

差异表达基因检测 数据分析研究

◎ 纪兆华 著

本专著得到北京市教育委员会科技计划一般项目支持，项目编号：
KM201910857002；项目名称：基于云计算技术的生物大数据可视化
分析平台构建及其关键技术研究。

差异表达基因检测 数据分析研究

纪兆华 著

 **北京理工大学出版社**

BEIJING INSTITUTE OF TECHNOLOGY PRESS

内 容 简 介

本文研究单基因水平基因表达谱数据的差异表达基因检测方法,提出了采用三均值和三均值绝对离差为统计量,利用样本信息和稳健性特征,用于差异表达基因检测,以较全面、稳定地表示样本信息;在ORT方法基础上提出了TriORT差异表达基因检测方法,利用四分位差判断基因芯片数据的异常差异值;提出了基于三均值的TriMOST差异表达基因检测方法,通过用三均值和三均值绝对差对样本表达值进行转换;对比分析了改进方法和已有差异表达基因检测方法在乳腺癌真实数据上的应用,为揭示疾病发病机制等提供算法帮助。

版权专有 侵权必究

图书在版编目(CIP)数据

差异表达基因检测数据分析研究/纪兆华著. —北京:北京理工大学出版社, 2021.5
ISBN 978-7-5682-9846-9

I. ①差…II. ①纪…III. ①差异性-基因表达-检测-数据处理-研究 IV. ①Q786
中国版本图书馆CIP数据核字(2021)第093172号

出版发行/北京理工大学出版社有限责任公司

社 址/北京市海淀区中关村南大街5号

邮 编/100081

电 话/(010)68914775(总编室)

(010)82562903(教材售后服务热线)

(010)68948351(其他图书服务热线)

网 址/<http://www.bitpress.com.cn>

经 销/全国各地新华书店

印 刷/唐山富达印务有限公司

开 本/710毫米×1000毫米 1/16

印 张/10

字 数/170千字

版 次/2021年5月第1版 2021年5月第1次印刷

定 价/48.00元

责任编辑/江 立

文案编辑/江 立

责任校对/周瑞红

责任印制/施胜娟

图书出现印装质量问题,请拨打售后服务热线,本社负责调换



近年来，生物数据随着生物技术的发展不断地增加，规模通常很大。高通量测序技术迅猛发展，使生物信息学进入了大数据时代，由此所引发的多组学海量生物数据更需要利用云的方式来解决存储和分析等问题。随着以高通量测序技术为代表的相关实验技术的不断发展和普及，科研人员可以更加容易和高效地获得大量的生物数据，其中蕴含着大量的“宝藏”等待人们探索。云计算正是一种通过 Internet，以服务的方式提供动态可伸缩、虚拟化的资源计算模式。云计算技术应用到生物信息学的大数据中，面向大数据的生物信息云有助于更好地应对生物信息大数据带来的新挑战，挖掘生物数据中蕴含的大量“宝藏”。

活体细胞内的基因通常按照一定的顺序进行基因表达，但在某些情况下，会因环境条件等因素的变化导致基因突变，并引起一定的表型异常变化，即所谓的差异表达基因。差异表达基因检测统计方法作为近几年迅速发展的生物学前沿技术之一，其主要目的是分析基因表达谱数据的生物学意义，并采用微阵列基因芯片技术同时、快速、准确地检测成千上万种基因是否有差异表达。

差异表达基因检测研究单基因水平的基因表达谱数据，利用统计学中的假设检验，从基因表达谱数据中筛选出潜在的、过表达的癌症样本，并研究有关基因和基因群组，发现癌症特异性基因。差



异表达基因检测的应用广泛，如研究适应药物作用的分子机制、寻找新药开发源头的药物靶标、筛选多靶点高通量药物、评价药物活性和毒性等，对揭示癌症疾病发生机制、开发抗癌药物等有重要的意义。基因芯片差异表达基因检测技术的核心方法通常基于统计学原理，即利用统计学中的假设检验从基因表达谱中筛选出潜在的特异性基因。传统差异表达基因检测的前提是假设整个癌症组样本的基因表达强度相对于正常组样本的基因表达强度都存在过表达的情况。2005 年 Tomlins 等人在 *Science* 杂志上撰文指出差异基因表达可以只出现在癌症组样本的某个子集中，而不是整个癌症组样本。近年来，大量的研究工作针对癌症组样本子集的差异基因表达问题展开，并且产生了多种用于解决这类问题的统计方法。

