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
现代生物技术与医药科技出版中心 出版发行

（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

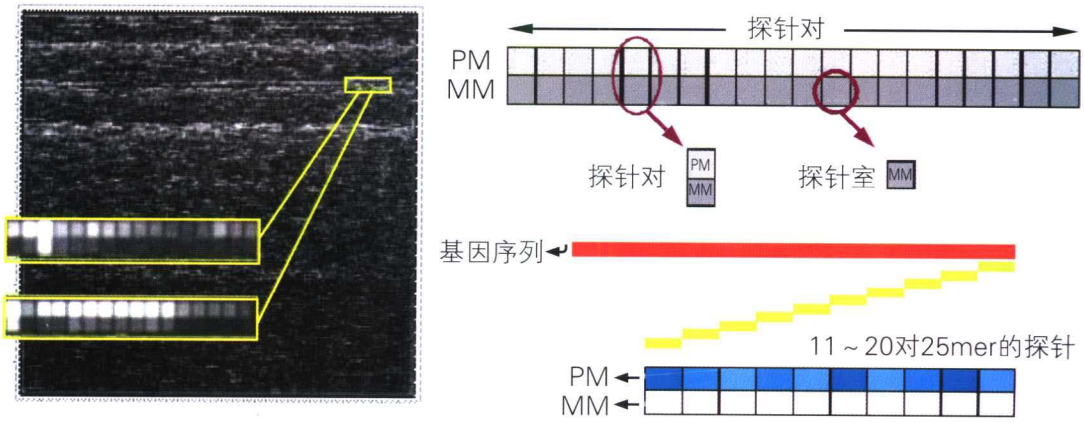


图 2－5　PM–MM探针设计方案

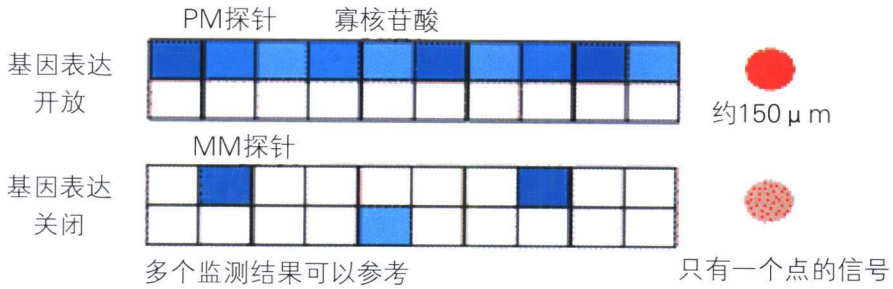
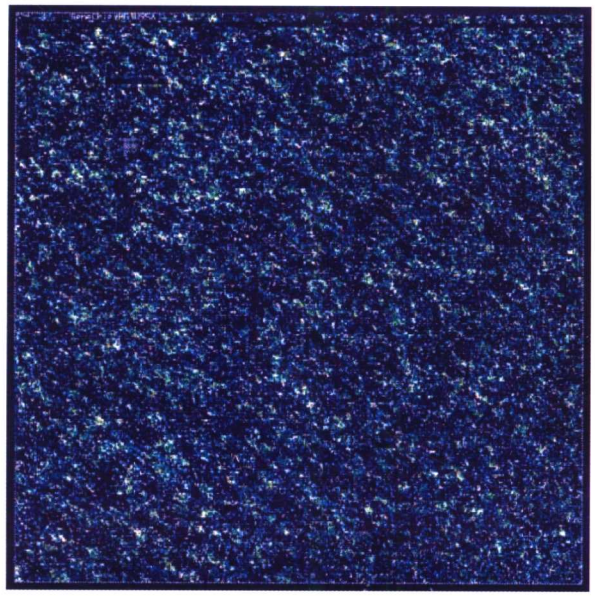
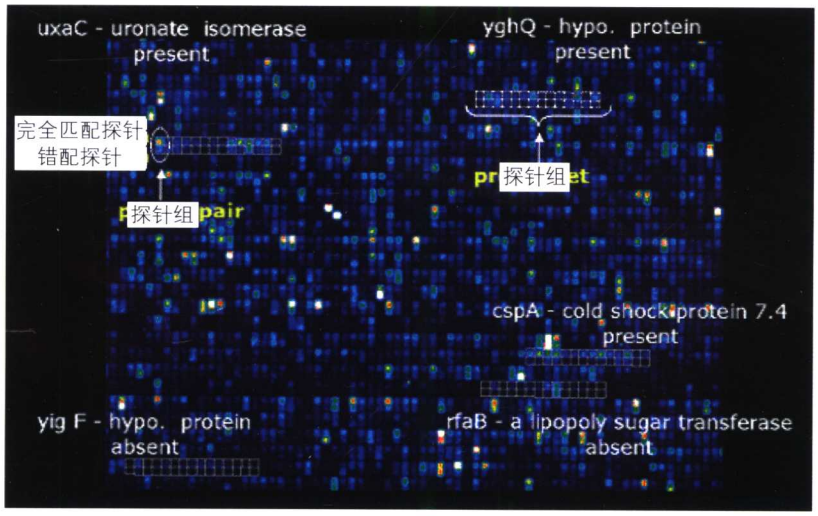


