Uso de microarrays e RNA-seq para a medida de níveis relativos de transcrição

Medidas dos níveis de mRNA

- O nível de mRNA de uma célula reflete (as vezes de forma grosseira) os níveis de proteínas da mesma.
- Os níveis de mRNAs irão variar em diferentes tecidos como reflexo das diferentes especializações destes.
- Detecção de diferenças dos níveis de mRNA em amostras diversas pode fornecer informações sobre os diferentes mecanismos moleculares que estão atuando sobre os mesmos.
- Existem diversos métodos para medida de níveis de RNA em larga escala, entre eles podemos destacar os métodos de microarray e RNA-seq

Microarrays

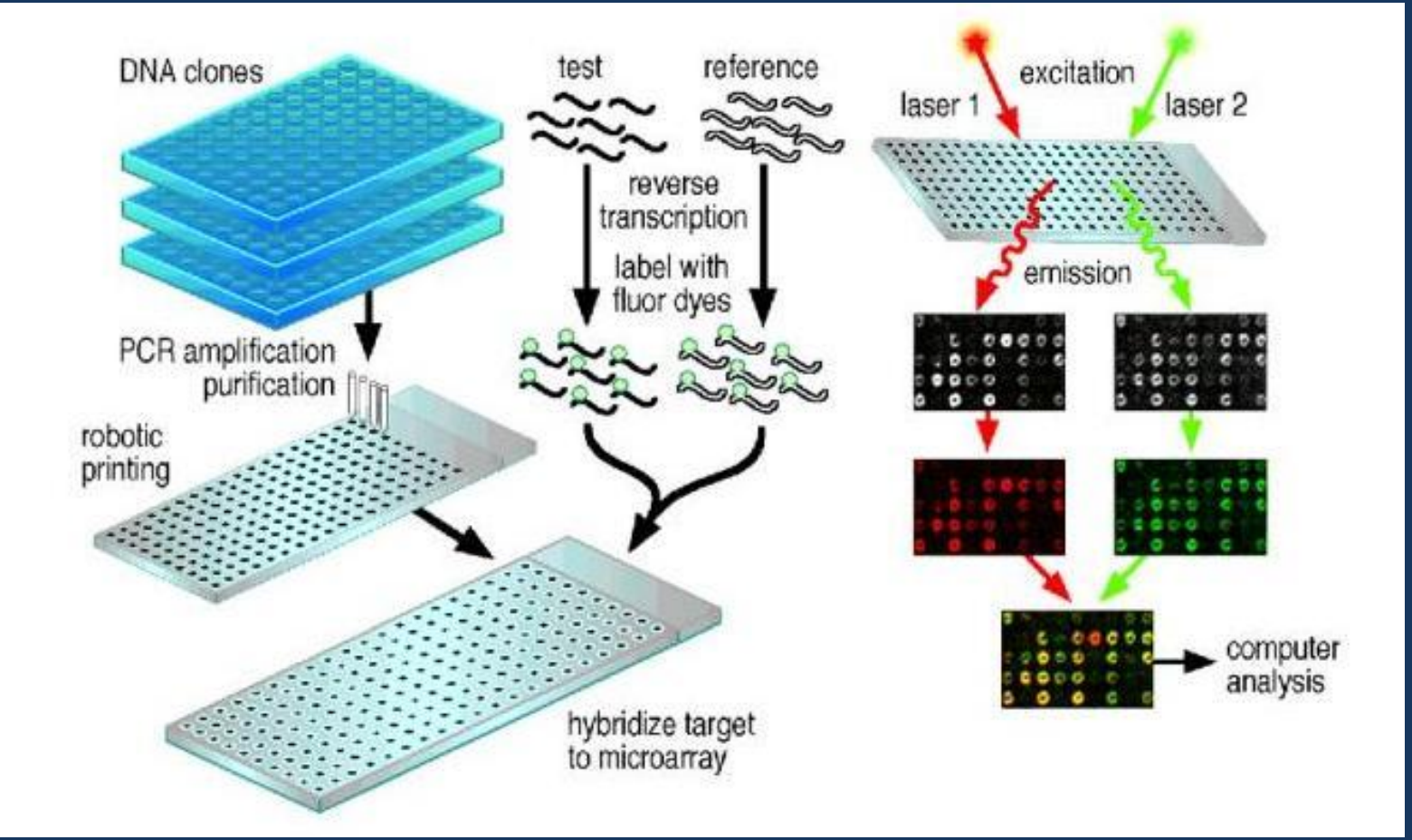
- Microarrays se utilizam de hibridização de sondas imobilizadas em um suporte solido que serão hibridizadas com amostras de cDNA permitindo a detecção destas moléculas.

- O cDNA alvos são marcados com moléculas florescente de modo que é possível realizar a sua detecção após a hibridização e o sinal obtido será proporcional ao numero de moléculas hibridizadas

- Desta forma, o sinal fluorescente obtido será proporcional a abundancia do transcrito na amostra estudada

