有关芯片数据写作方面的基本思路

生物芯片北京国家工程研究中心微阵列服务部高级主任　张亮博士 2007年5月

随着生物芯片技术在广大科研工作人员研究工作中的应用，在写作论文的时候，如何比较科学、直观地体现生物芯片实验的结果，是一些初步尝试利用芯片数据来写作论文的科研工作人员希望了解的环节。现就笔者在生物芯片技术领域积累的一些经验，做一个初步的总结。但由于生物技术发展很快，有很多内容可能遗漏，还请读者能及时指出不足之处。

1.挑选差异表达的基因

利用生物芯片进行研究的第一步，往往是要找到差异表达的基因。选择差异表达的基因，笔者建议采用生物芯片为科研工具的研究人员使用 SAM 软件（Significant Analysis of Microarray），它是由Standford大学开发的一个免费软件，目前广泛地被学术界所采用，进行挑选差异基因。SAM 软件可以作为插件在Office Excel 软件中进行应用，很容易被生物医学工作者掌握。SAM 软件进行分析的一个基本前提就是需要至少 3 次实验以上的重复。这里的重复可以是生物材料的重复，例如某种疾病包含多个病人；也可以是实验的重复，例如药物处理细胞做了 4 次实验。通过重复实验，才能从统计学意义上判断差异变化的基因。可以理解 SAM 软件和统计学 t-test 检验有类似之处。笔者从合作单位被编辑退回的稿件中了解到，有很多退稿是因为没有进行重复实验，例如对照和处理各一个样本，然后认为荧光信号值差异在 2 倍以上的基因就是差异的基因。审稿编辑的意见往往是需要加上重复实验进行统计分析。

举一个例子，要研究某种疾病 A 的人群和疾病 B 的人群血液中有核细胞基因表达的差异（疾病 A 和疾病 B 人群分别至少要有 3 个人以上）。若是使用了单通道的表达谱芯片，例如 Affymetrix 的芯片，你可能得到例如表 1 这样的数据：

表 1. 单通道芯片设计的例子（以信号值进行计算）

样 本 基因	病 A1	病 A2	病 A3 …	病 B 1	病 B 2	病 B 3…
NM_001192	122	453	278	1345	2315	1954
NM_004836	4566	3567	5632	5643	5689	6112
AK025431	11831	13432	12543	24231	21998	19888

在 SAM 软件进行差异基因筛选的时候，这种例子可以选择 two class unpaired（2 因素不配对）的算法。因为实验研究的就是 2 组样本，并且疾病 A 和疾病 B 的病人之间没有一一配对的关系。那么在进行 SAM 软件运算前，需要加一行样本识别标记（可参见表 2），让 SAM 程序知道哪些信号值是来自同一组病人的。

表 2. 单通道芯片数据在 SAM 软件中的格式

样 本 基因	病 A1	病 A2	病 A3 …	病 B 1	病 B 2	病 B 3…
	1	1	1 … …	2	2	2
NM_001192	122	453	278	1345	2315	1954
NM_004836	4566	3567	5632	5643	5689	6112
AK025431	11831	13432	12543	24231	21998	19888

另外加入的一行蓝色字体就是样本识别标记，这样 SAM 软件通过格式上的默认，就知道哪些数据是同一组病人的不同重复。目前发现单通道芯片有这样一种可能的缺陷：若一个实验进行的时间很长，例如 1-2 年以上，那么进行芯片实验的试剂之间有较大的差别，有时芯片杂交的信号强度差异并不是生物样品的差异，而是试剂不同带来的差异。

如果是使用双通道芯片，笔者不建议某个病人 A 个体和某个病人 B 个体的 RNA 混合在一起和一张芯片做杂交，因为你并不知哪两个病人应该配对。笔者建议取一个共同的参照物。选取共同参照物的基本要求就是比较容易得到该共同参照物样本，并且和所研究的因素没有关系。在下面这个例子中，可以取若干个