图 2－6　PM–MM的多对探针的结果与单个探针的结果比较

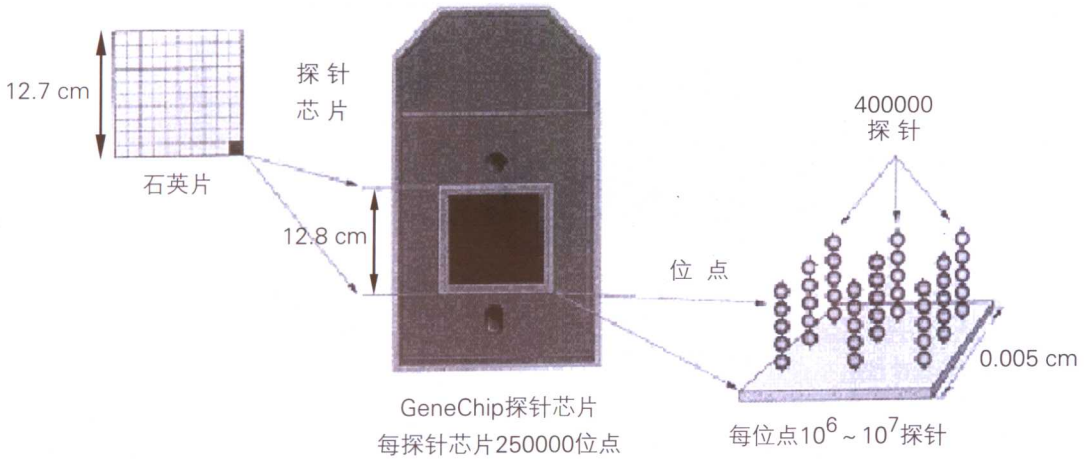


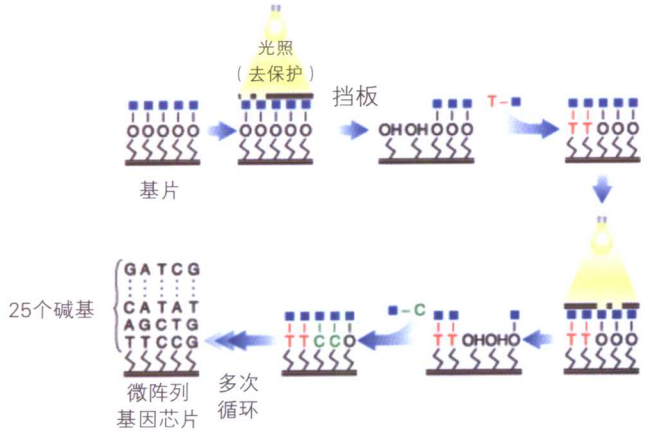
（a）整体图像



（b）局部放大

图 2－16　Affymetrix公司的高密度基因芯片图像





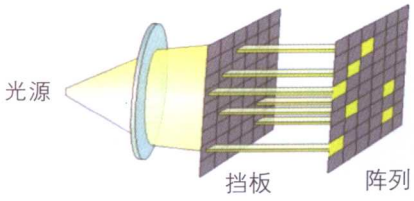


