(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 110878345 A
(43)申请公布日 2020.03.13

(21)申请号 201910945228.0

(22)申请日 2011.09.20

(30)优先权数据
61/385,001 2010.09.21 US
61/432,119 2011.01.12 US

(62)分案原申请数据
201180049727.3 2011.09.20

(71)申请人 安捷伦科技有限公司
地址 美国加利福尼亚州

(72)发明人 J·卡斯本 S·布伦那
R·奥斯本 C·利希藤斯坦
A·克拉斯

(74)专利代理机构 北京坤瑞律师事务所 11494
代理人 罗天乐

(51)Int.Cl.
C12Q 1/6869(2018.01)

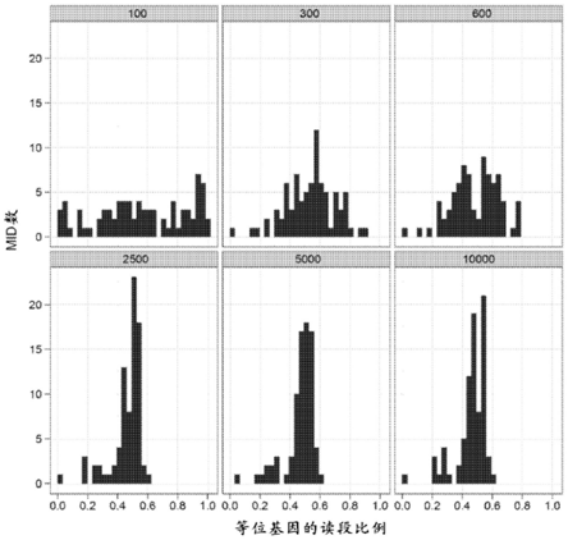
权利要求书1页 说明书24页
序列表1页 附图3页

(54)发明名称

通过分子计数提高等位基因调用的置信度

(57)摘要

本申请涉及通过分子计数提高等位基因调用的置信度。本发明的方面包括用于确定个体多核苷酸分子的数量的方法和组合物，所述个体多核苷酸分子来源于已在特定序列分析配置或过程中测序的相同原始样品的相同基因组区域。在本发明的这些方面，将简并碱基区域(DBR)附接到随后将进行测序(例如，在进行某些过程步骤后，例如扩增和/或富集步骤后)的起始多核苷酸分子。存在于测序运行中的不同DBR序列的数量可用于确定/估计已进行测序的不同起始多核苷酸的数量。DBR可用于增强许多不同的核酸序列分析应用，包括允许在基因分型应用中实现更高置信度的等位基因调用测定。



CN 110878345 A

1.一种用于确定个体多核苷酸分子的最小数量的方法,所述个体多核苷酸分子来源于已在特定序列分析配置或过程中测序的相同原始样品的相同基因组区域,该方法包括:

将简并碱基区域(DBR)附接到起始多核苷酸分子;

扩增衔接子附接的起始多核苷酸分子;

将扩增的多核苷酸分子测序,其中获得所述DBR及多核苷酸的一部分的序列;

确定附接到感兴趣多核苷酸的不同DBR的数量;

利用存在于测序运行中的不同DBR序列的数量来确定来源于已在该特定序列分析配置或过程中测序的相同原始样品的相同基因组区域的个体多核苷酸分子的最小数量;和

在基因分型测定法中确定无法单独通过读段数量获得的等位基因调用统计值。

2.如权利要求1所述的方法,其中DBR作为衔接子的一部分添加到起始多核苷酸。

3.如权利要求2所述的方法,其中衔接子还包含测序引物位点。

4.如权利要求1所述的方法,其中DBR存在于核酸合成引物中,使得当将引物用于聚合反应时将DBR添加到靶核苷酸。

5.如权利要求4所述的方法,其中所述核酸合成引物是PCR引物。

6.如前述任一项权利要求所述的方法,其中DBR的长度为3至10个碱基。

7.如前述任一项权利要求所述的方法,其中所述多核苷酸来自基因组DNA。

8.如前述任一项权利要求所述的方法,包括富集DBR附接的多核苷酸。

9.如前述任一项权利要求所述的方法,其中该方法包括将来源于多个原始来源的多核苷酸分子混合,其中来源于每个原始样品的多核苷酸分子包括多重标识(MID)标记,其中将每个原始样品与单一MID相关联,使得可以确定每个标记核酸片段所来源的原始样品。

通过分子计数提高等位基因调用的置信度

[0001] 本申请是2011年9月20日提交的申请号为201180049727.3(PCT申请号为PCT/IB2011/003160)、发明名称为“通过分子计数提高等位基因调用的置信度”的发明专利申请的分案申请。

[0002] 发明背景

[0003] 基因分型是遗传研究中用于对基因组作图以及对与遗传性状(如,遗传疾病)有关的基因进行定位的一项重要技术。受试者的基因分型通常包括基于从受试者DNA获得的测序数据确定一个或多个基因组座位的等位基因。二倍体基因组(例如人基因组)可归类为例如在基因组座位上纯合的或杂合的,具体取决于它们针对该基因座所具有的不同等位基因的数量,其中杂合个体具有针对该基因座的两个不同的等位基因,而纯合个体具有针对该基因座的相同等位基因的两个拷贝。当在需要以高统计置信度将基因型与表型相关联的大群中进行研究时,对样品作出正确的基因分型至关重要。

[0004] 在通过测序对二倍体基因组进行的基因分型分析中,将特定基因组座位的覆盖度(测序读段(sequencing reads)的数量)用于确定等位基因调用的置信度。然而,当在样品制备过程中引入偏差时,例如,当起始样品的量有限时和/或当将一个或多个扩增反应用于制备测序样品时,等位基因调用的置信度将显著降低。因此,在具有有限量DNA的样品中,由于扩增偏差(例如,仅从很少的或甚至一个多核苷酸分子扩增),可以观察到一条染色体上等位基因的覆盖度高于(即,高测序读段数量)不同染色体上等位基因的覆盖度。在这种情况下,当度量等位基因调用中的置信度时,仅依靠覆盖度可能会有误导。

[0005] 本发明可基于核酸序列分析用于提高等位基因调用以及其他应用中的置信度,尤其是在大群样品中研究基因型的背景下。

[0006] 发明概述

[0007] 本发明的方面包括用于确定个体多核苷酸分子的数量的方法和组合物,该个体多核苷酸分子来源于已在特定序列分析配置或过程中测序的相同原始样品的相同基因组区域。在本发明的这些方面,将简并碱基区域(DBR)附接到随后将进行测序(例如,在进行某些过程步骤后,例如扩增和/或富集步骤后)的起始多核苷酸分子。存在于测序运行中的不同DBR序列的数量可用于确定/估计个体多核苷酸分子的数量,该个体多核苷酸分子来源于已在特定序列分析配置或过程中测序的相同原始样品的相同基因组区域。DBR可用于改善许多不同核酸测序应用的分析。例如,DBR使得能够在基因分型测定法中确定无法单独通过读段数量获得的等位基因调用统计值。

[0008] 在某些实施方案中,本发明的方面涉及用于确定多个不同样品中所测序的起始多核苷酸分子的数量的方法。在某些实施方案中,该方法包括:(1)将衔接子附接到多个不同样品中的起始多核苷酸分子,其中每个样品的衔接子包含:该样品特定的单一MID;和简并碱基区域(DBR)(例如,具有选自以下的至少一个核苷酸碱基的DBR:R、Y、S、W、K、M、B、D、H、V、N及其修饰形式;(2)将多个不同的衔接子附接的样品混合以生成混合样品;(3)扩增混合样品中的衔接子附接的多核苷酸;(4)对多个扩增的衔接子附接的多核苷酸测序,其中获得该多个衔接子附接的多核苷酸的每一个的MID、DBR和多核苷酸至少一部分的序列;以及(5)确

定存在于来自每个样品的该多个测序衔接子附接的多核苷酸中的不同DBR序列的数量，以确定或估计已在测序步骤中测序的每个样品中的起始多核苷酸的数量。

[0009] 附图简述

[0010] 本发明通过以下详细说明在结合附图阅读时可以得到最好的理解。附图中包括以下各图：

[0011] 图1显示了由指定量的起始材料(各图的顶部;以纳克为单位)制备的样品中各MID的等位基因比。

[0012] 图2显示了在合成多态性位置与每个等位基因相关的各MID的DBR序列的分数。样品由指定量的起始材料(各图的顶部;以纳克为单位)制备。

[0013] 图3显示了使用具有DBR序列的引物在PCR的前两个循环中产生的产物。

[0014] 定义

[0015] 除非另有定义，否则本文所用的所有技术和科学术语具有与本发明所属领域的普通技术人员通常所理解的相同含义。另外，为了清楚和方便参考起见，下文将定义某些要素。

[0016] 本文所用的核酸化学、生物化学、遗传学和分子生物学的术语和符号遵循本领域中的标准专著和文献中的那些术语和符号，例如:Kornberg和Baker,DNA Replication,第二版(W.H.Freeman,New York,1992);Lehninger,Biochemistry,第二版(Worth Publishers,New York,1975);Strachan和Read,Human Molecular Genetics,第二版(Wiley-Liss,New York,1999);Eckstein编辑,Oligonucleotides and Analogs:A Practical Approach(Oxford University Press,New York,1991);Gait编辑,Oligonucleotide Synthesis:A Practical Approach(IRL Press,Oxford,1984)等等。

[0017] “扩增子”是指多核苷酸扩增反应的产物。也就是说，其为一群从一个或多个起始序列复制的通常为双链的多核苷酸。该一个或多个起始序列可以是相同序列的一个或多个拷贝，或者其可以是不同序列的混合物。扩增子可以通过多种扩增反应产生，这些反应的产物是一个或多个靶核酸的多个复制物。一般来讲，产生扩增子的扩增反应为“模板驱动的”，因为反应物(核苷酸或寡核苷酸)的碱基配对在形成反应产物所需的模板多核苷酸中具有互补序列。在一个方面，模板驱动的反应为通过核酸聚合酶进行的引物延伸或通过核酸连接酶进行的寡核苷酸连接。此类反应包括但不限于以下参考文献中所公开的聚合酶链反应(PCR)、线性聚合酶反应、核酸序列依赖性扩增(NASBA)、滚环扩增等等:Mullis等，美国专利4,683,195、4,965,188、4,683,202、4,800,159(PCR);Gelfand等，美国专利5,210,015(通过“TAQMAN™”探针进行的实时PCR);Wittwer等，美国专利6,174,670;Kacian等，美国专利5,399,491(“NASBA”);Lizardi，美国专利5,854,033;Aono等，日本专利公告JP 4-262799(滚环扩增)等，这些参考文献以引用方式并入本文。在一个方面，本发明的扩增子通过PCR产生。如果可以采用允许随着扩增反应的进行对反应产物进行测量的检测化学方法，则扩增反应可以是“实时”扩增，例如，如下所述的“实时PCR”，或Leone等,Nucleic Acids Research,26:2150-2155(1998)以及相似参考文献中所述的“实时NASBA”。如本文所用，术语“扩增”是指进行扩增反应。“反应混合物”是指含有进行反应所需的所有反应物的溶液，这些反应物可包括但不限于在反应中将pH维持在所选水平的缓冲剂、盐、辅因子、清除剂等。

[0018] 术语“评估”包括任何形式的测量，并包括确定某要素是否存在。术语“确定”、“测量”、“评价”、“估计”、“评估”和“测定”可互换使用，并包括定量和定性确定。评估可以是相对的或绝对的。“评估...的存在”包括确定存在的某物的量，和/或确定其是否存在。

[0019] “不对称标记的”多核苷酸具有不同的左衔接子结构域和右衔接子结构域。该过程一般称为不对称地附接衔接子或不对称地标记多核苷酸，例如多核苷酸片段。具有不对称衔接子末端的多核苷酸的产生可以通过任何适宜的方式实现。示例性的不对称衔接子见述于：美国专利5,712,126和6,372,434；美国专利公开2007/0128624和2007/0172839；以及PCT公开WO/2009/032167；所有这些专利均以引用方式整体并入本文。在某些实施方案中，所用的不对称衔接子为2009年4月29日提交的美国专利申请系列第12/432,080号中所述的那些，该专利以引用方式整体并入本文。

