

[12] 发明专利申请公开说明书

[21] 申请号 99816329.5

[43] 公开日 2002 年 2 月 13 日

[11] 公开号 CN 1335893A

[22] 申请日 1999.12.27 [21] 申请号 99816329.5
[30] 优先权
[32] 1998.12.28 [33] US [31] 09/222,596
[86] 国际申请 PCT/US99/30837 1999.12.27
[87] 国际公布 WO00/39339 英 2000.7.6
[85] 进入国家阶段日期 2001.8.22
[71] 申请人 罗斯塔英法美蒂克斯公司
地址 美国华盛顿州
[72] 发明人 R·斯托顿
H·戴

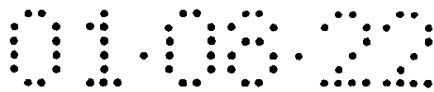
[74] 专利代理机构 中国专利代理(香港)有限公司
代理人 张广育 刘 玥

权利要求书 14 页 说明书 38 页 附图页数 8 页

[54] 发明名称 细胞表达特征的统计合并
[57] 摘要

一种在微排列实验中消除荧光团误差的方法,其中,在微排列实验中对颠倒使用荧光团。另外,一种用于计算与在标称重复微排列实验中所进行的每一项测定相关的单一误差的方法。该误差测定可选地与基于等级的方法结合,以便确定在受到扰动时,细胞成分是上调或下调的可能性。最后,一种用于确定在标称重复微排列实验中一种细胞成分表达水平的加权平均的置信度的方法。

ISSN 1008-4274



权 利 要 求 书

1. 一种消除荧光团误差的方法，包括以下步骤：

(a) 用第一种荧光团标记来自代表一种本底状态的生物系统的第一种遗传物质，以便获得第一种荧光团标记的遗传物质；

5 (b) 用第二种荧光团标记来自代表所述扰动状态的生物系统的第二种遗传物质，得到第二种荧光团标记的遗传物质；

(c) 用所述第二种荧光团标记来自代表所述本底状态的生物系统的第三种遗传物质，获得第三种荧光团标记的遗传物质；

10 (d) 用所述第一种荧光团标记来自代表所述扰动状态的生物系统的第四种遗传物质，得到第四种荧光团标记的遗传物质；

(e) 在能发生杂交的条件下，让所述第一种荧光团标记的遗传物质和所述第二种荧光团标记的遗传物质接触第一种微排列，并测定在此条件下结合于所述微排列上的所述第一种荧光团标记的遗传物质和在此条件下结合于所述微排列上的所述第二种荧光团标记的遗传物质之间的第一种颜色比例；

15 (f) 在能发生杂交的条件下，让所述第三种荧光团标记的遗传物质和所述第四种荧光团标记的遗传物质接触第二种微排列，并测定在此条件下结合于所述微排列上的所述第三种荧光团标记的遗传物质和在此条件下结合于所述微排列上的所述第四种荧光团标记的遗传物质之间的第二种颜色比例；和

20 (g) 通过平均所述第一种颜色比例和所述第二种颜色比例计算平均颜色比例。

2. 一种用于消除荧光团误差的计算机系统，该计算机系统包括一个处理器、和一个与该处理器连接的编码一种或多种程序的存储器，其中，所述一种或多种程序导致所述处理器完成一种方法，该方法包括：

(a) 用第一种荧光团标记来自代表一种本底状态的生物系统的第一种遗传物质，以便获得第一种荧光团标记的遗传物质；

30 (b) 用第二种荧光团标记来自代表所述扰动状态的生物系统的第二种遗传物质，得到第二种荧光团标记的遗传物质；

(c) 用所述第二种荧光团标记来自代表所述本底状态的生物系统的第三种遗传物质，获得第三种荧光团标记的遗传物质；

(d) 用所述第一种荧光团标记来自代表所述扰动状态的生物系统的第四种遗传物质，得到第四种荧光团标记的遗传物质；

(e) 在能发生杂交的条件下，让所述第一种荧光团标记的遗传物质和所述第二种荧光团标记的遗传物质接触第一种微排列，并测定在此条件下结合于所述微排列上的所述第一种荧光团标记的遗传物质和在此条件下结合于所述微排列上的所述第二种荧光团标记的遗传物质之间的第一种颜色比例；

(f) 在能发生杂交的条件下，让所述第三种荧光团标记的遗传物质和所述第四种荧光团标记的遗传物质接触第二种微排列，并测定在此条件下结合于所述微排列上的所述第三种荧光团标记的遗传物质和在此条件下结合于所述微排列上的所述第四种荧光团标记的遗传物质之间的第二种颜色比例；和

(g) 通过平均所述第一种颜色比例和所述第二种颜色比例计算平均颜色比例。

3. 一种消除荧光团误差的方法，该方法包括通过平均第一种颜色比例和第二种颜色比例而确定颜色比例，其中，所述第一种颜色比例和所述第二种颜色比例是这样测定的：

(a) 用第一种荧光团标记来自代表一种本底状态的生物系统的第一种遗传物质，以便获得第一种荧光团标记的遗传物质；

(b) 用第二种荧光团标记来自代表所述扰动状态的生物系统的第二种遗传物质，得到第二种荧光团标记的遗传物质；

(c) 用所述第二种荧光团标记来自代表所述本底状态的生物系统的第三种遗传物质，获得第三种荧光团标记的遗传物质；

(d) 用所述第一种荧光团标记来自代表所述扰动状态的生物系统的第四种遗传物质，得到第四种荧光团标记的遗传物质；

(e) 在能发生杂交的条件下，让所述第一种荧光团标记的遗传物质和所述第二种荧光团标记的遗传物质接触第一种微排列，并测定在此条件下结合于所述微排列上的所述第一种荧光团标记的遗传物质和在此条件下结合于所述微排列上的所述第二种荧光团标记的遗传物质之间的第一种颜色比例；和

(f) 在能发生杂交的条件下，让所述第三种荧光团标记的遗传物质和所述第四种荧光团标记的遗传物质接触第二种微排列，并测定

在此条件下结合于所述微排列上的所述第三种荧光团标记的遗传物质和在此条件下结合于所述微排列上的所述第四种荧光团标记的遗传物质之间的第二种颜色比例。

4. 一种用于消除荧光团误差的计算机系统，该计算机系统包括一个处理器、和一个与该处理器连接的编码一种或多种程序的存储器，其中，所述一种或多种程序导致所述处理器完成一种方法，该方法包括通过平均第一种颜色比例和第二种颜色比例而确定颜色比例，其中，所述第一种颜色比例和所述第二种颜色比例是这样测定的：

(a) 用第一种荧光团标记来自代表一种本底状态的生物系统的第一种遗传物质，以便获得第一种荧光团标记的遗传物质；

(b) 用第二种荧光团标记来自代表所述扰动状态的生物系统的第二种遗传物质，得到第二种荧光团标记的遗传物质；

(c) 用所述第二种荧光团标记来自代表所述本底状态的生物系统的第三种遗传物质，获得第三种荧光团标记的遗传物质；

(d) 用所述第一种荧光团标记来自代表所述扰动状态的生物系统的第四种遗传物质，得到第四种荧光团标记的遗传物质；

(e) 在能发生杂交的条件下，让所述第一种荧光团标记的遗传物质和所述第二种荧光团标记的遗传物质接触第一种微排列，并测定在此条件下结合于所述微排列上的所述第一种荧光团标记的遗传物质和在此条件下结合于所述微排列上的所述第二种荧光团标记的遗传物质之间的第一种颜色比例；和

(f) 在能发生杂交的条件下，让所述第三种荧光团标记的遗传物质和所述第四种荧光团标记的遗传物质接触第二种微排列，并测定在此条件下结合于所述微排列上的所述第三种荧光团标记的遗传物质和在此条件下结合于所述微排列上的所述第四种荧光团标记的遗传物质之间的第二种颜色比例。

5. 如权利要求 1 或 3 的方法，其中，所述第一荧光团和第二荧光团选自 Cy2-脱氧核苷三磷酸、Cy3-脱氧核苷三磷酸、Cy3.5-脱氧核苷三磷酸、Cy5-脱氧核苷三磷酸、Cy5.5-脱氧核苷三磷酸、Cy7-脱氧核苷三磷酸、荧光素、丽丝胺、藻红蛋白和罗丹明。

6. 如权利要求 2 或 4 的计算机系统，其中，所述第一荧光团和第二荧光团选自 Cy2-脱氧核苷三磷酸、Cy3-脱氧核苷三磷酸、Cy3.5-

脱氧核苷三磷酸、Cy5-脱氧核苷三磷酸、Cy5.5-脱氧核苷三磷酸、Cy7-脱氧核苷三磷酸、荧光素、丽丝胺、藻红蛋白和罗丹明。

5 7. 如权利要求 1 或 3 的方法，其中，所述第一种和第三种遗传物质是通过对从所述第一种生物系统中提取的 mRNA 进行逆转录而产生的 cDNA。

8. 如权利要求 1 或 3 的方法，其中，所述第二种和第四种遗传物质是通过对从所述第二种生物系统中提取的 mRNA 进行逆转录而产生的 cDNA。

10 9. 如权利要求 1 或 3 的方法，其中，所述平均颜色比例是通过以下公式计算的

$$\frac{1}{2} (\log(r_{X/Y}) - \log(r_{X/Y}^{(rev)}))$$

其中， $r_{X/Y}$ 表示所述第一种颜色比例，而所述 $r_{X/Y}^{(rev)}$ 表示所述第二种颜色比例。

15 10. 如权利要求 2 或 4 的计算机系统，其中，所述平均颜色比例是通过以下公式计算的

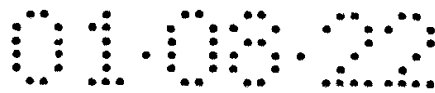
$$\frac{1}{2} (\log(r_{X/Y}) - \log(r_{X/Y}^{(rev)}))$$

20 其中， $r_{X/Y}$ 表示所述第一种颜色比例，而所述 $r_{X/Y}^{(rev)}$ 表示所述第二种颜色比例。

11. 如权利要求 1 或 3 的方法，其中，将所述平均颜色比例对所述第一和第二、第三和第四种荧光团标记的遗传物质与第三种微排列杂交的综合总强度作图。

25 12. 如权利要求 2 或 4 的计算机系统，其中，将所述平均颜色比例对所述第一和第二、第三和第四种荧光团标记的遗传物质与第三种微排列杂交的综合总强度作图。

13. 如权利要求 1 或 3 的方法，其中，将所述平均颜色比例对由荧光团标记的遗传物质与一种微排列杂交所产生的一定量的强度所测定的强度参数作图，其中，所述荧光团标记的遗传物质选自第一种



荧光团标记的遗传物质、第二种荧光团标记的遗传物质、第三种荧光团标记的遗传物质、和第四种荧光团标记的遗传物质。

14. 一种用于确定一种细胞成分的表达水平在多个成对的差异微排列实验中被一种扰动改变的可能性的方法，其中，在所述多个成对的差异微排列实验中的每一对差异微排列实验包括代表第一种生物系统的本底状态的第一种微排列实验，和代表所述第一种生物系统的扰动状态的第二种微排列实验，所述方法包括以下步骤：

5 (a) 通过将一种不依赖于强度的统计数字代入参考微排列实验对测定误差分布统计数字，其中，所述参考微排列实验对包括第一参考微排列实验，和作为所述微排列参考实验的标称重复的第二参考微排列实验；

10 (b) 从在所述多个成对的差异微排列实验中测定的一组细胞成分中选择所述细胞成分，并且，对于多个成对的差异微排列实验中的每一对差异微排列实验来说，用所述误差分布统计数字测定所述成对的差异微排列实验的所述第二种微排列实验和所述第一种微排列之间的所述细胞成分的表达水平的改变量；和

20 (c) 通过用基于等级的方法合并步骤(b)中所测定的所述多个成对的差异微排列实验中的每一对微排列实验的所述细胞成分的表达水平的改变量，确定在所述多个成对的差异微排列实验中所述细胞成分的表达水平被所述扰动改变的可能性。

15 15. 一种用于确定一种细胞成分的表达水平在多个成对的差异微排列实验中被一种扰动改变的可能性的计算机系统，其中，在所述多个成对的差异微排列实验中的每一对差异微排列实验包括代表第一种生物系统的本底状态的第一种微排列实验，和代表所述第一种生物系统的扰动状态的第二种微排列实验，所述计算机系统包括一个处理器、和一个与该处理器连接的编码一种或多种程序的存储器，其中，所述一种或多种程序导致所述处理器完成一种方法，该方法包括：

30 (a) 通过将一种不依赖于强度的统计数字代入参考微排列实验对测定误差分布统计数字，其中，所述参考微排列实验对包括第一参考微排列实验，和作为所述微排列参考实验的标称重复的第二参考微排列实验；

(b) 从在所述多个成对的差异微排列实验中测定的一组细胞成

分中选择所述细胞成分，并且，对于多个成对的差异微排列实验中的每一对差异微排列实验来说，用所述误差分布统计数字测定所述成对的差异微排列实验的所述第二种微排列实验和所述第一种微排列之间的所述细胞成分的表达水平的改变量；和

- 5 (c) 通过用基于等级的方法合并步骤 (b) 中所测定的所述多个成对的差异微排列实验中的每一对微排列实验的所述细胞成分的表达水平的改变量，确定在所述多个成对的差异微排列实验中所述细胞成分的表达水平被所述扰动改变的可能性。

10 16. 如权利要求 14 的方法，其中，所述误差分布统计数字是按以下公式计算的

$$\frac{(X-Y)}{\sqrt{\sigma_x^2 + \sigma_y^2 + f^2(X^2 + Y^2)}}$$

15 其中，X 表示所述参考微排列实验对的第一种微排列实验中细胞成分的强度，Y 表示所述参考微排列实验对的第二种微排列实验中细胞成分的强度， σ_x^2 是 X 的方差项，表示 X 的累计误差水平， σ_y^2 是 Y 的方差项，表示 Y 的累计误差水平，f 是分数倍增性误差水平。

17. 如权利要求 15 的计算机系统，其中，所述误差分布统计数字是按以下公式计算的

$$\frac{(X-Y)}{\sqrt{\sigma_x^2 + \sigma_y^2 + f^2(X^2 + Y^2)}}$$

20

其中，X 表示所述参考微排列实验对的第一种微排列实验中细胞成分的强度，Y 表示所述参考微排列实验对的第二种微排列实验中细胞成分的强度， σ_x^2 是 X 的方差项，表示 X 的累计误差水平， σ_y^2 是 Y 的方差项，表示 Y 的累计误差水平，f 是分数倍增性误差水平。

25 18. 如权利要求 16 的方法，其中，所述基于等级的方法包括按照权利要求 16 的公式导出的幅度，确定所述成对的差异微排列实验的第二种微排列实验和第一种微排列实验之间，所述细胞成分表达水平改变量，相对于所述多个成对的差异微排列实验中所有细胞成分的测

定值的等级。

19. 如权利要求 17 的方法，其中，所述基于等级的方法包括按照权利要求 17 的公式导出的幅度，确定所述成对的差异微排列实验的第二种微排列实验和第一种微排列实验之间，所述细胞成分表达水平改变量，相对于所述多个成对的差异微排列实验中所有细胞成分的测定值的等级。

20. 如权利要求 14 的方法，其中，所述基于等级的方法测定一种细胞成分针对一种扰动被上调的可能性。

21. 如权利要求 15 的计算机系统，其中，所述基于等级的方法测定一种细胞成分针对一种扰动被上调的可能性。

22. 如权利要求 20 的方法，其中，所述基于等级的方法具有以下公式

$$P(H_0^+) = \prod_i P_i$$

15 其中， P_i 是在所述多个成对的差异微排列实验 i 中一种细胞成分被上调的可能性， i 是选自所述多个成对的差异微排列实验的一个成对的差异微排列实验，而 P 是针对所述扰动所述细胞成分的表达水平被上调的可能性。

23. 如权利要求 14 的方法，其中，所述基于等级的方法测定一种细胞成分针对一种扰动被下调的可能性。

24. 如权利要求 23 的方法，其中，所述基于等级的方法具有以下公式

$$P(H_0^-) = \prod_i (1 - P_i)$$

25 其中， P_i 在所述多个成对的差异微排列实验 i 中一种细胞成分被下调的可能性，选自所述多个成对的差异微排列实验，而 P 是针对所述扰动所述细胞成分被下调的可能性。

25. 如权利要求 14 的方法，其中，所述多个成对的差异微排列实验中的每一对差异微排列实验是双荧光团微排列实验，其中，第一种

荧光团表示所述生物系统的本底状态，而第二种荧光团表示所述生物系统的扰动状态。

26. 如权利要求 14 的方法，其中，在所述成对的差异微排列实验中使用单一荧光团。

5 27. 如权利要求 14 的方法，其中，将第一种荧光团标记用于所述第一参考微排列实验，将第二种荧光团标记用于第二参考微排列实验。

28. 一种用于测定一种生物系统中一种细胞成分的表达水平针对扰动而发生的加权平均差异强度的方法，该方法包括：

10 (a) 通过将一种不依赖于强度的统计数字代入参考微排列实验对测定误差分布统计数字，其中，所述参考微排列实验对包括第一参考微排列实验和作为所述微排列参考实验的标称重复的第二参考微排列实验；

(b) 多次测定所述细胞成分的差异表达量；

15 (c) 对按照步骤 (b) 测定的每一个差异表达量来说，基于所述误差分布统计数字所得出的幅度计算相应的误差量；和

(d) 通过用在步骤 (b) 中测定的所述细胞成分的差异表达量除以在步骤 (c) 中测定的相应的误差量计算加权平均差异强度，按照以下公式计算

20

$$x = \frac{\sum (x_i / \sigma_i^2)}{\sum (1 / \sigma_i^2)}$$

其中， x 是所述细胞成分表达水平的加权平均差异强度， x_i 是通过步骤 (b) 测定的所述细胞成分差异表达量的测定值，而 σ_i^2 是通过步骤 (c) 测定的相应的 x_i 的误差量。