正常人血液有核细胞的 RNA 混合物做一个共同参照，也可以购买美国 Stratagene 公司的 Universal Human Reference RNA sample 作为一个共同参照物，这样得到的数据就将是一个比值（可参见表 3）：

表 3. 使用共同参照物实验设计的双通道芯片设计的例子（以比值进行计算）

样本 基因	病 A1/CK	病 A2/CK	病 A3/CK …	病 B1/CK	病 B2/CK	病 B3/CK…
	1	1	1 … …	2	2	2
NM_001192	0.34	0.28	0.35	1.12	1.43	1.22
NM_004836	4.44	3.67	5.65	5.66	3.54	6.43
AK025431	1.22	0.98	1.19	3.42	2.46	2.89

以上同样可以用 SAM 软件中 two class unpaired 的方法来计算疾病 A 组病人和疾病 B 组病人之间差异的基因。用比值进行计算的优点在于，各种试剂、操作产生的差异在比值中被消除掉了。

因此，在使用单通道芯片时，若需要在时间间隔比较长远的数据之间进行比较，目前存在一种趋势，即在某段时间内进行的单通道芯片实验，安排做一张共同参照物 RNA 的芯片，然后得到比值；经过较长时间以后再做芯片实验时，同时再安排一张共同参照物 RNA 的芯片，然后又得到比值，最后对不同时间段之间的比值进行比较。

另外一种常见的实验就是对动物或者细胞进行药物处理。例如选择一个细胞用药物处理后，观测药物处理引起的基因表达变化。实验重复了 3 次。这种实验，除了采用上述单通道芯片试验设计以及利用一个共同参照物来做双通道芯片的试验设计外，还可以把每次实验的处理和对照样品用不同的荧光素标记和一张芯片进行杂交。这样就得到一个纯粹的比值，数据格式如下：

表 4. 对照和处理同时杂交芯片实验设计得到的芯片数据格式双通道芯片数据

样 本 基因	处理 1/对照 1	处理 2/对照 2	处理 3/对照 3 …
NM_001192	0.23	0.12	0.15
NM_004836	3.55	4.24	3.56
AK025431	1.22	0.89	0.96

在 SAM 软件进行差异基因筛选的时候，这种例子可以选择 one class 的算法，大致的意思就是判断是否和比值=1 是否有显著性差别。

此时也需要另外在表格中插入一行，让 SAM 软件处理的时候知道只有一个因素。

表 5.对照与处理同时和芯片进行杂交的双通道芯片数据在 SAM 软件中的格式

基 因 样本	处理 1/对照 1	处理 2/对照 2	处理 3/对照 3 … …
	1	1	1
NM_001192	0.23	0.12	0.15
NM_004836	3.55	4.24	3.56
AK025431	1.22	0.89	0.96

由于在 SAM 软件进行差异基因选择时，可以通过调节参数来改变差异基因的数目。在文章中写作时，可以根据基因变化的倍数来选择变化的基因，并列出其他的一些参数，例如 False Discover Rate (FDR)，或者再加上 Local FDR，具体例子可参见图 2。

另外，对于有多因素分析的实验，例如比较多个组织，然后寻找在某个组织中特异表达的基因，也可以利用 SAM 软件中的 Multiclass 算法。

由于 SAM 软件通常需要根据需求或结果来调整参数，因此笔者认为利用芯片数据写作的研究人员最好能学会使用 SAM 软件。

2. 在论文写作中如何展示所得到的差异基因

首先可以考虑把 SAM 分析结果中的散点图展示出来，参见图 1。从这个散点图中，能大致看出变化基因在全部基因中所占的比率，以及上调和下调基因的数目情况（红色表示上调，绿色表示下调）。当然在图上还可以加上一些 SAM 分析的参数，例如 Significant, Median number of false positives, False Discovery Rate (%)等等。

