[19] Patents Registry
The Hong Kong Special Administrative Region
香港特別行政區
專利註冊處

[11] 1151069 B
CN 101910413 B

[12] STANDARD PATENT SPECIFICATION
標準專利說明書

[21] Application No. 申請編號
11105101.2

[22] Date of filing 提交日期
23.05.2011

[51] Int.Cl.[8] C12Q

[54] METHOD OF POOLING SAMPLES FOR PERFORMING A BIOLOGICAL ASSAY 用於進行生物測定的合併樣本的方法

[30] Priority 優先權
31.10.2007 EP 07119761.0

[43] Date of publication of application 申請發表日期
20.01.2012

[45] Publication of the grant of the patent 批予專利的發表日期
22.11.2013

CN Application No. & Date 中國專利申請編號及日期
CN 200880123442.8 31.10.2008

CN Publication No. & Date 中國專利申請發表編號及日期
CN 101910413 08.12.2010

Date of Grant in Designated Patent Office 指定專利當局批予專利日期
14.08.2013

[73] Proprietor 專利所有人
HENDRIX GENETICS B.V.
Spoorstraat 69
NL-5831 CK Boxmeer
NETHERLANDS
亨德里克斯基因有限公司
荷蘭

[72] Inventor 發明人
VEREIJKEN, ADRIANUS LAMBERTUS JOHANNUS 阿德里安烏斯・拉姆貝圖斯・約翰納斯・韋雷吉肯
JUNGERIUS, ANNEMIEKE PAULA 安內米克・波拉・容格烏斯
ALBERS, GERARDUS ANTONIUS ARNOLDUS 赫拉爾杜斯・安東尼厄斯・阿諾爾德斯・阿爾貝斯

[74] Agent and / or address for service 代理人及/或送達地址
Kangxin Partners PC (Hong Kong) Limited
Suite 501, Enterprise Place
No. 5 Science Park West Avenue
Hong Kong Science Park HONG KONG

(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 101910413 B
(45)授权公告日 2013.08.14

(21)申请号 200880123442.8
(22)申请日 2008.10.31
(30)优先权数据
07119761.0 2007.10.31 EP
(85)PCT申请进入国家阶段日
2010.06.29
(86)PCT申请的申请数据
PCT/NL2008/050687 2008.10.31
(87)PCT申请的公布数据
WO2009/058016 EN 2009.05.07
(73)专利权人 亨德里克斯基因有限公司
地址 荷兰博克斯梅尔
(72)发明人 阿德里安乌斯·拉姆贝图斯·约翰纳斯·韦雷吉肯
安内米克·波拉·容格乌斯
赫拉尔杜斯·安东尼厄斯·阿诺尔德斯·阿尔贝斯
(74)专利代理机构 北京康信知识产权代理有限责任公司 11240
代理人 李丙林 吴孟秋
(51)Int.Cl.
C12Q 1/68(2006.01)

(56)对比文件
WO 2005075678 A1,2005.08.18,
US 2003152942 A1,2003.08.14,
US 2002172965 A1,2002.11.21,
LINDROOS K,等.Multiplex SNP genotyping in pooled DNA samples by a four-colour microarray system.《NUCLEIC ACIDS RESEARCH》.2002,第30卷(第14期),E70-1.
HOH JOSEPHINE,等.SNP haplotype tagging from DNA pools of two individuals.《BMC BIOINFORMATICS》.2003,第4卷(第14期),

审查员 刘铮

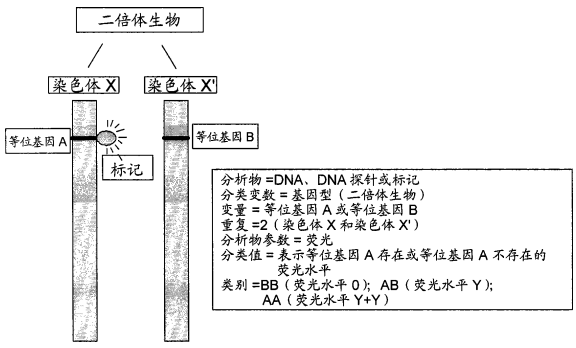
权利要求书1页 说明书24页 附图4页

(54)发明名称

用于进行生物测定的合并样本的方法

(57)摘要

本发明涉及一种合并样本以分析分类变数的方法,其中该分析涉及分析物的定量测量,所述合并样本的方法包括提供n个样本的池,其中该池中的个体样本的量是这样的使得该样本中的分析物以x^0:x^1:x^2:$x^{(n-1)}$的摩尔比存在,其中x代表分类变数的类别数,其为2或更大的整数。



CN 101910413 B

1. 一种合并样本以分析分类变数的方法，其中，所述分析涉及分析物的定量测量，所述合并样本的方法包括提供 n 个样本的池，其中所述池中的个体样本的量为这样的，使得所述样本中的所述分析物以 x^0 ： x^1 ： x^2 ： $x^{(n-1)}$ 的摩尔比存在，其中 x 代表所述分类变数的类别数，其为 2 或更大的整数，且其中，所述分析物是生物分子，而所述分类变数是所述生物分子的变量。

2. 根据权利要求 1 所述的方法，其中，所述生物分子是核酸。

3. 根据权利要求 2 所述的方法，其中，所述变量是所述核酸的核苷酸多态性。

4. 根据权利要求 3 所述的方法，其中，所述核苷酸多态性是 SNP。

5. 根据权利要求 2 所述的方法，其中，所述变量是特定核苷酸位置的碱基特征。

6. 根据前述权利要求中任一项所述的方法，其中，所述定量测量包含仪器信号的强度、峰高度或峰面积的测量。

7. 根据权利要求 6 所述的方法，其中，所述仪器信号是荧光信号。

8. 根据权利要求 1-7 中任一项所述的方法用于单倍体或多倍体个体的等位基因变量的基因分型的应用，其中，所述分类变数的类别数 (x) 等于 p+1，其中 p 代表倍性水平。

9. 根据权利要求 8 所述的应用，其中，对于二倍体个体中的等位基因变量的基因分型，x 是 3。

10. 一种对多个样本进行分析的方法，包含根据权利要求 1-7 中任一项所述的方法合并所述样本，从而提供合并样本并对所述合并样本进行所述分析。

11. 一种对多个样本进行分析的方法，包括对通过权利要求 1-7 中任一项所述的方法获得的一系列合并样本进行分析，其中对所述样本的分类变数进行分析，并涉及所述样本中的分析物的定量测量。

12. 根据权利要求 11 所述的方法，还包括从所述测量推断所述样本池中的所述个体样本的贡献。

用于进行生物测定的合并样本的方法

技术领域

[0001] 本发明涉及对生物样品进行测量的具有分类结果的测量领域，更特别的是涉及用于具有分类结果的生物测定的样本制备方法。本发明提供了合并样本的方法和所述方法在等位基因变量的基因分型中的应用。本发明还提供了对多个样品进行分析的方法、将多个样本合并成合并样本的合并装置、包含用于对一系列合并样本进行分析的处理器的分析装置、实施合并样本方法的计算机程序产品和实施对多个样本进行分析的方法的计算机程序产品。

背景技术

[0002] 生物测定是测定样本中的生物分析物的特性、浓度或存在的方法。生物测定是所有科学领域研究中的固有部分，最显著的是在生命科学领域，尤其是分子生物学领域。

[0003] 分子生物学中的一种特定类型的分析涉及基因分型和测序。基因分型和测序是指利用生物测定确定个体基因型的过程。目前的方法包括 PCR、DNA 和 RNA 测序、以及杂交至固定于各种载体（如玻璃片或珠）上的 DNA 或 RNA 微阵列。该技术对于父亲 / 母亲身份的检验、研究疾病相关基因的临床研究和旨在研究任意物种特性的遗传控制（例如扫描整个基因组搜寻 QTL（数量性状基因座））的其他研究是必需的。

[0004] 由于目前技术所限，几乎所有的基因分型只是部分的。也就是说，只确定了个体基因型的一小部分。在许多情况下，这并不是问题。例如，当检验父亲 / 母亲身份时，只研究 10 至 20 个基因组区域来确定是否具有亲缘关系，这 10 至 20 个基因组区域只是人类基因组的一小部分。

[0005] 单核苷酸多态性 (SNP) 是基因组中最丰富类型的多态性。随着高密度 SNP 标记图谱和高通量 SNP 基因分型技术的并行发展，SNP 已经成为许多遗传研究所选用的标记。在绘制图谱和关联研究或基因组筛选实验中都需要大量的样本。

[0006] 为了提供高通量基因分型能力，已经开发了阵列技术。这样的技术可由供应商处获得，例如 Affymetrix（基于微阵列的基因芯片（GeneChip®）映射阵列）、Illumina(BeadArray™)、Biotrove(OpenArray™) 和 Sequenom(MassARRAY™)。现在能够获得或在不远的将来能够获得许多物种（人、牲畜、植物、细菌和病毒）的大量的 SNP。创新的技术已经能够完成全基因组的基因分型或关联研究和用于植物和动物育种的相关全基因组筛选程序。但这样的方法的费用仍然非常高，如果样本单独地进行基因分型，则需要高达数百万美元的预算。因此，旨在确定任意物种的 SNP 的研究目前只涉及有限数量个体的分析。因此，由于本发明可使基因分型的费用明显降低，因此本发明非常重要。

[0007] 为了全面了解基因多变性，必须知道基因组（相关部分）的完整序列。但是，确定完整序列的费用甚至要高于前面段落所描述的基因分型的费用。除了费用以外，测序将取代基因分型从而提供个体基因型的全基因组或其特定区域还是期望的。本发明还提供了降低测序费用的方法。

[0008] 样本合并 (pooling) 经常作为减少分析费用的方法被用于分类性状

(categorical traits) 的研究。由几个样本的混合物构成的池 (pool) 中的特性的存在表明该池中至少一个样本具有该特性。例如，DNA 池被用于：

[0009] - 估算群体中的等位基因频率。

[0010] 通过从该群体选取适合的个体样本，等位基因 1 的粗略等位基因频率被计算为池中的等位基因 1 的结果与等位基因 1 的结果和等位基因 2 的结果之和的比值。

[0011] - 事件 (case)- 对照关联研究，其中事件和对照被分为不同的池，以及

[0012] - 在少数个体和少数 SNP 上重新构建单倍体型。

[0013] 根据在池中测量的等位基因频率，可通过不同的算法（如最大似然法）来估计单倍体型。术语“单倍体型频率”与术语“标记的联合分布”含义相同。

[0014] 样本合并的一个重要缺点是被测量的特性仅仅在作为整体的池中被确定，而不是在该池中的任意个体样本中被确定。一个例外是当建立了分别由两个个体（父亲 + 孩子和母亲 + 孩子）构成的两个池时的用于三个个体（父亲、母亲和孩子）基因分型的 DNA 池。在每个池中观察到的等位基因频率显示了所有 3 个个体的基因型。这种类型的样本合并使费用降低了 33％，但是只有是这样的三个个体时才是可能的。在所有的其他情况下，为了获得个体样本的结果，必须对合并样本中的个体进行重新分析。

[0015] 因此，提供三个个体以外的样本类型的样本池，并且仍能提供该池中的个体样本的测试结果是有利的。

发明内容

[0016] 现在，本发明的发明人已发现可合并随机个体，并且当该池中的每一个体样本的贡献对于每一其他样本的贡献是固定的比例时，即，当样本量不是以等摩尔 (equimolar) 而是以特定的比例提供时，能够从该池得到个体基因型。如果测试涉及分类变数的定量测量，即，该测试涉及被定量测量的分类或离散性状，可从合并的测试结果推断个体样本的结果。

