差异表达基因检测统计方法研究

Study on Statistical Methods
for Differential Gene Expression Detection

纪兆华 著

北京理工大学出版社
BEIJING INSTITUTE OF TECHNOLOGY PRESS

差异表达基因检测统计方法研究

纪兆华 著

北京理工大学出版社
BEIJING INSTITUTE OF TECHNOLOGY PRESS

图书在版编目（CIP）数据

差异表达基因检测统计方法研究/纪兆华著. —北京：北京理工大学出版社，2017.3

ISBN 978-7-5682-3840-3

Ⅰ.①差…　Ⅱ.①纪…　Ⅲ.①差异性-基因表达-检测②差异性-基因表达-统计方法　Ⅳ.①Q786

中国版本图书馆 CIP 数据核字（2017）第 058103 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／保定市中画美凯印刷有限公司
开　　本／710 毫米×1000 毫米　1/16
印　　张／9.5
字　　数／130 千字
版　　次／2017 年 3 月第 1 版　2017 年 3 月第 1 次印刷
定　　价／38.00 元

责任编辑／李秀梅
文案编辑／杜春英
责任校对／周瑞红
责任印制／王美丽

序

本专著是我博士、博士后期间研究工作的整合，感谢我的博士后导师——北京工业大学环境与能源工程学院彭永臻院士，彭老师学术精湛、为人师表，一直是我的楷模；感谢我的博士导师——吉林大学计算机科学与技术学院梁艳春教授，梁老师对学术精益求精、为人和蔼，培养我打下良好的学术基础。

感谢吉林大学、北京工业大学给了我温馨的生活环境和优良的学习平台，使我能不断汲取新知识，支持我有机会到北京大学、中国科学院北京基因组研究所等高校研究院所学习，也支持我有机会出国参加国际学术会议，让我拓宽视野。

感谢我的研究生徐行健、袁方方、刘芳对我的帮助和支持。

感谢我的母亲，母亲现已 85 岁高龄，依然关心、支持我的成长。

感谢我的爱人、儿子、亲人、朋友、同事、同学，你们给予了我无私的帮助、支持和鼓励，你们给了我温暖，也给了我丰富多彩、阳光灿烂的生活，你们是我前进的永恒动力。

感谢“北京市博士后工作经费资助项目”，资助本专著顺利出版。

感谢内蒙古自治区自然科学基金项目（2015MS0633）的资助。

纪兆华

2016 年 11 月于北京工业大学

前　言

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，即所谓的差异基因表达。基因芯片数据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。

基因芯片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，例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，在揭示癌症疾病发生机制、开发抗癌药物等方面有重要意义。

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样本组的基因表达强度相对于正常样本组的基因表达强度都存在过表达的情况。2005 年，Tomlins 等人在 *Science* 上撰文指出差异表达基因可以只出现在癌症样本组的某个子集中，而不是整个癌症样本组中。近年来，有大量的研究工作针对癌症样本组子集的差异表达基因问题展开，并且产生了多种用以解决这类问题的统计方法。

现代癌症临床医学的研究通常要分析基因表达谱数据，基因表达谱数据分析中把致癌基因表达值所表现出的过高表达或过低表达统称为“过表达”。差异表达基因检测就是研究单基因水平的基因表达谱数据，用来寻找不同实验条件下过表达的致癌基因。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，发现致癌基因及其生物学性能。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。

变点理论把统计控制理论、估计和假设检验理论、非贝叶斯方法和贝叶斯方法结合起来，所研究的统计推断问题能够对估计量的性质进行统计分析，在工业自动控制、医学、金融、信号过程和计算机等方面都有大量的应用。差异表达基因在基因芯片上的基因信号强度数值在表达上具有差异性和相关性，因此差异表达基因的表达强度值可以看作基因表达谱数据中的变点异常值。把统计学中变点理论知识应用到基因表达谱数据分析中，对变点理论的实践应用及差异表达基因检测的研究都提出了新的挑战。