Microarrays

Microarrays



Microarray

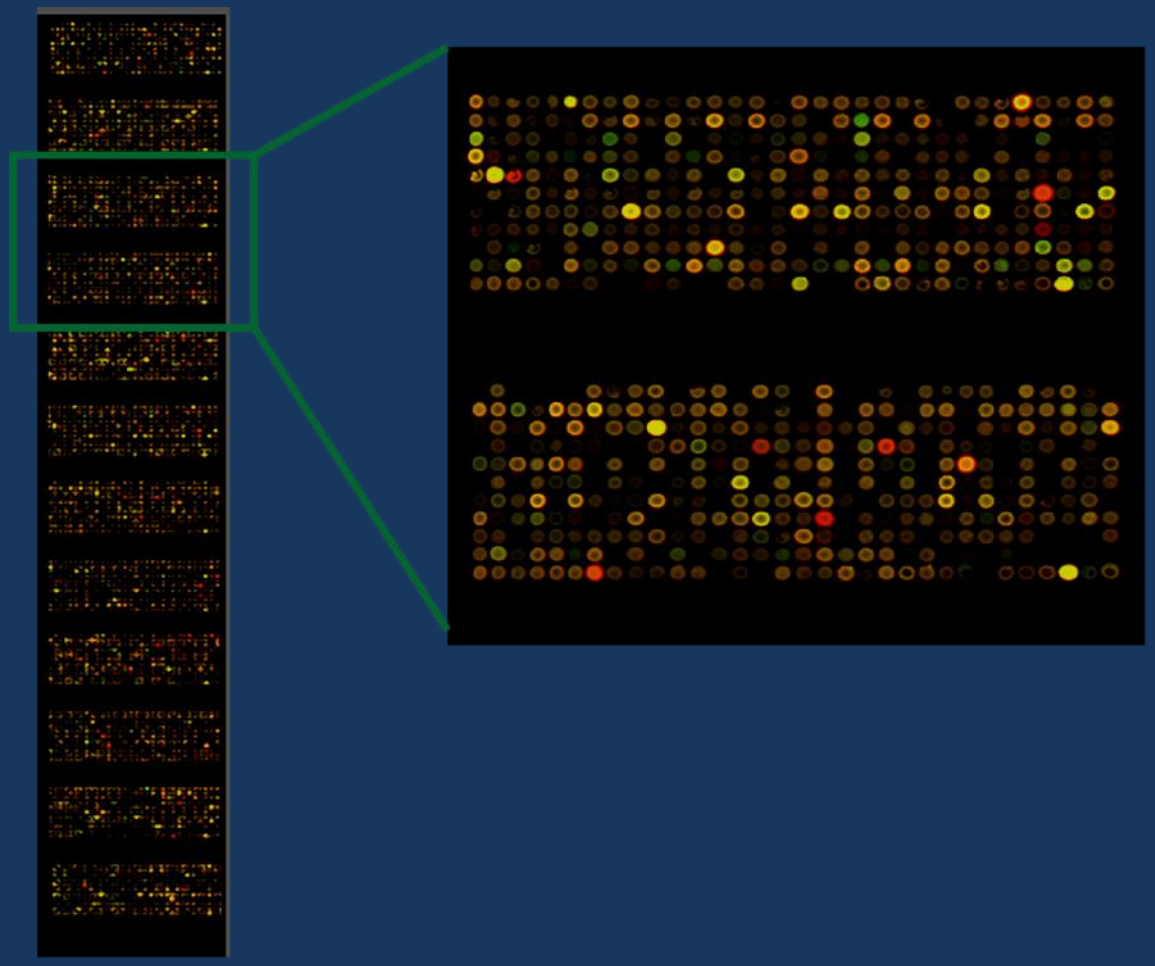
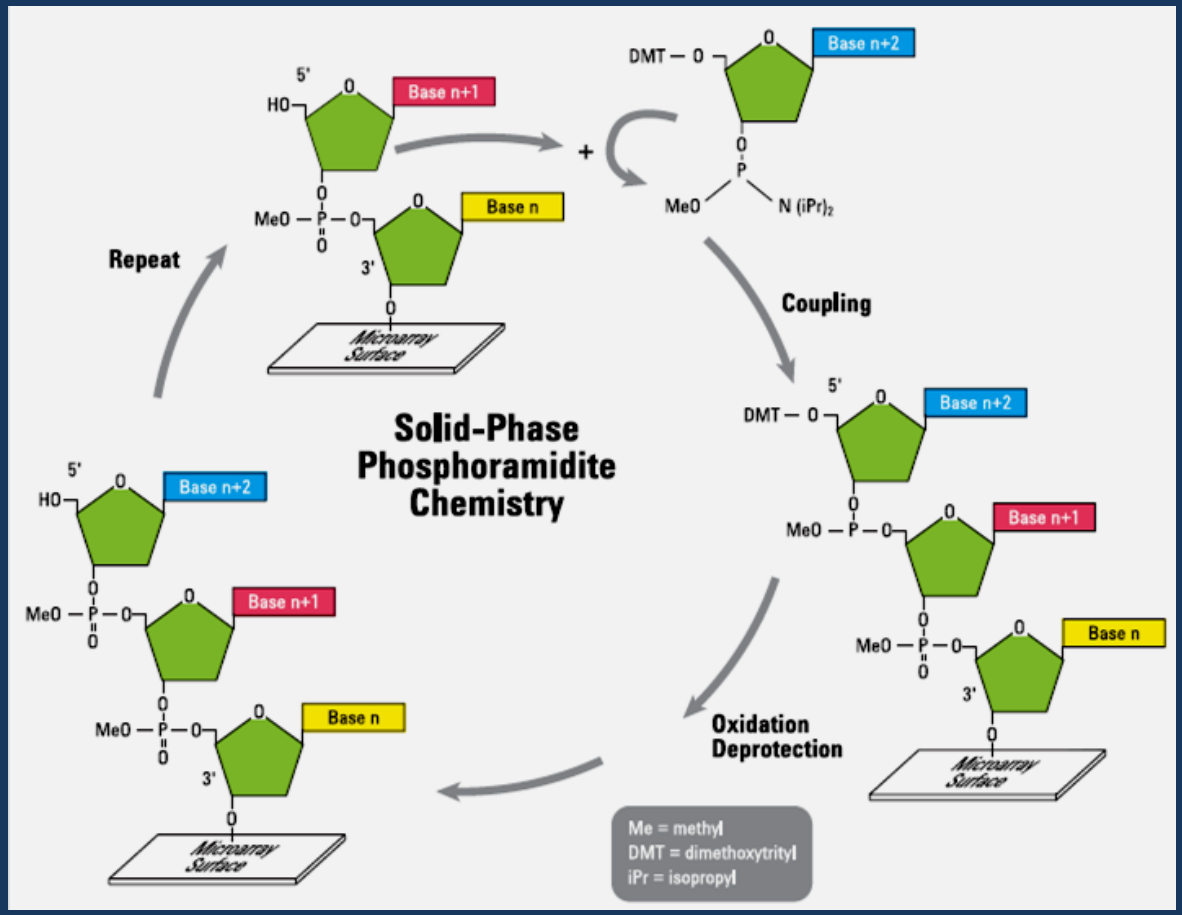


Imagem de um microarray construído a partir de cerca de 4 mil clones diferentes. Clones utilizados são derivados de um projeto de seqüenciamento de ESTs

Síntese de oligonucleotideos em fase solida



Oligoarrays

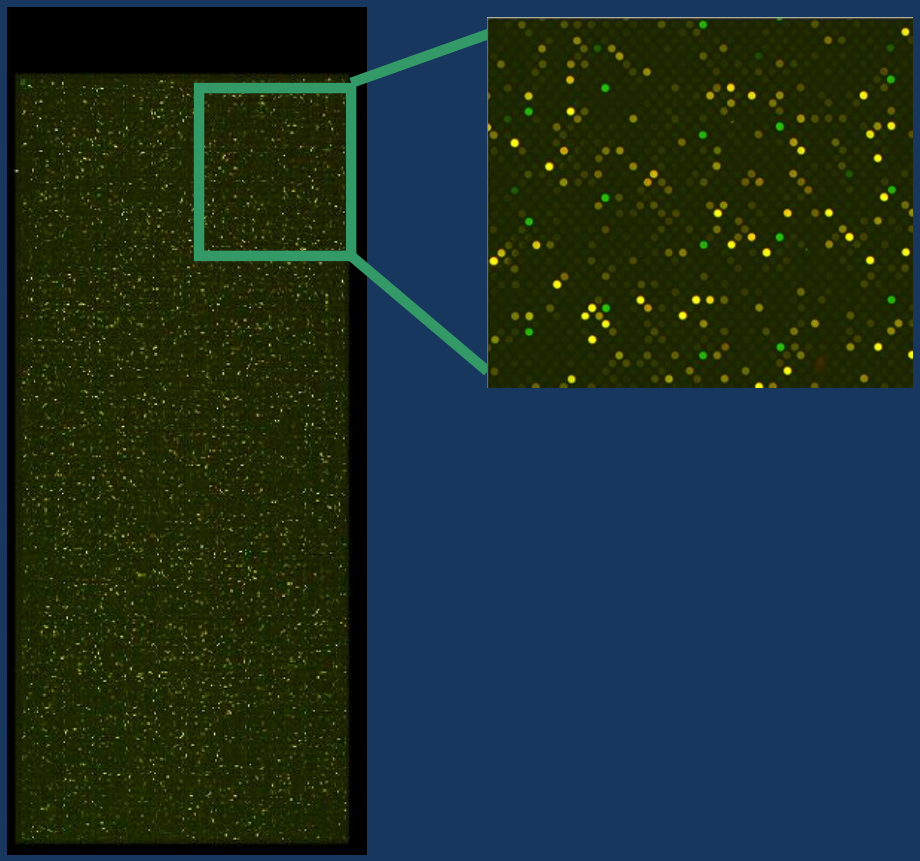
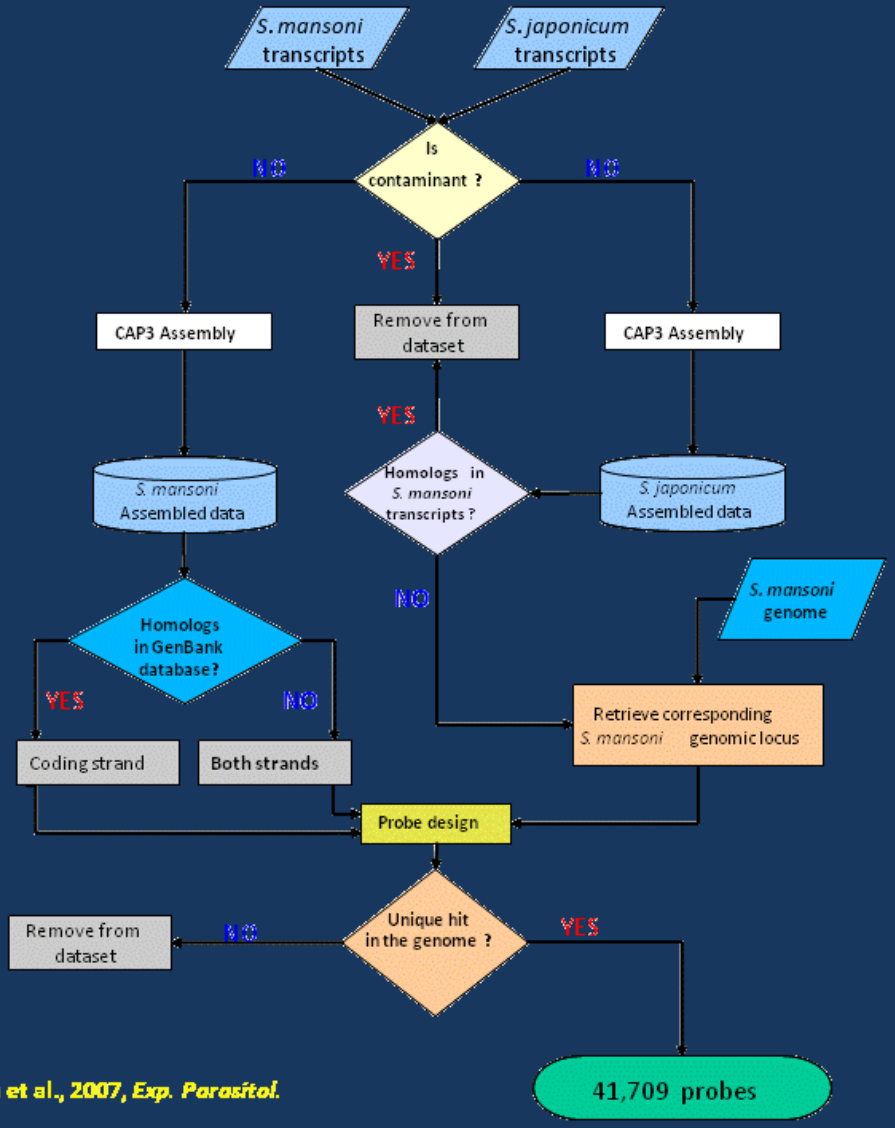