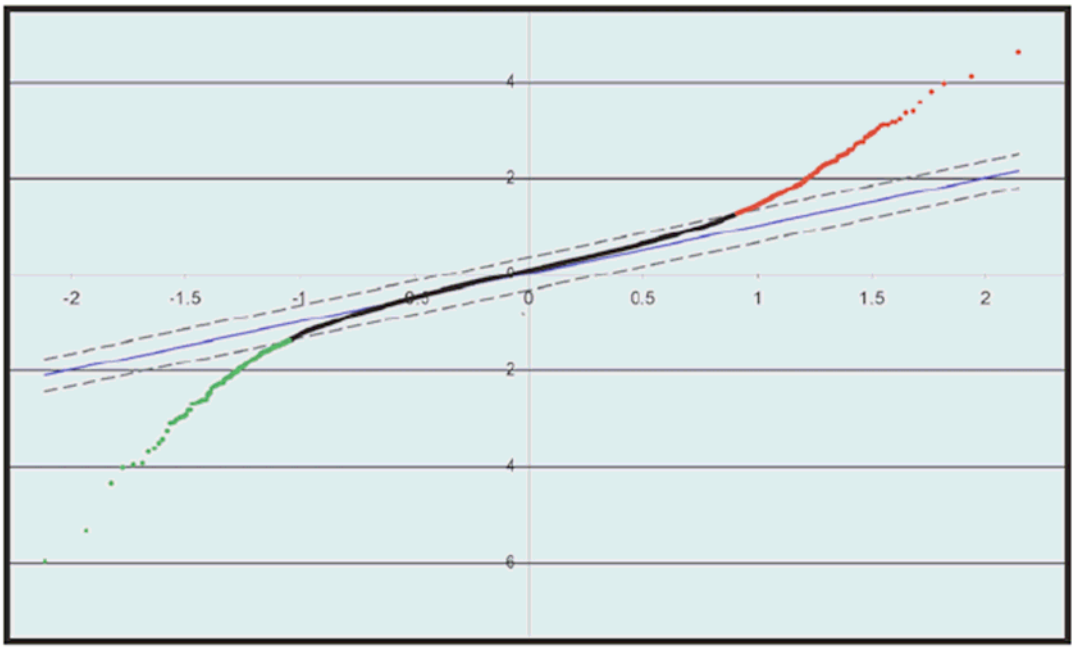

图 1. 在芯片结果中，用 SAM 软件分析得到的散点图展示变化的基因分布，数据引自 Forsberg,E.C., et al., PLoS Genetics, 2005,1(3):e28.

当突出的变化基因数目不多时，可以用列表方式来展示变化基因的具体信息，在表格中，除了列出变化的倍数，还可以加上一些 SAM 软件得到的统计学数据，参见图 2。

Table 1. Differentially expressed miRNAs in PTC tumors

MiRNA	Fold	Localfdr, %
hsa-mir-146	19.3	4.4
hsa-mir-221	12.3	3.2
hsa-mir-222*	10.9	2.7
hsa-mir-21*	4.3	6.7
hsa-mir-220	4.0	8.6
hsa-mir-181a	2.6	6.9
hsa-mir-181c	2.4	7.6
hsa-mir-181*	2.2	7.2
hsa-mir-155	2.2	9.1
hsa-mir-213	1.9	7.3
hsa-mir-34a	1.8	21.9
hsa-mir-24-2	1.7	10.1
hsa-mir-29a-2	1.7	21.7
hsa-mir-29b	1.7	23.7
hsa-mir-29c	1.6	21.6
hsa-mir-102	1.6	11.2
hsa-mir-24-1	1.5	14.5
hsa-mir-9-3	0.7	7.3
hsa-mir-219-1	0.6	5.7
hsa-mir-138-1	0.6	5.0
hsa-mir-138-2	0.6	4.9
hsa-mir-345	0.6	4.7
hsa-mir-26a-1	0.6	5.0

图 2. 在芯片结果展示时，通过列表来展示 SAM 分析得到的差异表达基因，数据引自 He et al., PNAS, 2005, 102:19075-80.

