[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
G01N 35/00 (2006.01)
C12Q 1/68 (2006.01)

[12] 发明专利说明书

专利号 ZL 02825717.0

[45] 授权公告日 2006 年 11 月 29 日

[11] 授权公告号 CN 1287155C

[22] 申请日 2002.12.23 [21] 申请号 02825717.0
[30] 优先权
[32] 2001.12.21 [33] US [31] 10/028,482
[86] 国际申请 PCT/US2002/041478 2002.12.23
[87] 国际公布 WO2003/060526 英 2003.7.24
[85] 进入国家阶段日期 2004.6.21
[71] 专利权人 阿菲梅特里克斯公司
地址 美国加利福尼亚州
[72] 发明人 珍妮特·沃林顿 尼拉·沙阿
审查员 边 昕

[74] 专利代理机构 中原信达知识产权代理有限责任公司
代理人 杨 青 樊卫民

权利要求书 2 页 说明书 33 页 附图 8 页

[54] 发明名称

利用高密度微阵列进行高通量的重新测序和变异检测

[57] 摘要

在本发明的实施方案中，提供了用于高通量基因型分析的方法和系统。 该系统包括样品制备方法，自动化的样品制备系统，样品追踪系统，自动化的阵列处理，以及用于基因型分析和数据分析的计算机系统。

1. 一种高通量的基因型检测系统，包括：

样品制备自动化系统；

样品追踪系统；

自动化的高密度探针阵列装入器；

真空辅助的洗涤室；以及

计算机系统，用来管理杂交数据和分析杂交数据，从而进行基因型查询。

2. 根据权利要求 1 所述的系统，其中样品追踪系统和计算机系统是联机的。

3. 根据权利要求 1 所述的系统，其中样品制备自动化系统能够对多个核酸样品进行长距离 PCR 扩增。

4. 根据权利要求 3 所述的系统，其中从长距离 PCR 扩增得到的扩增子是 3-15kb。

5. 根据权利要求 3 所述的系统，其中的样品制备自动化系统能够在 PCR 扩增之前对每个核酸样品进行反转录获得 cDNA。

6. 根据权利要求 1 所述的系统，其中的样品制备自动化系统包括机器人装置，它用来操纵多孔板。

7. 根据权利要求 1 所述的系统，其中样品追踪系统是条形码系统。

8. 根据权利要求 1 所述的系统，其中计算机系统包括中央处理器，以及和中央处理器配套的内存，内存中贮存了多个机器指令，这些机器指令能命令中央处理器执行分析杂交的方法步骤，从而确定基因型。

9. 根据权利要求 1 所述的系统，其中高通量的检测系统进一步包括高密度核酸探针阵列。

10. 根据权利要求 9 所述的系统，其中高密度核酸探针阵列的特征尺寸是 20×24 微米或以下。

11. 根据权利要求 9 所述的系统，其中的高密度核酸探针阵列能够同时筛选至少 30kb 的有义核酸序列，和至少 30kb 的反义链核酸序列。

12. 根据权利要求 9 所述的系统，其中高密度核酸探针阵列被重新测序或者是变异检测阵列。

13. 根据权利要求 9 所述的系统，其中高密度核酸探针阵列被设计来查询 SNP 的集合。

14. 根据权利要求 9 所述的系统，其中高密度核酸探针阵列包括的探针被设计来查询前面已经鉴定的 SNP 集合的等位基因。

15. 根据权利要求 9 所述的系统，其中连续的序列被平铺在高密度核酸探针阵列上。

16. 根据权利要求 1 所述的系统，其中样品追踪系统包括一维或多维的条形码系统。

17. 根据权利要求 1 所述的系统，其中样品追踪系统包括一个电磁编码系统。

利用高密度微阵列进行高通量的重新测序和变异检测

相关申请

本申请是 2001 年 12 月 21 日提交的美国专利申请 No.10/028482 的部分继续。上述申请以其全文引作本发明的参考。

发明背景

本发明涉及到基因型分析，实验室自动化，生物信息学和生物数据分析。具体地说，本发明提供了用于基因型分析的高通量方法和系统。

单核苷酸多态性(SNP)被广泛地应用于遗传分析。已经开发了快速而又可靠的以杂交为基础的 SNP 分析(参考 Wang 等，Science 280: 1077-1082(1998)；Gingeras 等，Genome Research 8: 435-448(1998)；Halushka 等，Nature Genetics 22: 239-247(1999)；Cutler 等，Genome Research 11(11): 1913-25(2001)(下文都写成 Cutler 等，2001)，所有这些文章以其全文引作本发明的参考)。

发明概述

在本发明的一个方面，提供了一种用于基因型高通量检测的系统。典型的系统包括样品制备方法，自动化的样品制备系统，样品追踪系统，自动的高密度探针阵列装入器，用于管理杂交数据的计算机系统，以及用于分析杂交数据进行基因型查询的计算机系统。

典型的自动化样品制备系统包括一个用于操纵多孔板的机器人装置。在一些实施方案中，通过一个机器可读的编码系统进行样品追踪，例如，通过一维或多维的条码系统或电磁编码系统。在一些实施方案中，样品追踪系统和计算机系统是联机的。

在一些实施方案中，典型的计算机系统包括一个处理器以及和处理器配套的内存，内存中贮存了多个机器指令，这些指令能命令处理器执行分析杂交数据的方法步骤，从而确定基因型，其中分析包括基因型查询。基因型的查询可以通过 GeneChip Data Analysis Software(GDAS)(加利福尼亚州的圣塔克莱拉的 Affymetrix 公司)或别的能够从杂交数据中确定基因型的软件进行。诸如 GDAS 的软件能计算出一套模型用于杂交的可能性，并根据这些模型的可能性检查碱基，其中杂交信号强度的分布被假设是高斯分布，而且正向链和反向链被当作独立的复制子对待。

在本发明的另一方面，提供了一种确定大量样品多态性基因型的方法。在典型的实施方案中，该方法包括制备多个核酸样品，确定每个核酸样品和高密度寡核苷酸探针阵列的杂交，其中的高密度寡核苷酸探针阵列能检测多态性；并确定每个样品中的多态性基因型，其中的分析包括利用计算机系统检查基因型。

在本发明的一个方面，该系统可以使两个实验人员每天获得至少 1.4Mb 大小序列的基因分型信息。例如，两个实验人员对可在一天之内对样品进行基因分型，该样品含有至少 40 个不同的个体，其中每一个至少含有 35kb 的序列。该样品制备方法可以包括对基因组 DNA 选定区域进行 PCR 扩增。可以设计引物去扩增选定区域。PCR 可以是长距离 PCR，这种 PCR 一个反应可以扩增 3 到 15kb。

假如该样品是 RNA，那首先得反转录成 cDNA，然后通过 PCR 扩增 cDNA。一方面，多个转录产物的相对丰度可以在 PCR 扩增之前，通过和阵列进行杂交来确定。不表达或低表达的目的序列可以鉴定出来。这些不表达或表达很弱的转录产物在 PCR 过程中不会有效地扩增。

附图简述

本发明前述的和别的目标、特点和优点在下列本发明优选实施方案的更具体的描述中将会明显，正如在附图中显示的那样，在这些附图中，同样的参考数字在不同的视图中指的是相同的部分。这些附图主要用来说明本发明的原理，因而没必要标度。

附图，被整合进本说明，并成为本说明的一部分，图示了本发明的实施方案，并与描述一起用来解释本发明的原理。

图 1 显示的是计算机系统的一个例子，这个系统被用来处理和分析杂交数据，从而检查基因型。

图 2 是图 1 计算机系统的系统块状简图。

图 3 所示的是计算机网络，适用于本发明一些实施方案。

图 4 所示的是典型的微阵列 SNP 发现过程。

图 5 所示的是定做的高密度重新测序阵列。扫描阵列获得的典型图像的放大部分显示在嵌入的图中。右边的放大图显示的是两个阵列中完全相同的部分，这两个阵列和两个不同个体的样品分别杂交，这两个个体样品的序列在第二个位置有变化。

图 6 所示的是 GeneChip®阵列扫描仪以及扫描仪的自动装入器。该扫描仪的自动装入器原型是一个冷却单元，它包括 8 个架子，每个架子上有 8 个阵列，还包括机器臂，这个机器臂可以在扫描仪上装载或卸下阵列。

图 7 所示的是高通量快速冲洗工作站。

图 8 所示的是等位基因频率和可信度之比，

发明详述

将对本发明的优选实施方案进行详细的提及。当本发明和优选的实施方案一起介绍时，并不是为了要将本发明局限在这些实施方案中，这应该可以理解。另一方面，本发明打算包括可替代的、修饰的和等同的实施方案，它们也包括在本发明的实质和范围之内。所有引用的参考文献，包括专利和非专利文献，在此以其全文引作本发明的参考，适用于任何目的。

本发明有许多优选的实施方案，它依赖于本领域技术人员所熟知的许多专利、申请和别的参考文献。因此，当一个专利、申请或别的参考文献被引用或在下面重复时，出于所有的意图以及重新引用它们的目的，它们以其全文被引作本发明的参考，这一点是可以理解的。

在用于本申请时，单数形式“一个”“一种”“所述”包括复数的意思，除非上下文有明确地说明。例如，术语“一种试剂”包括许多试剂以及它们的混合物。

个体不仅限于人类，它也可以是其他生物，这些生物包括但不限于哺乳动物、植物、细菌或从上述生物中衍生而来的细胞。

在整个公开的内容中，本发明的不同方面通过一种范围的方式提了出来。应该可以理解，用范围的方式进行叙述仅仅是为了方便和简洁，而不应该解释成对本发明范围的一种死板的限制。因此，对一个范围的描述应该被当作具体地涵盖了该范围内的所有亚范围以及每个数值。例如，对范围从 1 到 6 的叙述应该被认为具体地涵盖了从 1 到 3、从 1 到 4、1 到 5、2 到 4、2 到 6、3 到 6 等各亚范围，以及在该范围内的每个数值，例如 1、2、3、4、5 和 6。不论范围有多宽都是这样。

除非说明，本发明的执行可以采用有机化学、多聚体技术、分子生物学(包括重组技术)、细胞生物学、生物化学和免疫学的常规技术和介绍，这些都包括在本发明的技术范围内。这些常规技术包括多聚体阵列分析、杂交、连接和使用标记对杂交进行检测。合适技术的具体说明可以参考下面的实施例。然而，别的等价的常规的程序当然也可以使用。这些常规技术和说明可以在一些基本的实验室手册中找到，这些手册如《基因组分析：实验室手册系列》(I-IV 卷)，《利用抗体：实验室手册》，《细胞：实验室手册》，《PCR 引物：实验室手册》

和《分子克隆：实验室手册》(所有这些实验室手册均来自冷泉港实验室出版社)，Stryer，《生物化学》(Biochemistry)，第4版(1995年三月)，Gait，《寡核苷酸合成：实践方法》(oligonucleotide synthesis: A practical Approach)，1984，IRL出版社，伦敦，Nelson 和 Cox(2000)，Lehninger，《生物化学原理》(Principles of Biochemistry)第三版，W. H. Freeman出版社，纽约，和 Berg 等(2002)《生物化学》(Biochemistry)，第5版，W. H. Freeman 出版社，纽约，所有这些书以其全文引作本发明的参考，适用于任何目的。

在一些优选实施方案中，本发明能利用固体物质，包括阵列。应用到多聚体(包括蛋白)阵列合成的方法和技术在 U.S.S.N 09/536841, WO00/58516, 美国专利 Nos.5143854, 5242974, 5252743, 5324633, 5384261, 5424186, 5451683, 5482867, 5491074, 5527681, 5550215, 5571639, 5578832, 5593839, 5599695, 5624711, 5631734, 5795716, 5831070, 5837832, 5856101, 5858659, 5936324, 5968740, 5974164, 5981185, 5981956, 6025601, 6033860, 6040193, 6090555 和 6136269，PCT 申请 Nos.PCT/US99/00730 (国际公布号 WO99/36760)和申请 No.PCT/US01/04285，以及美国专利申请系列 Nos.09/501099 和 09/122216 中有描述，它们在这以其全文引作本发明的参考，并适用于任何目的。