[0017] 事实上，本发明的发明人已发现，对于二倍体动物中的某一位点处的某一等位基因存在的研究，以 1 ： 3 的比例混合在单个位点处具有 2 个可能的等位基因 (A 或 B) 的第一二倍体动物的 DNA 样本和在相同位点处也具有 2 个可能的等位基因 (A 或 B) 的第二二倍体动物的 DNA 样本，这导致该混合物中的任一等位基因存在 (2)+(2+2+2) ＝ 8 种可能性，其中单个等位基因（例如 A) 的预期定量仪器信号是最大样本信号强度的 12.5％。这表明，当测量的信号强度是最大样本信号强度的 37.5％，则该样本包含 3 倍 (3x) 的该等位基因 A，这表示该信号不可能是来源于第一二倍体动物，而只能来源于第二二倍体动物，这表示第一二倍体动物具有基因型 BB，而第二二倍体动物具有基因型 AB。同样，当测量的信号强度是最大样本信号强度的 50％时，所有的样本都具有基因型 AB。当测量的信号强度是最大样本信号强度的 0％时，则所有的样本都具有基因型 BB。该池中的两个个体总共具有 3*3 种可能的基因型。如果测量的精确度是至少 6.25％，每一测量可被分配 100％的八分之一 (1/8) 的值或其倍数的值。一般来说，每一种可能的测量结果可被分配 $1/(y*((p+1)^0+(p+1)^1+(p+1)^2+(p+1)^{(n-1)}))*100\%$ 的值，其中 y ＝ 2（等位基因 A 在一个位置上的两种可能的结果，等位基因存在或不存在），p 是倍性水平，n 是样本的数量，100％是最大样本信号强度。总体上，会有（倍性水平 +1)n 种可能的基因型。

[0018] 现在，当合并样本是比例为 1 ： 3 ： 9（也就是合并因子分别为 3）的 3 种动物（x、y 和 z）的合并样本时，理论上对于该混合物中的任一等位基因总共有 26 种可能，其中单个等位基因（例如 A）的预期定量信号是最大样本信号强度的 3.85％。这表示测量信号强度是最大样本信号强度的 12％，包含 3 倍（3x）等位基因 A 的样本显示动物 x 具有基因型 BB，动物 y 具有基因型 AB，且动物 z 具有基因型 BB。同样，当测量的信号强度是最大样本信号强度的 96％时，样本 x 具有基因型 AB，而样本 y 和 z 具有基因型 AA。如果测量的精确度是至少 1.9％，每一测量可被分配 100％的二十六分之一（1/26）或其倍数的值。（这样的合并实验的可能结果的综述请参见以下的实施例）。

[0019] 本发明的发明人已经示出了该法则可被用于涉及样本中的分析物的定量测量的大量分析，其中分析的结果是根据所述样本中的分析物的性质来分类的。

[0020] 在第一方面，本发明现提供了合并样本以分析分类变数的方法，其中该分析涉及分析物的定量测量，所述合并样本的方法包含提供 n 个样本的池，其中该池中的个体样本的数量是这样的使得样本中的分析物以 x^0 ： x^1 ： x^2 ： $x^{(n-1)}$ 的摩尔比存在，其中 x 代表分类变数（或合并因子）的类别数，x 是 2 或更大的整数，如 3、4、5、6、7 或 8，优选为 2 或 3，n 是样本的数量。x^0 ： x^1 ： x^2 ： $x^{(n-1)}$ 应被理解为表示 x^0 ： x^1 ： x^2 ： ... ： $x^{(n-1)}$ 或 x^0 ： x^1 ： x^2 ： x^i ；$x^{(n-1)}$，其中 n 是样本数目，i 是其值在 2 和 n 之间的逐渐递增的整数。

[0021] 对于合并多倍体个体，x 等于（倍性水平 +1），所以对于在一个位置具有两个可能的等位基因的单倍体而言 x ＝ 2，对于二倍体而言 x ＝ 3，对于四倍体而言 x ＝ 5，x 也等于可能的基因型的数目。

[0022] 假定具有三个可能的等位基因，则单倍体具有 3 种可能的基因型（x ＝ 3），二倍体具有 6 种可能的基因型（x ＝ 6），三倍体具有 10 种可能的基因型（X ＝ 10）。在一个二倍体个体中，第一等位基因可出现 0、1 或 2 次，第二和第三等位基因也是如此。这使得以如具有两个等位基因（x 也是多倍性水平 +1）相同的比例（x^0 ： x^1 ： x^2 ： $x^{(n-1)}$）合并成为可能。3 个等位基因的信号强度被四舍五入至最接近的结果点（$1/(y*((p+1)^0+(p+1)^1+(p+1)^2+(p+1)^{(n-1)}))*100\%$，其中 y ＝ 2（等位基因 1、2 或 3 存在或不存在），p ＝倍性水平，n ＝样本数量）从而得到合并样本中的等位基因数目。

[0023] 因此，池中的两个个体样本之间的比例（作为一个实施例）是这样的，使得其中的分析物以 1 ： x 的摩尔比存在，其中 x 是分类性状的类别的最大数目。

[0024] 其中池中的个体样本数量规定为公比为 3 的等比数列的方法尤其适合于二倍体个体中的等位基因变量的基因分型，其中每个个体具有三种可能的基因型。该基因型是具有三种可能的变量（AA、AB 和 BB）的分类性状。

[0025] 其中池中个体样本数量规定为公比为 2 的等比数列的方法尤其适合于单倍体个体中的等位基因变量的基因分型。对于其实施例，参考以下的实验部分。

[0026] 在另一方面，本发明涉及以上所描述的本发明的方法在单倍体或多倍体个体中的等位基因变量的基因分型中的应用，其中分类变数（x）的类别数目等于 p+1，其中 p 代表所述个体的倍性水平。例如，这样的应用可以用于进行二倍体或单倍体个体中的等位基因变量的基因分型。

[0027] 另一方面，本发明涉及对多个样本进行分析的方法，其包含根据以上所述的本发明的方法合并所述样本从而提供合并样本，并对所述的合并样本进行所述分析。然后将得

到的量化结果四舍五入至最接近的结果点（由理论区间的数目确定，其中最大样本信号强度根据每个可能的结果来划分，见下文），信号强度被分配为合并样本的分类变数的类别总数。由此，考虑到池中各种个体样本的比例，确定池中的每一个体样本分类变数。

[0028] 在另外一个方面，本发明提供了对多个样本进行分析的方法，其包含对通过本文以上所确定的合并样本的方法获得的一系列（或一组）合并样本进行分析，其中分析所述样本的分类变数，并且涉及所述样本中的分析物的定量测量。

[0029] 在本方法的一种优选实施方案中，进行分析的方法还包含从测量推导所述样本池中的个体样本的贡献的步骤。

[0030] 在另外一个方面，本发明提供了将多个样本合并成为合并样本的合并装置，其包含提供合并样本的样本吸出器 (aspirator)，还包含进行以上所述合并样本方法的处理器。

[0031] 在另外一个方面，本发明提供了分析装置，其包含用于对通过以上所述的合并样本的方法获得的一系列合并样本进行分析的处理器，其中所述的装置被设置用于分析所述样本的分类变数并进行所述样本中的分析物的定量测量。

[0032] 在该分析装置的一种优选实施方案中，该装置还包含合并装置，最优选是上文中所披露的合并装置。

[0033] 在另外一个方面，本发明提供了在计算机程序产品自身或在载体上的计算机程序产品，当该程序产品在计算机、编程的计算机网络或其他可编程设备中被加载 (load) 并且执行时，实施上文所述的合并样本的方法。

[0034] 在另外一个方面，本发明提供了在计算机程序产品自身或在载体上的计算机程序产品，当该程序产品在计算机、编程的计算机网络或其他的可编程设备中被加载并且执行时，实施以上所述的对多个样本进行分析的方法，所述方法包含对通过上文所述的合并样本的方法获得的一系列合并样本进行分析，其中分析所述样本的分类变数，并且涉及所述样本中的分析物的定量测量。

[0035] 在该计算机程序产品的一种优选实施方案中，所述方法还包含根据以上所述的合并样本的方法合并的步骤。

[0036] 通过利用本发明的方法，分析费用可大幅降低，即，一般降低 50％，甚至降低 66％或更多。

附图说明

[0037] 图 1 示出了基于合并数据的等位基因频率 (Y 轴）与基于个体测量的等位基因频率 (X 轴）之间相关性的曲线图。

[0038] 图 2 示出了个体测量的等位基因频率 (Y 轴）与池中的预测的等位基因频率 (X 轴）之间关系的曲线图。

[0039] 图 3 示出了池中的校正的等位基因频率 (Y 轴）与个体分型后测量的个体的等位基因频率 (X 轴）之间的关系的曲线图。

[0040] 图 4 示出了实验 1 中的预期的（基于个体分型）与对池 1 的预测的等位基因频率之间的差异的曲线图。

[0041] 图 5 示出了实验 2 中的预期的（基于个体分型）与对所有池的预测的等位基因频率之间关系的曲线图。

[0042] 图 6 表示了实验 2 中的预期的（基于个体分型）和对所有池预测的等位基因频率之间的差异的曲线图。

具体实施方式

[0043] 本文中使用的术语“分类变数”是指例如性质或性状的离散变数，例如分析物或其性质是否存在，或等位基因性状在分析物中是否以纯合或杂合的形式存在。“离散的”与“分类的”具有相同的含义，是指非线性的或不连续的。“变数”通常是指测量样本特性的（分类）性状。分类变数可以为二元的（由两类构成）。“类”是指可进行测量的组或类别。因此，纯粹的分类变数是可以分配类别的，分类变数取几个可能的类别（类）之一的值。尤其是，分类变数可能涉及遗传标记的存在，如单核苷酸多态性 (SNP) 或任意其他的遗传标记、等位基因、免疫反应、疾病、抗性能力、发色、性别、疾病感染状态、基因型或样本或生物体的任意其他性状或特性。虽然它们能够被量化地测量，例如作为可被分析装置接收、读取和 / 或记录的所产生的分析物信号，分类变数本身不具有数量意义而且类别不具有固有的排序。例如，性别是具有两个类别（男性或女性，通常被编码为 0 和 1）的分类变数，优选地代表了无序的类别。基因型也是具有多个优选无序的类别的分类变数 (AA、Aa 和 aa，有时候被编码为 2、1 和 0)。

[0044] 在本发明的一些方面，样本可以为测量了分类变数的任意样本。该样本可以为生物样本如动物（包括人）或植物的组织或体液样本，环境样本如土壤、空气或水样本。样本可以为（部分）纯化的或未经处理的（原始）样本。样本优选为核酸样本，例如 DNA 样本。

[0045] 在定量测试中测量其存在或形式的分析物可以为化学物质或生物体。在优选的实施方案中，分析物是生物分子，分类变数是所述生物分子的变体 (variant)。优选地，该生物分子是核酸，尤其是多核苷酸，如 RNA、DNA，并且该变量可以例如为所述多核苷酸中的核酸多态性，例如等位基因变量，最优选为 SNP，或特定核苷酸位置的碱基一致性。

[0046] 因此，本文定义的分析物可以是表现出一定分类变数的 DNA 分子（例如，该核酸分子中的特定核苷酸位置的碱基特征 (identity，同一性），具有 A、T、C 或 G 的分类值）。可利用定量测试来测量特定核苷酸位置的碱基特征，例如根据来自整合了荧光类似物的所述核苷酸的 cDNA 拷贝的荧光，如 DNA 测序领域中已知的。DNA 特定位置中的类似物放射出的、并通过分析装置进行测量的定量水平的荧光，为该核苷酸位置分配分类值，例如该位置为腺嘌呤。

[0047] 在确定特定核苷酸位置的碱基特征中，本发明涉及合并待确定其特定核酸的核苷酸序列的个体样本。当认识到序列测定涉及确定四种可能碱基中的任一种的信号（其中在例如测序凝胶中存在或不存在特定位置的任一特定碱基的信号对应于所述核酸中的特定核苷酸位置中存在或不存在该碱基特征）时，就能够理解本发明的方法适合用于序列测定（分析）。在运行测序凝胶 (sequence gel) 之前，以本文中所述的比例合并两个样本使得能够确定任意特定信号的来源，并由此确定每个个体核酸的序列。

[0048] “分析物”可以是多肽，例如蛋白质、肽或氨基酸。该分析物还可以为核酸、核酸探针、抗体、抗原、受体、半抗原和受体的配体或其片段，（荧光）标记、色原体、放射性同位素。事实上，分析物可以由可定量测量并且可用于确定分类变数的类别的任何化学或物理物质形成。

