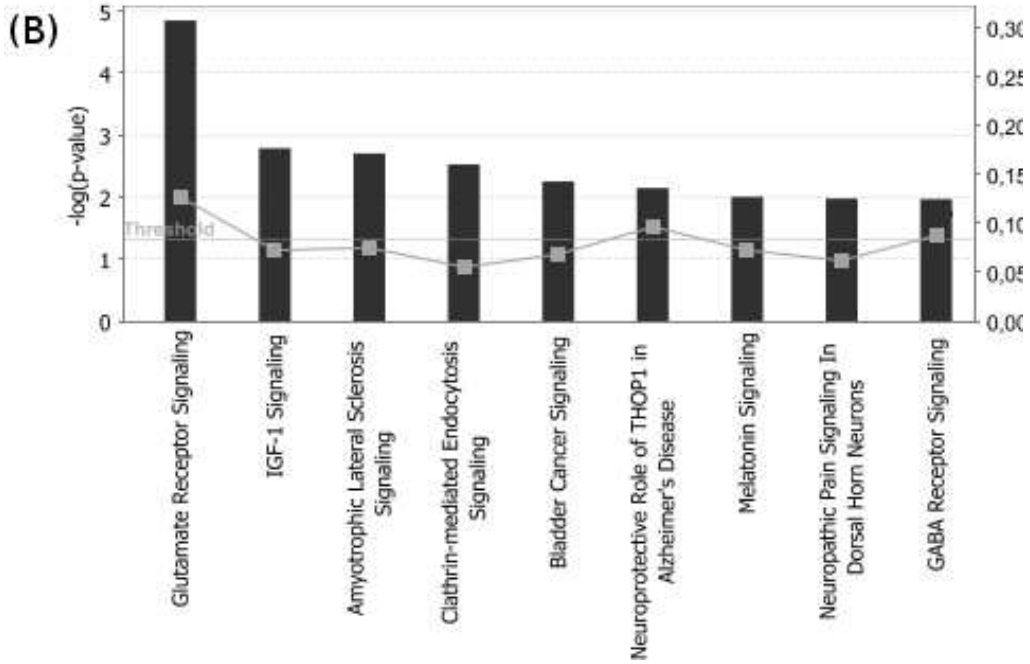
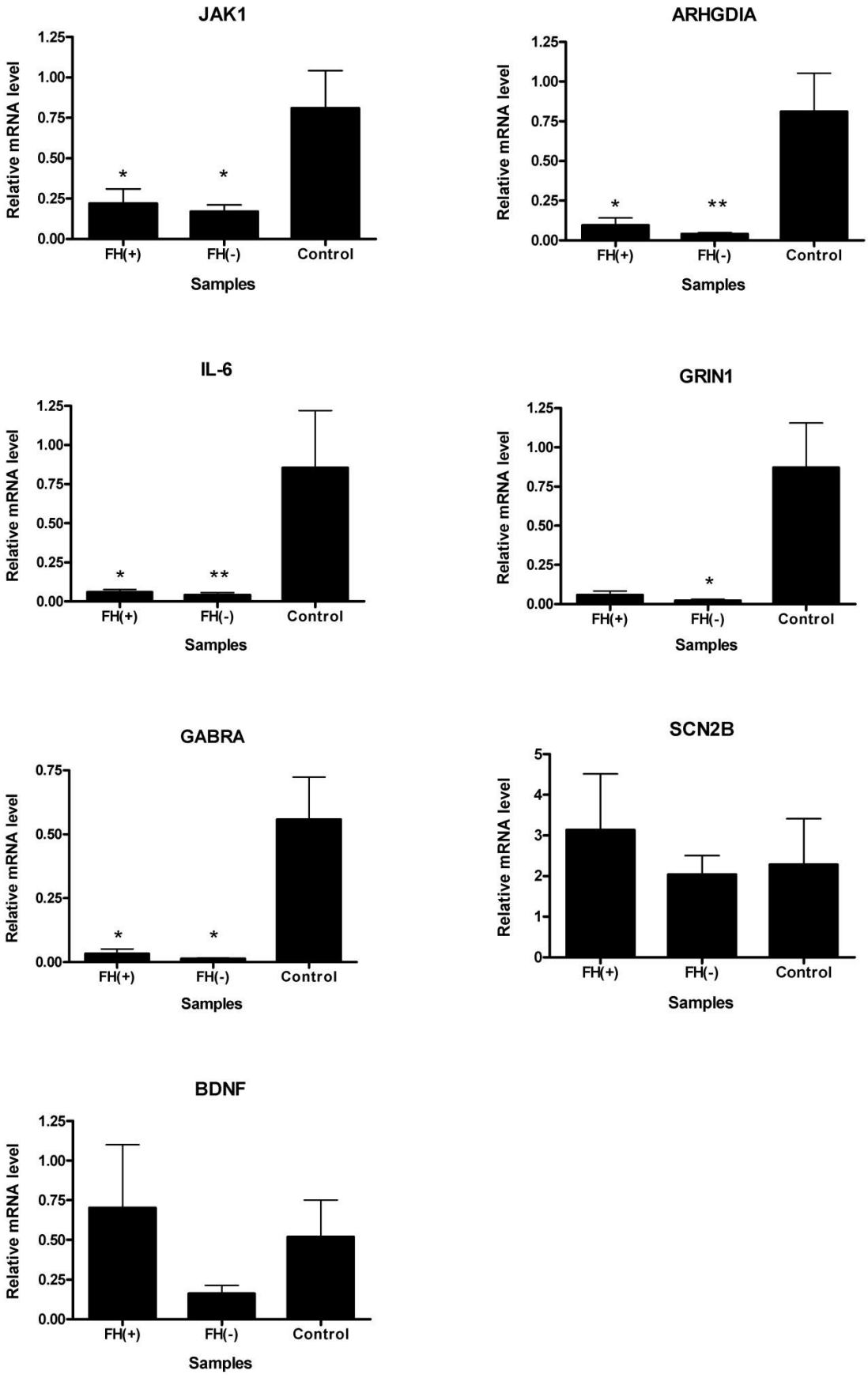
