[19]中华人民共和国国家知识产权局

[51]Int. Cl[7]
C12Q 1/68

[12]发明专利申请公开说明书

[21]申请号 02119281.2

[43]公开日 2002年11月27日

[11]公开号 CN 1381591A

[22]申请日 2002.3.29 [21]申请号 02119281.2

[30]优先权

[32]2001.3.30 [33]US [31]60/280530

[32]2001.8.17 [33]US [31]60/313264

[32]2001.10.5 [33]US [31]60/327006

[32]2001.11.26 [33]US [31]60/332550

[71]申请人 珀尔根科学公司

地址 美国加利福尼亚州

[72]发明人 N·帕蒂尔 D·R·科克斯

A·J·贝尔诺

D·A·海因兹

[74]专利代理机构 中国专利代理(香港)有限公司

代理人 张广育 刘 玥

权利要求书6页 说明书52页 附图19页

[54]发明名称 基因组分析方法

[57]摘要

本发明涉及鉴别人类基因组发生的变异的方法,及将这些变异与疾病和药物反应的遗传基础相关联的方法。具体地说,本发明涉及鉴别个体SNPs、确定SNP单倍型区块和模式,和进一步利用SNP单倍型区块和模式来分析疾病和药物反应的遗传基础。本发明的方法适用于全基因组的分析。

ISSN1008-4274

1．一种选择SNP单倍型模式的方法，包括：

从大量不同来源中分离出基本等同的核酸链用于分析；

确定每条核酸链中一个以上的SNP位点；

鉴别在所述核酸链上连锁的SNP位点，其中所述连锁的SNP位点形成一个SNP单倍型区块；

鉴别分离的SNP单倍型区块；

鉴别存在于每个SNP单倍型区块和分离的SNP单倍型区块中的SNP单倍型模式；和

选择存在于至少两条所述不同来源的基本等同的核酸链中的每一种鉴定出的SNP单倍型模式。

2．权利要求1的方法，其中通过greedy算法或最短-路径算法来确定所述的第一个鉴别步骤。

3．权利要求1的方法，其中所述SNP单倍型区块是非重叠的。

4．权利要求1的方法，其中所述基本上等同的核酸链至少从约10到约100个不同的来源获得。

5．权利要求4的方法，其中所述基本上等同的核酸链至少从约16个不同来源获得。

6．权利要求5的方法，其中所述基本上等同的核酸链至少从约25个不同来源获得。

7．权利要求6的方法，其中所述基本上等同的核酸链至少从约50个不同来源获得。

8．权利要求1的方法，其中所述基本上等同的核酸链是基因组DNA链。

9．权利要求1的方法，其中分离并分析至少10%的来源于一个生物体的基因组DNA。

10．权利要求1的方法，其中分离并分析所述基本等同的核酸链中的至少1×10^8个碱基。

11．权利要求1的方法，其中对从基本等同的核酸链上选择出的重复区域不进行分析。

12．权利要求1的方法，进一步包括：

在所述确定步骤后，鉴别所述大量等同的核酸链中只出现一次的SNP位点；和

分析时排除所述出现一次的SNP位点。

13．权利要求1的方法，进一步包括：

选择一种在所述基本等同的核酸链中出现频率最高的SNP单倍型模式；和

选择一种在所述基本等同的核酸链中出现频率次高的SNP单倍型模式；和

重复所述第二个选择步骤直到所述选择的SNP单倍型模式鉴别了一部分所述的基本等同的核酸链。

14．权利要求13的方法，其中所述的一部分是所述基本等同核酸链的约70％到99％。

15．权利要求14的方法，其中所述的一部分是所述基本等同核酸链的至少约80％。

16．权利要求13的方法，其中只选择了不超过三种SNP单倍型模式。

17．一种选择用于数据分析的SNP单倍型区块数据组的方法，包括：

比较SNP单倍型区块的信息量；

选择高信息量的第一SNP单倍型区块；

将所述第一SNP单倍型区块加入到所述数据组中；

选择高信息量的第二SNP单倍型区块；

将选择的所述第二SNP单倍型区块加入到所述数据组中；和

重复所述选择和加入步骤直到覆盖了核酸链上感兴趣的一段区域。

18．权利要求17的方法，其中所选择的SNP单倍型区块是非重叠的。

19．权利要求17的方法，其中采用greedy算法进行所述的选择步骤。

20．一种用于确定SNP单倍型模式中的信息型SNP的方法，包括：

确定SNP单倍型区块中的SNP单倍型模式；

将所述SNP单倍型区块中每一感兴趣的SNP单倍型模式与所述SNP单倍型区块中其他感兴趣的SNP单倍型模式进行比较；

在感兴趣的第一SNP单倍型模式中选择至少一个SNP，该SNP可以把

所述 SNP 单倍型区块中这种感兴趣的第一 SNP 单倍型模式与其他感兴趣的 SNP 单倍型模式区别开来，其中所选择的至少一个 SNP 是在所述 SNP 单倍型区块中所述的第一 SNP 单倍型模式中的信息型 SNP。

21．权利要求 20 的方法，进一步包括重复所述选择步骤直到选择了足够数量的信息型 SNPs 来区分 SNP 单倍型区块中的一部分 SNP 单倍型模式。

22．权利要求 21 的方法，其中所选择的一部分 SNP 单倍型模式占所述 SNP 单倍型区块中 SNP 单倍型模式的大约 70%到大约 99%。

23．权利要求 21 的方法，其中所选择的一部分 SNP 单倍型模式允许感兴趣疾病的鉴别。

24．一种确定 SNP 单倍型区块的信息量的方法，包括：

确定在所述 SNP 单倍型区块中 SNP 位点的数目；

确定区分所述 SNP 单倍型区块中感兴趣的 SNP 单倍型模式所需要的信息型 SNP 的数目；

用所述 SNP 位点的数目除以所述信息型 SNP 的数目产生一个商，其中所述商就是所述 SNP 单倍型区块的信息量。

25．一种确定 SNP 单倍型区块的信息量的方法，包括：

确定在所述 SNP 单倍型区块中 SNP 位点的数目；

确定区分所述 SNP 单倍型区块中每一个感兴趣的 SNP 单倍型模式所需要的信息型 SNP 的数目，其中区别感兴趣的 SNP 单倍型模式所需要的所述信息型 SNP 数目即所述 SNP 单倍型区块的信息量。

26．一种在预先不知道疾病相关遗传位点的序列或位置时，确定疾病相关遗传位点的方法，包括：

确定一个对照群体中至少 16 个个体的 SNP 单倍型模式；

确定一个患病群体中个体的 SNP 单倍型模式；和

将所述对照群体中所述 SNP 单倍型模式的频率与所述患病群体中所述 SNP 单倍型模式的频率进行比较，其中所述频率的差异指示了疾病相关遗传位点的位置。

27．权利要求 26 的方法，其中所述 SNP 单倍型模式在对照群体中的至少 50 个个体中确定。

28．权利要求 26 的方法，其中所述群体的所述 SNP 单倍型模式通过使

用信息型 SNPs 来确定。

29．一种使用多种全基因组构建 SNP 单倍型区块图谱的方法：

将所述全基因组的至少大约 10%中发现的 SNPs 分配到 SNP 单倍型区块中。

30．一种将 SNP 单倍型模式和一种感兴趣的表型性状进行关联的方法，包括：

用本发明的方法建立 SNP 单倍型模式的基线；

汇集来源于一个具有感兴趣的常见表型性状的群体的全基因组 DNA；和

鉴别与所述感兴趣的表型性状相关联的所述 SNP 单倍型模式。

31．权利要求 30 的方法，其中信息型 SNPs 用于所述建立和鉴别步骤。

32．一种鉴别诊断标记的方法，包括：

鉴别权利要求 20 的信息型 SNPs，其中所述的信息型 SNPs 是基于关联的诊断标记。

33．一种鉴别药物发现靶的方法，包括：

将 SNP 单倍型模式与一种疾病相关联；

鉴别所述关联 SNP 单倍型模式上的一个染色体位点；

确定所述染色体位点和所述疾病之间关联的性质；和

选择与所述疾病相关联的一个染色体位点或染色体位点的一种表达产物；其中所选择的与所述疾病相关联的染色体位点或染色体位点的一种表达产物是药物发现靶。

34．权利要求 33 的方法，其中基于一系列标准而排列所述关联染色体位点作为药物发现靶的优先次序，所述位点包括处于高度保守区域的位点和基因间区域的位点。

35．权利要求 33 的方法，其中信息型 SNPs 用于所述的关联步骤。

36．一种确定个体 SNP 单倍型模式的方法，包括：

检验至少一个信息型 SNP。

37．一种确定一个物种或一个物种的亚型的 SNP 单倍型模式的方法，包括：

鉴别在所述物种的多个生物体基因组中出现的 SNPs；

通过重复选择具有少量模糊位置的 SNP 单倍型模式，将所述 SNPs 分配

到SNP单倍型区块中。

38．一种包含来源于多种生物体基因组的SNP单倍型区块的数据库，其中所述数据库鉴别至少一种信息型SNP并且其中所述数据库保存在一种计算机可读介质上。

39．一种存在于计算机可读介质上的数据库，含有鉴别为与一个或多个特定表型性状关联的SNP单倍型模式。

40．一种存在于计算机可读介质上数据库，含有鉴别为与一个或多个特定表型性状关联的信息型SNPs。

41．权利要求38、39或40的数据库，进一步包含关于选自环境因素、其他遗传因素、相关因素，包括但不限于生化标记、行为，和/或其他多态现象，包括但不限于低频率的SNPs，重复、插入和缺失的一个或多个因素的信息。

42．一种诊断疾病、疾病易感性或治疗反应的试剂盒，包括检测患者基因组DNA的样本中SNP单倍型模式或信息型SNPs存在与否，及检测存在于计算机可读介质上的所述SNP单倍型模式或信息型SNPs与一个或多个特定表型性状的关联的数据组的工具。

43．一种分离的核酸，包括至少一种信息型SNP，其中所述信息型SNP指示一种根据本发明方法确定的SNP单倍型模式，其中所述信息型SNP与表型性状相关。

44．一种方法，包括：

鉴定大量个体的遗传变异；

鉴别个体的至少某些所述遗传变异，这些变异还同时伴有至少某些其他所述遗传变异；和

使用某些但不是全部的所述变异，所述变异还同时伴有至少某些其他与一种表型状态相关的所述遗传变异。

45．一种方法，包括：

确定一种生物体的序列；

扫描所述生物体的其他个体的所述序列的变异体；

鉴别在第一个群体中与其他所述变异体同时出现的一些所述变异体；

鉴别在第二个群体中与其他所述变异体同时出现的一些所述变异体；和

使用某些但不是全部的与具有一种表型状态的群体相关的所述第一和第二个群体中的所述变异体。

46．一种选择用于基因组分析的SNP单倍型区块的方法，包括：

从至少约五种不同的来源分离一条基本上等同的DNA链用于分析；

分析至少约五种不同来源的每条所述基本上等同的DNA链上至少约1×10^6个碱基；

确定每条DNA链上一个以上的SNP位点；

在所述DNA链中鉴别连锁的SNP位点，其中所述连锁的SNP位点形成一个SNP单倍型区块；

鉴别存在于每一个SNP单倍型区块中的SNP单倍型模式；和

选择存在于不同来源的任何所述基本等同的DNA链中的每个鉴别出的SNP单倍型模式。

47．一种在预先不知道药物基因组相关遗传位点的序列或位置的情况下，确定所述药物基因组相关遗传位点的方法，包括：

确定一个对照群体中至少16个个体的SNP单倍型模式；

确定对给予一种物质以变化方式产生反应的个体的SNP单倍型模式；和

将所述对照群体的SNP单倍型模式的频率，与所述对给予一种物质以变化方式产生反应的个体的所述SNP单倍型模式的频率进行比较，其中所述频率的差异指示药物基因组相关的遗传位点的位置。

48．权利要求47的方法，其中所述SNP单倍型模式在对照群体的至少50个个体中确定。

49．权利要求47的方法，其中使用信息型SNP来确定所述群体中的所述SNP单倍型模式。

基因组分析方法

技术领域

本发明涉及鉴别人类基因组发生的变异的方法，及将这些变异与疾病和药物反应的遗传基础相关联的方法。具体地说，本发明涉及鉴别个体 SNPs、确定 SNP 单倍型区块和模式，和进一步利用 SNP 单倍型区块和模式来分析疾病和药物反应的遗传基础。本发明的方法适用于全基因组的分析。

相关申请的交叉参考

本申请要求以下申请的优先权：2001 年 3 月 30 日提交的美国临时专利申请系列号 60/280,530，2001 年 8 月 17 日提交的美国临时专利申请系列号 60/313,264，2001 年 10 月 5 日提交的美国临时专利申请系列号 60/327,006，名称均为“鉴别人类 SNP 单倍型、信息型 SNPs 及其应用”，和 11/26/01 提交的美国临时专利申请系列号 60/332,550，名称为“基因组分析方法”，所有这些公开文件特别引入此处作为参考。

背景技术

构成人染色体的 DNA 提供了指导人体合成所有蛋白质的指令。这些蛋白质执行生命的重要功能。编码蛋白质的 DNA 序列的变异使其编码的蛋白质发生变异或突变，从而影响细胞的正常功能。尽管环境在疾病中常常起着重要的作用，但是个体 DNA 的变异或突变直接与几乎人类的所有疾病相关，包括传染病、癌症和自身免疫紊乱。而且，遗传学知识，尤其是人类遗传学，已经引导人们认识到许多疾病是由几种基因或其产物间复杂的相互作用，或者一个基因内任何数量的基因突变引起的。例如 I 型和 II 型糖尿病与多种基因相关，每型都有其各自的突变模式。相比之下，囊性纤维化可由单一基因内 300 个以上不同突变之一引起。

而且，当涉及药物反应——药物遗传学领域时，人类遗传学知识使得人们对个体间变异的理解有限。半个多世纪前，药物副作用与两个药物代谢酶——血浆胆碱酯酶和葡萄糖-6-磷酸脱氢酶的氨基酸变异有关。从此，将 35 个以上药物代谢酶的序列多态性（变异）、25 种药物靶和 5 种药物转运蛋白

与药物功效或安全性的折衷水平联系起来，进行仔细的遗传分析（Evans and Relling, Science 296:487-91(1999)）。在临床上，这样的信息正用于预防药物毒性；例如，根据硫嘌呤甲基转移酶基因的遗传学差异对患者进行常规筛选，该酶降低使 6-巯嘌呤或咪唑硫嘌呤的代谢。然而迄今有效的药物遗传学标记组只能充分解释观察到的药物毒性中的一小部分。甚至比毒性问题更常见的可能是这种情况，对一些个体证明安全和/或有效的药物，对其他个体的药物功效不充分或产生了预料不到的副作用。

除了理解人类遗传组成变异的影响的重要性外，理解其它非人类生物体——尤其是病原体的遗传学组成的变异，对于理解它们对人类的影响以及与人类间的相互作用非常重要。例如，病原菌或病毒的毒力因子的表达大大影响与这些生物体发生联系的人类的感染比例和严重性。另外，详细理解实验动物如小鼠、大鼠等的遗传学组成也非常有价值。例如，理解用于估计治疗的模型系统的动物遗传学组成的变异对于理解采用这些系统所获得的实验结果和用于人类的预值是很重要的。

由于任何两个人的遗传学组成的相似度是 99.9%，他们基因组的大部分 DNA 序列都相同。然而，个体间的 DNA 序列存在变异。例如，DNA 多碱基段的缺失，DNA 片段的插入，非编码区重复 DNA 元件的数量变异和基因组内单一含氮碱基位点的改变，即所谓的“单核苷酸多态性”（SNPs）。人类 DNA 序列变异占观察到的个体间差异的一大部分，包括对疾病的易感性。

尽管多数 SNPs 是罕见的，已经估计有 5300 万个频率均在 10—50%的常见 SNPs，它们占人群间 DNA 序列差异的绝大部分。在人类基因组中，每 600 个碱基对就存在着这样的 SNPs（Kruglyak and Nickerson, Nature Genet. 27:235，2001）。在亲密的自然近亲中形成的这种 SNPs 区块的等位基因（变异）常常相互关联，从而降低了遗传变异性，并限定了有限数量的“SNP 单倍型”，每一个该 SNP 单倍型都反应了它由单一的远古祖先染色体继承而来(Fullerton, et al., Am. J. Hum. Genet. 67:881，2000）。

人类基因组中局部单倍型结构的复杂性--和体单倍型延伸的距离—很少被确定。调查不同人群的基因组不同片段的研究经验揭示了局部单倍型结构的巨大变异性。这些研究表明突变、重组、选择、人群历史和随机事件以不可预知模式对单倍型结构改变起着相应的贡献，结果导致一些单倍型延伸仅几

千碱基(kb)，而另一些单倍型延伸超过100kb(A. G. Clark ，et al.，Am. J. Hum. Genet. ，63:595，1998)。

这些发现表明由常见 SNPs 定义的人基因组单倍型结构的任何全面描述，将需要对人类基因组中大量独立拷贝中的整套 SNPs 进行深入的经验学分析。这种全基因组分析将提供一个细致的基因图谱并探明特殊连锁区域。然而，对合理大小人群的每个个体的超过 3,000,000 SNPs 进行分型的实践和所需费用使人们在本发明前无法将这种努力付诸实践。在多种申请中，本发明允许用 SNP 单倍型进行全基因组关联分析。

发明内容

本发明涉及鉴定人类基因组所发生的变异以及将这些变异与疾病抵抗力，疾病易感性或药物反应等表型的遗传基础相关联的方法。“疾病”包括但不局限于生物体中需要改变的任何条件、性状或特征。例如，这些条件可以是物理的、生理的或心理的，并且可以有症状或没有症状。该方法提供了变异的鉴定，SNPs 的鉴定，SNP 单倍型区块的确定、SNP 单倍型模式的确定，以及更进一步，每一种模式的信息型 SNPs 的鉴定，信息型 SNPs 可压缩遗传资料。

因此，本发明的一个方面提供了用于资料分析的选择 SNP 单倍型模式的方法。这样的选择是通过以下步骤完成的：从大量个体中分离出基本等同（同源的）的核酸链；确定每条核酸链上 SNP 的位置；鉴定核酸链中连锁的 SNP 位点，在此处连锁 SNP 位点形成 SNP 单倍型区块；鉴定分离的 SNP 单倍型区块；鉴定出现在每个 SNP 单倍型区块中的 SNP 单倍型模式；以及选择出现在至少两条基本等同核酸链上的已鉴定的 SNP 单倍型模式。在一个优选实施方案中，使用了来自至少约 10 个不同个体或来源的核酸链。在一个更优选的实施方案中，使用来自至少 16 个不同来源的核酸链。更加优选的实施方案中，使用来自至少 25 个不同来源的核酸链，还要更优选的实施方案中，使用来自至少 50 个不同来源的核酸链。进一步，更优选的实施方案将确定来自至少约 100 个不同来源的核酸链中的 SNP 位点。另外，该方法可进一步包括选择在基本等同的核酸链中最频繁出现的 SNP 单倍型模式；选择基本等同的核酸链中次频繁出现的 SNP 单倍型模式；并重复选择直到所选 SNP 单倍型模式能鉴定出基本等同核酸链中感兴趣的一部分。在一个优选实施方案中，感

兴趣的部分占基本等同核酸链的 70%到 99%，并且，更优选的实施方案中，感兴趣的部分占基本等同核酸链的约 80%。可替换地，可期望将 SNP 单倍型模式的选择限制到每个 SNP 单倍型区块中只有不超过约 3 个 SNP 单倍型模式。

另外，本发明提供了用于数据分析的选择 SNP 单倍型区块数据组的方法，包括比较 SNP 单倍型区块中的信息量；选择高信息量的第一个 SNP 单倍型区块；将第一个 SNP 单倍型区块添加到数据组中；选择高信息量的第二个 SNP 单倍型区块；将第二个 SNP 单倍型区块添加到数据组中；重复选择和添加的步骤，直到覆盖了感兴趣的 DNA 链。在优选实施方案中，所选 SNP 单倍型区块无重叠。

本发明进一步提供了在一个 SNP 单倍型模式中确定至少一个信息型 SNP 的方法，包括首先确定 SNP 单倍型区块的 SNP 单倍型模式，接着将 SNP 单倍型区块中每个感兴趣的 SNP 单倍型模式与 SNP 单倍型区块中其它 SNP 单倍型模式进行比较，并在每个 SNP 单倍型模式中选择至少一个 SNP，它可将该感兴趣的 SNP 单倍型模式与 SNP 单倍型区块中其它 SNP 单倍型模式区分开来。所选 SNP（或 SNPs）是该 SNP 单倍型模式的信息型 SNP。

而且，本发明可快速扫描基因组区域并提供了确定疾病相关遗传位点或药物基因组相关位点的方法，而不需要预先知道疾病相关遗传位点或药物基因组相关位点的序列或位置。可通过确定对照人群中个体的 SNP 单倍型模式，然后确定例如患病人群中的个体或当给予药物时发生特殊方式反应的个体等实验人群个体的 SNP 单倍型模式来完成这个过程。比较对照和实验人群 SNP 单倍型模式的频率。这些频率间的差异指出了疾病相关遗传位点或药物基因组相关位点。

本发明的另一方面提供了在 SNP 单倍型模式和感兴趣的表型性状间建立联系的方法，包括：用本发明的方法建立对照个体的 SNP 单倍型模式的基线；汇集有共同感兴趣的表型性状的临床人群的全基因组 DNA；和确定与感兴趣表型性状相关的 SNP 单倍型模式。这样，本发明提供了基因组扫描用于鉴定与表型相关的多单倍型区块，当研究多基因性状时这一点尤其有用。

而且，本发明提供了鉴定药物发现靶的方法，包括：将 SNP 单倍型模式与疾病联系起来；确定相关 SNP 单倍型模式在染色体上的位置；确定染色体

位置和所述疾病关系的本质；和用该染色体位置的基因或基因产物作为药物发现靶。

附图说明

下列附图构成本发明详述的一部分，为进一步证明本发明的特定方面而引入。参考其中一幅或更多附图，结合这里的特殊实施方案详述可更好地理解本发明。

图 1 是本发明方法的一个实施方案的示意图，表示鉴定变异位点，以便使变异与表型关联，并用这种关联确定药物发现靶或作为诊断标记。

图2 显示根据本发明的样本 SNP 单倍型区块和 SNP 单倍型模式。

图3 是表示选择 SNP 单倍型区块的方法的实施方案示意图。

图4 是图3 所示方法的实施方案的简单例子。

图 5A 是选择最终 SNP 单倍型区块组的方法的一种实施方案的示意图。图 5B 是图 5A 所示方法的简单例子。图 5B 中标记“字母：数字”表示每个区块的“单倍型区块 ID：信息量值”。

图6 表示根据本发明的一个实施方案，如何选择信息型 SNPs 的例子。

图 7A 是解决变异体模糊和/或 SNP 单倍型模式模糊的实施方案示意图。图7B 显示了图7A 所示方法的简单例子。

图8 是在关联研究中使用本发明方法的一个实施方案的示意图。

图9 表示适于执行本发明一些实施方案的示范性计算机网络系统。

图10 是体细胞杂交体构建示意图。

图11 列表显示用 Affymetrix, Inc.的 HuSNP 基因芯片筛选仓鼠—人细胞杂交体的部分结果。

图12 表示用长距离 PCR 扩增人 22 号染色体和人 14 号染色体得到的多种扩增基因组区的例子。

图 13A 是表示针对 SNP 次要等位基因（变异体）的频率而描绘的 SNP 百分比的条形图。图 13B 是作为间隔中核苷酸多样性功能的 200kb 间隔所占百分比的图。图 13C 是表示针对间隔长度而描绘的所有间隔百分比的条形图。

图 14 显示了由人 21 号染色体上 147 个常见 SNPs 确定的 20 个独立总体多样染色体的单倍型模式。

图15 是所覆盖染色体比例作为该覆盖所需 SNPs 数量的函数的曲线。

具体实施方式

对本领域技术人员来说，在本申请公开的发明基础上可进行多种实施方案和改进，而不脱离本发明的范围和精神。这里提到的所有公开物的引入是为了描述和公开可能用到的与本发明相联系的试剂、方法学和概念。这些参考资料都不作为与此处所述的本发明相关的现有技术的证明。

说明书中使用的“一个”是指一个或更多。权利要求中使用的与单词“包括”相连时，单词“一个”是指一个或更多。其中使用的“另一个”是指至少第二个或更多。

当使用“不同起源”时，是指不同生物体的 DNA 链来自不同的起源。进一步，单个生物体的基因组中的每一 DNA 链来自不同的起源。在二倍体生物中，个体生物的基因组由一套成对的基本等同的 DNA 链组成。也就是说，单一个体将具有来自两个不同来源的基本等同的 DNA 链——两条中的一条 DNA 链来自母源，一条 DNA 链来自父源。认为两个或更多核酸序列——例如，两条或更多 DNA 链——是基本相同的，如果它们在核苷酸水平有至少约 70%的序列等同性，优选约 75%，更优选约 80%，更优选约 85%，更优选约 90%，甚至更优选约 95%，甚至更优选如果核酸序列在核苷酸水平表现出至少约 98%的序列等同性，认为核酸序列基本相同。两条或更多核酸序列间的相关的序列等同性程度将依赖于该核酸的宿主来源。例如，当进行相同物种比较时，超过 95%的序列等同性可能相关，而做交叉物种比较时，70%或甚至更少的序列等同性可能相关。当然，当涉及此处所述的 DNA，可包括 DNA 的衍生物，如扩增子、RNA 转录物、核酸模拟物等。

此处所述“个体”是指一个特定单个生物体，如单个动物、人、昆虫、细菌等。

此处所述 SNP 单倍型区块的“信息量”定义 SNP 单倍型区块提供有关基因区域信息的程度。

此处所述“信息型 SNP”是指 SNP 或 SNPs 亚型（一个以上）的遗传性变异体，倾向于将一个 SNP 单倍型模式与一个 SNP 单倍型区块中其它 SNP 单倍型模式区别开来。

