生物医学工程导论

第十六章 基因芯片与数据分析

基因芯片数据分析

1. 基因芯片(Microarray)简介

2. 图像处理与数据标准化

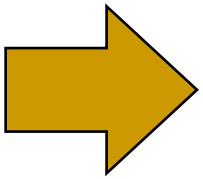
3. 基因芯片的数据分析

生物信息学在基因芯片中的应用

提取什么信息

如何提取信息

如何处理和利用信息



确定芯片检测目标

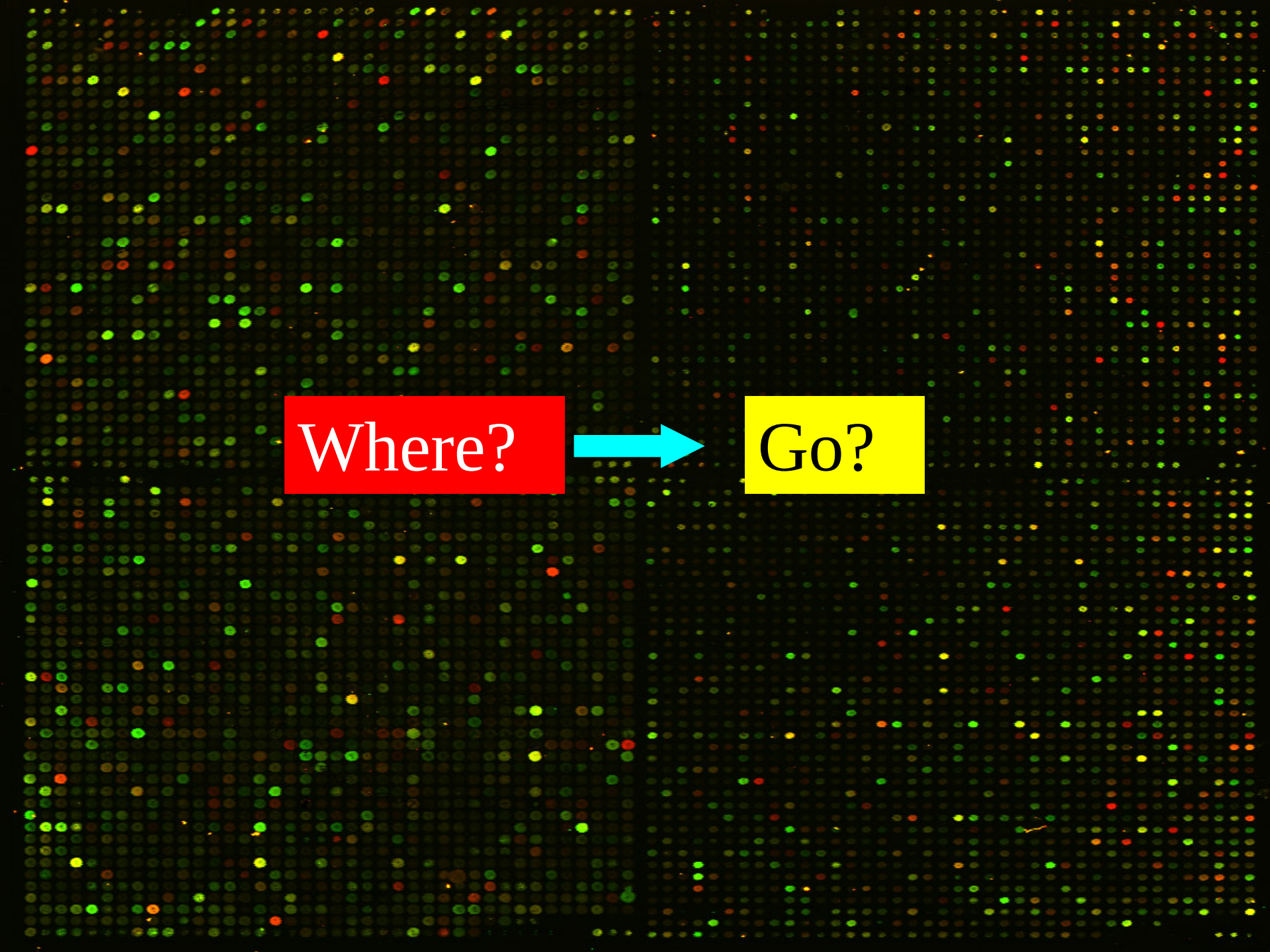
芯片设计

数据管理与分析

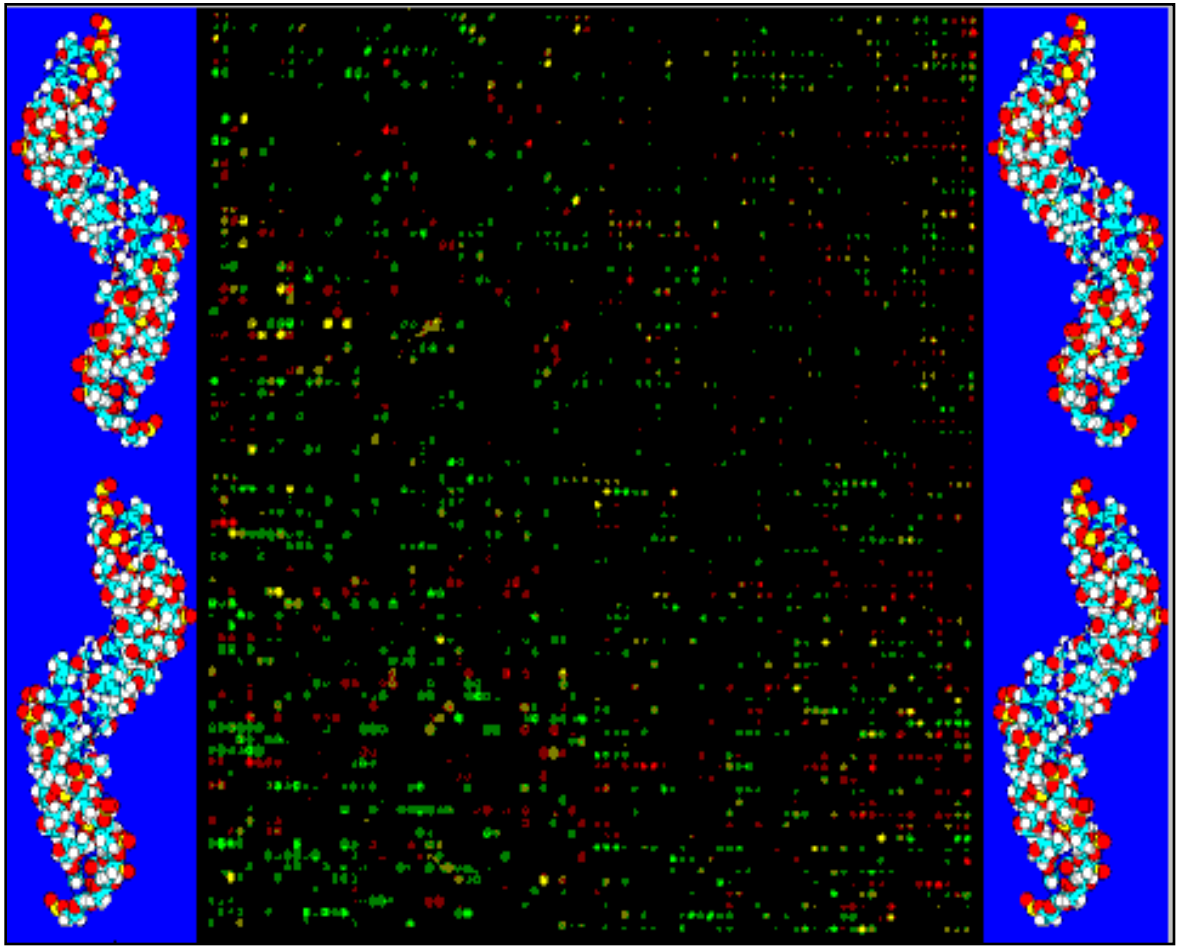
促进

生物信息学 ⇄ 基因芯片

丰富


Where?
Go?

1. 基因芯片简介

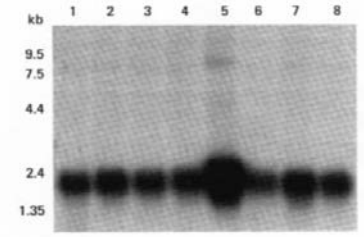


1. 基因芯片简介

- 基因芯片 (1987): 固定有寡核苷酸、DNA或cDNA等的生物芯片。基因芯片（gene chip）也叫DNA芯片、DNA微阵列（DNA microarray）、寡核苷酸阵列（oligonucleotide array）
- 是指采用原位合成（in situ synthesis）或显微打印手段，将数以万计的DNA探针固化于支持物表面上，产生二维DNA探针阵列
- 利用这类芯片与标记生物样品进行杂交，可对样品基因表达谱生物信息进行快速定性和定量分析。

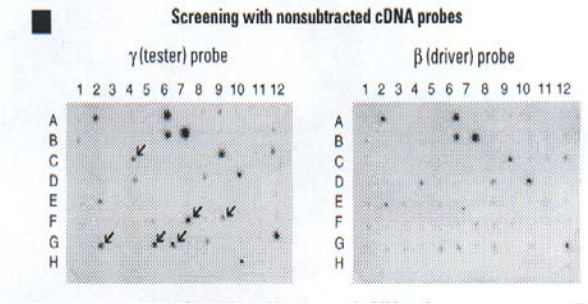
基因芯片的发展历史

Southern & Northern Blot

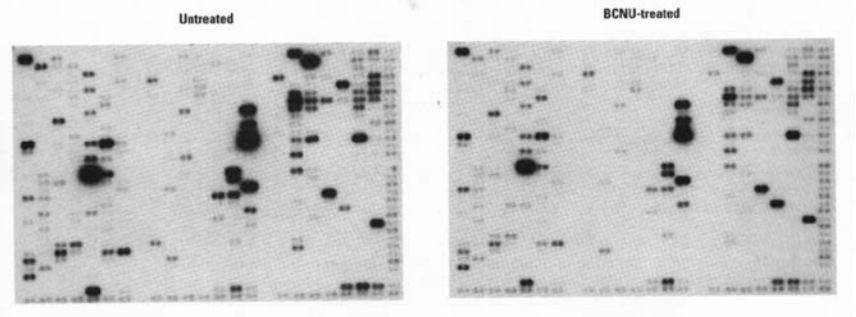


高通量、点阵以及Northern杂交
同时测定细胞内数千个基因的表达情况
将mRNA反转录成cDNA与芯片上的探针杂交

Dot Blot

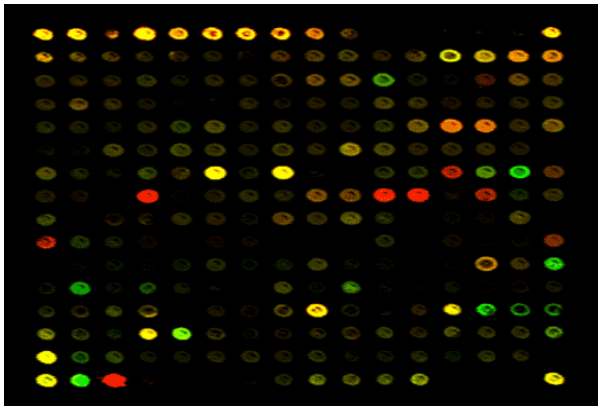


Microarray

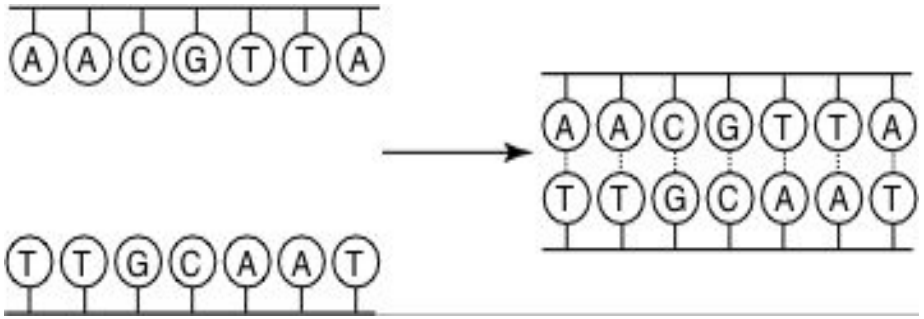


芯片的体积非常小：微量样品的检测
基因表达情况的定量分析

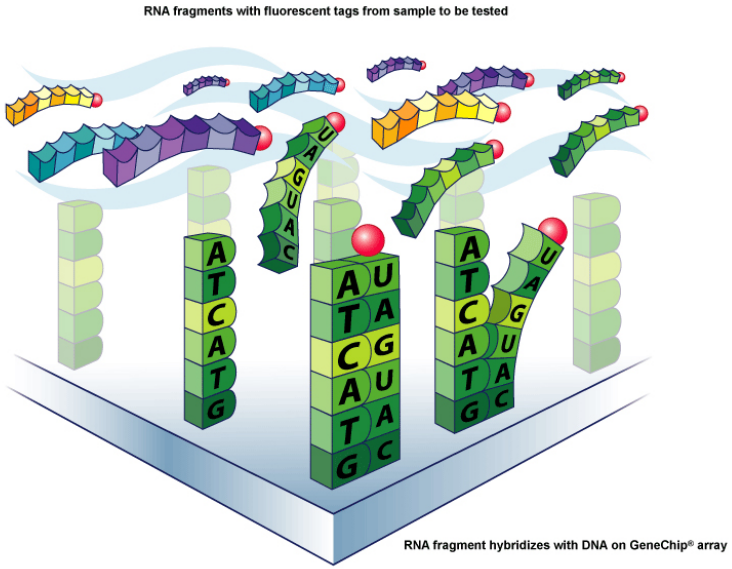
Microarray



基因芯片的密度：100-1 million DNA 探针/1cm^2



碱基互补



将样品中的DNA/RNA标上荧光标记，则可以定量检验基因的表达水平

基因芯片的分类

(一)按载体材料分类

玻璃芯片、硅芯片、陶瓷芯片。

(二)按点样方式分类

原位合成芯片、点样芯片。

(三)按基因芯片的使用功能

测序芯片、表达谱芯片、基因差异表达分析芯片。

（四）按技术手段、探针类型分类

1. Short oligonucleotide arrays (Affymetrix)
2. cDNA arrays (Brown/Botstein)
3. Long oligo arrays (Agilent)
4. Serial analysis of gene expression (SAGE)

