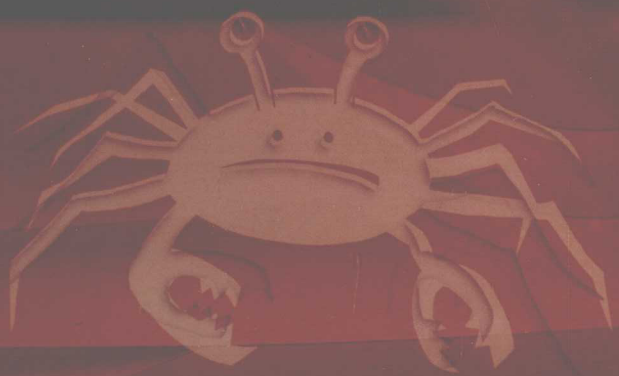
肿瘤研究前沿

第9卷

樊代明 主编



第四军医大学出版社

肿瘤研究前沿

9

肿瘤研究前沿

第9卷

樊代明 主编

第四军医大学出版社·西安

内容简介

本书是全面介绍肿瘤研究进展的系列著作——《肿瘤研究前沿》的第9卷。主要内容包括部分肿瘤的发生发展、血管生成和细胞周期调控的分子机制，还包括了受到广泛关注的生物信息学在肿瘤研究中的应用以及miRNA研究的一些最新进展，反映的内容都是当前肿瘤研究的热点和前沿。本书可作为相关专业研究人员的参考用书，也可供高校、医院的相关人员阅读使用。

图书在版编目(CIP)数据

肿瘤研究前沿(第9卷)/樊代明主编. —西安:第四军医大学出版社,2009.12
ISBN 978-7-81086-750-4

Ⅰ.肿… Ⅱ.樊… Ⅲ.肿瘤-研究 Ⅳ.R73

中国版本图书馆CIP数据核字(2010)第034563号

肿瘤研究前沿(第9卷)

主　　编　樊代明
责任编辑　富　明
出版发行　第四军医大学出版社
地　　址　西安市长乐西路17号(邮编:710032)
电　　话　029-84776765
传　　真　029-84776764
网　　址　http://press.fmmu.sn.cn
印　　刷　西安永惠印务有限公司
版　　次　2009年12月第1版　2009年12月第1次印刷
开　　本　850×1168　1/32
印　　张　12　彩插2页
字　　数　215千字
书　　号　ISBN 978-7-81086-750-4
定　　价　58.00元

编委会

主编简介

樊代明，1953 年出生，重庆市人。教授、主任医师，中国工程院院士。现任第四军医大学校长，西京消化病医院院长，肿瘤生物学国家重点实验室主任，国家临床药理基地主任，中华消化学会主委，中国抗癌协会副理事长，亚太胃肠病学会常务理事兼外事委员会主席，国家教育部长江学者计划特聘教授，973 项目首席科学家。担任国内外 25 家杂志的编委、主编或副主编。目前担任北京大学等 70 余所大学的客座教授或名誉教授。

长期从事消化系统疾病的基础及临床研究，特别是在胃癌的研究中作出突出成绩，先后承担国家 863、973、国家攻关、国家重大新药创制、国家杰出青年基金、国家自然科学基金等课题。获国家科技进步一、二、三等奖各 1 项，国家技术发明三等奖 1 项，主编专著 11 本。发表论文 310 余篇，其中在国外杂志发表论文 251 篇。

序

肿瘤是严重危害人类健康及生命的疾病。尽管国内外已投入大量的人力和财力进行研究，发表的论著也有成千上万，但至今对其病因和发病机制尚不清楚，多数肿瘤在临床诊断、治疗及预防方面也无重大突破。造成这种现状的根本原因除了肿瘤本身的复杂性外，还与各专业的研究者之间沟通较少、“各行其是”，对肿瘤研究的全貌及进展了解不够、顾此失彼，以及各专业在理论及技术上的协作欠佳有关。要解决这个问题，需要有人把各专业对肿瘤研究的重大进展及时进行整理总结并加以评述，从中找出相互间研究的生长点及解决办法，然后适时地介绍给正在或将要从事肿瘤研究的同事。《肿瘤研究前沿》将会适应这种需求，结合著者自己的科研成果，将目前世界上肿瘤研究的最新进展尽力以最通俗的语言介绍给同行及相关研究人员，每年一卷，各卷介绍的内容有所侧重，连续下去，坚持数年，必有好处。如无特殊情况，直至肿瘤被攻克之日。