书中针对差异基因表达样本表达强度值状态检测的特点，研究变点检测理论应用到差异表达基因检测的实用化方法，并设计基于变点原理的检测方法应用于差异表达基因检测。根据差异表达基因表达值分布状态的特点，模拟生成基因表达谱数据阵列，研究基于变点的差异表达基因检测问题的具体特征。给出变点的检测和估计（点估计和区间估计）以及估计量的渐近分布和收敛速度，最终实现检测差异表达基因的目的。讨论变点检测方法和 ROC 接受者操作曲线来研究变点检测方法的统计性能分析方法和参数优化的方法。通过蒙特卡洛马尔可夫方法对检测方法进行验证，证明基于变点的差异表达基因检测方

法的有效性。把极大似然比检验方法、最小二乘法应用到差异表达基因检测，对基因表达谱模型统计推断，考虑变点的检验或点估计。

引发该疾病的因素是多方面的，从生物信息学理论研究与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，对揭示多发疾病发病机制，开发药物，临床诊断等提供帮助。

目　录

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而可能发生基因突变，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，这些研究对揭示疾病发生机制、农作物基因育种等方面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，基因表达谱研究中把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半导体芯片的产物，是通过生物技术制作形成或者是在生物技术上得以应用的微处理器。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不同物质成分，生物芯片分为组织芯片、细胞芯片、蛋白质芯片与基因芯片。

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，使癌症基因的表达强度变化，造成过表达，诱发癌症的发生。微阵列基因生物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。基因芯片是将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 公司早在 20 世纪 80 年代末 90 年代初就开展了关于基因芯片的研究。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，因此技术上也比较成熟，并具有快速、高通量、并行化采集处理生物信息的特点，在应用上也是非常广泛的生物芯片产品。

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。从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（其各行表示表达基因，各列表示不同的实验条件、环境等，由矩阵形成的表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），基因芯片的制作流程如图 1.1 所示。

