(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 116368237 A
(43)申请公布日 2023.06.30

(21)申请号 202180072041.X
(22)申请日 2021.10.21
(30)优先权数据
63/094,750 2020.10.21 US
(85)PCT国际申请进入国家阶段日
2023.04.21
(86)PCT国际申请的申请数据
PCT/US2021/056094 2021.10.21
(87)PCT国际申请的公布数据
WO2022/087299 EN 2022.04.28
(71)申请人 ABS全球股份有限公司
地址 美国威斯康辛州
(72)发明人 詹妮弗·杰利尔 马修·坎贝尔
史蒂夫·朗斯利
帕特里斯·利内尔
(74)专利代理机构 上海翼胜专利商标事务所(普通合伙) 31218
专利代理师 翟羽
(51)Int.Cl.
C12Q 1/6806(2006.01)

权利要求书5页 说明书21页 附图4页

(54)发明名称

用于处理遗传样本以确定特性或检测污染的方法及系统

(57)摘要

用于处理来自麦管的精液样本,以确认遗传特性、测试样本的纯度、检测错误或污染、计算污染量,以及确定污染的特性的方法及系统。本文的方法可检测低水平的污染,例如约1%、2%等的污染。

CN 116368237 A

1.一种处理来自一测试样本的经提取的DNA的方法,其特征在于:所述方法包括步骤:

a.使用SNP引物对池,对来自所述测试样本的经提取的DNA进行核苷酸扩增,各个SNP引物对侧接一独特的基因座,所述独特的基因座包含单一目标SNP,所述单一目标SNP定义一第一等位基因及一第二等位基因,所述核苷酸扩增针对各个SNP引物对产生多个扩增子,生成多个SNP扩增子池;

b.对所述SNP扩增子池进行测序,以针对在所述池中各个扩增子生成一核苷酸序列;

c.针对各个SNP,计算所述第一等位基因及所述第二等位基因的频率;

d.将在所述目标SNP的子集中的各个SNP的第一等位基因及第二等位基因的频率与一参考序列进行比较,其中所述目标SNP的子集是预期为纯合的一组SNP;

其中倘若在所述目标SNP的子集中的各个SNP的第一等位基因及第二等位基因的频率与在所述参考序列中对应的SNP完全匹配,则所述测试样本与所述参考序列相同;

其中倘若在所述SNP的子集中的各个SNP的第一等位基因及第二等位基因的频率与在所述参考序列中对应的SNP不完全匹配,则计算在所述SNP的子集中的各个SNP的非匹配的等位基因的频率,其中倘若针对一特定的SNP的非匹配的等位基因的频率高于一预定的非匹配的阈值,则所述特定的SNP是一污染的SNP,其中倘若所述污染的SNP的数量高于一预定的污染的SNP的阈值,则所述样本被鉴定为被污染。

2.根据权利要求1所述的方法,其特征在于:所述测试样本是一精子样本。

3.根据权利要求2所述的方法,其特征在于:所述精子样本已经受用于确定动物性别的机器,及所述精子样本包括活精子及死精子。

4.根据权利要求1所述的方法,其特征在于:所述方法是多重分析,其中至少2个样本同时经受所述方法。

5.根据权利要求1所述的方法,其特征在于:所述引物池包括至少48个引物组。

6.根据权利要求1所述的方法,其特征在于:核苷酸扩增是PCR扩增,及

测序是下一代测序(NGS)。

7.根据权利要求1所述的方法,其特征在于:所述核苷酸扩增的步骤包括:使用引物对进行PCR扩增,及一随后的PCR步骤,以添加用于下一代测序的接头序列及条形码。

8.根据权利要求1至7中任一项所述的方法,其特征在于:所述方法进一步包括步骤:使用一分析系统以计算所述各个SNP的第一等位基因及第二等位基因的频率,以及将所述各个SNP的第一等位基因及第二等位基因的频率与在一序列库中的参考序列或参考序列组的频率进行比较。

9.根据权利要求1所述的方法,其特征在于:倘若在所述SNP的子集中,所述各个SNP的第一等位基因及第二等位基因的频率与在所述参考序列中相同的SNP的频率完全匹配,则所述测试样本的纯度为至少98%。

10.根据权利要求1所述的方法,其特征在于:倘若所述测试样本具有一或多个污染的SNP,则所述测试样本被确定为被污染。

11.根据权利要求1所述的方法,其特征在于:所述方法提供一或多个在遗传上不同的个体的污染率。

12.根据权利要求1所述的方法,其特征在于:所述方法进一步包括步骤:分析所述样本,以确保所述样本具有足够数量的遗传物质。

13.根据权利要求1所述的方法,其特征在于:在所述SNP的子集中,所述SNP的非匹配的等位基因的频率被计算为从0至1之间的数字。

14.根据权利要求1或13中任一项所述的方法,其特征在于:所述预定的非匹配的阈值为0.5%、1%、2%或5%。

15.根据权利要求1所述的方法,其特征在于:倘若所述样本在污染的SNP的SNP的子集中具有至少5个SNP,则所述样本被认为是一被污染的样本。

16.根据权利要求1所述的方法,其特征在于:倘若在所述SNP的子集中的至少5%、至少10%,或至少1%的SNP是污染的SNP,则所述样本被认为是一被污染的样本。

17.根据权利要求1所述的方法,其特征在于:各个SNP的等位基因的频率是通过计算读数的数量而被计算,所述读数在0至.5等级内包含各个等位基因,其中较小的等位基因是用于分子中;其中基因型是以0、1、2等级命名,其中根据所述参考序列,0是纯合的,及2是纯合的,但与所述参考序列相反,以及1是杂合的;其中倘若所述等位基因的频率大于或等于0.2,则所述基因型为1或是杂合的,倘若所述等位基因的频率<.2,且所述等位基因与所述参考序列相同,则所述基因型为0或纯合的,或者倘若所述等位基因与所述参考序列相反,则所述基因型为2。

18.根据权利要求1所述的方法,其特征在于:所述方法进一步包括步骤:识别所述污染的来源。

19.根据权利要求1所述的方法,其特征在于:确定在所述测试样本中的污染的来源包括步骤:将所述测试样本与一或多个替代参考序列进行比较,其中所述污染可追溯至一或多个替代参考序列。

20.根据权利要求19所述的方法,其特征在于:所述替代参考序列是来自一序列库、一公共数据库,或一产业数据库中的一者。

21.根据权利要求1所述的方法,其特征在于:所述方法进一步包括步骤:标记所述样本,以指示倘若所述样本与所述参考序列匹配,则可使用或出售所述样本。

22.根据权利要求1所述的方法,其特征在于:所述方法进一步包括步骤:倘若所述样本与所述参考序列匹配,则提供所述样本,以用于销售或使用。

23.根据权利要求1所述的方法,其特征在于:所述方法进一步包括步骤:标记所述样本,以表明倘若所述样本被鉴定为被污染,则无法使用或出售所述样本。

24.根据权利要求1所述的方法,其特征在于:所述方法进一步包括步骤:倘若所述样本被识别为被污染,则销毁所述样本。

25.根据权利要求1所述的方法,其特征在于:所述方法允许确认所述样本的特性、确定所述样本的纯度、检测在所述样本中的污染,以及确定在所述样本中的污染的来源。

26.根据权利要求1所述的方法,其特征在于:所述方法进一步包括步骤:确定所述样本是否具有足够的遗传物质以用于分析。

27.根据权利要求1所述的方法,其特征在于:所述样本是一错误标记的样本或一样本交换。

28.一种处理来自一测试样本的经提取的DNA,以检测一基因匹配的方法,其特征在于:所述方法包括步骤:

a.使用SNP引物对池,对来自所述测试样本的经提取的DNA进行核苷酸扩增,各个SNP引

物对侧接一独特的基因座,所述独特的基因座包含单一目标SNP,所述单一目标SNP定义一第一等位基因及一第二等位基因,所述核苷酸扩增针对各个SNP引物对产生多个扩增子,生成多个SNP扩增子池;

b.对所述SNP扩增子池进行测序,以针对在所述池中各个扩增子生成一核苷酸序列;

c.以及针对各个SNP,计算所述第一等位基因及所述第二等位基因的频率;

d.将在所述目标SNP的子集中的各个SNP的第一等位基因及第二等位基因的频率与一参考序列进行比较,其中所述目标SNP的子集是预期为纯合的一组SNP;

其中倘若在所述目标SNP的子集中的各个SNP的第一等位基因及第二等位基因的频率与在所述参考序列中对应的SNP完全匹配,则所述测试样本是基因匹配。

29.一种处理来自一测试样本的经提取的DNA,以检测污染的方法,其特征在于:所述方法包括步骤:

a.使用SNP引物对池,对来自所述测试样本的经提取的DNA进行核苷酸扩增,各个SNP引物对侧接一独特的基因座,所述独特的基因座包含单一目标SNP,所述单一目标SNP定义一第一等位基因及一第二等位基因,所述核苷酸扩增针对各个SNP引物对产生多个扩增子,生成多个SNP扩增子池;

b.对所述SNP扩增子池进行测序,以针对在所述池中各个扩增子生成一核苷酸序列;

c.针对各个SNP,计算所述第一等位基因及所述第二等位基因的频率;

d.将在所述目标SNP的子集中的各个SNP的第一等位基因及第二等位基因的频率与一参考序列进行比较,其中所述目标SNP的子集是预期为纯合的一组SNP;

其中倘若在所述SNP的子集中的各个SNP的第一等位基因及第二等位基因的频率与在所述参考序列中对应的SNP不完全匹配,则计算在所述SNP的子集中的各个SNP的非匹配的等位基因的频率,其中倘若针对一特定的SNP的非匹配的等位基因的频率高于一预定的非匹配的阈值,则所述特定的SNP是一污染的SNP,其中倘若所述污染的SNP的数量高于一预定的污染的SNP的阈值,则所述样本被鉴定为被污染。

30.根据权利要求27或28所述的方法,其特征在于:所述测试样本是一精子样本。

31.根据权利要求30所述的方法,其特征在于:所述精子样本已经受用于确定动物性别的机器,及所述精子样本包括活精子及死精子。

32.根据权利要求27或28所述的方法,其特征在于:所述方法是多重分析,其中至少2个样本同时经受所述方法。

33.根据权利要求27或28所述的方法,其特征在于:所述引物池包括至少48个引物组。

34.根据权利要求27或28所述的方法,其特征在于:核苷酸扩增是PCR扩增,及测序是下一代测序(NGS)。

35.根据权利要求27或28所述的方法,其特征在于:所述核苷酸扩增的步骤包括:使用引物对进行PCR扩增,及一随后的PCR步骤,以添加用于下一代测序的接头序列及条形码。

36.根据权利要求27或28所述的方法,其特征在于:所述方法进一步包括步骤:使用一分析系统以计算所述各个SNP的第一等位基因及第二等位基因的频率,以及将所述各个SNP的第一等位基因及第二等位基因的频率与在一序列库中的参考序列或参考序列组的频率进行比较。

37.根据权利要求27或28所述的方法,其特征在于:倘若在所述SNP的子集中,所述各个

SNP的第一等位基因及第二等位基因的频率与在所述参考序列中相同的SNP的频率完全匹配，则所述测试样本的纯度为至少98％。

38.根据权利要求27或28所述的方法，其特征在于：倘若所述测试样本具有一或多个污染的SNP，则所述测试样本被确定为被污染。

39.根据权利要求27或28所述的方法，其特征在于：所述方法提供一或多个在遗传上不同的个体的污染率。

40.根据权利要求27或28所述的方法，其特征在于：所述方法进一步包括步骤：分析所述样本，以确保所述样本具有足够数量的遗传物质。

41.根据权利要求27或28所述的方法，其特征在于：在所述SNP的子集中，所述SNP的非匹配的等位基因的频率被计算为从0至1之间的数字。

42.根据权利要求27或28所述的方法，其特征在于：所述预定的非匹配的阈值为0.5％、1％、2％或5％。

43.根据权利要求27或28所述的方法，其特征在于：倘若所述样本在污染的SNP的SNP的子集中具有至少5个SNP，则所述样本被认为是一被污染的样本。

44.根据权利要求27或28所述的方法，其特征在于：倘若在所述SNP的子集中的至少5％、至少10％，或至少1％的SNP是污染的SNP，则所述样本被认为是一被污染的样本。