（五） 按实验要求分类

1. 单通道 (Single Channel): 一次检验一种状态
2. 双通道 (Dual Channel): 差异表达基因的筛选

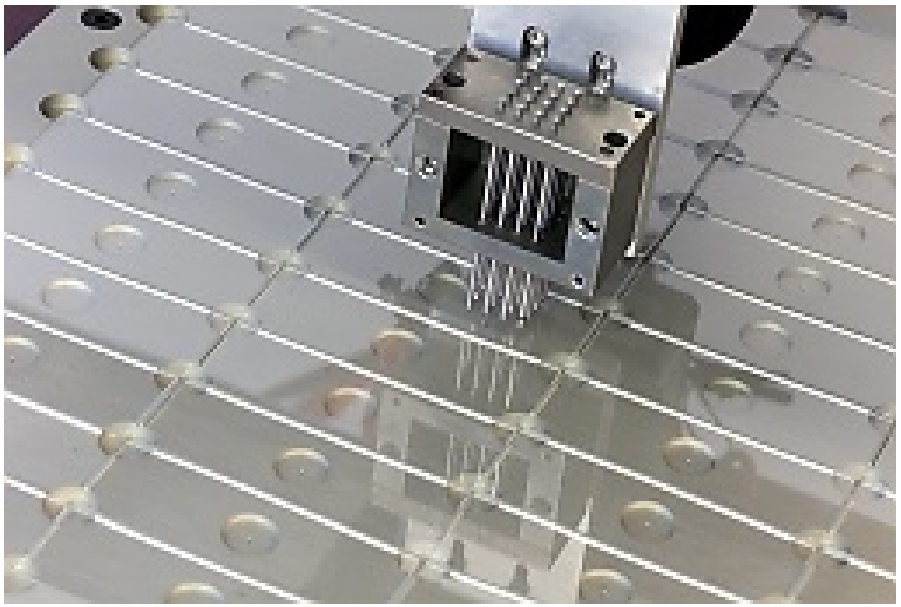
两类主流的DNA芯片

(1). cDNA microarrays: 将500~5,000bp的cDNA固载到介质上 (例如玻璃)。

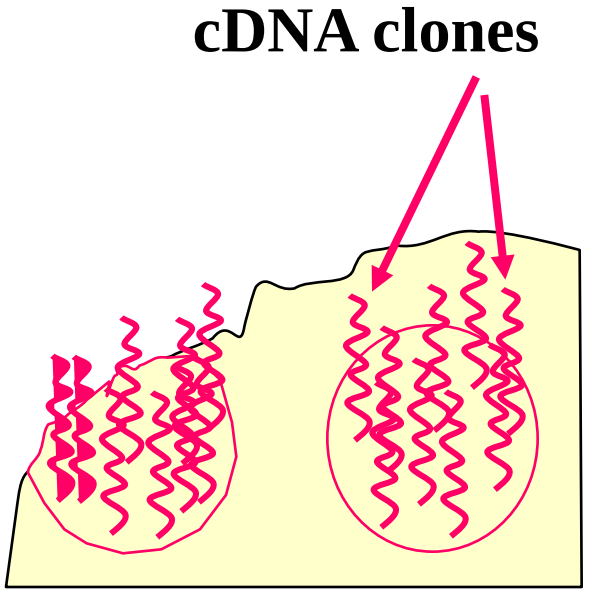
Stanford开发设计，通常为双通道，常用于差异表达基因的筛选。

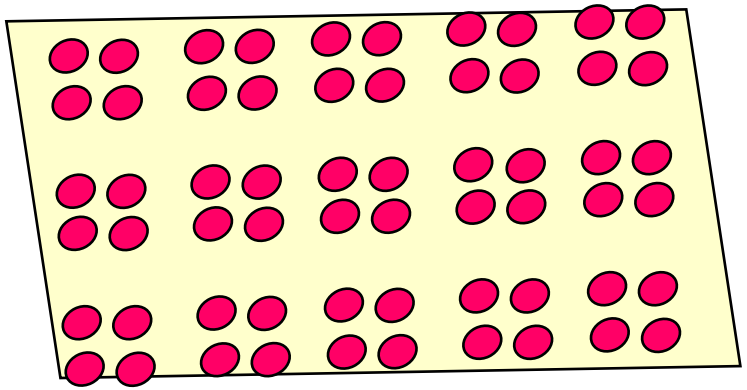
cDNA芯片是被众多实验室生产自己的芯片所使用的一门技术，因为采用双色标记杂交于同一芯片的技术，因而易于标准化和鉴定DEG。

(1) cDNA microarrays

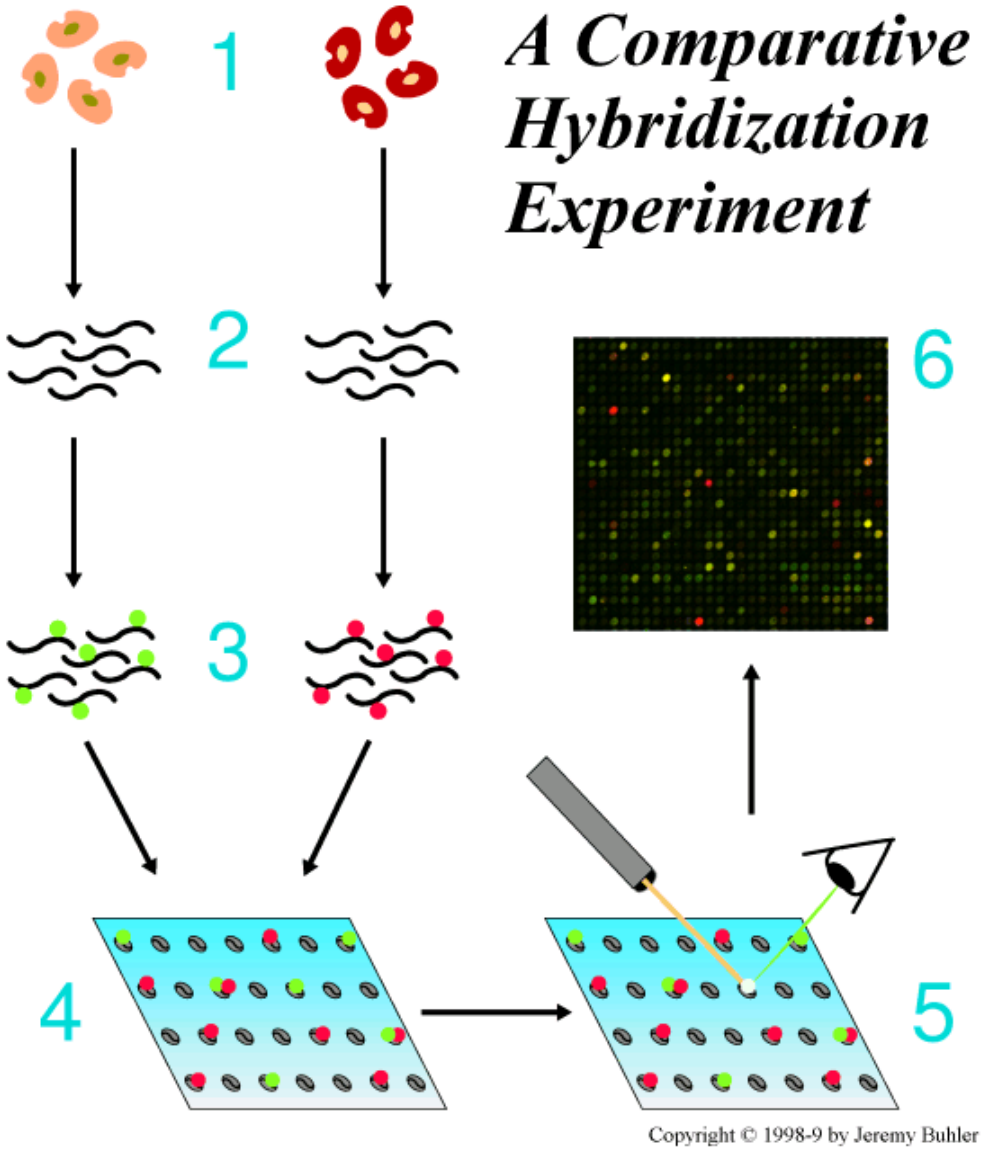


载玻片





差异表达基因的筛选



Treatment / control

Normal / tumor tissue

Brain / liver

…

荧光标记的靶基因

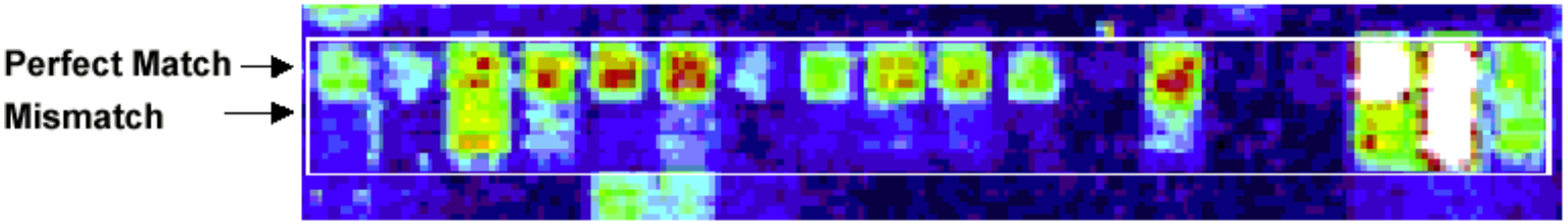
两类主流的DNA芯片

(2). DNA chips: 将寡核苷酸探针 (20~80-mer) 合成到芯片上。Affymetrix开发设计，通常为单通道，一次检验一种状态 。

寡聚核酸芯片是Affy公司原创的，通过原位合成技术制作，对芯片制作工艺技术要求很高，具有分辨率高，均一性好等优点，用于做 classification 等效果非常好，也很方便于通过统计学进行差异表达基因（DEG）筛选。

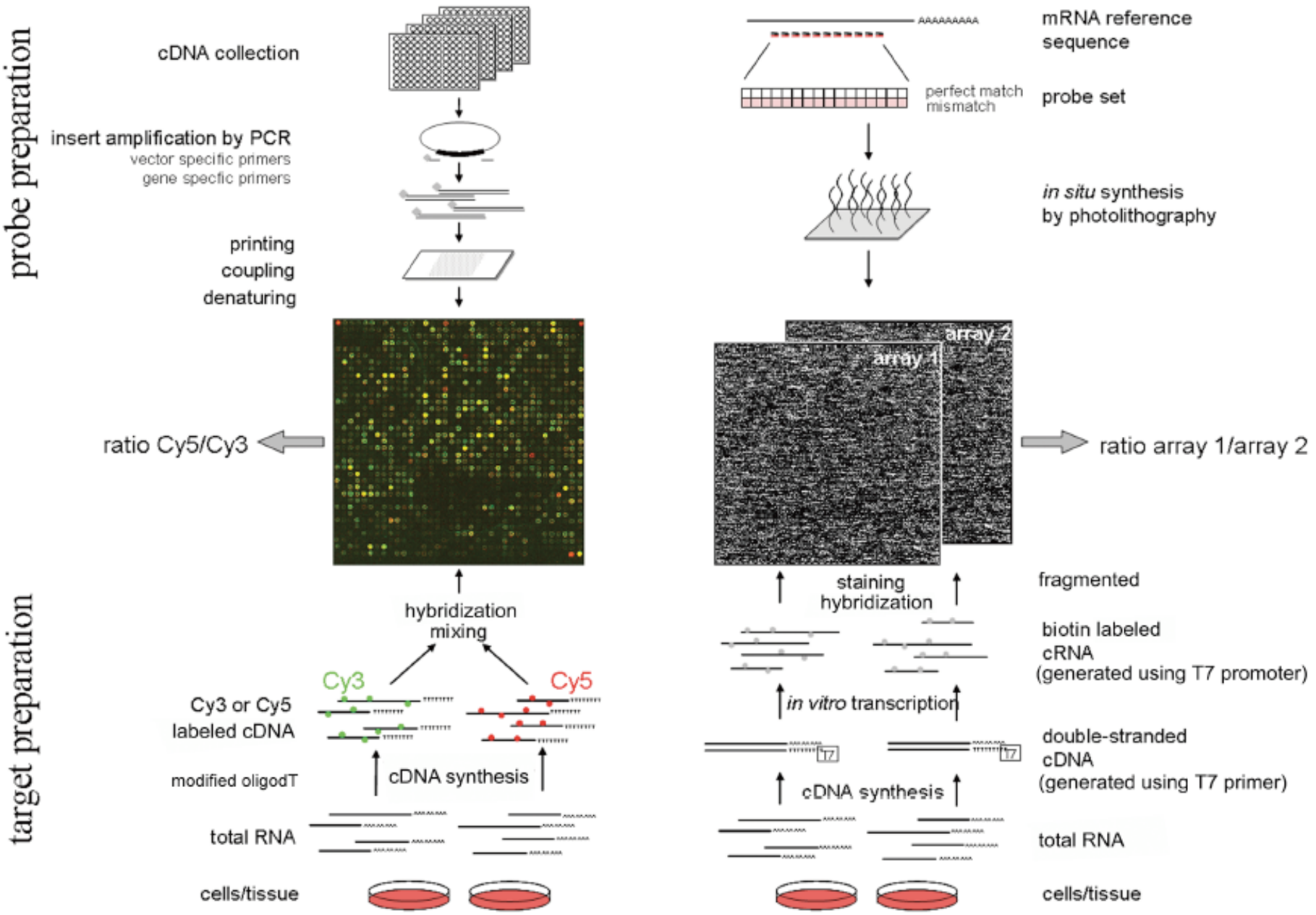
(2) DNA chips

A Probe Set (DNA Chip)