此处所述的术语“分离的 SNP 区块”是指由一个 SNP 组成的 SNP 单倍型区块。

此处所述的术语“连锁不平衡”、“连锁”或“LD”是指倾向于从一代到一代一起传递的遗传学位点；如非随机遗传的遗传学位点。

此处所述的术语“单一 SNP 单倍型”或“单一 SNP”是指出现在少于人群一定部分的特异性 SNP 等位基因或变异体。

此处所述的术语“SNP”或“单核苷酸多态性”是指个体间遗传学变异；如在生物体中 DNA 的可变单个含氮碱基位点。此处所述的“SNPs”是 SNP 的复数。当然，当涉及此处所述的 DNA，可包括 DNA 的衍生物，如扩增子、RNA 转录物等。

此处所述的术语“SNP 单倍型区块”是指一组不单独出现重组，在变异体或 SNPs 中能聚集在一起的变异体或 SNP 位点。

此处所述的术语“SNP 单倍型模式”是指 DNA 单链上一个 SNP 单倍型区块中一系列的 SNPs 基因型。

此处所述的术语“SNP 位点”是一条 DNA 序列中出现 SNP 的位点。

此处所述的术语“SNP 单倍型序列”是含有至少一个 SNP 位点的 DNA 链中的 DNA 序列。

用于分析的核酸的制备

用本领域技术人员所知的任何技术制备核酸分子用于分析。优选这种技术可产生足够纯的核酸分子，以确定该核酸分子的一个或更多位点的一个或更多突变的存在与否。这种技术可从，如 sambrook, et al., Molecular Cloning: A Laboratory Manual(Cold Spring Harbor Laboratory, New York)(1989), 和 Ausubel, et al., Current Protocols in Molecular Biology(John Wiley and Sons, New York)(1997)中找到，此处引入作为参考。

当细胞中存在感兴趣的核酸，需要首先制备细胞提取物，接着进行下列步骤——即，示差沉淀、柱层析、用有机溶剂等提取——以获得足够纯的核酸制剂。可采用本领域的标准技术如细胞的化学或机械裂解制备提取物。可进一步处理提取物，如采用过滤和/或离心和/或用离液盐如异硫氰酸胍或尿素，或用有机溶剂如苯酚和/或 HCCl3 使任何污染物和潜在干扰蛋白变性。当使用离液盐时，期望将含核酸样品中的该盐去除。这可通过使用本领域的标准技术来完成，如沉淀、过滤、大小排阻层析等。

在一些情况下，可能期望从细胞提取和分离信使 RNA。本领域技术人员

知道为这一目的所用的技术和材料，可包括使用附着于固相支持物如珠或塑料表面的寡dT。本领域技术人员知道合适的条件和材料，并可从上述Sambrook和Ausubel的参考文献中找到。可能需要使用如逆转录酶将mRNA逆转录为cDNA。适当的酶可从商业途径获得，如Invitrogen，Carlsbad CA。然后可以选择性地扩增由mRNA制备的cDNA。

一个尤其适合检查单倍型模式和单倍型区块的方法是用体细胞遗传学将染色体从二倍体状态分离为单倍体状态。在一个实施方案中，二倍体人类淋巴母细胞系可融入也是二倍体的仓鼠成纤维细胞系，这样人染色体被导入仓鼠细胞产生了细胞杂交体。检查所得杂交细胞，确定人哪一染色体被转移，以及，如果存在，可以确定哪一被转移的人染色体是单倍体状态（见，如，Patterson, et al., Annal.N.Y. Acad. Of Sciences, 396：69-81，1982）。

图10显示了步骤流程图。图10显示了胸苷激酶基因为野生型的二倍体人类淋巴母细胞系与含有胸苷激酶基因突变的二倍体仓鼠成纤维细胞系融合。所得细胞的亚群中，人染色体存在于杂交细胞中。使用HAT培养基（筛选培养基）筛选含人DNA的杂交细胞。只有稳定掺入了含野生型人胸苷激酶基因的人DNA链的杂交细胞能在含HAT的细胞培养基中生长。在所得杂交细胞中，一些细胞可能含有一些人染色体的两个拷贝，仅含人染色体的一个拷贝或没有特定人染色体的拷贝。例如，对于含有一个A或一个B等位基因座的人22号染色体，所得杂交细胞可能含有一个人22号染色体变异体（如“A”变异体）或其中一部分，一些可能含有另一种人22号染色体变异体（如“B”变异体）或其中一部分，一些可能含有这两种人22号染色体变异体或其中一部分，还有一些杂交细胞可能根本不含人22号染色体的任何部分。在图10中，只显示了所得杂交细胞群中的两个。一旦选择出合适的杂交细胞，可用如上述技术分离这些杂交细胞中的核酸，然后进行SNP检测，以及本发明的单倍型区块和单倍型模式分析。

扩增技术

在确定核酸中一个和更多变异的存在与否之前，可能需要扩增一个或更多感兴趣的核酸。核酸扩增增加了感兴趣核酸序列的拷贝数量。本领域技术人员所知的任何扩增技术可与本发明联用，包括但不限于，聚合酶链式反应（PCR）技术。PCR可使用本领域技术人员所知的材料和方法进行。

PCR 扩增通常包括使用核酸序列的一条链作为模板，以产生大量的该序列的互补序列。该模板可杂交于具有与模板序列的一部分互补的序列的引物，并接触合适的反应混合物，包括 dNTPs 和一种聚合酶。聚合酶使引物延伸产生原始模板的互补核酸。

为了扩增双链核酸分子的两条链，可使用两个引物，每个引物可含有与一条核酸链的一部分互补的序列。聚合酶作用下引物的延伸产生两个双链核酸分子，每个分子含有一条模板链和一条新合成的互补链。典型的引物序列选择使每个引物的延伸向着核酸分子中另一引物杂交的位点延长。

变性核酸分子链——例如，通过加热——重复这一过程，这段时间里，将在先步骤新合成的链作为随后步骤的模板。PCR 扩增实施方案可包括几个到多个变性、杂交和延伸反应的循环，以产生足够数量的期望核酸。

尽管典型 PCR 方法采用加热使链变性，允许随后的引物杂交，但可使用其它任何使核酸可以与引物杂交的方法。这种技术包括但不限于，物理、化学和酶学方法，如通过引入解旋酶，（见 Radding, Ann. Rev. Genetics 16:405-436(1982)）或用电化学方法（见 PCR 申请 Nos. WO92/04470 和 WO95/25177）。

PCR 中引物的模板依赖性延伸是在反应介质中至少 4 种脱氧核苷三磷酸（典型地选自 dATP，dGTP，dCTP，dUTP 和 dTTP）存在下，由聚合酶催化，反应介质中含有合适的盐、金属阳离子和 pH 缓冲系统。适当的聚合酶是本领域技术人员知道的，可以是克隆的或从自然来源分离，还可以是天然或突变形式的酶。只要酶保持延伸引物的能力，它们均可用于本发明的扩增反应。

本发明的方法中所用的核酸可被标记以利于后续步骤的检测。标记可在扩增反应过程中进行，将一个或更多标记过的三磷酸核苷酸和/或一个或更多标记的引物掺入到被扩增序列中。核酸可在扩增后进行标记，如通过一个或更多可检测基团的共价连接。可使用任何本领域技术人员所知的可检测基团，如荧光基团、配体和/或放射性基团。一个合适的标记技术是用末端脱氧核苷酸转移酶（TdT）将含标记的核苷酸掺入到感兴趣的核酸中。例如，将一个含有标记的核苷酸——优选双脱氧核苷酸——与待标记的核酸和足够量的将要掺入核苷酸的 TdT 一起孵育。优选核苷酸是附着生物素标记的双脱氧核苷酸——即 ddATP，ddGTP，ddCTP，ddTTP 等。

可使用优化长序列扩增的技术。这些技术对基因组序列很奏效。这些方

法在以下文献中公开：未决的 2001 年 9 月 5 日提交的 US 专利申请 USSN60/317,311；USSN（未授予申请号），代理案卷号 1011N-1，2002 年 1 月 9 日提交的名称为“选择引物对的运算法则”的申请；和 USSN（已授予申请号），代理案卷号 1011N1D1，2002 年 1 月 9 日提交的名称为“核酸扩增方法”的申请，这些方法尤其适合用在本发明的方法中扩增基因组 DNA。

扩增序列可在标记前或标记后进行其它扩增后处理。例如，在一些情况下，期望在与寡核苷酸阵列杂交前将扩增序列分成片段。一般可用本领域技术人员所知的物理、化学或酶学方法对核酸进行分段。合适的技术包括但不限于，迫使含核酸的液体样本通过狭窄的缝隙或用核酸酶消化 PCR 产物，给扩增核酸一个剪切力。适当的核酸酶的一个例子是脱氧核糖核酸酶 I。扩增后，可在核酸酶存在下孵育 PCR 产物一段时间，以产生合适长度的片段。片段的长度可根据期望而改变，例如，通过增加核酸酶的量或孵育持续时间以产生较小片段，或通过降低核酸酶的量或孵育持续时间产生较大片段。调整消化条件以产生期望大小的片段在本领域普通技术人员的能力范围内。然后可用上述方法标记所产生的片段。

检测 SNPs 的方法（SNP 发现）

可用任何本领域技术人员所公知的技术来确定核酸中一个或更多变异的存在和缺失。可使用任何允许精确确定变异的技术。优选技术允许使用需要最少样本处理的方法快速、精确确定多个变异。下面提供了一些适当技术的例子。

DNA 测序的几个方法是本领域技术人员熟知的并通常可获得的，可用于确定基因组中 SNPs 的位置。参见，如 Sambrook，et al.， Molecular Cloning: A Laboratory Manual(Cold Spring Harbor Laboratory，New York)(1989)和 Ausubel，et al.， Current Protocols in Molecular Biology(John Wiley and Sons，New York)(1997)，此处引入作为参考。这些方法可用于确定来自不同 DNA 链的相同基因区域的序列，比较该区域的序列，记录差异（链间的变异）。DNA 测序方法可使用这样的酶，如 DNA 聚合酶 I 的克列诺片段，测序酶（US Biochemical Corp, Cleveland, Ohio.），Taq 聚合酶（Perkin Elmer），耐热 T7 聚合酶（Amersham, Chicago, I11.），或联合使用聚合酶和校正核酸外切酶，如那些以 Gibco/BRL（Gaithersburg, Md.）为商标的延伸扩增系统中的酶。优选

地，用机器自动进行这个过程，如 Hamilton Micro Lab 2200（Hamilton，Reno，Nev.），Peltier 热循环仪（PTC200; MJ Research, Watertown, Mass.）和 ABI 催化剂和 373 和 377DNA 测序仪（Perkin Elmer，Wellesley，MA）。

另外，可用商业上可获得的毛细管电泳体系进行变异或 SNP 分析。尤其是，毛细管测序可使用电泳分离的可流动聚合体，激光激活的 4 种不同的荧光染料（每种染一种核苷酸），用电荷偶联装置照相机检测散发的波长。输出/光强度可用适当的软件（如基因型和序列导航器，Perkin Elmer，Wellesley，MA）转换为电信号，从载入样本到计算机分析和电子数据显示的整个过程可由计算机控制。再者，该方法可用于确定来自不同 DNA 链的相同基因组区域的序列，比较该区域的序列，记录差异（链间的变异）。

可选择地，一旦测序确定了来自一个参照 DNA 链的基因组序列，就可能使用杂交技术确定参照链和其它 DNA 链间的序列变异。这些变异可能是 SNPs。一个适当的杂交技术的例子包括使用 DNA 芯片（寡核苷酸阵列），如那些可从 Affymetrix, Inc. Santa Clara, CA.获得的 DNA 芯片。使用 DNA 芯片检测如 SNPs 的详细情况见授权给 Lipshultz 等的美国专利 No. 6,300,063 和授权给 Chee 等的美国专利 No.5,837,832，HuSNP 图谱方法，试剂盒和使用手册，Affymetrix 部分 No. 90094(Affymetrix, Santa Clara, CA)，在这里均引入作为参考。

在优选实施方案中，扫描参照序列和另一条 DNA 链的 10,000 多个碱基的变异。在更优选实施方案中，扫描参照序列和另一条 DNA 链的超过 1×10^6 个碱基的变异，甚至更优选扫描参照序列和另一条 DNA 链的超过 2×10^6 个碱基，更优选扫描 1×10^7 个碱基，更优选扫描超过 1×10^8 个碱基，更优选扫描参照序列和另一条 DNA 链的超过 1×10^9 个碱基的变异。在优选实施方案中，至少扫描外显子的变异，在更优选实施方案中，扫描内含子和外显子的变异。在更加优选实施方案中，扫描内含子、外显子和基因间序列的变异。在优选实施方案中，被扫描的核酸是基因组 DNA，包括编码区和非编码区。最优选实施方案中，这种 DNA 来自哺乳动物，如人类。在优选实施方案中，扫描来自生物体的 10%以上的基因组 DNA，在更优选实施方案中，扫描来自生物体的 25%以上的基因组 DNA，在更优选实施方案中，扫描来自生物体的 50%以上的基因组 DNA，在最优选实施方案中，扫描来自生物体的 75

%以上的基因组 DNA。在本发明的一些实施方案中，不扫描基因组中已知的重复区，并且不计入被扫描的基因组 DNA 的百分比。这种已知的重复区可以包括单一散布核元件（SINEs, 如 alu 和 MIR 序列），长散布核元件（LINEs, 如 LINE1 和 LINE2 序列），长末端重复（LTRs，如 MaLRs，Retrov 和 MER4 序列），转座子和 MER1 和 MER2 序列。

简言之，在一个实施方案中，采用——如加热到 95℃——使适当溶液中标记过的核酸变性，含变性核酸的溶液与 DNA 芯片共同孵育。孵育后，移去溶液，可用适当的洗涤液洗涤芯片以去除未杂交核酸，检测芯片上杂交核酸的存在。洗涤条件的严格性可根据需要调节，以产生稳定的信号。检测杂交核酸可直接进行，例如如果核酸含有荧光报告基团，荧光可以直接检测到。如果核酸上的标记不能直接检测到，例如，生物素，则可以在检测前加入含可检测标记，如，偶联于藻红素的链霉亲和素的溶液。其它可增强信号水平的试剂也可于检测前加入，例如，可以联合使用链霉亲和素特异性生物素化抗体和生物素、链霉素和素一藻红素检测体系。在一些实施方案中，本发明的方法中使用的寡核苷酸阵列中每个阵列含有至少 1×10^6 个探针。在优选实施方案中，本发明的方法中使用的寡核苷酸阵列中每个阵列含有至少 10×10^6 个探针。在更优选的实施方案中，本发明的方法中使用的寡核苷酸阵列中每个阵列含有至少 50×10^6 个探针。

一旦使用，如测序或微阵列分析法确定了变异位点（SNP 发现），有必要对对照和样品群的 SNPs 进行基因分型。刚才描述的杂交方法对这一目的很奏效，它提供了一种对多个样品中的 SNPs 进行精确和快速检测与基因分型的技术。此外，适于检测基因组 DNA 中 SNPs 的技术——无扩增——是 Invader 技术，可从 Third Wave Technologies, Inc., Madison , WI 得到。此技术用于检测 SNPs 的用途可在如 Hessner, et al., Clinical Chemistry 46(8):1051-56(2000)；Hall, et al.，PNAS 97(15):8272-77(2000)；Agarwal, et al.，Diag. Molec. Path. 9(3):158-64(2000)和Cooksey, et al., Antimicrobial and Chemotherapy 44(5):1296-1301(2000)中发现。在 Invader 步骤中，两个短 DNA 探针与靶核酸杂交形成一种可被核酸酶识别的结构。对于 SNP 分析，进行两个分离的反应——每一个检测一个 SNP 变异体。如果探针之一与该序列互补，核酶将对其切割以释放称为“瓣（flap）”的短 DNA 片段。瓣与荧光标记探针结合并形成另一种可被核酸酶

识别的结构。当酶切割标记探针时，探针散发出一种可检测的荧光信号，由此表示存在 SNP 变异体。

Invader 技术的可替换方法，滚环扩增使用与环状 DNA 模板互补的寡核苷酸，以产生扩增信号（见如，Lizardi, et al., Nature Genetics 19(3):225-32(1998); 和 Zhong, et al., PNAS 98(7):3940-45(2001)）。寡核苷酸延伸在长多联体内产生环状模板的多个拷贝。典型地，在延伸反应中，可检测标记掺入到延伸的寡核苷酸。延伸反应允许延续进行直到合成的延伸产物量可被检测到。

为了用滚环扩增检测 SNPs，使用三个探针和两个环状 DNA 模板。第一个探针—靶特异性探针—可构建为与靶核酸分子互补，使探针的 5'末端与靶核酸中紧邻 5'SNP 位点的核苷酸杂交。SNP 位点不是与第一探针配对的碱基。

其它两个探针——滚环探针——构建为含有两个 3'末端。这可通过多种方法完成，如在探针的中心部分引入一个 5'-5'键，使核苷酸序列的极性在那个点发生反转。每个探针的一个末端具有与不同的环状模板分子的一部分互补的序列，而另一个末端与靶核酸序列的一部分互补。靶序列互补末端构建为能使最 3'端的核苷酸与在 SNP 位点的核苷酸对比。探针之一可以包含与靶核苷酸 SNP 位点的核苷酸互补的核苷酸，而另一个包含非互补的核苷酸。在群体中出现两个或更多 SNP 变异体的情况下，探针可构建为含有与待检变异体互补的 3'-核苷酸。

探针——二者均是靶特异性和滚环探针——可与靶序列杂交并与连接酶接触。当滚环探针的最 3'端核苷酸与 SNP 位点的核苷酸形成碱基对时，这两个探针——靶特异性和滚环探针——被有效连接在一起。当滚环探针的最 3'端核苷酸不能与靶核酸的 SNP 位点形成碱基配对时，探针将不被连接在一起。未连接的探针被洗掉，样品接触模板循环、聚合酶和被标记核苷三磷酸。

另一个适于检测 SNPs 的技术是利用 DNA 聚合酶的 5'-核酸外切酶活性，通过消化探针分子而释放荧光标记核苷酸以产生信号。该方法常常指 Taqman 方法（如参见 Arnold, et al., BioTechniques 25(1):98-106(1998);和 Becker, et al., Hum. Gene Ther. 10:2559-66(1999)）。在一个与 SNP 位点杂交的探针分子存在下，含 SNP 的靶 DNA 被扩增。这个探针分子的 5'末端含有荧光报告分子标记的核苷酸，3'末端含有猝灭剂标记的核苷酸。选择探针序列，使探针中与靶 DNA 的 SNP 位点对比的核苷酸尽可能接近探针的中心，使正确配对探针

和错配探针间的解链温度差异最大。当进行 PCR 反应时，正确配对的探针与靶 DNA 的 SNP 位点杂交，并被 PCR 方法中使用的 Taq 聚合酶消化。消化使荧光标记核苷酸与猝灭剂物理分离，伴随荧光增加。在 PCR 反应的延伸过程期间，错配探针不再继续杂交，并且，因此不再被消化，荧光标记核苷酸继续保持猝灭。

用聚苯乙烯－二乙烯基苯反相柱的变性 HPLC 和离子配对活动相可以鉴定 SNPs。含有 SNP 的 DNA 片段被 PCR 扩增。扩增后，加热使 PCR 产物变性，并与 SNP 位点的核苷酸已知的第二变性 PCR 产物混合。PCR 产物被退火并在升温阶段用 HPLC 分析。温度选择为使在 SNP 位点错配的双链分子变性，而不使那些很好配对的分子变性。在这些条件下，异源双链分子一般早于同源双链分子被洗脱。使用这一技术的例子见 Kota, et al., Genome 44(4):523-28(2001)。

可用固相扩增和扩增产物微量测序来检测 SNPs。采用已经与引物共价结合的珠进行扩增反应。设计引物包括 II 型限制性内切酶的识别位点。扩增后产生与珠结合的 PCR 产物，用限制性酶消化产物。限制性酶切割产物产生一个包括 SNP 位点的单链部分和一个可延伸填充该单链部分的 3'-OH。在延伸反应中加入 ddNTPs 允许对产物直接测序。使用该技术鉴定 SNPs 的例子见 Shapero, et al., Genome Research 11(11):1926-34(2001)。

数据分析

图 1 显示本发明方法的一个实施方案的步骤示意图。一旦 SNPs（变异体）已经通过如 supra 描述的方法（图 1 的步骤 110）被定位或发现，SNP 单倍型区块，每个 SNP 单倍型区块的 SNP 单倍型模式和 SNP 单倍型模式的信息型 SNPs 即可被确定。可使用所有已定位的 SNPs 或变异体；可替换地，可仅集中分析已定位 SNPs 的一部分。例如，被分析的一组 SNPs 可排除 SNPs 转换形式 Cg<->Tg 或 cG<->cA。此外，在本发明的一个实施方案中，集中分析常见 SNPs。常见 SNPs 是那些在给定群体中其较不常见形式出现频率很低的 SNPs。例如，常见 SNPs 是那些在至少约 2%到 25%的群体中发现的 SNPs。优选实施方案中，常见 SNPs 是那些在至少约 5%到 15%的群体中发现的 SNPs。更优选实施方案中，常见 SNPs 是那些在至少约 10%的群体中发现的 SNPs。常见 SNPs 可能源于人进化早期的突变。集中分析常见 SNPs 缩小了

对照和实验人群间的系统等位基因或变体差异，这种差异看来与疾病或药物反应相关，但仅由迁移历史或杂交行为所致；即，集中分析常见 SNPs 降低现代人群异常造成的假阳性。而且，常见 SNPs 与人群中的更大部分相关，使本发明可更广泛地应用于疾病和药物反应研究。沿着相同线索，仅观察到一次变异的 SNPs 可从本发明的一些实施方案的分析中排除（例如单一 SNPs）。然而，以下的某些分析仍然可以进行，包括这些单一 SNPs 的一些或全部，尤其是当研究特异性子人群或已受迁移行为等影响的群体时。

在图 1 的步骤 120，将感兴趣的变异体或 SNPs 分配到单倍型区块用于估计。可对来自整个基因组或染色体的变异体或 SNPs 进行分析并将其分配到 SNP 单倍型区块。可替换地，可将仅仅来自于所关注的对一些疾病或药物反应机制呈特异性的基因组区域的变异体分配到 SNP 单倍型区块。

图 2 提供了在基因组单倍型区块中如何发生变异（通常是 SNPs）的图解，以及在每个单倍型区块中如何能产生一个以上的单倍型模式。如果 SNP 单倍型模式是完全随机的，预期 N 个 SNPs 的 SNP 单倍型区块中可观察到的 SNP 单倍型模式的数量将是 2^N。然而，观察到在执行本发明的方法时，每个 SNP 单倍型区块中的 SNP 单倍型模式数量少于 2^N，因为 SNPs 是连锁的（不是 4^N，由于变异体最常是双等位，如，仅发生两种形式之一，而不是所有可能的四种核苷酸碱基）。观察到某些 SNP 单倍型模式的频率远远高于在非连锁例子中期望出现的频率。因而，SNP 单倍型区块是倾向于作为一个单位而遗传的染色体区域，常见模式的数量相对较少。图 2 中的每一行代表不同个体的单倍体基因组序列的部分。如此处所示，个体 W 在 241 位有一个“A”， 242 位有一个“G”， 243 位有一个“A”。个体 X 在 241、242 和 243 位有相同的碱基。相反，个体 Y 在 241 和 243 位有一个 T，但在 242 位有一个 A。个体 Z 在 241、242 和 243 位与个体 Y 有相同的碱基。261 区块变异体倾向于一起发生。相似地，262 区块的变异体倾向于一起发生，区块 263 的那些突变体也是。当然，图 2 仅显示了基因组中的几个碱基。事实上，许多碱基与 245 和 248 位的碱基类似，个体与个体间不发生变化。

在一个例子中，分配 SNPs 到 SNP 单倍型区块，即图 1 的步骤 120，是重复的过程，包括从 SNP 位点沿着感兴趣的基因区域构建 SNP 单倍型区块。在一个实施方案中，一旦构建了初始 SNP 单倍型区块，即可确定构建的 SNP

单倍型区块中存在的SNP单倍型模式（图1的步骤130）。在一些特殊实施方案中，步骤130中，每个SNP单倍型区块中所选SNP单倍型模式数量不超过约5个。在另一个特殊实施方案中，每个SNP单倍型区块中所选SNP单倍型模式数量，等于鉴定所分析的50%以上DNA链中的SNP单倍型模式所需的SNP单倍型模式数量。换句话说，选择足够的SNP单倍型模式，例如，每个区块选择四个模式，使所分析的DNA链中至少一半有一个SNP单倍型模式与每个SNP单倍型区块中所选的四个模式之一匹配。在一个优选实施方案中，每个SNP单倍型区块中所选SNP单倍型模式数量，等于鉴定所分析的70%以上DNA链中的SNP单倍型模式所需的SNP单倍型模式数量。在一个优选实施方案中，每个SNP单倍型区块中所选SNP单倍型模式数量，等于鉴定所分析的80%以上DNA链中的SNP单倍型模式所需的SNP单倍型模式数量。此外，在本发明的一些实施方案中，将少于一定比例的所分析的DNA链中存在的SNP单倍型模式从分析中排除。例如，在一个实施方案中，如果分析10条DNA链，发现在10个样本中只有一个存在某种SNP单倍型模式，则将其从分析中排除。

一旦选择出感兴趣的SNP单倍型模式，即可确定这些SNP单倍型模式的信息型SNPs（图1的140步）。从初始的单倍型区块，可构建一组满足特定信息量标准的候选SNP区块（图1的150步）。图4和5显示了120、130、140和150步的更详细情况。

在图3中，步骤310提供了新的SNPs区块的评估。在一个实施方案中，所选的第一区块仅含有SNP单倍型序列的第一个SNP；这样在320步，将第一个单一的SNP加入该区区块。在330步，确定该区块的信息量。

在一个实施方案中，一个SNP单倍型区块的“信息量”定义为该区块提供有关基因区域信息的程度。例如，在本发明的一个实施方案中，信息量的计算可以用一个SNP单倍型区块中SNP位点数量除以该区块中为了区别每个考虑的SNP单倍型模式与其它考虑的SNP单倍型模式（信息型SNPs的数量）所需的SNPs数量。另一种信息测量方法可以是该区块中信息型SNPs的数量。本领域技术人员知道信息量可用任何数量的方法来测定。

