

基因芯片

数据分析与处理

李瑶 主编



化学工业出版社
现代生物技术与医药科技出版中心

基因芯片数据分析与处理

李 瑶 主编



化学工业出版社
现代生物技术与医药科技出版中心

· 北 京 ·

图书在版编目 (CIP) 数据

基因芯片数据分析与处理/李瑶主编. —北京: 化学工业出版社, 2006. 4
ISBN 7-5025-8564-8

I. 基… II. 李… III. ①基因-芯片-数据-分析
②基因-芯片-数据处理 IV. Q78

中国版本图书馆 CIP 数据核字 (2006) 第 037764 号

基因芯片数据分析与处理

李 瑶 主编

责任编辑: 周 旭

文字编辑: 陈 曦

责任校对: 李 林

封面设计: 胡艳玮

*

化 学 工 业 出 版 社 出版发行
现代生物技术与医药科技出版中心

(北京市朝阳区惠新里 3 号 邮政编码 100029)

购书咨询: (010)64982530

(010)64918013

购书传真: (010)64982630

<http://www.cip.com.cn>

*

新华书店北京发行所经销

北京永鑫印刷有限责任公司印刷

三河市东柳装订厂装订

开本 787mm×1092mm 1/16 印张 20½ 彩插 4 字数 522 千字

2006 年 7 月第 1 版 2006 年 7 月北京第 1 次印刷

ISBN 7-5025-8564-8

定 价: 49.00 元

版权所有 违者必究

该书如有缺页、倒页、脱页者, 本社发行部负责退换

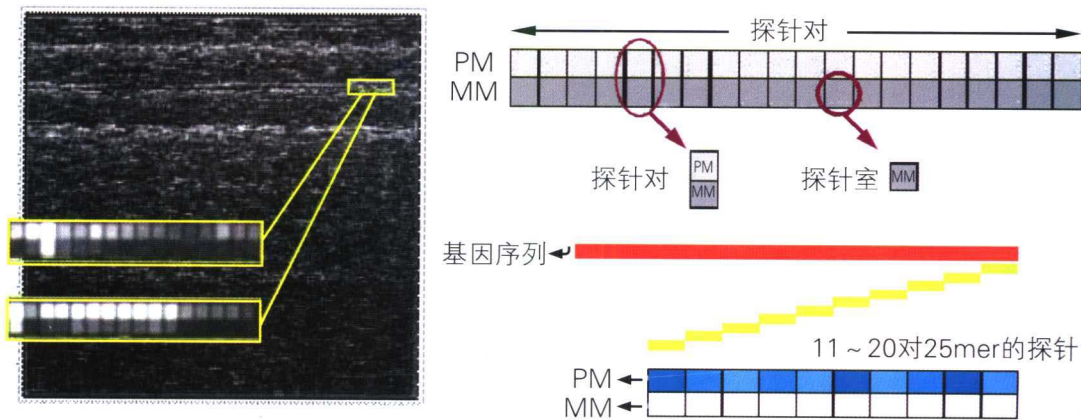


图 2-5 PM-MM探针设计方案

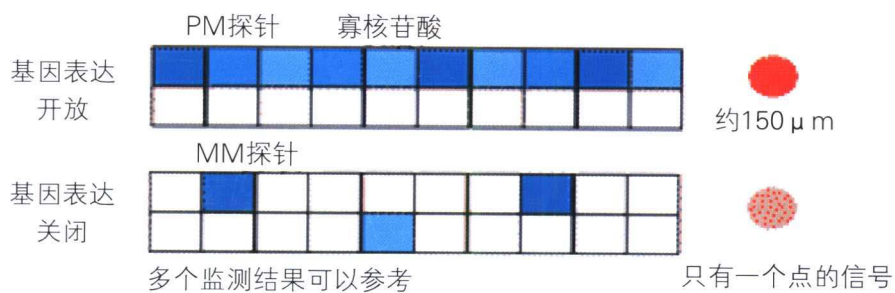
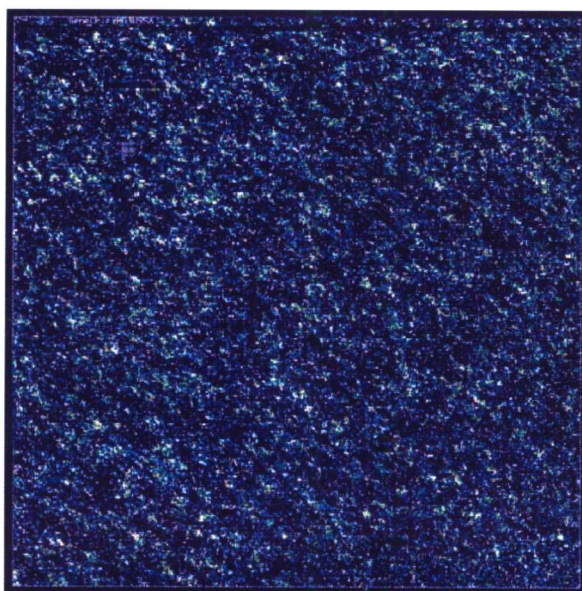
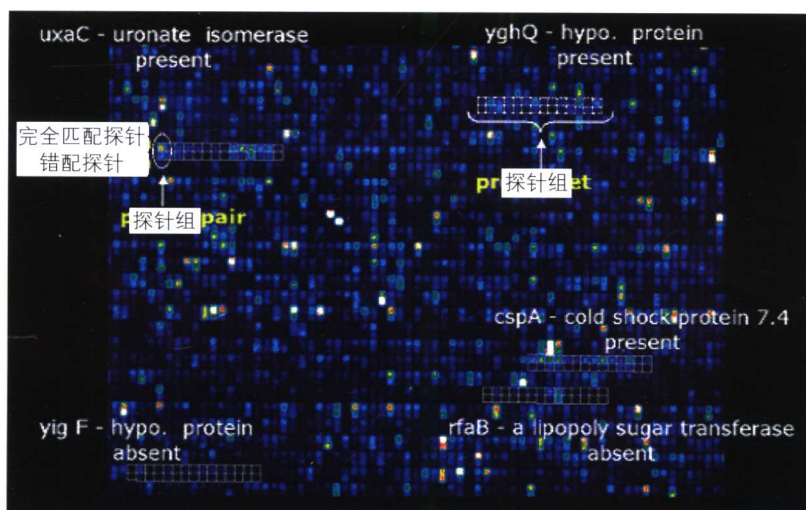


