综述

文章编号:1000-5404(2006)01-0080-03

基因芯片可靠性分析及数据处理

高利宏,曹 佳 (第三军医大学军事预防医学院卫生毒理学教研室,重庆 400038)

基因芯片(gene chip)又称为DNA微阵列(DNA microarray),其基本原理是将众多的靶基因序列或寡聚核苷酸片段有序而高密度地排列在玻璃、硅、尼龙膜等固相载体上,用待检测的标记样本分子与之杂交,并利用激光共聚焦显微扫描等技术对芯片上成千上万的杂交信号进行实时、灵敏而准确的检测,辅以计算机统计分析从而得到样本的基因表达信息。

自1992年美国Affymitrix公司制备出世界上第1张寡聚核苷酸生物芯片至今,仅短短10余年时间,基因芯片技术已广泛运用于生物学与医学的基础研究、疾病诊断、新药开发、环境保护等许多方面。在MEDLINE上以genechip和microarray为关键词检索,2000年以前仅数百篇相关文献,而到2004年已达4 000余篇文献。该技术最大的优点在于具备高通量平行检测的特点,能在更全面广泛的基因组水平上揭示不同基因之间内在的相互关系,使研究效率明显提高,并极大地降低了基因表达检测的平均成本,这对于继人类基因计划(human genome project, HGP)实施以来呈几何级数增加的基因序列资源的利用有至关重要的意义。众多学者认为,该项技术是继DNA重组、PCR技术以后又一项生物科学领域的革命性进展,将对21世纪生命科学的研究思维和方式产生重要影响。

但不可否认,作为一项新兴的技术,基因芯片也是处于不断探索和完善的过程中。就技术本身而言,可以把它形容为"强大但不完美"——基因芯片实验从前期设计到最后的数据提呈目前都还存在着一些技术难点和缺点,在应用过程中也非全能,而存在一定的局限性,需要人们对其有更全面和深刻的认识。目前世界各地芯片研究人员正致力于其方法学的改进、实验流程的规范以及共同实验标准的建立。现主要对芯片技术的可靠性和数据处理两方面进行综述。

1 芯片技术的可靠性分析

认识芯片技术的可靠性,首先应该了解该技术的定位。一般来说,基于芯片高通量和高灵敏的检测特点,在成千上万探针数量的基因表达谱水平上研究特定时间组织的多基因表达程度,看似既多又好,但这实际上是一把双刃剑。因为在实际情况中,动物组织、细胞的基因表达谱并非"均质性",不同个体或细胞株之间存在着一定的变异,而且表达谱对实验条件的变动也高度易感,在RNA抽提、cDNA逆转录等操作步骤上的微小差别,往往都能引起不同基因的表达改变,这样使微阵列检测所呈现的大量"阳性改变"基因中可能相当一部分难以判断确定的生物学意义。另一方面,微阵列实验是成百上千个核酸探针进行"同时杂交",能影响芯片探针和实验样本中相应的靶片段的结合效率的因素可来自多方面——序列自身因素(序列长度、碱基比例、互补性等)、杂交实验条件(探针和靶序列浓度、阳离子浓度、pH等)及非特异性杂交片段等。以目前的芯片制作工艺,还难以保证每个探针杂交反应能同时具备最佳反应条件,因此实际的芯片探针杂交往往会产生大量的假阳性和假阴性。对单个基因测定的精度而言,高通量性的微阵列并不如低通量的Northern、real-time PCR等技术更精确。还应该注意到,微阵列技术的另一大应用限制是其测定的mRNA水平的改变仅属于基因表达的中间产物,并非功能蛋白,还远不能直接解释在细胞和组织水平上主要由功能蛋白参与的多种生理、病理变化的机制。而且,即使是在微阵列表达谱上表现出明显变化的不同基因,它们之间的因果关系如何,单靠微阵列技术本身也不能单独进行确切判断,需要进一步利用相应实验技术进一步研究[1,2]。所以在目前阶段,DNA微阵列应主要定位于一种运用高通量手段在特定实验条件下观察基因组的整体性变化,以利于从纷繁的表达谱数据中寻找有效线索展开后续深入研究的定性实验。当然我们也应该看到,微阵列观察大批表达谱的特定的mRNA表达类型比传统分子生物学技术所检测的单个或少数基因表达更能全面地反映和预测相应的生物学机制,在蛋白质组技术普及应用以前,DNA微阵列技术还是在基因组水平研究基因表达最有效的方法。

