生物芯片和高通量测序技术及其应用

肖华胜博士

生物芯片上海国家工程研究中心
上海伯豪生物技术有限公司

Email：huasheng_xiao@shbio.com

主要内容

第一部分　概述

第二部分　基因芯片

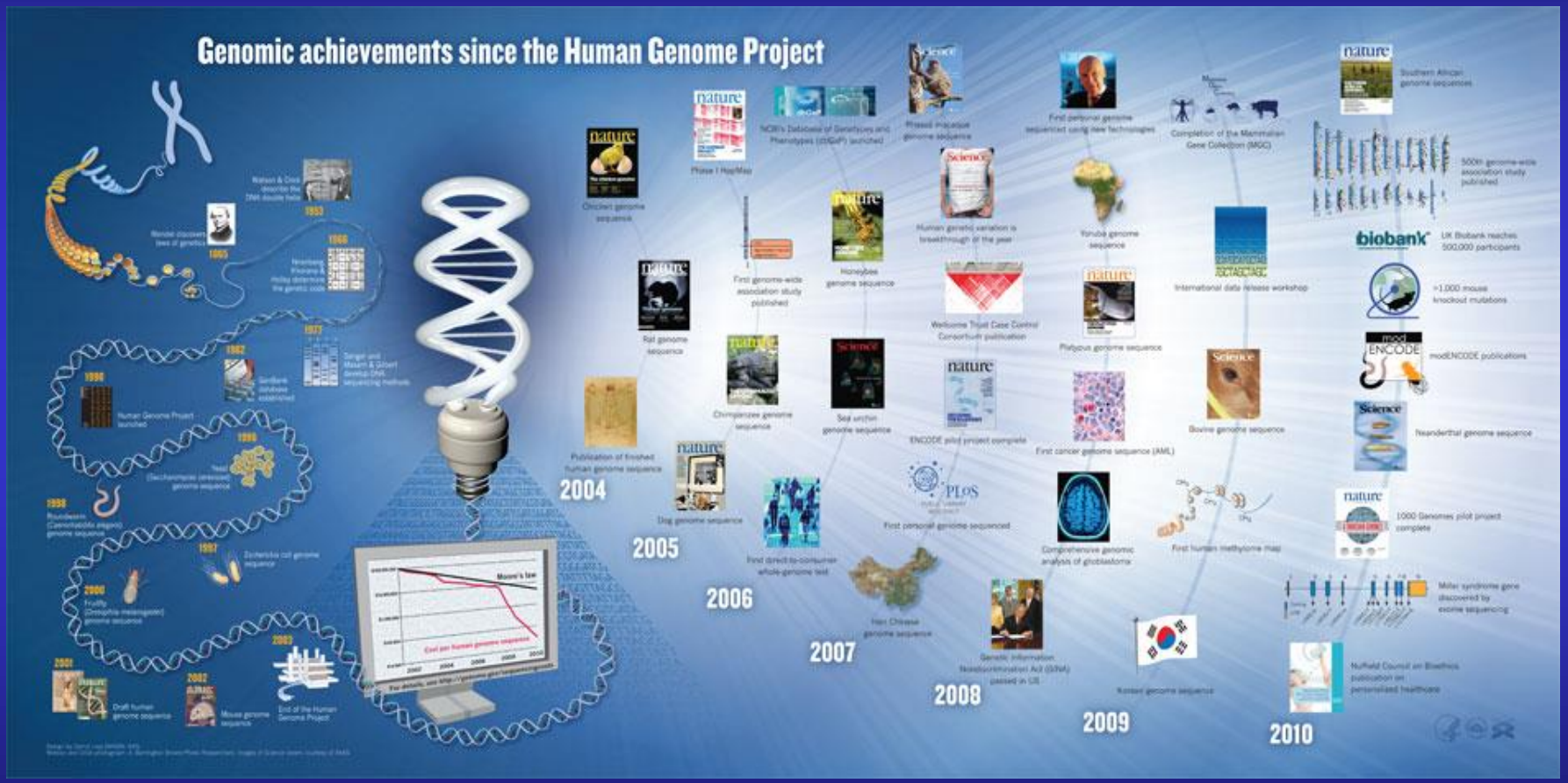
第三部分　蛋白芯片

第四部分　组织芯片

第五部分　高通量测序技术及其应用

第一部分　概　　述

“人类基因组计划”和“曼哈顿”原子弹计划，“阿波罗”登月计划，一起被誉为自然科学史上的“三大计划。



图片来源：Charting a course for genomic medicine from base pairs to bedside Nature Volume: 470, Pages: 204 – 213 Date published: (10 February 2011)

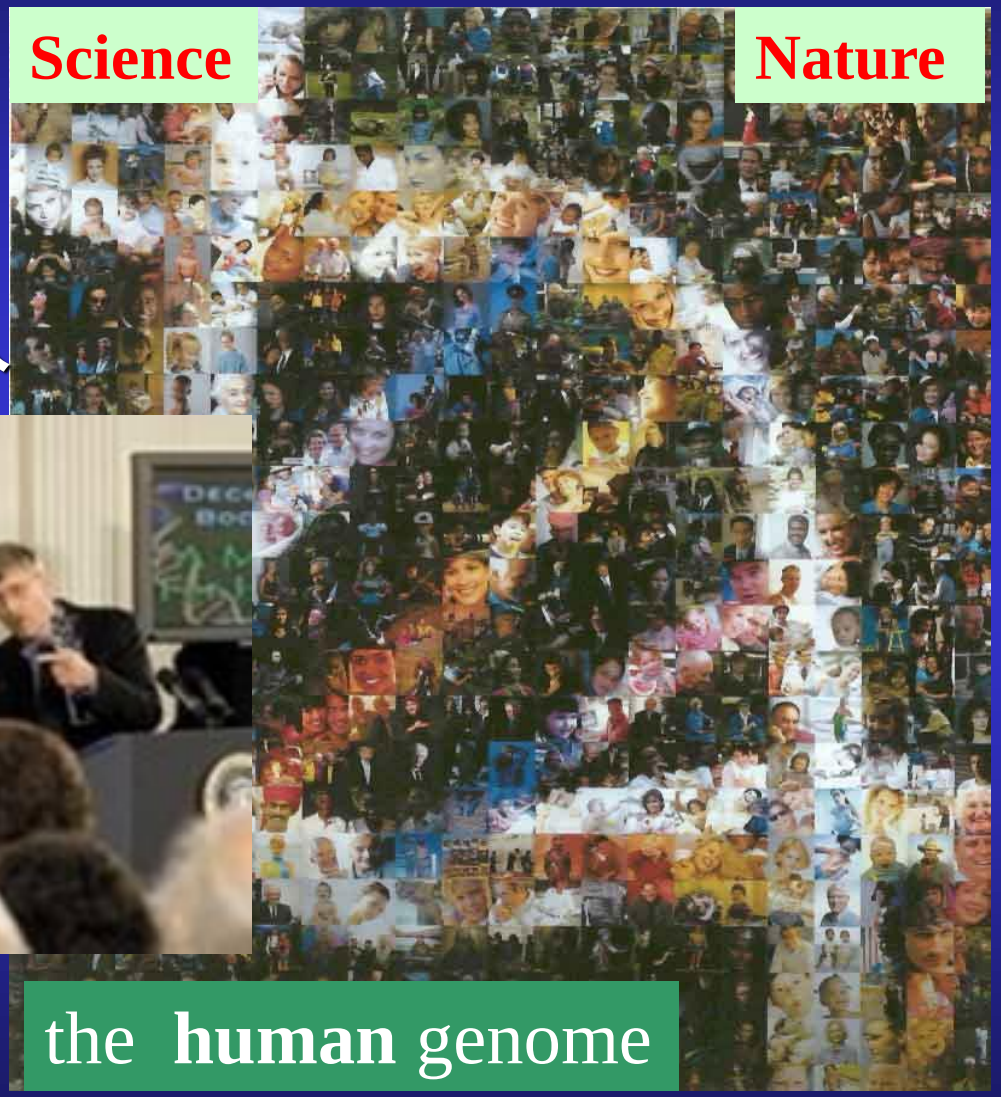
人类基因组研究战略和实验技术源源不断地产生了日益庞大及复杂的基因组数据，这些数据已被载入公共数据库，并改变了对几乎所有生命过程的研究。

人类基因组计划(HGP)

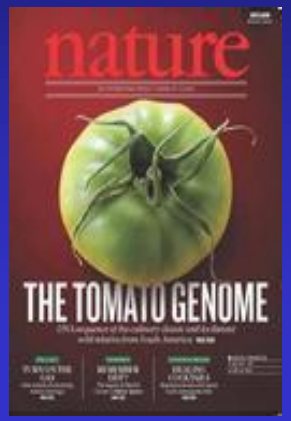
在1990年左右项目启动。

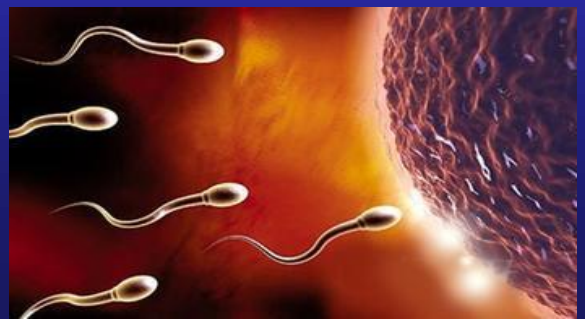
2000年6月26日，国际基因组组织宣布基因组草图完成。

2003年4月4日Collin Francs宣布人类基因组项目完成。



一些重要模式生物的基因组测序





已经完成的基因组测序

总的完成的基因组：13992
真核：2462（植物：218，动物：663，
真菌：1166，原生生物: 323，其他：13）
原核：50136
病毒：4940
质粒：6221
细胞器：7291

Source: http://www.ncbi.nlm.nih.gov/genome/browse/, 2015.10.26

已经完成超过百万个人的全基因组重测序

生物芯片和测序技术是系统生物学研究的重要技术

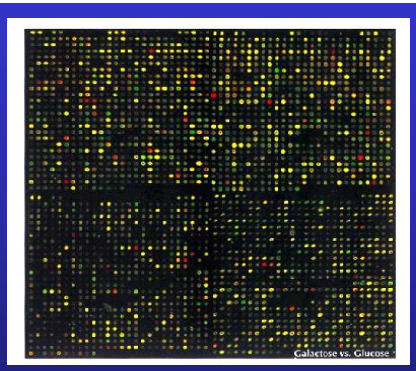
基因组
(DNA)

转录组
(RNA)

蛋白质组
(Protein)

功能基因组学

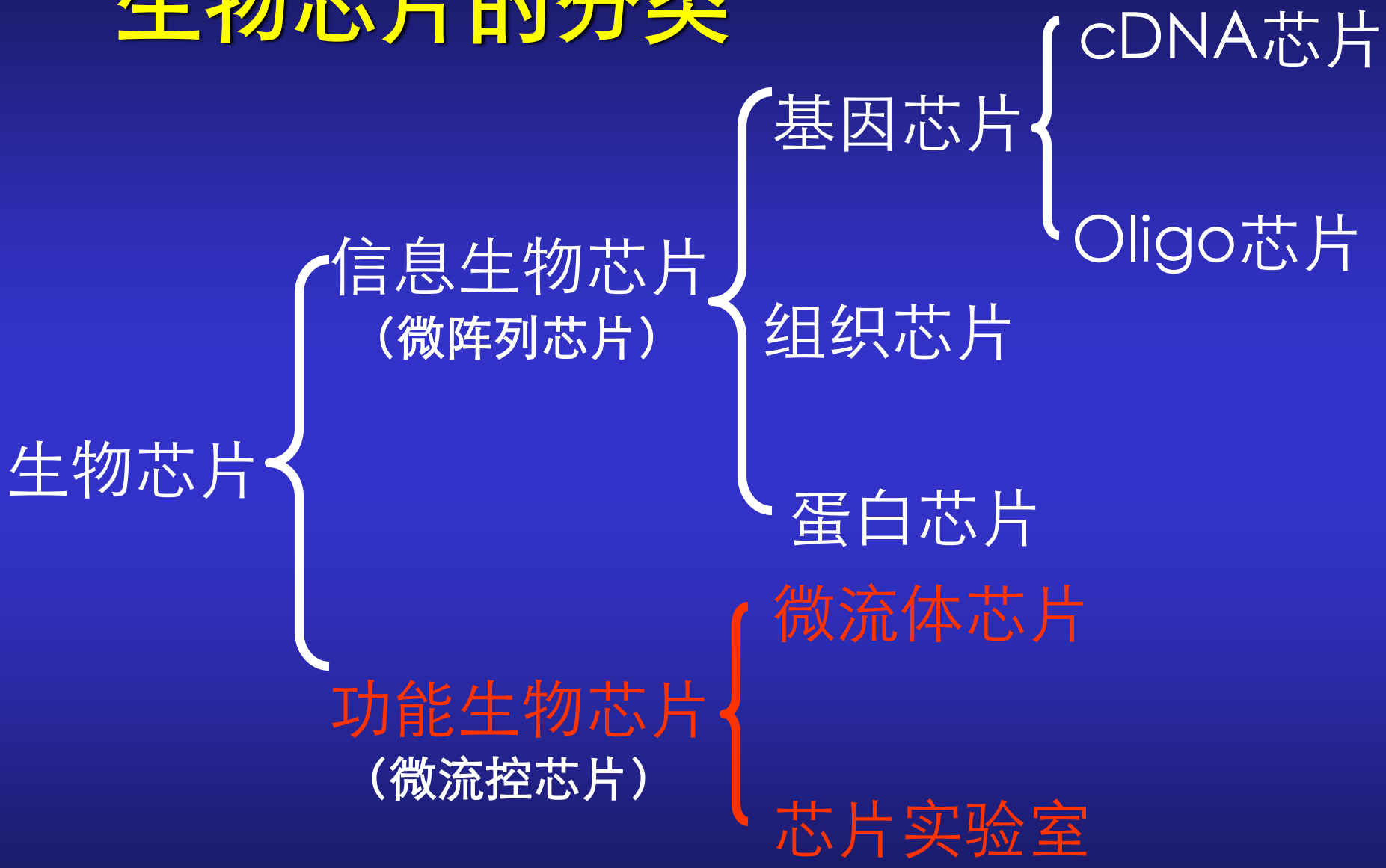
- Northern Blot
- 一代测序技术
- 新一代测序技术
- 差异显示、RDA（代表性差异分析）
- 基因芯片
- SAGE
- 蛋白质芯片



什么是生物芯片?

- 生物芯片的概念是Fodor等人于1991年提出(Fodor et al., 1991, Science)。
- 生物芯片的概念：是借用电子芯片的概念，是指能够快速并行处理多个样品并对其所包含的各种生物信息进行解剖的微型器件，它的加工运用了微电子工业和微机电系统加工中所采用的一些方法，只是由于其所处理和分析的对象是生物样品，所以叫生物芯片(Biochip)。
- 高通量，平行化，微量化的分析手段。

生物芯片的分类


生物芯片
信息生物芯片
（微阵列芯片）
基因芯片
cDNA芯片
Oligo芯片
组织芯片
蛋白芯片
功能生物芯片
（微流控芯片）
微流体芯片
芯片实验室

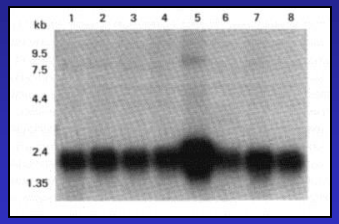
生物芯片技术的应用领域

第二部分 基因芯片

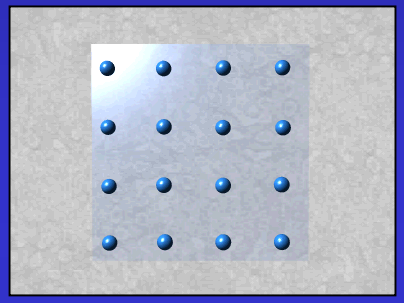
基因芯片(Gene chip)技术是指通过微阵列(Microarray)技术将高密度DNA片段通过高速机器人或原位合成方式以一定的顺序或排列方式使其附着在如膜、玻璃片等固相表面，以同位素或荧光标记的DNA探针，借助碱基互补杂交原理，进行大量的基因表达及监测等方面研究的最新革命性技术。

基因芯片发展过程

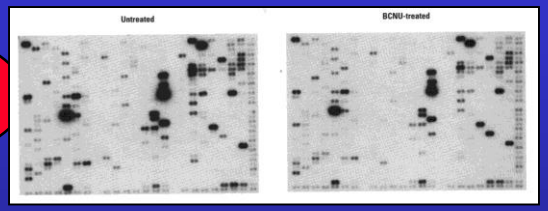
Southern & Northern Blot



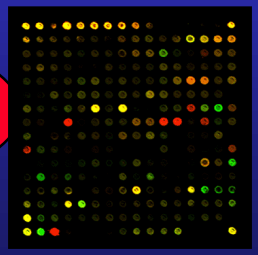
Dot Blot



Macroarray

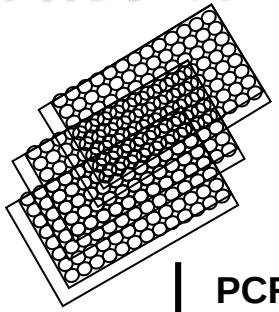


Microarray



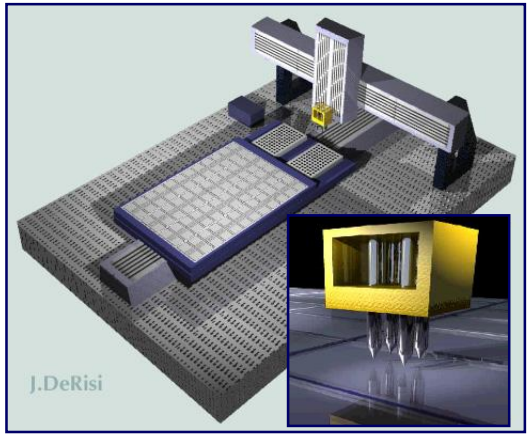
基因芯片的实验流程

cDNA clones
(target)



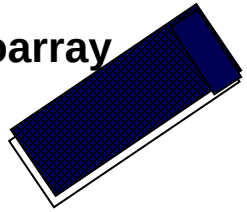
PCR product amplification
purification

printing

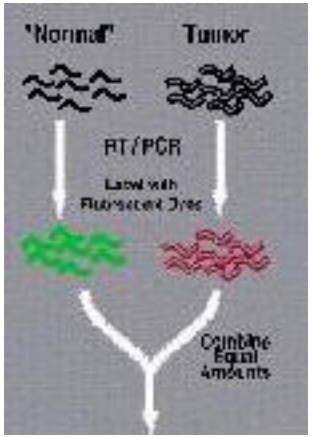


0.1nl/spot

microarray



mRNAprobe)