Perfect Match AGGCTATCGCACTCCAGTGG

Mismatch AGGCTATCGTACTCCAGTGG

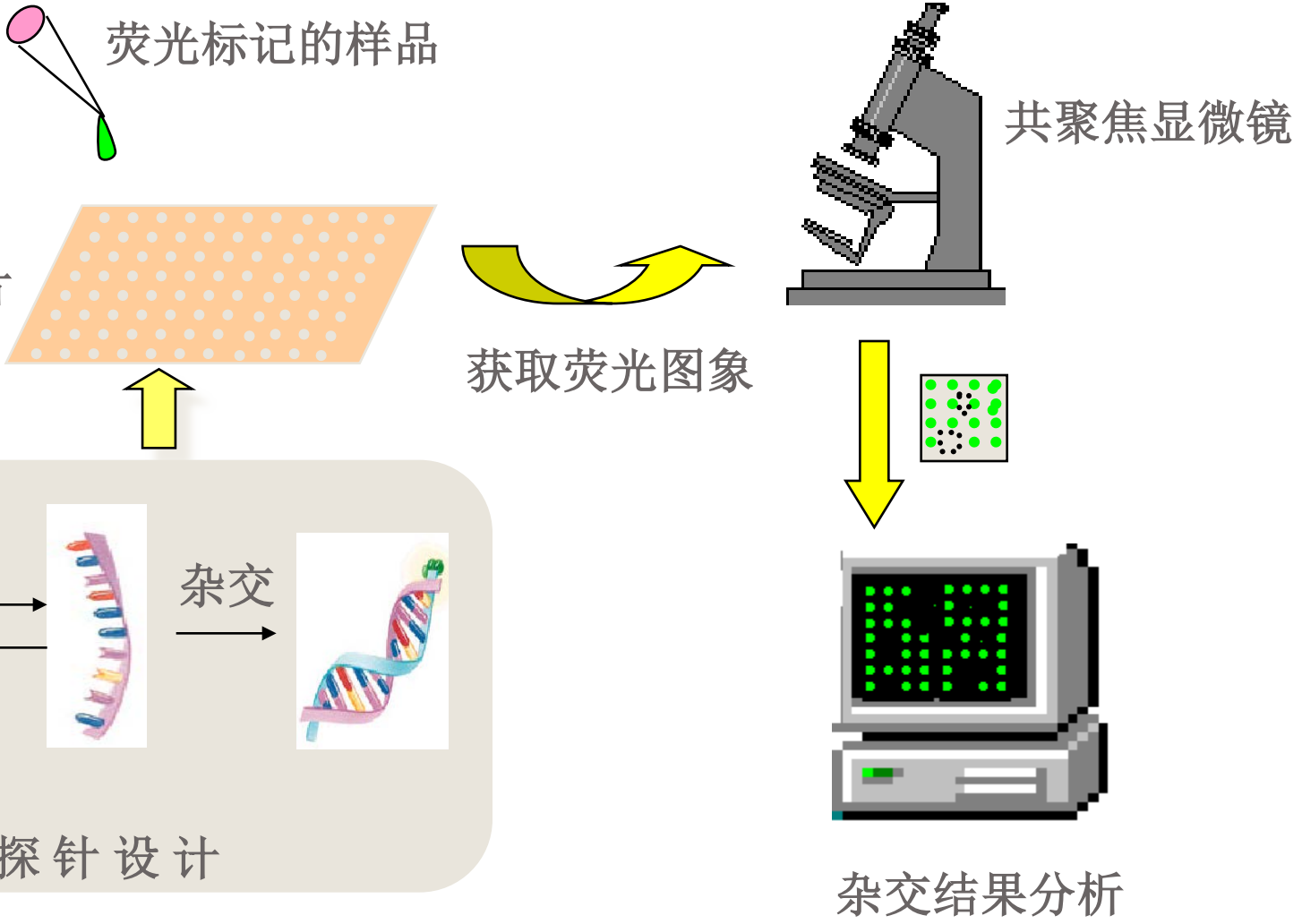

cDNA-microarrays
high-density oligonucleotide arrays
probe preparation
target preparation
cDNA collection
insert amplification by PCR
vector specific primers
gene specfic primers
printing
coupling
denaturing
ratio Cy5/Cy3
hybridization
mixing
Cy3
Cy5
Cy3 or Cy5
labeled cDNA
modified oligodT
cDNA synthesis
total RNA
cells/tissue
AAAAAAAA
mRNA reference
sequence
perfect match
mismatch
probe set
in situ synthesis
by photolithography
array 2
array 1
ratio array 1/array 2
staining
hybridization
fragmented
biotin labeled
cRNA
(generated using T7 promoter)
in vitro transcription
T7
double-stranded
cDNA
(generated using T7 primer)
cDNA synthesis
total RNA
cells/tissue

两种类型芯片

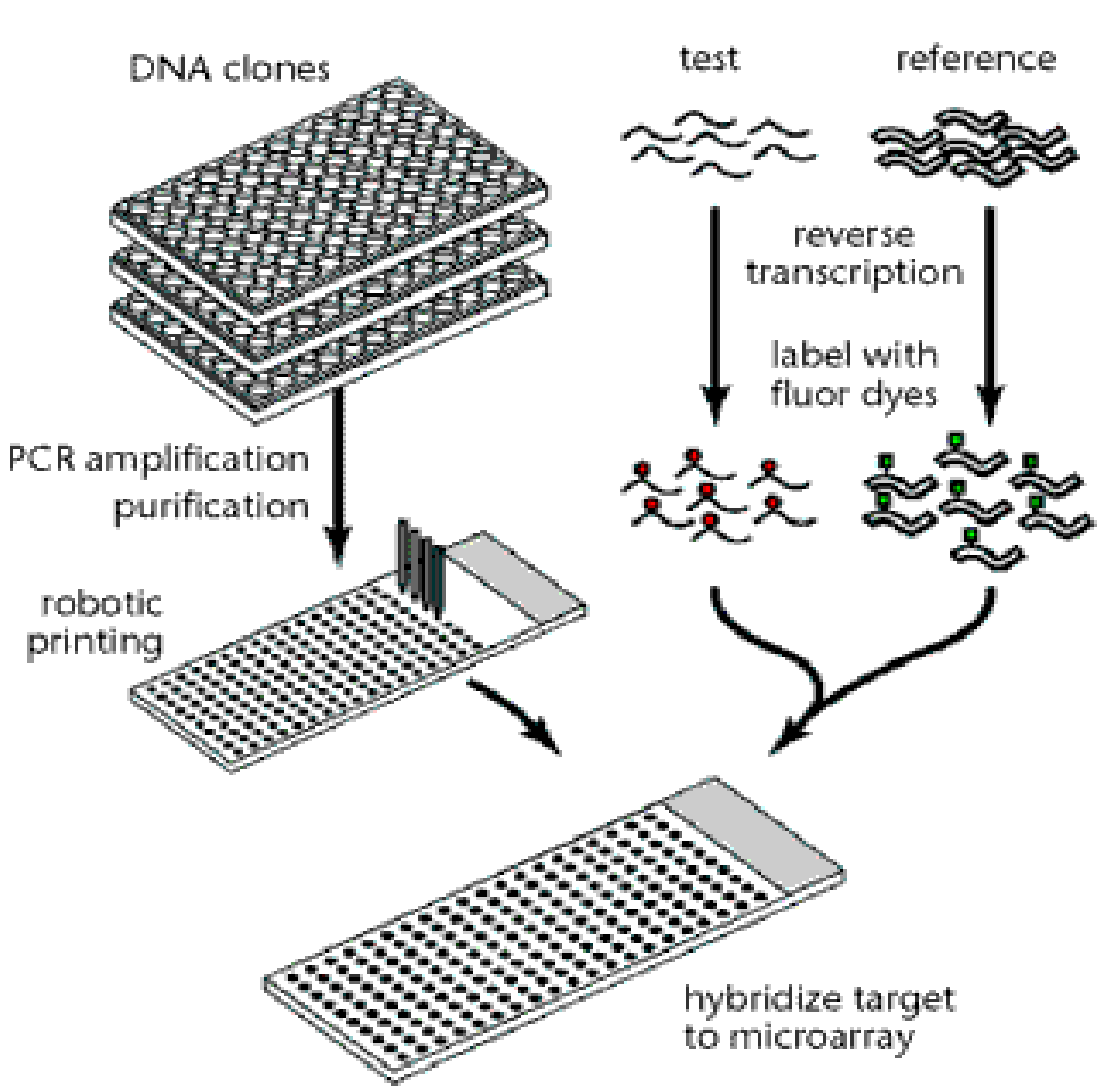
	寡核苷酸芯片	cDNA芯片
通道数	单通道	双通道
探针类型	25nt	cDNA长度
片段长度（全局）	基本都是25nt	基本上是每个基因cDNA的长度
基因重复	很普遍	一般没有
集成度	高	低
探针制作	原位合成	纯化探针，喷墨打印
精确度	高	低
非特异性杂交	可控制	难控制
结果输出	荧光强度	表达变化率
芯片制作	商业，自行合成昂贵	可自行制造

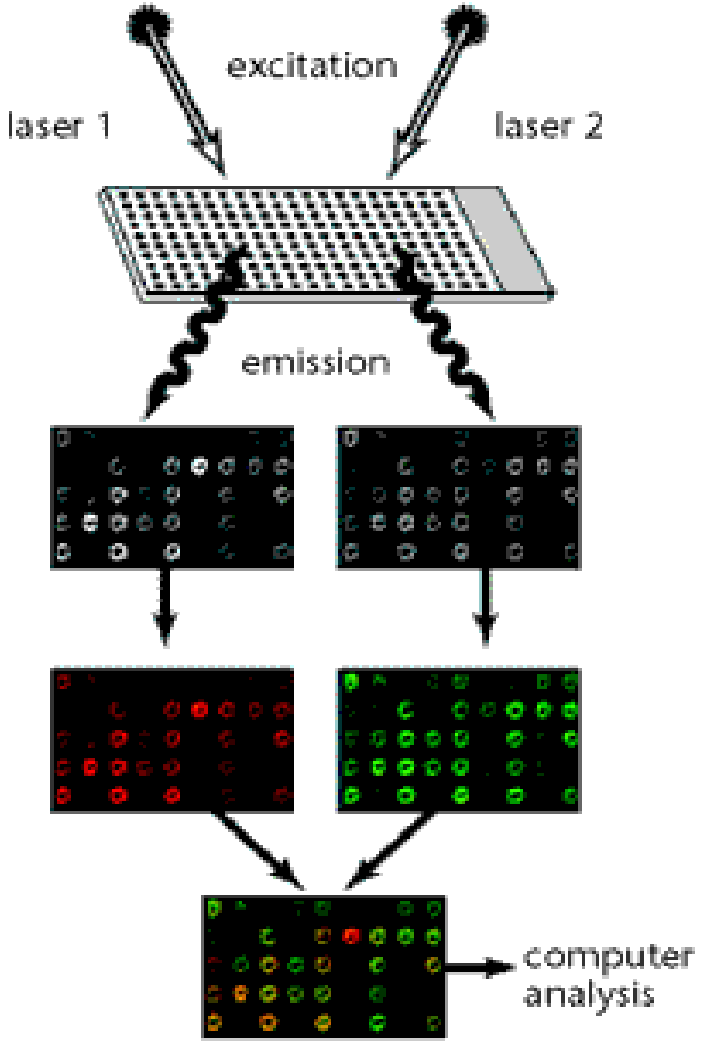
1. 基因芯片简介

- 组织芯片技术是近年来基因芯片(DNA芯片)技术的发展和延伸，与细胞芯片、蛋白芯片、抗体芯片一样，属于一种特殊生物芯片技术。
- 组织芯片技术可以将数十个甚至上千个不同个体的临床组织标本按预先设计的顺序排列在一张玻片进行分析研究，是一种高通量、多样本的分析工具。
- 芯片技术是近年来发展势头非常强劲的一门实验技术，无论在genome水平，还是transcriptome和proteome水平，芯片技术都得到应用。