一般来说,具备良好的灵敏度和较高特异度的核酸探针是保证芯片杂交数据可靠性的一个关键因素。目前,DNA微阵列根据探针的选择主要有两种类型——cDNA微阵列和寡聚核苷酸点阵芯片。

cDNA微阵列的探针一般来自于cDNA或EST文库的PCR扩增产物,长度为100~2 000 bp。这一类微阵列因其长链探针灵敏性高、造价低、技术平台易于获得而被许多实验室广泛使用。但就探针构建而言存在以下一些不可靠因素:①cDNA探针针对来自同一基因家族不同成员的靶序列的检测时特异性不高,在序列选择时需注意尽量挑选文库中带3'-末端的克隆片段以增强杂交特异性[3];②以不同荧光染料标记的探针进行杂交时存在标记效率不平衡的现象,常常需要试验组和对照组样本进行荧光交换标记重复试验;③双链cDNA探针相对于单链

基金项目:国家杰出人才基金资助项目(3012503),国家自然科学基金资助项目(30200354,30271136)
Supported by the National Science Fund for Distinguished Scholar (3012503) and the National Natural Science Foundation of China (30200354,30271136)

作者简介:高利宏(1971-),男,四川省自贡市人,博士研究生,讲师,中级职称,主要从事药理基因组学方面的研究。电话:(023)68752589

收稿日期:2005-04-12;修回日期:2005-07-24

寡聚核苷酸更容易产生交联而影响杂交；④大规模的 cDNA 克隆文库往往存在克隆交叉污染，影响 cDNA 微阵列探针质量[4]。

寡聚核苷酸点阵芯片的探针由 20～80 bp 的短链寡聚核苷酸构成，研究人员可以从 GenBank、EMBL 等的核酸序列数据库及 EST 数据库（dbEST）寻找感兴趣的序列数据作为参考直接合成，不需要如 cDNA 微阵列一样准备 cDNA 克隆文库。为提高短链寡聚核苷酸探针的特异鉴别能力和容错能力，通常此类芯片会针对每一个目标序列的不同区域设计多组“冗余”探针，提高杂交信噪比，并且如 Affymetrix 公司的 GeneChip 专利设计以一组完全匹配（perfect match，PM）及中间有一错误位点匹配（mismatch，MM）探针区分特异性结合与非特异性结合的靶片段。这样，使寡聚核苷酸点阵芯片的特异性可以达到能区分大部分基因家族成员序列的水平，除可进行表达谱研究外，还可以用于检测基因突变和多态性[5]。但此类芯片因制作成本相对较高，不如 cDNA 微阵列使用广泛。

另外需要注意的一个问题是，虽然根据目前的制作工艺，每张芯片高密度集成寡聚核苷酸探针可达 100 000 个以上，但探针数量水平应该根据实验目来选择。使用高密度探针的芯片有利于进行基因组筛选和基因表达谱全面性观察，以期对研究对象有一个更全面的认识或者发现新的线索——一种基于“假想发生”的“fish”科学[6]。但如果研究人员期望能更有针对性地研究某类代谢途径，或能更有效地对芯片数据结果进行统计分析处理，选择低密度的“focused”芯片往往可靠性更好，也更为经济[7]。

芯片的系统误差存在于实验的全过程，包括如前所述的样本的生物差异、样本荧光双色标记偏差、以及探针杂交和洗脱条件不一致、信号采集误差等等，因此在芯片实验的设计上，重复性是保证数据可靠性的一个必要原则。一般而言，芯片的重复性设计包括 3 个层次：①芯片内同一基因探针多次点样；②同一份 RNA 样本进行多次芯片杂交；③用来自同一品系或类型的不同生物学个体进行重复实验。前两种类型的重复性试验主要针对的是芯片实验的技术性误差，有利于提高芯片数据的精确度，而后一种类型还包含了样本的生物性变异，使实验的重复性结果更具有广泛的生物和统计学意义[8]。在样本的重复次数上存在着一定争论，Lee 等[9]认为 3 次比较适宜，Tilstone[10]认为重复次数的多少应该与想获得怎样的结果密切相关。而事实上，由于芯片昂贵的成本，很多实验室在具体的实验研究中却往往难以满足统计学分析所要求的重复例数，一些所发表的芯片实验论文也缺乏标准的统计学框架。在对芯片重复性进行了研究的文献中，Bartosiewicz 等[11]报道芯片内点间误差为 10%，片间误差 14%；Yue 等[12]报道片内误差为 5%～10%，片间为 10%～30%，由此看出使用芯片技术获得的数据结果仍然存在较大的波动。

2　芯片数据处理、分析和提呈

虽然芯片技术的优势是可以大范围高通量研究基因表达，但如果在完成杂交实验以后，如早期一些文献，数据分析仅停留于简单人工处理的原始数据列表上，大量有价值的信息和线索就会被湮没和浪费。因此生物信息学和统计学的有效参与，对芯片资料进行合理修正、规范分析、信息挖掘和提呈，往往是芯片实验研究中更为重要的部分，也是提高芯片数据可靠性的必要保证。

