

表达谱基因芯片*

徐伟文^{1,2)}** 李文全²⁾ 毛裕民¹⁾

(¹⁾复旦大学遗传所遗传工程国家重点实验室, 上海 200433; ²⁾第一军医大学热带医学研究所, 广州 510515)

摘要 表达谱基因芯片具有高通量、缩微、多参数、平行化等优点. 在功能基因组学研究中正发挥越来越重要的作用. 就其原理、应用、存在问题及解决策略等进行了综述.

关键词 表达谱基因芯片, 基因表达谱, 应用

学科分类号 Q503

人类基因组计划 (HGP) 的开展和深入, 产生了海量的基因组序列数据和信息. 如何利用这些研究成果进行基因功能的解析和开发, 已成为功能基因组学研究的重要任务. 基因芯片技术给基因功能研究一个全新的契机. 它自 1989 年由 Southern 提出后, 即受到多方重视和广泛应用, 被评为 21 世纪最有发展前途的 20 项高新技术之一. 表达谱基因芯片是用于基因组功能研究的一种应用型芯片, 正在功能基因组学研究中发挥广泛的作用.

1 原 理

表达谱基因芯片是基因芯片的一种, 其基本构件依然是由 DNA 或寡核苷酸组成的芯片 (chip) 或微阵列 (microarray), 反应原理也依然是核酸杂交, 所不同的是样品标记和检测系统. 通过双色荧光标记, 即对不同细胞或组织来源的 mRNA 在逆转录反应中分别用不同颜色的荧光标记成探针, 探针混合后与芯片或微阵列板上的基因进行严格杂交, 再通过不同波长的荧光扫描芯片, 将扫描所得每一点荧光信号值自动输入计算机并进行信息处理, 给出每个点在不同波长下的荧光强度值及其比值 (ratio 值), 同时计算机还给出直观的显色图. 这些信号就代表了样品中基因的转录表达情况^[1]. 目前常用的荧光染料是 Cy3 (绿色) 和 Cy5 (红色). 如果用 Cy3-dUDP 标记正常组织 mRNA, Cy5-dUDP 标记肿瘤组织 mRNA, 反应后那些在肿瘤组织中呈高表达 (或只在肿瘤组织中表达) 的基因其杂交点就会显红色, 相反, 那些在正常组织中高表达 (或只在正常组织中表达) 的基因其杂交点就会显绿色, 在两组织中表达水平相当的显黄色, 而均不表达的则为无色.

表达谱基因芯片可同时对两组样品中的基因表达水平进行平行检测, 若用同一参照, 可对多组样品间的基因表达情况进行分析比较, 从而可对来源于不同个体 (正常人与患者)、不同组织、不同细胞周期、不同发育阶段、不同分化阶段、不同病变、不同刺激 (包括不同诱导、不同治疗手段) 下细胞内的 mRNA 或逆转录产物 cDNA 进行大规模检测和分析. 因此, 在基因功能研究方面有着广泛的用途^[2,3].

目前, 商品化的表达谱基因芯片中 Affymetrix 公司研制生产的系列产品最具代表性. 如它的 U95 基因芯片, 点载有 760 000 种人类基因和 ESTs 片段, 可检测出单细胞中几个拷贝的 mRNA 转录物, 灵敏度达 1: 100 000, 能精确分辨出表达差异在 2 倍以上的基因, 检测跨度为 1 000. Affymetrix 公司甚至已把所有表达序列标签 (ESTs) 序列 (1 000 000 种探针) 放在 1 cm² 的芯片上, 用于基因表达分析.

2 应 用

2.1 肿瘤分型

这是表达谱芯片目前应用最广、取得进展最快的一个领域^[4]. 精确的肿瘤分型不仅对临床诊断和治疗有指导作用, 对治疗药物的筛选评估也极有意义. 譬如, 对一种抗肿瘤新药进行临床评估, 如果它只对 10% 的肿瘤患者有效, 则肯定会被淘汰. 但如果这 10% 的患者通过精确分型正好属于一种

* 上海市现代生物与新药产业发展基金资助项目 (984319121).

** 通讯联系人.

Tel: 021-65646221, E-mail: xu_sandy@mail.china.com

收稿日期: 2001-03-16, 接受日期: 2001-05-24

新的亚型,而不同于其他 90% 的患者,那么,这种新药应该是一种极有希望被开发的候选药.乳腺癌治疗药物 Herceptin 就是一个明证.