[0049] 本文中所使用的术语“核苷酸”是指包含连接糖（一般为核糖(RNA)或脱氧核糖(DNA)）的C-1碳的嘌呤（腺嘌呤或鸟嘌呤）或嘧啶（胸腺嘧啶、胞嘧啶或尿嘧啶）碱基的化合物，并且进一步包含一个或多个连接于该糖的C-5-碳的磷酸基。该术语包括核酸或多核苷酸的个体构建体(building block)，其中个体核苷酸的糖单元通过磷酸二酯桥相连从而形成具有待定的嘌呤或嘧啶碱基的磷酸糖骨架。

[0050] 本文中所使用的术语“核酸”包括单链或双链形式的脱氧核糖核苷酸或核糖核苷酸的聚合物，即多核苷酸，除非另有限定，该术语包含具有天然核苷酸的必要特性的已知类似物（例如，肽核酸），因为它们以与天然存在的核苷酸相似的形式与单链核酸杂交。多核苷酸可以为天然的或异源结构或调控基因的全长序列或子序列。除非另外指出，该术语包括特定的序列以及其互补序列。因此，具有为了稳定或其他原因而经修饰的骨架的DNA或RNA就是本文中的该术语意指的“多核苷酸”。此外，包含独特碱基，如次黄嘌呤核苷，或修饰的碱基，如三苯甲基化的碱基（只是定义了两个实例）的DNA或RNA是如本文所用的术语意指的多核苷酸。

[0051] 术语“定量测量”是指确定样本中的分析物的量。术语“定量”是指该测量可表达为数值的事实。该数值可涉及度量、尺寸、程度、数量、容量、浓度、高度、深度、宽度、广度、长度、重量、体积或面积。定量测量可涉及测量信号的强度、峰高或峰面积，例如显色或荧光信号，或任何其他的定量信号。一般来说，当确定分析物的存在或形式时，测量会涉及仪器信号。例如，当确定SNP的存在时，测量会涉及杂交信号，该测量通常会提供由荧光计测量的荧光强度。当确定免疫响应的存在时，该测量会涉及抗体效价的测量，该测量通常也可以以荧光强度提供。测量不需要提供连续的测量结果，但是会涉及离散区间或类别。测量也可为半定量的。只要能够在2^{n-1}、3^{n-1}或x^{n-1}偏微分(partial)中确定该测量，并且优选是最大样本信号强度（取决于该池是否分别以公比2、3或x的等差数列提供，其中n是池中样本的数量）的比例区间，该测量理论上就是合适的。

[0052] 本文中所使用的术语“合并(pooling)”是指为了最有利于使用者而将样本组合或汇合在一起。尤其是，术语“合并”是指制备多个样本的集合来代表一个具有加权值的样本。通常通过混合样本而将多个样本合并成一个单一样本。在本发明中，混合要求仔细称重单个样本的量，其中每个样本中存在的分析物的量是明确的。当样本A中的分析物的量为2g/L，样本B中的分析物的量为1g/L时，以1：6的体积比将这些样本合并从而提供1：3的分析物比例。

[0053] 当以例如1：3的比例将两个样本合并，或当样本以本发明的实施方案中描述的以1：3：9的比例将三个样本合并时，分别由12.5％和3.85％的区间端点来设定池中变量的可能频率。这些区间的端点在本文中被称为“结果点”，并且相当于定量测量值的逐步增加(stepincrements)，直到达到最大样本信号强度。

[0054] 术语“等比数列”是指其中任意两个连续项之间的比为相同的数列。换句话说，通过每次将前一项乘以相同的数字可获得数列中接下来的项。这个固定的数字被称为该数列的公比。在本发明的等比数列中，第一项为1，根据样本的类型，公比为2或3。

[0055] 术语“最大样本信号强度”是指当合并的所有样本产生阳性信号时（即，当100％的个体样本对于测试的分析物均为阳性时），从该池中得到的信号。可通过任意合适的方法来确定最大样本信号强度。例如，可以分别测量50个个体样本从而根据离散事件的数目来

确定它们在这些样本中存在的组成，随后可在合并实验中测量这些样本，其中所测量的合并样本的信号强度以相同的比例示出，通过累加所有的个体样本的信号强度获得。

[0056] 本发明的方法可以以任意数目的 n 个样本来进行。但是，在实践中，最大数目 n 根据测量方法的精确度而设定，也就是说，能够确定两个连续的结果点之间的合理的统计学差异的精确度。本方法的精确度（标准偏差）必须与此相符。

[0057] 本发明的方法的应用包括，但不限于，基因分型方法。基于合并 DNA 的基因分型具有多种应用。基因型可被用于所有物种的图谱绘制、关联和诊断。具体的基因分型的实例包括 a) 人类的基因分型，如医学诊断，以及病例 - 对照研究合并之后的追踪个体分型 ；b) 候选基因方法和基因组广泛筛选应用中的牲畜的基因分型，如 QTL 研究中的个体分型，和 c) 植物的基因分型，例如，为了绘制图谱和关联研究。

[0058] 当对人类、牲畜、植物、细菌、病毒进行序列测定时也可以使用合并。更具体而言，当想要比较两个或多个个体的序列时，合并个体样本进行测序是合适的。

[0059] 本发明的合并样本的方法包含从至少一个第一样本取子样本和从至少一个第二样本取子样本，其中所述第一和第二子样本被混合到同一容器中从而提供合并样本形式的两个子样本的混合物，其中根据本文所述的分析物的浓度，所述的合并样本中的所述的第一子样本和第二子样本的比例为 1 ： 3 或 3 ： 1。类似地，当三种样本被合并时（该说法是指混合三个子样本的事实），所得的合并样本中的第一、第二和第三子样本（任意顺序）的比例为如本文中所描述的为 1 ： 3 ： 9。根据 12.5％和 3.85％的区间端点分别设定池中的变量的可能频率。这些区间的端点在本文中是指“结果点”，并且相当于逐步增加 (step increment)，直到达到最大样本信号强度。

[0060] 本文定义的合并方法可通过（使用）合并装置来进行。这样的装置应当包含用于收集和递送确定量（例如以确定（但是可变的）体积的形式）的样本的样本收集器。合适的样本收集器是移液管操纵器 (pipettor)，如通常实验室中常用的自动样本递送和处理系统。这样的自动系统通常是台式设备，其应当包含微孔板处理器台、试剂操作台、过滤板吸气器和基于气体力学的自动移液器模块和一次性吸头中的一个或多个。这些样本自动系统非常适合于实施本发明的方法，因为它们从根本上被设计用于从不同样本将不同的液体体积合并到一个或多个反应管中。因此，它们是在技术人员的技能范围内的，从而将这样的自动移液管系统应用于执行从不同样本将不同液体体积合并成一个单一的合并样本的任务。但是，这样的自动移液管系统只是将多个样本合并为合并样本的样本合并装置中的一个合适的实施方案，所述的装置包含用于从多个样本瓶收集样本并且用于将样本递送到单个合并瓶中从而提供合并样本的样本收集器，并且还包含用于执行本文所定义的合并样本方法的处理器。本文中所使用的术语“处理器”意指包括任意的计算机设备，其中使用一个或多个执行单元（例如包含移液管装置和在样本瓶和自动移液管系统的合并瓶之间移动所述移液管装置的机械臂的部件）来执行存储的指令和从存储器或其他的存储设备检索的指令。术语“瓶 (vial)”应该是泛指，并可包括指阵列上的分析点。因此，本发明的处理器可包括，例如个人计算机、大型计算机、网络计算机、工作站、服务器、微处理器、DSP、专用集成电路 (ASIC)，以及其部分或组合和其他类型的数据处理器。设置所述处理器以用于接收上文中限定的合并装置上实施本发明的合并样本的方法的计算机程序的指令。

[0061] 合并样本以用于分析分类变数的方法，其中该分析涉及分析物的定量测量，所

述合并样本的方法包含提供 n 个样本的池，该池中的个体样本的量为样本中的分析物以 $x^0 : x^1 : x^2 : x^{(n-1)}$ 的摩尔比存在，其中 x 是 2 或更大的整数，其表示分类变数的类别数。

[0062] 虽然合并方法是非常直接的，并且能够以相对简单的公式表达，但本文所描述的分析合并样本的方法较为复杂。

[0063] 如本文所述，分类变数（例如基因型）可以取几个可能的类别 (BB、AB、AA) 中的一个类别的值。这些类别与结果区间的类相一致。通过对分析物 (DNA) 的参数（例如，荧光）实施定量测量可确定类别，并且根据分析结果的分类为这些参数分配类别，每一类代表了所述分类变数的一个变量（见图 7)。

[0064] 总的来说，可能的分析结果（输出）的总数取决于分类变数的性质。例如，在二倍体生物基因型的情况下，倍性水平决定了可能的分析结果的数目。一般来说，分类变数的性质可包括样本中存在不同数目的变量或系列分析物（仍参见图 7)。可能的分析结果的总数还取决于能够采用一次重复的可能的不同分类值。表 1 提供了可能的分析结果的数目的实施例。

[0065] 表 1. 当测量是由相同事件的重复构成时，其可能的分析结果（结局）的总数

[0066]

可能值	样本内的重复数目			
1 次重复	1	2	3	4
2	2	3	4	5
3	3	6	10	15
4	4	10	..	..
5	5	15	..	$\binom{n+k+1}{k}$

[0067] N 代表一次重复的可能的分类值或变量的数目，k 是样本内的重复数目。该表中提供的值是根据公式 $\binom{n+k+1}{k}$ 计算的。

[0068] 例如，二倍体个体（一个样本中一个等位基因有 2 个重复）的基因型为 3 种 (AA、AB 和 BB)，因为一个等位基因只能有两种不同的变量 (A 或 B)。三倍体（一个等位基因有 3 个重复）可具有 4 种不同的基因型 (AAA、AAB、ABB 和 BBB)

[0069] 个体的血型是具有四种不同的变量的一种重复 (A、B、AB 或 O)。

[0070] 表 1 中的公式对于测量的变量重复是不重要的情况来说是成立的。例如，对于基因型而言，基因型 AB 和基因型 BA 之间没有差异。但是，在重复的特征 (identity) 是重要的情况下，计算可能的分析结果的总数的公式为 n^k。则该公式替代了表 1 的公式 $\binom{n+k+1}{k}$。而且表中的所有的值随之相应地发生改变。对于有 2 个重复并且每个重复有 2 种可能的结果的情况，会有四种结果。对于有 3 个重复而且每个重复有 3 种可能的结果，则会有 9 种不同的结果。

[0071] 可能的分析结果的总数在本文中被用作合并比例（例如，1 ： 3 ： 9)，并且直接称为“合并因子”而被提供（在 1 ： 3 ： 9 的情况下为 3)。例如，当合并单倍体个体用于基因分型时，具有一个重复，每个重复有 2 种可能的变量。在这种情况下，合并因子等于 2（是表 1 中的结果数目）。

[0072] 合并 4 个个体则需要以 2^0 ： 2^1 ： 2^2 ： 2^3 的比例进行。

[0073] 当合并二倍体个体时，合并因子是 3。合并 3 个个体需要以 3^0 ： 3^1 ： 3^2 的比例进行。

[0074] 池中的结果的总数则等于以下公式：

[0075] 总合并结果＝合并因子^{样本数目}。

[0076] 则信号强度的增加（increment，或增量）等于：

[0077] 增加＝ 1/（合并因子^{样本数目}-1）＊ 100％

[0078] 或

[0079] 1/(y＊((合并因子)^0+(合并因子)^1+(合并因子)^2+.....+(合并因子)$^{(n-1)}$))＊ 100％，

[0080] 其中 n 是样本数目，y ＝合并因子 -1。

[0081] 如果测量强度对于一次重复的所有变量是存在的（为所有的值减去一，因为减少的一随后会被计算为 1 减去另外一个的强度），可遵循表 1 中的首行，因为这可以被视为该重复的每个值的存在或不存在，其对应于该重复的 2 个可能的结果。参见以上的实施例，其中假设有 3 种可能的等位基因而不是 2 个，并且可测量 3 种不同的光强度而不是 2 个（红和绿）。