45.根据权利要求27或28所述的方法，其特征在于：各个SNP的等位基因的频率是通过计算读数的数量而被计算，所述读数在0至.5等级内包含各个等位基因，其中较小的等位基因是用于分子中；其中基因型是以0、1、2等级命名，其中根据所述参考序列，0是纯合的，及2是纯合的，但与所述参考序列相反，以及1是杂合的；其中倘若所述等位基因的频率大于或等于0.2，则所述基因型为1或是杂合的，倘若所述等位基因的频率<.2，且所述等位基因与所述参考序列相同，则所述基因型为0或纯合的，或者倘若所述等位基因与所述参考序列相反，则所述基因型为2。

46.根据权利要求27或28所述的方法，其特征在于：所述方法进一步包括步骤：识别所述污染的来源。

47.根据权利要求27或28所述的方法，其特征在于：确定在所述测试样本中的污染的来源包括步骤：将所述测试样本与一或多个替代参考序列进行比较，其中所述污染可追溯至一或多个替代参考序列。

48.根据权利要求47所述的方法，其特征在于：所述替代参考序列是来自一序列库、一公共数据库，或一产业数据库中的一者。

49.根据权利要求27或28所述的方法，其特征在于：所述方法进一步包括步骤：标记所述样本，以指示倘若所述样本与所述参考序列匹配，则可使用或出售所述样本。

50.根据权利要求27或28所述的方法，其特征在于：所述方法进一步包括步骤：倘若所述样本与所述参考序列匹配，则提供所述样本，以用于销售或使用。

51.根据权利要求27或28所述的方法，其特征在于：所述方法进一步包括步骤：标记所述样本，以表明倘若所述样本被鉴定为被污染，则无法使用或出售所述样本。

52.根据权利要求27或28所述的方法，其特征在于：所述方法进一步包括步骤：倘若所述样本被识别为被污染，则销毁所述样本。

53.一种计算机实现的系统，其特征在于：所述计算机实现的系统用于进行根据权利要

求1至52中任一项所述的方法。

54.一种系统,其特征在于:所述系统用于进行根据权利要求1至52中任一项所述的方法中的一或多个步骤。

55.一种分析系统,其特征在于:所述分析系统用于为根据权利要求1至52中任一项所述的方法中的一或多个步骤编译及/或组织及/或进行数学及/或统计操作。

用于处理遗传样本以确定特性或检测污染的方法及系统

[0001] 相关申请交叉引用

[0002] 本申请要求于2020年10月21日提交的美国临时专利申请案第63/094,750号的优先权,其说明书通过引用整体并入本文中。

技术领域

[0003] 本发明涉及用于扩增、测序及分析遗传样本的方法,更具体地涉及用于处理遗传样本,例如,精子样本,以验证或确定样本的特性、检测其中的遗传污染的方法及系统,以及/或检测样本的错误标记。

背景技术

[0004] 在制备动物精液麦管时,可能会错误地标记麦管或意外地引入不需要的遗传物质。例如,样本可最先被分成多个等分试样,通过机器进行性别选择,之后进行重组。在重组步骤或其他步骤中,可能会引入污染或混合来自不同公牛的样本,导致小牛基于不正确的雄性亲畜而出生。

[0005] 用于评估在玉米、大豆或其他此种样本中的遗传纯度的现有方法(例如,当前的微阵列或芯片分析)主要集中于确定经评估的样本是否具有足够的纯度或足够低的污染水平,以用于额外的下游分析。现有方法亦在功能上检测各个品系的一个等位基因,以及可能只能识别在一样本中的给定品系的存在。不期望将本发明限制于任何理论或机制,据信此等方法通常不足以确定样本是否仅包括一组所需的遗传物质。

[0006] 然而,令人惊奇地发现,多重测序方法可被用于检测在麦管样本中的低水平的污染,例如低至1%(或小于1%)、2%、3%、4%、5%等。这是令人惊讶的,因为多重测序方法通常依赖于均匀的样本,而由于麦管样本包含死精子、紫外线照射的精子等,因此,麦管样本是非均匀的。

发明内容

[0007] 本发明的特征在于用于处理遗传样本,例如,精子样本(例如,来自精液麦管的样本)的方法及系统。本文的方法及系统允许验证或确定样本的特性(identity)。例如,本文的方法及系统允许确定/确认在样本中的遗传内容与参考动物匹配(例如,确认或确定是否存在100%的遗传匹配)。本文的方法及系统亦测试样本的纯度,以便检测可能的混合物或污染。在某些实施例中,本文的方法及系统的特征亦在于用于确定污染物的特性或遗传物质的来源的步骤(例如,在错误标记的样本的情况下)。

[0008] 不期望将本发明限制于任何理论或机制,据信本文的方法及系统是有利的,因为其允许简易、快速及高度灵敏的特性确认及/或污染的检测。多重分析有助于降低成本,及有助于实现一致性。

[0009] 本文的方法及系统有助于确定(在一定程度上)一测试样本仅包括一组所需的遗传物质。通过将精液麦管的遗传的纯度确定至此种程度,可更大程度地确保所述麦管包含

其声称的精子，从而避免代价高昂的错误。此外，某些司法管辖区要求高纯度标准，且本文的方法可帮助确保某些精子麦管可被出售。

[0010] 如前所述，当制备遗传物质的样本时，样本可能被错误标记或标签可能被错误读取。此外，混合或污染可能在拆分或分离样本，以及随后重组样本时发生，或者在处理遗传样本以进行基因分型时，可能发生污染。本文的方法及系统可用于帮助解决上述可能出现的关于遗传物质的加工、处理及质量控制的问题。

[0011] 不期望将本发明限制于任何理论或机制，据信本文的方法及系统是有利的，因为其改善了针对测序数据集的质量控制的现有方法，以及本文的方法及系统在精液处理的领域具有优势。现有的方法通常会检查在测序数据中的污染或样本切换，以确定测序数据是否合适。本发明提供一种方法，所述方法可识别在测序运行中的污染，以确认被测试的样本的特性(例如，确定在样本中的预期个体与在样本中所识别的实际个体之间是否匹配)，以及识别在在遗传样本(例如，精液麦管)中的任何污染，两者皆非通过现有方法提供的。

[0012] 简言之，使用SNP引物对池，对测试样本，例如遗传样本(例如，从所述测试样本中所分离或提取的DNA)进行扩增(例如，PCR扩增)，其中各个SNP引物对侧接一独特的基因座，所述独特的基因座包含单一目标SNP，所述单一目标SNP定义一第一等位基因及一第二等位基因。所述扩增(例如，PCR扩增)针对各个SNP等位基因产生多个扩增子，因此生成多个SNP扩增子池。所述方法进一步包括步骤：对多个SNP扩增子池进行测序(例如，下一代测序(next-generation sequencing，NGS))，其中测序针对在所述多个SNP扩增子池中的各个扩增子提供一核苷酸序列。

[0013] 将来自所述测序的结果提供至一分析系统，例如用于对从测序所获得的信息进行编译及/或组织及/或数学或统计操作的基于计算机的系统。例如，所述分析系统可计算各个SNP的第一等位基因及第二等位基因的频率。在某些实施例中，所述系统将各个SNP的序列及/或等位基因的频率与至少一参考序列(例如，文库序列、来自一已知个体，例如，一已知的公牛等的序列)的相应SNP进行比较。所述参考序列(例如，文库序列、来自一参考个体，例如，已知的公牛等的序列)的相应SNP的子集预期＝纯合的SNP。在某些实施例中，所述方法包括步骤：计算对应于在所述参考序列中的各个预期的纯合的SNP的测试样本中的各个SNP的非匹配的等位基因的频率。在一些实施例中，特定的计算是基于各个预期的纯合的SNP的非匹配的等位基因的频率，以及具有非匹配的等位基因的特定的频率的SNP的数量，从而允许确定一或多个在遗传上不同的的个体的污染率。所述分析有助于检测基因匹配(例如，确认特性)、检测混合物、确定遗传样本的特性，及/或确定污染物质的来源等。

[0014] 本文的方法及系统中所提及的遗传样本包括精子样本，例如，从公牛所获得的精子样本。然而，本发明并不限于精子样本，本发明亦不限于从公牛所获得的样本。如前所述，所述精子样本可储存在麦管中。在某些实施例中，所述精子样本可包括活精子及已经受紫外线照射的精子。在一些实施例中，所述精子样本包括活精子及死精子。在一些实施例中，所述精子样本已经受用于确定动物性别的机器。

[0015] 本文的方法被描述为自动化的多重方法。然而，本发明并不限于自动化及多样本的应用。多重分析是本领域普通技术人员所熟知的，其中至少2个样本同时经受所述方法，例如至少2个、至少6个、至少12个、至少24个、至少48个、至少96个等样本同时经受所述方法。

[0016] 本文所述的方法的特征可为具有额外的步骤，其中样本被重新测试，以确认结果，例如，非100％匹配的样本或标记为一潜在混合物的样本可被重新测试，以确认其被污染(或交换)，而非自动丢弃。在某些实施例中，样本在至少测试两次，以及两次或更多次显示为非100％匹配后被丢弃。

[0017] 本文的方法的特征亦可为用于分析所述样本的量的步骤。在某些实施例中，倘若确定所述样本的量不足，则可从来源(例如，精子麦管)获取另一样本。

[0018] 本发明亦包括将来自所述扩增及测序的结果提供至具有一用户界面的集成分析系统(例如，应用程序)。

[0019] 所述分析系统(例如，应用程序)可经由所述用户界面显示结果。结果的实例可包括但不限于：来自测序的SNP读取结果、测序结果的摘要、错误或警报等。例如，在某些情况下，所述分析系统可被编程为显示一警报(例如，经由所述用户界面)，例如，倘若无足够的遗传物质以进行分析、倘若存在疑似污染、倘若无基因匹配等。所述用户(例如，技术人员)可审查结果，及基于在所述用户界面上所显示的可视化数据(例如，从所述测序显示SNP读取结果的可视化数据)，确定是否需要手动审查所述数据及/或重新运行所述样本。

[0020] 本发明提供一种处理来自一测试样本的经提取的DNA的方法。在某些实施例中，所述方法包括步骤：使用SNP引物对池，对来自所述测试样本的经提取的DNA进行核苷酸扩增，各个SNP引物对侧接一独特的基因座，所述独特的基因座包含单一目标SNP，所述单一目标SNP定义一第一等位基因及一第二等位基因，所述核苷酸扩增针对各个SNP引物对产生多个扩增子，生成多个SNP扩增子池；对所述SNP扩增子池进行测序，以针对在所述池中各个扩增子生成一核苷酸序列；以及针对各个SNP，计算所述第一等位基因及所述第二等位基因的频率。可比较各个SNP的第一等位基因及第二等位基因的频率。例如，在某些实施例中，将在所述目标SNP的子集中的各个SNP的第一等位基因及第二等位基因的频率与一参考序列进行比较，其中所述目标SNP的子集是预期为纯合的一组SNP。倘若在所述目标SNP的子集中的各个SNP的第一等位基因及第二等位基因的频率与在所述参考序列中对应的SNP完全匹配，则所述测试样本与所述参考序列相同，例如，存在遗传匹配。倘若在所述SNP的子集中的各个SNP的第一等位基因及第二等位基因的频率与在所述参考序列中对应的SNP不完全匹配，则计算在所述SNP的子集中的各个SNP的非匹配的等位基因的频率(例如，所述SNP的子集并不完全匹配)。倘若针对一特定的SNP的非匹配的等位基因的频率高于一预定的阈值(例如，“非匹配的阈值”)，则所述特定的SNP被识别为一污染的SNP(例如，混合物或错误的潜在指示)。倘若所述污染的SNP的数量高于一预定的阈值(例如，“污染的SNP的阈值”)，则所述样本被鉴定为被污染(例如，所述样本被认为是一潜在的混合物或包含错误)。在一些实施例中，所述样本被确定为一错误标记的样本或交换。