Golub 等^[5]将表达谱基因芯片技术用于急性髓细胞性白血病 (AML) 和急性淋巴细胞性白血病 (ALL) 的分类研究,证明了该技术在肿瘤分型中的有效性和实用性.他们将研究分成两类:一是发掘基因类型,即确定以前未被认识的肿瘤亚型;二是预测肿瘤类型,即将特定的肿瘤标本归于已定义的可以反映当前状况和预后的肿瘤类型中.通过对 27 例 ALL 和 11 例 AML 病人基因表达谱的“neighborhood analysis”(相邻分析法)分析,他们建立了一种根据基因表达水平将 ALL-AML 分类的判别模式.在此基础上,他们选择与 ALL-AML 相关性极高的基因作为信息基因,并赋予不同的权重值,根据这些基因在新病例中的表达水平计算权重之和,以优胜方为诊断类型.测试反应证实了他们的判别模式是有效和准确的,而且这种不依赖于以往的生物学知识对肿瘤进行分型研究,与普通的病理分型相比,具有真实和可重复的优点.

Alizadeh 等^[6]用基因表达谱对弥漫性大 B 细胞淋巴瘤 (diffuse large B-cell lymphoma, DLBCL) 进行了成功的分型. DLBCL 是一种常见的非何杰金氏淋巴瘤,在临床上有着明显差异的转归:40% 的患者对现行的药物有效且有较长的存活期,其他的对治疗不敏感,预后极差.通过基因表达谱分析,Alizadeh 等发现可将该疾病分成两类:一类表达基因具有胚胎生发中心 B 细胞特征,另一类基因表达谱特征与体外正常诱导激活的外周血 B 细胞一致,前者的临床转归明显好于后者.该研究说明基于基因表达谱的肿瘤分子分型可鉴别以往无法细分的类别,并具有更好的临床指导意义.

2.2 细胞调控网络及生化代谢途径研究

个体生长发育、细胞或组织生理功能的发挥,往往是一个或几个功能基因群共同作用的结果.以往的研究只能检测有限的基因,无法做到同期化和平行化,因此无法系统地了解细胞的调控网络和生化代谢途径.表达谱基因芯片可同时检测成千上万个基因,使人们有可能对细胞或组织乃至机体在某一特定时点所有基因表达进行检测,从而明确细胞在某个阶段的调控网络或对某刺激的反应通路.

Holstege 等^[7]用酵母基因组制成的芯片研究了酵母细胞基因组调节周期.应用基因组水平的表达分析,发现 RNA 聚合酶 II、转录因子 TF II D 和

SAGA 染色体修饰复合物等均在基因的表达中有自己特定的作用位点.以此为基础,研究人员进一步绘出了酵母基因组控制图,并分析出各种调节因子的作用位点和分子机制. Tanaka 用小鼠基因的表达谱芯片研究了妊娠中期胎盘与胎儿组织间基因表达的差异,获取的差异表达基因的量是以往报道总数的 5 倍,显示了表达谱基因芯片在组织特异性基因筛选研究中的高效性和全面性.

2.3 疾病发生机制的研究

疾病的发生不是单基因的作用结果.任何生理状况的改变和表型的变化,最初的基础激发点是基因表达的改变.系统监测疾病发生过程中机体或主要靶器官、组织的基因表达变化,有可能揭示疾病发生过程中的多个作用环节,明确疾病的发生机制,同时也能为诊断或寻找治疗靶位提供线索.

Manger 等^[8]认为:表达谱基因芯片通过全面监测病原体感染机体后宿主及病原体基因表达的改变,不仅可回答宿主细胞或器官是如何认识病原体的,而且可揭示在何种水平上宿主能对不同因子进行辨别,以及宿主与病原体之间的交谈、潜在的治疗靶位和疾病转归的预后标志. Dempsey 等则用不同时期心血管病理标本构建多个 cDNA 文库,从中获取 57 000 余个 ESTs.在此基础上,用表达谱芯片技术研究心血管疾病的发生发展机制,为从遗传角度对该类疾病生物特性的研究、疾病相关基因的发现和数据库的开发打下了基础.