差异表达基因检测在医学领域中是一个重要问题，并可以在分子水平实现对癌症亚型的准确识别，具有广泛的应用，在研究适应药物作用的分子机制，寻找新药开发源头的药物靶标等方面有重要的意义。针对多发性癌症疾病，如女性多发的乳腺癌疾病，进行基因水平的数据分析，发现致癌基因和其生物学性能。基因芯片差异表达基因检测技术的核心方法通常利用统计学中的假设检验并采用微阵列基因芯片技术快速、准确地检测成千上万种基因是否有差异表达，从基因表达谱中筛选出潜在的特异性基因，分析基因表达谱数据的生物学意义。本书的主要工作如下。

(1) 对六种广泛应用的差异基因检测方法做了比较研究，包括 T 统计方法、PPST 方法、COPA 方法、OS 方法、ORT 方法和 MOST 方法。T 统计方法是传统的差异表达基因检测方法，它假定癌症组样本相对于正常组样本普遍呈现过表达，通过计算正常组样本和癌症组样本的均值及合并标准差，求得 T 统计值。PPST 方法通过识别在 A 组统计学意义样本中表达值强度超过 B 组统计学意义样本中表达值强度一个特定百分比的基因来检测差异基因。COPA 方法、OS 方法、ORT 方法和 MOST 方法用样本中值和中值绝对离差进行样本表达值的转换。OS 方法在 COPA 方法的基础上利用四分位数间距度



量数据分散性，ORT 方法和 OS 方法的区别在于 OS 方法的基因表达数据是正常组样本和癌症组样本的全部数据一起使用，而 ORT 方法是基于正常组样本数据定义的。MOST 方法隐性地考虑差异基因表达强度临界值所有可能的取值，通过确定其统计量最大值来确定阈值，从而检测差异表达基因。本书通过模拟实验和对真实数据的实验，比较和分析了 T 统计方法、PPST 方法、COPA 方法、OS 方法、ORT 方法和 MOST 方法等差异表达基因检测方法。

(2) 提出了两种用于差异表达基因检测的统计量，即三均值和三均值绝对离差。当微阵列基因芯片数据中存在差异表达基因值时，其均值易受差异表达基因值影响，中值具有较好的稳健性且受差异表达值的影响较小。三均值综合利用了上四分位数、下四分位数、中位数三个数据，对异常数据具有较强的抗干扰性。当需要充分利用样本信息和稳健性特征时，样本三均值和三均值绝对离差可以描述数据变化，不忽略距离中位数较远的信息，从而能够全面、稳定地表示样本的信息。

(3) 提出了针对癌症样本组子集的差异表达基因检测方法。在 ORT 方法的基础上提出了差异表达基因检测方法——TriORT 方法，TriORT 方法以三均值和三均值绝对离差为统计量表示数据转换的变化。TriORT 方法采用了中位数及少数其他次序统计量，能充分反映基因芯片中样本数据的特征，并且稳健性较强。通过启发式规则附加表达值，利用四分位差判断基因芯片数据的异常差异值，从而进行差异表达基因检测。实验结果表明，本书提出的基于三均值和三均值绝对离差的差异表达基因检测方法对于癌症组样本子集相对于正常组样本过表达的差异表达基因检测有效，并具有较好的敏感性和特异性。

(4) 提出了针对癌症样本组子集的差异表达基因检测方法。在 MOST 方法的基础上提出了一种基于三均值的差异表达基因检测方法，称为 TriMOST 方法。该方法将三均值引入癌症组样本子集相对于正常组样本过表达的差异表达基因检测方法中，通过三均值和三



均值绝对离差对样本表达值进行转换。当差异基因活跃的数目未知时，又引入了均值和方差，使用标准化的、代数形式的表达值间的差别确定差异基因的标准，尽可能较全面地考虑可能的阈值，将可能的值默认为差异表达阈值，从而使得检测的效果理想。

