(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 107002121 B
(45)授权公告日 2020.11.13

(21)申请号 201580050718.4
(22)申请日 2015.09.15
(65)同一申请的已公布的文献号
申请公布号 CN 107002121 A
(43)申请公布日 2017.08.01
(30)优先权数据
62/052,189 2014.09.18 US
(85)PCT国际申请进入国家阶段日
2017.03.20
(86)PCT国际申请的申请数据
PCT/US2015/050129 2015.09.15
(87)PCT国际申请的公布数据
WO2016/044233 EN 2016.03.24
(73)专利权人 亿明达股份有限公司
地址 美国加利福尼亚州
(72)发明人 J.布鲁安德 J.F.施莱辛格
(74)专利代理机构 北京市柳沈律师事务所 11105
代理人 张文辉
(51)Int.Cl.
G16B 20/30(2019.01)
G16B 30/00(2019.01)
(56)对比文件
US 2014163900 A1,2014.06.12
EP 2287307 B1,2014.04.09
WO 2013076586 A2,2013.05.30
WO 2014074611 A1,2014.05.15
US 6274317 B1,2001.08.14
WO 2013055817 A1,2013.04.18
US 2012053845 A1,2012.03.01
Sabine Michel等.Interpretation of low-copy-number DNA profile after post-PCR purification.《Forensic Science International: Genetics Supplement Series》.2009,第2卷542-543.
审查员 夏向东

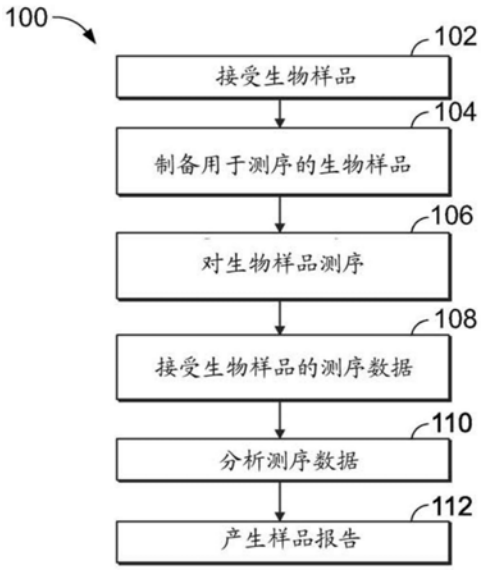
权利要求书2页 说明书40页
序列表5页 附图21页

(54)发明名称

用于分析核酸测序数据的方法和系统

(57)摘要

方法包括接收包含多个样品读段的测序数据,并将样品读段分配到指定基因座,所述样品读段具有相应的核苷酸序列。该方法还包括分析每个指定基因座的分配读段以鉴定分配读段内的相应的感兴趣区域(ROI)。每个ROI具有一个或多个系列的重复基序。该方法还包括基于ROI的序列分选分配读段,使得具有不同序列的ROI归为不同的潜在等位基因。该方法还包括针对具有多个潜在等位基因的指定基因座分析潜在等位基因的序列,以确定潜在等位基因的第一等位基因是否是潜在等位基因的第二等位基因的疑似打滑产物。



CN 107002121 B

1.方法,其包括:

接收包含多个样品读段的测序数据,所述样品读段具有相应的核苷酸序列;

基于所述核苷酸序列将所述样品读段分配到指定基因座,其中分配到相应的指定基因座的所述样品读段是所述相应的指定基因座的分配读段;

分析每个指定基因座的分配读段以鉴定所述分配读段内的相应的感兴趣区域(ROI),每个所述ROI具有一个或多个系列的重复基序,其中相应系列的每个重复基序包含相同的核苷酸集;

基于所述ROI的序列,对具有多个分配读段的指定基因座分选所述分配读段,使得具有不同序列的ROI归为不同的潜在等位基因,每个潜在等位基因具有与所述指定基因座内的其它潜在等位基因的序列不同的序列;和

针对具有多个潜在等位基因的指定基因座分析所述潜在等位基因的序列以确定所述潜在等位基因的第一等位基因是否是所述潜在等位基因的第二等位基因的疑似打滑产物(stutter product),如果已经在所述第一和第二等位基因之间添加或丢失了相应序列内的k个重复基序,其中k是整数,那么所述第一等位基因是所述第二等位基因的疑似打滑产物。

2.权利要求1的方法,其中针对所述具有多个潜在等位基因的指定基因座分析所述潜在等位基因的序列以确定所述第一等位基因是否是所述第二等位基因的疑似打滑产物包括比较所述第一和第二等位基因的ROI的长度以确定所述第一和第二等位基因的ROI的长度是否相差一个重复基序或多个重复基序。