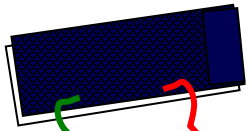
Hybridise target
to microarray

excitation

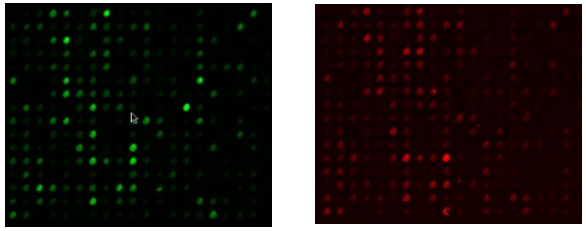
scanning

laser 2

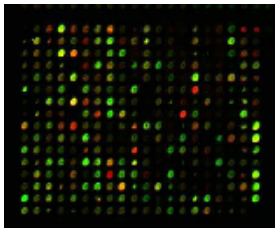
laser 1



emission

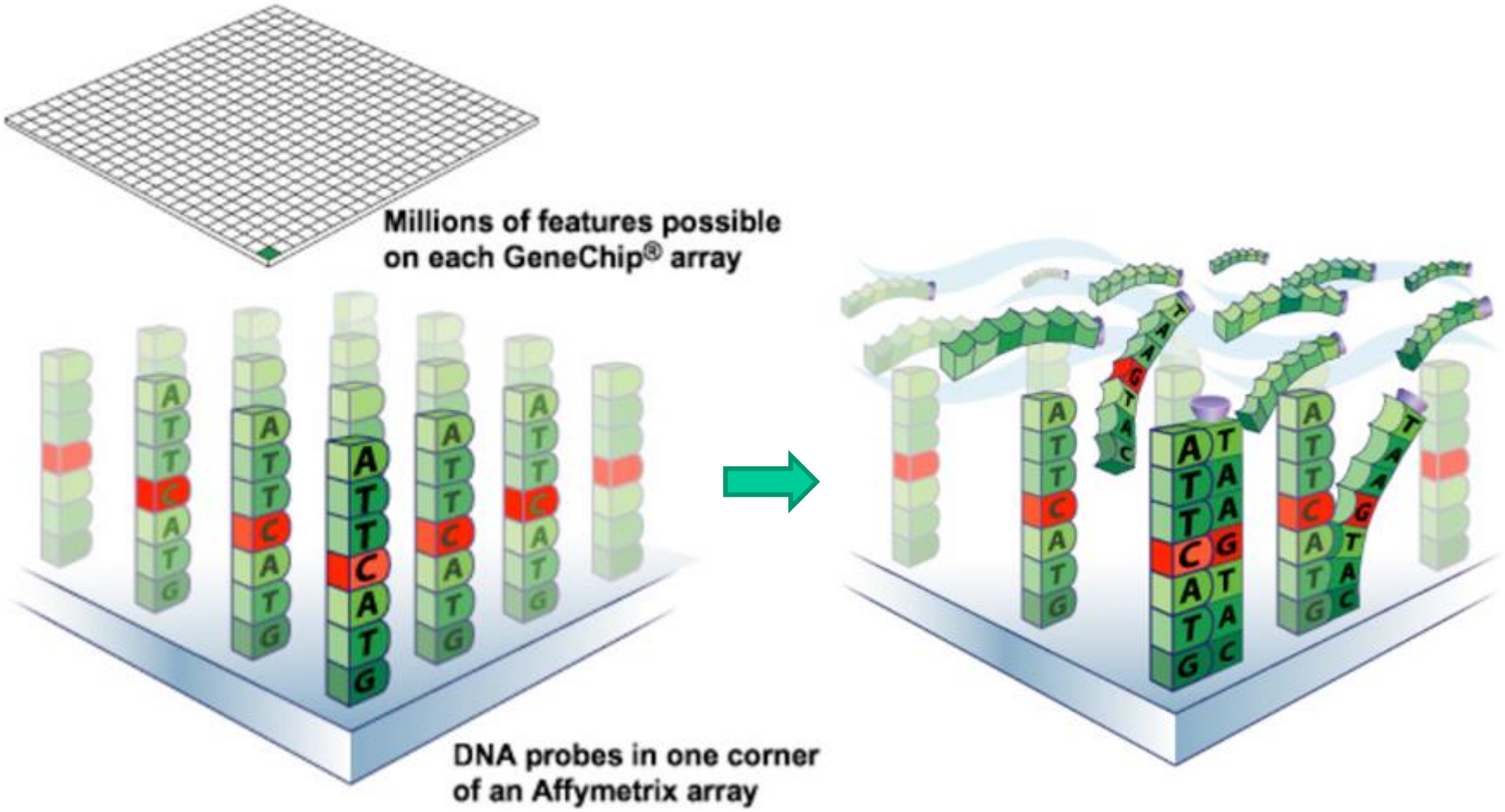


overlay images and normalise



analysis

基因芯片检测依赖核酸杂交原理



基因芯片的制作方法

- 原位合成
  - 光蚀保护 Affymetrix
  - Ink-jet Agilent
  - Digital Micromirror Device Nimbergene
  - 分子印章 东南大学
- 点样
  - cDNA PCR 产物
  - Oligo
  - 其它形式的基因产物

已有的微阵列制备技术

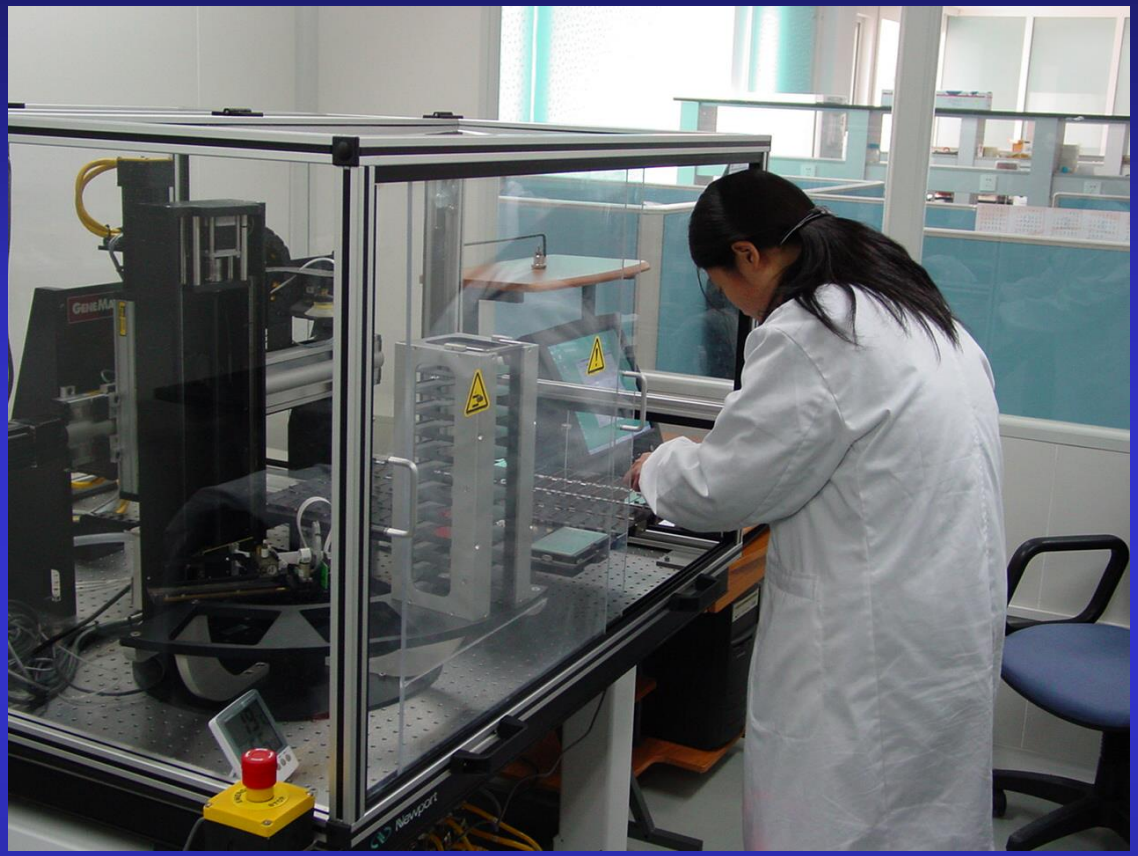
技术	探针类型	每个阵列的点数	实验室制备
接触点制	任意	~30,000	是
喷墨点制	任意	~20,000	是
无掩膜光刻技术	寡核苷酸	15,000	是
喷墨合成	寡核苷酸	240,000	否
基于掩膜的光刻技术	寡核苷酸	6,200,000	否
无掩膜光刻技术	寡核苷酸	2,100,000	否
微珠阵列	寡核苷酸	~50,000	否
电化学阵列	寡核苷酸	~15,000	否

主要微阵列供应商，以及他们是否提供定制芯片服务

公司	网站	定制芯片
Affymetrix	http://www.affymetrix.com/	是 *
Agilent Technologies	http://www.home.agilent.com/	是 *
Applied Biosystems	http://www.appliedbiosystems.com/	否
Applied Microarrays	http://www.appliedmicroarrays.com/	是 *
CombiMatrix Corporation	http://www.combimatrix.com/	是
Illumina	http://www.illumina.com/	是
NimbleGen Systems*	http://www.nimblegen.com/	是 *
Oxford Gene Technologies	http://www.ogt.co.uk/	是 *
Phalanx Biotech	http://www.phalanxbiotech.com/	是
SuperArray Bioscience	http://www.superarray.com/	是

* 已于2012年6月年关闭这个系统

点样制备芯片



- 接触式点样
- 非接触式点样

基因芯片使用样品的种类、处理和纯化

DNA　用常规的抽提方法的可以，要保证DNA的纯度，

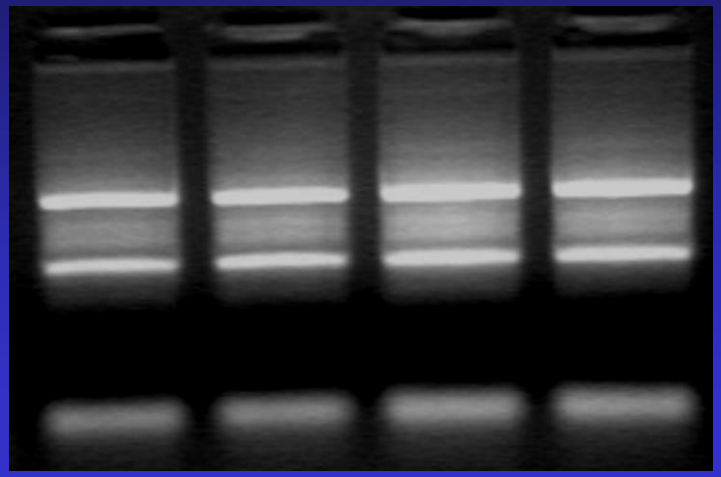
要求：A_{260}/A_{280}在1.6-1.8之间。

RNA

- 总RNA：完整性和纯度，忠实性最好。
- mRNA：完整性和纯度，操作比较繁琐，容易降解。
- cRNA：需要通过线性放大，适用于样品量较少，如血液中分离的细胞，激光显微切割获得的细胞等。

新鲜的组织样本，细胞，FFPE组织，显微切割组织等

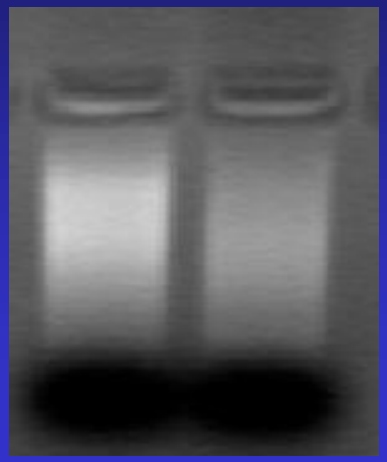
样本的质量控制



28S

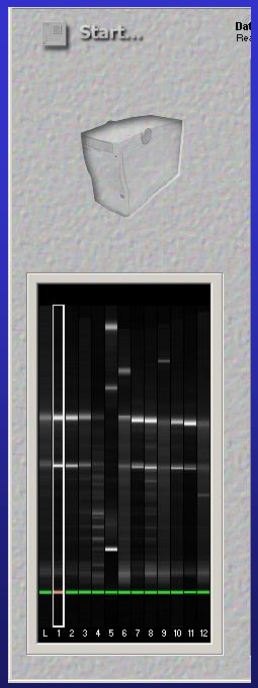
18S

28S:18S>2.0

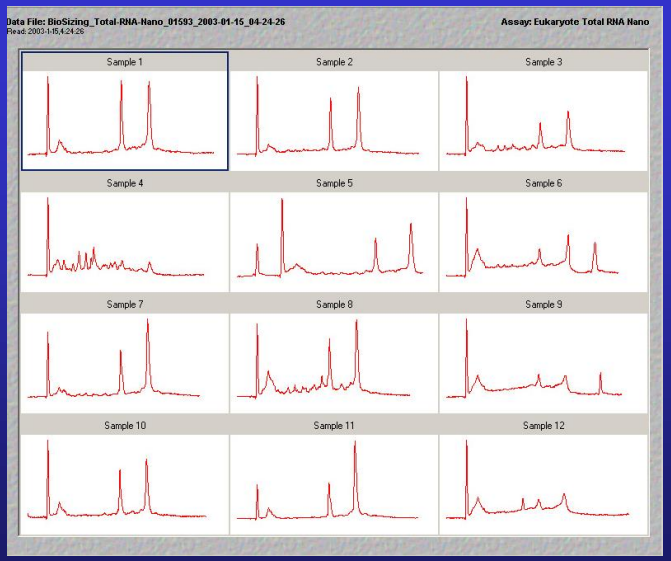


在0.5-2.0 Kb之间

Total RNA



mRNA



基因芯片的杂交

- 样品的标记

  标记方法

  DNA：PCR，随机引物，缺口翻译。

  RNA：逆转录，线性放大。

  标记物

  同位素

  荧光染料(Cy3, Cy5)

  化学发光

- 芯片杂交

  杂交体积(使核酸浓度增加10万倍)

  玻片: 2-200μl

  滤膜：5-50ml

  杂交液和杂交液的组份(

  杂交温度、时间

- 芯片的洗涤

  洗涤液的组成

  洗涤的温度、时间

基因芯片的扫描

- *激光共聚焦扫描*

  光源：特定波长的光

  激发面积：<100平方微米

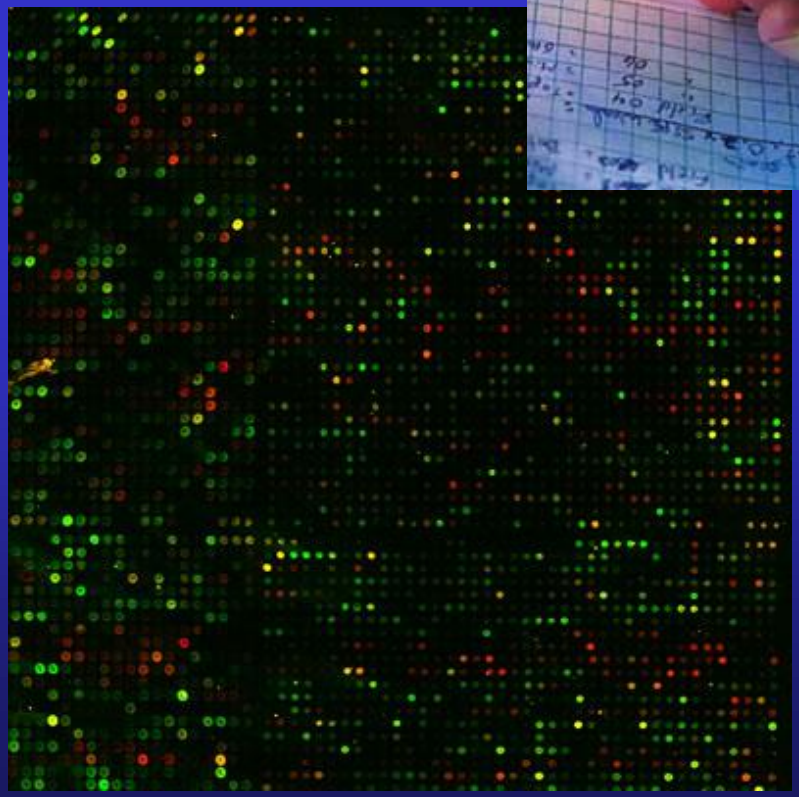
  如：Agilent Scanner

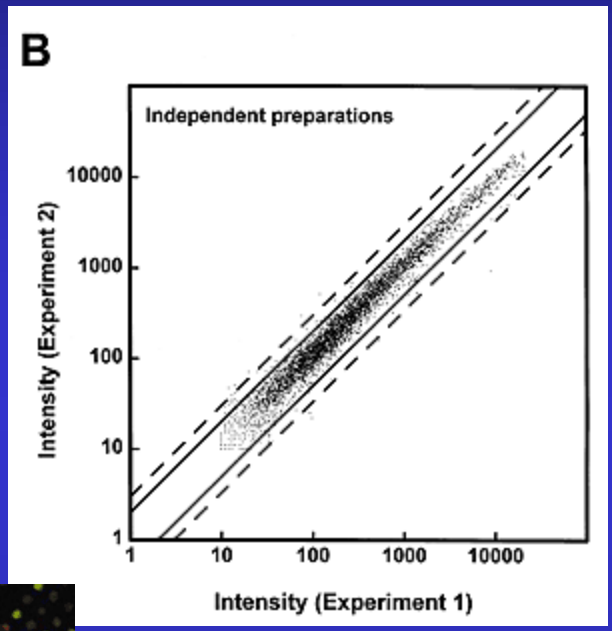
- CCD（电荷耦合器件）成像

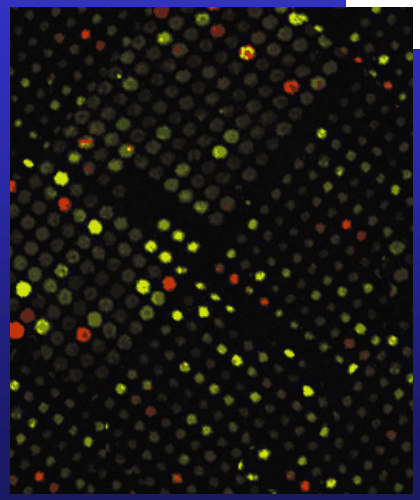
  光源：连续波长的光（如弧光灯）

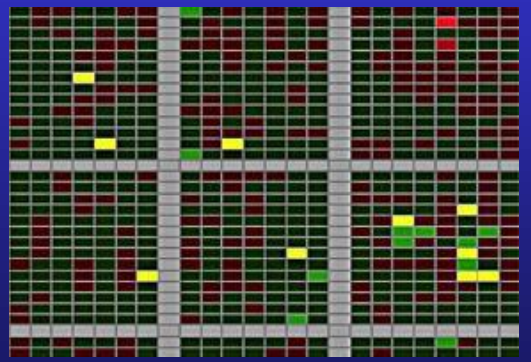
  激发面积：同时激发多个1平方厘米的面积

基因芯片的数据的处理和分析









基因芯片的数据解读和分析

1. 芯片图像的处理。
2. 芯片杂交后获得的数据与芯片的基因信息的连接。
3. 芯片数据的预处理及数据的可视化。
4. 数据处理和分析的算法。
5. 数据分析方法对新功能、疾病诊断和治疗等进行预测。
6. 芯片的数据与多种资源和信息建立有效连结。