图 2-6 PM-MM的多对探针的结果与单个探针的结果比较



(a) 整体图像



(b) 局部放大

图 2-16 Affymetrix公司的高密度基因芯片图像

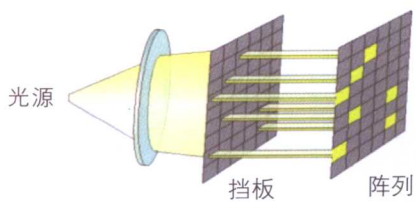
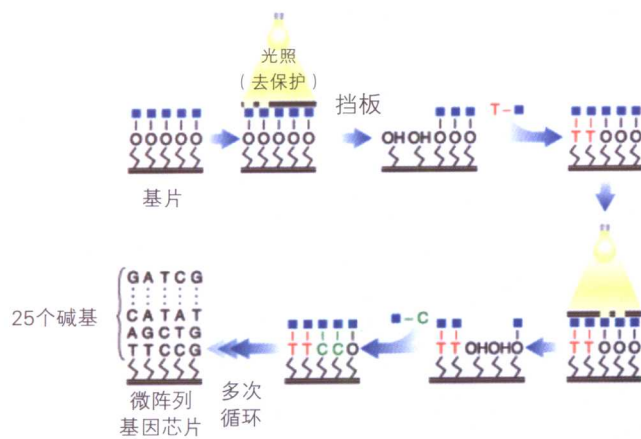
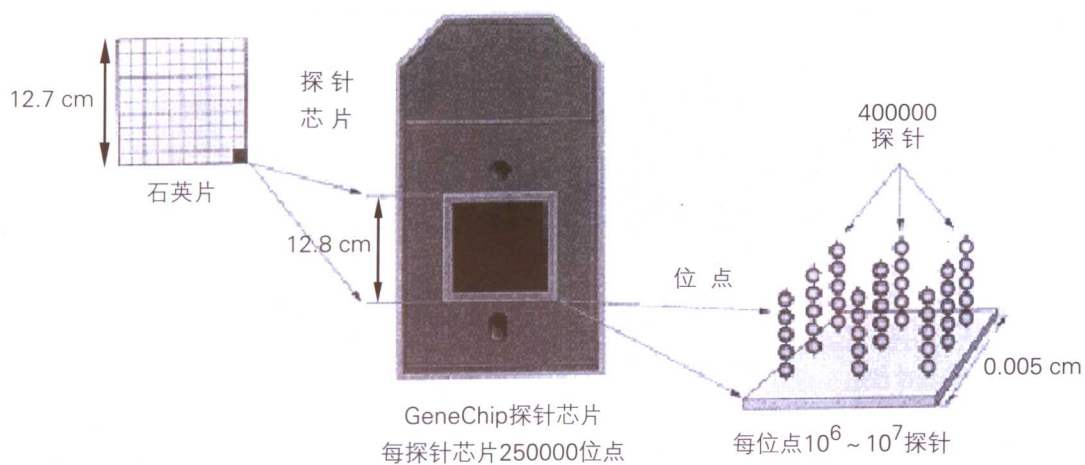