(5) 讨论并分析了改进的方法和已有差异表达基因检测方法在乳腺癌真实数据上的应用。为了进一步研究本书所提出的差异表达基因检测方法的性能，先将改进方法的仿真实验结果和已有方法的实验结果进行比较，然后将改进方法应用到乳腺癌真实数据集 West (2001)，再将得到的结果在 NCBI 数据资源库上进行验证，并针对验证结果对各方法的性能进行对比分析。检测乳腺癌差异表达基因，并认识相应基因群，这为乳腺癌疾病的治疗提供了有益的辅助信息。

(6) 对 RNA-Seq 数据分析软件的研究与开发一直是生物信息学中的研究热点。各个分析任务也都有传统的分析软件来完成各项工作，如 Bowtie、Tophat 和 Cufflinks 等。另外，一些研究人员将这些软件通过脚本程序组装成分析流程，如 PRADA、wapRNA 等。然而 OpenMP 或 Pthread 等传统的并行模型无法运行于云计算平台之上。云计算理念的出现使得分布式并行计算在解决大数据问题时的可用性和易用性得到了极大的提升和扩展。引发某种疾病的因素是多方面的，从生物信息学理论研究诱发同该疾病有关的致病基因，分析基因表达谱数据，检测差异表达基因，为揭示多发疾病发病机制、开发药物、临床诊断等提供帮助。

大数据为医疗领域带来了挑战和机遇，特别是对一些恶性疾病如癌症等的治疗。某种单一类型的肿瘤通常会伴随着产生差异基因表达数据的多样化基因突变，治疗投入增加，治疗靶点也增多。随着对大数据分析精度越来越高，对整个疾病发生过程的了解也会越来越深入，大数据分析成为治疗疾病的“利器”，更多的精准治疗方案会帮助人们做出更好的选择。



第1章 绪论	1
1.1 生物芯片概述	1
1.2 基因芯片简介	1
1.2.1 基因芯片的原理	2
1.2.2 基因芯片的分类	5
1.2.3 基因芯片的应用	5
1.3 差异表达基因检测统计方法	6
1.3.1 差异表达基因检测概述	6
1.3.2 传统的差异表达基因检测方法	8
1.3.3 其他相关的差异表达基因检测方法	9
1.4 变点问题的研究	11
1.5 变点理论应用于差异表达基因检测	12
1.6 微阵列基因芯片数据的网络资源介绍	12
1.7 基于云计算的差异表达基因检测	14
1.8 小结	15
第2章 差异表达基因检测样本子集过表达的统计方法	17
2.1 生物学背景	17
2.2 差异表达基因检测统计方法介绍	19
2.2.1 基于均值的差异表达基因检测方法	20
2.2.2 基于中值的差异表达基因检测方法	20
2.2.3 基于样本表达值特定百分比的差异表达基因检测方法	24



2.3 实验与分析	24
2.3.1 ROC 曲线比较	25
2.3.2 FDR 曲线比较	27
2.3.3 算法分析	29
2.4 小结	30
第3章 TriORT 差异表达基因检测统计方法	32
3.1 引言	32
3.2 TriORT 方法	33
3.3 实验与分析	34
3.3.1 仿真研究	35
3.3.2 真实数据研究	38
3.4 小结	41
第4章 TriMOST 差异表达基因检测统计方法	42
4.1 引言	42
4.2 TriMOST 方法	43
4.3 实验与分析	45
4.3.1 仿真研究	46
4.3.2 真实数据研究	50
4.4 小结	56
第5章 差异表达基因检测统计方法在乳腺癌数据上的应用	57
5.1 引言	57
5.2 差异表达基因检测方法	59
5.2.1 已有方法介绍	60
5.2.2 TriORT 方法和 TriMOST 方法	61
5.3 实验与分析	62
5.3.1 ROC 曲线仿真实验	63
5.3.2 FDR 曲线仿真实验	65
5.3.3 真实数据集实验	68
5.3.4 各方法找到的与乳腺癌相关的基因分析	72
5.4 小结	76
第6章 基于样本子集的差异表达基因检测方法比较	77
6.1 引言	77