25 29. 一种用于测定一种生物系统中一种细胞成分的表达水平针对扰动而发生的加权平均差异强度的计算机系统，该计算机系统包括一个处理器、和一个与该处理器连接的编码一种或多种程序的存储器，其中，所述一种或多种程序导致所述处理器完成一种方法，该方法包括：

30 (a) 通过将一种不依赖于强度的统计数字代入参考微排列实验

对测定误差分布统计数字，其中，所述参考微排列实验对包括第一参考微排列实验和作为所述微排列参考实验的标称重复的第二参考微排列实验；

(b) 多次测定所述细胞成分的差异表达量；

5 (c) 对按照步骤(b)测定的每一个差异表达量来说，基于所述误差分布统计数字所得出的幅度计算相应的误差量；和

(d) 通过用在步骤(b)中测定的所述细胞成分的差异表达量除以在步骤(c)中测定的相应的误差量计算加权平均差异强度，按照以下公式计算

10

$$x = \frac{\sum (x_i / \sigma_i^2)}{\sum (1 / \sigma_i^2)}$$

其中，x是所述细胞成分表达水平的加权平均差异强度， x_i 是所述细胞成分i的差异表达量，而 σ_i^2 是 x_i 的相应误差量。

30. 如权利要求28的方法，其中，步骤(b)还包括：

15

(i) 在所述微排列与来自表示本底状态的一种生物系统的第一种荧光团标记的遗传物质接触之后，测定该微排列上一个部位的第一强度，其中，所述微排列上的所述部位表示所述细胞成分；

(ii) 在所述微排列与来自表示扰动状态的一种生物系统的第二种荧光团标记的遗传物质一起培养之后，测定该微排列上一个部位的第二强度；

20

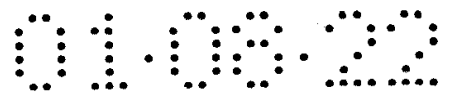
(iii) 通过从所述第一强度中扣除所述第二强度计算所述细胞成分的差异表达。

31. 如权利要求28的方法，其中，所述第一种和第二种荧光团标记的遗传物质包括通过逆转录由mRNA产生的cDNA。

25

32. 如权利要求28的方法，其中，所述误差分布统计数字是按以下公式计算的

$$\frac{(X - Y)}{\sqrt{\sigma_X^2 + \sigma_Y^2 + f^2(X^2 + Y^2)}}$$



其中， X 表示所述参考微排列实验对的第一种微排列实验中细胞成分的强度， Y 表示所述参考微排列实验对的第二种微排列实验中细胞成分的强度， σ_x^2 是 X 的方差项，表示 X 的累计误差水平， σ_y^2 是 Y 的方差项，表示 Y 的累计误差水平， f 是分数倍增性误差水平。

- 5 33. 一种用于对预定细胞成分 j 进行测定，确定多种细胞成分差异表达测定的加权平均置信度的方法，其中，每一种细胞成分差异表达测定是通过选自多个成对的微排列实验的一对差异微排列实验测定的，其中，每一对差异微排列实验包括代表一种生物系统的本底状态的第一种微排列实验，和代表所述一种生物系统的扰动状态的第二种微排列实验，所述方法包括以下步骤：

(a) 通过将一种不依赖于强度的统计数字代入参考微排列实验对测定误差分布统计数字，其中，所述参考微排列实验对包括第一参考微排列实验，和作为所述微排列参考实验的标称重复的第二参考微排列实验；

- 15 (b) 对所述多个成对的差异微排列实验中的每一对差异微排列实验来说，根据所述误差分布统计数字确定一个误差量；

(c) 根据所述多个成对的差异微排列实验，用以下关系式确定细胞成分 j 的分散值 s_j

$$s_j \cong \frac{1}{N-1} \sum_i (x_i - \bar{x})^2$$

20

其中， x_i 是通过成对的差异微排列实验 i 确定的细胞成分 j 的差异测定， $\bar{x}$ 是细胞成分 j 在所述多个成对的差异微排列实验中所有差异测定的非加权平均值，而 N 是所述多个成对的差异微排列实验的成对的差异微排列实验的数量；和

- 25 (d) 合并步骤 (b) 中所测定的每一对差异微排列实验的误差量和所述分散值 s_j ，以便确定对预定细胞成分 j 进行测定的所述多种细胞成分的差异表达测定的加权平均的置信度。

34. 如权利要求 33 的方法，其中，所述误差分布统计数字是按以下公式计算的

$$\frac{(X-Y)}{\sqrt{\sigma_x^2 + \sigma_y^2 + f^2(X^2 + Y^2)}}$$

其中，X表示所述参考微排列实验对的第一种微排列实验中细胞成分的强度，Y表示所述参考微排列实验对的第二种微排列实验中细胞成分的强度， σ_x^2 是X的方差项，表示X的累计误差水平， σ_y^2 是Y的方差项，表示Y的累计误差水平，f是分数倍增性误差水平。

35. 如权利要求 33 的方法，其中，细胞成分 j 的所述多种细胞成分的差异表达测定中的每一种细胞成分的差异表达测定是通过以下方法确定的：

10 (i) 在所述微排列与来自表示本底状态的一种生物系统的第一种荧光团标记的遗传物质接触之后，测定该微排列上一个部位的第一强度，其中，所述微排列上的所述部位相当于所述细胞成分 j；

(ii) 在所述微排列与来自表示扰动状态的一种生物系统的第二种荧光团标记的遗传物质一起培养之后，测定该微排列上一个部位的第二强度，其中，所述微排列上的所述部位相当于所述细胞成分 j；
15 和

(iii) 通过从所述第一强度中扣除所述第二强度计算所述细胞成分的差异表达。

36. 如权利要求 33 的方法，其中，步骤 (b) 还包括：

20 (i) 将所述误差统计数字在 X-Y 曲线上作图，其中，第一个轴线表示强度，而第二个轴线表示表达比例；和

(ii) 通过将所述第二种强度在所述第一轴线上作图确定沿所述第一轴线的位置并测定基于按照所述位置上的所述误差统计数字作图的 $\pm 1\sigma$ 网格线的宽度，确定所述误差的量。

25 37. 如权利要求 33 的方法，其中，步骤 (d) 还包括：

合并权利要求 33 的步骤 (b) 中所确定的每一对差异微排列实验的误差量和权利要求 33 的步骤 (c) 中的所述分散值 s_j ，按以下公式计算

$$\sigma_x = \frac{1}{N} \left[\left(\sqrt{\frac{1}{\sum_i \frac{1}{\sigma_i^2}}} \right) + (N-1) * s_j \right]$$

其中,

$$\frac{1}{\sum (\frac{1}{\sigma_i^2})}$$

5

是按照权利要求 33 的步骤 (b) 用所述误差分布统计数字确定的, N 是用于计算 S_j 的成对的差异微排列实验的数量, 而 σ_x 表示确定的所述预定细胞成分 j 的所述多种细胞成分差异表达测定的加权平均的置信度。

10

38. 一种用于对预定细胞成分 j 进行测定, 确定多种细胞成分差异表达测定的加权平均置信度的计算机系统, 其中, 每一种细胞成分差异表达测定是通过选自多个成对的微排列实验的一对差异微排列实验测定的, 其中, 每一对差异微排列实验包括代表一种生物系统的本底状态的第一种微排列实验, 和代表所述一种生物系统的扰动状态的第二种微排列实验, 其中, 所述计算机系统包括一个处理器、和一个与该处理器连接的编码一种或多种程序的存储器, 其中, 所述一种或多种程序导致所述处理器完成一种方法, 该方法包括:

15

(a) 通过将一种不依赖于强度的统计数字代入参考微排列实验对测定误差分布统计数字, 其中, 所述参考微排列实验对包括第一参考微排列实验, 和作为所述微排列参考实验的标称重复的第二参考微排列实验;

20

(b) 对所述多个成对的差异微排列实验中的每一对差异微排列实验来说, 根据所述误差分布统计数字确定一个误差量;

25

(c) 根据所述多个成对的差异微排列实验, 用以下关系式确定细胞成分 j 的分散值 s_j

$$s_j \equiv \frac{1}{N-1} \sum_i (x_i - \bar{x})^2$$

其中， x_i 是通过成对的差异微排列实验 i 确定的细胞成分 j 的差异测定， $\bar{x}$ 是细胞成分 j 在所述多个成对的差异微排列实验中所有差异测定的非加权平均值，而 N 是所述多个成对的差异微排列实验的成对的差异微排列实验的数量；和

(d) 合并步骤 (b) 中所测定的每一对微排列实验的误差量和所述分散值 s_j ，以便确定对预定细胞成分 j 进行测定的所述多种细胞成分的差异表达测定的加权平均的置信度。

39. 如权利要求 38 的计算机系统，其中，所述误差分布统计数字是按以下公式计算的

$$\frac{(X-Y)}{\sqrt{\sigma_X^2 + \sigma_Y^2 + f^2(X^2 + Y^2)}}$$

其中， X 表示所述参考微排列实验对的第一种微排列实验中细胞成分的强度， Y 表示所述参考微排列实验对的第二种微排列实验中细胞成分的强度， σ_X^2 是 X 的方差项，表示 X 的累计误差水平， σ_Y^2 是 Y 的方差项，表示 Y 的累计误差水平， f 是分数倍增性误差水平。

40. 如权利要求 38 的计算机系统，其中，细胞成分 j 的所述多种细胞成分的差异表达测定中的每一种细胞成分的差异表达测定是通过以下方法确定的：

(i) 在所述微排列与来自表示本底状态的一种生物系统的第一种荧光团标记的遗传物质接触之后，测定该微排列上一个部位的第一强度，其中，所述微排列上的所述部位相当于所述细胞成分 j ；

(ii) 在所述微排列与来自表示扰动状态的一种生物系统的第二种荧光团标记的遗传物质一起培养之后，测定该微排列上一个部位的第二强度，其中，所述微排列上的所述部位相当于所述细胞成分 j ；和

(iii) 通过从所述第一强度中扣除所述第二强度计算所述细胞

成分的差异表达。

41. 如权利要求 38 的计算机系统，其中，步骤 (b) 还包括：

(i) 将所述误差统计数字在 X-Y 曲线上作图，其中，第一个轴线表示强度，而第二个轴线表示表达比例；和

5 (ii) 通过将所述第二种强度在所述第一轴线上作图确定沿所述第一轴线的位置并测定基于按照所述位置上的所述误差统计数字作图的 $\pm 1\sigma$ 网格线的宽度，确定所述误差的量。

42. 如权利要求 38 的计算机系统，其中，步骤 (d) 还包括：

10 合并权利要求 38 的步骤 (b) 中所确定的每一对差异微排列实验的误差量和权利要求 38 的步骤 (c) 中的所述分散值 s_j ，按以下公式计算

$$\sigma_x = \frac{1}{N} \left[\left(\sqrt{\frac{1}{\sum_i \frac{1}{\sigma_i^2}}} \right) + (N-1) * s_j \right]$$

其中，

15

$$\frac{1}{\sum (\frac{1}{\sigma_i^2})}$$

是按照权利要求 38 的步骤 (b) 用所述误差分布统计数字确定的，N 是用于计算 S_j 的成对的差异微排列实验的数量，而 σ_x 表示确定的所述预定细胞成分 j 的所述多种细胞成分差异表达测定的加权
20 平均的置信度。

说明书

细胞表达特征的统计合并

1. 发明领域

5 本发明涉及用来自多次重复实验的数据产生每一个数据点的置信度值、提高灵敏度、并消除系统实验误差的方法。

2. 发明背景

2.1 细胞成分的定量测定

在“细胞成分”含量的定量测定形成方面目前正爆炸性地增加。
10 细胞成分包括生物系统中基因表达水平、编码特定基因的 mRNA 的丰度、和蛋白的表达水平。已知诸如 mRNA 编码基因和/或蛋白表达水平的各种细胞成分的含量会随着药物处理和细胞生物学状态的其它扰动而改变。因此，对多种此类“细胞成分”的测定包含了有关扰动对细胞生物学状态的影响的大量信息。通常将所述测定综合起来称作细
15 胞生物学状态的“特征”。

对哺乳动物细胞来说，有数量级在 100,000 上的不同的细胞成分。因此，特定细胞的特征通常是复杂的。生物系统的任何状态下的特征通常是在该生物系统受到扰动之后测定的。所述扰动包括与一种生物系统相关的实验或环境条件，让该系统接触一种药物候选物，导
20 入外源基因，从该系统中缺失基因，或改变培养条件。因此，综合测定细胞成分或基因和蛋白表达的特征，及其对细胞中扰动的反应具有多种用途，包括比较并理解药物作用的能力，诊断疾病，和优化患者药物方案。另外，还可将其用于基础生命科学研究。

在过去几十年里，所取得的若干技术进步使得有可能精确测定细
25 胞成分及其所产生的特征。例如，新技术提供了在任何时间监测大量转录物表达水平的能力（例如，参见 Schena 等，1995，用互补 DNA 微排列定量监测基因表达方式，科学 270: 467-470; Lockhart 等，1996，通过与高密度寡核苷酸排列杂交监测表达，自然生物技术 54: 1675-1680; Blanchard 等，1996，序列到排列：探测基因组秘密，
30 自然生物技术 14, 1649; US5,569,588; 于 1996 年 10 月 29 日授予 Ashby 等，题为药物筛选方法）。在整个基因组已知的生物中，可以分析特定细胞内所有基因的转录物。对于诸如人类的其它生物来说，

对其基因组的了解日益增加，有可能同时监测特定细胞内的大量基因。

在另一方面，通过用微柱反相液相层析电喷射离子化串联质谱分析 (LC/MS/MS) 直接鉴定混合物中所含的蛋白改善了蛋白丰度的直接测定。该技术有可能拓宽在生物系统中可以测定的蛋白丰度的动态范围。McCormack 等业已使用 LC/MS/MS 证实了可以用 30 倍不同的摩尔量方便地鉴定存在于系统混合物中的蛋白，所述鉴定是可再现，并且所述混合物里的蛋白可以在低的飞摩尔水平上鉴定。McCormack 等，1997，通过 LC/MS/MS 和数据库检索在低飞摩尔水平上直接分析并鉴定混合物中的蛋白，分析化学 69: 767-76。在有关串联质谱分析的综述中，Chait 指出了该技术的另一个优点，即该技术要远远比诸如 Edma 测序的常规技术快速。Chait, 1996, 在后基因组时代捕获蛋白，自然生物技术 14: 1544。

其它技术进步业已提供了用单个的遗传学突变对生物系统进行专一性扰动的能力。例如，Mortensen 等披露了一种生产胚胎干 (ES) 细胞系的方法，通过同源重组使两个等位基因同时失活。用 Mortensen 等的方法可以获得纯合的突变改变过的细胞，即 ES 细胞系的双剔除体。Mortensen 等提出他们的方法可一般性地应用于其它基因并应用于除 ES 细胞之外的细胞系。Mortensen 等，1992，用单一定向结构生产纯合突变 ES 细胞，细胞生物学 12: 2391-2395。

在另一种有希望的技术中，Wach 等提供了一种用于筛选酿酒酵母转化体的显性抗性模式，所述突变体完全由杂合 DNA 组成。该模式还可用于提供基于 PCR 的基因破坏。Wach 等，1994，用于酿酒酵母的基于传统或 PCR 基因破坏的新的杂合模式，酵母 10: 1793-808。

诸如使用微排列的技术发展便于用于药物发现 (例如，参见 Marton 等，1998，用微排列进行的药物目标确定和二级药物目标作用的鉴定，自然医学，待发表，Gray 等，1998，将化合物文库、结构、和基因组用于检索激酶抑制剂，科学 281: 533-538)。

将特征与数据库中的其它特征进行比较 (例如，参见 1998 年 7 月 7 月授予 Rine 等的题为产生并分析刺激响应输出信号机制的 US5, 777, 888) 或根据相似性将特征分类，可以提供有关药物分子目标及相关功能、药物候选物的效力和毒性和/或药物制剂的线索。所

述比较还可用于产生代表理想的药物活性或疾病状态的共有特征。特征比较还有助于在早期检测患者的疾病，并提供改善的临床结果，用于患者的疾病诊断。

2.2 荧光团误差

5 Shalon 等业已披露了两种荧光团的应用。Shalon, 1996, 用于通过双色荧光探针杂交分析复杂 DNA 样品的微排列系统, 基因组研究 6: 629-645. Shalon 所提出的该方法的问题是, 由于荧光标记分子与 mRNA 逆转录或与杂交效率或这两者的相互作用使得每一种 mRNA 分子在其测定的颜色比例方面存在误差。没有任何误差校正系统能校正该误差, 来自单一微排列实验的数据甚至是来自微排列实验的多个标称重复的数据 (其中将各个结果加以平均) 都会产生不可接受的误差率。在本文中, 术语标称重复或标称重复的实验是指在基本上相同或相似的实验条件下进行的实验, 以便可用于合并重复实验的结果。

2.3 细胞成分定量测定实验所固有的误差率

15 尽管技术进步业已容许对细胞成分的含量进行定量测定, 但这些实验的费用是昂贵的。一次微排列实验或一次凝胶电泳实验可能花费大约 100-1000 美元或更多。另外, 只是在将所述结果应用于实际商业需求的多次初步尝试之后才明白, 单一的实验在证实显著性方面存在高水平的伪阳性, 实际上这种显著性是不存在的。由于涉及到高的费用和高比例的伪阳性, 在现有技术中没有提供有关为了数据质量改善的表达目的而重复并统计合并多次、标称相同实验的有效方法。

20 用微排列获得的基因组范围的细胞特征分析的效力在于其在实际上所有细胞机制方面观察对扰动的反应的能力。不过, 在任何特定实验中, 通常仅有少数细胞成分在丰度方面可能有大的改变, 而大多数没有改变。存在的例外是, 尽管细胞对刺激具有特殊的、生物学上实际上的绝缘反应, 因此大多数特征涉及‘无变化’的大量成分, 只有少得多的成分要么上调要么下调。因此, 在测定方面即使小的伪报警比例也能严重损害其实用性。例如, 如果在典型的实验中 1% 的细胞成分实际上有反应, 测定时的分辨率为两倍, 而误差超过 1% 的两倍, 有与实际检测一样多的伪报警高于两倍的阈值。