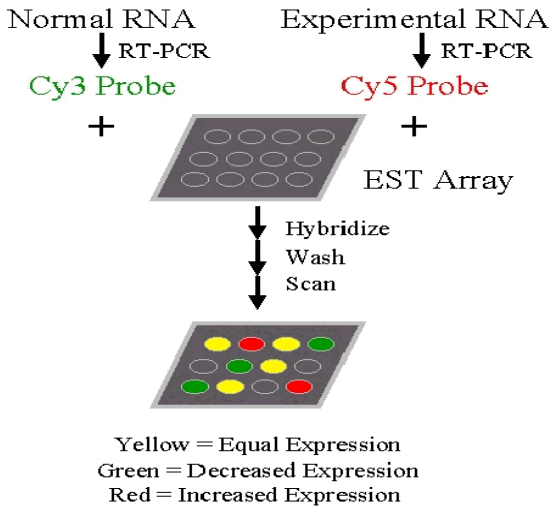

荧光标记的样品
基因芯片
共聚焦显微镜
获取荧光图象
杂交
探 针 设 计
杂交结果分析

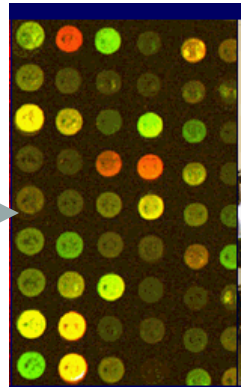
基因芯片原理


DNA clones
PCR amplification
purification
robotic
printing
test
reference
reverse
transcription
label with
fluor dyes
hybridize target
to microarray

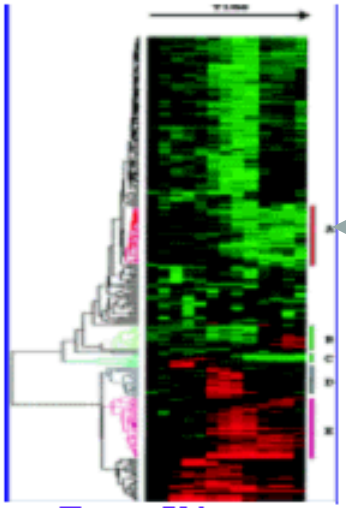

excitation
laser 1
laser 2
emission
computer
analysis

基因芯片的实验流程（双通道）





	A	B	C	D	E	F	G
1	YORF	NAME	GWEIGHT	spo0	spo30	spo2	spo5
2	EWEIGHT			1	1	1	1
3	YAL003W	EFB1	1	0.23	-1.79	-1.29	-1.56
4	YAL004W	YAL004W	1	0.41	-0.38	-0.89	-1.06
5	YAL005C	SSA1	1	0.61	-0.07	-1.29	-1.29
6	YAL010C	MDM10	1	0.16	-0.15	-0.76	-1.25



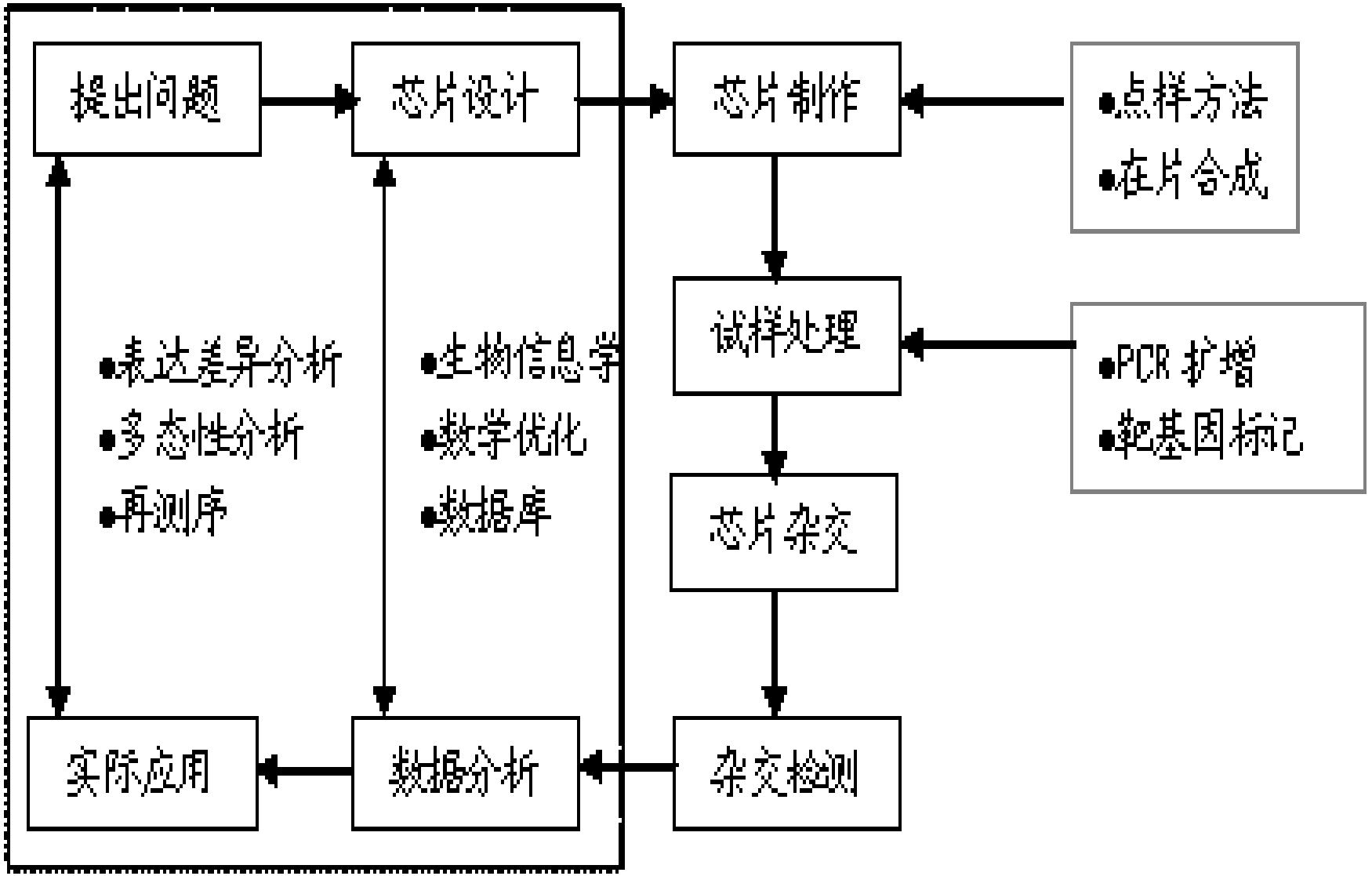


图 8.1 基因芯片相关技术示意图（陆祖宏等，2000）

基因芯片的基本技术

1、芯片方阵的构建：芯片制备是先将玻璃片或硅片进行表面处理，然后使DNA片断或蛋白质分子等生物分子按顺序排列在芯片上的过程。

2、样品的制备：生物样品往往是非常复杂的生物分子混合体，除少数特殊样品外，一般不能直接与芯片反应。可将样品进行处理，获取其中的蛋白质或DNA、RNA，并且加以标记，以提高检测的灵敏度。

3、生物分子反应：生物分子反应为芯片上的生物分子之间的反应，是芯片检测的关键一步。通过选择合适的反应条件使生物分子间反应处于最佳状态中，减少生物分子之间的错配率。

4、信号检测：常用的芯片信号检测方法是将芯片置入芯片扫描仪中，进行信号检测，以获得有关生物学信息。

基因芯片数据的获得

All microarray data and deep sequencing data used in this study have been deposited to GEO

http://www.ncbi.nlm.nih.gov/projects/geo

accession # GSE25899

Benjamin R. Carone.et al. Paternally Induced Transgenerational Environmental Reprogramming ofMetabolic Gene Expression in Mammals. *Cell*. Cell 143, 1094

基因芯片的数据的获得

Retrieve a specific record by entering a valid GEO accession number(GPLxxx,GSMxxx,GSExxx,GDSxxx)

基因芯片的数据的获得

- GEO上有四类数据GSM, GSE, GDS, GPL
- GSM是单个样本的实验数据
- GDS是人工整理好的关于某个话题的GSM的集合，一个GDS中的GSM的平台是一样的。
- GSE是一个实验项目中的多个芯片实验，可能使用多个平台。
- GPL是芯片的平台，如Affymetrix，Aglent等

基因芯片的数据的获得

Supplementary file	Size	Download	File type/resource
GSE25899_RAW.tar	661.3 Mb	(ftp)(http)[custom]	TAR (of CEL, GPR, TXT, BEDGRAPH)

Custom GSE25899_RAW.tar archive:

Supplementary file	File size
GSM635862.CEL.gz	4.0 Mb
GSM635863.CEL.gz	3.9 Mb
GSM635864.CEL.gz	3.9 Mb
GSM640576_s_1_seq.txt.gz	69.1 Mb
GSM640576_uLane1.F.bedGraph.gz	209.2 Kb
GSM640576_uLane1.R.bedGraph.gz	184.0 Kb
GSM640577_s_2_seq.txt.gz	86.9 Mb
GSM640577_uLane2.F.bedGraph.gz	285.6 Kb
GSM640577_uLane2.R.bedGraph.gz	251.1 Kb
GSM640578_s_3_seq.txt.gz	55.5 Mb
GSM640578_uLane3a.F.bedGraph.gz	246.5 Kb
GSM640578_uLane3a.R.bedGraph.gz	217.2 Kb
GSM640579_s_4_seq.txt.gz	62.1 Mb
GSM640579_uLane4a.F.bedGraph.gz	257.3 Kb
GSM640579_uLane4a.R.bedGraph.gz	228.7 Kb
GSM640580_s_5_seq.txt.gz	60.0 Mb
GSM640580_uLane5a.F.bedGraph.gz	264.6 Kb
GSM640580_uLane5a.R.bedGraph.gz	228.4 Kb

芯片设计的一般性原则

(1)互补性

(2)敏感性和特异性

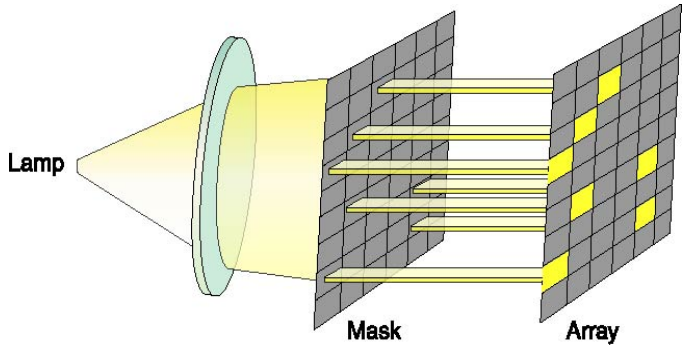
(3)容错性：采用冗余探针

(4)可靠性

(5)可控性：设置质量控制探针

(6)可读性：通过探针布局,使杂交信号便于观察理解

DNA chips的制备：Affymetrix photolitography



- 探针长度：25 bp
- 每个基因：22-40个探针
- Perfect Match (PM) vs. MisMatch (MM) probes

A. 选择硅片、玻璃片、瓷片或聚丙烯膜、尼龙膜等支持物

B. 采用光导化学合成和照相平板印刷技术在硅片等表面合成寡核苷酸探针; 或者通过液相化学合成寡核苷酸链探针，或**PCR**技术扩增基因序列，由阵列复制器（arraying and replicating device ARD），或阵列机（arrayer）及电脑控制的机器人，将不同探针样品定量点样于带正电荷的尼龙膜或硅片等相应位置上