[0021] 本文的方法及系统允许确认所述样本的特性、确定样本的纯度、检测在所述样本中的污染、确定在所述样本中的污染的来源，及/或确定所述样本是否具有足够的遗传物质以用于分析。

[0022] 在一些实施例中，所述核苷酸扩增的方法是PCR扩增。在一些实施例中，所述测序方法是下一代测序(next-generation sequencing，NGS)。在某些实施例中，所述核苷酸扩增的步骤包括：添加用于下一代测序的接头序列及条形码的步骤。

[0023] 在某些实施例中，所述方法进一步包括步骤：使用一分析系统以计算所述各个SNP

的第一等位基因及第二等位基因的频率，以及将所述各个SNP的第一等位基因及第二等位基因的频率与在一序列库中的参考序列或参考序列组的频率进行比较。

[0024] 在某些实施例中，所述测试样本是一精子样本。所述精子样本可已经受用于确定动物性别的机器，及所述精子样本包括活精子及死精子。所述精子样本可已经受紫外线照射。

[0025] 在一些实施例中，倘若在所述SNP的子集中，所述各个SNP的第一等位基因及第二等位基因的频率与在所述参考序列中相同的SNP的频率完全匹配，则所述测试样本的纯度为至少98％。在一些实施例中，倘若在所述SNP的子集中，所述各个SNP的第一等位基因及第二等位基因的频率与在所述参考序列中相同的SNP的频率完全匹配，则所述测试样本的纯度为至少99％。在一些实施例中，倘若在所述SNP的子集中，所述各个SNP的第一等位基因及第二等位基因的频率与在所述参考序列中相同的SNP的频率完全匹配，则所述测试样本的纯度为至少99.5％。

[0026] 在一些实施例中，倘若所述测试样本具有一或多个污染的SNP，则所述测试样本被鉴定为被污染。在一些实施例中，倘若所述测试样本具有2个或更多个污染的SNP，则所述测试样本被鉴定为被污染。在一些实施例中，倘若所述测试样本具有3个或更多个污染的SNP，则所述测试样本被鉴定为被污染。

[0027] 所述方法可以多重分析进行，例如，其中至少2个样本同时进行所述方法，其中至少48个样本同时进行所述方法，其中至少96个样本同时进行所述方法等。

[0028] 在某些实施例中，所述引物池包括至少48个引物组(例如，48个引物组、49个引物组、50个或更多个引物组等)。在某些实施例中，所述引物池包括24个或更多个引物组、30个或更多、36个或更多、40个或更多、42个或更多、45个或更多等。

[0029] 所述方法可进一步包括分析样本，以确保所述样本具有足够的量。在某些实施例中，倘若对于在引物池中的各个SNP(或特定的SNP的子集)，所述样本具有少于40个的读数，则所述样本具有不足量的遗传物质来进行分析。在某些实施例中，倘若对于在引物池中的各个SNP(或特定的SNP的子集)，所述样本具有少于35个的读数，则所述样本具有不足量的遗传物质来进行分析。在某些实施例中，倘若对于在引物池中的各个SNP(或特定的SNP的子集)，所述样本具有少于30个的读数，则所述样本具有不足量的遗传物质来进行分析。在某些实施例中，倘若对于在引物池中的各个SNP(或特定的SNP的子集)，所述样本具有少于25个的读数，则所述样本具有不足量的遗传物质来进行分析。在某些实施例中，倘若对于在引物池中的各个SNP(或特定的SNP的子集)，所述样本具有少于20个的读数，则所述样本具有不足量的遗传物质来进行分析。在某些实施例中，倘若对于在引物池中的各个SNP(或特定的SNP的子集)，所述样本具有少于45个的读数，则所述样本具有不足量的遗传物质来进行分析。在某些实施例中，倘若对于在引物池中的各个SNP(或特定的SNP的子集)，所述样本具有少于50个的读数，则所述样本具有不足量的遗传物质来进行分析。倘若所述样本中不具有足量的DNA以进行分析，则可获取另一DNA样本以进行重新测试。

[0030] 在某些实施例中，在所述SNP的子集中的SNP的非匹配的等位基因的频率被计算为从0至1的数字。

[0031] 在某些实施例中，预定的非匹配的阈值为0.5％。在某些实施例中，预定的非匹配的阈值为1％。在某些实施例中，预定的非匹配的阈值为2％。在某些实施例中，预定的非匹

配的阈值为5%。在某些实施例中，预定的非匹配的阈值为0.5%、1%、2%或5%。

[0032] 在某些实施例中，倘若所述样本在污染的SNP的SNP的子集中具有至少5个SNP，则所述样本被认为是一被污染的样本。在某些实施例中，倘若在所述SNP的子集中的至少5%的SNP是污染的SNP，则所述样本被认为是一污染的样本。在某些实施例中，其中倘若在所述SNP的子集中的至少10%的SNP是污染的SNP，则所述样本被认为是一污染的样本。在某些实施例中，其中倘若在所述SNP的子集中的至少15%的SNP是污染的SNP，则所述样本被认为是一污染的样本。

[0033] 所述方法可进一步包括步骤：确定在所述测试样本中的污染的来源。确定在所述测试样本中的污染的来源可包括步骤：将所述测试样本与一或多个替代参考序列进行比较，其中所述污染可追溯至一或多个替代参考序列。在某些实施例中，所述替代参考序列是来自一序列库、一公共数据库，或一产业数据库中的一者。

[0034] 在某些实施例中，各个SNP的等位基因的频率是通过计算读数的数量而被计算，所述读数在0至.5等级内包含各个等位基因，其中较小的等位基因是用于分子中；其中基因型是以0、1、2等级命名，其中根据所述参考序列，0是纯合的，及2是纯合的，但与所述参考序列相反，以及1是杂合的；其中倘若所述等位基因的频率大于或等于0.2，则所述基因型为1或是杂合的，倘若所述等位基因的频率<.2，且所述等位基因与所述参考序列相同，则所述基因型为0或纯合的，或者倘若所述等位基因与所述参考序列相反，则所述基因型为2。

[0035] 本发明亦提供一种处理来自一测试样本的经提取的DNA的方法。在某些实施例中，所述方法包括步骤：使用SNP引物对池，对来自所述测试样本的经提取的DNA进行PCR扩增，各个SNP引物对侧接一独特的基因座，所述独特的基因座包含单一SNP，所述单一SNP定义一第一等位基因及一第二等位基因，所述核苷酸扩增针对各个SNP引物对产生多个扩增子，生成多个SNP扩增子池；对所述SNP扩增子池进行下一代测序(next-generation sequencing，NGS)，以针对在所述池中各个扩增子生成一核苷酸序列，各个扩增子是SNP的第一等位基因或第二等位基因；以及针对各个SNP，计算所述第一等位基因及所述第二等位基因的等位基因的频率，以及将在SNP的子集中的各个SNP的第一等位基因及第二等位基因的等位基因的频率与在一参考序列中的彼等相同的SNP的子集进行比较，所述SNP的子集是预期为纯合的一组SNP。在某些实施例中，各个SNP的等位基因的频率是通过计算读数的数量而被计算，所述读数在0至.5等级内包含各个等位基因，其中较小的等位基因是用于分子中；其中基因型是以0、1、2等级命名，其中根据所述参考序列，0是纯合的，及2是纯合的，但与所述参考序列相反，以及1是杂合的；其中倘若所述等位基因的频率大于或等于0.2，则所述基因型为1或是杂合的，倘若所述等位基因的频率<.2，且所述等位基因与所述参考序列相同，则所述基因型为0或纯合的，或者倘若所述等位基因与所述参考序列相反，则所述基因型为2。在某些实施例中，倘若在所述SNP的子集中各个SNP的第一等位基因及第二等位基因的频率与在所述参考序列中的彼等相同的SNP完全匹配，则所述测试样本是与所述参考序列相同，以及所述测试样本的纯度为至少95%。在某些实施例中，倘若在所述SNP的子集中的各个SNP的第一等位基因及第二等位基因的频率不完全匹配，则计算非匹配的等位基因的频率，其中倘若所述非匹配的等位基因的频率高于一预定的非匹配的阈值，则所述SNP是一污染的SNP，其中所述方法进一步包括步骤：计算在所述测试样本中的污染的SNP的数量，其中倘若所述污染的SNP的数量高于一预定的污染的SNP的阈值，则所述样本被鉴定为具有一污染物。

[0036] 如前所述，本文的方法及系统允许确认所述样本的特性、确定所述样本的纯度、检测在所述样本中的污染、确定在所述样本中的污染的来源，及/或确定所述样本是否具有足够的遗传物质以用于分析。因此，本发明提供一种确认一样本的特性的方法；一种确定一样本的纯度的方法；一种检测在一样本中的污染物的方法；一种确定在所述样本中的污染的来源的方法；以及一种确定所述样本是否具有足够的遗传物质以用于分析。此等方法将上述步骤合并至处理样本的方法中，例如，使用SNP引物对池，对来自所述测试样本的经提取的DNA进行核苷酸扩增(其中各个SNP引物对侧接一独特的基因座，所述独特的基因座包含单一目标SNP，所述单一目标SNP定义一第一等位基因及一第二等位基因，其中所述核苷酸扩增针对各个SNP引物对产生多个扩增子，生成多个SNP扩增子池)；；对所述SNP扩增子池进行测序，以针对在所述池中各个扩增子生成一核苷酸序列；计算所述第一等位基因及所述第二等位基因的频率等。

[0037] 作为一实例，本发明提供一种处理来自一测试样本的经提取的DNA，以检测一基因匹配的方法。在某些实施例中，所述方法包括步骤：使用SNP引物对池，对来自所述测试样本的经提取的DNA进行核苷酸扩增，各个SNP引物对侧接一独特的基因座，所述独特的基因座包含单一目标SNP，所述单一目标SNP定义一第一等位基因及一第二等位基因，所述核苷酸扩增针对各个SNP引物对产生多个扩增子，生成多个SNP扩增子池；对所述SNP扩增子池进行测序，以针对在所述池中各个扩增子生成一核苷酸序列；以及针对各个SNP，计算所述第一等位基因及所述第二等位基因的频率。可比较各个SNP的第一等位基因及第二等位基因的频率。例如，在某些实施例中，将在所述目标SNP的子集中的各个SNP的第一等位基因及第二等位基因的频率与一参考序列进行比较，其中所述目标SNP的子集是预期为纯合的一组SNP。倘若在所述目标SNP的子集中的各个SNP的第一等位基因及第二等位基因的频率与在所述参考序列中对应的SNP完全匹配，则所述测试样本与所述参考序列相同，例如，存在一基因匹配。