[0082] 如果只进行单次测量，则可遵循表 1。

[0083] 如本文所述，本发明的用于分析合并样本的方法包含对所述合并样本上的所需的分析物实施测量。在记录测量结果（例如仪器信号）后，该分析包含一系列的步骤，在下文中提供的实施例中会详细解释这些步骤。

[0084] 对通过本发明的方法获得的一系列合并样本进行分析（其中分析所述样本的分类变数）涉及所述样本中分析物的定量测量。该分析物是化学或物理物质或实体，其参数显示所述分类变数的至少一个变量存在与否。例如，当具有变量等位基因 A 或 B 的生物基因型确定为分类变数时，分析物是生物体的 DNA、DNA 探针或遗传标记，该分析物的参数的绝对值与变量的存在（或不存在）直接相关。分析物的定量测量通常包括荧光强度、放射性同位素强度、或作为分析物参数的值的任意定量测量。超过一定阈值或分类值的测量值通常显示了变量的存在。因此，样本中分析物的定量测量是指分析物发出了在所述的样本中被分析的分类变数的变量存在或不存在的信号。

[0085] 基本上，在分析通过本文所述的合并样本的方法获得的合并样本的方法中，所述池中的个体样本的比例（即池中的个体样本的结果）确定如下。

[0086] 首先确定对 n 个样本的池进行的特定分析“A”的最大样本信号强度，并且设定为 100％信号。最大样本信号强度是当池中的 n 个样本的 100％样本对于分类变数为正 (positive) 的时候达到的信号强度。可通过提供 n 个阳性参考样本的测试池并且确定测量信号来确定最大样本信号强度，其中所述阳性参考样本对于分类变数为正，并且其中 n 是在其上进行分析“A”的池中的样本数目。记录分析“A”的最大样本信号强度或存储在计算机存储器中备用。接下来，通过分析“A”，在本发明的方法获得的合并样本中测量感兴趣的分析物，从而确定分析物的合并样本信号强度。记录合并样本中的分析物得到的信号强度，四舍五入至上文中确定的最接近的结果点，并且根据情况进行存储，然后与最大信号强度进行比较。该比较适合于这样进行。一般而言，为每个可能的测量结果分配值 1/

$(y*(3^0+3^1+3^2+3^{(n-1)}))*100\%$，其中 n 是合并样本的数目，y 是代表“A”存在与否的整数 2，100％是最大样本信号强度。$y*(3^0+3^1+3^2+3^{(n-1)})$ 应被理解为表示 $y*(3^0+3^1+3^2+3^i+3^{(n-1)})$，其中 n 是样本的数量，i 是具有 2 和 n 之间的值的递增的整数。例如，对于 y ＝ 2 个类别的分类变数（不存在或存在标记），以及具有 4 个样本的池，使用 4 种阳性参考样本将最大样本信号强度设定为 100％，总共有 $2*(3^0+3^1+3^2+3^3)$ ＝ 2+6+18+54 ＝ 80 个结果点，其中每个可能的测量结果可被分配 1/80*100％＝ 1.25％的值或其倍数。

[0087] 可从一个简单的结果表（其以计算机可读取的形式存储在计算机存储器中）中读取样本池中的每个样本的结果，该表为最大样本信号强度的 0％至 100％之间的递增步骤 $1/(y*(3^0+3^1+3^2+3^{(n-1)}))*100\%$ 的每一结果点分配了池中每个个体样本对应的值。例如，这样的结果表是以下的表 2 中提供的表。

[0088] 通过对所述合并样本中的各个子样本分配分类变数来完成分析。

[0089] 可通过分析装置来实施本文中定义的分析合并样本的方法。本发明的分析装置包含用于对通过以上所述的合并样本的方法获得的一系列合并样本进行分析的处理器，其中所述装置用于分析所述样本的分类变数并对所述样本中的分析物实施定量测量。如上文中所提到的，该分析装置的独特性质在于其用于分析所述池中的每个个体样本中的合并样本的分类变数，并对所述样本中的分析物实施定量测量。基本上，该分析装置用于测量和分析从合并样本获得的测量结果，并且从该结果推导出池中每个个体样本的分类变数。这样的装置应该包含用于测量合并样本中的分析物信号的信号读取单元。该分析装置还应当包含用于存储测量结果和以上所述的结果表的存储器。该分析装置还应包含用于从存储器和 / 或读取单元检索信息，并用于进行计算和进行迭代过程的处理器，其中使用上文中提到的结果表来将合并样本的测量结果与所述池中的个体样本的对应结果进行比较，并将合并样本的测量结果分配给个体样本的相应结果 ；将样本信息输入到存储器或处理器中的输入 / 输出界面 ；和连接于所述的处理器的显示器。处理器用于从计算机接收程序指令，其实施上文中所述的分析装置上的本发明的分析样本的方法。本文中所使用的术语“处理器”指包括其中使用一个或多个执行单元来执行从存储器或其他的存储装置检索到的指令的任何计算设备，例如接收合并样本并且通过确定样本或合并样本中分析物的信号进行所述分析物的测量的信号读取单元。

[0090] 本发明的分析装置还可包括本发明的合并装置。

[0091] 本发明还提供了在其自身或载体上的计算机程序产品，当该程序产品在计算机、编程计算机网络或其他可编程设备上加载并且执行时，可实施上文中所述的合并样本的方法。基本上，该计算机程序产品可被存储在本发明的合并装置的存储器中，并且所述处理器可通过为所述装置的处理器提供一系列对应于合并方法各处理步骤的指令来执行该程序。

[0092] 本发明还提供了在其自身或载体上的计算机程序产品，当该程序产品在计算机、编程计算机网络或其他可编程设备上加载并且执行时，可实施对多个样本进行分析的方法，所述方法包含对通过上文中所述的合并样本的方法获得的一系列合并样本进行分析，其中分析所述样本的分类变数，并且涉及对所述样本中的分析物实施定量测量。基本上，该计算机程序产品可被存储在本发明的合并装置的存储器中，并且所述处理器可通过为所述装置的处理器提供一系列对应于分析方法各处理步骤的指令来执行该程序。在用于进行分析的计算机程序产品中，嵌入到软件指令中的方法还可以包含如上文所述的合并样本的步

骤。

[0093] 现在通过以下的非限制性实施例来说明本发明。

[0094] 实施例

[0095] 实施例 1

[0096] 使用标准化的一个 50 个个体的池对于存在 SNP 的二倍体个体样本进行基因分型的实施例

[0097] 步骤 1) 单独检测 50 个个体

[0098] 对于每一 SNP 和每一个体，我们以微阵列形式使用两种不同的荧光染料，获得了红色荧光（存在等位基因）和绿色荧光（不存在等位基因）的强度。红色和绿色强度之间的比对于是纯合体动物不一定总为 1（或 0）或对于杂合体动物不一定总为 0.5。

[0099] 个体分型的数据被用于计算所有被分型的 SNP 的信号强度的校正系数。

[0100] 为了获得最重要的校正因子 (K)，通常使用校正因子来校正代表等位基因中的任意的不等效率的数据，我们使用来自杂合基因型的信号。如果不存在杂合基因型，我们就假定被研究的 SNP 在所研究的群体中是不分离的，因此应该忽略池中该 SNP 的结果。

[0101] 由于 50 个个体样本中不存在杂合型而忽略 SNP 会造成具有低 MAF（少数等位基因频率）的 SNP 上的信息丢失。对于许多应用（如基因组广泛选择）来说这是没有影响的，因为具有非常低的少数等位基因频率的 SNP 不会对精确性造成非常大的影响，因此可决定不使用这些 SNP 上的数据或者不采用校正系数。

[0102] 我们使用的第一校正因子 (K) 为；

[0103] K ＝ avg(Xraw/Yraw)

[0104] 其中 Xraw 是所测量的红色强度，Yraw 是所测量的绿色强度。该值是由具有基因型 AB 的个体基因分型的样本确定的。

[0105] 不使用一个基因型的所有微珠的平均结果，我们也可使用所有单独的微珠的结果。因此，我们使用来自一个样本的 Xraw 和 Yraw 或 X 和 Y 的平均结果，或者我们使用该样本的所有单独微珠的结果。

[0106] 其他的校正因子是 AAavg 和 BBavg。AAavg 是 AA 基因型的未校正等位基因频率的平均值。预期该值接近 1。BBavg 是 BB 基因型的未校正等位基因频率的平均值。预期该值接近 0。使用以下的公式来计算 AAavg 和 BBavg：

[0107] AAavg ＝ (avg(Xraw/(Xraw+Yraw)))

[0108] 和

[0109] BBavg ＝ (avg(Xraw/(Xraw+Yraw)))

[0110] 步骤 2) 构建一个测试池，其包括以上步骤 1) 的所有 50 个个体。为此，使用 NanoDrop 分光光度计 (NanoDrop Technologies,USA) 测量每个个体样本中的 DNA 浓度 (ng/μl)。然后，在合并成单一样本之前，将所有的 DNA 样本稀释到标准浓度 50ng/μl。在这样获得的测试池中，我们估算了未校正的或根据第一步骤得到的校正因子的等位基因频率。

[0111] 等位基因 A 的未校正等位基因频率被计算为红色强度除以两种强度之和的比值，如下：

[0112] 未校正等位基因频率＝ Xraw/(Xraw+Yraw)

[0113] 我们采用的等位基因频率的第一校正值为

[0114] 校正了的等位基因频率＝ Xraw/(Xraw+K ＊ Yraw)

[0115] 我们采用的第二校正值为归一化。

[0116] 归一化的等位基因频率＝（校正的等位基因频率 -BBavg)/AAavg

[0117] 对于校正和归一化，我们都使用了分别来自个体样本的每个 SNP 的所有 3 个基因型。

[0118] 估计的等位基因频率的准确性的顺序为 :归一化的（最准确）、校正的（两者之间）和未校正的（准确性最低）

[0119] 这表示，如果在步骤 1 中没有杂合个体，就将校正因子 K 设定为 0.5，如果没有纯合个体，就将校正因子 AAavg 和 BBavg 分别设定为 1 和 0。

[0120] 步骤 3) 我们比较了根据个体分型计算的和基于测试池中的结果的等位基因频率。由此我们估算了四次多项式，其中实际结果在 X 轴上。从图 1 可见单独测试的个体中和具有将近 18000 个 SNP 的池中的基因分型结果。使用 SNP 均匀分布在整个鸡基因组中 (vanAs 等，2007) 的 18K Chicken SNP iSelect Infinium 检测 (Illumina Inc，USA) 来进行基因分型。可在 Illumina 的网站 (http://www.illumina.com/pages.ilmn？ ID ＝ 12) 找到检测、操作流程和芯片的详细信息。

[0121] 当个体的已知频率为 0、0.05、0.1、0.15------0.9、0.95 和 1 时，我们通过这个多项式计算了测试池中的预计的等位基因频率。

[0122] 参见图 2，将第二幅图中的这些结果与 Y 轴上的实际频率放在一起，我们得到了第三校正步骤的校正因子。

[0123] 参见图 3，在应用这些校正因子之后，测试池中的等位基因频率表现出与实际频率的线性关系。

[0124] 在这个约 18,000 个 SNP 的实验中，与个体分型的结果相比，50 个个体的测试池中测量的超过 96％的等位基因频率（并且如上被校正）在 +6.25％或 -6.25％的范围内。

[0125] 对于本发明的应用，前面 3 个步骤优选在作为“校准”的实际分析之前进行，从而提高分析的精确性。但并不是每次都需要进行这些步骤。然后，（如果进行的话）测量校准随后实施以下步骤 :

[0126] 步骤 4) 以 1 ： 3、1 ： 3 ： 9 或 1 ： 3^1 ： 3^2 ： $3^{(n-1)}$ 的比例构建 2、3 或 n 个个体的DNA池，并且测量该池从而进行基因分型，其中使用18K Chicken SNP iSelect Infinium 检测（见上文）在芯片上确定红色和绿色的信号强度。