描述了在具体的实施方案中所用的合成技术的专利包括美国专利 Nos.5412087, 6147205, 6262216, 6310189, 5889165 和 5959098。在上述许多专利中都介绍了核酸阵列，但同样的技术也可应用到多肽阵列。

本发明也考虑黏附在固体物质上的多聚体的应用。这些应用包括基因表达监控、概貌分析、文库筛选、基因型分析以及诊断学。基因表达监控和概貌分析在美国专利 Nos5800992, 6013449, 6020135, 6033860, 6040138, 6177248 和 6309822 中介绍了。基因型分析及其应

用在 USSN10/013598 和美国专利 Nos.5856092, 6300063, 5858659, 6284460, 6361947, 6368799 和 6333179 中介绍了。别的应用体现在美国专利 Nos.5871928, 5902723, 6045996, 5541061 和 6197506 中。

本发明在某些优选的实施方案中也考虑到了样品制备的方法。例如，参看基因表达监控、概貌分析、基因型分析方面的专利，和上述的别的应用的专利，以及 USSN09/854317，Wu 和 Wallace，基因组学(Genomics)4: 560(1989)；Landegren 等，Science 241: 1077(1988)；Burg，美国专利 Nos.5437990, 5215899, 5466586, 4357421；Gubler 等，1985 生物化学和生物物理学报(Biochemica et Biophysica Acta)，“珠蛋白互补 DNA 的置换合成:序列扩增,转录扩增的证据”(Displacement sythesis of Globin Complementary DNA: Evidence for Sequence Amplification, transcription amplification)；Kwoh 等，Proc Natl Acad Sci. USA 86: 1173(1989)；Guatelli 等，Proc. Nati. Acad. Sci. USA 87: 1874(1990)；WO88/10315；WO90/06995；和 U.S.6361947。

本发明在某些优选的实施方案中也考虑到了配体之间杂交的检测。参考美国专利 Nos.5143854；5578832；5631734；5834758；5936324；5981956；6025601；6141096；6185030；6201639；6218803 和 6225625 以及 PCT 申请 PCT/US99/06097(以 WO99/47964 出版)，每一个专利以其全文引作本发明的参考，并适用于任何目的。

本发明也把各种计算机程序和软件用于不同的目的，如探针设计、数据管理、分析和仪器操作。参看美国专利 Nos5593839, 5795716, 5733729, 5974164, 6066454, 6090555, 6185561, 6188783, 6223127, 6229911 和 6308170。

此外，本发明有一些优选方案，包括在 Internet 网上提供遗传信息的方法。参看临时申请 60/349546。

在一些优选的实施方案中，提供了用于高通量基因型分析的方法。该方法使用高密度探针阵列，一个自动的样品制备系统，一个样品追踪系统，一个自动的阵列装入器，和一个用于管理和分析杂交数据的计算机系统，从而可以在一个选定的序列中鉴定出单核苷酸多态性(SNPs)。根据将要分析的序列，选择用于自动化的样品制备方法。

利用高密度探针阵列和检测基因型的高通量系统，本发明的不同方面将会在典型的实施方案中被介绍。

高密度探针阵列

在优选的实施方案中，本发明的方法和系统被用来分析获得的基因分型数据，这些数据是利用高密度探针阵列，如高密度核酸探针阵列获得的。

高密度核酸探针阵列，也叫“DNA 微阵列”，已经成为监控大量基因表达和检测序列变异、突变和多态性的一种选择方法。如本发明使用的那样，“核酸”包括任何核苷和核苷酸的多聚体和寡聚体(多核苷酸或寡核苷酸)，它们包括嘧啶和嘌呤碱基，优选分别是胞嘧啶、胸腺嘧啶和尿嘧啶、腺嘌呤和鸟嘌呤。(参看 Albert L.Lehninger，《生物化学原理》(PRINCIPLES OF BIOCHEMISTRY)，第 793-800 页(Worth 出版社，1982)和 L. Stryer，《生物化学》(BIOCHEMISTRY)，第 4 版(1995 年 3 月)，这两本书都被引作参考)。“核酸”包括任何去氧核糖核酸、核糖核酸或肽核酸成分，以及它们的化学变体，如这些碱基的甲基化、羟甲基化或糖基化形式等。这些多聚体和寡聚体在组成上可以是可以是非均质的或均质的，它们可以从自然来源中分离出来，或通过人工或合成的方法制造。此外，核酸可以是 DNA 或 RNA，或它们的混合物，它们可以永久地或暂时地以单链或双链形式、包括同源双链体、异源双链体和杂交体状态存在。

“靶分子”指的是目的生物分子。目的生物分子可以是配基、受

体、肽、核酸(RNA 或 DNA 的寡核苷酸或多核苷酸)，或任何别的列于美国专利 Nos5445934 第 5 栏第 66 行到第 7 栏第 51 行的生物分子，它们在这被引作参考，并适用于任何目的。例如，如果一个试验的目的是获得基因的转录产物，那转录产物就是靶分子。别的例子包括蛋白片段、小分子等。“靶核酸”指的是目的核酸(通常衍生自生物样品)。靶分子通常可以通过一或多个探针检测到。如本发明使用的那样，“探针”是一个用于检测靶分子的分子。它可以是任何和上面提及到的靶分子同类的分子。探针可以指核酸，如寡核苷酸，它能通过一或多种化学键，通常通过互补的碱基配对，和互补序列的靶核酸结合，配对碱基是通过氢键配对的。如本发明使用的那样，探针可以包括天然的(即 A.G.U.C 或 T)或修饰碱基(7-去氮杂鸟苷、肌苷等)。此外，探针中的碱基可以磷酸二酯键以外的键连接，只要这个键不干扰杂交。因此，探针可以是肽核酸，在肽核酸中，组成的碱基通过肽键而不是磷酸二酯键相连。别的探针例子包括用于检测肽或别的分子的抗体，用于检测结合受体的配体。当把靶序列或探针叫核酸时，应该可以理解，这些只是用作例证的实施方案，并不以任何方式对本发明进行限制。

在优选的实施方案中，可以将探针固定在一个物质上形成阵列。“阵列”包括一个带有肽或核酸或别的分子探针的固相支持物，这些肽或核酸或分子探针是黏附在固相支持物上。典型的阵列包括许多不同的核酸或肽探针，它们结合在物质表面的不同的、局域化的区域上。这些阵列，也叫“微阵列”，或通俗一点讲叫“芯片”，在本领域经常被介绍，例如在 Fodor 等，Science 251: 767-777(1991), 它在这引作参考，适用于任何目的。合成步骤最少的形成寡核苷酸，肽和别的多聚体序列的高密度阵列的方法已经公开，如在美国专利 Nos.5143854, 5252743, 5324633, 5384261, 5405783，5424186, 5429807，5445943，5510270，5677195， 5571639, 6040138，它们在这以其全文引作本发明的参考，并适用于任何目的。用不同的方法可以在固相支持物上合成寡核苷酸类似物，包括但不限于光引导化学耦合和机械引导耦合。参看 Pirrung 等人，美国专利 No.5143854，PCT 出版物 No.WO90/15070

和 Fodor 等人，PCT 出版物 Nos.WO92/10092 和 WO93/09668，美国专利 Nos.5677195, 5800992 和 6156501，它们公开了利用诸如光引导合成技术形成不同的肽阵列、寡核苷酸阵列和别的分子阵列的方法(参看 Fodor 等人，Science 251: 767-77(1991))。这些用来合成多聚体阵列的方法现在叫 VLSIPS™ 方法。

制造和使用分子探针阵列，尤其是核酸探针阵列的方法也公开了，例如在美国专利 Nos.5143854, 5242974, 5252743, 5324633, 5384261, 5405783, 5409810, 5412087, 5424186, 5429807, 5445934, 5451683, 5482867, 5489678, 5491074, 5510270, 5527681, 5527681, 5541061, 5550215, 5554501, 5556752, 5556961, 5571639, 5583211, 5593839, 5599695, 5607832, 5624711, 5677195, 5744101, 5744305, 5753788, 5770456, 5770722, 5831070, 5856101, 5885837, 5889165, 5919523, 5922591, 5925517, 5658734, 6022963, 6150147, 6147205, 6153743 和 6140044 中，所有这些专利都以其全文引作本发明的参考，适用于任何目的。

微阵列可以不同的方式使用。优选的微阵列含有核酸，被用来分析核酸样品。典型的核酸样品是从合适的来源中制备的，并用一个信号基团标记，如荧光标记。样品和阵列在合适的条件下杂交。冲洗阵列或用别的方法处理阵列，去掉没有杂交的核酸样品。然后通过检测标记在芯片上的分布来评估杂交。通过扫描阵列检测标记的分布，从而确定荧光强度分布。一般来说，每个探针的杂交可以通过若干像素强度反映出来。原始的强度数据可以贮存在灰度像素强度文件中。有几种贮存阵列密度数据的文件格式。最终的软件说明书可以在 www.gatcconsortium.org 上获得，以其全文引作本发明的参考。像素强度文件通常很大。例如，如果在水平和垂直轴上分别有大约 5000 个像素，而每个像素强度用 2 个字节，那么一个兼容的图象文件大约是 50Mb。这些像素可以分组成单元。(参看在 www.gatccomsortium.org 上的软件说明书)。单元中的探针被设计成具有相同的序列，即每个单元

都是一个探针区。CEL 文件包含一个单元的统计数据，如一个单元中像素强度的第 75 个百分点和标准偏差。单元像素强度的第 50，60，70，75 或 80 个百分点通常被用作单元的强度。

Affymetrix® Analysis Data Model(AADM)是 Affymetris 公司用来贮存实验结果的相关数据库计划。它包括支持作图的图表，配置的阵列和表达结果。Affymetrix 发布了 AADM 来支持开放进入通过 Affymetrix®软件产生和管理的实验信息，以便结果能够被相容的分析工具过滤和挖掘。也可以参考美国专利申请 No.60/396457 和美国专利申请 No.09/683982，它们在 2002 年 12 月 12 日出版，出版的申请号是 No.2002-0128993-AI。AADM 说明书(加利福尼亚州的圣塔克莱拉的 Affymetrix 公司)被引作本发明的参考，适用于任何目的。说明书可以从 http: //www.affymetrix.com/support/developer/aadm/content.affx，上获得，最后一次访问这个网站是在 2002 年 12 月 23 日。

利用高密度探针阵列进行基因型分析和多态性检测

基因型分析涉及确定个体的一个基因、基因组区或调节区的等位基因，或多态性标记身份。个体和群体基因型分析有许多用途。关于个体的遗传信息可以用于诊断某些与遗传因子有关的状态的存在和易感性。许多状态并不是起因于单个等位基因的影响，而是涉及到许多基因的共同作用。因此，确定若干基因组区的基因型对于诊断复杂的遗传状况是有好处的。

来自单个个体许多位点的基因型分析也能被用于法医，例如，根据个体的生物样品来鉴定个体。群体的基因型分析被用于群体遗传学。例如，对群体中不同等位基因频率的追踪能提供在很长一段时间内关于群体历史或遗传信息的重要信息。(对于基因型分析及其用处的总的综述请参看诊断《分子病理学：实践方法：细胞和组织基因型分析》(实践方法系列)，由 James O'Donnell McGee 和 C.S.Herrington 编辑，ISBN: 0199632383，和《SNP 和微卫星基因型分析：遗传分析标

记》(生物技术分子实验室方法系列)，编辑是 Ali Hajeer，Jane Worthington，和 Sally John，ISBN1881299384，这两本书以其全文引作本发明的参考)。

用寡核苷酸探针阵列可以确定基因组样品的基因型。这些阵列一般被“平铺”用来进行连续的测序或大量检测特定的多态性。在“平铺”进行连续测序的情况下，以前不知道的序列变异可以被发现和鉴定。