再次提到图2，SNP单倍型区块261含有三个SNPs和两个SNP单倍型模式（AGA和TAT）。三个存在的SNPs的任何一个可用于分别表达这些模式；

因而，可选择这些 SNPs 中的任何一个作为这个 SNP 单倍型模式的信息型 SNP。例如，如果确定了一个样本核酸在第一位置含有一个 T，同一样本将在第二位置含有一个 A，在第三位置含有一个 T。如果确定了第二个样本的第二个位置的 SNP 是 G，第一和第三 SNPs 将是 A。因而，根据信息量的一个测定方法，第一区块的信息量值是 3：即 3 个总 SNPs 除以区分各模式所需要的 1 个信息型 SNP。相似地，SNP 单倍型区块 262 含有三个 SNPs(两个位点无变异)和两个单倍型模式（TCG 和 CAC)。在先前分析的区块中，可估计三个 SNPs 中任何一个来区分一个模式与另一个之间的差异；因此，该区块的信息量是 3：即 3 个总 SNPs 除以区分各模式所需要的 1 个信息型 SNP。SNP 单倍型区块 263 含有五个 SNPs 和两个 SNP 模式（TAACG 和 ATCAC)。同样，五个 SNPs 的任何一个都可用于区别一个模式与另一个模式；因而，该区块的信息量是 5：即 5 个总 SNPs 除以区分各模式所需要的 1 个信息型 SNP。

图 2 提供了一个遗传学分析的简单例子。当一个 SNP 单倍型区块中存在几个 SNP 单倍型模式时，可能需要使用一个以上的 SNP 作为信息型 SNPs。例如，在一个例子中，一个 SNP 单倍型区块中含有，如六个 SNPs，两个 SNPs 是区分感兴趣模式所需的，该区块的信息量是 3：即 6 个总 SNPs 除以区分各模式所需的 2 个 SNPs。一般地说，多至 2^N 个不同的 SNP 单倍型模式可用 N 个适当选择的 SNPs 的基因型来区分。因此，如果在 SNP 单倍型区块中仅存在两个 SNP 单倍型模式，单一 SNP 应该能够将二者区分开来。如果有三个或四个模式，可能需要至少两个 SNPs，等。

在图 3 的步骤 340 中，一旦 SNP 单倍型区块的信息量确定，即可进行一个测验。该测验主要是基于选择的标准估计 SNP 单倍型区块（例如，一个单倍型区块是否满足信息量的阈测量值)，测验的结果决定是否，例如，向该区块中加入另一个 SNP 进行分析或是否用新的单倍型区块从不同的 SNP 位点开始进行分析。图 4 图解该过程的一个实施方案。

在图 4 中，假定有一个含有六个 SNP 位点的 DNA 序列。上述 SNP 单倍型区块的分析可以下述方式进行：SNP 单倍型区块 A 选择仅在 SNP 位置 1 含 SNP（图 3 的 310 步和 320 步)。计算该区块的信息量（330 步)，确定该区块的信息量是否满足信息型阈测定值（340 步)。在这种情况下，它“通过”并发生了两件事。一，一个 SNP 的单倍型区块（SNP 位置 1）被加入候选 SNP

单倍型区块组中（350 步）。二，另一个 SNP（这里，SNP 位置 2）被加入该单倍型区块（320 步）以创建一个新的单倍型区块，B，含有 SNP 位置 1 和 2，然后再分析。在这个例子中，B 区块也满足信息量的阈测量值（340 步），因此它将被加入候选 SNP 单倍型区块组中（350 步），另一个 SNP（这里，SNP 位置 3）被加入该单倍型区块（320 步）以创建一个新的单倍型区块 C，含有 SNP 位置 1、2 和 3，然后再分析。在这个例子中，C 区块也满足信息量的阈测量值（340 步），并被加入候选 SNP 单倍型区块组中（350 步），另一个 SNP（这里，SNP 位置 4）加入该单倍型区块（320 步）以创建一个新的区块 D，含有 SNP 位置 1、2、3 和 4，然后再分析。在这个图 4 的图解中，SNP 区块 D 不满足信息量的阈测量值。不将 SNP 区块 D 加入候选 SNP 单倍型区块组中（350 步），也不将另一个 SNP 加入区块 D 进行分析。而是选择一个新的 SNP 位点进行一轮 SNP 区块评估。

在图 4 中，单倍型区块 D 不满足信息量的阈测量值时，选择一个新的区块 E，仅在位置 2 含有 SNP。估计区块 E 的信息量，发现满足信息量的阈测量值，将其加入候选 SNP 单倍型区块组中（350 步），将另一个 SNP（这里，SNP 位置 3）加入该单倍型区块（320 步）以创建一个新的区块 F，含有 SNP 位置 2 和 3，然后再分析，等等。注意区块 H 不满足信息量的阈测量值，不将其加入候选 SNP 单倍型区块组中（350 步），也不将另一个 SNP 加入区块 H 进行分析。而是选择仅在位置 3 含有 SNP 的一个新的区块 I，等等。

一旦构建了一组候选 SNP 区块（图 3 的 350 步），对其进行分析，选择最终的 SNP 区块组（图 1 的 160 步）。可采用多种方法选择最终的 SNP 区块组。例如，回到图 4，可选择含 SNP 位置 1 的最大 SNP 区块，它通过了阈检测值（C 区块，含 SNPs1，2 和 3），放弃含相同 SNPs 的较小的区块（A 和 B 区块）。然后下一个被选区块可能是起始于 SNP 位置 4 的下一个区块，它是满足信息量的阈测量值的最大的区块（G 区块），放弃含相同 SNPs 的较小的区块（E 和 F 区块）。这种方法可得到一组跨越感兴趣基因组区域的无重叠最终 SNP 单倍型区块，它含有感兴趣的信息量水平高的 SNPs。这样，一旦评估了所有候选 SNP 单倍型区块，在一个优选实施方案中，结果将是一组涵盖了原始组所有 SNPs 的无重叠 SNP 单倍型区块。一些组中，称为分离组，可能仅由单一 SNP 组成，根据定义，其信息量为 1。其它组可能由一百或更多 SNPs

组成，含有30个以上信息量。

图5A和5B显示了一种筛选最终SNP单倍型区块组的替代方法。首先看图5A，在第一步510，分析候选SNP单倍型区块组（例如，由这里的图3和图4描述的方法得到）的信息量。在520步，选择完全候选组中信息量最高的候选SNP单倍型区块，加入到最终SNP单倍型区块组（530步）。一旦这个SNP单倍型区块被选为最终SNP单倍型区块组的成员，就从候选单倍型区块组中删除（540步），将与已选区块重叠的所有其它候选SNP单倍型区块从候选SNP单倍型区块组中删除（550步）。下一步，分析候选组中剩余的候选SNP单倍型区块的信息量（510步），选择信息量最高的候选SNP单倍型区块，加入最终SNP单倍型区块组（520和530步）。同前，一旦这个SNP单倍型区块被选为最终SNP单倍型区块组的成员，就从候选单倍型区块组中删除(540步)，将与已选区块重叠的所有其它候选SNP单倍型区块从候选SNP单倍型区块组中删除(550步)。这个过程持续到构建了一组最终的无重叠SNP单倍型区块，这些区块涵盖了原始组的所有SNPs。

图5B图示了图5A中所述的筛选最终SNP单倍型区块的方法的简单过程。在图5B中，根据本发明的方法分析5'到3'序列的SNPs，SNP单倍型模式和候选SNP单倍型区块。这个序列中所包含的候选SNP单倍型区块由其在序列下的位置指示，用一个字母指示。此外，在字母后，指出每一区块的信息量。例如，候选SNP单倍型区块A位于序列的最5'末端，信息量是1。候选SNP单倍型区块R位于序列的最3'末端，信息量是2。

根据图5A，在第一步510中，分析候选SNP单倍型区块的信息量，在520步中，选择信息量最高的SNP单倍型区块，加入到最终SNP单倍型区块组（520和530步）。在图5B的例子中，信息量为6的候选SNP单倍型区块M将是第一个被选择加入最终SNP单倍型区块组的候选SNP单倍型区块。一旦选择了SNP单倍型区块M，即从候选SNP单倍型区块中删除或去除（540步），所有其它与SNP单倍型区块M重叠的候选SNP单倍型区块（J，N，K，L，O和P区块）都从候选SNP单倍型区块组中删除（550步）。下一步，分析候选SNP单倍型区块组中剩余区块，即单倍型区块A、B、C、D、E、F、G、H、I、Q和R的信息量，在520步，信息量最高的剩余SNP单倍型区块，I，信息量是5，被选择加入最终SNP单倍型区块组（530），并从候选SNP

单倍型区块中删除或去除（540 步）。下一步，在 550 步，将与 SNP 单倍型区块 I 重叠的所有其它候选 SNP 单倍型区块，在这里，只有 H 区块，从候选 SNP 单倍型区块组中删除。再次分析候选 SNP 单倍型区块组中剩余 SNP 单倍型区块，即单倍型区块 A、B、C、E、F、G、Q 和 R 的信息量。在 520 步中，信息量最高的剩余 SNP 单倍型区块 F，信息量是 4，被选择加入最终 SNP 单倍型区块组（530），并从候选 SNP 单倍型区块中删除或去除（540 步）。下一步，与 SNP 单倍型区块 F 重叠的所有其它的候选 SNP 单倍型区块一这里是，E、G、C 和 D 区块一从候选 SNP 单倍型区块组中被删除，分析候选 SNP 单倍型区块组中剩余区块，即 SNP 单倍型区块 A,B,Q 和 R 的信息量，等等。

可使用其它方法从候选 SNP 单倍型区块组中选择用于分析的最终 SNP 单倍型区块组（图 1 的 160 步）。例如，为了这个目的可使用本领域知晓的计算法则。例如，使用最短路径计算法则（通常见，Cormen, Leiserson, and Rivest, Introduction to Algorithms(MIT Press)pp. 514-78(1994).）在最短路径问题中，给出了一个带有权重函数 w：E → R 的加权的、定向的曲线图 G=（V,E），权重函数将各边映射为真实权重。路径 P=（$v_0,v_1,\ldots v_k$）的权重是其组成边加权值的总和：

$$w(p) = \sum_{i=1}^{k} w(v_{i-1}, v_i).$$

如果存在 u 到 v 的路径，从 u 到 v 的最短路径权重由δ（u，v）定义，等于最小 w（p）：u → v；否则，δ（u，v）等于无穷大。接着顶点 u 到顶点 v 的最短路径由权重为 w（p）= δ (u,v) 的任何路径 p 定义。边权重可解释为多种度量标准：例如，距离、时间、价值、惩罚、损失或其它任何沿路径线性蓄积的期望最小的数量。本发明一个实施方案中所用的最短路径计算法则，每个 SNP 单倍型区块被认为是“顶点”，该单倍型区块的每个边界定义为“边”。每个 SNP 单倍型区块与每个其它 SNP 单倍型区块存在联系，每个边有“价值”。价值由选择的参数决定，如最高点的重叠（或其程度）或最高点间的距离。

单源最短路径问题集中在给定的图 G=（V,E），确定给定的源顶点 s ∈ V 到每个顶点 v ∈ V 的最短路径。而且，可应用多种单源的计算法则。例如，

可应用单一终点的最短路径解法，可发现从每个顶点 v 到给定终顶点 t 的最短路径。逆转图中每个边的方向使问题减为单源问题。可替换地，可应用单对最短路径问题，发现对于给定顶点 u 和 v，从 u 到 v 到最短路径。如果源顶点 u 的单源问题解决了，单源最短路径问题也就解决了。而且，可使用所有成对最短路径方法。在这种情况下，发现对于每对顶点 u 和 v，从 u 到 v 的最短路径——单源运算法则从每个顶点开始进行。

本发明方法可使用的一个单源最短路径运算法则是 Dijkstra's 运算法则。Dijkstra's 运算法则在加权的、定向的曲线图 G = （V，E）上解决了所有边权重是非负数情况下的单源最短路径问题。Dijkstra's 运算法则保留一组顶点 S，它们从源 s 的最终最短路径权重已经确定了。即，对所有顶点 v 是元件 S，d[v]= δ(s,v)。运算法则重复选择顶点 u 作为最短路径估计值最小的 V—S 的元件，，将 u 插入 S，释放所有由 u 发出的边。在执行过程中，保留包含 V—S 中所有顶点的优先队列 Q，以 d 值为基准。执行时，假定图 G 由邻接列表表示。

```
Dijkstra (G, w, s)
1    INITIALIZE-SINGLE SOURCE (G,s)
2    S ← Ø
3    Q ← V[G]
4    while Q ≠ Ø
5    do u ← EXTRACT-MIN (Q)
6    S ← S U {u}
7    for each vertex v ∈ Adj[u]
8    do RELAX (u,v,w)
```

这样，本例中的 G 是所分析的基因组序列的线性覆盖图，S 是选择的顶点组。一旦选择了一个覆盖基因组序列特定区域的顶点，可放弃其它与该序列重叠的顶点。

可采用其它运算法则来筛选 SNP 单倍型区块,包括 greedy 运算法则（还参见 Cormen, Leiserson, and Rivest, Introduction to Algorithms(MIT Press)pp. 329-55(1994).）。greedy 运算法则通过选择序列得到了解决问题的最佳方法。对于运算法则中每个决定点，选择此时看来最好的选择。这种启发式的策略不是总能产生最佳的解决方法。greedy 运算法则不同于动态编程中的运算法则，在动态编程中，每步都作出选择，但选择可依赖于子问题的解决方法。在 greedy 运算法则中，选择任何此时看来最好的选择，接着解决选择后产生的子问题。

因而，greedy 运算法则作出的选择依赖于到此为止作出的选择，但不依赖于任何未来的选择或子问题的解决方法。greedy 运算法则的一个可变方法是 Huffman 代码。Huffman greedy 运算法则构建了最佳前缀码，该法则建立了对应于自底向上方式中最佳代码的 T 树。它起始于一组|C|叶，执行一系列|C|-1 "合并"序列操作，以产生终树。例如，假定 C 是一组 n 个字符，每个字符 $c \in C$ 是具有确定频率 f[c]的对象；用以 f 为关键码的优先队列 Q 鉴定两个频率最低的合并在一起的对象。两个对象合并的结果是一个新的对象，其频率是被合并的这两个对象频率之和。例如：

```
1.  n←|C|
2.  Q←C
3.  for i←1 to n-1
4.  do z←ALLOCATE-NODE()
5.  x←left[z] ←EXTRACT-MIN(Q)
6.  y←right[z] ←EXTRACT-MIN(Q)
7.  f[z] ←f[x] + f[y]
8.  INSERT (Q,z)
9.  return EXTRACT-MIN(Q)
```

第 2 行用 C 中的字符初始化优先队列 Q。第 3-8 行的 for 循环重复提取队列中最低频率的两个节点 x 和 y，并在队列中用新节点 z 替代它们，表示它们的合并。在第 7 行中，Z 的频率计算为 x 和 y 频率的总和。节点 z 以 x 作为其左侧后代，y 作为其右侧后代。在 n-1 个合并之后，队列中左侧的一个节点——代码树的根——在第 9 行中被返回。

同样，这些方法产生了一组最终无重叠的 SNP 单倍型区块，它涵盖了在特定基因组区域中评估的所有 SNPs。根据本发明的方法，筛选 SNPs、SNP 单倍型区块和 SNP 单倍型模式的一个重要结果是，在一些实施方案中，计算 SNP 单倍型区块的信息量的过程中，可确定每个 SNP 单倍型区块和模式的信息型 SNPs。信息型 SNPs 允许压缩数据。在本发明的一个实施方案中，从每个含 p 模式的组中选择至少 $\log_2 p$ 个 SNPs（近似到最近的整数），提供了一组信息型 SNPs，它对于预示基因型/表型关联非常有利。本领域技术人员知道在其它分析中，不需要使用空间邻近组来确定这样的亚组。例如，在本发明的一些实施方案中，可能期望鉴定非邻近 SNPs 组，它们从统计学角度是以类似于 SNP 单倍型区块的方式传递，即使它们在 DNA 链中不是空间邻近的。

为了确定将精确用于关联研究的 SNP 单倍型区块（建立一个精确的 SNPs 和 SNP 单倍型区块和模式的基线），需要检测几个以上的个体 DNA 链。图 6 图示了检测至少约 5 个不同的 DNA 链对于确定 SNP 单倍型区块以及选择信息型 SNPs 的重要性。图 6 的顶部图解了 DNA 的假定片段序列，指出了变异位点并画出了变异区块边界；然而，开始不知道 SNP 单倍型区块边界。如图所示，通常在相对少量个体测序后，可能确定出大部分常见 SNPs。在图 6 的例子中，图 6 顶部显示的每个位点的 SNPs 已经被鉴定，由画勾的标记表示。

然而，如果不评价更多的个体，不可能在这个阶段正确鉴定区块的边界。例如，尽管在这个阶段可以画出 620 和 630 区块间的边界（注意第一个 C → G 变异预示第一个 G → A 变异，第一个 C → T 变异预示第二个 C → T 变异），在这个阶段不可能区分区块 630 和 640。在此阶段，看来第一个 C → T 变异将预示第一个和第二个 T → A 变异。因此，需要更多的统计学显著性样本组来绘出区块的边界。例如，在本发明的方法中，确定 SNP 单倍型区块、SNP 单倍型模式，和/或信息型 SNPs 所分析的 DNA 链的数量是大数目，例如至少约 5 或至少约 10。在优选实施方案中，分析的 DNA 链的数量至少是 16。在更优选实施方案中，分析确定 SNP 单倍型区块、SNP 单倍型模式，和/或信息型 SNPs 的 DNA 链的数量至少是 25。然而，一旦相关 SNPs 已经确定（即已进行了 SNP 发现），就可能仅对剩余样本中的变异位点进行基因分型，在不对完整基因组 DNA 片段进行测序的情况下，完成鉴定单倍型区块边界的过程。这种方法的例子见 2002 年 1 月 6 日申请的 USSN10/042,819，代理案卷号是 1016N-1，名称为“全基因组扫描”。

在图 6 中 650 显示了仅对另一个假定基因组样本中的 SNPs 进行基因分型处理的结果。如图所示，通过这个额外的分型步骤，现在可能看到 630 和 640 单倍型区块可区分开来。特别是现在可能看到第一个 C → T 变异不与第一个和第二个 T → A 变异相伴，而是，第一个 C → T 变异仅可用于预示第二个 C → T 变异（反之亦然），以及第一个 T → A 变异仅可用于预示第二个 T → A 变异（反之亦然）。

除了上述本发明的方面，本发明一个特殊的实施方案是它能用于解决模糊 SNP 单倍型序列的数据分析问题。例如，一个 SNP 可能是模糊的，因为凝胶测序操作资料或阵列杂交实验给出的结果不清晰。这种情况下，“解决”

可以指，如通过将 SNP 单倍型序列与该 SNP 单倍型序列关系最密切的 SNP 单倍型模式匹配，来解决 SNP 单倍型区块中的模糊 SNP 位点问题。此外，“解决”可以指从数据分析中去除模糊 SNP 单倍型序列。

解决模糊 SNP 单倍型序列的一个实施方案中，为了可能加入到模式组，将 SNP 单倍型序列放在数据组中。为了可能分配到 SNP 单倍型模式中，数据组将含有所有将要评价的 SNP 单倍型序列。现在提到图 7A，在 710 步，将数据组中的 SNP 单倍型序列与模式组中的模式序列进行，逐一比较。在一些情况下，开始时在模式组中没有模式，尽管在其它情况下，一些或全部模式序列已经预先知道。在 720 步，提出一个问题：来自数据组的 SNP 单倍型序列与模式组中的模式序列一致吗？如果答案是否定的，730 步提供了正在评估的 SNP 单倍型序列将加入到模式组中。如果答案是肯定的，则提出另一个问题：来自数据组的 SNP 单倍型序列是否与模式组中一个以上的模式序列一致？

如果答案是肯定的，来自数据组的 SNP 序列可能被丢弃，或者，在一些实施方案中，保留作进一步或不同的分析（750 步）。如果第二个问题的答案是否定的，那么，在 760 步中，将来自数据组的 SNP 序列与模式组中与其一致的模式序列比较。从这两个序列中，选择模糊数量最小的 SNP 序列并放到模式组中（770）。含模糊序列较多的 SNP 序列可能被丢弃，或，在一些实施方案中，保留作进一步或不同类型的分析。

参考图 7A 和 7B 可进一步理解解决过程。在图 7B 中，第一个 SNP 序列，TTCGA，与模式组中所含序列进行比较（710 步）。在这一点，在模式组中不含有模式序列，这样 TTCGA 与模式组中任何模式序列都不一致。接着从数据组中去除 SNP 序列 TTCGA 的出现（或保留作不同的分析），并加入到模式组（730）。现在模式组中有了一个模式序列，TTCGA。

再看图 7B，数据组中的第二个 SNP 序列，T?C??，与模式组中所含序列进行比较（710 步）。T?C??与序列一致（720 步）。当此时模式组中只有一个模式序列，TTCGA 时，第二个问题：SNP 序列 T?C??是否与模式组中的一个以上的模式序列一致？的答案（740）是否定的，因为现在模式组中只有一个模式序列，TTCGA。在 760 步中，T?C??与 TTCGA 比较，确定哪个序列更模糊。很清楚是 T?C??；因而，TTCGA 被留在模式组中（770），T?C??可

能被去除或保留作进一步分析。

图 7B 中数据组中的第三个序列是 C????。C????首先与 TTCGA 比较（710 步），发现与 TTCGA 不一致（720），然后加入到模式组（730）。图 7B 中的第四个序列是 CTACA。CTACA 与 TTCGA 和 C????比较（模式组中的模式序列，710 步），发现与 C????一致（720）。现在提出第二个问题（740）：CTACA 与 C????和 TTCGA 都一致吗？答案是否定的，因此接着比较 C????和 CTACA（760），以及，在这种情况下，模糊数量最少的序列，CTACA 保留在模式组中，C????被丢弃（从分析中去除），或保留作进一步分析（770）。

图 7B 数据组中的第五个序列是?T??A。将这个 SNP 序列与模式序列 TTCGA 和 CTACA（710）比较，发现与 TTCGA 和 CTACA 都一致。因而，问题 740 的答案是肯定的：?T??A 与模式组中一个以上的模式序列一致。在 750 步中，SNP 序列?T??A 保留作进一步的分析或丢弃（从分析中去除）。解决问题的另外的方法允许这种情况，比如，如果模式序列是 CCATT?，数据组的 SNP 序列是 C?ATTG，那么序列“联合”解决模糊问题（CCATTG），“联合”序列加入到模式组中。另外的阵列杂交，可使用测序或本领域所知的其它技术分析模糊 SNP 核苷酸位点。

SNP 单倍型区块和模式与表型的关联

鉴定出的 SNP 单倍型区块，SNP 单倍型模式和/或信息型 SNPs 可用于各种遗传分析。例如，一旦鉴定了信息型 SNPs，它们可用于大量不同的关联研究实验。例如，探针可设计为询问这些信息型 SNPs 的微阵列。其它作为示范的测定包括，如，前面描述的 Taqman 测定和 Invader 测定，及常规 PCR 和/或测序技术。

在一些实施方案中，如图 1 的 170 步所示，鉴定的单倍型模式可用于上面提到的测定，进行关联研究。通过确定带有感兴趣表型的个体（例如，表现出特殊疾病的个体或对特殊给药方式有反应的个体）的单倍型模式，并比较这些个体的单倍型模式频率和对照组个体的单倍型模式频率，可完成这种研究。这种 SNP 单倍型模式的鉴定优选是全基因组范围的；然而可以仅仅是基因组中感兴趣的特殊区域，及使用的这些特殊区域的 SNP 单倍型模式。除了这里公开的本发明的方法中的其它实施方案，这些方法额外提供表型“剖析”。也就是说，一个特殊表型可能由两个或更多不同的遗传碱基产生。例如，一

个个体的肥胖可能是 X 基因缺陷的结果，而不同个体中的肥胖可能是基因 Y 和基因 Z 突变的结果。因而，本发明的基因组扫描能力允许剖析相似表型的不同遗传碱基。一旦确定基因组的特殊区域与一个特殊的表型相关，这些区域可用作药物发现靶（图 1 的 180 步）或作为诊断标记（图 1 的 190 步）。

正如前段所述，进行关联研究的一个方法，是将带有感兴趣表型的个体的 SNP 单倍型模式频率与对照组个体的 SNP 单倍型模式频率相比较。在优选方法中，用信息型 SNPs 进行 SNP 单倍型模式比较。使用信息型 SNPs 的方法与迄今本领域所知的其它整个基因扫描或基因分型方法相比有极大的优势，不是阅读每个个体基因组的全部 30 亿个碱基—或甚至不阅读可能发现的 3—4 百万个常见 SNPs—仅需确定一个样本群的信息型 SNPs。正如上所述，阅读这些特殊的信息型 SNPs，提供了足够的信息，可从特殊实验群中提取统计学精确的关联资料。

图 8 显示了用本发明的方法确定遗传关联的方法的实施方案。在 800 步中，确定对照人群的基因组中信息型 SNPs 的频率。在 810 步，确定临床人群的基因组中信息型 SNPs 的频率。800 和 810 步可采用前述 SNP 测定而分析个体群中的信息型 SNPs。在 820 步中，比较 800 步和 810 步得出的信息 SNP 频率。可采用，例如，确定每群的每个信息型 SNP 位点中最小等位基因频率（带有特殊最小等位基因的个体数量除以全部个体数量）并比较这些最小等位基因频率的方法，进行频率比较。在 830 步中，选择在对照与临床人群发生频率间显示差异的信息型 SNPs 进行分析。一旦选择了信息型 SNPs，即可鉴定出含信息型 SNPs 的 SNP 单倍型区块，反过来可确定感兴趣的基因组区域（840 步）。用本领域所知的遗传学或生物学方法分析基因组区域（850 步），分析这个区域用作药物发现靶（860 步）或作为诊断标记（870 步）的可能性，如下详述。

鉴定基因组序列的用途

一旦基因组中遗传学位点或多个位点与一个特殊表型性状相关——如，疾病易感性位点——即可确定与该性状相关的一个或多个基因或调控元件。这样，这些基因或调控元件可被用作治疗该疾病的治疗靶，如图 1 的 180 步或图 8 的 860 步所示。用本发明的方法确定的基因组序列可以是基因或非基因序列。术语“基因”意指编码特殊多肽的开放阅读框架（ORF）、内含子区域，