[0127] 步骤 5) 通过步骤 1 和步骤 3 得到的校正因子，可从得到的池中的信号强度来计算等位基因频率。

[0128] 对于具有两个个体的池，预计的校正后的频率得到了 0％、12.5％、25.0％、37.5％、50.0％、62.5％、75.0％、87.5％和 100％的结果点。四舍五入至最接近的结果点。两个个体的基因型可从表 2 中显示的结果得到。

[0129] 对于具有 3 个个体的池，四舍五入至最接近的结果点，其中结果点之间的区间是 3.85％ (100/(3^3-1)) 等。

[0130] 连续的结果点之间的间隔越小，读取强度需要的精确度就越高，从而将特定的结果合理地分配于结果点之一。随着基因分型技术的进一步发展，更精确的读取会变得可行。

[0131] 对于一个池中具有 2 个个体的情况，可以决定只使用 SNP，其中池中的估计的和校

正的等位基因频率落在个体的实际频率的 ±6.25％的范围内（见图 3 中的红线）。

[0132] 表 2. 合并样本的等位基因频率的结果点和对于具有 A 和 C 等位基因的 SNP 在池中的两个个体的推导基因型

[0133]

合并样本中等位基因 A 的频率	个体 1 的推导基因型（在池中占 1 份）	个体 2 的推导基因型（在池中占 3 份）
0	CC	CC
12.5	AC	CC
25	AA	CC
37.5	CC	AC
50	AC	AC
62.5	AA	AC
75	CC	AA
87.5	AC	AA
100	AA	AA

[0134] 如果没有其他的信息来推导个体基因型，就应该省略合并结果和个体结果之间显示出差异大于 6.25％的 SNP（步骤 3）。

[0135] 推导个体基因型的其他信息可以来自于个体的谱系或该个体隶属的家族（或科，family）或群体（或种群，population）中的单倍体型的信息。

[0136] 根据校准因子的可重复性，已知检测条件相同的新的分析可完全跳过步骤 1、2 和 3。

[0137] 当遵循实例 1 的方法时，通过减少需要被分析的样本的总数，取得显著的节约，同时仍可得到原始个体样本的可靠结果。被分析样本的一般减少的总数在表 3 中示例性地示出。

[0138] 表 3. 当根据本发明的方法合并 2 或 3 个个体时，节约的被分析的样本数目

[0139]

基因分型的个体数目	当合并两个个体时的样本数目				当合并三个个体时的样本数目			
	个体数目+池	2个个体的池的数目	样本总数	被分析样本数目的减少(%)	个体数目+池	3个个体的池的数目	样本总数	被分析样本数目的减少(%)
250	50+1	100	151	39.6	50+1	67	118	52.8
500	50+1	225	276	44.8	50+1	150	201	59.8
1000	50+1	475	526	47.4	50+1	317	368	63.2
2000	50+1	975	1026	48.7	50+1	650	701	64.9
5000	50+1	2475	2526	49.5	50+1	1650	1701	66.0

[0140] 实施例 2

[0141] 使用标准化的 2 个个体的 25 个池的基因分型二倍体个体样本的实施例

[0142] 步骤 1) 如实施例 1 的步骤 1 单独测试 50 个个体。

[0143] 步骤 2) 以 1 ： 3 比例构建 25 个池，每个池中有 2 个样本，其包括以上的步骤 1) 的所有 50 个个体。在这些池中，估计未校正的或基于第一个步骤中得到的校正因子的等位基因频率。

[0144] 步骤 3) 将 2 个个体分型的等位基因频率的总和与具有 2 个个体样本的池中的估计频率进行比较。从这 25 个点计算回归线。然后可将回归系数和截距用于校正其他池的估计频率。

[0145] 步骤 4) 然后以 1 ： 3、1 ： 3 ： 9 或 1 ： 3^1 ： 3^2 ： $3^{(n-1)}$ 的比例构建 2、3 或 n 个样本的 DNA 池。

[0146] 步骤 5) 以步骤 1 和步骤 3 中得到的校正因子，计算池中得到的信号强度的等位基因频率。

[0147] 减少的样本数目与用于二倍体个体测序的表 8 中提到的减少数目一致。

[0148] 实施例 3

[0149] 对单倍体个体样本进行基因分型的实施例

[0150] 当合并两个单倍体样本并且测量等位基因 A 在基因组的某些位置的存在时，预期测量（峰高、表面积、强度）中的比例为；

[0151] 表 4. 合并样本的等位基因频率的结果点和具有 A 和 C 等位基因的 SNP 的池中的两个个体的推导的基因型

[0152]

合并样本的等位基因 A 的频率	个体 1 的推导基因型（在池中占 1 份）	个体 2 的推导基因型（在池中占 3 份）
0.00	C	C
0.33	A	C
0.67	C	A
1.00	A	A

[0153] 如果只使用两个样本的池，可以不需要校正因子。当合并更多的样本时，可能需要校正因子。则其可以通过具有相等量的分析物的模拟杂合和纯合二倍体个体的 2 个样本的池来计算。

[0154] 当以 1 ： 2 ： 4 的比例合并 3 个样本时，预期测量中的比例如下；

[0155] 表 5. 合并样本的等位基因频率的结果点和具有 A 和 C 等位基因的 SNP 的池中的三个个体的推导的基因型

[0156]

合并样本的等位基因 A 的频率	个体 1 的推导基因型（在池中占 1 份）	个体 2 的推导基因型（在池中占 2 份）	个体 2 的推导基因型（在池中占 4 份）
0.000	C	C	C
0.166	A	C	C
0.333	C	A	C
0.500	C	C	A
0.666	A	C	A
0.833	C	A	A
1.000	A	A	A

[0157] 实施例 4

[0158] 本发明在测序试验方案中的应用

[0159] 本发明中描述的合并方法可被应用于需要确定 2 个或多个个体的序列的情况。

[0160] 对合并个体、模板或 PCR 产物进行测序不是惯常操作，因为当分析双峰图 (double trace，双迹）时的重要问题是，在每一个位置都存在两个碱基，通过仅示例峰图 (the trace，踪迹）来辨别每一个碱基是来自哪个模板是不可能的。

[0161] 除了慎重地合并产生双峰图的模板之外，已知几种生物或生物技术情况会产生双峰图。在通过 RT-PCR 扩增的转录产物的选择性剪接区域、直接测序（未克隆）和随机插入突变实验中可观察到这些情况。

[0162] 已描述了几种追溯合并序列或双峰图的单倍体型的方法。Flot 等 (2006) 描述了提出用于找出个体的单倍体型的几种分子方法。例如，测序克隆的 PCR 产物（例如，Muir 等，2001）、SSCP（单链构象多态性）(Sunnucks 等，2000)、变性梯度凝胶电泳 (DGGE)(Knapp 2005)、极端稀释 DNA 至单个分子水平 (Ding & Cantor2003) 和等位基因特异性 PCR 引物的应用 (Pettersson 等，2003)。此外还提出了几种用于序列混合物的单倍体型重建的计算方法。

[0163] 但是，所述的所有方法都是非常昂贵和消耗时间的，只适用于特定目的（例如，重

新测序、选择性剪接、序列长度不同的两种产物的模板或 PCR 扩增混合物、参考基因组序列的可用性），而不是用于单倍体或二倍体样本的标准直接测序或完全未知序列的重新测序。

[0164] 遵循本发明描述的合并的序列模板的合并可应用于可在个体和合并样本中都获得相同的序列片段的情况。这表明，例如鸟枪测序（随机剪切片段）不适合用于合并。

[0165] 在以上所提到的所有应用中，如果基于一定目的应用合并，则合并等量的模板（样本、DAN、RNA 或 PCR 产物）。

[0166] 在本文中，我们描述了合并不等量的模板。对于该实施例，只描述了池由 2 个模板构成的情况，但是可以使用本发明以便对于二倍体生物以 1 ： 3、1 ： 3 ： 9、1 ： 3^1 ： 3^2 ： $3^{(n-1)}$ 的比例和对于单倍体生物以 1 ： 2、1 ： 2 ： 4、1 ： 2^1 ： 2^2 ： $2^{(n-1)}$ 的比例构建 2、3 或 n 个个体的 DNA（或 PCR 后产物）的池的情况。

[0167] 需要满足的一般条件是测序设备扫描模板（例如，荧光）并且得到的色谱图将 DNA 模板的序列表示为间隔规律、高度相似的一系列峰。

[0168] 步骤 1) 单独对 50 个个体进行测序反应

[0169] 个体测序反应的数据被用于从所有碱基（或核苷酸）位置的峰面积和峰高度计算校正因子。

[0170] 步骤 2) 对 2 个合并个体的 25 个池进行测序反应

[0171] 峰面积比被用于区别碱基和噪声峰处的第一峰和第二峰。第二峰是第一峰的一部分，并且阈值被用于区别峰和噪声峰。

[0172] 合并测序反应的数据被用于从所有碱基（核苷酸）位置处的峰面积和峰高度计算校正因子。

[0173] 步骤 3) 将步骤 1 和 2 的结果作图并建立回归线（计算回归系数和截距）。

[0174] 步骤 4) 构建 DNA（或 PCR 后产物）的池

[0175] 对于二倍体生物以 1 ： 3、1 ： 3 ： 9、1 ： 3^1 ： 3^2 ： $3^{(n-1)}$ 的比例和对于单倍体生物以 1 ： 2、1 ： 2 ： 4、1 ： 2^1 ： 2^2 ： $2^{(n-1)}$ 的比例构建 2、3 或 n 个个体的 DNA 的池。

[0176] 步骤 5) 以步骤 1、步骤 2 和步骤 3 得到的校正因子，能够由池中得到的信号强度来计算碱基判定 (basecalling)。

[0177] 在该实施例中，显示了在每个碱基位置只有 2 个可能的核苷酸 (A 和 C)，但是相同的原则可应用于作为遗传密码基础的 4 个可用的核苷酸中的 2 个的其他组合。“A”核苷酸的平均峰高度被设定为 100，而“C”核苷酸的平均峰高度被设定为 75。基于这些峰高度，表 6 中列出了对于两个单倍体样本池中的核苷酸的每一种可能的组合的相对峰高度。表 7 中提供了由两个二倍体模板构成的池的相对峰高度。

[0178] 表 6. 合并的和未合并的单倍体个体的等位基因频率的结果点和核苷酸序列中的随机位置的推导基因型

[0179]

推导的基因型		未合并的峰面积/高度		合并（比例 1:2）的峰面积/高度	
个体 1	个体 2	第一峰(A)	第二峰(C)	第一峰(A)	第二峰(C)
A		100			
C			75		
A	A			100	
A	C			33.3	50
C	A			66.6	25
C	C				100

[0180] 表 7. 合并的和未合并的二倍体个体的等位基因频率的结果点和核苷酸序列中的随机位置的推导基因型

[0181]

推导的基因型		未合并的峰面积/高度		合并（比例 1:3）的峰面积/高度	
个体 1	个体 2	第一峰(A)	第二峰(C)	第一峰(A)	第二峰(C)
AA		100			
AC		50	37.5		
CC			75		
AA	AA			100	0
AA	AC			62.5	28.125
AA	CC			25	56.25
AC	AA			87.5	9.375
AC	AC			50	37.5
AC	CC			12.5	65.625
CC	AA			75	18.75
CC	AC			37.5	46.875
CC	CC			0	100

[0182] 比较了本发明的合并方法与未合并情况，表 8 示出了减少的测序反应的数目。

[0183] 表 8. 当遵循本发明的方法合并 2 个个体时，减少的样本或测序反应的数目

[0184]

被测序的个体数目	使用本发明被测序的池或样本的数目			减少的测序样本数目（%）
	个体+池	2 个个体的池	样本总数	
250	50+25	100	175	30%
500	50+25	225	300	40%
1000	50+25	475	550	45%
2000	50+25	975	1050	47,5%
5000	50+25	2475	2250	49%

