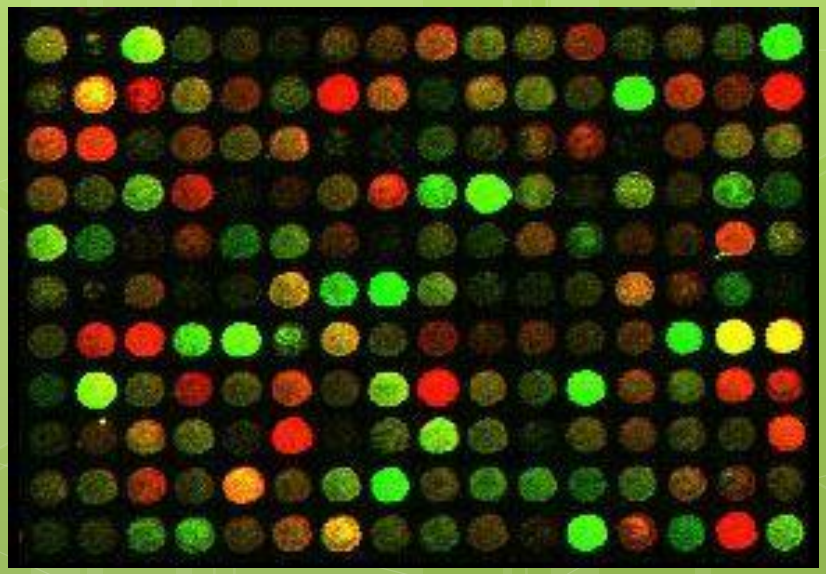
Análise de Expressão Gênica em Larga Escala



Microarrays
(microarranjos)

Diego R. Mazzotti, Ph.D.
Pesquisador
Laboratório de Biologia Molecular do Sono
Departamento de Psicobiologia
Universidade Federal de São Paulo

Supervisor
Molecular Core
Associação Fundo de Incentivo à Pesquisa (AFIP)

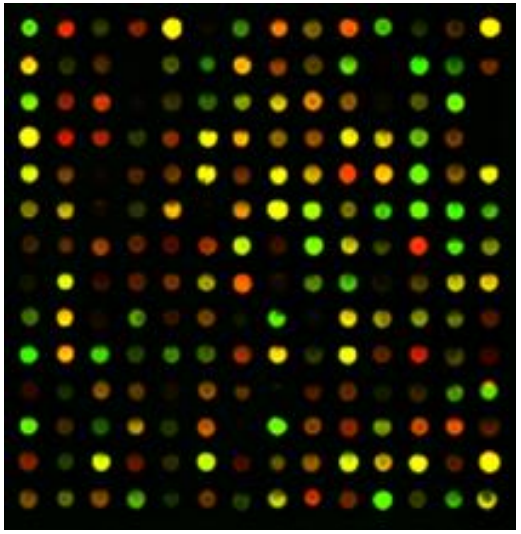
E-mail: mazzottidr@gmail.com

Tópicos de hoje

- Definição de *microarray*
- Principais aplicações
- Vantagens e desvantagens
- Modelos experimentais
- Introdução à análise dos dados de *microarray* de expressão gênica

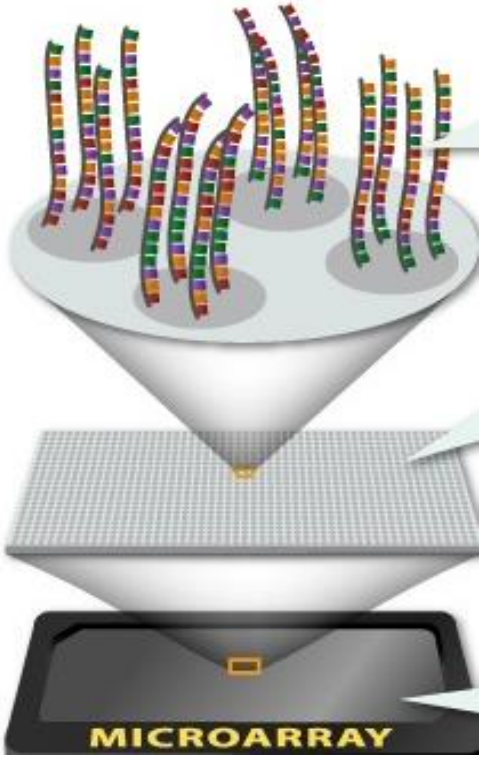
Microarrays (microarranjos)

- Desenvolvido na década de 1990
- Revolução na análise da expressão gênica
  - Monitoramento de RNA de milhares de genes de uma só vez
- Análise global da expressão gênica



Microarrays (microarranjos)

- Conjunto de centenas de milhares de sequências arranjadas em um suporte específico, interrogadas simultaneamente para uma amostra.

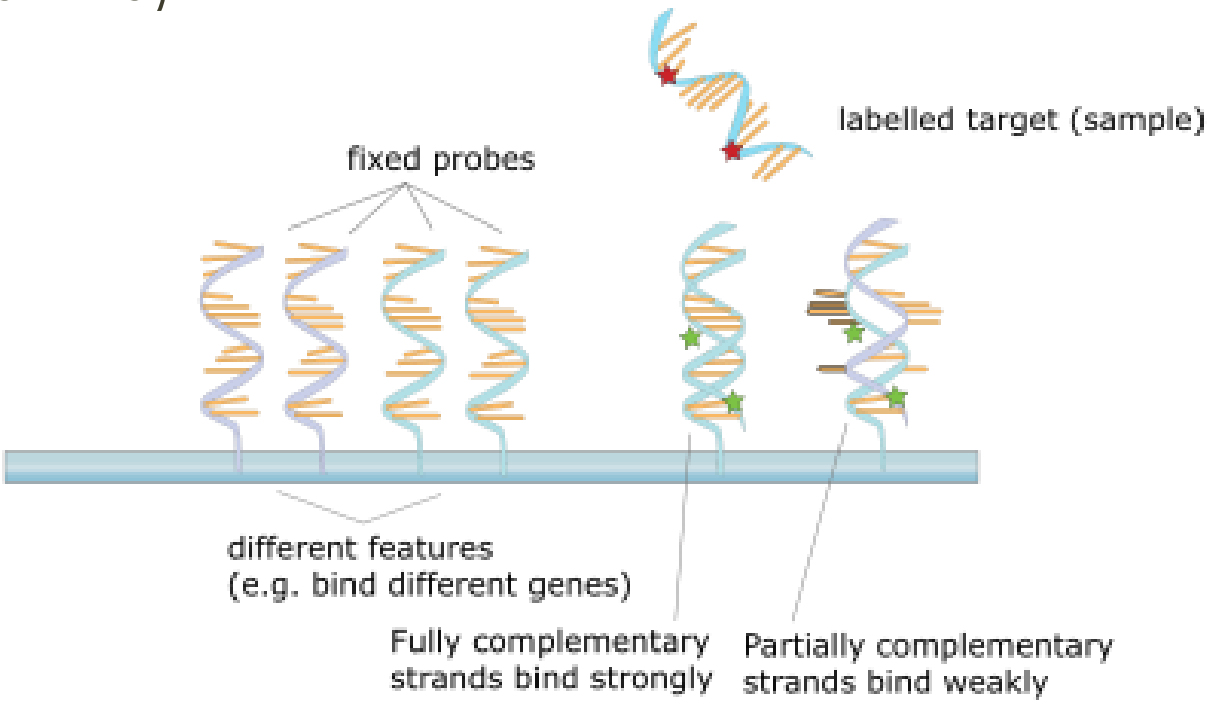


Sondas

Suporte

Microarrays (microarranjos)

- Metodologia baseada na hibridação de sequências contidas na amostra com oligonucleotídeos presentes e fixados em uma determinada superfície (chip ou lâmina)



http://en.wikipedia.org/wiki/File:NA_hybrid.svg

Principais aplicações

- Perfil de expressão gênica (transcriptoma)
- Genotipagem de polimorfismos em larga escala
- Alterações genômicas estruturais
- Detecção de número de cópias (CNVs)

Vantagens da metodologia

- Grande cobertura genômica
  - Análise simultânea de milhares de genes ou milhões de variações genéticas
  - Não restrito a genes candidatos
  - Possibilidade de integração de dados em nível genômico
- Análise de vias biológicas
- Informação qualitativa e quantitativa
- Protocolo fácil e rápido
- Permitiu o desenvolvimento de diversas ferramentas em bioinformática

Desvantagens da metodologia

- Limitado a um experimento de triagem somente
- Limitado às sequências gênicas colocadas no suporte
- Necessidade de validação em muitos casos
- Necessidade de domínio para análise completa dos resultados
- Plataforma para rodar experimentos é cara

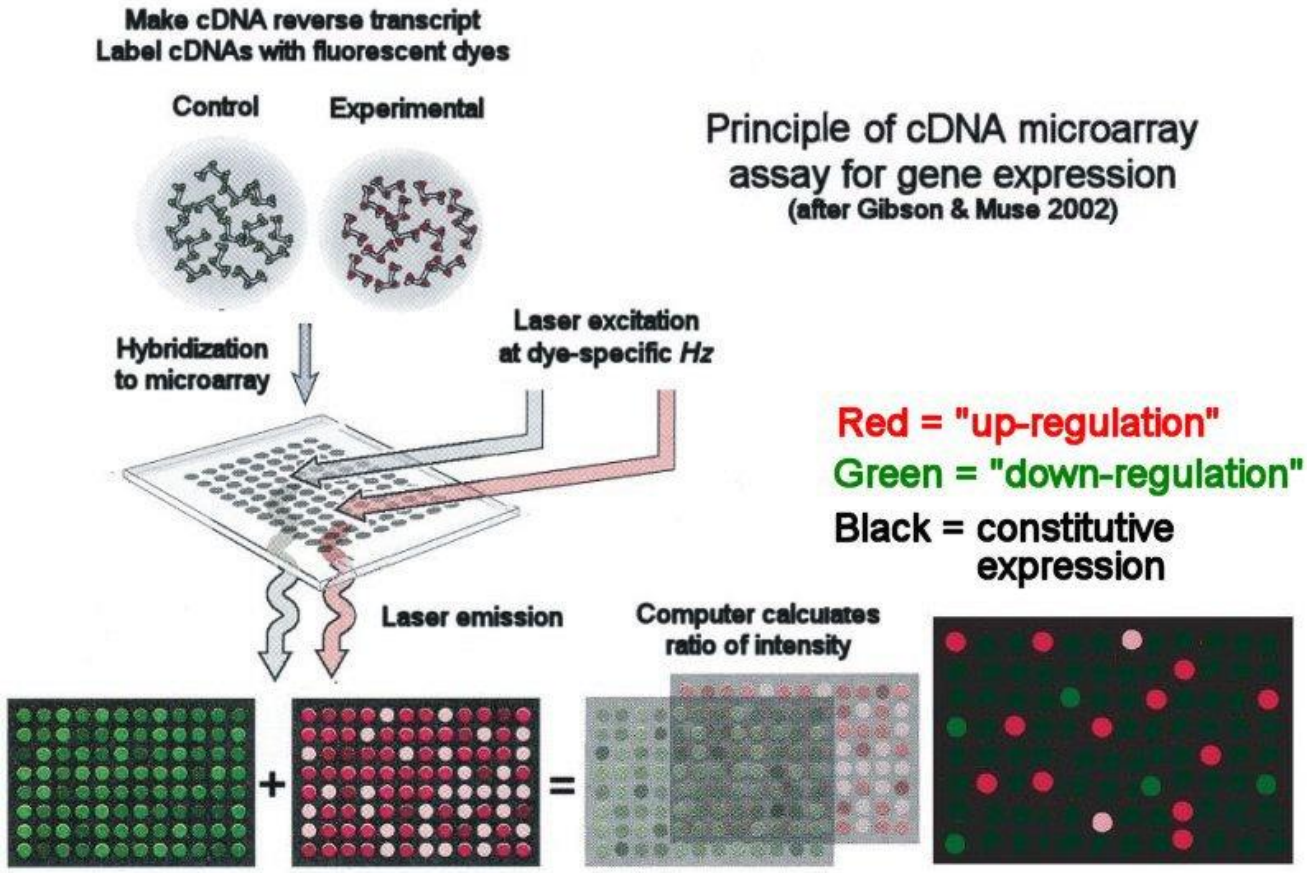
Tipos de experimentos

- *Microarrays* de dois canais (ou duas cores)
  - Perfil de Expressão gênica
  - Variações estruturais (*Array-CGH*)
  - Ex: *Agilent Technologies*

- *Microarrays* de um canal (ou uma cor)
  - Perfil de Expressão gênica
  - Genotipagem
  - Variações estruturais
  - Ex: *Affymetrix*

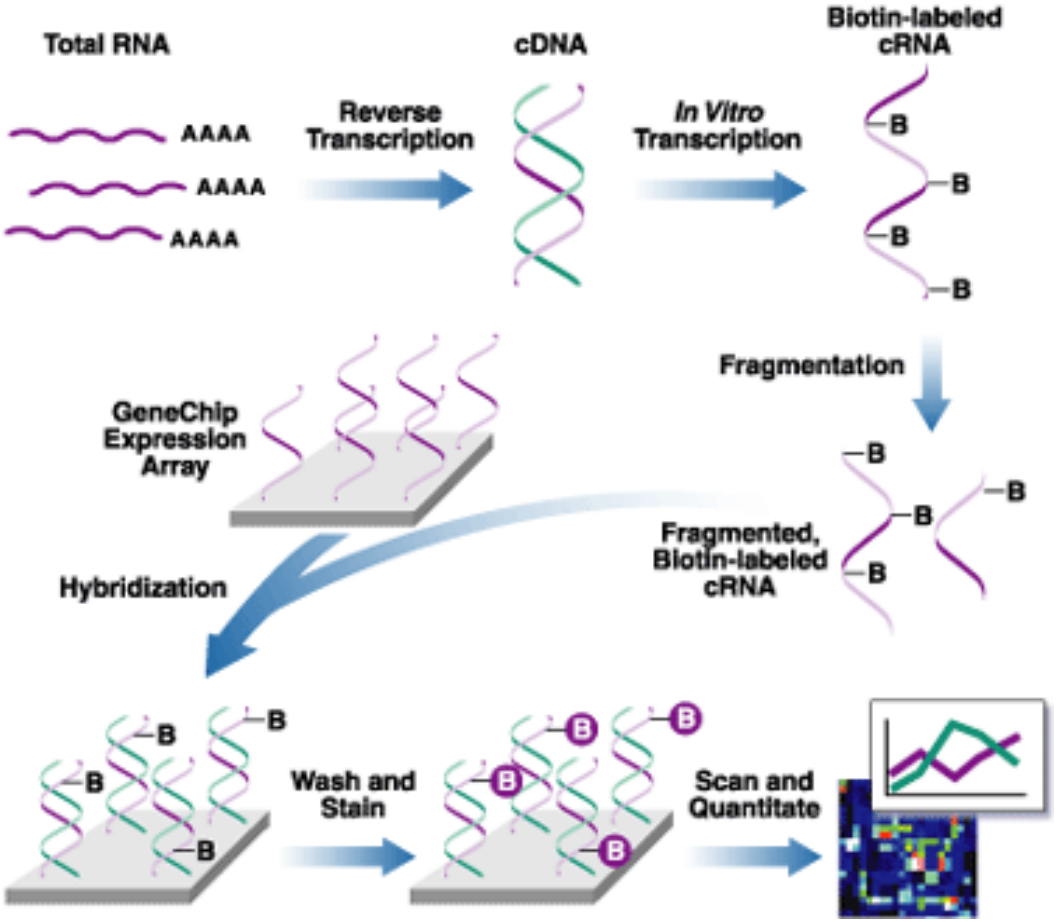
Tipos de experimentos

- *Microarrays* de dois canais (ou duas cores)