[0020] 作为一个实例，本发明的使用者可以使用不对称衔接子来标记多核苷酸。“不对称衔接子”是指这样的衔接子，当连接到双链核酸片段的两端时，它将导致产生引物延伸或产生具有位于感兴趣基因组插入片段侧翼的不同序列的扩增产物。连接后通常为后续处理步骤，以便生成不同的末端衔接子序列。例如，不对称衔接子附接的片段的复制产生多核苷酸产物，其中在末端衔接子序列之间存在至少一个核酸序列差异或核苷酸/核苷修饰。将衔接子不对称地附接到多核苷酸(例如，多核苷酸片段)产生在一个末端具有一个或多个衔接子序列(如，一个或多个区域或结构域，例如引物结合位点)的多核苷酸，该衔接子序列不存在或与另一末端上的衔接子序列相比具有不同的核酸序列。应当注意，称为“不对称衔接子”的衔接子本身不必在结构上不对称，也不仅仅是将不对称衔接子附接到多核苷酸片段后就立即使其成为不对称的。相反，在每一末端具有相同的不对称衔接子的不对称衔接子附接的多核苷酸产生的复制产物(或分离的单链多核苷酸)相对于在相对两端上的衔接子序列为不对称的(例如，在至少一轮扩增/引物延伸后)。

[0021] 任何适宜的不对称衔接子或不对称附接衔接子的过程均可用于实践本发明。示例性的不对称衔接子见述于：美国专利5,712,126和6,372,434；美国专利公开2007/0128624和2007/0172839；以及PCT公开WO/2009/032167；所有这些参考文献均以引用方式整体并入本文。在某些实施方案中，所用的不对称衔接子为2009年4月29日提交的美国专利申请系列第12/432,080号中所述的那些，该专利以引用方式整体并入本文。

[0022] “互补的”或“实质上互补的”是指核苷酸或核酸之间的杂交或碱基配对或双链体形成，诸如在双链DNA分子的两条链之间或在寡核苷酸引物与单链核酸上的引物结合位点之间。互补的核苷酸通常为A和T(或A和U)，或C和G。两个单链RNA或DNA分子在以下情况中被称为实质上互补的：当一条链的核苷酸(最优比对和比较并具有合适的核苷酸插入或缺失)与另一条链的核苷酸的至少约80％、通常至少约90％至95％以及更优选地约98％至100％配对时。作为另外一种选择，当RNA或DNA链将在选择性杂交条件下与其互补序列杂交时存在实质互补性。通常，选择性杂交将在至少14至25个核苷酸的片段上存在至少约65％互补性、优选至少约75％互补性、更优选至少约90％互补性时发生。参见M.Kanehisa Nucleic Acids Res.12:203(1984)，该文献以引用方式并入本文。

[0023] “双链体”是指完全或部分互补的至少两个寡核苷酸和/或多核苷酸在它们的所有或大部分核苷酸中发生Watson-Crick型碱基配对使得形成稳定的复合物。术语“退火”和“杂交”可互换使用以表示形成稳定的双链体。关于双链体的“完美匹配的”是指构成双链体

的多核苷酸或寡核苷酸链彼此之间形成双链结构，使得每条链中的每个核苷酸与另一条链中的核苷酸发生Watson-Crick碱基配对。稳定的双链体可包括双链体链之间的Watson-Crick碱基配对和/或非Watson-Crick碱基配对(其中碱基配对是指形成氢键)。在某些实施方案中，非Watson-Crick碱基配对包括核苷类似物，诸如脱氧肌苷、2,6-二氨基嘌呤、PNA、LNA等。在某些实施方案中，非Watson-Crick碱基配对包括“摆动碱基”，诸如脱氧肌苷、8-氧代-dA、8-氧代-dG等，其中所谓“摆动碱基”是指这样的核酸碱基，其可与互补核酸链中的第一核苷酸碱基进行碱基配对，但是当用作核酸合成的模板链时会导致将第二不同的核苷酸碱基并入合成链中(摆动碱基将在下文进一步详细描述)。在两个寡核苷酸或多核苷酸之间的双链体中的“错配”是指双链体中的一对核苷酸未能发生Watson-Crick键合。

[0024] 关于基因组或靶多核苷酸的“基因座位”、“基因座”或“感兴趣基因座”是指基因组或靶多核苷酸的连续亚区或片段。如本文所用，基因座位、基因座或感兴趣基因座可以指核苷酸、基因组中的基因或基因的一部分的位置，包括线粒体DNA或其他非染色体DNA(例如，细菌质粒)，或者其可以指基因组序列的任何连续部分，而无论其是否在基因内或与基因相关。基因座位、基因座或感兴趣基因座可以来自单个核苷酸至长度为几百个或几千个核苷酸或更长的片段。一般来讲，感兴趣基因座将具有与其相关的参考序列(参见下文的“参考序列”说明)。

[0025] “试剂盒”是指用于递送材料或试剂以进行本发明方法的任何递送系统。在反应测定法的背景下，此类递送系统包括允许储存、运输、或从一个位置到另一位置递送反应试剂(如，合适容器中的探针、酶等)和/或辅助材料(如，缓冲剂、进行测定法的书面说明等)的系统。例如，试剂盒包括装有相关反应试剂和/或辅助材料的一个或多个封闭物(例如，盒子)。此类内容物可一起或单独地递送给预期的受体。例如，第一容器可装有用于测定法的酶，而第二容器则装有探针。

[0026] “连接”是指在两个或更多个核酸例如寡核苷酸和/或多核苷酸的末端之间形成共价键或键合。键或键合的性质可差异巨大，并且连接可通过酶促或化学方式进行。如本文所用，连接通常通过酶促方式进行，以在一个寡核苷酸的末端核苷酸的5’碳与另一个寡核苷酸的3’碳之间形成磷酸二酯键合。多种模板驱动的连接反应在以下参考文献中有所描述，这些参考文献以引用方式并入：Whiteley等，美国专利4,883,750；Letsinger等，美国专利5,476,930；Fung等，美国专利5,593,826；Kool，美国专利5,426,180；Landegren等，美国专利5,871,921；Xu和Kool，Nucleic Acids Research，27:875-881(1999)；Higgins等，Methods in Enzymology，68:50-71(1979)；Engler等，The Enzymes，15:3-29(1982)；以及Namsaraev，美国专利公开2004/0110213。

[0027] 如本文所用的“多重标识”(Multiplex Identifier，MID)是指与多核苷酸相关的一个标记或标记组合，其同一性(例如，标记DNA序列)可用于区分样品中的多核苷酸。在某些实施方案中，多核苷酸上的MID用于识别多核苷酸的来源。例如，核酸样品可以是源自不同来源的多核苷酸(例如，源自不同个体、不同组织或细胞的多核苷酸，或在不同时间点分离的多核苷酸)的库，其中将来自每个不同来源的多核苷酸用单一MID标记。因此，MID提供多核苷酸与其来源之间的关联。在某些实施方案中，将MID用于单一标记样品中的每个个体多核苷酸。样品中单一MID的数量的确定可提供样品中存在多少个体多核苷酸的读出(或被操纵的多核苷酸样品源自多少原始多核苷酸，参见例如2009年5月26日公布的美国专利第

7,537,897号,该专利以引用方式整体并入本文)。MID通常由核苷酸碱基构成,并且长度可在2至100个核苷酸碱基或更多个碱基的范围内,以及可以包含多个亚基,其中每个不同的MID具有不同的同一性和/或亚基顺序。可用作MID的示例性核酸标记在以下专利中有所描述:2009年6月6日公布的并且名称为"Nucleic Acid Analysis Using Sequence Tokens"的美国专利7,544,473,以及2008年7月1日公布的并且名称为"Methods and Compositions for Tagging and Identifying Polynucleotides"的美国专利7,393,665,这两份专利中关于核酸标记及其在鉴定多核苷酸中的用途的说明以引用方式整体并入本文。在某些实施方案中,用于标记多个样品的MID集合不必具有任何特定的共同性质(例如,Tm、长度、碱基组成等),因为本文所述的方法可适应多种单一MID集合。这里应当强调,MID只需在给定的实验中是单一的。因此,相同的MID可用于标记在不同的实验中处理的不同样品。此外,在某些实验中,使用者可以使用相同的MID来标记同一实验内不同样品的子集。例如,源自具有特定表型的个体的所有样品可用相同的MID标记,例如,源自对照(或野生型)受试者的所有样品可用第一MID标记,而患有病症的受试者可用第二MID(不同于第一MID)标记。又如,可能有利的是,将源自相同来源的不同样品用不同的MID标记(例如,在不同时间得到的样品,或源自组织内的不同位点的样品)。另外,MID可通过多种不同的方式生成,例如,通过组合标记方法,其中一个MID通过连接而附接,第二个MID通过引物延伸而附接。因此,可按多种不同的方式设计并实施MID以在处理和分析过程中跟踪多核苷酸片段,并因此就这一点而言不旨在进行限制。

[0028] 如本文所用的"下一代测序"(NGS)是指能够以常规测序方法(例如,标准Sanger或Maxam-Gilbert测序方法)无可比拟的速度对多核苷酸测序的测序技术。这些无可比拟的速度通过平行地进行并读出几千至几百万个测序反应而实现。NGS测序平台包括但不限于以下这些:大规模平行签名测序(Lynx Therapeutics);454焦磷酸测序(454Life Sciences/Roche Diagnostics);固相可逆染料终止子测序(Solexa/Illumina);SOLiD技术(Applied Biosystems);离子半导体测序(Ion Torrent)和DNA纳米球测序(Complete Genomics)。某些NGS平台的说明可见于以下参考文献:Shendure等,"Next-generation DNA sequencing,"Nature,2008,第26卷,No.10,1135-1145;Mardis,"The impact of next-generation sequencing technology on genetics,"Trends in Genetics,2007,第24卷,No.3,第133-141页;Su等,"Next-generation sequencing and its applications in molecular diagnostics"Expert Rev Mol Diagn,2011,11(3):333-43;以及Zhang等,"The impact of next-generation sequencing on genomics",J Genet Genomics,2011,38(3):95-109。

[0029] 如本文所用的"核苷"包括天然核苷,包括2'-脱氧和2'-羟基形式,例如,如Kornberg和Baker,DNA Replication,第2版(Freeman,San Francisco,1992)中所述。关于核苷的"类似物"包括具有修饰碱基部分和/或修饰糖部分的合成核苷,例如,如Scheit,Nucleotide Analogs(John Wiley,New York,1980);Uhlman和Peyman,Chemical Reviews,90:543-584(1990)等文献中所述,前提是它们能够特异性杂交。此类类似物包括被设计为增强结合性质、降低复杂性、提高特异性等的合成核苷。包括具有增强的杂交或核酸酶抗性性质的类似物的多核苷酸在以下参考文献中有所描述:Uhlman和Peyman(上文引用);Crooke等,Exp.Opin.Ther.Patents,6:855-870(1996);Mesmaeker等,Current Opinion in

Structual Biology,5:343-355(1995);等等。能够增强双链体稳定性的多核苷酸的示例性类型包括寡核苷酸N3'→P5'磷酰胺酯(本文称为"酰胺化物")、肽核酸(本文称为"PNA")、寡-2'-O-烷基核糖核苷酸、含C-5丙炔基嘧啶的多核苷酸、锁核酸("LNA")以及类似化合物。此类寡核苷酸可商购获得或可使用参考文献中所述的方法合成。