及5'和3'的非编码核苷酸序列，它参与调节编码区上游最多到约10kb的基因表达，但可能进一步调节任意一个方向。由于鉴定到的基因的ORFs对蛋白结构的影响，它可影响疾病状态。可替换地，鉴定到的基因或非基因序列的非编码区可以通过影响蛋白表达的水平或表达的特异性而影响疾病状态。一般地讲，通过分离对基本不含其它核酸序列（其它核酸序列不包括基因序列）的鉴定基因而进行基因组序列的研究。DNA序列可用于多种途径。例如，DNA可用于检测或定量生物学样本的基因表达。文献中已经很好地建立了探测细胞中存在特殊核苷酸序列的方式，这里不需要再详述，然而，仍参见，如，Sambrook, et al., Molecular Cloing: A Laboratory Manual(Cold Spring Harbor Labortory, New York)(1989)。

此外，基因的序列，包括启动子区域和编码区侧翼，可发生本领域所知的各种方式的突变，在表达水平引起定向改变，或编码蛋白的序列发生改变，等等。序列改变可以是置换、插入、易位或缺失。缺失可包括大的改变，如完整区域或外显子缺失。克隆基因的体外诱变技术是已知的。位点特异性诱变方案的例子可在Gustin, et al., Biotechniques 14:22(1993); Barany, Gene 37:111-23(1985); Colicelli, et al., Mol. Gen. Genet. 199:537-9(1985); Prentki, et al., Gene 29:303-13(1984); Sambrook, et al., Molecular Cloning: A Laboratory Manual(Cold Spring Harbor Press)pp. 15.3-15.108(1989); Weiner, et al., Gene 126:35-41(1993); Sayers, et al., Biotechniques 13:592-6(1992); Jones and Winistorfer, Biotechniques 12:528-30(1992); 和Barton, et al., Nucleic Acids Res. 18:7349-55(1990)中找到。这种突变基因可用于研究蛋白产物的结构/功能之间的关系，或可用于改变影响蛋白的功能或调节的蛋白特性。

可用鉴定的基因来产生相应多肽的全部或部分。为了表达蛋白产物，使用一个掺入鉴定基因的表达盒。表达盒或载体通常提供一个转录和翻译起始区，可能是诱导型或组成型，编码区在转录起始区，转录和翻译终止区的转录控制下与其可操作性连接。这些控制区可以是鉴定基因的天然形式，或可以源自外源性来源。

肽可按照常规方法在原核或真核中表达，取决于表达的目的。为了大规模生产蛋白，单细胞生物体，如大肠杆菌、枯草芽孢杆菌、啤酒酵母、与杆状病毒载体组合的昆虫细胞，或较高级生物，如脊椎动物，尤其是哺乳动物的

细胞，如 COS7 细胞，可用作表达的宿主细胞。在许多情况下，期望在真核细胞表达基因，真核细胞将利于基因的天然折叠和翻译后修饰。小肽也可在实验室合成。在大量蛋白或其片段存在的情况下，可按照常规方法分离和纯化该蛋白。可制备表达宿主的裂解物，除了层析、凝胶电泳、亲和层析或其它纯化技术，可用 HPLC 纯化蛋白或其片段。

表达蛋白可用于生产抗体，短片段诱导特定多肽特异性抗体（单克隆抗体）的表达，以及大片段和完整蛋白允许产生超过多肽长度的抗体（多克隆抗体）。用常规方法制备抗体，表达的多肽或蛋白自身或与已知免疫原性载体，如 KLH,pre-S HbsAg，其它病毒或真核蛋白等结合作为免疫原。可使用多种佐剂，进行一系列适当的注射。对于单克隆抗体，一次或更多加强注射后，分离脾脏，通过细胞融合使淋巴细胞无限增殖，筛选高亲和抗体结合。然后可以展平无限增殖细胞，即产生所需抗体的杂交瘤。进一步的描述，见 Monoclonal Antibodies: A Laboratory Manual, Harlow and Lane, eds.(Cold Spring Harbor Laboratories,Cold Spring Harbor, N.Y.)(1988)。如果需要，可分离编码重链和轻链的 mRNA，在大肠杆菌中克隆诱变，混合重链和轻链，进一步增强抗体的亲和性。作为培养抗体的方法的体内免疫法的替换方法包括与噬菌体“显示”文库结合，常常联合体外亲和成熟。

鉴定出的基因、基因片段或编码蛋白或蛋白片段可能在基因治疗退行性和其它疾病中有用。例如，可用表达载体将鉴定基因导入细胞。这种载体通常在启动子序列附近具有方便的限制性位点，可在受体基因组中插入核酸序列。制备的转录盒可包括转录起始区域，靶基因或其片段，以及转录终止区域。转录盒可导入各种载体，如，质粒、逆转录病毒，如，慢病毒属；腺病毒等，在这里载体能够暂时或稳定存在于细胞中。该基因或蛋白产物可采用任何数量的途径，包括病毒转染、显微注射或小泡融合直接导入组织或宿主细胞。肌内给药可以采取喷射注射的方式，如 Furth, et al., Anal. Biochem,205:365-68(1992)所述。可替换地，如文献中所述，DNA 可包被到金微粒上，通过粒子轰击装置或“基因枪”将其皮内运送，（如见 Tang, et al., Nature,356:152-54(1992)）。

反义分子可用于下调细胞中鉴定到的基因的表达。反义试剂可能是反义寡核苷酸，尤其是有化学修饰的合成反义寡核苷酸，或能表达这种反义分子如

RNA 的核酸构建体。可联合给予反义分子，这里联合可包括多种不同的序列。

作为反义抑制剂的替换物质，催化性核酸化合物，如，核酶、反义缀合物等，可用于抑制基因表达。核酶可在体外合成并给予患者，或可编码在表达载体上，核酶在靶细胞中从表达载体合成（例如，见国际专利申请 WO9523225，和 Beigelman, et al., Nucl. Acids Res.23:4434-42(1995)）。具有催化活性的寡核苷酸例子在 WO9506764 中有述。反义寡核苷酸与金属复合物如三联吡啶铜（II）的缀合物，能够介导 mRNA 水解，在 Bashkin, et al., Appl. Biochem. Biotechnol.54:43-56(1995)中有述。

除了用鉴定的序列进行基因治疗，鉴定的核酸可用于产生遗传学修饰过的非人类动物从而构建疾病的动物模型，或在细胞系中产生位点特异性基因修饰，用于研究蛋白功能或调节。术语“转基因”意欲涵盖遗传学修饰的动物，它含有可在宿主细胞中稳定传递的外源基因，如基因序列可以改变，以产生一种修饰蛋白，或可以是与外源启动子可操作性连接的一个报告基因。转基因动物可通过同源重组得到，此时内源基因位点被改变、取代或破坏。可替换地，核酸构建体可能被随机整合到基因组中。稳定整合的载体包括质粒、逆转录病毒和其它动物病毒、YACs 等。感兴趣的转基因哺乳动物，如，牛、猪、山羊、马等，以及，尤其是啮齿动物，如，大鼠、小鼠等。

基因功能研究也可使用非哺乳动物模型，尤其是用那些生物学和遗传学已经被很好地区分的生物体，如雅致枝孢，D.melanogaster 和啤酒酵母。目的基因序列可用于剔除相应的基因功能，或弥补确定的基因病变，以确定参与蛋白功能的生理学和生物化学途径。药物筛选可与弥补和剔除研究联合进行，如，研究退行性疾病的进展、检测疗效，或药物发现。

此外，修饰过的细胞或动物在蛋白功能和调节的研究中有用。例如，可以在鉴别基因中进行一系列小的缺失和/或替代而确定不同酶活性结构域、细胞转移或定位等的作用。感兴趣的特异性结构包括但不局限于阻断基因表达的反义结构、显性失活基因突变的表达和鉴别基因的过度表达。也可以提供非正常表达细胞或组织中或在发育反常时期中鉴别基因或其变异体的表达。另外，通过提供细胞中非正常产生蛋白的表达，可以引起细胞行为的改变，这种行为改变可提供关于蛋白正常功能的信息。

可以通过测定蛋白分子推测其结构/功能参数。例如，通过提供大量鉴别

基因的蛋白产物，就可以鉴别与该蛋白产物结合、调节或模拟该蛋白产物活性的配体或底物。通过药物筛选，可以鉴别出，如那些在受影响细胞中能导致蛋白功能发生替换或增强的试剂，或是调节或灭活蛋白功能的试剂。这里使用的术语“试剂”代表了任何分子，如能直接或间接改变、模拟或掩饰一个鉴别基因或基因产物生理功能的蛋白或小分子。通常用多数对应于不同试剂浓度的测定混合物进行试验，以获得对不同浓度的差异性反应。一般用这些浓度中的一种作为阴性对照，即浓度为零或处于检测水平之下的浓度。

大量不同的测定法可用于此目的，包括标记的体外蛋白——蛋白结合测定、蛋白—DNA 结合测定、电泳迁移率转换测定、蛋白结合免疫测定等。而且，纯化的蛋白或其片段可以用于确定蛋白的三维晶体结构，用该结构可以确定蛋白或其部分的生物学功能、模拟分子间的相互作用及膜融合等。

候选试剂包括大量不同类的化学物质，虽然典型的是有机分子或复合体，但优选分子量大于 50，小于约 2500 道尔顿的小分子有机化合物。候选试剂含有与蛋白结构性相互作用必需的功能基团，尤其是氢键，一般包含至少一个氨基、羰基、羟基、羧基，并常常含有至少两个功能性化学基团。候选试剂常常含有碳环或杂环结构和/或芳香族或有一个或多个上述功能性基团替代的聚芳香族结构。候选试剂还包括但不局限于肽、糖类、脂肪酸、类固醇、嘌呤、嘧啶、衍生物、结构类似物或其联合的生物分子。

候选试剂可以通过不同的来源获得，包括合成或天然化合物文库。例如，可以利用多种方法来随机和直接合成大量的有机化合物和生物分子，包括随机寡核苷酸和寡肽的表达。或者，以细菌、真菌、植物或动物提取物的形式存在的天然化合物文库也可利用或易于产生。另外，可以通过常规化学、物理和生化手段很容易地对天然或合成文库和化合物进行修饰，并可用于产生组合文库。可以直接或随机地对已知药理试剂进行如酰化作用、烷基取代、酯化作用、酰胺化等化学修饰，从而产生结构类似物。

在筛选试验为结合测定时，一种或多种分子与一种标记偶联，该标记可以直接或间接地提供一种可检测信号。各种标记包括放射性同位素、荧光素、化学发光剂、酶、特异性结合分子、微粒，如磁性微粒等。特异性结合分子包括成对的，如生物素和链酶亲和素、地高辛和抗地高辛等。为了检测出特异性结合成分，通常根据已知步骤用一种提供检测的分子标记互补成分。在

筛选试验中还可以使用多种其它试剂。它们包括，如盐、中性蛋白（如白蛋白）、洗涤剂等，用于促进最佳的蛋白一蛋白结合和/或减少非特异性的或本底的相互作用。蛋白酶抑制剂、核酸酶抑制剂、抗微生物剂等试剂可以用来改进检测效率。

试剂可以结合可药用载体，该载体包括任何和所有的溶剂、分散介质、包衣、抗氧化剂、等渗和吸收延迟剂等。这种可以作为药用活性物质的介质和试剂的用途是本领域熟知的。除了作为任何与该活性成分不相容的常规介质和试剂，也考虑了它们在此处所述的治疗组合物和方法中的用途。附加的活性成分也可以加入组合物中。

配方可制备用于多种给药方式。配方可口服、吸入，或可注射，如血管内、肿瘤内、皮下、腹膜内、肌内注射等。治疗配方的剂量可以根据疾病的性质、给药的频率、给药的方式、药物从宿主的清除率等发生很大变化。首剂量可以大些，接着小剂量维持。给药可以不频繁，如一次给予、每周一次或两周一次，或者可分成较小剂量每天或每半周等给药一次，从而保持一个有效的剂量水平。在一些实例中，口服与静脉内给药需要不同的剂量。本发明确定的试剂可以掺入多种用于治疗给药的配方。更优选地，将复合物与适当的、可药用载体或稀释剂结合配伍成药物组合物，也可以将之配制成固体、半固体、液体或气态的制剂，如药片、胶囊、粉末、颗粒、药膏、溶剂、栓剂、注射剂、吸入剂、凝胶、微球体、气溶胶。这样，试剂的给药可以通过多种途径完成。试剂可在给药后扩散全身，或可以通过使用能使植入部位保持活性剂量的植入物使试剂局限。

下列方法和赋形剂仅仅是作为示范，并不以任何方式限制。对于口服制剂，一种试剂可以单独使用，或与适合的添加剂结合制成药片、粉末、颗粒或胶囊，如与传统的添加剂，如乳糖、甘露醇、玉米淀粉或马铃薯淀粉结合；和粘合剂，如结晶纤维素、纤维素衍生物、阿拉伯树胶、玉米淀粉或白明胶结合；和分解剂，如玉米淀粉、马铃薯淀粉或碳甲纤维素钠结合；和润滑剂，如云母或硬酯酸镁结合；如果需要还可以和稀释剂、缓冲剂、潮湿剂、防腐剂和调味剂混合。

另外，试剂可以通过溶解、悬浮或乳化于水或非水溶剂中配制成注射用制剂，这些溶剂包括植物性或其它相似的油、合成的脂肪酸甘油酯、高级脂

肪酸酯或丙烯乙二醇；如果需要，还可以添加传统的添加剂，如增溶剂、等渗剂、悬浮剂、乳化剂、固定剂和防腐剂。进而试剂可以用于气溶胶配方中，通过吸入方式给药。本发明确定的药剂可以配制成适合加压的推进剂，如二氯二氟甲烷、丙烷、氮等。还可以选择将试剂与不同基质如乳化基质和水溶性基质混合形成栓剂。进一步地，本发明确定的试剂可以通过一种栓剂直肠给药。栓剂可能含有一种载体如可可黄油、聚乙二醇和聚乙烯乙二醇，这些载体在体温下可以溶解，而处于室温时呈固态。

用于持续释放配方的植入物是本领域公知的。植入物可以和可生物降解的或非生物降解聚合物配制成微球体、胶块等。例如，乳酸和/或乙醇酸聚合物能形成易被宿主耐受的一种易蚀的聚合物。包含本发明确定的试剂的植入物可以置于作用位点临近处，这样活性试剂的局部浓度相对于身体的其他部位会增加。可以提供用于口服或直肠给药的制剂如糖浆、酏剂、悬浮液的单位剂量形式，其中一茶匙的量、一大汤匙的量、胶囊、药片或药栓等每一剂量单位含有预定量的本发明的成分。与之相似，用于注射或静脉内给药的单位剂量形式可以包含存在于组合物中，作为无菌水溶液、盐类或其它可药用载体的溶液的本发明的化合物。本发明新的单位剂量形式的规格取决于使用的特定化合物和需要产生的效果，及与宿主中每一种活性试剂相关的药代动力学。

可药用的赋型剂如媒介物、佐剂、载体或稀释剂，公众易于得到。而且，公众也容易得到可药用的辅助性物质，如 pH 调节和缓冲剂、张力调节剂、稳定剂、增湿剂等。

给患有疾病或紊乱的病人服用治疗剂量的鉴定试剂。根据特定的疾病进行局部用药、定位用药或是全身性用药。施用的化合物应保持一个有效的剂量，这样可以在一段适当的时期内基本阻止病情进展。体内使用时最好在医生的指导下开取和服用该组合物。剂量随特定的药剂和使用的配方、紊乱的类型、病人的状况等发生变化，这样可以充分地针对疾病或症状，同时减少副作用。治疗可以是短期的，如在外伤后，或是进行长期治疗，如精神分裂症的预防和治疗。

本发明鉴别的 SNPs 可以用于分析关联基因表达模式和与生物体的表型性状如疾病的易感性或药物反应性等相关的表达模式。可以确定多种组织的表

达模式并用来鉴别普遍存在的表达模式、组织特异性表达模式、暂时性表达模式和由各种外部刺激物如化学物质或电磁辐射诱导的表达模式。这种鉴别可以提供关于基因和/或其蛋白产物功能的信息。

新鉴别的序列还可以用作诊断标记，即预测表型特征，如对疾病的易感性或药物反应性。另外，本发明的方法可以用于对临床研究群体分层。同样地，这些基因或其片段可以用作探针来确定被检测生物体基因组中是否存在相同的核酸序列。另外，探针可以用来监控被测生物体或其部分，如特定组织或器官的 RNA 或 mRNA 的水平，从而确定与生物体内特定表型性状相关的标记表达水平。同样地，标记可在蛋白水平上分析，可以采用任何常规技术如免疫学方法——蛋白印迹、放射免疫沉淀等——或用活性基础分析测定一种与基因产物相关的活性。而且，当一种表型不能清楚区分有不同遗传基础的相似疾病时，可以用本发明的方法来正确鉴别疾病。

同样，本发明的方法显然能应用于除人类以外的生物体。例如，当该生物是动物时，本发明的方法可以用来鉴别与疾病抵抗力/或易感性、环境耐受力、药物反应等相关的位点，当该生物是植物时，本发明的方法可以用来鉴别与疾病抵抗力/或易感性、环境耐受力或除草剂抵抗力的相关位点。

可以理解本发明并不局限于所述的特定的方法、方案、细胞系、动物种或属、试剂，这些均可以发生改变。还需要说明的是，这里使用的术语只是为了描述特定的实施方案，并不限制本发明的保护范围，本发明的保护范围只能由所附权利要求限定。

数据库

本发明包括含有与变异相关信息的数据库，例如，这些信息包括有关 SNPs、SNP 单倍型区块、SNP 单倍型模式和信息型 SNPs 的信息。在一些实施方案中，本发明的数据库可以包含与一个或多个表型性状相关的一个或多个单倍型模式的信息。数据库还可以包含与给定变异相关的信息，如对发生变异的一般基因组域进行描述的信息，以及诸如变异是否发生在一个已知基因上，和附近是否有已知基因、基因同系物或调节区域等。

本发明数据库还可以包括其他信息，但不限于这些信息，SNP 序列信息、关于用于分析 SNP 单倍型模式的组织样品临床状况的描述性信息，或是与作为样品来源的患者的临床状况的描述性信息。数据库可被设计成包括不同的

部分，如变异数据库，SNP 数据库，SNP 单倍型区块或 SNP 单倍型模式数据库，信息型 SNP 数据库。数据库的构造和配置方法可广泛获得，见 Akerblom 等，(1999)的美国专利 US5,953,727，此处引入作为本发明的参考。

本发明的数据库可以链接外部或远程数据库，图 9 显示了一种适用于数据库和执行本发明软件的计算机网络实例。计算机工作站 902 通过一个局域网（LAN）如以太网 905 与应用/数据服务器 906 相连。打印机 904 可以直接与工作站或以太网 905 相连。LAN 可以与广域网（WAN）相连，如通过网关服务器 907 与因特网 908 相连，网关服务器 907 也可以作为 WAN908 和 LAN905 之间的防火墙。在优选实施方案中，工作站可以通过因特网 908 联系外部的数据资源，如 SNP 联盟（TSC）或国家生物技术信息中心 909。

可以用任何适宜的计算机平台进行 SNP 单倍型区块或模式、相关表型、数据库中的其他信息或由外部输入的信息间必要的比较。例如，可以从多个厂商，如 Silicon Graphics 购得大量计算机工作站。也可广泛获得客户服务器环境、数据服务器和网络，并用作本发明数据库的合适平台。

本发明的数据库还可以用于描述鉴别个体中 SNP 单倍型模式的信息，用这种描述可以来预测个体的一个或多个表型性状。该方法可以用于预测个体对疾病的易感性/抵抗力和/或对药物的反应。更进一步，本发明的数据库可以包括与本发明变异相关的一个或多个基因的表达水平有关的信息。

以下实施例具体描述了本发明的特定实施方案，但本发明例证的方法和材料并不限制发明的范围。

实施例 1：制备体细胞杂交体

用体细胞遗传学的标准方法将人类 DNA 链（染色体）从二倍体状态分离到单倍体状态。在本例中，一个具有野生型胸苷激酶基因的人二倍体类淋巴母细胞系与一个胸苷激酶基因突变的仓鼠二倍体成纤维细胞系融合，产生的细胞亚群体为包含人类染色体的杂交细胞。将仓鼠细胞系 A23 细胞吸进一个离心管中，管内有 10ml DMEM，其中添加了 10% 的胎牛血清（FBS）+1×Pen/Strep+10%谷氨酰胺，1500rpm 离心 5 分钟，后重悬于 5ml RPMI 中，吸移进一个含有 15ml RPMI 培养基的组织培养瓶中。类淋巴母细胞在 37℃生长至汇合。同时，将人类类淋巴母细胞吸移进一个离心管中，管内有 10ml DMEM，其中添加了 10% 的胎牛血清（FBS）+1×Pen/Strep+10%谷氨酰胺，

1500rpm 离心 5 分钟，后重悬于 5ml RPMI 中，吸移进一个含有 15ml RPMI 的组织培养瓶中。类淋巴母细胞在 37℃ 生长至汇合。

为了制备 A23 仓鼠细胞，吸走生长培养基并用 10ml PBS 漂洗细胞。接着用 2ml 的胰蛋白酶消化细胞，将这些细胞分开置于 3-5 个有新鲜培养基（没有 HAT 的 DMEM）的培养皿上，于 37℃孵育。转移培养基，将类淋巴母细胞移进离心管，以 1500rpm 离心 5 分钟，吸走生长培养基，将细胞重悬于 5ml RMPI 中，吸移 1—3ml 的细胞至 2 个装有 20ml 的 RPMI 的烧瓶中。

将约 8—10×10^6 个类淋巴母细胞 1500 rpm 离心 5 分钟以获得融合细胞。接着用 DMEM 重悬漂洗细胞沉淀，再次离心并吸走 DMEM，将类淋巴母细胞重悬于 5ml 于新鲜的 DMEM 中。受体 A23 仓鼠细胞生长至汇合，在融合前 3—4 天前，将其分开，此时细胞已经融合了 50—80%。移走原有的培养基后用 DMEM 漂洗细胞三次，用胰酶消化，最后将细胞重悬于 5ml DMEM。小心地把类淋巴母细胞移至受体 A23 细胞上，轻缓涡旋混合培养基，37℃孵育 1 小时。孵育后，接着小心地吸走 A23 细胞的培养基。一只手转动培养皿的同时，另一只手将一个移液管靠在培养皿的边缘缓慢地加入 2ml 室温 PEG1500，其中转动培养皿一周添加所有的 PEG 大约需要 1 分钟的时间。接着缓慢转动培养皿的同时沿培养皿边缘加入 8ml DMEM。小心地吸走细胞上层的 PEG/DMEM 混合物，用 8 ml DMEM 漂洗细胞。移走 DMEM，再加入 10ml 新鲜的 DMEM，在 37℃孵育细胞 30 分钟，吸走细胞上层的 DMEM，加入 10ml 10%FBCS 和 1×Pen/Strep 的 DMEM 后过夜孵育。

孵育后吸走培养基，用 PBS 漂洗细胞，胰蛋白酶消化细胞，分到含有选择性培养基（含有 10% 的 FBCS+1×Pen/Strep 链锁状球菌+1×HAT 的 DMEM）的培养皿中，这样每块培养皿中约达 10 万个细胞。铺板后第三天更换培养基。挑出克隆并置于 24 孔培养板中直到肉眼可见（9-14 天），如果挑出的克隆在 5 天内就融合，表明生长良好，胰蛋白酶消化细胞并移至 6 孔板中。

来源于 6 孔板培养基中的细胞来制备 DNA 和干杂交细胞培养基。胰蛋白酶消化细胞，在 100mm 大的含有 10ml 选择性培养基的板上分开细胞并将之置于Eppendof管中，管中细胞聚集成团，细胞重悬于200μl PBX 中，使用 Qiagen DNA mini 试剂盒分离 DNA，在每一旋转柱中细胞浓度少于 5 百万个。待

100mm 板上细胞生长至融合后，细胞继续培养于培养基中或进行冷冻保存。

实施例 2：选择单倍杂交体

使用 Affymetrix 的 HuSNP 基因芯片（Affymetrix，InC.，of Santa Clara，CA，HuSNP 图谱鉴定、试剂盒和使用手册，Affymetrix Part No.900194）评定人类染色体的存在、缺失和二倍体/单倍体状态，该芯片可以在单芯片杂交中测定 1494 个标记。作为对照，用 HuSNP 基因芯片杂交分析筛选仓鼠和人类二倍体类淋巴母细胞系。在每个融合细胞系中评定任何在亲本类淋巴母细胞二倍体细胞系为杂合型的 SNPs 的单倍性。假定“A”和“B”是在每个 SNP 位点可以选择的变异体，通过比较作为“AB”杂合型在亲本二倍体细胞系中出现的标记与作为“A”或“B”（半杂合）出现在杂交体中的相同标记，可以确定每个杂交系中单倍体状态的人类 DNA 链。

图 11 显示检测两个人类/仓鼠细胞杂交体（杂交体 1 和杂交体 2）中位于第 21 号人类染色体上的选择标记的结果。第 1 栏列出了 HuSNP 芯片标记名称，第 2 栏表明了当仓鼠细胞核酸（非融合）和 HuSNP 芯片杂交后是否获得了一种信号。如预测的一样，在仓鼠细胞样本中没有检测到任何标记的信号。第 3 栏表明了在二倍体亲本人类类淋巴母细胞系 CPD17 中检测到每种标记的变异体（“A”、“B”或“AB”）。在某些情况下，只有一种 A 变异体存在；在某些情况下，只有一种 B 变异体存在；在某些情况下，CPD17 细胞是杂合型（“AB”）变异体。最后两栏表明了来源于两种人类/仓鼠杂交体（杂交体 1 和杂交体 2）的核酸样本与 HuSNP 芯片杂交的结果。注意当只有 A 变异体出现在亲本 CPD17 细胞系中时，只有 A 变异体在细胞融合中转移；当只有 B 变异体出现在亲本 CPD17 细胞系中时，只有 B 变异体在细胞融合中转移；当 CPD17 细胞系为杂合型时，A 变异体转移进一些融合克隆，而 B 变异体转移进其他融合克隆。但必须要了解的是，通常只有部分染色体存在于融合过程产生的杂交细胞系中，一些杂交体可以是某些人类染色体或其部分的二倍体，一些杂交体可以是另一些人类染色体或其部分的单倍体，一些杂交体可以不含有某些染色体的任一种变异体。可以选择只含有特定人类染色体（如 21 号染色体）的一种变异体的杂交体用于分析。更优选的，将含有全部染色体（相对于只有其中部分而言）的杂交体选来分析。