[0038] 作为另一实例，本发明提供一种处理来自一测试样本的经提取的DNA，以检测可能的污染的方法。在某些实施例中，所述方法包括步骤：使用SNP引物对池，对来自所述测试样本的经提取的DNA进行核苷酸扩增，各个SNP引物对侧接一独特的基因座，所述独特的基因座包含单一目标SNP，所述单一目标SNP定义一第一等位基因及一第二等位基因，所述核苷酸扩增针对各个SNP引物对产生多个扩增子，生成多个SNP扩增子池；对所述SNP扩增子池进行测序，以针对在所述池中各个扩增子生成一核苷酸序列；以及针对各个SNP，计算所述第一等位基因及所述第二等位基因的频率。可比较各个SNP的第一等位基因及第二等位基因的频率。例如，在某些实施例中，将在所述目标SNP的子集中的各个SNP的第一等位基因及第二等位基因的频率与一参考序列进行比较，其中所述目标SNP的子集是预期为纯合的一组SNP。倘若在所述SNP的子集中的各个SNP的第一等位基因及第二等位基因的频率与在所述参考序列中对应的SNP不完全匹配，则计算在所述SNP的子集中的各个SNP的非匹配的等位基因的频率(例如，不完全匹配的SNP的子集)。倘若针对一特定的SNP的非匹配的等位基因的频率高于一预定的阈值(例如，“非匹配的阈值”)，则所述特定的SNP被识别为一污染的SNP(例如，混合物或错误的潜在指示)，以及所述污染的SNP的数量高于一预定的阈值(例如，“污染SNP阈值”)，则所述样本被鉴定为具有一潜在的污染物(例如，所述样本被认为是一潜在的混合物或包含错误)。

[0039] 本发明亦提供一种用于执行本文中所公开的方法的系统，例如，计算机实现的系统、分析系统。例如，所述系统可具有一处理器(例如，微处理器)的特征，以用于对在本文中公开的方法中的一或多个步骤进行数学及/或统计及/或分析操作。所述处理器(例如，微处理器)可为可操作地连接至所述系统的一或多个其他的组件，例如，测序系统、用户界面等。

[0040] 本文的方法可为计算机实现的方法。

[0041] 关于本文中所公开的方法，所述方法可进一步包括步骤：与销售及/或使用从中所获得的测试样本的所述样本相关的额外步骤，例如，所述测试样本所源自的精液麦管。在某些实施例中，所述方法进一步包括步骤："批准"或"合格"其中存在遗传匹配的样本，例如，以某种方式标记所述样本，以指示所述样本可被使用及/或出售。在某些实施例中，所述方法进一步包括步骤：提供一经批准(合格)的样本，以用于销售及/或使用。在某些实施例中，所述方法进一步包括步骤：倘若所述样本被认为可能被污染，则重新测试所述样本。在某些实施例中，所述方法进一步包括步骤："不合格"样本，其中存在污染物或错误，例如，以某种方式标记所述样本，以指示所述样本可能无法被使用及/或出售。在某些实施例中，所述方法进一步包括步骤：倘若确定所述样本被污染，则销毁所述测试样本所源自的样本。

[0042] 本文所述的任何特征或特征的组合皆包含在本发明的范围内，前提是包括在任何此种组合中的特征不会相互不一致，如同从上下文、本说明书及本领域普通技术人员的知识中显而易见。本发明的其他优点及方面在以下的详细描述及权利要求中是显而易见的。

附图说明

[0043] 通过考虑结合附图所呈现的以下详细描述，本发明的特征及优点将变得显而易见，其中：

[0044] 图1A显示一系列蓄意的混合的结果。

[0045] 图1B显示预期为纯合的单核苷酸多态性(single-nucleotide polymorphism，SNP)的第二等位基因的频率。

[0046] 图1C显示与来自第二头公牛的特定的污染部分的各种样本的背景杂讯相比较的混合的证据。

[0047] 图2显示特定的SNP的第二等位基因的频率的实例，所述特定的SNP预计在具有一污染的公牛的真实样本中是纯合的。SNP分为三类："0"代表第二等位基因的拷贝数为0；深灰色框(编号为37、30、34及35)代表所述第二等位基因的2个副本，及其余的编号框(编号为17、17、13、19、20、19及14)代表所述第二等位基因的一个副本。此用作识别所述污染的公牛的指纹。例如，公牛29HO17718是在所述组中唯一具有此特定基因型的公牛。

[0048] 图3A显示与本发明的方法相关的工作流程的示意图。本发明不限于在图3A中的工作流程。应注意的是，非100％匹配的样本在被销毁之前可经受重新测试。

[0049] 图3B显示与本发明的方法相关的工作流程的示意图。本发明不限于图3B中的工作流程。

[0050] 图4显示与本发明的方法相关的工作流程的示意图，其中所述测序数据是被存储在一数据库中，所述数据库集成或可操作地连接至具有一用户界面的分析系统。所述用户界面允许对所述样本进行审查及质量控制。

[0051] 图5显示一组假设麦管样本的输出数据的非限制性实例。

具体实施方式

[0052] 本发明的特征在于用于处理遗传样本及测试样本(例如,精液麦管)的纯度的方法及系统,例如,用于检测可疑的混合物或污染、用于检测一样本的错误标记(例如,样本交换)、用于确认一样本的特性、用于确定所述样本的特性(来源)、用于确定倘若存在的污染的特性(来源)等。

[0053] 本文用于处理遗传样本的方法的特征在于:扩增来自测试样本(例如,精子样本,例如,源自精液麦管的精子样本)的经提取的DNA的特定区域,对多个扩增子进行测序,以及分析所述多个扩增子。

[0054] 样本可作为经提取的DNA,以提供给一用户。或者,本领域普通技术人员能理解,在原始样本的情况下,可在进一步处理之前进行额外的DNA提取的步骤。

[0055] 扩增子的扩增及测序:

[0056] 参考本文的方法及系统,所述遗传样本(例如,经提取的DNA)经受扩增(例如,PCR扩增),其中所述扩增步骤产生特定的SNP的扩增子。DNA扩增的方法是本领域普通技术人员所熟知的。例如,扩增可指但不限于PCR扩增。在扩增过程中,SNP引物对的组用作引物,其中各个SNP引物对侧接一独特的基因座(一目标SNP),从一第一等位基因及一第二等位基因。

[0057] 在某些实施例中,所述引物池包括至少24个引物组(用于扩增24个目标SNP)。在一些实施例中,所述引物池包括至少36个引物组(用于扩增36个目标SNP)。在一些实施例中,所述引物池包括至少40个引物组(用于扩增40个目标SNP)。在一些实施例中,所述引物池包括至少48个引物组(用于扩增48个目标SNP)。在一些实施例中,所述引物池包括48个引物组。在一些实施例中,所述引物池包括49个引物组。在一些实施例中,所述引物池包括50个引物组。在一些实施例中,所述引物池包括24至48个引物组。在一些实施例中,所述引物池包括40至50个引物组。在一些实施例中,所述引物池包括48或49个引物组。在一些实施例中,所述引物池包括50至60个引物组。在一些实施例中,所述引物池包括多于50个引物组。

[0058] 在某些实施例中,至少10个目标SNP被预计是纯合的。在某些实施例中,至少15个目标SNP被预计是纯合的。在某些实施例中,至少20个目标SNP被预计是纯合的。在某些实施例中,至少25个目标SNP被预计是纯合的。在某些实施例中,至少30个目标SNP被预计是纯合的。在某些实施例中,至少35个目标SNP被预计是纯合的。在某些实施例中,至少10%的目标SNP被预计是纯合的。在某些实施例中,至少20%的目标SNP被预计是纯合的。在某些实施例中,至少25%的目标SNP被预计是纯合的。在某些实施例中,至少30%的目标SNP被预计是纯合的。在某些实施例中,至少40%的目标SNP被预计是纯合的。在某些实施例中,至少50%的目标SNP被预计是纯合的。在某些实施例中,至少60%的目标SNP被预计是纯合的。在某些实施例中,至少70%的目标SNP被预计是纯合的。在某些实施例中,至少75%的目标SNP被预计是纯合的。在某些实施例中,所有的目标SNP皆被预计是纯合的。

[0059] 在本发明的方法中所使用的引物对的实例描述于下表1中。应当理解,本文中所公开的引物仅作为一实例提供,以及本发明不限于本文中所公开的引物或SNP。基于本发明的公开内容,本领域的普通技术人员将能够选择其他的SNP,并为所述经选择的SNP设计引物。同样,本发明不限于测试公牛的精子样本;本文的方法及系统可应用于其他的样本类型及其他的物种(例如,其他的哺乳动物)。

[0060] 表1

[0061]

引物对	SEQ ID NO：	序列	染色体
1	1	TTTCCTTCTGTAGATGTTAACTGGT	5
	2	ACGGACAATAAACTGTAAATTTC	
2	3	CACTTGTATGTATTTCAGAAGTTTTC	25
	4	AGGATCACAAACAATGCCCT	
3	5	TCTGAAGGAATTGAAAATGTCTACCA	20
	6	ACTTCTACAATTAGCGATTAACTG	

[0062]

4	7	TCAGAGGAGAATGTCTAGTTTAGA	14
	8	TCAGTGTGACGGAGCTGC	
5	9	TCATGAGAAATCAGCCCACA	27
	10	ACCTGCTGTGTAATGTATTTAAAC	
6	11	AGGTTTAAGGGAATATTTGCACCT	19
	12	TTTGAATAACTGTACAGGGAATTC	
7	13	GCTTAAAGTTCTAAACCAATCAACA	1
	14	AGCTCTTCTAAAACAAGTAAAGCCA	
8	15	CCAAGAACCACTGTGATAGGAG	22
	16	TAAGAAAAGAGGGAACAAGACT	
9	17	ATGTGGCTTCCTGTATTCCCT	13
	18	GTTCAGCAAATAATTTATAGAGAACC	
10	19	CTGGCCCAAACTCATCACAC	8
	20	ACAAACCAACCACCAAACAGA	
11	21	TGAGTTTCAGAGAGGGCCAG	3
	22	GCCACTTGATGCCATAGGTC	
12	23	AGAAAGCCATACCCAGGGAG	11
	24	AGACAACAGTGAAGTTCAGGC	
13	25	CCCATGTATGTGTAGCTGGC	4
	26	CTGAGGAAGACAGGGAGAAGG	
14	27	ACATAAGTACATATCTACTGGCCT	2

[0063]

	28	TGAGAACCACTTAAGATAGGGT	
15	29	TACTAATTGTATATCTTGCTGCTCAG	3
	30	CTGCCTTCCCTCCTCAGTC	
16	31	AGGTATTATGAGTTGTGTGGGT	4
	32	ACATCCTCTTACCTAATCTGAGCT	
17	33	GGAAACCCTATGAGCCAGAGT	6
	34	TCCCTCAGCCCCTTTCAG	
18	35	GGGAGTGGAGAGTGGATTGG	13
	36	TGCCCCTTTTGTACAGATGG	
19	37	ATTCTCTTGACTTGCAGGCG	29
	38	GTCAATAACTCAATAAAAGCACTGA	
20	39	TGGGTTCTTGGGTCCAGAG	18
	40	AGGTCCCAGTCTCCTCCTTC	
21	41	GACTGCCTGGATCTGAGAGG	19
	42	CAACGGGACACTTGCCTTTC	
22	43	ATGTCTAGCCCTCACTCCCA	24
	44	CGTGATCAAGGCAGAACTACA	
23	45	AAGAGCAGCTGGTTTCCTAT	2
	46	AGGGGAGAAGATATGCAGACAG	
24	47	GTCCTCATCAGCCTCGTCAT	2
	48	ATATCCACCCTGCAAACCCA	

[0064]

25	49	AGTGCTATCATGTGCCTTGA	3
	50	ACACAGACTTTAAAGCAGCCT	
26	51	GAGGGGAGGGGAGGGTTG	5
	52	GAGAATTCAGGGGCCCTTCT	
27	53	GGATAATCTTTAGCAATCAGAGGC	7
	54	TCTCCCCGCTTGTAACAGTT	
28	55	AAAACTTGCTGCCAGGGAA	2
	56	CACGAGGCACATATCAGT	
29	57	CCTGTTACCTGGGCTTCATG	10
	58	ACGACTCGGCATAGATGATTTG	
30	59	CCTCTCTTTCAGGCCAAAATCC	8
	60	CTATTCTATTTTCCTCAAGTATCTGC	
31	61	ATTCAGATTGATGGTCCAGCA	15
	62	AGTGGTGATCTCAAAGAGGCT	
32	63	TAACTGCATCGTTAAACTGGCT	20
	64	CTTCTGGAGCTCACCCACC	
33	65	CACTCAGACGTCCCCAACC	10
	66	GGGTTCGGCAAGATTCAAGT	
34	67	GAGGCTGCCAGGTGTCAC	24
	68	TGAAAAGGTAGCCCAGGTGT	
35	69	GTTTTGCCTTCCACCCACC	13