若变化的基因数目很多，列表将占很大的篇幅，这时可以用聚类分析的图表展示变化的基因，并且在聚类图上标明研究者想突出的基因名称，参见图 3。

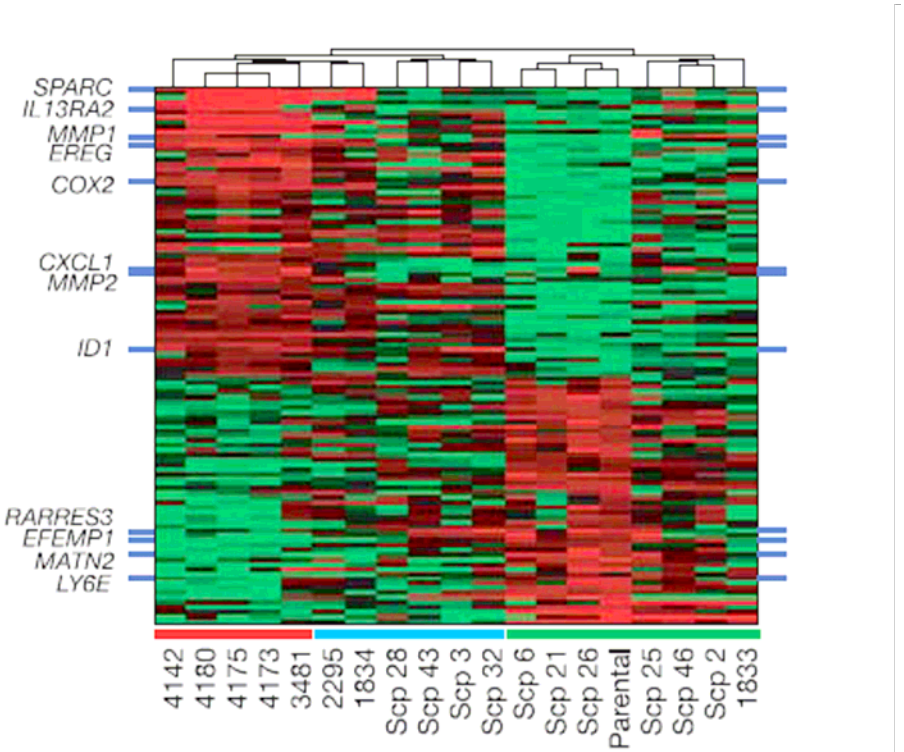

图 3. 以聚类图的方式体现变化的基因，可以在图上标明感兴趣的基因所在的位置。这种聚类图的表达方式非常直观。数据引自 Minn, A.J., et al., Nature, 2005, 436: 519-524.

3. 在论文写作中如何分析所得到的差异基因

得到较多数目差异表达基因后，若一个一个基因去查找文献将是一件非常累人的工作，而且还难以从总体上把握这些变化基因的内在联系。笔者认为比较合适的分析方法是，先对这些变化的基因进行批处理，分析出变化基因所在 pathway 的富集程度、基因按照功能分类的富集度、基因功能分类、是否存在共表达的基因、是否可能受到 microRNA 的调控等等，然后重点研究上述批处理结果中最感兴趣的部分。图 4 至图 7 列举了一些公开发表的例子，应用上述观点对差异表达基因批量进行分析。生物芯片北京国家工程研究中心为了更加方便生物芯片使用者对芯片数据的分析，汇集了分子生物学、生物信息和计算机相关人才开发了晶芯® 分子功能注释系统（MAS 系统），该系统整合了多个公共数据库的资源，以 Web 界面的形式免费提供给广大生物芯片使用者进行芯片数据分析整理，具体网址请见 http://bioinfo.capitalbio.com/mas/ 或者进入公司网站中文首页 http://www.capitalbio.com/index.asp 在导航条中的“客户通道”中选择“分子功能注释系统”，即能进行注册登陆后使用。

Table 2. Representative Fiber-Preferential Metabolic Pathways Identified by KOBAS