实施例 3：长距离 PCR

来源于仓鼠/人类细胞杂交体的 DNA 可以用于进行长距离 PCR 分析，这种分析一般可以从文献中获知，例如，Boehriger Mannheim 扩展长距离 PCR 试剂盒所述的标准长距离 PCR 方法，此处引入作为参考或用于所有目的。

通过以下途径设计用于扩增反应的引物：将一个给定序列，例如将 21 号染色体的 23 兆碱基的重叠群序列输入一个现有技术中已知的所谓“重复掩饰（repeat masker）”的软件程序，该掩饰可以识别基因组上的重复序列（如 Alu 和 Line 元件）（见 A.F.A.Smit 和 P.Green，www.genome.washington.edu/uwgc/analysistools/repeatmask，作为本发明的参考）。重复序列通过用“N”替换重复序列中的每个特定核苷酸（A、T、G 或 C）的程序被“掩饰”。接着将重复掩饰替代后产生的输出序列输入一个商业上可获得的引物设计程序（Oligo 6.23），利用该程序选择长度大于 30 个核苷酸、解链温度大于 65℃的引物。将 Oligo 6.23 设计的输出引物输入到一个程序，接着该程序“选择”出能 PCR 扩增基因组上的给定区域但与临近 PCR 产物重叠最小的引物对。商业上可获得的方法和这种引物设计进行长距离 PCR 的成功率至少为 80%，在人类染色体的某些部分上可以获得 95%以上的成功率。

长距离 PCR 的解释性实施方案使用 Boehringer Mannheim Cat# 1681834、1681842 或 1759060 扩展长模板 PCR 系统。在过程中，每个 50μL PCR 反应中需要两种主混合物。在特定的实施例中，在置于冰上的 1.5ml 的微量离心管中制备每个反应所需的主混合物 1，其中含有终体积为 19μL 的分子生物学级水（Bio Whittaker Cat.# 16-001Y）；2.5μL 10mM dNTP，含有的 dATP、dCTP、dGTP 和 dTTP 均为 10mM（Life Technologies Cat.#10297-018），每种 dNTP 的终浓度达到 400μM 和 50ng DNA 模板。

制备用于所有反应的主混合物 2，冰上保存。每个 PCR 反应的主混合物 2 包含一个终体积为 25μL 的分子生物学级水（Bio Whittaker）；5μL 10×PCR 缓冲液，其中含有 22.50mM $Mgcl_2$（Sigma, Cat.#M 10289）；2.5μL 10 mM $Mgcl_2$ (为了使 $Mgcl_2$ 的终浓度为 2.75mM)和 0.75μL 酶混合物（最后添加）。

将 6 微升预混合引物（含有 2.5μL 的主混合物 1）加入合适的试管中，然后在每管加入 25μL 主混合物 2，盖上管混合，轻缓离心，重置冰上。这时，开始按照以下程序进行 PCR 循环：步骤 1：94℃变性模板 3 分钟；步骤 2：94

℃ 30 秒；步骤 3：在引物适宜的退火温度退火 30 秒；步骤 4：68℃延伸 1 分钟/Kb 产物；步骤 5：重复第 2—4 步 38 次共 39 个循环；步骤 6：94℃30 秒；步骤 7：退火 30 秒；步骤 8：68℃延伸 1 分钟/Kb 产物加上额外的 5 分钟；和步骤 9：保持于 4℃。也可以选择使用两步 PCR：步骤 1：94℃变性模板 3 分钟；步骤 2：94℃ 30 秒；步骤 3：68℃退火和延伸 1 分钟/Kb 产物；步骤 4：重复第 2—3 步 38 次共 39 个循环；步骤 5：94℃ 30 秒；步骤 6：68℃退火和延伸 1 分钟/Kb 产物加上额外的 5 分钟；步骤 7：保存于 4℃。

人类第 14 和 22 号染色体上不同区域的长距离 PCR 扩增反应的结果可以通过溴化乙锭染色的琼脂糖凝胶电泳（图 12）显示。利用本发明长距离 PCR 扩增方法常规产生的扩增片段的平均大小约为 8Kb，只有在很少的情况下（见第 22 号染色体胶 G11）中没有扩增出基因组区域。

实施例 4：晶片设计、制造、杂交和扫描

寡核苷酸探针系列保存在一个寡核苷酸阵列（芯片或晶片）上，它是根据需要查询的人类 DNA 链序列设计的。寡核苷酸序列来源于公众可以利用的数据库报道的共同序列。一旦确定了探针序列，计算机运算法则就可以用于设计在加工含探针的阵列中应用的光刻掩模。通过一种光指示的化学合成过程来制造阵列，该过程结合了固相化学合成与光刻制作技术。见，例如，WO92/10092，或美国专利 5,143,854、5,384,261、5,405,783、5,412,087、5,424,186、5,445,934、5,744,305、5,800,992、6,040,138、6,040,193，在此引入其完整内容作为参考。利用一系列的光刻掩模来确定特定化学合成步骤后在玻璃底物（晶片）上的曝光位点，通过这个过程构建阵列上核苷酸探针的高密度区域，同时每个探针位于预定的位置。同时和平行合成多个探针区域。

合成过程包括通过将光透过一个光刻掩模，其中在未保护区域的化学基团被光激活，选择性照亮一个影像——保护的玻璃底层。接着选择性激活的底物晶片和一个选择的核苷酸一起孵育，在晶片激活区域发生化学偶联。一旦发生偶联，应用一个新的掩模模式与另一种选择的核苷酸重复偶联步骤，直到获得期望的探针系列。在一个特定的实施例中，应用 25-mer 的寡核苷酸探针，其中第 13 个碱基是待确定的碱基。用四个探针来查询存在于每个序列中的每个核苷酸——除了第 13 个碱基，一个探针与该序列互补，三个错配探针等同于互补探针。在某些情况下，每个阵列上至少有 10×10^6 个探针。

一旦阵列构成后，就与通过在仓鼠—人类细胞杂交体上进行长距离 PCR 反应产生的产物进行杂交。标记待分析的样品并将它与阵列一起孵育，直到样品与晶片上的探针杂交。

将杂交后的阵列插入一个共焦高性能的扫描仪，这样可以检测杂交的模式。当样品中 PCR 产物与探针结合，已掺入到 PCR 产物中的荧光报道基团发射荧光时，收集杂交数据。与那些含有错配的序列相比，样品中与晶片上探针互补的序列与晶片杂交的程度更强并能产生更强的信号。由于阵列中每个探针的序列和位置是已知的，通过互补性，可以鉴定出应用于探针阵列的样品核酸的变异。用于本发明的扫描仪和扫描技术是本领域技术人员已知的，并公开在：如描述了微阵列芯片的美国专利 US5,981,956，US6,268,388 和 US5,459,325 U.S.S.N。另外 2000 年 8 月 3 日提交的 60/223,278，及要求 2001 年 8 月 3 日提交的 USSN60/22,3278 的优先权的非临时专利申请描述了扫描仪和全晶片扫描的技术，在此引入其完整内容作为参考。

实施例 5：鉴定人类 21 号染色体的 SNP 单倍型

代表了非洲人、亚洲人和高加索人 21 号染色体的 20 个独立的染色体拷贝被用于分析 SNP 发现和单倍型结构。采用啮齿动物—人类体细胞杂交技术将每个个体 21 号染色体的两个拷贝物理分离（图 10），讨论如前。用于分析的参考序列由人类 21 号染色体基因组 DNA 序列组成，此序列由 32,397,439 个碱基组成。掩饰这种参考序列的重复序列，用高密度寡核苷酸阵列分析产生的 21,676,868 个碱基的（67%）的独特序列中的变异。将八个独特的寡核苷酸，每个长为 25 个碱基，用于查询独特样品中每一条 21 号染色体的碱基，共约 1.7×10^8 条不同的寡核苷酸。这些寡核苷酸分布于用前述策略（Chee，et al.，science，274:610，1996）设计的共八个不同的晶片上。光指示性寡核苷酸的化学合成是在 5 英寸×5 英寸的玻璃晶片（购自 Affymetrix，Inc，Santa Clara，CA）上进行的。

设计的独特寡核苷酸用于产生 3253 个最低程度重叠的长距离 PCR（LRPCR）产物，平均长度为 10kb 左右，跨越了邻近 21 号染色体 DNA 的 32.4Mb，制备如前所述。将相应的 LRPCR 产物混合，并用 Qiagen tip 500（Qiagen)纯化后用于与每个晶片杂交。用 37μl 10X One-Phor-All 缓冲液 PLUS（promega）和 1 单位 DNA 酶（生命技术/Invitrogen）将共 280μg 纯化的 DNA

分离成片段，反应总体积为370□L，置于37℃反应10分钟，99℃热灭活10分钟。成片段的产物用500单位的Tdt（Boehringer Manheim）和20nmol的生物素-N6-ddATP（Dupont NEN）进行末端标记，置于37℃反应90分钟，95℃热灭活10分钟。标记的样品与晶片杂交，该过程是在以下溶液中进行的：10mM Tris-HCL(PH 8)3M氯化四甲铵，0.01% Tx-100,10μg/ml变性的鲱精子DNA，总体积为每个晶片10ml，在50℃反应14-16个小时。在4×SSPE中粗略漂洗晶片，用6×SSPE洗涤三次，每次十分钟，用链酶亲和素R-藻红蛋白（SAPE，5ng/ml）室温染色10分钟。通过用抗链酶亲和素的抗体（1.25ng/ml）染色和重复SAPE染色步骤增强信号。

相应于一个单晶片上碱基的PCR产物被混合并作为一个单独反应与晶片杂交。为了寻找DNA序列变异，在160个晶片上合成共3.4×10^9个寡核苷酸用于扫描20个独立的人类21号染色体的拷贝。采用长距离PCR，由一个啮齿动物—人类杂交细胞系扩增得到每个独特的21号染色体。利用具有高适度严格参数的Oligo6.23引物设计软件设计LRPCR实验。所得引物长度典型为30个核苷酸，解链温度>65℃。扩增子长度范围在3kb—14kb之间。建立一个全染色体的引物数据库，用软件（pPicker）选择一套最小的非冗余的引物，这种引物能最大程度地涵盖21号染色体序列，并且在相邻扩增子之间重叠最小。作为选择，本发明也可以使用实施例3描述的选择引物的方法。使用较少改进的延伸长模板PCR试剂盒（Boehringer Mannheim）进行LRPCR反应。用常用共焦扫描仪扫描晶片。

用一个模式识别运算法则检测变化杂交时的SNPs。将先前描述的运算法则（Wang，et al.，Science，280：1077（1998））经过组合后用于检测基于变化杂交模式的SNPs。在20个染色体样本中鉴定出共35,989种SNPs。这些人类多态的序列和位点已经保存在GenBank的SNP数据库。双脱氧序列用于评定一个随机样本中的227种SNPs，这些SNPs存在于最初的DNA样本中，从而证实了分析的220（97%）种SNPs。为了获得3%的低假阳性SNPs，晶片上SNPs的检测需要严格的阈值，这样会产生高的假阴性率。晶片上约65%的碱基产生了高质量足够用于SNPs检测的数据，剔除了35%的呈假阴性的数据。分析样本中由于长距离PCR的失败导致35%假阴性率中的15%。剩下20%的假阴性分布于不产生高质量数据的碱基（10%）和只在分析的20条

染色体的一个片段上产生高质量数据的碱基（10%）之间。一般来说，碱基所处的序列的内容决定了它是否产生高质量的数据。发现约有 20%的碱基一贯提供差的数据，这与发现的在单个双脱氧测序时阅读 500 个碱基时质量评分太低不能用于可靠的 SNPs 检测（Altschuler，et al.，Nature，407：513，（2000））非常相似。在只有有限数量的被分析的样本产生了对于一个给定碱基的高质量数据的情况下，相对于较常见的 SNPs，发现罕见 SNPs 的能力发生不成比例地降低。结果，用这种方法进行的 SNP 发现偏向于常见 SNPs。

图 13A 表明了在总体多样染色体样本发现的所有 35,989 种 SNPs 中次要等位基因频率的分布。用两种核苷酸多样性方法估计针对样本中染色体数目标准化的遗传变异：每位点平均杂合性 π 和群体突变参数θ（见 Hartl 和 Clark，Principles of Population Genetics，（Sinauer，Massachusetts，1977））。完成的基因组 21 号染色体 DNA 的 32,397,439 个碱基被分成 200,000 碱基对的片段，检测每个片段中用于 SNP 发现的高质量碱基对。观察到的这些碱基的杂合性用于计算每个片段的平均核苷酸多样性（π）。总数据组（π＝0.000723 和θ＝0.000798）估计出来的平均核苷多样性与在 21 号染色体的 200,000 个邻接碱基对中测定的核苷多样性（图 13B）分布均处于先前描述值的范围内（国际 SNP 图谱工作组，Nature 409：928－33（2001））。

将 SNP 联盟（TSC）发现的 15,549 种 21 号染色体上 SNPs 的重叠程度与本发明发现的 SNP 进行比较。由于 TSC 发现的 SNP 中有 5,087 种位于重复 DNA 中，故没有将它们排在晶片上，在剩余的 10,462 种 TSC SNPs 中，鉴定了 4705 种（45%）。观察发现θ的估计值要大于所分析的邻近 DNA 序列的 162200kb 框中的 129kb 得出的 π 估计值。这种差异与最近的人口数量增长相一致，并与最近关于人类基因核苷酸多样性的研究相似（Stephens，et al.，Science 293：489（2001））。研究发现，与已知观察到的核苷酸多态性数量的中性模型期望的 43%单一相比， SNPs 中的 11,603（32%）种具有次要等位基因，在样本中只观察到一次（单一）。（Fu 和 Li，Genetics 133：693（1993））。观察值与期望值之间的差异可能是由于与上述讨论的与常见 SNPs 相比，本研究鉴别罕见 SNPs 的能力较低。

综上所述，对估计位于人类基因组上 32.4Mb 的 53000 种常见 SNPs 中的 47%进行了鉴定，这些 SNP 的等位基因频率大于等于 10%。将其与约 18－20

%的国际 SNP 图谱工作组和 SNP 联盟收集的所有这样的常见 SNPs 进行比较。涵盖差异可以通过本研究使用了较大数量染色体用于发现 SNP 来解释。为了评定发现的可重复性，进行 SNP 发现以便设计一个晶片，该晶片具有 21 号染色体的 19 个额外的拷贝，来源于和原始系列样本相同的多样性实验对象。利用两套样本鉴别了共 7188 种 SNPs，其中平均 66%在第一套样本中发现的 SNPs 也存在于第二套中，与先前的发现相一致（Marth，et al.，Nature Genet，27：37(2001)和 Yang，et al.，Nature Genet，26：13(2000)）。如预期的一样，一种 SNP 在第二套样本中没有复制大大依赖于等位基因的频率。那些次要等位基因在一套样本中只出现 2 到 3 次的 SNPs 中的 80%也出现在第二套样本中，而次要等位基因出现一次的 SNPs 的 32%出现在第二套样本中。这些结果表示，次要等位基因出现不止一次的集合中的 24,047 种 SNPs 在不同的总体样本中可高度重复，因而这些 SNPs 能用来定义常见总体单倍型。在 SNP 发现的过程中，鉴定了看来具有两个以上的等位基因的 339 种 SNPs，但这些 SNPs 不包括在本分析中。

除了 SNPs 在不同样本中可以重复，SNPs 集合中的连续 SNPs 之间的距离对于定义有意义单倍型结构是关键的。如果集合中连续的 SNPs 之间的距离很大程度上依赖于实际单倍型区块的大小，则可能不识别短至几 kb 的单倍型区块。即使重复序列不包括在 SNP 发现过程中，本研究中的 SNPs 集合也均匀地分布在染色体上。图 13C 表明了 SNPs 的分布覆盖了已完成的 21 号染色体 DNA 序列的 32,397,439 个碱基。一个间隔就是连续 SNPs 之间的距离。整个 SNP 组之间共有 35,988 个间隔，常见 SNP 组之间共有 24,046 个间隔（即次要等位基因在样本中出现多于一次的 SNPs）。考虑所有的 SNPs 时，连续 SNPs 之间的平均距离为 900 个碱基，当只考虑 24,047 个常见 SNPs 时，连续 SNPs 之间的平均距离为 1300 个碱基。对于这组常见的 SNPs 来说，93%位于包括重复 DNA 序列的基因组 DNA 的连续 SNPs 之间的间隔小于等于 4000 个碱基（同样，见图 13 C）。

从二倍体数据构建单倍型区块或模式是复杂的，因为任两个杂合 SNPs 等位基因之间的联系不能直接观察到。考虑带有两拷贝的 21 号染色体和两个等位基因，A 和 G，位于一条 21 号染色体 SNP，同时两个等位基因，A 和 G，位于第二条 21 号染色体 SNP。在这样一个例子中，不清楚是 21 号染色体的

一个拷贝在第一个 SNP 含有等位基因 A 以及，在第二个 SNP 含有等位基因 A，而 21 号染色体的另一个拷贝在第一个 SNP 含有等位基因 G，在第二个 SNP 含有等位基因 G，还是 21 号染色体在第一个 SNP 含有等位基因 A，在第二个 SNP 含有等位基因 G，而 21 号染色体的另一个拷贝在第一个 SNP 含有等位基因 G，在第二个 SNP 含有等位基因 A。目前的方法用于解决这个问题，包括统计估计单倍型频率，直接由家系数据得出推论，和对短片段进行等位基因特异性 PCR 扩增。

为了避免复杂性，本发明鉴定了从啮齿动物—人类体细胞杂交体分离的 21 号染色体的单倍体拷贝上有代表性的 SNPs，允许直接确定这些染色体的全部单倍型。这组带有在数据组中不只出现一次的一个次要等位基因的 24,047 个 SNPs 用于定义单倍型结构，如图 14 所示。表示了由 147 个常见人类 21 号染色体的 SNPs 定义的 20 种独立总体多样性染色体的单倍型模式。147 个 SNPs 覆盖 106kb 的基因组 DNA 序列。每一行有色框代表了一个 SNP，其中黑框代表 SNP 的主要等位基因，白框代表 SNP 的次要等位基因。任何位置没有框则代表数据丢失。每一列有色框代表了一条染色体，SNPs 按照它们的物理顺序排列在染色体上。在连续 SNPs 之间的没有变异的碱基没有在图中表示出来。147 个 SNPs 分成 18 个区块，由黑色水平线确定。确定一个区块开始和邻接区块结尾的 21 号染色体基因组序列中的碱基位置由垂直黑线左边的数字标示。图右面放大的框代表一个 SNP 块，由跨越基因组 DNA 上 19kb 的 26 个常见 SNPs 定义。在样本中表示的 7 个不同的单倍型模式中，四个最常见的模式包含了采样的 20 条染色体中的 16 条（即 80%的样本）。黑圈和白圈表明两种信息型 SNPs 等位基因的模式，可以明确区分这个区块中的四种常见单倍型。虽然没有两条染色体具有与这 147 个 SNPs 相同的单倍型模式，但在多条染色体的许多区域却有着一个相同的常见模式。通过扩展跨越 19kb 的 26 个 SNPs 定义的区域，可以进行更细致的分析（见图 14 放大的区域）。这个区块确定了 20 条染色体中的七种唯一的单倍型模式。除了由于不能通过数据质量阈值而丢失了一些数据以外，所有的情况下给定的 DNA 可以明确定位到这七种中的一种单倍型。四种最常见单倍型，每一种由三到四条染色体代表，占样本所有染色体的 80%。只需要 26 个中的两个“信息型”SNPs 来区别四种最常见的单倍型。在这个实施例中，仅使用这两个信息型 SNPs

中的信息将具有不常见单倍型的四条染色体分类为常见单倍型，这种分类是不准确的。然而，很明显 80%的整个总体样本中的单倍型结构是用区块中所有 SNPs 中的 10%定义的。几种可能性存在于可以被选择的三个信息型 SNPs 中，这样四种常见单倍型中的每一种可以由一个 SNP 来唯一确定。在混合样本分型的试验中，这些“三 SNP”选择中的一种将优于两 SNP 的组合，因为两 SNP 的组合将会限制在这种情形下四种常见单倍型频率的确定；因此，本发明提供了一种 SNP 作图的动态改进方法，该方法优于随机选择方法。

总而言之，尽管特定应用可以指导能获得单倍型信息的信息型 SNPs 的选择时，样本中大部分单倍型信息包含在所有 SNPs 的非常少的亚组中是显而易见的。另外从这个 SNPs 区块中随机选择的两到三个信息型 SNPs 将不能经常提供足够的信息来唯一分配一条染色体到四种常见单倍型中的一种。

当将定义单倍型结构时需要 SNPs 总数降至最少时，如何确定跨越 21 号染色体的 32.4Mb 的一组连续的 SNPs 区块是一个问题。在一个实施方案中，一个基于“greedy”策略的优化算法可以来解决该问题。考虑大小为一个 SNP 或更大的物理上连续的 SNPs 区块，将模糊单倍型模式视为缺失数据，并在计算涵盖百分率时不把它们包括在内。同时考虑到余下那些重叠的区块，选择在最少数量 SNPs 区块中总 SNPs 与唯一区别区块中不只一次出现的单倍型所需要的最小 SNPs 数目的比率最大的区块。任何与选择区块发生物理性重叠的剩余区块被放弃，重复这个过程直到选择了一组连接的，无重叠的覆盖 21 号染色体的 32.4Mb 序列而无缺口的区块，同时将每一个 SNP 分配到区块上。给定 20 条染色体样本大小，此算法产生每个区块 10 个常见模式的最大值，每种模式由两条独立染色体代表。

将这种算法应用于 24,047 个常见 SNPs 的数据组，确定了跨越 21 号染色体的 4135 个区块。其中有 14%的区块每一个含有 10 个以上 SNPs，共 589 个区块，包括总长 32.4Mb 的 44%。与之形成对照的是有 52%的区块每一个含有 3 个以下 SNPs，共 2138 个区块，只构成了染色体物理长度的 20%。最大的区块含有 114 个常见 SNPs，跨越基因组 DNA 的 115kb。总的来说，区块的平均物理长度为 7.8kb。区块的大小与它在染色体上的顺序没有关系，大区块与小区块散布在染色体上。每个区块平均有 2.7 个常见单倍型模式，定义为在多染色体上观察到的单倍型模式。平均说来，样本中 20 条染色体中的

9.6 条代表了区块中最常见的单倍型模式，第二常见单倍型模式由 4.2 条染色体代表，如果存在第三常见单倍型模式，则由 2.1 条染色体代表。总体多样性染色体的这样一个大片段由如此有限的单倍型多样性来代表的事实是值得注意的。当考虑单倍型模式的频率时，发现与观察相一致，在 313 个人类基因的集合中观察到的 82%的单倍型模式出现在所有种族中，只有 8%的单倍型是群体特异性的（Stephens，et al.，Science 293：489-93（2001））。通过几个试验测定单倍型算法中参数对产生的区块模式的影响。需要被常见单倍型覆盖的染色体比例可以改变，从起初的 80%到 70%和 90%。如所预期的，需要更完整的覆盖产生更大数量的短的区块。只用样本中次要等位基因频率至少为 20%的 16,503 个 SNPs 产生了一些更长的区块，但每个区块 SNPs 的数目并没有显著改变。在约 3Mb 的一个区域，分析了一个含有 38 条染色体的更稠的样本中的 SNPs 和频率至少为 10%的常见单倍型区块，使其与 20 条染色体的分析可比。产生的区块大小的分布接近于原始结果。而且，进行一项随机选择试验，改变每个 SNP 中已知等位基因顺序，然后用于单倍型区块发现。在这一项研究中，94%的区块含有 3 个以下的 SNPs，只有一个区块含有 5 个以上的 SNPs。这证实了由数据中看到的较大区块不是偶然联系产生的，或是本发明区块选择方法人为制造的结果。

为了确定基因是否按比例地表现于大区块和小区块中，用含有 10 个以上 SNPs、3—10 个 SNPs，3 个以下 SNPs 的区块中外显子碱基的数目来确定。外显子碱基比含有 3—10 个 SNPs 区块中的总碱基有些过表现（变换试验确定的 $P<0.05$）。

基于区块中单倍型结构的知识，可以选择 24,047 个常见 SNPs 的亚组来获得任何需要的常见单倍型信息的片段，常见单倍型信息定义出现超过一次，包括跨越了整个 32.4Mb 的样本的 80%以上的单倍型的信息。图 15 表明获得 32.4Mb 长的 21 号染色体上常见单倍型信息所需的 SNPs 的数量。对于每一个 SNP 区块，明确区分该区块中出现不只一次的区块所需的 SNPs 的最小数量（即常见单倍型信息）被确定下来。这些 SNPs 提供了由区块限定的全长物理距离的比例的常见单倍型信息。从为最大物理距离提供常见单倍型信息的 SNPs 开始，在物理覆盖（即覆盖比例）的累积增长被作图为与添加的 SNPs 的数量（即需要的 SNPs）相关。基因 DNA 包括开始于每个已知 21 号染色

体基因的第一个外显子的5'端10kb，延伸到该基因最后一个外显子的3'端10kb的所有基因组 DNA。例如，尽管获得所有常见单倍型信息需要最少的 SNPs为 4563 个，获得含有三个或更多 SNPs、覆盖 32.4Mb 中 81%的区块中的常见单倍型信息只需要 2793 个 SNPs。获得基因 DNA 中所有常见的单倍型信息共需要 1794 个 SNPs，代表了约二百二十个不同的基因。