30 一般, 本领域不理解在诸如微排列或蛋白凝胶实验的单一细胞成分定量实验中所存在的大量的误差。除了对特定扰动的实际孤立的响

应生物系统所造成的困难之外，在任何标称微排列实验中都存在大量的误差，这些误差由于假象而造成，如将 DNA 探针的斑点不均匀地印在微排列上，所述微排列上的划痕粉尘和伪迹、由于不均匀的 DNA 杂交而导致的通过该微排列的信号亮度的不均匀性、以及由于用于该微排列方法中的荧光团特有的误差所导致的颜色条。

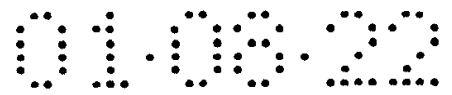
降低这项严重误差影响的一种方法是在相同条件下重复实验，并将数据加以平均。不过，将数据简单地进行平均，而不考虑实验误差背后的实质不能提供对实验误差所引起的问题的合适的解决方案。如果仅对所述数据进行简单地平均，可能需要过多次数的标称重复，以便将误差的影响降低到可以接受的水平。不过，由于在实施每一种细胞成分定量实验时所花费的费用，这不是一种可行的解决方案。因此，本领域需要用于合并重复的细胞成分定量实验的实验结果的有效方法，以便使用最少次数的标称重复就可以提供可接受的误差率。

本文对参考文献的讨论或引用并不意味着承认所述引用是本发明的现有技术。

3. 发明概述

本发明提供了减少为了产生可接受的误差水平的数据而必须重复进行的细胞成分定量实验次数的解决方案。因此，本发明的方法提供了一种消除荧光团误差的新方法。该方法能够将荧光团特有的误差减轻到可接受的水平，仅基于细胞成分定量实验的两次标称重复即可实现这一目的。本发明还提供了根据上调或下调的等级合并细胞成分定量实验的标称重复的方法。在所述方法中，通过细胞成分定量实验的标称重复确定的细胞成分的上调或下调可以通过用一种新的参数表达，该参数没有依赖强度的误差。在根据等级合并之前采用该参数，提供了一种消除弱表达细胞成分的误差的有效方法，而不必对费用昂贵的细胞成分定量实验进行过多次数的标称重复。

本发明的另一方面是一种用于计算标称重复的细胞成分定量实验中单一细胞成分测定的加权平均的改进方法。具体地讲，提供了一种用于计算与每一个细胞成分测定相关的误差的新方法。用这种新方法计算误差，在所述加权平均中的误差线显著降低。本领域技术人员可以理解的是，所述用于计算加权平均的改进方法可应用于双荧光团（双色）或单荧光团（单色）方案。



本发明的一种实施方案提供了一种消除荧光团误差的方法，包括以下步骤：

(a) 用第一种荧光团标记来自代表一种本底状态的生物系统的第一种遗传物质，以便获得第一种荧光团标记的遗传物质；

5 (b) 用第二种荧光团标记来自代表所述扰动状态的生物系统的第二种遗传物质，得到第二种荧光团标记的遗传物质；

(c) 用所述第二种荧光团标记来自代表所述本底状态的生物系统的第三种遗传物质，获得第三种荧光团标记的遗传物质；

10 (d) 用所述第一种荧光团标记来自代表所述扰动状态的生物系统的第四种遗传物质，得到第四种荧光团标记的遗传物质；

(e) 在能发生杂交的条件下，让所述第一种荧光团标记的遗传物质和所述第二种荧光团标记的遗传物质分别接触第一种微排列，并测定结合于所述微排列上的所述第一种荧光团标记的遗传物质和所述第二种荧光团标记的遗传物质之间的第一种颜色比例；

15 (f) 在能发生杂交的条件下，让所述第三种荧光团标记的遗传物质和所述第四种荧光团标记的遗传物质分别接触第二种微排列，并测定所述第三种荧光团标记的遗传物质和所述第四种荧光团标记的遗传物质之间的第二种颜色比例；和

20 (g) 通过平均所述第一种颜色比例和所述第二种颜色比例计算平均颜色比例。

本发明的另一种实施方案提供了一种确定一种细胞成分的表达水平在多个成对的差异微排列实验中被一种扰动改变的可能性的方法，其中，在所述多个成对的差异微排列实验中的每一对差异微排列实验包括代表第一种生物系统的本底状态的第一种微排列实验，和代表所述第一种生物系统的扰动状态的第二种微排列实验，所述方法包括以下步骤：

25 (a) 通过将一种不依赖于强度的统计数字代入参考微排列实验对测定误差分布统计数字，其中，所述参考微排列实验对包括第一参考微排列实验，和作为所述第一微排列参考实验的标称重复的第二参考微排列实验；

30 (b) 从在所述多个成对的差异微排列实验中测定的一组细胞成分中选择所述细胞成分，并且，对于多个成对的差异微排列实验中的

每一对差异微排列实验来说,用所述误差分布统计数字测定所述成对的差异微排列实验的所述第二种微排列实验和所述第一种微排列实验之间的所述细胞成分的表达水平的改变量; 和

- 5 (c) 通过用基于等级的方法合并步骤(b)中所测定的所述多个成对的差异微排列实验中的每一对微排列实验的所述细胞成分的表达水平的改变量,确定在所述多个成对的差异微排列实验中所述细胞成分的表达水平被所述扰动改变的可能性。

10 本发明的另一种实施方案是一种用于测定一种生物系统中一种细胞成分的表达水平针对扰动而发生的加权平均差异强度的方法,该方法包括:

(a) 通过将一种不依赖于强度的统计数字代入参考微排列实验对测定误差分布统计数字,其中,所述参考微排列实验对包括第一参考微排列实验和作为所述微排列参考实验的标称重复的第二参考微排列实验;

- 15 (b) 多次测定所述细胞成分的差异表达量;

(c) 对按照步骤(b)测定的每一个差异表达量来说,基于所述误差分布统计数字所得出的幅度计算相应的误差量; 和

- 20 (d) 通过用在步骤(b)中测定的所述细胞成分的差异表达量除以在步骤(c)中测定的相应的误差量计算加权平均差异强度,按照以下公式计算

$$x = \frac{\sum (x_i / \sigma_i^2)}{\sum (1 / \sigma_i^2)}$$

其中, x 是所述细胞成分的加权平均差异强度, x_i 是所述细胞成分 i 的差异表达测定, 而 σ_i^2 是平均差异强度 x_i 的相应的误差。

25 4. 附图简述

图 1 表示存在于微排列荧光图象上的测定误差的某些来源。(A) 表示不均匀印制的 DNA 探针斑点。(B) 表示划痕、粉尘、和伪迹的影响。(C) 表示斑点位置是如何漂移离开标称测定网格的。(D) 表示由于不均匀杂交所导致的在所述微排列上亮度的效果。(E) 表示

30 由于荧光团专一性误差所导致的所述微排列上的颜色条效果。

图 2 表示缺失酿酒酵母中决定钙调磷酸酶蛋白生产的基因 (CNA1 和 CNA2) 的效果。该图比较了两种酵母培养物的响应特征, 一种是天然培养物 (培养物 1), 另一种是业已缺失了 CNA1 和 CNA2 的培养物 (培养物 2)。水平轴线是从以上两种酵母培养物所获得的微排列上的单一杂交斑点强度的 $\log_{10}$, 因此表示 mRNA 的丰度。垂直轴线是所测定的一种荧光标记的强度 (培养物 1) 与所测定的其它标记强度 (培养物 2) 的比例的 $\log_{10}$ (表达比例)。CNA1/CNA2 突变的真实的标记基因被确定为明显偏离 $\log_{10}$ (表达比例) = 0 线的基因, 并作出标记。

图 3 表示由于荧光团光学检测效率误差和荧光团结合效率误差所导致的在细胞表达特征实验中所发生的取决于强度的误差。

图 4A 是一个实验中颜色比例/强度的曲线, 在该实验中, 两种培养物是相同背景的酿酒酵母菌株。对在红色和绿色荧光团之间存在明显误差的基因进行标记。图 4B 是与图 4A 相同的实验, 所不同的是颠倒使用红色和绿色荧光团。图 4C 表示本发明消除误差的方法, 其中, 将图 4A 和图 4B 合并以便产生没有荧光团专一性误差的响应特征。

图 5 比较了在相同实验条件下进行的两种相同的响应特征。该图表示实验误差的降低受强度 (表达水平) 的影响。不依赖于强度的高线表示本发明误差校正方法的一个分量。

图 6a 表示用药物环胞菌素 A 进行的单一实验的典型的标记曲线。

图 6b 表示将本发明方法的加权平均应用于图 6a 所示实验的 4 个重复中的结果。

图 7 表示可用于本发明实施方案中的计算机系统。

5. 发明详述

5.1 前言和一般定义

扰动: 在本文中扰动是与一种生物系统相关的实验或环境条件。扰动可以通过让一种生物系统接触药物候选物或药物制剂、将一种外源基因导入生物系统, 从所述生物系统中缺失基因、改变所述生物系统的培养条件、或本领域所认可的扰动生物系统的其它方法而实现。另外, 生物系统的扰动可以通过生物系统的疾病发作而实现。

遗传物质: 在本文中, 术语“遗传物质”, 是指核酸, 如 mRNA、

cDNA、基因组 DNA (gDNA)、DNA、RNA、基因、寡核苷酸、片段、及其任意组合。

荧光团标记的遗传物质：在本文中，术语“荧光团标记的遗传物质”是指用荧光标记探针（荧光团）标记过的遗传物质。荧光团包括，
5 但不限于荧光团、荧光素、丽丝胺、藻红蛋白、罗丹宁 (Perkin Elmer Cetus)、Cy2、Cy3、Cy3.5、Cy5、Cy5.5、Cy7、FluorX (Amersham) 及其它（例如，参见 Kricka, 1992, 非同位素 DNA 探针技术，学术出版社，San DIEGO, CA）。该 DNA 可以通过 mRNA 的逆转录或通过 (PCR/IVT) 或 (IVT) 而制备，使用本领域技术人员所了解的荧光团。
10 例如，参见 Gelder 等, 1990, 用有限量的外源 cDNA 合成的扩增 RNA, 美国科学院院报 87: 1663-1667。在本文中，术语 PCR 是指聚合酶链式反应。

生物系统：在本文中，术语“生物系统”被广义地定义为包括所有细胞、组织、器官或多细胞生物。例如，生物系统可以是细胞系、
15 细胞培养物、获自一个对象的组织样品、人类、哺乳动物、基本上与酿酒酵母同基因的酵母、或任何本领域所了解的其它生物系统。生物系统的状态可以通过其细胞成分的含量、活性或结构来测定。在本文中，生物系统的状态是根据所有细胞成分的状态来确定的，所述成分足于为了预期目的而鉴定所述细胞或生物，包括鉴定药物或其它扰动的效果。术语“细胞成分”包括任何类型的可测定的生物学变量。在
20 所述成分的所述状态下进行的测定和/或观察可以是其丰度（即在生物系统中的数量或浓度）、其活性、其改变状态（例如磷酸化）、或本领域所知的与一种生物系统的生理学状态相关的其它指标。在各种实施方案中，本发明包括用不同的细胞成分的集合进行所述测定和/或观察。所述细胞成分的不同的集合又被称作一种生物系统的生物学状态方面。

可用于在本发明中测定的所述生物系统的生物学状态（例如，细胞或细胞培养物）的一方面是其转录状态。生物系统的转录状态包括组成 RNA，特别是 mRNA 在特定条件下的细胞中的性质和丰度。通常，
30 测定所述生物系统中所有组成 RNA 的少部分，但至少测定足够的部分，以便鉴定药物或其它感兴趣扰动的作用。生物系统的转录状态可以通过用若干种现有基因表达技术的任一种测定 cDNA 丰度而方便地

进行测定。用于测定 mRNA 或大量基因的转录水平的 DNA 排列可用于确定一种系统的生物学状态。

可用于测定的生物系统的生物学状态的另一方面是翻译状态。生物系统的翻译状态包括组成蛋白在特定条件下的所述生物系统中的性质和丰度。优选对该生物系统中的所有组成蛋白的少部分进行测定，但至少对足够的部分进行测定，以便鉴定感兴趣的药物的作用。转录状态通常代表了所述翻译状态。生物系统的生物学状态的其它方面也可用于本发明。例如，生物系统的状态包括组成蛋白（还可选地包括具有催化活性的核酸）在特定条件下的该生物系统中的活性。正如本领域所公知，翻译状态通常代表了所述活性状态。

本发明还适用于生物系统的生物学状态的“混合的”方面，其中，将生物系统的生物学状态的不同方面的测定合并。例如，在一个混合方面，将某些 RNA 和某些蛋白的丰度与某些其它蛋白的活性的测定合并。另外，通过下面的说明可以理解，本发明还适用于可测定的生物系统的生物学状态的任何其它方面。

生物系统（例如，细胞或细胞培养物）的生物学状态可以用某种数量的细胞成分的特征表示。所述细胞成分的特征可以用矢量 S 表示。

$$S=[S_1, \dots S_i, \dots S_k]$$

其中， S_i 是第 i 种细胞成分的含量，例如，基因 i 的转录水平，或者蛋白 i 的丰度或活性水平。

细胞成分的定量测定：微排列 测定复杂样品中多种单一序列的相对丰度通常是用微排列进行的。例如，参见 Shalon 等，1996，用于通过双色荧光探针杂交分析复杂样品的微排列系统，基因组研究 6: 639-645。通常，转录排列是通过让代表存在于细胞中的 mRNA 转录物的可检测的标记过的多核苷酸（例如，由总的细胞 mRNA 合成的荧光标记过的 cDNA）与一种微排列杂交而产生的。微排列是具有用于结合细胞或生物的基因组中许多基因，优选大部分或几乎所有基因产物的结合（例如，杂交）位点的有序排列的表面。微排列是高度可再现性的，因此可以生产特定排列的多个拷贝，并将标称的拷贝彼此加以比较。优选的微排列是小型的，通常小于 5 平方厘米，并且是用在结合（例如，核酸杂交）条件下稳定的材料制成。所述微排列上的

特定结合位点或特定类型的结合位点能专一性地结合所述细胞中单一基因的产物。

当制备一种细胞的 RNA 的互补 cDNA 并在合适杂交条件下与微排列杂交时，与所述排列上相应于任何特定基因的位点的杂交水平反应了由该基因转录的 mRNA 在所述细胞中的含量。例如，当互补于总的细胞 mRNA 的可检测的标记过的（例如，用荧光团标记）cDNA 与一种微排列杂交时，所述排列上相应于在细胞中未转录的基因的位点（即，能够专一性地结合该基因的产物）将少有或没有信号（例如，荧光信号），而其所编码的 mRNA 占优势的基因将具有较强的信号。

微排列是优选的，因为代表两种不同核酸的核酸能够与一种微排列杂交，并且可以同时测定每一种核酸的相对信号。每一种核酸可能代表在扰动之前和之后一种生物系统的状态。例如，第一种核酸可能来自让一种细胞培养物接触一种药物制剂之前该细胞培养物的 mRNA，而第二种 cDNA 可能来自在让相同的培养物接触一种药物制剂之后的该培养物的 mRNA。另外，两种 cDNA 可能代表途径响应。因此，第一种 cDNA 文库可以来自接受过一种途径扰动的细胞培养物的第一等份试样（pool）的 mRNA，而第二种 cDNA 文库可能来自同一种细胞培养物的第二等份试样（pool）的 mRNA，其中，所述第二等份试样没有受到所述途径扰动。在本文中，微排列实验包括在本节所披露的实验，这项实验被称作“差异微排列实验”。本领域技术人员可以理解的是，除了本说明书所披露的差异微排列实验之外的许多形式的差异微排列实验属于“差异微排列实验”定义范围。另外，在本文中，术语“差异强度测验”是指在差异微排列实验中进行的测定。例如，一种差异强度测定可以是微排列上一个位置上的亮度之间的差异，它相当于感兴趣的细胞成分，所述差异出现在（i）所述微排列接触过来自代表一种本底状态的生物系统的 DNA，和（ii）所述微排列接触过来自代表受扰动状态的生物系统的 DNA 之后。另外，本领域技术人员可以理解的是，生物系统的本底状态可能表示该生物系统的野生状态。另外，生物系统的本底状态可以表示该生物系统的不同的扰动状态。差异微排列实验中的每一个微排列实验，或者重复的差异微排列实验优选采用相同或相似的排列。如果微排列是用大体上等基因的生物系统制备的并且每一个微排列上的大部分结合斑点是相同的话，该

微排列被视为是相似的。因此，用于重复的微排列实验中的微排列可以是同一性质的微排列，其中，在微排列实验之间对该微排列进行洗涤，或者用于重复的微排列实验中的微排列可以是彼此的复制品，或者它们可以彼此相似。在差异微排列实验中，无论两种 cDNA 的来源如何，如果选择双荧光团微排列方案的话，要对每一种 cDNA 进行不同的标记。本领域技术人员可以理解的是，本发明的某些方面不局限于双荧光团方案。通常，每一种 cDNA 是通过在存在 Cy3-（绿色）或 Cy5（红色）脱氧核苷三磷酸（Amersham）的条件下通过 polyA+RNA 的逆转录用衍生的荧光标记的 cDNA 标记的。当将两种 cDNA 混合并与所述微排列杂交时，测定所述微排列的每一个位点来自每一种 cDNA 的信号 5 的相对强度，并检测特定 mRNA 丰度的任何相对差异。

当使用诸如 Cy3 和 Cy5 的两种不同的荧光标记的探针时，在微排列的每一个位点上所发射的荧光可以使用激光扫描共焦显微术测定。在一种实施方案中，用合适的激发光线对所使用的两种荧光团的每一种进行独立的扫描。另外，可以使用激光使样品在对以上两种荧光团专一的波长下同时发光，并同时分析由这两种荧光团所发射的光线（例如，参见 Sahlon 等，同上）。所述微排列可以用装有计算机控制的 X-Y 平台和显微镜目镜的激光荧光扫描仪扫描。以上两种荧光团的顺序激发是用多光线、多线、混合的气体激光实现的，并根据波长 15 分离所发射的光线，并用两个光线倍增管检测。荧光激光扫描装置披露于 Schena 等（1996，基因组研究 6: 639-645）的文章和本文所引用的参考文献中。另外，还可以使用 Ferguson 等（1996，自然生物技术 14: 1681-1684）所披露的光学纤维束，以便同时监测大量位点上的 mRNA 丰度。