[0185] 实施例 5

[0186] 使用可选的校正方法使用标准的 1 个 50 个个体的池和 25 个 2 个个体的池对二倍体个体样本进行基因分型的实施例。该实施例描述了几个实验。

[0187] 步骤 1) 单独测试 50 个个体。

[0188] 与实施例 1 的步骤 1 相同，但是校正方法不同：使用归一化的强度 X 和 Y，而不是 Xraw 和 Yraw。

[0189] 使用 X 和 Y 计算第一校正因子 (K)。

[0190] K = avg(X/Y)

[0191] 其中 X 是等位基因 A（红色）的归一化强度，Y 是等位基因 B（绿色）的归一化强度。由基因型为 AB 的个体基因分型的样本来确定该值。

[0192] 其他的校正系数 AAavg 和 BBavg 也是基于 X 和 Y。AAavg 是 AA 基因型的未校正的等位基因频率的平均值。预期该值接近 1。BBavg 是 BB 基因型的未校正的等位基因频率的平均值。预期该值接近 0。使用以下的公式计算 AAavg 和 BBavg：

[0193] AAavg = (avg(X/(X+Y)))

[0194] 和

[0195] BBavg = (avg(X/(X+Y)))

[0196] 也可根据实施例 1 的步骤 1 中的 Xraw 和 Yraw 计算所有的校正因子 K、AAavg 和 BBavg。

[0197] 如果 50 个个体中没有基因型 AA，则 AAavg 被设定为 1。同样，如果没有基因型 BB，则 BBavg 被设定为 0。

[0198] 接下来的步骤是根据其中所有的 50 个个体都有结果的那些 SNP 的个体分型来计算等位基因频率。

[0199] 步骤 2) 如实施例 1 中的步骤 2 来构建一个来自步骤 1 的所有 50 个个体的池。

[0200] 等位基因 A 的未校正等位基因频率被计算为归一化的红色强度 (X) 除以两个归一化的强度 (X+Y) 之和的比。

[0201] 未校正的等位基因频率 = X/(X+Y)（称为 Raf）

[0202] 我们应用的等位基因频率的第一校正为

[0203] 校正的等位基因频率 = X/(X+K * Y)（称为 Rafk）

[0204] 如果没有杂合基因型，可不计算 K。在这种情况下，可应用以下的法则：

[0205] 如果 Raf < 0.1，则 Rafk 被设定为 0。

[0206] 如果 Raf ＞ 0.9，则 Rafk 被设定为 1。

[0207] 在所有其他的缺省 K 的情况下，Rafk 被设定为等于 Raf。

[0208] 当以归一化的强度 X 和 Y 开始时，不总是需要使用 AAavg 和 BBavg 进行归一化校正。如果以 Xraw 和 Yraw 开始，可如实施例 1 的步骤 2 一样应用使用 AAavg 和 BBavg 的归一化。

[0209] 如果应用归一化，则使用以下的公式；

[0210] 归一化的等位基因频率＝（校正的等位基因频率 -BBavg)/AAavg（称为 Rafn)

[0211] 步骤 3) 我们比较了在步骤 1 中对个体分型计算的预期 (expected) 等位基因频率和根据步骤 2 中对 50 个的池中的结果的观察（校正或未校正的）频率。我们使用以下模型计算了回归系数；

[0212] 预期的等位基因频率＝ b1 ＊观察到的频率 +b2 ＊观察到的频率 ^{2}+b3 ＊观察到的频率 ^{3}+b4 ＊观察到的频率 ^{4}，无截距

[0213] 校正的频率 (Rafk 和 Rafn) 或未校正的频率 (Raf) 被用作以上公式中的观察到的频率。

[0214] 通过比较预期的与从该模型预计的等位基因频率，可得到最佳的校正方法 (Rafk、Rafn 或 Raf)。

[0215] 此后，最佳校正方法的回归系数会被用于校正步骤 5a 中的 2 个个体的池的等位基因频率。

[0216] 步骤 4) 由 50 个个体以 1 ： 3 的比例建立 25 个 2 个个体的 DNA 池。应注意，池中的哪个个体使用了一次，而哪个个体使用了 3 次。

[0217] 步骤 5a) 基于 50 个个体的池的结果的校正。

[0218] 以步骤 1(K、AAavg 和 BBavg) 和步骤 3（回归系数 b1、b2、b3 和 b4) 中的得到的校正因子，可从步骤 4 中构建的池中得到的信号强度来计算等位基因频率。首先使用步骤 1 的校正因子 K、AAavg 和 BBavg（根据步骤 3 得到的最佳校正方法）来计算 Raf 或 Rafk 或 Rafn。

[0219] 使用步骤 3 得到的多项回归系数来计算 Rafc 或 Rafkc 或 Rafnc 为

[0220] 预期的等位基因频率＝ b1 ＊观察到的频率 +b2 ＊观察到的频率 ^{2}+b3 ＊观察到的频率 ^{3}+b4 ＊观察到的频率 ^{4}，其中观察到的频率＝ Raf 或 Rafk 或 Rafn。

[0221] 以池中的两个个体，预测的校正的频率应该提供结果点 0％、12.5％、25.0％、37.5％、50.0％、62.5％、75.0％、87.5％和 100％。四舍五入至最接近的结果点。两个个体的基因型可来自于实施例 1 的表 2 中示出的结果。

[0222] 步骤 5b) 基于 2 个个体的池的结果的校正

[0223] 根据步骤 4 构建的池的信号强度和步骤 1 中得到的校正因子 K、AAavg 和 BBavg 来计算 Raf、Rafk 和 Rafn。

[0224] 使用与步骤 3 相同的模型的多项回归系数，可根据 20 个池来计算实施例 5。该模型可被分别应用于每一个 SNP 或应用于所有的 SNP。

[0225] 根据这些回归因子来预测另外 5 个池中的等位基因频率为：

[0226] Rafkc ＝ b1 ＊ Rafk+b2 ＊ Rafk^{2}+b3 ＊ Rafk^{3}+b4 ＊ Rafk^{4}（来自于 Rafk 的回归模型）

[0227] Rafn = b1 * Rafn+b2 * $Rafn^2$+b3 * $Rafn^3$+b4 * $Rafn^4$(来自于 Rafn 的回归模型)

[0228] Rafc = b1 * Raf+b2 * Raf^2+b3 * Raf^3+b4 * Raf^4(来自于 Raf 的回归模型)。

[0229] 这可以以所有的样本用于预测一次的方式重复 5 次。然后将这些池中的预期等位基因频率与预测的等位基因频率进行比较,从而发现最好的校正方法。

[0230] 以具有两个个体的池,预测的校正的频率应该提供结果点 0%、12.5%、25.0%、37.5%、50.0%、62.5%、75.0%、87.5%和 100%。四舍五入至最接近的结果点。两个个体的基因型可来自于实施例 1 的表 2 中示出的结果。

[0231] 步骤 5c) 基于 2 个个体的池的结果的校正。

[0232] 可通过使用基于以下模型的对于光强度的 SNP(X 或 Xraw 和 Y 和 Yraw) 的多元线性回归系数来进行另一种方式的预测。

[0233] 预期的等位基因频率= b1 * X+b2 * Y

[0234] 或

[0235] 预期的等位基因频率= b1 * Xraw+b2 * Yraw。

[0236] 可以使用以下的公式利用这些多元线性回归因子来预测等位基因频率:

[0237] 预测的等位基因频率=截距 +b1 * X+b2 * Y

[0238] 或

[0239] 预测的等位基因频率=截距 +b1 * Xraw+b2 * Yraw。

[0240] 如上所述,基于 20 个池来计算多元线性回归系数。

[0241] 然后根据这些回归系数来预测另外 5 个池的等位基因频率。这可以以所有的样本用于预测一次的方式重复 5 次。然后将这些池中的预期等位基因频率与预测的等位基因频率进行比较,从而发现最好的校正方法。

[0242] 例如在步骤 5a 和步骤 5b 中,两个个体的基因型可来自于实施例 1 的表 2 中示出的结果。

[0243] 步骤 6) 由其他的个体样本以 1 : 3 的比例建立 2 个个体的 DNA 池。如步骤 4 需注意,池中的哪个个体使用了一次,哪个个体使用了 3 次。

[0244] 我们能够利用如所述的预测等位基因频率的最佳校正方法并利用实施例 1 的表 2 从这些池得到基因型。

[0245] 实验 1

[0246] 使用 Infinium 检测珠芯片技术 (Illumina,Inc.USA) 将实施例 5 中描述的方法应用于全基因组 SNP 分析。

[0247] 使用18K Chicken SNP iSelect Infinium检测(Illumina Inc,USA)(其中SNP均匀分布在整个鸡基因组中)对50个个体进行基因分型(van As等,2007)。在Illumina的网站上可找到检测、操作流程和芯片的详细信息 (http://www.illumina.com/pages.ilmn?ID=12)。

[0248] 为了检查频率是否被精确估算,将 8 个等位基因组合成一个池(来自 50 个独立基因分型的个体中的 4 种不同动物)。除了不使用表 2 将预测的等位基因翻译成基因型之外,进行实施例 5 中的步骤 1 至步骤 3 和步骤 5。

[0249] 在步骤 4 中,合并 4 个个体的等摩尔量的 DNA,而不是以 1 : 3 的比例合并 2 个个

体的 DNA。

[0250] 如果使用的是来自 2 种不同动物的 1 ： 3 比例，我们就可以认为这是将 8 个等位基因组合在一个池中。通过使用等摩尔量的 4 个个体，也可以组合 8 个等位基因。

[0251] 这样，就组成了 12 个池，以及如步骤 1 中的一个 50 个动物的池（在 4 个池中使用相同的样本加上 2 个额外的样本）。然后使用第二批 infinium 芯片对这 13 个池进行基因分型。

[0252] 如实施例 5 的步骤 1 来计算每一 SNP 的 K、AAavg 和 BBavg。

[0253] 然后如实施例 5 的步骤 2，计算 50 个动物的池的未校正和校正的等位基因频率。

[0254] 还如实施例 5 的步骤 3 计算多项回归系数。

[0255] 此外，如步骤 5b 和 5c 描述的，计算多项回归系数和多元线性回归系数。这是基于 11 个池进行的，然后使用回归因子来预测其余的池中的等位基因频率。

[0256] 在该实验中，对 X 和 Y（红色和绿色强度）的多元线性回归产生了最佳结果。最终结果参见图 4 和表 9。

[0257] 总共 4.6％的等位基因频率落入错误类别 (wrong class) 内。

[0258] 在以 1 ： 3 的比例合并 2 个个体的池的情况下，会产生 3.0％的基因分型误差。

[0259] 表 9. 与预期等位基因频率按类别进行比较的预测等位基因频率的数目。对角线上的数字会产生正确的基因型。对角线外但是在框中的等位基因频率会产生一个基因型误差。其他的结果会产生 2 个基因型误差。

[0260]

预期的等位基因频率 \ 预测的	0	12.5	25	37.5	50	62.5	75	87.5	100	总数
0	59489	144	13			2		1		59649
12.5	331	12888	452	11	3	1	1			13687
25	27	427	12060	897	10	1				13422
37.5	2		374	11342	1026	17	1			12762
50			4	671	11590	1098	27			13390
62.5	1			5	682	11074	727		1	12490
75			1		3	779	11421	494	29	12727
87.5			1		1	3	528	11172	416	12121
100	10			3	1	6	5	50	50896	50971

[0261] 实验 2

[0262] 使用 Veracode 检测技术 (Illumina, Inc.USA) 将实施例 5 中描述的方法应用于 SNP 分析。

[0263] 使用 96Chicken SNP Veracode,Golden Gate 检测 (Illumina Inc,USA)（其中 SNP 均匀分布于整个鸡基因组中）（步骤 1）来对 50 个个体进行基因分型。在 Illumina 的网站上可找到检测、操作流程和芯片的详细信息 (http://www.illumina.com/pages.ilmn？ ID＝6)。

[0264] 还构建了一个所有样本的池（如步骤 2）和比例为 1 ： 3 的 24 个具有 2 个个体的池（如步骤 4）。以第二批的化学物质来基因分型这 25 个池。

