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

基因芯片技术是交叉性很强的学科，尤其需要生物学家和计算科学家通过“双边对话”来完成实验设计、实验方法到数据分析和阐明。不同学科研究人员之间的沟通需要跨专业的复合型人才，而目前复合型人才非常缺乏。有鉴于此，我们编写本书，目的在于通过基因芯片技术及数据分析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、实验流程以及数据挖掘，以期更好地解决生物学命题。

本书共分为十六章，分属于三大部分。第一部分主要为基础知识部分，包括概述、微阵列基因芯片制备和检测技术、统计学基础3章；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、微阵列实验中的分类方法7章；第三部分主要为数据挖掘和应用相关内容，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6章。

通过阅读本书，生物学者和计算科学工作者都能从中获得他们各自所需的信息。从事统计学研究的人能对生物学和芯片技术有清楚的了解，生物学或医学领域的研究者能初步掌握基因芯片中所涉及的统计学知识。同时，本书也可作为各大专院校生物芯片技术和生物信息科学的学科建设的教学参考书。

ISBN 7-5025-8564-8
9 787502 585648 >

销售分类建议：生物

ISBN 7-5025-8564-8
定价：49.00元

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

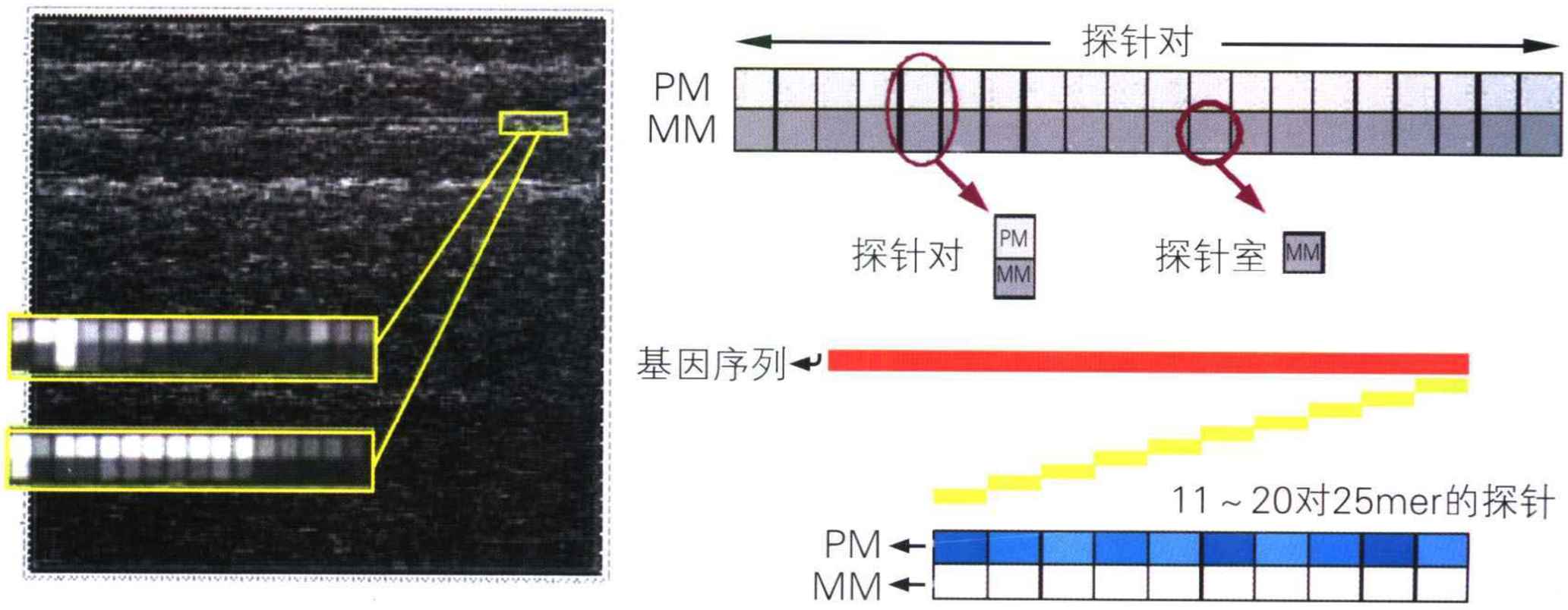


图 2－5　PM–MM探针设计方案

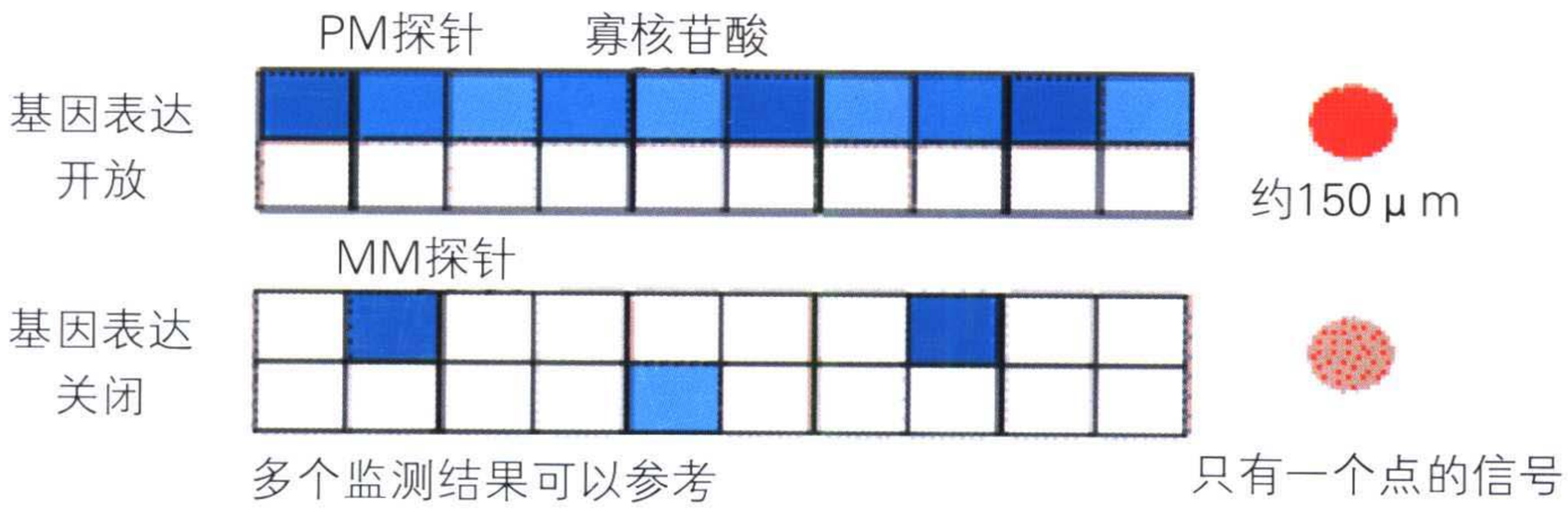
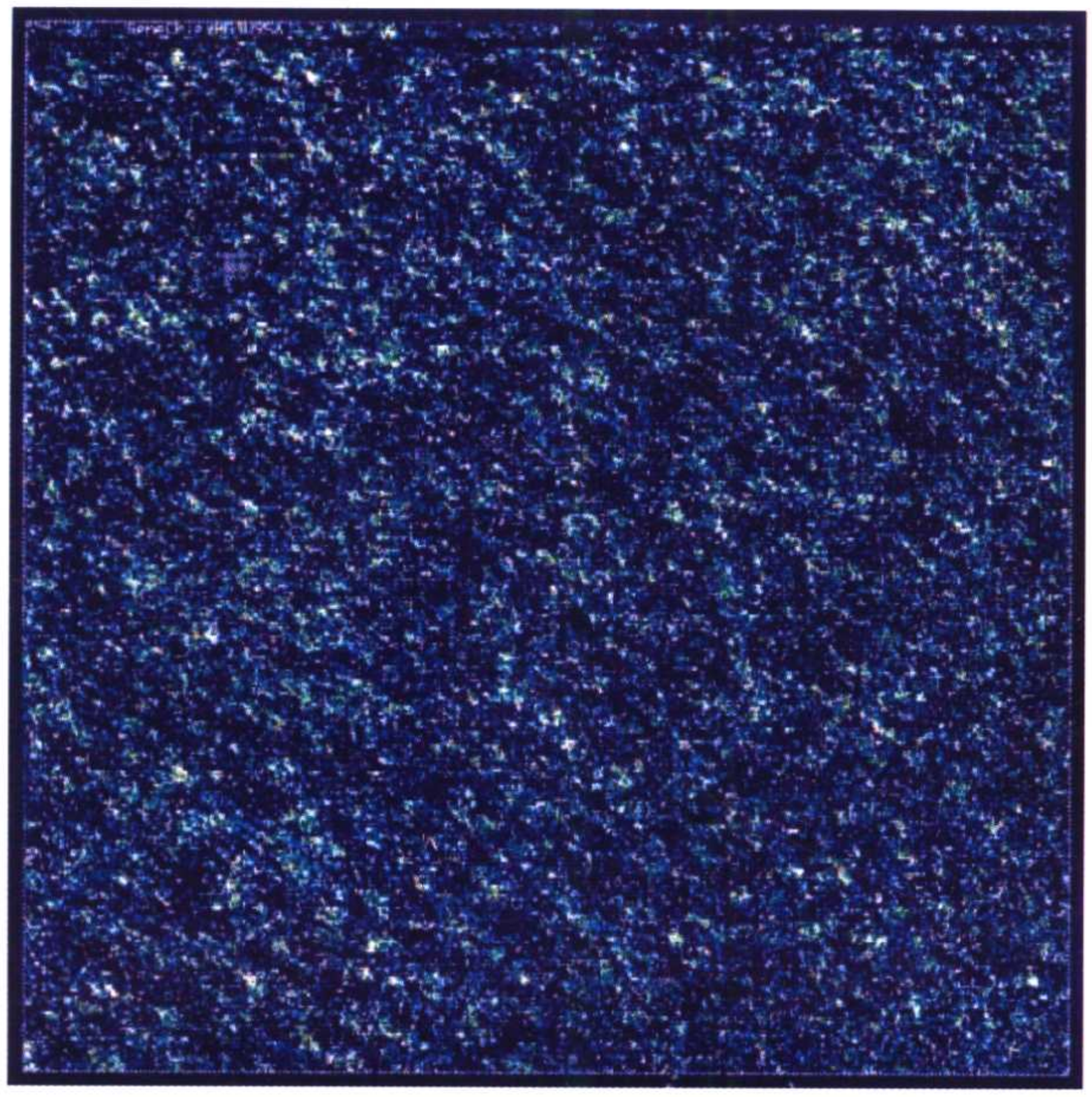
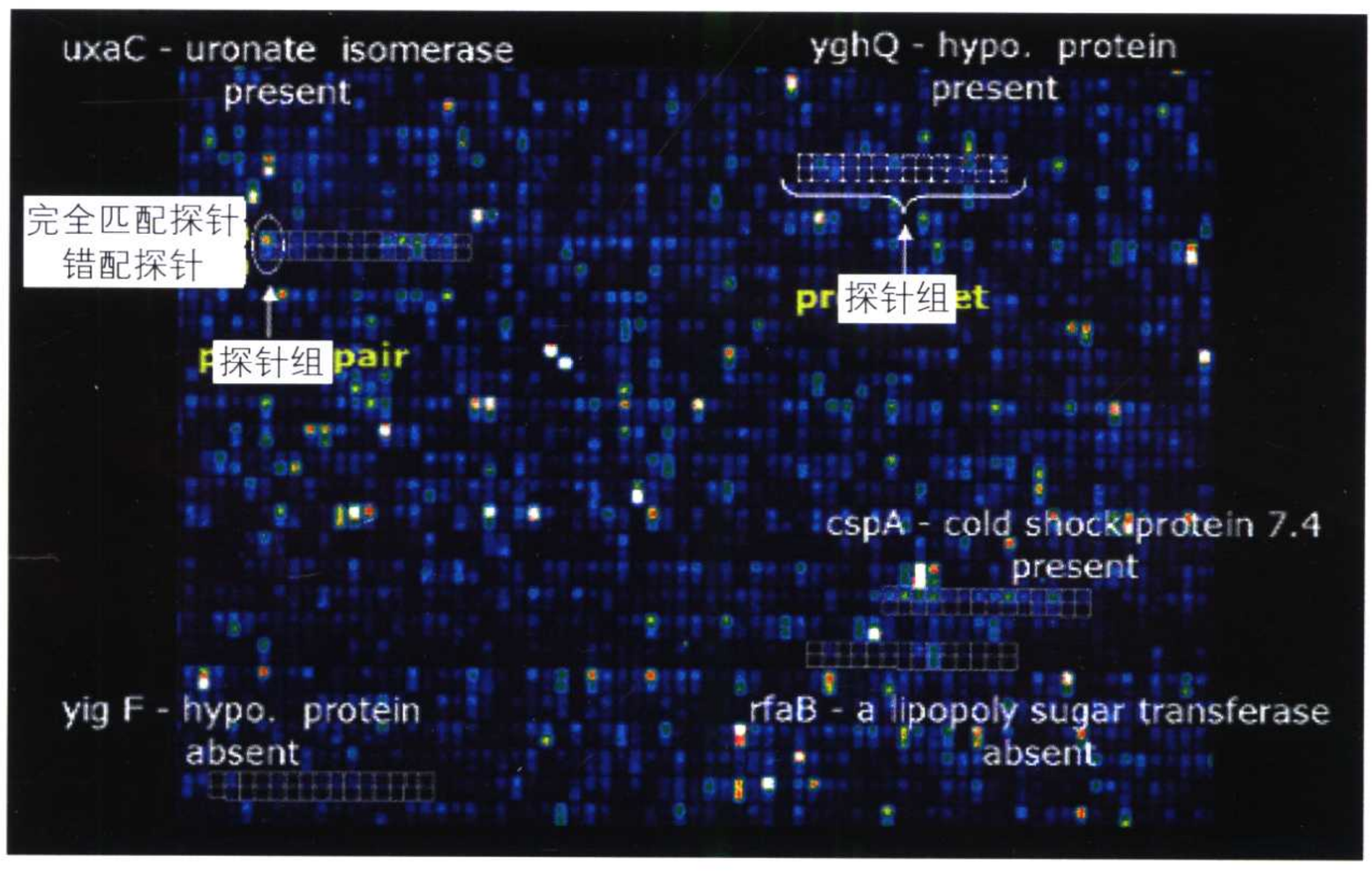