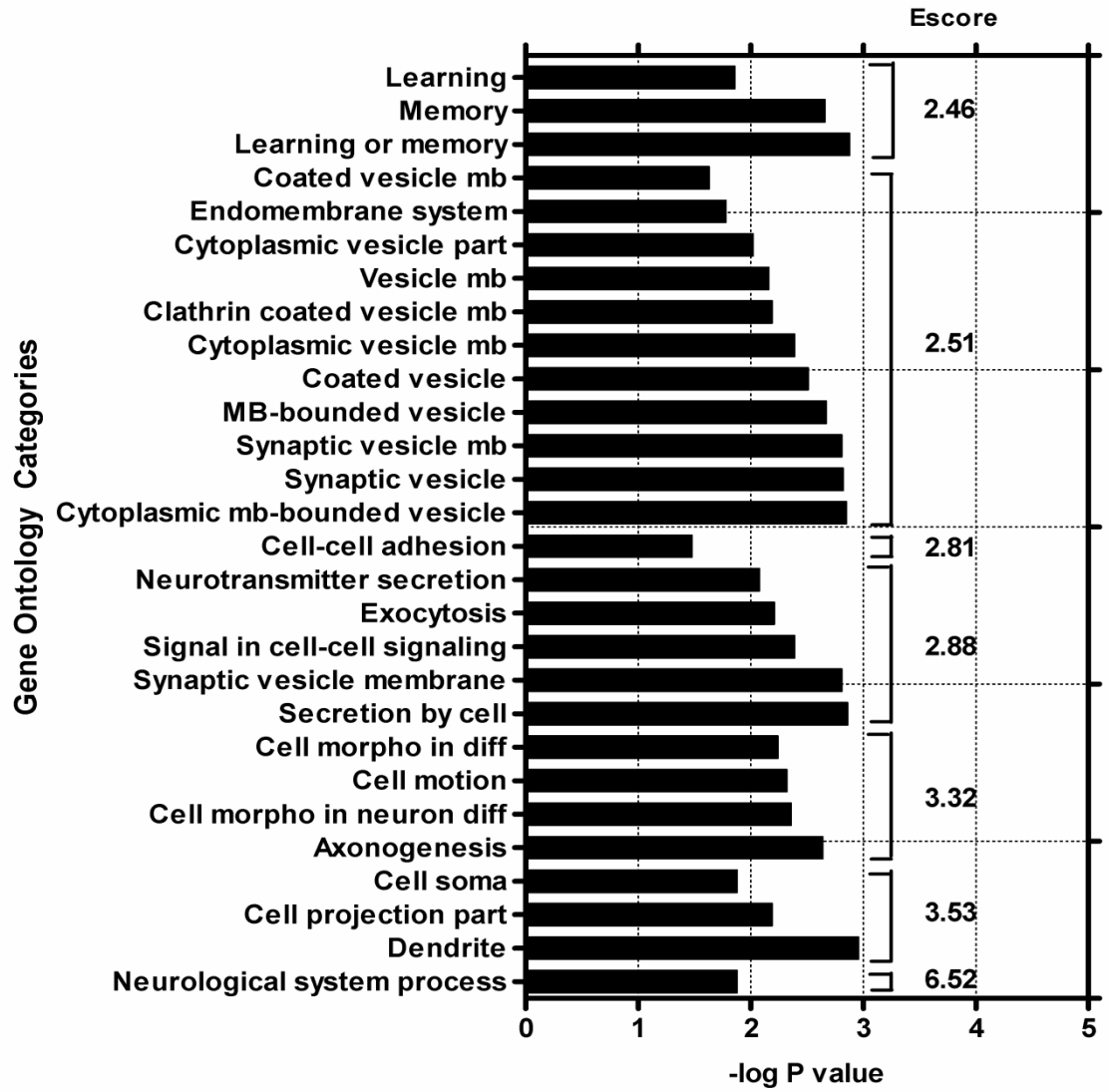