可以用计算机记录并分析信号，例如，使用 12 比特的模拟-数字板。可以使用绘图程序（例如，HiJaak 绘图软件）对扫描的图象进行脱斑，然后用图象分格程序进行分析，由此产生在每一个位点上的每一种波长下平均杂交的分布平面。如果必要，可以对以上两种荧光团的通道之间的干扰（或重叠）进行实验确定的校正。

在本文中，术语“微排列实验”是指在本节所披露的实验的总体类型。本领域技术人员可以理解的是，微排列实验可以包括使用单一的荧光团，而不是使用双荧光团，其例子见下文所述。另外，微排列

实验可以是成对的。如果是成对的，在该实验对中的第一种微排列可以代表表示本底状态的标称生物系统。该实验对中的第二种微排列实验可以代表在受到扰动之后的所述标称生物系统。因此，比较所述成对的微排列实验，可以揭示所述标称生物系统基于所述扰动的状态的变化。一般，如上文所述，所述微排列实验对被称作“差异微排列实验”。

细胞表达特征在微排列实验中使用两种不同 cDNA 的优点是可以对相应于两种细胞状态下每一种排列的基因进行直接的和内部控制的比较。这种技术和用于细胞成分定量测定的相关技术一般被称作细胞成分特征分析。细胞成分特征分析通常表现为在两种已知的细胞状态下绝对含量或含量的比例的改变，如对药物制剂处理本底状态的响应，如在上节中所披露的。

使用上一节所披露的实验方法，可以计算 DNA 转录物排列上的任何特定杂交位点上的由以上两种荧光团所发射光线的比例。该比例不取决于关联基因的绝对表达水平，但可用于其表达因为使用药物、基因缺失或任何其他扰动而受到明显调节的基因。如图 2-6 所示，双荧光团细胞表达特征通常在 x-y 曲线上作图。水平轴线表示所述微排列上的每一个部位的第一种和第二种 cDNA 之间的平均强度比例的 $\log_{10}$ （它大体上反应了由一个基因所产生的相应 mRNA 的表达水平），垂直轴线表示所述微排列上每一个杂交位点上测定的相应于所述第一种 cDNA 的一种荧光标记的强度与测定的相应于第二种 cDNA 的另一种荧光标记的强度的比例的 $\log_{10}$ 。

5.2 消除荧光团误差

正如在背景技术部分所述的，由 Shalon 等（同上）所提出的双色荧光杂交技术将误差引入了特征分析，因为用荧光团标记的每一种 mRNA 在其测定的颜色比例方面存在误差，这是因为荧光标记分子（荧光团）与 mRNA 的逆转录或与杂交效率或与这两种之间的相互作用所造成的。该误差可以用以下公式表示。如果我们用 k 表示特定 mRNA 的实际分子丰度，用 (k) 表示感兴趣的生物系统中细胞成分或基因 k ，无论荧光团误差的来源如何，探针 k 的颜色比例可以表示为：

$$r_{xy} = a_1(k) / a_2(k) \quad (1)$$

其中,

下标 1 和 2 表示两种分别提取的 mRNA 培养物, 其中, 对其丰度
5 进行比较;

$\alpha_1(k)$ 是 mRNA 培养物 1 中的 k 类的丰度;

$\alpha_2(k)$ 是 mRNA 培养物 2 中的 k 类的丰度;

下标 X 和 Y 表示使用的两种不同的荧光标记; 而 $r_{x/y}$ 是能理想地
反应丰度比例 α_1/α_2 的颜色比例。

10 公式 (1) 理想地代表了在图 2-6 的垂直轴线上作图的测定值。
不过, 使用荧光团标记的脱氧核苷三磷酸会影响 mRNA 逆转录成 cDNA
的效率, 并能影响荧光团标记的 cDNA 与所述微排列杂交的效率。特
定荧光团影响所述转录或杂交效率的精确程度在很大程度上取决于
所使用的荧光团的确切分子结构。因此, 在用不同的荧光团测定 α_1
15 (k) 和 $\alpha_2(k)$ 时, $\alpha_1(k)$ 与 $\alpha_2(k)$ 的直接比较不能算作是对转
录或杂交效率的荧光团专一性影响的比较。扫描仪在测定微排列上的
 $\alpha_1(k)$ 和 $\alpha_2(k)$ 的丰度时的效率也是荧光团专一性的。如果我们
用 E 表示特定荧光团在提取、标记、逆转录、杂交和光学扫描中的综
合效率的话, 对公式 1 中所出现的颜色比例的一种更实际的表达形
20 式为:

$$r_{xy} = a_1(k)E_x(k) / a_2(k)E_y(k) \quad (2)$$

其中,

$r_{x/y}$ 是颜色比例;

25 下标 1 和 2 如公式 1 中所定义的;

$\alpha_1(k)$ 与 $\alpha_2(k)$ 如公式 1 所定义的;

下标 X 和 Y 是两种荧光标记;

$E_x(k)$ 是荧光标记 X 的效率; 和

$E_y(k)$ 是荧光标记 Y 的效率。

在公式 2 中，用荧光团 X 分析培养物 1，而用荧光团 Y 分析培养物 2。现在，所述颜色比例与希望的丰度比例 α_1/α_2 相关，但包括由于所述荧光团专一性效率误差所导致的因素。如果进行第二种杂交实验，其中，培养物 1 用荧光团 Y 分析，而培养物 2 用荧光团 X 分析，那么

5 那么在第二种杂交实验中的颜色比例可以表示为：

$$r_{XY}^{(rev)} = a_2(k)E_X(k) / a_1(k)E_Y(k) \quad (3)$$

其中

$r_{X/Y}^{(rev)}$ 是所述颠倒实验中的颜色比例；和

10 $\alpha_2(k)$ 、 $\alpha_1(k)$ 、 $E_X(k)$ 、和 $E_Y(k)$ 如公式 (2) 中所述。

成对的进行杂交实验，在成对实验中的一个中颠倒了标记的使用，以便产生合并的平均测定，其中，所述荧光团专一性误差大大降低。例如，可以进行一对双荧光团杂交实验。第一种双荧光团实验可以按照公式 (2) 进行，而第二种双荧光团杂交实验可以按照公式 3

15 进行。如果取这两种实验的比例的对数，则合并的实验可以表达为：

$$\begin{aligned} (1/2) (\log(r_{XY}) - \log(r_{XY}^{(rev)})) &= \log(a_1(k)/a_2(k)) + (\log(E_X(k)/E_Y(k)) - \log(E_X(k)/E_Y(k))) \\ &= \log(a_1(k)/a_2(k)) \end{aligned} \quad (4)$$

它是理想的对数丰度比例。误差项 $\log(E_X(k)/E_Y(k))$ 和 $\log(E_X(k)/E_Y(k))$ 的取消取决于在每一个荧光团颠倒对中第一种

20 和第二种杂交实验之间的误差常数。公式 (4) 可以等同地书写成用公式 1-3 中所出现的比例代替 $\log$ 比例的差。不过，成分含量的改变最好表达为在构成所述差异测定的成对的条件下丰度比例的对数。这是因为倍数的改变比绝对的数学含量的改变更有意义。

这种误差消除方法特别适用于双色杂交实验。图 4 表示本发明的

25 误差消除方法。图 4a 是双色杂交实验的颜色比例与强度的曲线图，其中，所使用的两种培养物是酿酒酵母的标称相同背景的菌株。因为这两种培养物是标称相同的，预计在微排列上的每一个斑点上对所使用的荧光团来说，会发出相同强度的荧光。实验方法将在下面的实验部分介绍。不过，正如从图 4a 中可以看出的，所述微排列上的某些

斑点表现出荧光团专一性的强度。例如，所述微排列上相应于酿酒酵母的各个基因的斑点，其中，对红色荧光团的强度为相应的绿色强度2倍或更高的斑点进行标记，因为其具有强的荧光团专一性误差。图4b表示在图4a中作图的实验的荧光团颠倒形式的结果。在图4b中被标记的基因现在具有相反的误差。图4c表示按照上述本发明方法合并图4a和4b数据的结果。所述标记基因的误差被大大降低了。

上述误差消除方法可应用于其他场合，例如，如果培养物必须生长在培养箱中的特定位置，并按特定顺序收获的话，在下一次实验中可以颠倒两种培养物的位置和顺序，并按上述方法合并其结果，以便降低由于温度或潜在差异所导致的微小误差。

5.3 用基于等级的方法合并多个实验

现有技术没有提供用于优化合并多个微排列实验结果的明确方法。应当对若干实验的结果加以平均。不过，平均不能提供在所述微排列实验中每一个感兴趣的特定基因的任何特定测定的统计学显著性的信息。这一部分提出了一种复杂的方法，用于确定在多个微排列实验中测定的感兴趣的特定基因被上调或下调的统计学显著性。该方法可用于双荧光团微排列实验的标称重复。另外，所述方法可应用于成对实验的一个或几个，其中，在所述实验对中的第一个实验表示本底状态，而所述成对重复中的第二个实验表示在受到扰动之后的生物学状态。

如果感兴趣的基因在微排列实验的第一和第二标称重复中上调了5%的话，在两种排列中它被偶然上调的可能性仅为 $0.05 \times 0.05 = 0.0025$ 或 0.25%，表示系统误差被消除了。因此，重复测定可以提高证明感兴趣的基因被上调了的置信度。一般，如果在任何数目的重复实验中的表达比例表示为百分位数的话，任何（预定的）感兴趣的基因不被实际上调的几率 $P(H_0)$ 为

$$P(H_0) = \prod_i P_i \quad (5)$$

其中， P_i 是在第 i 个实验中的百分位数，表示为分数（ $5\% = 0.05$ ）。

该基因不被下调的可能性用以下公式表示

$$P(H_0) = \prod_i (1 - P_i) \quad (6)$$

所述基于等级的方法提供了一种通过重复测定降低伪报警的有效方式。例如，设定阈值为与覆盖酵母基因组（具有大约 6000 个基因）的探针杂交中的表达比例的上 5%，在一次实验中会产生大约 $6000 * 0.05 = 300$ 假性检测，但如果将相同的 5% 的阈值用在 4 个实验重复上平均仅有不到一次假性检测（ $6000 * (0.05)^4$ ）。优点是，除了假设没有系统误差之外，它不需要模拟在潜在杂交实验中的确切的误差行为。所述基于等级的方法是观察到的上调或下调的显著性的无参数统计学测验的一个例子。

诸如公式（5）和（6）的百分位数等级是基于所有基因的潜在误差行为是相似的这样一种假设。这不一定是真实的情形。例如，在图 5 中，绘制出了相同实验的两个标称重复的表达比例，其中的弱表达基因用 $\log_{10}$ （强度）表示，其 $\log_{10}$ （表达比例）偏离理想值 0。另外，如图 5 所示，一个特定基因的表达越弱，那么来自一个实验的两个标称重复的该基因的 $\log_{10}$ （表达比例）偏离 0 的倾向越高。因此，低丰度（弱表达，因此杂交强度低）基因比较高丰度基因更倾向于占据表达比例分布的尾部（即如图 5 所示，偏离 0）。

为了解决在诸如图 5 所示的杂交实验中所表现出的取决于强度的误差，可以设计一种测定上调和下调的方法，从而使得误差水平不取决于强度。这种不取决于强度的误差水平是利用统计学得到的，它能够鉴定在杂交实验中所出现的误差外套。这种误差外套在图 5 中用等高线表示。用于产生诸如图 5 所示曲线的实验所存在的误差的许多来源通常分成两类—累计性的和倍增性的。因此，以下统计学公式提供了特别适于代入所得到的误差的模式：

$$d = \frac{(X - Y)}{\sqrt{\sigma_x^2 + \sigma_y^2 + f^2(X^2 + Y^2)}} \quad (7)$$

其中，X 和 Y 是所述微排列上相应于 X 和 Y 荧光团的探针斑点的亮度，

σ_x^2 是 X 的误差项，表示 X 通道中的累计的误差水平， σ_y^2 是 Y 的误差项，表示 Y 通道中的累计的误差水平，f 是分数倍增性误差水平。另外，X 和 Y 是相应于来自一对单一荧光团实验的感兴趣的细胞成分的探针斑点亮度。在这样一种实验中，第一种荧光团 (X) 可以可选地代表处于本底状态的一种生物系统，而第二种荧光团 (Y) 可以代表处于扰动状态的该生物系统。无论是选择单一荧光团还是双荧光团实验，所述分数倍增性误差 f 是通过将测定的数据代入公式 7 作分母而凭经验产生的。公式 7 的分母是预期分子的预期标准误差，因此，d 具有单位误差。因此，d 是不取决于强度的误差分布途径，并因此可应用于等级方法。公式 7 的非参数特征的任何其他定义不能很好地改变以便用于所述等级方法。

根据本发明方法，公式 7 的分母被用于产生图 5 所示的不依赖于强度的等高线。因此，例如，在图 5 中，选择了在 ± 1 标准误差处绘成的等高线。因此，位于垂直轴线 ($\log(\text{表达比例})$ 等于 0) 上的 0 以上或以下的每一条等高线表示根据公式 7 的分母的误差的标准误差的增量完全是随意。所述等高线可以在任何方便的值，如 0.25σ 、 0.5σ 、 2σ 处绘成网格，只要所述等高线是按照公式 7 的分母或类似的误差的非参数表达绘制的即可。

从图 5 可以看出，所述等高线遵循所述误差外套。d 的值与所述等高线的数量成比例。一个特定的测定值偏离 $\log(\text{表达比例})$ 等于 0。因此，所述误差是相应于所述等高线类似地以低和高强度分布的，而 d 具有预期的特性。绘制等高线的一个优点是，与在所述微排列上测定的每一种细胞成分相关的误差的量可以根据所述微排列上所有细胞成分在多项测定中的误差所产生的信息计算。因此，通过使用图 5 所绘制的网格线，在双色荧光探针杂交实验中 X_i 和 Y_i 之间的任何误差的显著性可以用诸如公式 7 的分母的公式置于整个误差外套的范围内，其中，i 是一种特定细胞成分。由此提供了一种用于测定在包括双荧光团或单荧光团实验在内的微排列实验中测定的特定细胞成分的测定值的可靠性的不依赖于强度的方法。

除了取决于强度之外，误差水平还可以是基因专一性的，再次违反了公式 5 和 6 的假设。在这种情况下，我们可以类似于公式 7 的方式定义任何基因，

$$d = \frac{(X-Y)}{\sigma_{X-Y}} \quad (7.1)$$

其中, σ_{X-Y} 是与该基因相关的标准误差 (rms 不确定性)。这种不确定性可能来自重复的对照实验, 其中, X 和 Y 是来自相同的生物系统, 其中, σ_{X-Y} 是所述基因在所述实验组中所观察到的该基因的 X-Y 标准误差。d 的定义类似的分布于所述基因, 并且 (5) 和 (6) 可以用于对 d 进行分级。

5.4 用加权平均方法合并多个实验

可以合并重复的测定, 以便得到具有比单一测定小的误差线的定量表达水平或表达比例。按照所述单一实验误差水平, 加权平均方法需要了解或假设有关每一个测定量的误差行为的情况。一般, 具有最低误差的无误差加权平均是通过以下公式获得的

$$x = \frac{\sum (x_i / \sigma_i^2)}{\sum (1 / \sigma_i^2)} \quad (8)$$

其中, x 是被测定的细胞成分的加权平均值, X_i 和每一个 σ_i^2 是单一 X_i 的误差。例如, 参见公式 5-6, “物理科学的数据还原和误差分析”, 1969, Bevin-ton, McGraw-Hill, 纽约, 该文献被全文收作本文参考。

可以用多种方法测定公式 (8) 中的每一个 σ_i^2 。一种方法是用双荧光团微排列实验的两个标称重复计算微排列实验的误差外套, 其中, 以上两个实验的唯一差别是颠倒使用两种荧光团。例如, 参见图 4。另外, 可以只使用一种荧光团。因此, 为了确定误差外套在成对的两个标称重复之间可能根本就没有差别。这样的成对实验如图 5 所示。图 5 还表明与公式 7 的分母吻合的不取决于强度的等高线。为了确定每一个 σ_i^2 , 对每一个独立的测定 i 来说, 将强度 (X_i) 绘制在合适的参考曲线上, 如图 5 所示。例如, 在图 5 中, 单个测定的强度可沿水平轴线作图。一旦确定了水平位置, 就可以根据在所述参考曲线上在位置 X_i 处 $\pm 1\sigma$ 强度依赖性等高线的宽度计算 σ_i^2 。

用于计算所述平均值的不确定性的公式为

$$\sigma_x^2 = \frac{1}{\sum \left(\frac{1}{\sigma_i^2}\right)} \quad (9)$$

与上文所述“用于物理科学的数据还原和误差分析”的公式 5-10 吻合。应当指出的是，当与不同的标称重复测定的误差相等时，该误差的平均值为单一误差的 $N^{1/2}$ 倍。

在实施所述单一误差时， σ_i^2 本身是不确定的。检查如图 5 所示的对照实验发现了误差的大致分布，但不能表明在特定强度下的单个基因是否倾向于因为其 RNA 提取的特殊性甚至是在细胞中的生物学功能而具有较大误差。因此，为了更好地估算加权平均的误差，可以在公式 9 中增加一个分量，该分量表示在重复实验中的分散。如果我们用 s_j 表示所观察到的基因 j 的标准误差的话，其平均值的误差可以表示为：

$$\sigma_x = \frac{1}{N} \left[\left(\sqrt{\frac{1}{\sum \frac{1}{\sigma_i^2}}} \right) + (N-1) * s_j \right] \quad (10)$$