[0030] "聚合酶链反应"或"PCR"是指通过DNA互补链的同时引物延伸而体外扩增特定DNA序列的反应。换句话讲,PCR是制备侧翼为引物结合位点的靶核酸的多个拷贝或复制物的反应,此类反应包括以下步骤的一个或多个重复:(i)使靶核酸变性,(ii)使引物退火到引物结合位点,以及(iii)在三磷酸核苷存在下通过核酸聚合酶延伸引物。通常,使反应在热循环仪中通过针对每个步骤而优化的不同温度进行循环。特定的温度、每个步骤的持续时间以及步骤之间的变化速率取决于本领域普通技术人员熟知的许多因素,例如以下参考文献中示例性示出的因素:McPherson等编辑,PCR:A Practical Approach和PCR2:A Practical Approach(IRL Press,Oxford,分别为1991和1995年)。例如,在使用Taq DNA聚合酶的常规PCR中,可使双链靶核酸在>90℃的温度下变性,使引物在50-75℃范围内的温度下退火,以及使引物在72-78℃范围内的温度下延伸。术语"PCR"涵盖反应的衍生形式,包括但不限于RT-PCR、实时PCR、巢式PCR、定量PCR、多重PCR等。反应体积可在几纳升(例如2nL)至几百μL(例如200μL)的范围内。"逆转录PCR"或"RT-PCR"是指其前是将靶RNA转化为互补单链DNA(然后进行扩增)的逆转录反应的PCR,例如Tecott等的美国专利5,168,038,该专利以引用方式并入本文。"实时PCR"是指随着反应的进行对反应产物即扩增子的量进行监测的PCR。存在许多形式的实时PCR,其主要区别在于用于监测反应产物的检测化学,例如Gelfand等,美国专利5,210,015("TAQMAN™");Wittwer等,美国专利6,174,670和6,569,627(插入染料);Tyagi等,美国专利5,925,517(分子信标);这些专利以引用方式并入本文。实时PCR的检测化学在Mackay等,Nucleic Acids Research,30:1292-1305(2002)中进行了综述,该参考文献也以引用方式并入本文。"巢式PCR"是指两阶段PCR,其中第一PCR的扩增子变成使用一个新的引物集合的第二PCR的样品,这些引物中的至少一个结合到第一扩增子的内部位置。如本文所用,关于巢式扩增反应的"初始引物"是指用于生成第一扩增子的引物,而"第二引物"是指用于生成第二或巢式扩增子的一个或多个引物。"多重PCR"是指其中在同一反应混合物中同时进行多个靶序列(或单个靶序列和一个或多个参考序列)的PCR,例如Bernard等,Anal.Biochem.,273:221-228(1999)(双色实时PCR)。通常,将不同的引物集合用于每个扩增的序列。

[0031] "多核苷酸"或"寡核苷酸"可互换使用,每一个均为指核苷酸单体的线性聚合物。构成多核苷酸和寡核苷酸的单体能够通过正常模式的单体间相互作用特异性结合到天然多核苷酸,这些相互作用诸如为Watson-Crick型碱基配对、碱基堆积、Hoogsteen或反式Hoogsteen型碱基配对、摆动碱基配对等等。如下文将详细描述,所谓"摆动碱基"是指这样的核酸碱基,其可与互补核酸链中的第一核苷酸碱基进行碱基配对,但是当用作核酸合成的模板链时会导致将第二不同的核苷酸碱基并入合成链中。此类单体及其核苷酸间键合可天然存在,或可以为其类似物,例如天然存在的或非天然存在的类似物。非天然存在的类似物可包括肽核酸(PNA,例如,如以引用方式并入本文的美国专利5,539,082中所述)、锁核酸(LNA,例如,如以引用方式并入本文的美国专利6,670,461中所述)、硫代磷酸核苷酸间键合、含有允许附接标签(诸如荧光团或半抗原)的连接基团的碱基,等等。每当寡核苷酸或多

核苷酸的使用需要酶法处理时(诸如通过聚合酶延伸、通过连接酶连接等等),普通技术人员将会理解,在那些情况下的寡核苷酸或多核苷酸将不含核苷酸间键合、糖部分或者任何位置或一些位置的碱基的某些类似物。多核苷酸的大小通常在几个单体单元(例如5至40个,此时通常称之为"寡核苷酸")至几千个单体单元的范围内。每当多核苷酸或寡核苷酸由一系列字母(大小或小写)诸如"ATGCCTG"表示时,应当理解,核苷酸从左至右处于5'→3'顺序,并且"A"表示脱氧腺苷、"C"表示脱氧胞苷、"G"表示脱氧鸟苷,以及"T"表示胸苷、"I"表示脱氧肌苷、"U"表示尿苷,除非另外指明或上下文显而易见。除非另有说明,否则术语和原子编号惯例将遵循Strachan和Read,Human Molecular Genetics 2(Wiley-Liss,New York,1999)中所公开的那些。通常,多核苷酸包含通过磷酸二酯键合连接的四个天然核苷(例如对于DNA而言为脱氧腺苷、脱氧胞苷、脱氧鸟苷、脱氧胸苷或对于RNA而言为其核糖对应物);然而,它们还可包含非天然核苷酸类似物,例如包含修饰的碱基、糖或核苷酸间键合。对本领域的技术人员将显而易见的是,如果酶对于活性而言具有特定的寡核苷酸或多核苷酸底物要求,例如单链DNA、RNA/DNA双链体等,则选择寡核苷酸或多核苷酸底物的合适组成将在普通技术人员的知识范围内,尤其是通过专著中的指导原则,诸如Sambrook等,Molecular Cloning,第二版(Cold Spring Harbor Laboratory,New York,1989)等参考文献。

[0032] "引物"是指天然的或合成的寡核苷酸,其在与多核苷酸模板形成双链体时能够作为核酸合成的引发点,并从其3'末端沿着模板延伸使得形成延伸的双链体。在延伸过程中添加的核苷酸的序列由模板多核苷酸的序列确定。通常,引物通过DNA聚合酶延伸。引物通常具有与其在引物延伸产物合成中的用途相容的长度,并通常在8至100个核苷酸长度的范围内,诸如10至75个、15至60个、15至40个、18至30个、20至40个、21至50个、22至45个、25至40个等等,更通常地在18至40个、20至35个、21至30个核苷酸长的范围内,以及所述范围之间的任何长度。典型的引物可在10至50个核苷酸长的范围内,诸如15至45个、18至40个、20至30个、21至25个等等以及所述范围之间的任何长度。在一些实施方案中,引物通常不超过约10、12、15、20、21、22、23、24、25、26、27、28、29、30、35、40、45、50、55、60、65或70个核苷酸长。

[0033] 引物通常为单链的以便获得最大的扩增效率,但作为另外一种选择也可以为双链的。如果为双链的,则通常在用于制备延伸产物前首先对引物进行处理,以将其链分离。该变性步骤通常受加热影响,但作为另外一种选择可使用碱然后进行中和而进行。因此,"引物"与模板互补,并通过氢键或杂交与模板复合以得到用于通过聚合酶引发合成的引物/模板复合物,该复合物在DNA合成过程中通过添加连接到其与模板互补的3'末端的共价键合碱基而延伸。

[0034] 如本文所用的"引物对"是指第一和第二引物,其具有适于基于核酸的靶核酸扩增的核酸序列。此类引物对通常包括具有与靶核酸第一部分的序列相同或相似的序列的第一引物,以及具有与靶核酸第二部分互补的序列的第二引物,以提供靶核酸或其片段的扩增。在本文提及"第一"和"第二"引物是任意的,除非另有具体指明。例如,第一引物可被设计为"正向引物"(其从靶核酸的5'末端引发核酸合成)或设计为"反向引物"(其从通过正向引物引发的合成中产生的延伸产物的5'末端引发核酸合成)。同样,第二引物可被设计成正向引物或反向引物。

[0035] “读出”是指可转化成数字或数值的测量出和/或检测到的一个或多个参数。一些背景下，读出可以指此类采集或记录的数据的真实数字表示。例如，微阵列中荧光强度信号的读出是在微阵列每个杂交位点生成的信号的地址和荧光强度；因此，这样的读出可以按多种方式登记或存储，例如，作为微阵列的图像、作为数字表格等等。

[0036] “反射位点”、“反射序列”和等同形式用于表示存在于多核苷酸中的一个或多个序列，它们用于在多核苷酸中将某结构域以分子内的方式从其初始位置移动到不同位置。反射序列的使用在名称为“Compositions and Methods for Intramolecular Nucleic Acid Rearrangement”、于2011年2月24日作为WO/2011/021102公开并且以引用方式并入本文的PCT专利申请系列第PCT/IB2010/02243号中有详细描述。在某些实施方案中，对反射序列进行选择，以便与多核苷酸中的其他序列不同(即，与可能存在于多核苷酸中的诸如待处理的基因组或亚基因组序列的其他序列具有很低的序列同源性)。因此，应当选择反射序列，以使得在反射过程中所用的条件下除了其互补序列外不与任何序列杂交。反射序列可以是合成的或人工生成的序列(例如，在衔接子结构域添加到多核苷酸)或通常存在于进行测定的多核苷酸中的序列(例如，存在于进行测定的多核苷酸的感兴趣区域内的序列)。在反射系统中，在与反射序列相同的多核苷酸链上(例如，双链多核苷酸的相同链上或相同的单链多核苷酸上)存在(例如，插入衔接子结构域)反射序列的互补序列，其中将该互补序列置于特定的位置以便促进此特定链上的分子内结合和聚合事件。用于本文所述的反射过程的反射序列因此可具有多种长度和序列。反射序列的长度可在5至200个核苷酸碱基的范围内。

[0037] “固体载体”、“载体”和“固相载体”可互换使用并且是指具有一个或多个刚性或半刚性表面的一种材料或一组材料。在许多实施方案中，固体载体的至少一个表面将为实质上平坦的，但在一些实施方案中可能有利的是通过例如孔、凸起区域、柱、蚀刻槽等针对不同的化合物将合成区域在物理上分离。根据其他实施方案，固体载体将采取珠粒、树脂、凝胶、微球或其他几何构型的形式。微阵列通常包括至少一个平坦的固相载体，诸如玻璃显微镜载玻片。

[0038] 关于一个分子与另一个分子(诸如探针的标记靶序列)结合的“特异性的”或“特异性”是指两个分子之间的识别、接触以及形成稳定复合物，并且该分子与其他分子的识别、接触或复合物形成显著更低。在一个方面，关于第一分子与第二分子结合的“特异性的”是指就第一分子识别某反应中或样品中的另一分子并与其形成复合物而言，它与第二分子形成最大数量的复合物。优选地，该最大数量为至少百分之五十。一般来讲，在特异性结合事件中涉及的分子在其表面上或在腔体中具有引起彼此结合的分子之间发生特异性识别的区域。特异性结合的实例包括抗体-抗原相互作用、酶-底物相互作用、多核苷酸和/或寡核苷酸中双链体或三链体的形成、生物素-亲和素或生物素-链霉亲和素相互作用、受体-配体相互作用等等。如本文所用，关于特异性或特异性结合的“接触”是指两个分子足够近，使得弱非共价化学相互作用(诸如范德华力、氢键、碱基堆积相互作用、离子和疏水相互作用等)在分子相互作用中占主导地位。

[0039] 如本文所用，术语“T_m”关于“熔融温度”而使用。熔融温度是一群双链核酸分子变为半解离成单链时的温度(例如，以℃度量)。计算核酸T_m的多个公式在本领域中是已知的(参见例如Anderson和Young，Quantitative Filter Hybridization，in Nucleic Acid Hybridization(1985))。其他参考文献(例如Allawi，H.T.和SantaLucia，J.，Jr.，

Biochemistry 36,10581-94(1997))包括可供选择的计算方法,这些方法在计算T_m时考虑了结构和环境以及序列特性。

[0040] “样品”是指一定量的来自生物、环境、医疗或患者来源的材料,在其中寻求对靶核酸进行检测、测量或标记。在一方面,其旨在包括标本或培养物(例如,微生物培养物)。在另一方面,其旨在包括生物和环境样品两者。样品可包括合成来源的标本。生物样品可以是动物,包括人、液体、固体(如粪便)或组织,以及液体和固体食物和饲料产品及配料,诸如乳制品、蔬菜、肉类和肉类副产品,以及废弃物。生物样品可包括从患者采集的物质,包括但不限于培养物、血液、唾液、脑脊髓液、胸膜液、乳汁、淋巴液、痰、精液、针吸出物等等。生物样品可从家畜以及未驯服的或野生的动物的所有家族中获得,包括但不限于诸如有蹄类、熊、鱼、啮齿类等动物。环境样品包括诸如地表物质、土壤、水体和工业样品的环境材料,以及从食物和乳品加工仪器、装置、设备、器皿、一次性和非一次性物品获得的样品。这些样品不应视为限制适用于本发明的样品类型。