本书像专著，因为它含有著者的研究成果；它像综述，因为它介绍世界文献的最新进展；它像述评，因为它给出著者的观点及见解；它也像科普读物，因为它力求以最普通的文字面对读者。它以包容性、先进性、焦点争论为特色。这就是它既像什么又不完全是什么的缘故，这就是肿瘤研究的现状，也就是本书追逐的肿瘤研究的前沿。

樊代明

2001.8

目　录

第一章

生物信息学方法与肿瘤相关基因研究

近年来,利用高通量杂交阵列和基于测序技术的分子生物学实验已非常普及。这些技术要么被单一使用、要么被联合使用来评估大量 mRNA 和基因组 DNA 分子的信息。促成这种普及的主要因素是这些技术的平行、高通量特性及其伴随在时间上的高度保守性,即在极为相似的条件下同时(或者几乎同时)进行大量的分子样品实验所获得的信息资源。

随着人类基因组计划、蛋白质计划和生物芯片技术的开展,各种高通量表达谱实验技术如大规模的 EST 测序、SAGE 技术和芯片技术的发展与广泛应用,产生了大量的表达谱数据,这些数据中包含大量有用信息,其中包括各种细胞或组织在不同状态下(包括正常和肿瘤)的基因表达种类和丰度信息,对于寻找肿瘤相关基因起着巨大的推动作用。然而表达谱数据信息的挖掘却相对滞后,如何进一步探索数据中蕴含的信息成为肿瘤生物信息学研究的首要任务,表达谱数据库应运产生。表达谱数据库就是把高通量技术得到的组织和细胞表达谱数据加工、存储,形成利于电子传播和使用的一类生物信息数据库。表达谱数据库从技术原理上划分为三类:表达序列标签(EST)文库、基因表达系列分析(SAGE)文库和 cDNA 微阵列数据库。

如何利用各种日益增长的网络资源,通过网上实验室来克隆、

分析相关基因的结构与功能这个问题就摆在我们面前。在这种形势下,一门新兴的科学——生物信息学应运而生,它是利用数理和信息科学的理论和方法研究生命现象,组织和分析日益剧增的生物信息数据库的一门学科。它主要利用网络技术和不断发展的各种软件,研究遗传物质的载体——DNA 及其编码的功能大分子蛋白质,对日渐增多的序列和结构进行收集、整理、储存、发布、提取和加工,并从中分析发现新的基因序列,从而不断揭示人体生理和病理过程的分子基础,为人类疾病的预防、诊断和治疗提供依据。

一、差异表达基因高通量分析技术

差异表达基因高通量分析是研究肿瘤分子生物学如生化、药理等的有力武器。近年来发展了多种分析基因表达差异的技术,包括表达序列标签比较测序、差异展示、消减克隆、基因阵列的 mRNA 杂交和基因表达系列分析,有力地推动了肿瘤发生、发展机制的研究。

(一)基因表达系列分析(SAGE)

1. 概念

SAGE 是 1995 年由美国霍普金斯大学医学院 Velculescu 等发明的一种新的基因表达谱分析方法,广泛应用于比较疾病不同阶段或不同组织的基因表达差异的研究,以及信号转导途径中上、下游基因的筛选等领域。

2. 理论依据

SAGE 技术的主要理论依据有三个:①来源于转录体特定位置的 9 ~ 13bp 的核苷酸序列包含足够的信息,能确定一个转录

体，可以作为区别转录体的标签（tag）；②这些分离自不同转录体的标签可以被串连成一定长度的多联体（concatemer），克隆入载体进行测序；③同一标签的重复次数代表该转录体的表达水平，结合生物信息学方法可以确定表达的基因种类和基因的表达丰度。SAGE 可以分析不同细胞群的基因表达，从而对正常和疾病状态下组织、细胞的基因表达做定性和定量分析。值得一提的是，SAGE 对低丰度表达基因有较好检测效果。此外，还可以利用得到的未知基因转录产物的短序列发现新基因。缺点是费用昂贵，操作烦琐，结果分析数据庞大。