6.2 差异表达基因检测方法	77
6.2.1 T 统计方法	78
6.2.2 相对正常组样本的半 T 统计方法	79
6.2.3 相对癌症组样本的半 T 统计方法	79
6.2.4 增加参数的半 T 统计方法	79
6.3 实验和结果分析	80
6.3.1 基因表达强度相同, 差异表达基因数目变化	82
6.3.2 基因表达强度变化, 差异表达基因数目相同	83
6.3.3 基因表达强度相同, 差异表达基因数目相同, 基因数目不同 ..	84
6.3.4 基因表达强度相同, 差异表达基因数目相同, 基因数目相同, 样本数目不同	85
6.4 小结	86
第7章 基于样本子集的乳腺癌微阵列差异表达基因检测	87
7.1 引言	87
7.2 样本子集过表达的差异表达基因检测	88
7.2.1 基于基因表达强度分位数检测二分类差异基因表达方法	89
7.2.2 基于似然性方法检测二分类差异基因表达方法	92
7.2.3 基于基因表达分位数检测多分类的差异基因表达方法	93
7.3 实验比较	97
7.3.1 数据集说明	98
7.3.2 仿真研究二分类差异表达基因检测方法比较	99
7.3.3 仿真研究多分类差异表达基因检测方法比较	104
7.3.4 真实数据研究	107
7.4 实验总结	109
7.5 小结	110
第8章 基于云计算技术的差异表达基因鉴定流程	112
8.1 云计算技术	112
8.1.1 云计算的内涵	112
8.1.2 云计算的关键技术	113
8.1.3 云计算和大数据	115
8.2 基于大数据技术的机器学习算法	116
8.3 基于 Spark + Hadoop 的机器学习算法	116
8.3.1 Hadoop 技术和 Spark 技术	117



8.3.2 基于 Spark 的聚类算法	118
8.3.3 基于 Spark + Hadoop 的机器学习	119
8.4 RNA - Seq 数据分析	120
8.5 RNA 转录组的高通量全测序	122
8.5.1 高通量测序技术	122
8.5.2 RNA - Seq	122
8.5.3 RNA 转录组的高通量全测序	122
8.5.4 Hadoop 云计算框架设计实现 RNA - Seq 大数据分析流程	122
8.5.5 生物信息学用云的方式来解决存储和分析等问题	125
8.5.6 云计算理念提升了分布式并行计算解决大数据问题	125
8.6 基于云计算的基因差异表达分析流程方法	126
8.6.1 Reads 片段的 Mapping	126
8.6.2 计算各个基因的表达量	127
8.6.3 鉴定差异表达的基因	128
8.7 小结	129
第9章 总结与展望	130
9.1 总结	130
9.2 展望	131
参考文献	133
所主持的相关科研项目	147
致谢	148

第1章 绪 论

细胞内的表达基因由于环境条件等因素的改变可能发生基因突变，从而发生一定的基因表达异常变化，这种基因表达的异常变化可以采用微阵列基因芯片技术进行检测。统计学中的假设检验通常可以对单基因水平的基因表达谱数据进行研究，从而进行基因的差异表达检测。差异表达基因检测的统计方法研究在生命科学、医学治疗、新药开发、农业技术等领域有着非常重要的作用，对揭示疾病发生机制、农作物基因育种等有着重要的研究意义。

基因表达谱已广泛应用于癌症临床生物学的研究，在基因表达谱研究中，过高表达或过低表达统称为过表达。

1.1 生物芯片概述

生物芯片是电子技术和生物技术互相结合形成的半导体芯片，是通过生物技术制作形成或者在生物技术上得以应用的微处理器^[1]。生物芯片作为微型生物化学分析系统，可以对 DNA、蛋白质、细胞等生物成分进行准确的筛选和检测，其显著特点是高通量、大规模、微型化、自动化、平行操作和快速准确。根据其在固定载体上的物质成分不同，生物芯片分为有组织芯片、细胞芯片、蛋白质芯片与基因芯片^[2]。

1.2 基因芯片简介

基因芯片能同时测量成千上万种基因的表达，检测特定癌症类型的异位。异位可分为两类：一类导致蛋白质融合，对细胞有异常的影响；另一类使基因的启动区转移到癌症基因完整译码区，癌症基因的表达强度变化，造成过表达，诱发癌症的发生。微阵列基因芯片实验能通过癌症异常检测来研究染色体异位，进行癌症差异表达基因检测，发现基因芯片上癌症组样本子集相对于正常组样本有过高或过低表达的基因。