KEGG Pathways[a]	No. of Cotton Genes Located in Various Pathways	No. of Fiber-Upregulated Genes	P Value	FDR-Corrected P Value
Total	2914	162	–	–
Ethylene biosynthesis	5	3	0.0016	0.0295
γ-Hexachlorocyclohexane degradation	49	8	0.0049	0.0376
Cytoskeleton	75	10	0.0077	0.0376
Fatty acid biosynthesis and elongation	64	9	0.0080	0.0376
Glycosaminoglycan degradation	16	4	0.0099	0.0376
Stilbene, coumarine, and lignin biosynthesis	73	9	0.0183	0.0522
Ascorbate and aldarate metabolism	63	8	0.0217	0.0522
DDT degradation	12	3	0.0256	0.0522
Fluorene degradation	42	6	0.0267	0.0522
Androgen and estrogen metabolism	5	2	0.0275	0.0522
N-Glycan degradation	23	4	0.0356	0.0615
BR biosynthesis	6	2	0.0397	0.0629
Atrazine degradation	1	1	0.0556	0.0813

图 4. 对差异表达基因进行 pathway 富集度统计学分析后，定位差异基因最可能相关的 pathway。数据引自 Shi, Y. H. et al., The Plant Cell, 2006, 18: 651-664.

Overrepresented categories	U87		SF126	
	GOstat score	No. genes	GOstat score	No. genes
Genes with increased activity				
Cell cycle	7.00e−03	51	1.32e−09	60
Regulation of cell cycle	3.00e−03	34	2.23e−08	39
Transcription regulator activity				
Transcription factor binding	6.92e−04	20	8.87e−02	17
Transcription cofactor activity	1.82e−04	19	—	—
RNA polymerase II transcription factor	1.39e−03	16	7.12e−02	14
Nucleotide metabolism	4.32e−05	164	4.07e−03	181
Transcription from RNA polymerase II	3.74e−09	36	5.99e−06	35

图 5. 对差异表达基因进行 GO terms 富集度统计学分析后，定位差异基因最可能相关的 GO terms。数据引自 Lu et at., Cancer Research, 2006, 66: 1052-1061.

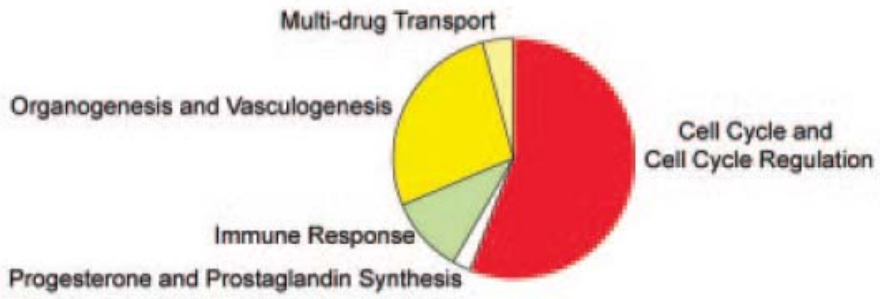

图 6. 对差异表达基因所在的功能分类 GO terms 做比例分配图。数据引自 Behbod, F., et al., Stem Cells, 2006: 24: 1065-1074.

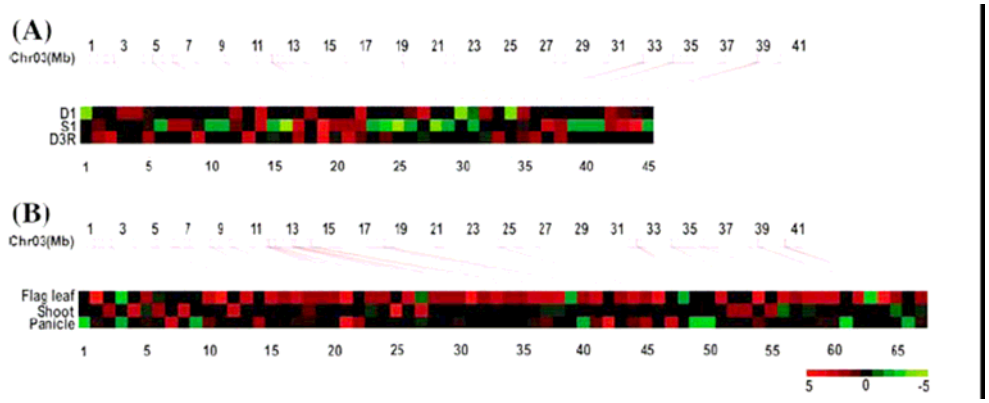

图 7. 在差异表达的基因中分析是否存在空间位置上共表达的基因。数据引自 Zhou J., et al., Plant Molecular Biology, 2007, 63: 591-608.