本发明使用的“平铺”指的是一套确定的寡核苷酸探针的合成，它由和目的序列互补的序列以及那个序列预先选出的变异组成，序列变异如在一或多个位置用一或多个基本单体如核苷酸取代。平铺策略已经在文献中详细讨论了，例如出版的 PCT 申请 No.WO95/11995 中，这个专利以其全文引作本发明的参考，适用于任何目的。

本领域的技术人员意识到本发明的方法、软件和系统不仅限于任何特定的平铺方式。

利用罩幕层来有效合成探针阵列的系统和方法在美国专利申请系列 No.09/824931 中介绍了，用来快速和灵活制造微阵列和网上订购的系统方法在美国临时专利申请系列 No60/265103 中详细介绍了，而不用罩幕层的光石版照相术的系统和方法在美国专利申请 No.6271957 和美国专利申请 No.09/683374 中详细介绍了，所有这些专利以其全文引作本发明的参考，适用于任何目的。

基因型分析数据分析系统

本领域的技术人员将会意识到许多计算机系统都适合执行本发明的方法。根据本发明实施方案，计算机软件可以在各种计算机系统中执行(对于基本计算机系统和计算机网络的介绍，请参考 Yale N.Patt, Sanjay J.Patel 的 Introduction to computing systems: From Bits and Gates

to C and Beyond, 第一版(2000 年 1 月 15 日), McGraw Hill Text, ISBN: 007236902; 以及《客户/服务器系统介绍：职业系统实践指南》, Paul E.Renaud, 第 2 版(1996 年 7 月), John Wiley&Sons; ISBN: 047133337, 这两本书以其全文引作本发明的参考，适用于任何目的)。

图 1 列举的是计算机系统的实施例，它可以用来执行本发明实施方案中的软件。图 1 所示的是计算机系统 101，它包括显示器 103，显示屏 105，机箱 107，键盘 109 和鼠标 111。鼠标 111 有一或多个按钮，用来和图形用户界面互动。机箱 107 含有磁盘驱动器 112，CD-ROM 或 DVD-ROM 驱动器 102，系统内存和硬盘驱动器(113)(也可以参看图 2)，硬盘驱动器可以用来贮存和恢复包含了用于执行本发明的计算机编码的软件程序，以及用在本发明中的数据等。尽管 CD114 被视为一种典型的计算机可读介质，但别的计算机可读贮存介质如磁盘驱动器，磁带，闪存，系统内存，和硬盘驱动器也可以利用。此外，包含在连线载波中的数据信号(例如在包括 Internet 在内的网络系统)也是计算机可读的贮存介质。

图 2 所示的是计算机系统 101 的系统块状简图，计算机系统 101 用来执行本发明实施方案中的软件。如图 1 所示，计算机系统 101 包括计算机监控器 201 和键盘 209。计算机系统 101 进一步还包括子系统如中央处理器 203(如 Intel 公司的奔腾 处理器)，系统内存 202，固定的存贮器 210(如硬盘驱动器)，可移动的存贮器 208(如磁盘或 CD-ROM)，显示适配器 206，音箱 204 以及网络界面 211。别的适合用于本发明的计算机系统可能还包括附加的或更少的子系统。例如，另外一个计算机系统包括多于一个的处理器 203 或高速缓冲存储器。适合用于本发明的计算机系统也可以被安置在测量仪器中。

图 3 所示的是典型计算机网络，它适合执行本发明的计算机软件。计算机工作站 302 和探针阵列扫描仪 301 相连接并控制它。从扫描仪上获得探针强度，显示在监控器 303 中。在工作站 302 处理强度，

并进行基因型检查(即根据探针强度确定基因型)。强度可以被处理并被贮存在工作站或数据服务器 306 中。工作站可以通过局域网如 Ethernet305 和数据服务器相连。打印机 304 可以直接和工作站或 Ethernet305 连接。局域网可以通过网关服务器 307 和广域网如 Internet308 相连，网关也可以作为 WAN308 和 LAN305 之间的防火墙。在优选的实施方案中，工作站可以通过 Internet 和外面的数据源，如国立生物技术信息中心发生交流。不同的方案，如 FTP 和 HTTP，可以被用于工作站和外面的数据库之间的数据交流。外面的遗传数据库，如 GeneBank310 是本领域技术人员众所周知的。GeneBank 和国立生物技术中心的全况可以从 NCBI 的网址上获得(http://www.ncbi.nlm.nih.gov)。

高通量基因型分析系统

图 4 所示的是高通量基因型分析过程的实施方案。选择基因或基因组区。设计引物并检测。有效的引物被用来进行 RT-PCR 或长距离 PCR。样品和高密度寡核苷酸探针阵列杂交。

在本发明的一方面，提供了一个基因型高通量检测系统。典型的系统包括样品制备方法；样品制备自动化系统；样品追踪系统；自动的高密度探针阵列装入器；和用来管理和分析杂交数据，从而进行基因型分析的计算机系统。

典型的样品制备方法包括选择基因或基因组区，设计和检测引物；如果样品是 RNA，如转录的 RNA，反转录该样品；通过 PCR 扩增，PCR 可以是长距离 PCR；收集扩增子；选择性地纯化扩增子；并片段化和标记。标记的片段可以和高密度探针阵列进行杂交。

典型的样品制备自动化系统包括一个机器人装置，它用来操纵多孔板，如 96 孔的微孔板。在一些实施方案中，利用机器可读编码系统进行样品追踪，如一维或多维的条形码系统或电磁编码系统。合适

的自动装入器也在美国专利申请 Nos.09/691702 和 60/396457 中介绍了，这两个专利以其全文引作本发明的参考。

自动装入器提供了一种机制，用于从扫描仪上转出或转进样品盒。本发明能方便地利用标准化的载体，载体能容纳许多样品盒，这些样品盒贮存在冷冻柜里。用一个双轴的机器将样品盒从扫描仪，加温室，容纳室中移进或移出。一个局域操作界面和网络连接可以提供给主机工作站，从而便利转运系统的操作。

使用盒式载体的优点在于它们能提供标准的方法来容纳多个样品盒。进一步说，盒式载体可以包括上锁的小孔，这是为了防止反向安装。使用具冷冻柜的机架使得样品盒在扫描前能贮存若干小时。然而，在一些实施方案中，温控柜是不必要的，这一点是可以理解的。移走后，加温室用来消除样品盒在送入扫描仪之前的冷凝作用。使用机器人也使得样品盒能在载体和不同的扫描仪室之间自动移动。本领域的普通技术人员也将意识到存在许多用来贮存和自动转运探针阵列盒的方法和部件。

别的自动装入器的实施例在美国临时专利申请系列 Nos.60/217246 中介绍了，题目是“盒式装入器和方法(CARTRIDGE LOADER AND METHODS)”，2000 年 7 月 10 提交；60/364731，题目是“生物材料高分辨率扫描系统，方法和产品”，2002 年 3 月 15 日提交；以及 60/396457，题目是“高通量微阵列扫描系统和方法”，2002 年 7 月 17 日提交；和美国专利申请系列 No.09/691702，题目是“盒式装入器和方法”，2000 年 10 月 17 日提交，这里的每一个专利申请以其全文引作本发明的参考，适用于任何目的。

条形码扫描仪可以方便地用来鉴定主机中的样品盒的内容。条形码可以作为样品追踪系统的一部分。一方面，利用网络界面联机转运系统，可以将当地用户的界面整合进去以便利装载或卸载样品盒。进

一步说，非插入式的排列机制可以用来非插入式的连接扫描仪。这种排列机制可被用作盒式装入器和扫描仪之间排列的唯一联系。可以方便地将盒式装入器设计得相对较小，以适合工作台的顶部，并能由一个人来安装。

在一些实施方案中，在阵列洗涤室中冲洗阵列。洗涤室可以从加利福尼亚州的圣塔克莱拉的 Affymetrix 公司购买到。参看美国专利 Nos.6114122，6391623 和 6422249，它们被引作本发明的参考。

在一些实施方案中，典型的计算机系统包括一个中央处理器，以及一个和中央处理器配套的内存，内存中贮存了多个机器指令，这些机器指令能使处理器执行杂交分析方法的步骤，从而确定基因型。

利用从 Affymetrix Variation Detection Arrays(VDAs)中获得的实验数据，用软件系统进行基因型分析，Affymetrix Variation Detection Arrays(VDAs)也叫 CustomSeq™ 阵列，它可以从加利福尼亚州的圣塔克莱拉的 Affymetrix 公司购买到。优选的软件是一种自动的统计系统，它用来确定单个 VDA 基因型，而不管这个位点是否具多态性。这个系统可以用于试验，在这个实验中，靶 DNA 序列可以是单倍体或双倍体。事实上，该系统使得研究者能利用 VDAs 去确定目的样品中的 DNA 序列。优选的软件是 GDAS，它出现在 Cutler 等人 2001 年的文章中(可以从 Aravinda Chakravarti 实验室获得，名字叫 ABACUS)。软件可以用诸如 ANSI 一类的标准码来运行。

基于 Cutler 等人 2001 年的文章和 GDAS 的算法的一个假设是观察到的荧光强度在功能点(feature)内是正常分布的。这个假设是根据中心限制法则确定的。每一个功能点 包含大约 1 百万个不同的组成同一的寡核苷酸。如果这些寡核苷酸中相当一部分在和标记的靶 DNA 结合的过程中相对独立，这些功能点 的整体荧光强度在强烈的中心限制法则下，应该是正常分布。在假设在靶样品中存在或不存在不同基

因型的情况下，开发了一系列统计模型。正向或反向链的给定基因型的每个统计模型的可能性可以独立计算出来，这种可能性和这个模型的整体可能性相结合。“质量分数”是最适模型和第二最适模型之间的可能性的对数(底为 10)的差，被赋值给每个 VDA 基因型。如果一个模型比别的模型更充分适合这个数据，就说一个位点基因型被“检查”。在所有的个体 VDA 基因型检查后，另外的启发式的、可靠的规则被采用。在完成这个程序的过程中，所有的位点都被赋了一个带相应质量分数的基因型。单个被认为不可靠的基因型定义为 N。系统分为六个阶段：阶段 1：数据完整性检查，阶段 2：建立具均匀背景的模型，阶段 3：比较模型，阶段 4：建了一个不均匀背景的模型，阶段 5：重复一个适应性的背景，以及阶段 6：采用最终的可靠规则。对于这六个阶段的详细介绍参看 Cutler 等人 2001 年的文章。

GDAS 也提供用于高通量基因型分析的软件。序列数据管理器具有分析发射强度值的功能，发射强度值包含在探针阵列数据文件中。数据管理器可以同时分析许多样品，例如 40 或更多的样品。

数据管理器可以执行基因型分析算法用于分析发射强度数据，如衍生自探针阵列的数据，该探针阵列设计用来探察 DNA 序列。为了获得可靠的数据，探针阵列在某些情况下需要许多拷贝的选定 DNA 序列。通过 PCR 可以复制许多拷贝的 DNA 序列。

基因型分析算法包括选定 DNA 序列的核酸组成的鉴定，单核苷酸多态性(后文称作 SNP's)，以及别的涉及基因组序列方面的特征。例如，一种算法可以包括来自 Affymetrix 公司的 CustomSeq™ 算法。CustomSeq™ 算法可以用来确定选定 DNA 每个序列位置的核酸组成。在本实施例中，算法可以使用来自探测装置的发射强度数据，探测装置放置在探针阵列上，探针阵列是设计用来探察特定基因组 DNA 或别的类型的序列。发射强度数据值包含在一或多个数据文件中，这些文件如*.cel 文件。

在一个可能的执行过程中，数据管理器可以通过很多步骤来执行该算法。第一步，数据管理器可以采用数据过滤器来鉴定不可靠的数据，或调整被认为是发射强度变异值的数据，发射强度接近检测极限。这里采用的术语“变异值”一般指的是数据离散性的度量。

数据过滤器可以使用一或多个来自样品的探针组将序列位置裁定为检测不到(n)，或者对探针阵列的变异值进行调整。例如，数据过滤器可以考虑两个探针组的发射强度，这两个探针组表示的是基因组序列上的同一个位置。例如，设计一个探针组来探察编码链上的序列位置，而设计另一个探针组探察非编码链上相应的序列位置。