Imagem de um oligoarray de alta densidade (44 mil pontos)

Analise bioinformática para desenho de sondas

- No caso de oligoarrays as sondas de DNA precisam ser planejadas pois somente a informação de seqüência é utilizada
- É necessário escolher um grupo de seqüências não redundantes e de interesse para ser representada na plataforma
- As sondas possuem tamanho e conteúdo de GC predefinidos é necessário procurar nas seqüências escolhidas trechos que possuam estar características e sejam únicas no genoma/transcriptoma do organismo
- Sondas são fita simples e portanto é necessário definir qual é a fita (+ ou -) que é de fato transcrita

Analise bioinformatica para desenho de sondas



Verjovski-Almeida et al., 2007, *Exp. Parasitol.*

60-mer
GC content = 35-55%
Tm = 68 to 76 °C
No repetitive seq
No low complexity (<7bases)
Most 3' probe in the contig

Filtragem de dados de microarray

Devido ao fato de nem todos os pontos do array terem hibridização positiva é necessário realizar uma filtragem de pontos com sinais muito fracos

A abordagem mais comum é utilizar como referencia negativa regiões do array que não possuem DNA depositado ou possuem DNA de outra espécie como um controle negativo

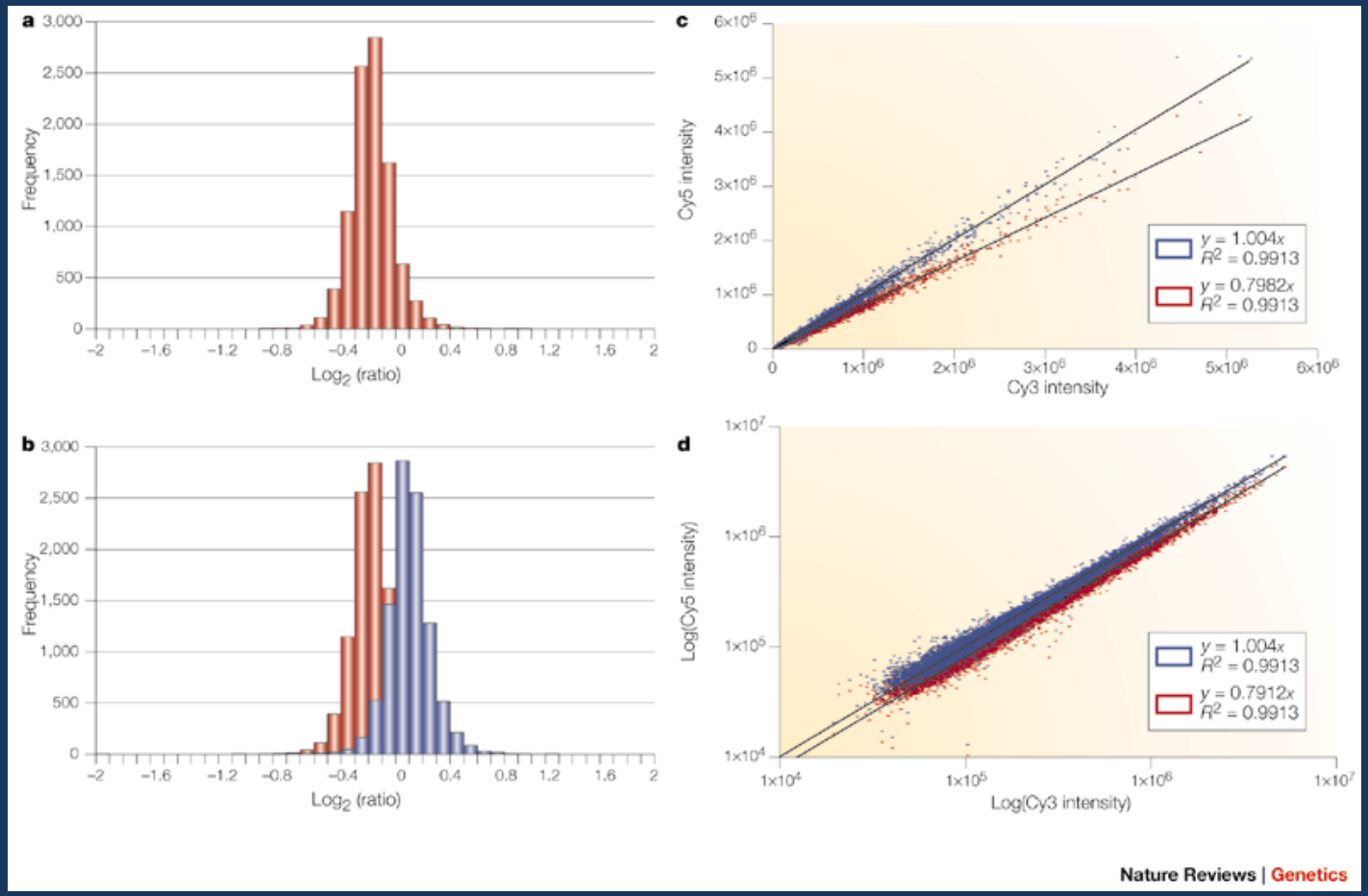
A intensidade de sinal destes controles negativos é medida e é realizado um calculo do valor médio e de desvio padrão. Dados com um valor abaixo da media+X desvio padrões do controle negativo são retirados, pois não apresentariam uma fluorescência significativa

Normalização de dados de microarray

Devido a possíveis diferenças na quantidade de cDNA marcado e nas propriedades de fluorescência das moléculas marcadoras utilizadas é necessário realizar uma normalização dos dados dos dois canis (diferentes comprimentos de onda) amostrados.