4. 如何展示定量 RT-PCR 的方法和芯片结果的吻合性

在利用芯片结果筛选到的基因进行生物学问题的阐述时，往往需要对支撑论文生物学发现的一些关键基因用其他检测基因表达的实验方法进行验证，其中常用的是定量 RT-PCR 法。若用定量 RT-PCR 验证的基因数目较多，例如 10 多个基因时，可以用散点图来展示两种方法检测的吻合度，参见图 8；当验证的基因数目较少，例如少于 5 个基因时，则可用柱状图来展示两种方法的吻合度，参见图 9。

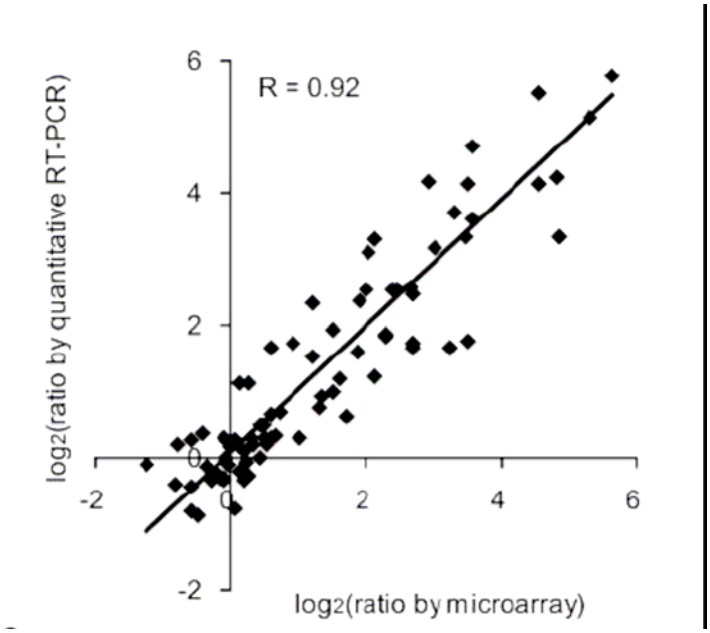

图 8. 用 X-Y 轴方式的散点图表明定量 RT-PCR 和芯片结果的吻合性。当验证的基因数目比较多时，适合用散点图格式。数据引自 Xiang,G-X., et al., Journal of Cellular Physiology, 2007, 212:126-136.

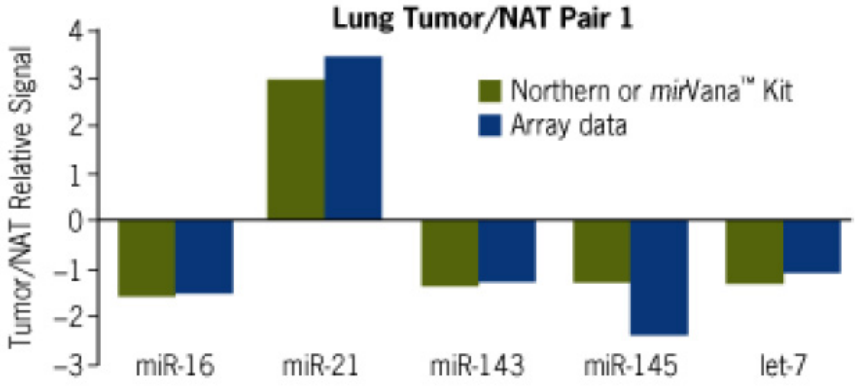

图 9. 当验证的基因数目不多时，可以用柱状图来展示两种方法的吻合度。尽管此图是比较 Northern 和芯片的结果，定量 RT-PCR 和芯片的结果的比较可以类似体现。若芯片结果和定量 RT-PCR 都有重复实验，则柱状图上可以加上误差线，科学性更强。数据引自 Ambion 公司的 TechNotes 11(6).