[0065]

	70	TCTCCAAAGCAAACAAGTTAGTG	
36	71	CCAACCTGTTCAGCACA	26
	72	AGATCTTCCCCAACAGTACT	
37	73	GGAAGTCCCTTGTGCTCATC	21
	74	AAGTAGACCACGCCCCTTTC	
38	75	AAAGTAATAAGGCTTGCCTTCCA	2
	76	TGAGAAATATTGACTACACGCCC	
39	77	AAGGGGCTCATAAGATAAAGC	1
	78	CCAGAGATGGGGCATAGAACT	
40	79	GTGCCTGGCTCTGTACA	3
	80	TGCCGACCTCACGTGG	
41	81	GTCCATCCATGTTGCTGCAA	10
	82	TACTGGCCAAGAACACCT	
42	83	ATTATGTTTTAATGCACTGCTGT	11
	84	TTGGCATTGGGAAAGGGAAA	
43	85	TGTGGGTTGTAGTAGCAGCA	14
	86	ACCCAGTCACATTCAGAGCT	
44	87	ACTAAAATACAAATGCAGGCAC	10
	88	GGTACCCTGTATTCTTTCAAATGAGT	
45	89	TGCTGCAGATTCTTTACTTGCT	21
	90	TCTCTAGCCCATGTCATGACT	

[0066]

46	91	TCTTTTCATCCTTACCATAGCTAGG	28
	92	TTGCAGATGTGATCAAGGCT	
47	93	AGGATTGTATTTAAGGTTGTCTGA	18
	94	TGATGCCCACAATACCAGGT	
48	95	CAGGCAGATGGTTTTAACACAC	28
	96	AGATAATATAGACTGTAGTGCTGGGT	

[0067] 所述引物对允许扩增特定的SNP,其中至少一SNP的子集可帮助确定所述测试样本的特性及纯度。

[0068] 进一步对扩增步骤产生的SNP扩增子池进行测序。测序的方法,例如但不限于下一代测序(next-generation sequencing,NGS),是本领域普通技术人员所熟知的。所述方法中的其他步骤可包括用于NGS的接头序列的连接。

[0069] 分析:

[0070] 将来自测序的结果提供至一分析系统,例如应用程序,例如,用于对从测序所获得的信息进行编译及/或组织及/或进行数学或统计操作的基于计算机的系统。例如,所述分析系统可计算各个SNP的第一等位基因及第二等位基因的频率。如本文所述,所述分析系统将各个SNP的序列及/或等位基因的频率与至少一参考序列(例如,文库序列、来自已知个体的序列,例如,已知的公牛等)的相应的SNP进行比较。所述参考序列的相应SNP的子集(例如,文库序列、来自一参考个体的序列,例如,已知的公牛等)是纯合的SNP。在某些实施例中,所述方法包括步骤:计算在所述测试样本中的各个SNP的非匹配的等位基因的频率,所述在所述测试样本中的各个SNP对应于在参考序列中的各个经预期的纯合的SNP。在一些实施例中,特定的计算是基于各个经预期的纯合的SNP的非匹配的等位基因的频率及具有非匹配的等位基因的特定的频率的SNP的数量,从而允许确定一或多个在遗传上不同的的个体的污染率。所述分析有助于检测基因匹配(例如,确认特性)、确定遗传样本的特性、检测潜在的混合物及/或确定污染物资的来源等。

[0071] 在一些实施例中,所述方法及系统的特征在于检查在所述样本中的DNA的数量,以确定是否有足够的遗传物质被扩增及测序,以便被适当地分析。倘若无足够的DNA进行适当的分析,所述系统可产生一错误信号,例如“低(LOW)”信号。在某些实施例中,重复进行DNA提取。

[0072] 关于所述样本的数量,所述分析系统可计算所有的目标SNP或特定的SNP的子集(例如,预期为纯合的SNP的子集等)的频率。对于在所述预定的SNP的子集(或所有的目标SNP,倘若所述系统将所有的目标SNP用于数量分析)中的各个SNP,此等方法及系统可需要一定数量的读数。例如,在某些实施例中,所述方法及系统需要各个SNP(或SNP的子集)的至少40个读数。在某些实施例中,各个SNP(或SNP的子集)至少有35个读数。在某些实施例中,各个SNP(或SNP的子集)的至少30个读数。在某些实施例中,各个SNP(或SNP的子集)的至少25个读数。在某些实施例中,各个SNP(或SNP的子集)的至少20个读数。在某些实施例中,各

个SNP(或SNP的子集)至少有45个读数。在某些实施例中,各个SNP(或SNP的子集)的至少50个读数。本发明不限于上述SNP的需求。

[0073] 在某些实施例中,用于数量分析的SNP的子集包括5个或更多个SNP。在某些实施例中,用于数量分析的SNP的子集包括10个或更多个SNP。在某些实施例中,用于数量分析的SNP的子集包括15个或更多个SNP。在某些实施例中,用于数量分析的SNP的子集包括20个或更多个SNP。在某些实施例中,用于数量分析的SNP的子集包括25个或更多个SNP。在某些实施例中,用于数量分析的SNP的子集包括30个或更多个SNP。在某些实施例中,用于数量分析的SNP的子集包括35个或更多个SNP。在某些实施例中,用于数量分析的SNP的子集包括40个或更多个SNP。在某些实施例中,用于数量分析的SNP的子集包括全部的目标SNP。在某些实施例中,用于数量分析的SNP的子集是预期为纯合的所有的SNP。

[0074] SNP的总体组成及其相对于定义的标准(例如,序列库、参考序列等)的比例允许测试特性,及倘若适用,估计一或多个遗传上不同的个体的污染程度。

[0075] 所述分析系统被配置为计算各个SNP(或至少SNP的子集,例如预期为纯合的SNP的子集)的等位基因的频率。例如,所述分析系统被配置为计算预期为纯合的目标SNP的第二等位基因(非匹配的等位基因)的频率。本文中地方法的纯度分析部分依赖于一参考基因型,以及着重于在所述经预期的基因型中应为纯合的SNP。倘若所述SNP应为纯合的,则在所述样本中,检测到此SNP的其他的等位基因(第二等位基因,非匹配的等位基因)将是非预期的(或者至少检测到大量其他的等位基因将是非预期的,因为某个特定的杂讯水平通常伴随测序)。因此,其他的等位基因的频率应低于一特定的预定的阈值(例如,如下所述的非匹配的等位基因的频率阈值)。

[0076] 等位基因的频率可被表示为百分比、从0至1的数字等。例如,倘若对于一特定的预期的纯合的SNP未检测到第二等位基因,则所述等位基因的频率可表示为0。倘若使用0至1的等级,例如0.03、0.05、0.1、0.18、0.3等,则第二等位基因的检测可导致等位基因的频率大于0(且高达1)。

[0077] 如前所述,所述分析系统可使用第二等位基因(或非匹配的等位基因)的频率阈值,其中检测到处于及/或高于所述预定的阈值的非匹配的等位基因的频率表明所述SNP是“污染的”SNP。在某些实施例中,大于0的非匹配的等位基因的频率表示一污染的SNP。换句话说,在一些实施例中,所述非匹配的等位基因的频率阈值可为大于0的任何值。在某些实施例中,所述非匹配的等位基因的频率阈值(例如,将指示SNP是一污染的SNP的第二等位基因的频率)为0.005,或0.5%。在某些实施例中,所述非匹配的等位基因的频率阈值(例如,将指示SNP是一污染的SNP的第二等位基因的频率)是0.01或1%。在某些实施例中,所述非匹配的等位基因的频率阈值(例如,将指示SNP是一污染的SNP的第二等位基因的频率)是0.02或2%。在某些实施例中,所述非匹配的等位基因的频率阈值(例如,将指示SNP是一污染的SNP的第二等位基因的频率)是0.03或3%。在某些实施例中,所述非匹配的等位基因的频率阈值(例如,将指示SNP是一污染的SNP的第二等位基因的频率)是0.04或4%。在某些实施例中,所述非匹配的等位基因的频率阈值(例如,将指示SNP是一污染的SNP的第二等位基因的频率)是0.05或5%。

[0078] 纯度检查通常使用来自一样本的称为基因型的纯合子,来将所述样本分类为具有一混合物。例如,在一些实施例中,干净的纯合子被预期为具有a(maf<0.02),以及干净的杂

合子被预期为具有a(maf>0.4)。一混合物可为称为具有a(0.02<maf<0.4)的纯合子的SNP,。

[0079] 检测到超过一预定的阈值的一定数量的污染的SNP表明在所述样本中存在一混合物或可能的混合物(例如,遗传上的污染)。例如,检测到1个或多个污染的SNP表明在所述样本中存在一混合物(例如,传上的污染)。在一些实施例中,2个或更多个污染的SNP表明在所述样本中存在一混合物(例如,遗传上的污染)。在一些实施例中,3个或更多个污染的SNP在所述样本中存在一混合物(例如,遗传上的污染)。在一些实施例中,4个或更多个污染的SNP在所述样本中存在一混合物(例如,遗传上的污染)。在一些实施例中,5个或更多个污染的SNP在所述样本中存在一混合物(例如,遗传上的污染)。在一些实施例中,6个或更多个污染的SNP在所述样本中存在一混合物(例如,遗传上的污染)。在一些实施例中,7个或更多个污染的SNP在所述样本中存在一混合物(例如,遗传上的污染)。在某些实施例中,其中至少1%的经预期的纯合的SNP被标记为污染的SNP的样本被认为是一混合或经污染的样本。在某些实施例中,其中至少2%的经预期的纯合的SNP被标记为污染的SNP的样本被认为是一混合或经污染的样本。在某些实施例中,其中至少3%的经预期的纯合的SNP被标记为污染的SNP的样本被认为是一混合或经污染的样本。在某些实施例中,其中至少4%的经预期的纯合的SNP被标记为污染的SNP的样本被认为是一混合或经污染的样本。在某些实施例中,其中至少5%的经预期的纯合的SNP被标记为污染的SNP的样本被认为是一混合或经污染的样本。在某些实施例中,其中至少10%的经预期的纯合的SNP被标记为污染的SNP的样本被认为是一混合或经污染的样本。在某些实施例中,其中至少15%的经预期的纯合的SNP被标记为污染的SNP的样本被认为是一混合或经污染的样本。在某些实施例中,其中至少20%的经预期的纯合的SNP被标记为污染的SNP的样本被认为是一混合或经污染的样本。在某些实施例中,其中至少25%的经预期的纯合的SNP被标记为污染的SNP的样本被认为是一混合或经污染的样本。

[0080] 在某些实施例中,通过取3个最高频率SNP的中值来报告纯度(或污染量)。在某些实施例中,通过取4个最高频率SNP的中值来报告纯度(或污染量)。在某些实施例中,通过取3个最高频率SNP的中值来报告纯度(或污染量)。在某些实施例中,通过取5个最高频率SNP的中值来报告纯度(或污染量)。本发明不限于上述的纯度或污染量的方法。

[0081] 在某些实施例中,一受污染的样本是具有最高三个污染的SNP的中值频率大于5%的样本。在某些实施例中,一受污染的样本是具有最高三个污染的SNP的中值频率大于10%的样本。在某些实施例中,一受污染的样本可为具有最高三个污染的SNP的中值频率大于15%的样本。在某些实施例中,一受污染的样本可为具有最高三个污染的SNP的中值频率大于20%的样本。在某些实施例中,一受污染的样本可为具有最高三个污染的SNP的中值频率大于25%的样本。