图 2－17　光导原位合成原理图

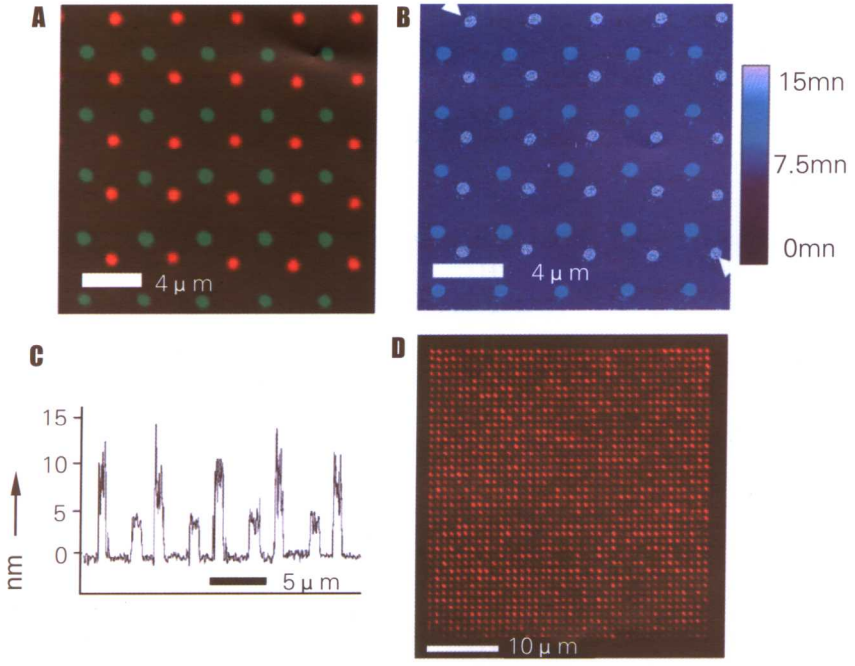


图 2－22　DPN制备的多 DNA“墨水”的图案

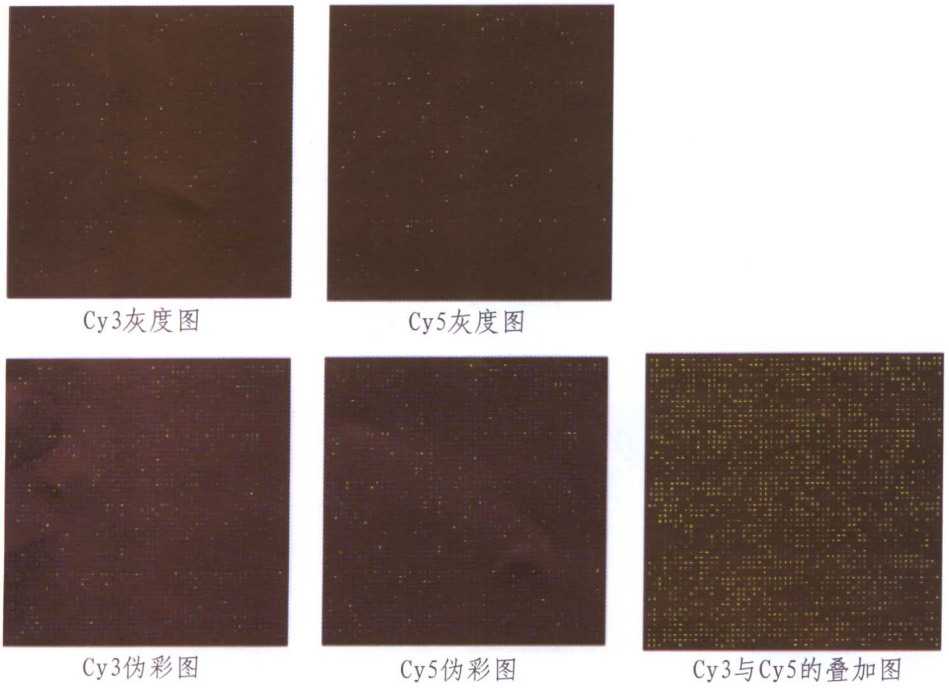


图 5–4　cDNA芯片的双色荧光图像

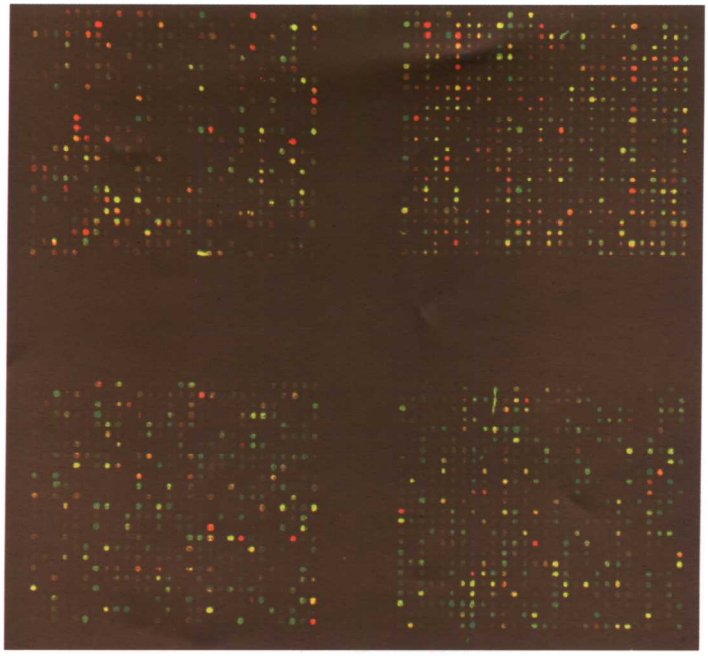


图 5-5　cDNA芯片的典型图像（图像中有四个子格子）