(B)
-log(p-value)
5
4
3
2
1
0
Threshold
Glutamate Receptor Signaling
IGF-1 Signaling
Amyotrophic Lateral Sclerosis Signaling
Clathrin-mediated Endocytosis Signaling
Bladder Cancer Signaling
Neuroprotective Role of THOP1 in Alzheimer's Disease
Melatonin Signaling
Neuropathic Pain Signaling In Dorsal Horn Neurons
GABA Receptor Signaling

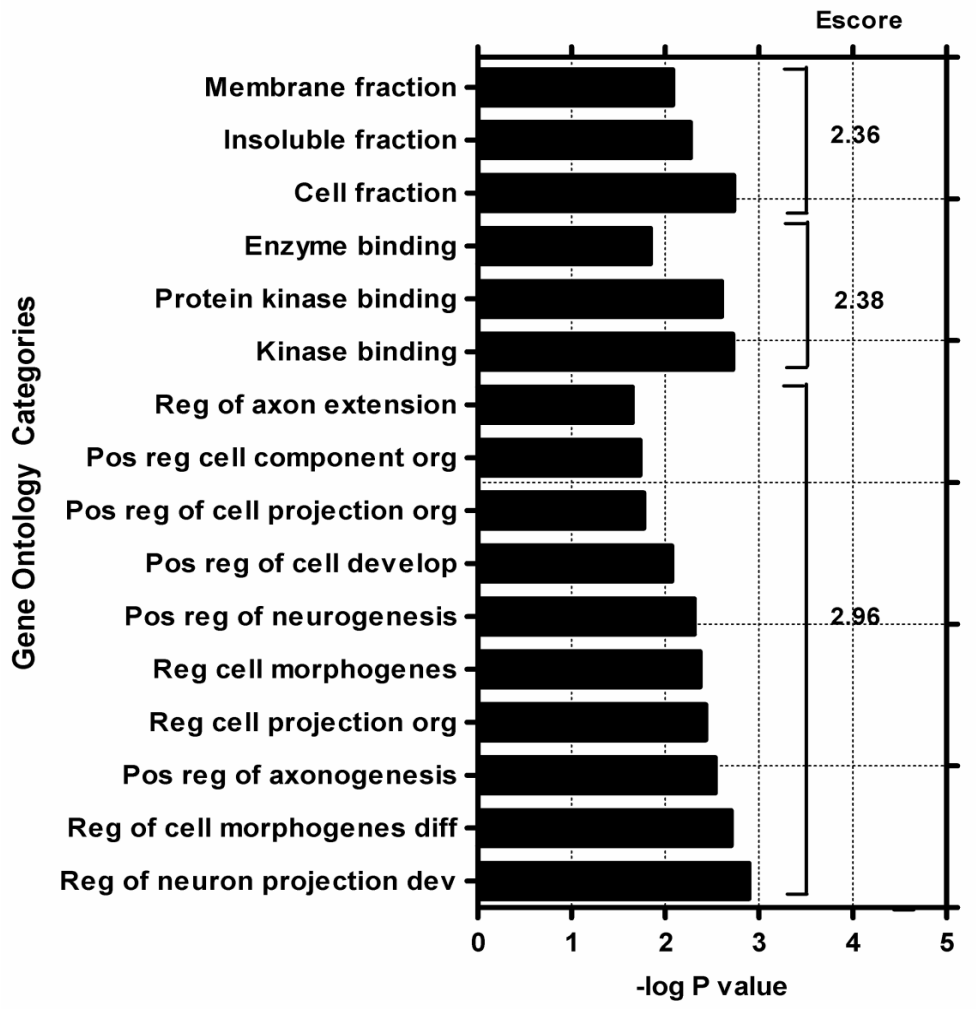
Figure 5


JAK1
Relative mRNA level
1.25
1.00
0.75
0.50
0.25
0.00
*
*
FH(+)
FH(-)
Control
Samples
ARHGDIA
Relative mRNA level
1.25
1.00
0.75
0.50
0.25
0.00
*
**
FH(+)
FH(-)
Control
Samples
IL-6
Relative mRNA level
1.25
1.00
0.75
0.50
0.25
0.00
*
**
FH(+)
FH(-)
Control
Samples
GRIN1
Relative mRNA level
1.25
1.00
0.75
0.50
0.25
0.00
*
FH(+)
FH(-)
Control
Samples
GABRA
Relative mRNA level
0.75
0.50
0.25
0.00
*
*
FH(+)
FH(-)
Control
Samples
SCN2B
Relative mRNA level
5
4
3
2
1
0
FH(+)
FH(-)
Control
Samples
BDNF
Relative mRNA level
1.25
1.00
0.75
0.50
0.25
0.00
FH(+)
FH(-)
Control
Samples

Supplementary Figure 1



Supplementary Figure 2



Apêndice 3: Artigo “MicroRNA Expression Profile in Murine Central Nervous System Development” que utilizou diversas ferramentas desenvolvidas neste trabalho.

J Mol Neurosci (2008) 35:331–337
DOI 10.1007/s12031-008-9068-4

MicroRNA Expression Profile in Murine Central Nervous System Development

Danyella B. Dogini · Patrícia A. O. Ribeiro · Cristiane Rocha · Tiago C. Pereira · Iscia Lopes-Cendes

Received: 20 February 2008 / Accepted: 13 March 2008 / Published online: 2 May 2008

Abstract MicroRNAs (miRNAs) regulate gene expression in a post-transcriptional sequence-specific manner. In order to better understand the possible roles of miRNAs in central nervous system (CNS) development, we examined the expression profile of 104 miRNAs during murine brain development. We obtained brain samples from animals at embryonic days (E) E15, E17, and postnatal days (P) P1 and P7. Total RNA was isolated from tissue and used to obtain mature miRNAs by reverse transcription. Our results indicate that there is a group of 12 miRNAs that show a distinct expression profile, with the highest expression during embryonic stages and decreasing significantly during development. This profile suggests key roles in processes occurring during early CNS development.

Keywords miRNA · Quantitative RT-PCR · Mouse · Brain development

Introduction

MicroRNAs (miRNAs) are a class of small endogenous noncoding RNAs that negatively regulate gene expression in a post-transcriptional sequence-specific manner (Bartel 2004). In drosophila, miRNAs associate with argonaute proteins in complexes that repress protein synthesis by a double interaction with target messenger RNA (Peters and Meister 2007). MiRNAs are predicted to regulate expression of at least one-third of all human genes and are known to be of great importance in a wide variety of biological processes, including cell cycle regulation, apoptosis, cell differentiation, maintenance of stemness, and imprinting (Ketting et al. 2001; Ambros 2004; Bartel 2004; Wienholds and Plasterk 2005; Wienholds et al. 2005; Lee et al. 2004).

Generation of cellular diversity during development requires coordination of gene expression mediated by both positive and negative (post-) transcriptional regulation (Wienholds and Plasterk 2005; Kloosterman and Plasterk 2006). Therefore, it is reasonable to predict that miRNAs characteristic temporal- and spatial-restricted expression pattern probably plays a role in brain morphogenesis and neuronal fate, which are key biological processes in central nervous system (CNS) development (Mansfield et al. 2004; Nelson et al. 2004; Vo et al. 2005; Krichevsky et al. 2003, 2006; Conaco et al. 2006; Wulczyn et al. 2007).