[0082] 在一些实施例中,在所述样本中的污染量为1%或更少。在一些实施例中,在所述样本中的污染量为2%或更少。在一些实施例中,在所述样本中的污染量为3%或更少。在一些实施例中,在所述样本中的污染量为4%或更少。在一些实施例中,在所述样本中的污染量为5%或更少。在一些实施例中,在所述样本中的污染量为1至2%或更少。在一些实施例中,在所述样本中的污染量为2至3%或更少。在一些实施例中,在所述样本中的污染量为2至4%或更少。在一些实施例中,在所述样本中的污染量为2至5%或更少。在一些实施例中,在所述样本中的污染量为5至10%或更少。在一些实施例中,在所述样本中的污染量为2至

15%或更少。在一些实施例中，在所述样本中的污染量为1%或更多。在一些实施例中，在所述样本中的污染量为2%或更多。在一些实施例中，在所述样本中的污染量为3%或更多。在一些实施例中，在所述样本中的污染量为5%或更多。

[0083] 作为一非限制性实例，在某些实施例中，至少20%的经预期的纯合的SNP必须具有至少1%的污染水平，方能使所述样本被标记为被污染。在某些实施例中，至少10%的经预期的纯合的SNP必须具有至少1%的污染水平，方能使所述样本被标记为被污染。在某些实施例中，至少20%的经预期的纯合的SNP必须具有至少2%的污染水平，方能使所述样本被标记为被污染。在某些实施例中，至少10%的经预期的纯合的SNP必须具有至少2%的污染水平，方能使所述样本被标记为被污染。在某些实施例中，至少20%的经预期的纯合的SNP必须具有至少3%的污染水平，方能使所述样本被标记为被污染。在某些实施例中，至少10%的经预期的纯合的SNP必须具有至少3%的污染水平，方能使所述样本被标记为被污染。

[0084] 如前所述，计算所述第二等位基因的频率(或非匹配的等位基因的频率)允许所述测试样本的特性，以确认所述精液是否来自正确的动物、以确定所述动物的起源于来自何者、识别一样本交换、识别污染的来源等。

[0085] 关于特性测试，在所述样本中的基因型可通过对照参考序列进行检查来确定，例如但不限于奶牛育种委员会(Council on Dairy Cattle Breeding，CDCB)数据库。倘若所述样本具有预期的基因型(例如，倘若所有经预期的纯合的SNP皆与所述经预期的基因型相匹配)，则所述样本可合格。倘若非如此，则可对照先前经测序的基因型(替代参考)来检查样本，例如但不限于在实验室中的先前经测序的样本(先前的QC样本)、其他产业样本、历史样本等(例如，公共数据库，尽管一些动物可能不存在于公共数据库中)。倘若所述样本是用作替代参考的参考序列，则所述样本可合格。倘若所述样本未合格，则可对其进行进一步分析及/或测试。

[0086] 所述方法及系统可能够基于在一序列库中的序列，来识别在一混合测试样本中的个体的组合(例如，动物#1加上动物#2、动物#1加上动物#2加上动物#3等)。在某些实施例中，所述样本可被确定为一样本交换或错误标记。

[0087] 本发明不限于用于计算在一样本中的污染水平的上述参数或参数组合。通常，SNP的总体组成及其相对于定义标准的比例可估计一或多个遗传上不同的个体的污染程度。最终结果的形式可为错配数量、所述样本是否为一样本交换、混合物百分比、污染百分比、潜在污染物列表等。其他的结果可包括各个SNP的等位基因的频率。如前所述，识别一被受污染的样本或混合物可导致重新测试一样本。之后可确定所述样本是纯的。不准确的纯度(例如，纯样本、混合的样本)可由于各种情况所造成，例如但不限于在所述测序过程中的错误、由一技术人员在样本制备的过程中的污染等。

[0088] 在某些实施例中，确定污染物的特性。例如，可对照参考基因型进行检索，所述参考基因型例如具有很高的污染物可能性的样本的参考基因型。在一些实施例中，可通过将结果与当天进行测序中的所有序列进行比较，以识别污染物。

[0089] 如前所述，在所述样本中的SNP的DNA序列可与一或多个参考序列进行比较。在某些实施例中，所述参考序列是在一序列库中的序列。在某些实施例中，所述参考序列是一组公牛的一或多个序列，所述一组公牛例如与被测试的精子样本匹配的公牛、与被测试的精

子样本非匹配的公牛、可为需对在麦管中的精子样本的污染负责的公牛、在一特定的组或群体中的公牛等。例如，一参考文库可包括来自一组公牛的序列(各个序列皆具有其独特的SNP概况)。如前所述，与参考序列的比较可有助于确认所述样本的特性、确定精子样本属于哪头公牛，及/或倘若存在污染，确定哪头公牛是污染的来源。

[0090] 图1A显示一系列蓄意的混合，说明经混合或被污染的样本的结果。图表102显示混合比为0。图表104显示混合比为1。图表106显示混合比为2.5。图表108显示混合比为5。图表110显示混合比为7.5。图表112显示混合比为10。图表114显示混合比为25。图表116显示混合比为50。在图1B中，图表120、122、124、126、128、130、132及134显示在混合比为0、1、2.5、2、5、7.5时，经预期为纯合的SNP的第二等位基因的频率，以及10个个体：布罗姆利、安弗拉姆、雷塔罗、匡滕、钱柏、哈特利、曼宁，以及都那利。图1C显示与来自一第二头公牛的特定的污染部分的各种样本的背景杂讯相比较的混合的证据。

[0091] 图2显示经预期为纯合的特定的SNP的第二等位基因的频率的实例。在图2的示例图表中，在第2、4、10号染色体的SNP等于或大于30％的第二等位基因的频率，以及在第2、3、5、8、18、21号染色体具有高于0％，但低于30％的第二等位基因的频率。此用作用于识别所述污染的公牛的指纹。例如，公牛29HO17718是在此组中唯一具有此特定的基因型的公牛。

[0092] 图3A、图3B，及图4显示本发明的方法及系统以及工作流程的示意图。应注意的是，非100％匹配(或标记为混合或可能混合)的样本在销毁前可经受重新测试。

[0093] 实例1：

[0094] 以下是本发明的非限制性实例。应当理解，所述实例不旨在以任何方式限制本发明。等同物或替代物是在本发明的范围内。

[0095] 首先，使用标准方法进行DNA提取。之后使用Qiagen Multiplex PCR Master Mix(货号：206145)及48对组合的引物进行多重PCR。之后将产物进行珠粒清洗，及验证在凝胶上的扩增。进行第二次PCR，以添加Illumina接头序列及条形码。之后将所述样本合并，并在下一代测序仪(next-generation sequencer，NGS)上进行测序，例如使用1x75 bp读数的Illumina测序仪。藉由bcl2fastq通过条形码对所述样本进行多重分解及分箱。使用BWA MEM，将fastqs与UMD3.1基因组对齐。各个SNP的等位基因的频率(Allele frequencie，AF)是通过计算读数的数量而被计算，所述读数在0至.5等级内包含各个等位基因，其中较小的等位基因被使用于分子中。基因型(Genotypes，GT)基因型是以0、1、2等级命名，其中0是纯合参考，及2是纯合替代。1是杂合的。倘若所述AF是大于或等于0.2，则所述GT为1(杂合的)。倘若所述AF是<.2且所观察到的等位基因是参考，则所述GT为0，否则GT为2。一旦所述基因型被命名之后，可通过将所观察到的基因型与在数据库中的基因型进行比较，以检查样本的特性。倘若所述样本与在所述数据库中的基因型非匹配，且非匹配少于4个，则扫描奶牛育种委员会数据库(Council of Dairy Cattle Breeding database，CDCB)，以寻找完美匹配。倘若发现>95％匹配的匹配度，则报告最高匹配度，并将所述样本标记为一完全交换。为了检测污染，首先选择经预期为纯合的SNP。倘若发生污染，预期此等SNP具有AF>0.02。通过检查此等非预期的AF的多个SNP，可确定一样本是否受到污染。使用前3个SNP，可估计污染水平。图3显示与本发明的方法相关的工作流程300的示意图。本发明不限于图3中的工作流程300。

[0096] 参考图3B中所提供的示例性工作流程300。步骤302提供一遗传样本的DNA提取。在

步骤304中，对所述样本的48个SNP进行多重PCR。在步骤306中，Illumina接头经由PCR添加至所述遗传样本中。在步骤308中，使用例如1x75 bp读数，在诸如Illumina测序仪的NGS上对所述样本进行测序。在步骤310中，来自步骤308的读数与基因组进行比对。在步骤312中，计算支持各个等位基因的读数，及命名遗传样本的基因型。在步骤314中，通过将步骤312的结果，所述被命名的基因型与已知基因型进行比较来识别样本交换，例如通过将被命名的基因型与在一序列库或数据库中的参考序列进行比较。在步骤316中，通过识别等位基因的比例，来识别样本混合物或污染，以及确定混合水平。在步骤318中，至少基于步骤314中的比较及在步骤316中的识别来识别一样本状态，所述样本状态例如是干净(例如，无污染、错误标记或错误识别的问题)、混合(例如，污染问题)或完全交换(例如，贴错标签或错误识别的问题)中的一者。

[0097] 实例2：

[0098] 以下是本发明的非限制性实例。应当理解，所述实例不旨在以任何方式限制本发明。等同物或替代物是在本发明的范围内。

[0099] 以下的实例描述本发明的方法及系统的特定的实施例的概况。

[0100] 从收集的精子细胞中提取DNA，此等精子细胞已经过加工及包装成麦管。以下的过程描述对照公牛的已知/参考DNA来分析DNA，以帮助确保在麦管中的DNA是打印在麦管上的公牛的DNA。可能的污染源是分析系统(演算法)对污染源的最佳猜测，选项是在单次运行中处理的所有公牛(例如，在某些实施例中，倘若真正的污染源不存在于测序运行中，则可能无法识别)。另应注意的是，测序运行可包括来自各种来源的DNA，以及可包括常规及性别鉴定的DNA。在某些实施例中，完整运行是96个DNA样本x6个盘。然而，本发明不限于此等参数。

[0101] 从生产实验室所接收的DNA被扩增。所述扩增过程扩增一组SNP，例如，48至49个SNP，包括已知的变异区域。对于公牛的SNP的独特的概况是众所周知的。各个亲本动物的基因型亦可拥有。此提供一种动物基因型库，包括各个动物的概况，其中包括SNP，例如，48至49个SNP。

[0102] 检验DNA如何在各个SNP位置进行读取(例如，确定读取为何)。将各个DNA读数与文库进行比较。首先，尝试确定100％匹配。倘若识别出100％匹配，则可结束所述过程。倘若无法识别出100％匹配，则所述系统将尝试在所述识别库中已知的概况的的组合。所述系统可识别所述样本是否包括例如动物#1加上其他物资(来自另一样本的污染)。可确定所述样本并非其所认为的(例如，“替代”或一样本交换或贴错标签)，或者可确定所述样本包括一些污染物。

[0103] 非100％匹配的估计可基于例如总目标SNP的10至20个SNP的子集。所述系统可识别近似的混合百分比，例如，+/-5％。在某些实施例中，系统可以识别一混合物以及样本与何者混合。然而，在某些实施例中，所述系统可识别一混合物，但不一定识别所述样本与何者混合(例如，可并非总是能够识别污染者)，例如，倘若另一个体不可用于比较。

[0104] 测序数据，例如NGS数据，可从所述测序仪路由至一内部数据库(例如，一分析系统，例如，应用程序)。数据可作为在应用程式(app)用户界面中的输出而被提供。来自应用程式的数据可作为一输出文件而被导出。所述输出文件可包括所有运行的批次及其排名。在一些实施例中，各个批次运行的标识在所述输出文件中指示，以及可包括附加信息的数个列或类别。例如，一类信息可为特性，其中“合格”是当所述特性被发现及确认时。其他的

