[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
C12Q 1/68 (2006.01)
C12P 19/34 (2006.01)
C07H 21/04 (2006.01)

[12] 发明专利申请公开说明书

[21] 申请号 200480010806.3

[43] 公开日 2006 年 6 月 21 日

[11] 公开号 CN 1791682A

[22] 申请日 2004.2.26
[21] 申请号 200480010806.3
[30] 优先权
[32] 2003.2.26 [33] US [31] 60/450,566
[86] 国际申请 PCT/US2004/006022 2004.2.26
[87] 国际公布 WO2004/076683 英 2004.9.10
[85] 进入国家阶段日期 2005.10.21
[71] 申请人 凯利达基因组股份有限公司
地址 美国加利福尼亚州
[72] 发明人 R·T·德拉曼尼克

[74] 专利代理机构 上海专利商标事务所有限公司
代理人 范 征

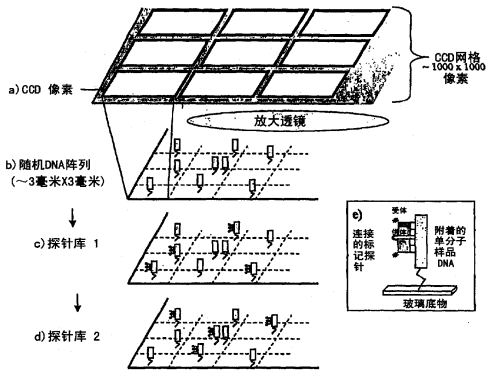
权利要求书 3 页 说明书 46 页 附图 13 页

[54] 发明名称

通过杂交进行的随机阵列 DNA 分析

[57] 摘要

本发明涉及分析单分子，如核酸的方法和装置。该单分子可来自天然样品，如不经过分离或富集单组分的细胞、组织、土壤、空气和水。本发明的某些方面，方法和装置是用于通过探针杂交进行的核酸序列分析。



1．一种分析核酸的方法，该方法包括：

a)将一群具有相同信息区序列的寡核苷酸探针杂交到单分子靶核酸阵列上，杂交在对全匹配序列的杂交比对错配序列的杂交平均更有效的条件下进行；

b)收集寡核苷酸分子与各靶分子多次连续杂交产生的信号；和

c)分析所述信号。

2．如权利要求 1 所述的方法，该方法还包括：

a)连接至少两个杂交到靶核酸分子的寡核苷酸探针。

3．一种分析核酸的方法，该方法包括：

a)将来自第一组可检测标记的寡核苷酸探针的一个或多个探针杂交到一个靶核酸的随机阵列上；

b)将来自第二组可检测标记或未标记的寡核苷酸探针的一个或多个探针杂交到所述靶核酸；

c)将与所述靶核酸杂交的各所述组的至少一个探针相连接；和

d)检测和分析所述连接的探针。

4．一种分析核酸的方法，该方法包括：

a)将第一组可检测标记的寡核苷酸探针群杂交到一个靶核酸阵列上；

b)将第二组可检测标记的寡核苷酸探针群杂交到所述靶核酸阵列上；

c)连接至少两个杂交到所述靶核酸分子的可检测标记的探针；和

d)检测所述标记的探针之间的荧光共振能量转移(FRET)信号。

5．如权利要求 4 所述的方法，该方法还包括：

a)收集多个 FRET 信号；和

b)将收集物与表征病原体的模式相比较。

6．如权利要求 1、2、3、4 或 5 所述的方法，其特征在于，所述探针或探针组连续杂交到阵列上。

7．如权利要求 1、2、3、4 或 5 所述的方法，其特征在于，所述探针选自不完全探针组、完全探针组、探针库和信息探针库。

8．如权利要求 1、2、3、4 或 5 所述的方法，其特征在于，所述探针包含至少一个修饰的碱基或通用碱基。

9．如权利要求 1、2、3、4 或 5 所述的方法，其特征在于，所述靶核酸被可检测地标记，所述标记选自:荧光团、纳米标记、化学发光标记、量子点、量子珠、荧光蛋白、具有荧光标记的树状聚体、微转发器、电子供体分子或分子结构、电子受体分子或分子结构和光反射颗粒。

10．如权利要求 1、2、3、4 或 5 所述的方法，其特征在于，至少一个探针被可检测地标记，所述标记选自:荧光团、纳米标记、化学发光标记、量子点、量子珠、荧光蛋白、具有荧光标记的树状聚体、微转发器、电子供体分子或分子结构、电子受体分子或分子结构和光反射颗粒。

11．如权利要求 1、2、3、4 或 5 所述的方法，其特征在于，用兆像素的 CCD 相机收集所述信号。

12．如权利要求 1、2、3、4 或 5 所述的方法，其特征在于，所述单分子靶核酸以 0.1-10μM 的密度随机排列在光学透明的底物上。

13．如权利要求 12 所述的方法，其特征在于，所述底物是载玻片。

14．如权利要求 1、2、3、4 或 5 所述的方法，其特征在于，所述单分子靶核酸以每个像素至少一个分子的密度随机排列在光学透明的底物上。

15．如权利要求 1、2、3、4 或 5 所述的方法，其特征在于，所述靶核酸获自病原体。

16．如权利要求 1、2、3、4 或 5 所述的方法，其特征在于，所述随机排列的靶核酸分子是原位扩增的。

17．如权利要求 1、2、3、4 或 5 所述的方法，其特征在于，所述分析包括用阵列中相同或共享序列对靶的存在计数。

18．一种扩增核酸的方法，所述方法包括下述步骤：

a)将侵入的寡核苷酸结合到靶 DNA 的一端；

b)将第一引物寡核苷酸杂交到靶 DNA 的第一可用单链位点上；

c)将第二引物寡核苷酸杂交到靶 DNA 的第二或相对单链位点上；和

d)重复步骤 1-3。

19．一种试剂盒，其包括与 FRET 受体分子相关的第一探针组、任选地与 FRET 供体分子相关的第二探针组、连接试剂和用于形成随机靶 DNA 阵列的底物。

20．一种试剂盒，其包括与 FRET 受体分子相关的第一探针组、与 FRET 供体分子相关的第二探针组和连接试剂。

21．一种组合物或混合物，其包括与 FRET 受体分子相关的第一探针组、与 FRET

供体分子相关的第二探针组和连接分子。

22. 一种装置，所述装置包括：

a)许多探针溶液的储器；和

b)一个用于杂交和展示随机靶 DNA 阵列的反应室。

23. 一种装置，所述装置包括：

a)一个样品储器或入口；

b)一个或多个化学试剂或溶液的储器，该化学试剂或溶液用于将 DNA 从样品中分离和/或片段化。

c)许多探针溶液的储器；

d)任选地包含一个混合室或管，试剂在其中与样品和/或探针混合；和

e)一个用于展示来自所述样品的随机靶 DNA 阵列的反应室。

通过杂交进行的随机阵列 DNA 分析

1. 相关申请的相互参照

本申请要求 2003 年 2 月 26 日提交的美国临时申请 60/450，566 的优先权，其题目为“通过杂交进行的随机阵列 DNA 分析”，代理人审理号 CAL-2。相关主题公开于共有、共待审的 2003 年 12 月 16 日提交的美国专利申请 10/738,108，题目为“通过用探针分子编制多个瞬时相互作用进行的单靶分子分析”，代理人审理号 CAL-3，其要求 2002 年 12 月 20 日提交的美国临时申请 60/435，539 的优先权，该临时申请的题目为“通过用探针分子编制多个瞬时相互作用进行的单靶分子分析”，代理人审理号 30311/39054。这些和所有其他专利和专利申请纳入本文作为参考。

2. 背景

2.1 技术领域

本发明涉及分析分子的方法和进行这种分析的装置。方法和装置对单分子核酸进行可信赖的分析。这种单分子可来自天然样品，如不分离或富集的单个组分细胞、组织、土壤、空气和水。在本发明的某些方面，方法和装置是用于进行核酸序列分析或核酸定量包括基因表达。

2.2 序列表

序列表列出了这里描述的多核苷酸序列，并以含有 2004 年 2 月 26 日 11:26:18AM 由 Windows 2000 操作系统的 IBM PC 创建的文件标签“CAL-2CIP PCT.txt”－8.00KB(8.192 字节)的光盘提交。名为“CAL-1CIP PCT.txt”的序列表以整体纳入本文作为参考。这里提交序列表“CAL-2CIP PCT.txt”的计算机可读形式(“CRF”)和 3 份副本(“拷贝 1”、“拷贝 2”和“拷贝 3”)。本发明申请人申明分别依照 37CRF §1.821(c)和(e)提交的序列表的 CRF 和拷贝 1、2 和 3 的内容是相同的。

2.3 背景

有三种建立的 DNA 测序技术。今天使用的主要测序方法是基于 Sanger's 双脱氧链终止法(Sanger 等，*美国国家科学院院刊* 74:5463(1977)，以整体纳入本文作为参考)，依赖于各种基于凝胶的分离设备，从手工系统到完全自动化的毛细管测序仪。Sanger 法在技术上是困难的，而且阅读长度限制在约 1 kb 或更短，要求多次阅读以

达到高准确度。第二种方法－焦测序(pyrosequencing)也用聚合酶产生序列信息，即通过监测在测试具体 DNA 碱基掺入生长链的连续循环期间产生的焦磷酸盐进行(Ronaghi,*基因组研究* 11:3(2001)，整体纳入本文作为参考)。该方法提供优秀的多孔板测定，但是仅用于非常短的 10-50 碱基片段的局部测序。此阅读长度限制对于基于序列的诊断是一个严重的限制。

上面两种技术代表直接测序法，其中链上每个碱基的位置是由直接试验顺序决定的。杂交测序(SBH)(美国专利 5,202,231；Drmanac 等，*基因组学* 4:114(1989)，将两者整体纳入本文作为参考)，使用互补核酸的碱基特异性杂交的基础生命化学，间接组装靶 DNA 的碱基顺序。在 SBH 中，将已知序列的重叠探针杂交至样品 DNA 分子，使用计算机算法所得的杂交模式用来产生靶序列(共有、共待审美国专利申请 09/874,772；Drmanac 等，*科学* 260:1649-1652(1993)；Drmanac 等，*Nat. Biotech.* 16:54-58(1998)；Drmanac 等，“用寡核苷酸探针杂交进行 DNA 测序和指纹分析”，*分析化学百科全书*，第 5232-5237 页(2000)；Drmanac 等，“杂交测序(SBH)：优点、成就及机会”，*生化工程/生物技术进展：芯片技术*，Hoheisel，J.(编辑)，第 76 卷，第 75-98 页(2002)；所有以整体纳入本文作为参考)。探针或 DNA 靶可以高密度阵列的形式排列(例如参见 Cutler 等，*基因组研究* 11:1913-1925(2001)，整体纳入本文作为参考)。SBH 法的优点包括实验的简单性，阅读长度较长、准确度较高和单一测定中可分析多个样品。

当前，迫切需要在复杂样品中可以快速并准确地检测、分析和鉴定所有潜在的病原体的新生物防御技术。现有病原体探测技术通常缺乏在样品中准确鉴定痕量的病原体的灵敏性和选择性，并且难以操作。此外，在它们的现有应用中，所有三个测序技术都需要大量的样品 DNA。通常用几种扩增方法之一制备样品，主要是 PCR。虽然与 DNA 扩增和序列制备有关的费用相当大，但是这些方法，尤其 SBH 可以提供良好的个体基因或 2-5 个基因混合物的基于序列的诊断。因此，所有现有测序方法都缺乏以接受的费用水平在复杂生物样品中进行全面的基于序列的病原体诊断和筛选提供所需的速度和效率。这在现有技术能力和新的测序需求之间造成大的差距。理想的是，合适的诊断方法应该允许对环境或临床样品中可能存在的所有重要病原体同时进行筛查，样品包括隐藏在生物体中的工程改造的病原体混合物。

全面病原体诊断的要求包括需要同时测序 10-100 个重要基因或整个基因组，从中筛查几百个病原体，并进行上千个样品的筛查。最终，对于进行连续系统测量的实验室来说，每个样品需要测序 10-100Mb 的 DNA，或每天需要测序 100Mb 至 10Gb 的

DNA。现有的测序方法比全面病原体诊断和症状前测量所需的要求，在测序通量上低100倍，而费用上高100倍。

现有的生物传感器技术使用各种分子识别策略，包括抗体、核酸探针、适体、酶、生物受体和其他小分子配体(Iqbal 等，*生物传感器和生物电子学* 15:549-578(2000)，整体纳入本文作为参考)。分子识别元件必须偶联报道分子或标记，以提供阳性检测事件。

DNA 杂交和基于抗体的技术都已广泛应用于病原体诊断。通常，基于核酸的技术比基于抗体的检测更特异性和更灵敏，但它是耗时的，而且不太耐用(Iqbal 等，2000，上述)。通常需要 DNA 扩增(通过 PCR 或克隆)或信号放大以达到可信赖的信号强度，还需要准确的先序列知识以构建病原体特异性探针。虽然单克隆抗体的开发增加了免疫测定的特异性和可靠性，但该技术相对昂贵，并容易产生假阳性信号(Doing 等，临床微生物学杂志 37:1582-1583(1999)；Marks，临床化学 48:2008-2016(2002)，将两者整体纳入本文作为参考)。其他分子识别技术如噬菌体展示、适体和小分子配体仍在其早期发展阶段，仍不能多方面足以解决所有病原体检测问题。

所有现有诊断技术的主要不利条件是它们缺乏在样品中检测和鉴定所有潜在病原体的灵敏性和多功能性。武器设计者可容易地设计新的生物战剂来阻挠大部分病原体特异性探针或免疫测定。显然迫切需要有效的基于序列的诊断。

为此，申请人开发了一种高效基因组测序系统，即通过杂交进行的基于随机 DNA 阵列的测序(rSBH)。rSBH 可用于对复杂微生物群落中存在的所有基因组进行基因组序列分析以及人类个体的基因组测序。rSBH 不需要 DNA 克隆或 DNA 分离，并减少了用本领域已知方法进行测序的费用。

4. 发明概述

本发明提供新颖方法、组合物或混合物和能够分析单分子 DNA 的装置，以对任何长 DNA 片段、片段混合物、整个基因、基因混合物、mRNA 混合物、染色体的长片段、整个染色体、染色体混合物、整个基因组或基因组混合物进行快速和准确地测序。另外，本发明提供在靶核酸内鉴定核酸序列的方法。通过连续瞬时杂交，从编制数据中获得准确而广泛的序列信息。在一个示范性实施例中，将单靶分子瞬时杂交到一种探针或探针群。杂交与一种或多种探针不再存在后，靶分子再次瞬时杂交到下一个探针或探针群。探针或探针群与以前的瞬时杂交探针可以相同或不同。编

制相同单靶分子与一种或多种同型探针分子的一系列连续结合提供了可靠的测量。因此，因为单靶分子与探针连续接触，可以提供足够数量的数据，以鉴定靶分子内的序列。通过编制数据，可以确定整个靶分子的核酸序列。

本发明还提供在单个生物体水平上分析和检测存在于复杂生物样品中的病原体，以及鉴定所有毒力控制基因的方法、组合物和装置。

本发明提供分析靶分子的方法，该方法包括下述步骤：

a)将靶分子与一种或多种探针分子在一系列连续结合反应中接触，其中每个结合都对靶分子或探针分子产生效应；和

b)编制系列连续结合反应的效应。

本发明还提供分析靶分子的方法，该方法包括下述步骤：

a)将靶分子与一种或多种探针分子在一系列连续杂交/解离反应中接触，其中每个结合都对靶分子或探针分子产生效应；和

b)编制系列连续杂交/解离反应的效应。

在某些实施例中，该系列包含至少 5、至少 10、至少 25、至少 50、至少 100 或至少 1000 个连续杂交/解离或结合反应。在一个实施例中，该系列包含至少 5 并少于 50 个连续杂交/解离或结合反应。

本发明包括实施例，其中探针分子序列或结构是已知或可确定的。这种实施例的一个优点是，它们用于在从编制一个或多个已知/可确定序列的探针效应的靶中鉴定序列。而且，当在一个靶分子内鉴定到多个重叠序列时，这种鉴定的重叠序列可用于对靶分子测序。

本发明还提供一种分析靶分子的方法，其中分析中包括的效应编制涉及时间测量(即检测到信号的时间长度或在预设时段信号的检测等)。在某些实施例中，用测定靶分子或探针分子产生荧光信号的时间来编制效应。

本发明也提供用检测仅在靶分子与探针的杂交或结合时产生的信号来编制效应的方法。方法包括通过测定信号产生的时段量来编制效应的的方法，和通过测定信号产生量来编制效应的的方法。在某些实施例中，靶分子包含荧光共振能量转移(FRET)供体和含有 FRET 受体的探针分子。在其他实施例中，靶分子包含 FRET 受体和含有 FRET 供体的探针分子。

本发明也提供对一种或多种探针的效应是探针修饰的方法。在某些实施例中，探针被连接，该方法还包括检测连接的探针。探针可用纳米标记进行标记。

在杂交或结合对探针的效应是探针修饰的实施例中，由全匹配杂交引起的修饰

比错配杂交引起的修饰更频繁，可通过检测存在相当高数量的修饰来确定全匹配。

本发明的方法包括：

a)核酸分子的片段化产生靶分子；

b)通过限制酶消化、超声波处理、氢氧化钠处理或低压剪切完成片段化；

c)可检测地标记靶分子；

d)用选自荧光标记、纳米标记、化学发光标记、量子点、量子珠、荧光蛋白、具有荧光标记的树状聚体、微转发器、电子供体分子或分子结构和光反射颗粒的标记可检测地标记靶分子和/或探针分子；

e)用电荷偶联装置(CCD)检测标记；

f)具有相同信息区的探针分子分别与相同可检测标记结合；

g)一种或多种探针分子包含多个标记；

h)将探针分子分至库中，每个库包含至少两种具有不同信息区的探针分子，每个库内的所有探针分子都与相同标记结合，该标记对与其他库相比的库是独特的；

i)通过排序杂交到靶分子的重叠探针序列，组装靶分子的序列；

j)通过排序重叠探针序列和从掺入的探针的杂交效率确定组装序列的分数/可能性/概率，组装靶分子的序列；

k)信息区中的探针各自独立，长度在 4 至 20 个核苷酸之间；

l)信息区中的探针各自独立，长度在 4 至 100 个核苷酸之间；

m)附着分子的靶序列具有的长度在约 20 至 2000 个碱基之间；

n)一种或多种探针由至少一种修饰或通用碱基组成；

o)一种或多种探针由至少一种在末端位置的通用碱基组成；

p)杂交条件有效使靶分子仅与靶分子的一部分完美互补的那些探针杂交；

q)接触包括具有互不相同的信息区的至少约 10、至少约 100、至少约 1000 或至少约 10,000 探针分子；和/或

r)用少于 1000、800、600、400、200、100、75、50、25 或 10 靶分子。

在一个实施例中，本发明的方法可用于在微生物生物薄膜中及其一定百分比的组合物中分析微生物基因组。生物薄膜群落包含的微生物包括嗜铁钩端螺菌(Leptospirillum ferriphilum)种系型、亚铁螺菌(Ferrospirillum sp.)、热氧化硫化杆菌种系型、古菌(包括 Ferroplasma acidarmanus、Aplasma、Geneplasma 种系型)和真核生物(包括 protests 和真菌)。

本发明还提供等温扩增的方法，该方法使用链取代酶，基于以侵入物寡核苷酸

引物退火为目的的单链 DNA 形成。

本发明还提供支持 rSBH 全基因组(复杂的 DNA 样品)并可处理多达 3Gbp 至 10Gbp 序列的软件。

本发明还提供试剂和试剂盒，以同时分析许多基因或诊断区、并从血样中处理和制备病原体 DNA。

本发明还包括包含探针、靶核酸和连接分子的混合物的组合物，以从血液、组织或环境样品中分析许多病原体基因或诊断区。

在考虑下面以目前优选实施例进行的本发明详述时，本领域技术人员将会明白本发明的许多其他方面和优点。

5. 附图说明

结合下面的附图可更好理解本发明的详述：

图 1 描述了接头连接和延伸。双链发夹接头(实线)被发夹端的交联碱基维持发夹形式。B 和 F 分别代表结合引物和固定的引物序列，它们的互补序列以小写字母表示。细线代表基因组序列。A)非磷酸化接头连接于基因组 DNA 在具有游离 3'末端(箭头)的链中产生缺口。B)从 3'末端延伸产生一条替代链和接头序列的复制。

图 2 描述了接头设计和附着于 DNA 片段，其中实心黑条代表基因组 DNA，F 代表游离的引物，B 代表结合的引物，f 和 b 分别代表它们的补体。

图 3 描述了芯片表面上 ampliot 的产生。A)接头捕获的基因组 DNA 解链后，一条链通过与结合引物 B 杂交而被捕获在载玻片表面上。从引物 B 的聚合酶延伸产生双链分子。B)通过加热和洗涤载玻片去除模板链，引入游离引物 F 并沿固定链延伸。C)F 引起的连续链替代扩增导致产生一条可以移动至附近的引物 B 杂交位点的链。D)将替代链用作模板，从新的引物 B 位点延伸。