C. 紫外线交联固定后即得到**DNA**微阵列或芯片

芯片制作

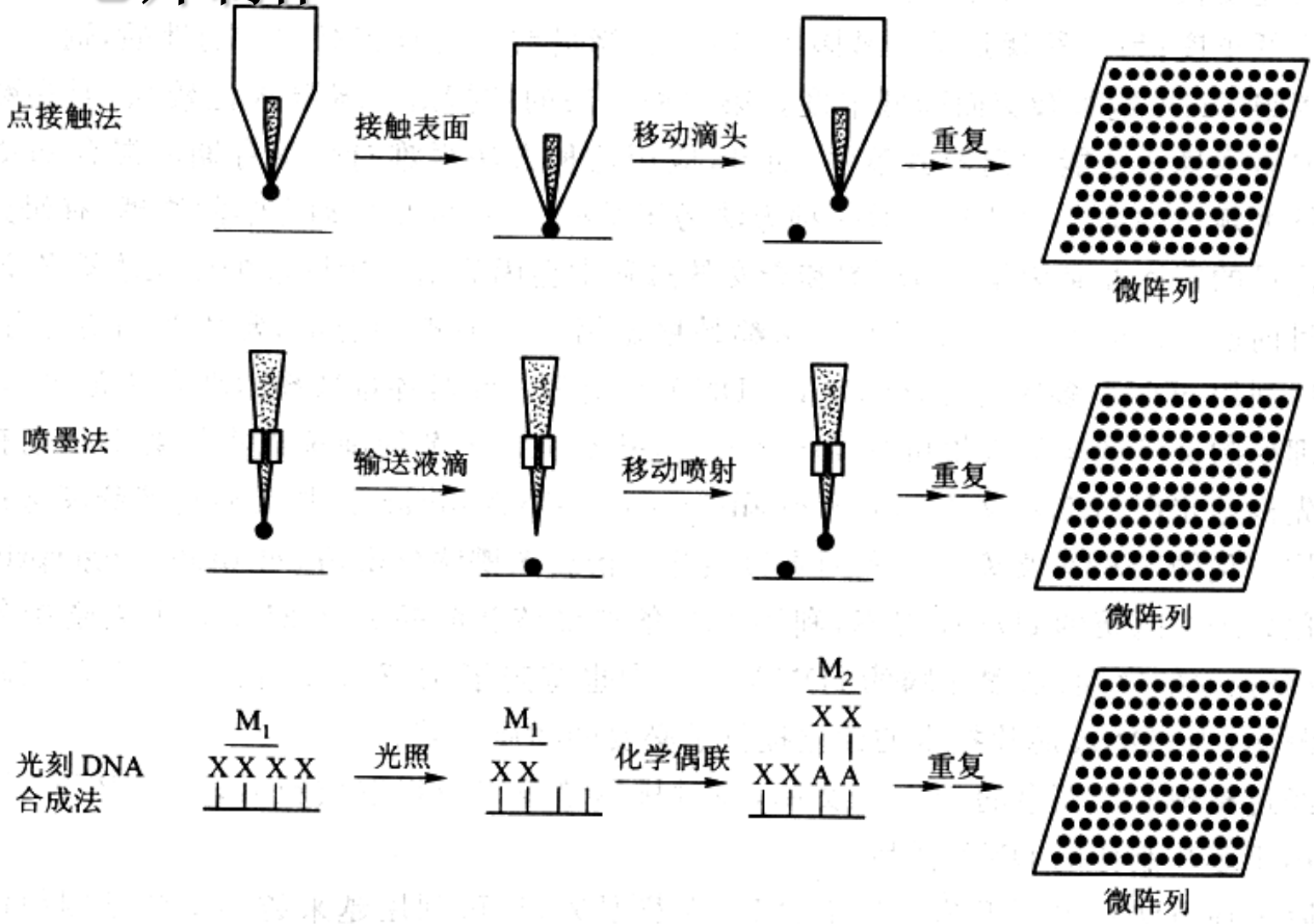


图 7-2　3 种制作生物芯片的方法　图中 M 表示照相掩蔽网，通过 M 来确定光脱保护的位置；X 表示光不稳定的保护基团；A 表示核苷酸。

样品准备

样品的分离纯化：

DNA，mRNA

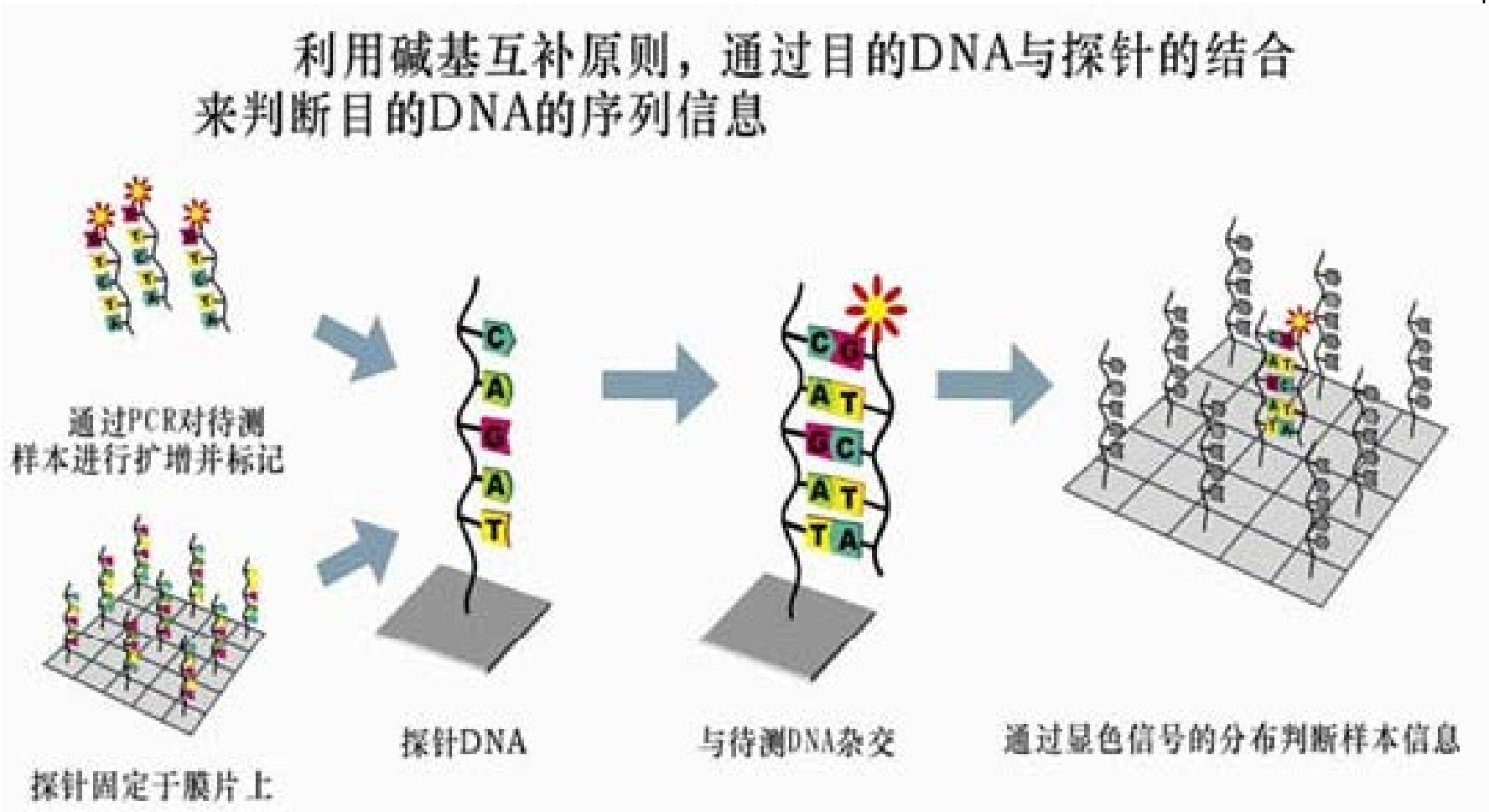
扩增：

PCR，RT—PCR，固相PCR

样品的标记：

荧光标记（常用Cy3、Cy5）

分子杂交


利用碱基互补原则，通过目的DNA与探针的结合
来判断目的DNA的序列信息
通过PCR对待测
样本进行扩增并标记
探针固定于膜片上
探针DNA
与待测DNA杂交
通过显色信号的分布判断样本信息

信号检测与结果分析


ScanArray® 3000
Biochip Scanner
Capture the
Full Potential of
Microarray Biochips

计算机“读片”机理

将样品中的DNA/RNA标上荧光标记，则可以定量检验基因的表达水平。

cDNA芯片、载有较长片段的寡核苷酸芯片采用双色荧光系统：目前常用 Cy3—dUTP（绿色）标记对照组mRNA，Cy5—dUTP（红色）标记样品组mRNA

用不同波长的荧光扫描芯片，将扫描所得每一点荧光信号值自动输入计算机并进行信息处理，给出每个点在不同波长下的荧光强度值及其比值，同时计算机还给出直观的显色图。

在样品中呈高表达的基因其杂交点呈红色，相反，在对照组中高表达的基因其杂交点呈绿色，在两组中表达水平相当的显黄色，这些信号就代表了样品中基因的转录表达情况。

基因芯片数据分析

1. 基因芯片(Microarray)简介

2. 图像处理与数据标准化

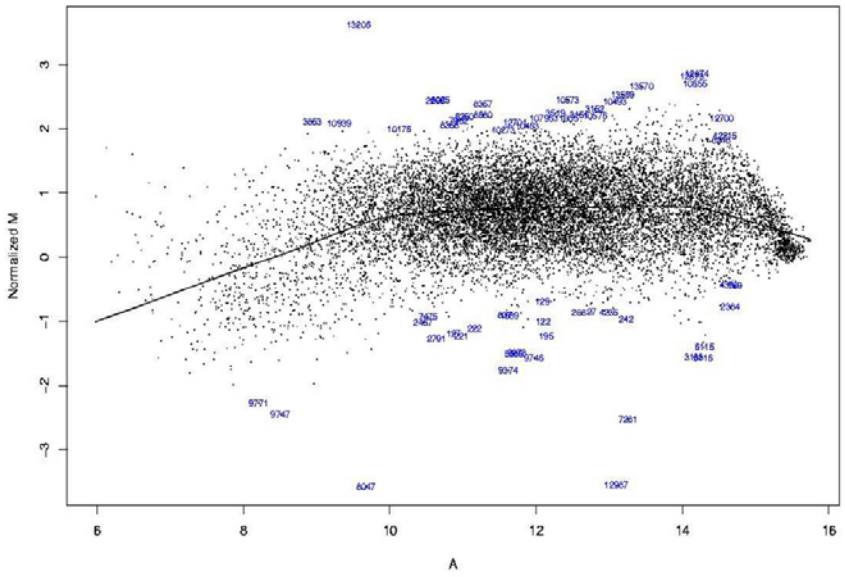
3. 基因芯片的数据分析

2. 图像处理与数据标准化

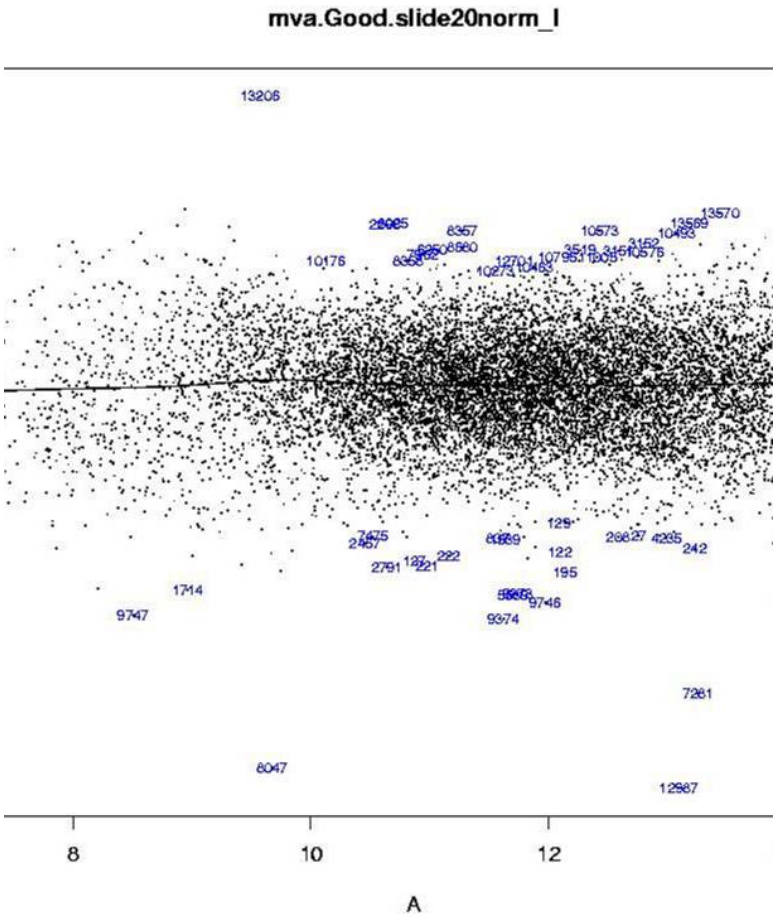
- 在基因芯片实验中,基因表达水平之间的相关性在推断基因间相互关系时起到非常重要的作用.
- 未经标准化处理的芯片数据基因之间往往都呈现出很强的相关性,这些高相关性一部分是由基因表达水平变化引起的,而另外一部分是由系统偏差引起的.
- 对芯片数据进行标准化处理的目的之一是消除系统偏差引起的高相关性,同时保留由真正生物学原因引起的基因表达水平高相关性

数据标准化

before

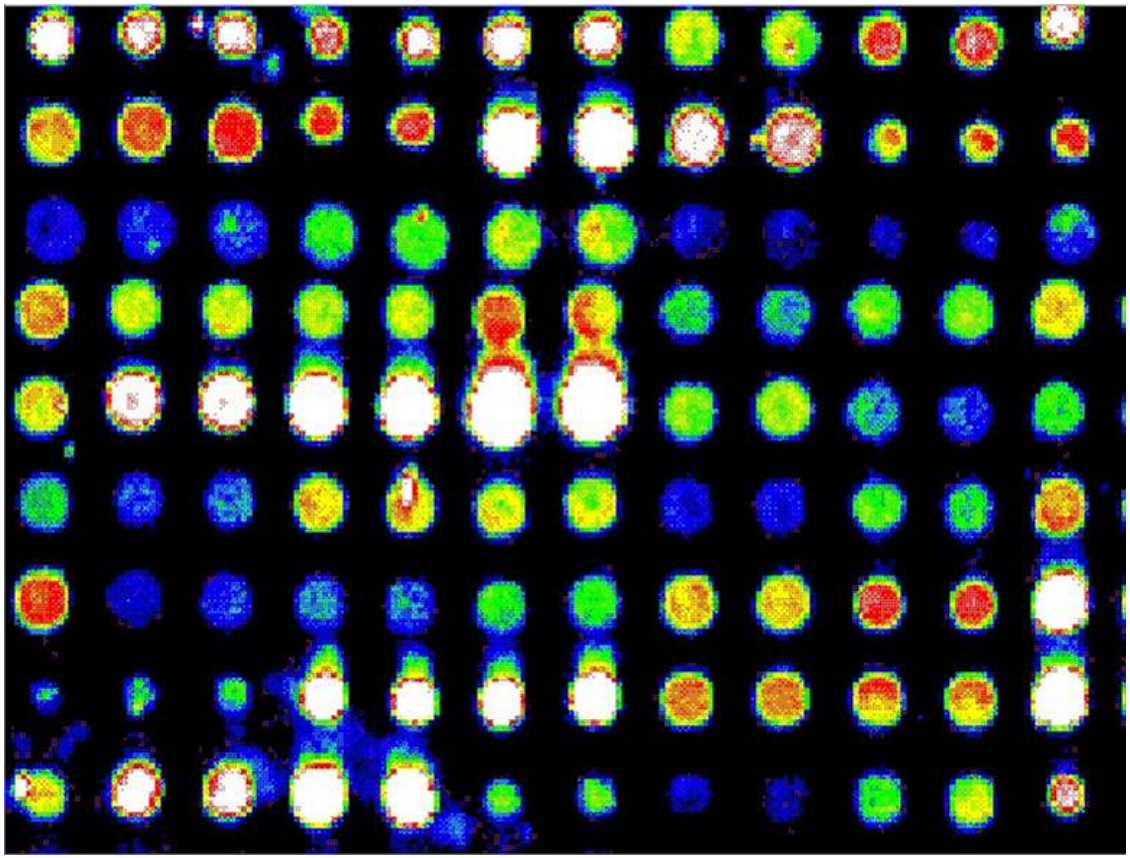


after



目的是消除系统偏差引起的高相关性，同时保留由真正生物学原因引起的基因表达水平高相关性。

数据标准化



单通道基因芯片

white (very high)

red (high)