图 2-17 光导原位合成原理图

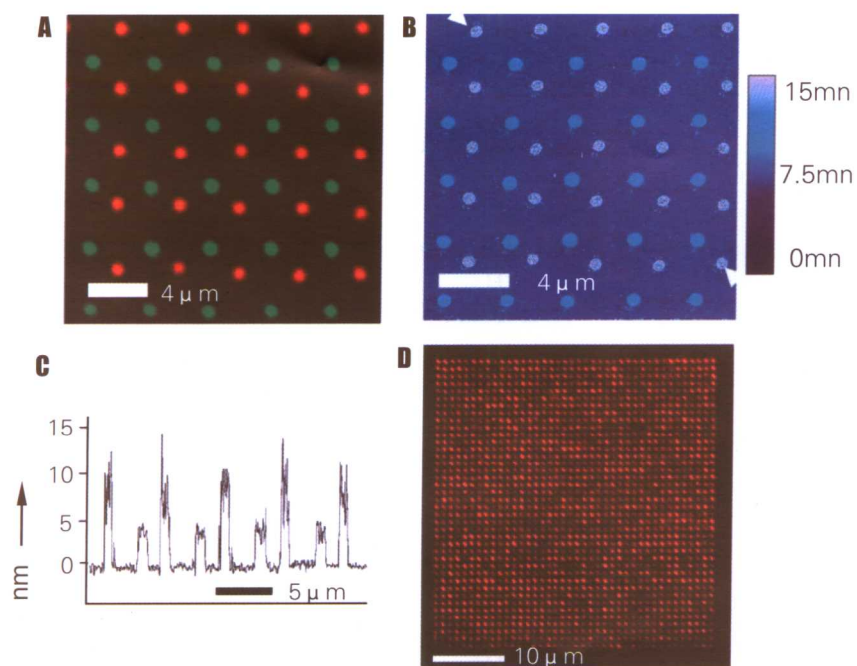


图 2-22 DPN制备的多 DNA “墨水” 的图案

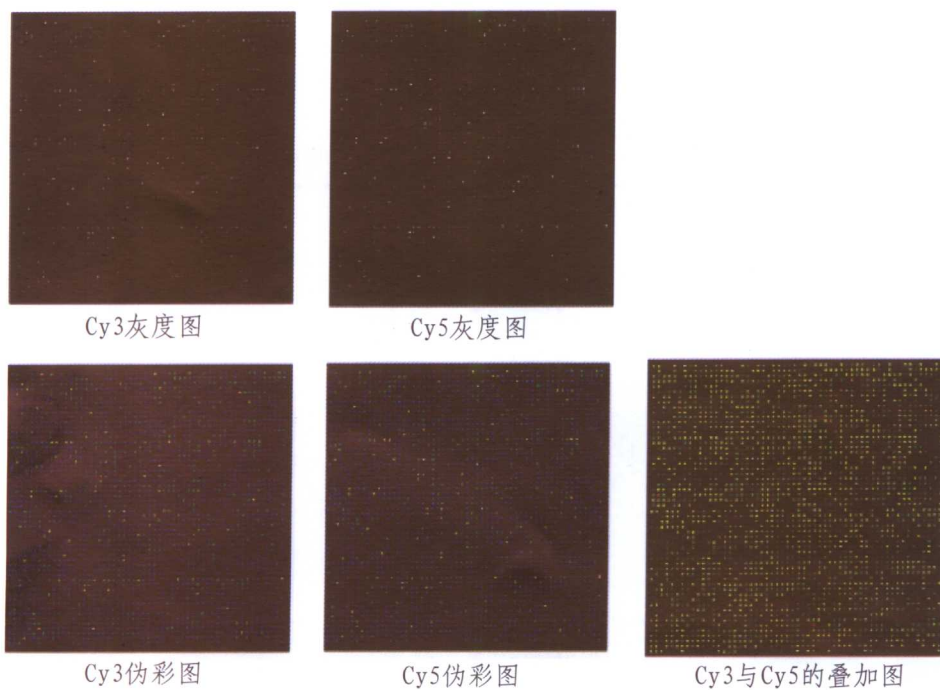


图 5-4 cDNA芯片的双色荧光图像

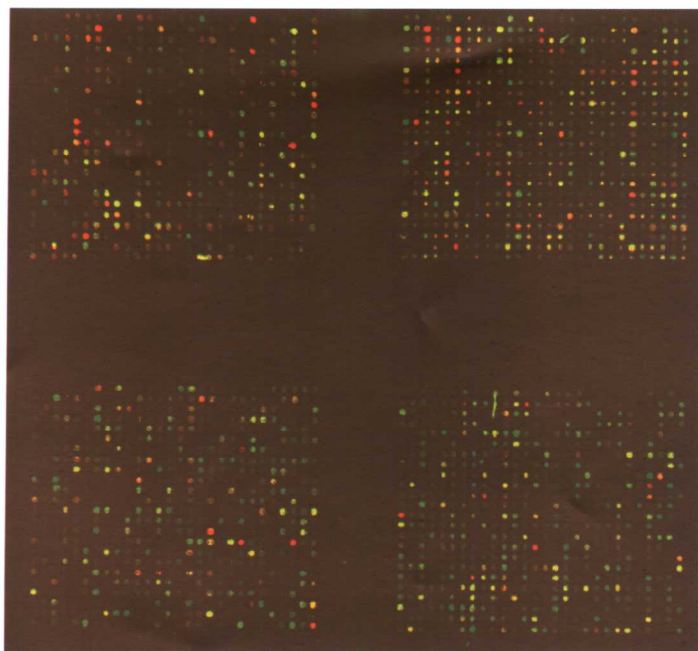


图 5-5 cDNA芯片的典型图像（图像中有四个子格子）

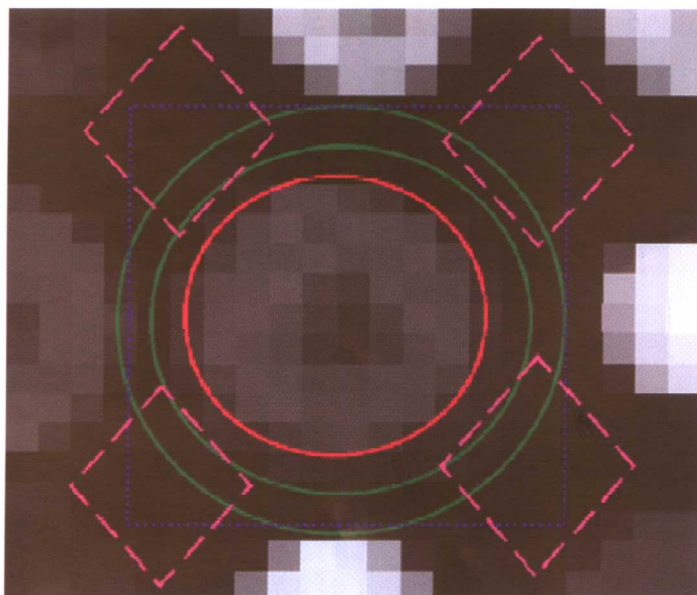


图 5-6 不同的背景值确定方法

红色圈所围区域是信号点框，其他颜色线条表示不同软件计算背景值时所选区域不同。