图 4 描述了用 RNA 中间体产生 ampliot。T7 代表 T7 噬菌体 RNA 聚合酶启动子。A)单链接头区杂交至结合引物 B，用 DNA 聚合酶延伸形成第二条链导致双链 T7 启动子的形成。B)T7 RNA 聚合酶产生 RNA 拷贝(虚线)。C)然后，RNA 结合至附近的引物 B，用逆转录酶产生 cDNA。然后用 RNA 酶 H 破坏双链 RNA。

图 5 是描述入侵物介导的等温 DNA 扩增过程的示意图。

图 6 描述了通过杂交(rSBH)方法进行的随机阵列测序。从上至下分别为：(a)将 CCD 照相机放置在反应平台上面，用一个透镜将平台上的 1 微米^{2}面积放大并聚焦于 CCD 照相机的一个像素上。(b)阵列(～3 毫米 x3 毫米)由 1 百万或更多个 1 微米^2

面积，它作为虚拟反应池(分别对应于 CCD 相机的单个像素)。每个像素对应底物上的相同位置。在一系列及时反应中，一个 CCD 像素可以结合几个反应的数据，因此产生虚拟反应池。将 DNA 样品随机消化并排列在反应平台的表面上，平均浓度为每像素一个片段。(c)用几个信息探针库之一对该阵列进行 rSBH 组合连接。记录每个像素的信号。(d)去除第一个库的探针，用不同库或探针对该阵列进行第二轮 rSBH 组合连接。(e)由于两个相邻并互补的探针连接而显示荧光共振能量转移(FRET)信号产生分子细节的插入物，所述探针的补体由靶代表。

图 7 描述了 rSBH 反应。总内反射显微术(TIRM)检测系统造成一个渐消失区，在这个区域，增强激发仅发生在玻璃底物的上方区域。FRET 信号在探针杂交到排列的靶时产生，接着连接，因此将 FRET 对定位在渐消失区内。未连接的探针无论在溶液中游离或者瞬时杂交到靶，均不产生可检测的信号。因此，TIRM 系统的渐消失区在减少从未反应探针产生的背景噪音的同时，在需要平面内提供强信号。

图 8 描述序列组装。通常，在 SBH 方法中，用重叠阳性探针组装靶序列。在本方法中，每个碱基都被阅读数次(即用 10-mer 探针阅读 10 次等)，这保证很高准确度，即使一些探针没有被正确记录。

图 9 描述用于 rSBH 方法的微流体装置的示意图。该装置集成了 DNA 制备、随机单分子 DNA 阵列形成、组合库混合和反应室的周期装载和洗涤。当样品管附着于芯片时，用预装载的试剂进行一系列反应，以分离并片段化 DNA，该 DNA 随机连接于芯片表面，密度约为每像素一个分子。然后用微流体装置将来自信息探针库(IPPs)的 5'和 3'组的两个探针库与反应溶液混合。一组探针库用 FRET 供体标记，另一组用 FRET 受体标记。然后，将含有 DNA 连接酶的混合库转移至单分子 DNA 上面的反应室。当两种探针(每库一个)在阵列表面上面狭窄的反射区域(～100 nm)内，杂交至靶 DNA 分子的相邻互补序列时，发生可检测的连接事件。在反射区域内 5'和 3'探针的连接产生 FRET 信号，用超灵敏 CCD 相机检测并记录该信号。记录连接事件后，用洗涤溶液去除每个库混合物，将预装载在微流体芯片上的相同 IPPs 组的第二对库组合，并引入反应室。通过将两组 IPPs 内所有可能库组合，记录阵列中每个靶分子用于两个探针组内存在的探针序列的每种可能组合的存在/不存在。

图 10 描述 TIRM 装置的基本光学和光路。(a)位于棱镜顶部的传统底物和产生渐消失区的光路描述。(b)和(c)显示使用电流计控制从激光器到棱镜组件的光路。

图 11 描述 rSBH 部件和过程的示意图，显示 rSBH 装置的组件和实验方法的逐步描述。不依赖该装置收集并制备样品(步骤 1 和 2)。通过样品集成模块(组件 A)进一

步处理产生的原始样品用于 rSBH 阵列形成(步骤 3)。随后将靶排列在反应盒(组件 B)内的底物模块上。通过探针模块(组件 C)输送 SBH 探针，用 SBH 探针对样品进行 SBH 连接测定(步骤 4)。处理产生的原始数据，产生序列数据的组装(步骤 5)和解释性分析(步骤 6)。

图 12 显示 4 个打点靶的全匹配连接信号。以范围 1 至 90 微摩尔的 7 个不同浓度打点四个不同的靶。连接探针浓度(5'探针:3'探针比例是 1:1)从 0.1 至 1 皮摩尔/20 微升不等。

图 13 显示将打点靶作为其他靶的捕获探针的图示。当载玻片直接用 Tg2-5'探针和 Tgt2-3'探针杂交/连接时(圆形)，当载玻片用靶 Tgt2-Tgt1-rc 预杂交、然后用 Tgt2-5'探针和 Tgt2-3'探针连接时(方形)，测量连接信号。

6. 优选实施例的详细描述

本发明提供单分子 DNA 分析方法和装置，以对任何长 DNA 片段、片段混合物、整个基因、基因混合物、mRNA 混合物、染色体的长片段、整个染色体、染色体混合物、整个基因组或基因组混合物进行快速和准确地测序。本发明方法允许在单个生物体水平上检测存在于复杂生物样品中的病原体，和鉴定毒力控制基因。本发明方法将杂交和尤其是，通过杂交测序技术(SBH)与总内反射显微术(TIRM)或用荧光、纳米粒或电学方法的其它灵敏的光学方法组合。本发明也提供了样品排列技术，该技术建立了与超灵敏电荷偶联装置(CCD)相机的个体像素相关的虚拟反应室。用荧光标记的寡核苷酸探针的完全/通用组信息库和组合连接方法，对排列的基因组进行重复探测，以便解码它们的序列。用生物信息学算法(共有、共待审的美国专利申请 09/874,772；Drmanac 等，*科学* 260:1649-1652(1993)；Drmanac 等，*Nat. Biotech.* 16:54-58(1998)；“用寡核苷酸探针杂交进行 DNA 测序和指纹分析”，*分析化学百科全书*，第 5232-5237 页(2000)；Drmanac 等，“杂交测序(SBH)：优点、成就及机会”，*生化工程/生物技术进展：芯片技术*，Hoheisel，J.(编辑)，第 76 卷，第 75-98 页(2002)；所有以整体纳入本文作为参考)将信息性荧光信号转换成组装的序列数据。该装置用定位在诊断实验室或小移动实验室的单一小型装置每小时可以测序超过 100 兆 DNA 碱基(30,000 碱基/秒)。由于随机单分子阵列的大容量，可用本发明方法检测、鉴定并测序复杂生物样品内痕量的病原体 DNA。因此，随机阵列 SBH(rSBH)提供必需技术，以使 DNA 测序除其它测序应用外，在对付生物战剂的防御中起重要的作用。

本发明提供单 DNA 分子分析方法，在病原体、宿主和环境 DNA 的复杂生物混合物中快速、准确地检测并鉴定任何病原体，并通常分析任何 DNA，包括人类个体 DNA。本发明方法允许在单个生物体水平上检测样品中存在的病原体，和鉴定所有毒力控制基因。本发明方法将小组通用信息探针库(IPPs)的组合杂交/连接方法直接或个体排列分子原位扩增约 10-或 100-、或 1000-或 10,000-倍后，用于随机单分子阵列。

在一个典型测试中，将获自一个样品的几百万随机排列单 DNA 分子与 IPPs 对杂交，该 IPPs 对代表所有长度为 8 至 10 碱基的可能探针序列的通用库。当两种探针杂交到靶 DNA 中相邻的互补序列时，它们连接产生一个靶分子的阳性记录，从重叠探针序列信息编制累计的这种记录组以组装靶序列。

在本发明的另一实施方式中，单个靶的特征或序列可用于组装整个基因或基因组的较长序列。此外，通过计算来自相同基因的相同分子或相同片段在阵列中出现多少次，可以得到定量的基因表达或病原体 DNA，这种数据可以与获得的序列组合。

SBH 是一种发展良好的技术，可通过很多本领域技术人员已知的方法实施。具体地，该技术涉及下述文献讨论的通过杂交进行测序，这些文献以整体纳入本文作为参考：Bains 和 Smith，*J. Theor. Biol.* 135:303-307(1988)；Beaucage 和 Caruthers，*Tetrahedron Lett.* 22:1859-1862 (1981)；Broude 等，*美国国家科学院院刊* 91:3072-3076 (1994)；Breslauer 等，*美国国家科学院院刊* 83:3746-3750 (1986)；Doty 等，*美国国家科学院院刊* 46 :461-466 (1990)；Chee 等，*科学* 274:610-614 (1996)；Cheng 等，*Nat. Biotechnol.* 16:541-546 (1998)；Dianzani 等，*Genomics* 11:48- 53 (1991)；Drmanac 的 PCT 国际专利申请 WO 95/09248；Drmanac 的 PCT 国际专利申请 WO 96/17957；Drmanac 的 PCT 国际专利申请 WO 98/31836；Drmanac 等的 PCT 国际专利申请 WO 99/09217；Drmanac 等的 PCT 国际专利申请 WO00/40758；PCT 国际专利申请 WO 56937；Drmanac 和 Jin 共有、共待审的美国专利申请 09/874,772；Drmanac 和 Crkvenjakov，*Scientia Yugoslaviea* 16:99-107 (1990)；Drmanac 和 Crkvenjakov，*Intl. J. Genome Res.* 1:59-79 (1992)；Drmanac 和 Drmanac，*Meth. Enzymology* 303:165-178 (1999)；Drmanac 等，美国专利 5,202，231；Drmanac 等，*Nucl. Acids Res.* 14 :4691-4692 (1986)；Drmanac 等，*Genomics* 4:114-128 (1989)；Drmanac 等，*J. Biomol. Struct. Dyn.* 8:1085-1102(1991)；Drmanac 等，“通过杂交进行部分测序:在基因组分析中的概念和应用”，第一届电泳、超级计算和人类基因组国际会议，第 60-74 页，世界科学出版公司，新加坡，马来西亚(1991)；Drmanac 等，第一届电泳、超级计算和人类基因组国际会议会刊，Cantor 等编辑，

世界科学出版公司，新加坡，7-59 页(1991)；Drmanac 等，*Nucl. Acids Res.* 19:5839-5842 (1991)；Drmanac 等，*Electrophoresis* 13:566-573 (1992)；Drmanac 等，*科学* 260:1649-1652 (1993)；Drmanac 等，*DNA and Cell Biol.* 9:527-534 (1994)；Drmanac 等，*Genomics* 37:29-40 (1996)；Drmanac 等，*Nature Biotechnology* 16:54-58 (1998)；Gunderson 等，*Genome Res.* 8 :1142-1153 (1998)；Hacia 等，*Nature Genetics* 14:441-447 (1996)；Hacia 等，*Genome Res.* 8:1245-1258 (1998)；Hoheisel 等，*Mol. Gen.* 220:903-14:125-132 (1991)；Hoheisel 等，*细胞* 73:109-120 (1993)；Holey 等，*科学* 147:1462-1465 (1965)；Housby 和 Southern，*Nucl. Acids Res.* 26:4259-4266 (1998)；Hunkapillar 等，*科学* 254:59-63 (1991)；Khrapko，*FEBS Lett.* 256:118-122 (1989)；Kozal 等，*Nature Medicine* 7:753-759 (1996)；Labat 和 Drmanac，“排序模拟和与少量寡聚探针杂交的随机 DNA 克隆的序列重建，第二届电泳、超级计算和人类基因组国际会议，第 555-565 页，世界科学出版公司，新加坡，马来西亚(1992)；Lehrach 等，基因组分析:遗传和物理作图 1:39-81 (1990)，冷泉港实验室出版社；Lysov 等，*Dokl. Akad. Nauk. SSSR* 303:1508-1511 (1988)；Lockhart 等，*Nat. Biotechnol.* 14:1675-1680 (1996)；Maxam 和 Gilbert，美国国家科学院院刊 74 :560-564 (1977)；Meier 等，*Nucl. Acids Res.* 26:2216-2223 (1998)；Michiels 等，*CABIOS* 3 :203-210 (1987)；Milosavljevic 等，*Genome Res.* 6:132-141 (1996)；Milosavljevic 等，*Genomics* 37:77-86 (1996)；Nikiforov 等，*Nucl. Acids Res.* 22:4167-4175 (1994)；Pevzner 和 Lipschutz，“向 DNA 测序芯片”，《计算机科学的数学基础》(1994)；Poustka 和 Lehrach，*Trends Genet.* 2:174-179 (1986)；Privara 等，编辑，第 143-158 页，第 19 届国际研讨会会刊，MFCS 94 年，Kosice，Slovakia，Springer-Verlag，柏林(1995)；Saiki 等，美国国家科学院院刊 86:6230-6234 (1989) ；Sanger 等，美国国家科学院院刊 74:5463-5467 (1977)；Scholler 等，*Nucl. Acids Res.* 23:3842-3849 (1995)；Southern 的 PCT 国际申请 WO 89/10977；Southern 的美国专利 5,700,637；Southern 等，*Genomics* 13:1008-1017 (1992)；Strezoska 等，美国国家科学院院刊 88 :10089-10093 (1991) ；Sugimoto 等，*Nucl. Acid Res.* 24:4501-4505 (1996)；Wallace 等，*Nucl. Acids Res.* 6:3543-3557 (1979)；Wang 等，*科学* 280 :1077-1082 (1998)；Wetmur，*Crit. Rev. Biochem. Mol. Biol.* 26:227-259 (1991).

rSBH 的优点：

rSBH 使附着在合适距离的两个靶 DNA 分子之间的靶-靶阻断相互作用减至最小

或消除。每点的 DNA 序列(200-300 间碱基)的低复杂性降低了可互相阻断的反向重复的可能性。在一些片段中，用平均每 20 个源 DNA 碱基一个切口，分开了回文序列和发夹臂，并连接至非互补引物 DNA。假阳性被最小化，因为重叠片段具有不同的重复和/或强的不匹配序列。探针-探针连接产物可通过洗涤除去杂交/连接特异性和差示全匹配/不匹配稳定性的组合用于由连接制成 11-13-mer 探针，对于产生更准确的数据是可能的。rSBH 提供了在溶液中用 3 探针连接的有效方法，包括短 DNA 的分析。模式化探针库可有效用于两种探针组分，以提供更具有信息的数据。另一优点是只需要非常小量的源 DNA。消除了对制备标准探针-点阵列的需要，因此成本降低。rSBH 提供高达 1000 样品的多重测序，样品用不同引物和接头标记。此外，本发明提供在高达一百万个体样品的库中对单个变体的检测。通过计算两个变体，可以检测杂合子。本发明在每个表面提供比标准阵列多 10 至 100,000 倍的信息。

6.1 多核苷酸的制备和标记

本发明的实施利用各种多核苷酸。一般地，一些多核苷酸被可检测地标记。本发明实施中所用的多核苷酸种类包括靶核酸和探针。

术语“探针”指相对短的多核苷酸，优选 DNA。探针优选比靶核酸至少短 1 个碱基，探针的长度更优选为 25 个碱基或更短，长度更优选为 20 个碱基或更短。当然，探针的最优长度将取决于分析的靶核酸的长度。在由约 100 或更少碱基组成的靶核酸的从头测序(不用参考序列)中，探针优选至少 7-mer；对于约 100-200 个碱基的靶核酸来说，探针优选至少 8-mer；对于约 200-400 个碱基的靶核酸来说，探针优选至少 9-mer；对于约 400-800 个碱基的靶核酸来说，探针优选至少 10-mer；对于约 800-1600 个碱基的靶核酸来说，探针优选至少 11-mer；对于约 1600-3200 个碱基的靶核酸来说，探针优选至少 12-mer；对于约 3200-6400 个碱基的靶核酸来说，探针优选至少 13-mer；对于约 6400-12,800 个碱基的靶核酸来说，探针优选至少 14-mer。对于靶核酸的长度每再增加两倍，最优探针长度增加一个附加碱基。

本领域技术人员将认识，对于 SBH 应用所用的连接探针来说，上述探针长度是连接后的。探针通常为单链，虽然在一些应用中可使用双链探针。

虽然探针一般由天然产生的碱基和磷酸二酯骨架组成，但它们不必如此。例如，探针可由一个或多个修饰的碱基组成，如 7-脱氮鸟苷或通用“M”碱基，或一个或多个修饰的骨架链接，如硫代磷酸酯。唯一的要求是探针能够杂交到靶核酸。已知各种修饰碱基和骨架链接可以与本发明联合使用，这对本领域技术人员来说将是显而易见的。

上述的探针长度指探针信息内容的长度，不一定是探针的真实物理长度。用于SBH的探针经常包含简并端，该简并端并不有助于探针的信息内容。例如，SBH应用经常用式$N_xB_yN_z$的探针混合物，其中N代表4种碱基中的任意一种，它因给定混合物中的多核苷酸而改变，B代表4种碱基中的任意一种，但对给定混合物中的各多核苷酸是相同的，x、y和z是整数。一般地，x和z是0和5之间的独立整数而y是4和20之间的整数。已知碱基B_y的数量定义了多核苷酸的“信息内容”，因为简并端不有助于探针的信息内容。包含固定多核苷酸的混合物的线性阵列在，例如通过杂交测序中是有用的。在这些简并探针混合物中错配的杂交区别仅指信息内容的长度，而并非全部物理长度。

可以用本领域中的公知技术制备本发明中使用的探针，例如用 Applied Biosystems 合成仪进行自动合成。此外，可以用 Genosys Biotechnologies 公司的方法制备探针，该方法使用大量多孔特氟隆圆片。对于本发明目的来说，所用的寡核苷酸探针来源并不重要，本领域技术人员将明白，用现在已知或以后开发的其他方法制备寡核苷酸也是足够的。

术语“靶核酸”指需要序列信息的一种多核苷酸或一种多核苷酸的一些部分，一般是在 SBH 测定中测序的多核苷酸。靶核酸可以是任何数量长度的核苷酸，取决于探针的长度，但一般长度约为100、200、400、800、1600、3200、6400或更多核苷酸。一个样品一般具有多于100、多于1000、多于10,000、多于100,000、多于一百万或多于一千万靶。靶核酸可由核糖核苷酸、脱氧核糖核苷酸或其混合物组成。一般地，靶核酸是DNA。虽然靶核酸可以是双链，但是它优选是单链。而且，靶核酸实际上可获自任何来源。在使用 SBH 测定前，根据它的长度，将其优选地剪切成上述大小的片段。与探针相似，靶核酸可由一种或多种修饰碱基或骨架链接组成。

靶核酸可获自任何合适的来源，如 cDNA、基因组 DNA、染色体 DNA、微切割染色体条带、粘粒或酵母人工染色体(YAC)插入物和 RNA，包括没有经过任何扩增步骤的mRNA。例如，Sambrook 等《分子克隆：实验室手册》，冷泉港出版社，纽约(1989)，整体纳入本文作为参考，该书描述了三种从哺乳动物细胞中分离高分子量 DNA 的方案(第9.14-9.23页)。

然后，一般用本领域技术人员已知的任何方法将多核苷酸片段化，这些方法包括，例如，如Sambrook等(1989)在第9.24-9.28页所述的使用限制酶消化、超声波剪切和NaOH处理。尤其适用于将DNA片段化的方法是，用两个碱基识别内切核酸酶CviJI，由Fitzgerald等，*Nucl. Acids Res.* 20:3753-3762(1992)所述，整体纳入

本文作为参考。

在一个优选实施方式中，制备靶核酸，使它们不能相互连接，例如通过用磷酸酶(即小牛肠磷酸酶)处理由酶消化或物理剪切获得的片段化的核酸。此外，样品核酸的非可连接片段可通过在样品核酸的 Sanger 双脱氧测序反应中用随机引物(即 N_5-N_9，其中 N=A、G、T 或 C)获得，该随机引物在 5'末端没有磷酸。

在大多数情况下，将 DNA 变性产生可用于杂交的单链是重要的。这可通过将 DNA 溶液在 80-90℃孵育 2-5 分钟完成。然后将该溶液迅速冷却至 2℃，防止 DNA 片段在与探针接触之前复性。

可以可检测地标记探针和/或靶核酸。事实上，任何产生可检测信号的标记和能够被固定在底物或附着于多核苷酸的标记均可以与本发明的阵列联合使用。产生的信号优适合于定量。合适的标记包括但不限于，放射性同位素、荧光团、发色团、化学发光部分。

由于它们易检测，优选标记荧光团的多核苷酸序列。书中已有适于标记多核苷酸的荧光团的描述，例如分子探针目录(Molecular Probes 公司，俄勒冈州 Eugene)和其中引用的参考文献。将荧光团标记连接到多核苷酸的方法是公知的，可在，例如 Goodchild，*Bioconjug. Chem.* 1:165-187(1990)中找到，以整体纳入本文作为参考。优选荧光团标记是 Cy5 染料，它可从 Amersham Biosciences 购得。