3.权利要求1或权利要求2的方法,其中针对所述具有多个潜在等位基因的指定基因座分析所述潜在等位基因的序列以确定所述第一等位基因是否是所述第二等位基因的疑似打滑产物包括鉴定已经添加或丢失的重复基序,并测定添加的或丢失的重复基序是否与所述相应序列中的相邻重复基序相同。

4.根据权利要求1的方法,其中k等于1或2。

5.根据权利要求1的方法,其中如果在所述第一和第二等位基因的ROI的序列之间不存在其它错配,那么所述第一等位基因是所述第二等位基因的打滑产物。

6.根据权利要求1的方法,其中所述方法还包括产生基因型序型(profile),所述基因型序型调用至少多个所述指定基因座的基因型,其中将具有疑似打滑产物的所述指定基因座指示为具有所述疑似打滑产物。

7.根据权利要求1的方法,其中所述方法还包括提供针对至少多个所述指定基因座的基因型调用物,其中所述基因型调用物中的至少一个指示对于所述至少一个基因型调用物的所述指定基因座存在疑似打滑产物。

8.根据权利要求1的方法,其还包括针对具有多个潜在等位基因的每个指定基因座计数对所述潜在等位基因调用的所述样品读段的总数,其中如果所述第一等位基因的样品读段小于所述第二等位基因的样品读段的指定阈值,那么所述第一等位基因是所述第二等位基因的打滑产物。

9.权利要求8的方法,其中所述指定阈值是所述第二等位基因的样品读段的40%。

10.根据权利要求8的方法,其中如果所述第一等位基因的样品读段超过所述第二等位基因的样品读段的预定百分比,那么将所述疑似打滑产物指定为来自另一贡献者。

11.根据权利要求8的方法,其中如果所述第一等位基因的样品读段小于所述第二等位基因的样品读段的预定百分比,那么将所述疑似打滑产物指定为噪音。

12.根据权利要求1的方法,其中所述分配读段包括具有位于其间的相应重复区段的第一和第二保守侧翼区,其中对于每个分配读段,所述方法还包括:

(a) 提供包含所述第一保守侧翼区和所述第二保守侧翼区的参考序列;

(b) 将所述参考序列的第一侧翼区的一部分与所述相应的分配读段对准;

(c) 将所述参考序列的第二侧翼区的一部分与所述相应的分配读段对准;并且

(d) 测定所述重复区段的长度和/或序列。

13.根据权利要求12的方法,其中所述在步骤(b)和(c)之一或两者中对准所述侧翼区的一部分包括:

(i) 通过使用与所述重复区段重叠或相邻的接种区的精确k聚体匹配来测定所述分配读段上的相应保守侧翼区的位置;并且

(ii) 将所述侧翼区与所述分配读段对准。

14.权利要求13的方法,其中所述接种区包含所述保守侧翼区的高复杂性区,其中所述高复杂性区以所述保守侧翼区中所有核苷酸的至少10%、15%、20%或25%的频率掺入四种核苷酸的每种。

15.权利要求14的方法,其中所述高复杂性区包括与所述重复区段充分不同以避免错误对准的序列。

16.权利要求14的方法,其中所述高复杂性区包含具有多种多样的核苷酸混合物的序列。

17.权利要求13的方法,其中所述接种区避免所述相应的保守侧翼区的低复杂性区,其中所述低复杂性区包含与所述重复区段具有超过30%、40%、50%、60%、70%或超过80%序列同一性的序列。

18.权利要求17的方法,其中所述低复杂性区具有类似于多个所述重复基序的序列。

19.权利要求13的方法,其中所述接种区直接邻近所述重复区段。

20.权利要求13的方法,其中所述接种区包括所述重复区段的一部分。

21.权利要求13的方法,其中所述接种区与所述重复区段偏移。

22.根据权利要求1的方法,其中所述样品读段是具有正向和反向引物序列的PCR扩增子。

23.根据权利要求1或22的方法,其中将所述样品读段分配到所述指定基因座包括鉴定与PCR扩增子的引物序列对应的所述样品读段的序列。

24.根据权利要求1或22的方法,其中所述测序数据来自合成测序(SBS)测定。

25.根据权利要求1或22的方法,其中所述ROI是短串联重复(STR)。

26.权利要求25的方法,其中所述STR选自由以下组成的基因座的组:CSF1PO, FGA, TH01, TPOX, VWA, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, D21S11, PENTA D和PENTA。