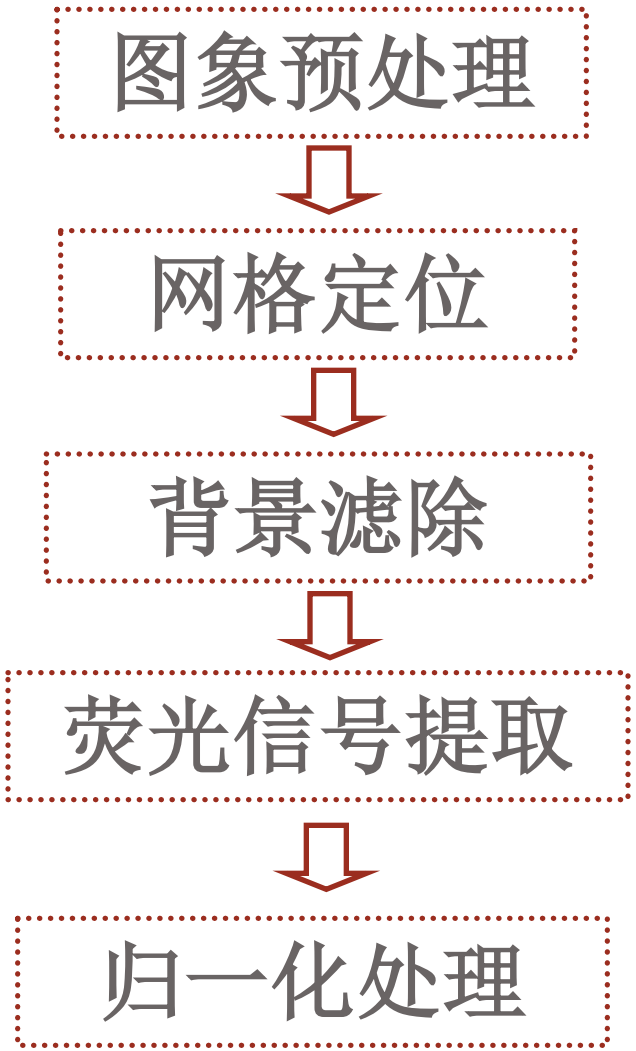
Yellow (a little high)

green (medium)

blue (low)

black (no)

数据标准化



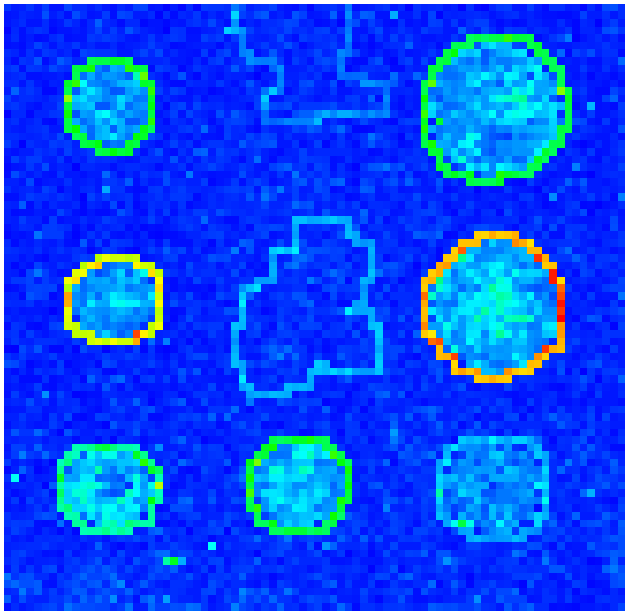
图像处理

栅格化：确定点的位置

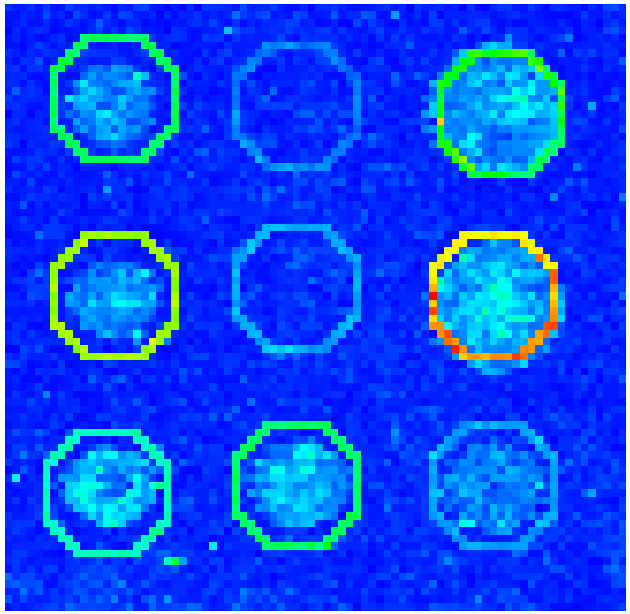
图象分割 (Segmentation)：将点从背景中分离出来。

抽提亮度：各个像素亮度的平均值 (mean)或中位数 (median)

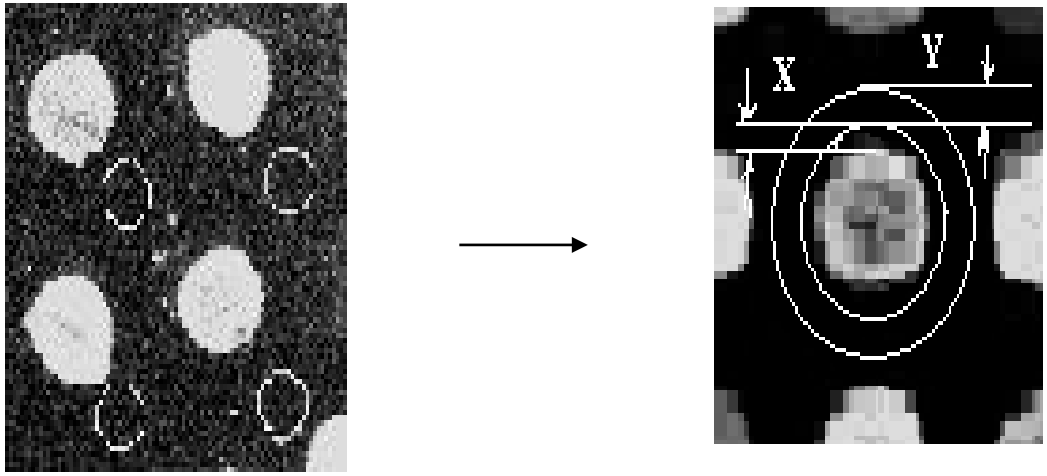
背景校正：局部或全局



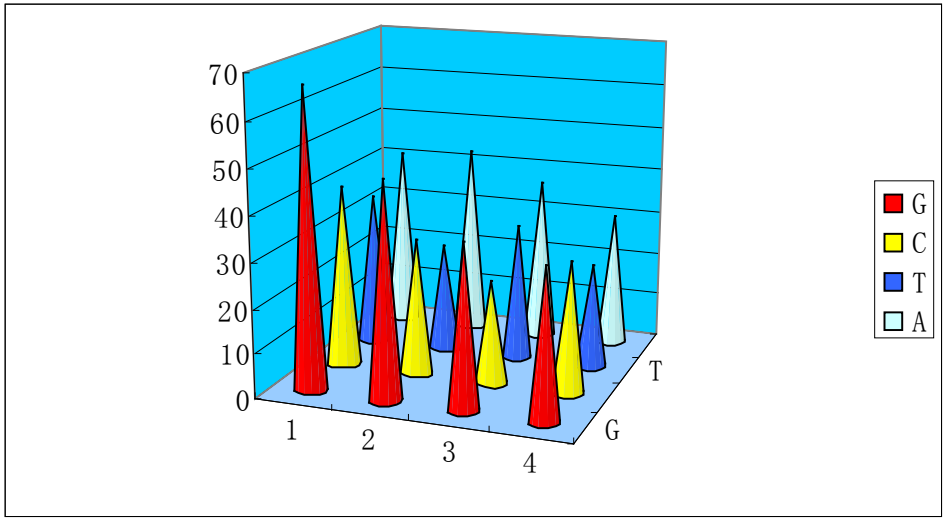
植根区域生长法(SRG)



Fixed Circle



背景滤除



荧光信号提取

基因表达量的定量

对于每个点，可以计算

Red intensity $= R_{fg} - R_{bg}$

fg = foreground, bg = background, and

Green intensity $= G_{fg} - G_{bg}$

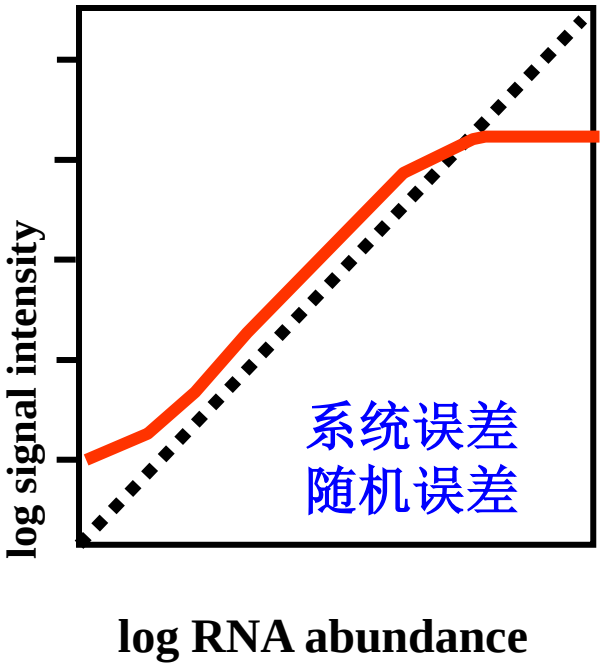
and combine them in the log (base 2) ratio

Log_2(Red intensity / Green intensity)