本发明特别涉及全基因组关联研究，对表型如普通疾病的基因作图。这种方法是基于一种假设，即对常见疾病的易感性是由常见的遗传变异引起的（Risch and Merikangas, Science 273:1516 (1996), Lander, Science 274:536 (1996)）。通过比较在不相关的病例和对照中的遗传变异频率，遗传关联研究可以识别在人类基因组中对引起疾病起重要作用的特定的单倍型。虽然这种方法已经成功地应用于与疾病相关的单个候选基因（Altschuler, et al. Nature Genet. 26:76 (2000)），近来人类 DNA 序列的确定提供了研究整个基因组的可能性，这使基因关联分析（Kruglyak, Nature Genet. 22:139(1999)）的能力极大提高。实现这种方法的主要限制是缺乏人类基因组单倍型结构方面的知识，而这方面的知识对于选择合适的遗传变异进行分析是必不可少的。本发明证明结合高密度寡核苷酸阵列与体细胞遗传样本制备可以提供一种高效解决办法，依据经验确定人类基因组的单倍型结构。

尽管具有一种简单单倍型结构的基因组区域的长度差异极大，但一组密集的常见 SNPs 还是可以帮助人们用系统的方法确定人类基因组区块，其中80%的总体人群仅由三种常见单倍型来描述。总而言之，当在这个实施方案中运用特定的算法时，任何区块中最常见的一种单倍型可以在 50%的个人中找到，第二常见的单倍型可以在 25%的人群中找到，而第三常见的单倍型可以在 12.5%的人群中找到。要注意的是，区块是基于它们的遗传信息内容来确定的，而非根据这种信息是如何产生或为何存在。正是如此，区块没有绝对的界限，可以通过不同的方式进行定义，这主要依据其特定的用途。这个实施方案中的算法仅仅提供了多种可行方法中的一种。结果表明，需要一组高密度的 SNPs 来捕获所有常见单倍型信息。另一方面，这种信息也可以用于识别更小亚组的 SNPs，从而有助于复杂的全基因组关联研究。

本领域的专业人员很容易想到将人类 21 号染色体上应用的这种技术运用到人类基因组中其他所有的染色体。在本发明的优选实施方案中，各代表性

的许多人群的多个全基因组被用于识别所有或大多数人常见的 SNP 单倍型区块。在一些实施方案中，SNP 单倍型区块是基于排除了出现频率低的 SNPs 后的亲本 SNPs。当亲本 SNPs 被保存在基因组中时，它很可能非常重要，因为亲本 SNPs 能传达给携带它的生物体一些选择优势。

实施例 6：使用关联基因进行基因治疗和药物发现

本预见性的实施例概括了一个使用本发明方法的例子。SNP 是在 20 个单倍体基因组中被发现的，采用本发明的方法分析了 50 个单倍体基因组，确定 SNP 单倍型区块、SNP 单倍型模式、信息型 SNPs 以及每个信息型 SNP 携带的次要等位基因频率。这 50 个单倍体基因组包含了本研究的对照基因组（见图 13 的步骤 1300）。

随后使用长距离 PCR 和如前所述的微阵列（见授权于 Lipshutz 等的美国专利 US6,300,063，和 Chen 等的美国专利 US5,837,832），可以分析来自于 500 个具有肥胖表型的个体的基因组 DNA，同时确定这个临床群体中每个信息型 SNP 的次要等位基因的频率（见图 13 步骤 1310）。比较两个群体中信息型 SNP 位点的次要等位基因频率，结果表明对照和临床群体在三个信息型 SNP 位点具有统计学上的显著差异（步骤 1320 和 1330），选择对照群体和临床群体在次要等位基因频率上具有最大差异的 SNP 位点用于分析。

选择的信息型位点存在于一个 SNP 单倍型区块中，此区块跨越编码区 5´端的 1kb 非编码序列和瘦素基因编码区的 4kb（步骤 1340）。分析这个区域里发生的变异表明这个区域里在一个 SNP 位点的一个 G 是负责破坏瘦素基因的启动子，使瘦素蛋白的表达相当缺乏。

成纤维细胞可以通过皮肤活组织切片检查获得。产生的组织置于组织培养基中，并分成小片。将小片组织放在装有培养基的组织培养瓶底部的潮湿表面，室温下 24 小时后，加入新鲜的培养基（如含有 10%FBS、青霉素、链霉素的 Ham's F12 培养基）。接着将组织在 37℃孵育约一个星期。在这段时间里，每隔几天更换一次培养基，同时加入新鲜的培养基。在培养额外的两个星期后，单层纤维原细胞形成，这些单层细胞被吸移进一个较大的培养瓶中。

将含有一个耐卡那霉素基因的来源于莫洛尼鼠白血病病毒的载体用限制性酶消化，从而克隆一个待表达的片段。消化的载体用牛肠磷酸酶处理，制

止自连接。去磷酸化的线性载体在一块琼脂糖胶上进行分馏和纯化。分离能够表达活性瘦素蛋白产物的瘦素 cDNA。如果需要对片段的末端进行修饰，将它克隆进载体。将相同摩尔量的莫洛尼鼠白血病病毒线性主链和瘦素基因片段一起混合，用 T4 DNA 连接酶连接。连接产物用于转化 E.coli 细菌，接着将细菌平铺在含有卡那霉素的琼脂上，卡那霉素表型和限制性分析证实了载体中适当插入了瘦素基因。

包装细胞在组织培养物中生长，达到在含有 10%牛血清、青霉素、链霉素的 Dulbecco's Modified Eagles 培养基（DMEM）中的融合密度。用标准技术将携带有瘦素基因的载体导入包装细胞，给细胞添加新鲜的培养基，在孵育一段时间后，从融合包装细胞平板中回收培养基，将含有感染性病毒颗粒的培养基滤过一个微孔过滤器，移走分离的包装细胞用于感染成纤维细胞。将培养物从一个亚融合成纤维原细胞平板转移，迅速以过滤后的培养基取代。Polybrene（Aldrich）可以包含在培养基中加强转导。在合适的孵育后，去除培养基，并用新鲜的培养基代替。如果病毒的滴度高，那么基本所有的成纤维细胞将被转染，并不需要选择。如果滴度低，则有必要使用一个逆转录病毒载体，该逆转录病毒载体带有一个选择性标记，如 neo 或 his，通过该标记可以选出转导的细胞用于细胞的扩展。

单独将加工的成纤维细胞导入个体，或在微载体珠如 Cytodex 3 珠上生长至融合后再导入。注射入的成纤维细胞产生瘦素产物，蛋白的生物活性被传输进宿主。

可以选择或额外将分离的瘦素基因克隆进一个表达载体，用它来产生瘦素多肽。表达载体含有合适的转录和翻译起始区域，以及转录和起始终止区域，如前公开。可以用这种方式产生分离的瘦素蛋白并将该蛋白用于识别与其结合的试剂；或者用表达这种加工的瘦素基因和蛋白的细胞在化验中鉴别试剂。试剂的鉴别可以通过例如，将一个候选试剂和一个分离的瘦素多肽接触一段足够长的时间，从而能形成一个多肽/化合物复合物，对复合物进行检测。如果检测到了一个多肽/化合物复合物，则结合在瘦素多肽上的化合物就能鉴别出来。通过这种方法鉴别的因子包括调节瘦素活性的化合物。用这种方式筛选出的试剂包括肽、碳水化合物、维生素衍生物，和其他小分子或药用制剂。除了通过生物化验来鉴别试剂外，还可以通过利用基于瘦素蛋白结

构的蛋白模拟技术选择的候选试剂对试剂进行预筛选。

除了鉴别与瘦素蛋白结合的试剂外，对能通过结合瘦素基因来控制基因表达的序列特异性或元件特异性的试剂也进行了鉴别。一类核酸结合结合试剂含有能与瘦素 mRNA 杂交、从而抑制翻译的碱基残基（如反义寡核苷酸）。另一类核酸结合试剂是那些能与 DNA 形成三螺旋结构来阻断翻译的试剂（三螺旋寡核苷酸）。这种因子通常含有 20－40 个碱基，是基于磷酸二酯、核糖核酸骨架、或可能是具有碱基附着能力的各种巯基或聚合衍生物。

另外，能与瘦素基因特异性杂交的等位基因特异性寡核苷酸、和/或能与变异瘦素蛋白特异性结合的试剂（如一种变异体特异性的抗体）可以用作诊断剂。制备和使用等位基因特异性寡核苷酸的方法，以及制备抗体的方法如前所述，是现有技术已知的。

在本说明书中提及的所有专利和公开文献提示了本领域技术人员本发明涉及的水平。所有的专利和公开文献均在此引入作为参考，与每一篇公开文献中特别和单独引入作为参考的程度相同。

本发明通过鉴别个体变异、确定 SNP 单倍型区块、确定单倍型模式、和进一步使用单倍型模式鉴别信息型 SNPs，从而提供了进行基因组范围关联研究的很大程度改进的方法。信息型 SNPs 可以用于以先前未知的实用的和经济有效的方式进一步分析疾病和药物反应的遗传基础。以上的描述是为了更好地理解发明，而不是对发明作的限制。本领域技术人员看了以上的描述，可以清楚地理解许多实施方案。因此，本发明保护的范围不应由以上描述的内容得出，而应该参考所附权利要求书，以及这些权利要求声明请求保护的等同物的全部范围。

序列表

<110> 珀尔根科学公司

<120> 基因组分析方法

<130> 054801-5001

<150> US 60/280,530
<151> 2001-03-30

<150> US 60/313,264
<151> 2001-08-17

<150> US 60/327,006
<151> 2001-10-05

<150> US 60/332,550
<151> 2001-11-26

<160> 7

<170> PatentIn version 3.1

<210> 1
<211> 13
<212> DNA
<213> 人工序列

<220>
<223> 样本 SNP 单倍型: W

<400> 1
agattcgata acg 13

<210> 2
<211> 13
<212> DNA
<213> 人工序列

<220>
<223> 样本 SNP 单倍型: X

<400> 2
agactacata acg 13

<210> 3
<211> 13
<212> DNA
<213> 人工序列

<220>
<223> 样本 SNP 单倍型: Y

<400> 3
tatttcgata acg 13

<210> 4
<211> 13
<212> DNA
<213> 人工序列

<220>
<223> 样本 SNP 单倍型: Z

<400> 4
tatctacaat cac 13

<210> 5
<211> 13
<212> DNA
<213> 人工序列

<220>
<223> SNP 序列

<400> 5
agtaacccct ttt 13

<210> 6
<211> 13
<212> DNA
<213> 人工序列

<220>
<223> SNP 序列

<400> 6
actgacccct ttt 13

<210> 7
<211> 13
<212> DNA
<213> 人工序列

<220>
<223> SNP 序列

<400> 7
agtgactctt taa 13

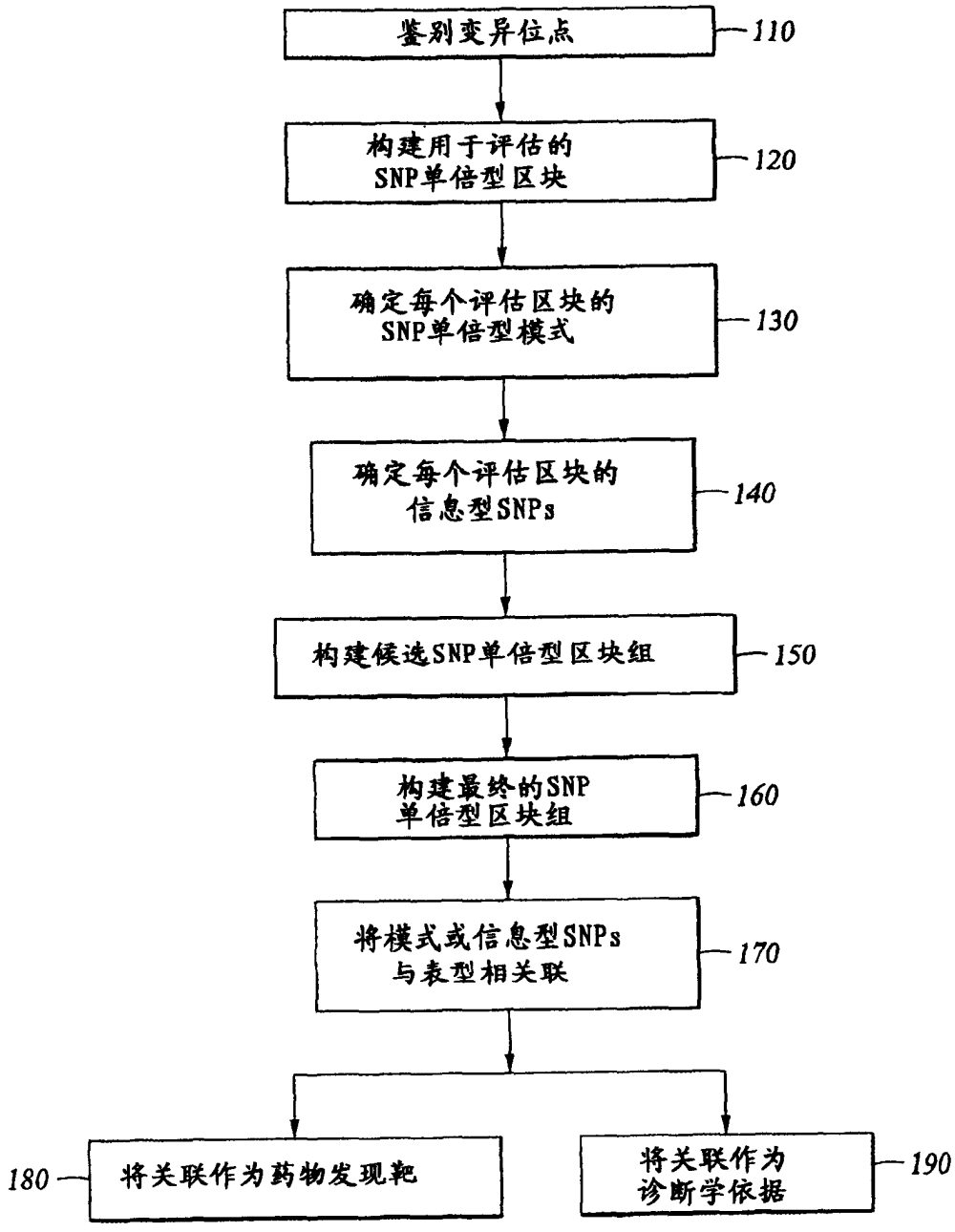

鉴别变异位点
110
构建用于评估的
SNP单倍型区块
120
确定每个评估区块的
SNP单倍型模式
130
确定每个评估区块的
信息型SNPs
140
构建候选SNP单倍型区块组
150
构建最终的SNP
单倍型区块组
160
将模式或信息型SNPs
与表型相关联
170
180
将关联作为药物发现靶
将关联作为
诊断学依据
190

图 1

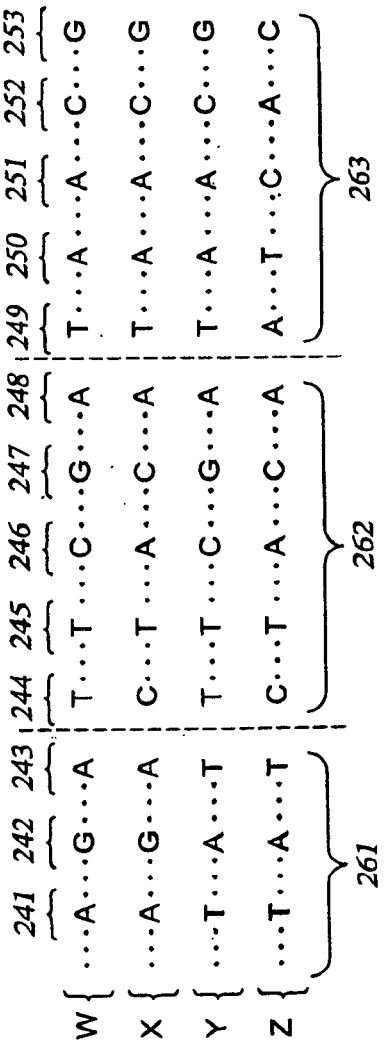

241 242 243 244 245 246 247 248 249 250 251 252 253
W ...A...G...A T...T...C...G...A T...A...A...C...G
X ...A...G...A C...T...A...C...A T...A...A...C...G
Y ...T...A...T T...T...C...G...A T...A...A...C...G
Z ...T...A...T C...T...A...C...A A...T...C...A...C
261
262
263

图 2

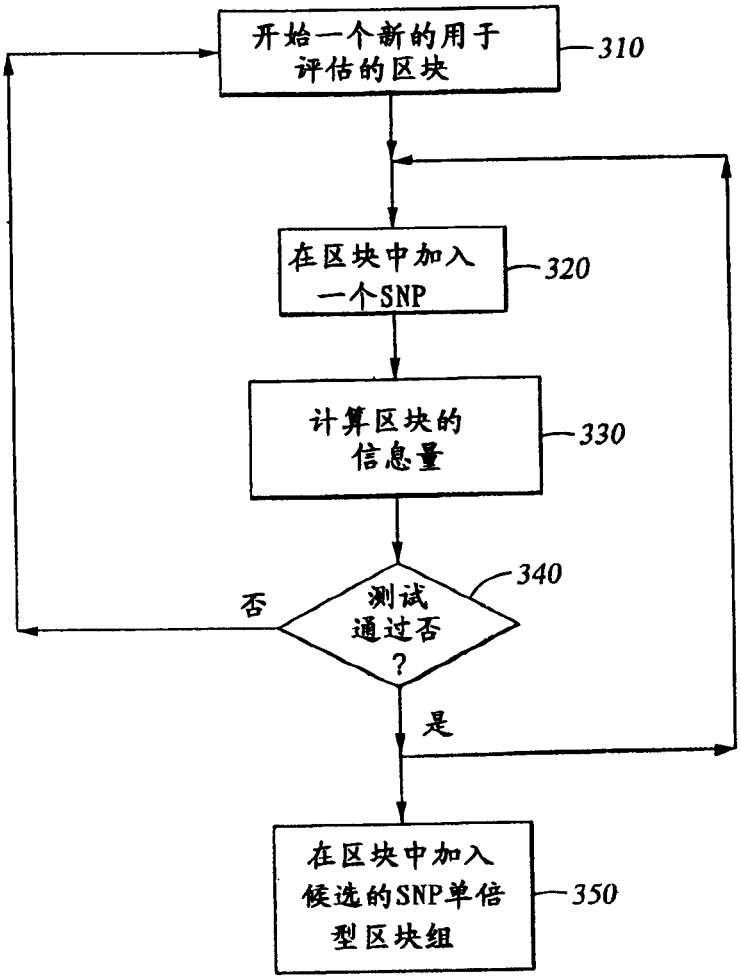

开始一个新的用于评估的区块
310
在区块中加入一个SNP
320
计算区块的信息量
330
测试通过否?
340
否
是
在区块中加入候选的SNP单倍型区块组
350

图 3

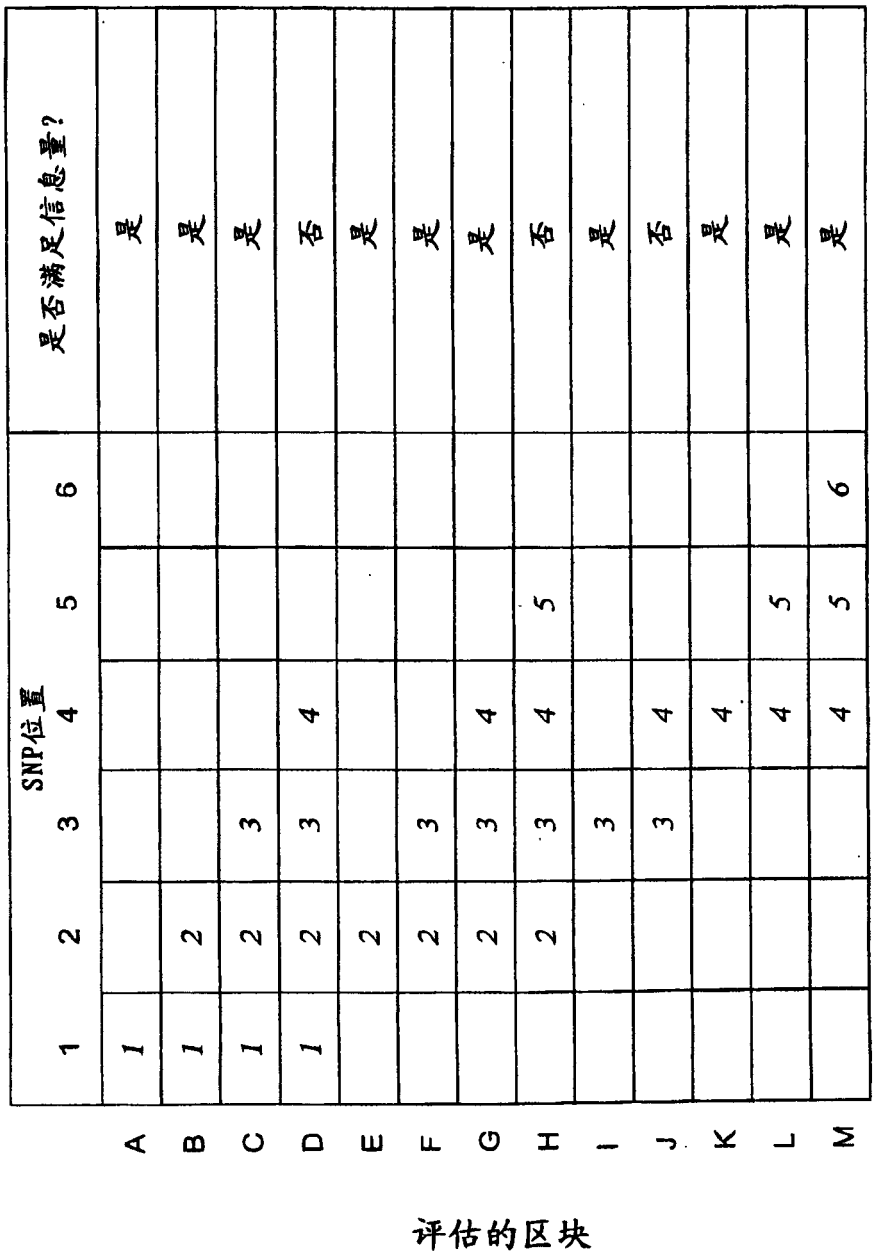

SNP位置
1 2 3 4 5 6
是否满足信息量?
A 1 是
B 1 2 是
C 1 2 3 是
D 1 2 3 4 否
E 2 是
F 2 3 是
G 2 3 4 是
H 2 3 4 5 否
I 3 是
J 3 4 否
K 4 是
L 4 5 是
M 4 5 6 是
评估的区块
选择作为候选组的区块: A B C E F G I K L M

图 4

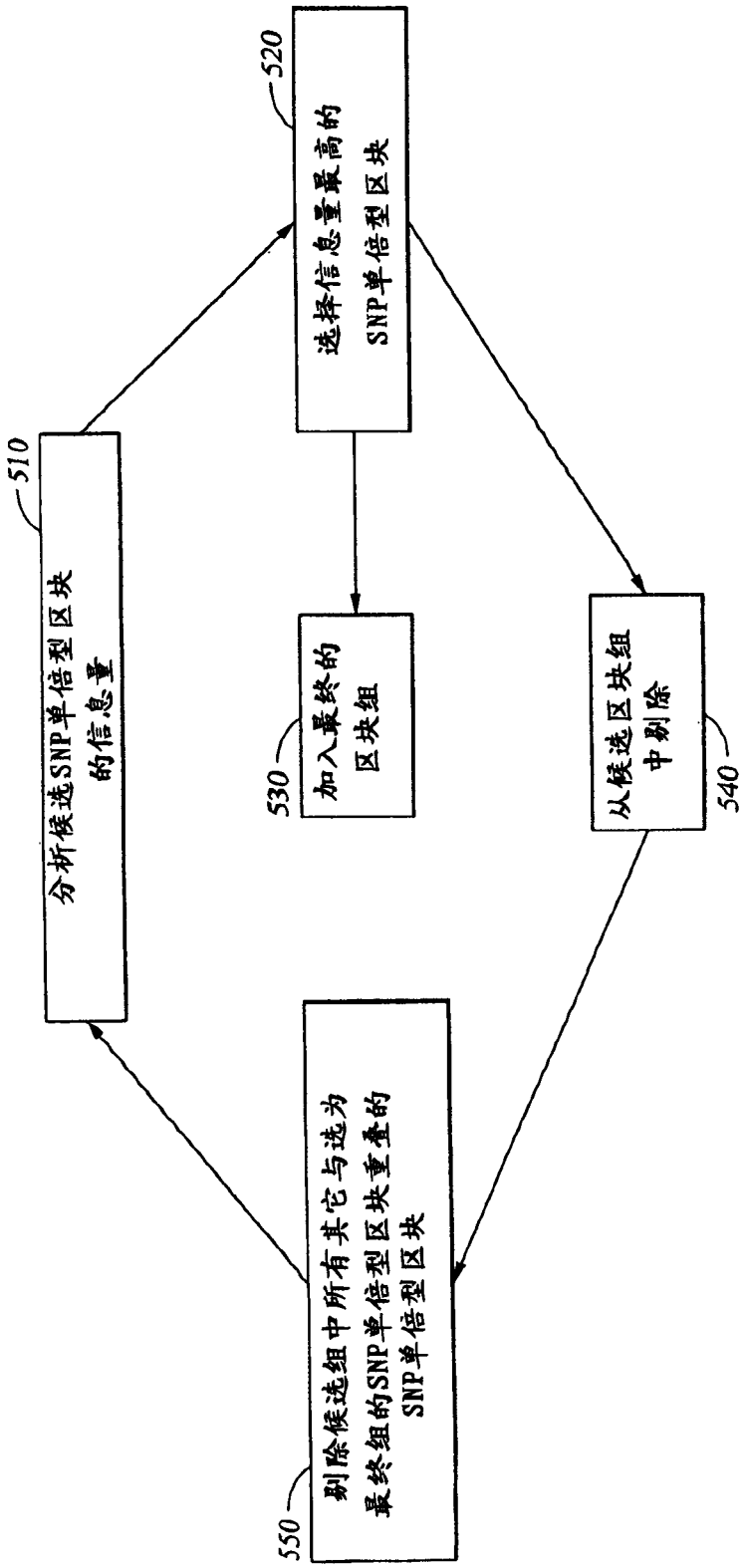

分析候选SNP单倍型区块的信息量
510
选择信息量最高的SNP单倍型区块
520
加入最终的区块组
530
从候选区块组中剔除
540
剔除候选组中所有其它与选为最终组的SNP单倍型区块重叠的SNP单倍型区块
550