Existem diversos métodos de normalização, mas o objetivo principal destes métodos é conseguir que a maioria dos dados amostrados apresentem um valor semelhante de modo que semente dados que representem verdadeiras diferenças sejam detectados como tais

Normalização de dados de microarray



Exemplo de normalização por lowess neste caso é realizada uma normalização para corrigir distorções relacionadas a intensidade do sinal (parte do pressuposto que a razão para a maior parte dos genes é aproxim. 1)

RNA-seq

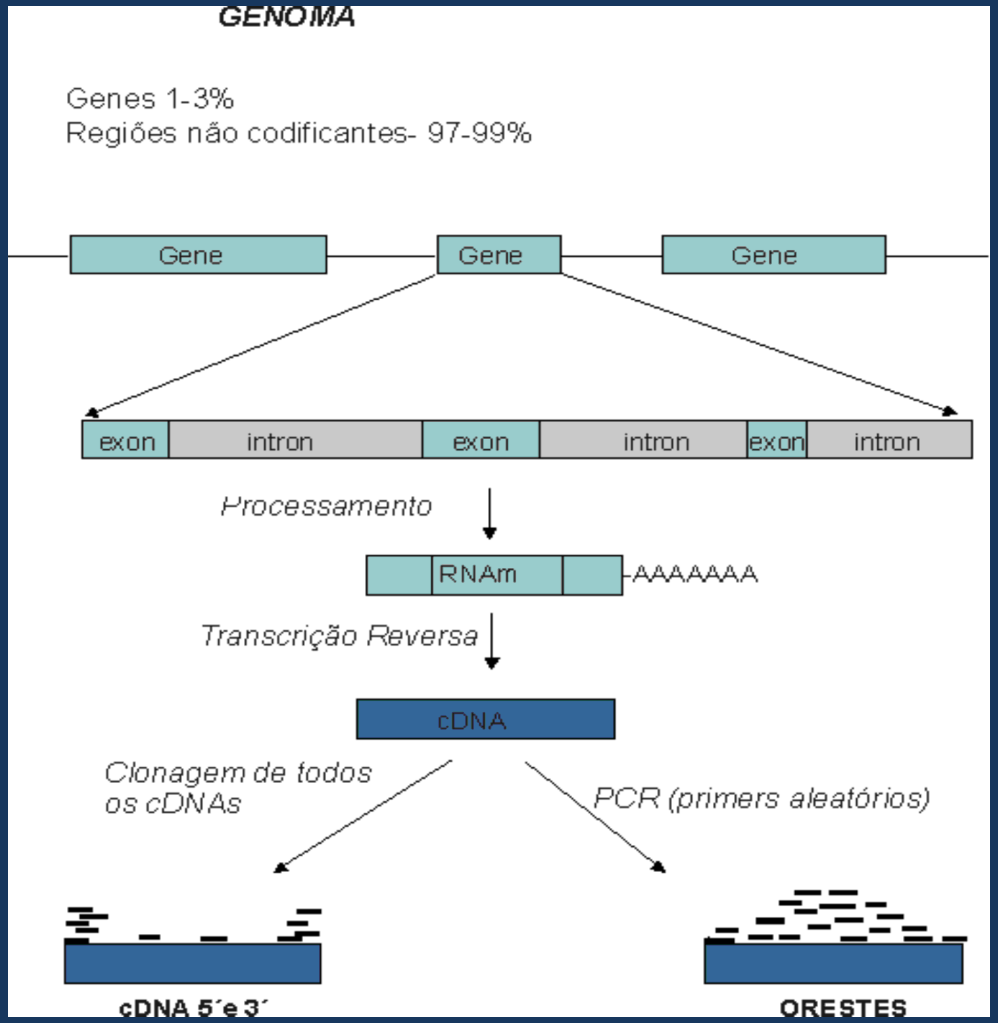
Seqüenciamento de transcriptomas

- Conforme mostrado no slide anterior organismos mais complexos tendem a possuir genes com um alto numero de exons. Além disso, o genoma destes organismos possuem uma alta quantidade de seqüências não-codificantes e portanto a predição da estrutura de genes não é trivial.

- Deste modo o seqüenciamento direto das moléculas de mRNA pode fornecer informações a respeito da estrutura de um gene, pois representa a molécula madura formada após os eventos de *splicing*

- Além disso, o seqüenciamento de mRNA permite a amostragem direta das seqüências codificantes permitindo com que um menor volume de seqüenciamento se obtenha maior informações sobre as proteínas deste organismo

Seqüenciamento de transcriptomas



Após o isolamento das moléculas de mRNA é realizada a reação de transcriptase reversa que irá gerar um fita de cDNA a partir de um mRNA molde.

Normalmente esta transcrição é realizada com um oligo-dT como primer o que permite com que o mRNA interio seja transcrito

Os cDNA produzidos são clonados e o conjunto de plasmideos produzidos é denominado biblioteca

Seqüenciamento de transcriptomas

Abundance distribution	
Total copies	55,172 (100)
>1000	5
501 to 1000	6
101 to 500	22
51 to 100	26
11 to 50	153
6 to 10	235
2 to 5	3,084
1	9,936

- Entretanto a abundancia de diferente mRNAs em uma célula varia muito. Existem alguns poucos mRNAs que possuem um numero de moléculas até 1000 X maior que a maioria dos mRNAs.
- Deste modo, seqüenciamentos em pequena escala de bibliotecas de mRNAs tendem a amostrar muito umas poucas moléculas e pouco um conjunto grande
- Além disso, nem todos os mRNA vão estar sendo expressos em um único tecido ou fase de vida do organismo e por isso para obter uma descrição completa dos mRNAs de um organismos vários destes deverão se amostrados

RNA-seq

- Com o advento das técnicas de sequenciamento em larga escala tornou-se muito mais fácil obter um volume de sequencias que permita obter uma amostragem razoável mesmo dos transcritos com menor nível de expressão.

- Partindo do pressuposto que transcritos mais abundantes irão produzir um maior numero de sequencias é possível realizar uma contagem do numero de sequencias produzidas para cada transcrito e correlacionar o numero relativo de sequencias de cada transcrito com a abundancia do mesmo.

Normalização de dados de RNA-seq

Para o numero de sequencias de um dado transcrito poder ser utilizado como medida de transcrição o mesmo deve sofrer algumas correções devido aos seguintes fatores:

-Numero de sequencias produzidas

-Comprimento dos transcritos (transcritos maiores tendem a produzir mais sequencias).

Um dos modos de correção é expressar a abundancia de sequencias na unidade de RPKM (reads per kilobase per million of sequences)

$$RPKM = \frac{\text{number of reads of the region}}{\frac{\text{total reads}}{1000000} \times \frac{\text{region length}}{1000}}$$

RNA-seq X microarray

Vantagens RNA-seq:

-Não é necessário conhecimento prévio do transcrito

-É possível realizar comparações de expressão entre diferentes genes

Vantagens Microarray:

-Custo de repetição de experimentos é menor que o RNA-seq o que permite com que um maior numero de replicas possa ser realizado.