In order to investigate the possible roles of miRNAs in murine CNS development, we examined the expression profile of 104 miRNAs during this process.

Materials and Methods

Programmed matings were carried out using specific pathogen free BALB/c/UNI mice (Mus musculus) in order to obtain embryos at specific developmental stages. Brains of three animals at the following ages were collected and frozen in liquid nitrogen: embryonic day (E) E15, E17, postnatal day (P) P1, and P7. The project was approved by the Ethics Committee on Animal Experimentation at the University of Campinas (Campinas, Brazil).

This paper was supported by the National Council for Scientific and Technological Development (CNPq) and the State of Sao Paulo Research Foundation (FAPESP).

D. B. Dogini · P. A. O. Ribeiro · C. Rocha · T. C. Pereira · I. Lopes-Cendes (✉)
Department of Medical Genetics, Faculty of Medical Sciences, University of Campinas—UNICAMP,
Tessália Vieira de Camargo, 126,
Campinas, 13084-971 Sao Paulo, Brazil
e-mail: icendes@unicamp.br

Total RNA was isolated from tissue using Trizol™ (Invitrogen, Carlbad, USA) and pools of three samples were used for each age studied. We obtained mature miRNAs by stem-loop reverse transcription as previously described (Chen et al. 2005) using the Human Panel Early Access™ kit (Applied Biosystems, Foster City, USA), which contains 104 of the 489 miRNAs listed in the Sanger miRBase database (http://microrna.sanger.ac.uk). This kit uses stem-loop primers to produce cDNAs from each expressed miRNA, by annealing to their specific 7-nucleotide seed regions. The glyceraldehyde 3-phosphate dehydrogenase (Gapdh) gene was used as an endogenous control and all experiments were performed in duplicates. An overview of the methodology is illustrated in Fig. 1.

The relative quantification (RQ) of miRNAs expression was calculated by the comparative threshold cycle, which is determined using the equation RQ ¼ 2 $^{-\Delta\Delta CT}$ (Livak and Schmittgen 2001). The values of relative gene expression were dispersed in a Gauss curve and the ANOVA test was used to access differences in gene expression among different developmental stages. Statistical analysis was carried out using the BioEstat 3.0 software (Ayres et al. 2003).

Bioinformatics studies were performed using two cluster analysis tools: the self-organizing map (SOM) and the hierarchical cluster, both using R—The R Project for Statistical Computing (R Development Core Team 2007). The SOM represents the result of a vector quantization algorithm that places a number of reference vectors into a high-dimensional input data space to approximate to its data sets in an ordered fashion and it is used to visualize metric ordering relations of input samples (Kohonen 1997). In hierarchical clustering, the data are not partitioned into a particular cluster in a single step, but instead a series of partitions takes place, which may run from a single cluster containing all objects to n clusters each containing a single object (Herrero and Dopazo 2002).

Prediction of potential mammalian miRNAs targets was carried out using three different programs simultaneously (TargetScans, miRanda and PicTar, in http://www.diana.pcbi.upenn.edu/cgi-bin/TargetCombo.cgi, Sethupathy et al. 2006). Only targets identified in the three programs used were recorded.

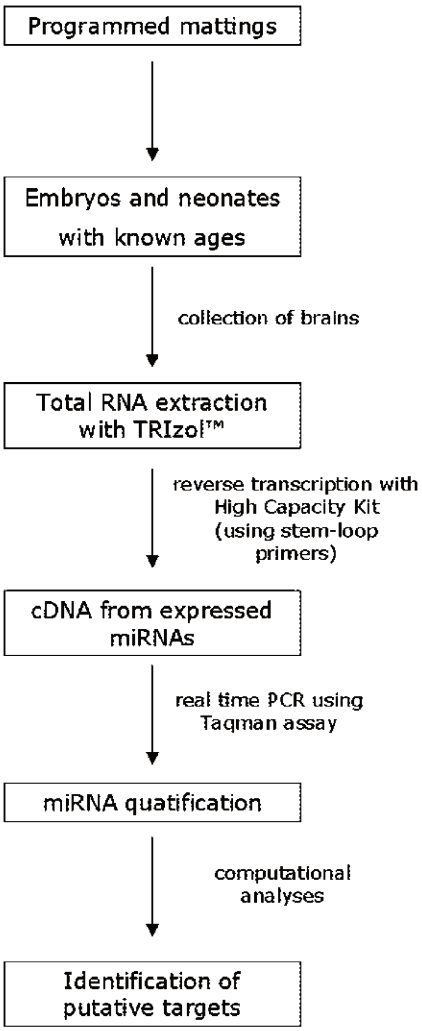


Figure 1 Overview of the methodology used for microRNA quantification. Each stem loop RT primer amplifies a specific miRNA using a 7-nucleotide complementary sequence. Generated cDNAs are quantified through real time PCR

Results

We determined the expression profile of 104 miRNAs during four stages of murine CNS development (Table 1). Values are presented as the fold change in gene expression normalized to the endogenous control and relative to the calibrator (Livak and Schmittgen 2001), which was chosen arbitrarily in our study as the expression of let7a sample in E15 stage.