数据过滤器可以根据某类特征特异地过滤发射强度数据，这些特征包括检测不到信号，信号弱，信号饱和，或高的信噪比。在某些例子中，如果发射强度数据不能满足在一或多类中指定的标准，数据过滤器可以将序列位置裁定为检查不到信号(n)。如果一个样品的序列位置被裁定为检查不到信号(n)，那信息可能被记录在样品基因型查询数据中。

没信号这一类包括的标准如被称作平均强度值的阈值。探针组的每个探针 功能点有一个唯一的平均强度值，被定义为探针功能点 中所有像素发射强度值的平均值。阈值包括预定义的值，可以是一个在零的两个标准差之间的数值。此外，阈值也可以是用户选择的值。本发明使用的术语“标准差”一般指的是变异值的平方根。在本执行方案中，对于一或多个样品中的序列位置，标准差衍生自一或多个探针组的每个探针功能点 的发射强度数据。另外，标准差可以衍生自探针功能点的子集 ，如功能点的类型 (A, C, G 或 T)，某个具体链(即编码或非编码链)的探针组，或来自探针阵列中的所有探针组。例如，如果任何探针组的任何一个探针功能点 的平均强度值低于阈值，则相应序列位置检测不到信号(n)。除非该类别满足了标准，否则检查不能被赋

值。

信号弱这一类包括的标准如称作最高平均强度值的阈值。最高平均强度值可被定义为探针功能点 的平均强度值，而这个平均强度值比探针组中别的探针功能点 的平均强度值要高。阈值可以包括预定义的值，这个值比同一条链(即编码或非编码链)的所有探针组的最高平均强度值的平均值要低 20 倍。此外，阈值也可以是用户选择的值。例如，如果探针组的最高平均强度值低于阈值，那给相应序列位置的赋值就是检测不到信号(n)。除非该类别满足了标准，否则检查不能被赋值。

信号饱和这一类包括的标准如一个阈值，当探针组中许多探针功能点不能达到该阈值时可以被赋值为检测不到信号(n)。这个阈值包括预定义的值。和前几类一样，用户也可选择阈值。标准差和用于无信号这一类的标准差可以一样，或由于衍生自另一套发射强度值而有所不同。为了给序列位置赋值检测不到信号(n)，这一类的第二个标准也包括不满足阈值标准的探针功能点 数。例如，一个序列位置对应于一条染色体，这条染色体可能处于单倍体状态(换句话说，单倍体状态指的是一条染色体，而二倍体指的是一对相似的染色体)。如果探针组有 2 或多个探针功能点 的平均强度值大于阈值，那么这个序列位置被赋值检测不到信号(n)。也是在本实施例，如果序列位置对应于二倍体状态，那么 3 或多个功能点 的平均强度必须比阈值高，才能赋值检测不到信号(n)。

信噪比这一类包括的标准如称作信噪比的阈值，它是用来指信号和噪音比率的值。本发明使用的“信噪比”一般指的是从杂交探针中产生的信号的发射强度值和噪音中的发射强度值的比率。噪音包括荧光发散，它是来自残余的未结合样品、与探针功能点非特异地结合的样品，或别的能产生荧光发散的过程，该荧光发散不包括样品和探针功能点的特异性结合。阈值包括预定义值。如果信噪比大于阈值，那

变异可以调整到相同或不同的阈值。在一个替代的实施例中，探针组的信噪比，或对应一个序列位置的一或多个探针组的信噪比比阈值大。在这样的实施例中，对应于一或多个探针组的变异可以被调整到阈值。

过滤的发射强度数据可以被分析模式比较器接收，进行该算法的下一步。比较器的分析过程是根据，至少一部分是根据若干模型进行的，开发这些模型是为了详细说明在选定的 DNA 序列的每个序列位置上特定核苷酸存在与否。根据不同的假设，这些数据采用了两套不同的模型。这些假设根据的是称作均匀背景或非均匀背景的东西，它们在下面将会详细介绍。

比较器可以计算某个具体的核酸在每个序列位置上适合模型的可能性。对编码链和非编码链，这种可能性可以独立的确定，而一个模型的最终可能性可以通过将编码链和非编码链的可能性数值相乘来确定。

对每一个模型，质量分数是根据，至少部分是根据概率值来计算的。可以为每一条链计算一个质量分数，也可以计算一个总的质量分数。例如，分别利用编码链、非编码链和总的概率值来计算质量分数。

正如被相关领域的普通技术人员所意识到的那样，均匀背景的假设是根据，至少部分是根据中心限制法则的。例如，探针 功能点的寡核苷酸被假设是同一组成的，而且在和标记靶序列结合时，是相对独立的。因此，如相关领域普通技术人员所意识到的那样，探针功能点的总体发射强度应该是正常分布的(换句话说，探针和样品结合的机会相同)。

模型包括检测不到信号模型，纯合体模型，杂合体模型。检测不到信号模型假设所有探针组具有同样的平均值，同一条链上的探针组

的变异相同(即编码或非编码链)，但是链与链之间的平均值和探针组的变异可以不一样。

纯合体和杂合体模型基础与检测不到信号模型基本相似，但假设有轻微不同。

在本执行方案中，杂合体模型可以仅用于二倍体数据，相关领域的普通技术人员都能够意识到其中原因。杂合体模型包括 A-C, A-G, A-T, C-G, C-T 和 G-T。该模型又一次和检测不到信号模型相似，但假设有差异。例如，对于 A-C 杂合体，编码链上 G 和 T 的背景特征被假设是独立和完全同样分布的。编码链上相似特征 A 和 C 也被假设是独立和同样分布的。这个模型反映的就是这个假设。

比较器计算所有均匀背景模型的概率和质量分数。根据质询的样品是单倍体还是二倍体，模型的数目可以变化。这里使用的术语“单倍体”和“二倍体”指的是出现在样品中的染色体的数目。单倍体一般指的是一条染色体，而二倍体指的是出现了两条染色体。对于单倍体数据，可以计算整个 5 种模型的概率和质量分数，换句话说，即检测不到信号，A, C, G 和 T 模型。对于二倍体数据，得补充六个模型，包括 AC, AG, AT, CG, CT 和 GT。

如果一个均匀背景模型几乎完全适合，而另外的均匀背景的模型适合程度相对低，即可以对序列位点进行基因型检查。

如果没有均匀背景模型完全适合，比较器可能根据次适合模型进行基因型检查。在一个例举的实施方案中，存在两个质量分数阈值，T_{Total} 和 T_{Strand}。两个阈值都可是预定义或用户确定型的，其中预定义阈值可以通过实验来确定。T_{Total} 用于次适合与完全适合的值可以是相同的，或者可以是一个不同的值。例如，预定义的阈值可以是通过实验为次适合专门确定的。在本实施例中，T_{Total} 的预定义值是 30，而 T_{Strand}

的预定义值是-2。

比较器下一步对另一套模型采用来自二倍体样品的发射强度数据，这个模型根据不同的一套假设。这些模型可被称作非均匀背景模型，在这些模型中，假设一条链上所有的探针组平均值和变异不是同样的。例如，能产生不同平均值和变异的情况包括交叉杂交或背景特征的非均匀性。在交叉杂交的实施例中，预测所有的样品在平均值和变异中表示出相同比率的非均匀性。

在一个执行方案中，非均匀背景模型包括那些能在样品间保持恒定非均匀比率的模型。表示平均值和变异比率的恒定值可以通过将所有样品中具同样基因型的每个序列位点的平均值和变异值平均来获得。相关领域普通技术人员将会意识到许多序列位点的基因型检查开始是未知的。在一个优选的执行方案中，当基因型检测改变时，可以用一个重复的方法来改变恒定值。这个重复方法可以继续直到基因型检测趋于相同，或可以通过一套重复的数据来进行，这套数据可以预定义或由用户来选择。

在一个执行方案中，为了完全符合或次符合而进行的非均匀模型基因型检测的标准和均匀模型的标准一样。也在本执行方案中，如果一个模型比其他模型都适合编码和非编码链，但不能满足次符合检测所必需的阈值，序列位点的基因型检测可以猜测。例如，如果一个给定模型的质量分数比 0 大而且该模型比其他模型更适合，则可以进行推测。

在均匀模型和非均匀模型两种情况下，如果对于一个给定的序列位点，一个模型不能检测或推测的话，那这个位点被归类为不能检测(n)。

为了测试基因型检测的可靠性，序列数据管理器可以向数据可靠

性测试器提出基因型检测结果。在一个优选的执行方案中，为了使其考虑得更可靠，基因型检测数据必须满足许多标准。这些标准包括但不限于下列介绍。

对每个序列位点，至少 50%周围位点必须进行基因型检测(即 A, C, G 或 T)或被裁定为检测不到(n)。周围位点的数目可以被预定义或由用户自己选择。例如，需要考虑的周围位点的数目由用户来选择是 20 个，这就意味着在序列位点每一侧都有 10 个位点被考虑。在本实施例中，如果在 20 个周围位点中有多于 10 个检测不到(n)，那么质询的序列位置就被裁定为检测不到(n)。

对于一个序列位置，如果在所有样品中，同一个序列位置有多于 50%的基因型检测被裁定为检测不到(n)，那这个序列位置就被裁定为检测不到(n)。

如果在 5 个序列位点中互相有 2 个 SNP's 被鉴定，那这两个 SNP's 就叫 SNP 双联体。例如，一个 SNP 叫 SNP1，而另一个叫 SNP2。对每个 SNP 的基因型检测，基因型检测的共同性越高，这种检测就可以叫野生型检测，而共同性越少，则叫突变型检测。相关领域的普通技术人员应该能意识到，前述的实施例只是为了例证，而在任何方式上都不应该作为限制。

确定 SNP 双联体的规则包括下列实施例。如果一个样品对 SNP1 而言是突变型，对 SNP2 而言则是野生型，而另一个样品对于 SNP1 是野生型，对于 SNP2 则是突变型。这两个突变 SNP 检测都确定是可靠的。如果一个样品在 SNP1 是突变的，在 SNP2 是野生型的，而所有别的样品在 SNP2 都是突变型的，在 SNP1 也是突变型的或是检测不到的(n)。那 SNP2 检测被确定是不可靠的，而所有样品则可以被认为在 SNP2 序列位点是检测不到的(n)。如果在样品中总是出现 SNP1 的突变体，而这些样品在 SNP2 也是突变型，或检测不到，反之亦然。

那带有少量检测不到的 SNP 被认为是可靠的，而别的 SNP 位点在所有样品中被称作检测不到(n)。

然后序列数据管理器可以将来自数据过滤器，分析模型比较器和数据可靠性测试器的结果集中到一或多个样品基因型检测数据文件中。数据可以包含对应于所用样品的结果，或者也可以是对应于每个样品都有一个单独的文件。例如，来自样品发射强度数据文件的基因型检测结果可以组合成一个样品基因型数据文件。在本实施例中，每个样品发射强度数据文件对应于单独的样品基因型数据文件。

输出管理器然后可以接受来自管理器的一或多个文件。输出管理器可以将来自每个样品的基因型检测排列，以图形化的用户界面呈现给用户。

实施例

这一部分介绍的是高通量系统，它利用的是高密度微阵列为发现 SNP 而进行重新测序。实施例展示了本发明的不同方面。在样品制备方法，杂交试验，阵列操作和分析方法上作了许多改进和补充。将来自三个不同种族的 40 个不相关个体的 DNA 扩增，标记，并与设计用来代表基因组的编码和调控区的探针阵列进行杂交。方案改进的地方包括使用长距离 PCR 和半自动化，标记和片段化花费降低。自动化改进的地方包括开发了一个用于阵列的扫描仪自动装入器，一个更快的阵列洗涤室，和一个相关的实验室追综和数据处理系统。这些改进使得在每个微阵列上能同时筛选 30kb 的有义和反义 DNA(图 5)，使每两个实验人员的通量上升到每天 1.4Mb。用于更小功能点尺寸，如 20×24 微米的有效的基因型分析软件也增加了通量。利用高密度重新测序和变异检测阵列(微阵列)在 8.3Mb 的人基因组中鉴定出多于 15000 的 SNPs。