[0265] 如实施例 5 的步骤 1 至步骤 3 描述的进行所有的校正。

[0266] 使用在步骤 3 中得到的多项回归因子，将步骤 5a 的校正应用于所有 24 个 2 个个

体的池。

[0267] 对于步骤 5b 和 5c，我们每次使用 23 个池来计算回归因子（步骤 5b 中的多项回归因子和步骤 5c 中的多元线性回归因子）从而能够预测其余池的等位基因频率。我们总共进行了 24 次，所有的池都被使用了一次从而来预测等位基因频率。

[0268] 使用 Rafk（根据归一化值 X 和 Y 计算）得到了最佳结果，然后使用得到 Rafkc 的步骤 5b 的多项回归因子进行校正。

[0269] 在个体中总计召集 (call) 了 84 个 SNP。而在某些个体上未召集某些 SNP。我们总计有 1906 个完整的池 *SNP 组合。

[0270] 表 10. 通过与预期等位基因频率按类别进行比较的预测等位基因频率的数目。对角线上的数字会产生正确的基因型。对角线外但是在框中的等位基因频率会产生一个基因型误差。其他的结果会产生 2 个基因型误差。

[0271]

预期基因型		预测的 CC	AC	AA	CC	AC	AA	CC	AC	AA	总数
		CC	CC	C	C	C	AC	AA	AA	AA	
CC	CC	312	9								321
AC	CC	4	156	4	2						166
AA	CC		13	39	7	3					62
CC	AC			10	129	7	1				147
AC	AC				9	228	12	1			250
AA	AC					24	144	5			173
CC	AA						4	49	9		62
AC	AA							7	135	1	143
AA	AA							1	5	576	582
											190
总数		316	176	54	147	265	159	64	148	577	6

[0272] 总共有 138(138/1906*100 = 7.2%）个错配（表 10)。因为每次观察由 2 个个体样本构成，这产生了 174 个基因型误差 (170/1906*2*100 = 4.46%)，参见见表 11、图 5 和图 6。

[0273] 确定该实施例中的最佳校正方法（如步骤 3（实施例 5) 和步骤 5a、5b 或 5c（实施例 5) 所进行的）的过程也提供了由于 SNP 错配数目的信息。这使得可以从系列中去除 SNP 从而以降低检出率为代价降低了错误的风险。

[0274] 表 11. 校正的预计基因型的数目

[0275]

预期的		预测的 CC CC	AC CC	AA CC	CC AC	AC AC	AA AC	CC AA	AC AA	AA AA	总数
CC	CC	624	9								633
AC	CC	4	312	4	0						320
AA	CC		13	78	0	0					91
CC	AC			0	258	7	1				266
AC	AC				8	456	12	0			477
AA	AC					24	288	0			312
CC	AA						0	98	9		107
AC	AA							7	270	1	278
AA	AA							1	5	1152	1158
总数		628	331	83	266	491	297	107	282	1153	3642

[0276] 实验 3

[0277] 使用其他基因型分型方法将实施例 5 的方法应用于 SNP 分析。

[0278] 实施例 5 中描述的方法也可以用于任何其他基因分型方法中，除了实验 1 和实验 2 中描述的方法，也可以使用如 Affymetrix 基因芯片 (Affymetrix Inc，USA) 或 Agilent Technologies。

[0279] 实施例 6

[0280] 本发明应用于如实施例 4 的测序方案，但使用其他的校正方法

[0281] 步骤 1) 单独对 50 个个体进行序列反应。

[0282] 使用等位基因 1 的峰高度和等位基因 2 的峰高度作为 Xraw 和 Yraw 值，或使用相对峰高度作为 X 和 Y。

[0283] 等位基因 1 的相对峰高度为 X ＝ X/(X+Y)，等位基因 2 的相对峰高度为 Y ＝ Y(X+Y)。

[0284] 然后以与实施例 5 的步骤 1 的基因分型相同的方法来计算 K、AAavg 和 Bbavg。

[0285] 步骤 2) 在一个所有 50 个个体的池中进行序列反应。

[0286] 如实施例 5 的步骤 2 来计算未校正和校正的等位基因频率。

[0287] 步骤 3) 从个体测序以及从池来计算频率。

[0288] 使用与实施例 5 的步骤 3 相同的模型得到多项回归系数。

[0289] 步骤 4) 进行 25 个具有 2 个合并个体的池的序列反应。

[0290] 步骤 5a) 将经校正的频率与基于所有 50 个个体的池的预期频率进行比较，从而得到最佳方法。

[0291] 步骤 5b) 使用利用以下模型的其他 20 个池中得到的多项回归因子来计算具有 2 个个体的 5 个池中的 Rafnc、Rafkc 和 Rafc。

[0292] 预期的等位基因频率＝ b1 ＊观察到的频率 +b2 ＊观察到的频率 ^{2}+b3 ＊观察到的频率 ^{3}+b4 ＊观察到的频率 ^{4}，无截距

[0293] 步骤 5c) 使用利用以下模型的其他 20 个池中得到的多元线性回归系数来计算具有 2 个个体的 5 个池的预测等位基因频率。

[0294] 预测的等位基因频率＝截距 +b1 ＊ X+b2 ＊ Y

[0295] 或

[0296] 预测的等位基因频率＝截距 +b1 ＊ Xraw+b2 ＊ Yraw

[0297] 通过以所有的池被用于预测等位基因频率（确认）的方式通过重复步骤 5b 和 5c 几次由步骤 3 和步骤 5 来确定最佳校正方法。

[0298] 如果需要的话，可使用用于确认的其他数目。例如，能够使用 24 个池来获得回归因子，然后使用这些因子来进行预测。

[0299] 总共需要重复 25 次。

[0300] 通过最佳的校正方法和所需的校正因子和回归因子，可以预测新的池的频率并读取表 2 中得到的等位基因。

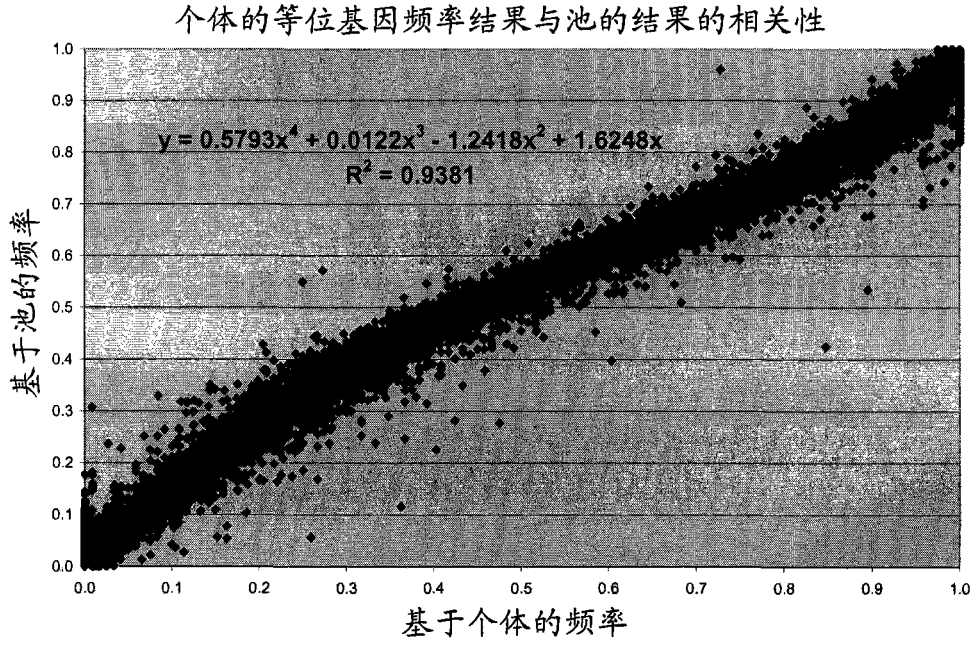

个体的等位基因频率结果与池的结果的相关性
y = 0.5793x4 + 0.0122x3 - 1.2418x2 + 1.6248x
R2 = 0.9381
基于池的频率
基于个体的频率
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

图 1

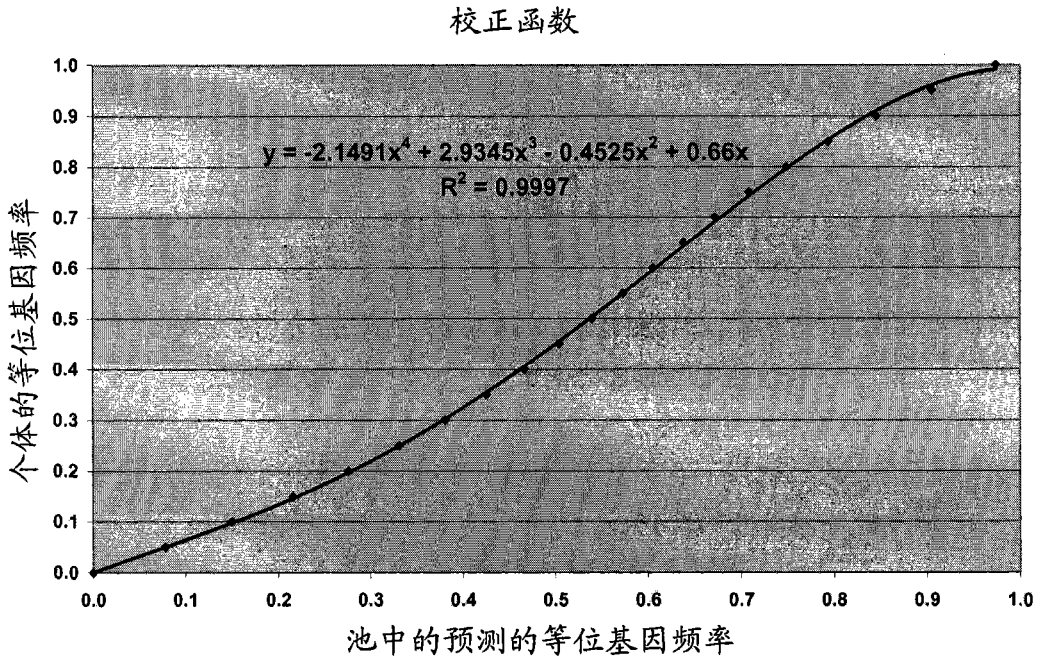

校正函数
y = -2.1491x4 + 2.9345x3 - 0.4525x2 + 0.66x
R2 = 0.9997
个体的等位基因频率
池中的预测的等位基因频率
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

图 2

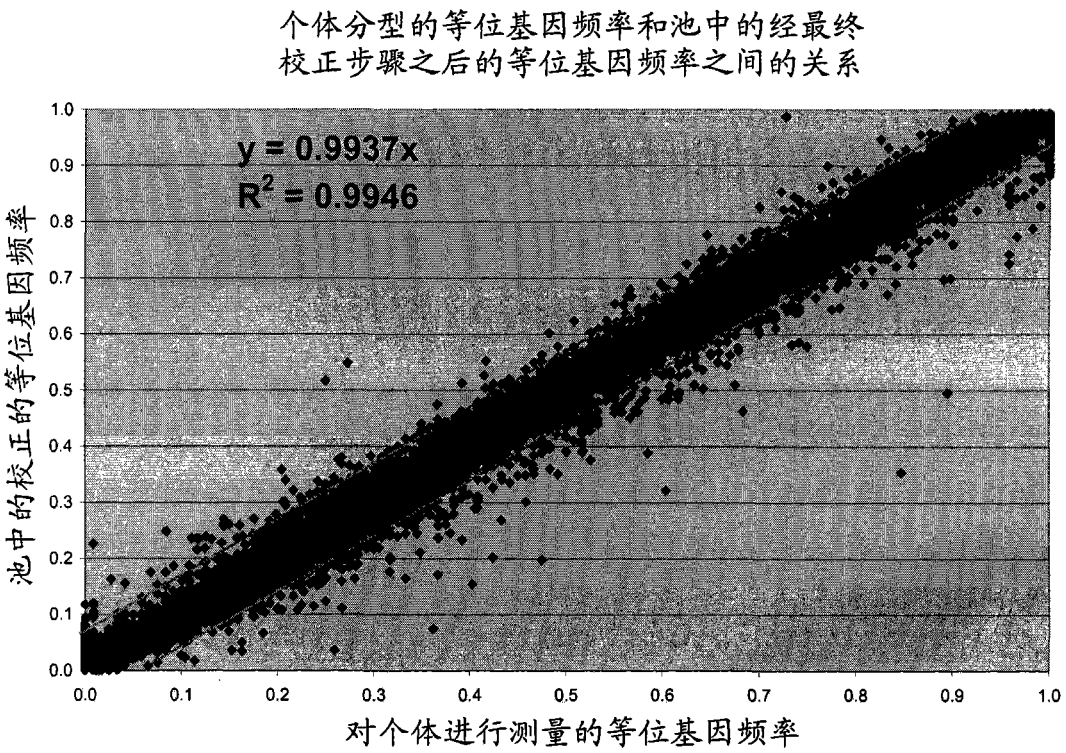

个体分型的等位基因频率和池中的经最终
校正步骤之后的等位基因频率之间的关系
y = 0.9937x
R2 = 0.9946
池中的校正的等位基因频率
对个体进行测量的等位基因频率
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