芯片图像的处理

- 芯片的图像信息存储在16位的Tiff图像里，每个点的强度都是根据在这个点位置的相应像素的强度值得到的。
- 图像处理的基本目的是得到芯片上样点的具体信息。
- 常用的芯片图像处理软件：

Biodiscovery Inc. 的 ImaGene

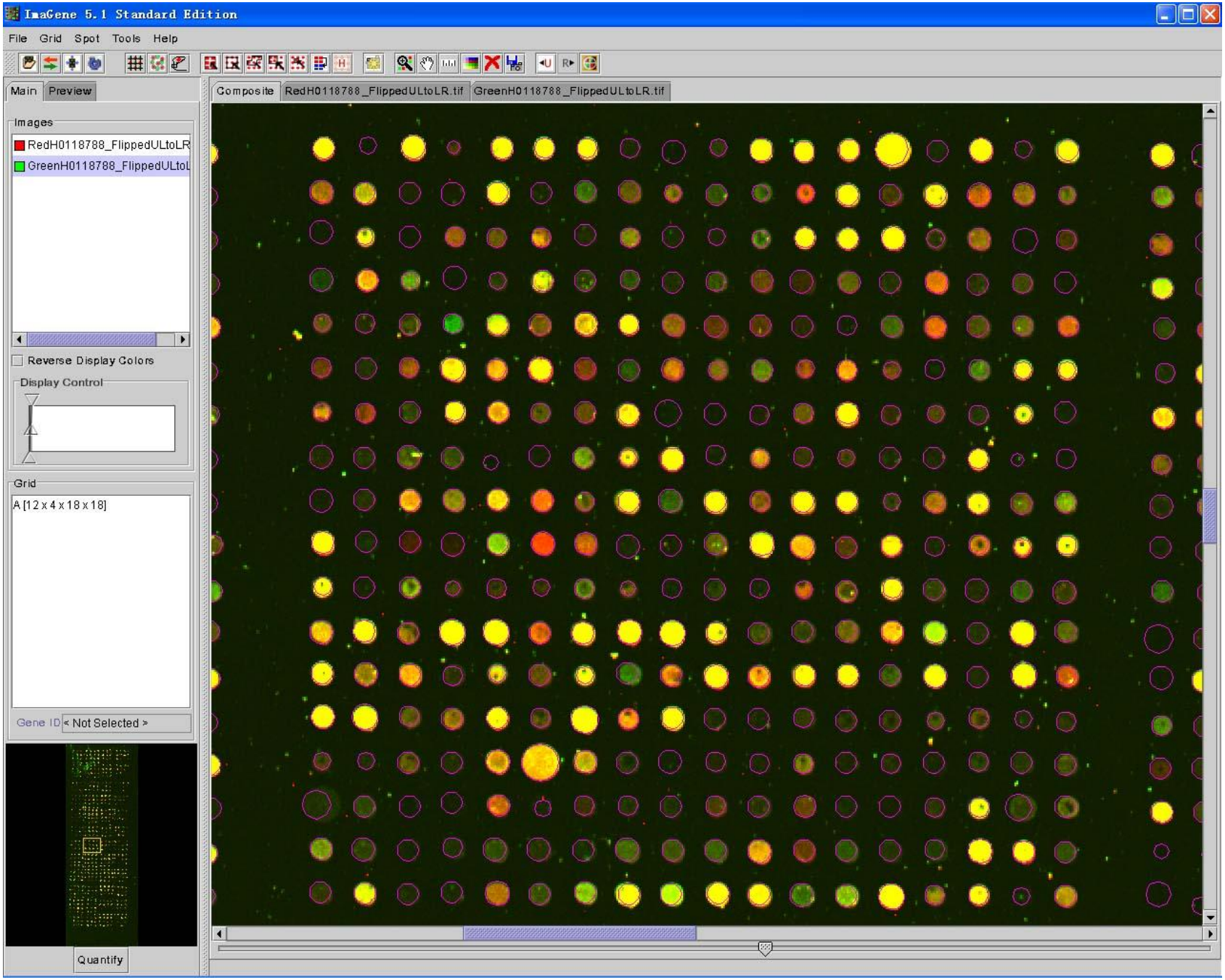
Imaging Research Inc. 的 ArrayVision

Axon Instruments Inc. 的 GenePix Pro

Packard Bioscience 的 QuantArray

Stanford University 的 Scanalyse

Media Cybernetic 的 ArrayPro Analyzer


ImaGene 5.1 Standard Edition
File Grid Spot Tools Help
Main Preview
Images
RedH0118788_FlippedULtoLR
GreenH0118788_FlippedULtoL
Reverse Display Colors
Display Control
Grid
A [12 x 4 x 18 x 18]
Gene ID < Not Selected >
Quantify
Composite RedH0118788_FlippedULtoLR.tif GreenH0118788_FlippedULtoLR.tif

下表是整理后数据的一部分

Array_ID	Cy3_SM	Cy3_BM	Cy3_BS	Cy3_signal	Cy3_SN	Cy5_SM	Cy5_BM	Cy5_BS	Cy5_signal	Cy5_SN	Flag
Sr01Sc01R01C01	2564	394	19.57607737	2170	110.8496	1964.5	120	12.26637544	1844.5	150.3704	0
Sr01Sc01R01C02	422	380	18.33430843	42	2.290787	132	119	13.36780439	13	0.972486	2
Sr01Sc01R01C03	4507	379	62.34876469	4128	66.20821	4069	119	27.876873	3950	141.6945	0
Sr01Sc01R01C04	373	366	15.82856721	0	0	122	117	12.48741749	0	0	2
Sr01Sc01R01C05	1023	365	18.82181909	658	34.95943	584	118	12.71648114	466	36.64536	0
Sr01Sc01R01C06	1079	361	16.60344801	718	43.24403	967.5	118	13.25245152	849.5	64.10135	0
Sr01Sc01R01C07	1194	357	17.0117534	837	49.20128	839	116	12.03354419	723	60.08205	0
Sr01Sc01R01C08	420	351	18.80657111	69	3.66893	163	116	13.04477968	47	3.602974	0
Sr01Sc01R01C09	364	359	15.80129994	0	0	118	116	11.83347663	0	0	2
Sr01Sc01R01C10	383	360	15.26872232	23	1.506347	136	117	12.29356769	19	1.545524	0
Sr01Sc01R01C11	4034	349	20.41211993	3685	180.53	2295	117	13.8323998	2178	157.4564	0
Sr01Sc01R01C12	1316	351	28.56571657	965	33.78175	1150	117	16.03960164	1033	64.4031	0
Sr01Sc01R01C13	660	346	19.56984928	314	16.04509	501	116.5	14.38882787	384.5	26.72212	0
Sr01Sc01R01C14	1022	343	20.75239347	679	32.71912	397.5	117	14.06295147	280.5	19.94603	0
Sr01Sc01R01C15	391	339	15.75445968	52	3.300653	128	115	12.07106232	13	1.076956	2
Sr01Sc01R01C16	3165	337	17.55824857	2828	161.0639	3052	115	12.76292006	2937	230.1198	0
Sr01Sc01R01C17	647.5	339	16.1425219	308.5	19.11102	521	116	13.31720769	405	30.41178	0
Sr01Sc01R01C18	497	338	15.1449806	159	10.49853	324.5	116.5	12.65218178	208	16.43985	0
Sr01Sc01R02C01	1593	388	17.65926338	1205	68.23614	2143	123	23.03825668	2020	87.68025	0
Sr01Sc01R02C02	520	381	20.98985167	139	6.622248	281.5	119	14.94907665	162.5	10.87024	0
Sr01Sc01R02C03	574	377	16.44852148	197	11.97676	328	120	15.48309511	208	13.43401	0
Sr01Sc01R02C04	485	367	16.03530107	118	7.358764	256	118	11.26796328	138	12.24711	0
Sr01Sc01R02C05	724.5	363	18.34423549	361.5	19.70646	417	119	12.28495537	298	24.25731	0
Sr01Sc01R02C06	453	363.5	17.6795157	89.5	5.062356	183	118	12.46740254	65	5.213596	0

芯片的基因信息

- 样点信息，包括Array ID, Clone ID, Probe ID 等。
- 对应的基因的序列信息或Probe序列。
- 基因的相关信息，包括GenBank ID, Unigene ID, Gene name, GO ontology，Pathway 等。
- 芯片制备的目的、条件、方法；样品的制备、杂交条件、环境条件和检测方法。

数据读取后的初步分析和可靠性评估

1、阳性信号检出率：取阴性质控点如DMSO的信号值，分别计算其Cy3、Cy5信号平均值与2倍平均方差的和（mean+2*STDEV），Cy3、Cy5信号值分别大于这一值的基因点为检出点。

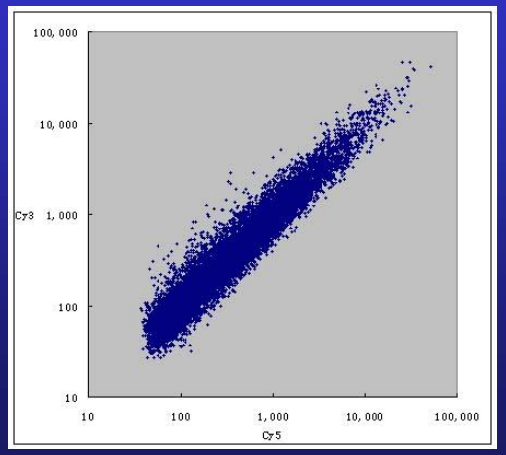
2、将一些质量很低，数据可能不准确的点予以滤出。

3、对芯片数据进行**标准化**处理。

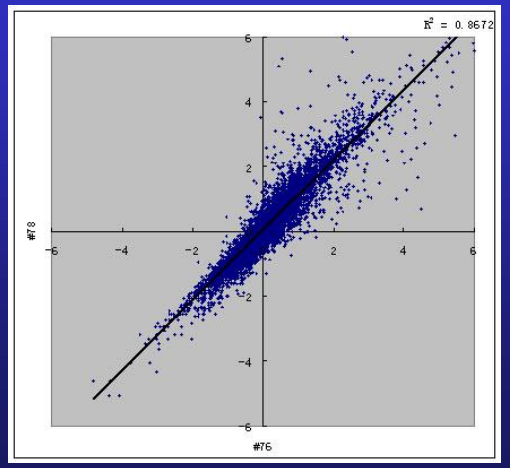
4、数据的分布：一般用散点图来表示。

5、重复性一般用R^2来表示。

片内



片间



数据的预处理 – 芯片数据的质量（1）

芯片数据的质量包括两个方面：整个芯片的质量和芯片上点的质量。

评价整张芯片质量的最简单也是最常用的方法是计算整个芯片的信噪比。信噪比低表示整张芯片背景高，芯片的质量差。

数据的预处理 – 芯片数据的质量（2）

信号点的质量影响体现在5个方面

- 信号点的大小和规则度
- 信噪比
- 信号点周围的背景强度
- 信号点背景的均一程度
- 信号的饱和程度

一般在分析时，会先确定质量低的点，并滤除这些点，这是数据处理的重要步骤。

数据的预处理 – 数据的标准化（1）

影响芯片原始数据的因素很多，在对芯片数据分析之前，必须进行数据的标准化（normalization)。数据的标准化是要减少芯片在处理过程中技术（系统）因素的影响，使检测的结果能真实地反映生物功能地差别，芯片的数据只有经过标准化处理后才具有可比性。

Normalization方法有按参与校正的基因分：

- 全部基因参与
- Housekeeping 基因参与
- Reference 基因参与

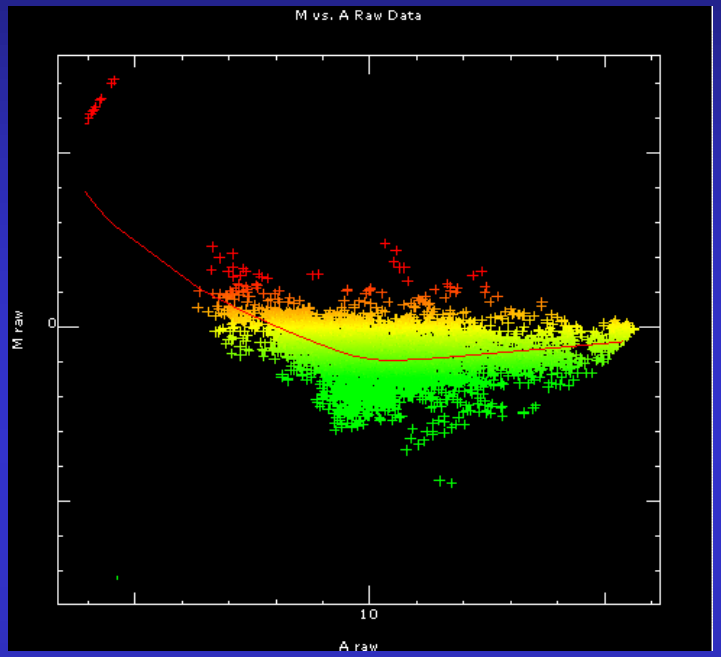
数据的预处理－数据的校正（2）

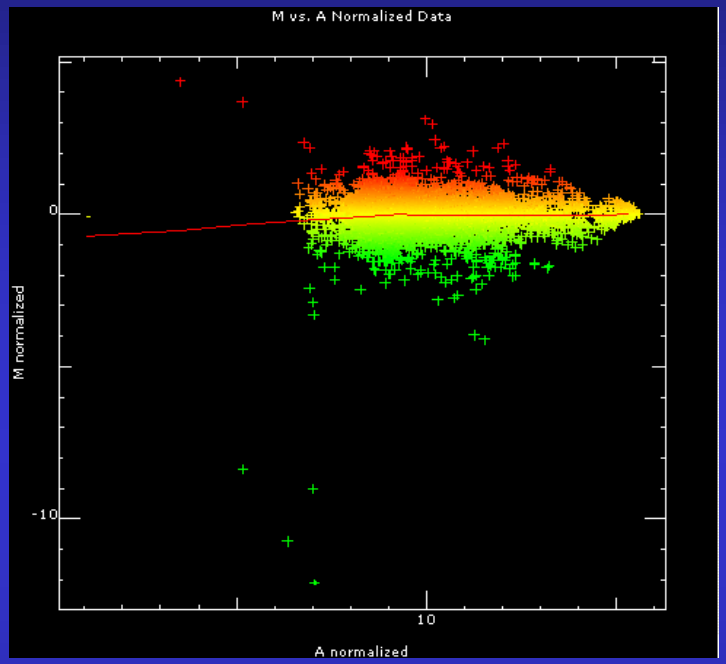
按值的估计方法分有3种：

- 全局法（global method）
- 线性回归（linear regression）
- 局部加权最小二乘法 （LOWESS)

无论那种方法，目的使调整芯片技术导致的误差，但不能纠正由于mRNA样本和芯片上探针基因本身所带来的误差。

数据的预处理 – 标准化前后的散点图





- 左图是未标准化处理的散点图
- 右图是经LOWESS处理的散点图

任何芯片数据进行分析前，都必须进行数据的标准化。

数据的初步分析 – 差异基因的选择

- 一般来说，ratio>2或ratio<0.5认为是在两种样本表达有差异。这方法没有考虑到差异表达的统计显著性。
- Z-score ((X- μ)/ó), Z>2表示比率在平均比率加两倍方差之外，差异基因有了统计意义。
- T-检验（ t-test ），从重复芯片中识别差异的表达基因。
- SAM。
- ANOVA。