[0041] 在描述核酸分子取向和/或聚合反应时的术语“上游”和“下游”在本文以本领域技术人员所理解的含义使用。因此,“下游”通常是指向5’至3’方向进行,即其中核苷酸聚合酶延伸序列时通常所采取的方向,而“上游”通常具有相反的含义。例如,在相同靶核酸分子上杂交第二引物“上游”的第一引物位于第二引物的5’侧上(并因此而言从第一引物发生的核酸聚合反应向第二引物进行)。

[0042] 还应当注意,可以起草权利要求书以排除任何可选要素。因此,该陈述旨在作为在结合权利要求要素的详述使用诸如“单独”、“仅”等排他性术语时或使用“负面”限制时的前置基础。

[0043] 发明详述

[0044] 在描述本发明前,应当理解本发明不限于所述的特定实施方案,因此当然可以发生变化。还应当理解,本文所用的术语只是为了描述特定实施方案,并非旨在进行限制,因为本发明的范围将只由所附权利要求书限定。

[0045] 如果提供了值的范围,则应当理解,介于该范围上限与下限之间的每个居间值直至下限单位的十分之一(除非上下文另有明确指出)也具体予以公开。介于所述范围中的任何所述值或居间值之间的每个更小的范围以及该所述范围中的任何其他所述或居间值被包括在本发明的范围内。这些较小范围的上限和下限可独立地被包括在所述范围内或被排除在所述范围外,并且其中两个限值之一包括在该较小范围内、两个限值均不包括在该较小范围内或者两个限值均包括在该较小范围内的每个范围也被包括在本发明的范围内,受所述范围内任何具体排除的限值的约束。如果所述范围包含所述限值的一个或两个,则排除这些所含限值一者或两者的范围也包括在本发明内。

[0046] 除非另有定义,否则本文所用的所有技术和科学术语具有与本发明所述领域的普通技术人员通常所理解的相同含义。虽然类似于或等同于本文所述的那些的任何方法和材料也可用于实践或检验本发明,但现在描述一些有潜力的和优选的方法和材料。本文提及的所有出版物以引用方式并入,以公开和描述与所引述的出版物相关的方法和/或材料。应当理解,当存在矛盾时,本公开将取代所并入的出版物的任何公开内容。

[0047] 必须注意,如本文和所附权利要求书中所用,单数形式“一”、“一个/种”和“该/所述”包括复数指代物,除非上下文另有明确指出。因此,例如,提及“一种核酸”包括多种这样

的核酸,而提及“该化合物”包括提及本领域技术人员已知的一种或多种化合物及其等同物,等等。

[0048] 除非另外指明,否则本发明的实践均可采用在本领域技术范围内的有机化学、聚合物技术、分子生物学(包括重组技术)、细胞生物学、生物化学和免疫学的常规技术和说明。此类常规技术包括聚合物阵列合成(polymer array synthesis)、杂交、连接和使用标记检测杂交。合适技术的具体例证可参考下文的实例。然而,其他等同的常规程序当然也可以使用。此类常规技术和说明可见于标准实验室手册,诸如Genome Analysis:A Laboratory Manual Series(第I-IV卷),Using Antibodies:A Laboratory Manual,Cells:A Laboratory Manual,PCR Primer:A Laboratory Manual以及Molecular Cloning:A Laboratory Manual(均得自Cold Spring Harbor Laboratory Press),Stryer,L.(1995)Biochemistry(第4版)Freeman,New York,Gait,“Oligonucleotide Synthesis:A Practical Approach”1984,IRL Press,London,Nelson and Cox(2000),Lehninger,A.,Principles of Biochemistry第3版,W.H.Freeman Pub.,New York,N.Y.以及Berg等(2002)Biochemistry,第5版,W.H.Freeman Pub.,New York,N.Y.,它们均以引用方式整体并入本文以用于所有目的。

[0049] 本文所述的出版物仅针对它们在本专利申请提交日之前的公开内容而提供。此处的任何信息均不应解释为承认本发明不能因为是在先发明而先于这些出版物。另外,所提供的公开日可能不同于实际的公开日,这可能需要独立地进行确认。

[0050] 如上所概述,本发明的方面涉及添加到进行序列分析的多核苷酸的简并核苷酸碱基(例如,在简并碱基区域或DBR中)的用途,其可用于确定个体多核苷酸分子的数量,该个体多核苷酸分子来源于已在特定序列分析配置或过程中测序的相同原始样品的相同基因组区域。在进行测序分析的多核苷酸中包括DBR可用于多种遗传分析,包括通过提供确定等位基因调用统计值(该值无法单独通过读段数量而得出)的机制而提高等位基因调用的置信度。DBR可以通过任何适宜的方式添加到多核苷酸,包括作为附接到进行测序的多核苷酸的衔接子(或衔接子库)的一部分,例如,DBR可位于还包括测序引物位点的衔接子中,或者DBR可存在于核酸合成引物(例如PCR引物)中,使得当将引物用于聚合反应时将DBR添加到靶核苷酸。

[0051] DBR还可用于对混合的多核苷酸样品进行遗传分析,其中该混合样品中的每个多核苷酸均包含其来源样品特定的MID(下文将详细描述)。这允许使用者确定经合并以生成混合样品的各来源样品中某一特定多核苷酸种类(或多个种类)的序列覆盖度。因此,本发明的实施方案包括对混合样品中的多核苷酸进行序列分析,其中每个多核苷酸均包含MID和DBR。

[0052] 核酸

[0053] 本发明(如下文详细描述)可用于操纵和分析来自几乎任何核酸来源的感兴趣核酸序列(或多核苷酸),包括但不限于基因组DNA、互补DNA(cDNA)、RNA(如信使RNA、核糖体RNA、短干扰RNA、微RNA等)、质粒DNA、线粒体DNA、合成DNA等。此外,任何有机体、有机材料或含核酸的底物均可用作将要根据本发明处理的核酸的来源,包括但不限于植物、动物(如爬行动物、哺乳动物、昆虫、蠕虫、鱼等)、组织样品、细菌、真菌(如酵母)、噬菌体、病毒、尸体组织、考古学/远古样品等。在某些实施方案中,核酸样品中的核酸源自哺乳动物,其中在一些

实施方案中该哺乳动物为人。

[0054] 在某些实施方案中，使核酸序列富集。所谓富集是指使核酸(例如，在多核苷酸样品中)接受降低核酸复杂性的过程，通常通过提高样品中特定核酸种类的相对浓度(例如，具有特定的感兴趣基因座、包含特定的核酸序列、不含某基因座或序列、在特定的大小范围内等等)。有多种方式可富集具有特定性质或序列的核酸，并且就这一点而言可以采用能实现该目的的任何适宜方法。富集(或降低复杂性)可在所述过程的多个步骤中的任何一个发生，并将由使用者的意愿决定。例如，富集可在个体亲本样品(例如，在连接衔接子前的未标记核酸)中或在多重化样品(例如，用编码MID的衔接子序列标记的核酸;MID将在下文进一步详细描述)中发生。

[0055] 在某些实施方案中，在分析前扩增核酸样品中的核酸。在这些实施方案的一些中，扩增反应还起到针对感兴趣序列或基因座富集起始核酸样品的作用。例如，可使起始核酸样品接受扩增一个或多个感兴趣区域的聚合酶链反应(PCR)。在某些实施方案中，扩增反应为指数式扩增反应，而在某些其他实施方案中，扩增反应为线性扩增反应。对起始核酸样品进行扩增反应的任何适宜方法均可用于实践本发明。在某些实施方案中，用于扩增反应的核酸聚合酶为具有校正能力的聚合酶(例如，phi29DNA聚合酶、嗜热高温球菌(Thermococcus litoralis)DNA聚合酶、激烈火球菌(Pyrococcus furiosus)DNA聚合酶等)。

[0056] 在某些实施方案中，进行分析的核酸样品源自单个来源(例如，单个有机体、病毒、组织、细胞、受试者等)，而在其他实施方案中，核酸样品为从多个来源提取出的核酸的库(例如，来自多个有机体、组织、细胞、受试者等的核酸的库)，其中所谓“多个”是指两个或更多个。因此，在某些实施方案中，核酸样品可包含来自2个或更多个来源、3个或更多个来源、5个或更多个来源、10个或更多个来源、50个或更多个来源、100个或更多个来源、500个或更多个来源、1000个或更多个来源、5000个或更多个来源、10,000个或更多个来源、25,000个或更多个来源等的核酸。

[0057] 在某些实施方案中，待与核酸片段混合的核酸片段源自多个来源(例如，多个有机体、组织、细胞、受试者等)，其中所谓“多个”是指两个或更多个。在此类实施方案中，源自每个来源的核酸包含多重标识(MID)，使得可以确定每个标记核酸片段的来源。在此类实施方案中，将每个核酸样品来源与单一MID相关联，其中所谓“单一MID”是指所用的每个不同的MID在至少一个特性(例如MID的核酸序列)方面可区别于所用的每个其他MID。可以使用任何类型的MID，包括但不限于在以下专利中所述的那些：于2007年1月22日提交的并且名称为"Nucleic Acid Analysis Using Sequence Tokens"的共同未决的美国专利申请系列第11/656,746号，以及于2008年7月1日提交的并且名称为“Methods and Compositions for Tagging and Identifying Polynucleotides”的美国专利7,393,665，这两份专利中关于核酸标记的说明及其在鉴定多核苷酸中的用途的内容以引用方式整体并入本文。在某些实施方案中，用于标记多个样品的MID集合不必具有任何特定的共同性质(例如，T_m、长度、碱基组成等)，因为不对称标记方法(以及许多标记读出方法，包括但不限于标记测序或标记长度测量)可适应多种单一MID集合。

[0058] 简并碱基区域(DBR)

[0059] 本发明的方面包括用于确定或估计个体多核苷酸分子的数量的方法和组合物，该

个体多核苷酸分子来源于已在特定序列分析配置或过程中测序的相同原始样品的相同基因组区域。在本发明的这些方面，将简并碱基区域(DBR)附接到随后将进行测序(例如，在进行某些过程步骤后，例如扩增和/或富集，例如PCR)的起始多核苷酸分子。如下详述，评价存在于测序运行中的不同DBR序列的数量(在一些情况下，为序列组合的数量)允许确定已针对特定多核苷酸(或感兴趣区域;ROI)测序的不同起始多核苷酸的数量(或最小数量)。该数量可用于例如给出等位基因基因调用中置信度的统计度量，并因此提高进行此类等位基因测定时的置信度(例如，当调用纯合等位基因时)。DBR还允许识别如果未检出时会对遗传分析造成负面影响的潜在测序或扩增误差。

[0060] DNA测序通常包括将衔接子附接到待测序的样品中的多核苷酸的步骤，其中衔接子包含测序引物位点(例如，通过连接)。如本文所用，“测序引物位点”是与测序引物(当为单链形式时)的序列相同或互补的多核苷酸区域或在测序引物序列与其互补序列之间形成的双链区域。测序引物位点的特定取向可由本领域普通技术人员从包含测序引物位点的特定多核苷酸的结构特征中推断出。

[0061] 除了测序引物位点外，还将简并碱基区域(DBR)作为含测序引物位点的衔接子的一部分或者独立地(例如，在附接到多核苷酸的第二衔接子中)附接到多核苷酸。任何适宜的将DBR附接或添加到多核苷酸的方法均可采用。DBR是与样品中的其他标记多核苷酸相比可具有可变碱基组成或序列(其可被视为“随机的”)的区域。在样品的一群多核苷酸中的不同DBR的数量将取决于DBR中碱基的数量以及可存在于每个位置的不同碱基的潜在数量。因此，一群附接有含2个碱基位置(其中每个位置可以为A、C、G和T中的任一者)的DBR的多核苷酸将可能具有16种不同的DBR(AA、AC、AG等)。DBR可因此包含1、2、3、4、5、6、7、8、9、10个或更多个碱基，包括15个或更多个、20个或更多个等等。在某些实施方案中，DBR的长度为3至10个碱基。此外，DBR中的每个位置可具有不同的碱基组成。例如，4个碱基的DBR可具有以下任一种组成:NNNN、NRSN、SWSW、BDHV(有关IUPAC核苷酸编码，参见下表1)。还应当注意，在某些实施方案中，DBR中的碱基可因具有可检测的修饰或附接到其上的其他部分而不同。例如，某些下一代测序平台(例如，Pacific Biosciences™)可用于在测序过程中检测碱基中的甲基化差异。因此，DBR中的非甲基化碱基可能不同于DBR中的甲基化碱基。因此，不旨在对DBR的长度或碱基组成进行限制。