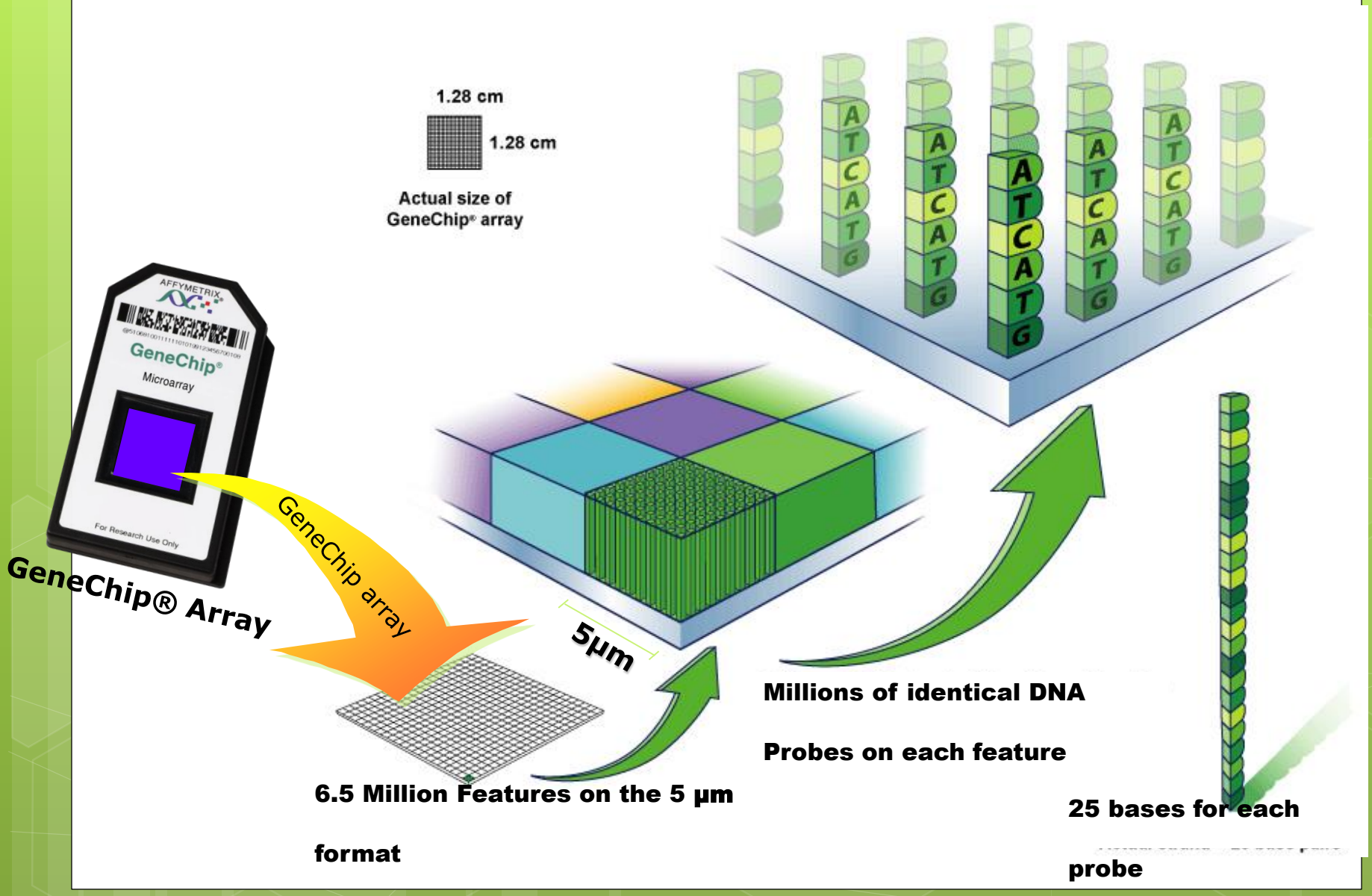


Tipos de experimentos

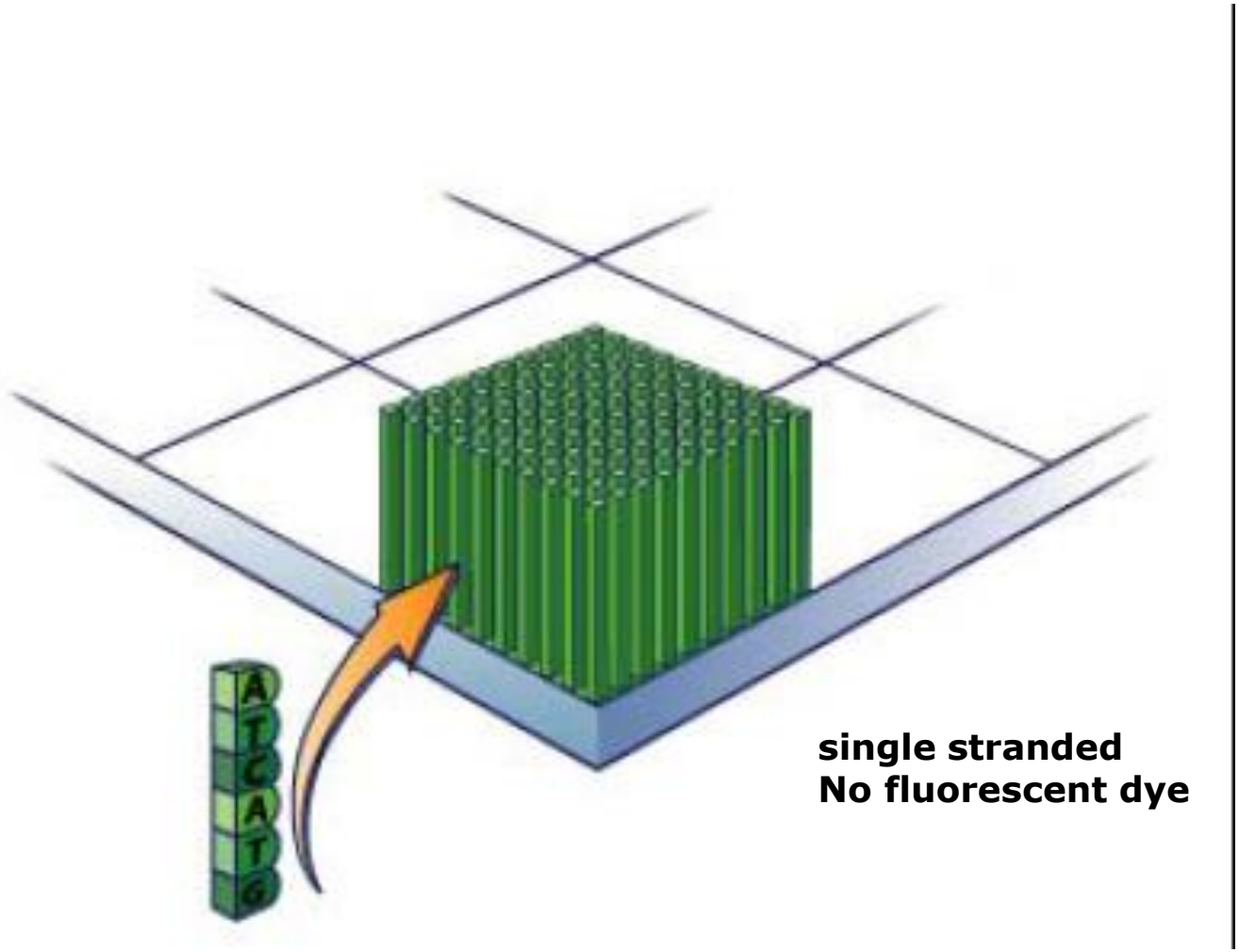
- *Microarrays* de um canal (ou uma cor)



Affymetrix *Microarrays*



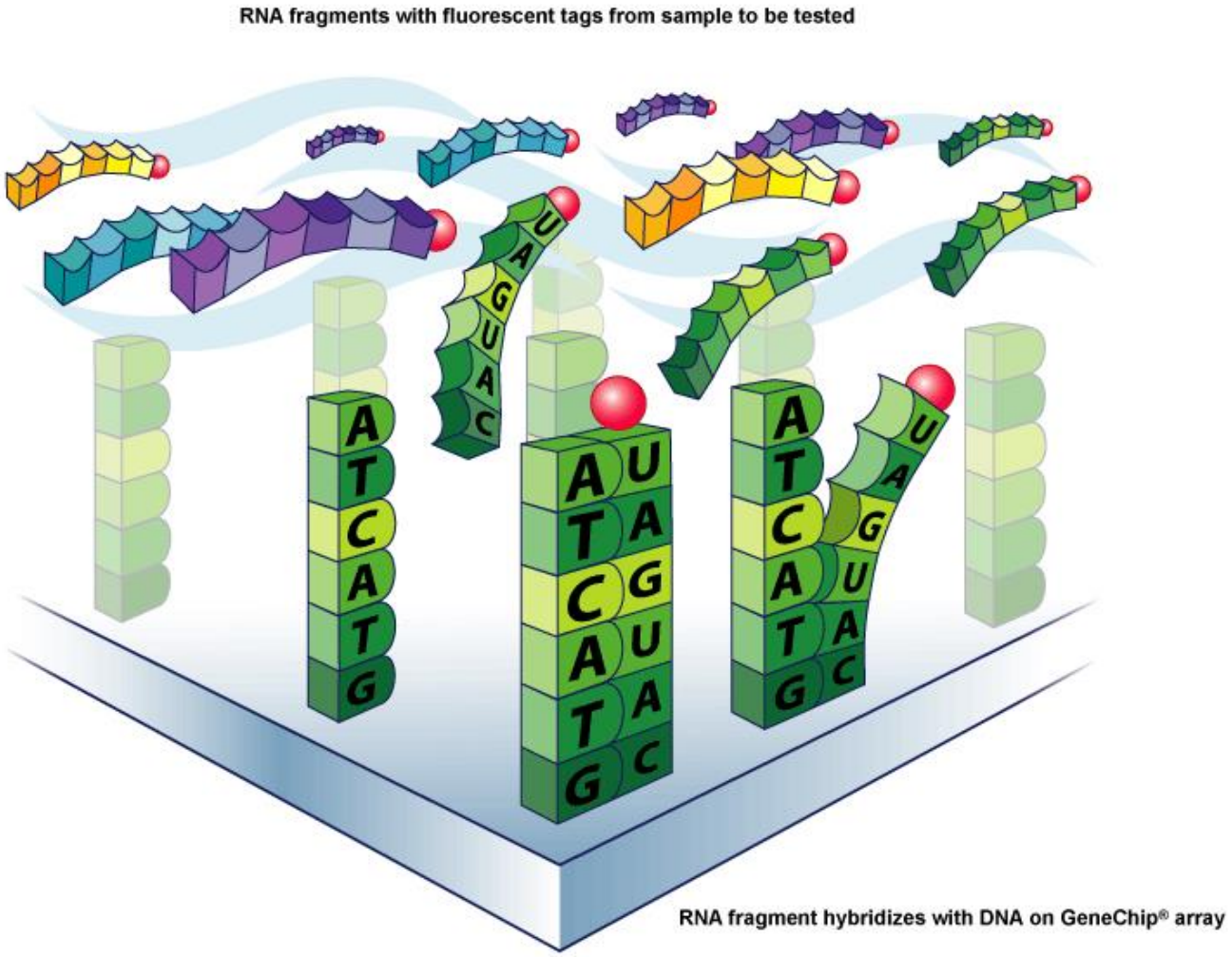
When a target sequence is not in the sample.



No fluorescence on that feature

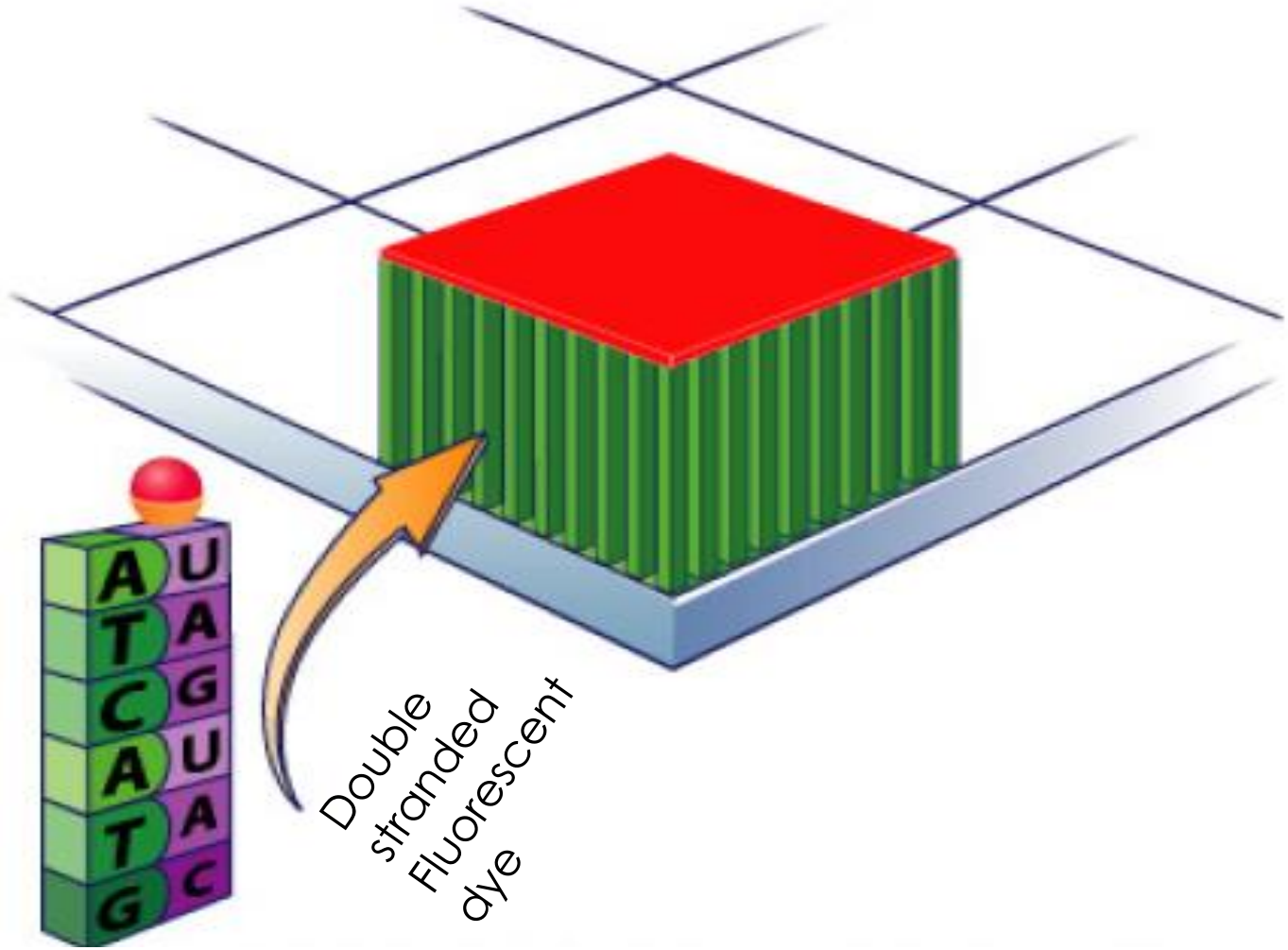
Affymetrix *Microarrays*

GeneChip® Array

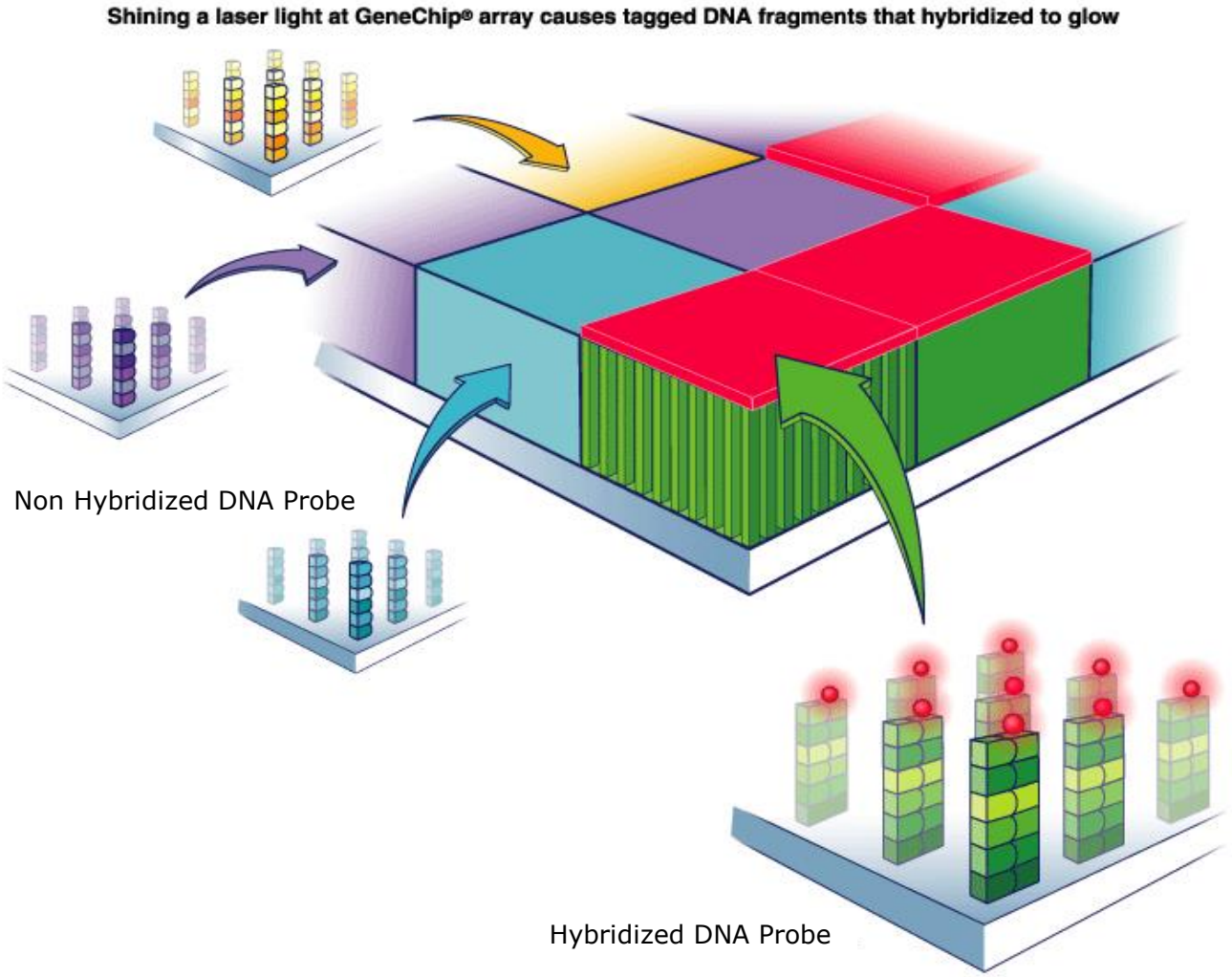


RNA/DNA labeled fragments hybridize with the DNA Probe on the GeneChip® array

Affymetrix *Microarrays*



Affymetrix *Microarrays*

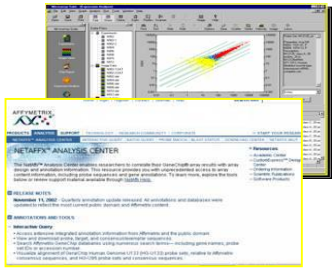


Affymetrix *Microarrays*

Reagents

Microarray
GeneChip®

Hybridization Oven 640



Software
Data Analysis

Scanner

Fluidics
Station

Modelos Experimentais

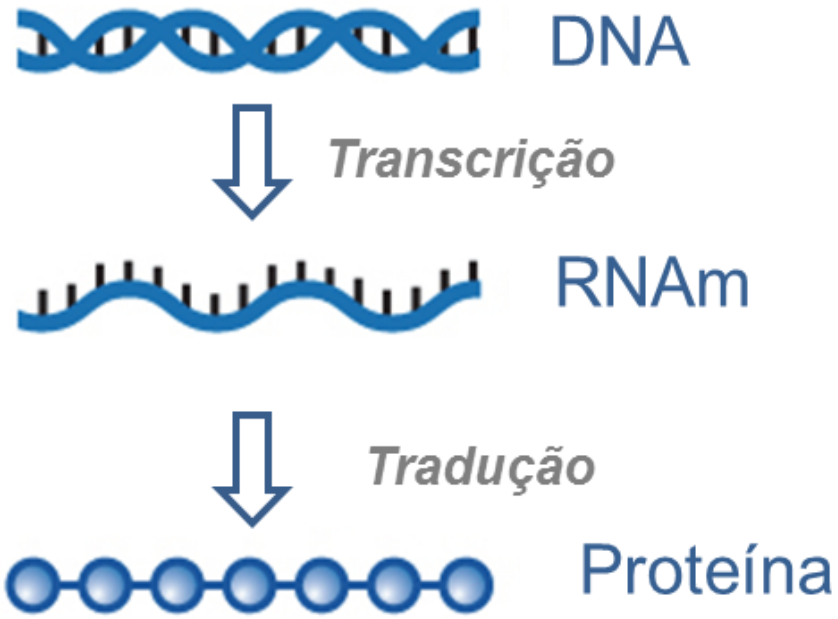
- Perfil de expressão gênica
- Genotipagem em larga escala
- Variações estruturais e número de cópias

Modelos Experimentais

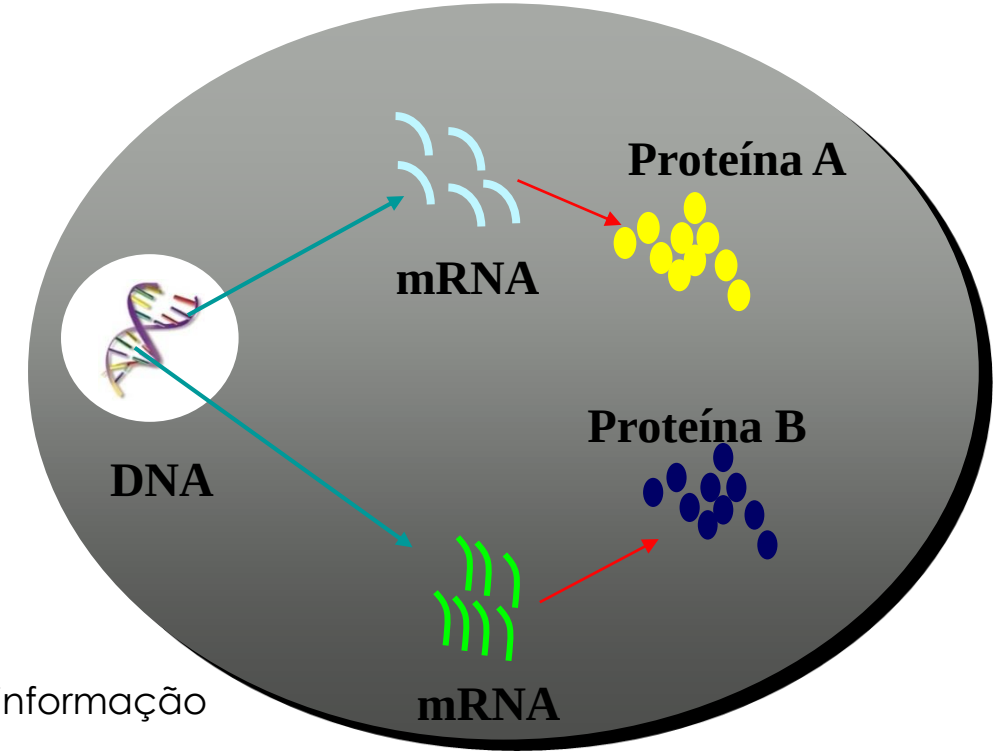
- **Perfil de expressão gênica**
- Genotipagem em larga escala
- Variações estruturais e número de cópias

Perfil de expressão gênica

- O "Dogma Central da Biologia Molecular"



Perfil de expressão gênica



Proteínas

Produtos gênico, componentes estruturais, enzimas, receptores, hormônios...