此外，探针或靶可用本领域任何其他已知技术标记。优选技术包括直接化学标记法和酶标记法，如激酶化和切口平移。标记的探针可容易地购自各种商业来源，包括 GENSET，而非合成。

通常，可将标记连接于探针或靶多核苷酸的任何部分，包括一个或多个碱基的游离末端。当标记通过多核苷酸连接于固体支持物时，它必须定位在可以用错配特异性内切核酸酶切割从固体支持物上释放的位置，如共有、共待审的美国专利申请 09/858,408 所述(整体纳入本文作为参考)。优选的是，该标记位置并不干扰标记的多核苷酸的杂交、连接、切割或其他杂交后修饰。

本发明的一些实施方式使用多种标记，即用许多可区别标记(如不同的荧光团)。多重标记允许在一次杂交反应中同时检测许多序列。例如，4 个颜色的多重标记通过附加因子 4 减少所需的杂交数。

其他实施方式使用探针信息库减少通常在 SBH 方案中发现的冗余，因此减少明确确定靶 DNA 序列所需的杂交反应数。在共有、待审的美国专利申请 09/479,608 中可以发现探针信息库及其使用方法，整体纳入本文作为参考。

6.2 多核苷酸附着于固体底物

本发明的一些实施方式需要将多核苷酸，例如靶 DNA 片段附着于固体底物。在优选实施方式中，可检测地标记适合的 DNA 样品，并随机地附着于固体底物，浓度为每像素 1 个片段。

固体底物的特性和几何学取决于各种因素，其中包括阵列类型和附着方式(例即共价或非共价)。通常，底物可由任何允许固定多核苷酸的材料组成，且这种材料不会在用来杂交和/或变性核酸的条件下解链或显著降解。此外，考虑到共价固定，底物应该可与活性基团激活，该活性基团能够与待固定的多核苷酸形成共价键。

很多适于用作本发明底物的材料在本领域中已有描述。在优选实施方式中，该底物由光学透明的物质，如载玻片组成。其他合适的示范材料包括，例如丙烯酸、苯乙烯-甲基异丁烯酸共聚物，乙烯/丙烯酸，丙烯腈-丁二烯-苯乙烯(ABS) ABS/聚碳酸酯，ABS/聚砜，ABS/聚氯乙烯，乙烯基丙烯，乙烯乙酸乙烯酯(EVA)，硝酸纤维素，尼龙(包括尼龙 6、尼龙 6/6、尼龙 6/6-6、尼龙 6/10、尼龙 6/12、尼龙 11 和尼龙 12)，聚丙烯腈(PAN)，聚丙烯酸酯，聚碳酸酯，聚对苯二甲酸丁烯酯(PBT)，聚对苯二甲酸乙二醇酯(PET)，聚乙稀(包括低密度、线性低密度、高密度、交联和超高分子量级)，聚丙烯均聚物，聚丙烯共聚物，聚苯乙烯(包括通用级和高冲击级)，聚四氟乙烯(PTFE)，氟化乙烯-丙烯(FEP)，乙烯-四氟乙烯(ETFE)，全氟化烷氧基乙烯(PFA)，聚氟乙烯(PVF)，聚偏氟乙烯(PVDF)，聚三氟氯乙烯(PCTFE)，聚乙烯-三氟氯乙烯(ECTFE)，聚乙烯醇(PVA)，硅苯乙烯-丙烯腈(SAN)，苯乙烯马来酐(SMA)，金属氧化物和玻璃。

通常，多核苷酸片段可通过合适活性基团结合于支持物。这种活性基团是本领域公知的，包括，例如，氨基(-NH2)、羟基(-OH)或羧基(-COOH)。可通过本领域技术人员已知的任何方法，用任何合适支持物，如玻璃来制备支持物连接的多核苷酸片段。可通过很多方法完成固定，包括，例如使用被动吸附(Inouye 和 Hondo，*J. Clin. Microbiol.* 28: 1469-1472 (1990)，整体纳入本文作为参考)、使用紫外光(Dahlen 等，*Mol. Cell Probes* 1: 159-168 (1987)，整体纳入本文作为参考)，或碱基修饰 DNA 的共价结合(Keller 等，*Anal. Biochem.* 170: 441-451 (1988)，Keller 等，*Anal. Biochem.* 177: 392-395 (1989)，两者整体纳入本文作为参考)或在探针和支持物之间形成酰胺基(Zhang 等，*Nucl. Acids Res.* 19: 3929-3933 (1991)，整体纳入本文作为参考)。

考虑到，进一步适合使用本发明的方法是 PCT 专利申请 WO 90/03382(Southern

等)中描述的方法，纳入本文作为参考。该制备结合到支持物的多核苷酸片段的方法包括将核苷 3'试剂通过磷酸基团的共价磷酸二酯键连接到支持物载有的脂肪族羟基基团。然后，在支持的核苷上合成寡核苷酸，在不从支持物切下寡核苷酸链的标准条件下，将保护基团从合成的寡核苷酸链去除。合适的试剂包括核苷亚磷酰胺和核苷磷化氢。

此外，可寻址激光激活的光去保护可用于寡核苷酸直接在玻璃表面上的化学合成，如 Fodor 等，*科学* 251: 767-773 (1991)所述，纳入本文作为参考。

一种具体的制备支持物结合的多核苷酸片段的方式是用 Pease 等，美国国家科学院院刊 91: 5022- 5026 (1994)(纳入本文作为参考)所述的光发生合成。这些作者使用现有的照相平版印刷技术产生固定了寡核苷酸探针的阵列，即 DNA 芯片。这些把光用于在高密度、小型化阵列中指导寡核苷酸探针合成的方法，使用了光不稳定性 5'-保护的 N-酰基-脱氧核苷亚磷酰胺、表面连接化学和各种组合合成策略。可用此方法产生 256 个空间位置确定的寡核苷酸探针的基质。然后用于如上所述的 SBH 测序。

在一个优选实施方式中，通过连接部分将本发明的 DNA 片段连接到固体基质。该连接可由能够形成至少两个共价键的原子组成，例如碳、硅、氧、硫、磷等，或由能够形成至少两个共价键的分子组成，例如糖-磷酸基团、氨基酸、肽、核苷、核苷酸、糖、碳水化合物、芳环、烃环、线性和分支烃等。在本发明的一个尤其优选的实施方式中，连接部分由烯化二醇部分组成。在优选实施方式中，将可检测标记连接到 DNA 片段(即靶 DNA)。

6.3 在固体支持物上形成可检测标记的双链体

在本发明的一个优选实施方式中，一个标记探针通过互补碱基配对相互作用连接到一个可检测的标记靶核酸，该核酸本身连接到固体支持物作为多核苷酸阵列的一部分，因此形成双链体。在另一优选实施方式中，将标记探针共价附着，即连接到另一探针，该探针通过互补碱基配对相互作用连接到一个靶核酸，该核酸本身连接到固体支持物作为空间可寻址的多核苷酸阵列的一部分，如果两种探针以连续方式杂交到靶核酸。

这里使用的核苷酸碱基“配对”或“互补”，如果它们在特定条件下形成稳定的双链体或结合对。一个碱基对于另一碱基的特异性是由碱基上的氢键供体和受体的有效性和方向指定的。例如，在通常用于杂交测定的条件下，腺嘌呤(A)与胸腺嘧啶(T)配对，而不与鸟嘌呤(G)或胞嘧啶(C)配对。相似地，G 与 C 配对，而不是 A 或 T。

以不太特异的方式相互作用的碱基如次黄嘌呤或通用碱基(M 碱基，Nichols 等，*自然* 369: 492-493 (1994)，整体纳入本文作为参考)与那些在特定条件下与其形成稳定双链体的碱基互补，或其他修饰碱基，例如甲基化碱基。未互相互补的核苷酸碱基被称为“错配”。

一对多核苷酸，如一个探针和一个靶核酸，如果在特定条件下核酸由互补核苷酸碱基配对介导的相互作用互相杂交、因此形成双链体，则称为“互补”或“配对”。在两个多核苷酸之间形成的双链体可以包括一个或多个碱基错配。这样的双链体被称为“错配双链体”或杂合双链体。杂交条件越宽松，越可能容忍错配和可形成相对稳定的错配双链体。

配对多核苷酸的亚类，称为“完美互补”或“完美配对”多核苷酸，由含有互相互补的连续碱基序列的成对多核苷酸组成，其中没有错配(即没有任何环境序列效应，形成的双链体具有该具体核酸序列的最大结合能)。“完美互补”和“完美配对”的含义也包括具有类似物或修饰核苷酸的多核苷酸和双链体。对于类似物或修饰核苷酸的“完美配对”是根据针对该类似物或修饰核苷酸选择的“完美配对规则”(例如具有具体类似物或修饰核苷酸的最大结合能的结合对)来判断的。

在使用上述具有 $N_xB_yN_z$ 类型的简并末端的探针库情况下，完美配对包括探针的信息内容区，即 B_y 区完美配对的任何双链体。N 区中的区别错配并不影响杂交实验的结果，因为这种错配不干扰由实验得来的信息。

在本发明尤其优选的实施例中，提供了多核苷酸阵列，其中在特定条件下提供在固体底物上形成靶 DNA 片段，该特定条件允许它们与至少一组溶液中提供的可检测标记的寡核苷酸探针杂交。在组内或组间，探针长度可以相同或不同。确定合适杂交条件的准则可以在文献中找到，例如 Drmanac 等，(1990)，Khrapko 等(1991)，Broude 等，(1994) (所有在上述已引用)和 WO 98/31836,以整体纳入本文作为参考。这些文章阐明了杂交温度范围、缓冲液和适用于 SBH 最初步骤中的洗涤步骤。探针组可以分别或同时应用于靶核酸。

杂交到靶核酸上的连续位点的探针是相互共价附着或连接的。连接可以通过化学连接剂(如水溶性碳二亚胺或溴化氰)、通过连接酶如市售的 T_4 DNA 连接酶、通过层积作用或通过任何其他引起相邻探针间形成化学键的方法进行。确定合适连接条件的准则可以在文献中找到，例如共有美国专利申请 09/458,900、09/479,608 和 10/738,108，以整体纳入本文作为参考。

6.4 随机阵列 SBH(rSBH)

本发明方法使用随机阵列 SBH(rSBH),它将组合连接方法延伸至单分子阵列，大大增加本发明方法的灵敏度和能力。rSBH 依赖通过标记寡核苷酸信息库对随机排列的 DNA 片段进行连续审查。在本发明方法中，待测序的复杂的 DNA 混合物展示在总内荧光反射显微(TIRM)平台的聚焦平面内的光学透明表面上，用超灵敏兆像素 CCD 相机连续监测。DNA 片段以约每平方微米 1 至 3 个分子的浓度排列，该面积与一个单独 CCD 像素对应。将 TIRM 用于显现焦点和被研究物体与连接表面间的紧密接触。在 TIRM 中，来自内反射激发源的渐消失区，在表面或表面附近选择性激发荧光分子，产生非常低的背景散射光和良好的信号背景比。可使背景与其相关噪声足够低，以在环境条件下检测单荧光分子(参见 Abney 等，*Biophys. J.* 61:542-552 (1992)；Ambrose 等，*Cytometry* 36:224-231 (1999)；Axelrod，*Traffic* 2:764-774 (2001)；Fang 和 Tan，“用渐消失波激发研究单分子成像和相互作用”，美国生物技术实验室(ABL)应用说明书，2000 年 4 月；Kawano 和 Enders，“总内反射荧光显微术”，国生物技术实验室(ABL)应用说明书，1999 年 12 月；Reichert 和 Truskey，*J. Cell Sci.* 96 (第 2 部分):219-230 (1990)，以整体纳入本文作为参考)。

用微流体技术，标记供体和受体荧光团的探针库对与 DNA 连接酶混合，且朝向随机阵列。当探针杂交到靶片段的相邻位点，它们连接在一起，产生荧光共振能量转移(FRET)信号。FRET 是距离依赖(10-100 埃之间)的两种荧光分子的电子激发态间的相互作用，其中激发由供体分子转移至受体分子而不发射光子(Didenko，*Biotechniques* 31:1106-1121 (2001)；Ha，*Methods* 25:78-86 (2001)；Klostermeier 和 Millar，*Biopolymers* 61:159-179 (2001-2002)，以整体纳入本文作为参考)。这些信号可以通过 CCD 相机检测，表明在片段内有配对序列。一旦检测到来自第一个库的信号，就去除探针并用连续循环测试不同探针组合。根据由几百个独立杂交/连接事件产生的荧光信号编制每个 DNA 片段的整个序列。

虽然只有一个可检测的颜色就能满足要求，但是多个颜色会增加多种组合并提高系统的效率。本领域现有状态提示可以同时使用四种颜色。除传统的直接荧光策略之外，也可使用时间分辨系统和时间分辨 FRET 信号系统(Didenko，*Biotechniques* 31:1106-1121 (2001)，整体纳入本文作为参考)。也可使用新的常规化学方法，如量子点增强的三重 FRET 系统。可以用树状聚体技术和相关的信号放大技术克服信号微弱。

不像传统杂交方法，本发明方法依赖杂交和连接的协同作用，其中来自两个库的短探针连接在一起产生具有更多信息容量的较长探针。例如，两组 1024 个 5-mer

寡核苷酸可以组合以检测超过一百万可能的 10-mer 序列串。信息探针库的使用(其中所有探针共享一个共同标记)大大简化了方法，允许几百万潜在探针配对随仅几百个库组合出现。多个重叠探针阅读连续碱基使得从获得的杂交模式准确测定 DNA 序列。通过将它们的用途延伸到单分子测序，增强了上述的组合连接和信息库技术。

6.5 结构的随机 DNA 制备

A. DNA 分离和最初片段化

用建立良好的基本方案(Sambrook 等,上述,1999;《新编分子生物学实验指南》, Ausubel 等编辑，John Wiley and Sons 公司，纽约，1999，将两者整体纳入本文作为参考)或商业试剂盒[如：可从 QIAGEN(Valencia，CA)或 Promega (Madison，Wl)购得的试剂盒]裂解细胞和分离 DNA。重要的要求是：1)DNA 不含 DNA 处理酶和杂质盐；2)整个基因组被同样地代表；和 3)DNA 片段长度在～5,000 和～10,000 个碱基之间。不需要对 DNA 进行消化，因为裂解和提取中产生的剪切力将产生所需范围内的片段。在另一实施方式中，通过酶片段化可产生较短片段(1-5kb)。10-100 拷贝的输入基因组数量将保证整个基因组重叠和允许阵列上靶的差捕获率。还有一个实施方式提供载体、在小量 DNA 情况下使用的循环合成的双链 DNA。

B. DNA 标准化

在一些实施方式中，环境样品的标准化对于减少普遍种类的 DNA 成分以使每阵列测序的不同种类的总数量最大是必需的。因为 rSBH 要求少至 10 个基因组等效物，所以可以进行彻底的 DNA 标准化或减少方法。用通常用于在 cDNA 文库产生期间标准化 cDNA 文库的方法可完成标准化。从样品收集的 DNA 一分为二，其中一份的质量比另一份大 10 倍。通过末端转移酶和 ddCTP 生物素化较大量的样品，并将其作为单链 DNA 连接到链亲和素柱或链亲和素包被的珠。另外，生物素化的随机引物可以用来产生连接到链亲和素的序列。也可应用整个基因组扩增方法(Molecular Staging 公司，New Haven，CT)。然后将得标准化的样品杂交到连接的分子，由于结合位点的数目较大，所以从溶液中优选地去除样品中有过多代表的那些分子。可以对相同样品应用几个杂交/去除循环，以获得完全标准化。另一实施方式提供了长双链 DNA 片段的有效杂交，而无需 DNA 变性，该杂交用定时的 lambda 核酸外切酶消化产生单链 DNA 的短末端区。

另一实施方式提供对难于用 DNA 标准化和 rSBH 的组合来分析低丰度成员的测序。一个样品对于另一个样品的标准化允许随条件变化监测共有结构中的变化并鉴定新成员作为条件变化。

C.二级 DNA 片段化和接头连接

本发明提供将通过剪切力产生的长 DNA 片段悬浮在溶液中，该溶液在位于载玻片的室内。调节 DNA 的浓度，使每个片段占据的体积为 50x50x50 微米。反应室包含限制性内切酶、T4 DNA 连接酶、链取代聚合酶和特别设计的接头的混合物。用限制性内切酶部分消化 DNA 产生平均长度为 250bp 的片段，它们具有一致的突出端序列。T4 DNA 连接酶将非磷酸化的双链接头与基因组片段末端通过互补粘性末端连接，产生稳定结构的基因组插入物，两端各有一个接头，但在一条链上有一个缺口，连接酶在此处不能催化形成磷酸二酯键(图 1)。T4 DNA 连接酶在大多数限制性内切酶缓冲液中是有活性的，但需要加入 ATP 和相对于基因组 DNA 一摩尔过量的接头，以促进在基因组分子的每端都连接接头。用非磷酸化的接头对于防止接头-接头连接是重要的。此外，接头包含两个引物结合位点，接头通过发夹端的交联碱基维持在一个发夹结构中，这能防止接头在高温解链期间解离。用链取代聚合酶，如 Vent 或 Bst 从 3'末端延伸导致两端具有接头序列的 DNA 链的产生。然而，一端的接头将被维持在发夹结构中，用于防止互补序列在 DNA 片段的另一端上结合。

本发明提供随机 DNA 阵列，以在一个测定中测序多个高度相似的样品(即来自病人的个体 DNA)，该测序在随机阵列形成之前，对每个样品的 DNA 片段进行标记。用于将引物序列掺入到 DNA 片段末端的一个或两个接头可以具有标记盒。不同的标记盒可用于各个样品。附着接头(优选通过连接)后，将所有样品的 DNA 混合，形成单一的随机阵列。完成片段的测序后，通过指定的标记序列识别属于各个样品的片段。用该标记法允许对来自约 10-1000 个样品的少量标记 DNA 区，在具有高达约一千万 DNA 片段的高容量随机阵列上进行有效测序。

D. DBA 附着和原位扩增

然后将接头连接的基因组 DNA 通过杂交到寡核苷酸与来自原始 5-100kb 片段的其他片段一起定位在载玻片上，该寡核苷酸与接头序列互补(引物 B)。接头连接和 DNA 延伸后，加热溶液，变性分子，该分子与附着在载片表面的高浓度引物寡核苷酸接触时，在重退火阶段杂交到这些互补序列。在另一实施方式中，不发生原位扩增，接头附着到支持物，DNA 片段被连接。由一个亲代分子产生的大部分 DNA 结构都定位到载片的一面，占 50x50 微米；因此如果限制酶消化一个亲代分子产生 1000 个分子，每个片段将平均占据 1-4 微米^{2}区域。1-4 微米^{2}区域可用 CCD 相机的单个像素观察，且代表一百万孔的阵列内的一个虚拟反应池。

在 50-100 微米厚的防止液体紊流的毛细管室中，短时间内不可能显著出现 DNA

片段穿过载片表面侧向扩散超过 50 微米。此外，高粘度缓冲液或凝胶可用于使扩散最小化。在另一实施例中，需要有限的紊流，在 50x50 微米表面扩散几百个取自单个 5-100kb 分子的短 DNA 片段。请注意，扩散并不必须是完美的，因为 SBH 可以在相同像素位点分析几个 DNA 片段的混合物。可以制备原始样品的几个具有更一致片段长度的部分(即 5-10kb、10-20kb、20-40kb、40-1000kb)，达到短片段之间相等间隔。而且，可将电场施加于库，以使短 DNA 片段附着于表面。具有局部混合短片段的部分结构的阵列几乎与完全结构的阵列同样有效，因为没有来自任何单个、长片段的短片段与约 10,000 个其他长起始片段产生的短片段混合。

本发明另一个实施方式提供将两个引物序列附着到 DNA 片段的连接方法。该方法基于靶向由双链 DNA 片段变性产生的单链 DNA。因为单链 DNA 具有独特的 5'和 3'末端，各端可连接特定的引物序列。设计了两个特定的接头各自包含两个寡核苷酸(图 2)，它们具有特定修饰的末端，其中 F 和 B 代表未结合的、无溶液的引物(F)和表面结合引物(B)序列，f 和 b 代表与这些引物序列互补的序列(即引物 f 与引物 F 互补)。只有连接到 DNA 片段必需的 3'-OH 基团在引物 F 上，其他寡核苷酸可以有双脱氧 3'末端(dd)，以防止接头-接头连接。除存在于引物 b 的 5'-磷酸基团(P)外，引物 B 也可具有用于接头连接后该引物降解的 5'-P 基团，以将用于杂交的引物 b 序列暴露到表面附着的引物/捕获探针 B。为允许接头连接到由源 DNA 通过随机片段化产生的任何 DNA 片段，寡核苷酸 f 和 B 具有几个(约 3-9 个，优选 5-7 个)简并碱基(Ns)。