总的说来，该方案的目的是减少重新测序阵列的费用，手段是在

该方案的每个方面进行一些变化。具体地说，目的是减少从阵列上获得信息所必需的时间和努力，方式是通过开发一套改进的、自动化的加工阵列的系统，包括开发花费较少的样品制备方法，如减少 PCR 引物的费用和样品的体积；在工作台上自动进行样品制备和芯片操作；为了控制阵列的功能添加一些内部对照；开发一套改进的能检测碱基的算法；以及改进碱基监测和 SNP 检测的精确度。逐步取得了一些进展，而且当通量上升以及发现 SNP 的费用降低的时候，数据的质量就改进了(Cargill 等人，Nat Genet22: 231-238(1999)；Lindblad-Toh 等人，Nature Genet.24: 381-386(2000)；Cutler 等人，2001)。

材料和方法

样品来源.来自国家卫生研究院柯瑞尔医学研究中心的不同小组的细胞系被用作基因组 DNA 或 mRNA 的来源 以制备 cDNA(Coriell Institute, Camden, NJ)。样品选自三个不同种族的 40 个男性和女性，其中北欧 11 个女性和 9 个男性，非洲 10 个女性，而亚洲是 4 个女性和 6 个男性。

引物设计.在目的基因或基因组区被鉴定后，为了进行长距离 PCR，利用一些公用或从商业上获得的程序即 Primer 3(www-genome.wi.mit.edu/cgi-bin/primer3-www.cgi)，Amplify 1.2(Engel 等人，Trends in Biochemical Science 18: 448-450(1993))，Oligo 6(SR Lifescience, www.lifescience-software.com) 去设计一些引物制备 3-15kb 的扩增子。根据这个方案，从三个不同的柯瑞尔样品、cDNA 或基因组 DNA 制备的 DNA 池被用来测试这些引物。

样品制备.使用标准的方法分离基因组 DNA(Moore 等人，〈基因组 DNA 的制备〉，在：Ausalel 等人编辑，《当代分子生物学实验方案》(Current Protocols in Molecular Biology).纽约：JohnWiley&Johns 公司，第 2.1.1-2.1.9 页(1984)。从 mRNA 制备 cDNA 的方法如前述(Mahadevappa 和 Warrington, Nat. Biotechnology 17: 1134-1136(1999))。

使用目的区的长距离 PCR 扩增样品，每个扩增子取等量进行电泳，确定扩增子大小和数量，然后用前述的方法收集(Cutler 等人，2001)。用 Multiples 型 MP EX 机器进行 PCR 扩增，扩增子收集，浓缩和纯化(Packard 仪器公司，Meriden CT)。

表达分析.为了优化 PCR 的成功，当用 cDNA 作为 PCR 的模板时，进行表达分析以确定每个转录产物的相对丰度和鉴定未表达的目的基因和转录产物，这些没表达的基因和转录产物丰度太低以至于不能从成淋巴细胞系中大量地扩增出来。在含有代表人类 6800 个全长基因探针的阵列、HuGeneFL 探针阵列(加利福尼亚州的圣塔克莱拉的 Affymetrix 公司) 上进行表达分析。样品制备和阵列杂交遵循制造商的说明(加利福尼亚州的圣塔克莱拉的 Affymetrix 公司)。如前述，通过将已知浓度的添加标准和它们的杂交强度关联起来以确定拷贝数(Lockhart 等人，Nat. Biotechnology.14: 1675-1680(1996))。假设每个细胞中平均有 300000 转录产物，每个转录产物大小是 1kb，来计算转录产物的丰度。

定制的重新测序阵列.设计高密度重新测序或变异检测阵列，即 SNP 发现阵列，对应于通过长距离 PCR 成功扩增出来的 DNA 片段。每个阵列包含一个 0.5kb 的用作内部试验对照的肌动蛋白序列，以及用于在制造中控制质量的一套标准对照。每一个定制的阵列含有 400000 个不同的探针，代表的是 30kb 的有义和反义链 DNA(图 2)。这 400000 个不同的探针位于 20×24 微米的功能点(feature)中，每个功能点含有同一个探针的数以百万计的拷贝。

自动化.

开发出来的自动化用于实验室，在自动化控制中，配置一些“岛”或室作为样品制备和试验的一部分。对于样品制备和扩增，每个室以 Packard Multiprobe Robot 为中心。所有的准备都是 96 孔格式的，而板是通过手工在室之间转移的。对于实验本身，几个 GeneChip®系统

包括杂交烘箱 320/640's、FS400 流动室和基因阵列扫描仪(加利福尼亚州的圣塔克莱拉的 Affymetrix 公司)被使用了。对 GeneChip®系统作了若干修改和改进。用于试验的扫描仪自动装入器、快速阵列洗涤室和连接的实验室追踪和数据管理系统被开发用来提高效率，减少失败分析时间、阵列操作时间和试剂的数量，最终减少总的费用。扫描仪自动装入器是一个冷冻的单元，含有由 8 个架子组成的旋转架，每个架子上是 8 个阵列(图 6)。一个机器臂将阵列从旋转架举起来，并把它降落入扫描仪中，这时相关的软件发信号扫描开始。一旦扫描完成，机器臂取回扫描的阵列，并把它重新放置在架子上，然后再去拿下一个阵列。所有的阵列信息都和条形码相联系，条形码放置在阵列盒中，并被自动装入器读出。一个快速洗涤室的原型(图 7)利用真空将水溶液引入阵列盒中，在短的保温期后将水引出来，能够处理 12-20 个阵列，时间和 FS400 洗涤工作站处理 4 个阵列的时间一样。此外，开发了一个特定的机器人装置器，使得 Multiprobe Robot 工作站在杂交之前，能自动地将样品装入 24 个阵列盒中。

带条形码阅读器和独特的条形码的扫描仪相结合能够唯一鉴定每个阵列，不管它是通过手工还是自动装入器装入的。条形码阅读器位于扫描仪的内部，可以阅读一或多个指代阵列的条形码。扫描仪控制和分析系统能使用条形码鉴定来将扫描的阵列盒和含有该探针阵列信息的实验文件相关联。正如相关领域的普通技术人员所熟知的那样，条形码是用条形和空间的组合来表示字母和数字，它可以一或多维格式表示。关于条形码的额外的讨论，请参考 02 年 7 月 17 日提交的美国临时专利申请 No.60/396457，和美国专利 No.6399365。

在一个优选的实施方案中，杂交工作站执行的程序可以将一或多个实验样品和许多探针阵列以高通量的方式进行杂交。别的信息请参考美国临时专利申请 No.60/417942，它是 02 年 10 月 11 日提交的，在这被引作本发明的参考。

探针阵列可以放置在一个表面，如载玻片上。杂交工作站可以将暴露的探针阵列浸入特定体积的样品中。另外，利用别的液体潴留的手段，可以将样品应用到探针阵列的表面。

此外，探针阵列可以装入一个封套或盒子中。杂交工作站能通过一或多个特定的端口将样品注入封套或盒子中。在一个可能的执行方案中，提供一个可以将材料注入封套或盒子的端口，以及将它们取出的端口。而别的执行方案包括一个端口，这个端口适用于这两种目的。例如，可执行文件能指导杂交工作站将特定体积的样品加入探针阵列。杂交工作站能够通过一根针从池中取出特定体积的样品，然后将这根针通过探针阵列封套中的一个指定的孔插入，并释放出样品。

杂交工作站可以利用管子将样品转移到另外一根针或别的转移装置，例如，管子能将池中的针和转运针连接在一起，通过把样品物理性沉积在另外一个表面上，或采用别的方式进行直接转移。另外一个转运装置包括一个叫双腔针的东西，双腔针可以被插入一个孔中。例如，一个腔可以被设计来传递样品、或别的流体到探针阵列，而另一个腔被设计来移出样品或别的流体。

杂交工作站包括检测系统，该系统能检测探针阵列封套中流体的存在。此外，该系统也能鉴定存在的流体的种类。

杂交工作站在可移动池中容纳了许多试验样品。池包括小瓶，小管，瓶子或别的适合容纳一定体积的容器。杂交工作站提供了一个容器或一系列容器，它能接受一或多个池。容器或一系列容器包括盘子，转盘或暗盒，这些容器中另外还包括独特的条形码或别的类型的标识符。

支架和一系列支架的位置是已知的，以至于一个试验样品能和一个位置相联系，并能和仪器控制软件交流。杂交工作站在每个支架上

也提供探测器，当池出现时，表示是可执行的。

杂交工作站能为样品中生物材料提供和探针阵列中的探针杂交的合适的条件。这样的条件包括温度，额外溶液的添加，气泡，搅动，震荡液面，或别的能促进生物样品和探针杂交的条件。在一个优选的执行方案中，杂交工作站在一个特定的间隔改变这些条件，从而优化杂交过程的效率。例如，超声波搅动能改进试验样品和探针阵列杂交的效率。封在盒子中的探针阵列可以被浸入液体溶液中，超声波搅动器能促进搅动在探针阵列上的均匀分散。杂交工作站可以提供一个气泡或封套，它们包括别的能增加液体在探针阵列上动荡的物理特点，从而通过混合，增加探针阵列暴露在试验样品成分的机会，进一步改进杂交的效率。在本实施例中，气泡包括环境中的气体或别的类型的气体，它们能改进样品杂交。

杂交工作站还可以进行杂交后操作，包括用缓冲液或试剂冲洗，以及用非严谨缓冲液加入探针阵列封套，从而使杂交的阵列在扫描之前保持完整性。别的杂交后操作包括本领域的普通技术人员通常叫做染色的过程。例如，染色包括将带荧光标记的分子导入，荧光标签能选择性的和探针阵列杂交的生物分子结合。在本实施例中，一或多个荧光标签标记的分子可以结合到每个生物分子上，从而能增加扫描过程中的发射强度。染色的过程也包括将杂交的探针暴露在带具不同特征的荧光标记的分子中。不同的特征包括能选择性地结合不同杂交生物分子或有不同激发和发射性质的荧光标记的分子。例如，当第一个荧光标记被暴露在第一波长的光中时，它会被激发，结果发射出第二波长的光。第二个荧光标记被第三波长的光激发，第三波长的光和第一个荧光标记的第二发射波长的光一样，同时发射第四波长的光。

杂交工作站允许中断运行，插入或移走探针阵列、样品、试剂、缓冲液或任何别的物质。中断后，杂交工作站可以扫描一些或所有的标识符，这些标志符和探针阵列，样品，旋转架或磁盘盒，用户输入

的标识符或别的在自动化过程中的标识符相联系。例如，一个用户可能希望中断这么一个过程，取走一个样品盘并插入一个新的盘子。

杂交工作站也会执行一些不会直接作用在探针阵列上的运行操作。这种功能包括对新鲜的和用过的试剂和缓冲液，试验样品，或别的在杂交操作中用过的材料进行管理。在本实施例中，样品具有条形码标记，这个标记是和它们相联系的唯一的标识符。可以用一个手提的阅读器扫描条形码，或者在杂交工作站中也可用一个内置的阅读器替代。此外，别的电子识别工具也可采用。用户可以将这些标识符和样品联系在一起，并把这些数据贮存成一或多个数据文件，例如这些文件可以包括试验数据。这些样品可以和特定的探针阵列类型相联系，这些探针阵列类型也用同样的方式贮存。

为本方案而建立的实验室和数据处理数据库，HTS2000，是一个双层的，分布式的客户/服务器应用程序，它是利用 ActiveX Data Objects(ADO)在 MS Visual Basic 6.0 和 Oracle8i 上开发出来的。具有 MS Outlook 的外表和感觉，界面的标准组件设计反映出高通量筛选和 SNP 发现过程的复杂，从序列和引物选择到证明引物测试凝胶结果和收集扩增子用于纯化，定量，片段化和标记 (参看美国专利 No.6484183，它在这被引作本发明的参考)。从样品制备到数据分析这个过程的每一步都被追踪，并和条形码相联系。也可以参考美国专利 Nos.09/682098 和 60/220587，它们以其全文引作本发明的参考，适用于任何目的。