[0062]

IUPAC 核苷酸编码	碱基
A	腺嘌呤
C	胞嘧啶
G	鸟嘌呤
T（或 U）	胸腺嘧啶（或尿嘧啶）
R	A 或 G
Y	C 或 T
S	G 或 C
W	A 或 T
K	G 或 T
M	A 或 C
B	C 或 G 或 T
D	A 或 G 或 T
H	A 或 C 或 T

[0063]

V	A 或 C 或 G
N	任何碱基

[0064] 这里应当注意，DBR可以是单个区域(即，具有所有彼此相邻的核苷酸碱基)或可存在于多核苷酸上的不同位置(即，DBR的碱基由非DBR序列分离，也称为断裂DBR)。例如，DBR可以在多核苷酸上的第一位置的第一衔接子中具有一个或多个碱基以及在相同多核苷酸上的第二位置的第二衔接子中具有一个或多个碱基(例如，DBR可具有存在于不对称标记的多核苷酸即具有不对称衔接子的多核苷酸的两端的碱基)。就这一点而言不旨在进行限制。

[0065] DBR可被设计为有利于检测在序列分析前进行的扩增期间在DBR中发生的误差和/

或在测序反应本身中发生的误差。在此类实施方案中，所用的DBR序列可设计为使得DBR序列中的误差不必定导致生成另一个可能的DBR序列(从而导致因DBR突变而不正确地将源自相同模板的复制子鉴定为来自不同的模板)。例如，考虑使用具有序列N的DBR。在N中的误差会将一个DBR变成另一个，这可能导致过高估计正确指定基因表型的概率。将此与具有序列Y的DBR进行比较。如果在此位置发现R，则可以知道存在误差。虽然正确的DBR并不一定能够被指定给该含有误差的DBR，但是可以检测其是因为误差所致(例如，在测序或扩增中)。

[0066] 在一些实施方案中，简并碱基序列可用作合并的MID-DBR，其既可(1)指定样品身份也可(2)进行分子追踪/计数。例如，考虑两个分子，一个用YYY标记，另一个用RRR标记。在测序反应中，观察到具有序列TAT的MID-DBR，其也不与合并的MID-DBR序列结构相符。需要一个突变将YYY转化成TAT。需要两个突变将RRR转化成TAT。因此，可以说MID-DBR为YYY而不是RRR的概率更高。

[0067] 示例性误差识别(或误差纠正)的说明可见于许多参考文献(例如，见述于名称为“Method and compositions for using error-detecting and/or error-correcting barcodes in nucleic acid amplification process”的美国专利申请公开US2010/0323348以及名称为“System and method for identification of individual samples from a multiplex mixture”的US2009/0105959，二者均以引用方式并入本文)。

[0068] 在DBR存在于包含其他功能结构域(例如，测序引物位点、MID、反射序列)的衔接子群内的某些实施方案中，衔接子群中的功能结构域将彼此相同，而DBR则将存在差异。换句话讲，不同于衔接子群中的其他结构域，DBR可具有可变(或随机)碱基组成。所谓“衔接子群”、“衔接子的群”等是指被设计为附接到样品中的多核苷酸的衔接子分子的样品。

[0069] 生成具有DBR的衔接子可通过任何适宜的方式实现，例如使用本领域熟知的DNA合成方法(参见上文定义章节中的引文)。

[0070] 一旦附接到亲本样品中的多核苷酸后，多核苷酸可接受进一步的处理和最终测序。可以进行的处理步骤包括使用者期望的任何过程步骤，例如富集、扩增等等。在测序步骤中，获得DBR的序列以及多核苷酸的一部分(例如，含有感兴趣区域)。一旦获得序列，即确定附接到感兴趣多核苷酸的不同DBR的数量。该数量可用于确定或估计在测序结果中表现的来自起始亲本样品的不同感兴趣多核苷酸的数量，其中在一些实施方案中，该确定的数量是在测序结果中表现的来自起始亲本样品的不同感兴趣多核苷酸的最小数量。

[0071] 例如，考虑具有碱基组成NN(其中N为任何脱氧核苷酸碱基，即A、G、C或T)的两个碱基的DBR用于对特定受试者样品的基因座测序，以进行等位基因调用(即，受试者在该基因座是纯合的还是杂合的)。虽然在寡核苷酸合成中可存在一些偏差，但是可以预期将在衔接子群中存在16种不同的DBR，而概率大致相等(如上所述)。当鉴定出潜在的纯合等位基因调用时，确定存在于测序运行中的DBR的数量可用于确定/估计实际测序的多核苷酸分子的数量(或最小数量)(并因此确定/估计在处理步骤期间扩增的数量)。

[0072] 对于二倍体基因组，等位基因调用(在理想或理论情况下)可通过二项分布建模。假定两个等位基因拷贝(X和Y)在某一位点不同，则观测所有X或所有Y的概率由式$(1/2)^c$给出，其中c为该位点观测值(读段)的数量。如果在该位点观察到X十次(并且无Y)，则可以说样品可能为类型X纯合的。该调用中误差的概率因此为$(1/2)^{10}$(略低于千分之一)。

[0073] 我们的实验表明，样品制备早期的低含量DNA可导致全部对应于一个等位基因的

读段的高覆盖度，并且这种情况的发生可比根据二项分布的预期情况多出许多次。这是由于少数DNA分子(或甚至单个分子)的扩增所致，该扩增导致了源自单条染色体(即，两条二倍体染色体中只有一条实际存在于感兴趣样品中)上基因座位的大量读段。其结果是随覆盖度而变化的误差大大偏离预测的二项误差。

[0074] 使用如本文所述的DBR将提高从具有有限量DNA的样品中进行等位基因调用的置信度。例如，如果某基因座位中一个等位基因的16个测序读段均包含DBR序列GA，则可能的是，所有这些读段均来自相同的亲本多核苷酸分子(并因此不能证明为纯合等位基因调用)。然而，如果这16个测序读段每一个均具有不同的DBR序列，则可以更可信地作出纯合调用，因为每个读段均来自不同的亲本多核苷酸分子。

[0075] 这里应当注意，在许多实施方案中，不可能得出具有相同DBR序列的多核苷酸源自相同亲本多核苷酸分子的结论，因为多个相同的DBR可存在于DBR附接的多核苷酸中。例如，如果将含有两个N碱基的DBR的衔接子群用于标记含有多于16个多核苷酸的样品，则标记的多核苷酸的子集将具有相同的DBR，并因此将不可能确定它们的序列源自不同的亲本多核苷酸分子。

[0076] 更准确地确定起始或亲本分子实际数量的一种示例性方式将是提高DBR的简并性(即，提高用于标记特定感兴趣样品的DBR中的单一序列的数量)，使得每一个分子可能具有不同的DBR。在任何情况下，我们可在示例性方法中使用观测的DBR数量或可能产生观测DBR数量的预期读段数量的概率分布。

[0077] 当计算特定等位基因调用是杂合子还是纯合子的估计值时，可以构建/采用合适的函数L(r,v)，给出r参考读段和v变体读段，它将返回基因型的可能性。当采用如本文所述的DBR时，可以采用计算估计值的修正函数L(r',v')，其中r'为参考读段的单一DBR的数量，v'为变体读段的单一DBR的数量。进行等位基因调用的任何适宜函数均可使用并进行修正，以采用如本文所述的关于DBR读段的数据。

[0078] 这里应当注意，本发明的方面可用于提高调用多核苷酸样品即基因组样品中的拷贝数变异时的置信度。拷贝数变异可包括基因组重排，诸如缺失、复制、倒位和易位或全染色体非整倍性，诸如单体性、二体性、三体性、四体性和五体性。例如，考虑其中亲本在给定的SNP具有基因型AC和CC并且先证者具有基因型ACC的复制事件。在具有基因型AC的亲本中，给定足够深度的测序覆盖度(即，足够的测序读段)，预计与C等位基因和A等位基因相关的DBR数量将相似。在先证者中，给定足够深度的测序覆盖度，预计与C等位基因相关的DBR数量将是与A等位基因相关的DBR数量的2倍，其提供复制事件涵盖C等位基因的证据。使用DBR而不是测序读段的数量在调用拷贝数变异中提供更高的置信度，因为DBR可用于识别源自不同多核苷酸分子的读段。

[0079] DBR的示例性应用

[0080] 如上文所详述，DBR允许进行异质样品中序列变体(包括复杂基因组或库)的统计验证。例如，DBR可用于分析肿瘤样品、微生物样品、环境样品等样品中的复杂基因组。

[0081] 下文提供示例性统计方法和DBR的示例性应用。下文的描述旨在仅用于示例性目的，不旨在限制在多核苷酸分析中采用DBR的范围。

[0082] 统计方法

[0083] 如上所述，在本发明的方面，将简并碱基运行(DBR)用于估计或定量测定在给定过

程中测序或分析的模板分子的实际数量。两个读段可具有相同的DBR,因为该读段来源于相同的模板分子,或者因为分子偶然接受了相同的DBR。进行测序的不同模板分子的可能数量在DBR数量到读段数量的范围内。来自多个起始分子的DBR的分布由占有分布给出[参见C.A.Charalambides和C.A.Charalambides.Combinatorial methods in discrete distributions.John Wiley and Sons,2005]。给定DBR的观测数量,则可使用最大似然估计或其他合适的技术计算起始分子的可能数量。作为另外一种选择,对于每个DBR,可使用具有该特定DBR的所有读段的共有序列估计最可能的模板分子。可组合这些方法以生成与特定变体相关的模板分子的数量。

[0084] DBR用于PCR扩增

[0085] DBR可用于估计或测定用作PCR反应模板的起始分子的数量。例如,可使用PCR引物对将起始多核苷酸样品通过PCR扩增第一循环或前几个循环,其中一个(或两个)引物包含通用引物序列和靶特异性序列5'的DBR。在初始一个或多个循环后,可将该含有DBR的PCR引物对移除或使其失活,然后用不具有DBR的PCR引物替换以用于剩余的循环。含DBR的引物的移除/失活可通过任何适宜的方式实现,例如通过物理或生物化学手段。例如,含DBR的引物可具有附接到其上的结合对的第一成员(例如,生物素),从而有利于通过将样品与附接到固体载体上的结合伙伴(例如,固体载体结合的链霉亲和素)接触并收集未结合的部分而移除这些引物。作为另外一种选择,可通过以下方式移除不含DBR的引物:通过将样品用单链特异性外切核酸酶(例如外切核酸酶I)处理,通过使引物不能参与进一步的引物延伸步骤(例如,通过在3'末端结合双脱氧核苷酸)或通过固相可逆固定(SPRI)过程(例如,Agencourt AMPure XP-PCR纯化,Beckman Coulter)。将第二PCR引物设计成包含存在于第一引物集合中每一个的5'末端上的序列,以便复制在从用于第一/前几个循环中的含DBR的PCR引物生成的模板中的DBR。因此,PCR的剩余循环将只扩增含有DBR的第一/前几个循环的产物。在另一个实施方案中,可将含有DBR的引物设计成具有比不含DBR的第二引物集合更高的Tm(即,第一PCR引物的靶特异性序列的Tm高于5'通用引物序列特定的第二PCR引物的Tm)。在该示例性情形中,含DBR的引物可以有限量存在,并且第一/前几个PCR循环在较高的Tm下进行使得只有含DBR的引物退火并参与核酸合成。由于含DBR的引物以有限量存在,因此它们将在第一/前几个PCR循环中用完。在较低的Tm下进行剩余的PCR循环将允许通过不含DBR但将从第一/前几个循环的产物中复制DBR(如上所述)的第二PCR引物集合进一步扩增。