用于分析核酸测序数据的方法和系统

[0001] 对相关申请的交叉引用

[0002] 本申请要求于2014年9月18日提交的题为“METHODS AND SYSTEMS FOR ANALYZING NUCLEIC ACID SEQUENCING DATA”的美国临时申请No.62/052,189的权益,其通过引用整体并入本文。

[0003] 发明背景

[0004] 已经鉴定了各种遗传基因座,其可用于区分物种群体(例如人类)内的个体或提供关于群体或群体内的个体的其它有用信息。例如,遗传基因座可以具有许多变体形式,称为等位基因,并且群体中的每个个体可以具有特定基因座的一个或多个等位基因。基因座的等位基因可以在长度(即,核苷酸总数)和/或核苷酸序列上与相同基因座的其它等位基因不同。存在分析遗传基因座的等位基因的各种遗传应用。这些遗传应用包括亲子关系测试、人鉴定(例如,法医分析)、嵌合体监测(例如组织移植监测)和植物和动物研究中的其它遗传应用。许多遗传应用分析包括短串联重复(STR)和/或单核苷酸多态性(SNP)的基因座。STR是包括重复基序的DNA的重复区域。重复基序的长度可以是例如2至6个核苷酸,尽管存在其它大小的重复基序。

[0005] 尽管近年来STR和/或SNP分析已经改善,但仍然存在挑战。例如,STR的分析通常不包括核苷酸的实际序列的分析。STR通常使用毛细管电泳(CE)系统分析。然而,CE系统仅测定等位基因的长度,并且不鉴定等位基因的序列。因此,当事实上个体具有两个具有相同长度但具有不同序列的不同等位基因时,CE数据可能会指示个体对于特定等位基因是纯合的。

[0006] 对分析核酸序列的系统也可能存在质量控制挑战。例如,一些测定法包括制备生物样品,扩增生物样品的STR等位基因,然后测序所得扩增子。在样品的制备和扩增后,可能的是一种或多种扩增子通过引物二聚体形成和/或包括来自超过一种来源(例如嵌合体)的核酸,使得相应的数据不可靠。如果不鉴定和过滤掉不想要的数据,那么可能更难以例如提供来源的准确遗传概况(profile)或鉴定存在有多个来源。如果鉴定出不想要的数据,那么通常将数据滤出并弃去,但不进一步分析。类似地,在测序期间发生的错误也可能使分析更加困难,并且通常弃去此类数据。最后,从未知来源可靠地测定个体的性别也可以是具有挑战性的。

[0007] 因此,需要用于分析测序数据的改进的方法和系统。

[0008] 发明概述

[0009] 在一个实施方案中,提供了方法,其包括:接收包含多个样品读段的测序数据,所述样品读段具有相应的核苷酸序列。方法还包括基于所述核苷酸序列将所述样品读段分配到指定基因座,其中分配到相应的指定基因座的所述样品读段是所述相应的指定基因座的分配读段。方法还包括分析每个指定基因座的分配读段以鉴定所述分配读段内的相应的感兴趣区域(ROI)。每个所述ROI具有一个或多个系列的重复基序,其中相应系列的每个重复基序包含相同的核苷酸集。方法还包括基于所述ROI的序列,对具有多个指定读段的指定基因座分选所述分配读段,使得具有不同序列的ROI归为不同的潜在等位基因。每个潜在等位

基因具有与所述指定基因座内的其它潜在等位基因的序列不同的序列。方法还包括针对具有多个潜在等位基因的指定基因座分析所述潜在等位基因的序列以确定所述潜在等位基因的第一等位基因是否是所述潜在等位基因的第二等位基因的疑似打滑产物(stutter product)。如果已经在所述第一和第二等位基因之间添加或丢失了相应序列内的k个重复基序,其中k是整数,那么所述第一等位基因是所述第二等位基因的疑似打滑产物。任选地,k等于1或2。

[0010] 在一个实施方案中,提供了方法,其包括接收测序数据,所述测序数据具有与遗传基因座集对应的扩增子的多个样品读段。所述样品读段包含读段对,其中相应扩增子的每个读段对包含所述相应扩增子的第一和第二读段。所述第一和第二读段中的每个具有相应的读段序列。方法还包括基于对所述第一读段的读段序列的分析,鉴定所述第一读段的潜在遗传基因座。潜在遗传基因座来自所述遗传基因座集。方法还包括针对具有至少一个潜在基因座的每个所述第一读段,确定所述第一读段是否与每个所述潜在遗传基因座的参考序列对准。如果所述第一读段与仅一个遗传基因座的参考序列对准,那么所述方法包括确定所述第一读段包括所述一个遗传基因座的潜在等位基因。如果所述第一读段与超过一个参考序列对准,那么所述方法包括确定第一读段包括具有与所述第一读段最佳对准的参考序列的遗传基因座的潜在等位基因。