ScanAlyze把蓝色矩形内且不在红圈内的像素作为背景信号；QuantArray把两个同心绿色圆间的像素作为背景值；GenePix则把4个粉色区域作为背景区域，这4个区域是芯片中的低谷，与围绕的周围的4个信号点距离最远

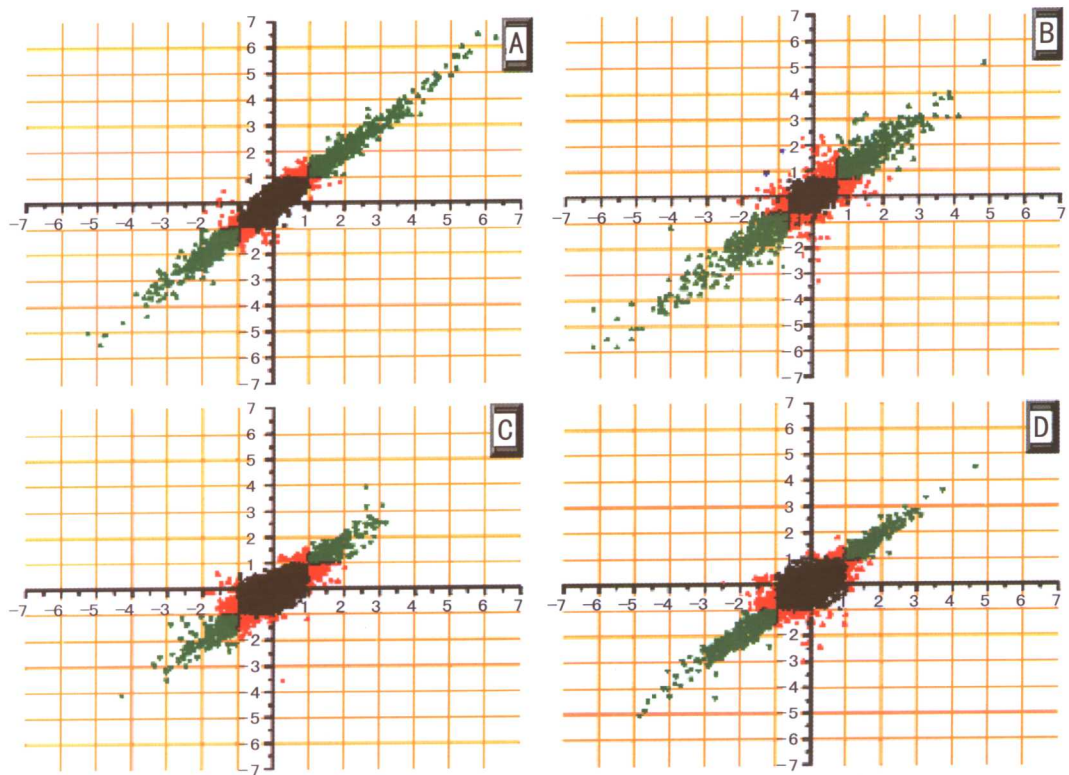


图 8-4 正向重复实验比值的散点图

横坐标为一次实验的Cy5/Cy3值的对数；纵坐标为另一次实验的Cy5/Cy3值的对数。

A: 同一张芯片的相同重复基因；B: 不同批次的两张芯片，用RNA同时标记后分别杂交；C: 相同批次两张芯片，RNA同时标记后分别杂交；D: 相同批次两张芯片，RNA分别标记后分别杂交。黑色的点代表两次实验中比值均小于阈值的基因，红色的点代表两次实验中有一次比值小于阈值的基因，绿色的点代表两次实验中比值都大于阈值的基因，这四组实验比值对数值的相关系数均在0.7以上。