Subsequently, we used the raw data obtained in the quantitative RT-PCR experiments to sort miRNAs according to their expression profile by hierarchical cluster programs. They classified them into three distinct groups depicted in Fig. 2. At the bottom section of Fig. 2, five miRNAs with the highest expression during embryonic stages can be found: miR214, miR130a, mi125a, miR9, and miR125b. In the mid-portion of Fig. 2, miRNAs with low expression in all developmental stages are depicted; the top section comprises miRNAs whose relative expression are

Table 1 Putative biological functions to target genes for microRNAs in cluster C1, according to three different programs (TargetScans, miRanda and PicTar—http://www.diana.pcbi.upenn.edu/cgi-bin/TargetCombo.cgi)

miRNA	Possible target gene	Targeted gene functions	Expression profile
miR-124a	NEUROD1 (Neurogenic differentiation factor 1)	Cell differentiation; neurogenesis	During cortical development (Przyborski et al. 2000); in humans, peak expression occurs in gestational week 19 (Franklin et al. 2001)
miR-125a	MYT1 (Myelin transcription factor 1)	Transcription factor activity; neurogenesis	Restricted to the developing nervous system—marker of primary neurogenesis (Bellefroid et al. 1996)
miR-125b	BAI1 (Brain-specific angiogenesis inhib. 1)	Inhibition of angiogenesis and neuronal differentiation; neurogenesis	Peak level in postnatal day 10, expressed in most neurons of cerebral cortex and hippocampus (Koh et al. 2001)
miR-130	GAP43 (Axonal membrane protein GAP-43)	Regulation of cell growth; neurogenesis	High expression in developing and regenerating nerve cells (Margolis et al. 1991)
miR-140	JAG1 (Jagged-1 precursor)	Regulation of cell migration; neurogenesis; control of cell specification and differentiation; DNA binding protein	In granule/germinate layer of neonatal mice (Gazit et al. 2004)
miR-205	NCAM1 (Neural cell adhesion molecule 1)	Cell-cell signaling; cell adhesion; neurogenesis	In neuronal precursor cells (Shanley and Sullivan 2007)
miR-9	CNTFR (Ciliary neurotrophic factor rec. alpha)	Enhances adult CNS neurogenesis; signal transduction; cytokine receptor; neurogenesis	Throughout the adult nervous system (Emsley and Hagg 2003)
miR-181a	SEMA4G (Semaphorin-4G precursor)	Cell differentiation; receptor activity; neurogenesis	During developing central and peripheral nervous systems as well as in several somatic tissues (Li et al. 1999)
miR-199a	ATXN7 (Spinocerebellar ataxia type 7 protein)	Nuclear organization; histone acetylation; neurogenesis	In adult mouse tissues, high expression level in brain (Strom et al. 2002)
miR-301	SNAP25 (Synaptosomal-associated protein 25)	Neurotransmitter uptake and secretion; neurogenesis	Highly expressed in adult brain (Zhao et al. 1994)

Only targets identified in the three programs used were recorded.

not as high as the group represented at the bottom, but are higher than those represented in the middle section: miR301, miR200a, miR205, miR199a, miR17–5p, miR181a, miR140, and miR124a.

The SOM analysis identified four distinct clusters of miRNAs according to their expression profile (C1, C2, C3, and C4; Fig. 3). However, only cluster C1 showed a significant difference ($p < 0.01$) in relative expression among the different ages examined. Cluster C1, which has 12 miRNAs (miR-9; miR-17–5p; miR-124a; miR-125a; miR-125b; miR-130a; miR-140; miR-181a; miR-199a; miR-205; miR-214; miR-301), has a very specific expression profile, showing high expression during embryonic stages (E15), decreasing progressively as the animals reach more mature posnatal stages (P7). The predicted mammalian targets for the 12 miRNAs in cluster 1 as well as data on their expression during development are summarized in Table 1.

Discussion

Post-transcriptional mechanisms such as alternative mRNA splicing, mRNA trafficking, and translational control are believed to play important roles in the regulation of neural gene expression (Tiedge et al. 1999; Musunuru and Darnell 2001; Maniatis and Tasic 2002; Steward and Schuman 2003). It is becoming clearer that miRNAs are an important part of a novel regulatory pathway which needs further exploration.

The nervous system is a rich source of miRNAs that are involved in regulation of biological processes (Krichevsky et al. 2006; Miska et al. 2004). There is little evidence that miRNAs are involved in the early stages of neural induction (Zhang et al. 2006). Nevertheless, the specific expression of certain miRNAs in stem cells, where they might have a role in differentiation by negatively regulating the expression of key genes, suggests that neural induction might be a rich development point for more studies (Kosik 2006; Kosik and Krichevsky 2005). Following neural induction, the evidence for a role of miRNAs in CNS development is rather strong (Zhang et al. 2006; Kosik 2006).

We reported here the expression profile of 104 miRNAs, which are distributed according to specific temporal expression patterns during murine brain development. We want to draw special attention to the 12 miRNAs identified