虽然 rSBH 检测设计用于单分子检测，但一些实施方式原位扩增了各 DNA 靶。本发明方法提供在一个微米大小的定位扩增子内的等温指数扩增，这里称为“ampliot”(定义为扩增子位点)(图 3)。通过用结合至表面的引物(引物 B)和一个溶液中的游离引物(引物 F)完成扩增。引物 B 首先与原始靶序列杂交，然后延伸复制靶序列。将非附着链解链，并洗涤掉。加入新试剂成分，包括具有链取代特性的 DNA 聚合酶(如 Bst DNA 聚合酶)，dNTSs 和引物 F。然后用连续扩增反应合成新链并取代之前合成的补体。

连续指数扩增反应产生了一条取代链，它包含与捕获阵列寡核苷酸互补的序列，因此，反过来被捕获并用作进一步扩增的模板。该链取代方法需要引物能够连续引发聚合。本领域中有几种描述的策略，如 ICAN™ 技术 (Takara BioEurope,Gennevilliers，法国)和 SPIA 技术(NuGEN，San Carlos，Ca；美国专利 6,251,639，整体纳入本文作为参考)。一旦延伸开始，允许另一引物杂交并引发聚合和链取代，就用 RNA 酶 H 在 RNA/DNA 双链体中降解 RNA 的特性去除引物。在一个优选实施方式中，将引物 F 位点设计在接头中的 A/T 富集区，以使双链 DNA 能够经

常变性并允许 F 引物在所选 DNA 聚合酶的最优化温度下结合。ampliot 中约 100 至 1000 拷贝是通过连续指数扩增产生的，而无需热循环。

本发明的另一实施方式将 T7 启动子掺入接头，并合成 RNA 作为中间体(图 4)。载片表面上首先用切口平移或链取代聚合酶产生双链 DNA。将新形成链用作 T7 聚合酶的模板，通过从引物 B 延伸也形成必需的双链启动子。从启动子的转录产生可以杂交到附近的表面结合引物的 RNA 链，反过来可以用反转录酶反转录。该线性扩增方法可以产生 100-1000 靶拷贝。然后，产生的 cDNA 可以通过用 RNA 酶 H 将 RNA/DNA 双链体中的 RNA 链降解或通过碱和热处理转化成单链 DNA。为了引物 B 序列在 RNA 分子中的分子内杂交最小化，引物 B 的一半序列可来自 T7 启动子序列，因此将产生的互补序列量减少到约 10 个碱基。

两种扩增方法都是等温的，以确保合成链仅在 ampliot 内有限扩散。ampliot 大小约 2 微米，但它可以大至 10 微米，因为扩增的 DNA 信号可以补偿每 CCD 像素总表面背景增加 25 倍。而且，引物 B 附着位点是以约 10 纳米间隔(10,000/微米 ^{2})分开，这提供取代 DNA 的立即捕获。封闭的毛细管反应室几乎清除了缓冲液紊流。

本发明的另一实施方式提供了用链取代酶进行等温扩增的方法，链取代酶基于侵入物寡核苷酸形成用于引物退火的单链 DNA(参见图 5)。可以用两种引物在恒温下扩增双链 DNA，一个侵入物寡核苷酸或其他试剂，和链取代聚合酶，如 Klenow 片段聚合酶。侵入物寡核苷酸的浓度等于或高于相应的引物浓度。起始时，靶 DNA 约比引物浓度低 100 至 1 亿倍。用侵入物寡核苷酸进行的等温扩增法包括下述步骤：

1) 通过侵入方法将侵入物核苷酸(可由 LNA 或 PNA 或提供与 DNA 较强结合的其它修饰物部分制备)结合到靶 DNA 的 5'末端序列之一。侵入核苷酸可以是单链或双链突出端(Ds)。可以通过接头加到相应靶 DNA 末端的$(TA)_x$或相似序列的低双链体稳定性可以帮助入侵。

2) 将引物 1 杂交到可用的单链 DNA 位点并引发引物延伸和用聚合酶取代一条 DNA 链。侵入物核苷酸与引物 1 部分互补。为避免完全阻断引物，将在所用温度和浓度下互补部分的大小和结合效率设计为提供约 9:1 的结合/不结合平衡比。约 10%的游离引物 1 是超过靶 DNA 的量。

3) 将引物 2 杂交到单链 DNA 的另一端，用聚合酶产生一个新的双链 DNA。

4) 由于起始和新的 dsDNA 分子连续引发步骤 1-3，重复步骤 1-3。

E. 探针和库设计

可以用一个或多个可检测颜色；然而多个颜色会降低连接循环的数量并提高系

统效率。本领域的现状提示可以同时使用四个颜色。本发明的优选实施方式使用基于 FRET 的系统，时间分辨系统、时间分辨 FRET 信号系统(Didenko，2001，上述)。也考虑使用常规化学方法，如量子点增强的三重 FRET 系统，以及树状聚体技术。

优选实施方式中使用了基于 FRET 检测的两组通用探针。用前面在共有美国专利申请 09/479,608 和 10/608,298(整体纳入本文作为参考)中描述的探针设计，用 1024 或更少的单个合成产生所有 4096 个可能的六聚物。在用于实验之前对探针进行筛选(matriculation)和 QC(质量控制)处理方案(Callida Genomics 公司，Sunnyvale，CA)。设计探针使效率差异最小化，各个探针与全匹配或错配靶的实际表现是由 QC 测定确定，并通过高级碱基调用系统(Callida Genomics 公司)使用。

6.6 核心技术

本发明方法依赖三个核心技术：1) 通用探针，它允许通过杂交来自任何生物体的 DNA 和检测任何可能的序列改变来完成测序。这些探针是用统计原理设计的，而并不参考已知的基因序列(参见共有、共待审美国专利申请 10/608,293，整体纳入本文作为参考)；2) 组合连接，其中通过 DNA 连接酶的“酶校对”提供两个通用小组短探针组合产生成千上万个具有优良特异性的长探针序列(参见美国专利申请 10/608,293)；3) 信息探针库(IPPs)，几百个标记相同而序列不同的探针的混合物简化了杂交过程，而不会对序列确定产生负面影响(参见美国专利申请 09/479,608，整体纳入本文作为参考)。

本发明方法使用几百万单分子 DNA 片段随机排列在光学透明的表面上，作为杂交/连接来自 IPPs 的荧光标记探针对的模板。用一个具有先进光学的灵敏的兆像素 CCD 相机，同时检测整个阵列上的几百万单个杂交/连接事件(图 6)。将 DNA 片段(长度为 25 至 1500bp)排列成密度约每 CCD 像素 1 个分子(每平方微米底物 1 至 10 个分子)的阵列。各 CCD 像素限定一个约 0.3 至 1 微米的虚拟反应池，包含一个(或几个)DNA 片段和几百个标记的探针分子。SBH 分析样品混合物和组装所包含各片段的能力对于随机阵列大有好处。DNA 密度可以调节到 1-3 个片段，在超过 90%的所有像素中可以有效地分析。各反应的体积约 1-10 毫微微升。一个 3x3 毫米的阵列具有容纳 1 亿个片段或约 1 千亿个 DNA 碱基(相当于 30 个人类基因组)的能力。

6.7 组合 SBH

如上所述，标准 SBH 比竞争性基于凝胶的测序技术具有显著优点，包括样品阅读长度的改进。然而，标准 SBH 方法最终受限于需要使用指数较大的探针组以测序越来越长 DNA 靶。

组合 SBH 克服了标准 XBH 技术的很多限制。在组合 SBH(Drmanac 的美国专利 6,401,267，整体纳入本文作为参考)中，在 DNA 连接酶存在下将两个完整、通用短探针组与靶 DNA 接触。一般地，一组探针与固体支持物连接，如载玻片，而此外一组标记了荧光团的探针在溶液中游离(图 6 和 7)。当附着和标记的探针在精确相邻的位置与靶杂交时，它们被连接，产生与表面共价连接的长标记探针。洗涤去除靶和未附着的探针后，用标准阵列阅读仪对各个阵列位置上的荧光信号进行记分。在给定位置，阳性信号说明在靶内存在一个与两种探针互补的序列，它们结合产生信号。组合 SBH 比标准 SBH 方法具有巨大的阅读长度、成本和材料优势。例如，在标准 SBH 中，对长度为 10-100kb 的靶 DNA 准确测序(为了发现突变)需要超过一百万 10-mer 探针的全组探针。相反，用组合 SBH，通过将两小组 1024 个 5-mer 探针结合产生相同的 10-mer 探针组。通过大大降低实验复杂性、成本和材料要求，组合 SBH 在 DNA 阅读长度和测序效率上取得了引人注目的改进。

6.8 信息探针库

通过使用信息探针库(IPPs)，组合 SBH 的效率进一步增强。IPPs 是统计选择的探针组，它们在杂交过程中汇集以使必须测试的组合数目最小化。设计包含 4 至 64 个不同库的一组 IPPs 以明确确定任何给定的靶序列。每个库组包含一个通用探针组。库的大小一般为 16 至 256 个探针。当这些探针中的一个或多个产生阳性信号时，库中所有探针接受阳性记分。用来自任何独立 IPP 配对的记分产生一个对各碱基位置的组合概率分数。准确的序列数据实际上是确实的，因为各碱基位置的记分是由分别在不同库中的十个或更多重叠探针组合产生的。对于一个探针的假阳性记分容易通过很多其他来自不同库的正确分数校正。此外，对互补 DNA 链的独立测序使库相关的假阳性探针的影响减至最小，因为在不同库中，对于各互补链的真实阳性探针偶然趋于降低。较长探针的 IPPs 实际上比单独记分的较短探针提供更多信息并提供更准确的数据。例如，对于 2kb 的 DNA 片段，16,000 个 64 个 10mer 的库比 16,000 个单独的 7-mer 提供的假阳性少 100 倍。

IPPs 组将用于从阵列的 DNA 靶中获得序列信息。IPPs 是仔细选择的给定长度的寡核苷酸库，一般各库包含 16 至 128 个单个探针。所有这种长度的可能的寡核苷酸都在各组 IPPs 中至少出现一次。用供体荧光团标记一组 IPPs，用受体荧光团标记此外一组。当来自供体和受体组的探针间发生连接时，这些一起作用产生 FRET 信号。该连接事件仅在两个探针同时杂交到靶上相邻的互补位点时发生，因此鉴定了其中 8-10 碱基长的互补序列。每像素可分析的 DNA 片段长度是探针长度、库大小和测试

探针库对的数量的函数，一般范围在 20 至 1500bp。通过增加库和/或探针的数量，可以测序几千个碱基的靶 DNA。可以用小亚组的 IPPs 或者甚至单个探针对完成 1-10kbDNA 片段的部分测序和/或特征分析。如果使用多个荧光标记，可以在连续杂交循环中或同时测试 IPP 对。CCD 相机相对于阵列的固定位置保证了对单个靶分子连续杂交的准确追踪。

设计的 IPPs 是促进强 FRET 信号和序列特异性连接的。一般的探针设计包括，第一组 IPPs 为 5'-F_x-N_{1-4}-B_{4-5}-OH-3'和第二组为 5'-P-B_{4-5}-N_{1-4}-F_y-3'，其中 F_x和 F_y是供体和受体荧光团，B_n是特异性(信息)碱基，N_n是简并(随机混合)碱基。简并碱基的存在增加了有效探针长度，而没有增加实验复杂性。各探针组需要合成 256 至 1024 个探针，然后将它们混合建立每库 16 个或更多探针的库，每组总共 8 至 64 IPPs。单个探针可按需存在于一个或多个库中，以使实验灵敏度、柔性和冗余度最大化。将来自供体组的库与阵列杂交顺序地与来自受体组的库在 DNA 连接酶的存在下杂交。一旦来自供体组的库与来自受体组的库配对，则记录 8-10 个碱基信息序列的所有可能组合，由此鉴定各像素上靶分子内的互补序列。该技术的能力在于用两小组合成的寡核苷酸探针组合建立并记录可能是几百万的较长序列串。

该方法的精确的生化反应依赖于序列特异性杂交和两个短寡核苷酸的酶连接，使用单个 DNA 靶分子作为模板。虽然随时仅探查单一靶分子，但是各靶可用到几百个相同序列的探针分子，以进行快速连续探查，提供统计显著性测定。连接过程中的酶效率与最优化反应条件结合提供了相同单靶分子的快速多重探查。在相当高的探针浓度和高反应温度下，单个探针迅速杂交(2 秒以内)，但是解离更快(约 0.5 秒)，除非它们是连接的。此外，在最优化温度下连接的探针保留杂交到靶的时间约为 4 秒，连续产生可被 CCD 相机检测的 FRET 信号。通过以每秒 1-10 像帧监测各像素 60 秒，在配对靶序列上平均出现 10 个连续的连接事件，在该位置，60 秒中产生约 40 秒的光信号。在错配靶的情况下，连接效率低约 30 倍，因此很少产生连接事件，在 60 秒反应时间中产生小信号或不产生信号。

主要的监测挑战是背景信号的最小化，这可由所需过量的标记探针分子产生。除了使 CCD 像素聚焦于最小的可能底物面积外，我们对此问题的基本解决方式依赖于表面接近和 FRET 技术的协同组合(图 7)。仅在一对探针排列在相同靶分子上极接近照明面时(例如由总内反射产生的 100 纳米宽的渐消失区)，一个探针上报道标记的长时间激发才会发生。因此，背景信号并不会从溶液中过量的未杂交探针产生，因为或者供体离照明面太远，或者受体离供体太远不能引起能量转移。此外，可以

用多种染料分子(由分支树状聚体连接)标记探针分子，比普通系统背景增加探针信号。

在测试了所有 IPPs 后，用 SBH 算法和软件(共有、共待审的美国专利申请09/874,772;Drmanac 等,*科学* 260:1649-1652 (1993);Drmanac 等,*Electrophoresis* 13:566-573(1992)；Drmanac 等，*J. Biomol. Struct. Dyn.* 8:1085-1102(1991)；Drmanac 等，*Genomics* 4:114-128 (1989)；Drmanac 等的美国专利 5,202,231 和5,525,464，以整体纳入本文作为参考)进行单个分子的序列组装。这些先进的统计学方法定义了匹配具有最高可能性的连接数据的序列。将用 CCD 相机测定的亮度作为给定探针对的全匹配序列存在于该像素/靶位点的概率处理。因为来自不同库的几个阳性重叠探针以正确的序列独立“阅读”各碱基(图 8)，这些探针的结合概率提供了准确的碱基测定，即使几个探针失败。此外，不正确序列相对应的多个独立探针不能与靶杂交，则认定该序列的结合概率低。即使在几个不正确序列相对应的探针显示阳性的情况下也会发生，因为它们恰好存在于具有真阳性探针的 IPP 中，该真阳性探针匹配真实序列。

6.9 rSBH 方法

本发明 rSBH 方法的核心涉及建立并分析含有几百万基因组 DNA 片段的高密度随机阵列。该随机阵列去除了在底物表面上排列探针的昂贵、耗时的步骤，也去除了对个别制备几千个测序模板的需要。代之以，它们提供了快速和有成本效益的方式来分析单个测定中包含 10Mb 至 10Gb 的复杂 DNA 混合物。

本发明的 rSBH 方法结合了下述优点：1) 在溶液中两个 IPPs 的组合探针连接产生序列特异性的 FRET 信号；2) 用来分析一个测定中 DNA 混合物的组合方法的准确性、长阅读长度和能力；3) TIRM，一种高灵敏、低背景的荧光检测方法；4)市售的具有单光子灵敏度的兆像素 CCD 相机。本发明方法提供在单靶分子上检测连接事件的能力，因为仅当两个连接探针杂交到附着的靶，使供体和受体荧光团互相在 6-8 纳米内，并且在阵列表面上产生 500 纳米宽的渐消失区内时，才会产生长时间信号。

本发明方法一般用几千至几百万单分子 DNA 片段，随机排列在光学透明的表面作为杂交/连接来自 IPPs 的荧光标记探针对的模板(图 6)。用供体和受体荧光团标记的探针库对与 DNA 连接酶混合，并提供给随机阵列。当探针杂交到靶片段上的相邻位点时，它们被连接在一起，产生 FRET 信号。用具有先进光学的灵敏的兆像素 CCD 相机同时检测整个阵列上的几百万单个杂交/连接事件。各配对序列可能产生几个独立的杂交/连接事件，因为连接探针对最终从靶中扩散开来并被新杂交的供体和受体

探针取代。靠近互相杂交的未连接对可瞬间产生 FRET 信号，但并不保持结合到靶足够长时间以产生显著的信号。

一旦检测到来自第一个库的信号，就去除探针并用连续连接循环测试不同探针组合。CCD 相机相对于阵列的固定位置保证了对连续测试的 256 个 IPP 对(16x16 IPPs)的准确追踪，且花费 2-8 个小时。根据由几百个独立杂交/连接事件产生的荧光信号编制每个 DNA 片段的整个序列。

将 DNA 片段(长度为 15-1500bp)以密度约为每平方微米底物 1 个分子排列。各 CCD 像素限定约 1x1 至 3x3 微米的虚拟反应池，包含一个(或几个)DNA 片段和几百个标记的探针分子。本发明方法有效利用 SBH 的能力来分析样品混合物并组装混合物中各片段的序列。各反应的体积约 1-10 毫微微升。一个 3x3 毫米的阵列具有容纳 100-1000 万片段或约 10-100 亿个 DNA 碱基的能力，上限是 3 个人类基因组的等效物。

每像素可分析的 DNA 片段长度是探针长度、库大小和测试探针库对的数量的函数，一般范围在 50 至 1500bp。通过增加库和/或探针的数量，可以测序几千碱基的靶 DNA。可以用小亚组 IPPs 或者甚至单个探针对完成 1-10kb DNA 片段的部分测序和/或特征分析。

本发明的 rSBH 方法保留了组合 SBH 的所有优点，包括连接方法的高度特异性。同时，它增加了几个重要益处，这些益处来自 DNA 片段而不是探针的附着。DNA 附着产生使用容量比常规探针阵列更大的随机 DNA 阵列的可能性，并通过溶液中两个标记探针的连接进行 FRET 检测。此外，溶液中具有两种探针组件允许将 IPP 策略延伸至两个探针组，而在常规组合 SBH 中是不可能的。

6.10 方法步骤

rSBH 全样品分析包括下述处理步骤，可以归并到单微流体芯片中(图 9)：

1)简单样品处理或 DNA 分离(如果需要)，包括一种在病原体混合物柱上收集病原体 DNA 的有效方法；

2)随机 DNA 片段化，产生合适长度的靶

3)DNA 的直接末端附着到活性底物表面，例如通过连接到通用锚；

4)阵列洗涤，去除所有未结合 DNA 和样品中存在的其他分子；

5)以合适探针浓度从两个 IPP 组中引入第一个 IPP 对，以及 T4 连接酶和一些其他(即热稳定性)DNA 连接酶；

6)照明的同时孵育少于 1 分钟，以每秒 1-10 帧进行信号监测。

7)洗涤去除第一个 IPP 对，然后引入第二个 IPP 对；和

8) 在测试了所有 IPP 对后，计算机程序将产生各片段的特征或序列，然后将它们与特征或序列的综合数据库进行比较，并报告样品中存在的 DNA 的性质。

6.11 装置大小和特征

本发明方法所用的装置是基于共有、共待审的美国专利申请 10/738,108 中的描述，整体纳入本文作为参考。本发明的装置包括三个主要部件：1)操作(混合、引入、去除)IPPs 的操作子系统，考虑到此模块可以扩展引入“芯片上”的样品制备，2) 反应室-能装载任何底物的温控流通室，和 3) 照明/检测子系统(图 10)。这些子系统一起工作，提供单荧光团检测灵敏度。

本发明的装置操作具有一个槽和口的插入式反应室，槽用于放置阵列底物，口用于连接探针模块，如果 DNA 附着和/或原位扩增在该室内完成，那么口也用于连接一个可能的阵列制备模块。

盒包括多达 64 个单独的储器，用于多达 32 个 FRET 供体和多达 32 个 FRET 受体库(图 11)。盒包括一个与各库的储器连接的混合室，该连接通过单一的微流体通道和一个整体真空/压力启动的微型阀实现。

6.11.1 反应室

底物一旦附着到反应室，则形成杂交室的底面。该室控制杂交温度、提供向室中加入探针库的口、从渐消失区去除探针库、探针库在整个室中再分布以及洗涤底物。将标记的探针库溶液引入反应室，在给定时间与靶 DNA 杂交(数秒)。通过在杂交溶液中建立电压势，从渐消失区中拔出没有参与杂交事件的探针。用能够检测单光子的高灵敏度 CCD 相机，通过反应室顶部的窗口监测底物，检测 FRET 杂交/连接事件(Ha，*Methods* 25:78-86(2001)，整体纳入本文作为参考)。以约 30 秒的规则间隔拍摄底物的图像。然后冲洗室，去除所有探针，引入下一个探针库。此过程重复 256-512 次，直到测定了所有探针库。

6.11.2 照明子系统

照明子系统基于 TIRM 背景降低模型。TIRM 在两个光学特性不同的材料的界面建立了 100-500 纳米厚的渐消失区(Tokunga 等，*Biochem. Biophys. Res. Commun.* 235:47-53(1997)，整体纳入本文作为参考)。本发明装置使用的照明法消除了测定中光束的高斯分布的任何效应。本发明装置的这个子系统中的激光器和所有其他组件都安装到光学台上。通过移动安装在电流计 1 和 2 上的镜子建立一条 1 厘米的扫描线(图 10)。然后用电流计 3 将扫描线通过棱镜 1 导入底物。调整电流计 3，使扫描线在其临界角交叉玻璃/水界面。光束进行全内反射，在底物上建立渐消失区。该