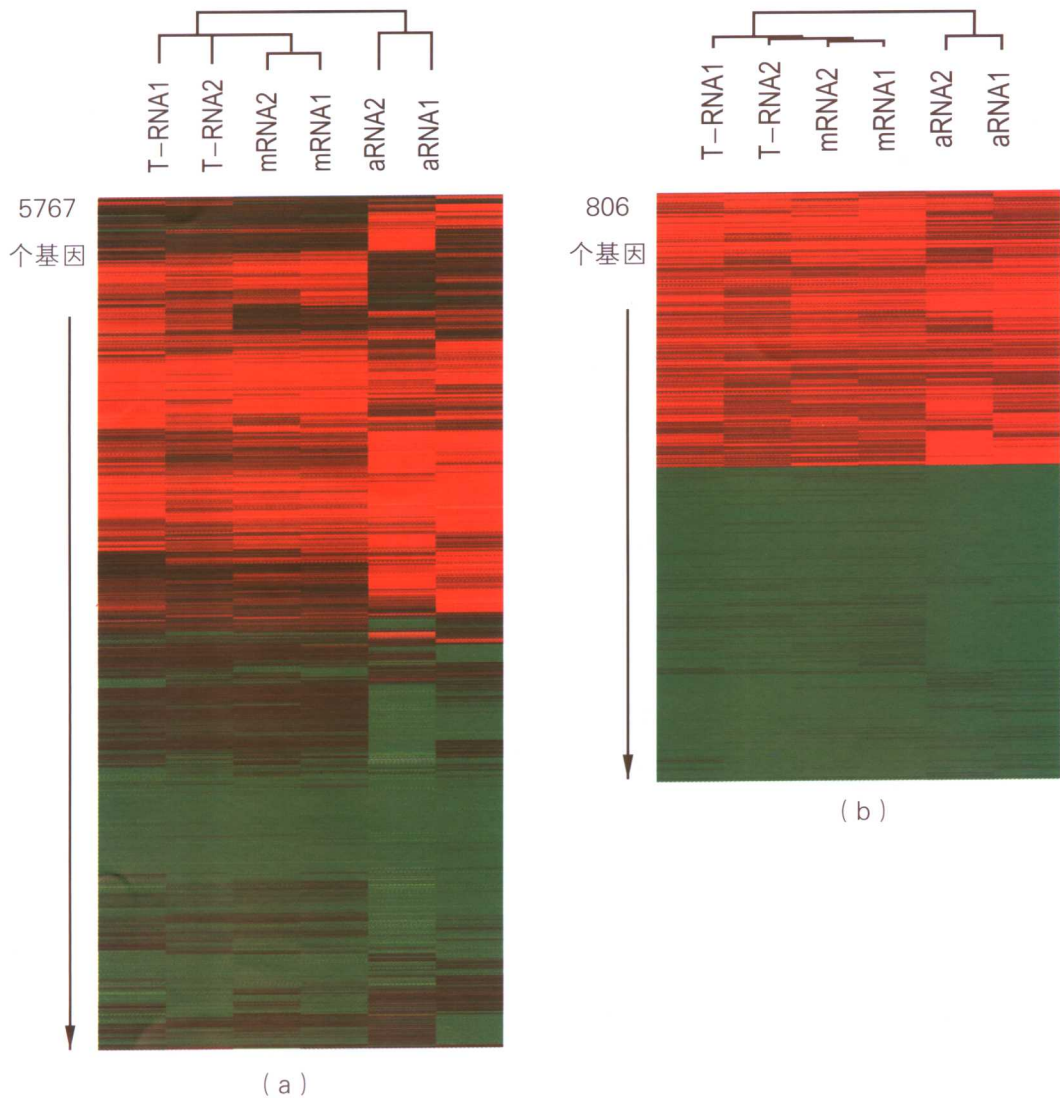


图 8-16 三种标记方法共六次实验的表达谱数据的层级聚类结果(hierarchical cluster)