[0086] 应当注意,存在许多可用于完成上述DBR PCR扩增的PCR引物和扩增条件组合。例如,此类反应可包含3个引物,其中将引物1(靶多核苷酸特异性的并包含DBR和5'通用引发序列的正向引物)和引物2(靶多核苷酸特异性的并且不含DBR的反向引物)用于在第一/前几个循环中扩增靶标,将引物3(引物1的5'通用引发序列特异性的正向引物)和引物2用于剩余的循环。

[0087] 还应当注意,用DBR标记PCR产物的两端(即,其中用于第一/前几个循环的两个引物均包含DBR)可在估计扩增的起始多核苷酸数量时提供更高的置信度。应当注意,如果使用多于2个PCR循环来附接DBR,则必须在使用DBR时的数据分析过程中采取另外的预防措施,以追踪扩增产物的初始模板(或起始)分子。这是由于存在以下可能性:在第3个PCR循环中,具有DBR的PCR引物可在现有DBR位点上结合在之前生成的PCR产物中,从而引入新的DBR

序列。如下所述，前三个PCR循环的理论分析表明可能追踪分子的谱系。应当注意，以下分析可在理论上用于添加DBR序列的任何数量的PCR循环，但测序深度将必须足够深。

[0088] 下述方法允许在使用含DBR的PCR引物进行的1个以上DBR添加循环后对序列读段进行分组。表1显示了由单个双链模板在三个PCR循环每一个中生成的PCR产物的每一个(该模板具有上游链A和下游链B，如图3的第0个循环中所述)。在表I中，存在于每个循环中的每条链(用字母A至P表示)与其相应的模板链(即，在标示链的合成中用作模板的链)以及存在于链上的5’DBR和3’DBR(若有，用数字1至14表示)一起示出。所谓“5’DBR”是指当作为PCR引物的一部分而并入的DBR序列。所谓“3’DBR”是指5’DBR序列的互补序列(即，因跨现有5’DBR序列的引物延伸而生成的)。在第3个循环中，会发现可能发生DBR覆盖(在最右列中示出，参见例如在第3个循环中产生的K和N链)。

[0089] 表I在单个双链模板(A和B)的PCR反应的第0至3个循环中的DBR标记。

[0090]

循环	链	5' DBR #	3' DBR #	模板链	DBR 覆盖
第 0 个循环	A	-	-	-	
	B	-	-	-	
第 1 个循环	A	-	-	-	
	B	-	-	-	
	C	1	-	A	
	D	2	-	B	
第 2 个循环	A	-	-	-	
	B	-	-	-	

[0091]

		C	1	-	A	
		D	2	-	B	
		E	3	-	A	
		F	4	1	C	
		G	5	2	D	
		H	6	-	B	
	第 3 个循环	A	-	-	-	
		B	-	-	-	
		C	1	-	A	
		D	2	-	B	
		E	3	-	A	
		F	4	1	C	
		G	5	2	D	
		H	6	-	B	
		I	7	-	A	
		J	8	3	E	
		K	9	4	F	DBR 1 被 DBR 9 覆盖
		L	10	1	C	
		M	11	2	D	

[0092]

	N	12	5	G	DBR 2 被 DBR 12 覆盖
	O	13	6	H	
	P	14	-	B	

[0093] 上表I和图3显示了已在整个PCR过程中累积的链。(注意第0个循环至第1、2和3个循环中A链和B链的产物遗留(carry-over);第1个循环至第2和3个循环中C链和D链的产物遗留;等等)。

[0094] 给定足够的测序深度,可将DBR用于追溯起源分子,即使发生了DBR覆盖。例如,K链具有5'DBR#9和3'DBR#4。DBR#4与具有5'DBR#4和3'DBR#1的F链共有。DBR#1与C链共有。因此,K和F链最初源自C链。相似地,N链具有5'DBR#12和3'DBR#5。DBR#5与具有5'DBR#5和3'DBR#2的G链共有。DBR#2与D链共有。因此,N和G链最初源自D链。

[0095] 如上所述,将含有DBR的PCR引物在前几个循环后移除(例如,在完成如表I所示的第3个循环后)。

[0096] 图3显示了如表I所示的单个双链模板的前2个PCR循环的示意图。在第0个循环,只存在双链模板链,即上游链A和下游链B。注意,图3中每条链上的箭头方向表示5'至3'方向。在第一个PCR循环(第1个循环)中,通过PCR引物对(包含DBR序列的引物对的两个成员)产生了2个产物:具有DBR#1的第一产物(C)(如图中所示的"1")和具有DBR2的第二产物(D)(如图中所示的"2")。在第二个PCR循环(第2个循环)中,四个模板(A、B、C和D)产生4个产物(E、F、G和H),每一个具有附接的后续DBR(分别为DBR#3、#4、#5和#6)。注意,从模板C和D生成的产物(F和G)现在在两端均具有DBR。然后,第3个循环(未在图3中示出)使用来自第2个循环的8个模板产生8个产物,每一个具有附接的另外的DBR(参见表I中所示的产物)。第3个循环是其中可发生DBR覆盖的第一个循环(即,用后续DBR标记的PCR引物引发和延伸模板F和G将覆盖DBR#1和DBR#2;这些在表I中示为K和N链)。

[0097] 在分析其中可能发生DBR覆盖的多核苷酸的DBR时,根据5'和3'DBR序列对读段分组,并追踪亲本分子的谱系。

[0098] 如上文所述将显而易见的是,DBR可用于(1)识别在早期循环中发生的PCR误差,以及(2)准确确定等位基因调用/拷贝数。对于误差识别,显然可允许对独立的引发事件进行分组。鉴于引发事件不一定代表独立的启动分子,准确计算等位基因调用的复杂性略微升高。然而,可合理假设,引发事件在任一等位基因上是等可能事件,并且该分析因此可用于提高等位基因调用的准确性。

[0099] 对于极低的初始模板拷贝数,使用多个DBR添加循环将是有利的。例如,在极低的DNA浓度下,可能无法回收足够数量的DBR以使用标准方法给出准确的基因型。在此情况下,允许在相同的模板分子上发生多个引发事件可通过提供更多的数据为进行等位基因调用给出足够的置信度。

[0100] 分析最终扩增产物中的DBR可用于估计在反应中扩增的起始分子的数量。此类分析将允许使用者确定PCR反应的产物是否代表只有少数(或甚至一个)起始多核苷酸发生选择性扩增和/或有助于确定已在扩增期间发生的PCR误差(例如,如上所述)。

[0101] DBR用于异质肿瘤样品

[0102] DBR还可用于评估肿瘤样品中染色体异常的异质性,例如,在单个肿瘤内或在受试者中的不同肿瘤之间。例如,可从单个肿瘤(例如,在肿瘤内或周围的不同位置)和/或从受试者中的不同肿瘤获得一个或多个肿瘤样品并分析在一个或多个染色体位置的遗传变异。在某些实施方案中,样品可得自一个肿瘤(或受试者)的不同时间。此类变异可包括如本领域所知的特定碱基变化、缺失、插入、倒位、复制等等。DBR可用于在识别特定遗传变异之前标记一个或多个肿瘤样品中的多核苷酸,从而提供进行统计分析以验证识别出的任何变体的方法。例如,统计分析可用于确定检测到的变异是否代表肿瘤细胞亚群中的突变、是否是受试者中特定肿瘤的特异性变异、是否是存在于个体非肿瘤细胞中的变异,或者是否是识别变体时的过程假象(例如,PCR假象)。

[0103] DBR用于评估微生物多样性

[0104] DBR分析也可用于确定单个样品中或不同样品之间(例如,在不同时间点或从不同位置采集的样品)一群微生物/病毒的遗传变异/多样性。例如,在感染的过程中从个体采集的样品可使用本文所述的DBR分析感染过程期间的遗传变异。然而,不旨在对微生物/病毒样品的类型进行限制,并因此样品可来自任何来源,例如,来自感染的受试者、来自环境来源(土壤、水体、植物、动物或动物废物等)、来自食物来源,或者需要在一个或多个基因座位或区域确定样品中微生物的遗传多样性的任何其他样品。在实践该方法时,将源自样品的多核苷酸用如本文所述的DBR标记(在富集步骤之前或之后)并进行处理,以识别一个或多个感兴趣遗传位点或基因座的遗传变异。然后可进行DBR的分析,以提供在一个或多个感兴趣基因座测定的样品中微生物遗传多样性的置信度。此类分析可对从多种来源和/或在一个来源的多个时间点采集的样品进行。可用于评估微生物多样性的示例性基因座位包括但不限于核糖体RNA,例如16S核糖体RNA、抗生素抗性基因、代谢酶基因等。

[0105] DBR用于评估样品中不同多核苷酸种类的水平

[0106] DBR分析还可用于评估样品中不同多核苷酸种类的水平。具体地讲,由于DBR分析可确定(或估计)样品中亲本多核苷酸的数量,因此可评估特定多核苷酸种类的相对量或定量量以及确定此类种类时的置信度。例如,使用DBR分析cDNR样品可用于评估样品中不同cDNA种类的相对或定量水平,从而提供一种确定它们的相对基因表达水平的方法。

[0107] DBR用于分析混合样品

[0108] DBR的另一种应用是对混合的多核苷酸样品进行遗传分析,其中该混合样品中的每个多核苷酸均包含其来源样品特定的MID(上文已详细描述)。这允许使用者确定经合并以生成混合样品的各来源样品中某一特定多核苷酸种类(或多个种类)的序列覆盖度。这提供了确保来自混合样品中各起始样品的多核苷酸得到充分表现的机制。因此,本发明的实施方案包括混合样品中多核苷酸的序列分析,其中每个多核苷酸均包含MID和DBR。应当注意,在这些实施方案中,相同的DBR设计可结合所有亲本样品/MID使用,因为它是用于样品特异性序列分析中的MID/DBR组合。

[0109] 使用MID和DBR的混合样品分析可用于多种遗传分析,包括进行等位基因调用、序

列误差纠正、相对和定量基因表达分析等等。应当注意，在根据本发明的方面分析混合样品中的多核苷酸时，重要的是在所采用的工作流程的每个步骤中保持MID和DBR两结构域，因为丧失一个或另一个结构域将对所得结果的置信度造成负面影响。

[0110] 还应当注意，在遗传分析中使用MID和DBR结构域在与下一代测序(NGS)平台相结合时尤其有效，其中许多下一代测序平台提供存在于待进行测序的样品中的各个个体多核苷酸的序列数据。与其中对多核苷酸的个体克隆进行单独测序的常规测序方法相比，NGS平台同时提供样品中多个不同多核苷酸的序列。这一差异允许完成样品特异性统计分析，而不受限于必须进行克隆并单独地对每个多核苷酸测序。因此，本文所述的MID/DBR结构域分析与NGS平台协同，从而提供改进的统计方法，以分析得自混合样品的极大量的序列数据。

[0111] 试剂盒和系统

[0112] 本本发明还提供了用于实践主题方法(即使用DBR确定已针对特定多核苷酸测序的不同起始多核苷酸的数量(或最小数量))的试剂盒和系统。因此，系统和试剂盒可包含：含有DBR(例如衔接子)的多核苷酸以及任何其他如本文所述的感兴趣功能结构域(例如测序引物位点、MID、反射序列等等)。系统和试剂盒还可以包含：在将衔接子附接到亲本样品中的多核苷酸时用于进行任何步骤的试剂，用于制备亲本样品以附接衔接子/DBR的试剂，和/或用于进行测序反应的试剂(例如连接酶、限制酶、核苷酸、聚合酶、引物、测序引物、dNTP、ddNTP、外切核酸酶等等)。根据需要，系统和试剂盒的各种组分可存在于单独的容器中或者某些相容的组分可预先合并到单个容器中。

[0113] 主题系统和试剂盒还可以包含用于根据主题方法制备或处理核酸样品的一种或多种其他试剂。这些试剂可以包括一种或多种基质、溶剂、样品制备试剂、缓冲剂、脱盐试剂、酶试剂、变性试剂，其中也可以提供校准标准品，诸如阳性和阴性对照。因此，该试剂盒可以包括一个或多个容器(诸如小瓶或瓶子)，而每一个容器容纳单独的组分以根据本发明进行样品处理或制备步骤。