渐消失区是超过玻璃/水接界面几百纳米(通常是 100-500 纳米)的光束能量的延伸。

6.11.3 检测子系统

本发明装置使用能进行光子计数的高灵敏度 CCD 相机(如 Andor Technology (Hartford, CT)的 DV887, 512x512 像素),将它悬挂在杂交室上方。该相机通过反应室的窗口监测底物。相机的透镜提供足够的放大倍数,使各像素接受来自 3 平方微米底物的光线。在另一实施方式中,相机可以是水冷的,以供低噪音应用。

高灵敏度的电子倍增 CCD(EMCCD)检测器使高速检测单个荧光团成为可能。假设 532 纳米处的 1 瓦特激发激光(用于 Cy3/Cy5 FRET),可以计算每秒由激光器发出的光子数目,也可以估计每秒到达检测器的光子数目。用等式 $e=hc/\lambda$,其中 λ 代表波长,一个 532nm 波长的光子具有能量 3.73e-19 焦耳。假设激光器输出功率是 1 瓦特,或 1 焦耳/秒,预计每秒从激光器发出 2.68e18 个光子。将此数量的能量扩展穿过 $1cm^2$ 的底物面积,预计每平方纳米将接受约 1e-15 焦耳的能量,或约 26,800 个光子。假设荧光团的量子产量为 0.5,预计每秒约输出 13,400 个光子。用高精度透镜,用 CCD 应该能够捕获收集到总输出的约 25%或总共 3350 个光子。Andor 的 DV887 CCD 在 670-700 纳米处具有的量子效率为约 0.45, Cy5 在此发射。每秒产生约 1500 个光子被各像素记录下来。在每秒 10 帧,每帧记录 150 个数。在-75℃时,相机的无照电流约是 0.001 电子/像素/秒,一秒内每 1000 像素平均有 1 个假阳性计数。即使假设每秒每像素 1 个假阳性计数,每帧每像素 0.1 个,也获得了 1500:1 的信噪比。结合 TIRM 照明技术,检测器的背景实际上为零。

6.11.4 装置的小型化

在另一实施方式中,本发明方法可以在一个微型装置中进行。一个简单的、仅需要几个现成组件的物理装置就可以进行整个过程。照明和检测组件形成该系统的核心。该核心系统仅由一个 CCD 相机、一个激光器或其他光源、0 至 3 个扫描电流计、用于底物的石英或等效支持物和一个反应室组成。将所有这些组件置于 1 立方英尺的装置中是可能的。微型流体操作机器人或微流体芯片实验室装置(图 9)将通过访问 IPPs 对进行该测定,该 IPPs 对来自两个 8 至 64 个 IPPs 的库,这种装置可占据约 0.5 英尺^3。高密度多孔板或具有 64 储器的芯片实验室将可以供该库的超小型存储使用。单板计算机或便携式电脑可以运行该装置并进行分析。该系统易于运输,并可装入几乎任何车辆中,以用于野外环境调查或者对急救队或生物危害工作者作出反应。

该系统的组件包括: 1) 微型个人电脑(1 英尺 x 1 英尺 x 6 英寸), 2) 机器人

或芯片实验室流体操作系统(1 英尺 x 1 英尺 x 2 英寸)，3) 激光器(6 英寸立方体)，4) 具有散热器的扫描电流计(3 英寸立方体)，5) 载片/杂交室组件(3 英寸 x 1 英寸 x 2 英寸)，6) CCD 相机(4 英寸 x 4 英寸 x 7 英寸)和 7)流体储器(约 10-1000 毫升容量)。

本发明装置的另一实施方式集成了一个基于模块微流体的底物，在它上面进行病原体检测的所有测定(图 11)。该消耗型底物是以集成的“反应盒”的形式。反应盒的底物组分必须接受三种不同种类的一次性集成模块，包括：探针库模块、样品集成模块和反应底物模块。所有机器功能都作用于此盒产生测定结果。该底物需要集成的流体，如提供反应盒和相关模块的快速连接。

将微流体引入底物以便在底物的检测表面上处理信息探针库。使用模块方法，其中开发的起始探针操作模块不依赖底物，可以用“即插即用”法将最终设计加到标准底物反应盒中。该反应盒包含多达 64 个单独的储器，用于多达 32 个 FRET 供体库和多达 32 个 FRET 受体库(参见图 11)。可以将较大数量的 IPPs 储存到一个或一组反应盒上，例如 2x64、或 2x128、或 2x256、或 2x512 或 2x1024 个 IPPs。该反应盒具有与主通道连接的混合室，该连接通过其自己的微流体通道和一个整体真空/压力启动的微型阀实现。当阀打开时，应用真空将库移动入混合室。然后关闭阀，重复该过程，加入第二个库。混合室与洗涤泵成一行，洗涤泵用于搅动库并将它们推入反应室。

6.12 软件部件和算法

行数据代表每像素中每对库(即 IPPs)在不同时间/温度点的约 3-30 个强度值。通过统计处理 10-100 CCD 测量(优选每秒 5-10)获得各值。各个片段有 512 组 3-30 个强度值。有一百万片段的阵列包含约一百亿强度值。可以对几百个像素的组进行信号标准化。如果组不符合预计特点，那么将丢弃给定 IPP 对的所有数据点。将丢弃无有用数据(即没有足够的阳性或阴性数据点)的各个像素(其中大部分具有合适的 DNA)。确定强度值在其他像素中的分布并用来调整碱基调用参数。

所有单个短片段可用记录特征作图为一个相应的参比序列，也可用比较测序方法分析或用从头 SBH 函数进行序列组装。从引物序列开始，由约一百万可能的 10-mers 组装各个约为 250 碱基的片段。组装过程通过评价组合 10-mer 记录进行，从几百万局部候选序列变体的重叠 10-mers 中计算。

一组来自一个阵列区域的片段代表了一个长的连续基因组片段，该片段与几组来自其他阵列区域的片段组具有显著的重叠序列。这些组也可以通过短片段序列与

参比序列的比对来识别，或作为一个包含被空像素环绕的像素的 DNA 岛。将短片段定向到组，尤其在部分结构的阵列中是一个引人感兴趣的算法问题。

一组内的短片段来自一个片段化的单 DNA 分子，且不重叠。但是短序列在相应组间会重叠，代表常、重叠 DNA 片段，长片段的组装是通过与鸟枪测序方法中粘粒或 BAC 克隆的序列组装相同的方法进行的。因为 rSBH 方法中的长基因组片段在 5-100kb 范围内变化，代表 5-50 基因组等效物，在所有相关水平提供作图信息，以指导准确的毗连群组装。该方法可容许将小片段分配到长片段中的遗漏和误差以及单个组中约 30-50%随机缺少片段。

本发明的 rSBH 方法提供了稀有生物体的检测，或细胞数量或各微生物基因表达的定量。当优势种有 1x 基因组覆盖时，约 10 个基因组片段代表该种以 0.1%水平存在。DNA 标准化可以进一步将检测灵敏度提高到超过 10,000 个细胞中的一个细胞。DNA 定量是通过计数代表一个基因或一个生物体的 DNA 片段存在的数量完成的。不含克隆步骤意味着，rSBH 应该比常规测序提供各 DNA 序列类型发生率的更定量的估计。为了量化研究，直接将样品片段化到 250 bp 片段，形成标准(非结构的)随机阵列是足够的。可用部分标准化使差异最小化，但仍然存在差异；可用标准化曲线计算原始频率。一个一百万个片段的阵列足够定量几百个基因种和它们的基因表达。

6.12.1 rSBH 软件

本发明提供了支持 rSBH 全基因组(复杂 DNA 样品)测序的软件。该软件可以按比例增加到分析整个人类基因组(～3Gb)或高达～10Gb 的基因组混合物。考虑了在几个 CPU 上进行并行计算。

rSBH 装置可以高达 10/秒或更快的速度生成一组 tiff 图像。各图像代表靶与合并的标记引物对的杂交。各杂交可以产生多幅图像，以确定信号平均值。将靶片段化为长度约 100-500 个碱基的多个片段。这些片段附着在随机分布中的玻璃底物表面。杂交和洗涤未杂交探针后，用 CCD 相机拍摄表面的图像。最后，图像的各像素可包含一个片段，虽然有些像素可以是空的，而另外一些可含有两个或更多片段。该装置可能成像 100-1000 万，或者甚至更多片段。

总的装置运行时间是由杂交/洗涤/成像循环(～1 分钟)乘以使用的库组的数目决定的。用 1024 库组(产生 1024 图像)，该运行将持续约 17 小时；两种颜色可减少该过程的一半时间。图像分析软件以近实时处理图像，并将数据发送到碱基调用分析软件。

A. 并行处理

rSBH 分析理想地适于并行处理。因为各“点”杂交到不同片段，碱基调用分析可以在各点并行进行，而无需分析之间的通讯。在整个分析中仅有的通讯是在控制模件(GUI)和分析程序之间进行的。为避免竞争情况，需采取非常次要的步骤。在实践中，CPU 的数目限制了并行处理的数目。对于一百万个片段来说，一台有 100 个处理器的计算机将把该工作分成 100 个并行的碱基调用程序，每一个连续分析 10,000 或更多的片段。

一个 200 个片段组可以在一个处理器上运行，然而它也可以在几个 CPU 上运行。如果没有突变或突变试验(升级功能)，最优化的碱基调用程序可以在～100 毫秒内完成。该时间包括数据加载和标准化。对于最长的参比序列(见下文)来说，可以加～100 毫秒的参比查找时间。参比序列查找时间与其长度成正比，对短长度可以忽略。分析多突变可以将运行时间延长到约 1 分钟/多突变位点。如果平均分析时间是每片段 1 秒，一百万个片段用 100 个 CPU 就可以在 10,000 秒内分析。类似地，200 个片段用一个 CPU 可以在 200 秒内分析，或者用 10 个 CPU 在 20 秒内分析。使程序速度最优化需要每个 CPU 有大量的 RAM。如下所述，如果各 CPU 具有～2GB 至 8GB，取决于 CPU 的数目和片段的数目，该软件就不会受存储器限制。当前，为每个系统购买 32GB 以上的 RAM 是可能的。

B.数据流

在一个 CPU 上运行 GUI 和图像分析程序，而在几个(N)CPU 上运行碱基调用分析程序。在启动时，图像分析程序装有数目 N，监测 CCD 相机写入 tiff 图像的目录。对各 tiff 文件来说，它获得各片段的记录并将其分组到 N 个文件，一个对一个分析 CPU。例如，如果有 200 个片段和 10 个 CPU，图像分析程序把第一个 20 个片段记录写入一个文件，用第一个碱基调用分析 CPU 处理，第二个 20 个片段记录写入第二个文件，用第二个碱基调用分析 CPU 处理，依此类推。也考虑到用其他的通讯方式，例如插座或 MPI。因此，可以把文件的输入/输出定位于一个模块，这样它后来容易被换出。

随着时间推移，针对连续增长的 tiff 文件数量会产生大量的图像分析文件。本发明为各碱基调用分析 CPU 提供了分离的图像分析目录。碱基调用分析 CPU 各自监测它们各自的图像分析目录，一旦数据可用则进行加载。存储所有图像数据必需的 RAM/CPU 数量是[2 字节 x 片段数量 x 图像数量÷N]。对于 1 百万片段，1024 幅图像和 1 个 CPU 来说，这是～2GB/CPU；或者对于 10 个 CPU，这是 200KB/CPU。

其他大量(在 RAM 方面)输入到碱基调用分析程序的数据是参比序列(长度 L)。

为了优化速度，把参比转换成 10-mer(和 11-mer，12-mer)位置的矢量，提供快速查找各片段的最高记录探针(见下文)。这是在每个碱基调用分析 CPU 中存储参比序列位置数据的最快方法。存储参比序列位置数据需要的存储量是 2 字节 x L 或 2 字节 $x4^{12}$ 中较大的。最大的 RAM 是 2 字节 x10GB＝20GB。也必须存储实际参比本身，但这可以 1 字节/碱基储存，甚至压缩到 0.22 字节/碱基。

各片段的分析产生调用序列结果。这些结果被串接到写入图像分析目录的文件，图像分析目录与各 CPU 相关。当碱基调用完成时，GUI 处理调用的序列文件。它从不同 CPU 中加载所有文件并重新排列片段的位置，产生最终完全的调用序列。注意到，该重新排列并不重要，因为在前面的参比序列查找步骤中已经将各片段定位。GUI 也可以提供该调用序列的可视化工具。此外，GUI 可以显示最终序列的强度图。在这种情况下，碱基调用程序也必须输出强度文件(串接为调用序列数据)。

现有的碱基调用程序输出一种基于参比和点记录(例如来自 HyChip™)的短报告文件。这对于 rSBH 可能不是有用的，因为各片段的点分布在很多杂交载片中。代之以，对各杂交产生了一种新的“短报告”，它比 HyChip 的短报告更加简要。具体地，新报告可以列出各载片上的全匹配数目(N)和最高的 N 记录的中位数。它也可给出任何对照点的中位数，如标准参照物或空点，如果存在的话。该新报告的优点是可在经常升级的 GUI 桌面上实时观测各个图像。这将在早期(和运行全程)告诉用户，rSBH 系统是否正在产生有用数据，而不是等待一天才看到最终结果。该新报告的高级用途是，允许用户向 rSBH 装置反馈。例如，从 GUI 暂停/停止运行，或者如果任何一个库组失败，重复运行一个库组。GUI 也可以在运行中实时显示装置参数，如杂交和洗涤温度。最终，产物可以将该装置集成到用户 GUI 的命令和控制模块中。

C.碱基调用

因为合并的探针库对于每个片段都是相同的，所以 rSBH 碱基调用程序可以在合并的探针中对所有片段只阅读一次。碱基调用程序需要输入参比序列。对于 rSBH，参比来自对最高的几百个记录群集的分析。最高记录位置的简单存储仓(binning)算法是最有效的，因为它要求一次通过存储仓位置找到最大存储仓计数。最大存储仓计数的窗口定位了参比序列中的片段位置。用 250bp 片段和 1024 个测定，1/4 的片段记录是阳性的(即全匹配杂交记录)。然后，由于合并的探针的复杂性，1/4 的 10-mer 代表阳性记录。而且，对于长于 4^{10} 的参比序列来说，探针被重复，使参比序列中所有 10-mers 的 1/4 是阳性。对于 11-mers 和 12-mers 同样应用，所有参比探针的 1/4 是阳性。对于 1 毫微秒中可以把一个探针放入存储仓的处理器来说，为一个片段找

到参比序列需要[L÷4÷10^9]秒。对于极值 L＝10,000,000,000，用一个 CPU 需要 2.5 秒/片段。对于一百万个片段和 100 个 CPU 来说，找到参比序列的总时间为～25,000 秒(6-8 小时)。

将最高 L 记录放入存储仓的替代方法是对各片段进行从头型的序列组装，将上面实例中使用的探针数量减少到大大少于 250。如果该从头算法是快的(如少于 1 毫秒)，它将加速片段查找过程。一种快速从头算法可包括找到几组 10 个或更多具有重叠探针的最高 250 个记录，并可减少所需的时间一个数量级或更多。

D.碱基调用算法

1．读探针库文件

2．读参比序列(长度 RL)并储存到参比序列对象中。

2a. 产生参比序列位置数据结构。

3．读亮度文件(实时，因为它们由图像分析产生)。

3a. 把值存储到存储数据结构中。

4．累加各片段的最高 L 记录(中位数长度 L)

5．各片段的分析环

5a.为最高 L 记录建立一个参比序列中的位置列表

5b.建立长度为[RL÷(m×L)]的向量，将最高 L 记录位置放入存储仓中。这产生了 m×L 的存储长度，其中 m 应该是～1.5，以在片段的任何一侧提供边界。

5c.将最高 L 记录的位置放入存储仓向量中。

5d.找到总存储计数最高的区域。这给出片段参比序列到(m-1)×L 碱基位置内。

5e.用片段参比序列进行碱基调用。

5f.将调用序列串接成一个文件：称为“序列”(包括位置信息)。

6.13 附加实施方式

本发明方法允许通过多种机制设计探针和 IPPs。在一个实施方式中，通过改变每库探针数量来设计探针和 IPPs，更具体地，每库为 4 到 4096 探针范围。在第二个实施方式中，通过改变每组库的数量来设计探针和 IPPs，更具体地，每组为 4 到 1024 探针范围。探针可含有 2 至 8 个信息碱基，提供总共 4-16 个碱基。在另一实施方式中，用一些位置上的简并合成制备探针作为库。另一实施方式包括两组 IPPs 的两个组装物，其中不同探针混合在一个库内。

一个 20 到几百探针的小组可以提供单个核酸片段的独特杂交特征。将杂交模式与序列配对以鉴定病原体或任何其他核酸，例如计数 mRNA 分子。本发明方法的一个实施方式使用特征以在不同随机阵列上识别相同的分子。在不同阵列上杂交相同探针组后，这可以产生特征，由相同样品制备的不同阵列上不同亚组的测试探针杂交，然后结合每个体分子的数据。

本发明方法的另一实施方式进行不用组合连接、仅用单个 IPPs 组或单个探针的单分子 DNA 分析。在该实施方式中，通过用供体荧光团标记靶和具有受体荧光团的探针，或用受体荧光团标记靶和具有供体荧光团的探针检测 FRET 信号。可以个别地或作为在特定位置含有简并碱基(混合)的库，来合成 5'-N_x-B_{4-16}-N_y-3'形式的探针。在另一实施方式中，通过掺入一个或多个标记的核苷酸，基于聚合酶的杂交探针的延伸与探针/探针库杂交结合，其核苷酸一般是区别标记的。

本发明方法的另一实施方式利用探针去除，以便用相同探针序列完成一个靶分子的多个测试，可以用电场、磁场或溶液流从和向支持物表面重复去除探针分子。循环从每 1-10 秒至 20-30 秒发生。仅在探针去除开始后或仅在探针去除完成后，记录循环中各相的荧光信号。去除与温度循环偶联。在该实施方式中，探针去除不需要 FRET 标记，而是依赖来自一个标记的直接荧光。此外，FRET 反应在标记的探针和附着于靶分子的染料分子之间发生。

本发明方法涉及重复测试探针序列的另一实施方式，采用从容器外将相同探针重复加载到反应室中。然后，首先用不去除全匹配杂交体(如果用连接则是两个探针的连接产物)但去除游离探针的洗涤缓冲液，快速去除以前的探针负荷。在后续探针负荷引入之前，用第二次洗涤解链所有杂交体。

在另一实施方式中，仅测量一次各探针与靶分子的相互反应。该过程依赖在阵列内不同位置的相同 DNA 区段的多余代表，和/或依赖一次连接事件的准确性。

除在上样于支持物上形成阵列之前制备最终片段外，在本发明方法的另一实施方式中使用了两级切割程序。首先将样品 DNA 随机切割形成较长片段(约 2-200kb 或更长)。将这些片段的混合物装到支持物上，该支持物可以由包含约 10x10 微米^{2}大小单元的栅格形式的疏水材料组成。调节样品浓度使各单元中主要存在一个或几个长片段。这些片段将被进一步原位随机片段化，使最终片段长度约 20-2000 个碱基，并附着于支持物表面。最优单元大小取决于每单元引入的 DNA 总长度和最终片段的优选长度。本发明的这个片段化方法提供长范围的作图信息，因为一个单元中所有的短片段都属于一个或几个来自长重叠片段的长片段。这个推论简化了长 DNA 序列

的组装，并可提供全染色体单倍型结构。

在本发明的另一实施方式中，用例如柱从复杂的样品中捕获选择的靶 DNA，该柱含有特定基因或生物体的均等数量的 DNA 分子。例如，选择的病毒或细菌基因组，或部分基因组可以附着的单链 DNA(ssDNA)的形式在这些柱上出现。如果通过杂交到固定的 DNA 捕获了双链 DNA(dsDNA)和互补链，那么样品 DNA 被解链。洗涤掉过量的互补 DNA 或任何其他不相关的 DNA。然后用高温或化学变性去除捕获的 DNA。在感染剂的诊断中，可以用该方法去除人或其他复杂的 DNA。也提供了降低有过多代表的试剂浓度的方法，以检测在较小阵列上以低拷贝数存在的其他试剂。可以在管、多孔板的孔或微流体芯片中进行捕获过程。