图 2－6　PM–MM的多对探针的结果与单个探针的结果比较



（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

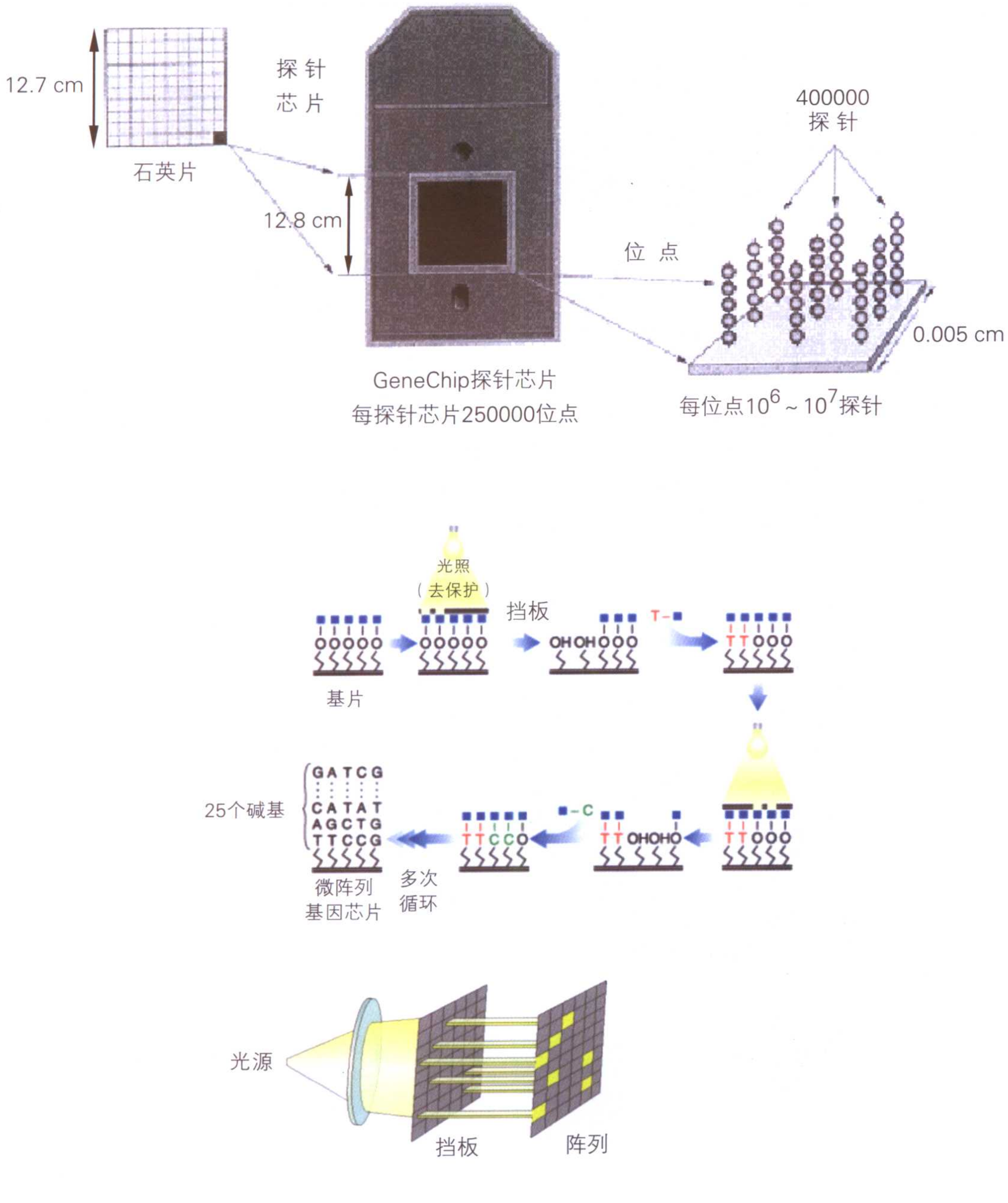


图 2－17　光导原位合成原理图

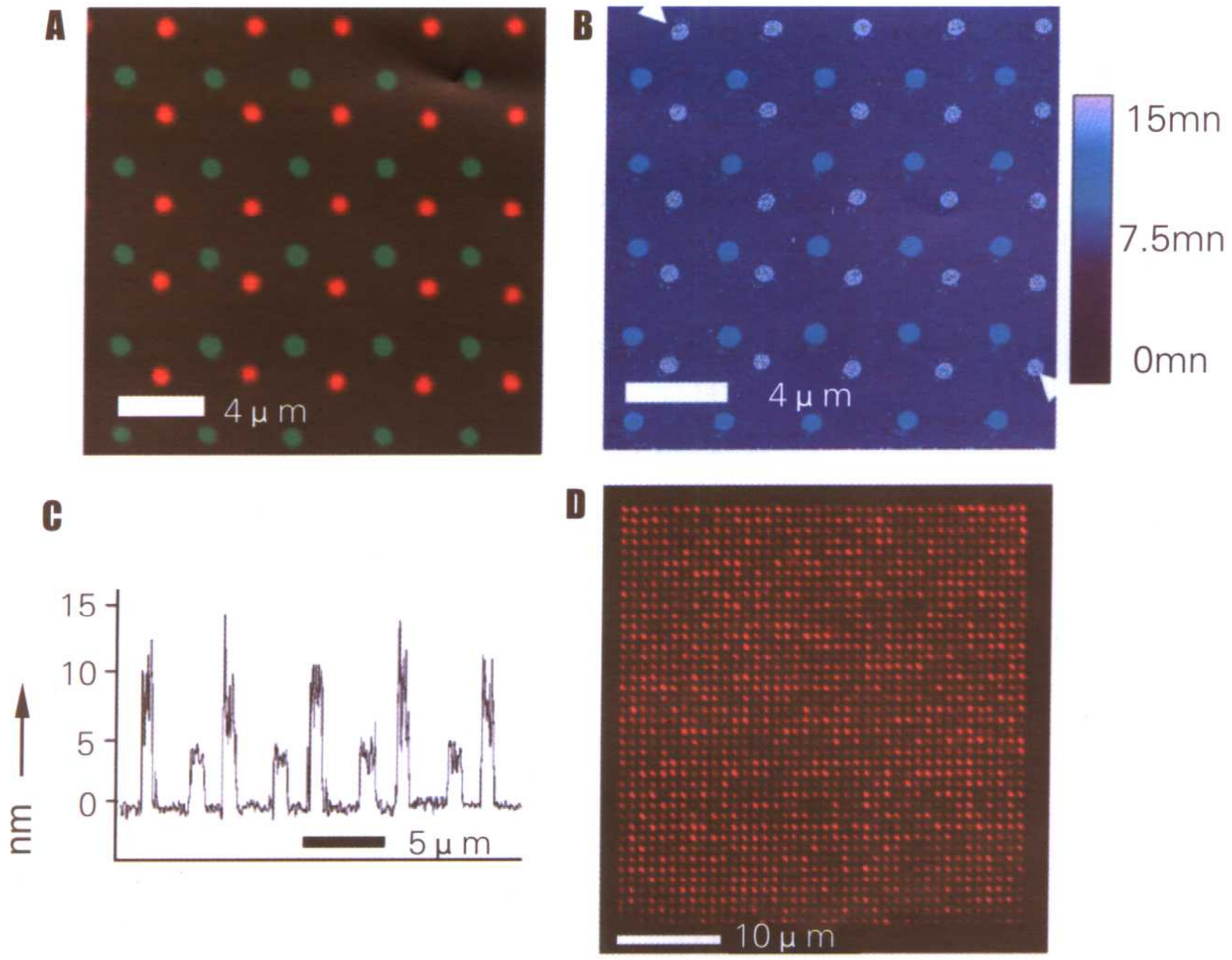


图 2－22　DPN制备的多 DNA“墨水”的图案

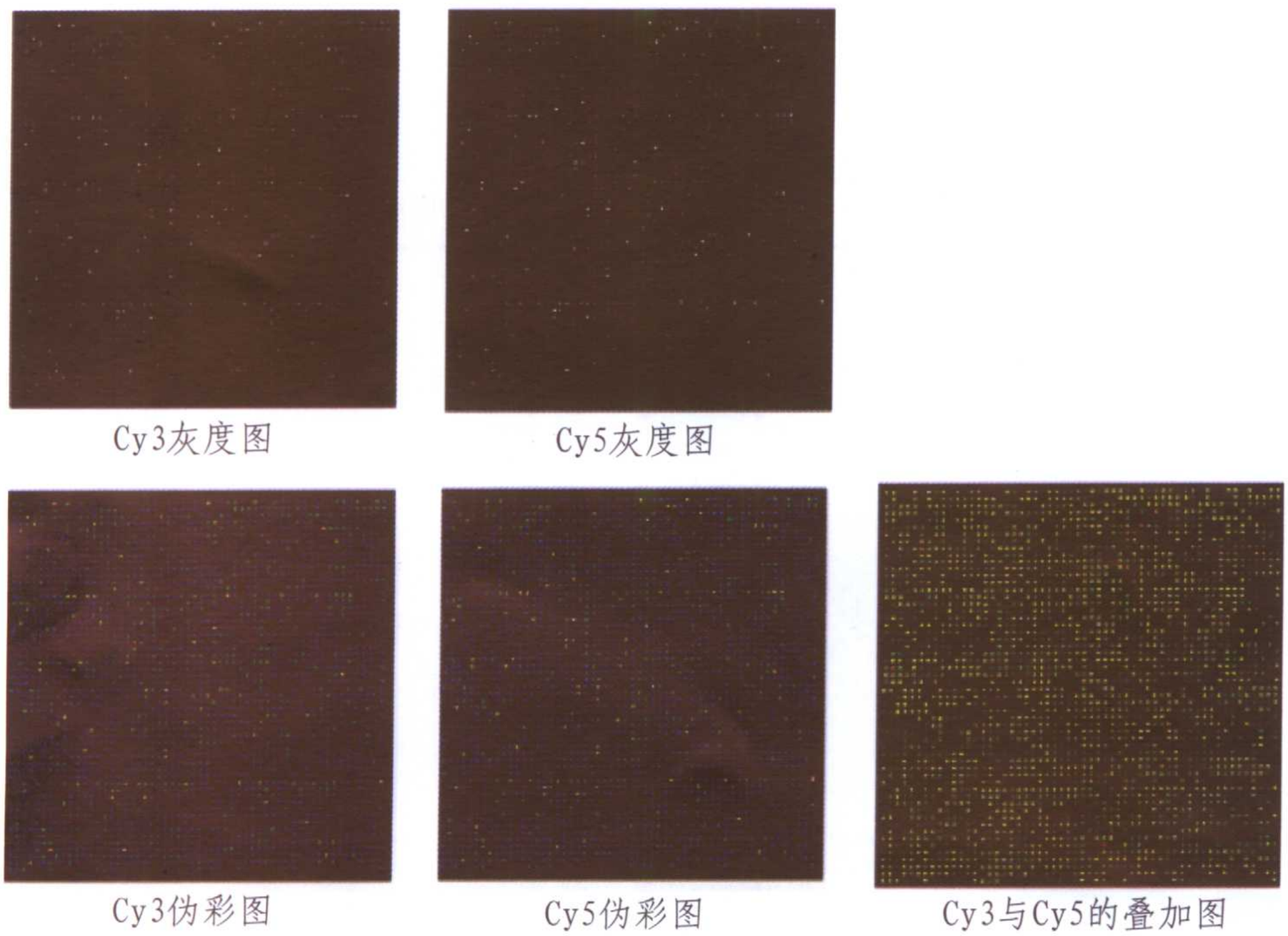


图 5-4　cDNA芯片的双色荧光图像

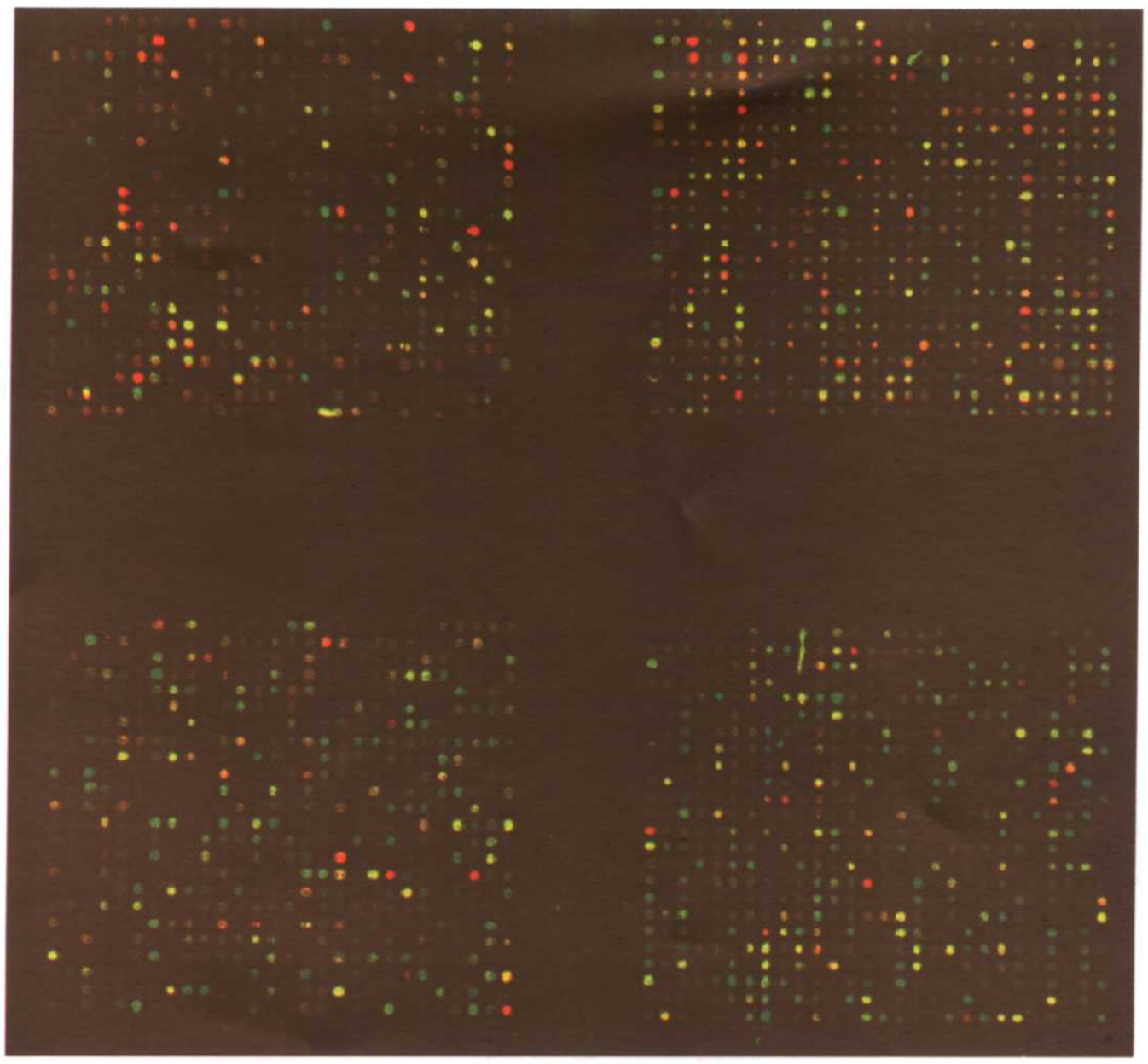


图 5-5　cDNA芯片的典型图像（图像中有四个子格子）

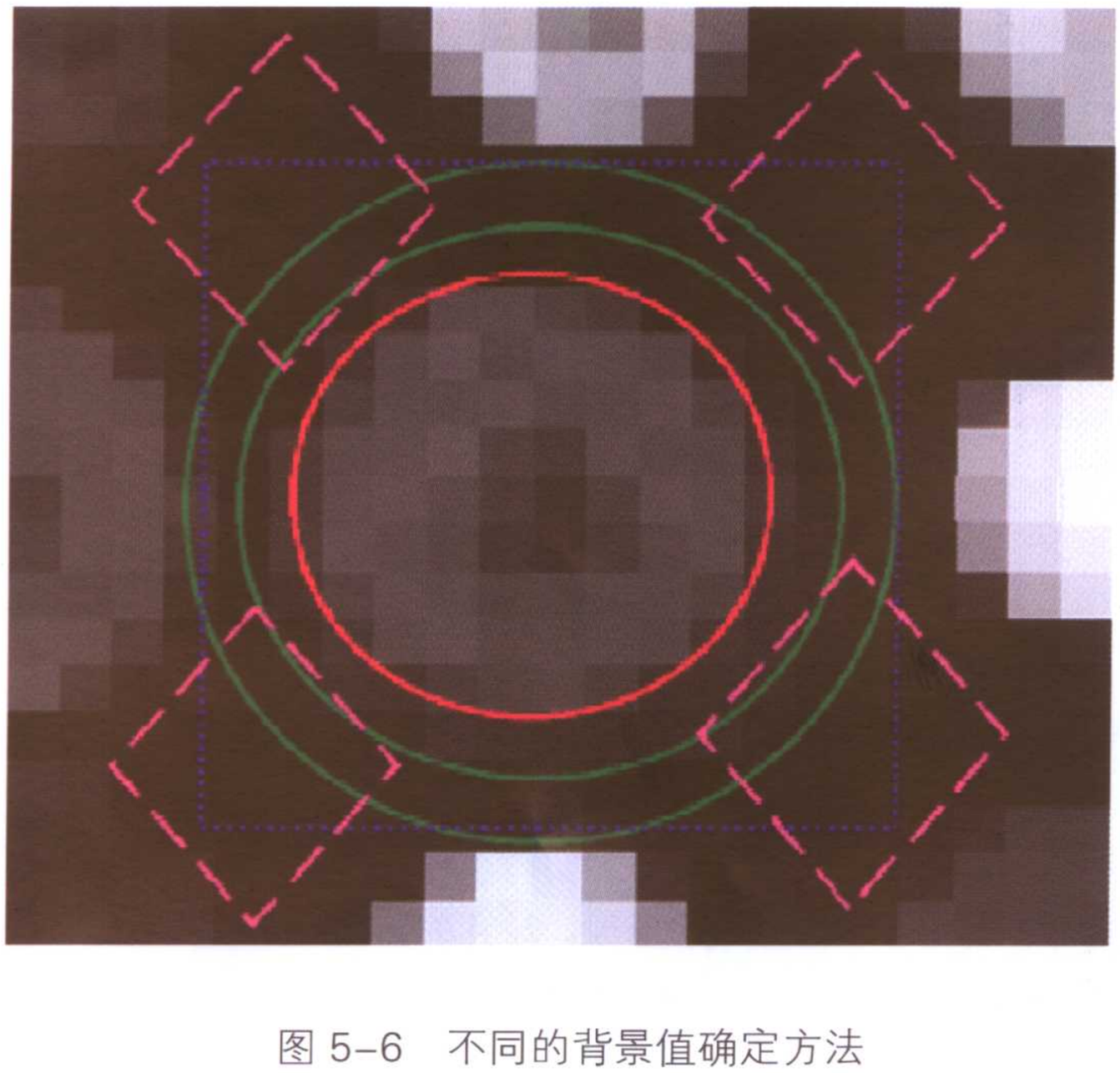


图 5-6　不同的背景值确定方法

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

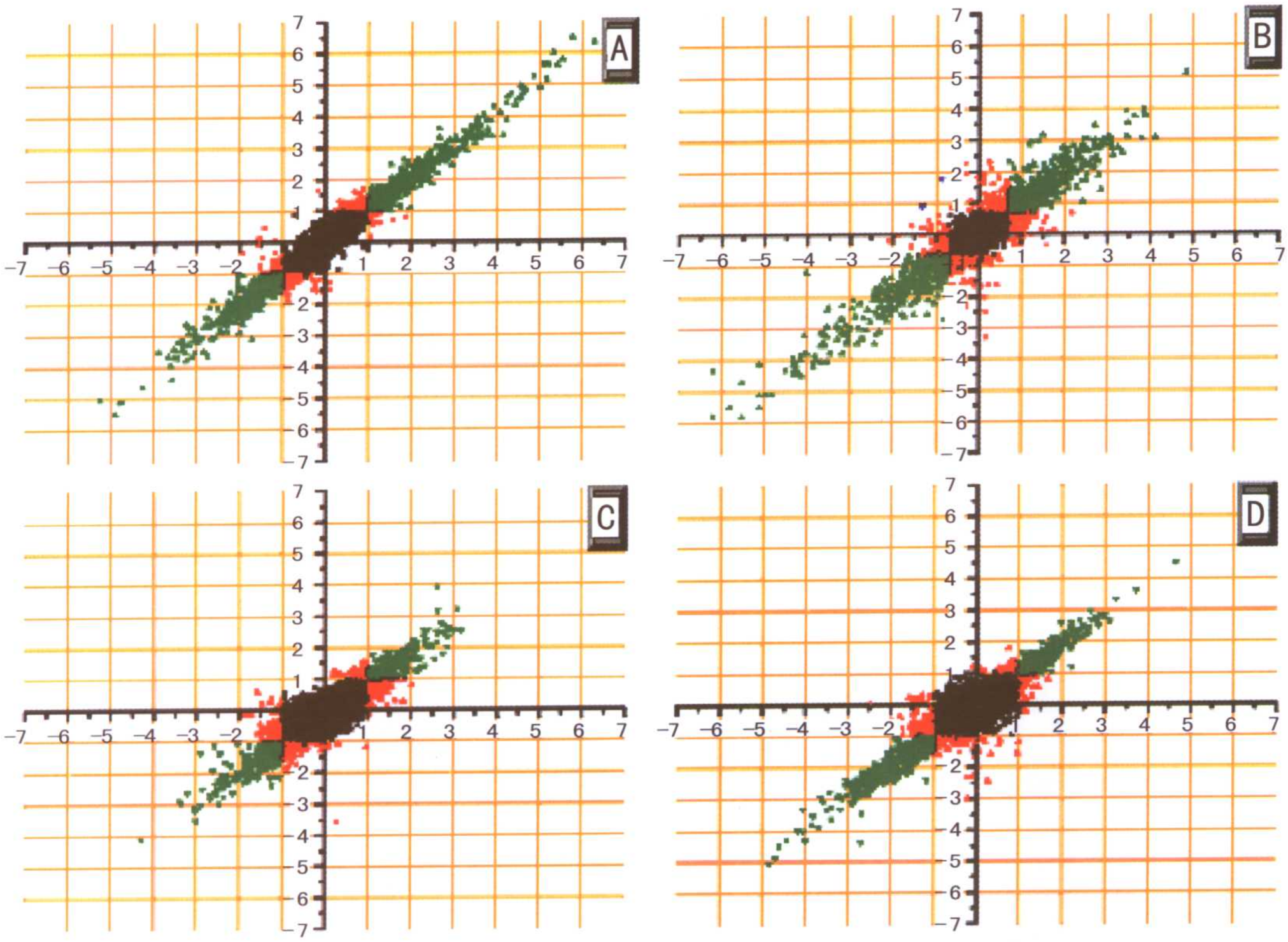


图 8-4　正向重复实验比值的散点图

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。

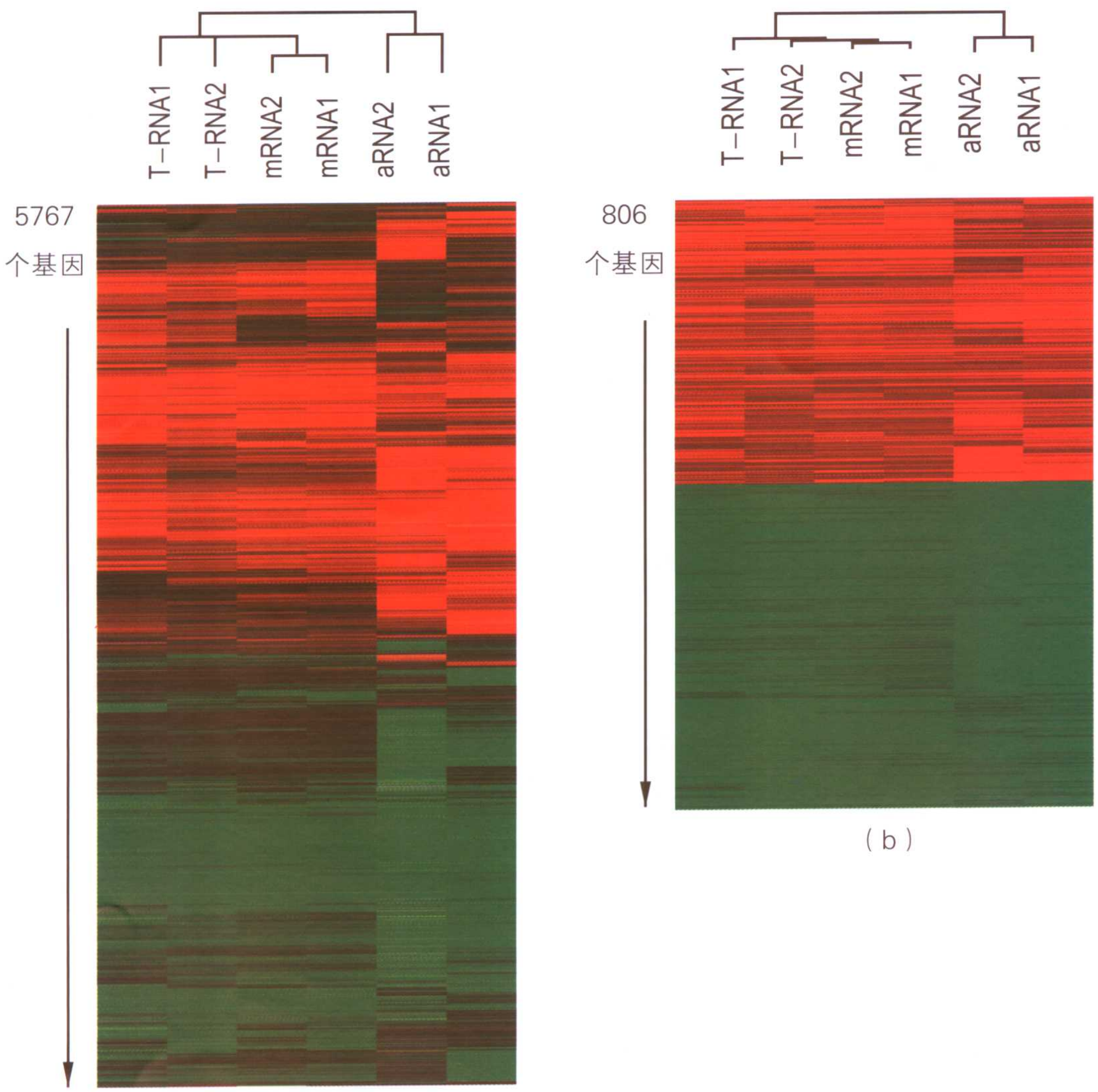


图 8–16　三种标记方法共六次实验的表达谱数据的层级聚类结果(hierarchical cluster)

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

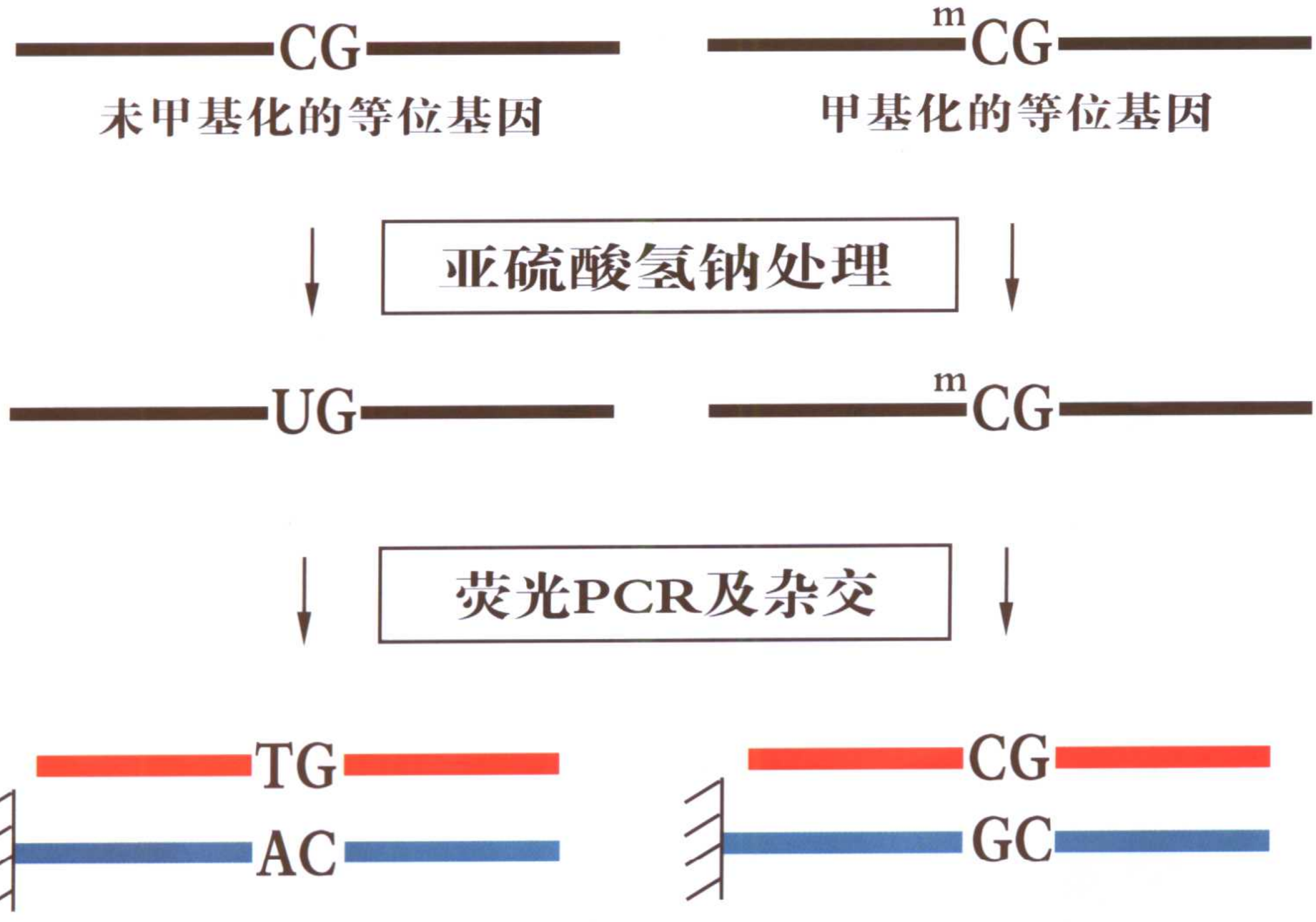


图14-5 甲基化特异性寡核苷酸芯片方法原理示意图

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

第一章　概　　述

第一节　分子生物学技术及基因、基因组科学发展历史简介

科学和技术两个概念的内涵不同，但却像一对“双胞胎”，总是相互伴随、相互促进，在科学理论指导下产生的新技术往往进一步促进了科学的发展，而历史上每次科学的突破都会促进新技术的产生，并永久性地改变科学实践。

生物科学是一门实验科学，它的发展极大程度上依赖于生物实验技术的发展，在过去的50年内，遗传学的一些基本的发现和一些生物技术的进步伴随着计算机科学的发展，使得分子生物学和基因组研究成为可能，从而开创了分子生物学研究的新时代。

下面简单地介绍一下生物学中的某些基础和技术革新所产生的结果，可以有助于了解基因芯片在生物学中的地位。

早在第二次世界大战开始时，美国华盛顿州卡内基研究所的 R. B. Roberts 等物理学家和化学家开创性地使用了放射性同位素来阐明代谢途径，主要研究大肠杆菌的生物合成。这项开创性的工作结合早期的遗传学和生理学研究，在以后的 20 年内共同确立了大肠杆菌为生物学研究的模式生物。在此期间，科学家阐明了大部分中间代谢产物的生物合成和代谢途径，并发展了新的生物化学和遗传学方法来鉴别参与这些途径的酶，这些科学家推动了现代分子生物学的前身——酶学的发展。

在酶学研究时期，遗传学家用生化学家提供的信息来发展模式系统的遗传研究，如细菌、酵母、果蝇和小鼠等，最早将基因突变、蛋白质变异和疾病联系在一起的是 Pauling 和他的同事（1949），他们发现患镰刀形细胞贫血症的病人和正常人相比，其血红蛋白在电泳时的迁移发生了改变，作者分析这是由于血红蛋白的表面电荷发生改变所致；通过进一步分析，他们发现血红蛋白基因的变化导致了蛋白质的改变，这在以后的基因测序中得到证实。这项研究成果发表在《科学》杂志上，它促进了分子生物学的研究，并奠定了一系列分子生物学实验技术的理论基础。

1953 年，James D. Watson 和 Francis H. C. Crick（Watson-Crick）通过综合化学研究和模型的数据，提出了著名的 DNA（脱氧核糖核酸）分子双螺旋结构模型假说，以后的生物化学研究和结构研究证实了他们的假说。他们这一发现获得了 1962 年的诺贝尔奖，这是生命科学发展史上的一个重要里程碑，是现代遗传学的基础，生命科学也因此深入到分子研究水平。在这个理论基础上，DNA 复制、蛋白质合成、基因表达、遗传物质的交换和重组等分子机制迅速得以阐明。1961 年，Jaob 和 J. Monod 建立了蛋白质介导的基因表达调控机制模型。

20 世纪 50 年代，洛克菲勒研究所（现洛克菲勒大学）的 Bruce Merrifield 建立了固相合成技术，该技术最早是用于多肽合成，将第一个氨基酸固定在不溶的聚合物载体上，然后其他氨基酸依次连接到固定的氨基酸上，当合成终止时，所合成的链从固相载体上脱落下来并进行纯化。该技术很快被用于寡核苷酸的合成并得以发展完善，广泛用于分子生物学的各个方面，包括以后发展形成的原位合成基因芯片，大大推动了分子生物学的进程。Merrifield 所做的工作及作出的贡献，使他获得了 1984 年的诺贝尔化学奖。

Kornberg 和他的同事在 20 世纪 50 年代发现了 DNA 聚合酶，几乎同一时期，Ochoa 发现了 RNA 聚合酶，他们因为发现了核糖核酸和脱氧核糖核酸的生物合成机制而共同获得 1959 年的诺贝尔奖。核酸聚合酶在以后的分子生物学中应用十分广泛。1970 年，Baltimore 和 Temin 等在研究 RNA 病毒时发现了一种特殊的聚合酶——逆转录酶（RT），它能够以 RNA 为模板合成 DNA，这一发现进一步完善了中心法则，即 RNA 可以逆转录成 DNA。Baltimore、Temin 和 Dulbecco 因为这一发现共同获得了 1972 年的诺贝尔奖。这些研究成果是发展基因芯片技术必不可少的前提条件。

20 世纪 70 年代初，Daniel Nathans 和 Hamilton Smith 在研究为什么大肠杆菌在感染噬菌体后能获得抗后继感染的免疫能力时，发现了限制性内切酶和 DNA 连接酶，它们可以在 DNA 序列特异位点上切割和重新连接。1972 年，斯坦福大学的 Paul Berg 和他的同事发明了 DNA 重组技术，构建了含有 λ 噬菌体基因的动物病毒 SV40 载体，可以将外源 DNA 导入大肠杆菌，这种克隆和表达外源基因的技术作为基因工程的基础技术一直沿用到现在，包括应用于最新的基因组研究领域，他们也因此获得了 1980 年的诺贝尔化学奖。