图 3

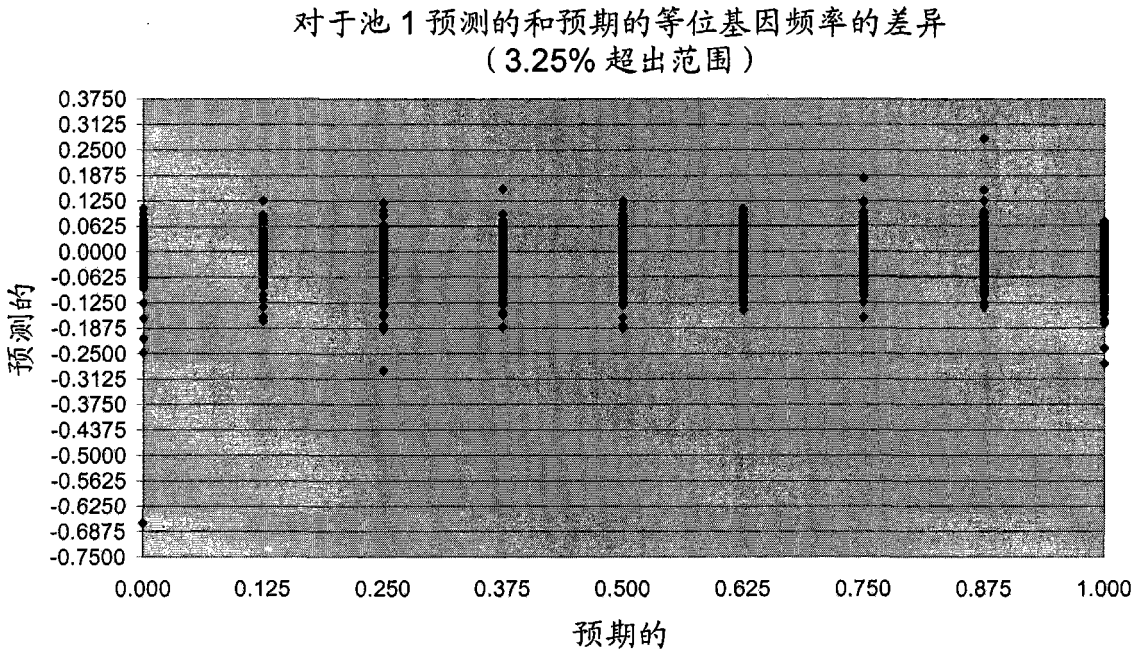

对于池 1 预测的和预期的等位基因频率的差异
（3.25% 超出范围）
预测的
预期的
0.3750
0.3125
0.2500
0.1875
0.1250
0.0625
0.0000
-0.0625
-0.1250
-0.1875
-0.2500
-0.3125
-0.3750
-0.4375
-0.5000
-0.5625
-0.6250
-0.6875
-0.7500
0.000
0.125
0.250
0.375
0.500
0.625
0.750
0.875
1.000

图 4

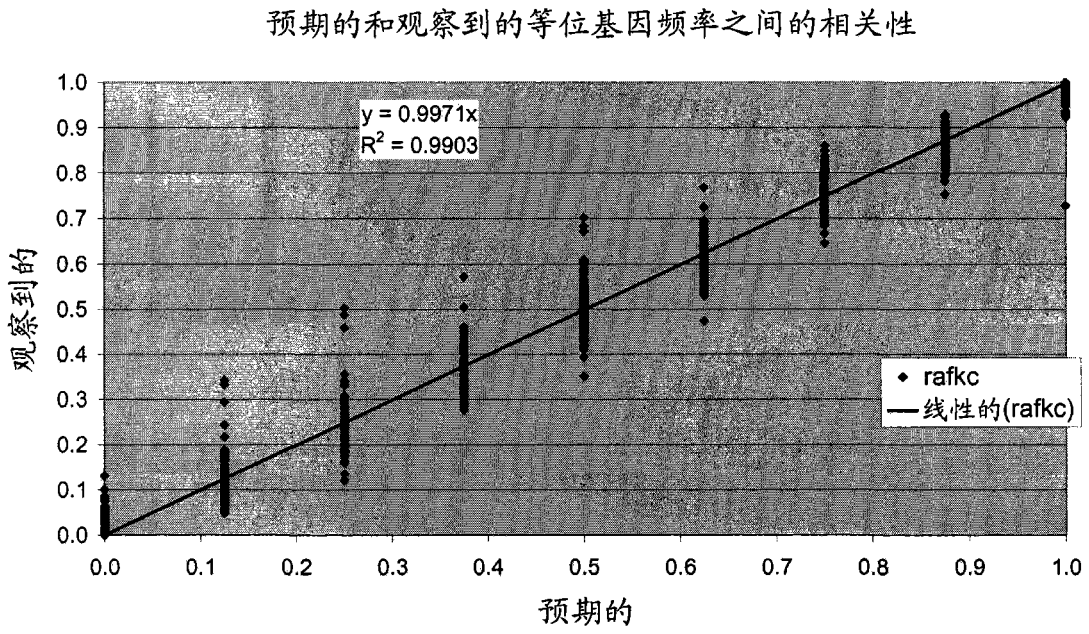

预期的和观察到的等位基因频率之间的相关性
y = 0.9971x
R2 = 0.9903
rafkc
线性的(rafkc)
观察到的
预期的
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

图 5

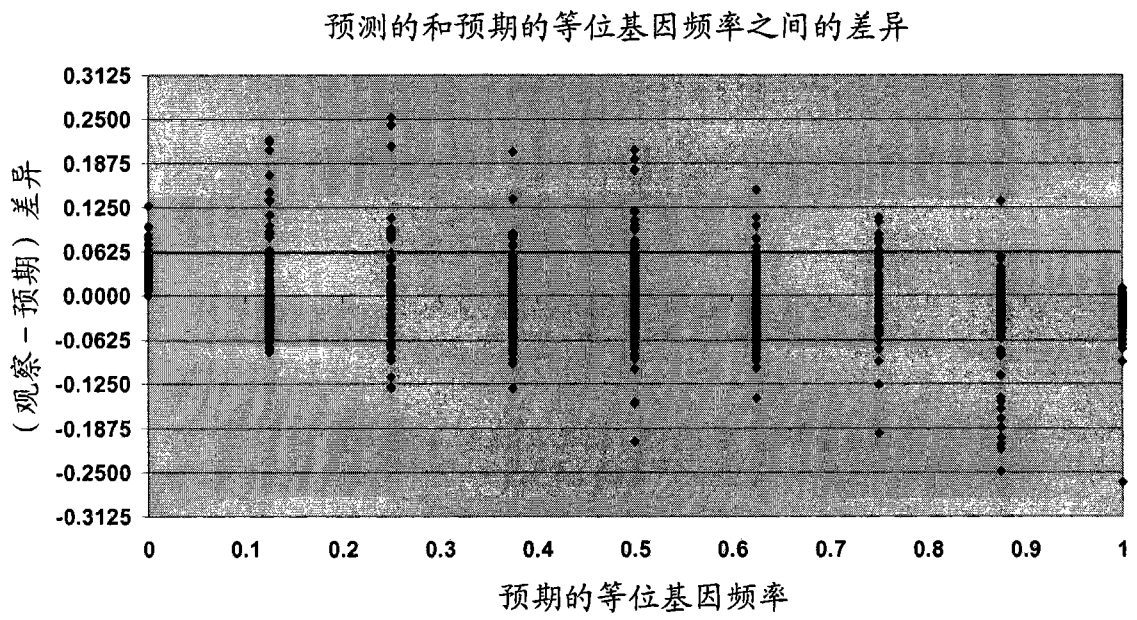

预测的和预期的等位基因频率之间的差异
（观察－预期）差异
预期的等位基因频率
0.3125
0.2500
0.1875
0.1250
0.0625
0.0000
-0.0625
-0.1250
-0.1875
-0.2500
-0.3125
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

图 6

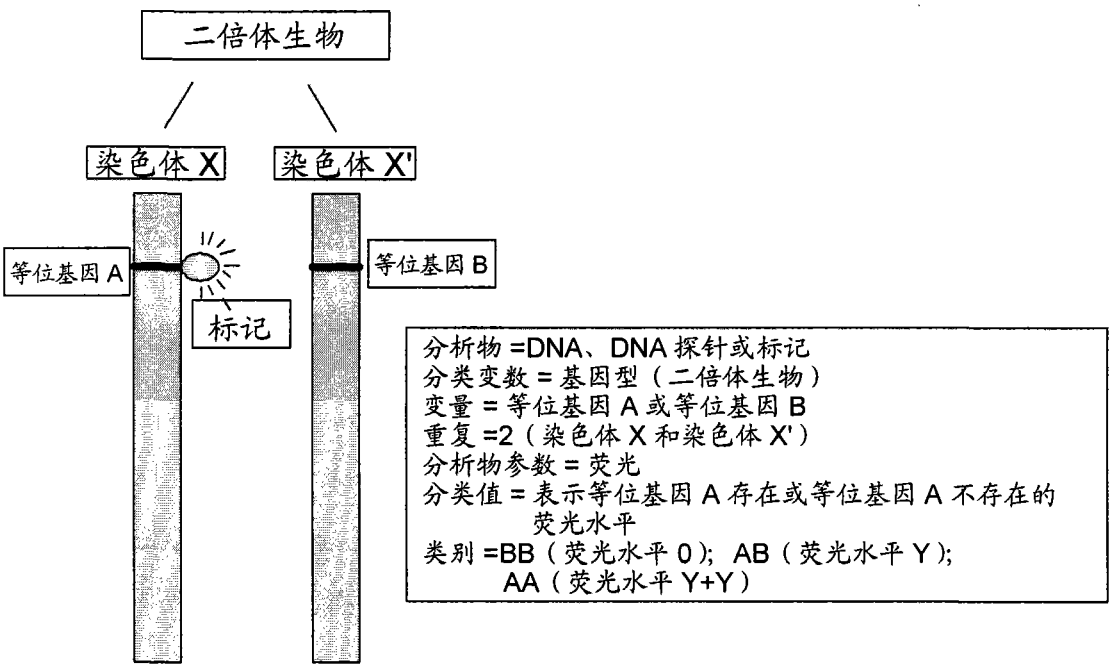

二倍体生物
染色体 X
染色体 X'
等位基因 A
标记
等位基因 B
分析物 =DNA、DNA 探针或标记
分类变数 = 基因型（二倍体生物）
变量 = 等位基因 A 或等位基因 B
重复 =2（染色体 X 和染色体 X'）
分析物参数 = 荧光
分类值 = 表示等位基因 A 存在或等位基因 A 不存在的荧光水平
类别 =BB（荧光水平 0）；AB（荧光水平 Y）；AA（荧光水平 Y+Y）

图 7