另外，当基因变化倍数较大时，例如超过 5 倍以上，选择半定量的 RT-PCR 方法未尝不是一个好方法。通过在凝胶电泳上展示 RT-PCR 终产物的灰度值，结果非常直观明了，在很多高档次的文章中都有使用。参见图 10。

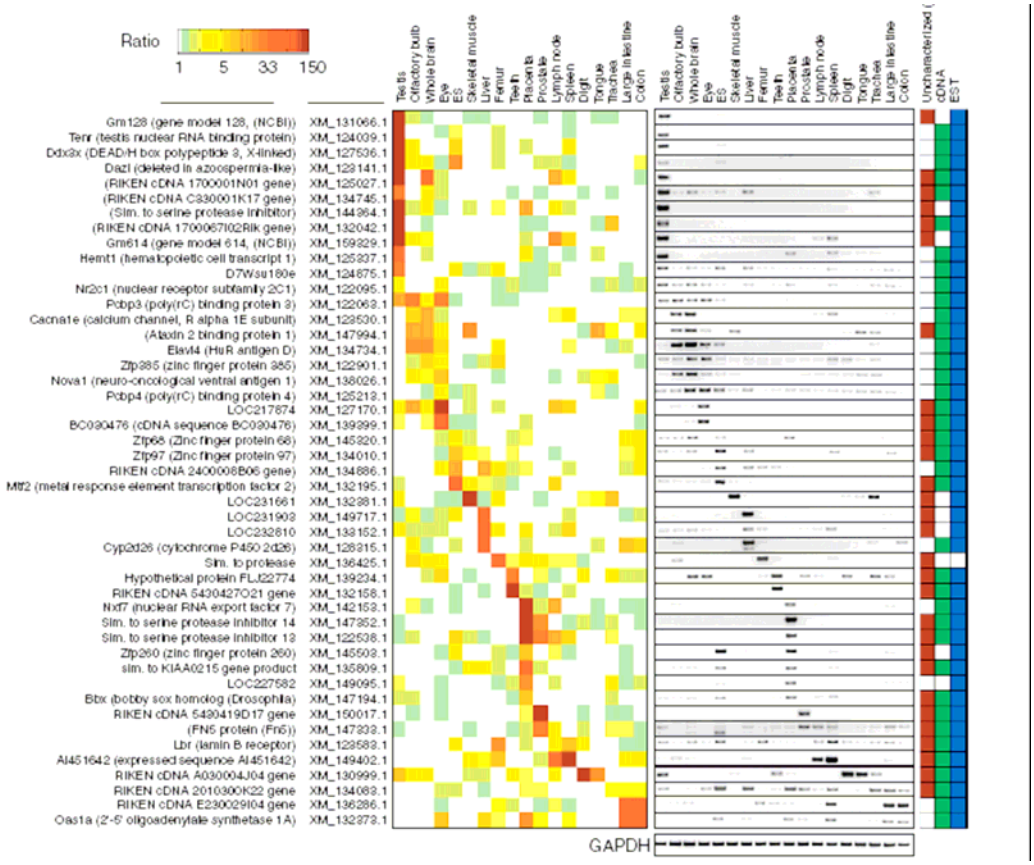

图 10. 通过半定量 RT-PCR 来验证芯片的数据。图左边的彩图是芯片的结果，右图是半定量

RT-PCR 的凝胶电泳图。这种比较方法非常直观明了。数据引自 Zhang, W., et al., Journal of Biology, 2004, 3:21.

总的说来，芯片技术是一个大规模的研究方法手段，技术方法最终是为解决生物学问题而服务的，在文章写作时，需要把握这个总体目标。