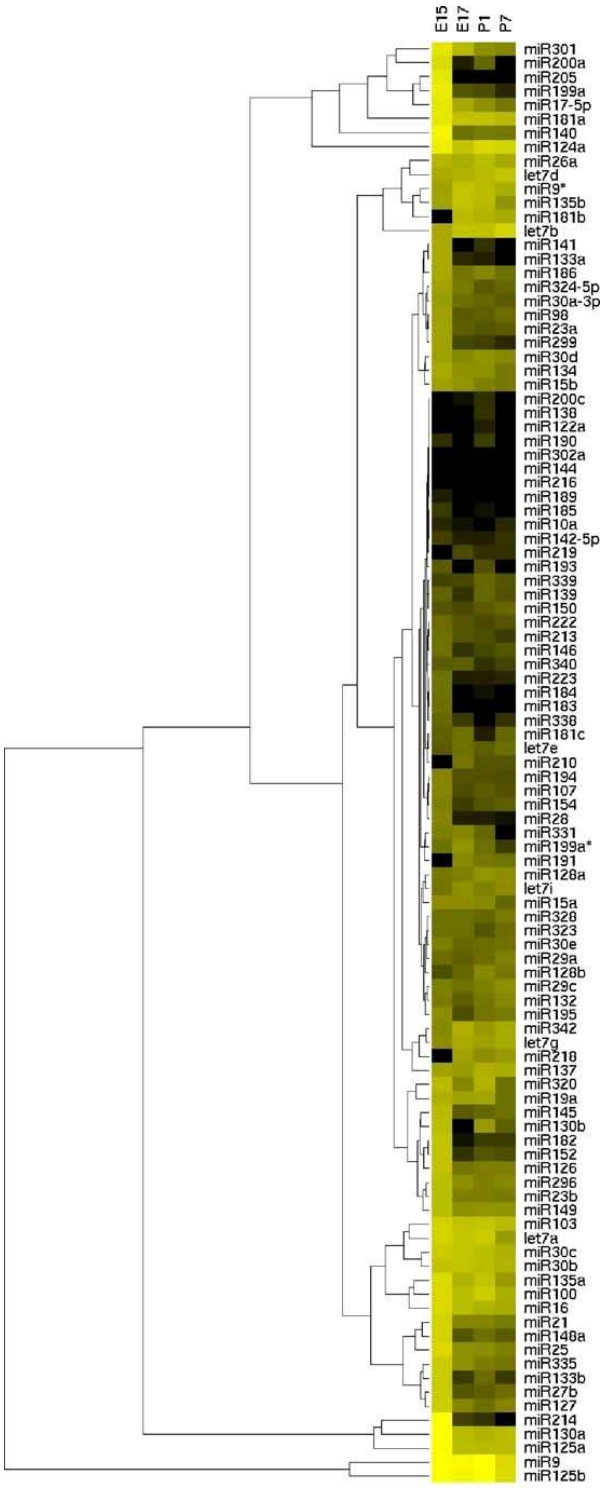


Figure 2 Results of hierarchical cluster analysis with dendrogram. Each column represents an age: E15 (embryonic day 15); E17 (embryonic day 17); P1 (postnatal day 1); and P7 (postnatal day 7). MiRNAs with the highest relative expression at E15 are at the bottom section. MiRNAs with low relative expression in all developmental stages are represented in the middle portion and miRNAs whose expressions are not as high as the bottom group, but higher than those represented in the middle section, are on the top section

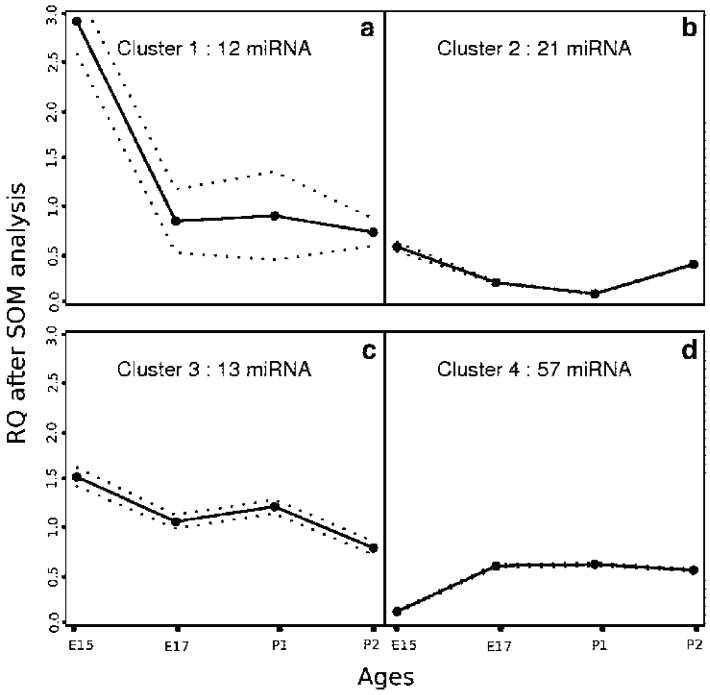


Figure 3 Results of self organization map (SOM) which generated four different clusters of miRNA expression: A cluster 1 (12 microRNAs); B cluster 2 (21 miRNAs); C cluster 3 (13 miRNAs); D cluster 4 (57 miRNAs). The standard deviation (SD) curve is represented as a dotted line (1SD)

in cluster 1, due to their distinct expression pattern: high expression during embryonic stage and decreasing as the brain develops. This profile is indicative of possible key roles in processes occurring during early CNS development. Many of these miRNAs, such as miR-124a and miR-125a, have already been reported to be up-regulated during neuronal maturation. Smirnova et al. (2005) verified in embryonic stem cells significant differences in the temporal expression pattern of a panel of highly expressed neural miRNAs. They have shown that neuron-specific miRNAs (miR-124a, miR-125a, and miR-128) gradually accumulated in parallel to neuronal maturation.

During oogenesis, the egg is loaded with several mRNAs which are needed for early development of the embryo. MiR-430 is the only abundant miRNA in the first 4–8 h after fertilization, promoting the down-regulation of maternal transcripts and allowing the zygotic genome expression. Lack of this miRNA causes a mix between the two sets of transcripts (maternal and zygotic), leading to the generation of embryos with subtle defects in gastrulation and brain morphogenesis (Cohen and Brennecke 2006). This abundant and early expression of miR-430 is conserved among vertebrates and indicate that miR-430 function in the maternal-zygotic transition may be a feature of vertebrate embryogenesis.

Strauss et al. (2006) show for the first time that cells express a miRNA signature characteristic of their developmental stage. Using RNA from mouse embryonic stem cells (ES), embryoid bodies (EBs), day 11 mouse embryos, and mature somatic tissues (heart, brain, kidney, liver, and lung), they showed that both ES/EB present a less complex miRNA expression signature than do cells that are developmentally advanced and highly restricted. They observed that the brain expressed the most complex miRNA signature and liver the least one. Since the miRNA expression signature is correlated to a particular stage of differentiation, it seems that specific miRNA expression signatures reflect commitment to particular developmental lineages.