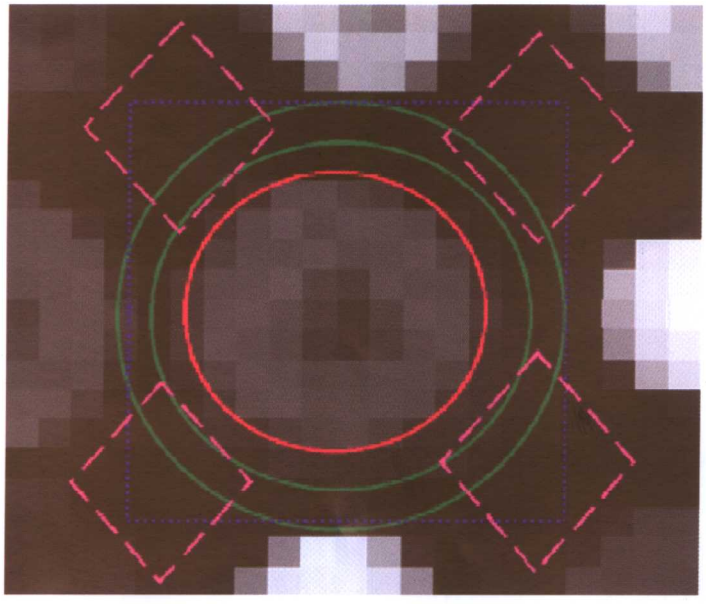


图 5-6　不同的背景值确定方法

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

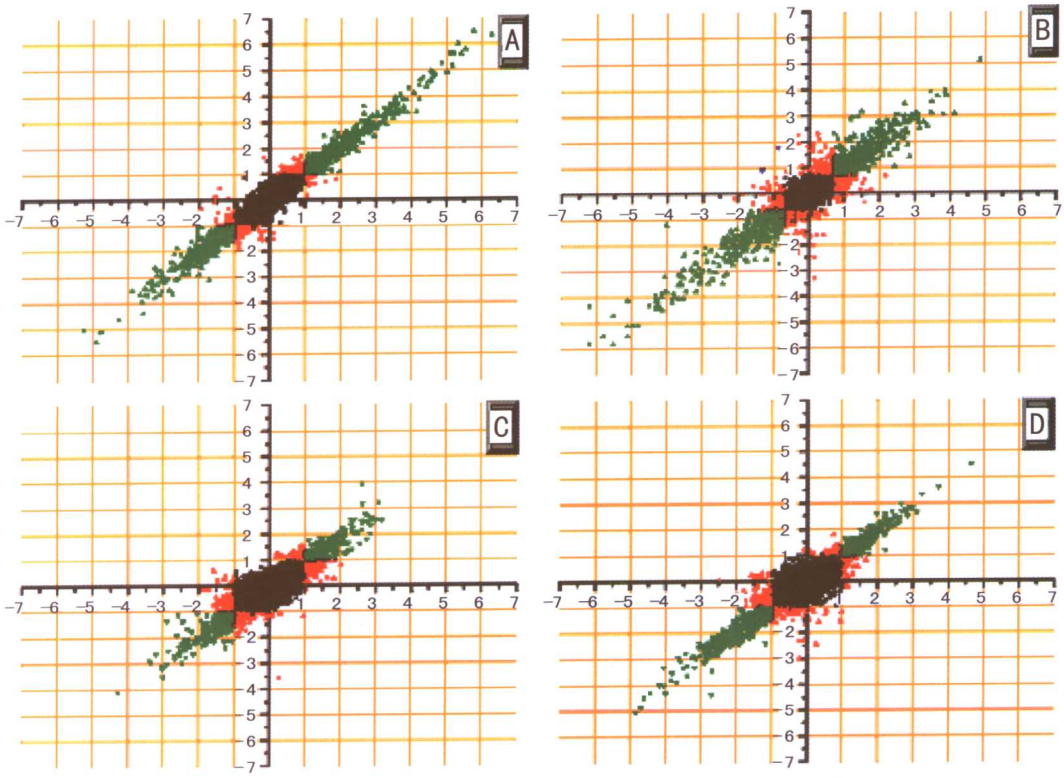


图 8-4　正向重复实验比值的散点图