DNA

Contém a informação

Transcrição

RNA

Moléculas transientes que representam atividade gênica

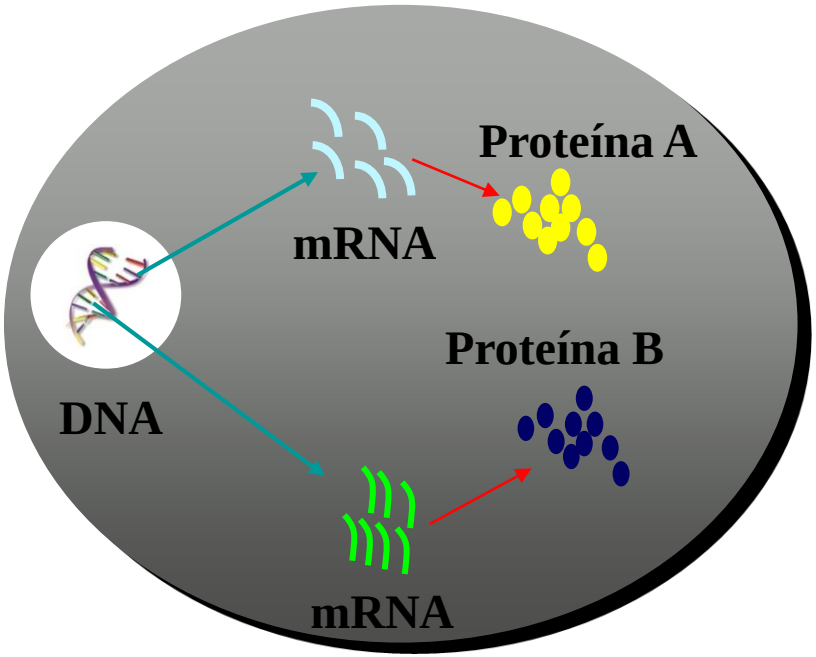
Tradução

Perfil de expressão gênica

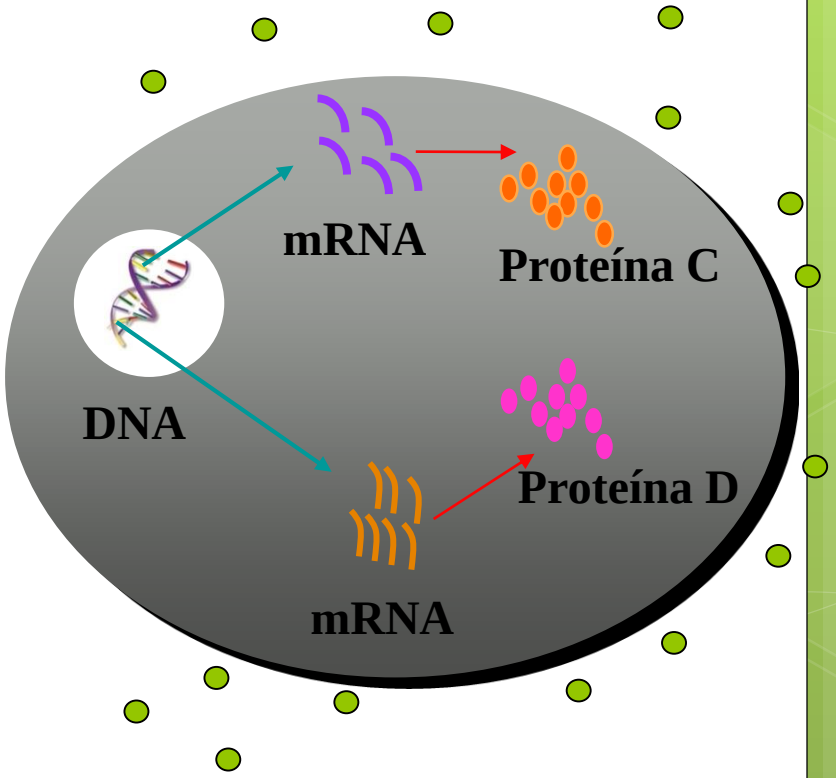
- O que acontece se:
  - Manipularmos o ambiente em que a célula/ tecido se encontra?
    - Estudos experimentais
  - Avaliarmos o perfil de expressão de um tecido em determinada situação (ex: doença)?
    - Estudos observacionais – identificação de biomarcadores

Perfil de expressão gênica

Célula em condição normal

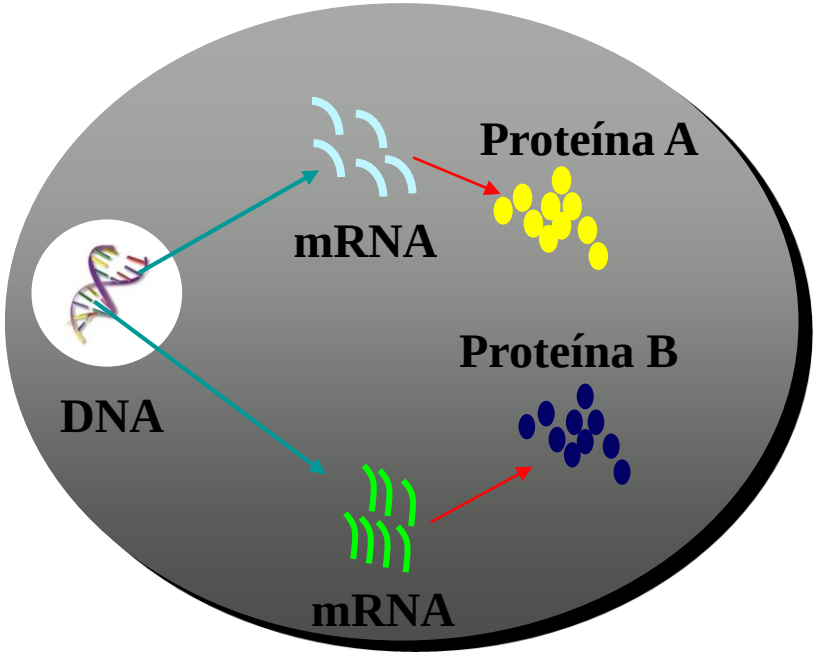


Célula na presença de droga

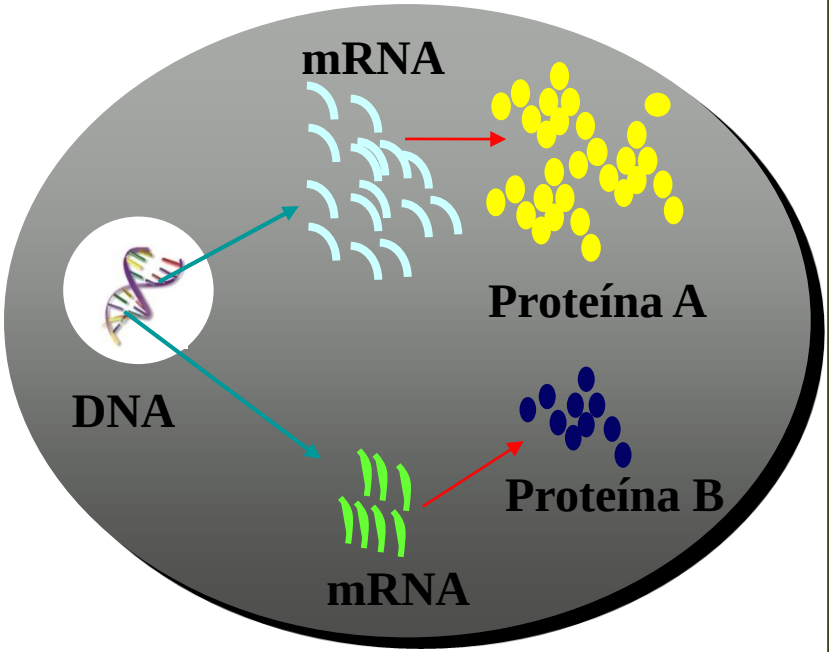


Perfil de expressão gênica

Célula Controle

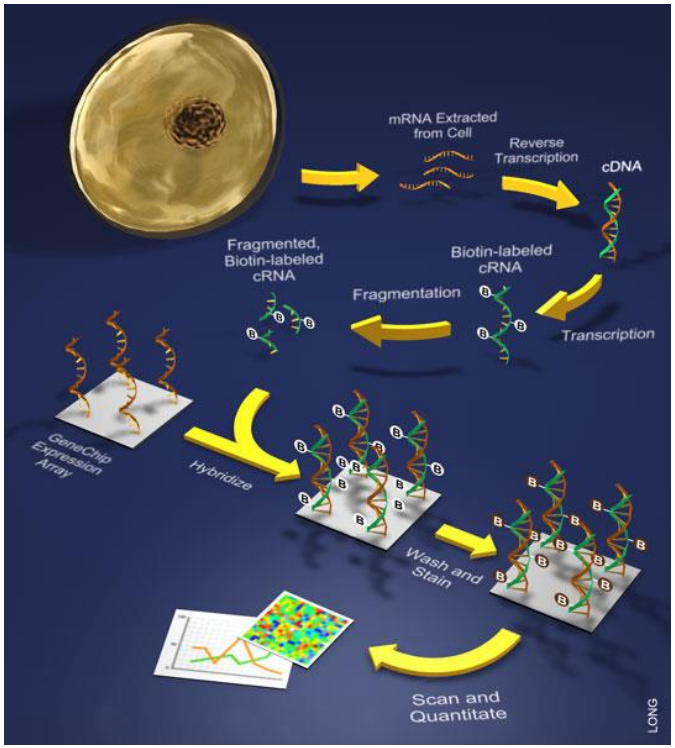


Célula Tumoral



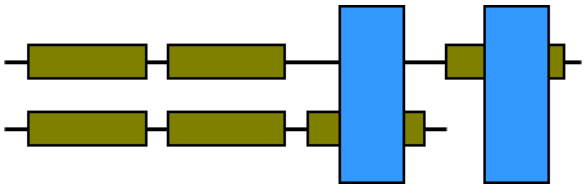
Perfil de expressão gênica

- *Microarrays* de expressão → avaliar o perfil transcricional da célula/ tecido em determinada condição

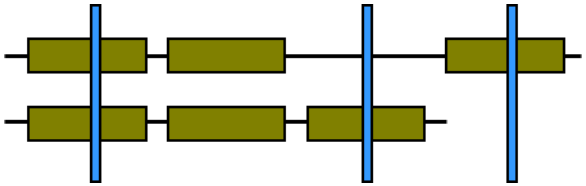


Como é mensurada a expressão gênica pelo *microarray*?

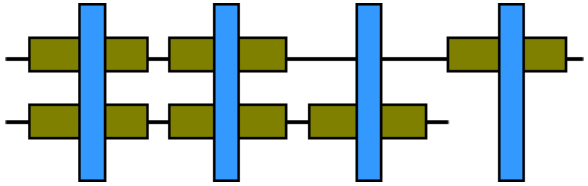
- Desenho das sondas → diferentes estratégias de análise



Sondas 3' (Ex: HG-U133)

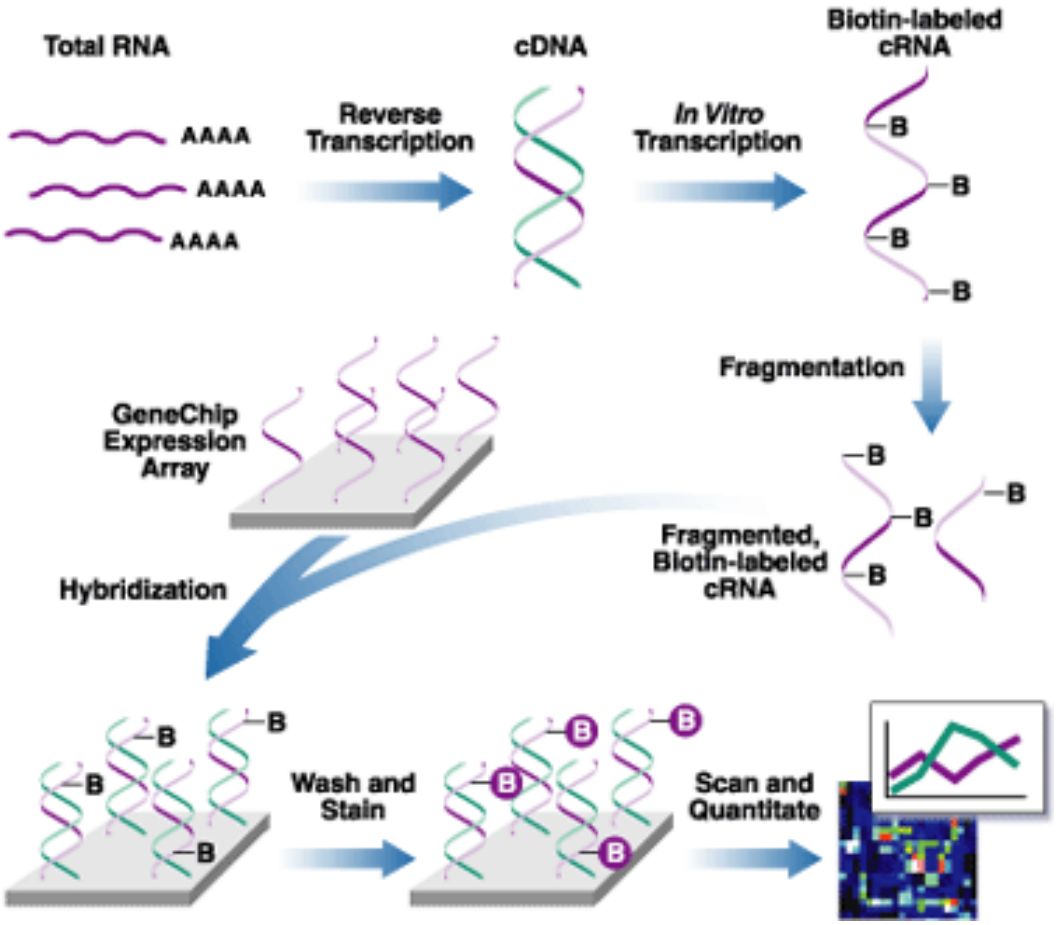


Sondas em Genes (Ex: Human Gene 1.0 ST)



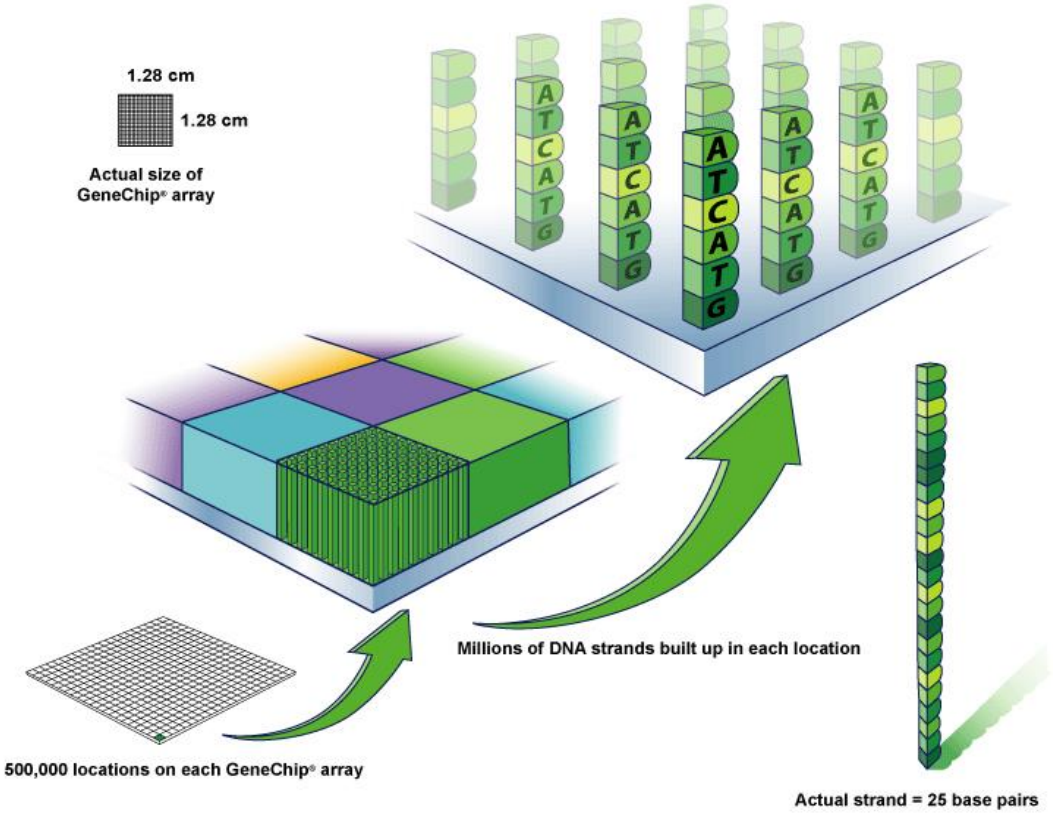
Sondas em Éxons (Ex: Human Exon 1.0 ST)

Como é mensurada a expressão gênica pelo *microarray*?

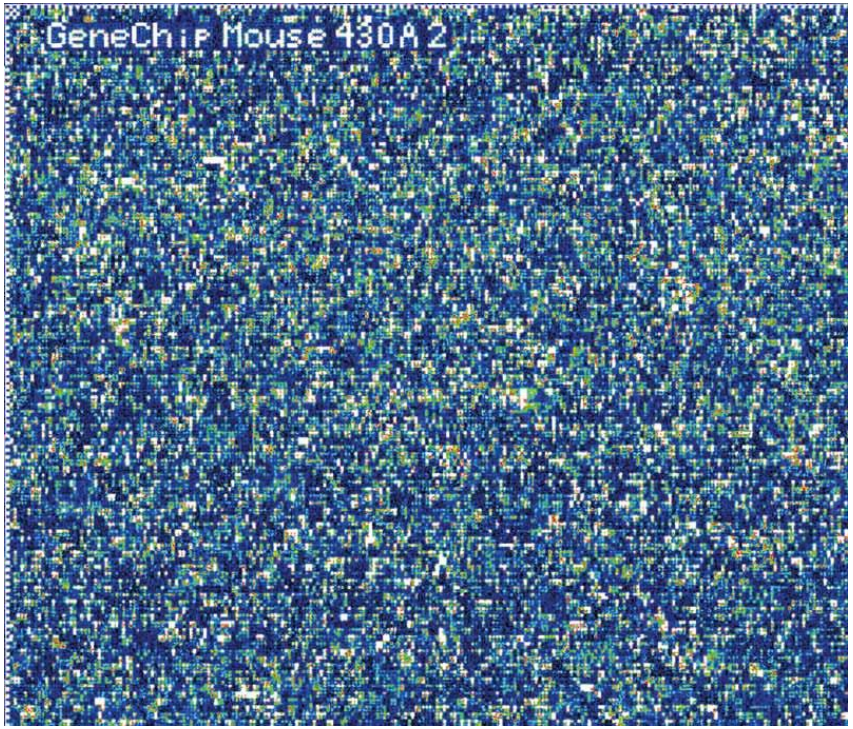


Como é mensurada a expressão gênica pelo *microarray*?