1.2.1 基因芯片的原理

基因芯片 (Gene Chip) 技术是前沿生物技术之一, 是将 DNA 或互补 DNA 序列等按微阵列方式固定在微型载体上制成的生物芯片, 又称为 DNA 芯片 (DNA Chip) 或 DNA 微阵列等, 利用基因芯片可以快速定性和定量分析样本的基因表达谱生物信息。美国 Affymetrix 公司在 20 世纪 80 年代末 90 年代初就开展了关于基因芯片的研究^[3]。基因芯片是半导体微电子技术、分子生物学技术、物理学、激光化学和计算机科学等多种学科技术的有机结合。相对于其他生物芯片而言, 基因芯片的研究开发时间较早, 因此技术上也很成熟, 并具有快速、高通量、并行化采集处理生物信息等特点^[4], 是应用非常广泛的生物芯片产品^[5,6]。

微阵列基因芯片数据是经过杂交的阵列形成的扫描图像, 能够显示每一个点的杂交信号强度。该图像能够通过双通道荧光标记、比色标记或同位素等方法形成^[7]。微阵列基因芯片实验产生的原始数据将形成基因表达矩阵 (各行表示表达基因, 各列表示不同的实验条件、环境等, 矩阵中的数据代表反映各个基因相对表达水平的基因表达信号的强度), 基因芯片的制备流程如图 1.1 所示。

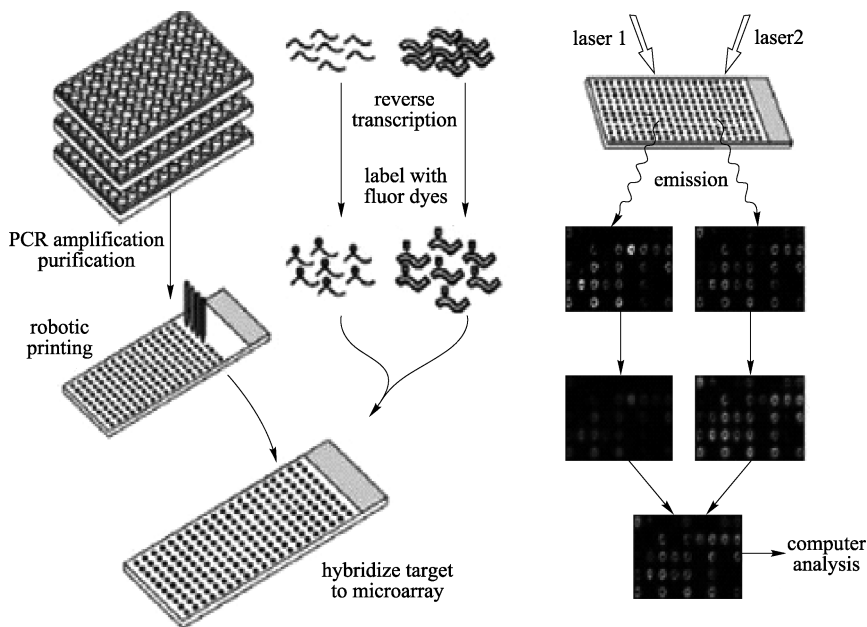


图 1.1 基因芯片的制备流程 (引自文献 [7])

以人类乳腺癌多柔比星耐药细胞株 MCR-7/ADM 和 MCF-7 (MCR-7/ADM 的亲本细胞株) 为例, 图 1.2 和图 1.3 分别是人类乳腺癌细胞 MCF-7 的表现形态、MCR-7/ADM 细胞株的表现形态 (细胞来源于 ATCC: American Type Culture Collection)。

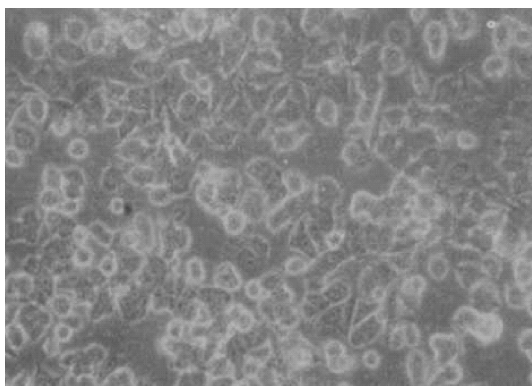


图 1.2 人类乳腺癌细胞 MCF-7 的表现形态 (引自文献 [8])