图 5A

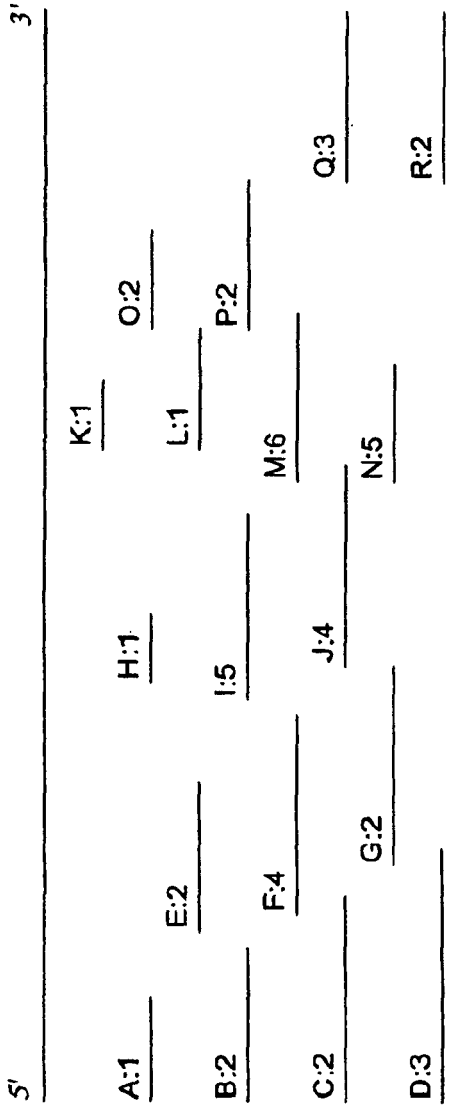

5'
3'
A:1
B:2
C:2
D:3
E:2
F:4
G:2
H:1
I:5
J:4
K:1
L:1
M:6
N:5
O:2
P:2
Q:3
R:2
M 放弃 J,N,K,L,O & P
I 放弃 H
F 放弃 E,G,C & D
Q 放弃 R
B 放弃 A

图 5B

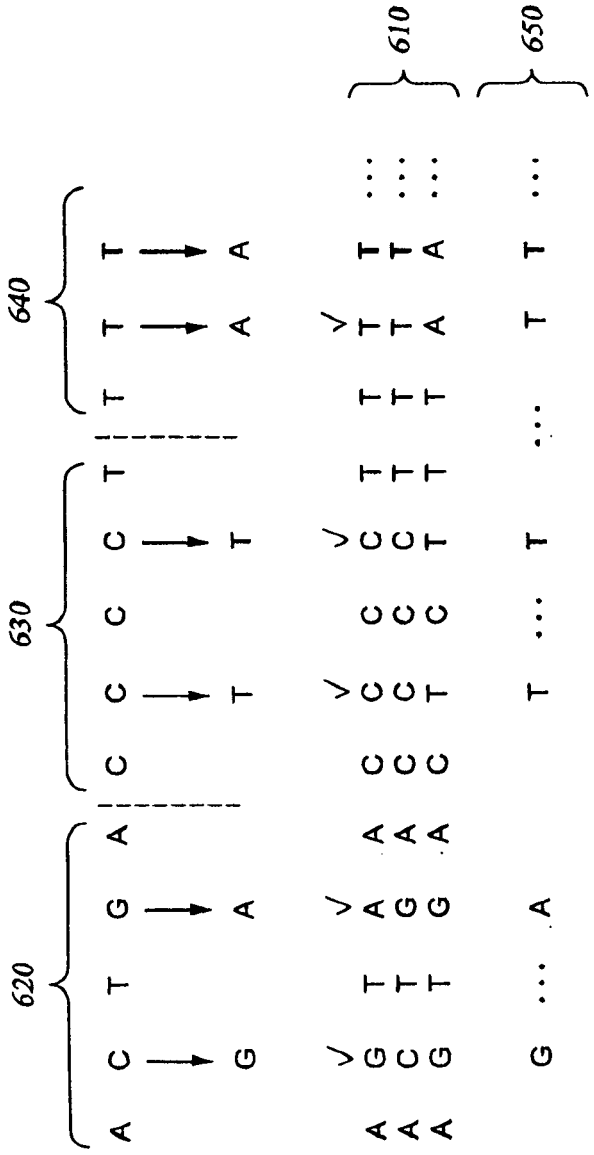


图 6

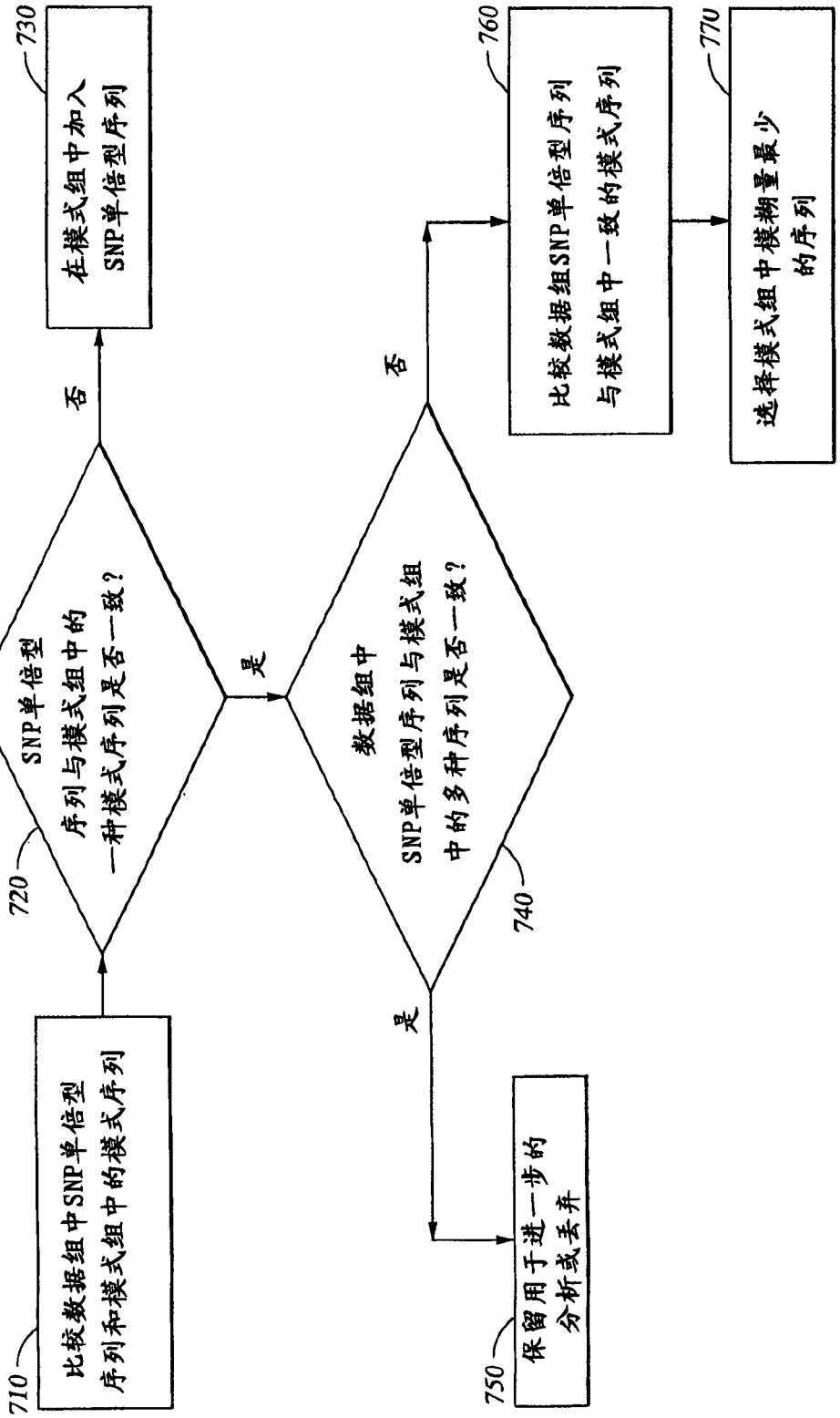

比较数据组中SNP单倍型序列和模式组中的模式序列
710
SNP单倍型序列与模式组中的一种模式序列是否一致？
720
是
否
在模式组中加入SNP单倍型序列
730
数据组中SNP单倍型序列与模式组中的多种序列是否一致？
740
是
否
保留用于进一步的分析或丢弃
750
比较数据组SNP单倍型序列与模式组中一致的模式序列
760
选择模式组中模糊量最少的序列
770

图 7A

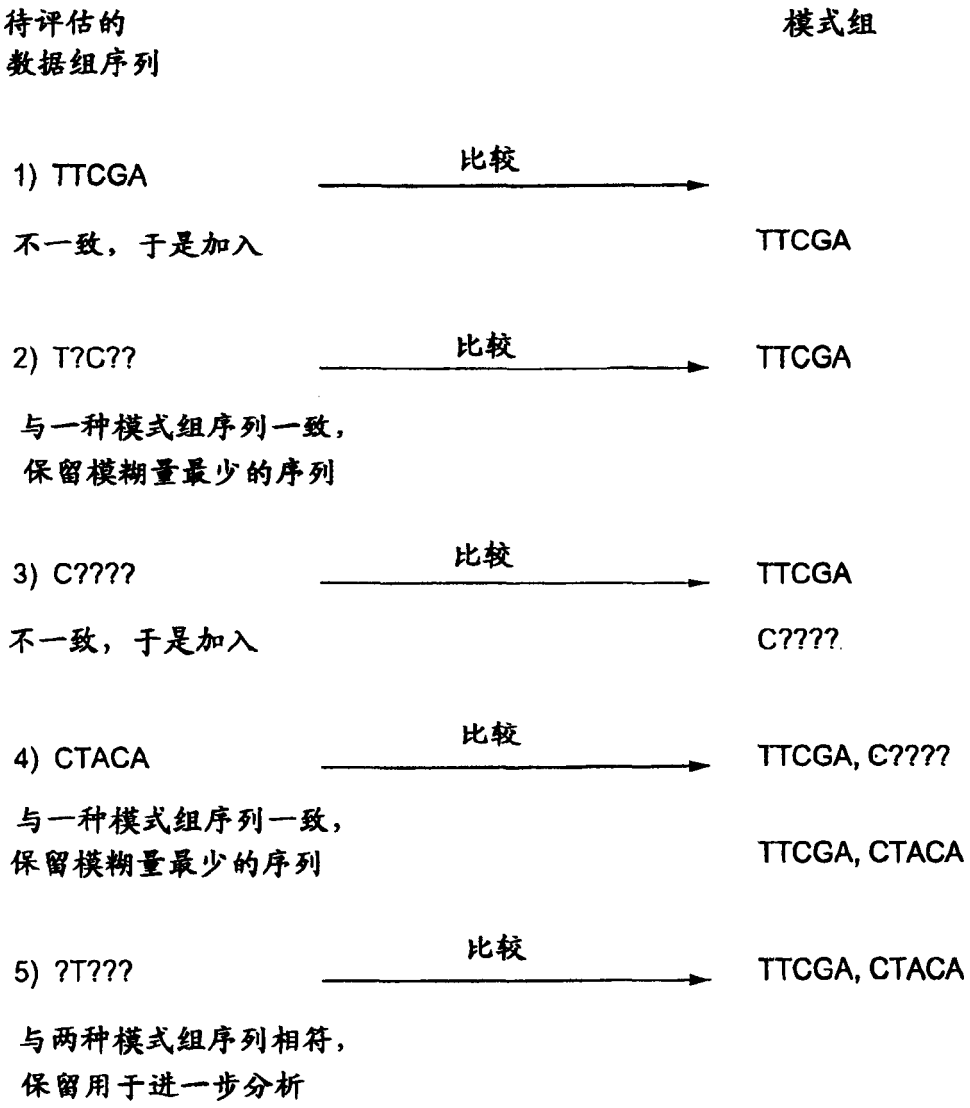

待评估的
数据组序列
模式组
1) TTCGA
比较
不一致，于是加入
TTCGA
2) T?C??
比较
TTCGA
与一种模式组序列一致，
保留模糊量最少的序列
3) C????
比较
TTCGA
不一致，于是加入
C????
4) CTACA
比较
TTCGA, C????
与一种模式组序列一致，
保留模糊量最少的序列
TTCGA, CTACA
5) ?T???
比较
TTCGA, CTACA
与两种模式组序列相符，
保留用于进一步分析

图 7B

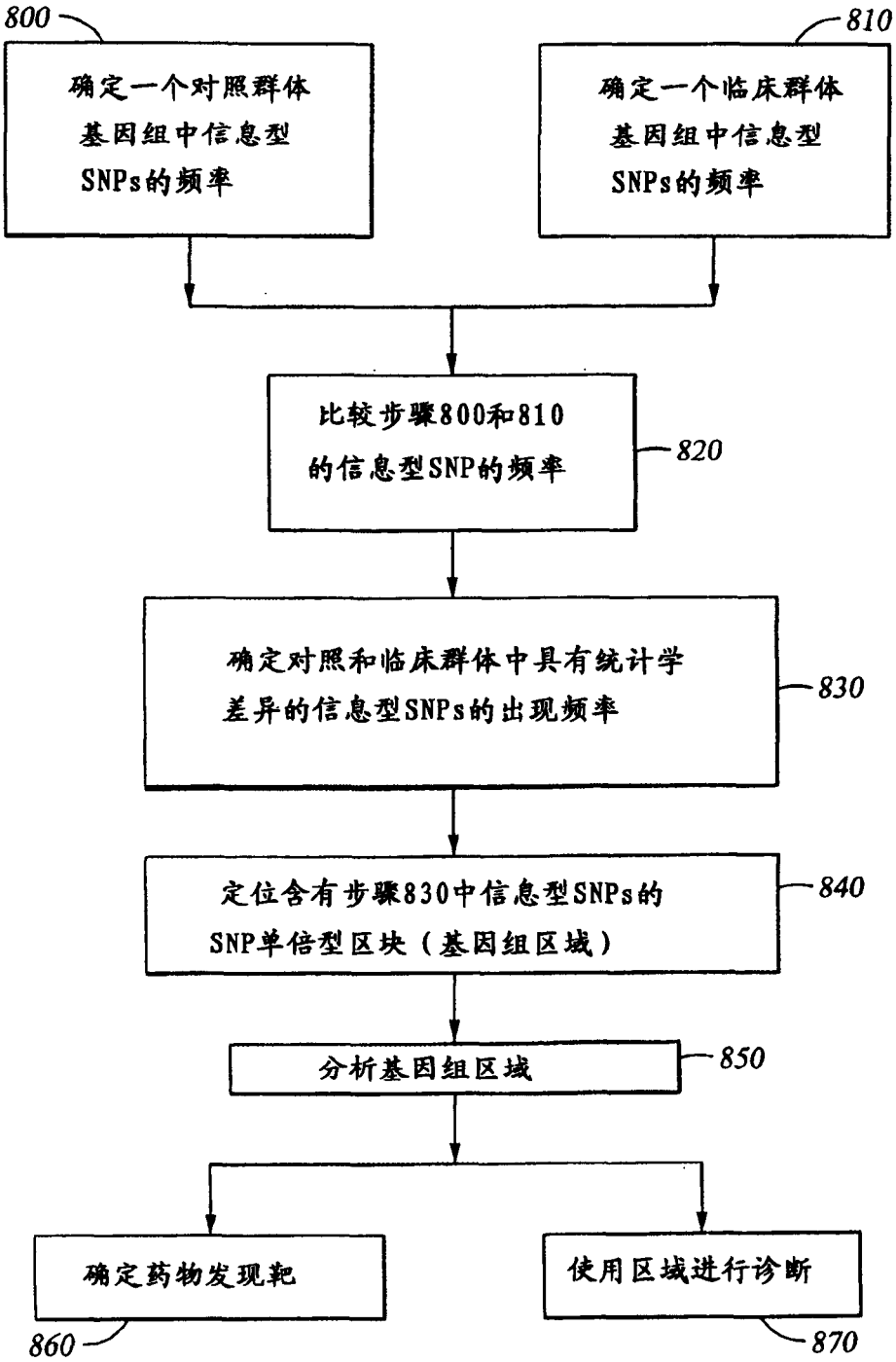

800
确定一个对照群体基因组中信息型SNPs的频率
810
确定一个临床群体基因组中信息型SNPs的频率
比较步骤800和810的信息型SNP的频率
820
确定对照和临床群体中具有统计学差异的信息型SNPs的出现频率
830
定位含有步骤830中信息型SNPs的SNP单倍型区块（基因组区域）
840
分析基因组区域
850
确定药物发现靶
860
使用区域进行诊断
870

图 8

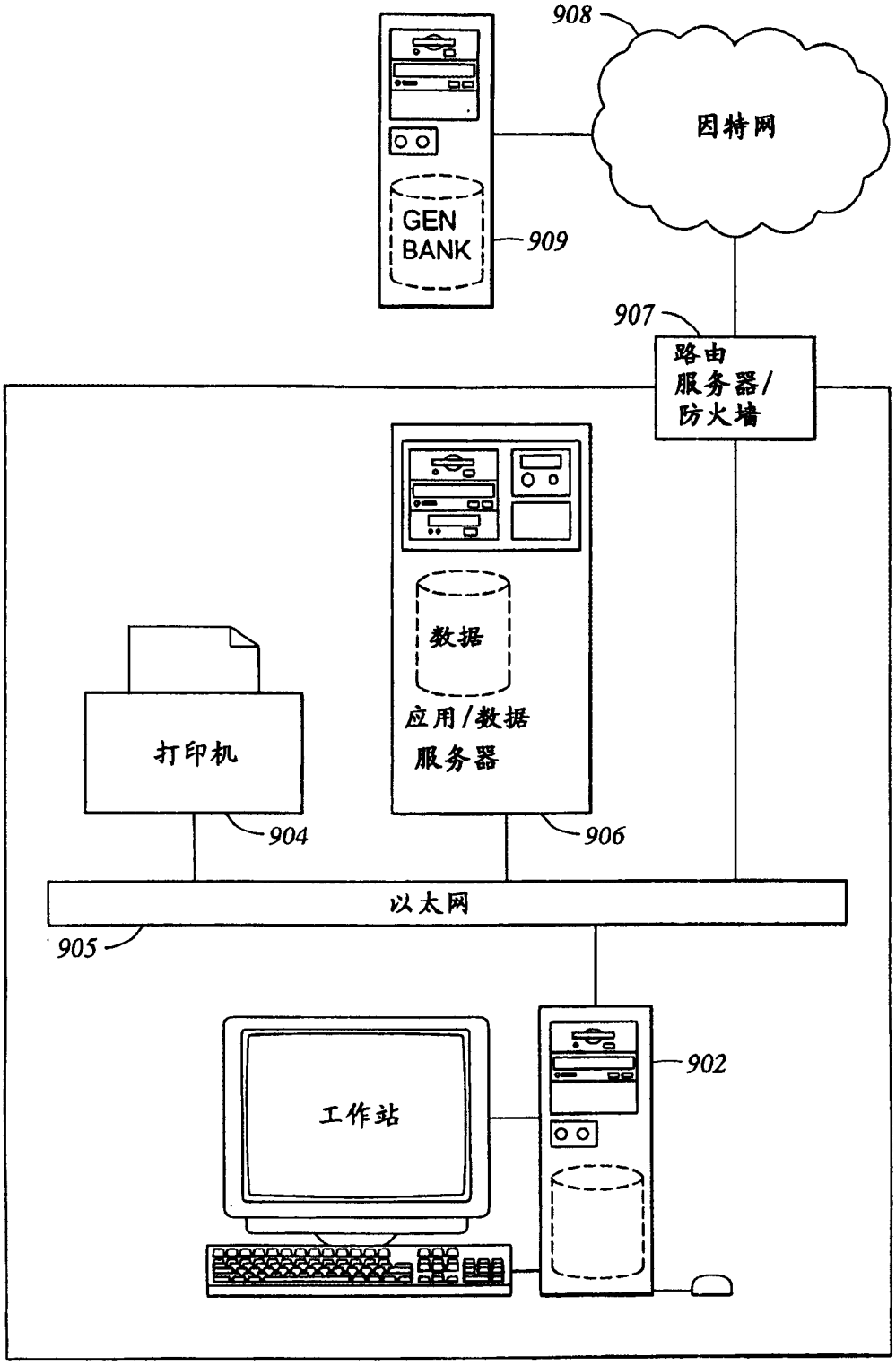

908
因特网
GEN
BANK
909
907
路由
服务器/
防火墙
数据
应用/数据
服务器
打印机
904
906
以太网
905
工作站
902

图 9

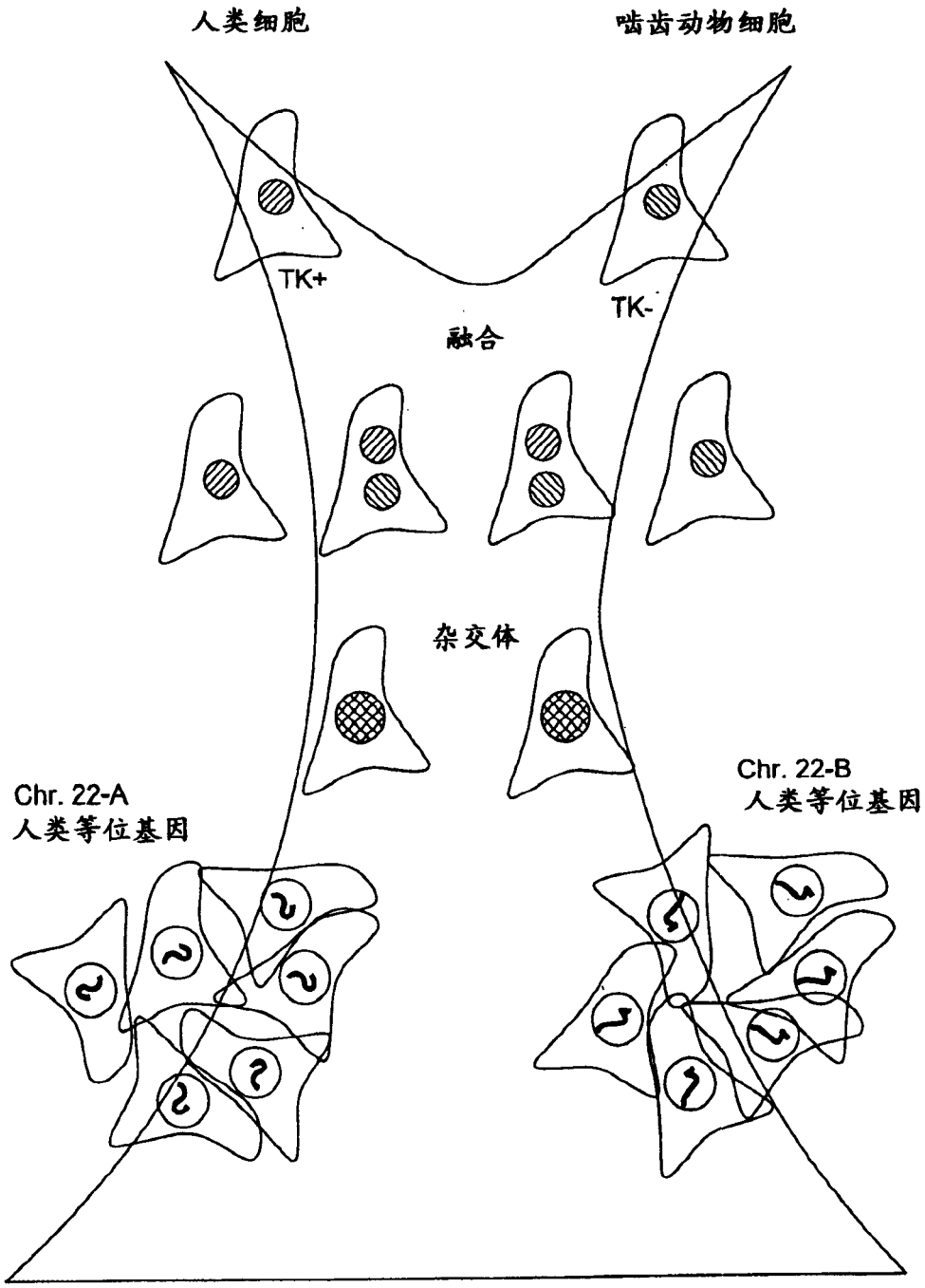


图 10

人类21号染色体上的SNP标记	仓鼠	CPD17	杂交体1	杂交体2
WIAF-3497	无信号	A	A	A
WIAF-3498	无信号	AB	A	B
WIAF-599	无信号	A	A	A
WIAF-3562	无信号	无信号	A	B
WIAF-559	无信号	AB	B	A
WIAF-4546	无信号	AB	B	A
WIAF-3508	无信号	B	B	B
WIAF-624	无信号	B	B	B
WIAF-1500	无信号	A	A	A
WIAF-3496	无信号	AB	A	B
WIAF-1943	无信号	A	A	A
WIAF-2477	无信号	无信号	无信号	A
WIAF-1538	无信号	B	无信号	B
WIAF-3479	无信号	A	A	无信号
WIAF-2436	无信号	A	A	A
WIAF-1857	无信号	AB	B	A
WIAF-899	无信号	AB	A	B
WIAF-1682	无信号	B	B	B
WIAF-2214	无信号	AB	A	B
WIAF-2643	无信号	无信号	A	无信号
WIAF-4514	无信号	B	B	B