Analise de dados de expressão diferencial

Determinação de genes diferencialmente expressos

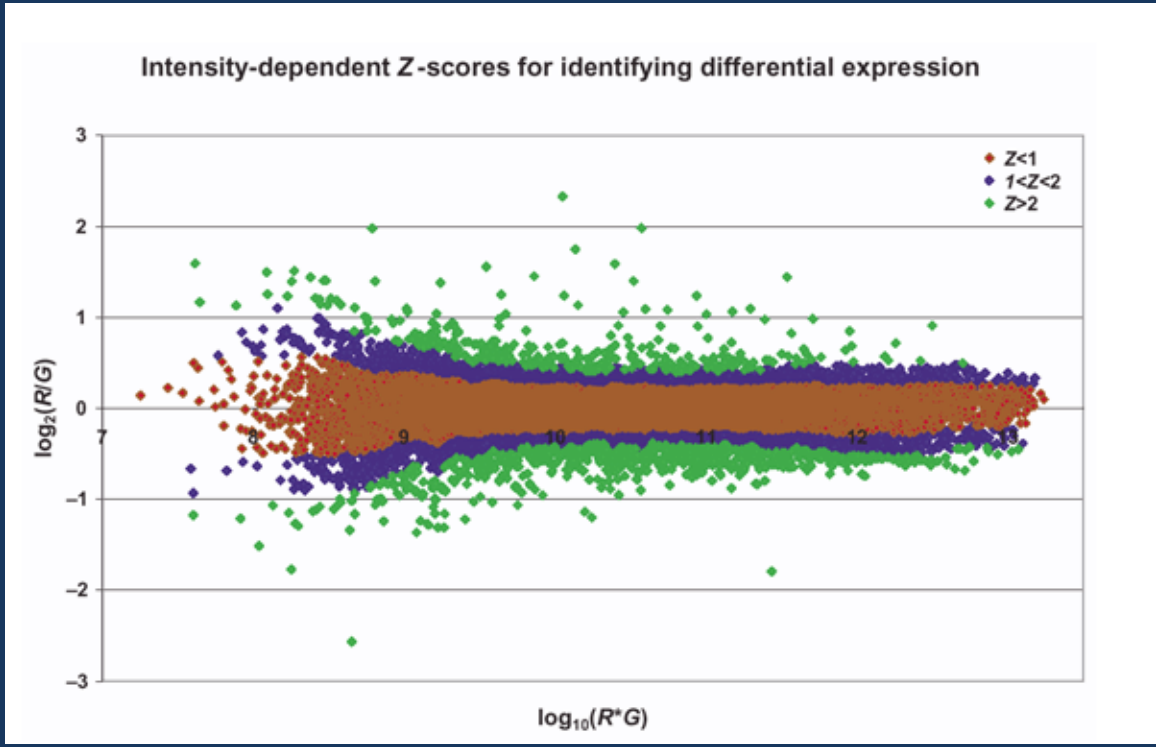
- Dados de microarray apresentam uma variação intrínseca devido a variação de condições experimentais, portanto é necessário a aplicação de métodos que permitam a distinção desta variação "natural"

- A abordagem mais simples é simplesmente estabelecer limites arbitrários a partir do qual a variação é considerada significativa

- Diferentes métodos estatísticos podem ser empregados para tentar diferenciar o ruído de uma variação estatisticamente significativa

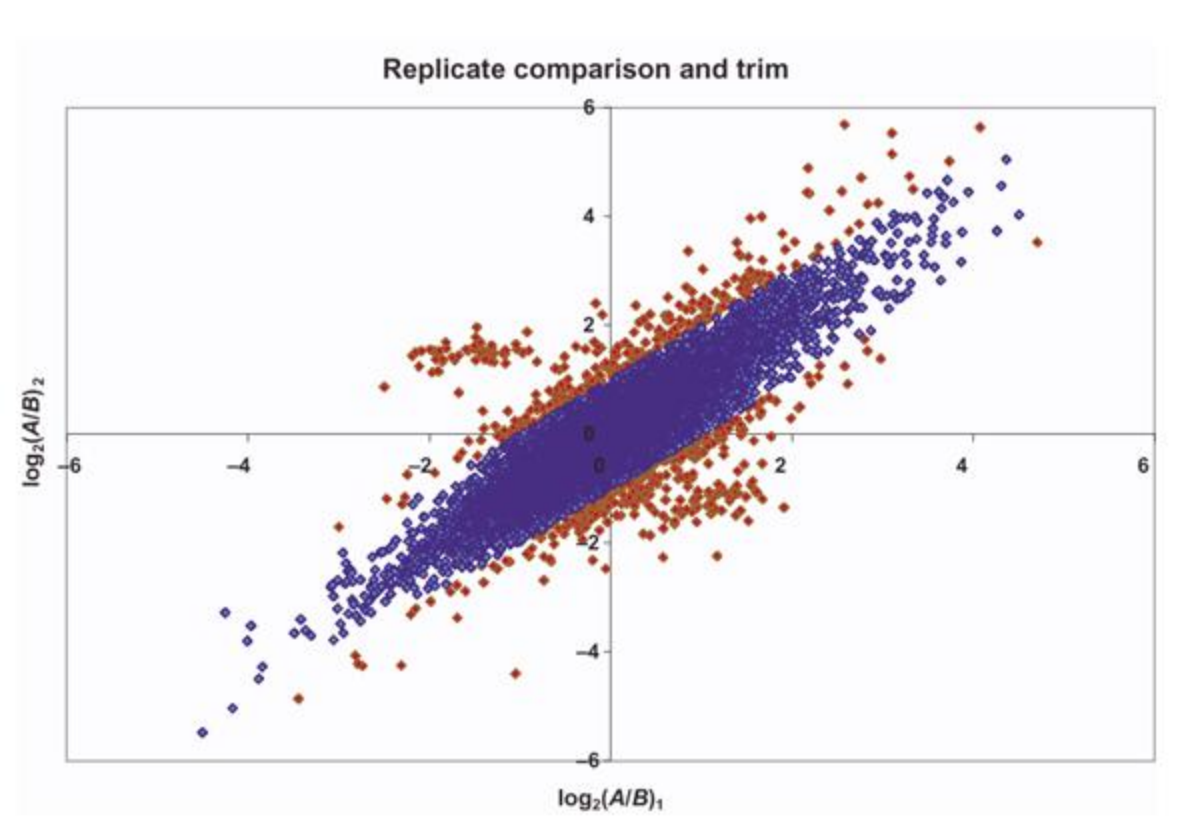
- Realização de replicas do experimento permite uma analise estatística mais robusta

Determinação de genes diferencialmente expressos



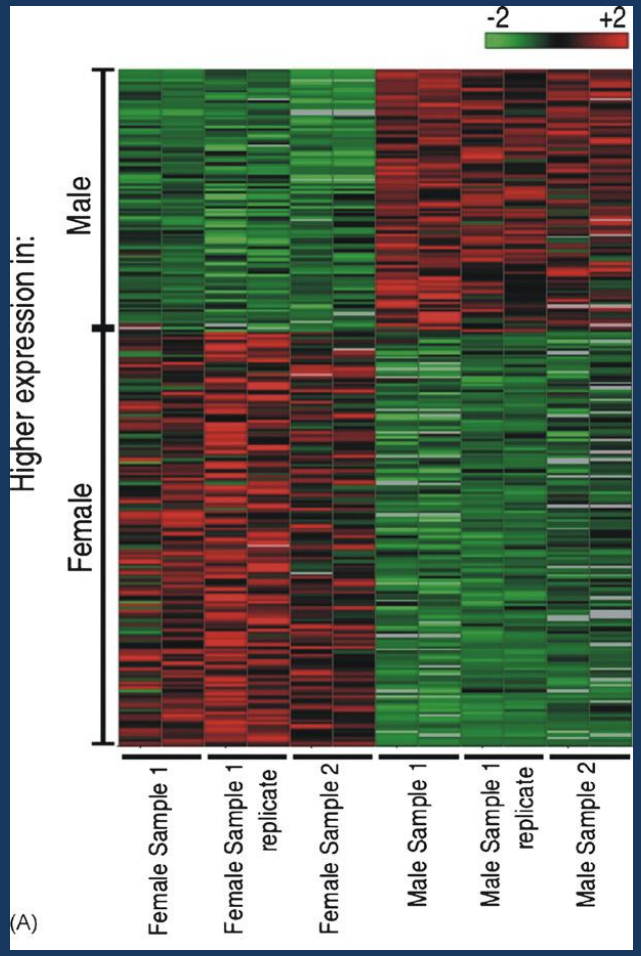
Neste caso o corte é dado pelo z-score (numero de desvios padrões divergentes da media) de cada faixa de intensidade de sinal