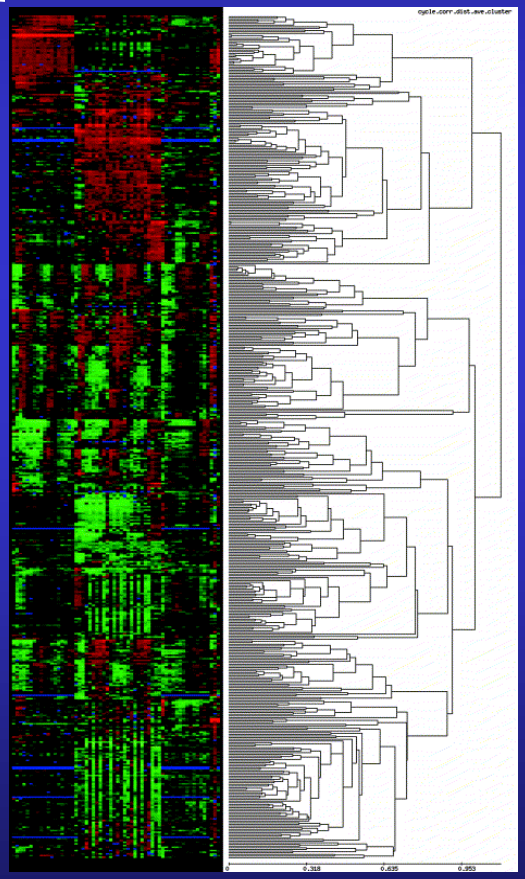
现在识别差异基因方法已经有很多方法，它仍是数据处理中的一个热点。

数据的可视化(1) – 聚类分析

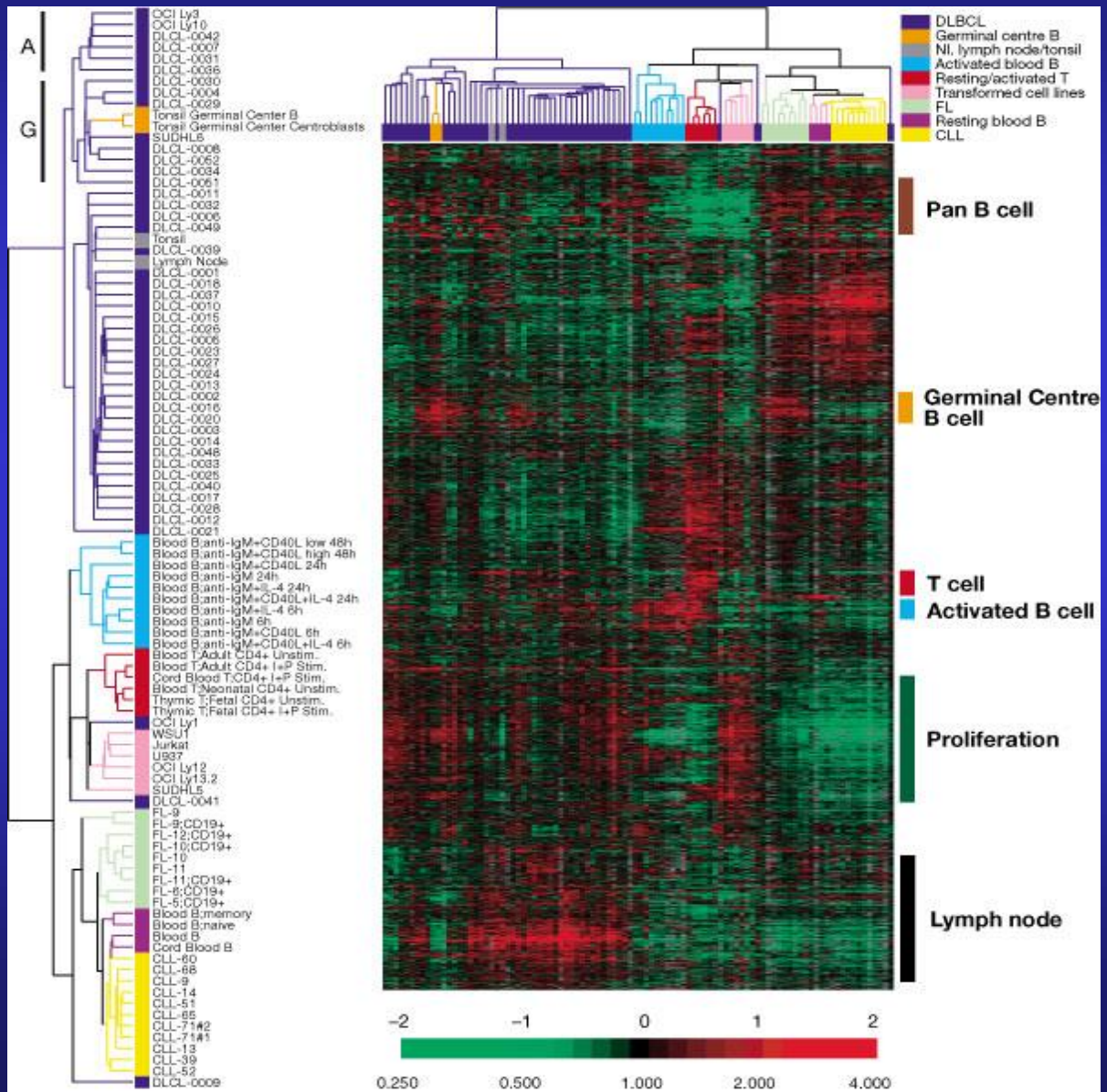
基因聚类分析的主要任务是确定相似表达模式的基因，相似的基因可能具有共同的特征。

探索完全未知的数据特征的方法

- 层次聚类(hierarchical cluster)
- K-means 基于向量的
- SOM
- 主成分分析(PCA)

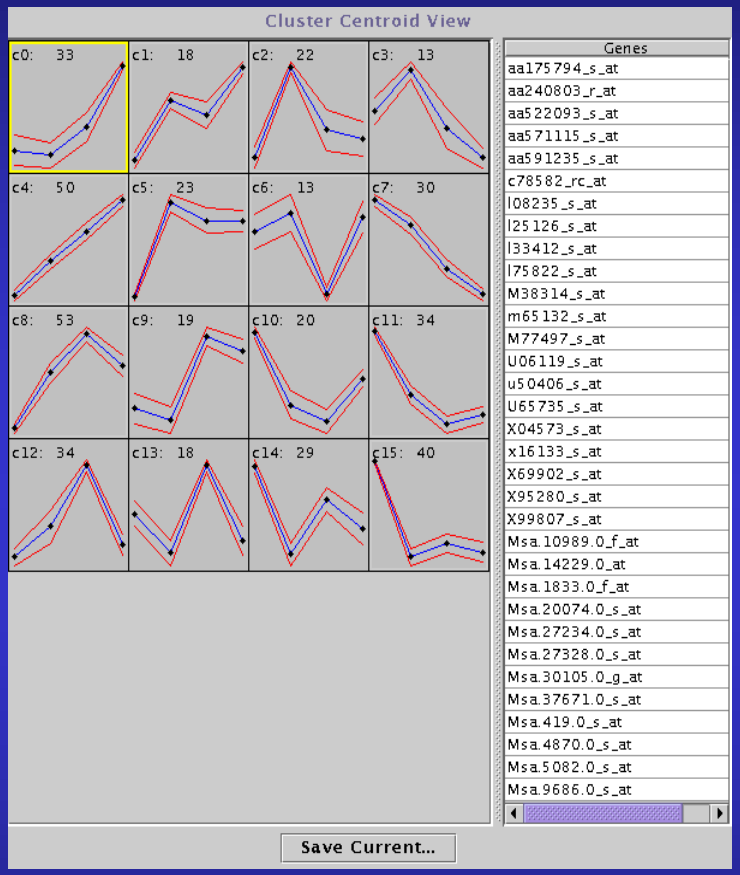


一个实例

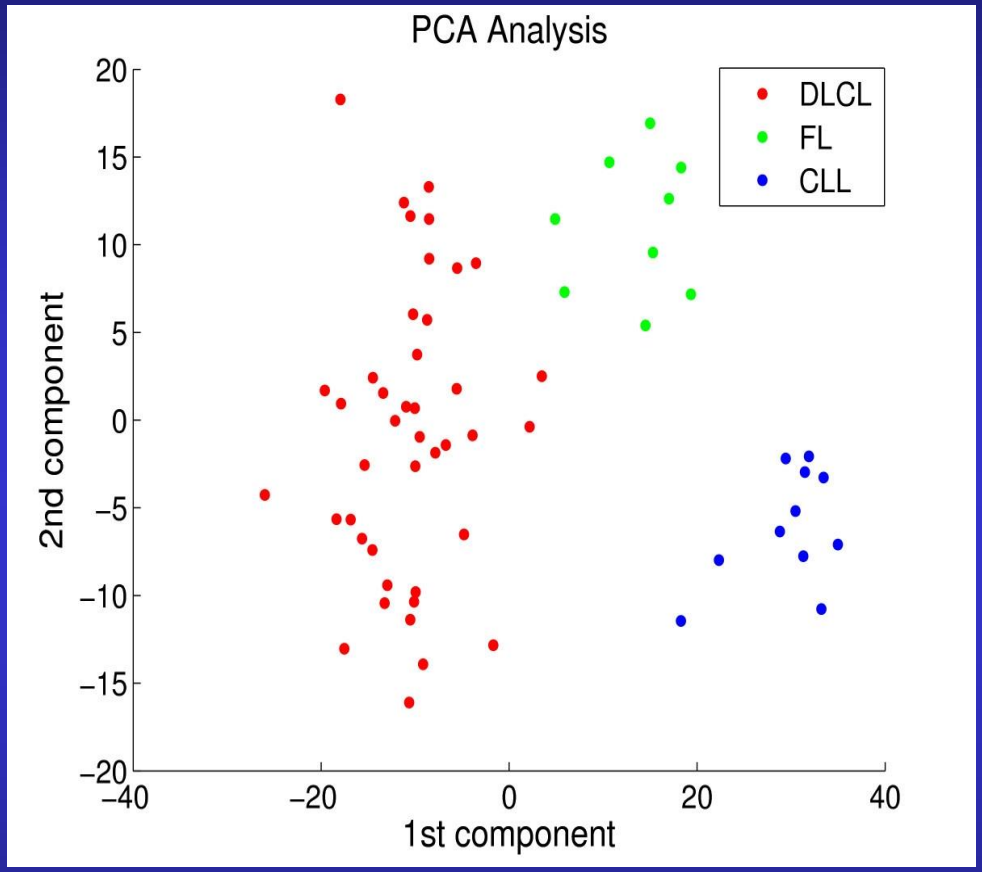


Nature Feb, 2000 Paper by Allzadeh. A *et al* *Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling*

SOM


Cluster Centroid View
c0: 33
c1: 18
c2: 22
c3: 13
c4: 50
c5: 23
c6: 13
c7: 30
c8: 53
c9: 19
c10: 20
c11: 34
c12: 34
c13: 18
c14: 29
c15: 40
Genes
aa175794_s_at
aa240803_r_at
aa522093_s_at
aa571115_s_at
aa591235_s_at
c78582_rc_at
l08235_s_at
l25126_s_at
l33412_s_at
l75822_s_at
M38314_s_at
m65132_s_at
M77497_s_at
U06119_s_at
u50406_s_at
U65735_s_at
X04573_s_at
x16133_s_at
X69902_s_at
X95280_s_at
X99807_s_at
Msa.10989.0_f_at
Msa.14229.0_at
Msa.1833.0_f_at
Msa.20074.0_s_at
Msa.27234.0_s_at
Msa.27328.0_s_at
Msa.30105.0_g_at
Msa.37671.0_s_at
Msa.419.0_s_at
Msa.4870.0_s_at
Msa.5082.0_s_at
Msa.9686.0_s_at
Save Current...

PCA


PCA Analysis
DLCL
FL
CLL
2nd component
1st component
20
15
10
5
0
−5
−10
−15
−20
−40
−20
0
20
40

数据的深入分析(2) – Pathway 分析

它是聚类分析的延伸，若有一些基因在聚类分析中被认为是同一群，我们可以判断 他们可能是同一个代谢途径

KEGG:

Kyoto Encyclopedia of Genes and Genomes

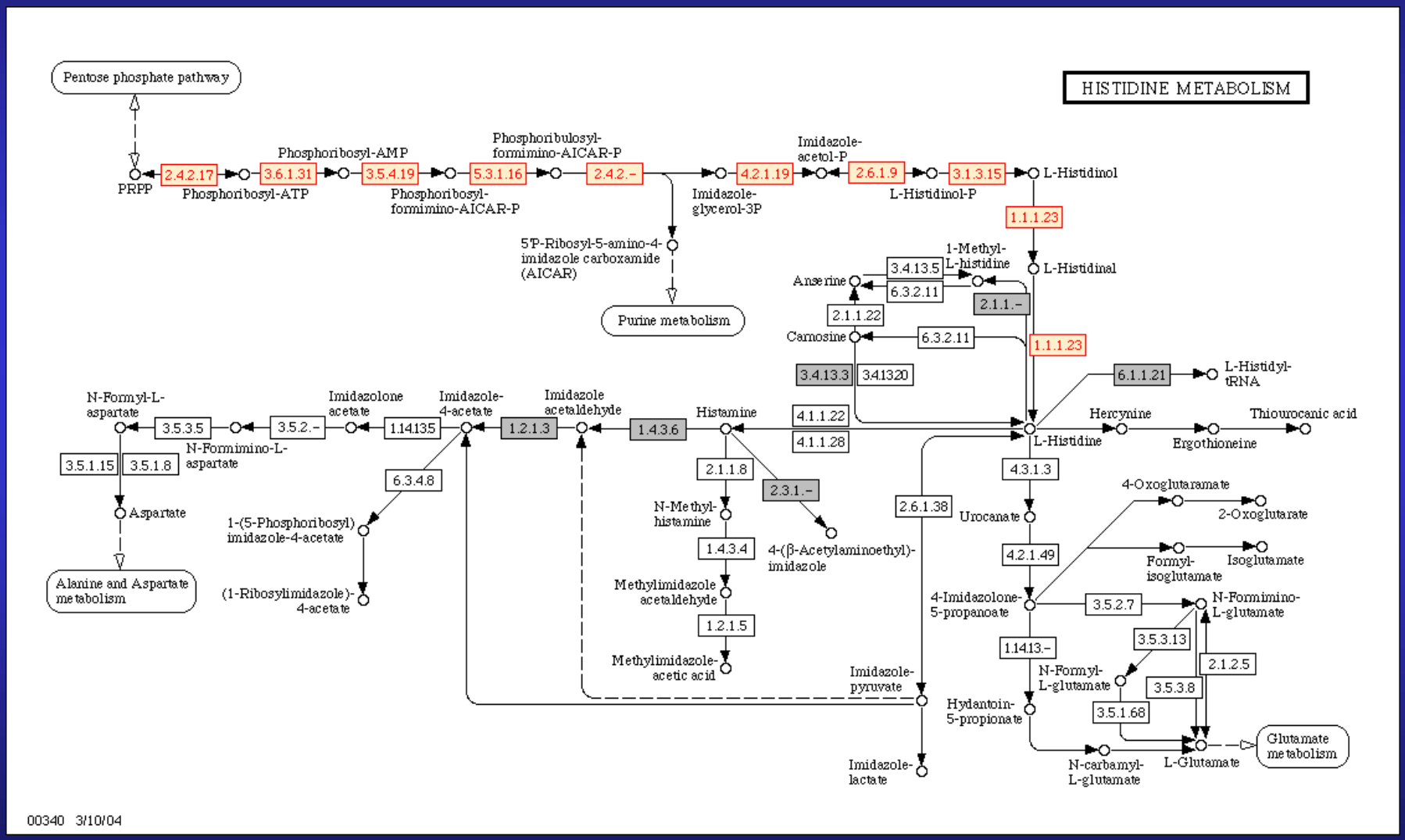
Software :

Pathway Assist；Pathway Editor；Pathway Finder

Network/pathway：建议文献

Pilpel Y，etc. Identifying regulatory networks by combinatorial analysis of promoter elements. Nat Genet 2001 Oct;29(2):153-9.

Pathway 分析


HISTIDINE METABOLISM
Pentose phosphate pathway
PRPP
2.4.2.17
Phosphoribosyl-ATP
3.6.1.31
Phosphoribosyl-AMP
3.5.4.19
Phosphoribosyl-formimino-AICAR-P
5.3.1.16
Phosphoribulosyl-formimino-AICAR-P
2.4.2.-
5'P-Ribosyl-5-amino-4-imidazole carboxamide (AICAR)
Purine metabolism
4.2.1.19
Imidazole-glycerol-3P
2.6.1.9
Imidazole-acetol-P
3.1.3.15
L-Histidinol-P
L-Histidinol
1.1.1.23
L-Histidinal
3.4.13.5
6.3.2.11
1-Methyl-L-histidine
Anserine
2.1.1.22
Carnosine
6.3.2.11
2.1.1.-
1.1.1.23
3.4.13.3
3.4.13.20
6.1.1.21
L-Histidyl-tRNA
L-Histidine
4.1.1.22
4.1.1.28
Histamine
1.4.3.6
1.2.1.3
Imidazole acetaldehyde
1.14.13.5
Imidazole-4-acetate
Imidazolone acetate
3.5.2.-
N-Formimino-L-aspartate
3.5.3.5
N-Formyl-L-aspartate
3.5.1.15
3.5.1.8
Aspartate
Alanine and Aspartate metabolism
6.3.4.8
1-(5-Phosphoribosyl)-imidazole-4-acetate
(1-Ribosylimidazole)-4-acetate
2.1.1.8
N-Methyl-histamine
1.4.3.4
Methylimidazole acetaldehyde
1.2.1.5
Methylimidazole-acetic acid
2.3.1.-
4-(β-Acetylaminoethyl)-imidazole
2.6.1.38
Imidazole-pyruvate
Imidazole-lactate
4.3.1.3
Urocanate
4-Oxoglutaramate
2-Oxoglutarate
4.2.1.49
Formyl-isoglutamate
Isoglutamate
4-Imidazolone-5-propanoate
3.5.2.7
N-Formimino-L-glutamate
1.14.13.-
3.5.3.13
N-Formyl-L-glutamate
2.1.2.5
3.5.3.8
Hydantoin-5-propionate
3.5.1.68
N-carbamyl-L-glutamate
L-Glutamate
Glutamate metabolism
Hercynine
Thiourocanic acid
Ergothioneine
00340 3/10/04

数据的深入分析(3)-基因的分类（GO）

分子功能

the tasks performed by individual gene products; examples are *carbohydrate binding* and *ATPase activity*

生物过程

broad biological goals, such as *mitosis* or *purine metabolism*, that are accomplished by ordered assemblies of molecular functions

亚细胞定位

subcellular structures, locations, and macromolecular complexes; examples include *nucleus*, *telomere*, and *origin recognition complex*

数据的深入分析(4)-基因的序列和结构域的分析

- 通过差异筛选、聚类分析，从数据中选出感兴趣的基因，就面临深入了解这些基因的功能信息问题。
- 基因序列分析通常是通过公共数据库查询和序列生物信息分析，如：NCBI中的GenBank、LocusLink、UniGene库，欧洲的EMBL核酸库；日本的DDBJ库，蛋白质序列数据库SWISS-PROT、PSD。
- 该部分包括：序列比对和数据库搜索，核酸序列和蛋白质序列功能预测。

数据的深入分析(4)-基因的序列和结构域的分析

- 序列比对和数据库搜索：两两比对，找出相似区域和保守性位点，寻找可能的分子进化关系。常见的工具有FASTA和BLAST。
- 核酸序列功能预测：在序列中寻找功能位点，以及标记已知序列模式等过程。
- 蛋白质序列功能预测：通过分析识别蛋白的分子量、等电点、二级结构、三级结构、四级结构、膜蛋白跨膜区域、酶的活性点、蛋白间作用等。

数据的深入分析(5)-文献的挖掘 (Text Mining or Literature mining)

对筛选出来的差异表达基因，通过前面的一系列分析，已经有了很好的提示。利用NCBI的文献库，可以用基因的GenBank ID或者Unigene ID查询相关的文献，也可以用基因名的主题词进行查询，查阅相关的主要文献，尤其是综述，总结基因的功能。用芯片的数据和基因的已知功能去解释观察到的现象的分子机制。