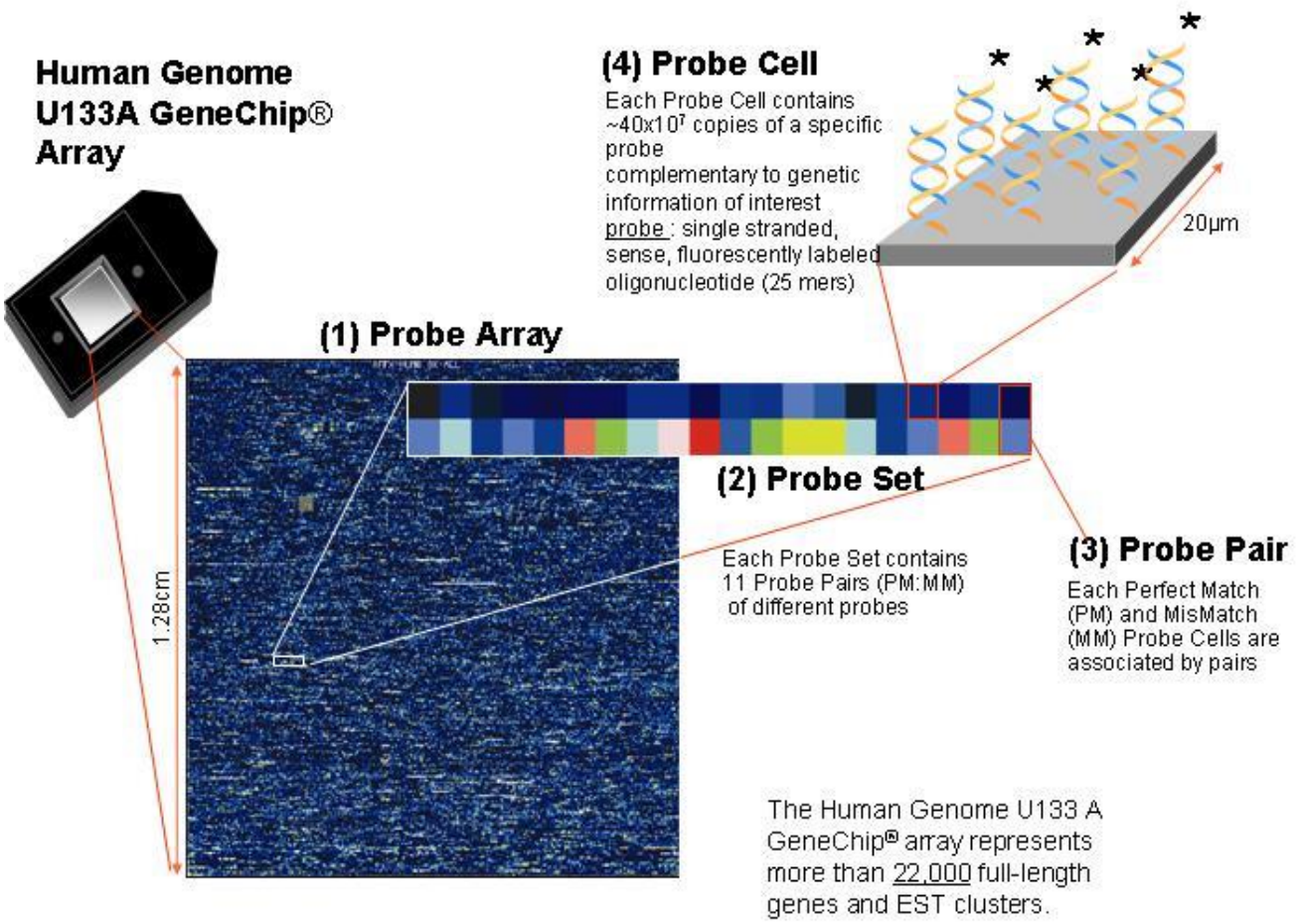
- Organização das sondas



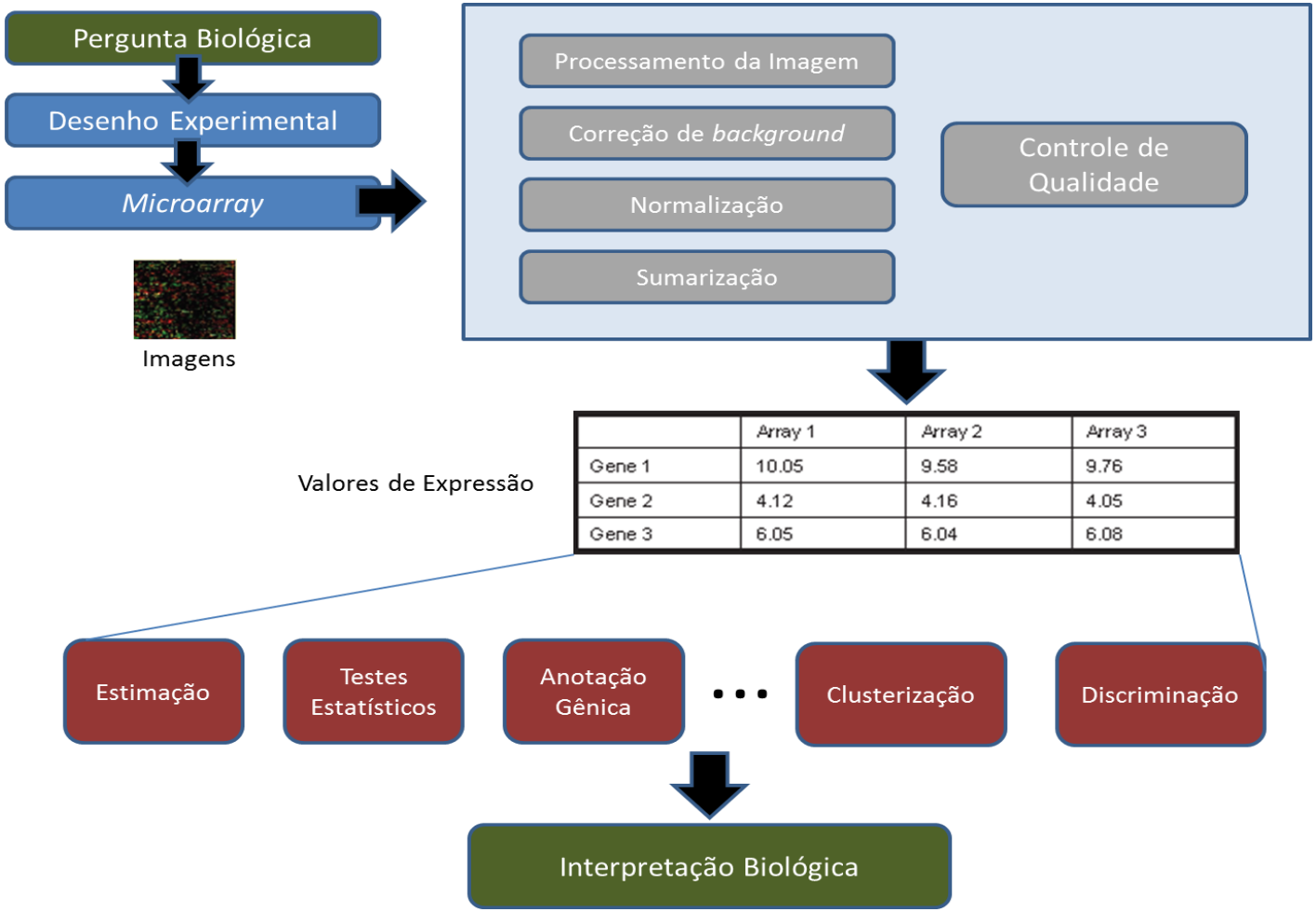
Como é mensurada a expressão gênica pelo *microarray*?



Como é mensurada a expressão gênica pelo *microarray*?



Análise dos dados de expressão gênica



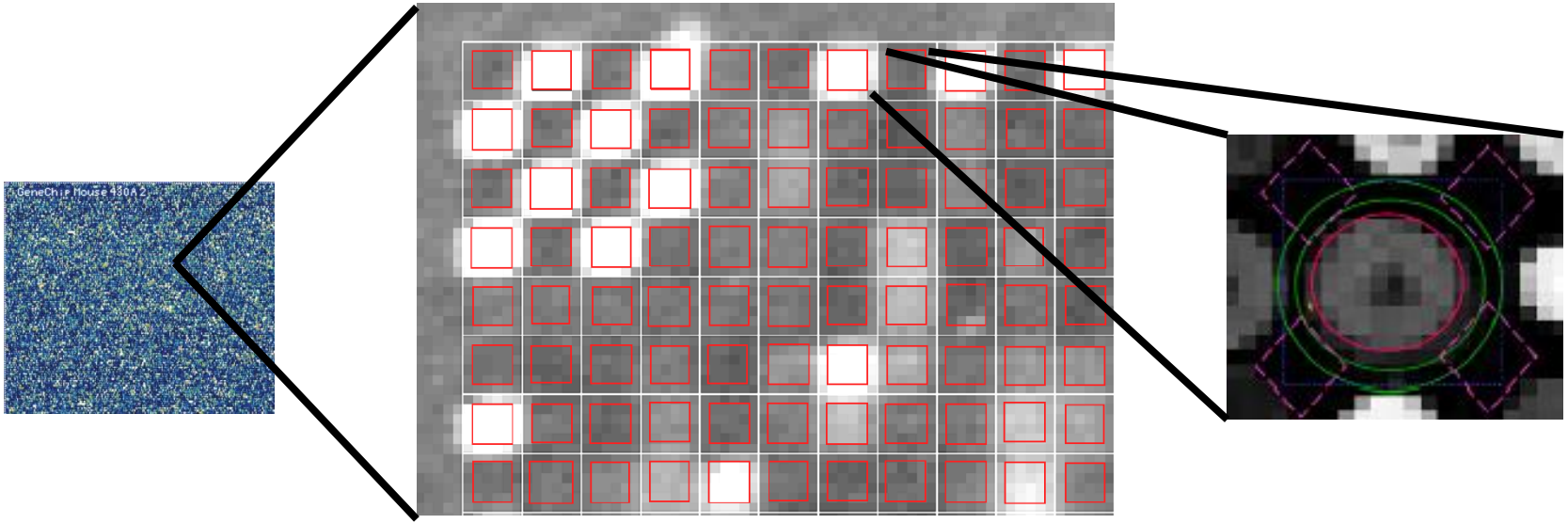
Adaptado de Bolstad, B. Methodologies for Pre-processing Microarray Data, 2006

Análise dos dados de expressão gênica

- **Principais etapas da análise:**
  - Processamento da imagem
  - Pré-processamento dos dados
    - Correção de *background* (ruído)
    - Normalização
    - Sumarização
  - Checagem do pré-processamento (controle de qualidade)
  - Análise exploratória dos dados
  - Identificação dos genes diferencialmente expressos
  - Análises funcionais (vias, ontologia gênica, etc.)

Análise dos dados de expressão gênica

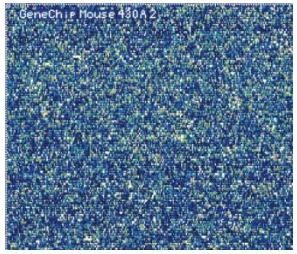
- Processamento da imagem



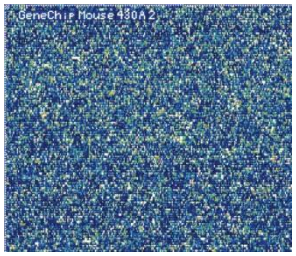
- Etapa automatizada (Affymetrix), sem interferência do analista

Análise dos dados de expressão gênica

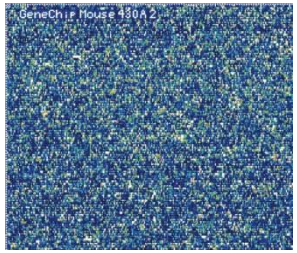
- Sinal é mensurado para cada "*spot*" (que representa uma sonda no *microarray*)



Amostra 1



Amostra 2



Amostra 3

	Amostra1	Amostra2	Amostra3	...
sonda1	0,957041	1960,81	0,149155	0,552391
sonda2	0,980226	959,1938	0,069077	0,673049
sonda3	0,959119	1649,556	0,076062	0,634061
...	0,933941	2570,302	0,155557	1,297719

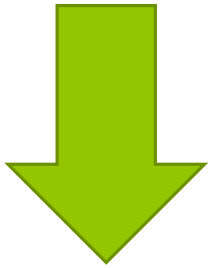
"Dados Brutos"

Análise dos dados de expressão gênica

- Um valor mensurado é atribuído para cada sonda
- No entanto, cada gene representado no *microarray* é composto por um conjunto de sondas (***probeset***)
- Dados brutos não leva em considerações "ruídos"

Análise dos dados de expressão gênica

- Um valor mensurado é atribuído para cada sonda
- No entanto, cada gene representado no *microarray* é composto por um conjunto de sondas (***probeset***)
- Dados brutos não leva em considerações "ruídos"



Pré-processamento

Análise dos dados de expressão gênica – Pré-processamento

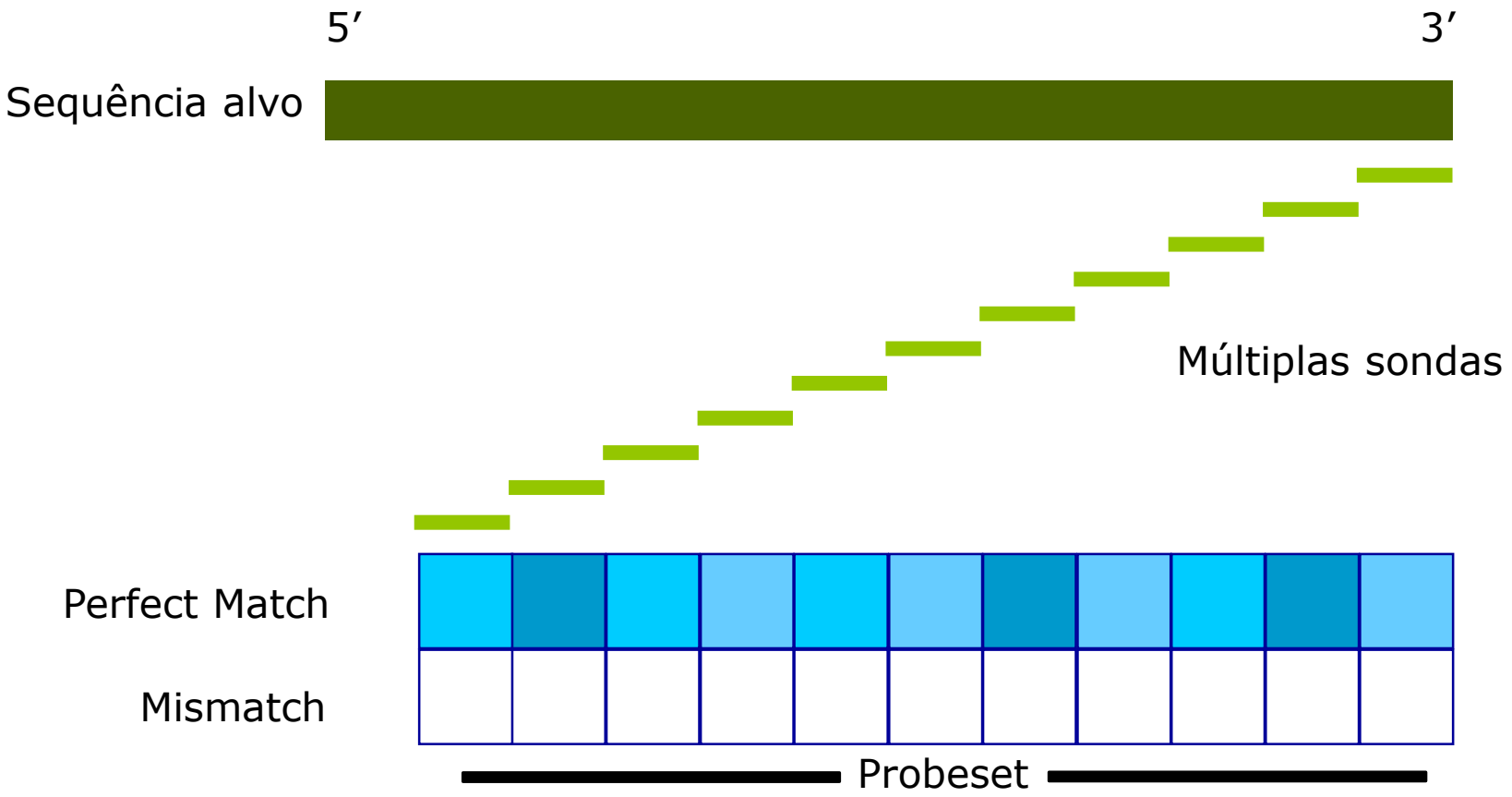
- Etapas do pré-processamento:
  - Correção de *background*
  - Normalização
  - Sumarização

Análise dos dados de expressão gênica – Pré-processamento

- Etapas do pré-processamento:
  - **Correção de *background***
  - Normalização
  - Sumarização

Análise dos dados de expressão gênica – Pré-processamento

- Organização das sondas



Análise dos dados de expressão gênica – Pré-processamento

- Organização das sondas – implicações na análise

Sinal mensurado = expressão do gene + ruído

Análise dos dados de expressão gênica – Pré-processamento

- Organização das sondas – implicações na análise

Sinal mensurado = expressão do gene + ruído

Perfect Match

Fluorescência específica

Mismatch

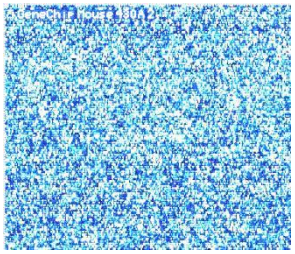
Fluorescência inespecífica

- **Correção de background** → descontar o "ruído" de todos os sinais mensurados

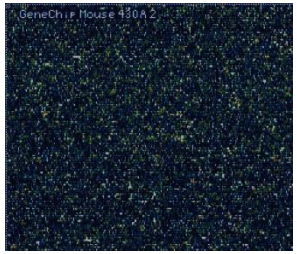
Análise dos dados de expressão gênica – Pré-processamento

- Etapas do pré-processamento:
  - Correção de *background*
  - **Normalização**
  - Sumarização

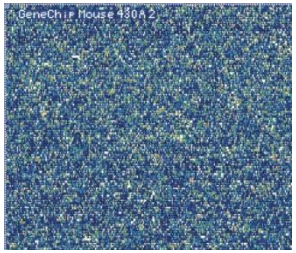
Análise dos dados de expressão gênica – Pré-processamento



Array 1



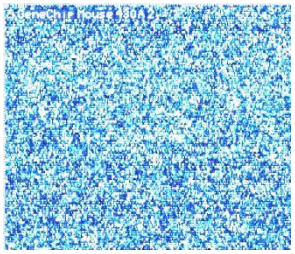
Array 2



Array 3

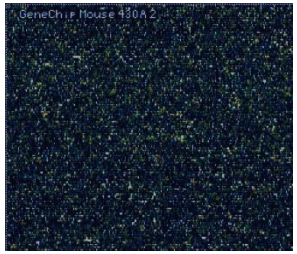
Análise dos dados de expressão gênica – Pré-processamento

- Nem todos os *arrays* são comparáveis



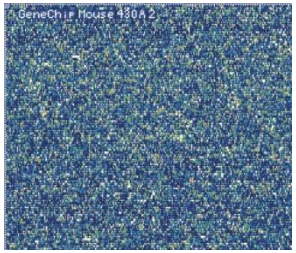
Array 1

Será que há mais RNAm colocado no Array 1 e menos no Array 2?



Array 2

Será que a expressão global de todos os RNAm é maior na amostra colocada no Array 1 do que no Array 2?



Array 3

Será que o ar condicionado quebrou no dia que foi feita a leitura do Array 2?