每一行代表一个基因，每一列代表一次芯片杂交实验。黄色和绿色分别代表肝癌种上调表达和下调表达的基因，黑色代表没有表达差异的基因，灰色代表由于各种原因缺失的数据

(a) 用在一种以上的标记方法中有差异的5767个基因进行聚类；

(b) 用在在三种标记方法的六次实验中都有差异的806个基因进行聚类

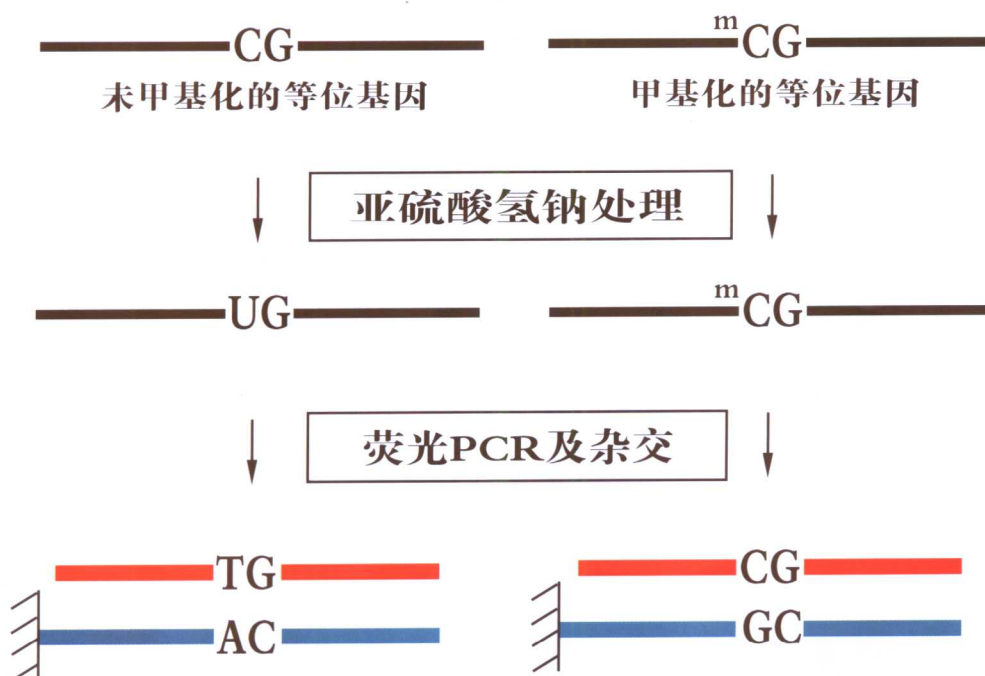


图14-5 甲基化特异性寡核苷酸芯片方法原理示意图

蓝线代表固定在玻片上的等位基因特异性探针，红线代表荧光标记的靶基因，黑线代表未标记的靶基因

《基因芯片数据分析与处理》 编写人员

主 编 李 瑶

副 主 编 贺 佳 陈一东 曹勇伟 许 可

参编人员 (以姓氏笔画为序)

付旭平 (复旦大学生命学院遗传所)

刘三震 (上海博星基因芯片有限责任公司)

许 可 (美国 Pfizer 公司)

孙美倩 (复旦大学生命学院遗传所)

李 平 (美国 Monsanto 公司)

李 瑶 (复旦大学生命学院遗传所)

杨 晓 (美国 Monsanto 公司)

吴 海 (复旦大学生命学院遗传所)

吴 骋 (第二军医大学统计教研室)

陈一东 (National Human Genome Research Institute,
National Institute of Health, 美国)

贺 佳 (第二军医大学统计教研室)

贺宪民 (第二军医大学统计教研室)

曹勇伟 (美国 Monsanto 公司)

魏 庆 (复旦大学生命学院遗传所)

前 言

基因芯片技术是交叉性很强的学科，它不仅体现在实验技术本身所涉及的众多领域，如物理、化学、材料、生物等，而且在数据挖掘方面需要数学、统计学、计算机科学、生物学和医学方面的专业人才，需要不同学科的研究者共同努力，尤其需要生物学家和计算科学家通过“双边对话”共同完成实验设计、实验方法改进以及数据分析和阐明。生物学家必须依赖从事计算的人员来发展计算方法，而计算工作者也依赖生物工作者提供实验数据，不同学科研究人员之间的沟通需要复合型人才，然而目前复合型人才非常缺乏。

基因芯片数据挖掘方面已有不少英文版著作出版，但笔者还未见有相关的中文著作，因此我们认为有必要编写一本中文的参考书。编写这本书的目的就在于通过对基因芯片技术及数据分析的基本原理的深层描述，培养有多种技能的复合型人才，从提出生物学命题开始，经过合理的实验设计、实验流程，进行数据挖掘，以期解决相关领域的生物学命题。我们在统计分析方法方面作了重点的介绍，使生物学家和计算科学工作者都能从中获得他们各自所需的信息。使学统计的人能对生物学和芯片技术有大致地了解，也使生物学或医学领域的研究者能大致了解基因芯片中所涉及的统计学知识。这本书针对双方各自的需求，希望能成为“催化剂”，降低学科间对话的难度，促使双方更好地交流。但本书的阐述更偏重于数据分析和处理的部分，这是因为目前我们还没有发现有同类的中文书籍出版。

本书共分为十六章，分属于三大部分，第一部分为基础知识，包括概述、微阵列基因芯片制备和检测技术以及统计学基础；第二部分内容是数据处理方法，包括实验设计、图像的获得和数据的前处理、数据的预处理和归一化、差异表达基因分析、芯片数据的可靠性分析、聚类分析和可视化以及微阵列实验中的分类方法；第三部分主要是与数据挖掘和应用相关的内容，包括微阵列技术的标准化、基因芯片数据的基因注释和功能分析、系统生物学及基因调控网络、基因芯片技术的应用——从基因筛选到临床诊断、主要数据分析软件的介绍和展望。

目前，国际上的一些大学已开始对研究生开设有关基因芯片数据挖掘的课程，而据我们了解在中国还没有开设相关课程。此书的编写为将来开设课程打下基础，可以作为各大专院校的教学参考书，该书也可为生物学和医学领域的科技工作者提供科研参考。

本书的编写工作除了有国内同行的积极参与以外，还荣幸地邀请到多位在美国长期从事生物信息学研究和基因芯片数据挖掘的华人同行参与，没有他们的大力支持，很难在短时间内完成此书。本书的编写过程中，上海博星基因芯片有限公司的刘青女士和复旦大学的马允胜先生帮助制作了部分图表，同时还得到了博星公司多位同仁的支持，在此一并表示感谢。

本书由长期从事基因芯片技术和生物信息学研究的中外科学工作者共同编写完成，作者均为在科研第一线的工作人员，由于作者的学科背景不同、写作经验不足、知识面有限，而且该领域发展又相当迅速，因此本书中难免有不足之处，敬请专家和读者指正。