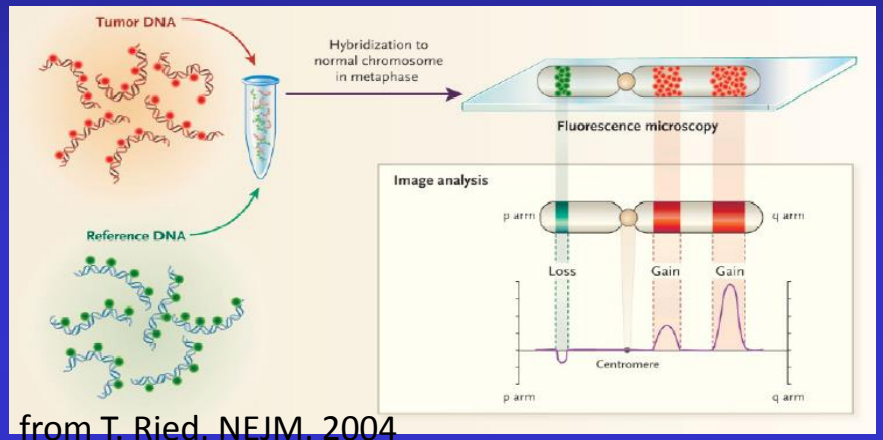
一些常用的网址

- NCBI www. ncbi.nlm.nih.gov
- Affymetrix www.affymetrix.com
- Michael Eisen Lab at LBL (hierarchical clustering software “Cluster” and “Tree View” (Windows))
  rana.lbl.gov/
- Stanford MicroArray Database (“Xcluster” (Linux))
  genome-www4.stanford.edu/MicroArray/SMD/
- Review of Currently Available Microarray Software
  www.the-scientist.com/yr2001/apr/profile1_010430.html
- Microarray DB
  www.biologie.ens.fr/en/genetiqu/puces/bddeng.html

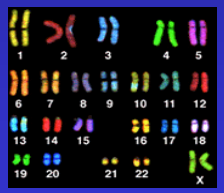
基因芯片在生命科学和医学研究中的应用

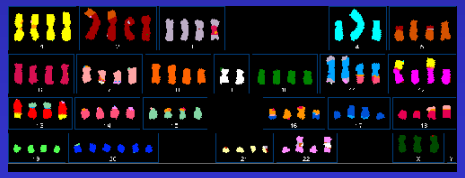
在基因组水平的应用

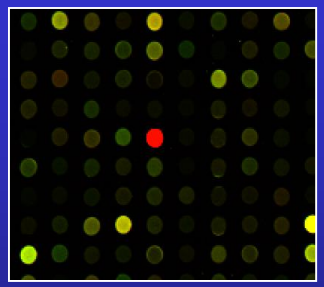
aCGH： Array Based Comparative Genomic Hybridization



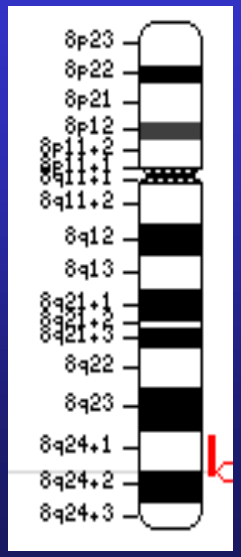
from T. Ried, NEJM, 2004







Oligo aCGH



- High-resolution
- High-throughput
- Quantitative
- Highly-flexibile

Mechanisms of disease onset and progression,
Identification of novel therapeutic targets,
Drug resistance mechanisms and patient stratification
Biomarkers for diagnostics and prognosis

Huang J et al., FEBS Letters. 2006

Correlation between genomic DNA copy number alterations and transcriptional expression in hepatitis B virus-associated hepatocellular carcinoma

1、HBV-induced HCC have unique pattern of DNA copy number alterations

（1）HCC（n＝41） and HCC cell lines （n＝12） occur significant DNA copy number alterations as compared with normal livers （n＝2）, and adjacent non-cancerous livers （n＝5）

C=Cancer
N=non-cancerous liver
NL=normal liver
Red=amplification(gain)
Green=deletion(loss)

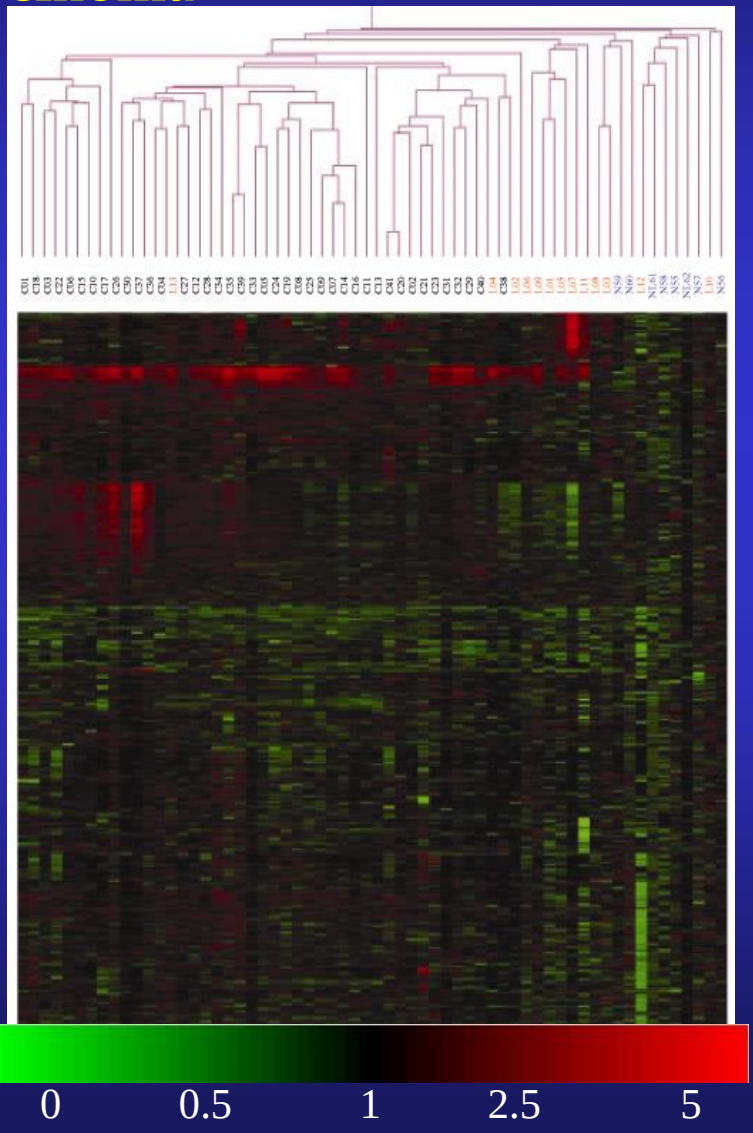
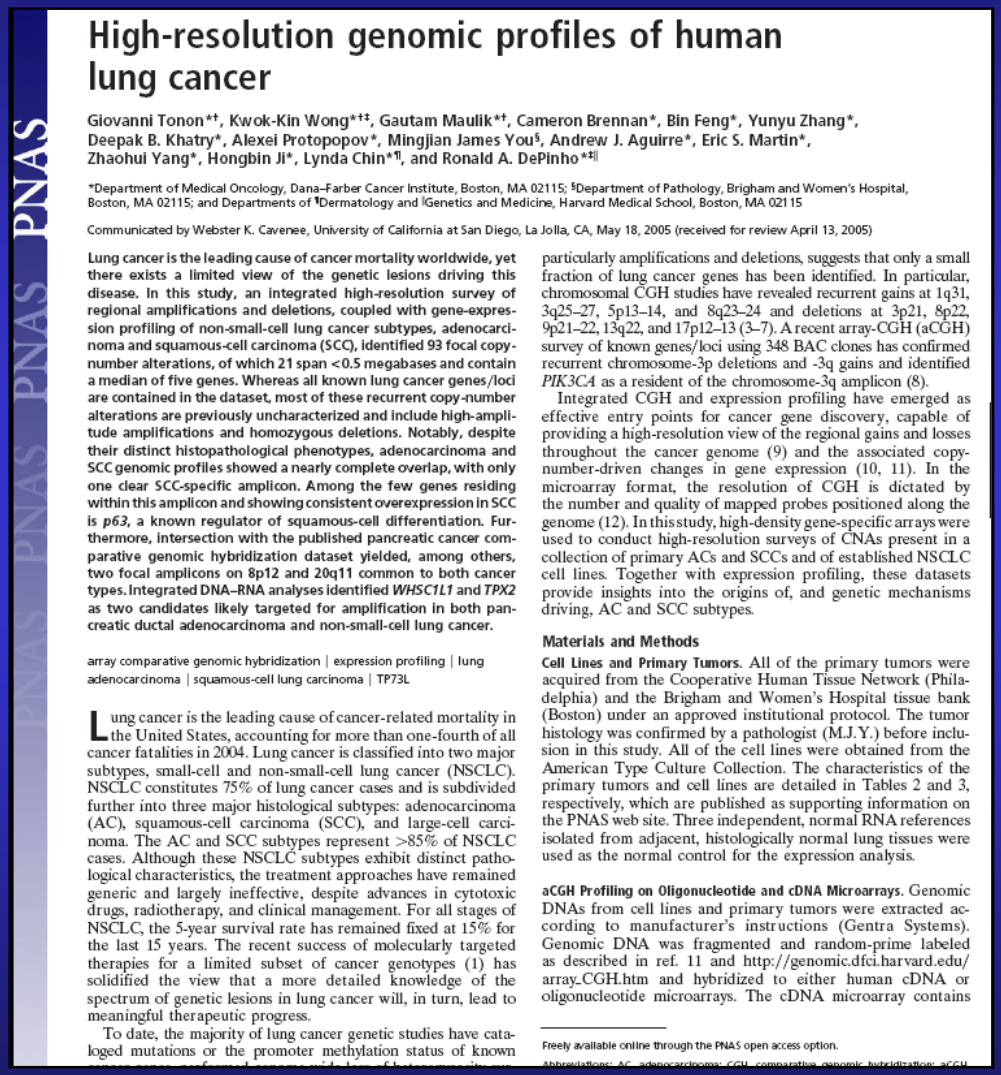


Figure 1

High-resolution Survey of Human Lung Cancer



High-resolution genomic profiles of human lung cancer

Giovanni Tonon*†, Kwok-Kin Wong*†‡, Gautam Maulik*†, Cameron Brennan*, Bin Feng*, Yunyu Zhang*, Deepak B. Khatry*, Alexei Protopopov*, Mingjian James You§, Andrew J. Aguirre*, Eric S. Martin*, Zhaohui Yang*, Hongbin Ji*, Lynda Chin*¶, and Ronald A. DePinho*‡‖

*Department of Medical Oncology, Dana–Farber Cancer Institute, Boston, MA 02115; §Department of Pathology, Brigham and Women's Hospital, Boston, MA 02115; and Departments of ¶Dermatology and ‖Genetics and Medicine, Harvard Medical School, Boston, MA 02115

Communicated by Webster K. Cavenee, University of California at San Diego, La Jolla, CA, May 18, 2005 (received for review April 13, 2005)

Lung cancer is the leading cause of cancer mortality worldwide, yet there exists a limited view of the genetic lesions driving this disease. In this study, an integrated high-resolution survey of regional amplifications and deletions, coupled with gene-expression profiling of non-small-cell lung cancer subtypes, adenocarcinoma and squamous-cell carcinoma (SCC), identified 93 focal copy-number alterations, of which 21 span <0.5 megabases and contain a median of five genes. Whereas all known lung cancer genes/loci are contained in the dataset, most of these recurrent copy-number alterations are previously uncharacterized and include high-amplitude amplifications and homozygous deletions. Notably, despite their distinct histopathological phenotypes, adenocarcinoma and SCC genomic profiles showed a nearly complete overlap, with only one clear SCC-specific amplicon. Among the few genes residing within this amplicon and showing consistent overexpression in SCC is *p63*, a known regulator of squamous-cell differentiation. Furthermore, intersection with the published pancreatic cancer comparative genomic hybridization dataset yielded, among others, two focal amplicons on 8p12 and 20q11 common to both cancer types. Integrated DNA–RNA analyses identified *WHSC1L1* and *TPX2* as two candidates likely targeted for amplification in both pancreatic ductal adenocarcinoma and non-small-cell lung cancer.

array comparative genomic hybridization | expression profiling | lung adenocarcinoma | squamous-cell lung carcinoma | TP73L

Lung cancer is the leading cause of cancer-related mortality in the United States, accounting for more than one-fourth of all cancer fatalities in 2004. Lung cancer is classified into two major subtypes, small-cell and non-small-cell lung cancer (NSCLC). NSCLC constitutes 75% of lung cancer cases and is subdivided further into three major histological subtypes: adenocarcinoma (AC), squamous-cell carcinoma (SCC), and large-cell carcinoma. The AC and SCC subtypes represent >85% of NSCLC cases. Although these NSCLC subtypes exhibit distinct pathological characteristics, the treatment approaches have remained generic and largely ineffective, despite advances in cytotoxic drugs, radiotherapy, and clinical management. For all stages of NSCLC, the 5-year survival rate has remained fixed at 15% for the last 15 years. The recent success of molecularly targeted therapies for a limited subset of cancer genotypes (1) has solidified the view that a more detailed knowledge of the spectrum of genetic lesions in lung cancer will, in turn, lead to meaningful therapeutic progress.

To date, the majority of lung cancer genetic studies have cataloged mutations or the promoter methylation status of known

particularly amplifications and deletions, suggests that only a small fraction of lung cancer genes has been identified. In particular, chromosomal CGH studies have revealed recurrent gains at 1q31, 3q25–27, 5p13–14, and 8q23–24 and deletions at 3p21, 8p22, 9p21–22, 13q22, and 17p12–13 (3–7). A recent array-CGH (aCGH) survey of known genes/loci using 348 BAC clones has confirmed recurrent chromosome-3p deletions and -3q gains and identified *PIK3CA* as a resident of the chromosome-3q amplicon (8).

Integrated CGH and expression profiling have emerged as effective entry points for cancer gene discovery, capable of providing a high-resolution view of the regional gains and losses throughout the cancer genome (9) and the associated copy-number-driven changes in gene expression (10, 11). In the microarray format, the resolution of CGH is dictated by the number and quality of mapped probes positioned along the genome (12). In this study, high-density gene-specific arrays were used to conduct high-resolution surveys of CNAs present in a collection of primary ACs and SCCs and of established NSCLC cell lines. Together with expression profiling, these datasets provide insights into the origins of, and genetic mechanisms driving, AC and SCC subtypes.

Materials and Methods

Cell Lines and Primary Tumors. All of the primary tumors were acquired from the Cooperative Human Tissue Network (Philadelphia) and the Brigham and Women's Hospital tissue bank (Boston) under an approved institutional protocol. The tumor histology was confirmed by a pathologist (M.J.Y.) before inclusion in this study. All of the cell lines were obtained from the American Type Culture Collection. The characteristics of the primary tumors and cell lines are detailed in Tables 2 and 3, respectively, which are published as supporting information on the PNAS web site. Three independent, normal RNA references isolated from adjacent, histologically normal lung tissues were used as the normal control for the expression analysis.

aCGH Profiling on Oligonucleotide and cDNA Microarrays. Genomic DNAs from cell lines and primary tumors were extracted according to manufacturer's instructions (Gentra Systems). Genomic DNA was fragmented and random-prime labeled as described in ref. 11 and http://genomic.dfci.harvard.edu/array_CGH.htm and hybridized to either human cDNA or oligonucleotide microarrays. The cDNA microarray contains

Freely available online through the PNAS open access option.

Abbreviations: AC, adenocarcinoma; CGH, comparative genomic hybridization; aCGH

Application: Copy Number (aCGH)

- Integrating high-resolution gene amplification and deletion data with gene expression profiling
- 93 focal copy-number alterations
  - *74 amplifications*
  - *19 deletions*
  - *median size = 1.53 MB*
  - *21 within highly focal subset with median size < 0.5 MB (spanning ~ 5 genes)*
  - *7 cross-tumor-type candidates*
- Rational starting point for productive gene-discovery efforts

DANA-FARBER CANCER INSTITUTE

Autism Association

ScienceXpress Report

Strong Association of De Novo Copy Number Mutations with Autism

Jonathan Sebat,[1]* B. Lakshmi,[1] Dheeraj Malhotra,[1†] Jennifer Troge,[1†] Christa Lese-Martin,[2] Tom Walsh,[3] Boris Yamrom,[1] Seungtai Yoon,[1] Alex Krasnitz,[1] Jude Kendall,[1] Anthony Leotta,[1] Deepa Pai,[1] Ray Zhang,[1] Yoon-Ha Lee,[1] James Hicks,[1] Sarah J Spence,[4] Annette T. Lee,[5] Kaija Puura,[6] Terho Lehtimäki,[7] David Ledbetter,[2] Peter K. Gregersen,[5] Joel Bregman,[8] James S. Sutcliffe,[9] Vaidehi Jobanputra,[10] Wendy Chung,[10] Dorothy Warburton,[10] Mary-Claire King,[3] David Skuse,[11] Daniel H Geschwind,[12] T. Conrad Gilliam,[13] Kenny Ye,[14] Michael Wigler[1]*

[1]Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA. [2]Department of Human Genetics, Emory University School of Medicine, Atlanta, GA 30322, USA. [3]Department of Medicine and Genome Sciences, University of Washington, Seattle, WA 98195–7720, USA. [4]Pediatrics and Neurodevelopmental Psychiatry Branch, National Institute of Mental Health, National Institutes of Health, Bethesda, MD 20892–1255, USA. [5]Feinstein Institute for Medical Research, North Shore–Long Island Jewish Health System, Manhasset, NY 11030, USA. [6]Department of Child Psychiatry, University of Tampere, Medical School, Tampere, Finland. [7]Department of Clinical Chemistry, University Hospital of Tampere and University of Tampere, Medical School, Tampere, Finland. [8]Fay J. Lindner Center for Autism and Developmental Disorders, North Shore–Long Island Jewish Health System, 4300 Hempstead Turnpike, Bethpage, NY 11714, USA. [9]Center for Molecular Neuroscience, Vanderbilt University, Nashville, TN 37232–8548, USA. [10]Departments of Genetics and Development, and Pediatrics, Columbia University, New York, NY 10027, USA. [11]Behavioural and Brain Sciences Unit, Institute of Child Health, University College London, 30 Guilford Street, London, WCIN 1EH, UK. [12]Interdepartmental Program in the Neurosciences, Program in Neurogenetics, Neurology Department, David Geffen School of Medicine, University of California at Los Angeles, Los Angeles, CA 90095–1769, USA. [13]Department of Human Genetics, The University of Chicago, 920 East 58th Street, Chicago, IL 60637, USA. [14]Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, NY 10461, USA.