Uso de replicas em experimentos



Comparação entre dados de replicas a partir das mesmas amostras
Dados apresentando grande divergência entre as replicas (marrom) são descartados devido a sua baixa consistência

Visualização de dados de expressão diferencial



Heat map- cores representam razão ou z-score dos genes (linhas) em diferentes amostras (colunas)

Analise de perfis de expressão gênica

- Seja através do uso de microarray ou RNA-seq existe um grande interesse em descobrir grupos de genes que apresentam co-variação em diversas amostras e que podem estar representando módulos funcionais

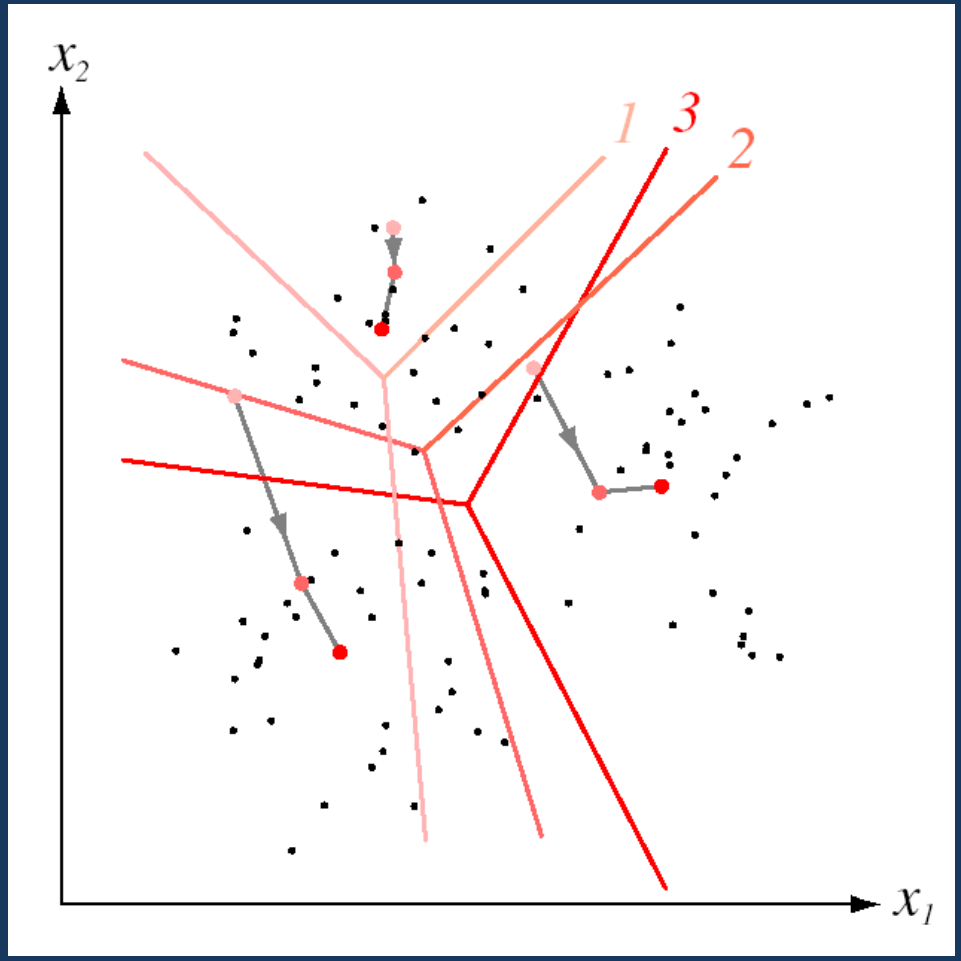
- Estes módulos seriam genes que possuiriam regiões de regulação da transcrição semelhantes e que portanto seriam regulados por um mesmo conjunto de fatores de transcrição

- Deste modo foram desenvolvidas técnicas de analise para tentar realizar um agrupamento destas seqüência e definição destes módulos

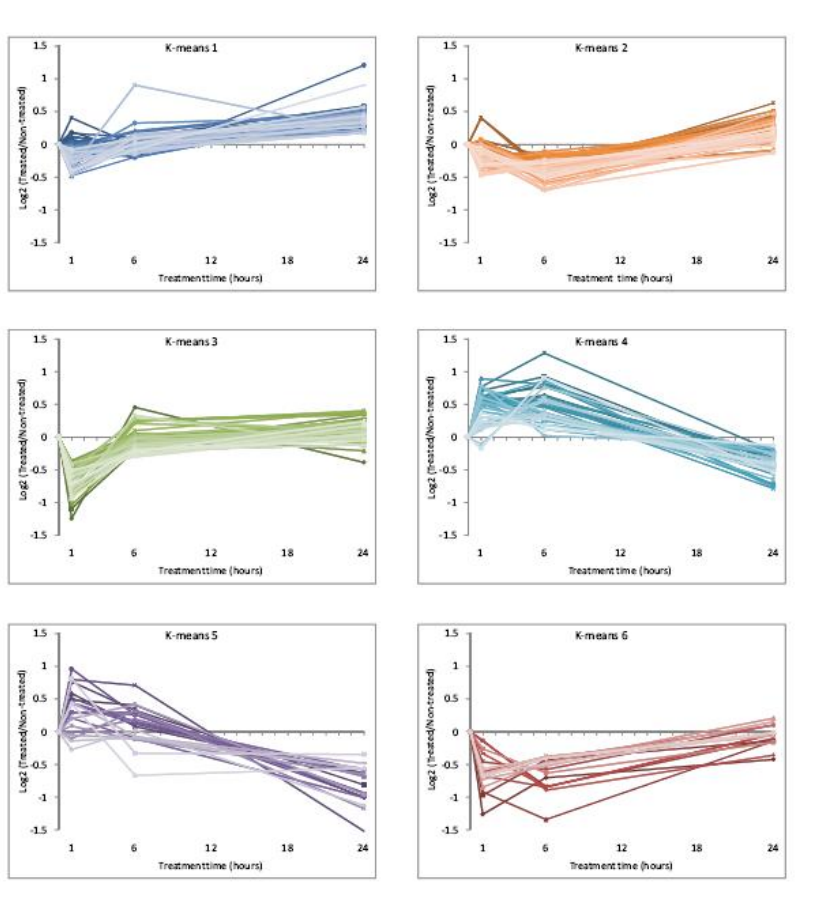
Agrupamento por *k*-medias

- Um numero inicial arbitrário de centro de agrupamentos (*k*) são gerados em posições randômicas
- Cada conjunto de dados é atribuído ao centro mais próximo, formando-se assim um numero de agrupamentos idêntico ao numero de centros.
- Um novo centro de agrupamento é recalculado para ser o ponto médio entre os membros daquele grupo.
- Com base nestas novas posições o calculo a atribuições dos grupos e cálculos são refeitos