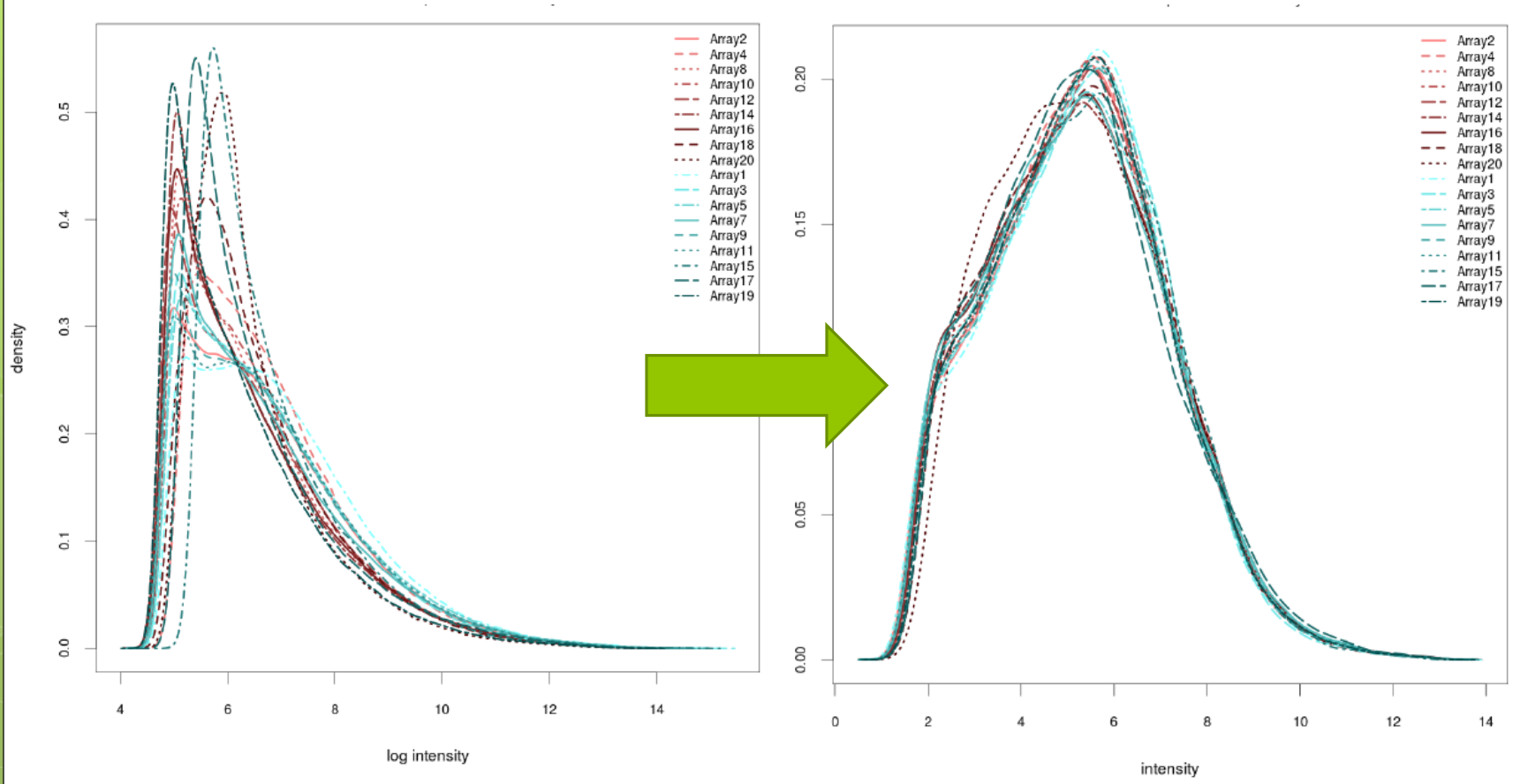
Análise dos dados de expressão gênica – Pré-processamento

- Necessidade de torná-los comparáveis
  - Objetivo
    - Identificar diferenças de expressão dos genes entre as condições (efeito biológico)
    - Não confundir efeito biológico com efeito técnico
  - Solução → **normalização** dos dados

Análise dos dados de expressão gênica – Pré-processamento

- Estratégias de normalização:
  - *Dye Bias* (arrays de duas cores)
  - *Median Scaling*
  - *Loess (LOcally WEighted Scatterplot Smoothing)*
  - *Quantile Normalization* (mais usado)
  - ...

Análise dos dados de expressão gênica – Pré-processamento

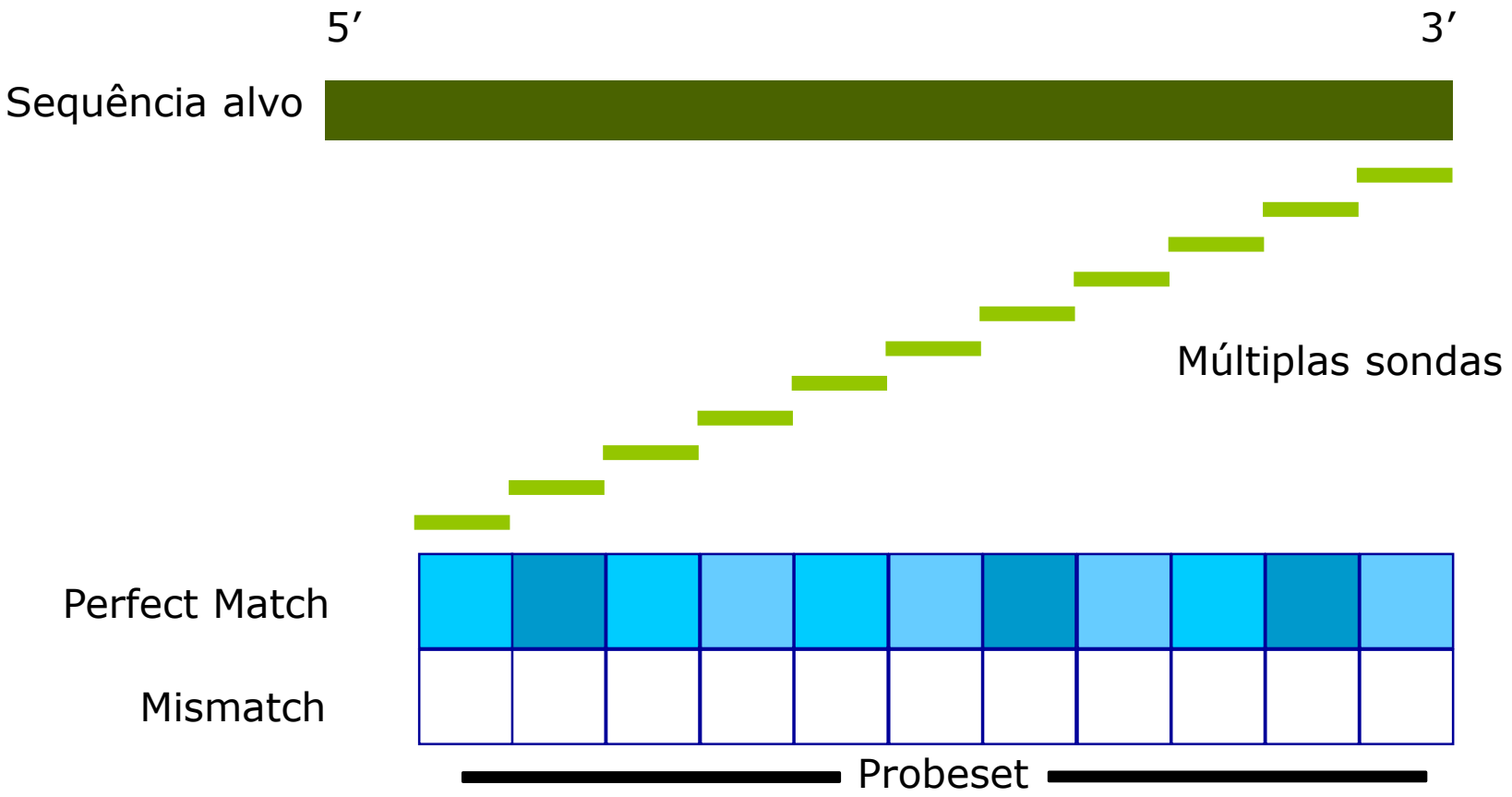


Análise dos dados de expressão gênica – Pré-processamento

- Etapas do pré-processamento:
  - Correção de *background*
  - Normalização
  - **Sumarização**

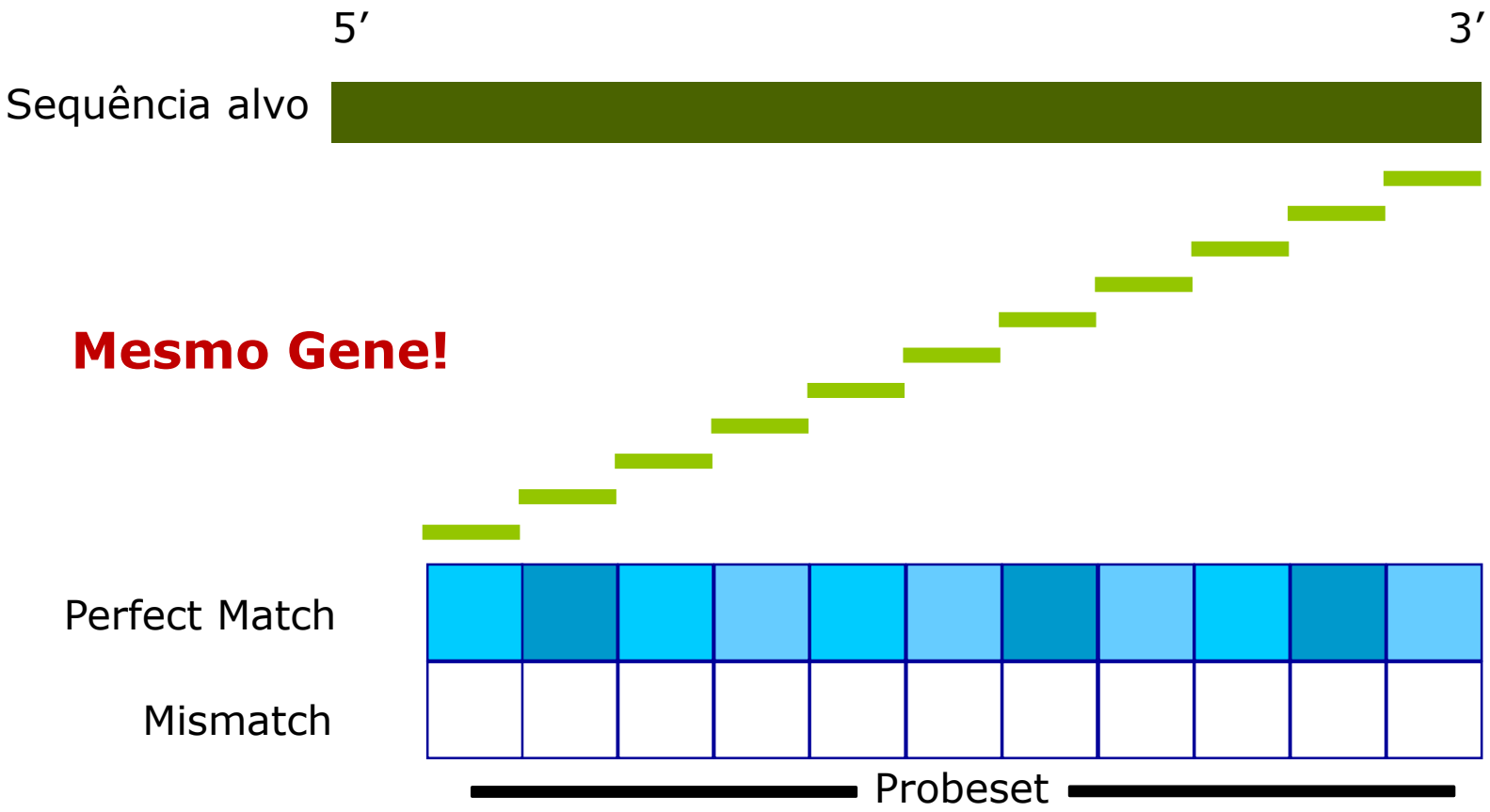
Análise dos dados de expressão gênica – Pré-processamento

- Sumarização



Análise dos dados de expressão gênica – Pré-processamento

- Sumarização



Análise dos dados de expressão gênica – Pré-processamento

- Sumarização:
  - É preciso "sumarizar" todas as sondas para que deem um valor único por *probeset*
  - É preciso "sumarizar" todos os *probesets* para que deem uma valor único por *gene*

Análise dos dados de expressão gênica – Pré-processamento

- Sumarização:
  - É preciso "sumarizar" todas as sondas para que deem um valor único por *probeset*
  - É preciso "sumarizar" todos os *probesets* para que deem uma valor único por *gene*

	Amostra1	Amostra2	Amostra3	...
sonda1	0,957041	1960,81	0,149155	0,552391
sonda2	0,980226	959,1938	0,069077	0,673049
sonda3	0,959119	1649,556	0,076062	0,634061
	0,933941	2570,302	0,155557	1,297719

Número de linhas = milhões (spots no array)

	Amostra1	Amostra2	Amostra3	...
Gene1	1,196301	2451,013	0,186444	0,690489
Gene2	1,225283	1198,992	0,086346	0,841311
Gene3	1,198899	2061,945	0,095078	0,792576
...	1,167426	3212,878	0,194446	1,622149

Número de linhas = 30.000 (transcritos do genoma)

Análise dos dados de expressão gênica

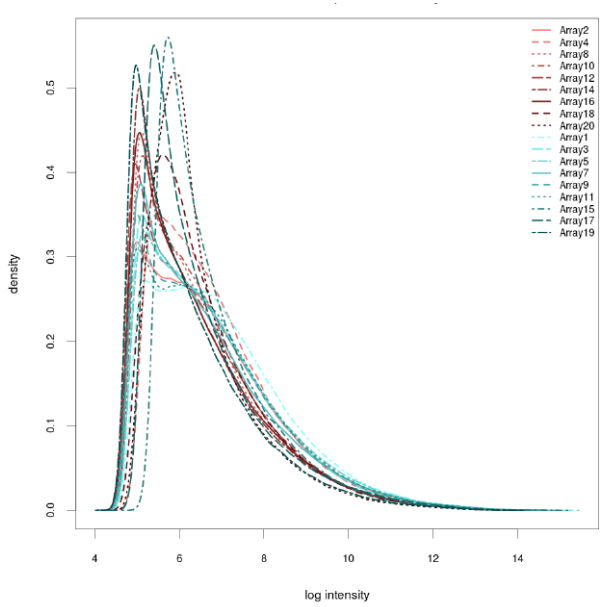
- Principais etapas da análise:
  - Processamento da imagem
  - Pré-processamento dos dados
    - Correção de *background* (ruído)
    - Normalização
    - Sumarização
  - Checagem do pré-processamento (controle de qualidade)
  - Análise exploratória dos dados
  - Identificação dos genes diferencialmente expressos
  - Análises funcionais (vias, ontologia gênica, etc.)

Análise dos dados de expressão gênica – Controle de Qualidade

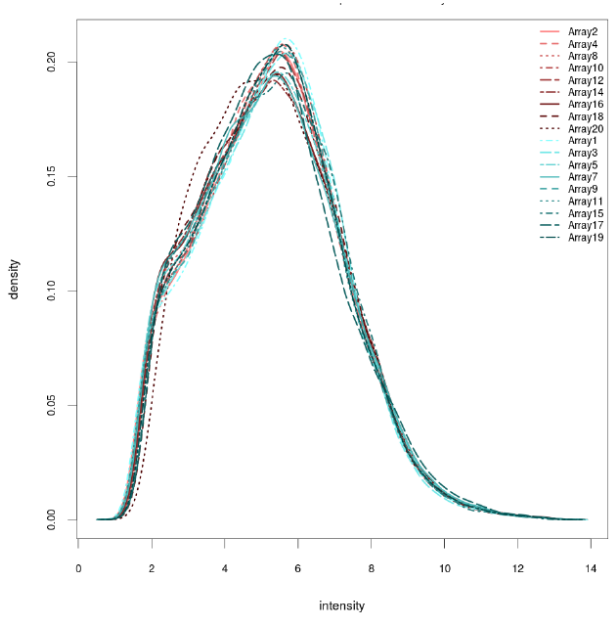
- Controle de qualidade
  - Será que o pré-processamento foi suficiente?
  - Existem *outliers*?
  - Existem grupos de amostras que se diferenciam drasticamente de outras?

Análise dos dados de expressão gênica – Controle de Qualidade

- *Density plots*



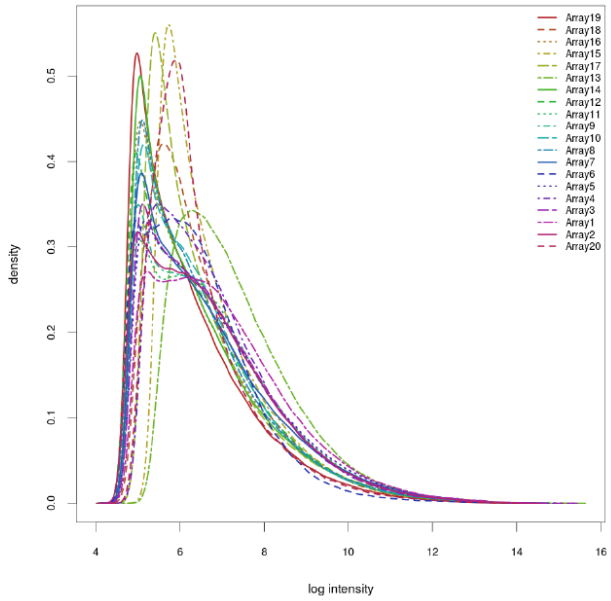
Antes



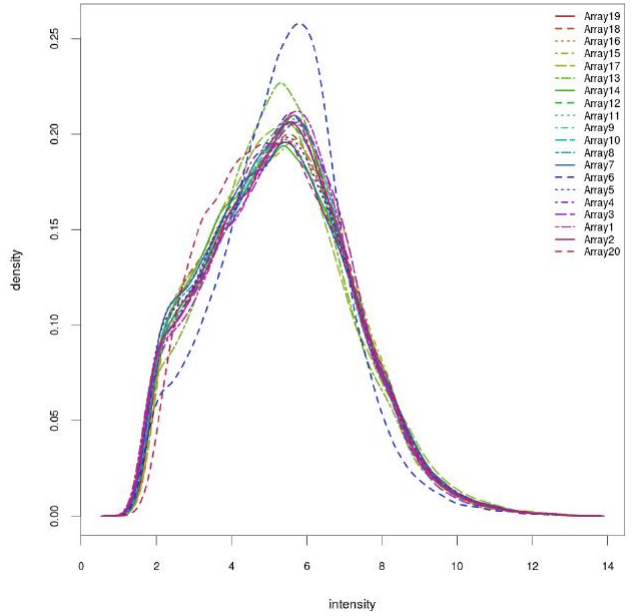
Depois

Análise dos dados de expressão gênica – Controle de Qualidade

- *Density plots*



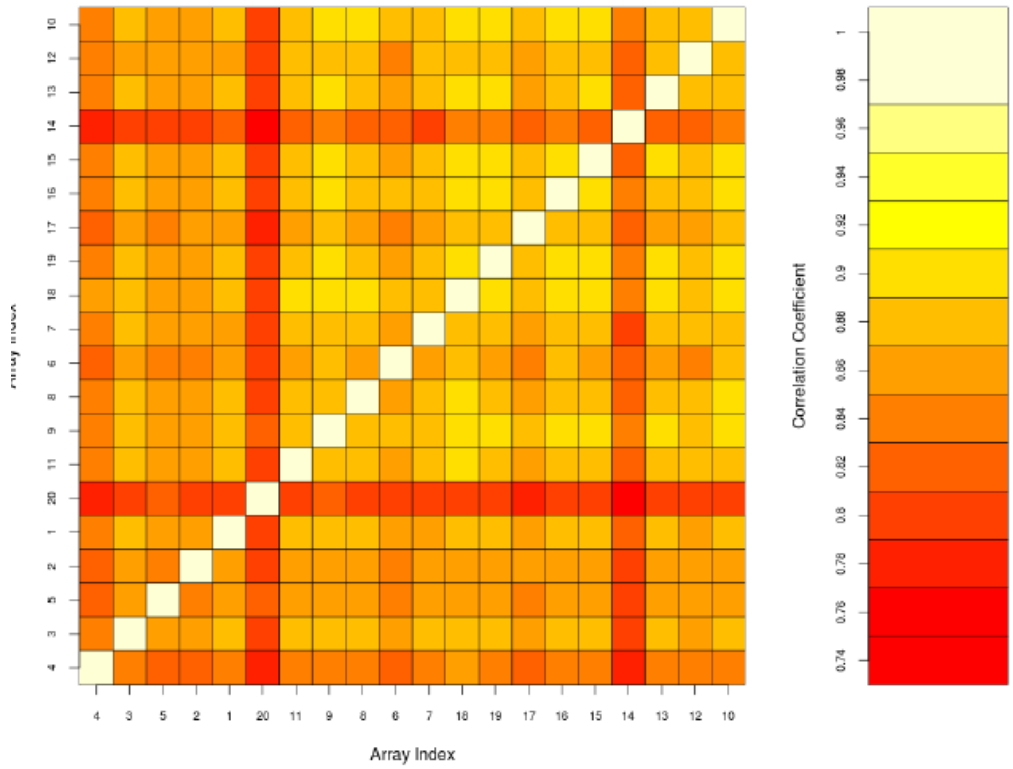
Antes



Depois

Análise dos dados de expressão gênica – Controle de Qualidade

- *Correlation plot*



Correlação:

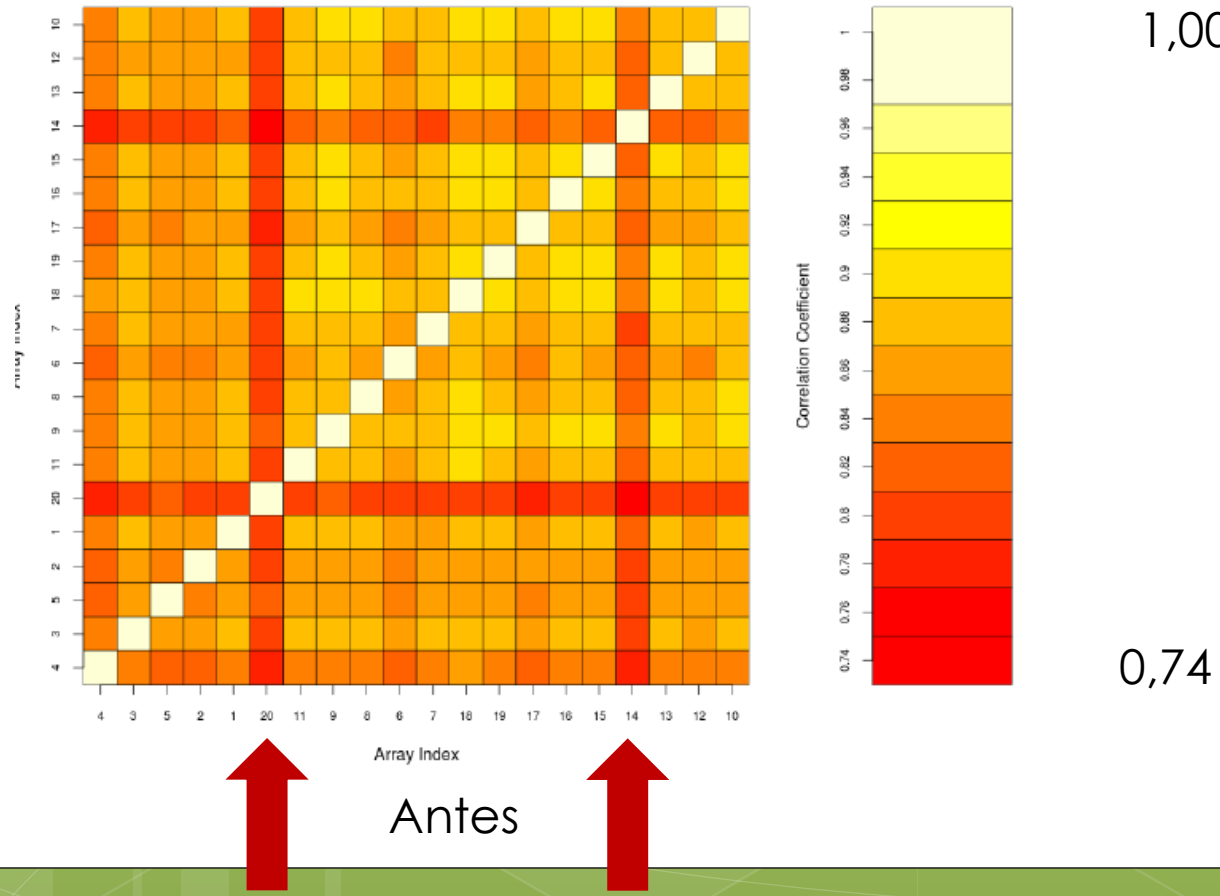
1,00

0,74

Antes

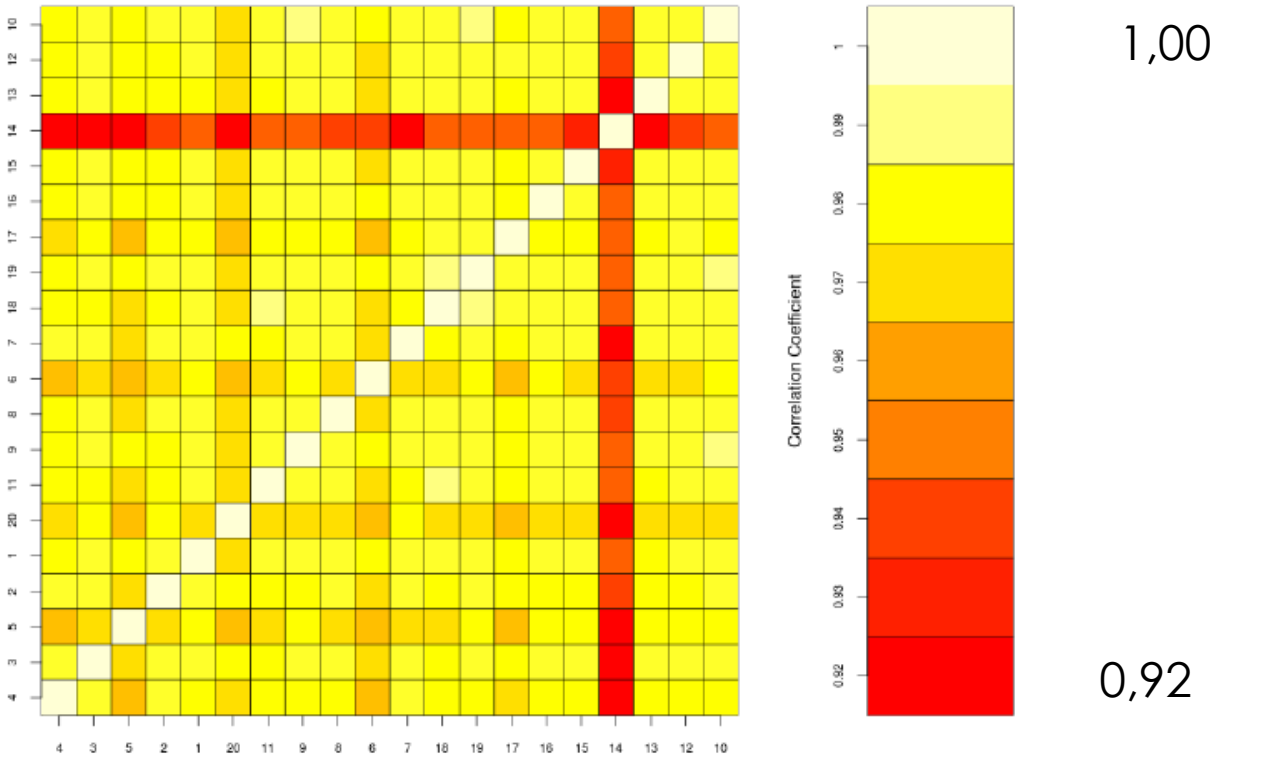
Análise dos dados de expressão gênica – Controle de Qualidade

- *Correlation plot*



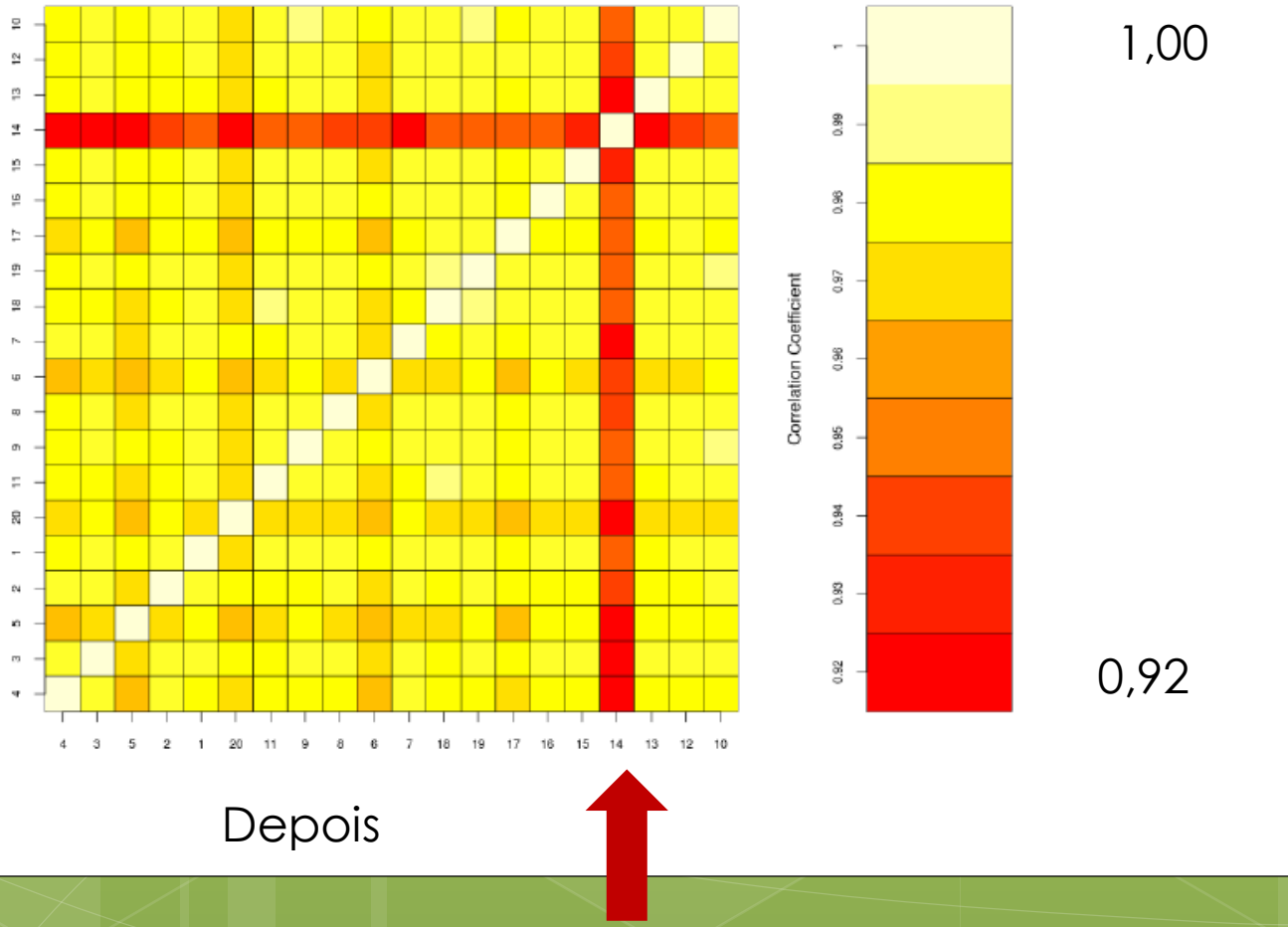
Análise dos dados de expressão gênica – Controle de Qualidade

- *Correlation plot*



Análise dos dados de expressão gênica – Controle de Qualidade

- *Correlation plot*



Análise dos dados de expressão gênica

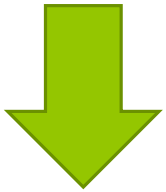
- Principais etapas da análise:
  - Processamento da imagem
  - Pré-processamento dos dados
    - Correção de *background* (ruído)
    - Normalização
    - Sumarização
  - Checagem do pré-processamento (controle de qualidade)
  - Análise exploratória dos dados
  - Identificação dos genes diferencialmente expressos
  - Análises funcionais (vias, ontologia gênica, etc.)

Análise dos dados de expressão gênica – Análise exploratória

- Análise exploratória dos dados
  - Como é o perfil de expressão gênica global em todas as amostras?
  - Será que as amostras se agrupam de acordo com o perfil de expressão?
  - Quão "próximas" estão as amostras?

Análise dos dados de expressão gênica – Análise exploratória

- Análise exploratória dos dados
  - Como é o perfil de expressão gênica global em todas as amostras?
  - Será que as amostras se agrupam de acordo com o perfil de expressão?
  - Quão "próximas" estão as amostras?



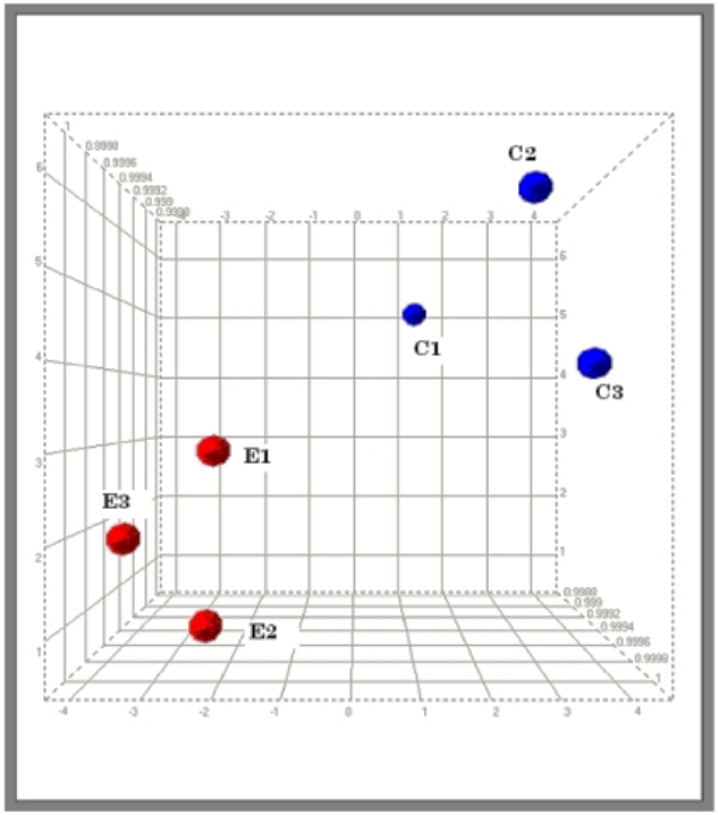
1. **Redução de dimensionalidade**
2. **"Clusterização"**

Análise dos dados de expressão gênica – Análise exploratória

- Redução de dimensionalidade
  - 30.000 transcritos → poucas variáveis
  - Exemplo de metodologia: Análise dos Componentes Principais (PCA)
    - Aplicada para agrupar amostras (arrays) diferentes

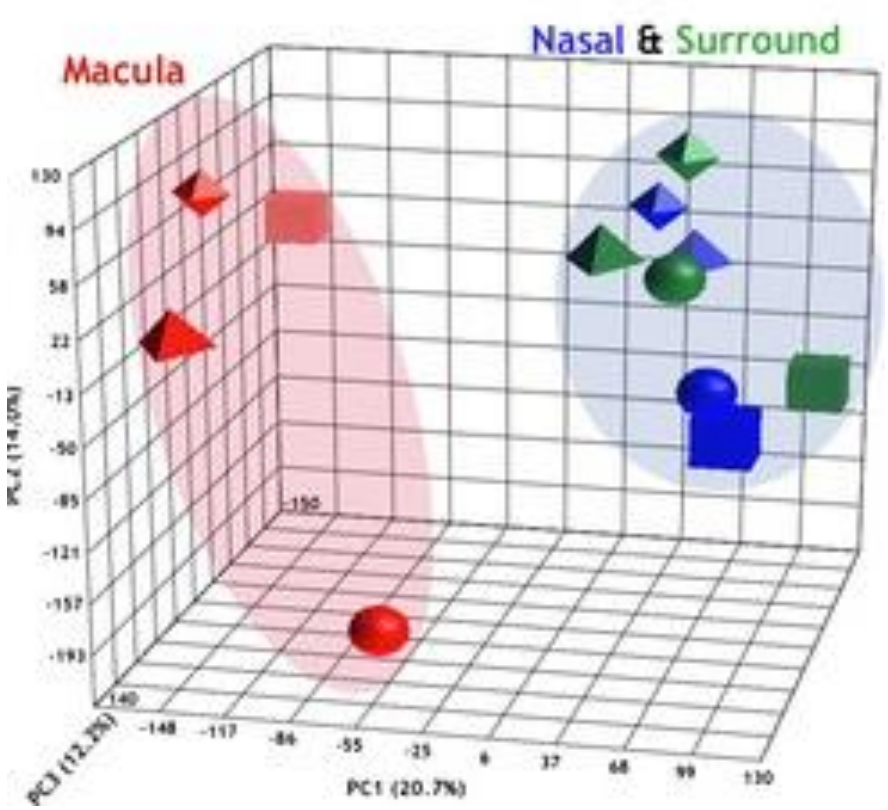
Análise dos dados de expressão gênica – Análise exploratória

- Análise dos Componentes Principais
  - 3 variáveis
  - 3 dimensões



Análise dos dados de expressão gênica – Análise exploratória

- Análise dos Componentes Principais

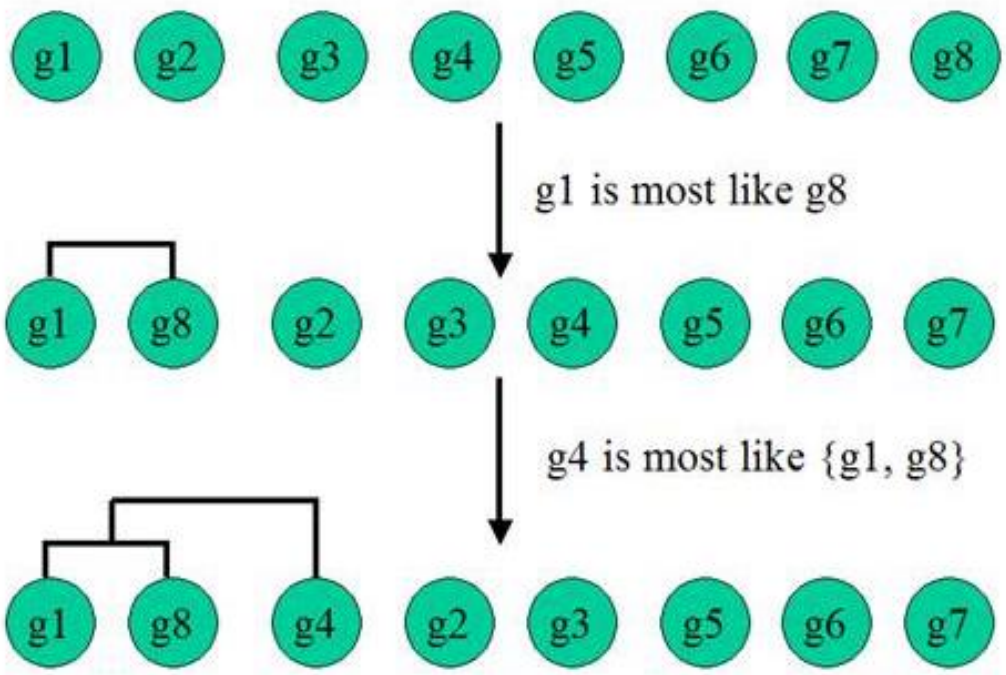


Análise dos dados de expressão gênica – Análise exploratória

- "Clusterização"
  - Agrupamento em categorias não previamente definidas
  - Diversas metodologias:
    - *Hierarchical clustering*
    - *Self organizing maps*
    - *K means clustering*
  - Pode ser feita tanto em relação aos genes quanto às amostras (*arrays*)