基因组研究的兴起也离不开技术的进步，在众多技术中最有代表性的是 DNA 测序技术和 PCR 技术。1977 年，哈佛大学的 Maxam 和 Gilbert、英国医学研究委员会（MRC）的 Sanger 等分别发明了 DNA 测序技术，Gilbert 和 Sanger 共同获得 1980 年的诺贝尔奖。自动化测序技术的产生大大提高了测序速度，在不到一周的时间内可以完成几百万个碱基的微生物全基因组的测序。目前已完成的原核生物和真核生物的全基因组序列不下几十种。20 世纪 80 年代初，Cetus 公司的 Mullis 等利用耐高温 DNA 聚合酶发明了 PCR 技术，可使得微量的生物样本得以大量扩增。

到了 20 世纪的最后 10 年，分子生物学研究发生了很大的变革，从单个基因或蛋白质的研究转向大规模研究基因，从而产生了基因组学、功能基因组学等新学科。虽然在 1984 年已经完成了几种微生物基因组的部分作图和测序，但基因组时代到来的真正标志是 1986 年在新墨西哥的 Santa Fe 召开的国际会议，此次会议由美国能源部健康和环境研究办公室主办。会上，国际顶尖的科学家无可争议地签署了完成人类基因组计划的必要性和可行性报告。这次会议使得 1988 年美国国家研究理事会将题为“人类基因组作图和测序”的计划推荐给国家，支持人类基因组计划（HGP），并提出多个阶段的大致计划。同年，在 Lawrence Berkeley、Lawrence Livermore 和 Los Alamos 国家实验室同时成立了三个基因组研究中心。同时，在 James Wyngaarden 的领导下，美国国立卫生研究院（NIH）成立了基因组研究办公室，并于 1989 年成为国立人类基因组研究中心，由 James D. Watson 领导。在此后的 10 年内，由于自动测序技术的进一步完善，全世界更多的研究机构加入到人类基因组计划的行列中，如波士顿的 Whitehead 研究所和英国剑桥的 Sanger 中心，这些工作促进了计算方法和信息学方法的发展，以满足处理大规模测序所产生的日益膨胀的数据的需要。

1991 年，美国国立卫生研究院（NIH）的 Craig Venter 创立了一种发现人类基因的方法，该方法不需要全部的基因组测序。该方法基于约 3%的基因组组成了编码 mRNA 的基因的推测，他认为只有部分的 DNA 具有转录活性，这些 DNA 表达的片段被酶转化和加工成 mRNA 分子，因此最有效的获取基因的方法是利用细胞内加工的 mRNA，在逆转录酶的作用下逆转录成互补的 cDNA，这些稳定的 cDNA 被称作表达序列标签，简称 EST，通过 EST 重叠区的拼接形成更长的片段，最终得到完整的基因。1992 年，Venter 离开 NIH 创立了基因组研究所（TIGR）。到 1995 年研究人员分离到 170000 个人类 EST，占全部人类基因的一半以上。1998 年，Venter 加入 Perkin-Elmer 公司（波士顿，MA），并创立了 Cel-

era 基因组公司（Rockville，MD）。

1998 年，人类基因组计划宣布将在 Watson-Crick 的 DNA 双螺旋结构发现 50 周年庆典时完成人类基因组测序，当初的目标是：

2001 年完成至少覆盖基因组 90%的工作草图；

2001 年底完成 1/3 的人类 DNA 测序；

2003 年完成全部的人类基因组序列测定，全部组装序列并可以自由查询。

2000 年 6 月 26 日，美国总统克林顿会见人类基因组计划负责人 Francis Collins 和 Celera 公司的 Craig Venter，宣布他们都提前两年完成了人类基因组计划的工作草图，草图发表于 2001 年《科学》和《自然》特刊上，序列公布在美国 NIH 医学图书馆的国立生物技术情报中心（NCBI）网站上。

人类基因组最初的计划是用至少 15 年时间、花费 3 亿美元，最后由于技术的改进提早了 3 年，且花费大大降低。人类基因组计划为大规模阐明生命的遗传调控机制铺平了道路，为药物基因组学、诊断学和药靶鉴定打下基础，也是基因芯片技术产生的前提和动力。

最初的基因组学重点在于研究有机体基因组的遗传图谱、物理图谱和 DNA 序列，以测序为基础的基因组学是结构基因组学（structural genomics）。但序列的解码仅仅是获得了一部“天书”，要读懂这本“天书”还需要解码这些 DNA 的生理功能，只有了解其功能才能真正体现 HGP 的价值，才能破译人类基因这部天书并造福人类。目前对于基因功能的了解远远落后于对基因序列的了解，因此为实现这一目标，近年来提出了后基因组计划，后基因组研究的重点主要集中在基因功能的研究，因此又称为功能基因组研究（functional genomics）。功能基因组学是以全面研究所有基因功能为中心，并结合基因功能解决生物医学中的基础和应用问题，功能基因组学除了转录组学、蛋白质组学外，还包括在此基础上产生的不同分支学科，如疾病基因组学、药物基因组学等，即以“-omics”为后缀的新学科。另外，基因多态性研究也是后基因组研究的内容之一，它虽然属于结构基因组学的范畴，但与功能基因组学密不可分，重点是研究基因多态性与表型的关系，因此是功能基因组研究中必不可少的组成部分。

在后基因组学研究方面，我国相继启动了基因组计划、SNP 计划、单倍型（haplotypes）计划、功能基因组计划和蛋白质组学计划。我国在 2002 年启动的“十五”科技重大专项“功能基因组和生物芯片”中，6 个专题中有 5 个是与人类疾病密切相关的，如人类重大疾病相关基因研究、中华民族单核苷酸多态性的开发应用以及与人类重大疾病及重要生理功能相关的蛋白质研究等，并最终在基因组药物和药物靶标的开发上取得一定的进展，这些研究将有助于基因组学向纵深发展。在基因多态性研究方面，我国已在 SNP 研究领域中开展了中华民族基因组特点与序列多态性研究的 SNP 目录建立工作，还参加了国际合作 HapMap 的 SNP 项目并承担了 10%的任务，在此任务中我国将提供全部基因样品的 1/6，并负责 3 号、21 号和 8 号染色体短臂的单体图谱构建和近 2000 万个基因多态性位点的测定。

第二节　基因芯片技术简介

生物芯片分析技术在生物学的历史上是独特的，首先它是在基因组学和后基因组学的基础上产生的，它提供了一种高通量和系统性的研究手段，很好地迎合了后基因组学研究的需求；第二，没有一种技术像生物芯片技术这样涉及如此多的学科，需要不同领域的专家共同努力；第三，它不仅对基因功能的基础研究具有重要的价值，而且也具有明显的产业化

前景。

一、基因芯片的基本概念

基因芯片是最主要的且发展最早、最快的生物芯片，它的概念源自于计算机芯片。计算机芯片是指将不同功能单元集成在一块微型器件上，基因芯片借用了计算机芯片的集成化特点，运用缩微技术，把核酸密集有序地排列固定在固相平面载体预先设置的区域内，形成微型的检测器件，固相载体通常是硅片、玻片、聚丙烯或尼龙膜等，其中玻片是最常用的载体材料。将待测样本标记后同芯片进行杂交，检测原理是利用核酸配对原理，样本中的标记分子与芯片上的配对探针分子特异性结合，通过激光共聚焦荧光扫描仪或其他检测手段获取信息，经电脑系统处理、分析得到信号值，信号值代表了结合在探针上的待测样本中特定大分子的信息，从而检测对应片段是否存在、存在量的多少。其中最成功的典型的基因芯片是在介质表面有序地排列 DNA 点阵，因此狭义的基因芯片又叫 DNA 微阵列（DNA microarray）。微阵列（microarray）是一个新的科学名词，来源于希腊文字 mikro（small，小）和法国文字 arayer（arranged，安排），主要包括 cDNA 微阵列和寡核苷酸微阵列。

广义的基因芯片还包括微流体芯片和“芯片实验室（lab-on-a-chip），即能对核酸分子进行快速并行处理和分析的厘米见方的固体薄型器件。将微阵列技术与生物微机电技术相结合，通过微加工技术和微电子技术在固体基片表面构建微型的生物化学分析系统，以实现对核酸准确、快速、大信息量的检测。

由于芯片上可以固定成千上万的探针，因此可以同时检测样本中成千上万的序列，而传统的检测方法一次只能检测一个或几个序列，因此一次芯片实验就完成了成千上万个传统实验，即一次生物芯片反应是多次传统实验的集成。基因芯片也可以将生命科学中许多独立的反应过程集中在芯片上，使其可以连续、迅速地完成，即将分离、标记、反应、检测等多个步骤进行集成，使这些分析过程连续化、微型化、集成化和自动化。它可以用于基因的功能研究和基因组研究、疾病的临床诊断和检测等众多方面。

二、基因芯片技术的产生和发展

人类基因组计划推动了后基因组或功能基因组研究，要同时研究生物体成千上万条基因的功能，特别是研究基因与基因之间表达与调控的复杂网络关系，显然传统的以杂交或电泳为基础的基因表达、测序、突变检测和多态性分析等研究方法效率太低，无法适应基因组与功能基因组研究的要求。如果利用传统的方法，全世界的科学家一同工作，也需要数百年的时间才能完成。要想高效地研究数万条基因，迫切需要高效的方法和工具，能大规模、高通量地检测众多基因在各种生理状态下的表达全貌。基因芯片技术正是在这种环境下应运而生的，它为满足人类对数以万计基因的研究和应用的迫切需要而发明，被评为 1998 年度世界十大科技突破之一 。

基因芯片技术是在不同学科和技术的基础上产生的，是典型的多学科、多技术交叉的结晶，它涉及物理学、化学、材料科学、生物化学、核酸化学、分子生物学、遗传学、毒理学、电子工程、机械工程、光学、统计学及计算机科学等，这些学科的研究和先进技术的发展都直接或间接促进基因芯片技术的发展。

基因芯片技术的产生与生物技术在 50 年中的发展密不可分，包括 DNA 体外聚合、重组 DNA 技术、PCR（聚合酶链反应）技术等，其中对于基因芯片来说，DNA 杂交技术的发明起了关键性的作用，它是在传统的膜杂交技术上发展起来的。Francis Crick 和 James D. Waston 于 1953 年在《自然》杂志上发表的那篇里程碑式的论文中写道：“我们不能不指出，由我们设想的配对原则直接预示着遗传物质可能的复制机制。”事实确实如此，核酸专

一性配对方式这一性质已广泛应用到核酸研究的各个领域，Southern blot 等杂交、测序、PCR 等关键的核酸技术无一不是在创造性地应用碱基配对这一基本性质。基因芯片技术则是在一个新的角度再一次创造性地应用了碱基配对原理。

早期的杂交实验一般采用膜作为介质，20 世纪 70 年代，基于硝酸纤维素膜和尼龙膜的杂交技术在斯坦福大学诞生。1975 年，斯坦福大学的 Grunstein 和 Hogness 发表了第一篇关于 DNA 阵列的文章，作者用硝酸纤维素膜的细菌克隆阵列分离果蝇基因。Stanford 大学的 Davis 等用硝酸纤维素膜阵列检查细菌的嗜菌斑，并第一次用于研究高等生物的基因差异表达。但是由于核酸样品在滤膜上容易扩散，因此单位面积上的点样密度受到限制。同时，由于滤膜面积较大而且需要较多探针，故检测灵敏度较低。为了提高点样密度和检测灵敏度、降低探针用量，以玻璃、硅片等材料为载体的 DNA 芯片应运而生。

荧光染料的应用也是基因芯片诞生不可缺少的因素，荧光染料在生物技术中的应用也已有 30 年左右的历史。早在 20 世纪 70 年代，Waggoner 和 Stryer 将荧光染料用于研究生物膜，Pinkel 等在 20 世纪 80 年代末至 90 年代初发展了双色荧光标记和荧光显微镜检测策略，并将其用于染色体分析，Yu 等 1994 年报道用花青素（cyanine）荧光染料通过酶反应标记 DNA。这些先驱的工作成果后来都用于基因芯片的标记和检测中。

杂交测序法的提出和实施直接促进了基因芯片的形成。20 世纪 80 年代末，俄罗斯科学院恩格尔哈得分子生物学研究所（Engelhardt Institute of Molecular Biology）的 Mirzabekov 等（1989）提出了用杂交法测定核酸序列（SBH）的想法。他们将八聚体寡核苷酸探针固定在玻璃介质上进行杂交测序。几乎与此同时，英国牛津大学生化系的 Southern 等也取得了在载体上固定核苷酸及杂交法测序的国际专利。此后，Affymatrix 公司的 Fodor 等（1991）、牛津大学的 Maskos 和 Southern（1992）、Baylor 大学的 Eggers 等（1994）、Wisconsin 大学的 Smith 等（1994）也分别报道了这方面的研究。最早将机械手臂用于阵列制备的是英国皇家癌症研究基金（Imperial Cancer Research Fund，ICRF）的 Hans Lehrach 等（Nizetic 等，1991）。20 世纪 80 年代末，他们利用机械手臂和实心针将人的基因组 DNA 克隆点在大的尼龙膜上，制备成阵列。这些技术的产生为基因芯片技术的产生奠定了基础。

基因芯片产生的两个标志性事件是 1991 年 Affymetrix 公司生产的第一块原位合成的基因芯片和 1994 年斯坦福大学用直接点样法制备的第一块微阵列 cDNA 芯片。DNA 芯片技术的发展还与寡核苷酸的高密度空间合成技术的建立有密切关系，而照相平版印刷技术以及激光共聚焦扫描技术的引入则是 DNA 芯片实现工业化生产的条件。美国 Affymetrix 公司在 20 世纪 80 年代末至 90 年代初征集了多位计算机科学、数学和分子生物学专家率先开展了这方面的研究。1991 年，Affymetrix 公司运用半导体照相平版技术，在 $1cm^2$ 左右的玻璃片上原位合成寡核苷酸片段，得到了世界上首张寡核苷酸基因芯片（DNA chip），并且很快就应用于 DNA 序列分析中。1994 年，俄罗斯科学院研制出了一种基因芯片，用于检测 β-地中海病人血样的基因突变，筛选了一百多个 β-地中海贫血已知的突变基因。这种基因芯片的基因译码速度比传统的 Sanger 法和 Maxam Gilbert 法快，是一种新的快速测序方法。这段时间基因芯片（DNA chip）专指这种寡核苷酸芯片，但是专利保护和芯片制备工艺的难度限制了这一先进技术的迅速推广。1994 年，斯坦福大学 P. O. Brown 实验室综合了多种技术发展了直接点样的基因芯片制备技术，设计了机械手臂将 PCR 扩增的 cDNA 点在经过修饰的载玻片上，从而制备了第一张 cDNA 微阵列芯片，并创造性地将双色荧光杂交系统应用到 cDNA 微阵列上。该方法用于比较拟南芥 mRNA 和人 AchR mRNA 基因表达水平差异，这

也是首次将基因芯片用于大规模基因表达差异分析，初步显示了基因芯片技术在基因表达研究方面的巨大潜力。由于直接点样法方法较简便、易推广，而且对探针的要求低，可以是任意长度的单链或双链的核酸，还可以是蛋白质或其他任何种类的分子，甚至是细胞或组织，因此当这一技术公之于众后，吸引了众多的研究者和公司，使该技术迅速得到普及和推广，使基因芯片技术步入了全面研究和应用的新时代。

此后短短 10 年的时间，在 DNA 微阵列技术的基础上，又逐渐出现了以蛋白质、组织和细胞等为材料的芯片，统称为生物芯片，包括蛋白质芯片、组织芯片、微球体芯片（microspheres）、微流体芯片和芯片实验室（lab-on-a-chip）等。国际上出现了生物芯片热，世界上专门从事生物芯片技术研究的单位已经有几百家，进行生物芯片研究和生产的公司也已经有几百家，生物芯片技术如同 PCR 技术一样，已成为生物医学研究必不可少的实验手段，深入到几乎所有的生物学研究领域之中，已经开始在生命科学研究中发挥重要作用。

三、基因芯片的应用领域

与其他传统的基因检测技术相比，基因芯片技术的最大特征在于能同时定量或者定性地检测成千上万的基因信息。基因芯片技术具有微型化、自动化和网络化等特点。所谓微型化，一方面是指成千上万种探针分子仅仅点在几平方厘米的介质上，另一方面则是指样品和药品消耗量小；所谓自动化，是指点样、杂交、图像处理和数据处理都可以用计算机指导的自动化系统自动或者半自动地完成；所谓网络化，是指点样和数据处理都需要利用 Internet 上庞大的生物信息数据库，如 GeneBank。基因芯片技术具有传统的生物技术不可比拟的高效、快速、多参量等特点，是生物技术发展史上的一次飞跃，已成为生命科学领域一项最强大也最具有应用前景的生物技术之一。

就像几个世纪以前显微镜的发明一样，基因芯片为研究复杂的生物系统提供了一个新的视野，成为目前使用非常广泛的技术。概括地讲，其主要用途有两大方面。第一，它可以从转录水平测定一种细胞在特定时间内的基因表达概貌，由于基因表达与基因功能密切相关，是基因型和表型之间的基本连接点，它对于研究基因调控、发育、疾病等复杂的生物系统和过程起了主要的作用，因此可以研究基因功能、生长发育、疾病的产生等生物学命题。基因芯片表达谱技术将为同时检测生物体所有基因在特定组织、特定条件下 RNA 表达水平的整体面貌提供可能。第二，基因芯片还可以用于高通量基因组分析，如比较基因组杂交（CGH）、SNP 和甲基化分析等，其中典型的用途是用于检测基因序列及其变异，用于基因分型。基于基因芯片的基因分型技术将为同时检测生物样本中成千上万位点的基因型提供一线光明，这将为人类从整个基因组范围内研究复杂的基因型差异对遗传的影响，特别是对疾病产生与治疗的原因的理解产生深远影响。用于疾病基因的突变筛选时，基因芯片技术将可能使个体等位基因的突变型与具体疾病的治疗直接联系起来，这种主要针对 SNP（single nucleotide polymorphism，单核苷酸多态性）位点的研究已经成为一个新的研究热点。

在所有基因芯片相关论文中，基因表达研究方面的应用目前占基因芯片论文的 81%，其他的应用包括基因分型、组织分析和蛋白质研究。基因芯片的出现正在给生命科学研究、疾病诊断、新药开发、生物武器战争、司法鉴定、食品卫生监督、航空航天等领域带来一场革命，下面就列举几个主要的应用领域。

1. 发育

多细胞生物中每个细胞含有相同的遗传物质，但为什么细胞的形状、大小和功能如此多样？由于遗传程序的控制，改变了不同细胞的基因表达指令。通过基因芯片进行全基因组表

达谱研究，可以了解一个细胞如何发育成含有约 10^{15} 个细胞的复杂个体，可以建立各种细胞和组织在不同时期的基因表达数据库，这对研究发育过程中的分子调控很有意义。

2. *疾病研究*

疾病的产生是多种因素作用的综合结果，包括遗传、饮食、环境和感染因素等，疾病与基因表达调控异常有密切关系，不同疾病导致不同的基因表达异常。基因芯片技术可以用于所有的疾病研究中，其中癌症是应用基因芯片研究最多的疾病，占发表论文的 84%左右。这是因为癌症的病因复杂，涉及基因组水平、基因调控水平和表观遗传水平的改变，常规方法很难全面展开研究，采用基因芯片技术有明显的优势。该技术的限制是样本的需求量较大，其他很多疾病不易获得足够的细胞，而肿瘤细胞量大，材料容易获得，这是有关肿瘤的研究报道多的另一个原因。其他疾病，如糖尿病、心血管疾病、囊肿性纤维化、艾滋病、帕金森症、孤独症和贫血病等，也利用芯片技术进行了广泛的研究。通过比较正常组织和疾病组织的表达谱差异，可以了解疾病发生的分子基础，从而更好地预防和治疗疾病。

3. *药物发现*

很多药物通过与细胞的特定靶标结合从而抑制蛋白质的功能并影响基因的表达。从理论上可通过比较病人疾病发生的过程及用药过程基因表达变化情况，从而实现药物的发现、毒理研究和临床药效研究。这样可以大大降低药物研究的成本、缩短研发周期、降低药物的毒副作用并增强用药的安全性。

4. *遗传筛查和诊断*

遗传物质的小的改变（基因突变或基因多态性）可以导致蛋白质不能行使正常功能，从而导致疾病的发生或导致对疾病的易感性。目前已发现的疾病基因与疾病相关基因有上千个，其中包含着大量的与疾病相关的基因突变或 SNP，另外，不同的基因型会导致对药物的不同耐受能力。人类基因多态性在阐明人体对疾病、毒物的易感性与耐受性，疾病临床表现的多样性以及对药物治疗的反应性上都起着重要的作用。因此，如果能阐明基因型（genotype）与表型（phenotype）之间的联系，对每个个体的遗传体质进行筛查，针对每个个体的特征进行个性化诊断，根据诊断结果采用不同的治疗方案、选择不同的药物或不同的用药剂量，就能实现个性化治疗。个性化诊断还可以对疾病进行早期诊断和预测，从而更好地预防疾病。例如，不同的人每天维生素 C 的摄入量可能有 3 倍的差异，而几乎所有的营养品和药物对于不同个体的最适摄入量还未知，DNA 芯片研究提供的信息将有助于定量分析这些营养品或药物对人体健康的影响。

第三节　生物信息学与基因芯片的数据挖掘

一、生物信息学的兴起

随着实验技术的不断革新和发展，生物学数据日益膨胀。以 DNA 序列为例，1997 年的 GeneBank 数据库有 176 万条序列，到 2002 年，数据量就增加到 2232 万条。数据量的积累使得人们有可能从基因组的角度研究生物，这就形成了以基因组和功能基因组为主要研究对象的生物信息学。

生物信息学是一门正在兴起的新学科，它顺应了基因组学和功能基因组学研究的需求，现在的测序、基因芯片等实验手段能产生大量的数据，这些数据必须经过加工，才能读懂并得出结论，如揭示新基因的序列、定位、功能注释、途径的阐明及其在细胞组织中的调控机制和网络等。为了实现这些目的，首先需要建立各种数据库，再通过具备专门知识的研究人员根据原始数据逐一地加以整理和分析，并发展数学统计算法和模型。我们希望将来对知识

的操纵就像现在货物的生产一样，最终目标是实现自动分析加工，使数据成为“知识”。

生物信息学就是为了解决这一问题而产生的，它是介于计算机和生物学之间的学科，它是利用信息技术的方法、运算法则和工具，解决生命科学的问题，利用计算机将生物的信息整理成生物知识。最初生物信息学仅用于数据的贮存，后来逐渐赋予生物信息学更多的内容，包括解决生物学问题的运算法则和技术，目前使用的方法来自于计算机科学、统计学和数学等。以前生物信息学研究的内容主要有序列分析、蛋白质结构预测和复杂生物系统的建模，随着技术的发展，不断产生新的数据，如蛋白质-蛋白质相互作用、蛋白质-DNA 相互作用、酶和生化途径、群体规模的序列数据、大规模的基因表达和生态以及环境数据等，因此研究的范围逐渐扩展，最终从整体和系统的角度来研究生命的规律。

未来的挑战是分析、阐明和理解所有产生的数据，生物学家的任务是如何发现和理解基本的生物现象；计算机学家的任务是发展新的算法和技术来支持这些科学发现。

二、基因芯片的数据挖掘

基因芯片技术目前是研究人类基因组和其他各种模式生物基因组复杂性的最强有力的工具，这一技术已经广泛应用于生物学和医学的各个研究领域。在技术迅速发展的同时，数据也在不断地增加，如何有效地处理和管理芯片实验所产生的海量数据越来越引起研究者们的广泛关注，基因芯片的数据分析已成为生物信息学的一个新的重要的分支。

基因芯片数据分析简单来说就是对芯片的高密度杂交点阵图像进行处理并从中提取杂交点的荧光强度信号进行定量分析，通过有效数据的筛选和相关基因的聚类，最终整合杂交点的生物学信息，发现基因的序列或表达谱与功能可能存在的联系。芯片数据挖掘包括众多方面，如图像分析、标准化问题、差异基因确定、聚类分析及基因功能注释等。基因芯片数据分析流程如图 1-1 所示。