其中， N 是重复测定的数量。随着重复数量 N 变大，公式 10 从公式 (9) 转变成所观察到分散值 s_j 。要指出的是， s_j 是按照传统统计学方法计算的，因此

$$s_j \equiv \frac{1}{N-1} \sum_i (x_i - \bar{x})^2 \quad (11)$$

其中， N 是测定次数， X_i 是在特定微排列实验中基因 j 的强度的单一的测定，而 $\bar{x}$ 是单一测定取样平均值。例如，参见上文所述“用于物理科学的数据还原和误差分析”中的公式 2-10，其中， $s_j = \sigma_i^2$ 。如公式 10 所述平均值误差的估计 x 是必要的，因为诸如公式 11 的公式需要大量的标称重复 (N)，以便真实反应误差。基于公式 (9) 的误差估算没有考虑如图 1 所示的特殊测定所容易受到影响的误差以及基

因专一性反常。本领域技术人员可以理解的是，能够实现公式 9 向公式 10 转变的其他公式也是可行的。

图 6 表示通过重复实验所获得的误差的降低，以及随后所获得的信息。图 6a 是用药物 CsA 所进行的单一实验的标记曲线，是按下文实验部分所述方法获得的。根据公式 7 的分母确定了一个 σ 误差线，累计性和倍增性误差水平的值是从对照实验中获得的。只有基因偏离曲线 $\log(\text{比例})=0$ 超过 1.5σ ，即只有当其以 95% 以上的置信度被上调或下调时，该基因标记为 $1-\sigma$ 误差线。图 6b 表示产生 4 次重复 ($N=4$) 的加权平均的结果。在这里将 1.5σ 的相同标准用于标记误差线 (但有更多的基因被标记)。与图 6a 比较发现，以 95% 的置信度检测的数量从 4 个增加到 20 个基因以上。因此，该例子表明如果用诸如公式 10 的公式对测定误差进行合适的模拟的话，可以通过重复测定获得有关药物反应的额外信息。

5.5 响应特征

生物系统对诸如药物制剂的扰动的响应可以通过观察该生物系统在特定生物学状态下的变化来测定。一种响应特征是细胞成分变化的集合。一种生物系统 (例如细胞或细胞培养物) 对扰动 m 的响应特征可以定义为矢量 $v^{(m)}$ ：

$$v^{(m)} = [v_1^{(m)}, \dots, v_i^{(m)}, \dots, v_k^{(m)}] \quad (12)$$

其中， $v_i^{(m)}$ 是细胞成分 i 在扰动 m 作用下的响应幅度。在响应特征的某些实施方案中，对使用药物试剂的药物学响应是通过诱导至少 2 个基因，优选 10 个以上基因，更优选 100 个以上基因，最优选 1000 个以上基因的转录物产生的改变而测定的。

在某些实施方案中，生物学响应特征包括在扰动之前和之后生物学变量之间的简单的差别。在某些优选实施方案中，所述生物学响应被定义为在施加扰动之前和之后细胞成分的比例。

在某些优选实施方案中，如果基因 i 的响应低于某种阈值幅度或者通过了解测定误差行为所确定的置信度水平的话，将 $v_i^{(m)}$ 设定为 0。在所述实施方案中，将其测定响应值低于所述阈值的细胞成分规定为

响应值为 0，而其测定的响应值大于所述阈值的细胞成分保留其测定的响应值。当预计大多数较小的响应会被测定误差而大大提高时，这种响应矢量的截短是合适的。在所述截短之后，所述响应矢量 $v^{(n)}$ 还适合于用于检测类似扰动存在的‘匹配探测器’（例如，参见 Van Trees, 1968, 检测、估算、和调节理论, 1 卷, Wiley&Sons）。本领域技术人员可以理解的是，可以根据检测的目的和测定误差设定截短水平。例如，在某些实施方案中，将转录水平的改变低于 2 倍、更优选低于 4 倍的基因设定为 0 值。

在响应特征的某些优选实施方案中，以几种不同的强度水平施加扰动。例如，可将不同量的药物用于一种生物系统，以便观察其响应。在所述实施方案中，所述扰动响应可以通过所述扰动强度 u 的单一参数化‘模拟’化函数模拟每一个值的方式通过插入法求出。接近于模拟转录状态数据的一种代表性模拟函数是 Hill 函数，它具有可调整的参数 a 、 u_0 、和 n 。

$$H(u) = \frac{a(u/u_0)^n}{1 + (u/u_0)^n} \quad (13)$$

所述可调整的参数是针对每一种细胞成分的扰动响应分别选择的。优选的是，所述可调整的参数是为每一种细胞成分选择的，以便降低在每一种扰动强度下模拟函数（例如，Hill 函数，公式 13）和相应的实验数据之间的差异的平方的总和。这种优选的参数调整方法在本领域中被称为最小二乘拟合。其他可行的模拟函数是基于多项式拟合。有关模拟拟合和生物学响应的详细的说明披露于以下文献中：Friend 和 Stoughton, 用基因表达特征测定蛋白活性水平的方法，美国临时专利申请流水号 60/084,742，申请日为 1998 年 5 月 8 日，该文献被以全文形式从各个方面收作本文参考。

5.6 投射特征

本发明的方法可用于比较放大的特征，这些特征包括任意数量的响应特征和/或投射特征。在对基因集进行讨论之后可以更好地理解投射特征，所述基因集是共调节基因。投射特征可用于分析包括基因集在内的许多类型的细胞成分。

5.6.1 共调节基因和基因集

用基因集表示投射特征披露于这一节以及下面的小节中，另外，还详细披露于以下专利文献中：美国专利申请流水号 09/179,569，申请日为 1998 年 10 月 27 日，题为“用共调节基因集增强基因表达的测定和分类的方法”，申请人为 Friend 等，和美国专利申请流水号（待定）（律师档案号 9301-039-999）申请日为 1998 年 2 月 23 日，申请人为 Friend 等，题为“用共调节基因集增强基因表达的测定和分类的方法”，这两份申请均以全文形式收作本文参考。某些基因倾向于成群的增强或减弱其表达。当某些基因具有相似的调节序列形式，即转录因子结合位点时，这些基因倾向于同时增强或减弱其转录速度。这是对特定信号输入作出协调响应的机制（例如，参见 Madhani 和 Fink, 1998, MAP 激酶信号传导专一性之谜，遗传学事项 14: 151-155; Arnong 和 Davidson, 1997, 发育的硬件：基因组调控系统的组织和功能，发育 124: 1851-1864）。能产生必需蛋白或细胞结构不同成分的独立基因倾向于协变。复制的基因（例如，参见 Wazner, 1996, 由复制所引起的遗传学冗余性及其在转录调节子网络中的进化，Biol. Cybern. 74: 557-567）倾向于协变到这样的程度，突变不会导致所述调控区的功能异化。另外，由于调控序列是模量（例如，参见 Yuh 等, 1998, 基因组顺调节逻辑：海胆基因实验和计算分析，科学 279: 1896-1902），两个基因所共有的模量越多，预计其转录速度发生协变的条件的种类就越多。模量之间的差别也是一个重要决定因素，因为还涉及到共激活剂。因此，总而言之，对于任何有限类型的条件来说，预期基因不会发生独立的变化，并且基因和蛋白的简化的亚类会发生协变。这些协变类型的基因构成了数学意义上的完整的基础，通过它可以说明在所述有限的类型条件下的所有特征变化。

5.6.2 通过聚类分析进行基因集分类

对很多用途来说，需要发现在多种条件下共调节的基础基因集。用于鉴定所述基础基因集的一种优选实施方案涉及聚类算法（例如，有关聚类算法的综述可以参见 Fukunaga, 1990, 统计学方式识别，第 2 版，学术出版社，San Diego; Everitt, 1974, 聚类分析，伦敦：海曼教科书；Hartigan, 1975, 聚类算法，纽约，Wiley; Sneath 和 Sokal, 1973, 数字分类学，Freeman; Anderberg, 1973, 应用的

聚类分析，学术出版社，纽约。

在某些采用聚类分析的实施方案中，大量基因的表达是作为受到多种扰动影响的生物系统来监测的。包括有基因表达测定值的数据表格被用于聚类分析。为了获得包括能在多种条件下协变的基因的基础基因集，采用了多种扰动或条件。对具有 $m \times k$ 项的数据的表格进行聚类分析，其中， m 是条件或扰动的总数，而 k 是测定基因的数量。

有多种聚类算法可用于聚类分析。在形成聚类时，聚类算法利用主体之间的不相同性或距离。在某些实施方案中，所使用的距离是多维空间的欧几里德距离：

10

$$I(x, y) = \left\{ \sum_i (X_i - Y_i)^2 \right\}^{1/2} \quad (14)$$

其中， $I(x, y)$ 是基因 X 和 Y 之间的距离， X_i 和 Y_i 是在扰动 i 的作用下的基因表达响应。可以对所述欧几里德距离进行平方，以便在进一步分离的目标上逐渐施加更大的权数。另外，所述距离测定可以是曼哈顿距离，例如，基因 X 和 Y 之间的距离，该距离是用以下公式提供的

15

$$I(x, y) = \sum_i |X_i - Y_i| \quad (15)$$

20

同样， X_i 和 Y_i 是在扰动 i 作用下的基因表达。某些其他的距离定义可以是 Chebchev 距离，乘幂距离、和百分不符合性。百分不符合性被定义为 $I(x, y) = (X_i \neq Y_i \text{ 的数量}) / i$ ，它特别适用于本发明的方法，如果对所述各项的数据进行性质分类的话，另一种特别适用于细胞响应场合的有用的距离定义为 $I = 1 - r$ ，其中， r 是响应矢量 X 、 Y 之间的相关系数，又被称为标准化点乘积 $X \cdot Y / |X| |Y|$ 。

25

各种聚类连接规则可用于定义基因集。单一连接-最近邻方法能测定两个最近主体之间的距离。相反，完全连接方法可以通过不同聚类中的任意两个主体之间的最大距离测定距离。该方法特别适用于当基因或其他细胞成分形成天然的独特“堆积”的场合。另外，非加权

成对的平均确定两个不同聚类中所有主体对之间的平均距离。该方法同样非常适用于基因或其他细胞成分的聚类，以便形成天然的独特堆积。最后，还可以使用加权重对的平均方法。该方法与非加权平均成对方法相同，所不同的是将相应聚类的大小用作一个权数。该方法特别适用于其聚类大小容易发生大的变化的实施方案（Sneath 和 Sokal, 1973, 数字统计学, 旧金山: W.H. Freeman&Co.）。其他聚类连接规则，如非加权和加权重对重心和 Ward's 方法也可用于本发明的某些实施方案，例如，参见 Ward, 1963, 美国统计学会杂志 58: 236; Hartigan, 1975, 聚类算法, 纽约: Wiley.

随着聚类集中扰动种类的增加，明确可分辨的基因集变得更小数量也更多。不过，即使在大的实验组中，仍有小的基因集保留其一致性。这些基因集被称为不可诱导的基因集。通常，施加大量不同的扰动，以便获得所述可诱导的基因集。

通常，基因集的聚类是通过图形表示的，并称为‘树’。基因集可以根据一棵树的许多较小的分支确定，或者通过在不同的高度上切割该树形成的少量的较大的分支。切割高度的选择可以与预期的不同响应途径的数量匹配。如果少有或没有可供利用的有关途径数量的信息，那么应当将所述树分割成与真实差别一样多的分支。“真实差别”可以被定义为单个分支之间的最小距离值。典型的值在 0.2-0.4 范围内，其中 0 是完全相关，而 1 是零相关，不过，对于较差质量的数据或所述训练组中较少的实验来说可以更大一些，或者对于较好的数据和在所述训练组中较多的实验来说可以更小一些。

优选的是，“真实差别”可以用所述大树上的每一个分叉点的统计学显著性的客观试验来确定。在本发明的一个方面，将所述实验组中的每一种细胞成分响应的实验指数的 Monte Carlo 随机化用于确定客观试验。

在某些实施方案中，所述客观试验按以下方式确定：

设定 p_{ki} 为细胞成分 k 在实验 i 中的响应。设定 $II(i)$ 为该实验指数的随机扰动。然后对大量（大约 100-1000）不同随机扰动的每一种建立 $p_{kII(i)}$ 。对于所述原始树的每一个分支来说，对于每一种扰动：

(1) 用与在原始非排列数据上所使用的相同算法（在这种情况下

下为 ‘hclust’) 进行谱系聚类;

(2) 计算在从一个聚类到两个聚类的过程中相应于聚类中心的总的分散的分数改进 f

$$f = 1 - \Sigma D_k^{(1)} / \Sigma D_k^{(2)} \quad (16)$$

5

其中, D_k 是成分 k 相对所确定的其聚类中心 (平均值) 的距离测定值的平方。上标 1 或 2 表示它是相对于整个中心的分支或者相对于两个亚类中的一个合适亚类的中心。用于所述聚类方法的距离函数 D 的定义有相当大的自由度。在这些例子中, $D=1-r$, 其中, r 是所述实验组中一种成分的响应与另一种成分的响应 (或与平均聚类响应) 之间的相关系数。

从 Monte Carlo 方法中所获得的分数改进的分布是该分布在设定分支为不显著的无效假设条件下所述分布的估计值。然后具有所述未排列数据的分支的实际分数改进与来自所述无效假设的累计可能性分布加以比较, 以便确定其显著性。通过将所述无效假设分布代入对数标准模型求出标准误差。使用该方法大于大约 2 的标准误差表示所述分支的显著性的置信水平为 95%。通过聚类分析确定的基因集通常具有潜在的生物学显著性。

聚类分析方法的另一方面提供了用于在下面章节中所披露的特征投射的基础矢量的定义。

一组基础矢量 V 具有 $k \times n$ 项, 其中, k 是基因数量, n 是基因集数量

$$V = \begin{bmatrix} V_1^{(1)} & \cdot & V_1^{(n)} \\ \cdot & \cdot & \cdot \\ V_k^{(1)} & \cdot & V_k^{(n)} \end{bmatrix} \quad (17)$$

$V^{(n)}_k$ 是基础矢量 n 中基因指数 k 的贡献幅度。在某些实施方案中, 如果 k 是基因集 n 的一页的话, $V^{(n)}_k=1$, 如果 k 不是基因集 n 的一页的话, $V^{(n)}_k=0$ 。在某些实施方案中, $V^{(n)}_k$ 与基因集 n 中基因 k 在用于确定该基因集的训练数据组中的反应成比例。

在某些优选实施方案中，因素 $V_k^{(n)}$ 是规范化的，因此每一个 $V_k^{(n)}$ 具有除以基因集 n 中的基因数量的平方根的单位长度。由此所产生的基础矢量不仅是正交的（通过分割聚类树所产生的基因集是分离的），而且是标准化的（单位长度）。通过选择这种规范化，在投射到 $V_k^{(n)}$ 上的特征的随机测定误差其大小倾向于对每一个 n 来说是相当的。规范化能避免大的基因集支配相似性计算的结果。

5.6.3 基于调控机制的基因集分类

基因集还可以根据基因的调控机制确定。其调控区具有相同的转录因子结合位点的基因更倾向于是共调节的。在某些优选实施方案中，用多重排比分析比较感兴趣基因的调控区，以便解释可能共有的转录因子结合位点（Stormo 和 Hartzell, 1989, 鉴定非排比 DNA 片段的蛋白结合位点，美国科学院院报 86: 1183-1187; Hartz 和 Stormo, 1995, 鉴定非排比 DNA 和蛋白序列上的共有方式：触发缺口的大误差统计学基础，第 3 届生物信息和基因组研究国际会议公报，Lim 和 Cantor 著，世界科学出版有限公司，新加坡，201-216 页）。例如，下文的例 3 所示，在 20 个基因上对 Gcn4 有响应的共同的启动子序列可能负责这 20 个基因在多种扰动下的共调节。

所述基因的共调节不局限于具有针对相同转录因子的结合位点的基因。共调节（协变）基因可以处于上游/下游关系，其中，上游基因的产物能调节下游基因的活性。本领域技术人员所熟知的是，具有多种基因调控网络。本领域技术人员还能理解的是，本发明的方法不限于任何特定类型的基因调控机制。如果它来自所述调控机制，就是说两个基因在其针对扰动的活性改变方面是共调节的话，可以将这两个基因聚类成一个基因集。

由于对感兴趣的基因的调控缺乏全面了解，通常优选调控机制知识合并聚类分析，以便得到更好的确定的基因集。在某些实施方案中，当感兴趣的基因的调控是部分已知的时，可将 K-平均值聚类用于聚类基因集。在基因集的数量是通过了解其调控机制而预定的情况下是特别有用的。一般，K-平均值聚类被局限于准确地产生所希望的聚类数量。因此，如果启动子序列比较表明测定的基因应该分成 3 个基因集，可以用 K-平均值聚类准确地产生 3 个基因集，在各聚类之间具有最大可能的差别。

5.6.4 表示投射的特征

可以将基因表达值转换成基因集的表达值。该过程被称为投射。在某些实施方案中，该投射过程如下：

$$P = [P_1, \dots, P_i, \dots, P_n] = p \bullet V \quad (18)$$

5

其中， p 是表达特征， P 是投射特征， P_i 是基因集 i 的表达值， V 是预定的基础矢量集。该基础矢量是事先在公式 17 中确定的：

$$V = \begin{bmatrix} V_1^{(1)} & \dots & V_1^{(n)} \\ \vdots & \ddots & \vdots \\ V_k^{(1)} & \dots & V_k^{(n)} \end{bmatrix} \quad (19)$$

10 其中， $V^{(n)}_k$ 是基础矢量 n 的细胞成分指数 k 的幅度。

在一种优选实施方案中，基因集的表达值是该基因集中所有基因表达值的简单平均值。在某些其他实施方案中，对所述平均值进行加权，以便高效表达的基因不会支配该基因值。该基因集表达值的综合就是所述投射特征。

15 5.6.5 特征比较和分类

一旦选择所述基础基因集，就可以获得任一组特征指数 i 的投射特征 P_i 。 P_i 之间的相似性比原始特征 p_i 之间的相似性看得更清楚，这是由于两种原因。首先，外来基因的测定误差已经被排除或平均了。其次，基础基因集倾向于捕获特征 P_i 的生物学，其单一响应因素的匹配的探测器也是如此。特征的分类和聚类都是基于一种客观相似性参数 S ，其中，一个有用的定义为

20

$$S_{ij} = S(P_i, P_j) = P_i \bullet P_j / (\|P_i\| \|P_j\|) \quad (20)$$

该定义为矢量 P_i 和 P_j 之间的一般化角余弦。它是 p_i 和 p_j 之间的常规相关系数的投射形式。特征 p_i 似乎最近似于其 S_{ij} 最大的特征 p_j 。新特征可以按照其与具有已知生物学价值的特征的相似性来分