2.4 基因功能研究

表达谱基因芯片的另一应用领域是对基因功能的研究.通过全面监测细胞在一定条件下基因的表达并进行聚类分析,根据功能类别对归入其内的未知基因以功能线索的提示.或者,绘制基因在不同组织或条件下表达的特征图谱,并根据与其图谱一致(类似)或成镜像的已知基因的功能,推测未知基因的功能.目前这方面的研究报道也非常多.如 Iyer 等^[9]将成纤维细胞置于无营养的环境中,使绝大部分的基因活性关闭;两天后,加入 10% 的血清,24 h 内分 6 个时间点,用成纤维细胞的 8 600 个基因片段制成的基因芯片观察基因的活化情况.结果表明,在所有被监测的基因中,约有 500 个基因最为活跃,而使细胞保持不分裂状态的基因活性被抑制.其中最早被活化的是那些转录调控基因.在活化的基因中,有 28 个基因共同作用,控制细胞的增殖;8 个与免疫反应的激活有关;19 个与血管重建有关;另有许多基因,与血管新生密切

相关。

2.5 药物作用及毒理学研究

在传统药物研究中,需要对每一种新药进行毒性实验。无论是急性毒性实验还是慢性毒性实验,均需要大量动物,而且慢性毒性实验还需要很长的观察时间,花费了大量的财力与人力,更延缓了一种也许有效的药物进入临床。基因组研究的深入和芯片技术的发展,使人们有可能在实验早期从基因整体水平监测毒性作用的有无,从而缩短药物开发的周期。Pennie 所在的实验室在这方面作了尝试。他们选择了约 600 个与毒性有关的标志基因,对每一基因的生化功能、在疾病中担任的角色及等位变异等信息进行了详细的资料汇集,设计并研制了系列芯片 (ToxBlot) 专门用于毒理学研究^[10]。

表达谱基因芯片也可用于药物作用的研究^[11]。通过用药前后不同时点细胞或组织内基因表达水平的监测,可揭示药物作用通路上的基因及其调控网络,解析药物作用机制。这类研究对药物作用靶位基因的筛选十分有效。根据已知疗效的药物筛选药物作用靶位,进而将同类药物筛选而得的系列靶位基因及其他功能基因(如毒性相关基因、疾病相关基因,等)制成芯片,配以合理的判别模式,用于药物的筛选,是高通量药物筛选及研究开发疗效更显著、毒副作用更小或无的药物的好策略。

2.6 其他

表达谱基因芯片除上述疾病及药物领域的应用外,在其他许多方面均有用武之地,如基因改良食用植物的安全性评估、食用组分的功能分析等^[12]。

3 存在问题及解决策略

表达谱基因芯片综合了分子生物学、半导体微电子、激光、化学等领域的最新科学技术,从出现至今短短几年时间里,就得到了广泛的应用。但任何一种新技术的产生,都有一个不断完善和发展的过程,也不可能一时取代其他相应的技术。在表达谱基因芯片得以广泛应用的同时,一些问题相继暴露并亟待解决。

3.1 加载量及布阵设计

根据研究目的,人们有时希望在同一张芯片上点载有某一生物或与某功能相关的所有基因,这样才有可能真正做到“整体分析”(analysis as a whole, analysis on whole genome)。虽然人类基因组计划开展顺利而且一再提速,30 多种模式生物的全基因组测序工作已经完成,但很多基因的功能依

然是一无所知的。随着基因组计划和后基因组研究的广泛开展和深入,会有越来越多的基因序列信息和功能信息被揭示。应用目前已全序列的模式生物(如酵母、结核分枝杆菌),人们已研制出加载有他们全基因(开放阅读框)的芯片,通过比较不同条件下表达谱的变化,可揭示基因功能和调控网络。在已知基因信息有限的情况下,设计多个功能板块进行布阵,也利于解决一些问题。可以建立表达谱基因芯片与基因功能研究之间的良性循环。另一方面,按目前的芯片点样技术,虽已可以在芯片上加载 2 万多个点,要将人类所有的约 4 万多条基因加载于同一芯片上,还有一定的困难。不少公司正致力于超微型芯片(supermicroarrays)的研制,如 Packard 仪器公司负责开发的压电喷头系统,目前已可将小到 300 pL 的 DNA 样品液滴定位到芯片上。芯片实验室(Lab-on-Chip)研究的开展,更展示了缩微和高通量完美结合的前景^[13]。