编 者
2006. 1

目 录

第一章 概述	1	一、总体与样本	41
第一节 分子生物学技术及基因、基因组		二、资料的统计描述	42
科学发展历史简介	1	三、随机变量、概率与分布	43
第二节 基因芯片技术简介	3	四、统计量	45
一、基因芯片的基本概念	4	第二节 假设检验	46
二、基因芯片技术的产生和发展	4	一、假设检验的基本原理	46
三、基因芯片的应用领域	6	二、假设检验的步骤	47
第三节 生物信息学与基因芯片的数据		三、假设检验的基本方法	47
挖掘	7	第三节 方差分析	54
一、生物信息学的兴起	7	一、完全随机设计资料的方差分析	54
二、基因芯片的数据挖掘	8	二、随机区组设计资料的方差分析	55
参考文献	9	三、多个样本均数间的多重比较	57
第二章 微阵列基因芯片实验技术	11	第四节 聚类分析与判别分析简介	57
第一节 基因芯片的价值和分类	11	一、聚类分析	58
一、基因芯片的价值	11	二、判别分析	59
二、基因芯片的分类	12	参考文献	61
第二节 基片的制备	15	第四章 实验设计	62
一、基片的类型和性质	15	第一节 样品配对模式	62
二、玻璃基片表面的修饰方法	17	一、基因芯片实验的分类	62
第三节 点样探针的制备	18	二、样品配对方案概述	64
一、cDNA 探针的制备	19	三、样品配对模式的选择	66
二、基因组 DNA 探针	19	第二节 样品的重复及合并	69
三、寡核苷酸探针	19	一、实验误差的来源及重复样品的使用	69
四、独特的 PM-MM 探针设计	20	二、样品重复数量的确定	70
第四节 基因芯片点样	22	三、样品合并	70
一、芯片点样仪和点样方式	22	第三节 总结	72
二、点样后处理	27	参考文献	72
三、基因芯片的质量标准	28	第五章 基因芯片图像的采集和处理	74
第五节 原位合成及纳米结构的基因芯片		第一节 基因芯片图像的采集	74
制备	28	一、激光共聚焦扫描仪	74
一、原位合成法制作基因芯片	28	二、CCD 扫描仪	78
二、纳米结构的基因芯片制备	31	三、扫描仪的技术指标	79
第六节 表达谱基因芯片的检测方法	34	第二节 基因芯片图像的处理	81
一、样本选择、处理和 RNA 的分离	35	一、划格	83
二、mRNA 样本标记	35	二、分割	84
三、芯片杂交	38	三、信息提取	87
参考文献	39	四、质量评估	88
第三章 统计学基础	41	第三节 一些芯片扫描仪和芯片图像处理	
第一节 统计学的基本概念	41	软件的介绍	88

一、激光共聚焦扫描仪	90	二、马氏距离	163
二、激光非共聚焦扫描仪	91	三、Chebychev 距离	164
三、CCD 基因芯片检测仪	92	四、Mahalanobis 距离	164
参考文献	96	五、Minkowski 距离	164
第六章 数据的预处理和归一化	98	六、平均点积	164
第一节 数据的预处理	98	七、向量间的角度	165
一、背景的校正	98	八、协方差	165
二、弱信号的处理	99	九、Pearson 相关距离	165
三、数据的对数转换	101	十、Spearman 秩相关	166
四、重复数据的合并	102	十一、互信息	166
五、缺失数据的处理	103	十二、Kendall's Tau	167
第二节 数据的归一化	104	第二节 聚类算法	167
一、cDNA 芯片数据的归一化	105	一、系统聚类	168
二、Affymix 芯片数据的归一化	115	二、分割聚类	172
参考文献	118	第三节 二维聚类	177
第七章 差异表达基因分析	120	一、耦联二维聚类	177
第一节 差异表达基因的挑选	120	二、区组聚类	177
一、倍数法	120	第四节 主成分、SVD 和基因修剪	178
二、Z 值法	121	一、主成分	178
三、重复实验的判别方法	121	二、奇异值分解	178
四、其他方法	124	三、基因修剪	179
五、总结	125	参考文献	179
第二节 研究差异表达基因的意义	126	第十章 微阵列实验中的分类方法	181
一、在基因组研究中的作用	126	第一节 概述	182
二、在药物研究中的作用	127	一、利用基因表达谱数据进行生物样本	
三、在医学基础研究中的作用	129	分类	183
参考文献	131	二、分类的背景	183
第八章 芯片数据的可靠性分析	133	三、基因表达谱数据	184
第一节 数据的评价	133	第二节 不同分类方法的概述	184
一、差异表达基因的可靠性	133	一、分类及统计决策论	184
二、芯片数据重复性评价	139	二、费歇线性判别分析	186
第二节 误差来源分析	142	三、线性判别和二次判别分析	186
一、生物学差异来源	142	四、线性判别分析的扩展	188
二、实验系统误差	144	五、最近邻分类器	188
第三节 基因芯片的质控体系	149	六、决策树	190
一、直接点样的基因芯片的质控体系	149	七、BP 神经网络分类法	194
二、Affymetrix 的寡核苷酸芯片质控		八、支持向量机	197
体系及其产品质量评估	151	九、Parzen 窗	204
第四节 信号线性扩增技术及其评估	154	第三节 分类中的一般问题	205
一、信号线性扩增技术	154	一、特征选取	205
二、信号扩增方法的可靠性评价	154	二、标准化和距离函数	206
参考文献	161	三、缺失值填充	207
第九章 聚类分析和可视化	162	四、多分类问题	208
第一节 相似性（或距离）的度量	162	第四节 性能评价	209
一、欧氏距离	162	一、偏差、方差和误差率	209