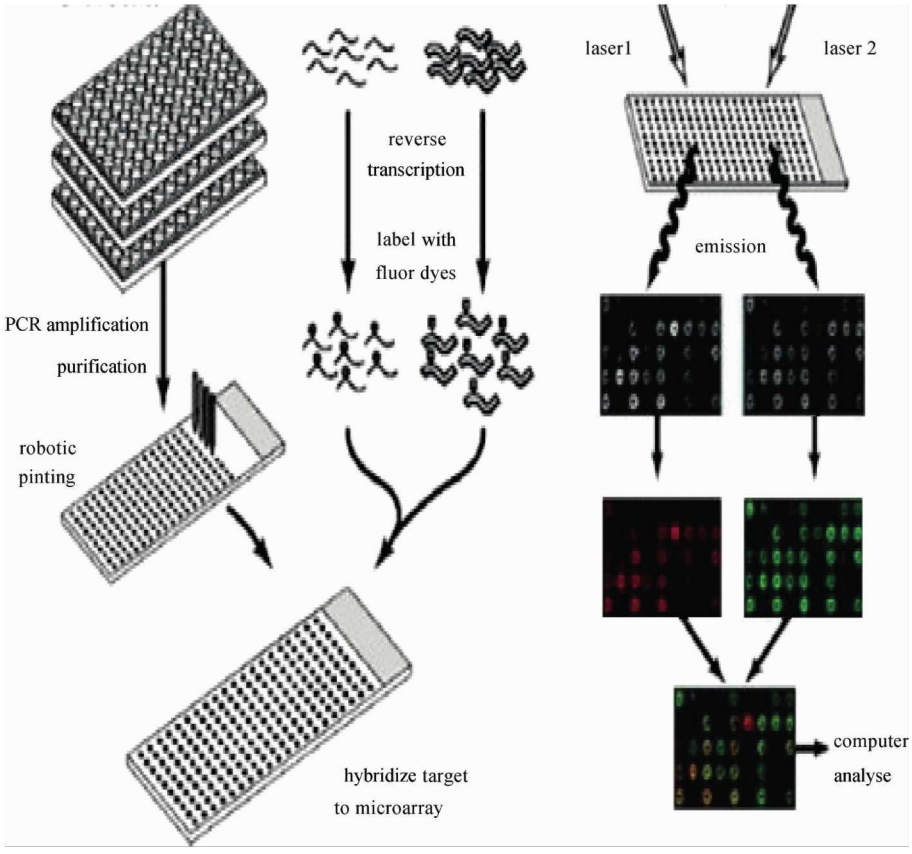

图 1.1　基因芯片制作流程

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 分别显示人类乳腺癌细胞系 MCF－7 的表现形态和 MCR－7/ADM 细胞株的表现形态（细胞来源于 ATCC：American Type Culture Collection）。