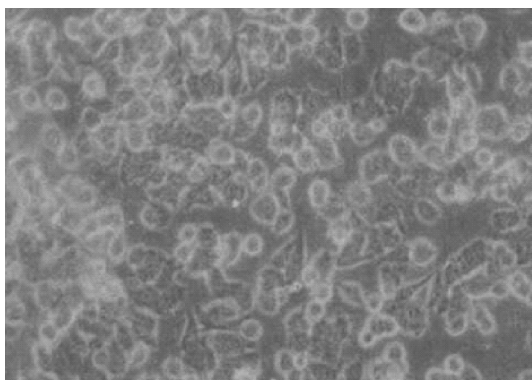


图 1.3 MCR-7/ADM 细胞株的表现形态 (引自文献 [8])

对基因芯片杂交并进行扫描, 呈现红色的 Cy5 标记为 MCF-7, 呈现绿色的 Cy3 标记为 MCR-7/ADM。基因芯片如图 1.4 所示。

为使基因芯片更清晰可见, 把基因芯片局部放大, 显示可见的 Cy3 信号和 Cy5 信号 (计算机叠加图像处理的基因图片来源于南京凯基生物公司), 如图 1.5 所示。

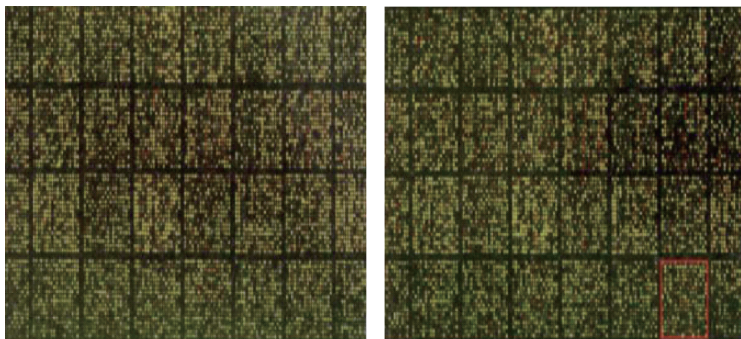


图 1.4 基因芯片 (引自文献 [8])

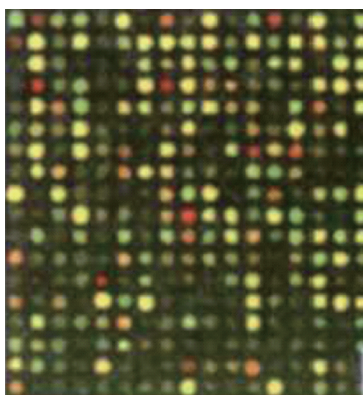


图 1.5 基因芯片的局部放大显示 (引自文献 [8])

图 1.5 反映每个基因在正常组织和癌症组织中表达的差异,从图中可以看到红色、绿色和黄色三种颜色的信号。其中,红色代表差异基因表达强度值上调,绿色代表差异基因表达强度值下调,黄色代表该基因表达强度值在两种组织中接近。从图 1.5 中也可以看出大部分的基因表达水平基本接近,小部分的基因表达发生了改变,呈现红色或绿色^[9]。

把微阵列基因芯片实验产生的原始数据转换成基因表达矩阵,该矩阵能反映成千上万的基因相对表达水平(见图 1.6),基因表达矩阵的每一个基因在一定条件下的差异基因表达强度值,对微阵列基因芯片的数据进行分析,即把这些实验数据值按表达模式进行差异表达基因检测^[10]。例如,通过基因芯片检测到肿瘤的差异表达基因,因为差异表达基因在肿瘤的耐药疗效发展过程中起着重要作用,为有效预测化疗效果和确定潜在靶向分子开辟了新的途径^[8]。