归纳起来，芯片数据的整理和分析一般有以下的步骤。

2.1　芯片扫描提取杂交信号

用于芯片荧光扫描的扫描仪目前主要有两类：基于光电倍增管（photomultiplier tube，PMT）的激光共聚焦显微镜检测系统；基于电荷偶连装置（charge-coupled devices，CCD）摄像原理检测光子。

2.2　芯片信号的背景扣减

其意义在于减除芯片杂交信号中属于非特异性的背景噪音部分。以前多以图像处理软件划格后每个杂交点周围区域的背景平均值来计算，但存在芯片不同区域背景扣减不均匀的缺点，也可利用芯片最低信号强度的点（代表非特异性的样本与探针结合值）或综合整个芯片非杂交点背景所得平均值做为背景扣减[13]。

2.3　数据校正

以求校正样本提取、荧光标记及芯片杂交等过程中的系统误差对实验数据的影响。很多文献采用“稳定”的看家基因的表达比率作为校正标准，但目前认为它们在不同实验条件下同样存在改变。为克服此缺陷，有研究人员改用平衡对照和实验点的芯片整体信号强度整体平均值或中位数做标准，以及 Lowess 密度依赖性校正方法[14]。

2.4　数据分析

通过以上计算机的图像分析和标准化处理，我们得到代表芯片上每个基因信号强度数值的电子数据表，下一步工作是如何在其中挖掘寻找众多基因在表达上的差异性和相似性规律，进而发现其所代表的生物学意义。分析微阵列上基因的差异表达，很多文献都采用根据处理和对照组相应基因的信号比率，用人为界定的阈值确定之——Ratio 分析（Ratio Analysis）（多为处理/对照 <0.5 作为基因显著下调标准，处理/对照 >2 作为基因显著上调标准）。该方法简单直观，但其阈值的划分主观性较强、缺乏生物学和统计学支持，尤其对于分析样本中的低拷贝或高拷贝转录子，容易产生假阳性和假阴性问题。Ideker 等[15]阐述了通过重复性实验中的误差评估（variability and Error Ass essment）和基因定标后特异性 t 检验（significance analysis of microarrays）来判断差异表达基因的方法，以及其他一些报道的方案[16]，但目前尚无统一性标准，芯片后验证性实验（RT -PCR、荧光定量 RT-PCR、Northern 等）是确定样本基因差异表达的金标准。寻找基因表达水平的相似性规律时则常用聚类统计分析对基因表达谱数据统计归类，探索代表不同生物学意义的分类标准、同类基因的共同功能以及在基因表达水平上预测新的生物模式等。主要策略有监督分析和非监督分析两类，前者根据特定样本或基因的已知生物学信息对表达谱建立分类器，进而对各基因进行功能分类和预测，后者则通过计算和比较表达谱各基因统计学距离，聚类“相似性”样本或基因。代表性的数学模型有层次聚类（hierarchical clustering）、自组织作图（self-organizing maps）、k-means、主元分析方法（principle component analysis）、LDA（linear discriminant analysis）

等[17]，不过 Zhou 等[18]认为，一些相似功能的基因并不总是表现相似的表达谱，针对此他们提出了“过度共表达基因”概念及相应的数学模型鉴定表达谱中此类基因，作为聚类分析的补充。在 Internet 网络上许多商业和学术机构所提供的大量芯片数据统计分析软件包等资源可供研究人员参考使用（http://www.lab-on-a-chip.com/suppliers/inform.html 及 http://genome-www5.stanford.edu/restech.shtml 等）。

2.5 芯片数据的管理和交流

进行芯片的数据分析以后并不标志着实验的结束，研究人员逐渐认识到，要对呈数量级增长的实验数据进行有效管理、交流和验证，需要建立起通行的数据储存和交流平台，以及一套科学的策略和统一的标准化管理方案。Brazma 的研究小组在 2001 年提出记录和报告芯片实验数据的建议标准——MIAME（minimum information about a microarray experiment 最小化阵列表达信息）[19]，该方案主要从 6 个方面对芯片实验的描述进行了规划：整体实验规划和设计、芯片阵列的设计、样本收集提取和标记的方案、芯片杂交的流程和参数、影像数据的测量和规范、数据标准化校正分析，以期统一芯片报告的格式和整合相关资讯。至今为止，MIAME 策略已得到较为广泛的响应、认同和发展，尤其以学术界和商界组成的微阵列基因表达数据（MGED）协会加快了其应用普及，一些公共的生物芯片信息数据库如 EBI 的 ArrayExpress、NCBI 的 GEO、日本的 CIBEX 等均采用 MIAME 标准接纳芯片数据 *Nature*、*The Lancet* 等一些专业杂志已经把 MIAME 格式作为接受芯片研究论文的必要条件（http://www.nature.com/nature/submit/policies/），许多著名的芯片及软件生产商，如 Affymetrix 公司、Rosetta Biosoftware 公司、Iobion Informatics 公司等也纷纷将 MIAME 标准整合到相关产品中。从 MIAME 建立的意义上来说，它是一个随着芯片技术研究进展而不断发展的指导策略，而并非固定教条（2001 年文献提出的是 1.0 版本，2002 年已改进为 1.1 版本）。关于 MIAME 较为近期和深入的讨论与进展可以查阅 http://www.mged.org/Workgroups/MIAME/miame.html。