通过用有下游切割的限制酶切割 DNA 和连接匹配接头来完成特异性基因或其他基因组片段的选择(共有、共待审的美国专利申请 10/608,293 中描述，整体纳入本文作为参考)。不被接头捕获的片段将被破坏或去除。另一实施方式使用 6-60 个碱基的核苷酸，或更优选 10-40 个碱基，或甚至更优选 15-30 个碱基，设计与具有一个或多个错配的给定序列配对，允许使用错配识别与切割酶一起切割 DNA，可以设计两个寡核苷酸用于切割具有约 1-20 个碱基移动的互补链，产生连接接头或连接载体臂的粘性末端。可以获得和捕获来自基因组片段的两对所述寡核苷酸切割模板，或为用特异性接头捕获而进行末端修饰。合成切割模板，或设计一个或多个短寡核苷酸文库以提供任何 DNA 的必需切割模板的通用源。256 个寡核苷酸的文库可用下列共有序列表示：nnnbbbnn、nnbbbbnn 或 cggnnnbbbbnn、nnbbbnn、nnbbbnnncac，其中 n 代表四个碱基的混合物或通用碱基，b 代表一个特定碱基，bbbb 代表 256 个可能的 4-mer 序列中的一个，cgg 和 cac 代表该文库中所有成员共享的特定序列的实例，它们可用来建立切割模板。为了建立切割模板，可以用组装模板连接两个或三个选自相应核苷酸文库的成员，该组装模板为 nnnnnnnnnnnnnnnnn 或 gccnnnnnnnnnnnnnnnnnnnnnnnnnngtg。

除了各种化学附着方法之外，使用附着到锚、接头、引物、其他附着到表面的特异性结合物的片段，将通过随机切割或特异性切割制备的 DNA 片段附着到表面。一个实施方式使用具有长度约 1-10 碱基的粘性末端的随机附着锚，将 ssDNA 片段或 dsDNA 片段与配对的粘性末端连接。附着到 DNA 片段的接头可以提供粘性末端。该方法提供了将底物面与具有不同粘性末端的锚用于鉴定附着片段的末端序列的可能性。另一实施方式将引物连接到支持物上，该引物与连接到 DNA 片段上的接头互补。在 ssDNA 与引物杂交后，用聚合酶延伸引物。将产生的 dsDNA 解链，去除没有附着

到支持物的链，用于下述的 DNA 扩增。另一实施方式中，用特异性结合物(例如环肽)覆盖表面，该结合物能识别 DNA 片段的 3'或 5'端并与它们高亲和性结合。

分析一侧或两侧上附着接头的短片段，可以帮助读遍回文结构和发夹结构，因为当回文结构/发夹结构内有一个切口时，新接头序列将与剩余序列不互补的序列连接。接头允许用所有重叠探针阅读靶 DNA 的每个碱基。

在另一实施方式中，通过使用单分子的随机阵列，然后原位、局部扩增(Drmanac 和 Crkvenjakov，1990，上述，纳入本文作为参考)增加了检测的准确度和效率，在相同像素区域内产生多达 10、多达 100、多达 1000、多达 10,000 个附着的复制分子。在这种情况下，无需要求单分子灵敏度，因为探针的多个记录是不必需的，即使仍可用 FRET 和 TIRM。扩增过程包含下述步骤：1) 用覆盖了一种引物(约 1000-50,000 个引物分子/微米 ^{2})的支持物，2) 用连接接头修饰的样品 DNA 片段和溶液中的第二引物。需要使混合和扩散最小化，例如通过用毛细管室(与支持物仅有 10-100 微米空间的盖玻片)或将靶和第二引物嵌入凝胶。由单个靶分子扩增产生的分子群将形成一个点，或“扩增子”，它的大小应该小于 10-100 微米。杂交或连接事件的扩增也可以用来增加信号。

一个优选实施方式使用连续等温扩增(即不同类型的链取代)，因为无需用高温变性 dsDNA，高温会引起大量扩散或紊乱，除了连接引物，该取代链不结合其他互补 DNA，可以产生局部的高浓度 DNA。另一使用等温扩增的实施方式是设计至少一个接头(针对靶 DNA 的一端)与具有低解链温度的核心序列(即用具有 3-13 个 TA 重复的 TATATAT...序列)，引物与此核心序列基本配对。在聚合酶能够进行此反应中所用的链取代的最佳温度下，TATATA...位点的 dsDNA 会局部解链，允许进行引物杂交并起始一个新的复制循环。可以调节核心的长度(即稳定性)，以适应 30-80℃之间的温度。在这个连续扩增反应(CAR)中，进行以前合成的酶从引物位点一移动，就开始合成新链，花费约数秒。用该方法只要使用一个引物，就可以从 dsDNA 开始产生高浓度的 ssDNA。对于一个引物附着于表面的扩增来说，低温解链接头应该是针对非附着端的，相应的引物在溶液中游离。除了聚合酶外，CAR 并不需要任何其他酶。为在可能需要高温解链的源 dsDNA 上进行两个或多个起始扩增循环，通过与 DNA 片段连接或靶特异性引物的尾延伸引入接头。

上述的核酸分析方法是单独基于探针/探针库杂交，或与碱基延伸或两个探针连接到样品 DNA 片段的随机阵列结合，该方法用于各种应用，包括：较长 DNA(包括细菌人工染色体(BACs)或整个病毒、整个细菌或其他复杂基因组)或 DNA 混合物的测序；

选择基因的诊断性序列分析；新生婴儿的全基因组测序；精确了解新作物和动物的遗传基因组成的农业生物技术研究；单个细胞表达监测；癌症诊断；DNA 计算的测序；监测环境；食品分析；和发现新的细菌和病毒生物体。

本发明方法从单个标记探针产生了足够信号，而将背景降低到检测的阈值以下。特殊的底物材料或包被物(如金属化)和先进光学系统被用来降低高系统背景，此高系统背景阻碍了从 1 厘米$^{-2}$表面并行检测几百万单分子。另外，在样品引入或在 DNA 附着过程中产生的背景可以通过样品的特殊处理降低，包括亲和柱、改良的 DNA 附着化学方法(如连接)或排除 DNA 特异性的结合分子(如环肽)。在一些情况下，降低由溶液中的未结合探针复合物或底物上的组装物所产生的背景，需要循环去除未杂交/连接的探针，该去除可通过下述方法进行：电场脉冲；特别工程改造的连接酶，酶具有最优化的热稳定性和全匹配特异性；或三重 FRET 系统，该系统具有第三种染料(如量子点)连接于靶分子。

在另一实施方式中，本发明方法需要用电场提高支持物上 DNA 分子的浓度，以便从随机阵列表面上的染色体或基因组中捕获所有片段。允许正确组装的染色体片段化可能需要间隔化的底物和原位片段化，原位片段化将起始单个的 100kb 到 1Mb 的 DNA 片段，获得较短的 1-10kb 片段的连接的组。

得到的快速杂交/连接允许用一对引物库在少于 60 秒/循环中多次探查靶，可能需要使用最优化的缓冲液和/或主动探针操纵，这可能使用电磁场。用激发性质与 DNA 稳定性兼容的荧光染料(或树状聚体)和照明(纳秒激光脉冲)的精确控制来增加系统(包括排列的靶 DNA 分子)的化学和物理稳定性，以允许几个小时的照明。

快速实时图像处理和来自重叠探针的单个片段和来自重叠的 DNA 片段的整个基因组的组装可能需要可编程的逻辑阵列或多处理器系统用于高速计算。

本发明方法依赖下述过程：通过标记探针和 DNA 酶产生可见荧光信号，来进行互补 DNA 序列的特异性分子识别。通过依赖自然进化的序列识别和酶校对过程，rSBH 消除了显著的技术挑战，该技术挑战是在物理上区别单个 DNA 碱基，它们仅有 0.3 纳米大小并且互相的差别仅几个原子。本发明方法也有非常简单的样品制备和处理，包括染色体或其他 DNA 的随机片段化，形成每平方微米包含约一个 DNA 分子小(1-10 毫米$^{-2}$)随机单分子阵列。本发明方法同时从几百万单 DNA 片段中高速收集数据。用 10 个荧光颜色和 10 兆像素的 CCD 相机，单个 rSBH 装置每秒可以阅读 10^5个碱基。本发明的阅读长度是可调的，从每片段约 20-20,000 个碱基，每个随机阵列上单个实验总共高达 1 千亿个碱基。通过单个长片段的起始片段化和相应短片段组连接到分

离的随机亚阵列，rSBH 方法的有效阅读长度可以高达 1Mb。对于各个测试的单 DNA 分子来说，每个碱基获得 100 个独立检测(即 10 重叠探针序列，平均各测试 10 个相同 DNA 分子的连续连事件)，这保证了最大的测序准确度。

用 IPPs 的组合 SBH 对 PCR 扩增几千个碱基长度的样品提供准确度超过 99.9%的序列数据。该阅读长度比目前使用的基于凝胶的方法获得的长度长许多倍，并在单个测定中提供全基因测序。本发明方法将基于杂交的 DNA 分析的并行性、准确性和简单性的优点与小型化的效率和单分子 DNA 分析的低材料成本相结合。应用通用探针组，组合连接和信息探针库允许对任何和所有 DNA 分子进行有效和准确的分析，并用单个小组的寡核苷酸探针库检测其中的任何序列变化。本发明方法用一个集成系统应用公知的生物化学和信息学，处理超高密度、随机单分子阵列，以获得比现有的凝胶和 SBH 测序方法显著高 1,000 到 10,000 倍的测序通量。本发明方法将允许对所有存在于复杂生物样品中的核酸分子进行测序，包括未经 DNA 扩增或几百万克隆操作的细菌、病毒、人和环境 DNA 的混合物。使样品处理和低化学品消耗最小化，以及完全集成的方法将每个碱基的测序成本降低至少 1,000 倍，或更多。本发明方法能够在一天内，在单个阵列上对整个人类基因组进行测序。

容易制备短 DNA 片段的随机阵列，该阵列比目前使用的大多数标准 DNA 阵列密度高 100 倍。探针与该阵列杂交和先进的光学系统允许用兆像素 CCD 相机进行超快速并行数据收集。阵列中各像素监测不同 DNA 分子的杂交，以每秒 1-10 帧的速率提供数千万的数据点。随机阵列可以在 3x3 毫米表面上包含超过 1 千亿个碱基对，在 10-100 个像素单元中代表每个 DNA 片段。SBH 方法提供的固有冗余度(其中几个独立重叠探针阅读每个碱基)帮助保证了最高的最终序列准确度。

为达到本发明连接方法的全部容量，即每分子阅读高达 1000 个碱基，必须同时处理多个 IPP 试剂。本发明连接方法不需要共价修饰每个需要分析的靶分子。因为 SBH 探针不和靶共价结合，所以它们在循环间可以被容易地去除或光致漂白。此外，包含聚合酶保证了在任何给定的 DNA 分子中，一个碱基只能被测试一次。本发明的杂交/连接方法允许用每个给定探针多次探查，并通过几个重叠探针多次探查各碱基，为每个碱基提供的测量数量增加了 100 倍。此外，连接酶比聚合酶允许利用较大的标记结构(即具有多个荧光团或量子点的树状聚体)，这进一步提高检测准确度。

本发明方法可以用常通用探针的较小不完全组产生长 DNA 分子的通用特征分析。每个像素可以分析长达 10kb 的单分子。片段长度为 10,000bp，1 千万片段的阵列包含一万亿(10^{12})DNA 碱基，相当于 300 个人类基因组。用一个 10 兆像素 CCD 相机

分析这个阵列。在 10-100 分钟内获得信息特征，取决于多重标记的水平。分析较之小 10-或 100-或 1000-倍的阵列在特征或测序或定量应用中是非常有用的。

在一个实施方式中，在阵列中用 10-10,000 片段代表一个病原体细胞或病毒，因此不需要 DNA 扩增。本申请的单分子特征方法提供了对病原体基因组每个区域的综合调查，说明比在标准探针阵列上分析成千个 DNA 扩增子的多重扩增有显著改进。DNA 扩增是一个非线性过程，在单分子水平并不可靠。不是每个病原体扩增几个区段，用病原体亲和柱将不需要或污染 DNA 的浓度降低，可以分析收集病原体的整个基因组。可以从上千人体细胞中收集单个病毒或细菌细胞，收集的细胞由 10-1000 像素上的 1 至 10kb 的片段代表，提供准确鉴定和精确的 DNA 分类。

在另一实施方式中，用本发明方法检测并抵御生物战剂。rSBH 鉴定结标记，这允许在致病性和症状发生之前，在单个生物体水平对生物剂进行立即检测。SBH 对参与病原体的攻击方式、毒性和抗生素敏感性的任何或所有基因提供综合分析，以便迅速地了解参与的基因和如何回避任何和所有这些基因。rSBH 可以分析包含病原体、宿主和环境 DNA 混合物的复杂生物样品。此外，本发明方法采用快速、低成本综合检测方法可用来监测环境和/或员工，并且可以制成便于携带的产品。

6.14 试剂盒

本发明也提供作为产品的 IPP 试剂盒，可以装载到盒或有预加载探针的盒，试剂盒任选地包括含有缓冲液和酶的连接混合物。

本发明也提供病原体/基因-特异性样品制备试剂盒，和从样品如血样中分析病原体的方案，病原体如炭疽杆菌(*Bacillus anthracis*)和鼠疫耶尔森氏菌(*Yersina Pestis*)。本发明提供将样品制备 DNA 产物集合到底物中，形成本发明的 rSBH 阵列。描述了每像素产生单个靶阵列的逐步方法，和任选的每像素产生 10-1000 拷贝的原位扩增。产生了靶 DNA 的随机阵列，该阵列进行 rSBH 的序列分析。本发明中制备底物的模块化方法允许早期形式的底物具有简单的样品应用位点，而最终开发的底物可以具有一个“即插即用”阵列制备模块。

满足最小纯度和量规格的 DNA 样品，将用作用 rSBH 样品排列技术进行的真实样品整合的起始材料。样品整合以酶消化(限制酶或核酸酶消化)原始样品产物开始，产生提供特异性(或随机)粘性末端的长度大约 250bp 的片段。这个酶混合物代表产品试剂盒中或许提供的几个组分之一。

消化的排列涉及粘性末端与排列在表面上补体的连接。按照下述方法从其原始玻璃表面修饰阵列表面：1)形成氨丙基硅烷单层；2)用对称的二异硫氰酸盐激活；

3)用新的氨基化寡核苷酸混合物(包括捕获探针、引物探针和间隔探针)，用异源的单层探针修饰激活的阵列表面。

所有的附着探针共享保守的设计(>90%)，因此阻止了间隔和捕获探针被隔离的同源岛形成。捕获探针与所有其他探针的比产生等于 1 互补连接位点(样品和捕获探针)/平方微米的平均密度，通过超灵敏 CCD 的单个像素观察各平方微米。下一步，将消化的 DNA 样品加入预先形成的阵列表面，用 T4 连接酶连接到捕获探针，获得由每像素一个靶组成的新 rSBH 反应位点。从阵列表面去除多余样品，通过加热和附加洗涤，dsDNA 产生 ssDNA。这里，在捕获探针设计中使用磷酸化策略，以保证实际上仅有一条链共价连接到 rSBH 阵列，另一条链被洗涤去除。

在适合的公知技术(Andreadis 和 Chrisey，*Nucl. Acids Res.* 28: E5 (2000)；Abath 等，*Biotechniques* 33: 1210 (2002)；Adessi 等，*Nucl. Acids Res.* 28: E87 (2000)，以整体纳入本文作为参考)的检测中，局部原位扩增靶对于产生满意的检测信号是必需的。等温链取代技术可能是最合适于局部低拷贝数扩增的。为了隔开捕获探针，需要掺入间隔探针和引物探针。这些探针共享一些保守序列和结构，每种探针都起到其名称描述的作用。因此，捕获探针捕获靶 DNA，间隔探针帮助形成间隔合适的单层探针，如果需要，引物探针的存在是为了进行原位扩增。所有靶逐渐清除相同的阵列引物序列，简化任务。一旦样品与阵列连接，阵列 DNA 的游离末端就会获得通用引物用于扩增。用标准方案和材料(即引物、聚合酶、缓冲液、NTPs 等)在阵列内的分子上进行原位扩增。虽然 10-1000 个拷贝可能足够，但是只需要约 50 个拷贝即可。可用不同效率扩增各靶，而并不影响序列分析。

总之，样品整合和 rSBH 阵列形成需要 DNA 消化原始样品制备的产物、分离和集合到底物，以形成 rSBH 阵列。本发明提供了与消化、分离和连接步骤有关的试剂和试剂盒。

7. 实施例

7.1 对一个细菌基因组进行测序

对一个通常无毒实验室菌株的整个细菌基因组进行测序。选择了已经很好表征且序列已知的大肠杆菌菌株。在一个一天测定中对整个基因组进行测序。该测定说明诊断性系统的全部操作，以及确定与系统输入和输出设计中相关的重要规格和对于原始样品分离和制备的一般要求。

来自划线平板的单一的菌落或几微升液体培养物提供足量材料。裂解细胞和分

离 DNA，用本领域公知的方案(参见 Sambrook 等，《分子克隆:实验室手册》，冷泉港实验室出版社，纽约州(1989)或 Ausubel 等，《新编分子生物学实验指南》，John Wiley & Sons，纽约，纽约州(1989)，二者以整体纳入本文作为参考)。产率并不重要，重要的因素是 DNA 的质量。本实施例中定义的样品规格应用于所有其他样品。将拷贝数为 10-100 的基因组用于最终分析。本测定的附加要求是：1)DNA 不含 DNA 处理酶；2)样品不含杂质盐；3)整个基因组被平均代表，且构成大部分总 DNA；4)DNA 片段长度在 500 至 50,000 个碱基之间；和 5)样品以无菌 DNA 溶液提供，浓度已知(例如 1.0 微克/毫升，1 微升就足够)。

10-100 的输入拷贝数保证整个基因组的重叠，并容许阵列上捕获靶差。用 10-100 拷贝获得足够的重叠片段，以保证碱基调用和高准确度的足够成功。rSBH 样品的质量约为 1-10 皮克，其中大部分用于样品的鉴定和定量。通过连续稀释鉴定的产物，获得用于分析的样品。

DNA 必须不含蛋白，尤其是核酸酶、蛋白酶和其他酶。用基于苯酚的提取，如 PCl 将大部分蛋白去除并失活(Sambrook 等，1989，上述；Ausubel 等，1989，上述)。低渗裂解或基于去垢剂的裂解(用核酸酶抑制剂混合物，如 EDTA 和 EGTA)，接着 PCl 提取是一种快速有效的样品消化和一步 DNA 分离的方法。锁相提取(可用于 3'5')简化了这个任务，产生纯净的 DNA。此时不需要 DNA 消化，因为裂解和提取期间的剪切力产生了所需长度范围的片段。通过严格纯化 DNA(即后续的氯仿提取、乙醇沉淀和大小排阻)完成苯酚去除。用苯酚留下的紫外(UV)光谱特征进行纯度检测和 DNA 定量。

DNA 必须不含杂质盐和有机物，悬浮在 SBH 相容性 Tris 缓冲液中。这通过大小排阻层析或微量透析完成。

原始 DNA 样品范围在 500bp 到 50,000bp。低于 500bp 的片段很难在分离和纯化中复性，且也影响阵列过程。大于 50,000bp 的片段很难溶解，并且可以不可逆地聚集。

以 1 微克/毫升的无菌溶液提供样品至少 1 微升。原始 DNA 的总需要量仅为～1 纳克至 1 皮克，比进行测序量的 1%还少。

对于最终样品制备来说，消化 DNA 产生预期平均长度约 250bp 的片段，该片段带有可用于在组合阵列表面上排列分子的粘性末端。间隔这些分子，使每平方微米出现一个分子，这是通过 CCD 相机的单个像素观察到的，且代表几百万个孔的阵列内的一个虚拟反应孔。这需要消除自组装单层(SAM)效应。使用一种酶驱动流程，该流程将样品与间隔排列在组合阵列单层内的特异性位点连接，所述单层化学附着到

检测底物的表面上。通过 SAM 化学方法驱动捕获阵列，但是在末端互补突出端中的小变化不应该产生类似序列的岛。因此，用捕获阵列制备底物，将样品通过合适突出端的酶连接附着到底物表面。

此外，需要原位扩增各靶，产生“扩增子”。用通用引物接头完成扩增，该接头通过未在起始捕获连接中附着的末端连接到靶序列上。用 DNA 聚合酶和 NTPs 合成一条新链，取代原始的补体，在捕获阵列上提供具有互补元件的取代链，因此依此捕获并连接。预期通过线性扩增产生～10 个拷贝。此外，可用指数扩增策略每微米产生 100-1000 个拷贝。

用专用探针和集成微流体将阵列样品、单分子或者局部扩增子进行 rSBH 循环测序。将生物信息学完全整合用于数据收集、储存、分析和序列比对。用碱基调用和准确度的统计分析候选生物体的基因组序列报告结果。

7.2 从炭疽杆菌和鼠疫耶尔森氏菌细胞培养物或血样制备样品

7.2.1 全基因组分析

从原始样品中分离具体的病原体需要从原始样品中分离或富集细胞，然后裂解产生具体的基因组。标准生物化学和细胞生物学实验室技术，例如分部离心、过滤、培养或亲和层析用来分离细胞，然后提取基因组。一般地，大部分病原体比人细胞至少小两个数量级，而比大部分生物分子结构大几个数量级，因此合理地允许通过传统物理技术进行容易的分离。优选用市售的抗体或其他已用于某些靶的亲和工具来层流分离，并将风险减至最小，该亲和工具如病毒外被蛋白。生物体富集时，用标准方案将生物体裂解并分离 DNA。