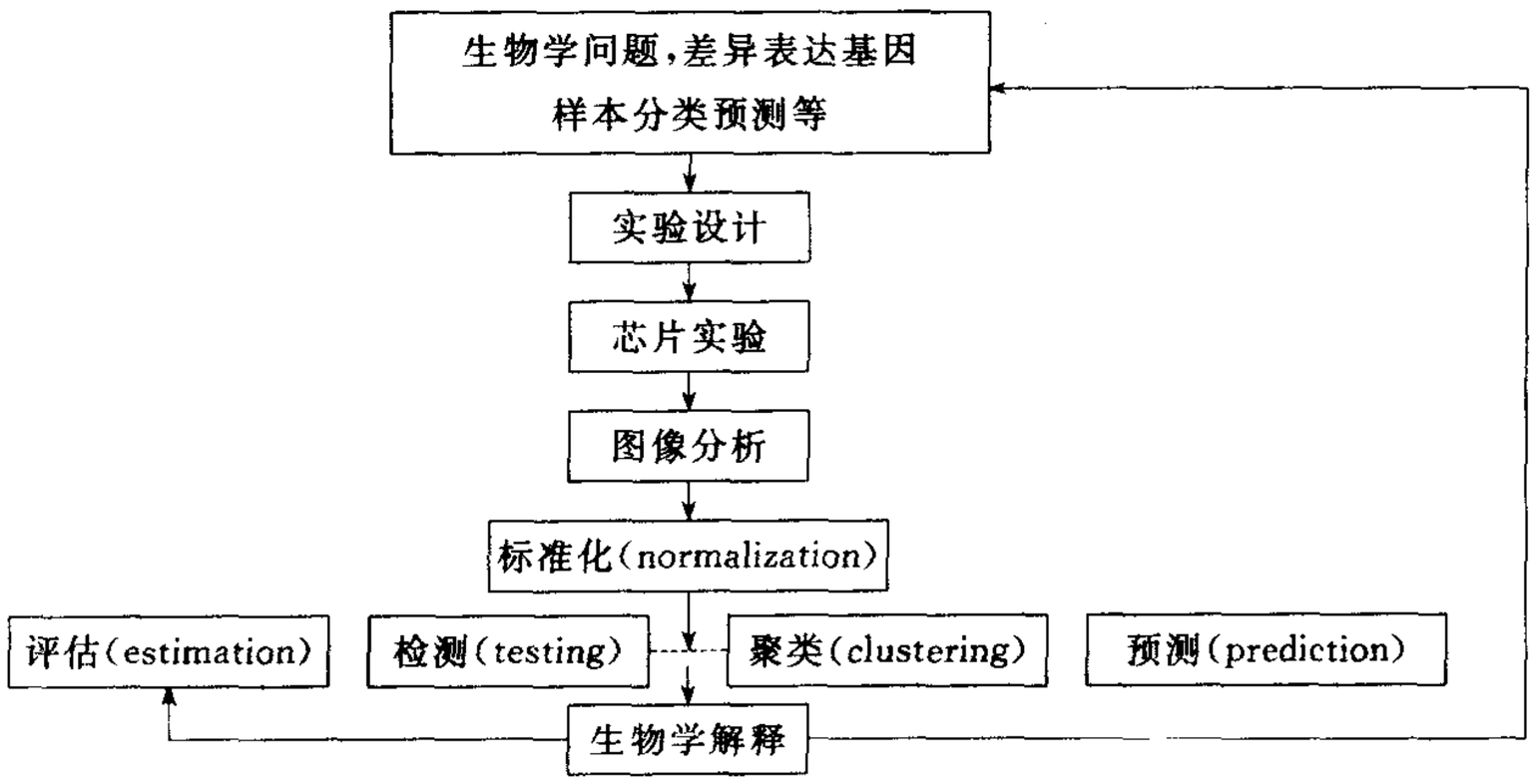


图 1-1　基因芯片数据分析流程

目前 DNA 微阵列最广泛的应用是大规模研究基因的表达，将对照组和实验组样本的 mRNA 分别标记后，与芯片上的互补序列配对杂交，通过杂交信号的强弱来反应表达的丰度及其表达改变的情况。为了研究生命现象，需要比较不同情况，如正常个体与疾病个体之间、同一个体的不同时间段之间的表达水平。通常，一次基因芯片实验会产生成千上万个数据，随着数据的大量出现，新的问题随之而来。由于这些数据不能直接揭示生命现象，如果将所有获得的数据集中起来，我们能否将未知功能的新基因归类到已知的功能分类中，能否将基因表达与基因功能联系起来，能否发现新类型的共调控基因，能否从芯片表达数据中发现完整的基因调控网络，能否揭示疾病或其他表型的分子机理，这些都需要强有力的分析方

法来挖掘这些数据、管理分析芯片数据，从中挖掘有用的信息。但是掌握这些分析方法的数学家、计算机专家和统计学家往往缺乏生物学的基本知识，而有很好生物学知识的生物学家又不具备相应的分析技能来挖掘这些数据。

现代的基因芯片技术与17世纪的显微镜技术的显著不同在于输出结果的不同，虽然都产生图像，但显微镜的图像可以直接显示结果，而基因芯片的图像不能直观地阐明结果，必须转化成数字并加以贮存，最终产生以基因为行、实验样本为列的矩阵数据，由于基因芯片上含有成千上万条基因，而样本的数量却往往在100例以下，因此该阵列不同于通常的统计数据，常规的统计往往只有少数几个参数，但样本量大，数百例甚至上千例；基因芯片的数据矩阵则含有成千上万行而只有几十列，因此在数据统计分析上也不完全等同于常规的生物统计学。

需要强调的是，DNA芯片技术还处于发展的早期阶段，从实验技术到数据形式都没有统一规范，有关芯片数据的可靠性、噪声、校正和统计显著性分析等问题还需要完善和改进，这些问题没有解决和标准化，即使对于简单的原核系统也难以完成复杂的遗传调控网络的研究，而有关调控网络研究的进展主要从这些简单的生物着手而获得最初的知识。

由于以上的原因，迫切需要发展适合基因芯片的生物信息学工具，有关基因芯片的生物信息学研究包括数据库、数据标准化、弱信号处理、均一化、数学分析方法、聚类方法、基因注释等问题的研究，最终解决生物科学的基础理论命题，如帮助人们揭示基因型、表型、环境之间复杂的网络关系；在人类健康方面，基因芯片帮助人们了解疾病，建立新的诊断方法，并可在未来实现个性化诊断和治疗。

（李瑶）

参 考 文 献

1 Baltimore D. RNA-dependent DNA polymerase in virions of RNA tumour viruses. Nature, 1970, 226: 1209～1211

2 Benton W, Davis R. Screening lambdagt recombinant clones by hybridization to single plaques in situ. Science, 1977, 196: 180～182

3 Blanchard A, Keiser R, Hood L. Synthetic DNA arrays. Biosensors and Bioelectronics, 1996, 11: 687～690

4 Fodor S, et al. Multiplexed biochemical assays with biological chips. Nature, 1993, 364: 555～556

5 Fodor S, Rava R, Huang X, et al. Light-directed, spatially addressable parallel chemical synthesis. Science, 1991, 251: 767～773

6 Grunstein M, Hogness D. Colony hybridization: a method for the isolation of cloned DNA that contain a specific gene. PNAS, 1975, 72: 3961～3965

7 Guo Z, Guilfoyle R, Thiel A, et al. Direct fluorescence analysis of genetic polymorphisms bu hybridization with oligonucleotide arrays on glass supports. Nucleic Acids Res 1994, 22: 5456～5465

8 Maskos M, Southern E. Oligonucleotide hybridization on glass supports: a novel linker for oligonucleotide synthesis and hybridization properties of oligonucleotides synthesized in situ. Nucleic Acids Res, 1992, 20: 1679～1684

9 McGall G, Labadie J, Brock P, et al. Light-directed synthesis of high-density oligonucleotide arrays using semiconductor photoresists. PNAS USA, 1996, 93: 13555～13560

10 Mullis K. The unusual origin of the polymerase chain reaction. Sci Am, 1990, 262: 56～61

11 Pauling L, Itano H, Singer S, et al. Sickle cell anemia: a molecular disease. Science, 1949, 110: 543～548

12 Sanger F, Nicklen S, Coulson A. DNA sequencing with chain-terminating inhibitors. PNAS, 1977, 74: 5463～5467

13 Schena M, Shalon D, Heller R, et al. Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes. PNAS USA, 1996, 93: 10614～10619

14 Schena M. DNA microarrays, a practical approach. Oxford University Press, 1999, 12～20

15 Southern E. Arrays of complementary oligonucleotides for analysing the hybridization behavior of nuclei acids. Nuleic Acids Res, 1994, 22: 1368～1373

16 Watson J, Crick F. Molecular structure of nucleic acid. A structure for deoxyribose nucleic acid. Nature, 1953, 171: 737～738

17 Weier H, Lucas J, Poggensee M, et al. Two-color hybridization with high complexity chromosome-specific probe DNA allows unambiguous discrimination between symmetrical and asymmetrical translocations. Chromosome, 1991, 100: 371～376

第二章　微阵列基因芯片实验技术

前面的章节中已经介绍了微阵列基因芯片的基本概念、起源和主要应用。具体到基因芯片的应用，可以分为两大类：特定基因的定性检测和基因表达（量）的分析（表达谱基因芯片）。Southern 杂交、Northern 杂交和斑点杂交主要用于特定基因的检测，特异性和灵敏度是主要的检测指标，在这个方面基因芯片也不例外。基因芯片由于具有更高的探针密度，允许设置足够多的对照探针以及更小的杂交体积，其特异性和灵敏度都有很大程度的提高，使得基因芯片在单纯的特异基因的检测方面得到迅速发展。目前应用的主要有基因检测（诊断）芯片、基因分型芯片等。这类芯片涉及的数据分析主要是阈值（cut off 值）确定、灰值区间设定和阴性或阳性判断等，属于简单类型的数据处理。本书讨论的重点是表达谱基因芯片的数据分析与处理，是在简单数据处理基础之上的复杂统计分析。表达谱基因芯片是在 mRNA 水平检测组织或细胞中各个基因的表达情况，基因芯片一般不直接检测各基因表达的绝对量，而是通过各基因之间表达量的相对变化来间接反映。这也是我们称之为“基因表达量的分析”而不能称之为“基因表达的定量分析”的原因。虽然基因芯片制备的工艺和质量是芯片实验的关键部分，但是优秀的芯片设计可以为后期的数据分析提供良好的基础。在第四章中会有芯片实验设计的详细介绍，本章节重点介绍微阵列基因芯片的制备和检测技术。

现在一般称传统的核酸杂交技术为正向杂交，即把目的基因或基因组/PCR 片段的酶切产物转移到固相支持物上，用标记的探针与之杂交（探针一般指已知序列的特定 DNA 片段）。而基因芯片采用的是所谓反向杂交技术，是将各种不同的探针按有序的方式固定到固相支持表面上，再与样本中标记的靶基因进行杂交，靶基因一般指待检测的 DNA 序列。这样做起码有以下好处：①由于建立了基因探针与芯片点阵一一对应的关系，基因芯片几乎可以满足涵盖所有核酸杂交领域的实验要求，不仅能够针对某些特定基因“点”进行检测，而且更能在组织、器官等基因表达谱“面”的研究上大显身手；②基因芯片的商业化发展和现代网络信息技术的应用，能使世界各地实验室的实验数据实现共享，而这一点越来越显示出基因芯片的巨大发展前景，因为某些疾病样本的珍稀性和地域性，数据共享无疑是非常重要的，而且即使所用的芯片种类不同，也可以根据相关共有基因的表达（如持家基因），将两组不同芯片的数据“接轨”，从中得到更多有用的信息，无疑此举将大大地加快生物医药领域的研究步伐。

第一节　基因芯片的价值和分类

一、基因芯片的价值

基因芯片的制造工艺和方法各不相同，因此各类芯片的制造和检测成本也千差万别。价格一般是应用基因芯片的研究人员首先需要考虑的因素，如检测芯片的裸价、检测费用和后续服务等，下面就以基因芯片的价值为总的评估指标进行简单的分析。

研究人员首先关心就是芯片上有多少点（spot）。需要注意的是，芯片上点、探针和基因数量不具有同等的意义，比如说有个 10000 点的基因芯片，这 10000 个点可能是代表 10000 条基因的不同探针，在这种情况下，三者的概念是一致的。但多数情况下这 10000 点

只是5000条基因探针重复点样一次。如果每条基因分别有5个不同的探针来表示，并且都重复点样一次的话，这10000点的芯片实际是有5000条不同探针，而只代表了1000个基因。因此，了解基因芯片的实际情况是研究人员应用商业固定芯片的第一步。

从经济性角度讲，“大”芯片的单价似乎很低，如Affymetrix的2.5万全基因组芯片约1000美元，每条基因约4美分；而一个定制约200条基因的pathway芯片可能售价超过200美元，平均每条基因要1美元，但对于专项研究者来说仍然划算，既可以每张芯片节约800美元，又可在后续的数据分析时大大减少工作量。随着基因研究的深入发展，这种“小”芯片的需求量将越来越大。

二、基因芯片的分类

基因芯片制备就是根据实验目的和要求，将设计好的检测探针（详细请看实验设计的相关内容）通过一定的方法固定到固相支持物上。根据固相支持物（一般称为基片或片基）、探针来源或固定方法等的不同可以将基因芯片进行分类。

1. 根据芯片的制备方式

根据芯片的制备方式，可以将基因芯片分为两大类：原位合成芯片和直接点样（又叫合成后点样）芯片。原位合成法由美国Affymetrix公司率先研究并应用于基因芯片的制作，是在固相介质表面特定区域逐个碱基地合成已知序列的寡聚核苷酸探针。由于Affymetrix公司的专利保护，原位合成法并没有得到普及，目前仅有Affymetrix、安捷伦等少数的基因芯片生产商采用此类技术制造基因芯片。

由斯坦福大学实验室率先研究发明的点样法由于相对简单易行，无须昂贵的特殊设备(仅对较低密度芯片而言)，很快在其他各大实验室得到普及和应用。合成后点样主要有两种方法：接触式点样和非接触式点样。前者通过刚性的点样针接触基片表面形成样点，是使用最多的方法。点样针有实心的、裂隙的和毛细管型的，还有比较复杂的针与环结构。非接触式点样占的比例较小，由微量液体分配系统喷出一定体积的液体形成样点，点样针不接触基片表面。主要有微螺线阀技术和压电技术两种。

图2-1是对目前主要的基因芯片制作方法的分类。商业制作由于高额投入可以根据不同的要求选用不同的制作方法，而自制主要采用直接点样法，尤其是一些小型实验室，仍然采用以膜为支持片基的低密度的微阵列。

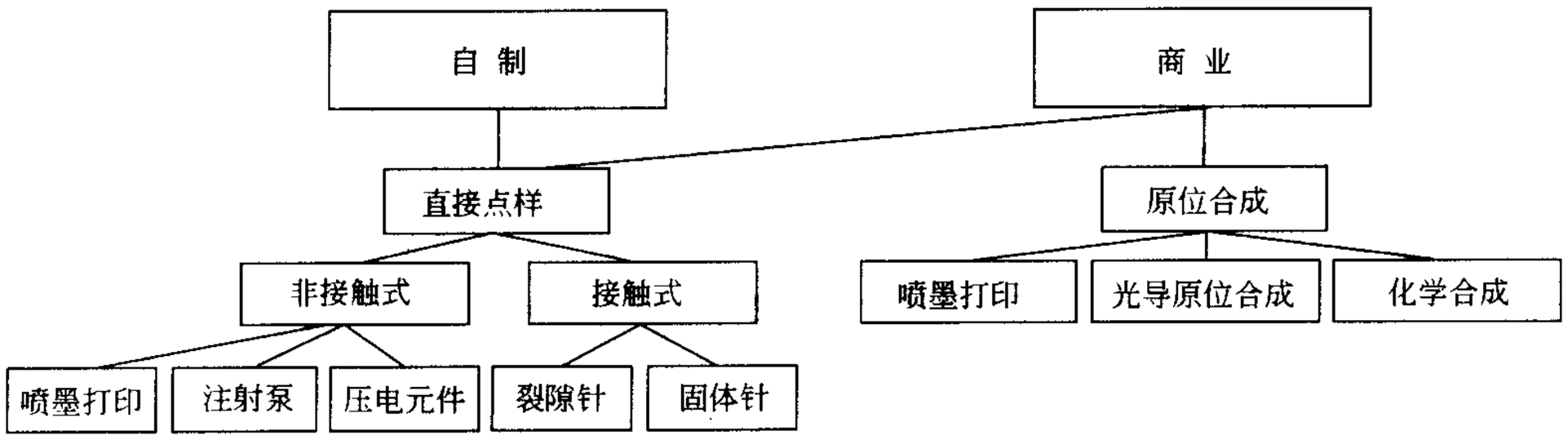


图2-1　目前主要的基因芯片制作方法

原位合成法和直接点样法存在较大的区别，各自有优势和局限。原位合成法的优点如下。

① 可直接从cDNA数据库中得到信息合成寡核苷酸，避免了cDNA样品制备中的不确定因素。直接点样法的探针样品必须事先制备和保存。

② 原位合成减小了芯片批间的差异，可保证芯片质量的高精确度。

③ 原位合成法制备的芯片密度高，目前最高可达 400000 个寡核苷酸片段/1.6cm^2。而直接点样法目前最高只有 64500 个基因/6.5cm^2，通过技术的改进将来可望达到 100000 个基因/6.5cm^2。

原位合成法与直接点样法相比，也有以下的缺点。

① 成本高。

② 设计和制造烦琐、耗时多。

③ 固相合成的寡聚合苷酸长度有限、特异性差，而且随长度的增加合成错误率随之增高。

④ 原位合成芯片仅能用于检测已知序列的基因，而直接点样芯片还可以用来发现和研究新的基因的功能。

由于直接点样法芯片成本低、容易操作、适用面广，而且方便快捷，技术、设备都较为经济，不但适合于商业化生产，也可以满足科研工作者自制芯片的要求，因此在基因芯片生产厂家和自制芯片的科研工作者中得到广泛应用。表 2-1 是原位合成法和直接点样法的比较。

表 2-1　原位合成法和直接点样法的比较

项　　目	原位合成法	直接点样法
探针预合成	不需要	需要
探针类型	寡核苷酸	DNA、抗原、抗体、细胞等
探针长度	约 25nt	不限制
可否自制	不可以	可以
探针固定方式	共价键	共价键、离子键
密度	高，最高可达 40 万/1.6cm^2	较低，最高 6.45 万/6.5cm^2
制作成本	高	低
应用	基因表达，突变检测	基因表达，突变检测，CGH

对于已经制备好的基因芯片，无论是采用原位合成法还是直接点样法，应用上都没有太大的区别，需要注意的是基因芯片采用的材质、探针的类型以及基因芯片的检测范围。商业制作的微阵列基因芯片一般采用配套试剂盒的形式出售，或是提供芯片检测、基本的数据分析等一条龙的服务，大大简化了研究人员的具体操作，使他们能有更多的精力投入到后期的数据分析和判断中。

2. 根据芯片的介质分类

直接点样芯片根据固相支持物（基片）的种类不同，可以分为玻璃芯片、膜芯片、塑料芯片等；其中以玻璃和膜介质最为常见。早期的点样芯片以膜基片为主，是反向斑点杂交的直接应用。现在的点样芯片以玻璃基片为主，与传统的膜介质相比，玻璃类介质具有以下主要优点。

① DNA 样品可以共价结合到表面修饰后的玻璃表面，膜表面的结合则通常是通过物理吸附或者电荷相互作用，前者更牢固。

② 玻璃一般都可以承受高温处理和高离子强度溶液的清洗，膜则需要用温和的清洗条件。

BioChip Arrayer 将压电元件和一末端连有高精确性注射泵的石英毛细管整合在一起。注射泵用来吸取一定体积的液体，通过压电元件的变形压缩，液体能被精确分配。注射泵又被用来复原或洗去毛细管中的液体以完成吸取-分配的循环过程。BioChip Arrayer 目前以 9mm 的间距配置四根点样针（图 2-14）。但是 PerkinElmer 公司已经研发出新机型，配置 8 根针，可将多个控制板连接起来生成多达 48 根或 96 根分配针的 Nanoliter Module 液体分配单位。

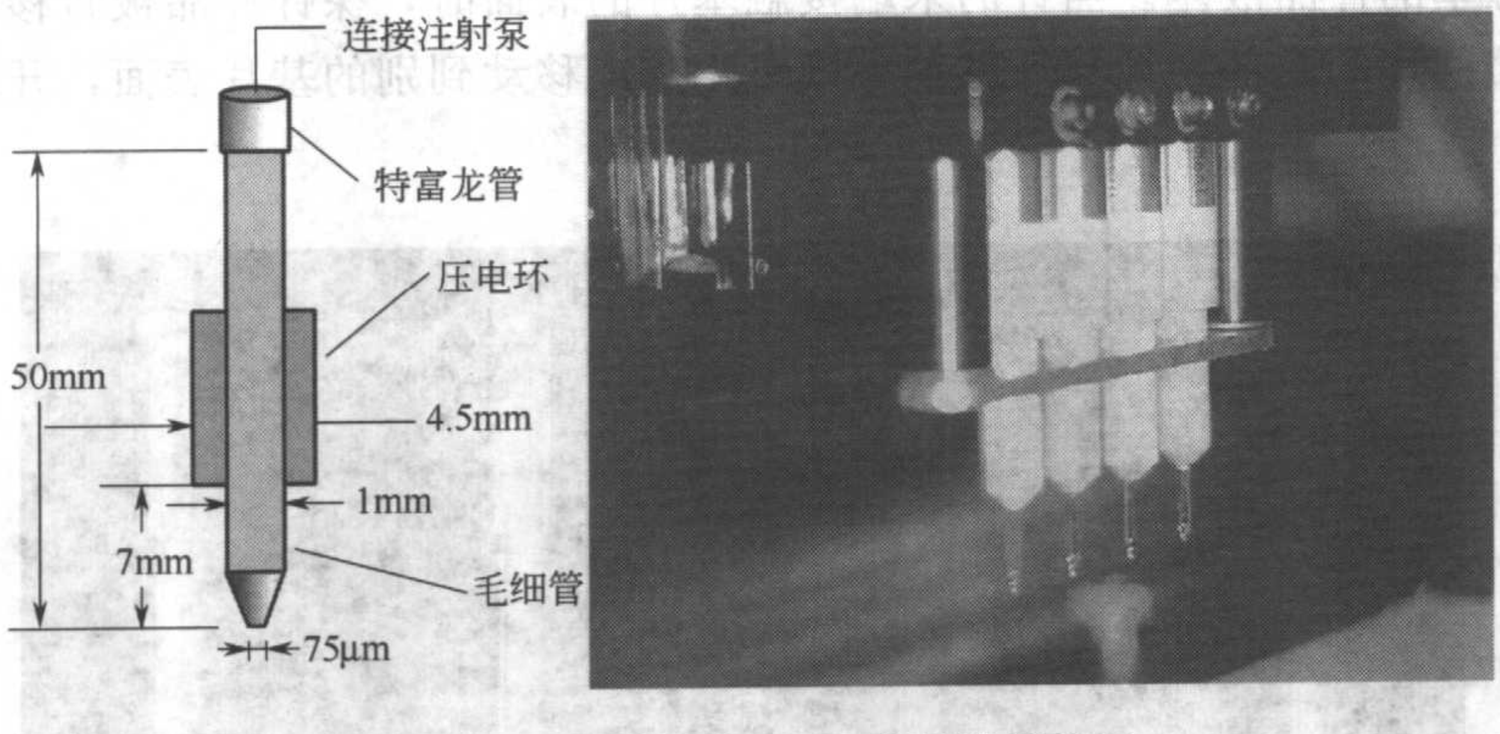


图 2-14　BioChip Arrayer 点样针结构图

（2）非接触式点样中需要注意的因素　相对于接触式点样，非接触式液体分配方法还有一些优势，首先非接触式针头一次可以喷出数滴液体而不用在垂直方向作任何移动。这一优点可大大节省机器平台的空载时间。其次接触式点样的样品盛放需要开放的环境，因此接触式点样都需要湿度控制系统，而实际上湿度的控制很难实现，在较长的点样过程中，点样环境中的蒸发会严重影响样品的浓度，造成芯片上的样品浓度不同。非接触式点样的样品是封闭的，只在点样针末端形成液滴，液滴离开针孔落在基片上的时间很短，具有很好的重复性。

一般认为非接触式点样需要考虑以下几个关键因素。

① 液滴形成　液滴的体积取决于点样针孔的直径和压力波的调制频率，压力波的快速调节可以避免喷嘴变湿，增强液滴形成的重复性。对压电元件系统来说，压力波产生的加速度非常大（约 100000g），推动液滴以 0.5～4m/s 的速度从针孔内射出，由于液滴的质量很小，避免了接触基片表面时飞溅或破碎。

② 液体性质　非接触式点样适合均一性样品的点样，因此要求同一次点样的各种探针纯净度、浓度、长度等都比较一致，才能保证液体黏度/表面张力的一致性，而在正式点样过程中，这些条件往往难以达到或需要很大的代价才能达到。