25

类，如在特定生物学途径中对已知药物或扰动的响应形式。可以用以下距离参数对新特征组进行聚类

$$D_{ij} = 1 - S_{ij} \quad (21)$$

- 5 其中，该聚类类似于整套响应测定的原始较大空间的聚类，但具有刚才所提到的优点，即降低测定误差影响，并提高相关生物学的获得。

任何观察到的相似性 S_{ij} 的统计学显著性可以用在不相关的无效假定下产生的经验可能性分布进行评估。该分布是通过实施所述投射产生的，将公式 (19) 和 (20) 用于所述原始特征 p 中的成分指数的许多不同随机排列。就是说，用 $p_{II(k)}$ 取代有序的集 p_k ，其中， $II(k)$ 是排列，有大约 100-1000 种不同的随机排列。随机产生相似性 S_{ij} 的可能性为这些排列的分数，这些排列的相似性 S_{ij} (排列过的) 超过用原始非排列数据所观察到的相似性。

5.7 测定生物学响应特征的方法

- 15 这一部分提供了用于测定生物学响应的某些代表性方法，以及制备用于所述方法的试剂所必需的方法。

5.7.1 微排列的制备

微排列为本领域所公知，并包括一个表面，在序列上与基因产物 (例如，cDNA、mRNA、cRNA、多肽及其片段) 相关的探针可以专一性地杂交或结合在该表面上的一个已知点。在一种实施方案中，所述微排列是一种排列 (即一种基质)，其中，每一个位点代表一个用于结合由一个基因所编码的产物 (例如蛋白或 RNA) 的独立结合位点，并且，其中存在有关生物基因组的大部分或几乎所有基因产物的结合位点。在一种优选实施方案中，所述“结合位点” (以下称之为位点) 是核酸或核酸类似物，一种特定的同源 cDNA 可以与它专一性地杂交。例如，所述结合位点的核酸或类似物可以是合成寡聚体、完整长度的 cDNA、不够完整长度的 cDNA、或基因片段。

25 尽管在优选实施方案中所述微排列包括目标生物基因组中所有或者几乎所有基因产物的结合位点，但这种理解并不是必须的。通常，所述微排列所具有的结合位点相当于所述基因组中基因的至少大

约 50%，通常至少大约 75%，更通常至少大约 85%，更通常高于大约 90%，最通常至少大约 99%。优选的是，所述微排列具有与感兴趣的药物或感兴趣的生物学途径的作用相关基因的结合位点。“基因”是优选具有至少 50、75、或 99 个氨基酸的开放读框（ORF），在生物（例如，如果是单细胞的话，）或在多细胞生物的某些细胞中由它转录成 mRNA。一个基因组中基因的数量可以根据该生物所表达的 mRNA 的数量来估计，或者根据该基因组业已鉴定过的部分来推测。当业已对感兴趣生物的基因组进行测序之后，可以通过分析其 DNA 序列确定 ORF 的数量，并鉴定 mRNA 编码区。例如，业已对酿酒酵母基因组做过全面测序，并报导具有大约 6275 个开放读框，其长度超过 99 个氨基酸。对这些 ORF 进行的分析表明，有 885 个 ORF 有可能编码蛋白产物（Goffeau 等，1996，生命，有 6000 个基因，科学 274: 546-567，以上文献被全文收作各个方面的参考）。相反，估计人类基因组包括大约 10^5 个基因。

5.7.2 制备微排列的核酸

如上文所述，特定同源 cDNA 能专一性杂交的“结合位点”通常是连接在所述结合位点上的核酸或核酸类似物。在一种实施方案中，所述微排列的结合位点是相应于一种生物基因组中每一个基因的至少一部分的 DNA 多核苷酸。例如，所述 DNA 可以通过聚合酶链式反应（PCR）扩增来自基因组 DNA 的基因片段、cDNA（例如，通过 RT-PCR）或克隆序列获得。PCR 引物是根据所述基因或 cDNA 的已知序列选择的，它能导致特殊片段的扩增（即与所述排列上任何其他片段都没有 10 个以上碱基的连续相同序列的片段）。计算机程序可用于设计具有所需专一性和最佳扩增特性的引物。对于相应于极长基因的结合位点来说，有时候需要扩增该基因 3'末端的片段，因此当寡聚 dT 启动的 cDNA 探针与所述微排列杂交时，不够完整长度的探针将能够有效结合。通常，所述微排列上的每一个基因为大约 50-大约 2000bp，更通常为大约 100-大约 1000bp，一般为大约 300-大约 800bp 长度。PCR 方法是众所周知的，并且披露于 Innis 等的著述中，1990，PCR 方法：方法和应用指南，学术出版公司，San Diego，CA，该书被全文收作本文各个方面的参考。制备所述微排列的核酸的另一种方法是合成合成多核苷酸或寡核苷酸，例如，用 N-磷酸或亚磷酰胺化合法合成

(Froehler 等, 1986, 核酸研究 14: 5399-5407; McBride 等, 1983, Tetrahedron Lett. 24: 245-248)。合成序列的长度为大约 15-大约 100 个碱基, 更通常为大约 20-大约 50 个碱基。在某些实施方案中, 合成核酸包括非天然碱基, 例如肌苷。如上文所述, 核酸类似物可以

5 用作杂交的结合位点。合适的核酸类似物的一个例子是肽核酸 (例如, 参见 Egholm 等, 1993, 遵循沃森克里克氢结合规则的 PNA 与互补寡核苷酸的杂交, 自然 365: 566-568; 同样参见 US5, 539, 083)。

在另一种实施方案中, 所述结合 (杂交) 位点是由质粒噬菌体克隆、cDNA (例如, 表达序列标记)、或其插入片段制备的 (Nguyen

10 等, 1995, 通过排列 cDNA 克隆的定量杂交测定鼠胸腺中差异基因表达, 基因组学 29: 207-209)。在另一种实施方案中, 所述结合位点的多核苷酸是 RNA。

5.7.3 将核酸连接在固体表面上

将核酸或类似物连接在固体表面支持物上, 该支持物可以是用玻

15 璃、塑料 (例如, 聚丙烯、尼龙)、聚丙烯酰胺、硝酸纤维素、或其他材料制成。一种用于将核酸连接在表面上的优选方法是通过印染到玻璃板上而实现的, 总体上如 Schena 等所述 (1995, 用互补微排列定量监测基因表达方式, 科学 270: 467-470)。该方法特别适用于制备 cDNA 微排列。还可参见 DeRisi 等, 1996, 用微排列分析人类癌

20 症的基因表达方式, 自然遗传学 14: 457-460; Shalon 等, 1996, 用于通过双色荧光探针杂交分析复杂 DNA 样品的微排列系统, 基因组研究 6: 639-645; 和 Schena 等, 1995, 平行人类基因组分析; 1000 个基因的基于微排列的表达, 美国科学院院报 93: 10539-11286。

制备微排列的第二种优选方法是通过制备高密度寡核苷酸排列

25 而进行的。用原位合成的照相平板印刷技术 (参见 Fodor 等, 1991, 光学定向空间可寻质平行化学合成, 科学 251: 767-773; Pease 等, 1994, 用于快速 DNA 序列分析的光学定向寡核苷酸排列, 美国科学院院报 91: 5022-5026; Lockhart 等, 1996, 通过与高密度寡核苷酸杂交监测表达, 自然生物技术 14: 1675; US5, 578, 832; 5, 556, 752;

30 和 5, 510, 270, 以上所有文献均以全文形式收作本文各方面的参考) 或用于快速合成和沉积特定寡核苷酸的其他方法 (Blanchard 等, 1996, 高密度寡核苷酸排列, 生物传感器和生物电子学 11: 687-90)

在一个表面的特定部位产生含有数千个互补于特定序列的寡核苷酸的技术是公知的。当采用以上方法时，在诸如衍生化玻璃载玻片的表面上直接合成具有已知序列的寡核苷酸（例如，二十聚体）。通常，所产生的排列含有针对每一种目标转录物的多种探针。可以选择寡核苷酸探针，以便检测剪接过的 mRNA 或是用作各种类型的对照。

制备微排列的另一种优选方法是使用喷墨印刷方法直接在固体相上合成寡核苷酸，例如，在以下专利文献中所披露的：待批美国专利申请流水号 09/008,120，申请日为 1998 年 1 月 16 日，Blanchard 等，题为“用溶剂微滴进行化学合成”，该专利被以全文形式收作本文参考。

还可以使用用于制备微排列的其他方法，例如，掩饰方法（Maskos 和 Southern, 1992, 核酸研究 20: 1679-1684）。原则上讲，可以使用任何类型的排列，例如，存在于尼龙杂交膜上的斑点印迹（例如，参见 Sambrook 等，分子克隆-实验室手册，第 2 版，1-3 卷，冷泉港实验室，冷泉港，纽约，1989），不过，正如本领域技术人员所了解的，优选极小的排列，因为小排列的杂交体积也较小。

5.7.4 制备标记过的探针

用于制备总的和 poly(A)+RNA 的方法是众所周知的，并总体上披露于 Sambrook 等的著述中（同上），在一种实施方案中，RNA 是从本发明所感兴趣的各種类型的细胞中提取的，用硫代氰酸胍裂解，然后用氯化铯离心（Chirgwin 等，1979，生物化学 18: 5294-5299）。poly(A)+RNA 是通过用寡聚 dT 纤维素筛选选择的（参见 Sambrook 等，同上）。感兴趣的细胞包括野生型细胞、药物处理过的野生型细胞、修饰过的细胞、和药物处理过的修饰细胞。

通过寡聚 dT 启动的或随机启动的逆转录用 mRNA 制备标记过的 cDNA，这两种方法都是本领域所熟知的（例如，参见 Klug 和 Berger, 1987，酶学方法 152: 316-325）。逆转录可以在与可检测的标记缀合的 dNTP 存在下进行，最优选荧光标记的 dNTP。另外，通过在标记过的 dNTP 存在的条件下进行双链 cDNA 的体外转录将分离的 mRNA 转变成标记过的反义 RNA（Lockhart 等，1996，通过与高密度寡核苷酸排列杂交监测表达，自然生物技术 14: 1675，该文献以全文形式收作本文各方面的参考）。在另一种实施方案中，可以在缺少可检测

标记的条件下合成 cDNA 或 RNA 探针，并可以随后进行标记，例如通过整合生物素化的 dNTP 或 rNTP，或某些类似方法（例如，生物素的补骨质素衍生物 RNA 的光交联），然后添加标记过的链霉亲和素（例如，藻红素缀合的链霉亲和素）或其他等同物。

- 5 当使用荧光标记的探针时，有许多合适的荧光团是已知的，包括荧光素、丽丝胺、藻红素、罗丹明（Perkin Elmer Cetus），Cy2、Cy3、Cy3.5、Cy5、Cy5.5、Cy7，FluorX（Amersham）及其他物质（例如，参见 Kricka, 1992, 非同位素 DNA 探针技术，学术出版社，San Diego, CA）。可以理解的是，选择具有不同发射光谱的成对的荧光团，以便容易加以区分。

- 10 在另一种实施方案中，采用除荧光标记以外的标记，例如，可以使用放射性标记、或具有不同发射光谱的成对的放射性标记（参见 Zhao 等，1995，高密度 cDNA 过滤分析：大规模定量分析及表达的新方法，基因 156: 207; Pietu 等，1996，通过高密度 cDNA 排列的定量杂交发现的在人类肌肉中优势表达的新型基因转录物，基因组研究 6: 492）。不过，由于放射性颗粒的扩散，以及随后需要使用大空间的结合位点，因此使用放射性同位素是不太理想的实施方案。

- 15 在一种实施方案中，标记过的 cDNA 是通过将含有 0.5mM dGTP、dATP 和 dCTP+0.1mM dTTP+荧光脱氧核糖核苷酸（例如，0.1mM 罗丹明 110（Perkin Elmer Cetus）或 0.1mM Cy3 dUTP（Amersham））的混合物与逆转录酶（例如，SuperScript™II，LTI 公司）一起在 42℃ 下培养 60 分钟而合成的。

5.7.5 与微排列杂交

- 25 对核酸杂交和洗涤条件进行优化选择，以便探针能与特定排列位点进行“专一性结合”或“专一性杂交”，即所述探针能与具有互补核酸序列的序列排列位点杂交、形成双链或结合，但不能与具有非互补核酸序列的部位杂交。这样的多核苷酸序列被视为与另一种多核苷酸序列互补：如果较短的多核苷酸低于或等于 25 个碱基，使用标准碱基配对规则没有错配的话，或者如果较短的多核苷酸长度超过 25 个碱基，其错配不超过 5% 的话。优选的是，所述多核苷酸是完全互补的（无错配）。通过进行包括负对照在内的杂交实验可以方便地证实能产生特异杂交的特殊杂交条件（例如，参见 Shalon 等，同上，

和 Chee 等, 同上)。

最佳杂交条件取决于标记过的探针和固定的多核苷酸或寡核苷酸的长度(例如, 寡聚体与多核苷酸长于 200 个碱基)和类型(例如, RNA、DNA、PNA)。核酸的特定(即严格)杂交条件的一般参数披露于 Sambrook 等的著述中(同上), 以及 Ausubel 等的著述中(1987, 当代分子生物学方法, Greene Publishing 和 Wiley-Interscience, 纽约)。当使用 Schena 等的微排列时, 典型的杂交条件为在 65℃ 下在 5XSSC+0.2%SDS 中杂交 4 小时, 然后在 25℃ 下在低严格洗涤缓冲液(1XSSC+0.2%SDS)中洗涤, 再在 25℃ 下在严格洗涤缓冲液(0.1XSSC+0.2%SDS)中洗涤 10 分钟(Shena 等, 1996, 美国科学院报, 93: 10614)。有用的杂交条件还披露于例如, Tijessen 的著述中(1993, 与核酸探针的杂交 Elsevier 科学出版社)和 B.V. 和 Kricka 的著述中(1992, 非同位素 DNA 探针技术, 学术出版社, San Diego, CA)。

5.8. 计算机实施

在以上各部分所披露的分析方法可以用下面的计算机系统并按照下面的程序和方法完成。图 7 表示适于执行本发明分析方法的代表性计算机系统。所示计算机系统 501 包括内部元件并与外部元件连接。该计算机系统的内部元件包括与主存储器 503 相互连接的处理器元件 502。例如, 计算机系统 501 可以是 Intel8086-, 80386-, 80486-, 奔腾®、或基于奔腾的处理器, 优选具有 32MB 或更高的主存储器。

所述外部元件包括主记忆装置 504。该主记忆装置可以是一个或多个硬盘(它通常与所述处理器或存储器包装在一起)。所述硬盘优选具有 1GB 或更大的储存能力。其他外部元件包括用户界面装置 505, 它可以是一组监视器, 同时还有一个输入装置 506, 它可以是一个“鼠标”, 或其他图形输入装置(未示出), 和/或键盘。打印装置 508 也可以与计算机 501 连接。

通常, 计算机系统 501 还与网络链接 507 连接, 它可以是与其他局部计算机系统、远程计算机系统、或宽域连接网络, 如互联网连接的 Ethernet 的一部分。该网络连接使得计算机系统 501 能与其他计算机系统共享数据和处理任务。

在操作该系统期间将若干软件成分加载到储存器中，所述软件既是本领域标准化的又是本发明所特有的。这些软件成分共同导致该计算机系统按照本发明的方法工作。所述软件成分通常储存在主记忆装置 504 上。软件成分 510 表示有关操作系统，它复杂管理计算机系统 501 及其网络相互连接。例如，该操作系统可以是微软 Windows' 家族，如 Windows3.1，Windows95，Windows98，或 WindowsNT。软件成分 511 表示通常存在于该系统上的通用语言和功能，以便协助程序实施本发明特有的方法。许多高级或低级计算机语言可用于对本发明的分析方法进行编程。指令可以在运行期间翻译或编辑。优选的语言包括 C/C++、FORTRAN 和 JAVA®。最优选的是，本发明的方法用数学软件包进行编程，该软件包能够通过符号输入公式和处理的高级说明，包括要使用的算法，从而使用户没有必要程序性地对单一的公式或算法进行编程。所述软件包包括 Mathworks (Natick, MA) 的 Matlab, Wolfram Research 的 Mathematical (Champaign, IL), 或 Math Soft 的 S-Plus (Cambridge, MA)。因此，软件成分 512 和/或 513 表示用程序语言或符号包编程的本发明的分析方法。在一种代表性实施方案中，为了实施本发明的方法，用户首先将差异微排列实验的数据输入计算机系统 501。这些数据可以由用户从监视器 505、键盘 506，或从通过网络连接 507 连接的其他计算机系统，或诸如 CD-ROM、软盘（未示出）、磁带 drive（未示出）、ZIP®drive（未示出）或储存介质上或通过网络（507）直接输入。然后，用户开始执行表达特征分析软件 512，由该软件完成本发明的方法。

在另一种代表性实施方案中，用户首先将微排列实验数据输入所述计算机系统。该数据是从储存介质（504）或从远程计算机，优选从动态基因集数据库系统通过网络（507）输入所述存储器的。然后，用户开始启动软件，由它实施荧光团误差消除、本发明的基于等级的方法或本发明的加权平均方法的步骤。

用于实施本发明的分析方法的其他计算机系统 and 软件对本领域技术人员来说是显而易见的，并且被视为包括在所附权利要求书范围内。具体地讲，所附权利要求书被视为包括用于实施本发明方法的其他程序结构，这些程序结构对本领域技术人员来说是显而易见的。