Computational analysis for target identification revealed that 10 of the 12 miRNAs present in cluster 1 have been reported to be involved in related processes, such as cell differentiation and cell adhesion. However, neurogenesis is a major biological process which was identified in our target search and is directly related to CNS development. According to Table 1, 10 of the 12 miRNAs in cluster 1 are predicted to regulate genes whose functions could be related to different pathways involved in neurogenesis. It is interesting to note that the decrease in expression of the 12 miRNAs in cluster C1 in the latter stages of development correlate with the increase in expression of the target genes listed in Table 1.

Nevertheless, it is important to point out that although the evidence exists, no clear functional role for miRNAs in

mammalian neurogenesis has been proved at the experimental level. In 2006, Krichevsky et al. demonstrated simultaneous high expression of miR-9 and miR-124a in the transition from neural precursor (NP) to neural differentiation (ND) cell stage transition. Since miR-124a is preferentially expressed in embryonic neurons, whereas miR-9 is expressed in both neurons and glia, these results suggest that early overexpression of miR-124a in NPs prevents gliogenesis, whereas miR-9 expression contributes to neurogenesis. In addition, an inhibition of miR-9 alone or in combination with miR-124a caused a reduction of neuronal differentiation from precursor cells. These observations are indications that indeed miRNAs could be important in gene regulation at this early neuronal differentiation stage.

It is also important to consider that miRNAs that are induced during CNS development could have multiple targets and interconnected combinatorial effects. In addition, several miRNAs may cooperatively target mRNAs with similar functions, and these events may imply unlimited number of regulatory combinations created by a network of co-expressed miRNAs that contribute to a highly complex cell response. By contrast, it is also conceivable that some of these miRNAs could have the same target gene (Kim 2005).

In conclusion, we identified a cluster of 12 miRNAs with a specific expression profile in CNS during mouse development. Further studies should be carried out addressing the specific role of these miRNAs in gene regulation of biological processes involved in CNS development.

References

Ambros, V. (2004). The functions of animal microRNAs. Nature, 431, 350–355.

Ayres, M., Ayres, M., Jr., Ayres, D. L., & Dos Santos, A. S. (2003). BioEstat 3.0: Aplicações estatísticas nas áreas das Ciências Biológicas e Médicas. Belém: Sociedade civil Mamiruá, Brasília.

Bartel, D. P. (2004). MicroRNAs: Genomics, biogenesis, mechanism and functions. Cell, 23, 281–297.

Bellefroid, E. J., Bourguignon, C., Hollemann, T., Ma, Q., Anderson, D. J., Kintner, C., et al. (1996). X-MyT1, a Xenopus C2HC-type zinc finger protein with a regulatory function in neuronal differentiation. Cell, 87(7), 1191–1202.

Chen, C., Ridzon, D. A., Broomer, A. J., Zhou, Z., Lee, D. H., Nguyen, J. T., et al. (2005). Real-time quantification of micro- RNAs by stem-loop RT-PCR. Nucleic Acids Research, 33(20), e179.

Cohen, S. M., & Brennecke, J. (2006). Developmental biology. Mixed messages in early development. Science, 312(5770), 65–66.

Conaco, C., Otto, S., Han, J. J., & Mandel, G. (2006). Reciprocal actions of REST and a microRNA promote neuronal identity. Proceedings of the National Academy of Sciences of the United States of America, 103(7), 2422–2427.

Emsley, J. G., & Hagg, T. (2003). Endogenous and exogenous ciliary neurotrophic factor enhances forebrain neurogenesis in adult mice. Experimental Neurology, 183(2), 298–310.

Franklin, A., Kao, A., Tapscott, S., & Unis, A. (2001). NeuroD homologue expression during cortical development in the human brain. Journal of Child Neurology, 16(11), 849–853.

Gazit, R., Krizhanovsky, V., & Ben-Arie, N. (2004). Math1 controls cerebellar granule cell differentiation by regulating multiple components of the Notch signaling pathway. Development, 131 (4), 903–913.

Herrero, J., & Dopazo, J. (2002). Combining hierarchical clustering and self-organizing maps for exploratory analysis of gene expression pattern. Journal of Proteome Research, 1, 467–470.

Ketting, R. F., Fischer, S. E., Bernstein, E., Sijen, T., Hannon, G. J., & Plasterk, R. H. (2001). Dicer functions in RNA interference and in synthesis of small RNA involved in development timing in C. elegans. Genes & Development, 15(20), 2654–2659.

Kim, V. N. (2005). MicroRNA biogenesis: Coordinated cropping and dicing. Nature Reviews. Molecular Cell Biology, 6(5), 376–385.

Kloosterman, W. P., & Plasterk, R. H. (2006). The diverse functions of microRNAs in animal development and disease. Developmental Cell, 11(4), 441–450.

Koh, J. T., Lee, Z. H., Ahn, K. Y., Kim, J. K., Bae, C. S., Kim, H. H., et al. (2001). Characterization of mouse brain-specific angiogenesis inhibitor 1 (BAI1) and phytanoyl-CoA alpha-hydroxylase-associated protein 1, a novel BAI1-binding protein. Molecular Brain Research, 87(2), 223–237.

Kohonen, T. (1997). Self-organizing maps. New York, NY: Springer.

Kosik, K. S. (2006). The neuronal microRNA system. Nature Reviews. Neuroscience, 7(12), 911–920.

Kosik, K. S., & Krichevsky, A. M. (2005). The elegance of microRNAs: a neuronal perspective. Neuron, 47(6), 779–782.