[0114] 除了上述组分外，主题试剂盒通常进一步包含使用试剂盒组分以实践主题方法(例如，采用如上所述的DBR)的说明。实践主题方法的说明通常记录在合适的记录介质上。例如，该说明可印刷在基材上，诸如纸张或塑料等上。因此，该说明可作为包装说明书存在于试剂盒中、存在于试剂盒容器或其组件(即，与包装或分包装相关的)的标签上等。在其他实施方案中，该说明作为电子存储数据文件存在于合适的计算机可读存储介质上，例如CD-ROM、磁盘等。在其他实施方案中，实际的说明不存在于试剂盒中，但提供从远程来源获得说明的方法，例如通过互联网。该实施方案的实例为包含网址的试剂盒，可在该网址查看说明和/或可从该网址下载说明。与说明一样，获得说明的该方法也记录在合适的基材上。

[0115] 除了主题数据库、编制程序和说明外，试剂盒还可以包含一种或多种对照样品和试剂，例如，用于测试试剂盒的两种或更多种对照样品。

实施例

[0116] 方法

[0117] 用衔接子标记了小鼠基因组DNA的两个相同的样品。一个样品使用具有由7个碱基(RYBDHVB)组成的冗余合成(redundantly synthesized)区域的衔接子，其每一个可以是两个(对于R和Y位置)或三个(对于B、D、H或V位置)碱基(或总共972个不同的序列)中的一个接

着为碱基ACA;第二个样品使用具有由7个碱基(RYBDHVB)组成的冗余合成区域的衔接子,其每一个可以是两个(对于R和Y位置)或三个(对于B、D、H或V位置)碱基(或总共972个不同的序列)中的一个接着为碱基ACG。注意,黑体带下划线的碱基对应于合成的多态性位点。在这些衔接子中,序列RYB用作DBR区域,DHVB用作MID。因此,存在12个可能的DBR(2×2×3)编码和81个不同的MID(3×3×3×3)。

[0118] 然后将两个样品等量混合在一起,以实际上形成MID(即,DHVB序列)下游的A和G三个碱基的完美50/50杂合子。将不同量的混合物(100ng、300ng、600ng、2500ng、5000ng和10,000ng)进行杂交拉下(pull-down)反应,接着通过用TiA和TiB引物进行的10个PCR循环扩增。所用的捕获单针为5'-生物素化60-mer反相净化滤芯纯化的寡核苷酸(BioSearch)。在用TiA和TiB扩增后,将与TiA和用TiB(5'-CCTATCCCCTGTGTGCCTTGGCAGTCTCAGGGACACCCAGCCAAGACAGC-3')(SEQ ID NO:1)加尾的序列特异性引物的二次PCR反应用于扩增特定片段。将杂交拉下/PCR中由每个样品生成的PCR片段送往进行454Ti鸟枪法测序(shot sequencing),以确定DBR、MID和A/G等位基因。

[0119] 杂交拉下/PCR中的扩增子序列如下所示(SEQ ID NO:2)。DBR区域带下划线,MID为黑体,等位基因(R,其对应于A或G)为黑体带下划线。

[0120]
CCATCTCATCCCTGCGTGTCTCCGACTCAGRYB**DHVB**AC**R**TAGAATGTGCATGG
ATCGTATGAGCACCTGTGGGCAGGGCAAGTGGCAGATGCCTTAGTGGATCTCAC
TGGAAGCCTGGCAGAAAGGTGGAGCTTGAAGGATGTAACGAAAGCCAGCGGCC
AGCAGGACAGACCCAGTGGTGGGGAGCACAGAACTTGTCGGCAGCTACTCCAC
CTGAAGGACCGGTGTCTAATCAGCTGCTCTGTGCTTAGCCCCAGAGCAGGTACA
GCTATGGCTACAACTCCCTCCACCATTAGCTTGTTACAGAGAAGGAAATCGGTC
CTTGAGAGGCTGTCTTGGCTGGGTGTCCCTGAGACTGCCAAGGCACACAGGGGA
TAGG

[0121] 结果

[0122] 图1显示了样品中各MID的等位基因比。在6张图中每一张的顶部的数字显示了所用的基因组DNA的输入质量(以纳克为单位)。横坐标针对任何特定的MID显示了A测序读段或调用(即,在合成的SNP位置多态性碱基之一的读段数量)与总调用数的比值,例如,(A调用)/(A调用+G调用)的比率。这称为等位基因比。纵坐标是特定等位基因比所观察到的MID数量(如上所述的MID总数为81)。由于已知输入的DNA的比率为50/50A/G,因此每个样品的等位基因比应为0.5。

[0123] 在低输入质量下,等位基因比失真,要么过高要么过低,因为进入第一PCR步骤的分子数有限,因此优先观察到两个等位基因中的一个。随着进入第一步骤的质量提高,两个等位基因均将被观察到,并且越来越接近于预期比率0.5。该分析表明在100ng、300ng和600ng时存在相当多的等位基因脱扣(allele drop out),而在较高的输入水平时很少或没有发生等位基因脱扣。

[0124] 图2显示了与每个等位基因相关的各MID的DBR序列的分数。由于输入材料在合成多态性位置名义上为50/50A和G,因此可以预计81个有效MID的每一个与50%A读段和50%G读段(如上所述)相关。此外,由于12个DBR为随机的,并与81个不同MID的每一个相关,因此可以预计,通过足够的输入DNA拷贝,对于A等位基因将观察到所有12个DBR,对于G等位基因

也将观察到所有12个。因此，在理想情况下，特定碱基所观察到的DBR分数对于每个等位基因而言将是12/24或0.5。在进入处理步骤的分子数量不足的情况下，对于每个MID等位基因可能存在少于12个DBR，并因此该比率会偏离理想情况。

[0125] 在图2中，横坐标显示了对于任何特定的MID，每个等位基因实际观察到的DBR除以实际观察到的DBR总数得到的比例。纵坐标是特定比例或与等位基因相关的DBR所观察到的MID数。存在总共81个不同的MID，并且应当见到总共81个不同的MID。我们观察到在低输入质量下，常常会见到比例失真，要么过高要么过低。这可能是因为输入第一PCR步骤的分子数量有限，并因此优先观察到两个等位基因中的一个。随着进入第一步骤的质量增加，两个等位基因均将更加频繁地观察到，并且与每个等位基因相关的是观察到越来越多的DBR，并因此比例更接近0.5。

[0126] 图1和图2中数据的比较显示出采用如本文所述的DBR的另外特征。对于真实的杂合子，在DBR分析(图2)中较早地(即，在较低质量下)开始看到在预期比率0.5附近的分布聚类。这很容易理解，因为，例如A等位基因观察到12个DBR中的6个，并且G等位基因观察到12个DBR中的4个，则将导致(6/[6+4])＝0.60的比例，其事实上相当接近于预期值0.5。总的效果是使用DBR在存在真实杂合子或真实纯合子的情况下给出高得多的置信度。

[0127] 虽然出于清晰理解的目的已经通过举例说明和实施例比较详细地描述了前述发明，但是按照本发明的教导对于本领域的普通技术人员将显而易见的是，在不脱离所附权利要求书的精神和范围的情况下可对本发明作出某些变化和修改。

[0128] 因此，前文仅仅阐述了本发明的原理。应当理解，本领域的技术人员将能够设想出虽然本文未明确描述或示出但体现本发明原理并包括在其精神和范围内的各种安排。此外，本文详述的所有实例和条件语言主要是为了有助于读者理解本发明的原理以及使发明人贡献的概念促进本领域的发展，并且应当视为对此类具体详述的实例和条件没有限制。此外，本文详述本发明原理、方面和实施方案及其具体实施例时的所有陈述旨在涵盖其结构和功能等同物两者。另外，此类等同物旨在包括当前已知的等同物以及将来开发的等同物，即，进行相同功能的所开发的任何要素，而无论结构如何。本发明的范围因此不旨在限于本文所示和所述的示例性实施方案。相反，本发明的范围和精神由所附权利要求书体现。

[0129] 示例实施方案：

[0130] 1.一种估计多个样品中所测序的起始多核苷酸分子的数量的方法，所述方法包括：

[0131] 将衔接子附接到多个不同样品中的起始多核苷酸分子，其中每个样品对应的所述衔接子包含：

[0132] 所述样品特定的单一MID；和

[0133] 简并碱基区域(DBR)，其包含选自以下的至少一个核苷酸碱基：R、Y、S、W、K、M、B、D、H、V、N及其修饰形式；

[0134] 将所述多个不同的衔接子附接的样品混合以生成混合样品；

[0135] 扩增所述混合样品中的所述衔接子附接的多核苷酸；

[0136] 对多个所述扩增的衔接子附接的多核苷酸测序，其中获得所述多个衔接子附接的多核苷酸的每一个的所述MID、所述DBR和所述多核苷酸至少一部分的序列；以及

[0137] 确定存在于来自每个样品的所述多个测序衔接子附接的多核苷酸中的不同DBR序

列的数量,以确定已在测序步骤中测序的每个样品中的起始多核苷酸的数量。

[0138] 2.根据项1所述的方法,其中所述衔接子进一步包含测序引物位点。

[0139] 3.根据项1所述的方法,其中所述衔接子进一步包含以下一者或多者:反射序列和启动子位点。

[0140] 4.根据项1所述的方法,其中所述DBR包含至少2个核苷酸碱基,其中所述至少2个核苷酸碱基的每一个选自:R、Y、S、W、K、M、B、D、H、V和N。

[0141] 5.根据项4所述的方法,其中所述DBR包含3至20个核苷酸碱基,其中所述3至10个核苷酸碱基的每一个选自:R、Y、S、W、K、M、B、D、H、V和N。

[0142] 6.根据项1所述的方法,其中将已在测序步骤中测序的每个样品中的多核苷酸的确定数量用于等位基因调用方法。

[0143] 7.根据项1所述的方法,其中所述测序步骤包括进行下一代测序过程。

[0144] 8.根据项1所述的方法,其中所述多个样品中的每一个为基因组DNA样品。

[0145] 9.根据项8所述的方法,其中所述多个不同的基因组DNA样品中的每一个源自不同的受试者。

[0146] 10.根据项9所述的方法,其中所述受试者为人。

[0147] 11.根据项1所述的方法,而且其中在所述附接步骤前针对感兴趣区域富集所述样品。

[0148] 12.根据项1所述的方法,其进一步包括针对感兴趣区域富集所述衔接子附接的多核苷酸。

[0149] 13.根据项4所述的方法,其中所述衔接子为不对称衔接子,其中第一衔接子结构域存在于所述多核苷酸的第一末端并且第二衔接子区域存在于所述多核苷酸的第二末端,其中所述DBR为包含所述第一衔接子结构域中的所述至少2个核苷酸碱基的一个或多个以及所述第二衔接子结构域中的所述至少2个核苷酸碱基的一个或多个的断裂DBR。