6 实验

以下部分详细说明了制备用于图 2-6 所示实验所用试剂的方法。

酵母菌株的构建、生长和药物处理

用于本研究的菌株是用标准技术构建的。例如，参见 Schiestl
5 等，1993，通过转化将 DNA 导入酵母，方法：酶学方法手册 5: 79-
85。对于涉及 FK506 的实验来说，按照以前由 GARRETT-Engle 等所
披露的方法（1995，钙调磷酸酶， Ca^{2+} /钙调蛋白依赖型蛋白磷酸酶是
具有细胞完整性缺陷的酵母突变体和缺乏有功能的液泡 H^{+} -
ATPase 的突变体所必需的，分子细胞生物学 55: 4103-4114），让
10 细胞在补充了 10mM 氯化钙的 YAPD 培养基（YPD+0.004%腺嘌呤）中
生长 3 代，使密度达到 1×10^7 细胞/毫升。在接种所述培养物 0.5 小
时之后按照说明添加 FK506 达到 1 微克/毫升的终浓度。添加环孢菌
素 A (CsA) 使浓度为 30 微克/毫升。用具有以下改进的标准方法破
碎细胞（例如，参见 Ausubel 等，当代分子生物学方法，John
15 Wiley&Sons 公司，纽约，12.12.1-13.12.5）。将细胞沉淀重新悬浮
在裂解缓冲液（0.2mM Tris-HCl, pH7.6, 0.5M 氯化钠，10mMEDTA,
1%SDS），在设定为 8 的条件下在存在 60%玻璃珠（425-600 微米目；
Sigma）和苯酚：氯仿（50: 50, v/v）的条件下，在一台 VWR 多试
管涡旋搅拌器上涡旋搅拌 2 分钟。在分离之后对水相进行重新提取，
20 并用乙醇沉淀。通过在寡聚 dT 纤维素（NEB）上用已建立的方法通过
两次连续的层析纯化分离 poly(A)+RNA。例如，参见 Ausubel 等，
同上。

标记样品的制备和杂交

大体上按 DeRisi 等所披露的方法（1997，在基因组规模上研究
25 基因表达的代谢和遗传学控制，科学 278: 680-686）制备、纯化并
杂交荧光标记的 cDNA。简单地讲，在逆转录（SuperscriptII, LTI
公司）期间将 Cy3-或 Cy5-dUTP（Amersham）结合到 cDNA 中，并通
过用 Microcon-30 微量浓缩仪（Amicon）浓缩到少于 10 微升进行纯
化。将成对的 DNA 悬浮在 20-26 微升杂交溶液（3XSSC, 0.75 微克/
30 毫升 poly(A)DNA, 0.2%SDS）中，并加样到所述微排列上，用 22X30mm
的盖玻片覆盖，在 63℃ 下放置 6 小时，所有操作均按 DeRisi 等所述
进行（1997），同上。

微排列的制备和扫描

用含有共同的 5'和 3'序列的 PCR 产物 (Research Genetics) 作模板, 使用氨基修饰的正向引物和未修饰过的反向引物 PCR 扩增来自酿酒酵母基因组的 6065 个 ORF。第一代的成功率为 94%。得到出乎预料大小产物的扩增反应在随后的分析中被排除在外。不能用购买的模板扩增的 ORF, 用基因组 DNA 扩增。将来自 100 微升反应物的 DNA 样品用异丙醇沉淀, 重新悬浮在水中, 用 3XSSC 调整到总体积为 15 微升, 然后转移到 384 孔微量滴定板上 (Genetix)。按照以下文献的说明用机器人结构将 PCR 产物印染到 1X3 英寸的聚赖氨酸-处理过的玻璃载玻片上 (Schena 等, 同上; DeRisi 等, 1996, 用微排列发现并分析与炎症疾病相关的基因, PNAS USA, 94: 2150-2155; 和 DeRisi 等, 1997)。在印染之后, 按照公开方法处理载玻片。参见 DeRisi 等, 1997。

用 Applied Precision 公司 (Seattle, WA) 正在开发的原型多框 CCD 相机对微排列进行照相。每一个 CCD 相框大约为 2 毫米见方。在从一个相框转移到下一个空间上连续的相框之前, 对每一个相框上在 Cy5 通道中暴光 2 秒 (白色光线通过 Chroma618-648nm 激发过滤器, Chroma657-727nm 发射过滤器), 并在 Cy3 通道中暴光 1 秒钟 (Chroma535-560nm 激发过滤器, Chroma570-620nm 发射过滤器)。Cy3 和 Cy5 通道之间的颜色分离为大约 100: 1 或更好。在软件中将所述相框编制在一起, 以便构成完整图象。通过框背景扣除和每一个通道中的强度平均, 用 10 微米的象素对斑点 (大约 100 微米) 强度进行定量。所得斑点强度的动态范围通常为最亮的斑点和背景扣除累计误差水平之间的千分之一。所述通道之间的规范化是通过将每一个通道对所有基因的平均强度进行规范而实现的。这一方法几乎与使用基因组 DNA 斑点强度定额的通道之间的规范化相同 (参见 DeRisi 等, 1997), 但可能更可靠一些, 因为它是基于分布在所述排列上的数千个斑点的强度做出的。

确定标记相关系数及其置信度极限

用以下公式计算各个实验的标记 ORF 之间的相关系数

$$\rho = \sum_k x_k y_k / (\sum_k x_k^2 \sum_k y_k^2)^{1/2}$$

其中， x_k 是在 x 标记中第 k 个基因的表达比例的 $\log_{10}$ ，而 y_k 是 y 标记中的第 k 个基因的表达比例的 $\log_{10}$ 。其总和是用在任一个实验中以 95% 的置信度水平被上调或下调了的基因求出的。这些基因各自具有不到 5% 的几率被实际上调（仅仅由于测定误差而使表达比例偏离一致性）。该置信度水平是根据误差模拟确定的，它给每一个基因表达比例确定一个对数正态可能性分布，其特征宽度基于在其重复测定中观察到的分散值（在相同标称实验条件下的重复排列）并基于单一排列杂交质量。后一种依赖性来自对照实验，其中，Cy3 和 Cy5 样品来自相同的 RNA 样品。对于大量重复测定来说，误差降低到观察到的分散值。对于单一测定来说，误差是基于排列的质量和斑点强度。

在 x 和 y 标记中的随机测定误差倾向于使相关性偏向于 0。在大多数实验中，大多数基因没有受到明显影响，但确实表现出小的随机测定误差。仅选择 95% 置信度的基因用于相关性计算。而不是用整个基因组计算，能降低该误差并使得实际生物学相关性更明显。

通过定义使特征和它本身之间的相关性统一。根据单一测定误差线和假设的不相关误差，相关性误差的极限为 95% 的置信度极限。它不包括上面提到的误差；因此， ρ 与统一性的偏移并不一定意味着潜在的生物学相关性是不完美的。不过，例如， 0.7 ± 0.1 的相关性与 0 的不同是极显著的。在表格和文本中的小的（幅度为 $\rho < 0.2$ ）但正式的显著相关性可能是由于在 Cy5/Cy3 比例上的小的系统误差所导致的。它证实了用于产生 95% 置信度极限的独立测定误差的假设。因此，这些小的相关值应当被作为不显著处理。未纠正过的系统误差的一个可能的来源是它有差别地影响 Cy3 和 Cy5 检测通道。

将 1 微升/毫升 FK506 处理标记与 40 个以上不相关的缺失突变型或药物标记比较。所述对照的特征与 FK506 特征的相关系数分布在 0 左右（平均 $\rho = -0.03$ ），标准误差为 0.16（数据未发表），并且没有一个相关性大于 $\rho = 0.38$ 的。类似地，钙调磷酸酶突变型标记与 CsA-处理标记十分相关（ $\rho = 0.71 \pm 0.04$ ），但不与负对照标记相关（平均 $\rho = -0.02$ ，标准误差为 0.18）。

质量控制

通过分析在重复杂交中的误差不断检测表达比例的测定精确性，所述杂交使用 Cy3 和 Cy5 标记过的同一种 mRNA 进行，还可以用从相关标称菌株和条件的独立培养物中的分离的 Cy3 和 Cy5 mRNA 样品进行。在该方法中的检测到的误差，如基因专一性误差可能是由于 Cy3 和 Cy5-dUTP 结合到 cDNA 中的差异所导致的，该误差可以通过成对地进行颠倒荧光团的杂交而降低，其中，在与一个实验相应的另一个实验中，生物学条件的 Cy3/Cy5 标记颠倒了。每一个基因的表达比例就是成对的两个实验之间的比例。通过算法数字 detrending 消除其他误差。在缺乏 detrending 和荧光团颠倒的条件下，所述误差的幅度通常在 30% 比例的数量级上，但对某些 ORF 来说也可以高出 2 倍。

表达比例是基于每一个点上的平均强度。平均而言，偶然出现的较小的斑点具有较少的像素。只要像素的数量不低于 10，就不会明显降低精确性，在这种情况下，将所述斑点从数据集中排除。通过图象处理软件在排列亚区上适当跟踪斑点位置相对标称网格的移动。对于自动化定量算法来说，在一个亚区中不均等的斑点移动超过斑点空间的一半会出现问题；在这种情况下，对所述移动的人工检查将该斑点从分析中排除。从所述数据集中排除所有部分重叠的斑点。通常有不到 1% 的斑点因为以上原因而被排除。

7. 所引用的参考文献

本文所引用的所有文献以其全文形式收作本文参考，在任何情况下都以相同的程度作为参考，就如同每一份文献或专利或专利申请被专门和单独以其全文收作本文参考一般。

正如本领域技术人员可以理解的是，在不脱离其实质和构思的前提下可以对本发明作出各种改进和改变。本文所披露的具体的实施方案仅仅是用于举例的，而本发明只是由所述权利要求书来限定的，以及等同于所述权利要求所包括的所有范围。