二、再置换估计	210	二、研究转录因子及其调控基因的实验	
三、倍数交叉验证法	210	方法	254
四、解靴带估计	210	三、基因调控网络与图形	254
第五节 实例分析	211	第三节 用高斯图形模型推导基因调控	
一、基因表达谱数据	211	网络	257
二、数据预处理	212	第四节 贝叶斯网络模型在基因芯片	
三、支持向量机软件应用	213	数据中的应用	259
参考文献	216	一、贝叶斯网络简介	259
第十一章 微阵列技术的标准化	218	二、学习贝叶斯网络	261
第一节 MIAME 规则	218	三、贝叶斯网络方法在基因芯片数据	
一、MIAME 规则的具体内容	219	方面的应用	262
二、MIAME 表单	221	第五节 从时间序列数据中推导基因调控	
三、MIAME 的目前与将来	222	网络	266
第二节 Affimetrix 芯片系统与 MIAME		一、基因调控网络模型的“事件模型” ..	266
规则	223	二、关于基因调控网络的“动态	
一、遵循 MIAME 规则	224	概率模型”	268
二、Affimetrix 实验的 MIAME 表单	225	第六节 通过基因扰动来推导基因调控	
三、Affimetrix 的 RNA 抽提、清洗、		网络的反义工程方法	270
标记和杂交规范	225	第七节 结论	271
参考文献	227	参考文献	272
第十二章 基因芯片数据的基因注释和		第十四章 基因芯片技术的应用——	
功能分析	228	从基因筛选到临床诊断	274
第一节 单一基因的注释	228	第一节 基因表达谱研究与临床肿瘤学	274
一、一般的注释	228	一、确定肿瘤亚型	275
二、关于疾病的信息	233	二、识别肿瘤的组织来源	276
三、蛋白质家族的信息	234	三、预后分析	276
第二节 转录因子调节的分析	235	四、存在问题	277
一、Transfac 数据库	236	第二节 微矩阵芯片和遗传多态性	278
二、转录因子研究中的统计学检验	238	一、单核苷酸多态性简介	278
第三节 Gene Ontology 数据库中基因		二、基因多态性与疾病易感性	279
功能分类的分析	240	三、基因多态性作为遗传标记的应用	279
一、Gene Ontology 数据库	240	四、基因多态性与个性化用药	280
二、GO 数据库相关分析的工具	241	五、基因多态性和基因芯片检测技术	281
第四节 生物学通路和生物学相互作用的		第三节 微矩阵和基因拷贝数变化	282
分析	243	一、cDNA 阵列 CGH	283
一、生物学通路中的基因分析	244	二、基因组阵列 CGH	283
二、生物学网络中的基因分析	249	第四节 微矩阵和感染性疾病	284
三、基因芯片数据中使用者自己定义的		一、微生物的鉴定和分型	285
基因集的分析	250	二、耐药性研究	286
参考文献	251	三、致病机理研究	287
第十三章 系统生物学及基因调控		第五节 微矩阵芯片的其他应用	288
网络	252	一、微矩阵芯片和 DNA 甲基化分析	288
第一节 系统生物学简介	252	二、转录因子结合位点分布	290
第二节 基因转录调控网络的构成	253	三、展望	291
一、基因转录过程简介	253	参考文献	292

第十五章 主要数据分析软件的介绍	295	一、斯坦福大学 SMD/SOURCE	309
第一节 分析软件在基因芯片技术中的		二、UCSC 基因组浏览器	309
地位	295	三、MySQL 客户	310
第二节 主要图像和数据处理软件	296	参考文献	311
一、基因芯片图像分析软件		第十六章 展望	312
GenePix Pro	296	第一节 后基因组研究的趋势——系统	
二、Affymetrix GCOS 系统	297	生物学	312
三、Cluster 和 TreeView 程序	298	一、系统生物学的启动	312
四、GeneSpring	300	二、系统生物学的发展趋势	313
五、SpotFire DecisionSuite	300	第二节 后基因组应用研究发展的	
六、SAM 和 PAM	302	趋势——基因组医学	314
七、R 平台及生物导体	303	第三节 基因芯片技术在系统生物学和	
八、MATLAB 生物信息工具箱	304	基因组医学中的地位	316
第三节 基因表达谱公共数据库	304	一、基因芯片及数据挖掘在基础研究中	
一、NCBI-Gene Expression Omnibus		的地位	316
(GEO) 基因表达数据专用库	304	二、基因芯片技术在基因组医学分子	
二、EBI ArrayExpress 和 SMD	307	诊断中的应用趋势	316
三、微阵列数据库的建立和管理	307	参考文献	318
第四节 基因注释数据库的访问	308		