分析软件.一旦阵列被扫描，栅格被排列，将 X, Y 坐标分配到信号强度上，这是每个 功能点上产生的信号强度，以便随后进行分析。对于 SNP 发现和基因型分析，更多样品是必需的，因此，使用了一个自动化的批处理栅格排列工具(也可以参看美国临时申请，Nos.60/408848 和 60/393926，它们在这被引作本发明的参考)。

数据分析

自动化的 SNP 检查和可信度排列排除了对每个检查进行单独复查和评估的需要，这样显著改进了一致性，精确性和通量，而分析时间和费用却下降了。为了改进可重复性和精确性，尤其是杂合体检查的可重复性和精确性，诸如 GDAS 这样的分析软件(或别的在 Cutler 等人 2001 年的文献中所示的软件)可以作为高通量基因型分析系统的一部分，并在别处被详细介绍。

结果

以前的样品制备方法都是从 cDNA 或基因组 DNA 中通过扩增小于 1kb 的短片段来获得样品，或扩增平均小于 200bps 的短的序列标记位点来获得样品。多个短的扩增子，50-6000，被汇集到一块用于杂交。精确的测量和汇集等摩尔量的大量扩增子不是一个繁琐的事，而且足够准确地进行这个过程，并防止数据质量的不利效果是困难的。在汇集的扩增子浓度存在高和低，并且和一个阵列进行杂交的情况下，很难从背景和噪音中辨别低丰度信号。例如，一个杂合体突变体样品杂交强度在两个探针之间是分开的，没精确定量的样品，如浓度低的样品将会产生一些不显著强于背景的信号，从而使得精确的碱基检查变得不可能。此外，在汇集之前，对每个样品的 50-6000 个扩增子进行电泳的时间和费用是过高的。然而，没有这些质量控制步骤，将会导致不完全样品的杂交。扩增子丢失通常是因为不精确的定量和汇集，用来获得 cDNA 的低丰度转录产物所引起的 PCR 失败，以及在配对区存在 SNPs 或存在单拷贝的质量差的样品 DNA 而导致的不充分的退火。这都会导致一些样品的一些片段的数据丢失，从而导致分析中能力的损失。

人类基因组全部序列的获得提供了附加的序列信息，使得基因组 DNA 和长距离 PCR 能用于样品产生。长距离 PCR 样品制备提供了许多优点，包括减少必需的引物数，这样能减少试剂和 PCR 相关处理步骤的费用。通过这个方法，可以获得更少的可以估计数量和汇集的扩

增子，导致阵列中信号强度更一致，从而获得更好的数据质量。利用基因组 DNA 和长距离 PCR，可以在每个样品中汇集平均 6kb 长度的 5 个扩增子。这在 PCR 反应数，跑胶数，以及可以估算数量和汇集的扩增子数上减少了 10 倍或更多。当基因组 DNA 的长距离 PCR 被用作模板，PCR 扩增成功率典型地多于 80%，或多于 90%。

利用 Chee 等人的算法的改编版本进行 SNP 发现分析(Chee 等人，Science 274: 610-614(1996)。做了一些修饰，从而可以补偿由小的特征阵列所产生的低强度信号，以进行杂合体的碱基检查。修改的分析方法产生候选的 SNPs，这些 SNPs 由两个受过培训的分析家独立评估。在证实和验证由这种方法获得的结果的努力中，将这个结果和 328 个片段的单向测序结果作比较，这些片段已经在高的，适度的或低的可信度作了检查。对来自 2 个样品中的每个片段做单向测序，2 个样品分别是参考的纯合体和纯合体或杂和体等位基因。使用 Chee 等人的修改的算法鉴定的 81%的 SNPs 是同样的。最难鉴定的 SNPs 是稀有等位基因，这是鉴定的最大一类 SNPs。在这类中，只有 66%的 SNPs 被鉴定了(图 8)。因为手工分析所要求的时间量以及验证效果差，因此通量上的改进和 SNP 检查精确性将会受益于自动分析方法的改进，这是很明白的。

任何一个本领域的技术人员都会意识到，任何统计算法都必须利用实际的基因型分析数据，选择合适的算法和开发不同的用于算法的参数进行评估。利用基因型分析数据，GDAS 和 Cutler 算法都被开发出来和执行。这两种自动化的碱基检查都会得到一个质量分数，而且通过利用概率模型方法都可以鉴定 SNPs。对于纯合体考虑了四种模型。如果样品是纯合体 G，假设有义链该位置上代表另外 3 个可能核苷酸的符号是独立的(C, T, A)，而且是同样分布的，那么 G 的强度信息有不同的平均值和变异。对于纯合体中别的三个可能的检查用同样的方式进行。对于杂合体，数据用四个上述的纯合体模型和 6 个杂合体模型，G-C，G-T，G-A，A-T，A-C，C-T(参看 Cutler 等人 2001 的

文章)进行评估。针对两条链上每个碱基的每个模型的可能性独立评估，这些可能性结合起来确定这些模型适合得怎样好，以及是否模型比别的模型适合得足够更好。如果一个模型比别的模型显著适合某个数据，就进行检查，这个模型必须都适合有义和反义链的位置，而不能比别的位置更显著的适合模型的位置叫 N。在分析软件中使用别的规则试图去鉴定 PCR 失败的原因，这种失败会导致不正确的碱基检查。这些规则的阈值可以由用户来设置。默认的设置要求扩增子中多于 50%的碱基应该是可检查的，换句话说至少 10/20 的周围碱基是可检查的。在多于 50%被调查样品中，一个位点也必须是明确可检查的，没有 N's。当然该位点在这些样品中不能有同样的碱基检查。碱基检查是完全自动的，它去掉了分析者的偏见，并且显著降低了分析时间。每个检查的碱基获得一个可信度，这样就为以后的研究提供了一种评估特定 SNPs 相对风险的工具。可信度是最适合模型和第二适合模型之间可能性对数的差。对于 GDAS 的别的介绍，请参看美国临时专利申请 No.60/408848，2002 年 9 月 6 日提交，在这被引作参考。

进行了两类确认研究来评估这个过程，碱基检查或基因型分析精确性和 SNP 检查精确性。为了评估碱基检查精确性，进行了确证研究，将阵列为基础的重新测序和通过 4-8X 为 1938 个碱基对获得的数据进行比较。99.998%(1935/1938 碱基对)的检查具有 Abacus 可信度数值为 1：100000，显示了高的可信度。为了证实通过重新测序发现的 SNPs，选择了含有 117 个碱基对的子集，它们中的 100%是通过标准测序方法证明的。

样品制备自动化使得试剂体积减少，试剂费用降低 33%。自动化的阵列控制和分析使通量的可能性加倍。最终，高通量系统使得两个熟练的研究人员能在一天程序化地和可重复地制备样品，杂交和分析 40 个阵列。在这两年的过程中，3 个不同人种的 40 个不相关个体中，包括启动子区在内的 25051 个人基因(8.3Mb)的所有或部分被筛选，产生了总数超过 15000 的 SNPs，它们被放置在 dbSNP(http:

//www.ncbi.nlm.nih.gov/SNP)。

别的典型信息可在 Warrington 等人的人类突变(human mutation) 19: 402-409(2002)中找到。

本发明的范围不应该局限在上面的介绍中，而应该由附录的权利要求来确定，以及由这些权利要求所授权的相当的东西的全部范围来确定。

所有引用的参考文献，包括专利和非专利文献以及网址，都被引作参考，适用于任何目的。

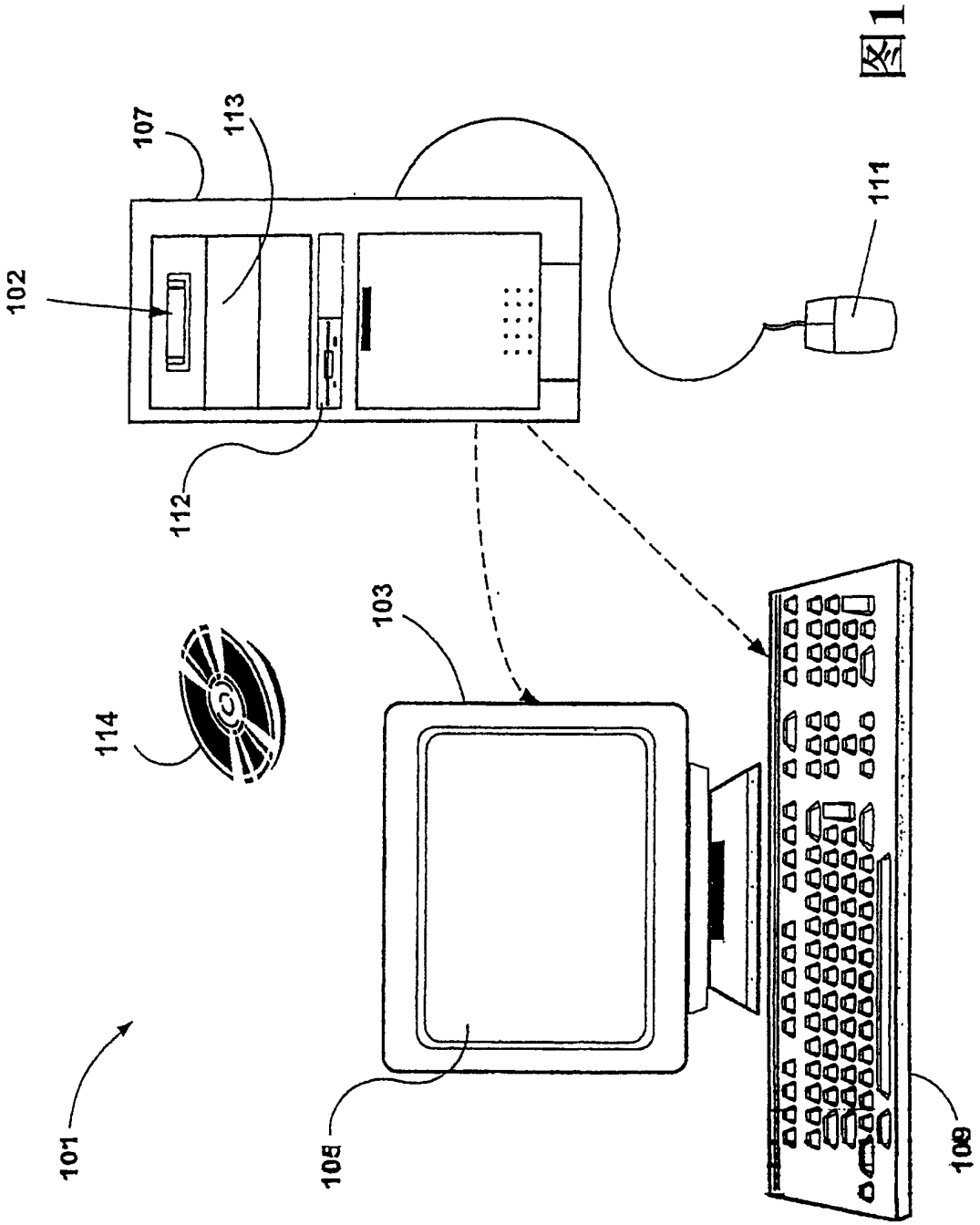


图1

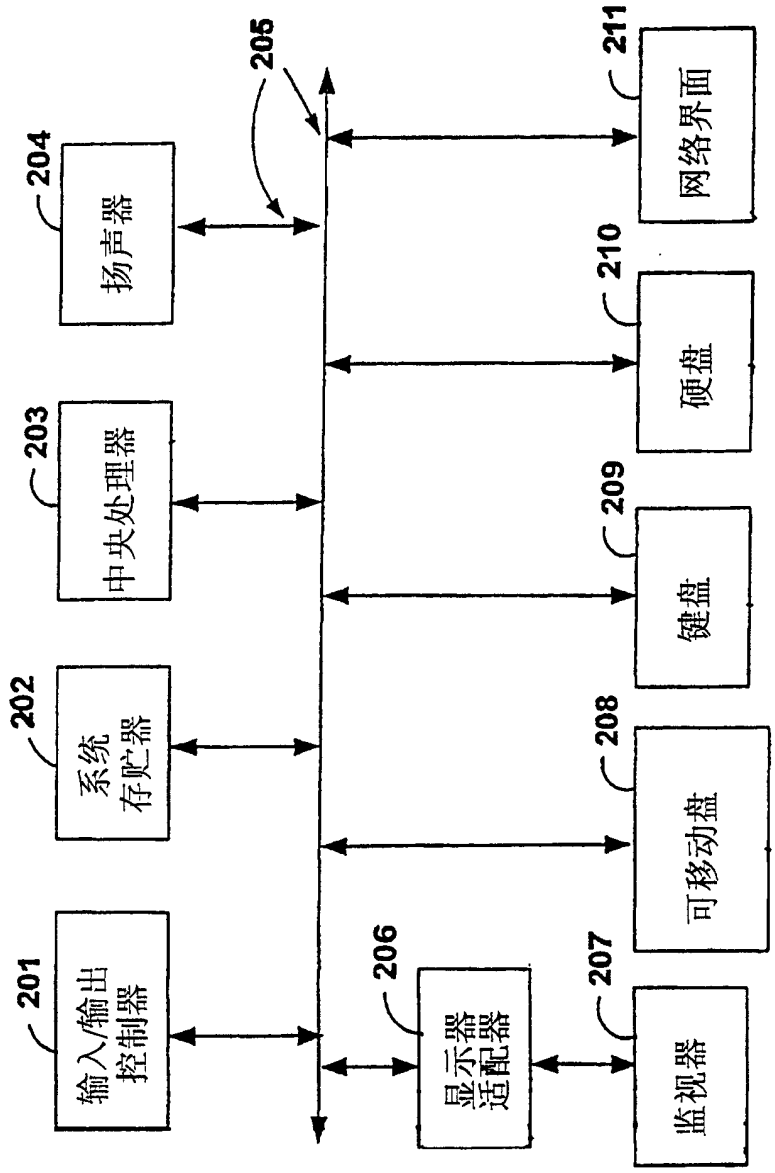

201
输入/输出控制器
202
系统存贮器
203
中央处理器
204
扬声器
205
206
显示器适配器
207
监视器
208
可移动盘
209
键盘
210
硬盘
211
网络界面

图2

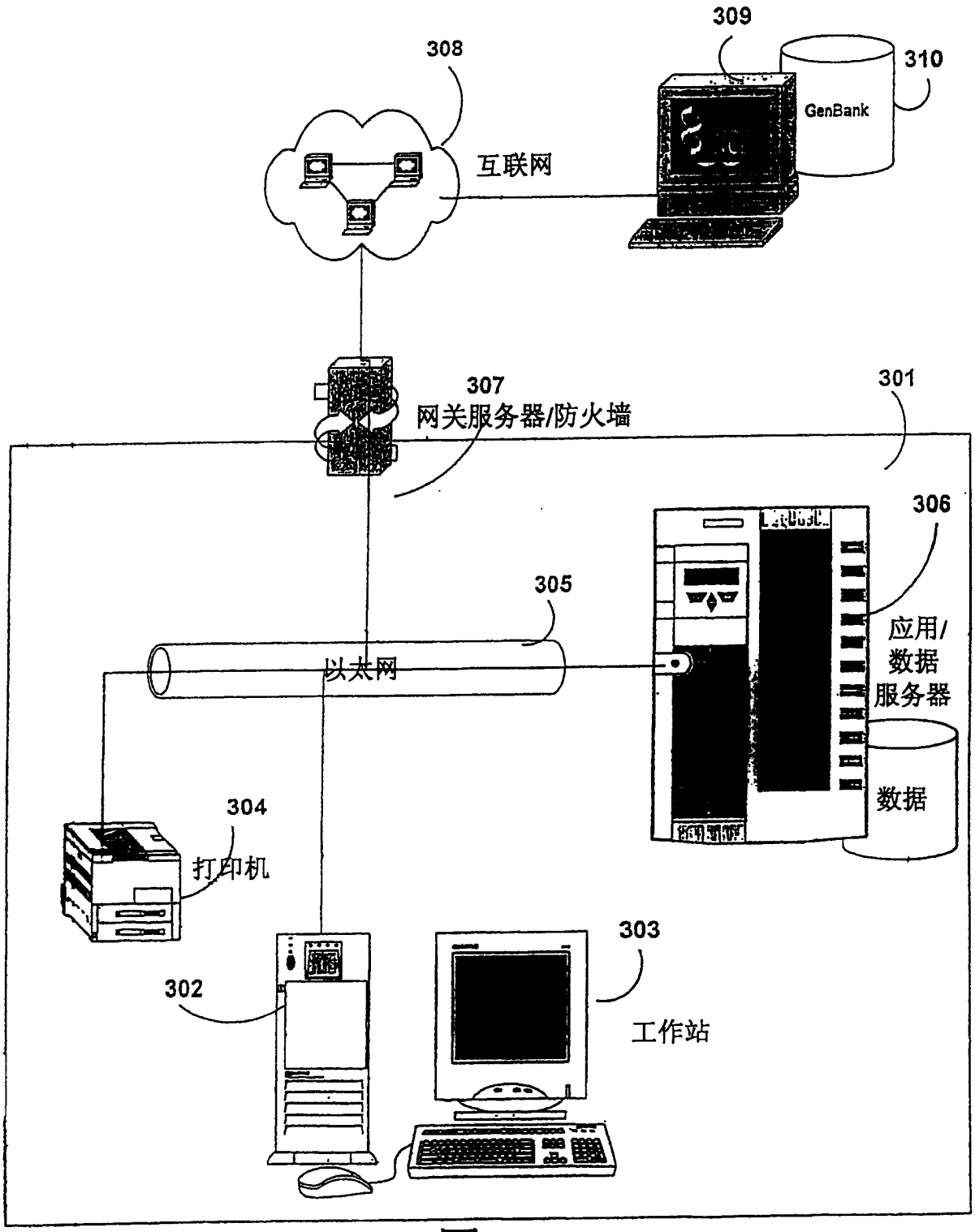

309
308
310
GenBank
互联网
307
网关服务器/防火墙
301
306
305
应用/
数据
服务器
以太网
数据
304
打印机
303
302
工作站