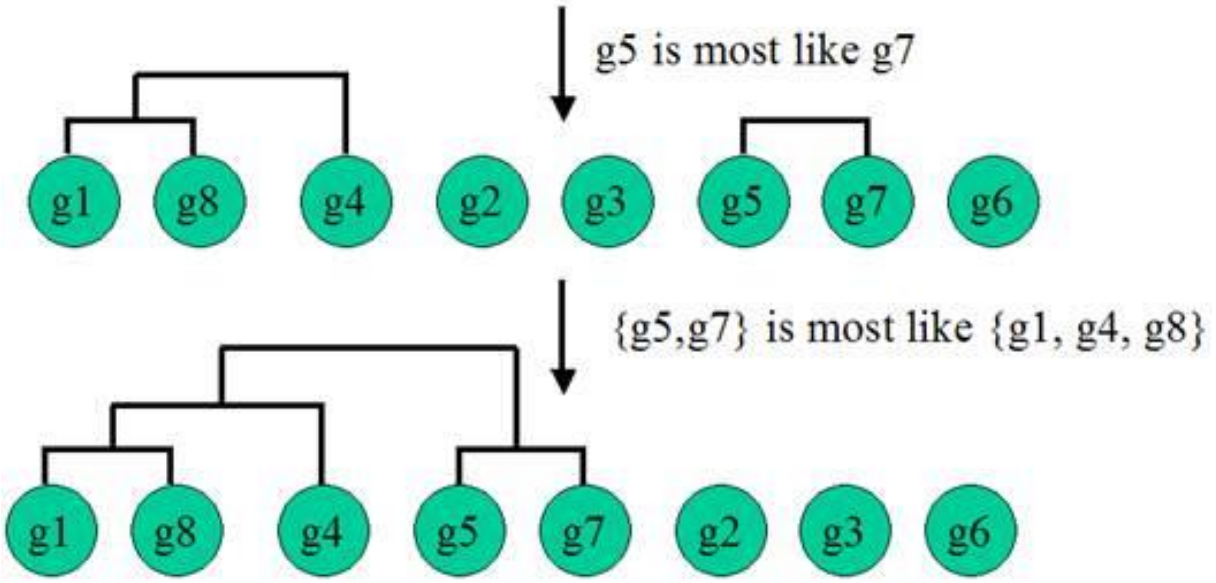
Análise dos dados de expressão gênica – Análise exploratória

- Agrupamento hierárquico



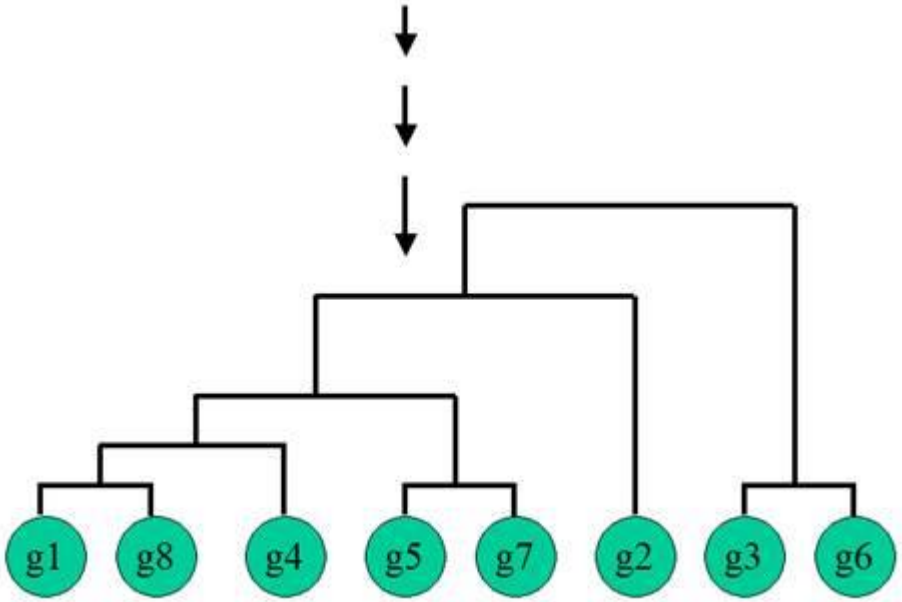
Análise dos dados de expressão gênica – Análise exploratória

- Agrupamento hierárquico



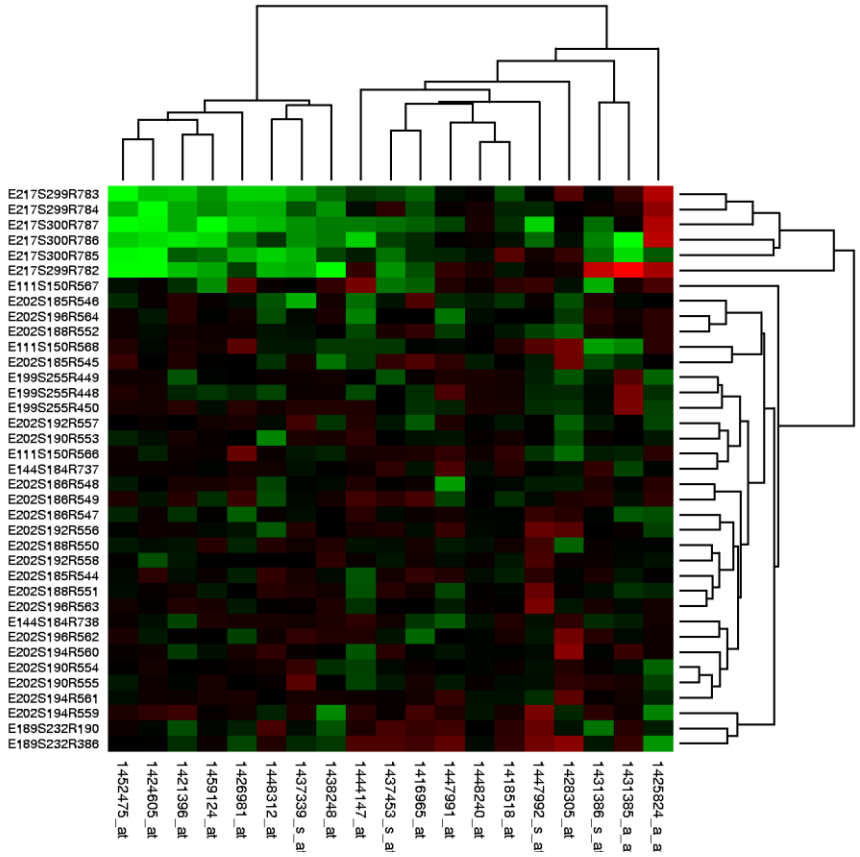
Análise dos dados de expressão gênica – Análise exploratória

- Agrupamento hierárquico



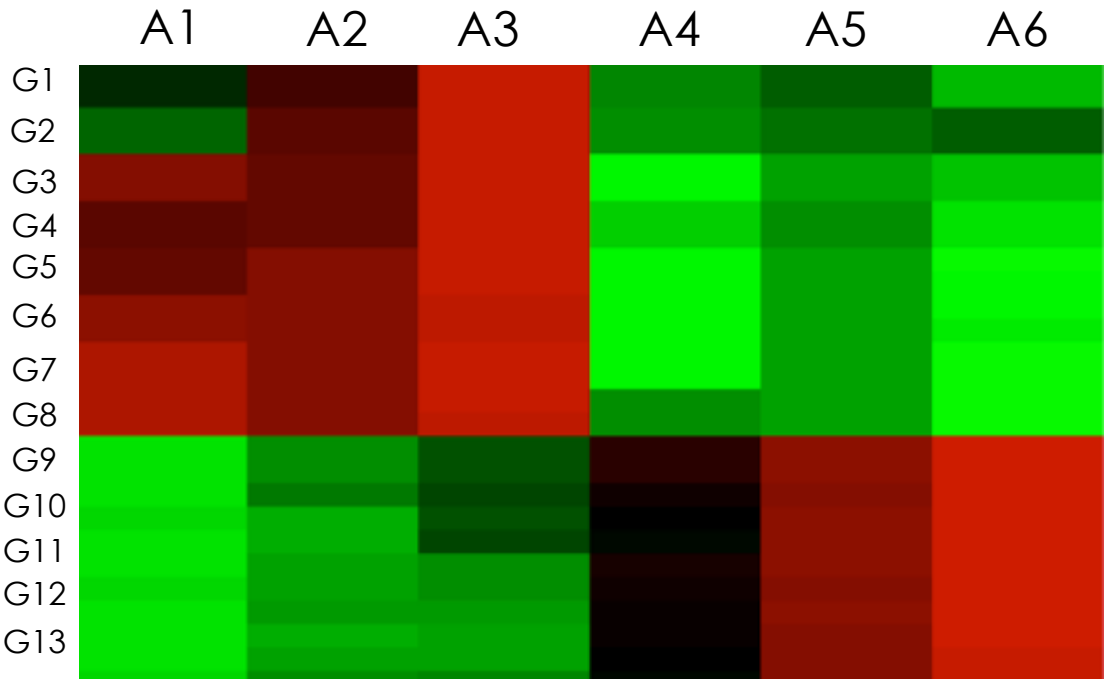
Análise dos dados de expressão gênica – Análise exploratória

- Agrupamento hierárquico + visualização (*heat map*)

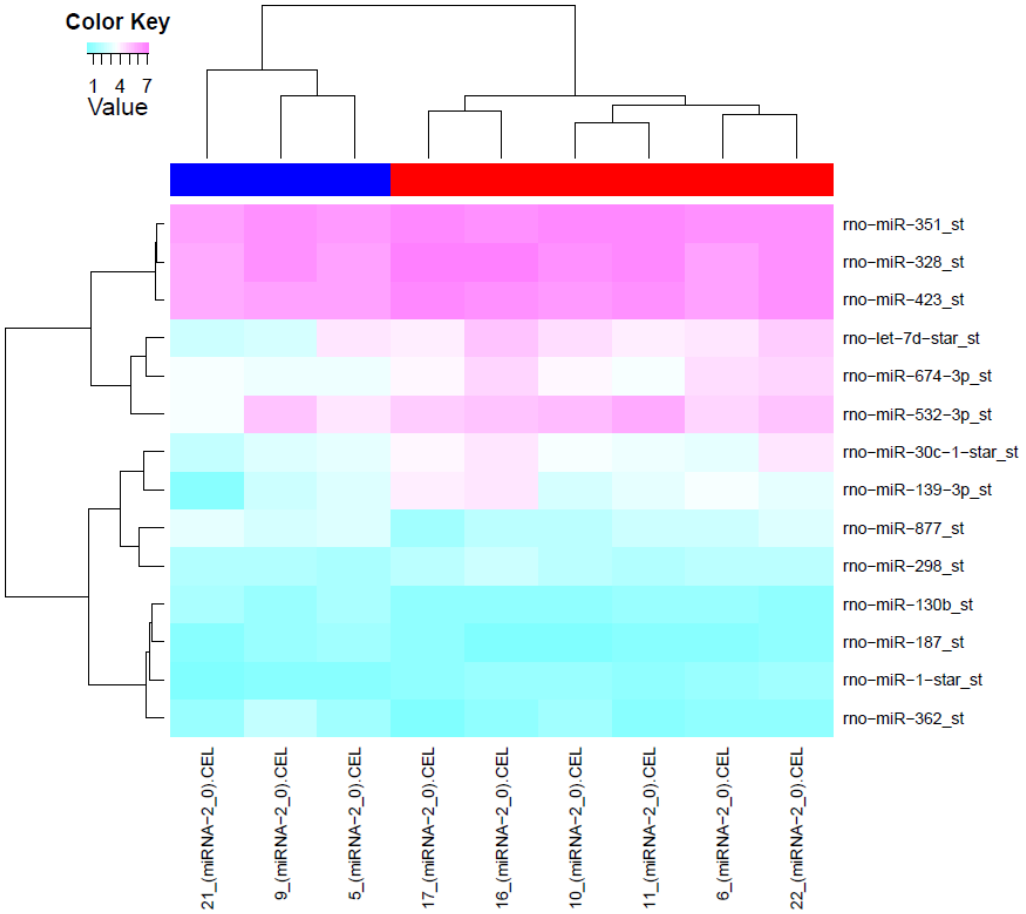


Análise dos dados de expressão gênica – Análise exploratória

- Visualização (*heat map*)



Análise dos dados de expressão gênica – Análise exploratória



Análise dos dados de expressão gênica

- Principais etapas da análise:
  - Processamento da imagem
  - Pré-processamento dos dados
    - Correção de *background* (ruído)
    - Normalização
    - Sumarização
  - Checagem do pré-processamento (controle de qualidade)
  - Análise exploratória dos dados
  - Identificação dos genes diferencialmente expressos
  - Análises funcionais (vias, ontologia gênica, etc.)

Análise dos dados de expressão gênica – Genes Diferencialmente Expressos

- Quais, dentre os 30.000 transcritos estão diferencialmente expressos?
  - Uso de testes estatísticos apropriados
  - Correção para múltiplos testes
  - Efeito da magnitude
- Identificação de um grupo de genes que atendem estes critérios
- "Lista de Genes"

Análise dos dados de expressão gênica – Genes Diferencialmente Expressos

- Testes mais usados:
  - Teste t / ANOVA
  - Testes não-paramétricos (Rank-Prod)
  - Modelos lineares (LIMMA)
  - Significance Analysis of Microarray (SAM)

- Correção para múltiplos testes
  - *False Discovery Rate*
  - Correção de Bonferroni

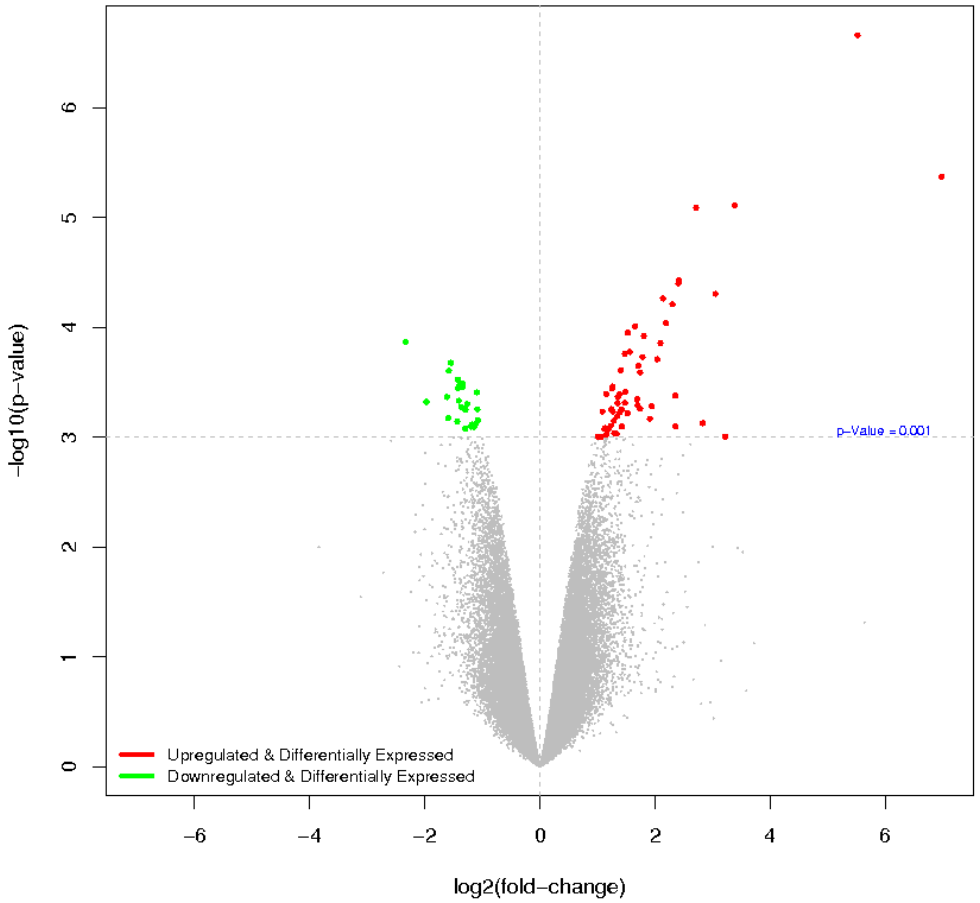
Análise dos dados de expressão gênica – Genes Diferencialmente Expressos

- Caracterização de gene como diferencialmente expresso:
  - Valor de p menor do que nível de significância escolhido
  - *Fold change* (expressão relativa)

Análise dos dados de expressão gênica – Genes Diferencialmente Expressos

- Genes diferencialmente expressos – *Volcano plot*

-log10(valor de p)



Fold Change

Análise dos dados de expressão gênica – Genes Diferencialmente Expressos

	Amostra1	Amostra2	Amostra3	...
Gene1	1,196301	2451,013	0,186444	0,690489
Gene2	1,225283	1198,992	0,086346	0,841311
Gene3	1,198899	2061,945	0,095078	0,792576
...	1,167426	3212,878	0,194446	1,622149

Número de linhas = 30.000 (transcritos do genoma)

Gene1	1,196301	2451,013	0,186444	0,690489
Gene2	1,225283	1198,992	0,086346	0,841311
Gene3	1,198899	2061,945	0,095078	0,792576
Gene4	1,167426	3212,878	0,194446	1,622149
Gene5	1,168725	3018,344	0,148763	1,673193
Gene6	1,157424	3333,199	0,152037	1,947817
Gene7	1,146123	3648,053	0,155311	2,222442
...	...	...	...	...
Gene112	1,123522	4277,763	0,161859	2,771691
Gene113	1,112221	4592,618	0,165133	3,046315

Número de linhas = 113 genes diferencialmente expressos

Análise dos dados de expressão gênica – Genes Diferencialmente Expressos

Caracterização dos Genes Diferencialmente Expressos

Agrupamento ("supervisionado")

Hierarchical clustering

Self organizing maps

K means clustering

Análise Funcional

Pathway Analysis

Gene Ontolgy

Análise dos dados de expressão gênica

- Principais etapas da análise:
  - Processamento da imagem
  - Pré-processamento dos dados
    - Correção de *background* (ruído)
    - Normalização
    - Sumarização
  - Checagem do pré-processamento (controle de qualidade)
  - Análise exploratória dos dados
  - Identificação dos genes diferencialmente expressos
  - Análises funcionais (vias, ontologia gênica, etc.)

Análise dos dados de expressão gênica – Análise Funcional

- O que fazer com uma lista de genes diferencialmente expressos?
- Como dar significado biológico para estes genes?
- Os genes participam de processos celulares em comum?
- Os genes interagem em uma cascata biológica?