此外，可以用含有可逆亲和标记的特异性引物(用于异源的原始样品)或通用引物组(用于分离的细胞类型)完成基因组扩增。将样品进行裂解，如果需要，并进行原始 DNA 的分离。将引物和扩增混合物同时加入原始样品，通过可逆标记和亲和捕获分离产物。

7.2.2 基因组足迹分析

该方法包括扩增特定组的足迹基因，该足迹基因对感兴趣的生物体特异。通过同时检测多个基因区域，可以区分相同病原体的不同菌株，或者可筛选大量的不同病原体。下述文献(Radnedge 等，*App. Env. Micro.* 67:3759-3762 (2001)；Wilson 等，*Molecular and Cellular Probes* 16:119-127 (2002)；Radnedge 等，*Microbiology* 148:1687-1698 (2002)；Radnedge 等，*Appl. Eizv. Micro.* 印刷中(2003)，以整体纳入本文作为参考)中描述了可用于检测各种生物威胁性病原体的测定。鉴定了对感

兴趣的病原体特异的 DNA 区域，而该病原体的近亲中不存在这些区域。然后设计引物，检查环境样品中 DNA 产物的扩增。将炭疽杆菌和鼠疫耶尔森氏菌用作模式生物。确定量的病原体细胞与人血混合，以确定检测的灵敏度。早期症状的病人血液中含有这些病原体中任何一个，浓度>10^4细胞/毫升血液。目的是在其到达症状期之前检测病原体。检查具有 10^1到 10^5细胞/毫升的血样，以确定检测的准确度。用 QiaAmp 组织试剂盒 250 (Qiagen 公司，Valencia，CA)或 NucleoSpin Multi-8 血液试剂盒(Macherey-Nagel 公司，Düren，德国)提取基因组 DNA。通过平板接种对数中期细胞并用血球计数板显微计数来测定病原体浓度，显微计数中将 10 微升稀释细胞加入到 190 微升人血中，接近症状前浓度。然后提取基因组 DNA，准备用于诊断靶和基因的扩增。

7.3 从无生物危害性的炭疽杆菌和鼠疫耶尔森氏菌样品中制备 100 个诊断靶的测定

选择靶，以鉴定可能的抗生素抗性区、毒性基因的突变和遗传工程中可作参考的载体序列。这种靶，尤其是毒性和抗生素抗性基因，一般并不对具体病原体独特，但提供了附加的定性信息。将用 50 个引物对扩增靶 DNA，以探查各病原体中相关的独特和定性区域。将产物汇入一个样品中用于 SBH 分析。可以用多个引物对以简化靶序列的扩增。

使用的引物具有一个可切割标记，用于从原始复合 DNA 混合物中分离扩增子。优选地，该标记是基于生物素/链亲和素的，具有一个 DTT 可切割二硫键，或者引物内特别地基因工程的限制性位点。通过亲和标记分离扩增子，并作为纯化的 DNA 样品释放。进一步通过大小排阻纯化产物，去除任何不需要的盐和有机物，然后定量用于下游整合。

7.4 从微生物生物薄膜测序样品

用结合现场研究和 FISH 的 rSBH 检查生物薄膜群落基因组。用 rSBH，在多于一个时间点和从不同栖息地对生物薄膜群落测序，以确定基因种。样品间的 DNA 标准化简化了分析，以突出在群落结构的基因种水平中的差异，并提供对低丰度基因种基因组的显著覆盖。根据非常确实的流程构建各样品的 16S rDNA 克隆文库。区别种系型的 FISH 探针和靶向 SNP 以区别种系型内细微变体用来绘制分布模式图，并提供 SBH 确定的基因种和 16S rDNA 种系型分布间的相关性。从物理和化学性质不同的栖息地收集样品，在样品收集时间测量主要环境参数，包括 pH、温度、离子强度、氧化还原态(即 Fe^{2+}/Fe^{3+}比)和溶解的有机碳、铜、锌、镉、砷和其他离子的浓度。

7.5 碱基调用模拟测试

用 90%重叠的 250bp(平均长度)片段的大肠杆菌产生模拟数据。用标准单碱基改变调用分析首先的 10,000 个片段。这个量超过了足够检查准确度和计时的量。参比序列查找在全部 4Mbp 参比基因组中成功找到 10,000 个片段位置。此外，在每个片段上进行的碱基调用都是正确的。将各片段依全部 4Mbp 参比序列存储，确证查找计时和准确度与受测片段的数目无关。参比序列查找和碱基调用所需的时间是 0.8 秒/片段。碱基调用包括单个碱基改变的测试和用于优化准确度的标准化。在参比序列查找中允许片段两侧出现边缘，这增加了分辨时间。

7.6 使单个 Q-点阵列和成像

将 2 微升 0、8、160 和 400 皮摩尔的结合链亲和素的 Q-点(Qdot 公司，Hayward，CA)沉积在生物素修饰的盖玻片(Xenopore 公司，Hawthorne，NJ)表面(在盖玻片中心)，2 分钟。通过真空去除液滴。施以 10 微升去离子水，再以相同方式去除。该洗涤重复 4 次。将盖玻片颠倒处理，置于干净的载玻片上。用 1 微升水将载玻片与表面粘合。将少量物镜浸油放在盖玻片边缘，以通过盖玻片周围产生密封阻止蒸发。

用 Zeiss Axiovert 200 显微镜，通过 Plan Fluar 100x 浸油物镜(1.45 na)表面照明成像。用标准色品度 Cy3 滤光片组从 Q 点成像的 655nm 发射。色品度 Cy3 发射滤光片的透射光谱与用于 655nm Q 点的发射滤光片重叠。用 Roper Scientific CoolSNAP$_{HQ}$™ 相机(Roper Scientific 公司，Tuscon，AZ)记录图像，曝光时间是 50 毫秒。从这些图像可以明显看出，较高的 Q 点浓度产生更可见的点。由于各种污染，用水点沾的对照盖玻片仅含几个可见点。除了看到 Q 点组具有预计颜色的稳定荧光之外，也看到个别亮度和颜色都不同的闪亮点。这些特征说明这些小点是单 Q 点。遥远的波长、聚焦平面外或个别颗粒之间的活性变化可以解释亮度差异。这些结果的显著性是，单个分子如果标记了 Q 点，就可以用先进的显微技术检测到。用 TIRF 系统进一步降低背景，希望通过激光能更有效地激发，以进行单个荧光分子的常规准确检测。

7.7 连接信号和点样靶和寡核苷酸

设计这些实验以证明：1) 点样靶可以用作两个探针连接的模板，具有良好的全匹配特异性；2) 点样寡核苷酸可以用作引物(或捕获探针)，以将靶 DNA 连接到表面。

A.装配载玻片

可以用作靶或引物或捕获探针的 4 个 5'-NH_2-修饰的寡核苷酸(序列参见表 1)，以 7 个不同浓度(1、5、10、25、50、75、90 皮摩尔/微升)点在 1, 4-亚苯基异硫氰

酸盐衍生的载玻片上，各浓度重复 6 次。长的 Tgt2-Tgtl-rc 寡核苷酸包含整个 Tgt2 序列和 Tgtl 互补序列的一部分(下划线部分是反平行方向中互补的)。将 Tgt2-Tgtl-rc 用作可以被 Tgtl 捕获的测试靶，通过把 2-探针连接与直接点样并被 Tgtl 捕获的 Tgt2 序列相比，可以测定捕获效率。

表 1 寡核苷酸用作靶或引物或捕获探针

引物名称	序列	SEQ ID NO:
Tgtl	NH-C6-C18-C18-CCGATCTTAGCAACGCATACAAACGTCAGT-3' (30mer)	1
Tgt2	NH-C6-C18-C18-TTCGACACGTCCAGGAACGTGCTTCAATGA-3' (30mer)	2
Tgt3	NH-C6-C18-C18-GTCAACTGTACCTATTCAGTCACTACTCAT-3' (30mer)	3
Tgt4	NH-C6-C18-C18-CAGCAGTACGATTCATACTTGCATAT-3' (26mer)	4
Tgt2-Tgtl-rc	TTCGACACGTCCAGGAACGTGCTTCAATGAACTGACGTTTGTA TGCGTTG-3'	5

B. 实验 1

在一个封闭室中，室温下进行杂交/连接 1 小时。反应溶液包含 50 毫摩尔 Tris、0.025 单位/微升 T4 连接酶(Epicentre, Madison, WI)和 0.1 毫克/毫升 BSA、10 毫摩尔 $MgCl_2$、1 毫摩尔 ATP、pH 7.8 和不同量的连接探针库(见表 2)，探针库从 0.005 至 0.5 皮摩尔/微升。反应后，用 3x SSPE 在 45℃下洗涤载玻片 30 分钟，然后用重蒸馏水冲洗 3 次，离心干燥。然后在 Axon GenePix4000A 上扫描这些载玻片，PMT 设置在 600 毫伏。

表 2

库 1	FM 库	SMM1 库	SMM2 库
Tgt1-5'-探针	5'-NNNTGTATG (SEQ ID NO: 6)	5'-NNNTGTAAG (SEQ ID NO: 7)	5'-NNNTGTATG (SEQ ID NO: 6)
Tgt1-3'-探针	5'-CGTTGNN-* (SEQ ID NO: 8)	5'-CGTTGNN-* (SEQ ID NO: 8)	5'-CGATGNN-* (SEQ ID NO: 9)
Tgt2-5'-探针	5'-NNNCACGTT (SEQ ID NO: 10)	5'-NNNCACGAT (SEQ ID NO: 11)	5'-NNNCACGTT (SEQ ID NO: 10)
Tgt2-3'-探针	5'-CCTGGNN-* (SEQ ID NO: 12)	5'-CCTGGNN-* (SEQ ID NO: 12)	5'-CCAGGNN-* (SEQ ID NO: 13)
Tgt3-5'-探针	5'-NNNGACTGA (SEQ ID NO: 14)	5'-NNNGACTCA (SEQ ID NO: 15)	5'-NNNGACTGA (SEQ ID NO: 14)
Tgt3-3'-探针	5'-ATAGGNN-* (SEQ ID NO: 16)	5'-ATAGGNN-* (SEQ ID NO: 16)	5'-ATCGGNN-* (SEQ ID NO: 17)
Tgt4-5'-探针	5'-NNNGTATGA (SEQ ID NO: 18)	5'-NNNGTATCA (SEQ ID NO: 19)	5'-NNNGTATGA (SEQ ID NO: 18)
Tgt4-3'-探针	5'-ATCGTNN-* (SEQ ID NO: 20)	5'-ATCGTNN-* (SEQ ID NO: 20)	5'-ATGGTNN-* (SEQ ID NO: 21)

注：*说明 标记的 Tamra，下划线碱基说明单碱基错配的位置。

C.实验 2

用 4 个 NH_2-修饰的 26-32 mers 点样载玻片，用溶解在 20 微升 50 毫摩尔 Tris 的 1 皮摩尔长靶 Tgt2-Tgt1-rc(表 1)和 0.1 毫克/毫升 BSA、10 毫摩尔 $MgCl_2$，pH 7.8，室温下杂交该载玻片 2 小时。载玻片于 45℃用 6x SSPE 洗涤 30 分钟。然后在 0.5 单位/20 微升的 T4 连接酶存在下，与连接探针(Tgt2-5'探针和 Tgt2-3'探针，表 2)在室温下孵育 1 小时。反应后，如上所述洗涤并扫描载玻片。

D.结果

1.连接信号取决于反应溶液中点样靶的浓度和 5'探针和 3'探针的浓度

图 12 显示连接信号对溶液中点样靶和连接探针的依赖性。当点样靶浓度约为 75 皮摩尔/微升，20 微升反应溶液中连接探针(探针-5'和探针-3')约是 1 皮摩尔时，获得最高信号。这些依赖性说明观察的信号实际上是连接依赖的信号，点样靶可以用作连接模板。全匹配连接探针和单碱基错配探针之间的区别约是 4-20 倍(表 3)。

表 3：连接信号的全匹配和单错配区别

靶	5'-探针的 FM/SMM	3'-探针的 FM/SMM
Tgt1	14	20
Tgt2	7	12
Tgt3	9	16
Tgt4	4	4

2.点样寡核苷酸可以用作引物(或捕获探针)，以有效地连接靶 DNA

点在载玻片上的寡核苷酸 1(Tgt1)用作靶 Tgt2-Tgt1-rc 的捕获探针，在其 3'-侧，Tgt2-Tgt1-rc 包含一个 Tgt1 的反向互补序列，在其 5'-侧，包含一个 Tgt2 的反向互补序列。在杂交/捕获 Tgt2-Tgt1-rc 后，连接探针(Tgt2-5'探针和 Tgt2-3'探针)杂交连接到 Tgt2 靶的点上，也杂交到 Tgt1 靶的点上。图 13 显示了观察到的连接信号。很清楚，在这个条件下，点样靶可以用作引物(或捕获探针)，以可以用来杂交/连接短探针的形式来连接靶 DNA，用于序列测定。

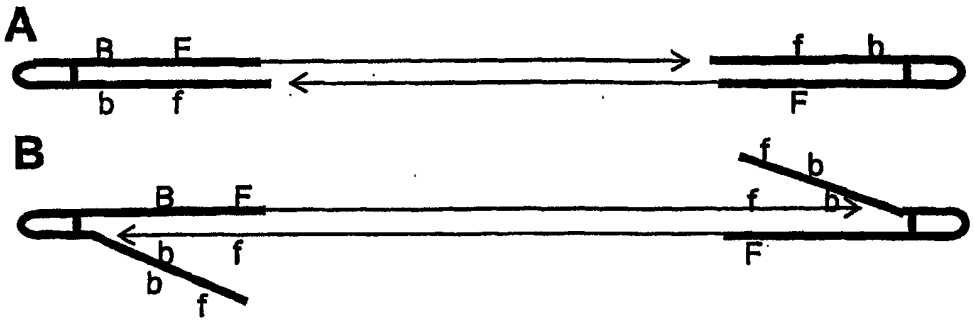


图 1

接头序列:

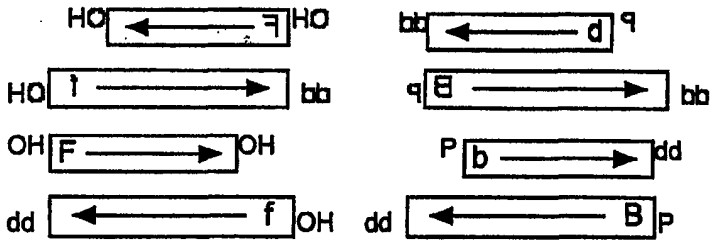
Ad-F 5'OH-GGGGTTACACAATATCATCTACTGCACTGA-3'OH (SEQ ID NO: 22)
Ad-f 3'dd-CCCCAATGTGTTATAGTAGATGACGTGACTNNNNNNNN-5'OH (SEQ ID NO: 23)

Ad-b 5'P-TCAGTAATAGCCTTAGACCGATTTCAGAAC-3'dd (SEQ ID NO: 24)
Ad-B 3'dd-NNNNNNNAGTCATTATCGGAATCTGGCTAAAGTCTTG-5'P (SEQ ID NO: 25)

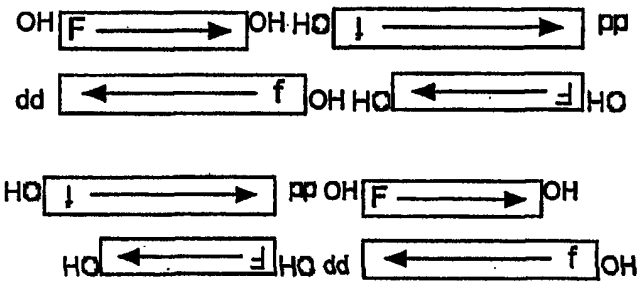
接头和基因组DNA(黑条)的排列:

***箭头从5' 指向3'

F和B接头不能互相连接:



F不能自连接:



B不能自连接:

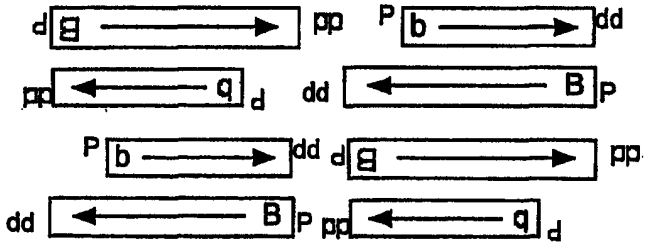


图 2

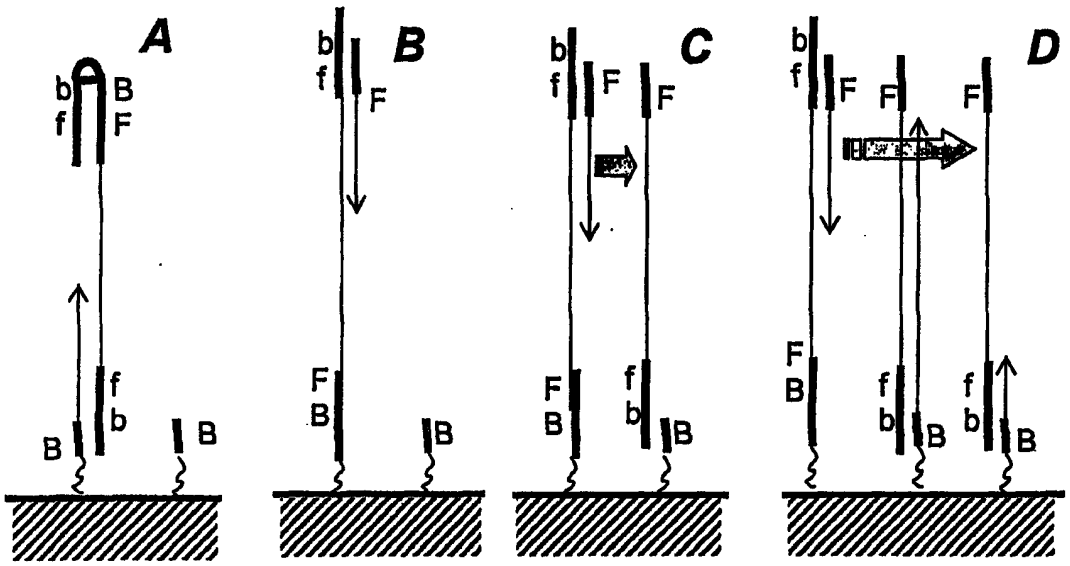


图 3

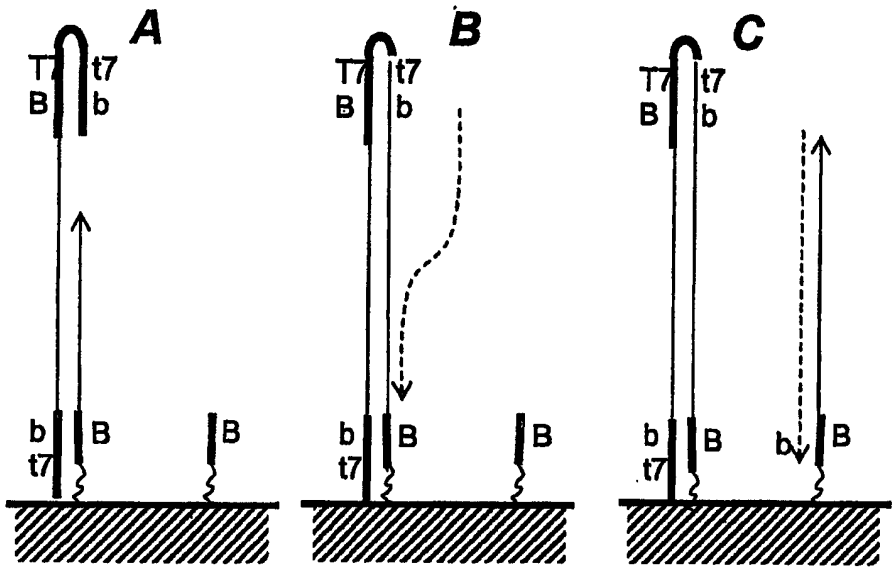


图 4

侵入物介导的等温DNA扩增图解

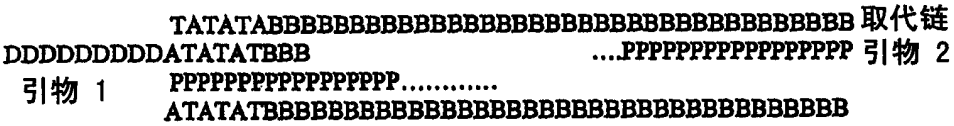


图 5

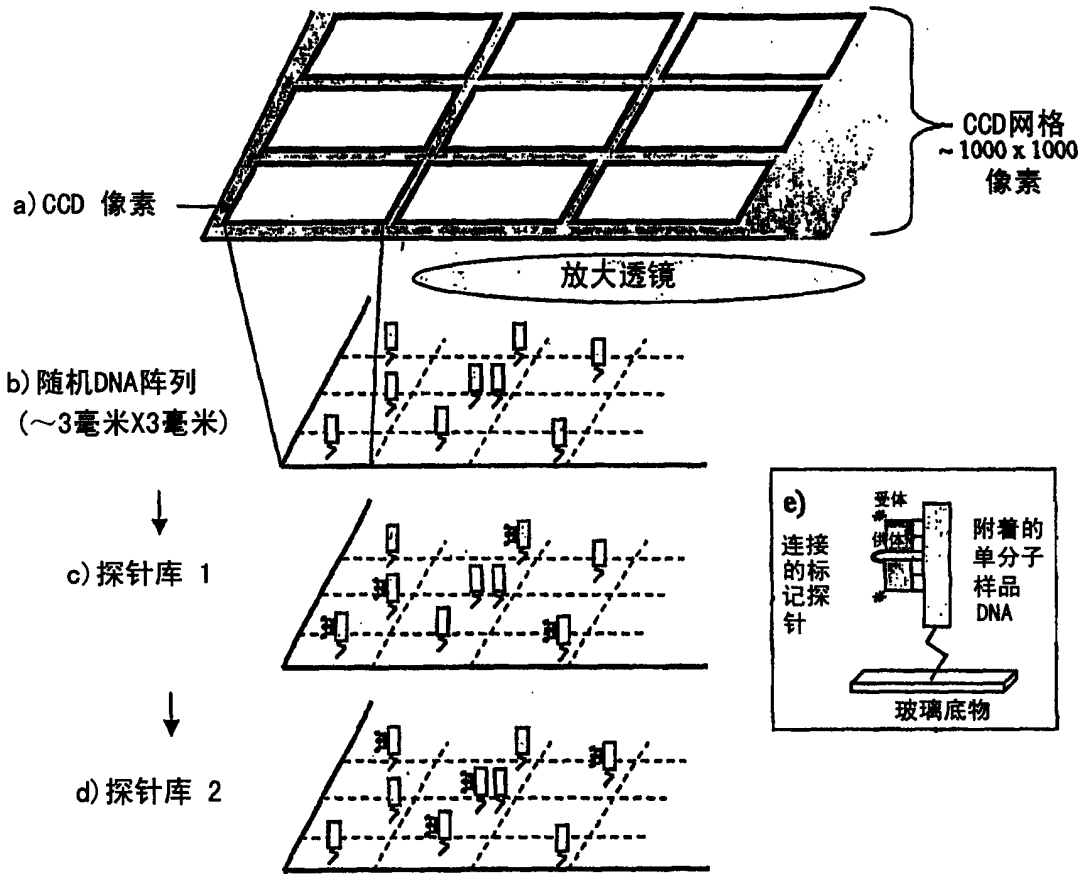

a) CCD 像素
CCD网格
~1000 x 1000
像素
放大透镜
b) 随机DNA阵列
(~3毫米X3毫米)
c) 探针库 1
d) 探针库 2
e)
连接
的标
记探
针
受体
供体
附着的
单分子
样品
DNA
玻璃底物