*To whom correspondence should be addressed. E-mail: sebat@cshl.edu; wigler@cshl.edu

†These authors contributed equally to this work

We tested the hypothesis that de novo copy number variation (CNV) is associated with autism spectrum disorders (ASDs). We performed comparative genomic hybridization (CGH) on the genomic DNA of patients and unaffected subjects to detect copy number variants not present in their respective parents. Candidate genomic regions were validated by higher-resolution CGH, paternity testing, cytogenetics, fluorescence in situ hybridization, and microsatellite genotyping. Confirmed de novo CNVs were significantly associated with autism ($P = 0.0005$). Such CNVs were identified in 12 out of 118 (10%) of patients with sporadic autism, in 2 out of 77 (2%) of patients with an affected first-degree relative, and in 2 out of 196 (1.0%) of controls. Most de novo CNVs were smaller than microscopic resolution. Affected genomic regions were highly heterogeneous and included mutations of single genes. These findings establish de novo germline mutation as a more significant risk factor for ASD than previously recognized.

Autism spectrum disorders (ASDs) [MIM 209850] are characterized by language impairments, social deficits and repetitive behaviors. The onset of symptoms occurs by the age of 3, and usually requires extensive support for the lifetime of the afflicted. The prevalence of ASD is estimated to be 1 in 166 (*1*), making it a major burden to society.

Genetics plays a major role in the etiology of autism. The concordance rates in monozygotic twins are 70% for autism and 90% for ASD, while the concordance rates in dizygotic twins are 5% and 10% respectively. Previous studies suggest autism displays a high degree of genetic heterogeneity. Efforts to map disease genes using linkage analysis have found evidence for autism loci on 20 different chromosomes. Regions implicated by multiple studies include 1p, 5q, 7q, 15q, 16p, 17q, 19p and Xq (*2*). Moreover, microscopy studies have identified cytogenetic abnormalities in >5% of affected children, involving many different loci on all chromosomes (*3*). In some rare syndromic forms of autism, such as Rett syndrome (*4*) and tuberous sclerosis (*5*), mutations in a single gene have been identified. Otherwise, neither linkage nor cytogenetics has unambiguously identified specific genes involved.

Genetic heterogeneity poses a considerable challenge to traditional approaches for gene mapping (*6*). Some of these limitations are overcome by methods which rely on the direct

Utilized ROMA and Agilent 244k arrays for CGH analysis.

Key conclusions:

- Each de novo CNV was rare in patient population
- Lesions at different loci can contribute to autism
- Clear evidence that the two classes of autism, familial and sporadic, are genetically distinct
- Methods for detecting CNVs genome wide are a powerful alternative to traditional gene mapping approaches

CNV variation in copy number in the human genome

nature Vol 444|23 November 2006|doi:10.1038/nature05329

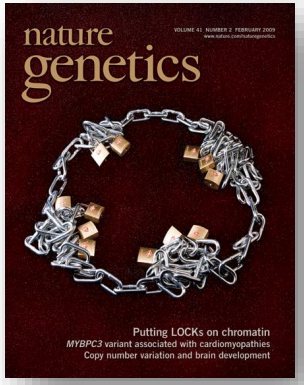
ARTICLES

Global variation in copy number in the human genome

Richard Redon[1], Shumpei Ishikawa[2,3], Karen R. Fitch[4], Lars Feuk[5,6], George H. Perry[7], T. Daniel Andrews[1], Heike Fiegler[1], Michael H. Shapero[4], Andrew R. Carson[5,6], Wenwei Chen[4], Eun Kyung Cho[7], Stephanie Dallaire[7], Jennifer L. Freeman[7], Juan R. González[8], Mònica Gratacòs[8], Jing Huang[4], Dimitrios Kalaitzopoulos[1], Daisuke Komura[3], Jeffrey R. MacDonald[5], Christian R. Marshall[5,6], Rui Mei[4], Lyndal Montgomery[1], Kunihiro Nishimura[2], Kohji Okamura[5,6], Fan Shen[4], Martin J. Somerville[9], Joelle Tchinda[7], Armand Valsesia[1], Cara Woodwark[1], Fengtang Yang[1], Junjun Zhang[5], Tatiana Zerjal[1], Jane Zhang[4], Lluis Armengol[8], Donald F. Conrad[10], Xavier Estivill[8,11], Chris Tyler-Smith[1], Nigel P. Carter[1], Hiroyuki Aburatani[2,12], Charles Lee[7,13], Keith W. Jones[4], Stephen W. Scherer[5,6] & Matthew E. Hurles[1]

Copy number variation (CNV) of DNA sequences is functionally significant but has yet to be fully ascertained. We have constructed a first-generation CNV map of the human genome through the study of 270 individuals from four populations with ancestry in Europe, Africa or Asia (the HapMap collection). DNA from these individuals was screened for CNV using two complementary technologies: single-nucleotide polymorphism (SNP) genotyping arrays, and clone-based comparative genomic hybridization. A total of 1,447 copy number variable regions (CNVRs), which can encompass overlapping or adjacent gains or losses, covering 360 megabases (12% of the genome) were identified in these populations. These CNVRs contained hundreds of genes, disease loci, functional elements and segmental duplications. Notably, the CNVRs encompassed more nucleotide content per genome than SNPs, underscoring the importance of CNV in genetic diversity and evolution. The data obtained delineate linkage disequilibrium patterns for many CNVs, and reveal marked variation in copy number among populations. We also demonstrate the utility of this resource for genetic disease studies.

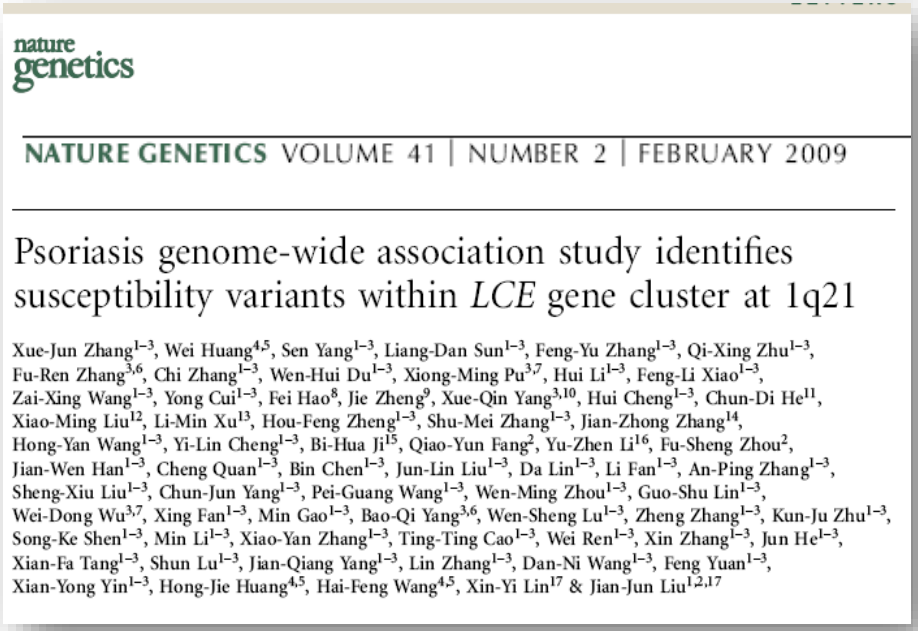
GWAS应用案例：银屑病易感基因LCE研究



研究背景：

银屑病的发病率占世界人口的0.1%～3%，黄种人发病率为0.1%～0.3%。截至2007年，我国银屑病患者已经达到458万人，但发病的基因背景尚不清楚。SBC与安徽医科大学合作，采用全基因组关联分析手段，在大量样本中进行了中国汉族人银屑病易感基因的搜寻和鉴定工作，探讨这类疾病的发病机理和遗传易感性。

Nat Genet. 2009 Feb;41(2):205-10.


nature genetics

NATURE GENETICS VOLUME 41 | NUMBER 2 | FEBRUARY 2009

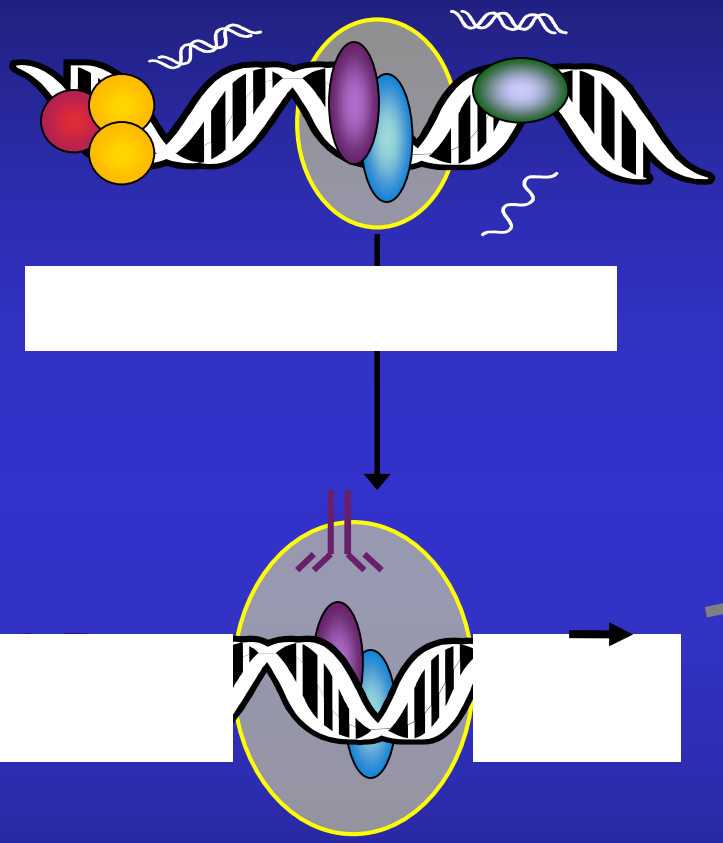
Psoriasis genome-wide association study identifies susceptibility variants within *LCE* gene cluster at 1q21

Xue-Jun Zhang[1–3], Wei Huang[4,5], Sen Yang[1–3], Liang-Dan Sun[1–3], Feng-Yu Zhang[1–3], Qi-Xing Zhu[1–3], Fu-Ren Zhang[3,6], Chi Zhang[1–3], Wen-Hui Du[1–3], Xiong-Ming Pu[3,7], Hui Li[1–3], Feng-Li Xiao[1–3], Zai-Xing Wang[1–3], Yong Cui[1–3], Fei Hao[8], Jie Zheng[9], Xue-Qin Yang[3,10], Hui Cheng[1–3], Chun-Di He[11], Xiao-Ming Liu[12], Li-Min Xu[13], Hou-Feng Zheng[1–3], Shu-Mei Zhang[1–3], Jian-Zhong Zhang[14], Hong-Yan Wang[1–3], Yi-Lin Cheng[1–3], Bi-Hua Ji[15], Qiao-Yun Fang[2], Yu-Zhen Li[16], Fu-Sheng Zhou[2], Jian-Wen Han[1–3], Cheng Quan[1–3], Bin Chen[1–3], Jun-Lin Liu[1–3], Da Lin[1–3], Li Fan[1–3], An-Ping Zhang[1–3], Sheng-Xiu Liu[1–3], Chun-Jun Yang[1–3], Pei-Guang Wang[1–3], Wen-Ming Zhou[1–3], Guo-Shu Lin[1–3], Wei-Dong Wu[3,7], Xing Fan[1–3], Min Gao[1–3], Bao-Qi Yang[3,6], Wen-Sheng Lu[1–3], Zheng Zhang[1–3], Kun-Ju Zhu[1–3], Song-Ke Shen[1–3], Min Li[1–3], Xiao-Yan Zhang[1–3], Ting-Ting Cao[1–3], Wei Ren[1–3], Xin Zhang[1–3], Jun He[1–3], Xian-Fa Tang[1–3], Shun Lu[1–3], Jian-Qiang Yang[1–3], Lin Zhang[1–3], Dan-Ni Wang[1–3], Feng Yuan[1–3], Xian-Yong Yin[1–3], Hong-Jie Huang[4,5], Hai-Feng Wang[4,5], Xin-Yi Lin[17] & Jian-Jun Liu[1,2,17]

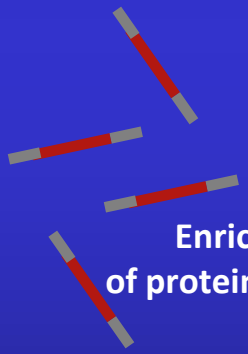
研究结果：

该研究发现了一个新的位于染色体1q21的LCE易感基因变异体，同时验证了已知的两个易感位点：MHC和IL12B。LCE基因编码表皮终末分化角质外膜蛋白，该研究发现LEC基因变异与皮肤表皮细胞更新速度异常有关，其基因变异可以增加患病风险。

研究DNA和蛋白的相互作用

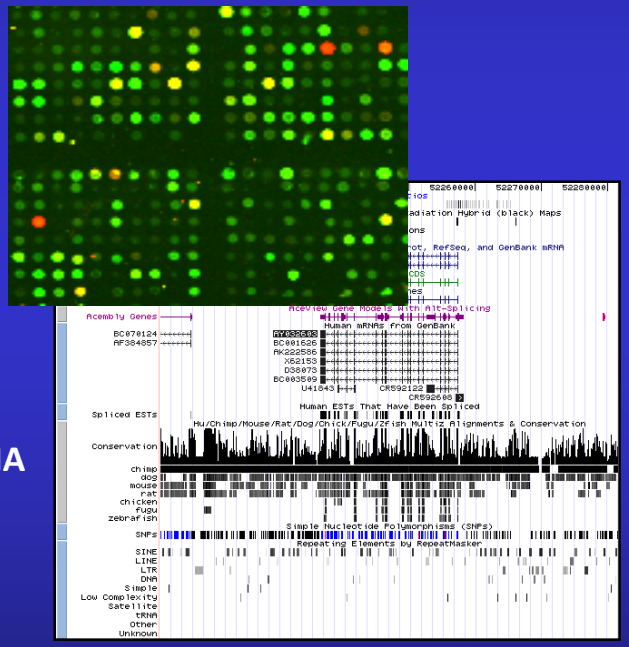


Crosslink protein-DNA complexes
IP chromatin to enrich and capture bound DNA
Purify and amplify DNA fragments