Análise dos dados de expressão gênica – Análise Funcional

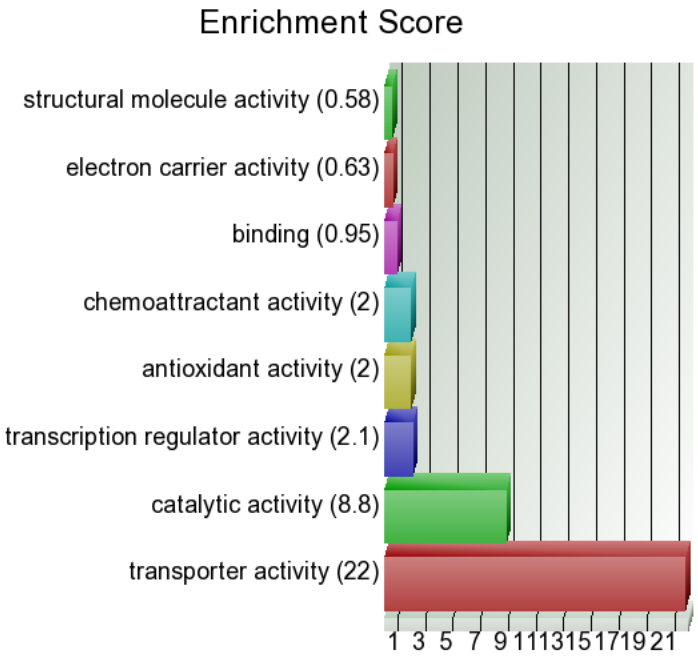
- Principais Ferramentas:
  - *Gene Ontology Analysis* (www.geneontology.org/)
  - *Pathway Analysis*
  - *Gene Set Enrichment Analysis*

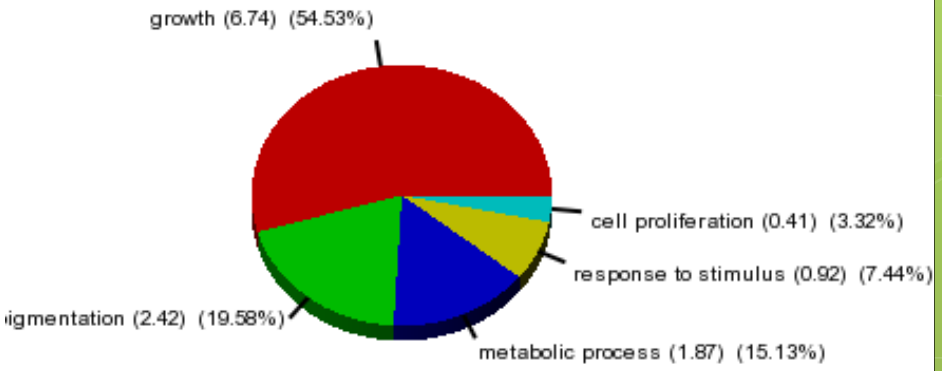
Análise dos dados de expressão gênica – Análise Funcional

- *Gene Ontology*:
  - Banco de dados que mantém anotações de genes com base em 3 categorias (ontologias):
    - *Molecular function* (ex: fator de transcrição)
    - *Biological process* (ex: mitose)
    - *Cellular components* (ex: núcleo)
  - Lista de genes é confrontada com banco de dados para identificar quais dessas ontologias estão mais representadas

Análise dos dados de expressão gênica – Análise Funcional

- *Gene Ontology*:



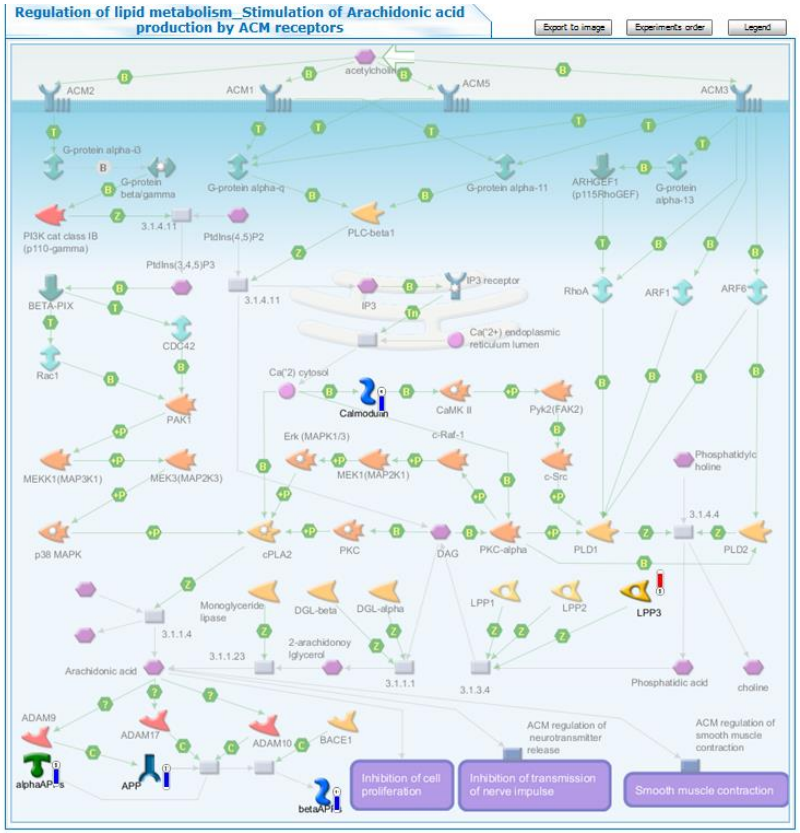


Análise dos dados de expressão gênica – Análise Funcional

- *Pathway Analysis*
  - *Biological Pathway* ou via biológica → série de moléculas que interagem direta ou indiretamente como parte de um evento biológico (ex: cascata enzimática)
  - Lista de genes é confrontada para verificar quais vias biológicas estão enriquecidas de acordo com os genes presentes

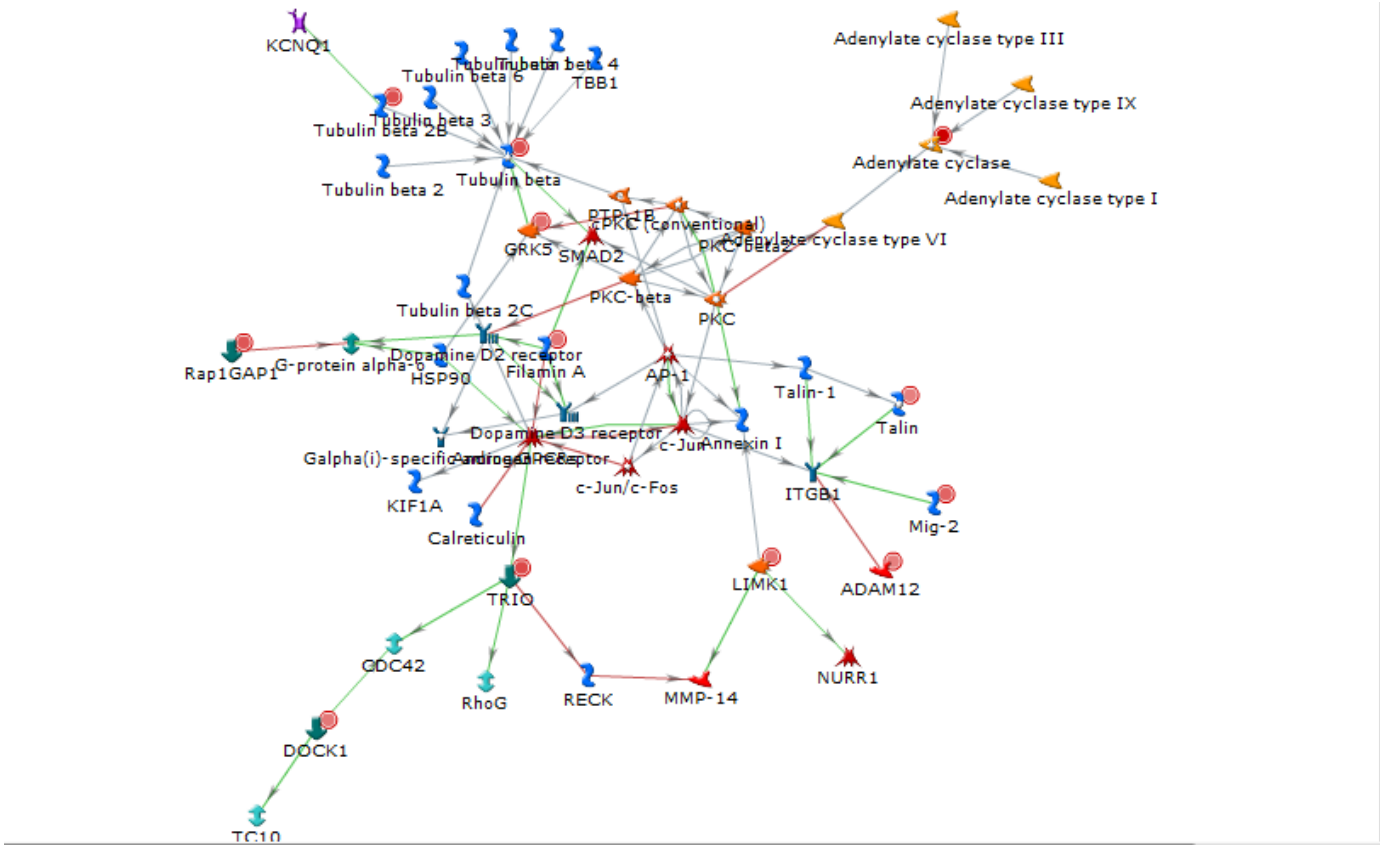
Análise dos dados de expressão gênica – Análise Funcional

- *Pathway Analysis*



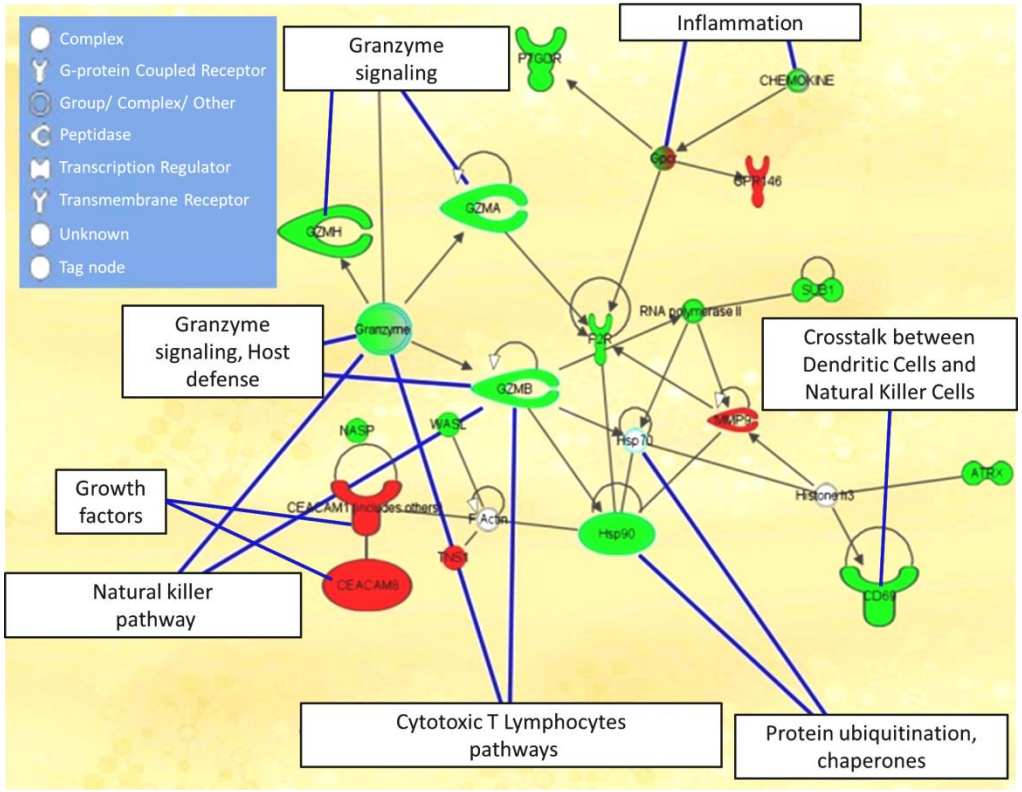
Análise dos dados de expressão gênica – Análise Funcional

- *Network Analysis*



Análise dos dados de expressão gênica – Análise Funcional

- Exemplos da vida real – privação de sono em humanos



Análise dos dados de expressão gênica – Análise Funcional

- *Gene Set Enrichment Analysis (GSEA)*
  - Subramanian et al, 2005
  - Estratégia semelhante, mas "sets" de genes são caracterizados por conhecimento biológico prévio
    - Informações publicadas sobre vias bioquímicas, coexpressão de moléculas, etc

Análise dos dados de expressão gênica – Análise Funcional

Collections

The MSigDB gene sets are divided into 6 major collections:

c1 **positional gene sets** for each human chromosome and cytogenetic band.

c2 **curated gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.

c3 **motif gene sets** based on conserved cis-regulatory motifs from a comparative analysis of the human, mouse, rat, and dog genomes.

c4 **computational gene sets** defined by mining large collections of cancer-oriented microarray data.

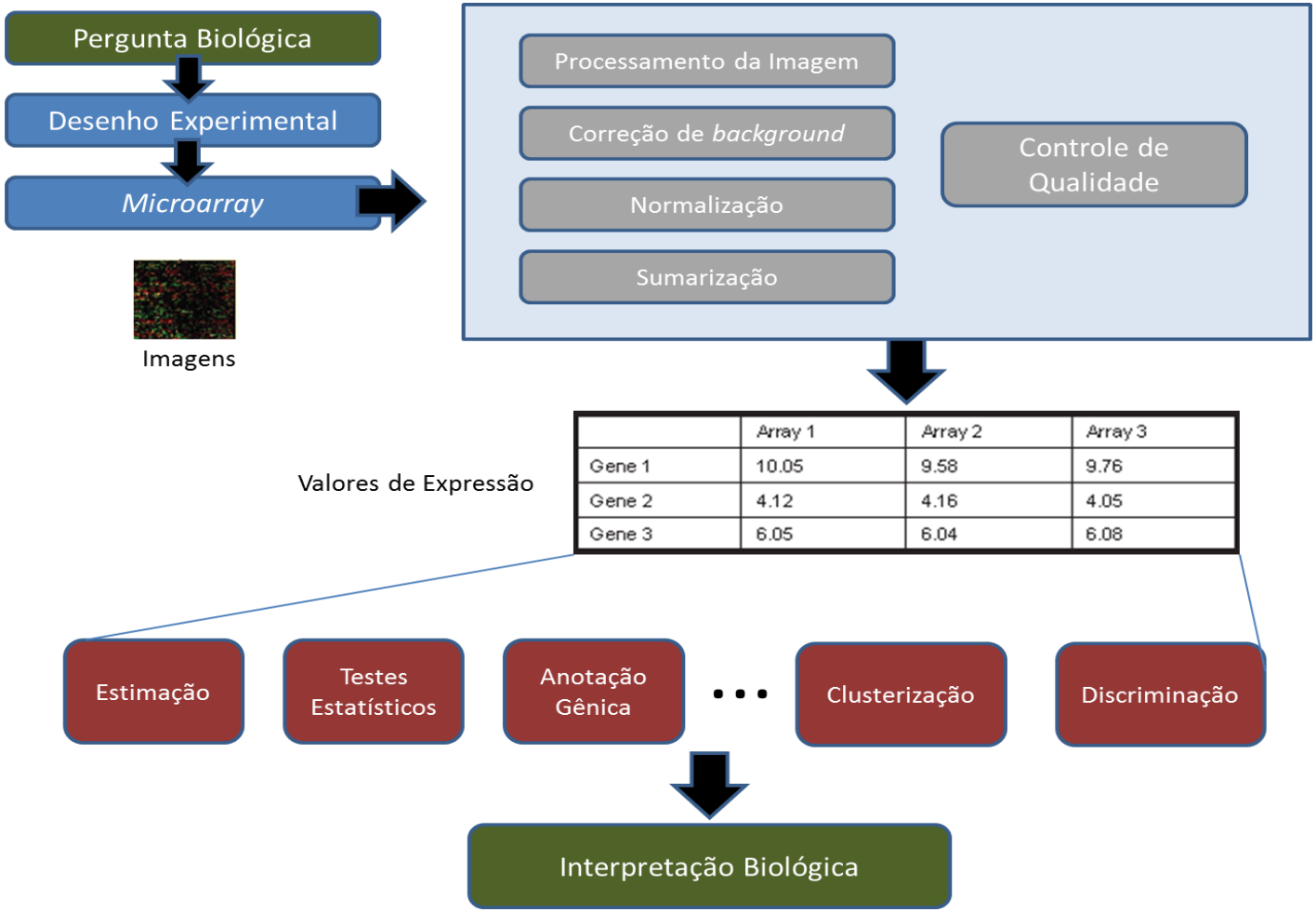
c5 **GO gene sets** consist of genes annotated by the same GO terms.

c6 **oncogenic signatures** defined directly from microarray gene expression data from cancer gene perturbations.

Análise dos dados de expressão gênica

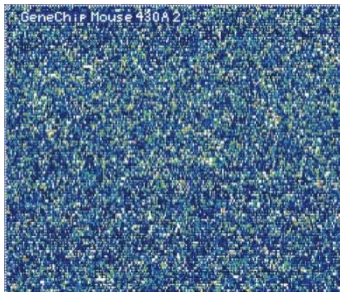
- Principais etapas da análise:
  - Processamento da imagem
  - Pré-processamento dos dados
    - Correção de *background* (ruído)
    - Normalização
    - Sumarização
  - Checagem do pré-processamento (controle de qualidade)
  - Análise exploratória dos dados
  - Identificação dos genes diferencialmente expressos
  - Análises funcionais (vias, ontologia gênica, etc.)

Análise dos dados de expressão gênica



Adaptado de Bolstad, B. Methodologies for Pre-processing Microarray Data, 2006

Análise dos dados de expressão gênica



		MR*	COR*	symbol	function
0	☑			PSMD14	proteasome (prosome, macropain) 26S subunit, non-ATPase, 14
1	☐	2.0	0.68	PSMC2	proteasome (prosome, macropain) 26S subunit, ATPase, 2
2	☐	2.5	0.66	PSMD12	proteasome (prosome, macropain) 26S subunit, non-ATPase, 12
3	☐	2.6	0.64	PSMD1	proteasome (prosome, macropain) 26S subunit, non-ATPase, 1
4	☐	4.5	0.62	OLA1	Obg-like ATPase 1
5	☐	4.7	0.61	EIF3J	eukaryotic translation initiation factor 3, subunit J
6	☐	5.0	0.64	PSMB7	proteasome (prosome, macropain) subunit, beta type, 7
7	☐	5.5	0.63	VBP1	von Hippel-Lindau binding protein 1
8	☐	5.7	0.68	RAN	RAN, member RAS oncogene family
9	☐	6.9	0.64	ZC3H15	zinc finger CCCH-type containing 15
10	☐	7.8	0.63	MRPL3	mitochondrial ribosomal protein L3
11	☐	7.8	0.64	PSMA1	proteasome (prosome, macropain) subunit, alpha type, 1
12	☐	8.8	0.64	PSMA5	proteasome (prosome, macropain) subunit, alpha type, 5
13	☐	9.0	0.63	MRPL47	mitochondrial ribosomal protein L47
14	☐	10.9	0.63	TIMM17A	translocase of inner mitochondrial membrane 17 homolog A (yeast)
15	☐	11.4	0.54	USP14	ubiquitin specific peptidase 14 (tRNA-guanine transglycosylase)