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

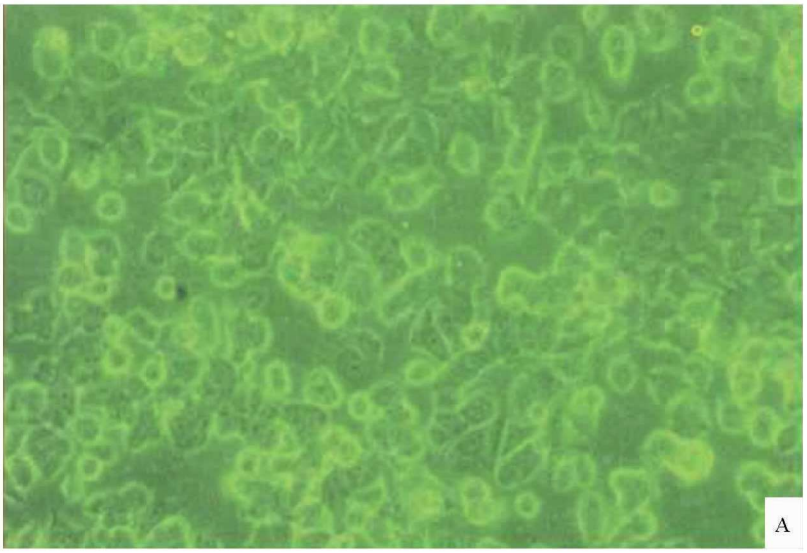

图 1.2　人类乳腺癌细胞系 MCF－7 的表现形态

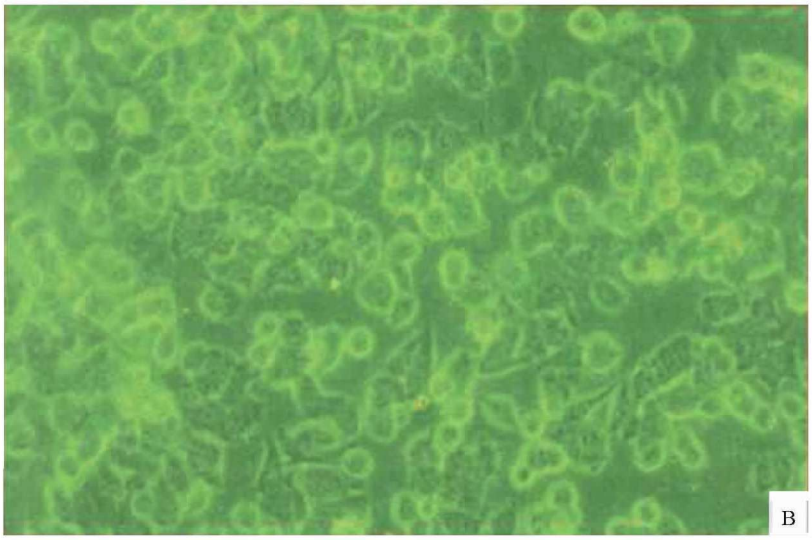

图 1.3　MCR－7/ADM 细胞株的表现形态

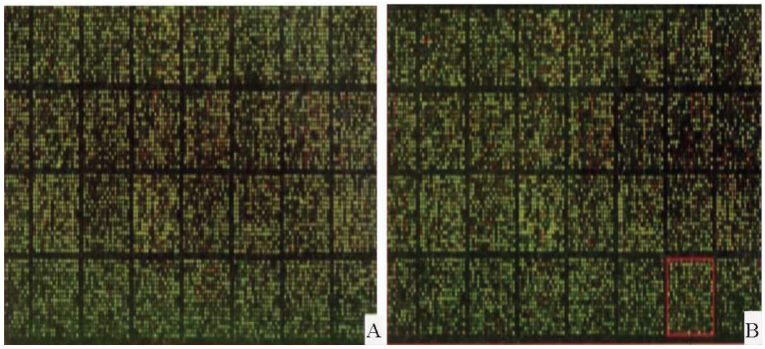

图 1.4　基因芯片图示

为使基因芯片清晰可见，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（基因芯片来源：南京凯基生物公司），如图 1.5 所示。

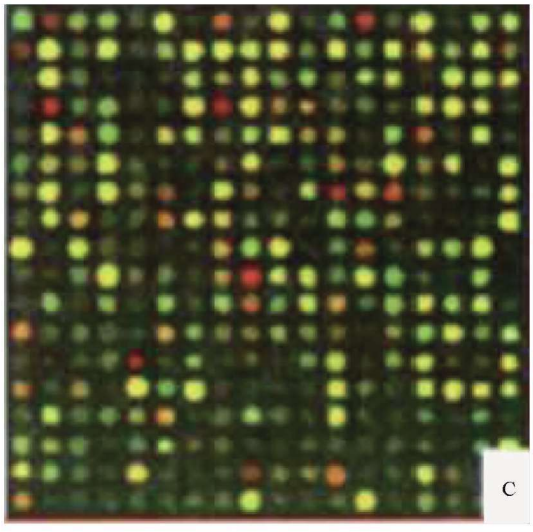

图 1.5　基因芯片的局部放大显示

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，看到有红色、绿色和黄色三种颜色的信号。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在两种组织中表达接近。从图 1.5 中也可以看出，大部分基因表达水平基本接近，小部分基因表达发生了改变，呈现红色或者绿色。

把微阵列基因芯片实验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及每一个基因在一定条件下的表达值，对微阵列基因芯片的数据进行分析，就是把这些实验数据值按表达模式进行差异表达基因检测。例如，通过基因芯片检测到肿瘤的差异表达基因，因为差异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治疗潜在靶向分子的确定开辟了新的途径，如图 1.6 所示。

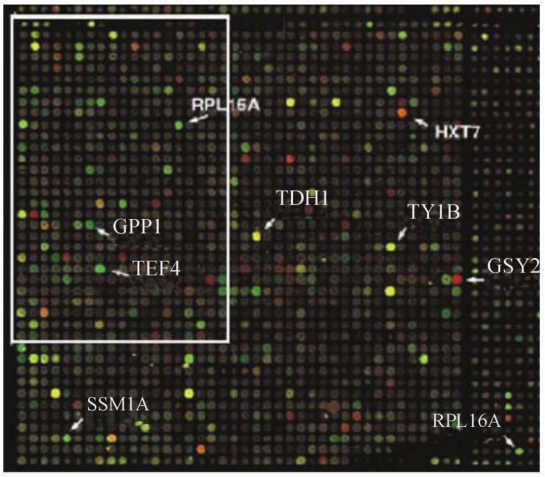

图 1.6　基因芯片反映了基因的表达水平

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，其应用范围十分广泛。依据制作方法和所使用的载体材料不同，基因芯片的类型多种多样，划分方法也有所不同。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、cDNA 微阵列芯片等；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。