总之，在人类基因组计划初步完成，开始进入后基因组时代的今天，作为高通量筛选技术的代表——基因芯片存在着巨大的科研需求，将可能成为未来生命科学研究的主要分析手段之一。因此，寻找芯片实验技术上的突破，有效提高其检测效能和可靠性，并对微阵列所提供的巨大数据进行有效的分析和管理，是基因芯片技术能否向前突破的关键。相信在世界各地研究人员的共同努力下，基因芯片所存在的技术性缺陷和壁垒将被逐一克服，这项具有时代意义的技术会很快成熟起来。

关键词：基因芯片；微阵列；可靠性；数据处理

中图法分类号：R318.04　　**文献标识码**：A

参考文献：

[1] VRANA K E, FREEMAN W M, ASCHNER M. Use of microarray technologies in toxicology research [J]. Neurotoxicology, 2003, 24(3): 321-332.

[2] ANDERSON N L, MATHESON A D, STEINER S. Proteomics: applications in basic and applied biology [J]. Curr Opin Biotechnol, 2000, 11(4): 408-412.

[3] HUANG J, LIH C J, PAN K H, *et al*. Global analysis of growth phase responsive gene expression and regulation of antibiotic biosynthetic pathways in Streptomyces coelicolor using DNA microarrays [J]. Genes Dev, 2001, 15(23): 3183-3192.

[4] HALGREN R G, FIELDEN M R, FONG C J, *et al*. Assessment of clone identity and sequence fidelity for 1189 IMAGE cDNA clones [J]. Nucleic Acids Res, 2001, 29(2): 582-588.

[5] LIPSHUTZ R J, FODOR S P, GINGERAS T R, *et al*. High density synthetic oligonucleotide arrays[J]. Nat Genet, 1999, 21(1 Suppl): 20-24.

[6] SHIODA T. Application of DNA microarray to toxicological research [J]. J Environ Pathol Toxicol Oncol, 2004, 23(1): 13-31.

[7] STEARS R L, MARTINSKY T, SCHENA M. Trends in microarray analysis [J]. Nat Med, 2003, 1(9): 140-145.

[8] YANG Y H, SPEED T. Design issues for cDNA microarray experiments [J]. Nat Rev Genet, 2002, 3(8): 579-588.

[9] LEE M L, KUO F C, WHITMORE G A, *et al*. Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations[J]. Proc Natl Acad Sci U S A, 2000, 97(18): 9834-9839.

[10] TILSTONE C. DNA microarrays: vital statistics [J]. Nature, 2003, 424(6949): 610-612.

[11] BARTOSIEWICZ M, TROUNSTINE M, BARKER D, *et al*. Development of a toxicological gene array and quantitative assessment of this technology [J]. Arch Biochem Biophys, 2000, 376(1): 66-73.

[12] YUE H, EASTMAN P S, WANG B B, *et al*. An evaluation of the performance of cDNA microarrays for detecting changes in global mRNA expression [J]. Nucleic Acids Res, 2001, 29(8): E41.

[13] BROWN C S, GOODWIN P C, SORGOR P K. Image metrics in the statistical analysis of DNA microarray data. [J]. Proc Natl Acad Sci U S A, 2001, 98(16): 8944-8949.

[14] YANG Y H, DUDOIT S, LUU P, *et al*. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation [J]. Nucleic Acids Res, 2002, 30(4): E15.

[15] IDEKER T, THORSSON V, SIEGEL A F, *et al*. Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data [J]. J Comput Biol, 2000, 7(6): 805-817.

[16] DRAGHICI S. Statistical intelligence: effective analysis of highdensity microarray data [J]. Drug Discov Today, 2002, 7(11 Suppl): S55-S63.

[17] ALON U, BARKAI N, NOTTERMAN D A, *et al*. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays [J]. Proc Natl Acad Sci U S A, 1999, 96(12): 6745-6750.

[18] ZHOU X, KAO M C, WONG W H. Transitive functional annotation by shortest-path analysis of gene expression data [J]. Proc Natl Acad Sci U S A, 2002, 99(20): 12783-12788.

[19] BRAZMA A, HINGAMP P, QUACKENBUSH J, *et al*. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data [J]. Nat Genet, 2001, 29(4): 365-371.

（编辑　陈聪连）