Agrupamento *k*-medias

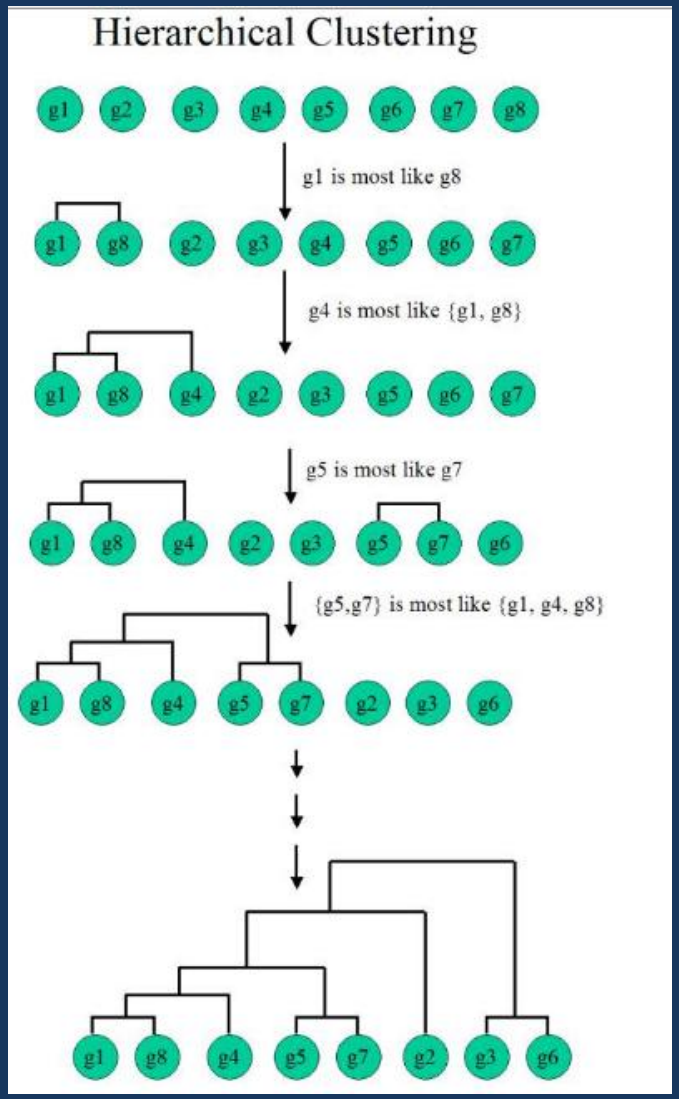


Agrupamento *k*-medias



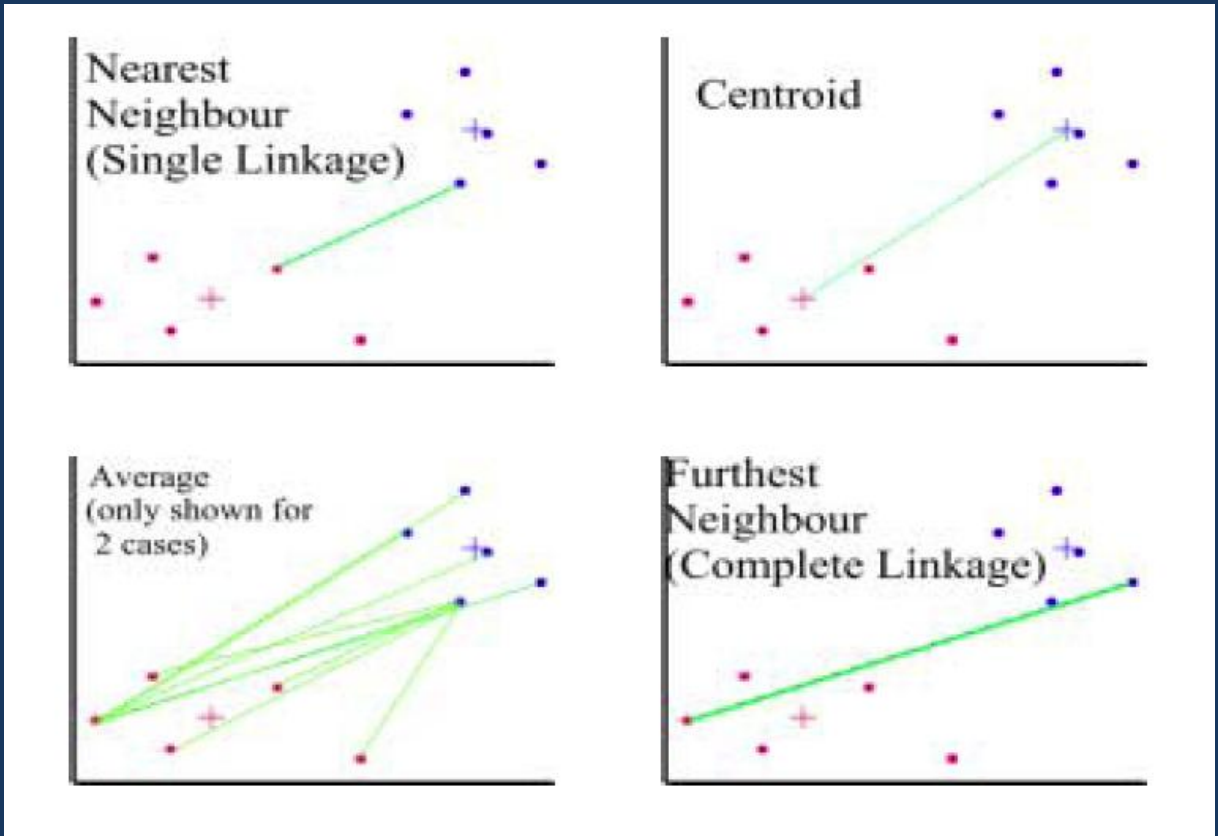
- Exemplo de agrupamentos formados a partir de um experimento temporal de tratamento com uma droga

Agrupamento hierárquico



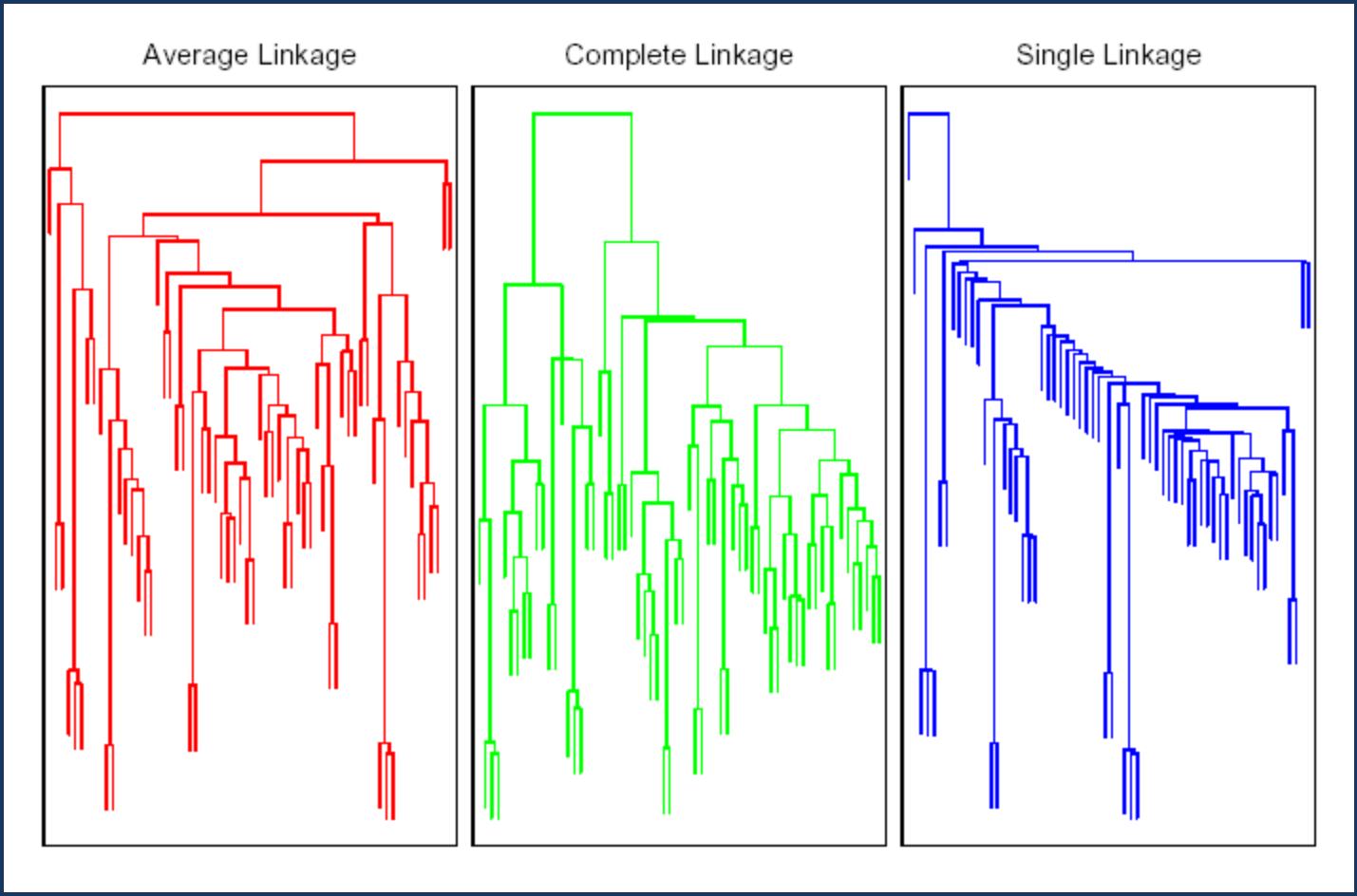
- Agrupamento sucessivo através de medidas de distancias
- Gera um dendograma unindo amostras ou genes