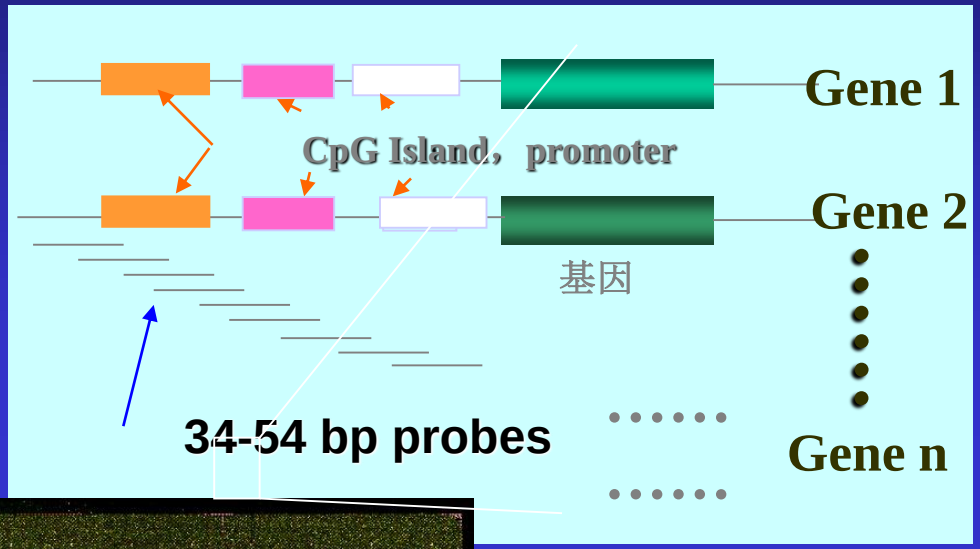
Enriched pool
of protein-bound DNA

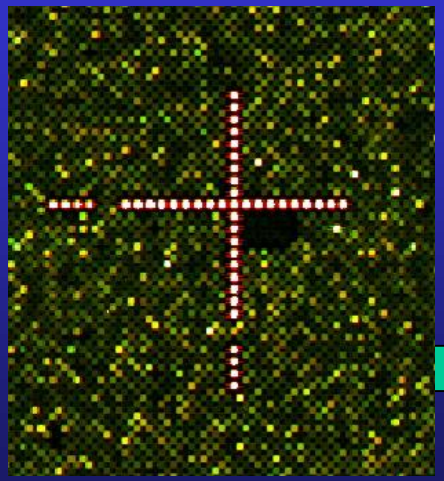
Hybridize to microarray
for detection



ChIP Analytics output

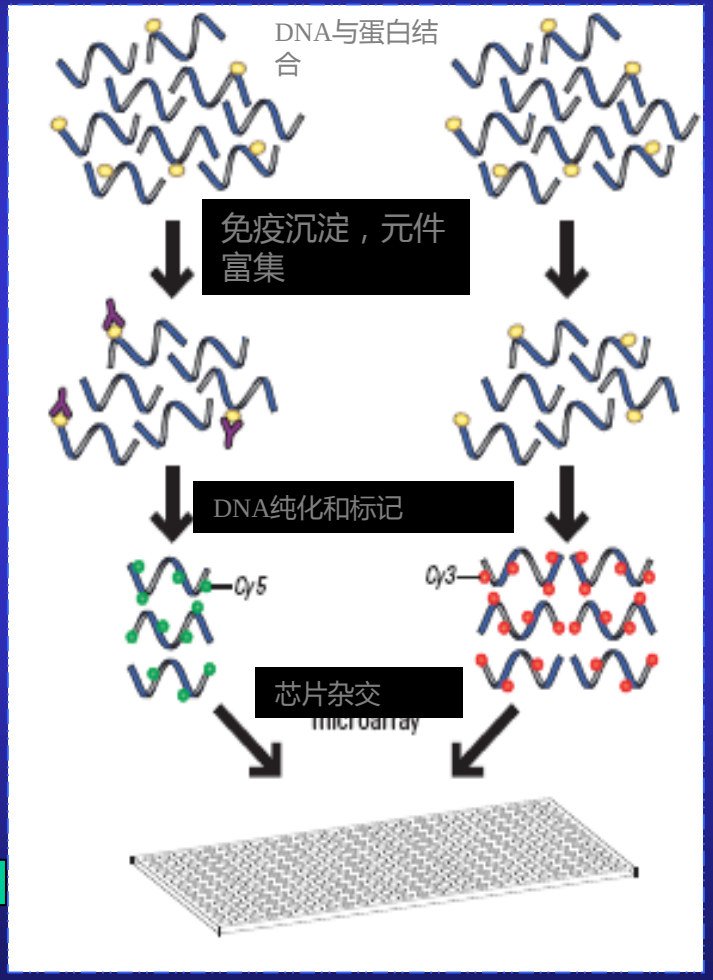
DNA 甲基化芯片





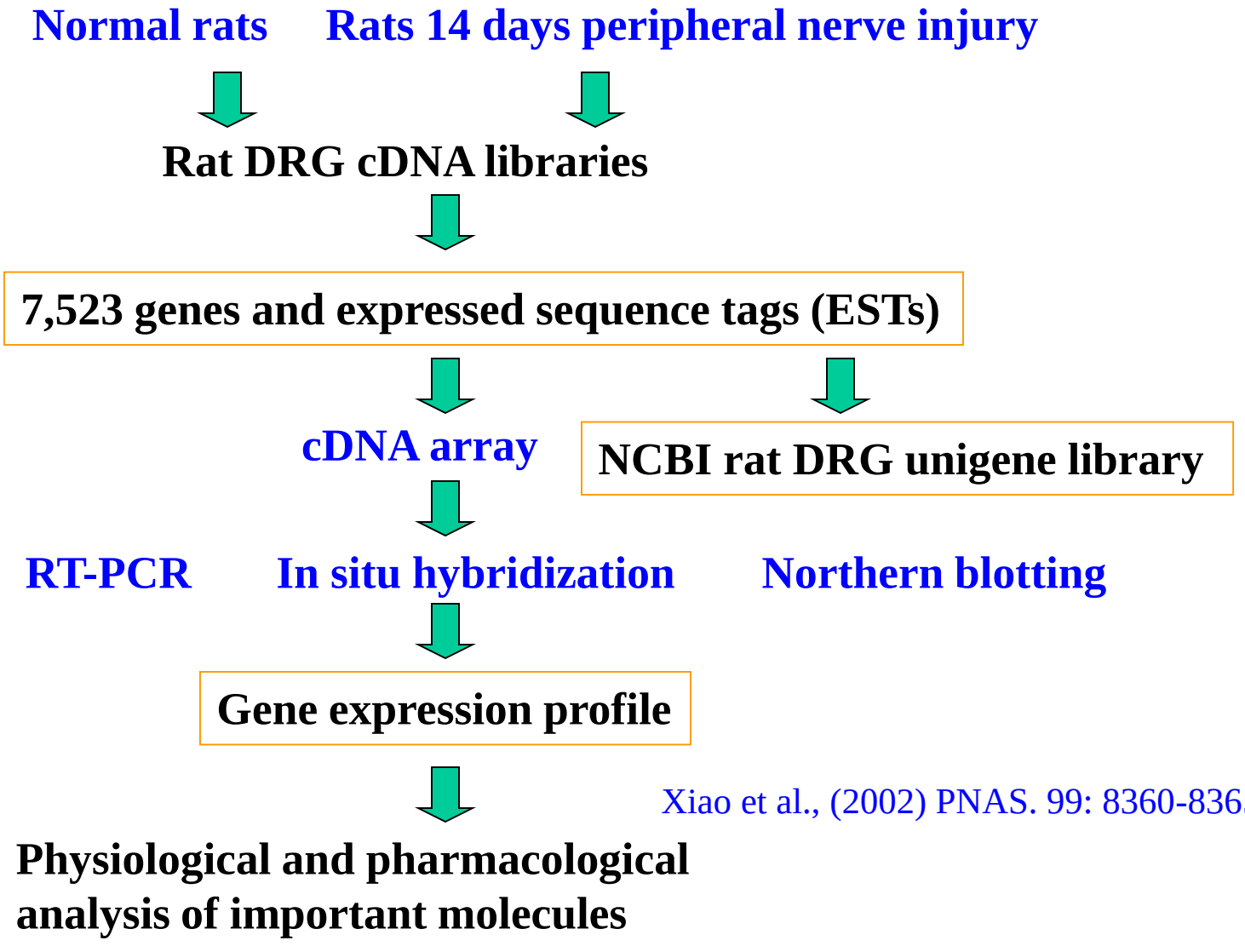
Test

Reference



基因芯片在基因表达水平的应用

Identification of genes expression profile of dorsal root ganglion in the rat peripheral axotomy model of neuropathic pain



Xiao et al., (2002) PNAS. 99: 8360-8365

Markedly regulated 122 genes and 51 novel ESTs

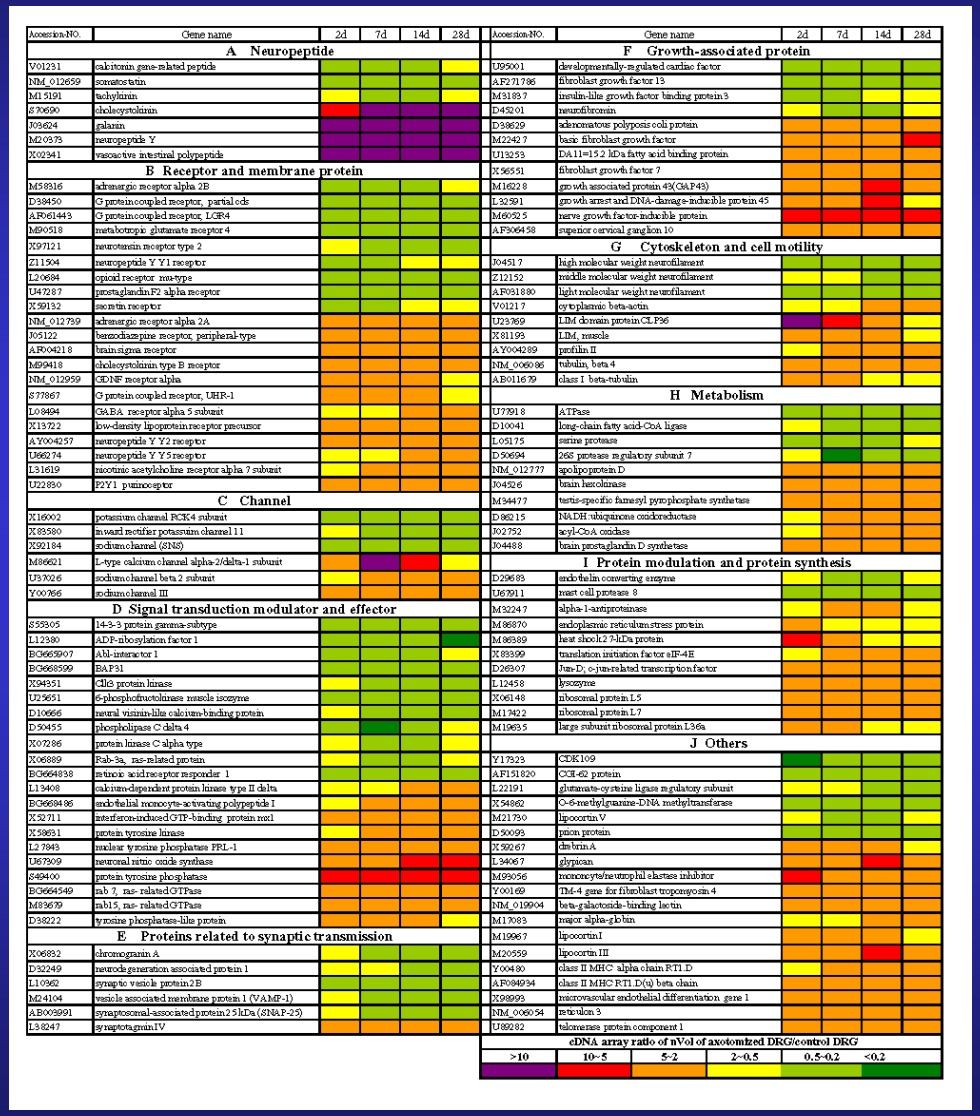
Neuropeptides

Receptors, membrane proteins

Channels

Signal transduction modulators and effectors

Synaptic transmission



Accession-NO.	Gene name	2d	7d	14d	28d
A Neuropeptide					
V01231	calcitonin gene-related peptide				
NM_012659	somatostatin				
M15191	tachykinin				
S70690	cholecystokinin				
J03624	galanin				
M20373	neuropeptide Y				
X02341	vasoactive intestinal polypeptide				
B Receptor and membrane protein					
M58316	adrenergic receptor alpha 2B				
D38450	G protein-coupled receptor, partial cds				
AF061443	G protein-coupled receptor, LGR4				
M90518	metabotropic glutamate receptor 4				
X97121	neurotensin receptor type 2				
Z11504	neuropeptide Y Y1 receptor				
L20684	opioid receptor mu-type				
U47287	prostaglandin F2 alpha receptor				
X59132	secretin receptor				
NM_012739	adrenergic receptor alpha 2A				
J05122	benzodiazepine receptor, peripheral-type				
AF042118	brain sigma receptor				
M99418	cholecystokinin type B receptor				
NM_012959	CIDNF receptor alpha				
S77867	G protein coupled receptor, UHR-1				
L08494	GABA receptor alpha 5 subunit				
X13722	low-density lipoprotein receptor precursor				
AY004257	neuropeptide Y Y2 receptor				
U66274	neuropeptide Y Y5 receptor				
L31619	nicotinic acetylcholine receptor alpha 7 subunit				
U22830	P2Y1 purinoceptor				
C Channel					
X16002	potassium channel RCK4 subunit				
X89580	inward rectifier potassium channel 11				
X92184	sodium channel (SNS)				
M86621	L-type calcium channel alpha-2/delta-1 subunit				
U37026	sodium channel beta 2 subunit				
Y00766	sodium channel III				
D Signal transduction modulator and effector					
S55305	14-3-3 protein gamma-subtype				
L12380	ADP-ribosylation factor 1				
BC065907	Abl-interactor 1				
BC068599	BAP31				
X94351	Clic6 protein kinase				
U25651	6-phosphofructokinase muscle isozyme				
D10666	neural visinin-like calcium-binding protein				
D50455	phospholipase C delta 4				
X07286	protein kinase C alpha type				
X06889	Rab-3a, ras-related protein				
BC066838	retinoic acid receptor responder 1				
L13408	calcium-dependent protein kinase type II delta				
BC069486	endothelial monocyte-activating polypeptide I				
X52711	interferon-induced GTP-binding protein mx1				
X59631	protein tyrosine kinase				
L27843	nuclear tyrosine phosphatase PRL-1				
U67909	neuronal nitric oxide synthase				
S49400	protein tyrosine phosphatase				
BC066549	rab 7, ras-related GTPase				
M83679	rab15, ras-related GTPase				
D38222	tyrosine phosphatase-like protein				
E Proteins related to synaptic transmission					
X06832	chromogranin A				
D32249	neurodegeneration associated protein 1				
L10362	synaptic vesicle protein 2B				
M24104	vesicle associated membrane protein 1 (VAMP-1)				
AB003991	synaptosomal-associated protein 25 kDa (SNAP-25)				
L38247	synaptotagmin IV				

Accession-NO.	Gene name	2d	7d	14d	28d
F Growth-associated protein					
U95001	developmentally-regulated cardiac factor				
AF271786	fibroblast growth factor 13				
M31837	insulin-like growth factor binding protein 3				
D45201	neurofibromin				
D38629	adenomatous polyposis coli protein				
M22427	basic fibroblast growth factor				
U13253	DA11=15.2 kDa fatty acid binding protein				
X56551	fibroblast growth factor 7				
M16228	growth associated protein 43(GAP43)				
L32591	growth arrest and DNA-damage-inducible protein 45				
M60525	nerve growth factor-inducible protein				
AF506458	superior cervical ganglion 10				
G Cytoskeleton and cell motility					
J04517	high molecular weight neurofilament				
Z12152	middle molecular weight neurofilament				
AF031880	light molecular weight neurofilament				
V01217	cytoplasmic beta-actin				
U23769	LIM domain protein CLP36				
X81193	LIM, zyxcis				
AY004289	profilin II				
NM_006086	tubulin, beta 4				
AB011679	class I beta-tubulin				
H Metabolism					
U77918	ATPase				
D10041	long-chain fatty acid-CoA ligase				
L05175	serine protease				
D50694	26S protease regulatory subunit 7				
NM_012777	apolipoprotein D				
J04526	brain hexokinase				
M34477	testis-specific farnesyl pyrophosphate synthetase				
D88215	NADH-ubiquinone oxidoreductase				
J02752	acyl-CoA oxidase				
J04488	brain prostaglandin D synthetase				
I Protein modulation and protein synthesis					
D29685	endothelin converting enzyme				
U67911	mast cell protease 8				
M32247	alpha-1-antiproteinase				
M86870	endoplasmic reticulum stress protein				
M98589	heat shock12 74kDa protein				
X83399	translation initiation factor eIF-4E				
D26307	Jun-D; c-jun-related transcription factor				
L12458	lysozyme				
X06148	ribosomal protein L5				
M17422	ribosomal protein L7				
M19635	large subunit ribosomal protein L36a				
J Others					
Y17323	CDK109				
AF151820	CGI-62 protein				
L22191	glutamate-cysteine ligase regulatory subunit				
X54862	O-6-methylguanine-DNA methyltransferase				
M21730	lipocortin V				
D50093	prion protein				
X59267	drebrin A				
L34067	glypican				
M99056	monocyte/neutrophil elastase inhibitor				
Y00169	TM-4 gene for fibroblast tropomyosin 4				
NM_019904	beta-galactoside-binding lectin				
M17083	major alpha-globin				
M19967	lipocortin I				
M20559	lipocortin III				
Y00480	class II MHC alpha chain RT1.D				
AF049494	class II MHC RT1.D(u) beta chain				
X98993	microvascular endothelial differentiation gene 1				
NM_006054	reticulon 3				
U89282	telomerase protein component 1				

cDNA array ratio of nVol of axotomized DRG/control DRG

>10	10~5	5~2	2~0.5	0.5~0.2	<0.2

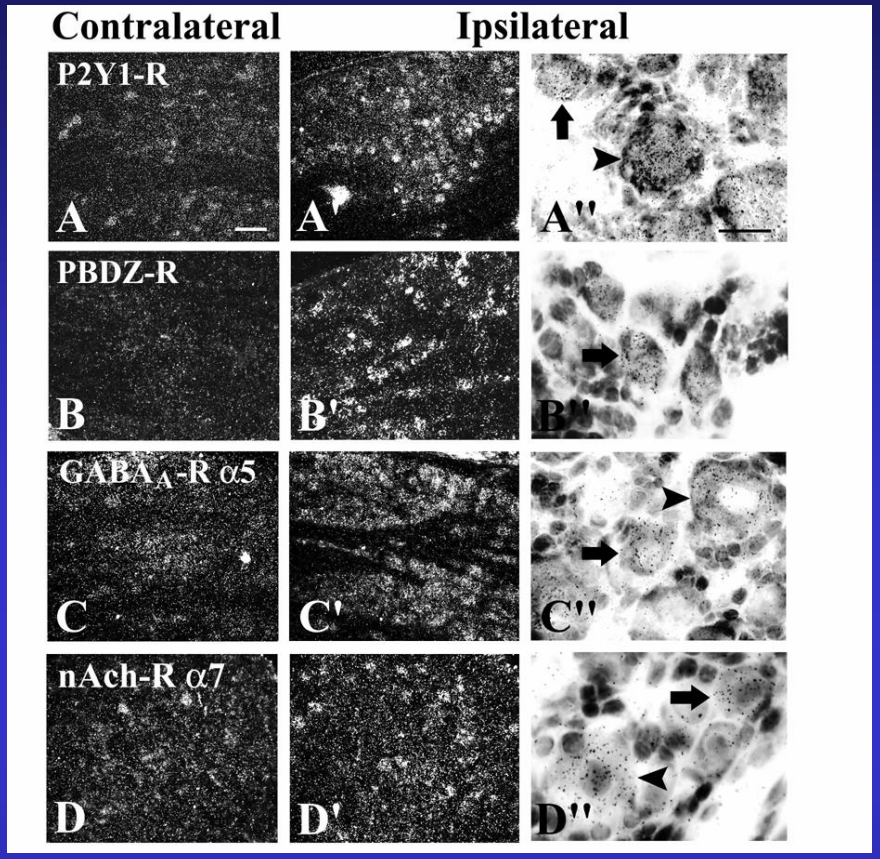
Growth-associated proteins

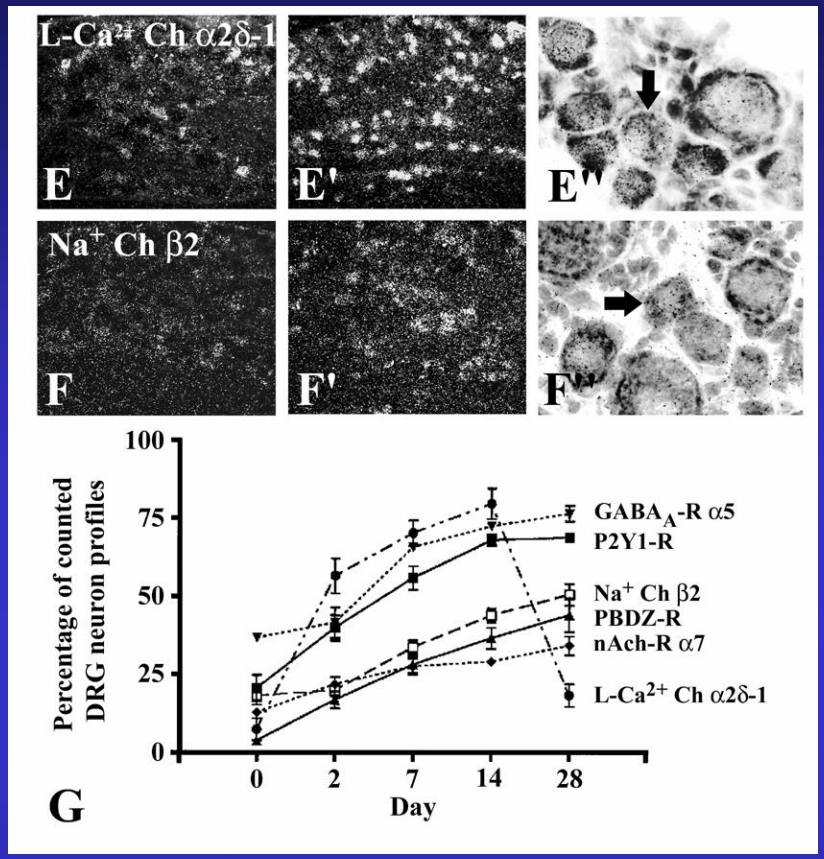
Cytoskeleton and cell motility

Metabolism

Protein modulation and protein synthesis

Others





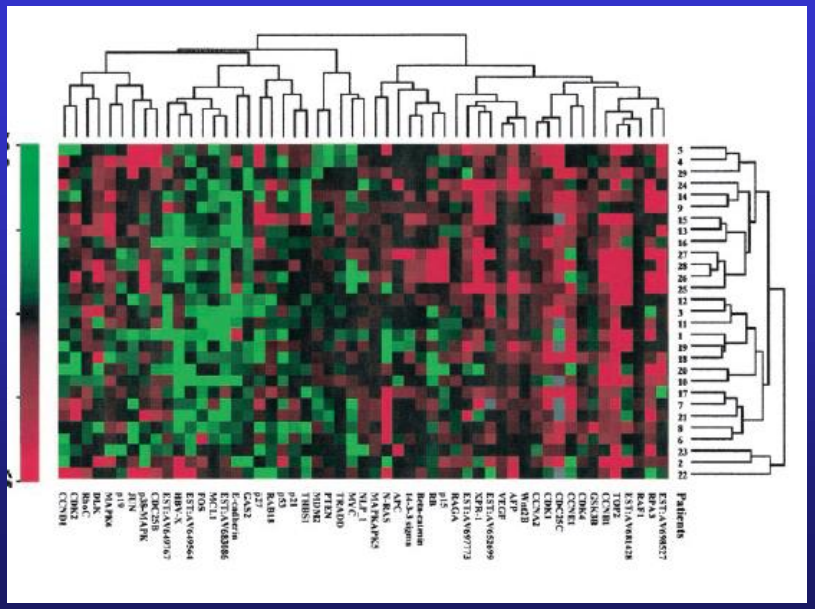
Unexpectedly several antidepressant, antianxiety and anticonvolsant drug target moleculars were up regulated in sensory neuron after axotomy. These provided the candidate genes for developing new drugs and therapeutic methods to treat neuropathic pain with minimal side effects.