③ 卫星液滴　当液滴从分配针孔喷出时，在液滴的头部和针孔之间会形成一个尾巴，尾巴拉长时会断裂。如果断尾的残端缩回针孔，能避免形成卫星液滴。但如果尾部同时断裂成两个或多个部分，则将会形成卫星液滴，跟随主液滴到达基片，但卫星液滴有时会脱离轨道，落在靶点以外的地方。解决方法一般选择合适的针孔直径，仔细调节压电元件的收缩时间和幅度以及针头内部的负压，从而有效避免形成卫星液滴，液体的分类比值也有助于定义这些参数。

BioChip Arrayer 降低了卫星液滴形成的概率，在质量控制测试中保证从 0.5mm 孔径的分配针中射出的卫星液滴融合在主液滴中（图 2-15）。

④ 液滴喷射轨道　制作点样针头的原料及表面处理要求很高，其疏水性要非常好，以

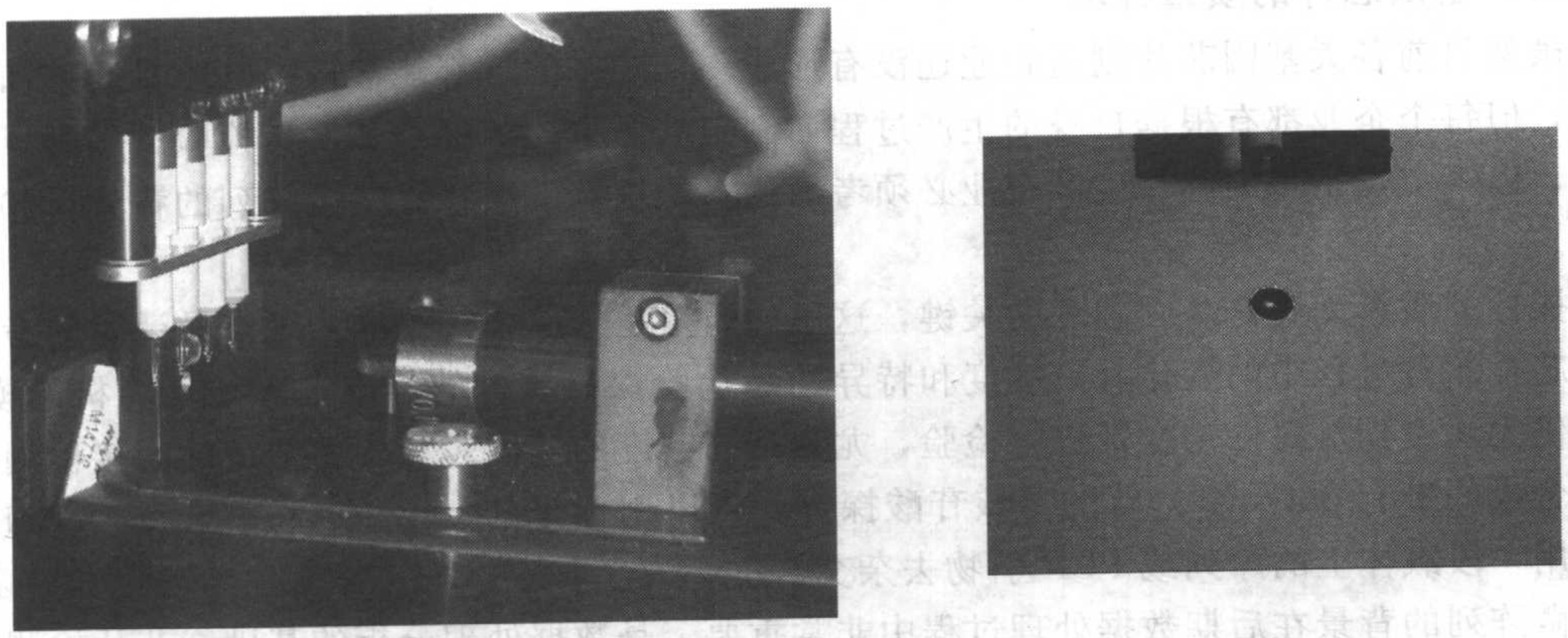


图 2-15　BioChip Arrayer 点样针在接受频闪监测仪检测点样参数（右图为实际点样照片）

防止在点样过程中喷出的液滴黏附在针尖上而造成漏点、大点。点样针针孔形态不规则或点样针上的电极组合不理想是导致液滴从针孔射出后偏离预定轨道的两个主要原因。这种情况发生的可能性随着液体浓度的增大而增加，当液体中含有长链 DNA（大于 10kb）或大蛋白质分子时特别明显。某些溶液会慢慢在分配针管内外表面形成一层膜，影响液体的正确分配。因此清洁点样针的末端对于非接触式液体分配仪器准确地分配液滴是非常重要的。另外，使用者必须控制诸如空气流动和稳定电流等环境因素，以保证非接触式液体分配技术高质量地完成。

⑤ 其他因素　点样针孔会被灰尘、液体沉积物或特殊物质堵塞是非接触式液体分配系统的又一个问题。通过将非接触式点样仪安置在一个清洁的环境中，过滤和离心每一种液体，定期清洁通向点样头的液体回路等方法来解决。同时，点样过程中缩短分配流程中清洗过程的时间间隔、降低某些液体的浓度也是解决这个问题的有效途径。

二、点样后处理

在微矩阵点样完成后，还要进行必要的后处理过程。后处理是芯片制备过程中最后一道工序，也是非常关键的一道步骤，用以增强探针样点在基片上的固定效率，洗去未结合的探针片段，同时还对载体上未与探针结合的游离活性基团进行封闭，以避免在杂交过程中非特异性的吸附对实验结果（特别是背景）造成影响。因此，芯片点样后处理直接影响芯片实验的检测结果，并可在一定范围内提高杂交时的灵敏度。

为了使探针分子与基片表面牢固结合，首先采用潮湿的环境保证化学反应充分进行（水合反应），还可以用紫外交联的方法增强离子键的作用；其次用含 SDS 的溶液洗去未结合在基片上的多余探针，以免杂交时产生干扰；最后用封闭剂（如琥珀酸酐、硼氢化钠等分别用于封闭氨基和醛基）对载体上未与探针结合的游离活性基团进行封闭，以降低芯片背景。

点样后处理的参考方法如下。

玻璃基片的后处理步骤为：水合 30min（提供适宜的湿度环境以保证样品与基片表面在溶液状态下充分反应）或在 4℃冰箱中过夜→紫外交联（巩固样品与基片之间形成的化学键）→室温晾干（如果空气中的湿度非常大，80℃烤干 30min 能有效提高探针的固定效率）→洗涤剂洗涤（0.2% SDS 溶液旋转洗涤 20min）→双蒸水洗涤→室温晾干→封闭液封闭→双蒸水洗涤→室温晾干。

膜芯片的后处理步骤为：80℃烘干（固定样品）→缓冲液中和→双蒸水洗涤→室温晾干。

三、基因芯片的质量标准

虽然目前各大基因芯片制造企业还没有形成一个统一的、都能被认可的质量标准和检测手段，但每个企业都有根据自身的生产过程而制定的质量控制标准和质量检验的细则。归纳一下，以下三点是基因芯片生产企业必须考虑的：①探针的质量；②微阵列的背景；③可重复性。

探针的质量无疑是基因芯片的关键，这里的质量是个综合的标准，既有设计时的考虑又包含芯片制备过程中的变化。灵敏度和特异性是检验探针质量的标准，除了计算机的模拟，现在仍无法做到对每一条探针进行检验，尤其是高密度的基因芯片。而常用的间接检测手段有：合成的带有 poly（T）臂的寡核苷酸探针，可以用标记的 poly（A）去杂交；cDNA 探针可用一段载体上的序列或 PCR 引物去杂交。

微阵列的背景在后期数据处理过程中非常重要，是数据处理分析的基础。芯片检测对背景的要求就是干净、均一。

可重复性一般是指片间重复。片内重复比较容易达到较高的标准，一般在 95%～98% 之间，但片间重复甚至是批间重复，点样芯片能够达到 90%就不错了。而原位合成芯片具有无可比拟的优势，尤其是 Affymetrix 的 PM-MM 探针的应用，有效提高了芯片的可重复性。但这也仅仅是对表达谱芯片而言，对于其他的检测芯片，无论是原位合成的还是直接点样的寡核苷酸芯片，差别并不太大。

第五节　原位合成及纳米结构的基因芯片制备

一、原位合成法制作基因芯片

美国 Affymetrix 公司的寡核苷酸光导原位合成技术是原位合成生产高密度基因芯片的核心技术，图 2-16（彩图见插页）为 Affymetrix 公司的高密度基因芯片图像。下面将详细讨论该法的技术原理及设计原则。另外将简要介绍新近出现的 NimbleGen 公司的无掩膜原位合成技术。

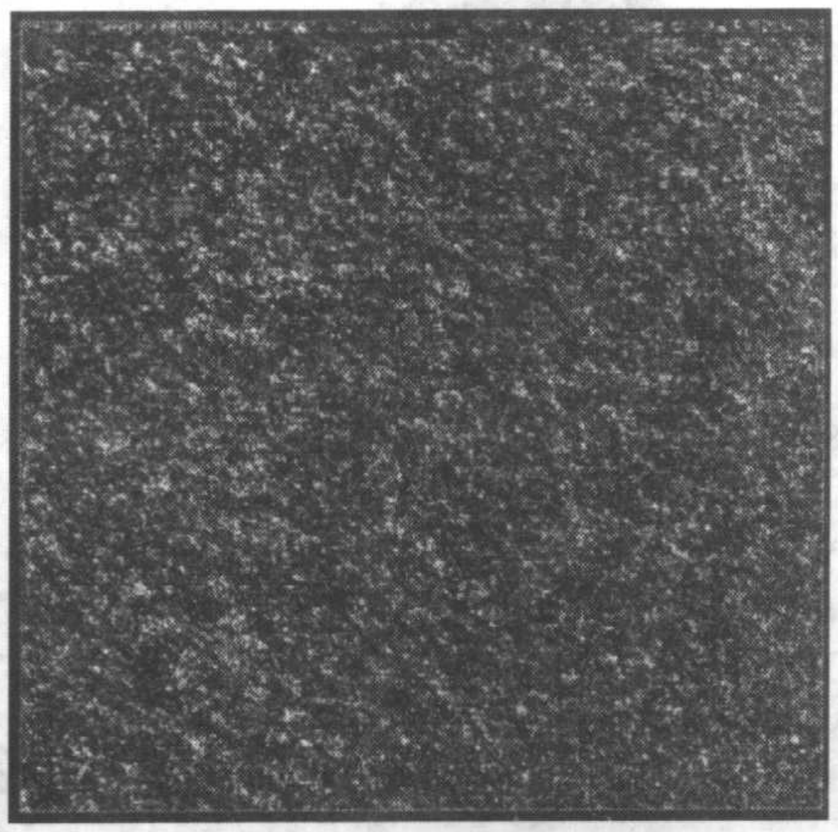
1. 光导原位合成

寡核苷酸的光导原位合成是美国 Affymetrix 公司生产高密度寡核苷酸基因芯片的核心技术。其原理如图 2-17（彩图见插页）所示，这种方法结合了照相平版印刷技术（photolithography）和组合化学技术（combinatorial chemistry），得到位置确定、高度多样性的化合物集合，由这种方法得到的芯片通常称为 GeneChip™。因为合成的步骤是由微矩阵上探针的长度而不是探针的数量来决定的，因此这种方法的效率极高。通过优化算法可以减少蔽光掩膜（mask）的数量，降低生产成本。由于 GeneChip 芯片的密度可以很高（可以高达 40 万点以上），研究者可以使用很小体积的样品获得高质量全基因组的数据。

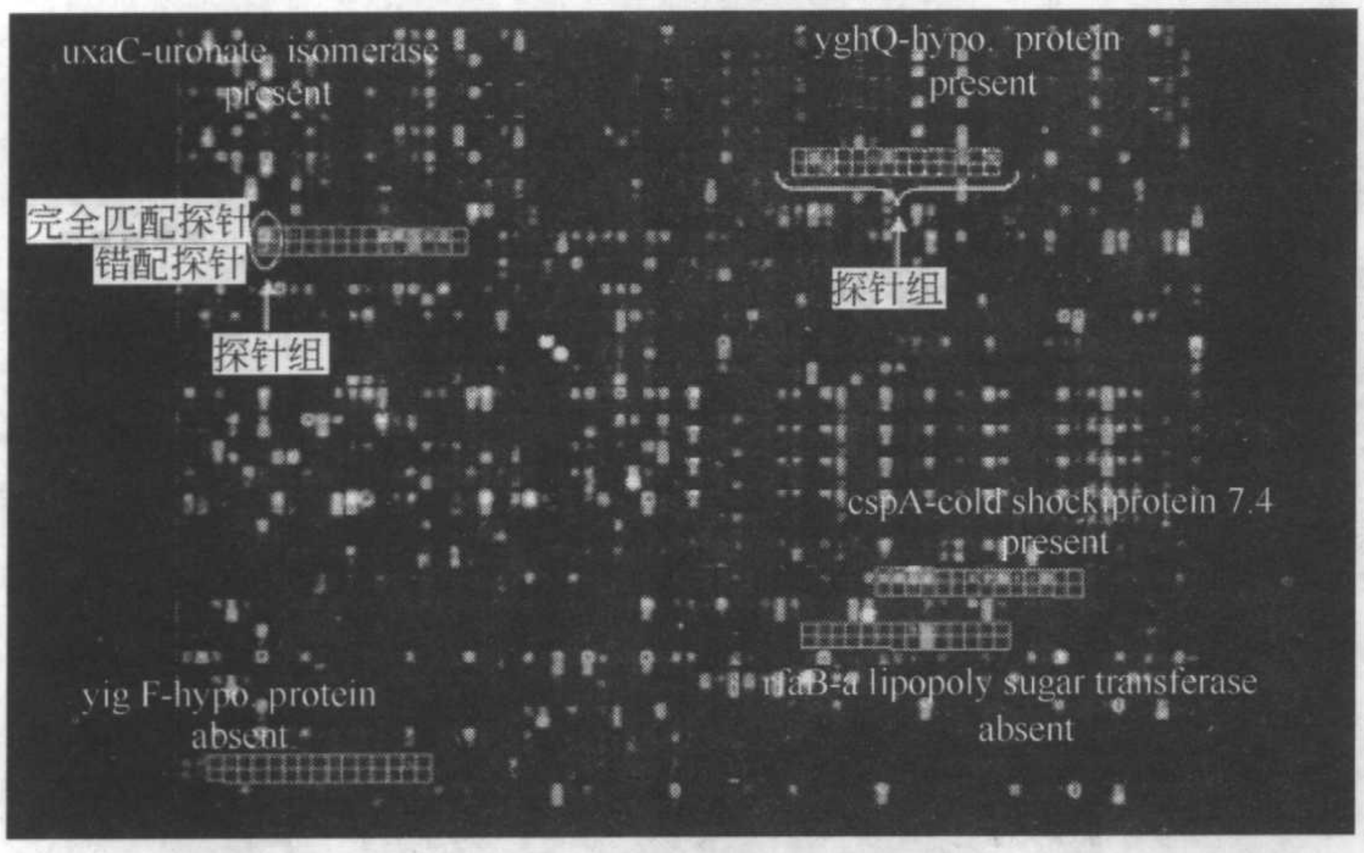
GeneChip 以边长为 12.7cm（5in）的正方形石英片为材料进行生产。表面洁净的石英片是天然羟基化的基质，具有结合其他化合物的良好特性，通过结合带有光敏保护基团的连接分子，可以把探针固定在芯片上。

芯片制备中首先将石英片进行硅烷浴，使之与石英片上的羟基发生作用，形成一层共价结合的连接分子膜。硅烷膜为探针合成提供了密度均匀的羟基。附着在硅烷膜上的连接分子则提供了可以在空间上被光激活的表面。硅烷分子间的距离决定了探针的固定密度，可以在边长 1.28cm 的正方形的面积内固定多于 500000 个探针位点（feature），每一个位点上容纳了数百万个相同的 DNA 分子。

原位合成芯片的每条探针是平行合成的，为了确定每一步合成需在哪一条寡核苷酸链上



(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

接上一个核苷酸，一张带有许多边长 18～20μm 的正方形窗口的蔽光掩膜被放在涂有保护剂的石英片的对应位点上。掩膜上窗口的分布是由所需合成的每一条探针的序列决定的。当第一步合成开始时，紫外线照在覆盖掩膜的石英片上，窗口暴露的位点连接分子去保护，可以进行核苷酸偶联。这一步骤的关键因素是在每一次合成步骤前，必须将蔽光掩膜在石英片上精确定位。为了准确完成这一步骤，在掩膜和石英片上都要进行精确标记。

一旦所需要合成的位点被活化，在覆盖掩膜的石英片表面上加上 5′端带光敏保护基团的单一核苷酸溶液。核苷酸结合在活化载体上，开始合成步骤。因为合成所用的单体分子一端按传统固相合成方法活化，另一端受光敏保护基的保护，所以发生偶联的部位反应后仍旧带有光敏保护基团。因此，每次通过控制蔽光掩膜的图案（透光与不透光）决定哪些区域应被活化以及所用单体的种类和反应次序就可以实现在特定位点合成大量预定序列寡聚体的目的。由于少数活化的分子未能和新的核苷酸发生偶联，为了避免这些漏掉一个核苷酸的 DNA 分子成为探针，用一步加帽步骤截断它们。

接下来的合成步骤里，别的掩膜被加在石英片表面，发生新的一轮去保护和核苷酸偶联。这些步骤不断重复直至合成全长的探针，通常是 25 个核苷酸。

因为每条探针序列的每一个位置上都有可能是 4 个碱基中的任何一种，那么合成 25mer

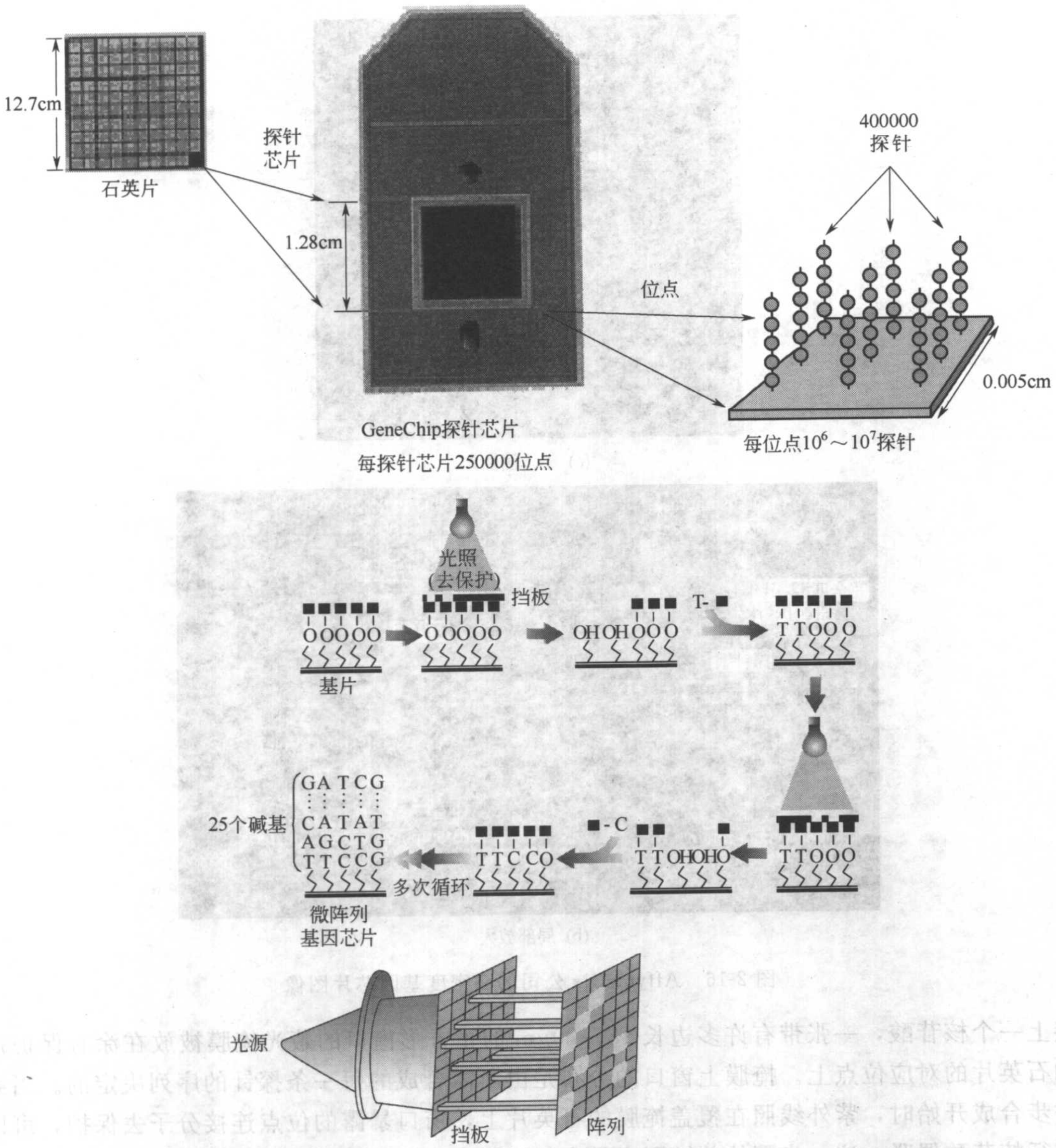


图 2-17　光导原位合成原理图

的所有探针需要 25×4，即 100 种蔽光掩膜才能完成所有的合成步骤。但是经过算法设计后，可以大大减少掩膜的需要量，从而降低生产成本。如通过安排探针的最佳排列，可以使某些掩膜使用多次。

当合成过程全部完成之后，将石英片去除保护基团，切割成小块，每一块即为一张芯片。根据每张芯片上的探针位点数不同，一块单张石英片可生产 49～400 张芯片。每张芯片被装在一个塑料舱中，一方面便于贮存和操作，另一方面也提供了在扫描仪中精确、重复定位芯片的装置。样品室深 1mm，可以容纳 200μL 杂交液体。

光导原位合成法虽然有很多优点，但也有局限。由于需要预选设计、制造一系列蔽光掩膜，光导原位合成法的造价较高。另外，制造过程中采用光脱保护方式，掩膜孔径较小时会发生光衍射现象，制约了探针密度的进一步提高，而且光脱保护不彻底，每步产率只有

92%～94%，因此这种方法只能合成 30nt 左右的寡核苷酸探针，同时探针区域由于存在大量不成功的合成片段，造成杂交背景较高，不适于定量检测。同时，对研究者而言，每次实验只是使用商品化芯片探针中的一部分，探针浪费较为严重。

McGall 将光导合成技术与半导体工业所用的光敏抗蚀技术相结合，以酸作为脱保护剂，每步产率可提高到 98%，解决了由于光衍射造成的脱保护不彻底及对 DNA 点阵密度的制约。目前可以制备探针位点为 8μm 的方形区域的基因芯片，预计可以将每种探针位点缩小为 1μm，这样就可以制备每平方厘米 10^{10} 种探针的高密度基因芯片。

2. 无掩膜原位芯片合成技术

NimbleGen 公司制作的高密度基因芯片是利用其专利的无掩膜合成技术（maskless array synthesizer，MAS）技术原位合成 DNA。MAS 系统利用的是被称为照相沉积化学（photo deposition chemistry）的技术，其核心原理是数字微镜晶片设备（digital micromirror devices，DMD）。利用固态的微小氧化铝镜片阵列反射由多达 786000 个单一像素光线构成的光束。数字微镜晶片在本质上相当于一种光学掩膜，代替了常规原位合成法中的铬掩膜。

DMD 晶片包含了多达上百万个正方形微镜片，每个镜片都由铰链固定在同一平面上。每个镜片的尺寸仅为人头发直径的 1/5，对应于投射图像中的一个像素（图 2-18）。