图 6

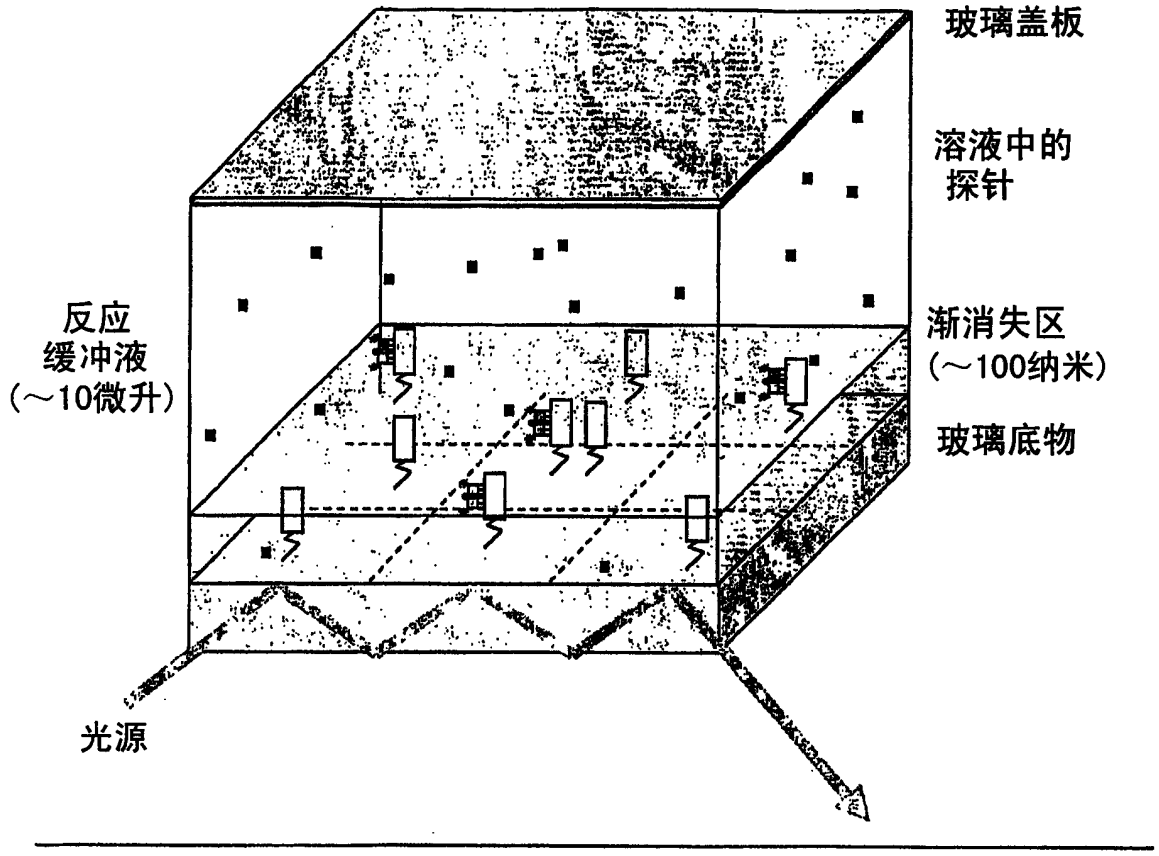


图 7

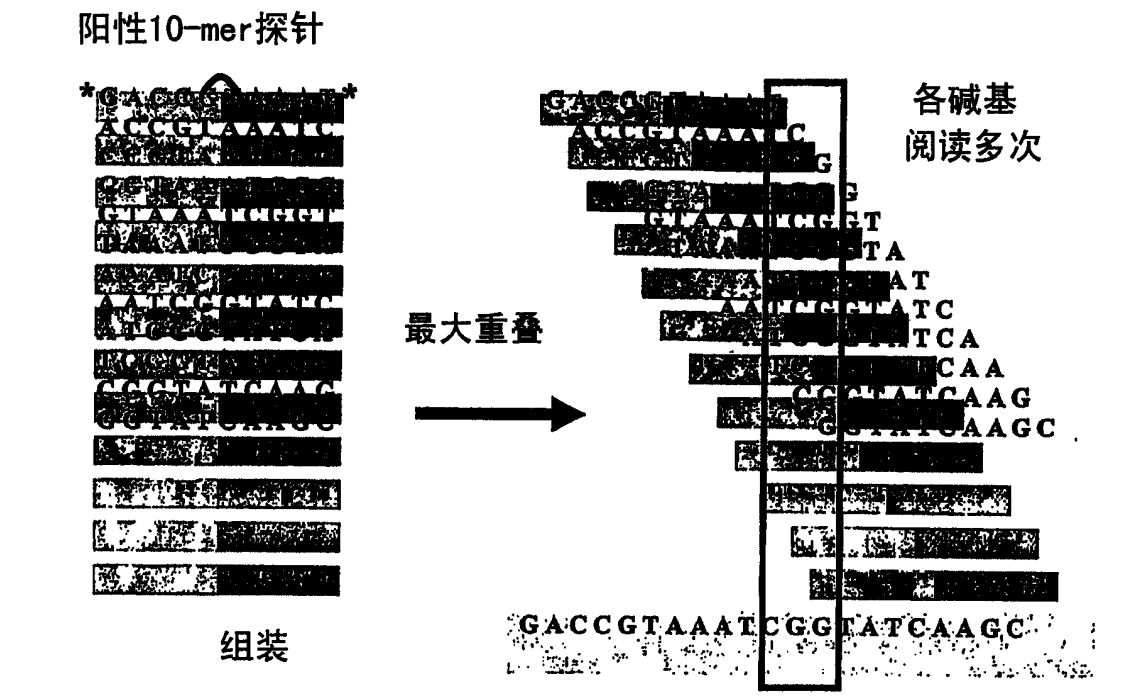


图 8

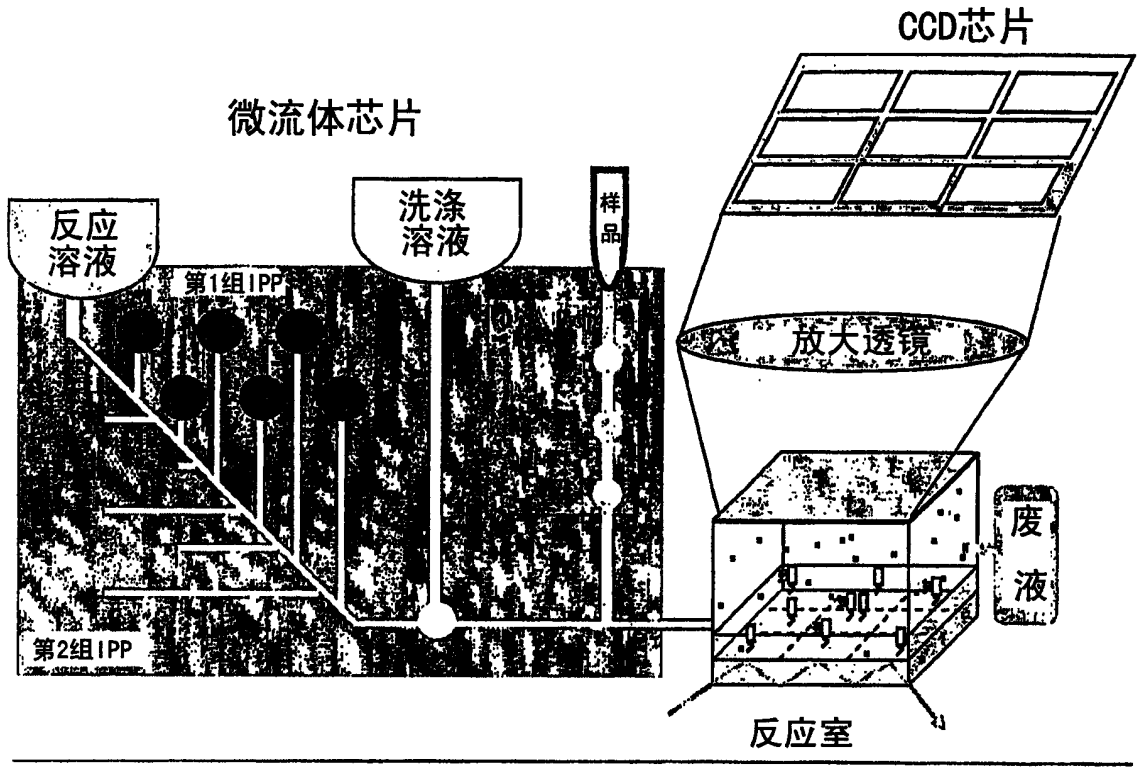


图 9

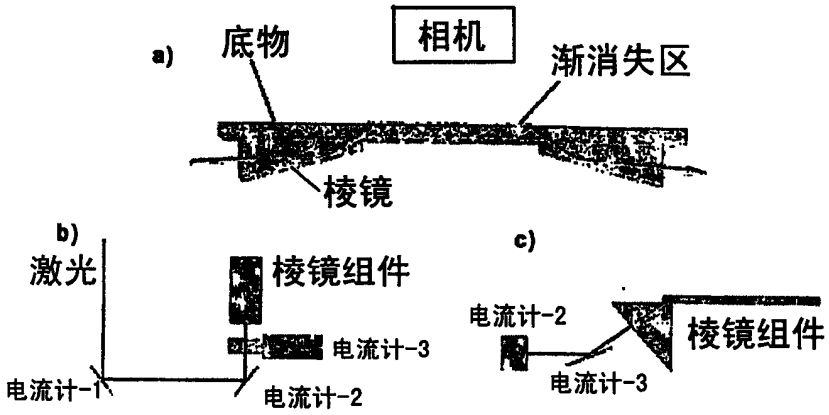


图 10

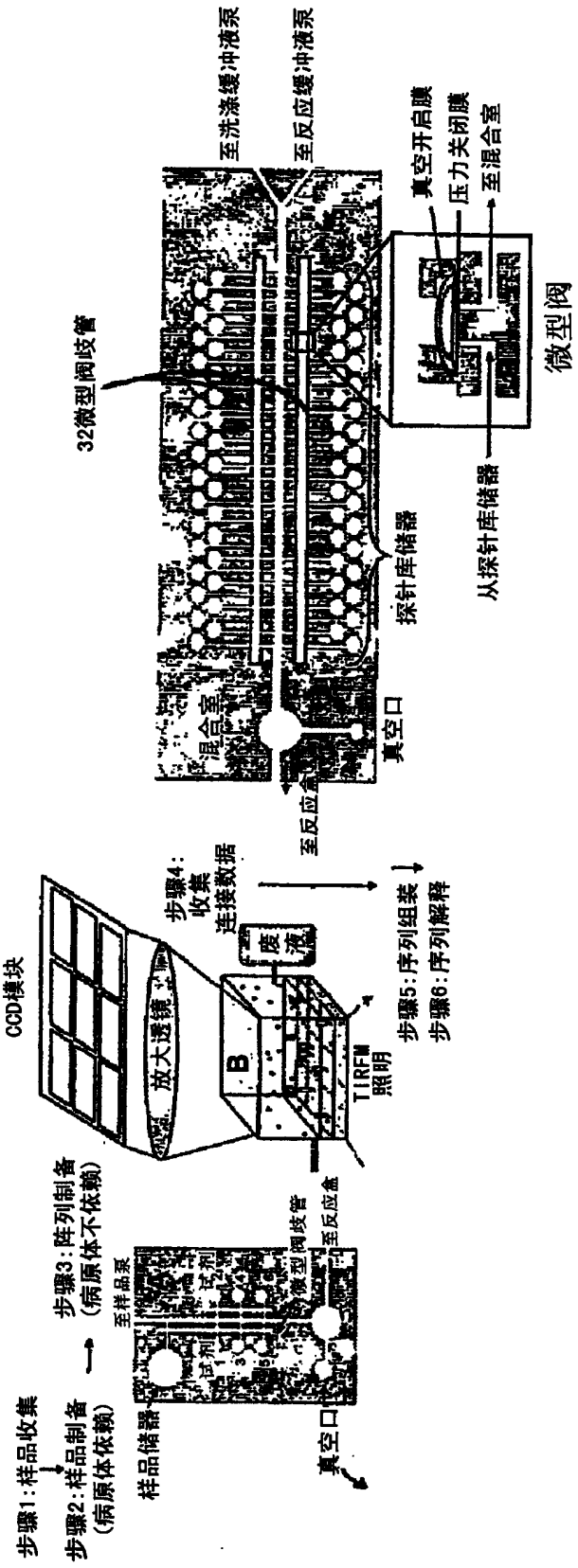

步骤1:样品收集
步骤2:样品制备
(病原体依赖)
样品储器
至样品泵
步骤3:阵列制备
(病原体不依赖)
真空口
至反应盒
CCD模块
放大透镜
B
TIRFM
照明
废液
步骤4:
收集
连接数据
步骤5:序列组装
步骤6:序列解释
至反应盒
真空口
混合室
32微型阀歧管
至洗涤缓冲液泵
至反应缓冲液泵
探针库储器
真空开启膜
压力关闭膜
至混合室
从探针库储器
微型阀

图 11

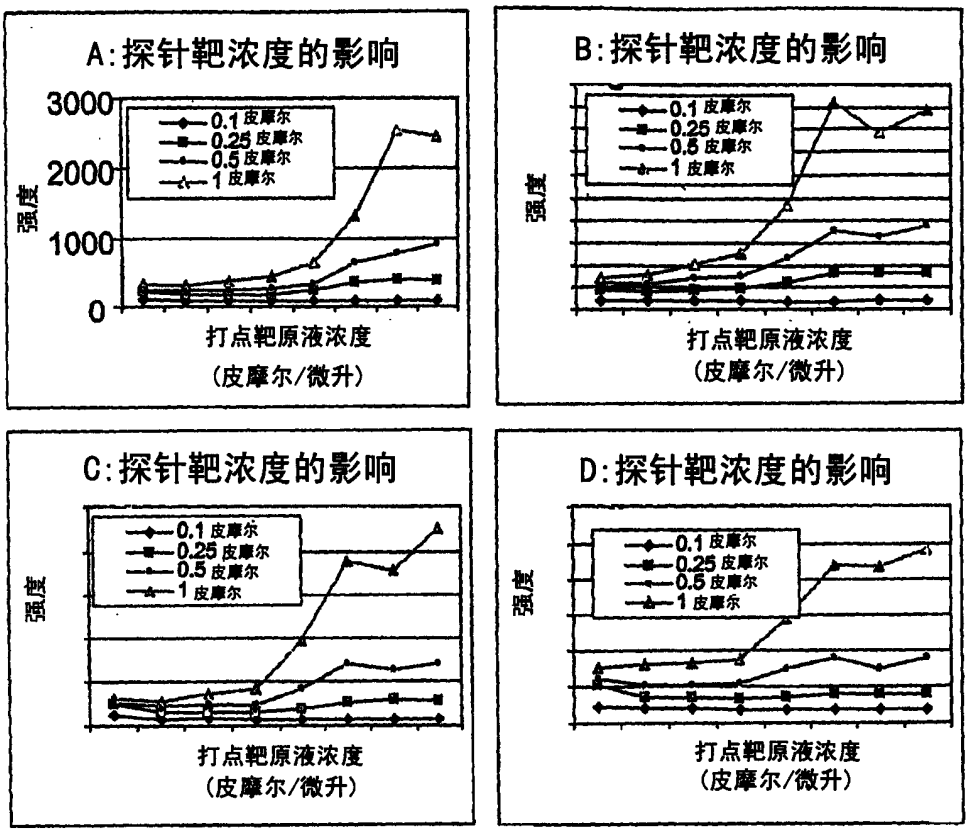

A:探针靶浓度的影响
3000
2000
1000
0
强度
0.1 皮摩尔
0.25 皮摩尔
0.5 皮摩尔
1 皮摩尔
打点靶原液浓度
(皮摩尔/微升)
B:探针靶浓度的影响
强度
0.1 皮摩尔
0.25 皮摩尔
0.5 皮摩尔
1 皮摩尔
打点靶原液浓度
(皮摩尔/微升)
C:探针靶浓度的影响
强度
0.1 皮摩尔
0.25 皮摩尔
0.5 皮摩尔
1 皮摩尔
打点靶原液浓度
(皮摩尔/微升)
D:探针靶浓度的影响
强度
0.1 皮摩尔
0.25 皮摩尔
0.5 皮摩尔
1 皮摩尔
打点靶原液浓度
(皮摩尔/微升)

图 12

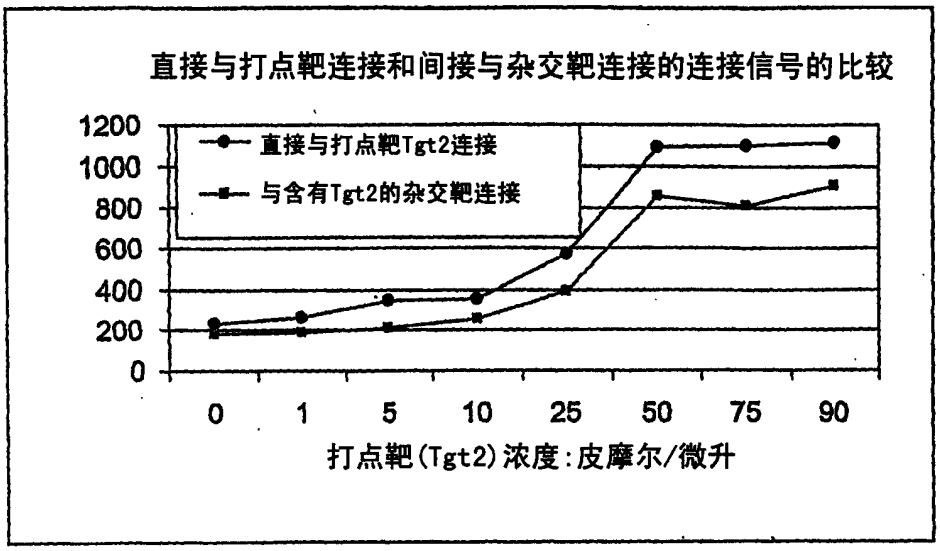

直接与打点靶连接和间接与杂交靶连接的连接信号的比较
1200
1000
800
600
400
200
0
直接与打点靶Tgt2连接
与含有Tgt2的杂交靶连接
0
1
5
10
25
50
75
90
打点靶(Tgt2)浓度:皮摩尔/微升

图 13