Green intensity (medium): ~1

Microarray: 误差的来源

1. 图像分析
2. 扫描
3. DNA杂交过程 (温度、时间、混合均匀程度等)
4. 探针的标记
5. RNA的抽提
6. 加样
7. 其他



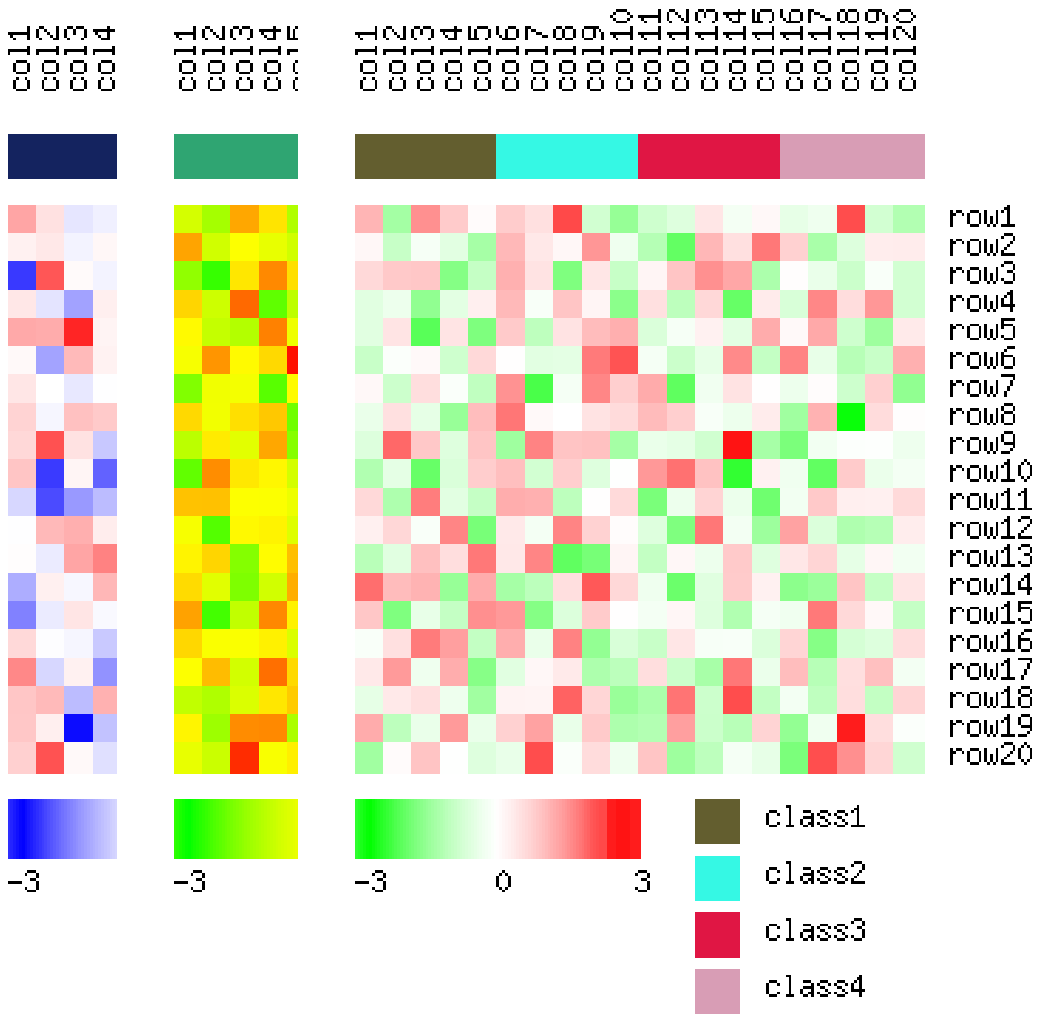
基因芯片数据分析

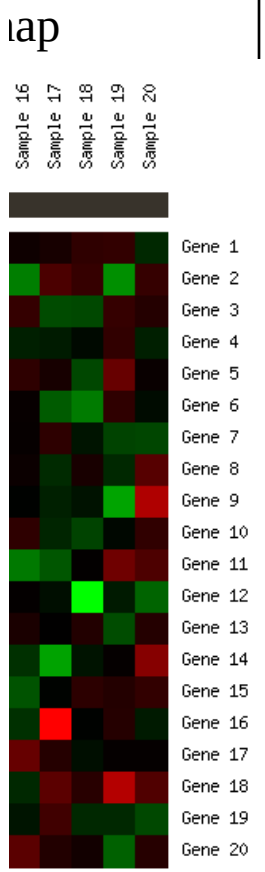
数据结构（表达

	Array
Gene1	0.5
Gene2	-0.2
Gene3	2.(
Gene3	1.1
Gene5	1.5
Gene6	2.4

用到的术语

1.Treatment
的荧光值
2.Control: 对
3.Ratio: trea
4.表达值: lo





基因芯片数据分析

1. 基因芯片(Microarray)简介

2. 图像处理与数据标准化

3. 基因芯片的数据分析

3. 基因芯片的数据分析

(1) 差异表达基因的分析

(2) 基因表达数据的分类

(3) Map to GO

(4) Gene regulatory network

(1) 差异表达基因的分析

- 经过预处理，探针水平数据转变为基因表达数据。为了便于应用一些统计和数学术语，基因表达数据仍采用矩阵形式。
- 差异表达基因的分析: 寻找处理前后表达上调或者下调的基因
- Are the treatments different?
- 使用标准的统计学方法检验 (t-test or f-test)，发现统计显著性差异表达的基因，如果处理本身并不显著，则结果无意义

(1) 差异表达基因的分析

倍数分析方法：倍数变换fold change，单纯的case与control组表达值相比较，对没有重复实验样本的芯片数据，或者双通道数据采用这种方法（该方法是对基因芯片的ratio值从大到小排序，即cy5/cy3比值，一般0.5-2.0之间内的基因不存在差异表达，范围之外存在差异表达。缺点是倍数选取具有任意性，可能不恰当）。

(1) 差异表达基因的分析

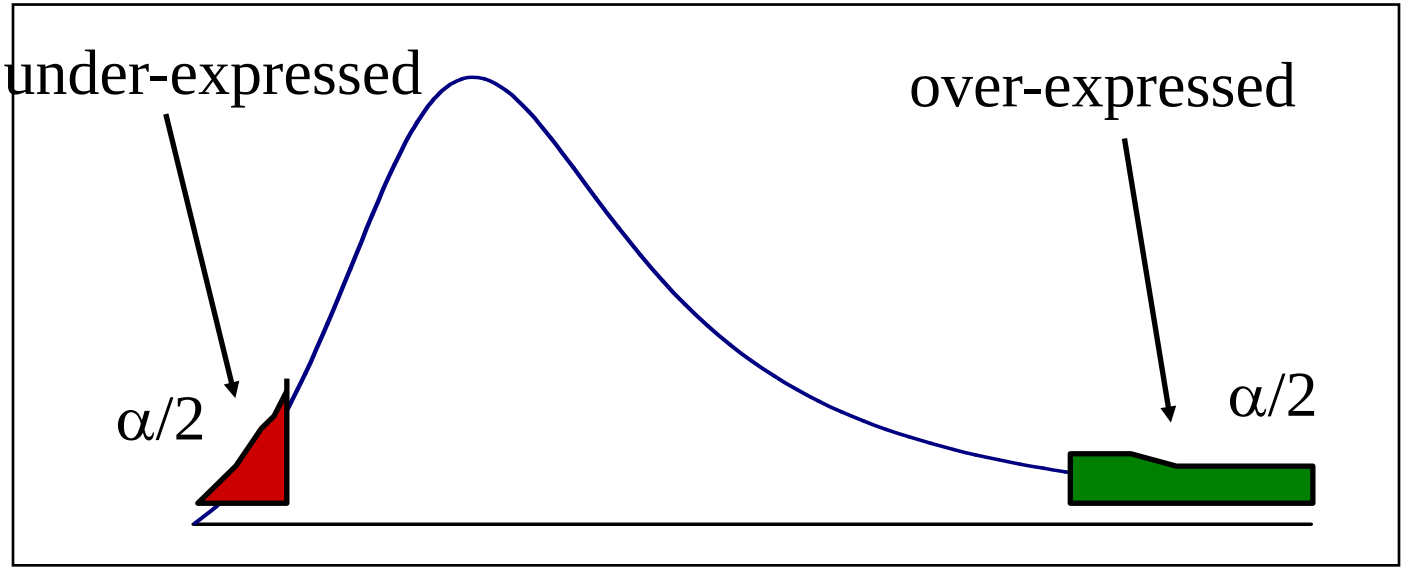
参数法分析（t检验）：当t超过根据可信度选择的标准时, 比较的两样本被认为存在着差异。但小样本基因芯片实验会导致不可信的变异估计，此时采用调节性T检验 。

非参数分析：由于微阵列数据存在“噪声”干扰而且不满足正态分布假设，用t检验有风险。非参数检验并不要求数据满足特殊分布的假设，所以可使用非参数方法对变量进行筛选。

如经验贝叶斯法、芯片显著性分析SAM法。常用的利用R的limma包使用t检验筛选差异表达基因，利用R的siggenes包使用SAM方法筛选差异表达基因

统计学分析

- Fold change, 一般2-fold increase or decrease (平行实验的样本较少)
- p-value (平行实验的样本较多)



False Discovery Rate (FDR)

- 在基因芯片的实验中，每一个基因/探针，都是一个独立的实验。
- 基因芯片：高通量，>1,000个基因/探针。
- 因此，无论怎么比较，总会有一些基因会是统计显著性差异表的 —— 可能是随机产生的。
- 如何评估表达差异基因预测的有效性？
- FDR = p-value * No. of Genes

例：1,000个探针的双通道芯片，以p-value < 0.01为域值，发现7个上调基因，5个下调基因，分析结果是否具有统计学意义？计算： FDR= 0.01* 1,000=10 (随机) 。7个上调基因，5个下调基因 < 10，因此上例计算的结果无统计学意义。

FDR必须远小于发现的差异表达基因数目。

(2) 基因表达数据的分类

- 根据基因表达的数据将样本分成两类或多类
- 督导学习 (supervised learning)：根据发现的模式进行预测
- 应用：
  - 癌症 vs. 正常组织
  - 癌症的亚型、不同阶段 (良性的 vs. 恶性的)
  - 对药物的敏感性 (tamoxifen for breast cancer)
- 基因芯片用于基因诊断：

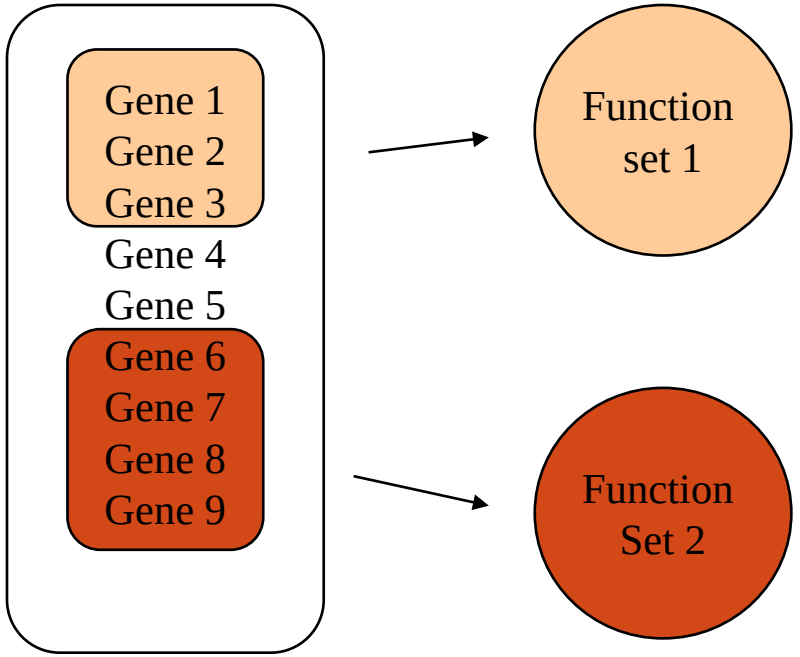
  通过基因芯片诊断地中海贫血患者的血液细胞发现在B-珠蛋白基因中存在3个明确的突变位点

(3) Map to GO

- 通过基因芯片，找到了一批“interesting”的基因
- 生物学功能上是否存在关联？
- 基因本体（Gene Ontology, GO）：GO数据库把基因的功能分为三类：分子功能，生物学过程和细胞组分。在每一个分类中，都提供一个描述功能信息的分级结构。

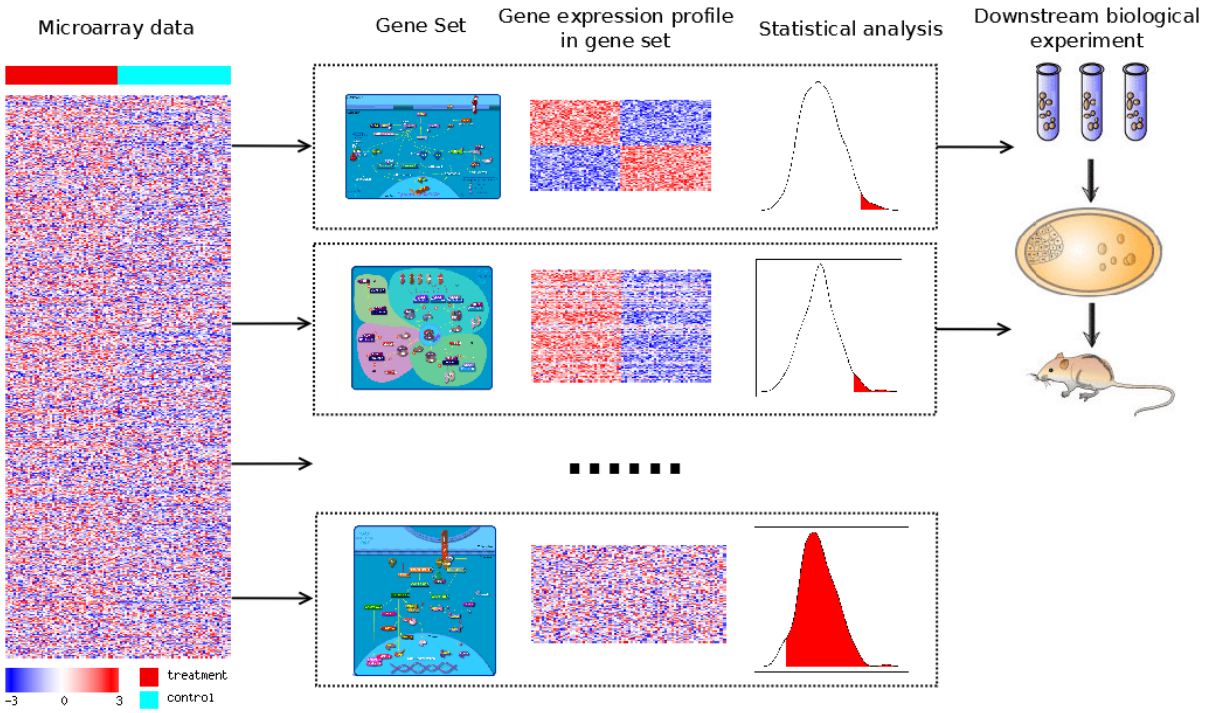
基因集合分析

- 基因集合是一系列具有相似生物学属性的基因
- Gene ontology
- Pathways
- Chromosome locus
- Regulatory motif
- Cancer related
- Tissue specific
- Network modules
- Cluster
- …



基因集合分析的优点

- 基因不是单独起作用的
- 有助于更好的理解基因的功能
- 也有助于理解关键基因是如何影响关键通路的
- 找到显著的功能，可以缩小寻找关键基因的范围



研究者可以通过GO分类号将分类与具体基因联系起来，从而对基因的功能进行描述。

在芯片的数据分析中，研究者可以找出哪些变化基因属于一个共同的GO功能分支，并用统计学方法检定结果是否具有统计学意义，从而得出变化基因主要参与了哪些生物功能。

- 比较著名的基于GO分类法的芯片数据分析网络平台有七十多个：

Name	Internet Site
Onto-Tools	http://vortex.cs.wayne.edu/projects.htm
GOToolBox	http://burgundy.cmmt.ubc.ca/GOToolBox/
GOstat	http://gostat.wehi.edu.au/
GFINDer	http://www.medinfopoli.polimi.it/GFINDer/
EASE	http://david.abcc.ncifcrf.gov/ease/ease.jsp

基因集合分析 – Gene Ontology数据库

Term Lineage

Switch to viewing term parents, siblings and children

Filter tree view

Filter Gene Product Counts

Data source: All, AspGD, CGD, dictyBase

Species: All, Anaplasma phagocy..., Arabidopsis thaliana, Bacillus anthraci...