图3

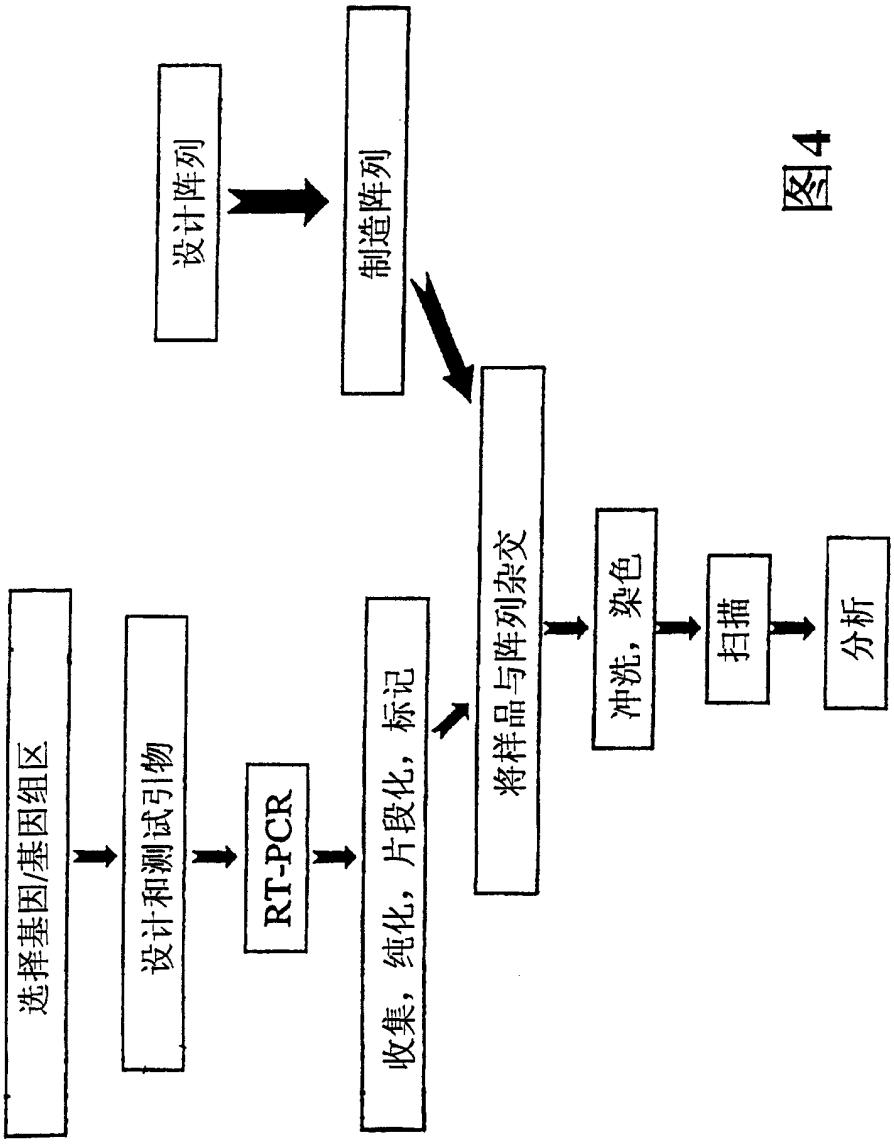

选择基因/基因组区
设计和测试引物
RT-PCR
收集，纯化，片段化，标记
设计阵列
制造阵列
将样品与阵列杂交
冲洗，染色
扫描
分析

图4

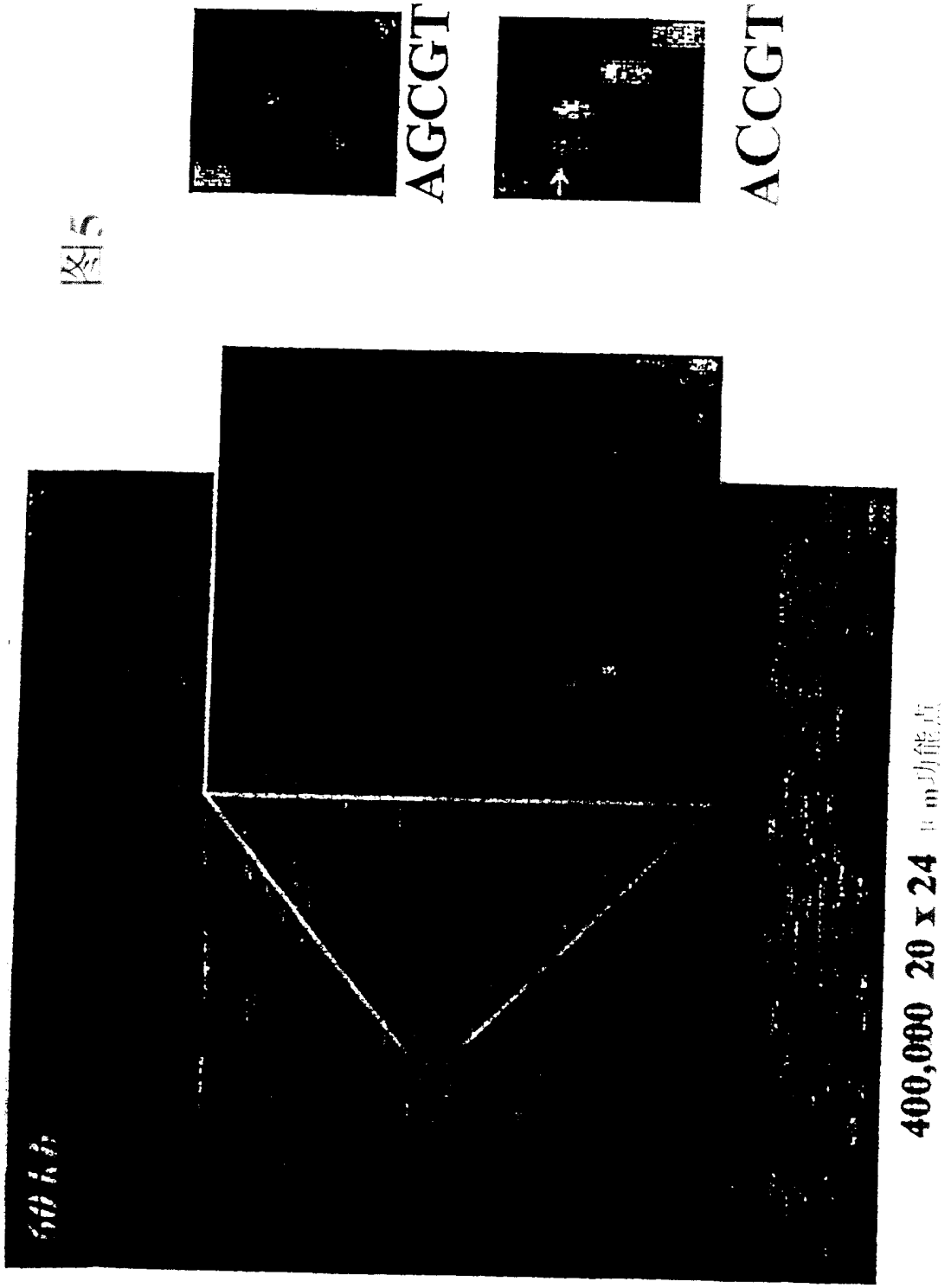


图5

图6

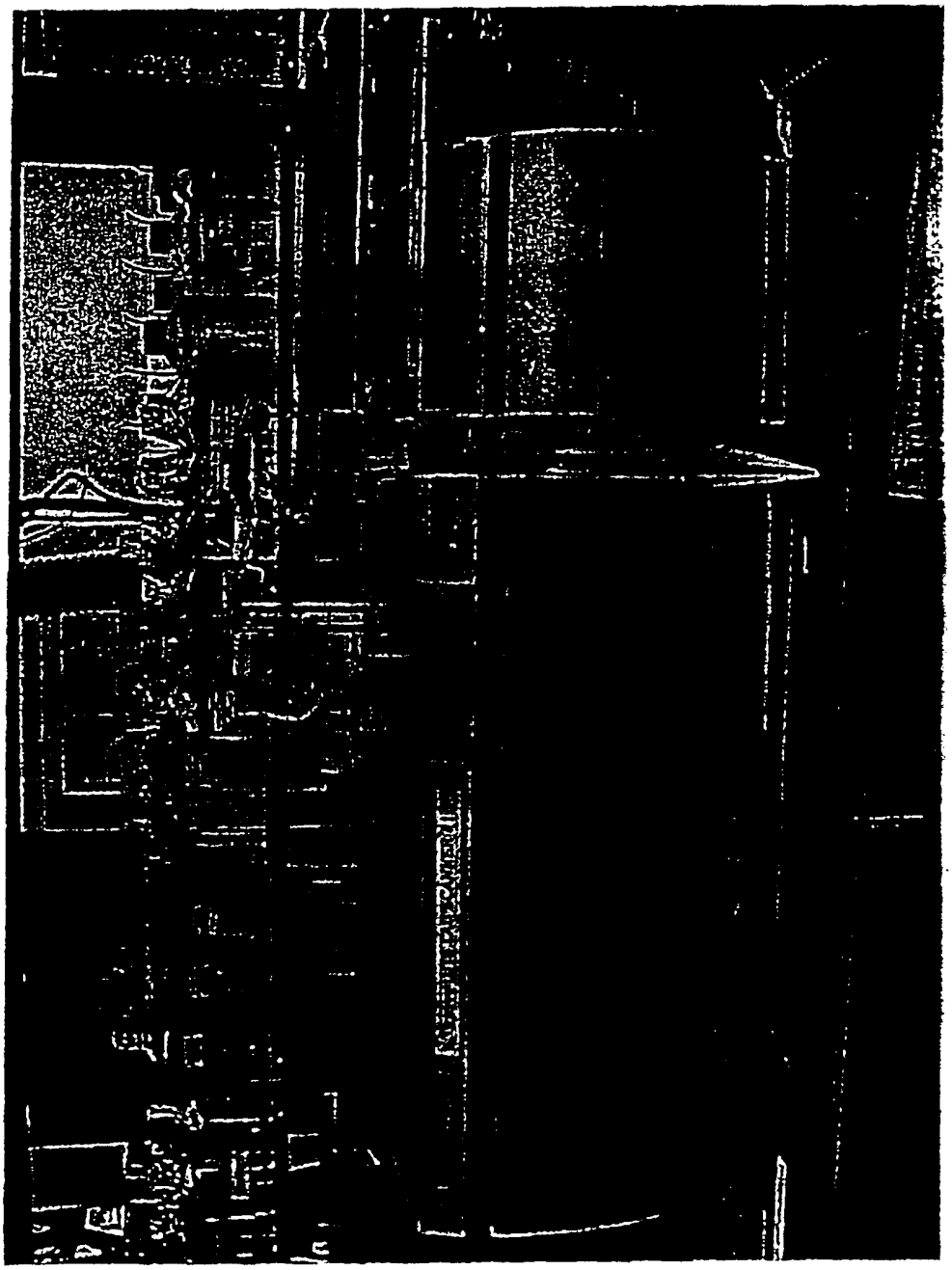


图7

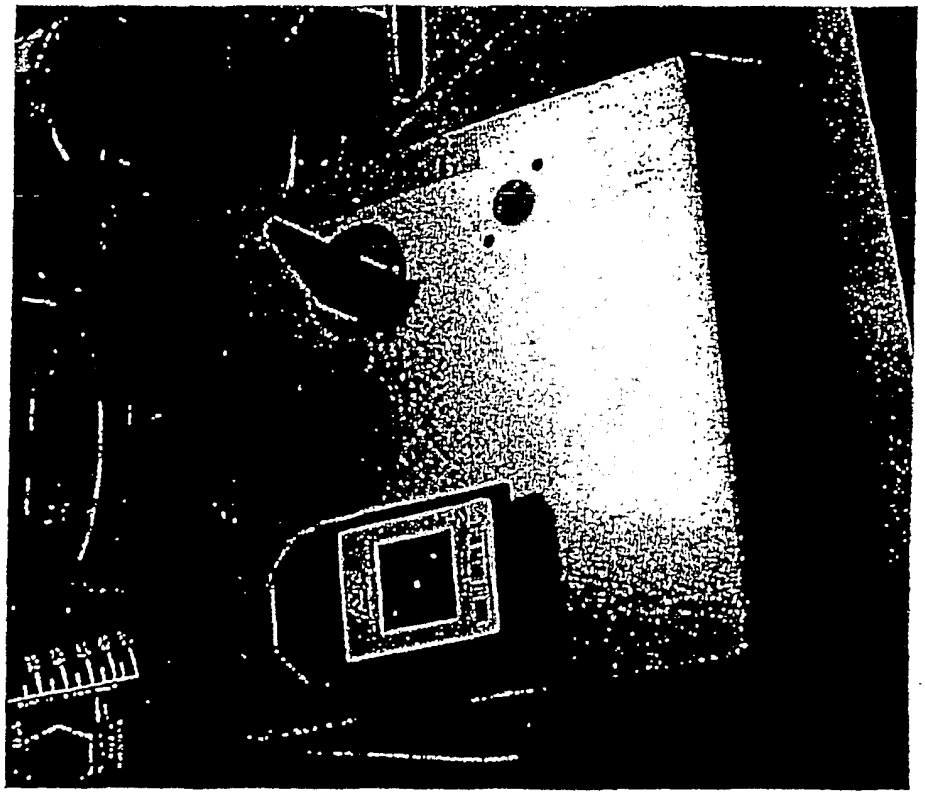
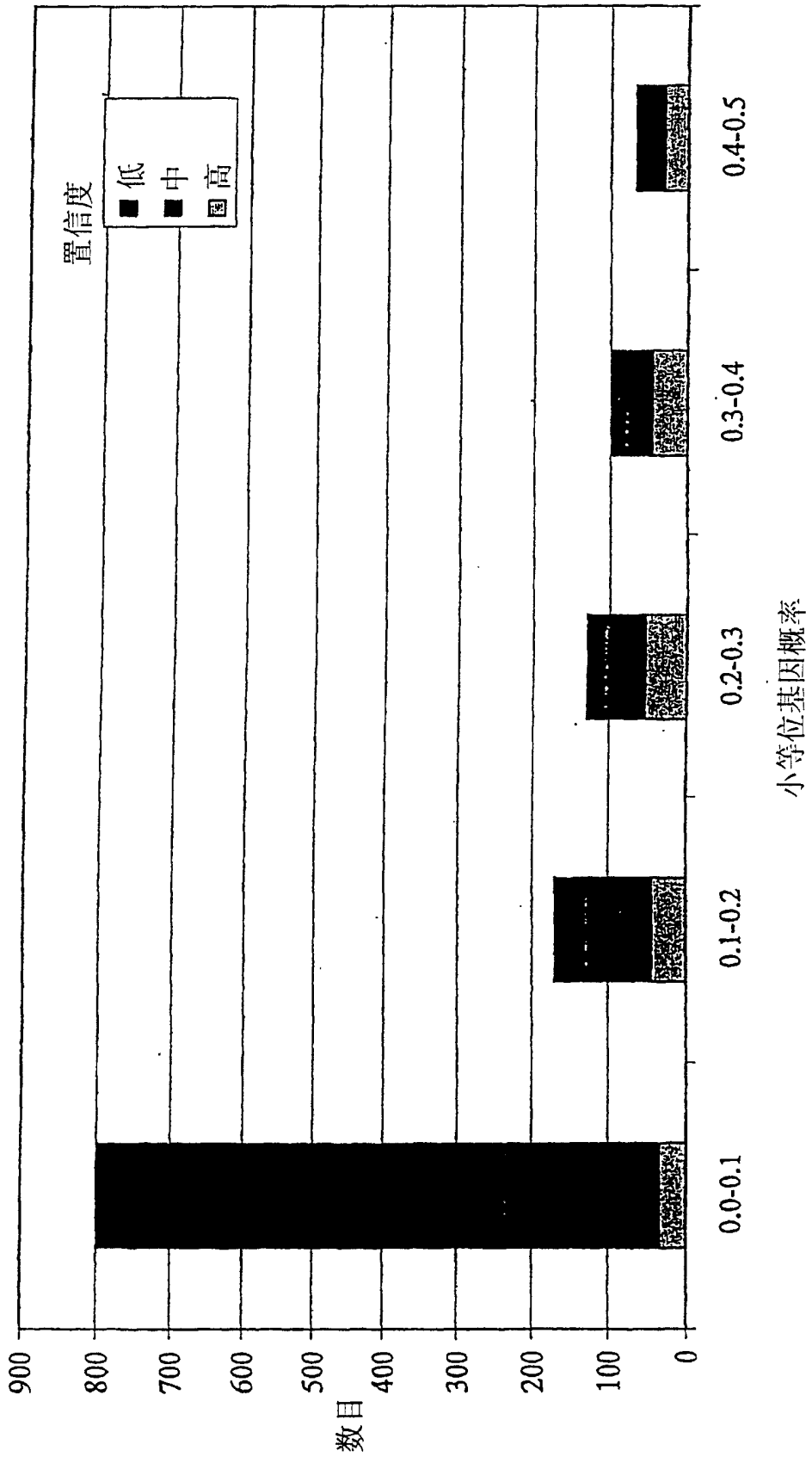


图8


置信度
低
中
高
900
800
700
600
500
400
300
200
100
0
数目
0.0-0.1
0.1-0.2
0.2-0.3
0.3-0.4
0.4-0.5
小等位基因概率