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。

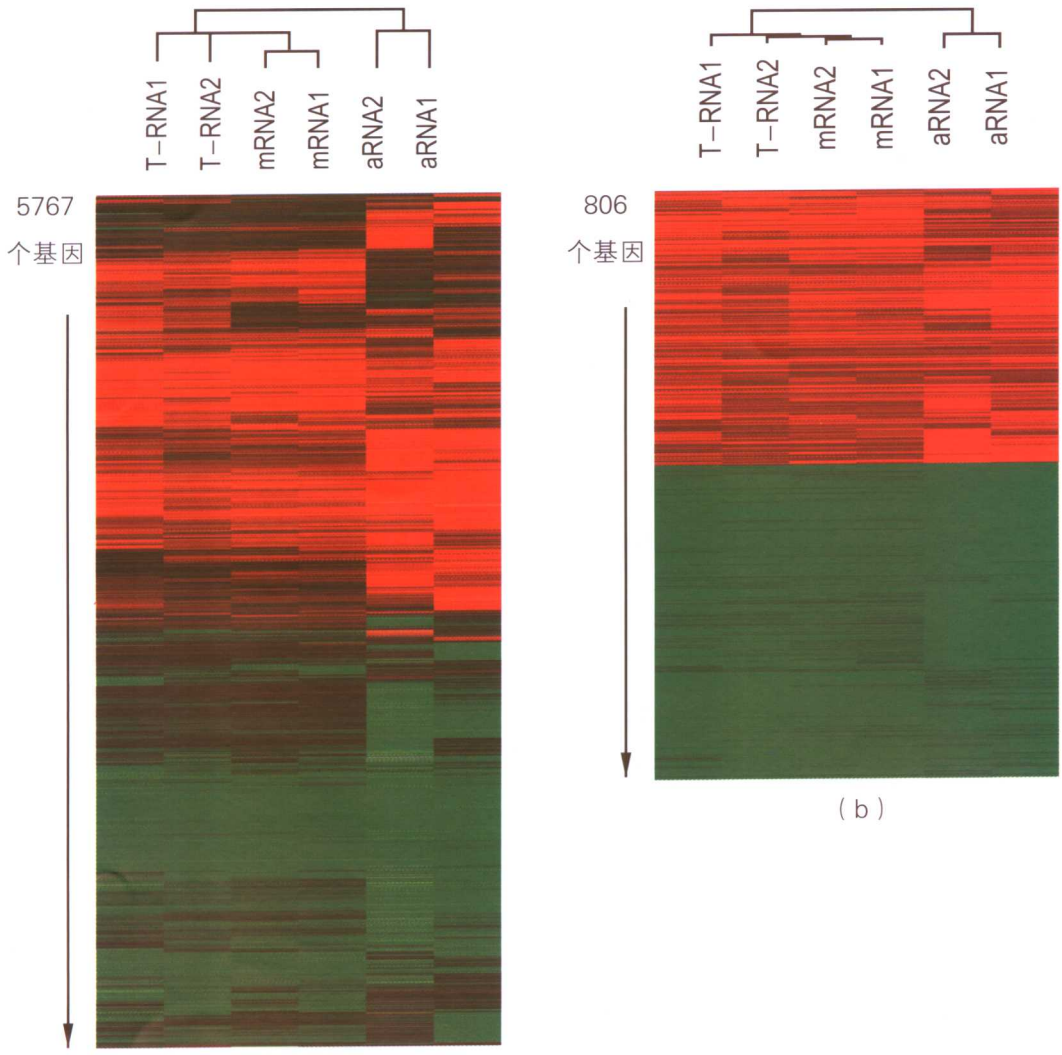


图 8-16　三种标记方法共六次实验的表达谱数据的层级聚类结果(hierarchical cluster)

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

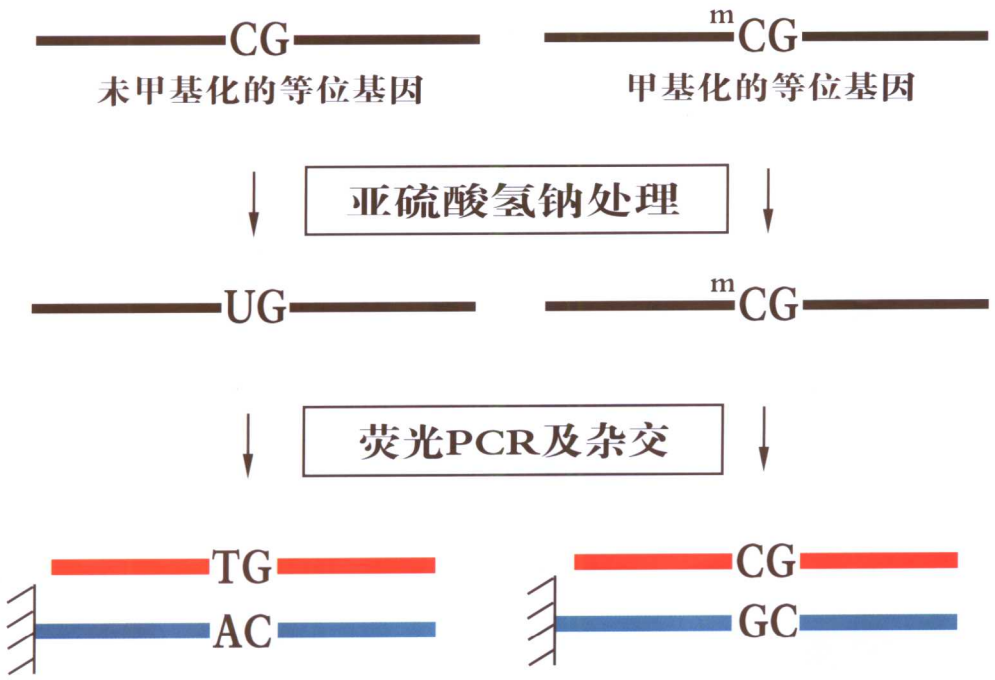


图14-5 甲基化特异性寡核苷酸芯片方法原理示意图

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）

刘三霞（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的了解，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