3. 基本步骤

SAGE 方法的基本步骤为：①提取 mRNA 并将其与生物素酰化的寡聚糖核苷酸 oligo（dT）混合，纯化出带有 polyA$^+$ 的 mRNA，以其为模板经逆转录得到双链 cDNA；②用一个序列特异性（专门识别 4bp 碱基）的锚定酶（anchoring enzyme）如 NIa Ⅲ，消化双链释放 cDNA 序列，而生物素酰化的 5′端仍被吸附在链霉亲和素的磁珠（streptavidin - coatedbeads）上；③将上述样品均匀分成两份，分别与接口分子（adapter）A 和 B 连接，接口分子 3′端含有一段特定的 ⅡS 型限制性内切酶（如 BsmF1）的识别位点，接口分子 3′端可与 NIa Ⅲ 消化产物的黏性末端互补结合，接口分子的 5′端分别含有一段用于 PCR 扩增的引物 P1 和 P2 的互补片断；④用 ⅡS 型内切酶［又称为标签酶（tagging enzyme），酶切位点一般位于识别后 20bp 处］处理样品，释放带有接头的 SAGE 标签（约 9 ~ 13bp）；⑤利用 DNA 聚合酶（klenow）补平 SAGE 标签的黏性末端，并将这些分别含有接口分子 A 和 B 的片段连接成双标签（ditag）序列；⑥利用PCR 扩增双标签序列，再次利用锚定酶 NIa Ⅲ 切除序列上多余的接口分子，得到尾尾相连的 SAGE 双标签，双标签侧翼有锚定酶切点；⑦去除接头的 SAGE 双标签彼此连接成长短不一的

多联体,电泳分离后收集大小适中的片段克隆到高拷贝的质粒载体中,构建 SAGE 文库(SAGE library);⑧随机挑选 SAGE 文库中的克隆进行测序,并用专门的 SAGE 软件对标签进行分类和计数,生成相应的报告和丰度指标,并与 NCBI 的 SAGEmap、Gene Bank 及 dEST 等公用数据库进行比对,比较 SAGE 文库标签的类型及丰度,分析其意义。

4. 数据分析

对 SAGE 数据的分析主要包括从原始的序列中得到标签列表,比较来自于不同组织细胞或不同生理状态乃至不同物种的标签及其出现频率,在相应数据库中搜索匹配序列,进行基因功能的分析或发现新的基因。

目前用于 SAGE 数据分析的应用软件很多并且在不断发展,如 SAGEmap 公用标签数据库提供的 SAGE300、xProfile 和 Kampen 等提出的可进行标签数据在线分析的 USAGE(http://www.cmbi.kun.nl/usage/)。作为由 NIH 发起的 CGAP 的一部分,SAGEmap 公用标签数据库包含了来源于人类的 47 个标签库的 200 万个标签以及其他生物的 SAGE 资料,而且可将 SAGE 标签和UniGene联系起来,将 SAGE 标签对应于 UniGene 的序列集合以确定一个唯一的转录物。

(二)表达系列标签(EST)

1. 概述

20 世纪 90 年代初 Graig Venter 提出了 EST 的概念,基本思路很简单,是从已建好的 cDNA 库中随机取出一个克隆,从 5′末端或 3′末端对插入的 cDNA 片段进行一轮单向自动测序,所获得的约 60 ~ 500bp 的一段 cDNA 序列,即代表特定转录子的标签。用于 EST 测序的 cDNA 文库分标准化和非标准化两种,标准化文库

所含每种转录子的量基本相同,这就减少了高表达基因的重复测序,提高了稀有转录子的检出;未扩增的非标准化文库的特点是克隆的出现频率准确反映来源细胞或组织中转录子的丰度,常用于确认高表达的未知基因,并比较不同标本中表达的丰度。通过自动序列分析系列分析所得序列,包括 Gene Bank 和 UniGene 数据库中的 Blastn 程序和来自 Swiss - Prot、Gene Bank 和 UniGene 非丰余序列蛋白数据库,再用 PHRAP 序列组合程序将 EST 按相似性分簇,确认虽然未知但在数据中多次出现的序列,最后得到一个代表基因(定量)表达谱的数据库。虽然公共数据库已经很大,强大的 EST 测序仍能发现未见报道或已报道但无功能描述的基因,但是发现多少新基因及数据的统计意义依赖于测序了多少克隆和所分析组织的复杂性。

2. 表达序列标签文库(EST 文库)

由于 EST 来源于 cDNA,因此每一条 EST 均代表了文库建立时所采样品特定发育时期和生理状态下的一个基因的部分序列。EST 数据有不足之处:①ESTs 很短,没有给出完整的表达序列;②低丰度表达基因不易获得;③由于只是一轮测序结果,出错率达 2% ~5%;④有时有载体序列和核外 mRNA 来源的 cDNA 污染或是基因组 DNA 的污染;⑤有时出现镶嵌克隆;⑥序列的冗余,导致所需要处理的数据量很大。

EST 文库可能包含来自同一个 mRNA 的一个或许多拷贝的相同标签,这意味着挖掘一个特定的 EST,可能得到许多标签,这些标签可能代表一个基因。ESTs 聚类将来自同一个基因或同一个转录本的具有重叠部分的 ESTs 整合至单一的簇(cluster)中,UniGene 的出现是为了解决序列表达标签分析具有的重复、冗长的问题而出现的。

EST 文库是最早建立的一类表达谱数据库,所以,收录的数