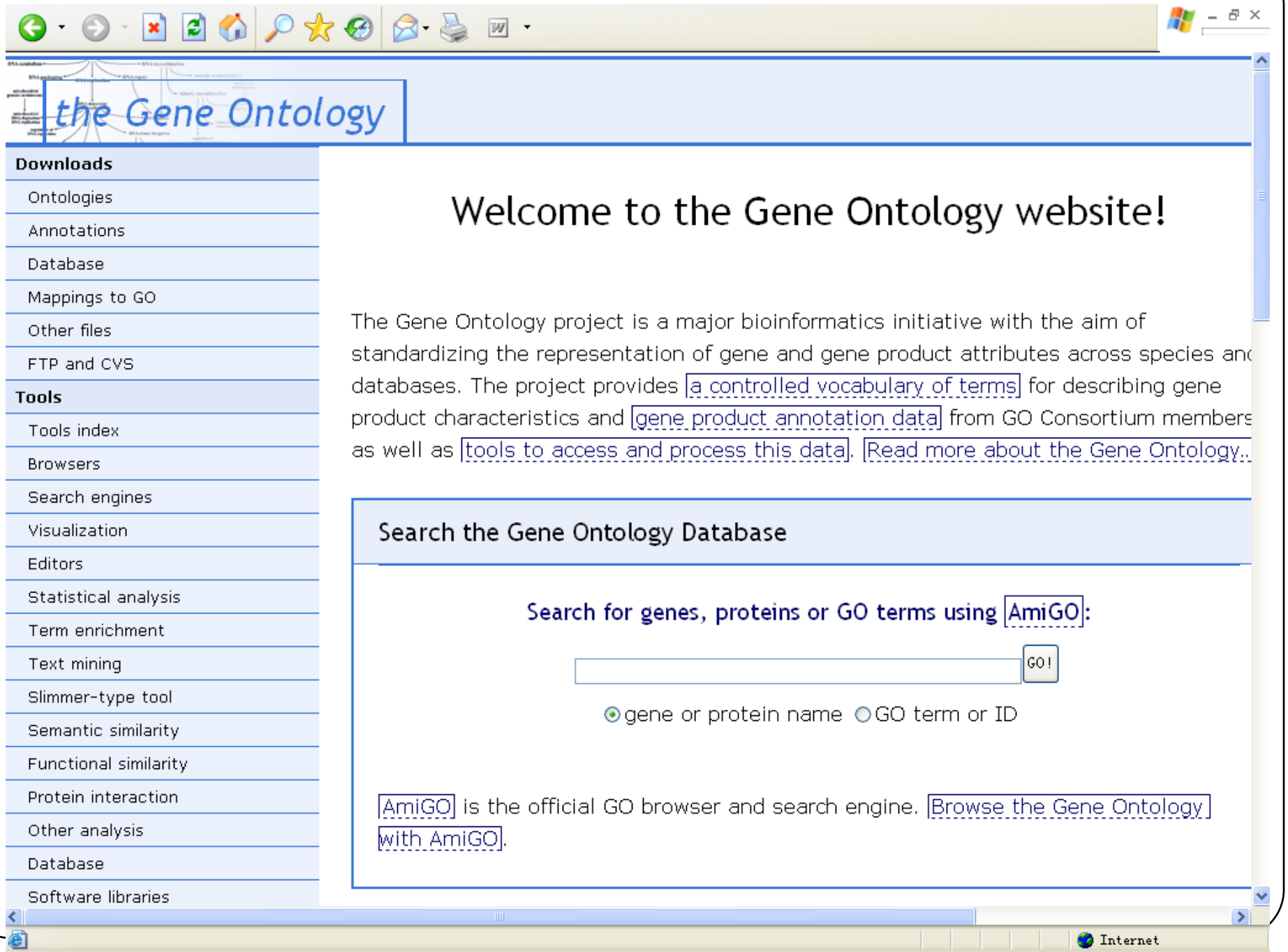
View Options

Tree view: Full / Compact

Set filters

Remove all filters

- all : all [377382 gene products]
  - GO:0008150 : biological_process [270820 gene products]
    - GO:0050896 : response to stimulus [30457 gene products]
      - GO:0009605 : response to external stimulus [5585 gene products]
        - GO:0009611 : response to wounding [2289 gene products]
          - GO:0006954 : inflammatory response [1173 gene products]
            - GO:0002526 : acute inflammatory response [427 gene products]
              - **GO:0002532 : production of molecular mediator of acute inflammatory response** [44 gene products]
      - GO:0006950 : response to stress [16147 gene products]
        - GO:0006952 : defense response [4501 gene products]
          - GO:0006954 : inflammatory response [1173 gene products]
            - GO:0002526 : acute inflammatory response [427 gene products]
              - **GO:0002532 : production of molecular mediator of acute inflammatory response** [44 gene products]
        - GO:0009611 : response to wounding [2289 gene products]
          - GO:0006954 : inflammatory response [1173 gene products]
            - GO:0002526 : acute inflammatory response [427 gene products]
              - **GO:0002532 : production of molecular mediator of acute inflammatory response** [44 gene products]

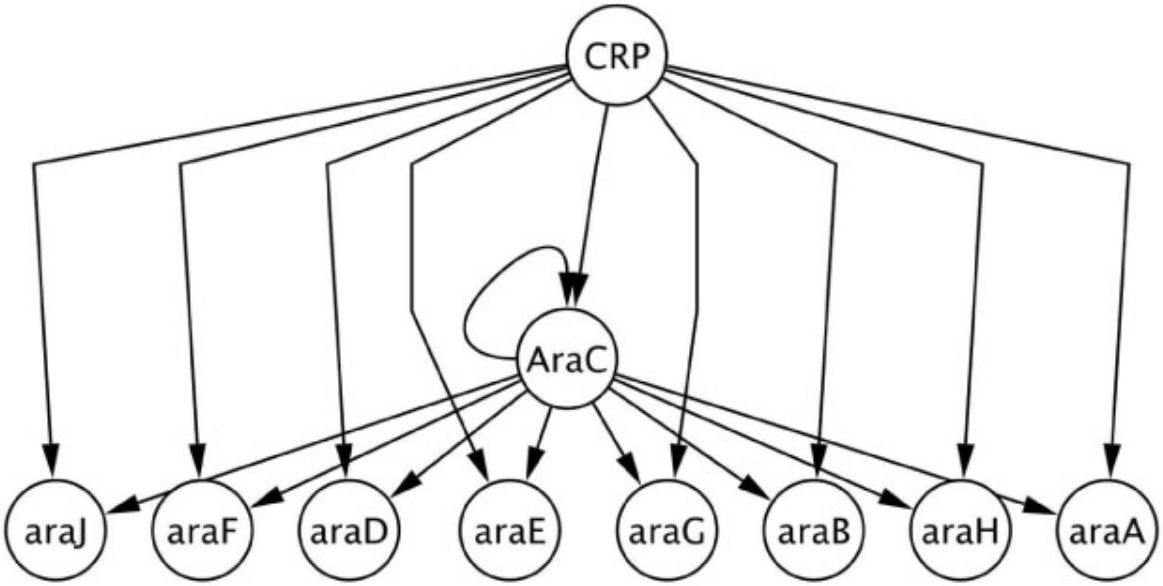

the Gene Ontology
Downloads
Ontologies
Annotations
Database
Mappings to GO
Other files
FTP and CVS
Tools
Tools index
Browsers
Search engines
Visualization
Editors
Statistical analysis
Term enrichment
Text mining
Slimmer-type tool
Semantic similarity
Functional similarity
Protein interaction
Other analysis
Database
Software libraries
Welcome to the Gene Ontology website!
The Gene Ontology project is a major bioinformatics initiative with the aim of standardizing the representation of gene and gene product attributes across species and databases. The project provides a controlled vocabulary of terms for describing gene product characteristics and gene product annotation data from GO Consortium members as well as tools to access and process this data. Read more about the Gene Ontology..
Search the Gene Ontology Database
Search for genes, proteins or GO terms using AmiGO:
GO!
gene or protein name
GO term or ID
AmiGO is the official GO browser and search engine. Browse the Gene Ontology with AmiGO.
Internet

(4) Gene regulatory network

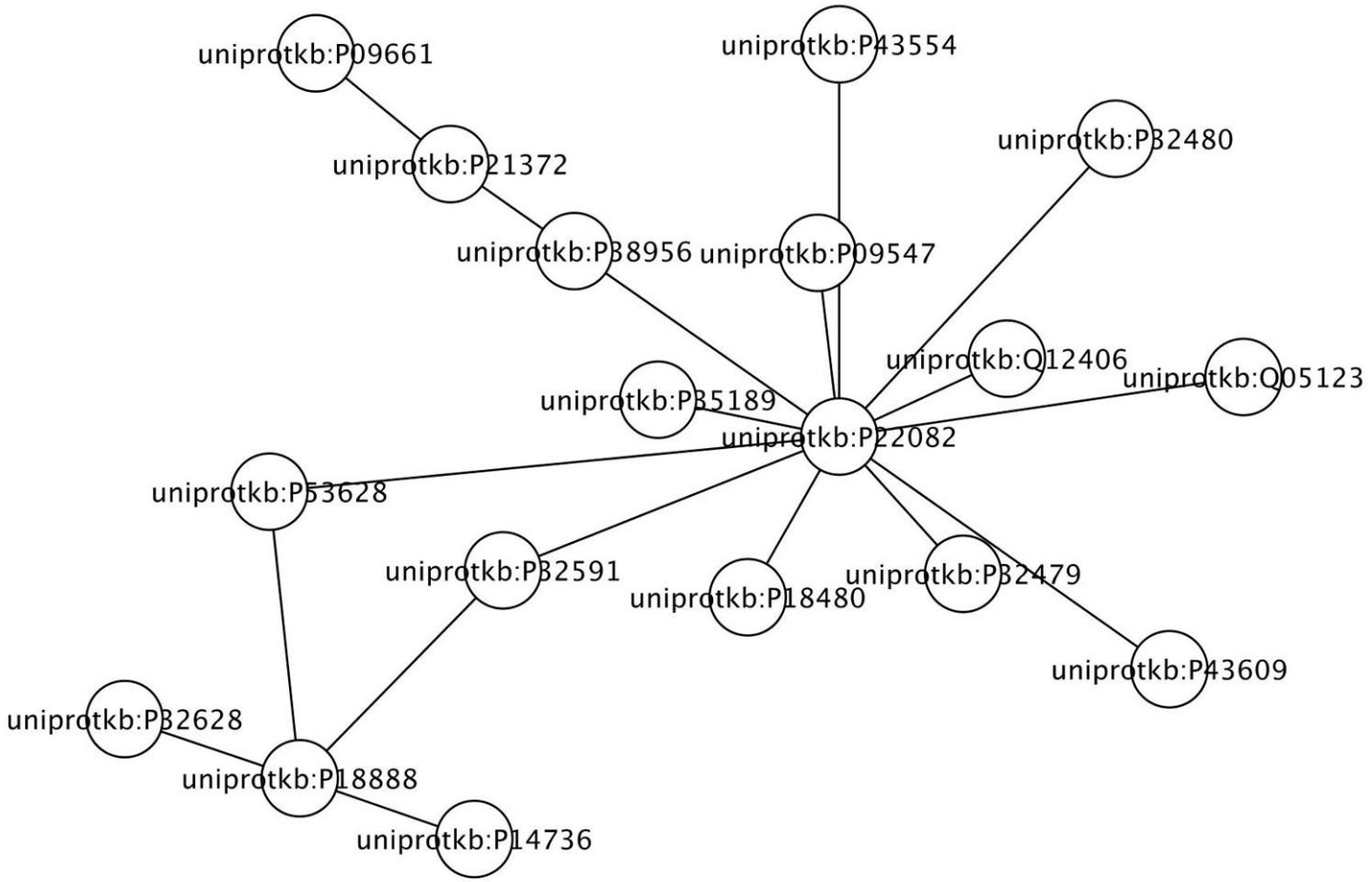
- 早期观点：表达谱相似的基因可能存在功能上的关联，可能有相互作用... (直接作用)。
- 当前的观点：表达谱相似的基因可能具有共同的调控元件 (基因UTR区域存在共同的Promotor), 能够被同一个上游因子所调控。

基因转录调控网络

基因转录调控网络是以转录因子和受调控基因作为节点，以调控关系作为边的有向网络。



蛋白质互作网络


uniprotkb:P09661
uniprotkb:P43554
uniprotkb:P32480
uniprotkb:P21372
uniprotkb:P38956
uniprotkb:P09547
uniprotkb:Q12406
uniprotkb:Q05123
uniprotkb:P35189
uniprotkb:P22082
uniprotkb:P53628
uniprotkb:P32591
uniprotkb:P18480
uniprotkb:P32479
uniprotkb:P43609
uniprotkb:P32628
uniprotkb:P18888
uniprotkb:P14736

代谢网络和信号传导网络

代谢通路 是指细胞中代谢物在酶的作用下转化为新的代谢物过程中所发生的一系列生物化学反应。

代谢网络 是指由代谢反应以及调节这些反应的调控机制所组成的描述细胞内代谢和生理过程的网络。

信号传导 是指细胞将一种类型的生物信号或刺激转换为其它生物信号最终激活细胞反应的过程。

信号传导网络 是指参与信号传导通路的分子和酶以及其间所发生的生化反应所构成的网络。

急需解决的问题

1. 生物芯片的重复利用。
2. 生物芯片的多重用途。
3. 统一的行业标准。
4. 定量分析。
5. 降低检测生物芯片的仪器的价格。