[0150] 14.根据项1所述的方法,其中将所述衔接子在扩增反应中附接到所述多核苷酸分子,其中所述DBR存在于用于所述扩增反应的合成引物中。

[0151] 15.根据项14所述的方法,其中所述扩增反应为PCR。

[0152] 16.根据项14所述的方法,其中所述方法进一步包括确定在所述PCR反应中扩增的起始多核苷酸的数量。

[0153] 17.根据项1所述的方法,其中所述方法涉及确定所述样品中所述多核苷酸的遗传异质性。

[0154] 18.根据项17所述的方法,其中所述样品包含源自肿瘤组织的多核苷酸。

[0155] 19.根据项17所述的方法,其中所述样品包含源自微生物和/或病毒的多核苷酸。

[0156] 20.根据项1所述的方法,其中所述方法涉及确定存在于所述多个不同样品中的所述多核苷酸的遗传异质性。

[0157] 21.根据项20所述的方法,其中所述多个不同的样品源自肿瘤的不同切片。

[0158] 22.根据项20所述的方法,其中所述多个不同的样品源自受试者的不同肿瘤。

[0159] 23.根据项22所述的方法,其中所述多个不同的样品源自在不同时间的受试者。

[0160] 24.根据项23所述的方法,其中所述受试者存在感染,其中确定一种或多种感染原在不同时间的遗传异质性。

序列表

<110> 安捷伦科技有限公司

<120> 通过分子计数提高等位基因调用的置信度

<130> CGLC-025WO

<150> 61/385,001

<151> 2010-09-21

<150> 61/432,119

<151> 2011-01-12

<160> 2

<170> FastSEQ Windows 4.0版

<210> 1

<211> 50

<212> DNA

<213> 人工序列

<220>

<223> 合成序列

<400> 1

cctatcccct gtgtgccttg gcagtctcag ggacacccag ccaagacagc 50

<210> 2

<211> 380

<212> DNA

<213> 人工序列

<220>

<223> 合成序列

<400> 2

ccatctcatc cctgcgtgtc tccgactcag rybdhvbacr tagaatgtgc atggatcgta 60

tgagcacctg tgggcagggc aagtggcaga tgccttagtg gatctcactg gaagcctggc 120

agaaaggtgg agcttgaagg atgtaacgaa agccagcggc cagcaggaca gacccagtgg 180

tggggagcac agaacttgtc ggcagctact ccacctgaag gaccggtgtc taatcagctg 240

ctctgtgctt agccccagag caggtacagc tatggctaca actccctcca ccattagctt 300

gttacagaga aggaaatcgg tccttgagag gctgtcttgg ctgggtgtcc ctgagactgc 360

caaggcacac aggggatagg 380

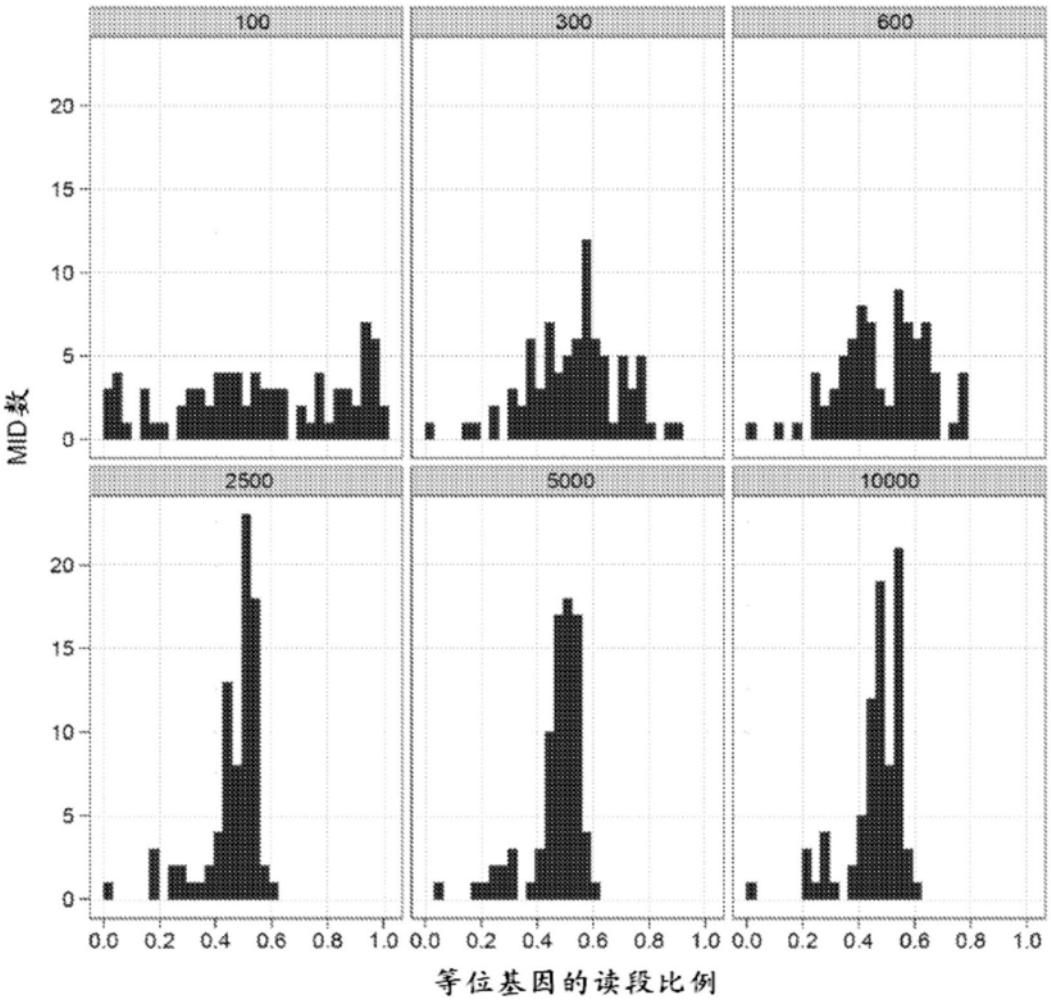

100
300
600
2500
5000
10000
MID数
0
5
10
15
20
0.0
0.2
0.4
0.6
0.8
1.0
等位基因的读段比例

图1

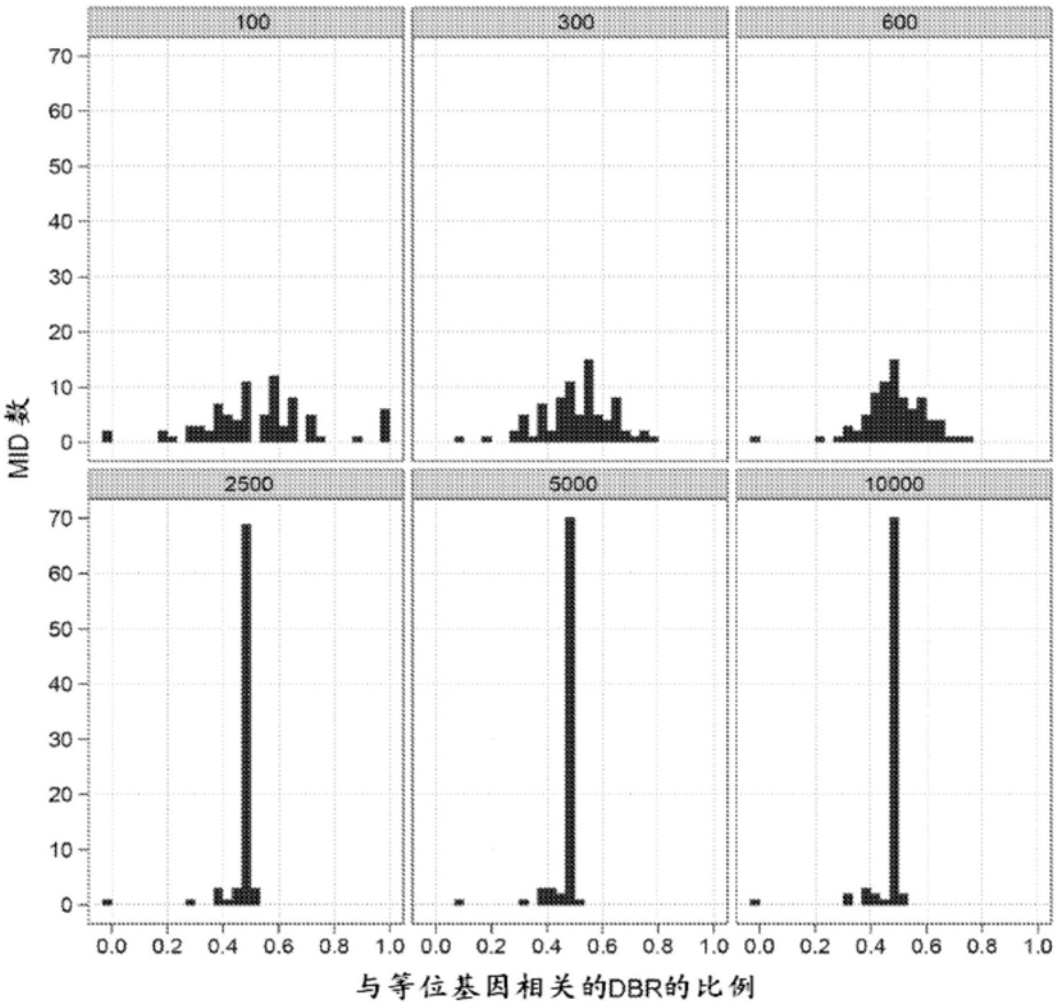

100
300
600
2500
5000
10000
MID 数
0
10
20
30
40
50
60
70
0.0
0.2
0.4
0.6
0.8
1.0
与等位基因相关的DBR的比例

图2

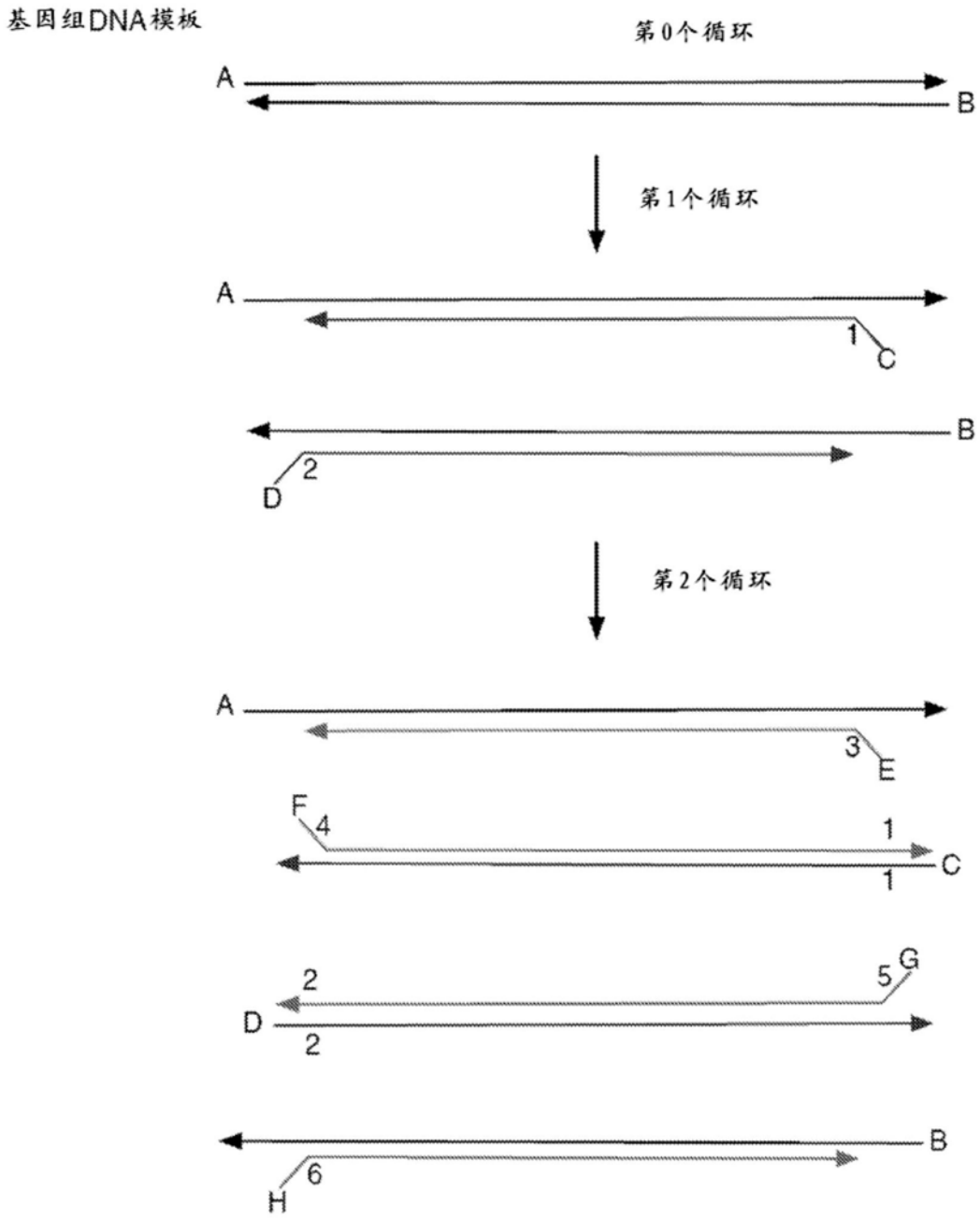


图3