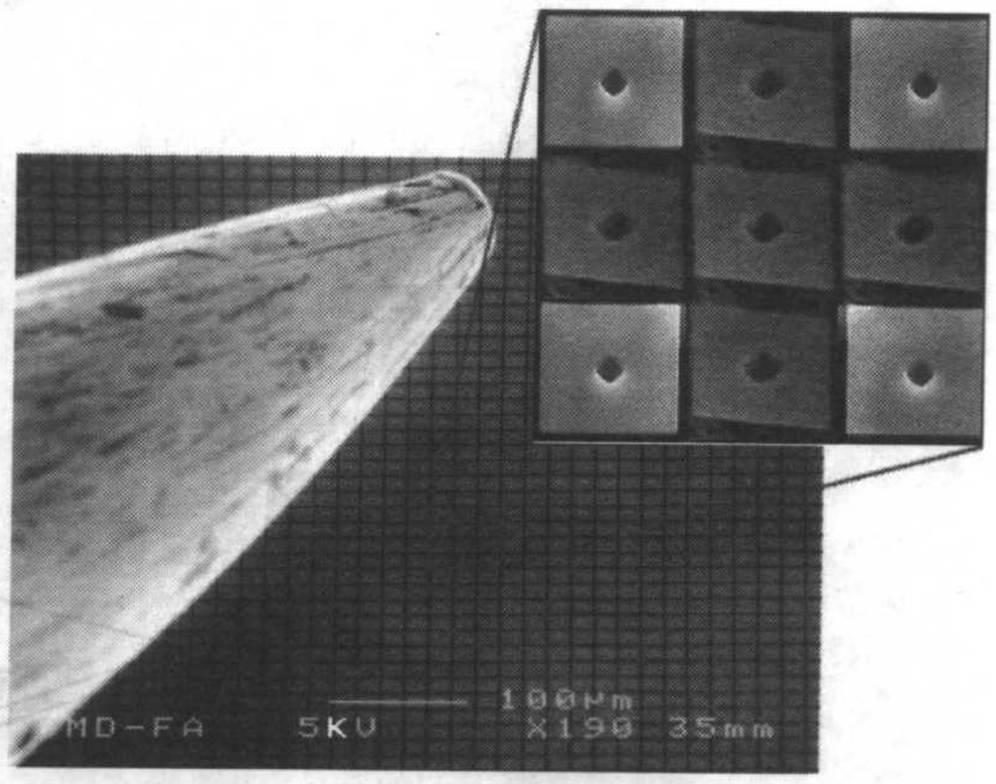


图 2-18　数字微镜晶片及局部放大图

当 DNA 晶片和投射仪偶联时，微镜片可以将全数字的图像反射到一个屏幕或其他表面上。DMD 微镜片由铰链固定，可以朝向（ON）或背向（OFF）系统中的光源，在投射图像上形成一个亮的或暗的像素。实际上就起到了掩膜的作用，在亮的地方相当于透光的地方，光敏基团将去保护，发生核酸结合反应；暗的地方就是不透光的地方，光敏基团保持不变（图 2-19）。数字微镜晶片系统的镜片直径非常小，并且克服了光线通过普通掩膜小孔时发生衍射的可能，从而大大提高了芯片的密度。

二、纳米结构的基因芯片制备

原位合成的基因芯片比点样芯片提高探针密度约一个数量级（一张原位芯片约包含 10 万～40 万个点），这样高密度的芯片基本满足了表达谱基因芯片的设计要求。以人的基因组为例，预测全部有大约 3 万～5 万个基因，包含“错配”及“校对”探针的冗余设计，原位合成的基因芯片基本可以实现全基因组的基因表达谱的研究，剩下来关键问题似乎就是分析与解读。然而随着单核苷酸多态性（SNP）领域研究的深入，对基因芯片密度的要求又提高了大约 1000 倍，满足全基因组 SNP 的研究需要微矩阵技术的一场革新。

原子力显微镜（AFM）是 20 世纪 80 年代初问世的扫描探针显微镜（scanning probe microscope，SPM）的一种。这种显微镜的放大倍数远远超过以往的任何显微镜，原子力显微镜的放大倍数能高达 10 亿倍，比电子显微镜分辨率高 1000 倍，可以直接观察物质的分子和原子，是人类探索微观世界的理想工具。AFM 以下的特点使之在生物学中得以迅速应用及发展：①AFM 的样品制备简单，无需对样品进行特殊处理，因此，其破坏性较其他生物学常用技术（如电子显微镜）要小得多；②AFM 能在多种环境（包括空气、液体和真空）

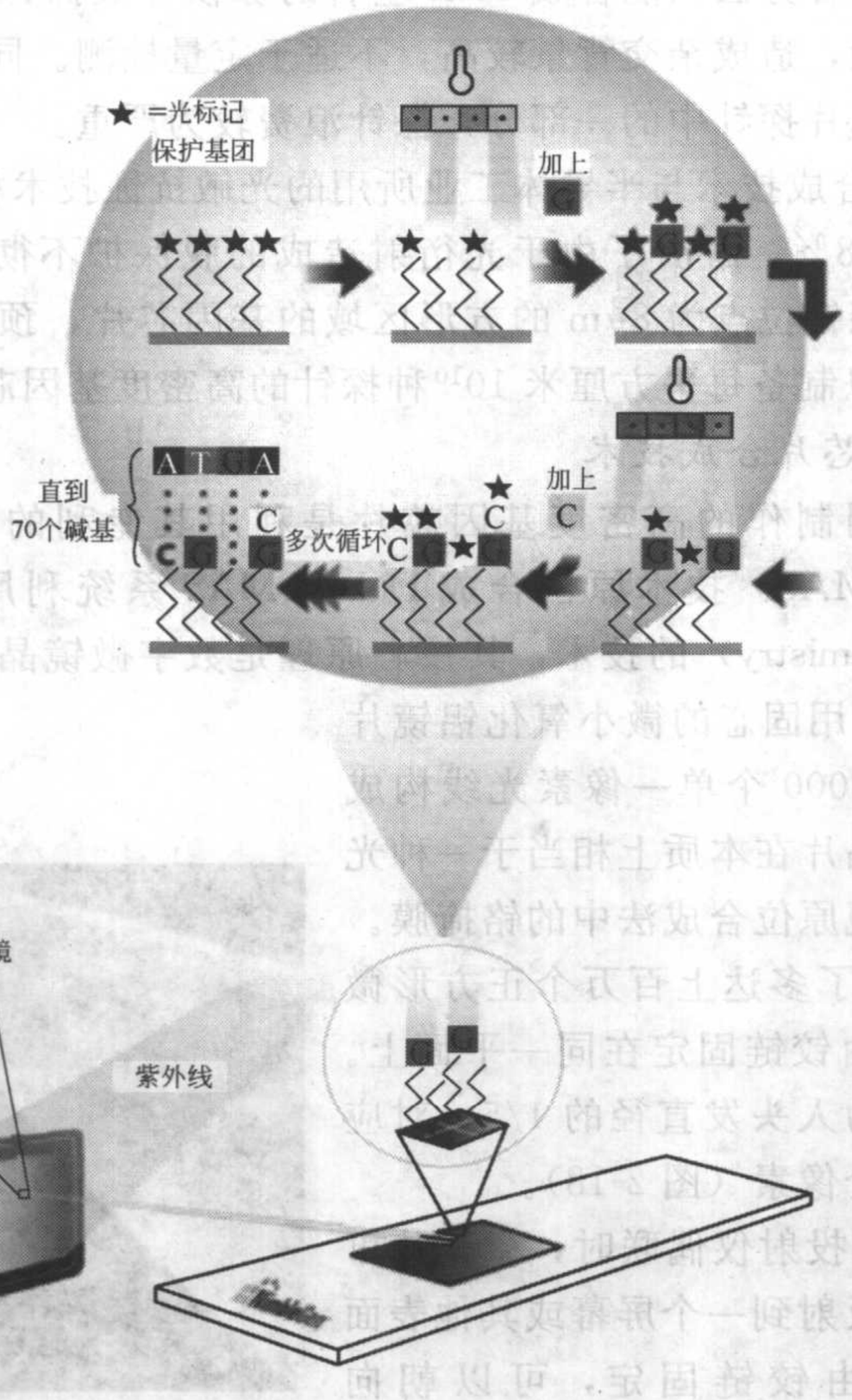


图 2-19　无掩膜原位合成原理图

中运作，生物分子可在生理条件下直接成像，也可对活细胞进行实时动态观察；③AFM 能提供生物分子和生物表面的分子/亚分子高分辨率的三维图像；④AFM 能以纳米尺度的分辨率观察局部的电荷密度和物理特性，测量分子间（如受体和配体）的相互作用力；⑤AFM能对单个生物分子进行操纵。正是基于此点，人们已在实验室内制作出类似基因芯片的纳米微矩阵。

Dip-Pen 纳米刻写术（Dip-Pen nalithography™，简称 DPN）由美国西北大学的 Mirkin 教授研究小组开发。DPN 是一种简单方便的从 AFM 针尖到基底传输分子的方法，其分辨率可与电子束刻蚀等方法相比，对纳米器件的功能化更为有用。其基本原理图如图 2-20 所示。用 AFM 的针尖作“笔”，固态基底作“纸”，能与基底有化学作用力的分子作“墨水”，分子通过凝结在针尖与基底间水滴的毛细作用直接“书写”到基底表面。

作为一项“直接书写”技术，DPN 特别适合在固体表面排列生物分子，并且在分辨率方面超越任何点样技术和照相印刷技术。因此在已经普及了的基于“芯片”概念的各式检测系统中，DPN 有望在 DNA、蛋白质以及生物小分子的芯片领域发挥出威力。首先，DPN 技术能够将目前的芯片密度提高 1000～100000 倍；如此更高密度意味着更大范围的检测对象和更小样本体积和更少的耗时，将来有可能在一张芯片上检测人基因组的所有 SNP。其次，AFM 可以实时动态地扫描杂交结果，可以有效提高检测的通量。

DPN 除了可以制备核苷酸纳米微矩阵、蛋白质纳米微矩阵外，还可以制作单细胞甚至单病毒微矩阵，弥补高密度蛋白质芯片和细胞芯片的不足。

激光系统扫描仪的分辨率主要是由激光的聚焦直径决定的，分辨率一般为 5～10μm。CCD 系统扫描仪的分辨率主要取决于 CCD 像素的数目，一次成像型 CCD 扫描仪的分辨率一般大于或等于 20μm，多次扫描拼接成像型则可达到 2.5μm。

除了以上提及的参数外，扫描速度、扫描视野、扫描图像定位的准确程度等指标也是设计和制造微阵列扫描仪所需考虑的技术指标。微阵列扫描仪之间的比较是一个复杂的过程，不应只对单个性能参数进行比较，通常需要对扫描样品进行综合测试。

基因芯片扫描仪的作用是得到采集杂交后的芯片荧光信号，形成芯片图像，并检测图像上每个点的荧光信号强度值。只有在充分了解扫描仪的原理和技术参数的基础上，才能对芯片进行正确的扫描。概括地讲，基因芯片扫描仪的操作过程如下：

① 将杂交、洗脱后的芯片放置于载片台上；

② 设置扫描参数，如激发光的光强，光探测器的强度（如 PMT）；

③ 确认其已置于初始位置，一般采用粗扫（一般采用 50μm 或 100μm 的分辨率）确定矩阵区域；

④ 调整扫描参数和扫描区域进行精扫（一般采用 5μm、10μm 或 20μm 的分辨率，根据扫描仪的精度以及用户的需求调节），获得芯片图谱；

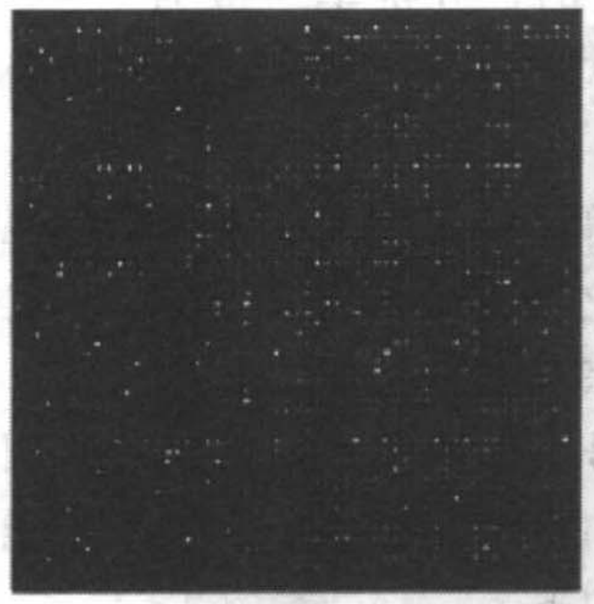
⑤ 软件分析得到的每个基因的信号值，再进一步分析其表达特性。

值得注意的是，荧光染料分子在激发光的照射下，其产生荧光的强度随着时间的延长逐渐变弱消失。这种光漂白现象几乎存在于所有的荧光染料中，光漂白的程度随光照强度的增加和照射时间的延长而增强。在激光系统扫描仪中，光强较大，因此光漂白的作用较强。CCD 系统虽然激发光较弱，但 CCD 扫描时通常采用长时间曝光方式，因此也同样存在光漂白现象。因此在扫描芯片时应尽量减少扫描次数。

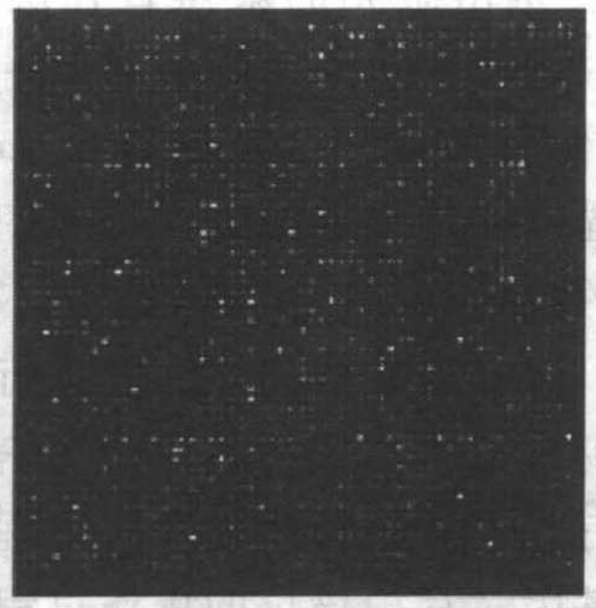
第二节　基因芯片图像的处理

基因芯片实验后的所有原始信息都贮存在芯片图像中，芯片图像通常都是 16 位的 TIFF、JPEG、RAW 等格式的图像（也有些扫描仪能得到更高的 24 位的图像），显现的是灰度值，每个像素的灰度值在 0～65535 的范围内，每个灰度值都反映了图像所对应芯片位置的荧光分子相对强度信息。在芯片图像中，每个点的像素强度总和就对应着相应核酸序列的杂交量。芯片图像处理的目的是定位每个点，将每个点所对应的不同形状和强度的杂交量化，并将得到的一系列数值（例如强度和通道比率等）形成表格。以 cDNA 芯片为例，cDNA 芯片是两张 16 位的 TIFF 灰度图像，每个通道代表一个 TIFF 文件。cDNA 芯片图像处理的主要目的是量化芯片上每个点在两个通道中的前景（foreground）和背景（background）。背景值用来校正可能带有局部差异的前景值，校正后的各个点的信号值就是后面用来进一步分析的主要数据值。在比较完善成熟的方法中，图像的处理还包括评估每个点的质量，计算每个点相应数据的可信程度，标记出数据不可信的点，并对在芯片的制造和杂交过程可能出现的问题提出预警。基因芯片包括很多种，但芯片图像处理的原理是一样的，cDNA 芯片是双荧光芯片，具有两个荧光通道，更具有代表性，下面将以 cDNA 芯片图像为例来讲述芯片图像处理的原理。

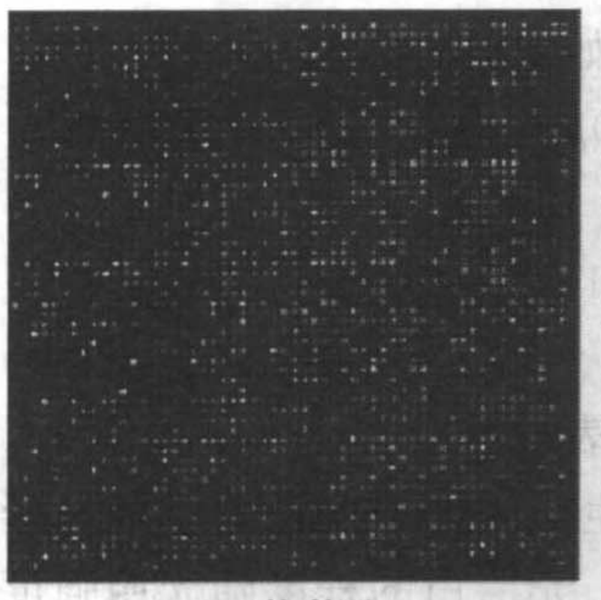
在 cDNA 芯片中，由于是双色荧光标记，一张片子扫描后得到两个图像文件，对应不同的荧光图像，假如都用黑白的灰度图像表示不易区分。计算机处理时，将两张图像叠加，同时对两个图像进行处理，读出每种荧光的强度。为了更直观，通常将不同的荧光用相应的颜色（即伪彩色）来表示，以示区分，并使图像看起来更悦目（图 5-4，彩图见插页）。例



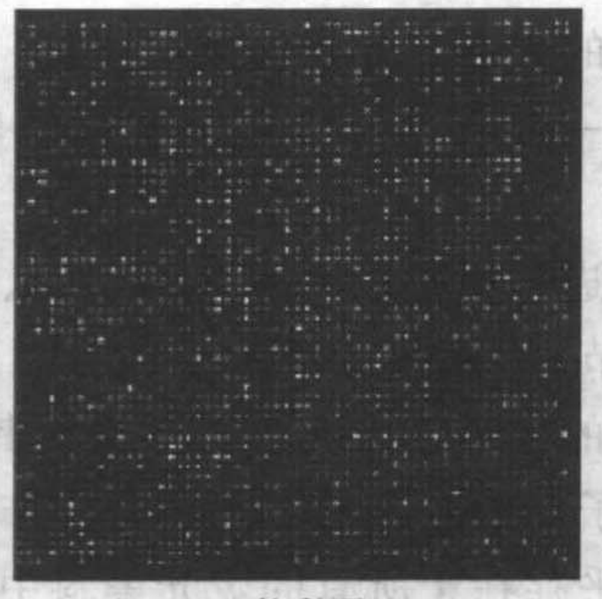
Cy3灰度图



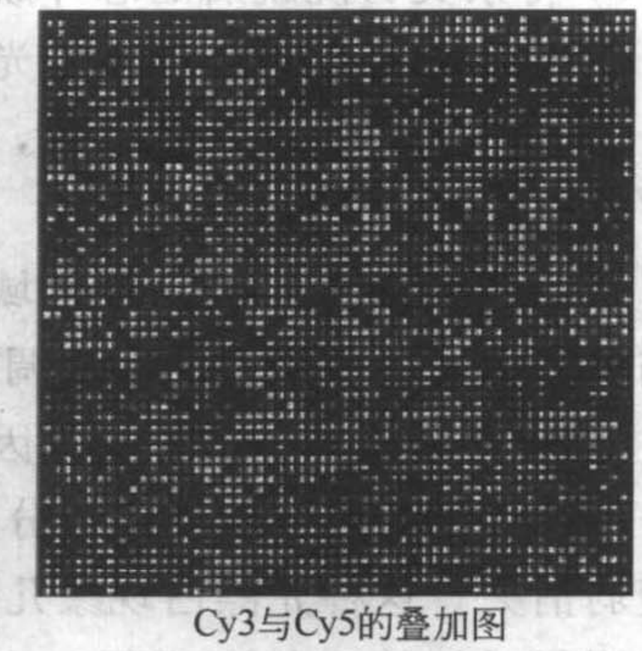
Cy5灰度图



Cy3伪彩图



Cy5伪彩图



Cy3与Cy5的叠加图

图 5-4　cDNA 芯片的双色荧光图像

如，Cy3 用绿色图像表示，Cy5 用红色图像表示。所有的视图处理工具均不更改原始图像文件，而只是影响图像的视觉效果，否则芯片的数据信息便会遭到破坏。有时候为了便于扫描完毕就能大致判断芯片中是否有上调或下调的差异基因，有些扫描或分析软件中还提供叠加图的功能。它是将同一张芯片所对应的两种不同杂交样本扫描所得图谱分别转变成绿色和红色的图谱。假如对照样本（通常用 Cy3 标记）被转变为绿色，而实验样本（通常用 Cy5 标记）被转变为红色。当两张图谱叠加在一起时，每个基因点的最终色彩由两种颜色的比率而决定：如果最终该点的颜色为红色，即 Cy5 的信号比 Cy3 强，说明该基因在实验样本中的表达量比对照样本中的表达量高，该基因为上调基因；如果最终基因点的颜色为绿色，即 Cy5 的信号比 Cy3 弱，则该基因在实验样本中的表达量比对照样本中的表达量低，该基因为下调基因；如果最终基因点的颜色为黄色，则该基因在两种样本中的表达量相同，该基因没有显著变化。

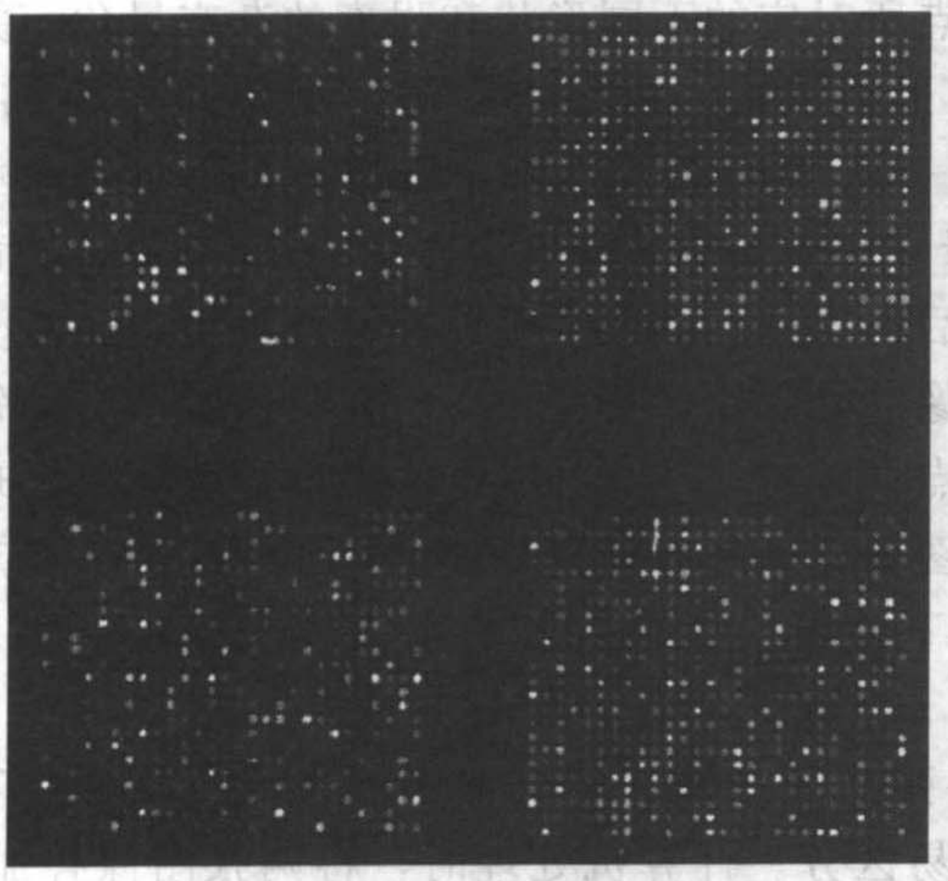


图 5-5　cDNA 芯片的典型图像
（图像中有 4 个子格）

芯片图像上的点像格子一样排列。有些芯片的图像还包括几个部分，每个部分含有相同行和相同列的点的子格，并且每个子格之间的距离大概相等，所有的子格合在一起就是一整张芯片。图 5-5（彩图见插页）是一张 cDNA 芯片图像。理想情况下的芯片图像是容易处理的。将设定大小、距离和一定个数的圆放到图像上，

计算是通过邻近区域背景值的加权和得到，权重与距给定区域的距离平方的倒数呈正比。如，区域 k 对于芯片坐标为（x,y）的格子的权重为：

$$w_k(x,y)=\frac{1}{d_k^2(x,y)+c}$$

式中，c 为平滑常数，用于确保分母不会变得太小而接近于 0。对于每个格子，假设其坐标为（x，y），其加权背景为：

$$b(x,y)=\frac{1}{\sum_{k=1}^{K}w_k(x,y)}\sum_{k=1}^{K}w_k(x,y)b_{Z_k}$$

式中，b_{Z_k} 为区域 Z_k 的背景值。

进行背景校正时，通常使用格子的强度减去背景的强度，但常有格子的强度低于根据以上公式计算得到的背景值，使得到的背景校正值为负值，这将给后续的数据处理带来问题，如无法对负值进行对数的转换等。为解决此问题，可以按照相同的方法计算局部噪声值：

$$n(x,y)=\frac{1}{\sum_{k=1}^{K}w_k(x,y)}\sum_{k=1}^{K}w_k(x,y)n_{Z_k}$$

式中，n_{Z_k} 为区域 Z_k 的局部噪声值，是该区域格子最低 2%背景值的标准差。

单个格子的强度可以调整为：

$$I'(x,y)=\max[I(x,y),0.5]$$

式中，$I(x,y)$ 为格子（x，y）的原始强度。这样背景校正强度值可以表示为：

$$I_c(x,y)=\max[I'(x,y)-b(x,y),\mathrm{NF}n(x,y)]$$

式中，NF 为选择的一个全局背景变异的比例，通常选择 0.5。

（二）信号的计算

对于 cDNA 微阵列，预处理的目标是修正基因特异性强度值，使得到的值能够准确反映溶液中的转录量。在 Affymetrix 阵列中由于使用多个探针对来测量基因的表达水平，需要把这些值合并为一个值。非配对序列上的杂交量被认为代表了非特异性杂交，应从完全配对的强度值中减去非配对的强度值，对应于一个基因有多对探针，计算一个基因所有探针对差值的均数，就得到了平均差（average difference，AD）。AD 越大，基因的表达水平越高。由 Affymetrix 公司提供的软件还计算另外一个值 call，它是关于基因的三重判断：无（A）、边际存在（M）和存在（P）。