类别的信息可包括:纯度;基因组学,其中所述信息与所述测试样本是否为混合物有关;数量,其中所述信息与是否有足够的样本进行运行及分析有关;时间,其中所述信息与动物样本运行的次数有关(例如,32次中的32次,其意味所述动物样本已运行32次,以及各次皆返回正确的结果;167次中的155次,其将意味所述动物的样本已运行155次,但12次样本返回时是混合的等)等。图5显示所述输出文件的假设实例,其中样本EX1至EX3被标记为潜在的混合,其中样本EX4至EX7被显示为匹配。检查样本EX1至EX3,以确定错误来源。样本EX4至EX7通过所有的测试并释放以被使用。

[0105] 如前所述,可修改数量及纯度的阈值。因此,诸如在图5中的输出数据是所述预定的阈值的函数。

[0106] 可在以下情况下进行识别为混合的项目的人工审查:无基因组数据可与之比较,所述样本为7%混合。在一些实施例中,重新测试两个麦管。倘若检测到相同百分比的污染及相同数量的污染的SNP,则所述样本不合格。

[0107] 在某些实施例中,倘若被适当地识别为另一头公牛,则完全交换(例如,一样本的错误标记)可成为合格。

[0108] 实例3:

[0109] 以下是本发明的非限制性实例。应当理解,所述实例不旨在以任何方式限制本发明。等同物或替代物是在本发明的范围内。

[0110] 以下的实例描述所述分析系统的概况,例如,用于分析及质量控制的应用。图4显示与所述测序仪及内部数据库以及质量控制集成的分析系统的示意图。

[0111] 一用户(例如,技术人员)可登录至所数分析系统(例如,应用程序)的用户界面。所述用户可使用所述系统,以审查某些样本,例如,确认基因匹配、确定一样本是否可为混合物等。

[0112] 所述分析系统可以视觉指示器作为特征,其提醒一用户注意特定的问题(例如,可能的混合、数量不足、与预期的参考动物非100%匹配等)。

[0113] 一批次称为一次运行,以及一次运行是由数个盘所组成,各个盘包括数个样本。

[0114] 在无足够的序列信息或无足够的读数来确定纯度或特性的情况下,可能会发生失败。

[0115] 在某些情况下,一技术人员将检查混合物的列表。结果可显示为图表,其中X轴是基因座(SNP)。倘若所述样本是混合的,就会有明显的图案(例如,参见图1A、图1B)。在某些实施例中,与个体公牛的共通/已知的DNA相匹配的SNP可在一第一设计中被标记(例如,完整的、用圆圈填充的;第一种颜色等);与个体公牛的共通/已知的DNA非匹配的SNP可在一第二设计中被标记(例如,完整的空心圆圈;第二种颜色等);以及无共通/已知的DNA进行比较的SNP可在一第三设计中被标记(例如,使用X标记;第三种颜色等)。一干净的样本不存在指示污染或非预期的结果的点(例如,红点)。

[0116] 所述分析系统,例如绘图,倘若样本可能被污染,允许确定在所述过程中何时存在超出阈值的混合物。所述系统有助于可视化过程的输出何时出错。

[0117] 所述分析系统可允许与过去的样本进行比较。例如,一用户可选择最近的运行以进行回顾。所述系统有助于轻松地查看异常值。所述系统可帮助识别在一定数量的样本(麦管)上出现的混合物(例如,低水平的混合物),例如,超过20个麦管、30个、40个、50个等,其

可表明在过程中存在问题。

[0118] 所述系统可允许设置合格(例如,识别为基因匹配)、不合格(例如,识别为错配或可能的混合),以及混合(例如,何者被认为是一混合)的阈值。在某些实施例中,所述阈值可通过图形或绘图中的线而被可视化。

[0119] 如前所述,所述系统可提供指示诸如污染或DNA太少的问题的方法。在某些实施例中,所述读数及读数百分位数是所述测序进行如何的指标,以及所述分析系统可提供所述数据的置信水平。

[0120] 作为非限制性实例,所述系统可使用在图上的红点作为污染或非预期的结果的指标。图1C显示一实例,其中指示为“混合证据”的点是需要评估的潜在问题的指标。例如,杂合失败可能是其中一种混合与SNP的预期不完全相同;其他的失败可能会显示为在所述绘图或界面的中间的点(例如,红点)。在某些实施例中,所述应用程序允许比较分层的样本数据,以查看个体动物的一致失败的区域。

[0121] 虽然已显示及描述本发明的优选实施例,但是对于本领域的技术人员而言显而易见的是,可对其进行修改而不超出所附权利要求的范围。因此,本发明的范围仅受所附权利要求的限制。在一些实施例中,本专利申请中所呈现的附图是按比例绘制的,包括角度、尺寸比例等。在一些实施例中,附图仅是代表性的,且权利要求不受附图尺寸的限制。在一些实施例中,在本文中所描述的本发明的描述使用短语“包括”可被描述为“基本上由...组成”或“由...组成”的实施例,因此满足使用短语“基本上由...组成”或“由...组成”来要求保护本发明的一或多个实施例的书面描述要求。

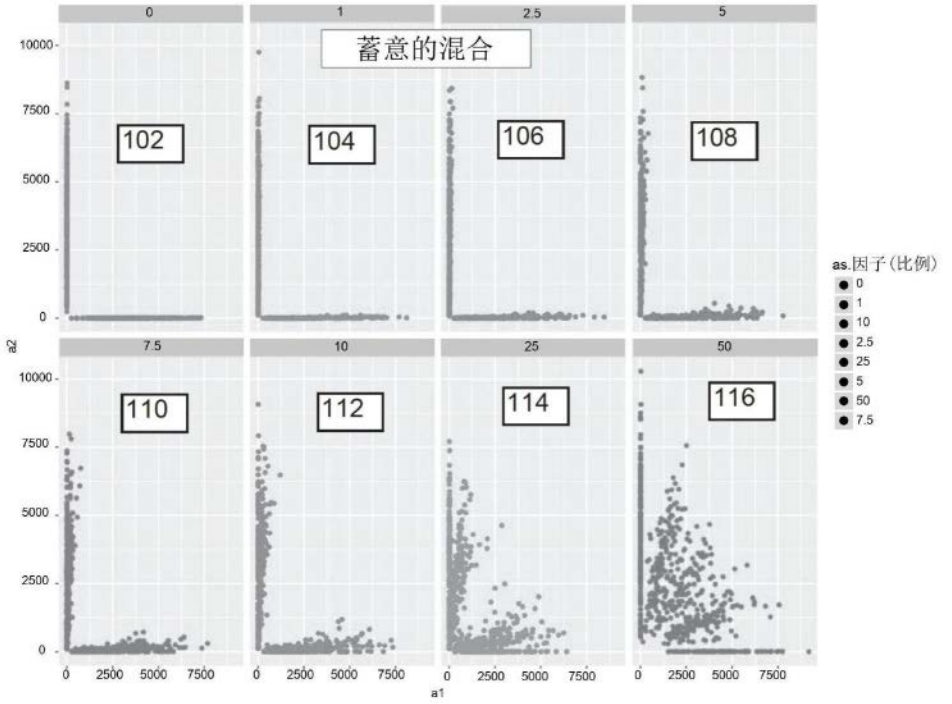

蓄意的混合
0
1
2.5
5
7.5
10
25
50
102
104
106
108
110
112
114
116
a2
a1
as.因子(比例)
0
1
10
2.5
25
5
50
7.5

图1A

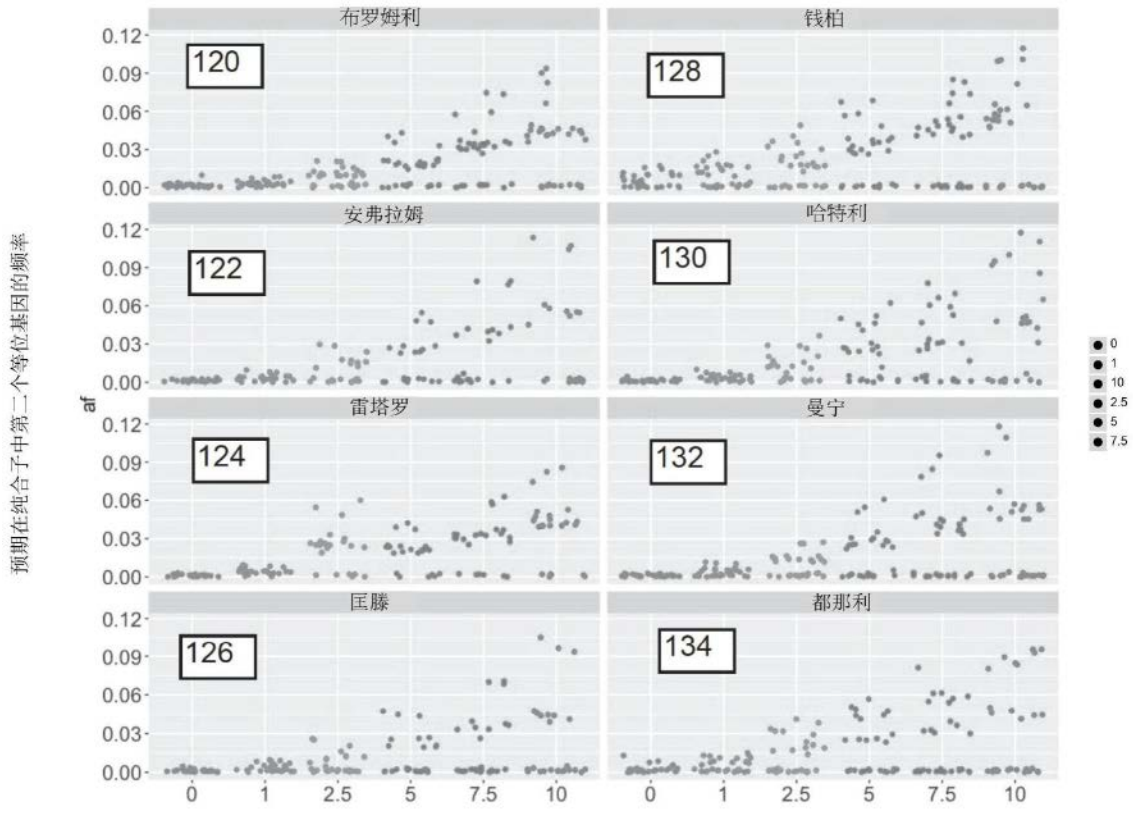

布罗姆利
钱柏
安弗拉姆
哈特利
雷塔罗
曼宁
匡滕
都那利
120
128
122
130
124
132
126
134
af
预期在纯合子中第二个等位基因的频率
0.12
0.09
0.06
0.03
0.00
0
1
2.5
5
7.5
10
0
1
10
2.5
5
7.5