3.2 检测域

现行的表达谱基因芯片虽然可同期监测成千上万个基因,但检测灵敏度一般才达 1:100 000,完成一次杂交,所用 mRNA 量起码需从 10^6 细胞中提取,这对于象果蝇或线虫这样的微小生物、检测标志在细胞中稀有表达的情况、或微量的临床病理标本,就很难满足检测的需要。目前有两种方法用于解决小样本量问题。a. 基于 PCR 技术的方法:PCR 技术在理论上可扩增单拷贝基因,所以很自然地人们就用该技术对小样品量样本的遗传物质进行扩增以满足芯片杂交的需要。RT-PCR、巢式 PCR(nested-PCR)、结合加尾反应(homomeric tailing)的通用 PCR、以及根据 3' 端信息全面扩增的 three-primer-end amplification PCR (TPEA),相继登台。但不论何种 PCR,都存在不同程度的信息损失和代表性降低的缺陷。b. 不基于 PCR 技术的方法:该技术不用 PCR 扩增,它对核酸量的扩大基于多轮的“cDNA 合成-互补 RNA 的体外转录(*in vitro* transcription, IVT)”反应。每一轮反应后,RNA 的量就可增加 100~1 000 倍。这样,只需总 RNA 1~50 ng 的样本(单个细胞),通过多轮 cDNA/IVT 扩增反应,就可满足芯片杂交的量,而且复制的忠实性好(相关系数大于 0.97),引入的偏倚也较基于 PCR 扩增的方法小。另外还有些提高检测域的方法,如引入高效的 T7 RNA 聚合酶扩增 mRNA、杂交后用“夹心”酶联反应进行靶效应的放大、用质谱进行杂交检测等。

Affymetrix 公司通过对芯片灵敏度及特异性影响因素的系列研究,发现探针长为 25 个碱基时,有最佳的完全匹配 (perfect match, PM) 和错配 (mismatch, MM) 之间的平衡,使灵敏度提高至 1: 300 000, 特异性也最佳(www.affymetrix.com/, Affymetrix GeneChip Technical Note)^[14~16].

3.3 质控及标准化

对任何一项检测技术,人们总希望具有高灵敏度和特异性,而把假阳性率和假阴性率控制到最低.据研究报道,目前的表达谱基因芯片,其假阳性率< 2%,对同源性达 70%~80% 的基因能进行有效的鉴别.在最近一期《Nature》上, Knight 则发布调查消息说,由于人为的误差及系统误差,芯片的错误率可高达 30%,因此提醒研究者对芯片结果的解释要慎重.另一方面,随着芯片的广泛使用和大量表达谱数据的获得,对表达谱数据的公开化及加快公共数据库建设的呼声越来越高,NCBI 的 Gene Expression Omnibus (GEO) 和 EBI 的 ArrayExpress 等数据库中的数据正日渐增加.合理比较分析来自不同实验室的表达谱数据,充分利用数据库资源进行后基因组学研究,质控和标化问题就显得十分重要.质控和标化贯穿芯片设计、制作和使用的全过程.在设计中, Affymetrix 采用一组探针 (16~40 个) 作为每个靶基因的质控对照,每个探针之间只相差一个碱基,以精确判别信噪比.对用于点样的 PCR 产物进行重复测序及纯度测试与定量、用不同方法检测点样的均一性及探针标记的有效性、标准曲线的绘制、激光交互反应的测试及 RNA 完整性的检测等,均是质控所需要考虑的,也是芯片数据可靠与否的前提.在芯片使用中,标准化还涉及样品的选择、提取、标记及杂交检测等实验过程^[17,18].

3.4 数据分析

和其他基因组或后基因组研究类似,表达谱基因芯片在研究中也面临超大量数据的处理和分析.如何解读芯片上成千上万个基因点的杂交信息,将无机与信息数据与有机的生命活动联系起来,阐释生命特征和规律以及基因的功能,是生物信息学研究的重大课题.目前,基于电荷偶合装置 (charge-coupled device, CCD) 技术或光电倍增管 (photomultiplier tube, PMT) 检测系统的荧光扫描仪,可精确识别并给出每一点的荧光图像,生物信息学也已提供一些分析软件对复杂的扫描图像及庞大的信号数据进行处理分析,如各类聚类分析软

件、自我管理图像分析系统 (self-organizing maps, SOM)、Singular value decomposition (SVD) 等.生物学家更提出“knowledge discovery in databases (KDD)”方案.在该方案中,对电子解码的有意义或有用的关联,专家和计算机将对其有效性、新颖性、实用性、简单性等进行进一步地发掘、验证和评估.但各类分析技术还有待改进和完善,专用于表达谱数据分析的软件也有待开发.数据的可靠性、检测基因的代表性及标准化问题,也是数据分析首先要考虑和解决的^[19,20].