图 11

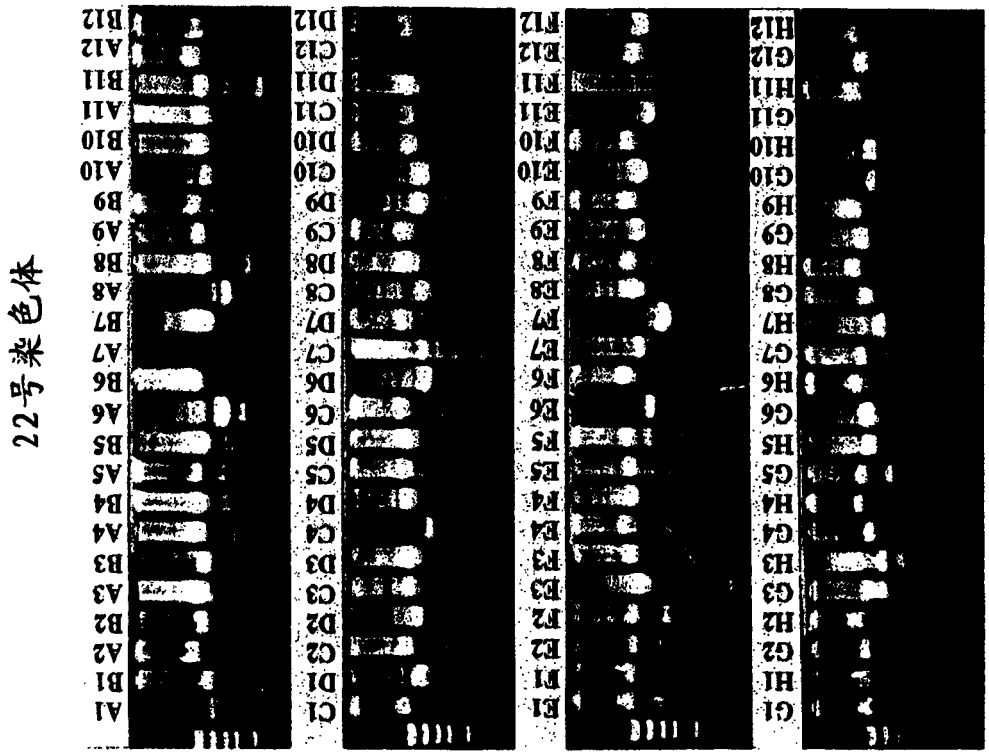
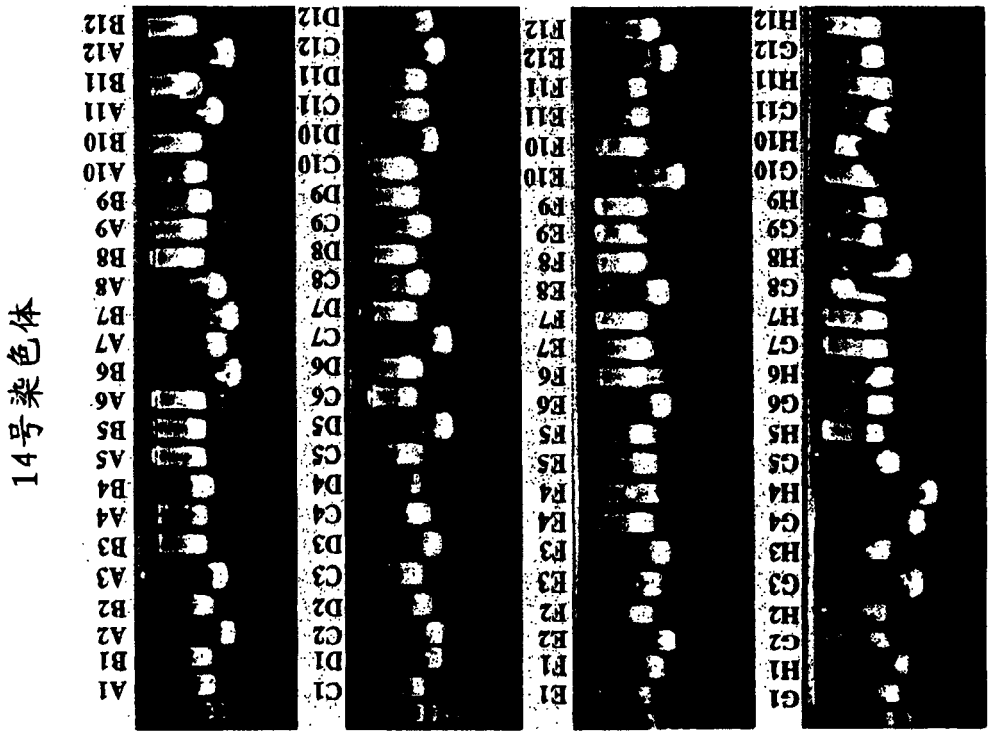

22号染色体
A1 B1 A2 B2 A3 B3 A4 B4 A5 B5 A6 B6 A7 B7 A8 B8 A9 B9 A10 B10 A11 B11 A12 B12
C1 D1 C2 D2 C3 D3 C4 D4 C5 D5 C6 D6 C7 D7 C8 D8 C9 D9 C10 D10 C11 D11 C12 D12
E1 F1 E2 F2 E3 F3 E4 F4 E5 F5 E6 F6 E7 F7 E8 F8 E9 F9 E10 F10 E11 F11 E12 F12
G1 H1 G2 H2 G3 H3 G4 H4 G5 H5 G6 H6 G7 H7 G8 H8 G9 H9 G10 H10 G11 H11 G12 H12

图 12


14号染色体
A1 B1 A2 B2 A3 B3 A4 B4 A5 B5 A6 B6 A7 B7 A8 B8 A9 B9 A10 B10 A11 B11 A12 B12
C1 D1 C2 D2 C3 D3 C4 D4 C5 D5 C6 D6 C7 D7 C8 D8 C9 D9 C10 D10 C11 D11 C12 D12
E1 F1 E2 F2 E3 F3 E4 F4 E5 F5 E6 F6 E7 F7 E8 F8 E9 F9 E10 F10 E11 F11 E12 F12
G1 H1 G2 H2 G3 H3 G4 H4 G5 H5 G6 H6 G7 H7 G8 H8 G9 H9 G10 H10 G11 H11 G12 H12

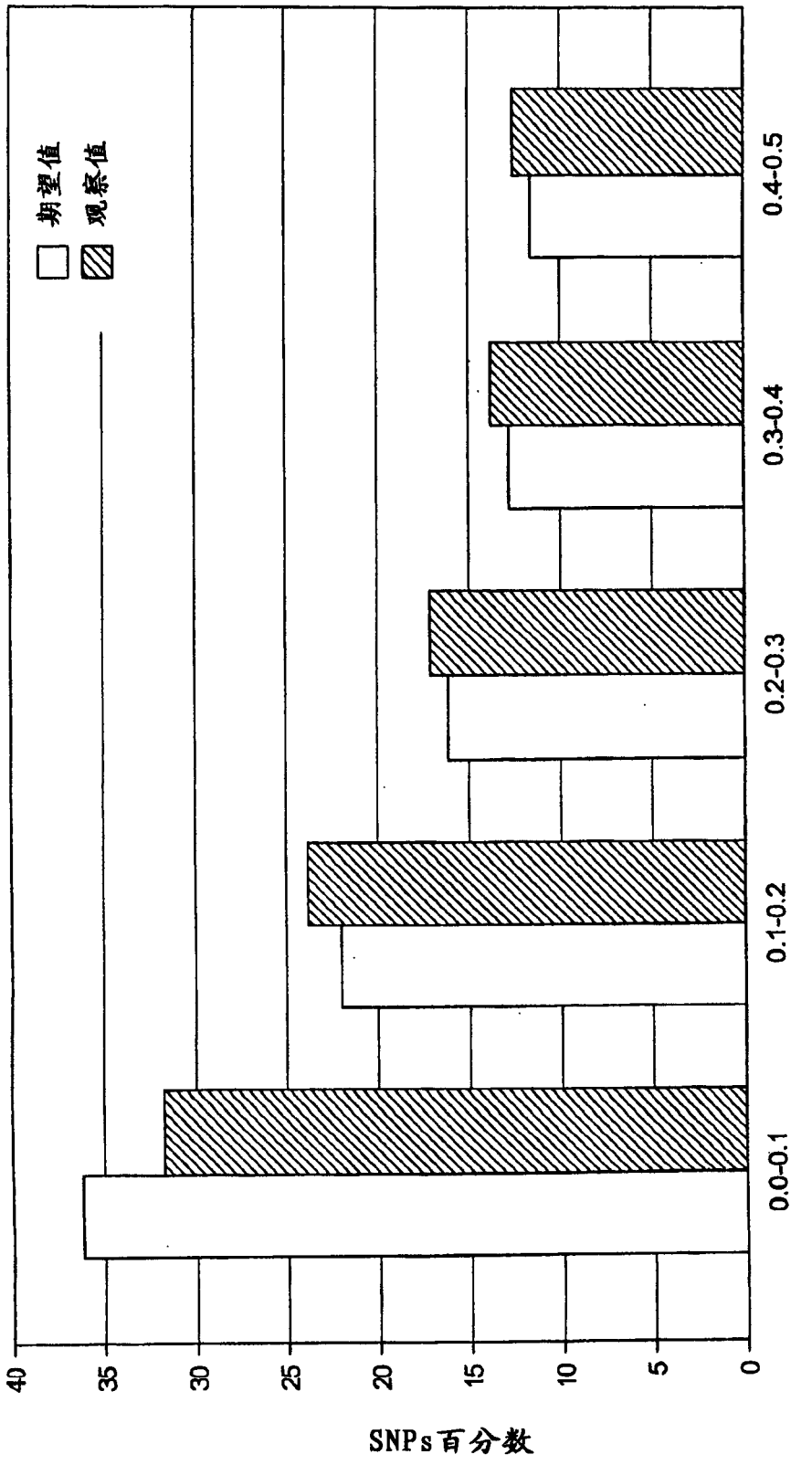

期望值
观察值
0.0-0.1
0.1-0.2
0.2-0.3
0.3-0.4
0.4-0.5
次要等位基因频率
0
5
10
15
20
25
30
35
40
SNPs百分数

图 13A

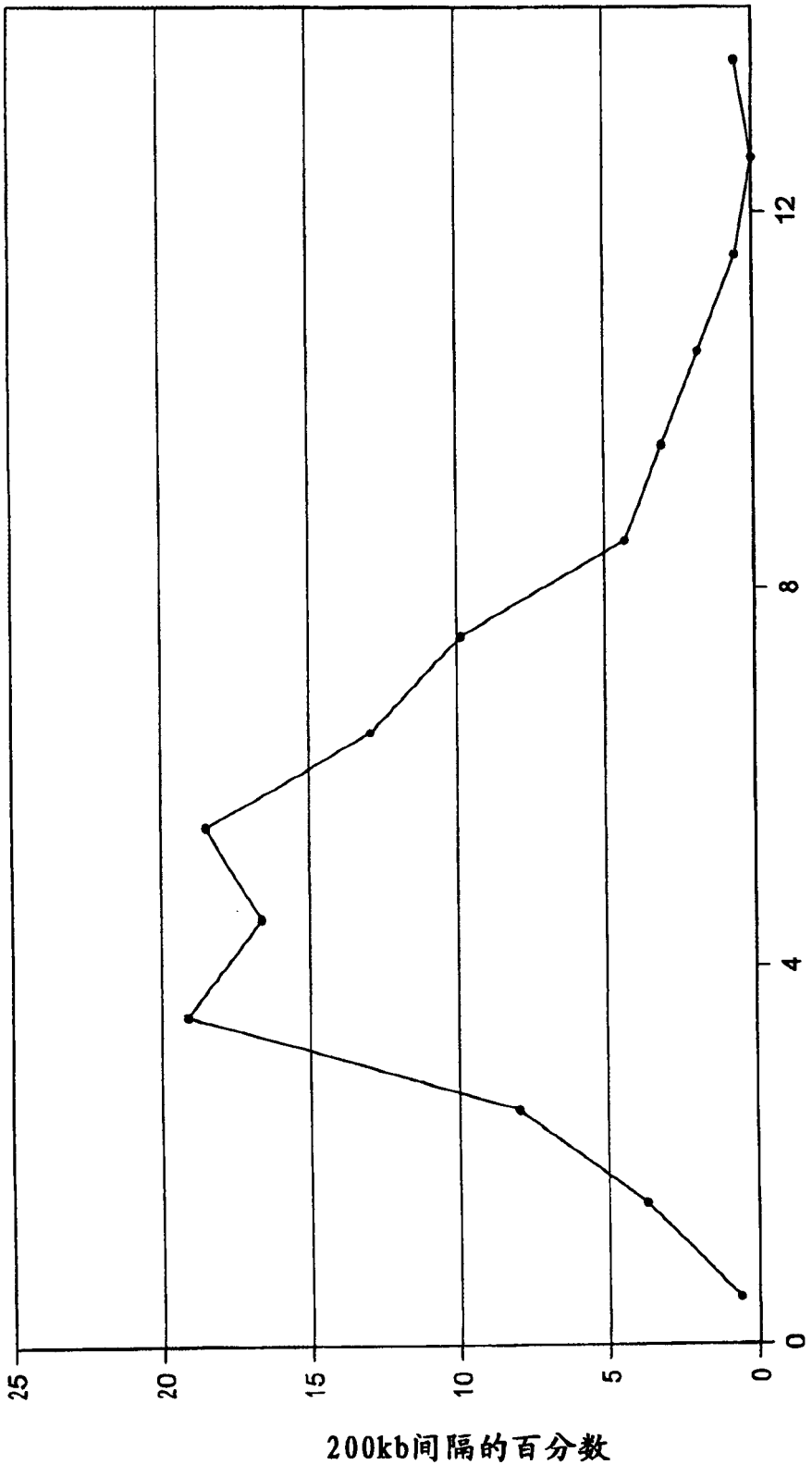


图 13B

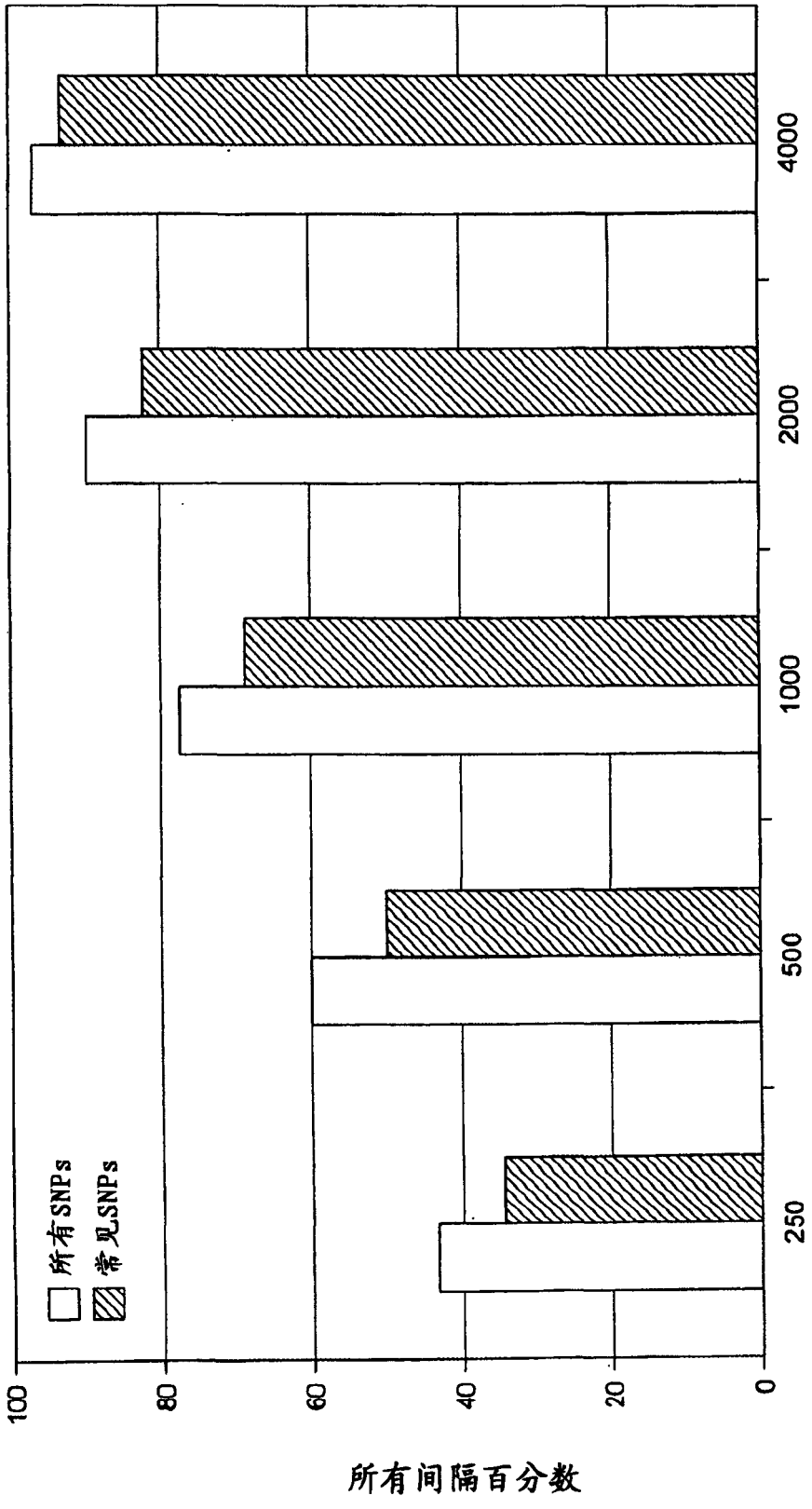

所有SNPs
常见SNPs
100
80
60
40
20
0
250
500
1000
2000
4000
最大间隔长度（bp）
所有间隔百分数

图 13C

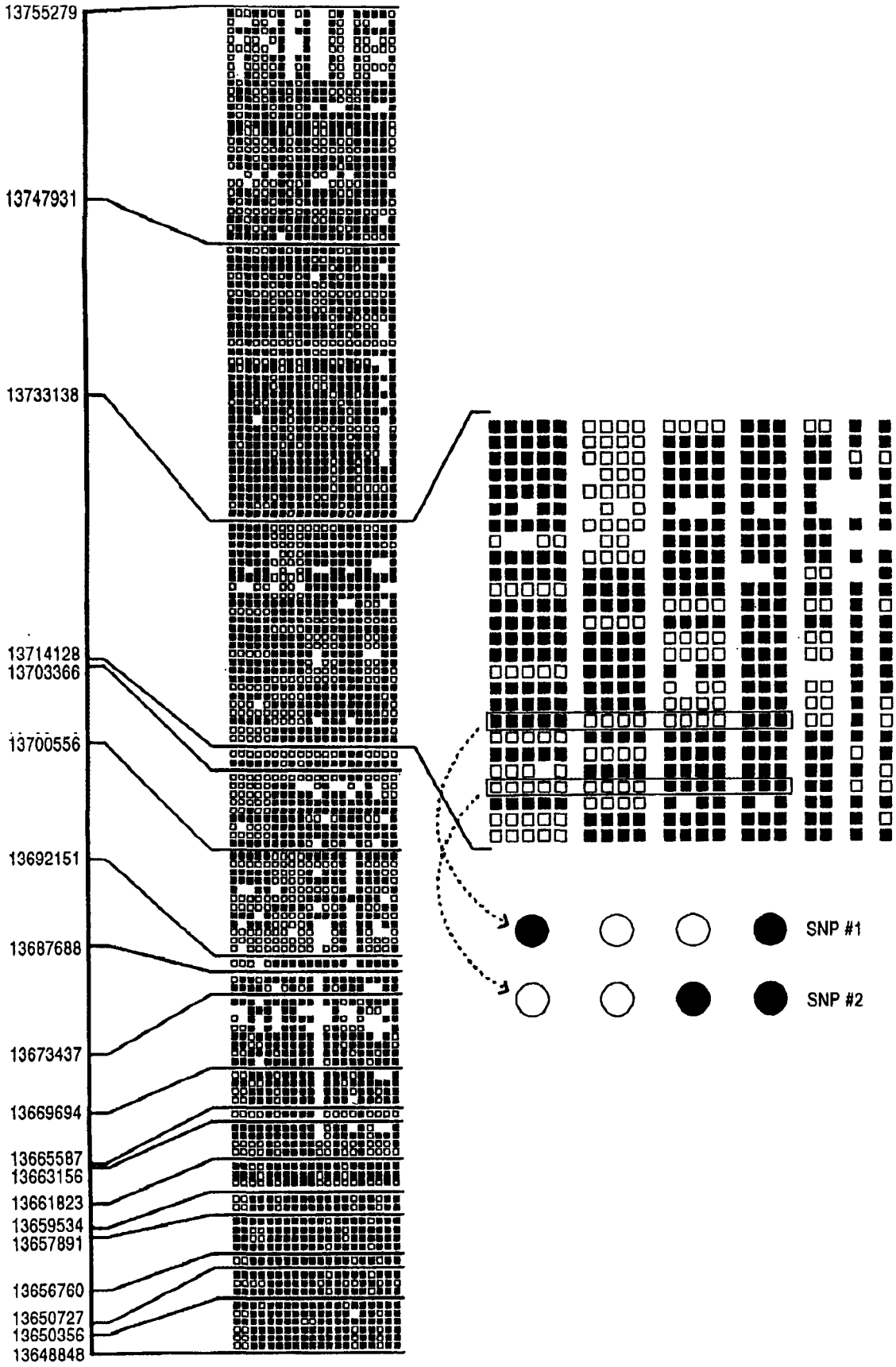

13755279
13747931
13733138
13714128
13703366
13700556
13692151
13687688
13673437
13669694
13665587
13663156
13661823
13659534
13657891
13656760
13650727
13650356
13648848
SNP #1
SNP #2

图 14

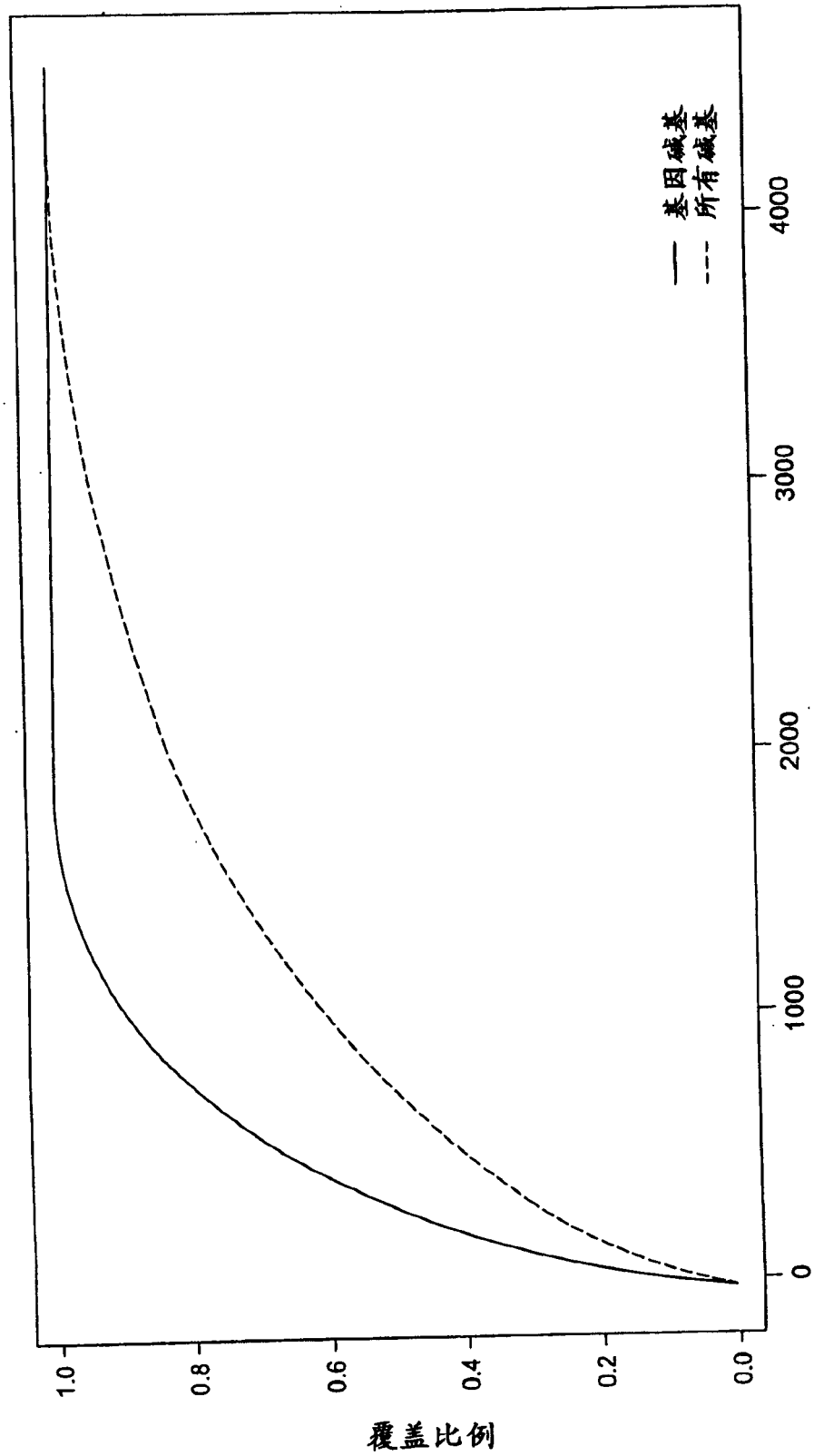


图 15

专利名称(译)	基因组分析方法		
公开(公告)号	CN1381591A	公开(公告)日	2002-11-27
申请号	CN02119281.2	申请日	2002-03-29
[标]发明人	N帕蒂尔 DR科克斯 AJ贝尔诺 DA海因兹		
发明人	N·帕蒂尔 D·R·科克斯 A·J·贝尔诺 D·A·海因兹		
IPC分类号	G01N33/53 C07H21/00 C07H21/02 C07H21/04 C12N15/09 C12P19/34 G01N33/00 G01N33/566 G06F19/22 G06F19/24 C12Q1/68		
CPC分类号	G16B30/00 C12Q1/6827 G16B40/00 Y10T436/143333		
代理人(译)	刘玥		
优先权	60/280530 2001-03-30 US 60/313264 2001-08-17 US 60/327006 2001-10-05 US 60/332550 2001-11-26 US		
外部链接	Espacenet SIPO		

摘要(译)

本发明涉及鉴别人类基因组发生的变异的方法,及将这些变异与疾病和药物反应的遗传基础相关联的方法。具体地说,本发明涉及鉴别个体SNPs、确定SNP单倍型区块和模式,和进一步利用SNP单倍型区块和模式来分析疾病和药物反应的遗传基础。本发明的方法适用于全基因组的分析。