说明书附图

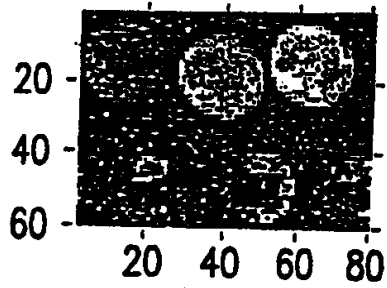


图 1A

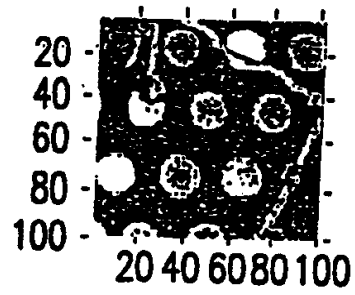


图 1B

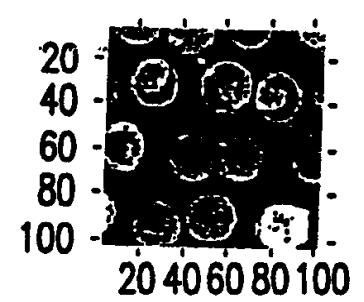


图 1C

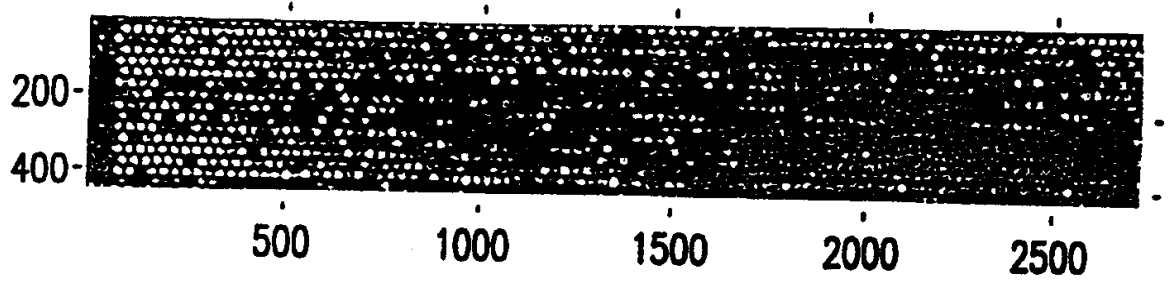


图 1D

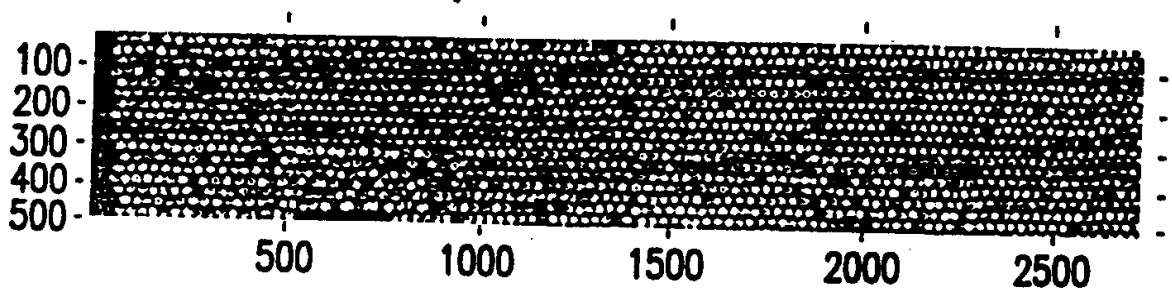


图 1E

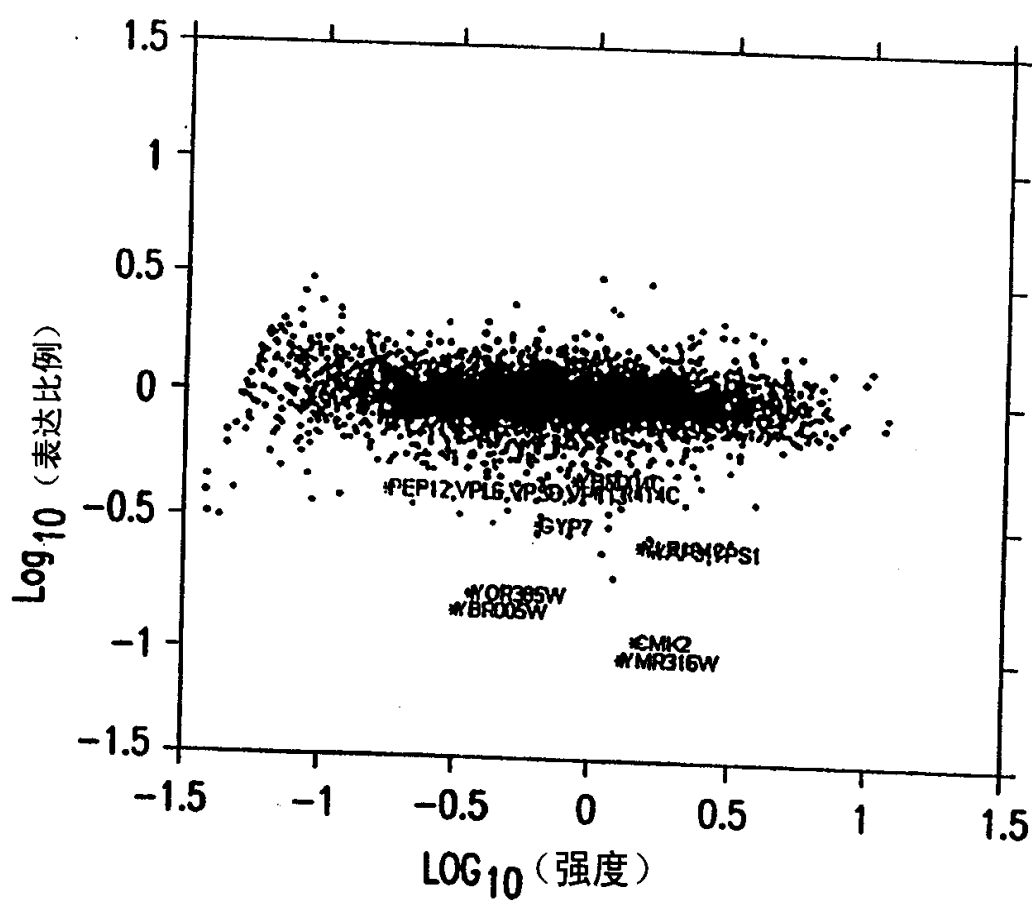


图 2

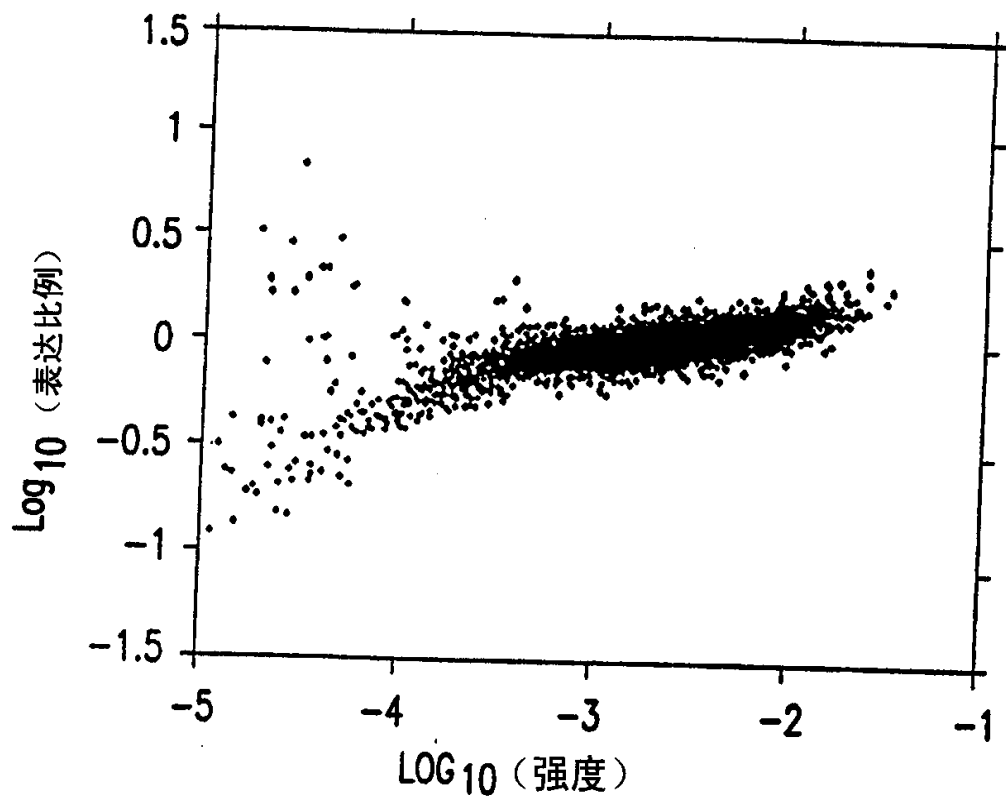


图 3

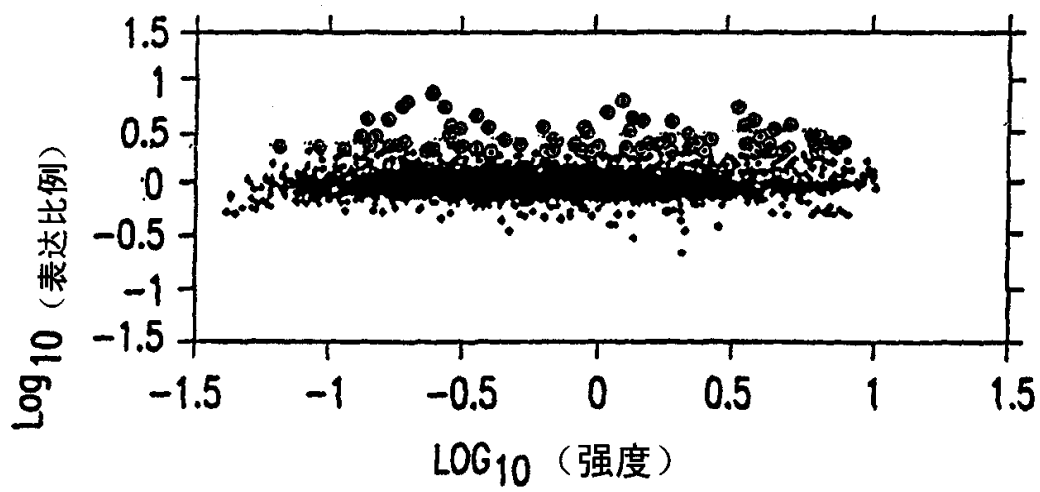


图 4a

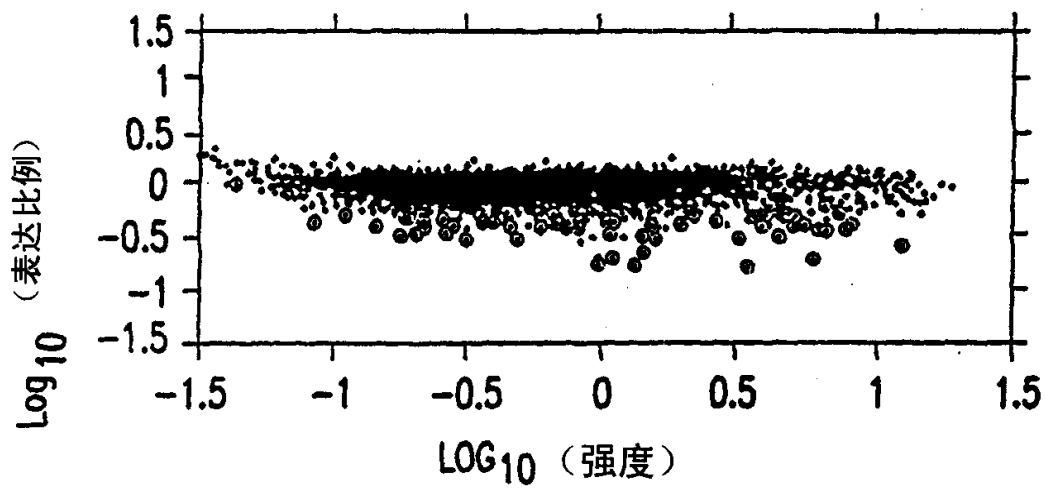


图 4b

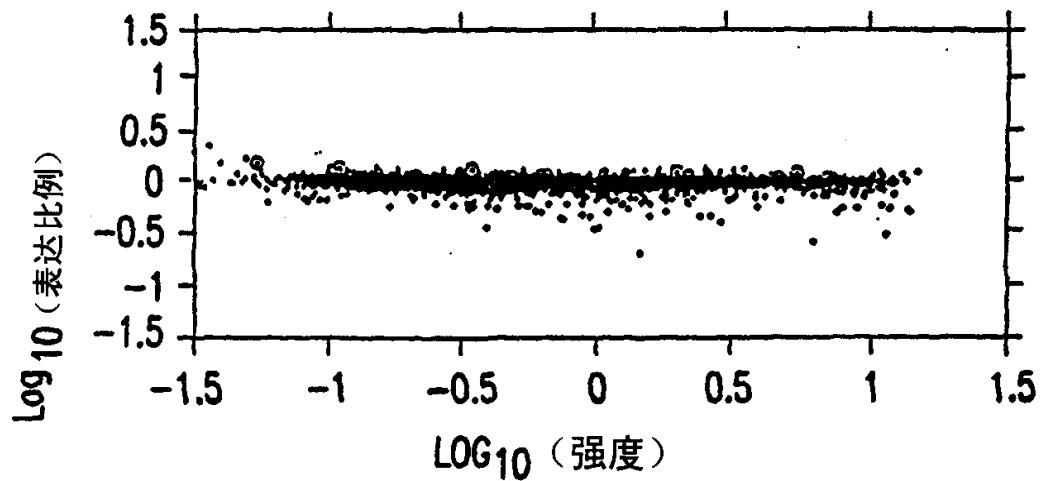


图 4c

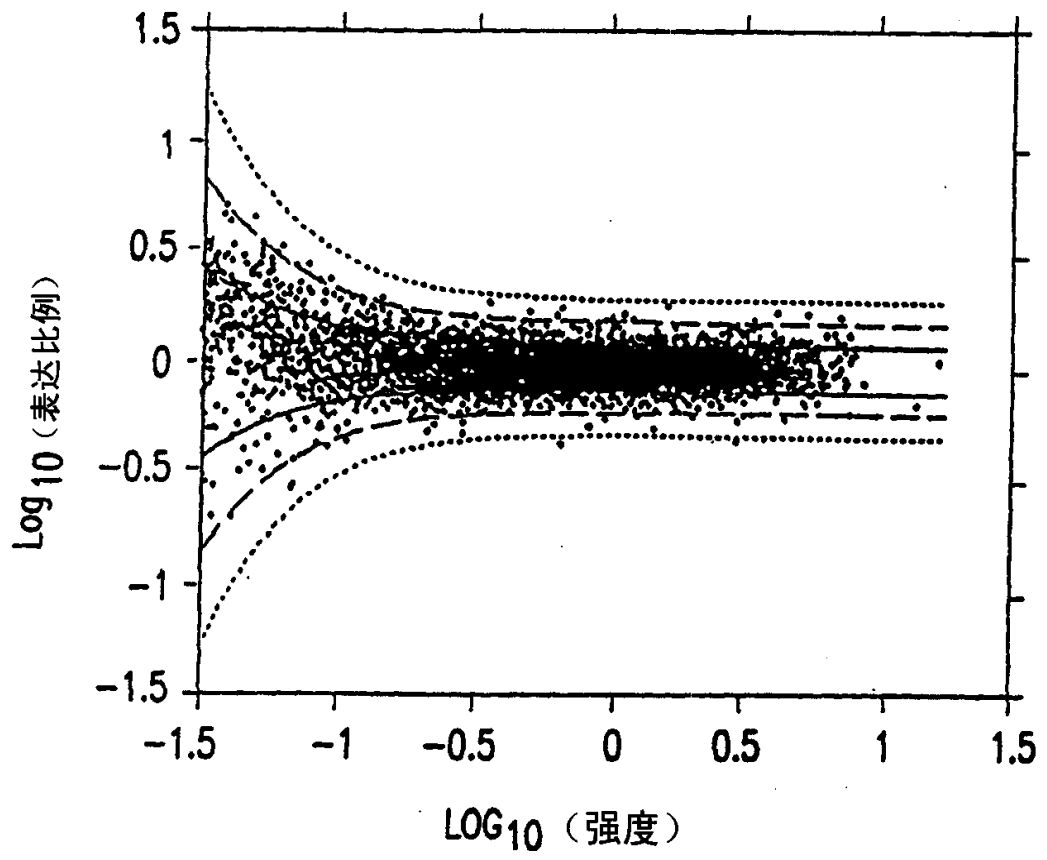


图 5

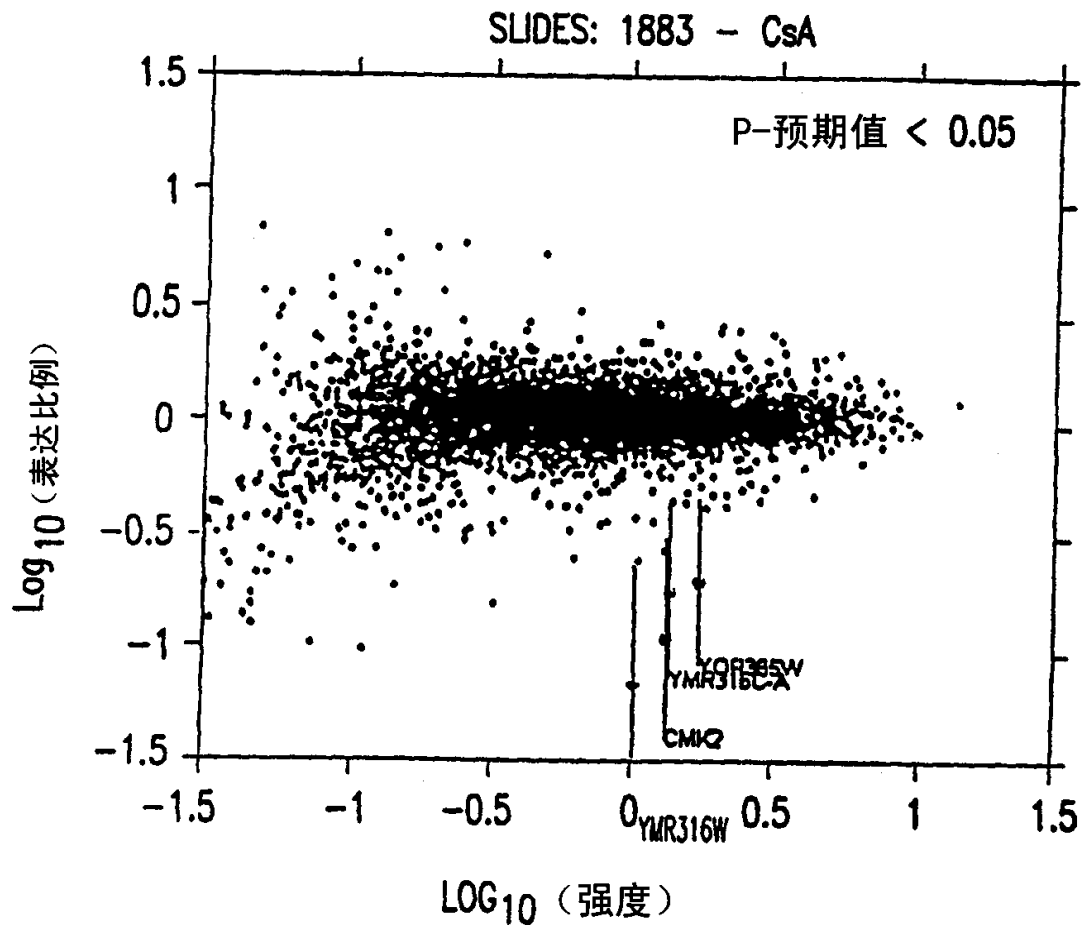


图 6a

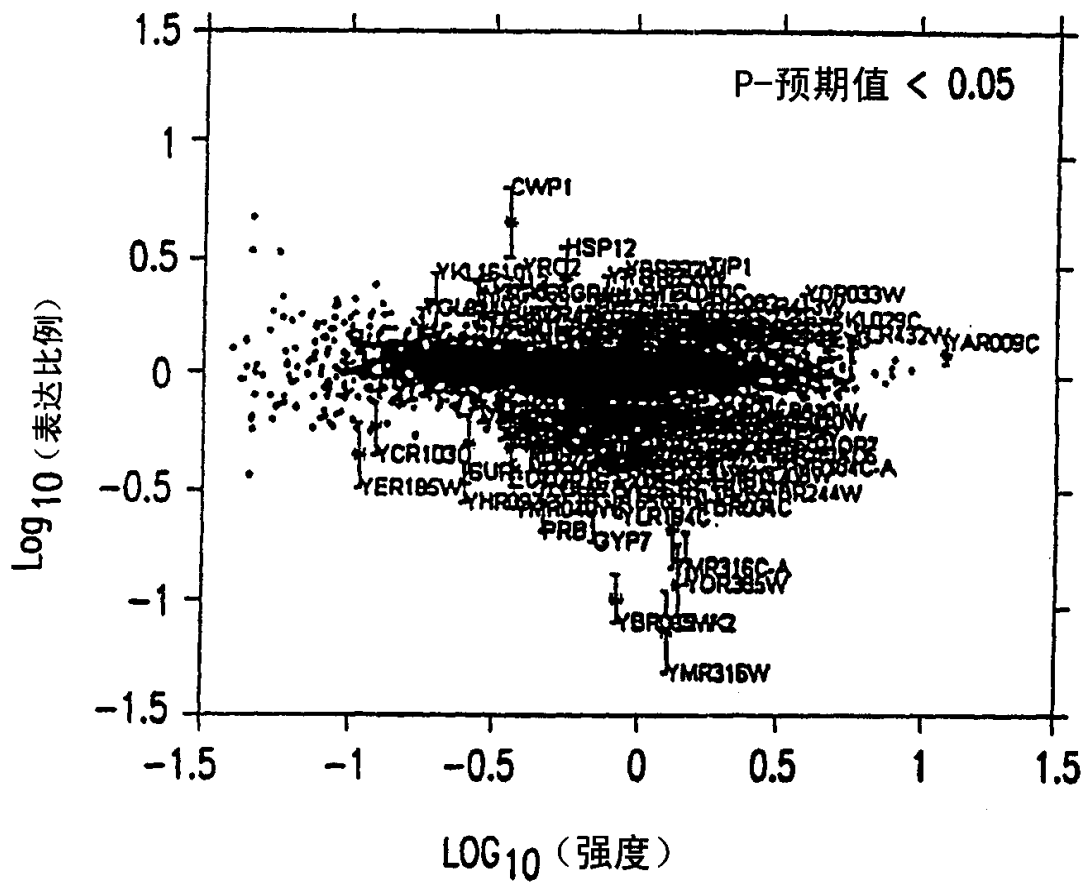


图 6b

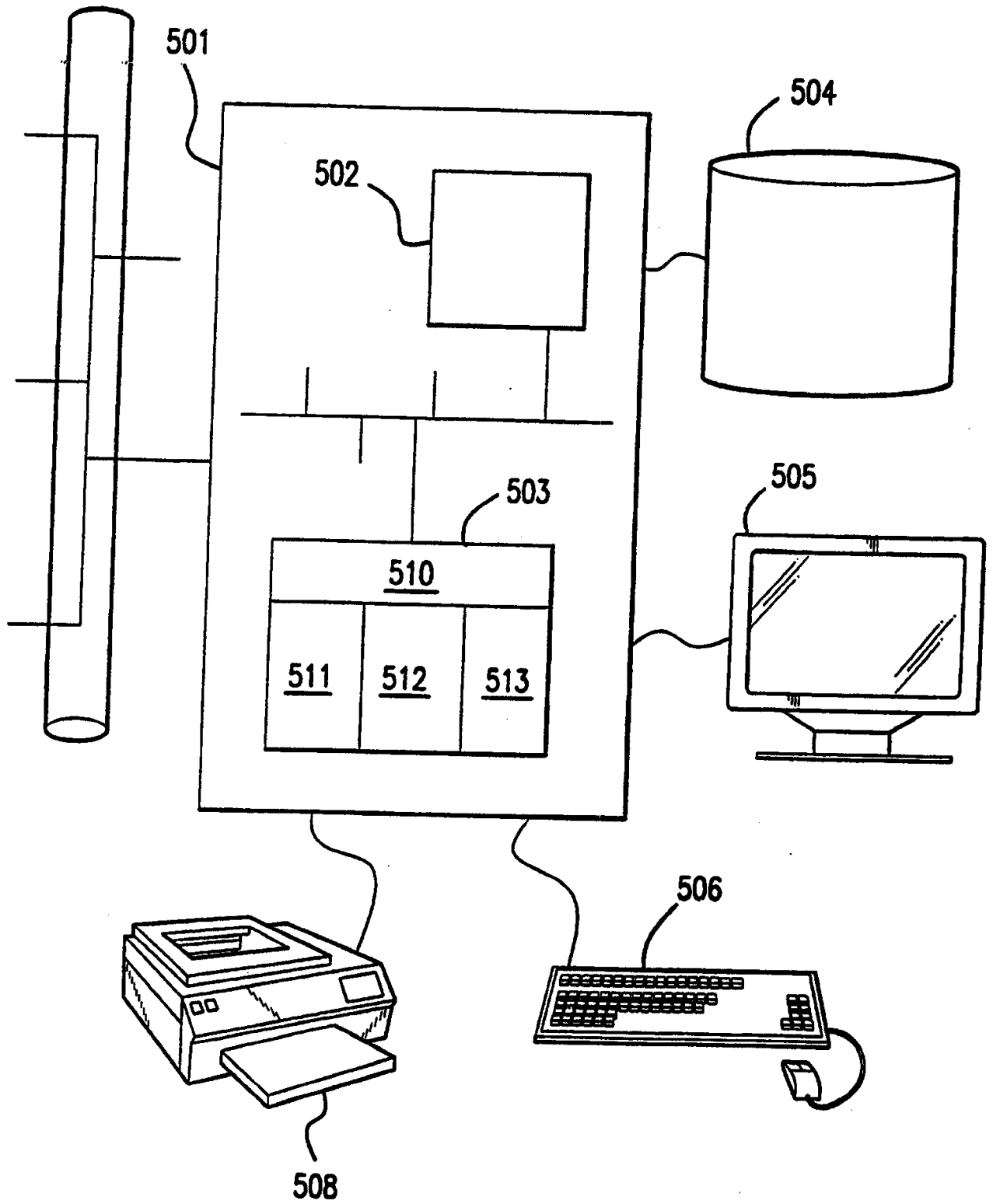


图 7