参 考 文 献

- 1 Brown P O, Botstein D. Exploring the new world of the genome with DNA microarrays. *Nature Genetic (suppl)*, 1999, **21** (1): 33~37
- 2 Xiang C C, Chen Y. cDNA microarray technology and its application. *Biotechnology Advances*, 2000, **18** (1): 35~46
- 3 Schena M, Heller R A, Thériault T P, *et al.* Microarrays: biotechnology's discovery platform for functional genomics. *Trends Biotechnol*, 1998, **16** (7): 301~306
- 4 Perou C M, Brown P O, Botstein D. Tumor classification using gene expression patterns from DNA microarrays. *Mol Med Today*, 2000, 67~76
- 5 Golub T R, Slonim D K, Tamayo P, *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 1999, **286** (5439): 531~537
- 6 Alizadeh A A, Eisen M B, Davis R E, *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 2000, **403** (6769): 503~511
- 7 Holstege F C, Jennings E G, Wyrick J J, *et al.* Dissecting the regulatory circuitry of a eukaryotic genome. *Cell*, 1998, **95** (5): 717~728
- 8 Manger I D, Relman D A. How the host 'see' pathogens: global gene expression responses to infection. *Cur Opin Immunol*, 2000, **12** (2): 215~218
- 9 Iyer V R, Eisen M B, Ross D T, *et al.* The transcriptional program in the response of human fibroblast to serum. *Science*, 1999, **283** (5398): 83~87
- 10 Pennie W D. Use of DNA microarrays to probe and understand the toxicological consequences of altered gene expression. *Toxicology Letters*, 2000, 112~477
- 11 Debouck C, Goodfellow P N. DNA microarrays in drug discovery and development. *Nature Genetic (suppl)*, 1999, **21** (1): 48~50
- 12 van Hal N L, Vorst O, van Houwelingen A M, *et al.* The application of DNA microarrays in gene expression analysis. *J Biotechnol*, 2000, **78** (3): 271~280
- 13 Lipshutz R J, Fodor S P A, Gingeras T R, *et al.* High density synthetic oligonucleotide arrays. *Nature Genetics (suppl)*, 1999, **21** (1): 20~24
- 14 Dixon A K, Richardson P J, Pinnock R D, *et al.* Gene expression analysis at the single-cell level. *TiPS*, 2000, **21** (2): 65~70
- 15 Lockhart D J, Winzler E A. Genomics, gene expression and DNA arrays. *Nature*, 2000, **405** (6788): 827~836
- 16 Duggan D J, Bittner M, Chen Y D, *et al.* Expression profiling using cDNA microarrays. *Nature Genetics (suppl)*, 1999, **21** (1): 10~14

- 17 van Berkum N L, Holstege F C. DNA microarrays: raising the profile. *Curr Opin Biotechnol*, 2001, **12** (1): 48~ 52
- 18 Celis J E, Kruhøffer M, Gromova I, *et al.* Gene expression profiling: monitoring transcription and translation products using DNA microarrays and proteomics. *FEBS Letters*, 2000, **480** (1): 2~ 16
- 19 Brazma A, Vilo J. Gene expression data analysis. *FEBS Letters*, 2000, **480** (1): 17~ 24
- 20 Raychaudhuri S, Sutphin P D, Chang J T, *et al.* Basic microarray analysis: grouping and feature reduction. *Trends in Biotechnology*, 2001, **19** (5): 189~ 192

DNA Microarrays for Gene Expression Profiles^{*}

XU Wei Wen^{1,2)}**, LI Wen-Quan²⁾, MAO Yu-Min¹⁾

⁽¹⁾State Key Laboratory of Genetic Engineering, Institute of Genetics, Fudan University, Shanghai 200433, China;

⁽²⁾Institute of Tropical Medicine, The First Military Medical University, Guangzhou 510515, China)

Abstract DNA microarray for gene expression profiles is a high throughput technique. It allows expression monitoring of tens of thousands genes in parallel. It has become an increasingly popular tool to investigate the function of genes. Here the principle and applications of this technique were reviewed. The problems it faced with and the resolve strategies are also mentioned.

Key words DNA microarrays for gene expression profiles, gene expression profiles application

^{*} This work was supported by a grant from Shanghai Modern Molecular & New Drug Industrial Development Fund (984319121).

^{**} Corresponding author. Tel: 86-21-65646221, E-mail: xu_sandy@mail.china.com

Received: March 16, 2001 Accepted: May 24, 2001