Xu XR et al., PNAS, 98(26):15089-15094

Insight into hepatocellular carcinogenesis at transcriptome level by comparing gene expression profiles of hepatocellular carcinoma with those of corresponding noncancerous liver

Xiang-Ru Xu*†, Jian Huang*, Zhi-Gang Xu*, Bin-Zhi Qian*, Zhi-Dong Zhu*, Qing Yan*, Ting Cai*, Xin Zhang*, Hua-Sheng Xiao*, Jian Qu*, Feng Liu*, Qiu-Hua Huang*, Zhi-Hong Cheng*, Neng-Gan Li*, Jian-Jun Du*, Wei Hu*, Kun-Tang Shen*, Gang Lu*, Gang Fu*, Ming Zhong*, Shu-Hua Xu*, Wen-Yi Gu*, Wei Huang*, Xin-Tai Zhao‡, Geng-Xi Hu§, Jian-Ren Gu‡, Zhu Chen*†¶, and Ze-Guang Han*‖

*Chinese National Human Genome Center at Shanghai, 351 Guo Shou-Jing Road, Shanghai 201203, China; †Shanghai Institute of Hematology, Rui Jin Hospital, 197 Rui Jin Road II, Shanghai 200025, China; ‡National Laboratory for Oncogenes and Related Genes, Shanghai Cancer Institute, 25 Lane 2200, Xietu Road, Shanghai 200032, China; §Shanghai Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 320 Yueyang Road, Shanghai 200031, China; and ¶Morgan-Tan International Center for Life Science, Institute of Genetics, Fudan University, 220 Han Dan Road, Shanghai 200433, China



Identified 2,253 genes or ESTs as candidates with differential expression by EST sequencing and cDNA microarray. Many genes involved in cell cycle regulation such as cyclins, cyclin-dependent kinases, andcell cycle negative regulators were deregulated in most patients with HCC. Aberrant expression of the Wnt--catenin pathway and enzymes for DNA replication also could contribute to the pathogenesis of HCC.

Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring

T. R. Golub,[1,2*†] D. K. Slonim,[1†] P. Tamayo,[1] C. Huard,[1] M. Gaasenbeek,[1] J. P. Mesirov,[1] H. Coller,[1] M. L. Loh,[2] J. R. Downing,[3] M. A. Caligiuri,[4] C. D. Bloomfield,[4] E. S. Lander[1,5*]

Science, 1999, 286:531-537

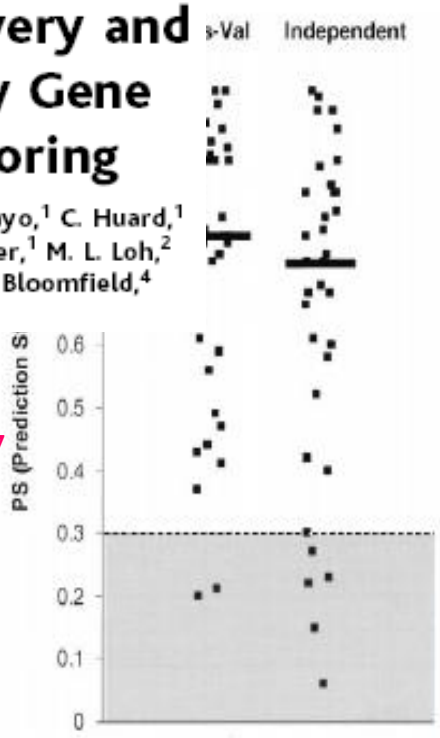
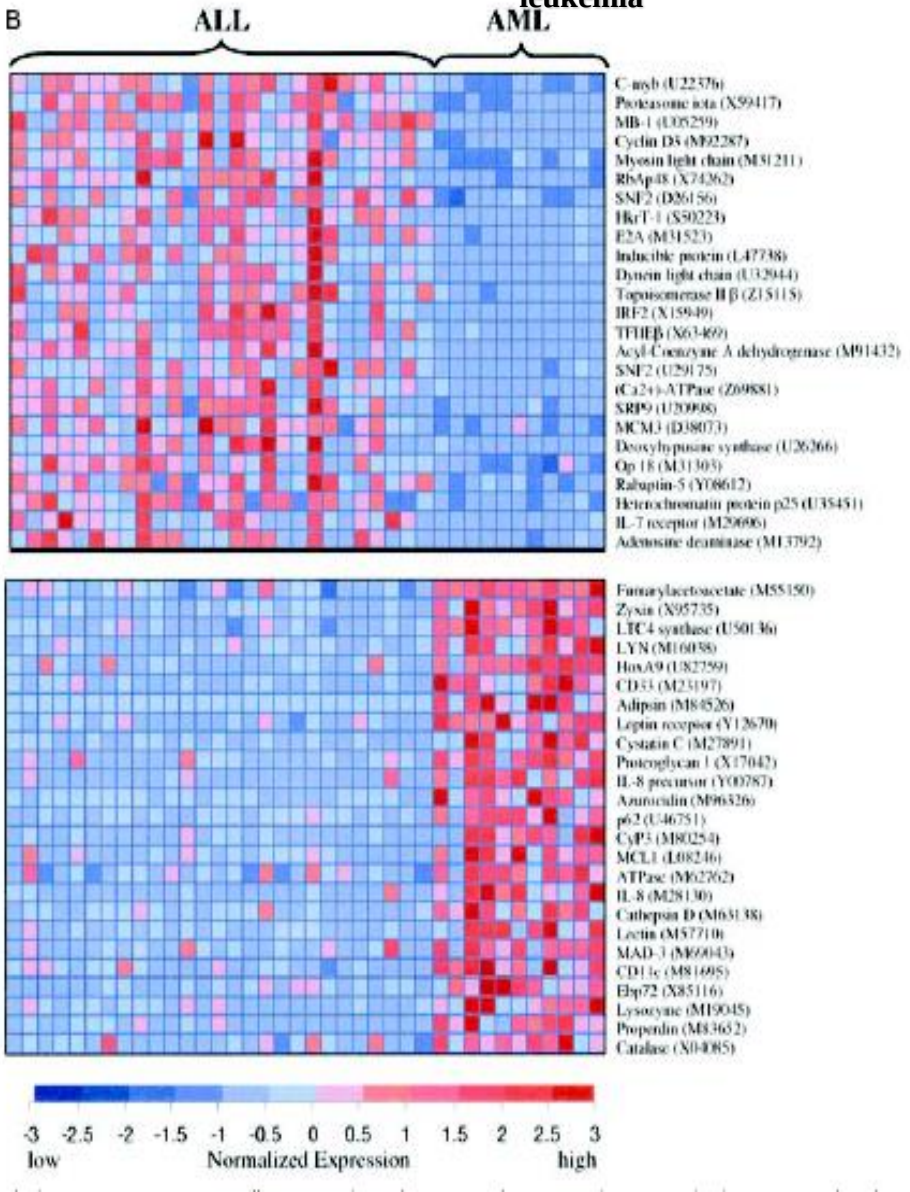


Fig. 3. (A) Prediction strengths. The scatter-plots show the prediction strengths (PSs) for ...d on the ...s is de-...with PS ...) Genes ...) genes ...ML class ...onds to ...ding to Expres-...d across the samples such that the mean is 0 and the SD is 1. Expression levels greater than the mean are shaded in red, and those below the mean are shaded in blue. The scale indicates SDs above or below the mean. The top panel shows

acute myeloid leukemia　　acute lymphoblastic leukemia



不同类型疾病的基因表达特征研究

Gene Expression Signatures for Breast Cancer

The New England Journal of Medicine

VOLUME 347 DECEMBER 19, 2002 NUMBER 25

A GENE-EXPRESSION SIGNATURE AS A PREDICT
IN BREAST CANCER

MARC J. VAN DE VIJVER, M.D., PH.D., YUDONG D. HE, PH.D., LAURA J. VAN 'T V
AUGUSTINUS A.M. HART, M.SC., DORIEN W. VOSKUIL, PH.D., GEORGE J. SCHREIBER,
CHRIS ROBERTS, PH.D., MATTHEW J. MARTON, PH.D., MARK PARRISH, DOU
ANNUSKA GLAS, PH.D., LEONIE DELAHAYE, TONY VAN DER VELDE, HARRY
SJOERD RODENHUIS, M.D., PH.D., EMIEL T. RUTGERS, M.D., PH.D., STEPH
AND RENÉ BERNARDS, PH.D.

ABSTRACT

Background A more accurate means of prognostication in breast cancer will improve the selection of patients for adjuvant systemic therapy.

Methods Using microarray analysis to evaluate our previously established 70-gene prognosis profile, we classified a series of 295 consecutive patients with primary breast carcinomas as having a gene-expression signature associated with either a poor prognosis or a good prognosis. All patients had stage I or II breast cancer and were younger than 53 years old; 151 had lymph-node–negative disease, and 144 had lymph-node–positive disease. We evaluated the predictive power of the prognosis profile using univariable and multivariable statistical analyses.

Results Among the 295 patients, 180 had a poor-prognosis signature and 115 had a good-prognosis signature, and the mean (±SE) overall 10-year survival rates were 54.6±4.4 percent and 94.5±2.6 percent, respectively. At 10 years, the probability of remaining free of distant metastases was 50.6±4.5 percent in the group with a poor-prognosis signature and 85.2±4.3 percent in the group with a good-prognosis signature. The estimated hazard ratio for distant metastases in the group with a poor-prognosis signature, as compared with the group with the good-prognosis signature, was 5.1 (95 percent confidence interval, 2.9 to 9.0; P<0.001). This ratio remained significant when the groups were analyzed according to lymph-node status. Multivariable Cox regression analysis showed that the prognosis profile was a strong independent factor in predicting disease outcome.

Conclusions The gene-expression profile we studied is a more powerful predictor of the outcome of disease in young patients with breast cancer than standard systems based on clinical and histologic criteria. (N Engl J Med 2002;347:1999-2009.)

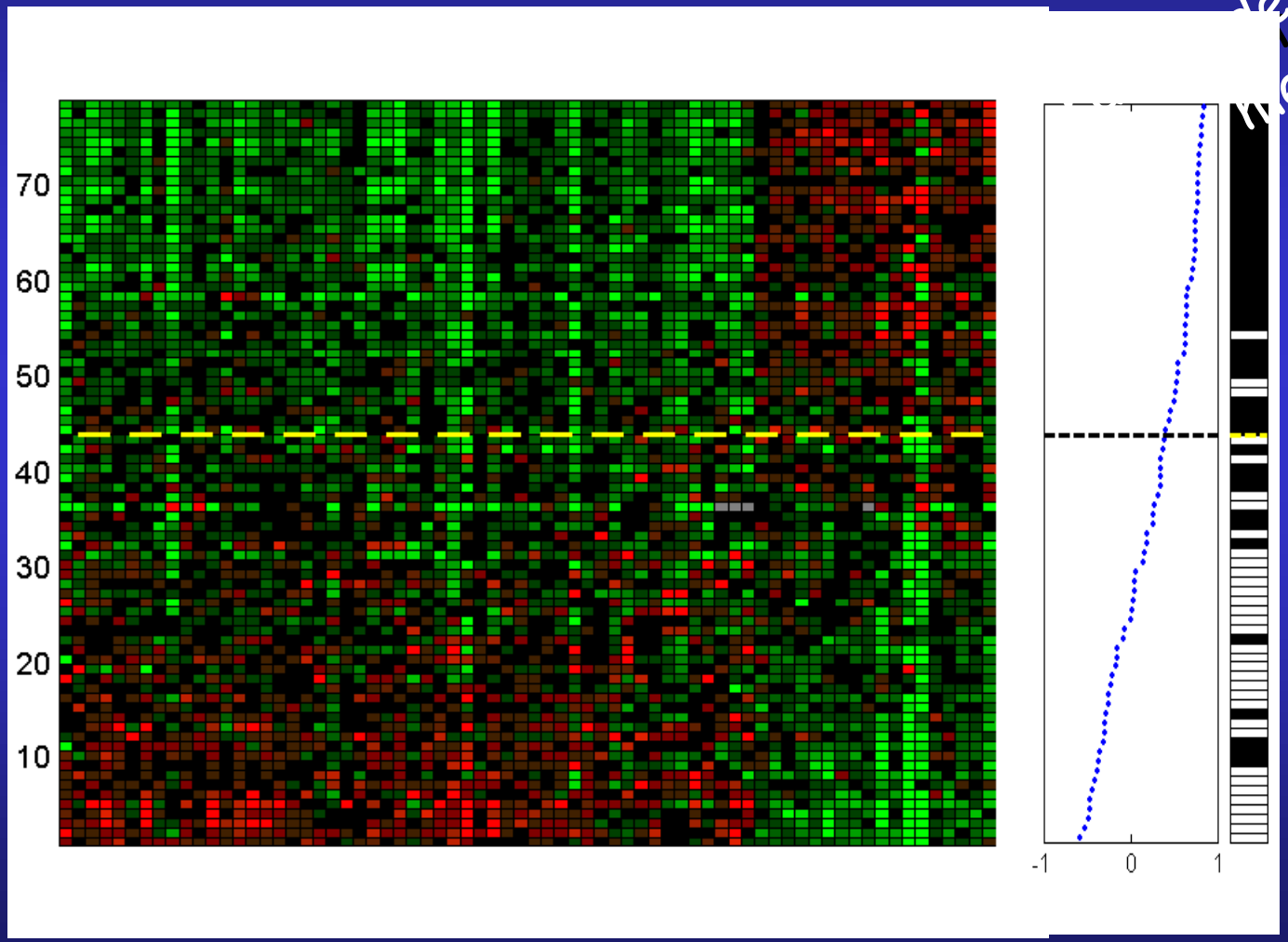
Laura J. Van't Veer, PhD
COO and Co-founder, Agen
Head, Family Cancer Center
Netherlands Cancer Institute

Agendia™

Application: Gene Expression

- Identifying gene expression signature that reflects biological behavior of a tumor
  - *70-gene prognostic profile tested*
  - *295 breast cancer samples*
  - *outperformed all clinical variables for predicting patient survival*
- Microarray classifiers to direct customized therapy
- Starting point for targeted and rational drug development

Supervised Classification Prognosis



good signature

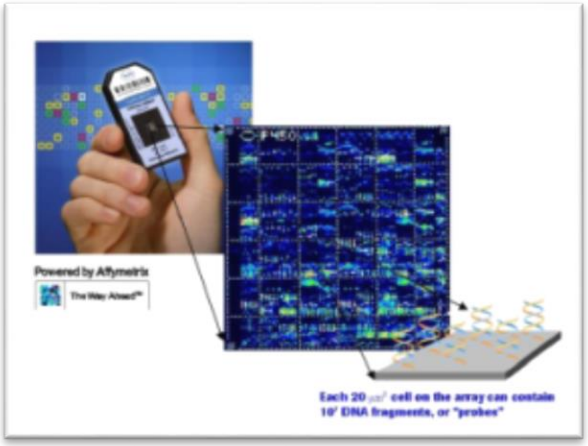
3 false negative

poor signature

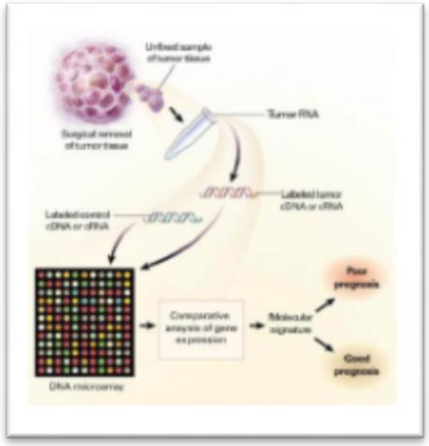
12 false positive

走向临床诊断的基因芯片

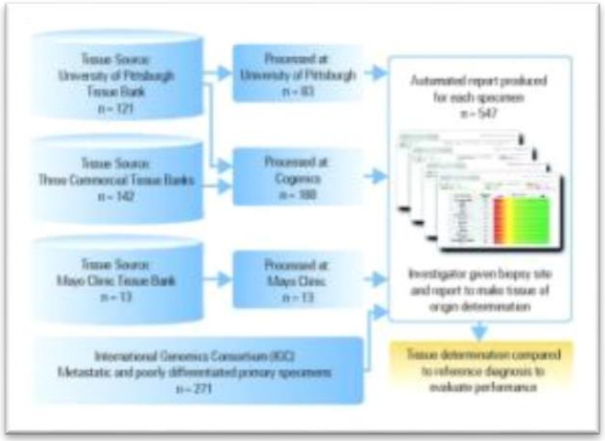
通过FDA或CE审核的诊断型芯片



2004年12月：罗氏公司 AmpliChip CYP450 Test 芯片诊断产品通过FDA审核。



2007年2月：荷兰Agendia公司乳腺癌预后诊断芯片MammaPrint通过FDA审核。



2010年：美国Pathwork公司的TOO芯片（Tissue of Origin Test）通过FDA审核。

2004年，Affymetrix专门用于体外诊断芯片的仪器系统——GeneChip System 3000Dx（GCS 3000Dx）先后通过CE和FDA的审核。