图1B

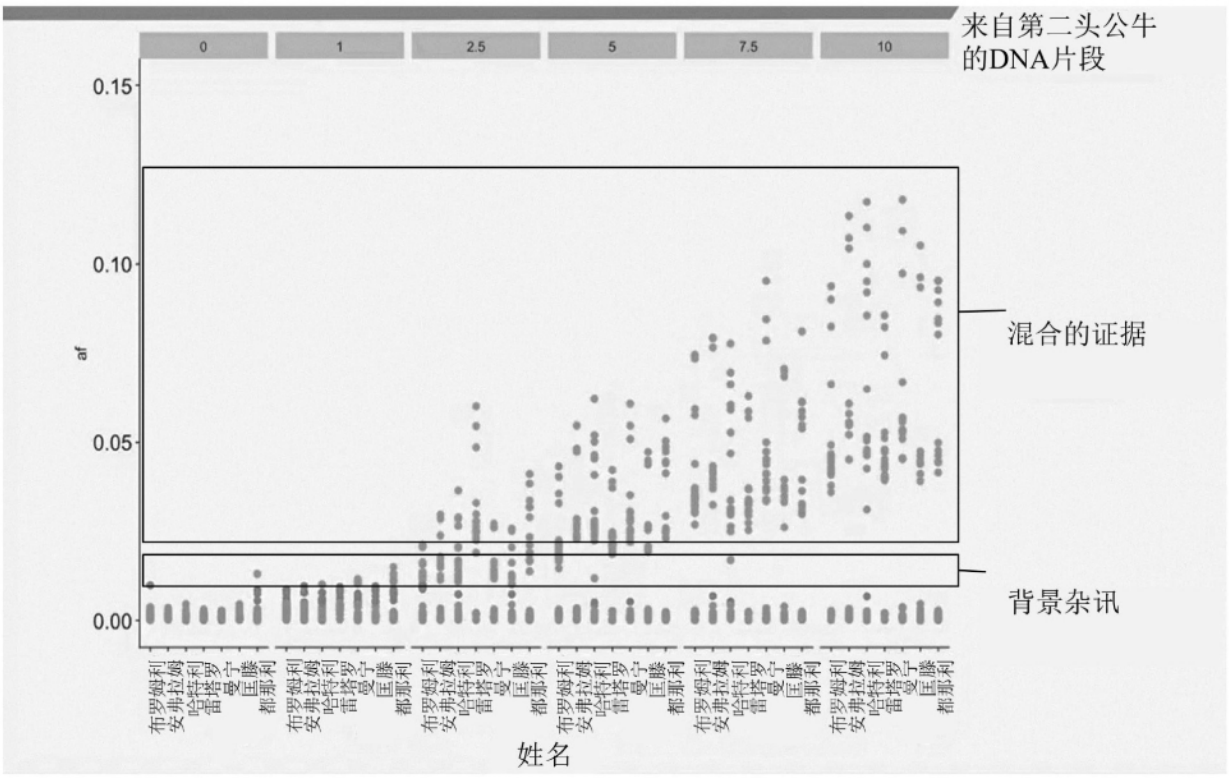


图1C

染色体	1	1	2	2	2	2	2	3	3	3	3	4	4	5	5	6	7	8	8	10	10	10	10	11
629HO18246-8170	TT	CC	AG	AG	TT	AA	CT	CC	GG	CC	CC	GG	GG	GT	GG	AG	TT	TT	CC	-	CT	GT	CC	GG
纯合的	TT	CC			TT	AA		CC	GG	CC	CC	GG	GG		GG		TT	TT	CC				CC	GG
第二等位基因的频率(%)	0	0			37	17		0	17	0	13	30	34		19		0	20	0				35	0
29HO17718	TT	CC			GG	GA		CC	AG	CC	CT	AA	AA		AG		TT	CT	CC				TT	GG

染色体	11	13	13	13	14	14	15	18	18	19	19	20	20	21	21	22	24	24	25	26	27	28	28	29
629HO18246-8170	AG	CC	AG	TT	CT	TT	AG	CT	TT	GG	AG	CT	AG	GG	AA	AA	TT	AG	CT	CT	CG	CC	AA	AG
纯合的		CC		TT		TT			TT	GG				GG	AA	AA	TT					CC	AA	
第二等位基因的频率(%)		0		0		0			19	0				14	0	0	0					0	0	
29HO17718		CC		TT		TT			CT	GG				GT	AA	AA	TT					CC	AA	

图2

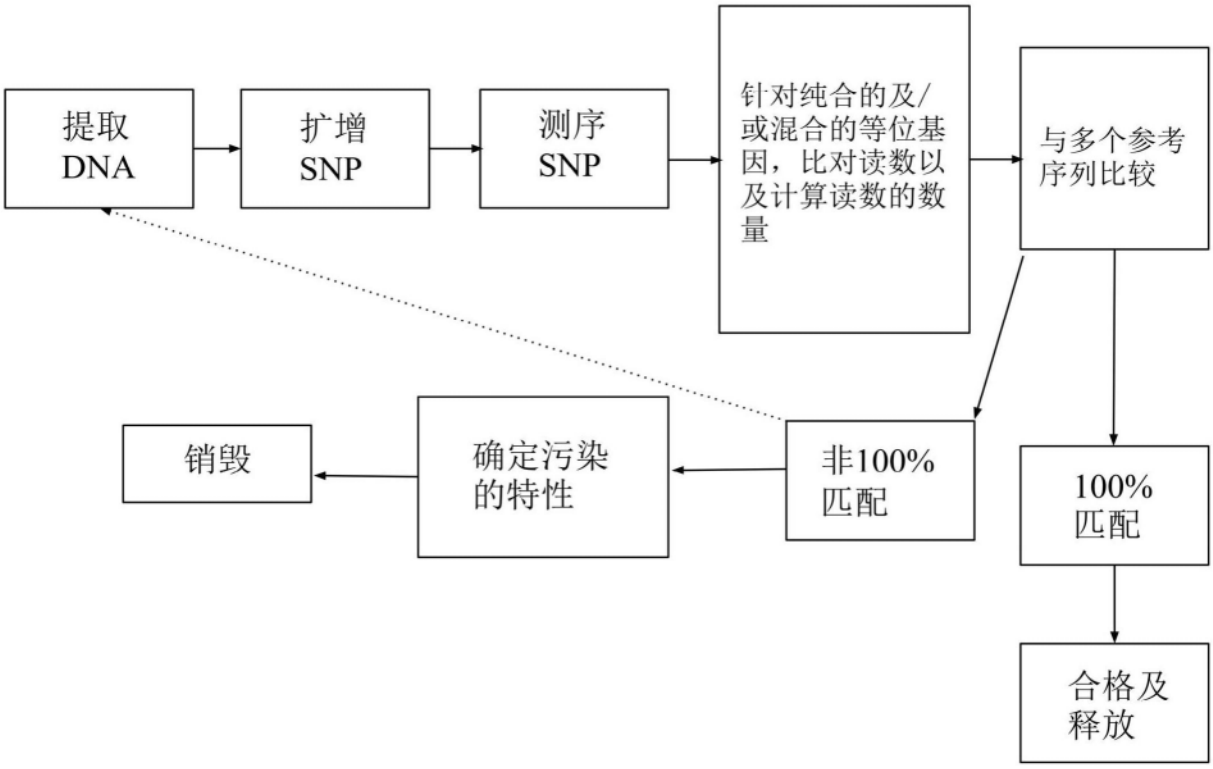

提取DNA
扩增SNP
测序SNP
针对纯合的及/或混合的等位基因，比对读数以及计算读数的数量
与多个参考序列比较
销毁
确定污染的特性
非100%匹配
100%匹配
合格及释放

图3A

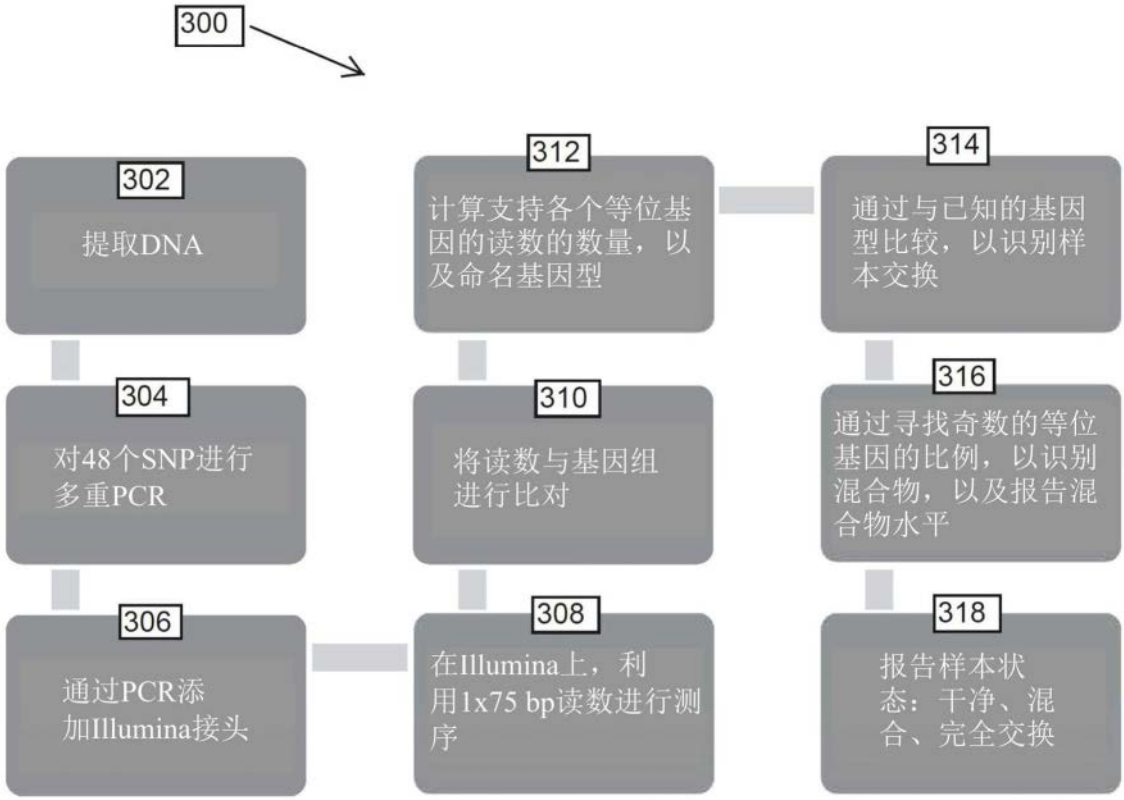

300
302
提取DNA
304
对48个SNP进行多重PCR
306
通过PCR添加Illumina接头
308
在Illumina上，利用1x75 bp读数进行测序
310
将读数与基因组进行比对
312
计算支持各个等位基因的读数的数量，以及命名基因型
314
通过与已知的基因型比较，以识别样本交换
316
通过寻找奇数的等位基因的比例，以识别混合物，以及报告混合物水平
318
报告样本状态：干净、混合、完全交换

图3B

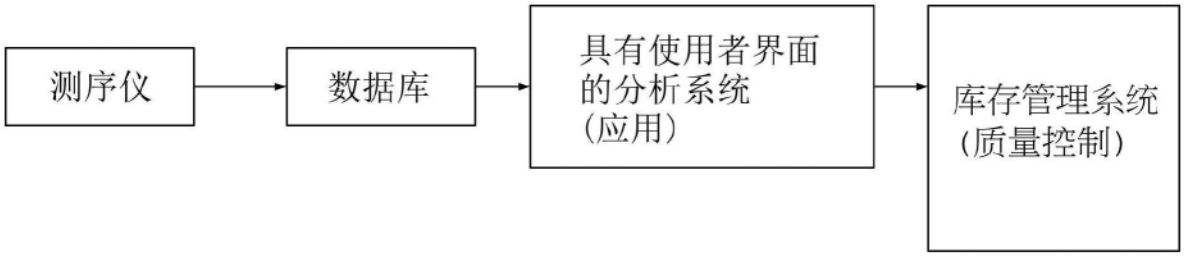


图4

麦管标记（代码）	盘	数量	特性	先前的样本	纯度	估计的污染	基因组学
EX1	1	合格	不匹配	6次中的0次	混合的	0.05805	审查
EX2	1	合格	不匹配	9次中的8次	混合的	0.4574	审查
EX3	1	合格	不匹配	255次中的224次	混合的	0.0698	审查
EX4	1	合格	合格	32次中的32次	合格	N/A	合格
EX5	1	合格	合格	234次中的227次	合格	N/A	合格
EX6	1	合格	合格	126次中的120次	合格	N/A	合格
EX7	1	合格	合格	17次中的15次	合格	N/A	合格

图5