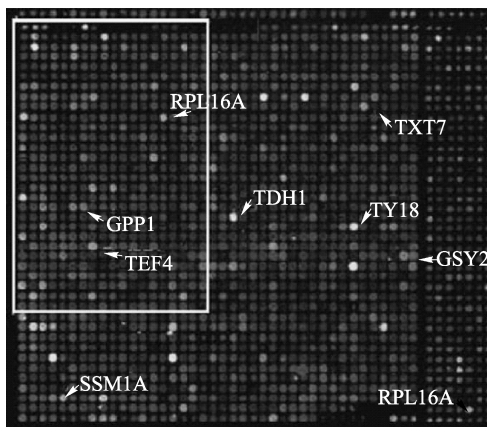


图 1.6 基因芯片反映了基因的表达水平（引自文献 [9]）

1.2.2 基因芯片的分类

基因芯片数据通常可以用来做差异表达基因检测分析、聚类分析和判别分析等^[11]，其应用范围十分广泛。基因芯片的类型是多种多样的，依据制作方法和所使用的载体材料各不相同，划分方法也有所不同^[12]。根据使用的载体材料，基因芯片可以分为硅芯片、膜芯片、玻璃芯片和陶瓷芯片等；根据制作方法，基因芯片可以分为直接点样芯片和原位合成芯片等；根据基因芯片所用探针类型，基因芯片可以分为寡核苷酸探针芯片^[13,14]、cDNA 微阵列芯片^[15-17]等；根据基因芯片的应用，基因芯片可以分为 DNA 测序芯片、甲基化芯片、诊断类芯片和基因表达谱芯片等^[18]。

1.2.3 基因芯片的应用

基因芯片在生命科学、医学研究、环境保护、司法、军事、农业技术等诸多方面皆有非常重要的应用，是当今时代备受人们关注且很有发展前景的高新技术之一^[19-22]。

在农作物的基因育种方面，基因芯片技术已经成为育种工作的一项重要手段，利用基因芯片技术可以在多样本、高通量群体中进行基因筛选，从而找到携带目的基因的优良个体，充分利用有利的基因序列资源，创造方便、快捷的育种工作环境。在基因功能的研究方面，基因表达谱能够高通量、高效率地检测不同条件、不同组织、不同细胞的差异表达基因，同时研究大量基因在不同环境条件下受到不同刺激的差异表达，探索相关基因的功能，寻



找未知的基因等^[23-25]。

在临床医学的诊断方面，基因芯片用于从分子水平进行疾病的诊断，将某个基因同一种或者多种疾病的多样性联系起来，确定具体染色体的合适位点，以找到与疾病相关的基因，具有快速、省时、无污染和自动化操作等特点。目前，基因芯片在医学研究领域，如遗传病的遗传机制、病毒病原体的诊断和基因分型等诸多方面有着十分广泛而有效的研究^[26,27]。

在药物的筛选和开发方面，将基因芯片和组合化学联系起来，可以有效寻找药物靶点，提高药物的疗效；快速、大规模地筛选和研制新药，缩短了药物筛选的时间周期^[28-30]。

在司法鉴定方面，便携式 DNA 芯片检测设备可以直接对犯罪现场上遗留的毛发、血液、唾液、精液等进行分析，并进行快速比对，及时侦破案件，提高办案速度和质量。

在军事领域，通过基因芯片可以研制生物战保护剂，开发生物战病原体检测等^[31]。

在环境保护和监测方面，基因芯片可以检测地表环境，快速检测到污染地表的微生物和有毒化合物的成分及其危害程度，及时寻找保护性基因，通过研制基因工程药物来防止危害发生^[18]。

基因芯片是基因测序、基因突变分析以及基因表达研究的重要高效手段之一，其中基本的功能是对基因表达技术的研究。近几年来，以基因芯片技术为代表的生物芯片技术迅速发展，使人类科学技术的研究有了阶段性的进步，对科学技术的发展产生了巨大的影响^[32-34]。

1.3 差异表达基因检测统计方法

基因工程在科学实验中有着广泛的应用，基因工程中的差异表达基因检测方法是功能基因组学研究的重要工具^[35-38]。

1.3.1 差异表达基因检测概述

微阵列基因芯片技术可以同时检测成千上万个生物样本的基因表达水平，从而得到大量的基因表达谱数据。基因表达谱不仅数据量非常庞大，而且在数据中蕴含着丰富的生物学知识^[39,40]。通过分析基因表达谱数据，比较相同基因在不同环境条件、不同组织（如癌症组织和正常组织）中的不同表达强度，发现那些有可能与疾病诊断相关的基因，这就需要从微阵列基因芯片样