Krichevsky, A. M., King, K. S., Donahue, C. P., Khrapko, K., & Kosik, K. S. (2003). A microRNA array reveals extensive regulation of microRNAs during brain development. RNA, 9 (10), 1274–1281.

Krichevsky, A. M., Sonntag, K. C., Isacson, O., & Kosik, K. S. (2006). Specific microRNAs modulate embryonic stem cell-derived neurogenesis. Stem Cells, 24(4), 857–864.

Lee, Y. S., Nakahara, K., Pham, J. W., Kim, K., Sontheimer, E. J., & Carthew, R. W. (2004). Distinct roles for Drosophila Dicer-1 and Dicer-2 in the siRNA/miRNA silencing pathways. Cell, 117(1), 69–81.

Li, H., Wu, D. K., & Sullivan, S. L. (1999). Characterization and expression of sema4g, a novel member of the semaphorin gene family. Mechanisms of Development, 87(1–2), 169–173.

Livak, K. J., & Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta C}$ method. Methods, 25(4), 402–408.

Maniatis, T., & Tasic, B. (2002). Alternative pre-mRNA splicing and proteome expansion in metazoans. Nature, 418(6894), 236–243.

Mansfield, J. H., Harfe, B. D., Nissen, R., Obenauer, J., Srineel, J., Chaudhuri, A., et al. (2004). MicroRNA-responsive 'sensor' transgenes uncover Hox-like and other developmentally regulated patterns of vertebrate microRNA expression. Nature Genetics, 36(10), 1079–1083.

Margolis, F. L., Verhaagen, J., Biffo, S., Huang, F. L., & Grillo, M. (1991). Regulation of gene expression in the olfactory neuroepithelium: A neurogenetic matrix. Progress in Brain Research, 89, 97–122.

Miska, E. A., Alvarez-Saavedra, E., Townsend, M., Yoshii, A., Sestan, N., Rakic, P., et al. (2004). Microarray analysis of microRNA expression in the developing mammalian brain. Genome Biology, 5(9), R68.

Musunuru, K., & Darnell, R. B. (2001). Paraneoplastic neurologic disease antigens: RNA-binding proteins and signaling proteins in neuronal degeneration. Annual Review of Neuroscience, 24, 239–262.

Nelson, P. T., Hatzigeorgiou, A. G., & Mourelatos, Z. (2004). miRNP: mRNA association in polyribosome in a human neuronal cell line. RNA, 10(3), 387–394.

Peters, L., & Meister, G. (2007). Argonaute proteins: Mediators of RNA silencing. Molecular Cell, 26(5), 611–623.

Przyborski, S. A., Morton, I. E., Wood, A., & Andrews, P. W. (2000). Developmental regulation of neurogenesis in the pluripotent human embryonal carcinoma cell line NTERA-2. European Journal of Neuroscience, 12(10), 3521–3528.

R Development Core Team (2007). R: A language and environment for statistical computing. R Foundation for Statistical Comput- ing, Vienna, Austria. (http://www.R-project.org.ent).

Sethupathy, P., Megraw, M., & Hatzigeorgiou, A. G. (2006). A guide through present computational approaches for the identification of mammalian microRNA targets. Nature Methods, 3, 881–886.

Shanley, D. K., & Sullivan, A. M. (2007). Expression of the cell surface markers mAb 2F7 and PSA-NCAM in the embryonic rat brain. Neuroscience Letters, 424(3), 165–169.

Smirnova, L., Grafe, A., Seiler, A., Schumacher, S., Nitsch, R., & Wulczyn, F. G. (2005). Regulation of miRNA expression during neural cell specification. European Journal of Neuroscience, 21 (6), 1469–1477.

Steward, O., & Schuman, E. M. (2003). Compartmentalized synthesis and degradation of proteins in neurons. Neuron, 40 (2), 347–359.

Strauss, W. M., Chen, C., Lee, C. T., & Ridzon, D. (2006). Nonrestrictive developmental regulation of microRNA gene expression. Mammalian Genome, 17(8), 833–840.

Strom, A.-L., Jonasson, J., Hart, P., Brannstrom, T., Forsgren, L., & Holmberg, M. (2002). Cloning and expression analysis of the murine homolog of the spinocerebellar ataxia type 7 (SCA7) gene. Gene, 285, 91–99.

Tiedge, H., Bloom, F. E., & Richter, D. (1999). RNA, whither goest thou? Science, 283(5399), 186–187.

Vo, N., Klein, M. E., Varlamova, O., Keller, D. M., Yamamoto, T., Goodman, R. H., et al. (2005). A cAMP-response element binding protein-indiced microRNA regulates neuronal morphogenesis. Proceedings of the National Academy of Sciences of the United States of America, 102(45), 16426–16431.

Wienholds, E., Kloosterman, W. P., Miska, E., Alvarez-Saavedra, E., Berezikov, E., de Bruijn, E., et al. (2005). MicroRNA expression in zebrafish embryonic development. Science, 309(5732), 310–311.

Wienholds, E., & Plasterk, R. H. (2005). MicroRNA function in animal development. FEBS, 597(26), 5911–5922.

Wulczyn, F. G., Smirnova, L., Rybak, A., Brandt, C., Kwidzinski, E., Ninnemann, O., et al. (2007). Post-transcriptional regulation of the let-7 microRNA during neural cell specification. FASEB Journal, 21(2), 415–426.

Zhang, B., Pan, X., & Anderson, T. A. (2006). MicroRNA: A new player in stem cells. Journal of Cellular Physiology, 209(2), 266–269.

Zhao, N., Hashida, H., Takahashi, N., & Sakaki, Y. (1994). Cloning and sequence analysis of the human SNAP25 cDNA. Gene, 145 (2), 313–314.