Agrupamento hierárquico

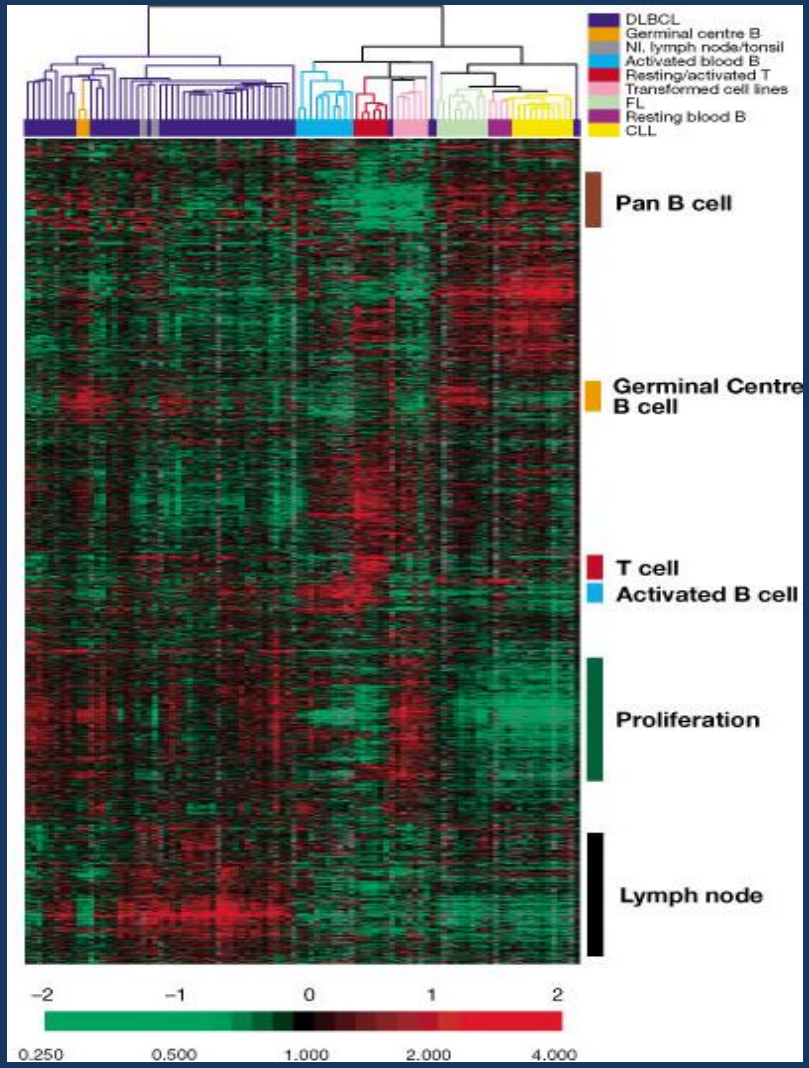


- Distancias entre agrupamentos podem ser calculadas através de diferentes metodos

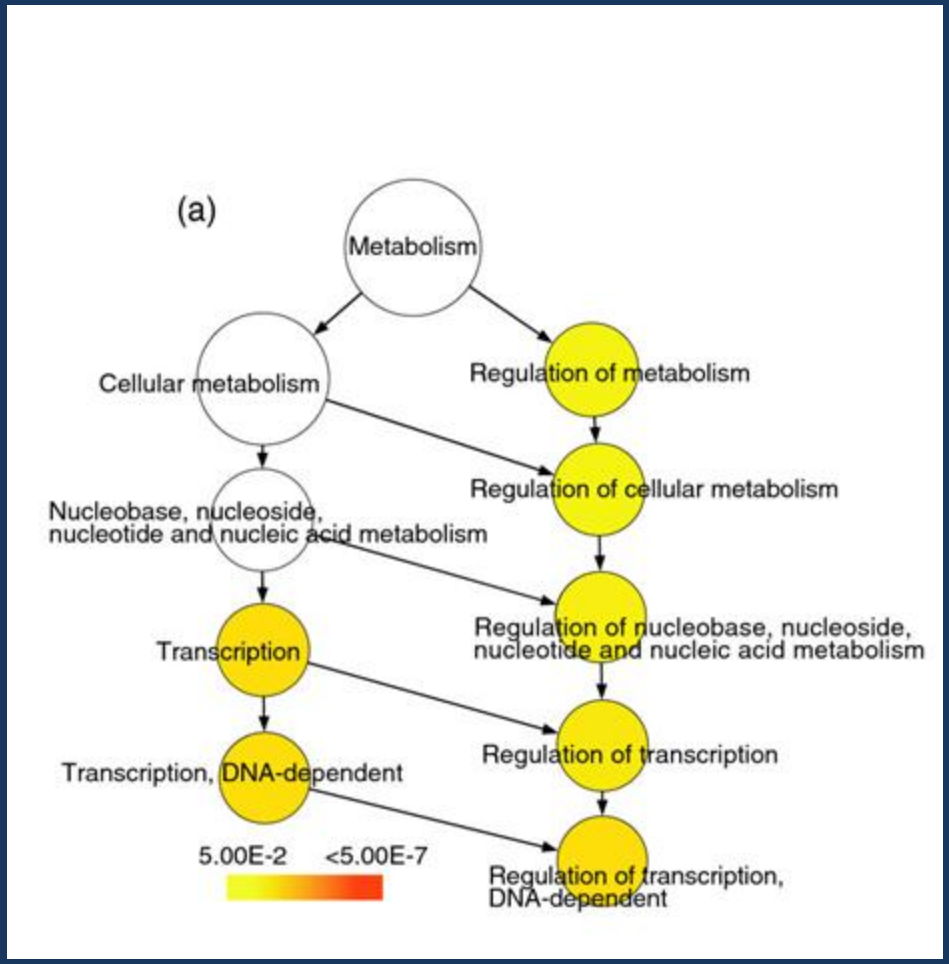
Agrupamento hierárquico



Agrupamento hierárquico



Utilização de GO para identificação de módulos funcionais diferencialmente expressos



- Através da classificação de GO de todas as sondas utilizadas é possível constatar se existe enriquecimento de algum termo