1. 理想的非配对

在 Affymetrix 中包含 MM 探针的原因是提供一个估计非特异性杂交和其他影响 PM 的偏离信号的值。有以下三种情况。

① 若 MM 值小于 PM 值，MM 值被认为是背景估计值，理想的非配对（ideal mismatch，IM）等于 MM：

$$\mathrm{IM}_{i,j}=\mathrm{MM}_{i,j}$$

② 若对于某个探针 MM 大于 PM，它就不能用于背景的估计。此时，IM 的估计使用同一基因的其他探针的 PM 和 MM 的差值进行：

$$\mathrm{SB}_i=T_{bi}[\log_2(\mathrm{PM}_{i,j})-\log_2(\mathrm{MM}_{i,j})],\ j=1,\cdots,n$$

式中，T_{bi} 相当于两差值的权重，计算方法可查阅相关文献。如果 SB_i 大于界值 τ_c（contrast tau），IM 的计算公式为：

$$\mathrm{IM}_{i,j}=\frac{\mathrm{PM}_{i,j}}{2^{\mathrm{SB}_i}}$$

③ 如果 $SB_i < \tau_c$，IM 的估计方法为：

$$IM_{i,j} = \frac{PM_{i,j}}{\frac{\tau_c}{2^{1+\frac{\tau_c - SB_i}{\tau_s}}}}$$

式中，τ_s 为另外一个界值（scale tau）。通常两个的界值为 $\tau_c = 0.03$，$\tau_s = 10$。

2. 探针值

当每个探针的 IM 计算得到后，探针值就可以计算得到：

$$V_{i,j} = PM_{i,j} - IM_{i,j}$$

对数转换后的探针值（probe value，PV）为：

$$PV_{i,j} = \log_2(V_{i,j}),\ j = 1, \cdots, n$$

由一个基因的多个探针对来计算反映其转录量的信号对数值（signal log value，SLV）：

$$SLV = T_{bi}(\log_2 V_{i,j}, \cdots, \log_2 V_{i,n_i})$$

式中，T_{bi} 为单步 Tukey 双权重估计值（one-step Tukey's biweight estimate），相当于权重。对于给定的探针配对，PM 和 IM 可能非常接近，两差值的对数是一个非常大的负数，将严重影响加权均数的计算。为了避免此问题，通常对此差值设置一个低限界值 $\delta = 2^{-20}$，任何 $\log_2(V_{i,j})$ 低于此 δ 值将设置为此低限界值。

3. 标准化探针值

计算截尾均数可以有效提高所得数值的可信性。截尾均数的计算是首先去除极大和极小的部分数值，然后再计算剩余数值的均数。前面计算得到的信号值为对数值，因此截尾均数的计算应首先进行反对数的变换。用 TM 表示截尾均数：

$$TM(2^{SLV}, 0.02, 0.98)$$

式中，参数 0.02 和 0.98 定义分布的双侧尾部被截取数值的比例（极小值和极大值分别截取 2%）。最终，由软件输出每个基因的 RV 值，其计算方法为：

$$RV_i = NF \cdot SF \cdot 2^{SLV}$$

式中，NF 为归一化因子（normalization factor）；SF 为尺度因子（scaling factor）。

对于一个目标信号 TS，其尺度因子的计算为

$$SF = \frac{TS}{TM(2^{SLV}, 0.02, 0.98)}$$

计算归一化因子 NF 的目的是使两个阵列可以直接进行比较，如参照和实验之间进行比较。归一化因子的计算可以通过两个截尾均数的简单比值得到。

$$NF = \frac{TM(2^{SLV_{reference}}, 0.02, 0.98)}{TM(2^{SLV_{experiment}}, 0.02, 0.98)}$$

因为比较分析是在探针对水平上进行的，单个的探针对的值也要进行相应的归一化和标度因子的转换。标准化后的探针值 SPV 为

$$SPV_{i,j} = PV_{i,j} + \log_2(NF \cdot SF)$$

因此，Affymetrix 阵列数据的预处理总结为以下几步：

① 通过对邻近区域背景的加权平均对每个格子的背景强度进行背景的校正；

② 计算理想的非配对值 IM，然后从 PM 强度中减去 IM；

③ 校正后的 PM 值进行对数转换；

④ 经对数转换后的值进行稳健性均数估计，然后进行反对数转换；

⑤ 对信号值通过截尾均数进行标准化。

检方法。

二、Affymetrix 的寡核苷酸芯片质控体系及其产品质量评估

（一）Affymetrix 芯片的质控体系

Affymetrix 的寡核苷酸芯片的质控包括三个环节：设计、合成和终产品的信号强度。

为此，他们借用了两个相关的成熟技术的质控方法：寡核苷酸合成和半导体工业的技术标准。所用的其他原材料的标准也与寡核苷酸合成相同，不同之处在于是用光而不是用酸进行去保护来合成探针。与半导体工业相比，他们借鉴了已有的技术体系并根据自身的特点对基因芯片生产过程进行了整合，包括底物准备、光刻及包装等。

每个基因的探针序列经过严格的设计确定下来，根据这些序列来设计光刻掩膜。从设计到完成合成的整个原位合成过程，都采用自动化软件系统控制操作和质量，如通过自动设计检测过程来保证掩膜的设计无误，过程控制软件保证在正确的时间将正确的试剂加到一定的位置上，进行正常的循环，以保证在确定位置上（*X* 轴、*Y* 轴）有正确的探针序列。

为了保证每个步骤的准确性，必须正确识别每个掩膜，他们采用了光特征识别（optical character recognition，OCR）来确保特定时间采用正确的掩膜和晶片。

另外，他们在芯片上设计了特殊的对照探针，由于不可能对每个探针进行质量检测，他们只需对对照探针进行质检就能很好地显示整张芯片的质量。图 8-11 是一个例子，上面包含长度为 4 个碱基的 4 条探针，可以用 8 个独立的化学循环和 8 个掩膜完成这 4 条探针的合成，探针 1 的合成通过循环 1、2、4 和 6 实现，探针 2 的合成通过循环 3、5、7 和 8 实现，因为探针 1 和探针 2 的合成覆盖了所有的循环（8 个），因此要确定这 4 条探针的合成精确性，只需分析探针 1 和探针 2 的质量就能保证其他两条探针的质量。同样地，探针 3 和探针 4 的合成也覆盖了所有的循环，也能用探针 3 和探针 4 来进行质检。运用这一原理，两套探针中的任何一套都可以用来作为质控探针分析合成过程的所有循环，而不必对所有探针分别进行质检。实际的芯片合成过程远比以上的例子复杂，但可利用相同的原理选择一套对照探针来监控化学误差和光刻错误。

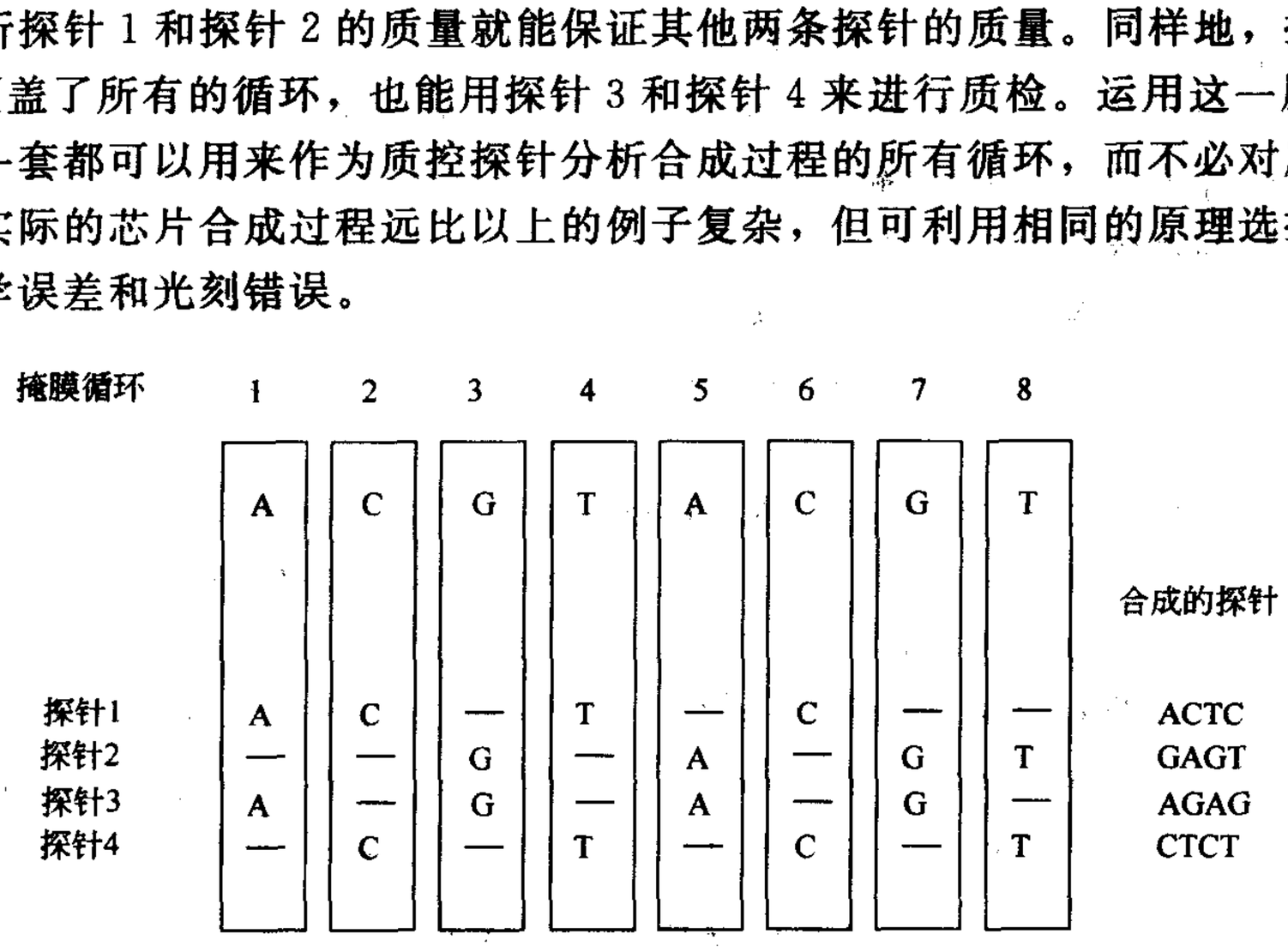


图 8-11　原位合成探针的过程

信号的判断是通过合成完成后的杂交实验来实现的，在芯片上设计了另外一组探针进行信号的质量监控，用来检查芯片是否能产生足够的信号，例如在人 U133（HG-U133）芯片中设计了一组对照探针，其中包括 *bioB* 基因、*bioC* 基因、*bioD* 基因和 *cre* 基因的探针，基因 *bioB*、基因 *bioC*、基因 *bioD* 是大肠杆菌中生物素合成途径中的基因，*cre* 基因是来自 P1 噬菌体的重组酶基因。这组探针用来检验芯片是否能产生足够强的信号，从而保证实验的成

功率。

通过以上的质控体系，就能保证序列合成的正确性和合成效率，如果合成的序列发生错误或合成效率低，就不能全部通过以上质控体系的检验。

（二）Affymetrix 芯片的质量评估

对于芯片的性能情况，主要从两个方面进行评估：灵敏度和重复性。

1. 输出结果中的基因分类

为了检测芯片的灵敏度和重复性，Affy 芯片采用了特有的方式来进行量化。由于 Affymetrix 芯片为单通道芯片，因此根据其用途，数据的分析分为以下两种方式。①单张芯片的结果可以反映基因表达的绝对丰度。根据信号的强弱，在输出的基因报告中将基因记录成 P、A、M：信号值大于背景的基因，即有效检测基因用 P（present）表示；那些弱信号的无效点用 A（absent）表示；另外，还有少量的基因介于 P、A 之间的临界点，用 M（marginal）表示。②基因差异表达比较分析。与 cDNA 芯片类似，计算基因表达改变的倍数，当两种样本分别与两张芯片杂交时，两张芯片信号强度的比值即反映了表达改变的倍数。据此，Affymetrix 也发展了一种简单的判别方式：在输出的报告中将基因记录成上调（increase）、下调（decrease）或没有差异（no）三类。

2. 灵敏度测定

灵敏度的检测是通过预先标记的转录物（靶 DNA）的杂交来实现的，可以检测的最低转录物浓度为 1.5pmol/L，该浓度相当于 100000 个 mRNA 中的一个 mRNA 分子，或一个细胞中的 3.5 个拷贝的 mRNA 分子，即 Affymetrix 芯片检测的灵敏度为 1/100000 的 RNA 丰度。基因差异表达比较分析中的灵敏度测定是通过检测 3pmol/L 和 1.5pmol/L 的预先标记转录物（两倍变化）检出能力来确定的。

对于单张芯片的结果，检测的灵敏度标准（阈值）设置在对于 1.5pmol/L 转录物的检出阳性率大于或等于 70%，即当输入 1.5pmol/L 转录物时，作为 P（present）记录的次数占总的检测次数的百分率大于或等于 70%。

在基因差异表达比较分析实验中，灵敏度标准是一张芯片中输入 1.5pmol/L 标记转录物，另一张芯片输入 3pmol/L 的标记转录物，其比值有 80%及以上的概率被判断为上调（increase）。

例如，采用 HG-U133 芯片进行灵敏度评估，样本选用了两种人的 RNA 转录物，其中分别加入 4 种浓度为 1.5pmol/L 或 3.0pmol/L 的大肠杆菌对照转录物，每组转录物重复 3 批，每批做 3 张芯片，共 9 张芯片。采用 MAS 5.0 软件，灵敏度的计算基于 1.5pmol/L 或 3.0pmol/L 的大肠杆菌对照转录物所产生的信号，通过分别计算 1.5pmol/L 大肠杆菌对照转录物为 present 记录的百分数及 3.0pmol/L 比 1.5pmol/L 转录物记录为上调（increase）的百分数来确定。平均值的计算是基于 9 次重复实验获得。结果如表 8-4 所示。

表 8-4　灵敏度测定的结果

项　　目	HG-U133A	HG-U133B	项　　目	HG-U133A	HG-U133B
平均的 P 记录百分数	79%	77%	平均的上调记录百分数	94%	87%
P 记录的标准差	8%	3%	上调记录的标准差	3%	7%

3. 重复性评估

对于重复性的评估，Affymetrix 采用了两个标准：假阳性率（false change）和重现率

二、费歇线性判别分析

判别分析（discriminat analysis）的任务是根据已掌握的一批分类明确的样本（基因），建立较好的判别函数（discriminat function），使产生错判的对象最少，进而对给定的一个新样本（基因），判断它来自哪个总体。

根据所采用的判别准则不同，经典的判别分析方法有费歇判别（Fisher discriminat）和贝叶斯判别（Bayes discriminat）。

Fisher 线性判别分析（FLDA）又称典则判别（canonical discriminant），其基本思想是投影，将多维问题简化为一维问题来处理。选择一个适当的投影轴使所有的样本点都投影到这个轴上得到一个投影值。对该投影轴的方向要求是：使每一类内的投影值所形成的类内离差尽可能小，而不同类间的投影值所形成的类间离差尽可能大。

假定已知 $1\times G$ 特征向量 $x=(x_1,\cdots,x_G)$，称为判别指标（或变量）。Fisher 线性判别分析就是要找出一个线性组合 xa，使类间与类内平方和的比值较大。对于一个 $G\times n$ 的训练集数据矩阵 X，矩阵 X 中行的线性组合 $a'X$ 的类间与类内平方和的比值可以表示为 $a'Ba/a'Wa$，其中 B 和 W 分别表示类间和类内平方和以及交叉乘积的 $G\times G$ 维矩阵。$a'Ba/a'Wa$ 的极端值可以由 $W^{-1}B$ 的特征值和特征向量得到。矩阵 $W^{-1}B$ 具有最多 $s=\min(K-1,G)$ 个非零的特征值，$\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_s$，分别对应于线性独立的特征向量 $v_1,v_2,\cdots,v_s$。判别变量（discriminant variables）定义为 xv_l，$l=1,\cdots,s$，且 $a=v_1$ 时可以使 $a'Ba/a'Wa$ 最大化。

对于特征向量 $x=(x_1,\cdots,x_G)$，令 $d_k^2(x)=\sum_{l=1}^{s}[(x-\bar{x}_k)v_l]^2$ 表示其欧式距离的平方，根据判别变量，针对训练集 L 从类 k 中 $1\times G$ 的向量样本均数 $\bar{x}_k=(\bar{x}_{k1},\cdots,\bar{x}_{kG})$ 得到[这里，$\bar{x}_{kg}=\sum_i I(y_i=k)x_{gi}/n_k$]。对于新的特征向量 x 所预测的类别就是在判别变量中，其均数向量 $\bar{x}_k$ 与 x 最接近的那一类，即 $C(x;L)=\arg\min_k d_k(x)$。

因此，FLDA 中两个主要的步骤是：

① 特征选取或者降维，以此识别出 s 个判别变量；

② 分类，根据变量在降维之后的空间中离类均数的距离对样本进行分类。

FLDA 是一种非参数方法，但也出现于参数设置中。对于 $K=2$ 的情况，FLDA 可以得到与具有相同的协方差矩阵的多元高斯分类密度的样本的极大似然判别法则相同的分类器。

三、线性判别和二次判别分析

当每一类中的特征向量都服从正态分布时，特征向量 x 上的线性和二次判别法则就是贝叶斯法则或者极大似然（ML）判别法则。令 $X\mid Y=k\sim N(\mu_k,\sum_k)$，其中 $\mu_k=(\mu_{k1},\cdots,\mu_{kG})$，$\sum_k$ 分别表示期望值和类 k 中特征向量的 $G\times G$ 协方差矩阵。贝叶斯法则为：

$$C(x)=\arg\min_k\{(x-\mu_k)\sum_k^{-1}(x-\mu_k)'+\lg|\sum_k|-2\lg\pi_k\} \tag{10-6}$$

一般来说，这是一个二次判别分析（quadratic discriminant analysis，QDA）。判别法则中主要的量是 $(x-\mu_k)\sum_k^{-1}(x-\mu_k)'$，即样本 x 离类 k 的均数向量 μ_k 的马氏距离的平方。下面将对先验概率相同的情况进行讨论，即在不同的 k 中 π_k 不变。

1. 线性判别分析（linear discriminant analysis，LDA）

当类密度具有相同的协方差矩阵时，$\sum_k=\sum$，判别法则基于马氏距离的平方，与 x

成线性关系，定义为：

$$C(x) = \arg\min_k (x-\mu_k) \sum^{-1} (x-\mu_k)' = \arg\min_k (\mu_k \sum^{-1} \mu_k' - 2x \sum^{-1} \mu_k')$$

2. 对角线二次判别分析（diagonal quadratic discriminant analysis，DQDA）

当类密度具有对角协方差矩阵时，$\Delta_k = diag(\sigma_{k1}^2, \cdots, \sigma_{kG}^2)$，判别法则由每个特征向量的附加二次贡献给出，即 $C(x) = \arg\min_k \sum_{g=1}^{G} \left\{ \frac{(x_g - \mu_{kg})^2}{\sigma_{kg}^2} + \lg\sigma_{kg}^2 \right\}$。

3. 对角线线性判别分析（diagonal linear discriminant analysis，DLDA）

当类密度具有相同的对角线协方差矩阵时，$\Delta = diag(\sigma_1^2, \cdots, \sigma_G^2)$，判别法则是线性的，即 $C(x) = \arg\min_k \sum_{g=1}^{G} \frac{(x_g - \mu_{kg})^2}{\sigma_g^2}$。

4. 最近质心（nearest centroid）

在最简单的情况下，假定 $\sum_k = I_G$ 为 $G \times G$ 单位矩阵。样本根据其距离类均数向量的欧式距离进行分类。

DLDA 和 DQDA 对应于用于高斯分类条件密度的朴素贝叶斯法则。正如任何明确地估计贝叶斯法则的分类器，类的后验概率可用于评价对单个样本的预测可信度。应该注意到，虽然 LDA 和 QDA 同 ML 或用于高斯分类密度的贝叶斯法则一样，是在参数部分进行的介绍，但是这些方法有着更为广泛的应用。

对于样本贝叶斯或 ML 判别法则，总体的均数向量和协方差矩阵由训练集 L 估计得到，分别用样本均数向量和协方差矩阵进行估计：$\hat{\mu}_k = \bar{x}_k$，$\sum_k = S_k$。对于常数协方差矩阵的情况，协方差矩阵的估计是：$\hat{\Sigma} = S = \sum_k (n_k - 1) S_k / (n-k)$。

上述一般法则可以进行简单的调整以允许不相等的类别先验概率出现，即在不同的 k 中有不同的 π_k；先验概率的估计可以用样本的类别比例进行估计，$\hat{\pi}_k = n_k / n$。LDA 和 QDA 的折中方法是调整的判别分析（regularized discriminant analysis），它将 QDA 的协方差矩阵 S_k 缩减为 LDA 中的一般协方差矩阵 S：$S_k(\alpha) = \alpha S_k + (1-\alpha) S$，其中参数 α 可以以优化预测准确性为目的进行选取，例如使用交叉验证法（cross-validation）。

5. Golub 等提出的加权基因投票设计

作为最早应用于解决基因表达谱数据分类问题的方法，Golub 等人于 1999 年提出了用于解决二分类问题的加权基因投票法（weighted gene voting scheme）。这一方法被证实为 DLDA 或者朴素贝叶斯分类的另一种表达方式。对于两分类问题，$k=1$ 和 2，当且仅当 $\sum_{i=1}^{n} \frac{(x_i - \bar{x}_{2i})^2}{\hat{\sigma}_i^2} \geqslant \sum_{i=1}^{n} \frac{(x_i - \bar{x}_{1i})^2}{\hat{\sigma}_i^2}$，即 $\sum_{i=1}^{n} \frac{\bar{x}_{1i} - \bar{x}_{2i}}{\hat{\sigma}_i^2} \left(x_i - \frac{\bar{x}_{1i} + \bar{x}_{2i}}{2} \right) \geqslant 0$ 时，样本 DLDA 法则会将具有基因表达模式 $x = (x_1, \cdots, x_n)$ 的样本分配给类 1。

判别函数可以改写为 $\sum_i v_i$，其中 $v_i = a_i (x_i - b_i)$，$a_i = (\bar{x}_{1i} - \bar{x}_{2i}) / \hat{\sigma}_i^2$，且 $b_i = \bar{x}_{1i} + \bar{x}_{2i})/2$。这与 Golub 等人所用的函数几乎是相同的，除了 Golub 等人将 a_i 定义为 $a_i = (\bar{x}_{1i} - \bar{x}_{2i}) / (\hat{\sigma}_{1i} + \hat{\sigma}_{2i})$。$\hat{\sigma}_{1i} + \hat{\sigma}_{2i}$ 是差别标准误的估计值，使用标准差而不是 a_i 的方差。对于分类器得到的预测值，Golub 等人还定义了一个预测强度（prediction strength，PS），它代表“成功的余地（margin of victory）”：

（4）样本标记信息

① 被标记的样本总量。

② 使用的标记物（诸如 A-Cy3、G-Cy5、33P 等）。

③ 标记的详细过程。

（5）样本标记过程中的对照

① 芯片中和标记的探针杂交的靶点。

② 标记类型（例如寡核苷酸、质粒 DNA、转录本）。

③ 标记特征（例如浓度、期望比率）。

3. 杂交实验的信息

每一份杂交实验信息应该包括如下内容。

（1）标记样本的具体信息（哪个样本、标记信息）及使用的微阵列的信息（微阵列设计信息、微阵列编号）。

（2）杂交过程的详细信息

① 杂交体系的信息，包括溶液的浓度等。

② 阻遏剂信息。

③ 清洗流程。

④ 标记物使用量。

⑤ 杂交时间，体系浓度，温度。

⑥ 杂交仪器的描述等。

4. 杂交获得的实验数据及后续数据分析的信息。MIAME 规则从三个层面来考量微阵列杂交实验获得的数据：原始数据、图像分析结果和均一化及分析后的结果数据。

（1）原始数据

① 扫描相关数据、扫描使用的软件和硬件信息、扫描参数。

② 扫描获得的结果图片。

（2）图像分析和图像分析结果

① 图像分析软件信息，包括软件名称、版本号、软件所使用的参数。

② 对于每一个扫描获得的图片用以上软件分析后获得的分析结果。

（3）均一化和分析后数据（基因表达差异）

① 数据处理流程，包括均一化算法等。

② 分析后获得的基因表达差异数据。

各个研究院所在进行微阵列研究过程中，都建议参照以上的 MIAME 规则。这样，一个微阵列系统的基本实验信息都已经涵盖在其中了，研究人员可以根据自己研究的特殊性，增加一些具体的信息。

二、MIAME 表单

MIAME 表单（checklist）是一种工具，可使发表著作的相关芯片数据更容易共享。它其实是 MIAME 规则的一种精华版，目的是使相关文著的芯片数据达到 MIAME 的要求，使得数据更好地得到利用。它能引导和帮助文章作者、科学杂志期刊的审稿专家和编辑更好地相互协作、配合，使公布的芯片实验结果数据在格式上具有可获得性，能够清晰解释数据并且有验证结论的可能。在 MGED 的努力下，多家杂志已经建议或要求研究人员在提交基因芯片相关文章时，所提交的数据须符合 MIAME 规则。并且 MGED 建议将芯片数据递交到一个公共数据库以使数据能够被共享。这个表单的主要内容如下。

1. 实验设计

① 实验目的——最多一行，例如相关报道的标题。

② 实验的简单描述，例如相关报道的摘要。

③ 关键字，例如时间序列、细胞系比较。

④ 实验参数——实验参数或条件，例如时间、剂量、遗传差异。

⑤ 实验设计——样本、实验处理、RNA 抽提、标记以及芯片设计。

⑥ 质量控制步骤，例如重复芯片和荧光交换实验。

⑦ 与相关著作、网址或者数据库接收号的链接。

2. 样本的使用、抽提和标记

① 每个生物样本的来源（例如物种的名称、样本的提供者）和特征（例如性别、年龄、发育阶段、外力或者疾病阶段）。

② 生物样本的处理和使用流程（例如生长条件、处理条件、分离技术）。

③ 每个样本、每个实验的参数（例如在时间序列实验中的时间为 30min）。

④ 抽提和标记的技术流程（例如 RNA 或 DNA 抽提和纯化的流程）。

⑤ 外界标记。

3. 杂交进程和参数

杂交的流程和所用的条件，阻滞、清洗、染色等后期处理步骤。

4. 数据和标准

（1）数据

① 原始数据　即扫描图像信息提取的结果（可以提供图像）。数据应当与每个实验设计相对应（图像中每行每列的输出结果应当与芯片上的样点相对应）。

② 归一化和整理后的数据　即作者得到结论所用到的一些芯片上的量化数据（基因芯片实验中的表达谱数据矩阵，可能是由一些归一化后的对数比率值组成）。数据必须和单个实验对应（尤其是整理后的数据每行应当与一些生物注释如基因名称对应）。

（2）数据提取和处理流程

① 图像扫描的硬件和软件、处理流程和参数。

② 数据归一化、变换和滤取过程以及参数。

5. 芯片设计

（1）常用的芯片设计内容

包括平台种类（芯片是否玻片或是原位合成的芯片等），表面参数和固定参数，点样流程（自制芯片），商业芯片产品代号（名字或者参考序号）。

（2）芯片特征和序列描述　通常是形成一个表格，包括样点位置（行、列）和在这个位置的序列（可能同一序列在不同位置会出现）分子的明确信息，主要有以下几点。

① 序列作用——控制基因还是待测基因。

② 寡核苷酸碱基序列。

③ 长序列（例如 cDNA 序列或者 PCR 产物）来源，准备过程和数据库编号。

④ 每个序列合适的生物注释，例如基因名称（可能不同序列有相同的生物注释）。

三、MIAME 的目前与将来

目前，一些比较著名的基因芯片数据库系统，如斯坦福大学的基因芯片数据库（Stanford Microarray Database，SMD）、欧洲生物信息研究所（European Bioinformatics Institu-

很多生命形式中的极为复杂的基因转录调控网络，因为这些重复出现的网络形式是很多更复杂网络的构建基础。

Shen-Orr 等（2002）研究了大肠杆菌（*E. coli*）中基因转录调控网络中的网络基序 motif。他们利用统计方法识别了若干经常发生的网络基序 motif，这些是通过比较真实的大肠杆菌（*E. coli*）网络结构与具有同样特征的随机调控网络得到的。他们发现有三种重复出现的网络基序 motif：前馈环形环、单一输入模式（SIM）和密集重复调控模式（DOR）如图 13-3 所示。下面对这三种情形进行简单的介绍。

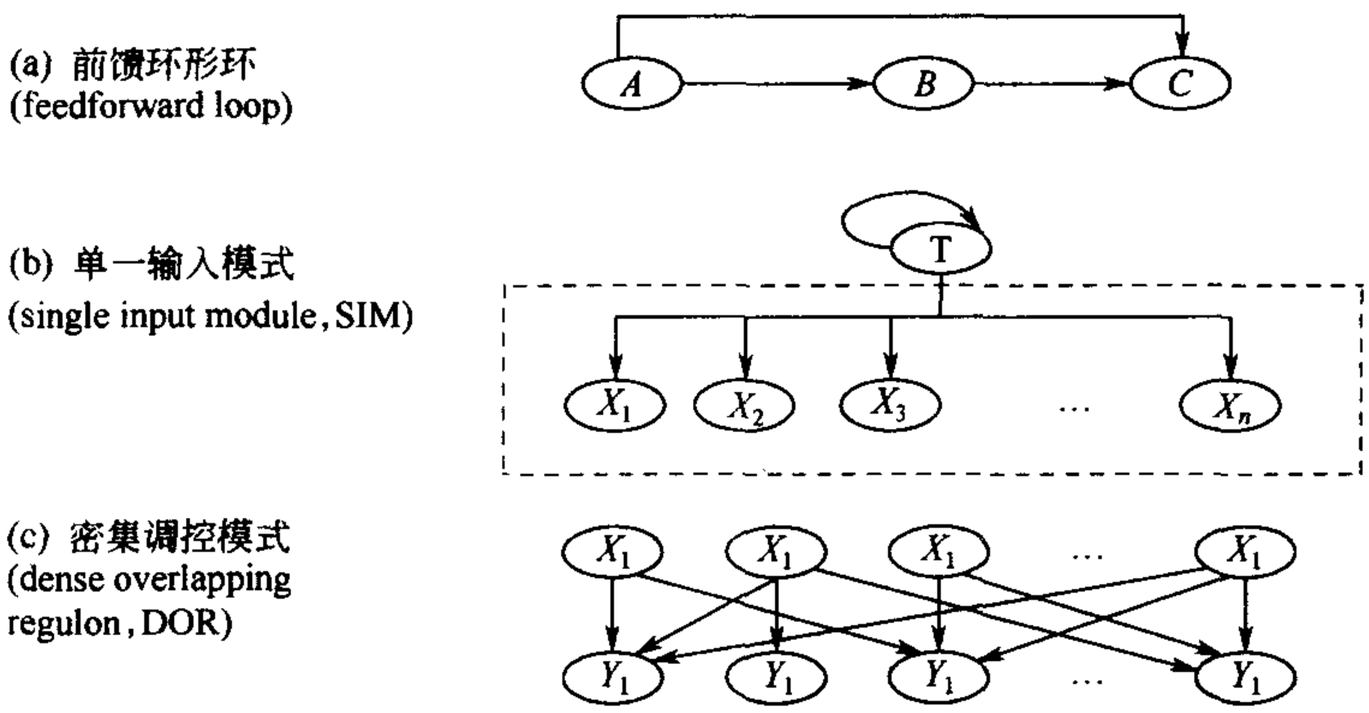


图 13-3　从 *E. coli* 转录调控网络中发现的三种主要的网络基序 motif（Sen-Orr 等，2002）

第一种是前馈环形环。在这种模式下，第一个转录因子 A 调控第二个转录因子 B，而基因 *A* 和基因 *B* 同时调控第三个基因 *C*［图 13-3（a）］。基因 *A* 也被称为总的调控因子，而基因 *B* 则是具体的调控因子，基因 *C* 是被调控基因。在这种情形下，还可以区分以下情况。一种前馈环形环是理性的，即总的调控因子对被调控基因的影响与具体的调控因子对被调控基因的影响是相似的（同是正面影响或是负面影响）。例如，如果基因 *A* 和基因 *B* 同时正面地调控基因 *C*，基因 *A* 同时正面地调控基因 *B*，这样一个网络基序 motif 就是理性的。与之相反，如果基因 *A* 负面地影响基因 *B*，那么这个网络基序 motif 就是非理性的。应该指出的是，这种调控模式非常普遍地存在于很多生命形式中。

第二种调控模式被称为单一输入模式（SIM）。简单而言，就是一系列的基因由同一个转录因子所调控，这种模式其实是一种一对多的关系［图 13-3（b）］。在这种情况下，所有的基因是由同一个转录因子所控制，而且控制的方式是一样的（都是正面影响或负面影响），除此以外没有其他的基因调控这些基因。

第三种调控模式是密集调控模式（DOR），它描述了一系列的调控基因与一系列的被调控基因之间的调控关系［图 13-3（c）］。实际上这是一种多个转录因子对多个被调控基因之间的多个对多个的关系。

这些基因转录调控网络基序 motif 的发现具有极为重要的意义，因为对其他复杂生命形式中的基因调控网络的研究和了解离不开这些基本的网络结构，因为人们普遍认为在细胞中的很多功能都是由众多的高度模式化的调控网络所控制的（Hartwell 等，1999）。尽管这些基本调控模式是从简单的生命形式中得来的，但是从理论上讲，它们也应当可以适用于其他复杂的生命形式，如动物及植物等。尤为重要的是，在今后用各种计算方法来创建各种调控网络时，可以用各种简单的网络调控模式来组建复杂的调控网络结构。

第三节　用高斯图形模型推导基因调控网络

尽管分子生物学在最近几年中有了很大的技术进步（如人类基因组测序的完成、大规模基因芯片技术的出现等），但从当前的众多数据中推导基因调控网络仍然面临众多的挑战。除了前面提到的很多小规模的实验方法外，研究人员们也开发出若干计算方法，试图从各种生命形式的数据中推导基因调控网络。绝大多数的方法和模型都是根据当前的大规模基因表达数据（基因芯片数据）来了解转录因子与它们所调控的基因同时出现的情况。当前的文献中有很多关于这些方法的介绍，这里将简单介绍比较有代表性的几种模型。第一种模型是一些相对简单的方法，如利用一个转录因子与它所调控基因的基因表达数据之间的相关系数来识别它们之间的调控关系，它还包括利用偏相关系数和高斯图形模型来识别这种关系的方法（Schafer、Strimmer 等，2004）。第二种模型建立在更系统的概率模型上，这些模型更多的是用贝叶斯网络模型从非时间序列数据中寻找基因调控网络。这方面的研究成果包括 Friedman 等（2000），Pe'er 等（2001）与 Segal 等（2003）发表的结果。第三种数量模型是针对当前的很多时间序列基因芯片数据而开发的，它包括动态贝叶斯网络模型（Friedman 等，2004；Friedman，2004）和事件模型（Kwon 等，2003）。这些方法用统计概率模型来描述转录因子与它们所调控基因的表达的时间次序，从而可以精确地推导它们之间的调控关系。第四种模型是根据基因扰动（gene perturbation）数据来推导基因调控关系（Yeung 等，2002；Tegner 等，2003）。这是一种很精确的方法，但是它的广泛应用目前还受制于基因扰动数据的数量。下面就对每一种模型作简单的介绍，同时让读者对这个领域的研究进度有一定的了解。首先向读者介绍高斯图形模型（GGM）在基因芯片数据方面的应用。

高斯图形模型（GGM）由 Shafer 和 Strimmer 于 2004 年提出，类似的思想已经出现在以前的文献中。这一方法主要被用来分析大规模基因芯片数据中基因相互作用的关系，它最初是用来描述多元数据中的相互依赖（dependency）结构，但同时也让人们对这些结构作更严格的统计验证。

基因芯片数据可以用 $N\times G$ 的矩阵来表示，这里，N 是实验中所用的 mRNA 样品的数量，G 代表这个实验中所用的基因的数量。值得注意的是，对当前的基因芯片实验来说 N 远远小于 G，这是 GGM 模型能够被用于这里的主要原因。在高斯图形理论中，这些数据可以看做是一个从多元正态分布中的随机取样。这个多元正态分布有均值向量 $\mu=(\mu_1, \mu_2, \cdots, \mu_G)^T$ 和协方差矩阵 $\sum=(\sigma_{ij})$，这里 $i, j=1, 2, \cdots, G$。在这些假设下，这些基因之间的相关矩阵 $P=(\rho_{ij})$，可以从以下公式得出：

$$\rho_{ij}=\sigma_{ij}/\sqrt{\sigma_{ii}\sigma_{jj}}$$

这里是表示任何两个基因之间的相关系数。两个基因之间的较高的相关系数表明它们之间可能有直接的相互作用或非直接的相互作用，或是它们被同一个基因所调控。但是，对识别和创建一个基因调控网络而言，研究者只对有相互作用的情况感兴趣，因为在这样一个网络中，两个基因之间的相互作用意味着这两个接点（或基因）之间有一条边相连。在高斯图形模型下，两个接点（基因）之间的这种关系可以用偏相关系数矩阵 $\Pi=(\pi_{ij})$ 来描述。这个矩阵可以从上面提到的相关系数矩阵 P 通过如下关系得到：

$$\pi_{ij}=-\omega_{ij}/\sqrt{\omega_{ii}\omega_{jj}}$$

这里，$\Omega=P^{-1}=(\overline{\omega}_{ij})$，是相关系数矩阵的逆矩阵。这些偏相关系数 π_{ij} 描述了任何两个基因在给定其他所有基因的情况下的相关系数。

醛处理使蛋白质与和它结合的 DNA 交联，超声波处理使染色体断裂成小片段，通过抗体沉淀蛋白质-DNA 交联复合体，然后解除交联、扩增 DNA 并标记上荧光物质，最后，荧光标记的扩增片段与包含有启动子序列的微矩阵杂交。通过和不同类型荧光标记的参照比较，就能判断蛋白质的结合位点。因为酵母基因组相对简单，较容易扩增并得到酵母基因组内包含顺式调节信息的基因间 DNA 片段，所以相对于其他真核生物，制备包含酵母调控序列的微矩阵芯片更为容易。

利用 ChIP-to-chip 技术从全基因组范围内分析转录因子的结合模式非常重要。

第一，通过 ChIP-to-chip 技术，初步筛选转录因子在体内直接调控的下游基因。如果比较正常细胞和特定基因被敲除或高表达细胞的表达转录谱，试图通过分析其他基因的表达寻找该基因的直接调控基因，这个方法很难辨别出基因的差异表达是直接调控、间接调控或是多种效应共同作用的结果。如果这个基因对于细胞的存活是必需的，就很难得到存活的突变体。*Esa1* 基因是酵母细胞唯一的组蛋白乙酰转移酶，它是细胞存活的必需基因，在运用 ChIP-to-chip 技术以前，该基因在体内的生物学功能并不被了解，采用该技术后发现 *Esa1* 基因具有调节核小体蛋白基因转录的功能。

第二，蛋白质在全基因组范围内的结合分布提供了大量关于 DNA 和蛋白质在体内相互作用特异性的数据。这样就有可能比较这些结合 DNA 的序列，识别蛋白质和 DNA 结合特异性的决定序列和决定结合特异性的染色质结构特征。例如，细胞周期调节因子 SBF、MBF 和 Rap1 蛋白通用的蛋白质结合位点序列，就是通过比较它们的调节序列得知的。有趣的是，即使在基因编码区含有该通用序列，蛋白质依然偏向于结合调节区的通用序列，这提示了可能有其他因素在影响蛋白质和染色质的结合，比如染色质结构。

第三，通过了解转录因子和其他蛋白质在全基因组范围内的结合分布，有助于了解基因组特征。例如，在染色体特定区域的蛋白质结合位点簇。

第四，同时分析转录因子在全基因组范围内的结合分布和全基因组表达谱数据，有助于揭示细胞复杂生理进程的基因调控网络。结合转录因子的基因很有可能就是转录激活蛋白（activator）或者是阻遏蛋白（repressor），初始信号就是通过一系列的激活或阻遏调控最终形成级联反应。

ChIP-to-chip 技术已经成功应用于蛋白质在全基因组上结合位点分布图的研究，实验结果可以通过 PCR 的方法进行验证。相当多的研究致力于揭示哺乳动物基因组蛋白质结合位点分布图。应用 ChIP-to-chip 技术研究激活蛋白 E2F 结合位点，部分已知人类基因组靶 DNA 的研究结果证明，该技术适合于寻找哺乳动物蛋白质结合位点。但是，哺乳动物基因与基因之间的距离较长，一般有 10 个到几百个碱基，所以几乎不可能制备覆盖所有基因间区域的全基因组微矩阵芯片。能否利用 cDNA 微矩阵作为替代芯片还是值得考虑的，因为，如果转录因子或其他蛋白质的结合位点靠近基因起始位点，ChIP 扩增的 DNA 片段应该包含内含子的序列片段，这样就可以利用 cDNA 微矩阵芯片来获知所研究蛋白质的结合区域。

三、展望

微矩阵芯片作为一门新兴技术，虽然在很多技术细节上还有待完善，并需要结合新的实验手段，但在短短十来年的应用研究中已显示出巨大作用和应用潜力。更重要的是，微矩阵芯片技术敲开了高通量研究技术的大门，未来，各种形式的高通量技术会不断涌现，今后的生物学技术必然由此走上高通量、微量化的道路。在临床上，疾病的诊断和治疗将摆脱仅仅依据疾病的表面现象和少数分子标记的历史，在技术可靠性得到保证的基础上，高通量手段将应用于临床常规检测，为疾病的诊断和治疗提供丰富的信息。目前，利用生物信息学技术

整合基因组信息和基因表达谱信息，研究人员可以获得全基因表达和基因组动力学的全貌，包括蛋白质结合位点分布图、DNA 修饰和拷贝数的变化等信息。对于肿瘤学研究，全方位的信息量是深入了解肿瘤发病机理的有力保证。本章花了很多笔墨来讨论微矩阵芯片在临床上的应用，但微矩阵芯片只是高通量技术的开始，而非全部，希望今后在常规临床实验室里能够发生真正的革命性变化。

（刘三震　李瑶）

参　考　文　献

1 Adorjan P, Distler J, Lipscher E, et al. Tumour class prediction and discovery by microarray-based DNA methylation analysis. Nucleic Acids Res, 2002, V (30): e21

2 Alizadeh AA, Eisen MB, Davis RE, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature, 2000, 403: 503～511

3 Armstrong SA, Staunton JE, Silverman LB, et al. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nat Genet, 2002, 30: 41～47

4 Augenlicht LH, Wahrman MZ, Halsey H, et al. Expression of cloned sequences in biopsies of human colonic tissue and in colonic carcinoma cells induced to differentiate in vitro. Cancer Res, 1987, 47 (22): 6017～6021

5 Azuaje F. In silico approaches to microarray-based disease classification and gene function discovery. Ann Med, 2002, 34 (4): 299～305

6 Baylin SB, Herman JG. DNA hypermethylation in tumorigenesis: Epigenetics joins genetics. Trends Genet, 2000, 16: 168～174

7 Ben Mamoun C, Gluzman IY, Hott C, et al. Co-ordinated programme of gene expression during asexual intraerythrocytic development of the human malaria parasite Plasmodium falciparum revealed by microarray analysis. Mol Microbiol, 2001, 39: 26～36

8 Bodrossy L, Sessitsch A. Oligonucleotide microarrays in microbial diagnostics. Curr Opin Microbiol, 2004, 7 (3): 245～254

9 Bruder CE, Hirvela C, Tapia-Paez I, et al. High resolution delection analysis of constitutional DNA from neurofibromatosis type 2 (NF2) patients using microarray-CGH. Hum Mol Genet, 2001, 10: 271～282

10 Buckley PG, Mantripragada KK, Benetkiewicz M, et al. A full-coverage, high-resolution human chromosome 22 genomic microarray for clinical and research applications. Hum Mol Genet, 2002, 11 (25): 3221～3229

11 Campbell CJ, Ghazal P. Molecular signatures for diagnosis of infection: Application of microarray technology. J Appl Microbiol, 2004, 96 (1): 18～23

12 Chittur SV. DNA microarrays: Tools for the 21st Century Comb Chem High Throughput Screen, 2004, 7 (6): 531～537

13 Chizhikov V, Rasooly A, Chumakov K, et al. Microarray analysis of microbial virulence factors. Appl Environ Microbiol, 2001, 67: 3258～3263

14 Chizhikov V, Wagner M, Ivshina A, et al. Detection and genotypingof human group A rotaviruses by oligonucleotide microarray hybridization. J Clin Microbiol, 2002, 40: 2398～2407

15 Chung CH, Bernard PS, Perou CM. Molecular portraits and the family tree of cancer. Nat Genet, 2002, 32: 533～540

16 Churchill GA. Fundamentals of experimental design for cDNA microarrays. Nat Genet, 2002, 32 Suppl 2: 490～495

17 Daigo Y, Chin SF, Gorringe KL, et al. Degenerate oligonucleotide primed-polymerase chain reaction-based array comparative genomic hybridization for extensive amplicon profiling of breast cancers : A new approach for the molecular analysis of paraffin-embedded cancer tissue. Am J Pathol, 2001, 158: 1623～1631

18 Dean FB, Hosono S, Fang L, et al. Comprehensive human genome amplification using multiple displacement amplification. Proc Natl Acad Sci, 2002, 99: 5261～5266

19 Enders G. Gene profiling——chances and challenges. Acta Neurochir Suppl, 2004, 89: 9～13

预防措施，阻止病情恶化。如早期诊断出肿瘤可以采取预防性手术或化疗等措施；如育龄夫妇为遗传病基因携带者，则采取选择性的生育计划或辅助生育技术，并进行相关的产前诊断，以避免患儿出生，实现优生优育。

随着后基因组研究的深入，必将揭示更多的基因分型与疾病发生、发展和预后的关系，利用高通量的基因芯片就可以对人的健康状况提供指导。

当然，要最终实现个性化治疗还有很长的一段路要走，随着基因组医学研究的深入和生物技术的不断完善，人们的目标正一步步地实现。相信在不久的将来，这些研究成果会对人类疾病的治疗和预防产生深远的影响。

（李瑶）

参 考 文 献

1 Banerjee N, Zhang MQ. Functional genomics as applied to mapping transcription regulatory networks. Curr Opin Microbiol, 2002 6, 5 (3): 313～317

2 Finkelstei D, Ewing R, Gollub J, et al. Microarray data quality analysis: Lessons from the AFGC project. Arabidopsis Functional Genomics Consortium. Plant Mol Biol, 2002, 1, 48 (1-2): 119～131

3 Oliver DJ, Nikolau B, Wurtele ES. Functional genomics: High-throughput mRNA, protein, and metabolite analyses. Metab Eng, 2002, 1, 4 (1): 98～106

4 Schena M, Heller RA, Theriault TP, et al. Microarrays: Biotechnology's discovery platform for functional genomics. Trends Biotechnol, 1998, 7, 16 (7): 301～306

5 Joos L, Eryuksel E, Brutsche MH. Functional genomics and gene microarrays——the use in research and clinical medicine. Swiss Med Wkly, 2003, 1, 133 (3-4): 31～38

6 Cambien F. The epidemiologist, genetics and system biology. Eur J Epidemiol, 2004, 19 (3): 201～203

7 Attur MG, Dave MN, Tsunoyama K, et al. "A system biology" approach to bioinformatics and functional genomics in complex human diseases: Arthritis. Curr Issues Mol Biol, 2002, 10, 4 (4): 129～146

8 Zaiou M. The future of genetic and genomic medicine in health risk assessment and disease: A path toward individualized medicine. Pharmacogenomics, 2005, 1, 6 (1): 7～12

9 Knaup P, Ammenwerth E, Brandner R, et al. Towards clinical bioinformatics: Advancing genomic medicine with informatics methods and tools. Methods Inf Med, 2004, 43 (3): 302～307

10 Garber K. Related Articles, Links Genomic medicine. Gene expression tests foretell breast cancer's future. Science, 2004, 3, 303 (5665): 1754～1755

11 Hall WD, Morley KI, Lucke JC. The prediction of disease risk in genomic medicine. EMBO Rep, 2004, 10, 5: 22～26

12 Zavras AI. Related Post-marketing drug safety in the era of genomic medicine. J Dent Res, 2005, 2, 84 (2): 105～106