[0011] 如果所述第一读段不与参考序列对准,那么所述方法包括将所述第一读段指定为未对准读段,并分析所述未对准读段以从与所述未对准读段最佳拟合的潜在遗传基因座鉴定遗传基因座。方法还包括产生遗传概况,其包含至少多个所述遗传基因座的调用的基因型,其中所述调用的基因型基于所述相应遗传基因座的潜在等位基因。遗传概况还包含具有未对准读段的遗传基因座的一个或多个通知。

[0012] 在一个实施方案中,提供了方法,其包括接收测序数据,所述测序数据具有与遗传基因座集对应的扩增子的多个样品读段。样品读段包含读段对,其中相应扩增子的每个读段对包含所述相应扩增子的第一和第二读段。第一和第二读段中的每个具有相应的读段序列。方法还包括基于对所述第一读段的读段序列的分析,鉴定所述第一读段的潜在遗传基因座。潜在遗传基因座来自所述遗传基因座集。方法还包括针对具有至少一个潜在基因座的每个所述第一读段,确定所述第一读段是否与每个所述潜在遗传基因座的参考序列对准。方法还包括将不与参考序列对准的所述第一读段指定为未对准读段。方法还包括分析所述未对准读段以从与所述未对准读段最佳拟合的所述潜在遗传基因座鉴定遗传基因座。方法还包括分析未对准读段以测定对于所述最佳拟合遗传基因座是否存在潜在的等位基因退出(dropout)。

[0013] 在一个实施方案中,提供了方法,其包括接收多个遗传基因座的每个遗传基因座的读段分布。读段分布包含多个潜在等位基因,其中每个潜在等位基因具有等位基因序列和读段计数。读段计数表示测定为包括所述潜在等位基因的来自测序数据的样品读段的数目。方法还可以包括针对所述多个遗传基因座的每个遗传基因座,鉴定具有最大读段计数的所述读段分布的潜在等位基因之一。方法还可以包括针对所述多个遗传基因座的每个遗传基因座,测定所述最大读段计数是否超过解读阈值。如果所述最大读段超过所述解读阈值,那么所述方法包括分析对应遗传基因座的潜在等位基因以调用所述遗传基因座的基因型。如果所述最大读段小于所述解读阈值,那么所述方法包括产生所述遗传基因座具有低

覆盖的警报。方法还包括产生遗传概况和具有低覆盖的遗传基因座的警报，所述遗传概况具有调用基因型的每个所述遗传基因座的基因型。

[0014] 在一个实施方案中，提供了方法，其包括：(a)接收遗传基因座的读段分布。读段分布包含多个潜在等位基因，其中每个潜在等位基因具有等位基因序列和计数得分。计数得分基于测定为包含所述潜在等位基因的来自测序数据的样品读段的数目。方法还包括：(b)基于一个或多个所述潜在等位基因的计数得分测定所述遗传基因座是否具有低覆盖。如果所述遗传基因座具有低覆盖，那么所述方法包括产生所述遗传基因座具有低覆盖的通知。如果所述遗传基因座不具有低覆盖，那么所述方法包括分析所述潜在等位基因的计数得分以测定所述遗传基因座的基因型。方法还包括：(d)产生包括遗传基因座的基因型的遗传概况或遗传基因座具有低覆盖的警报。

[0015] 在一个实施方案中，提供了方法，其包括接收遗传基因座的读段分布。读段分布包含多个潜在等位基因，其中每个潜在等位基因具有等位基因序列和读段计数。读段计数表示分配到所述遗传基因座的来自测序数据的序列读段的数目。方法还包括测定每个所述潜在等位基因的计数得分。计数得分可以基于所述潜在等位基因的读段计数。方法还可以包括测定所述潜在等位基因的所述计数得分是否通过分析阈值。如果相应的潜在等位基因的计数得分没有通过所述分析阈值，那么所述方法包括弃去所述相应的潜在等位基因。如果相应的潜在等位基因的计数得分通过所述分析阈值，那么所述方法包括将所述潜在等位基因指定为所述遗传基因座的指定等位基因。

[0016] 在一个实施方案中，提供了方法，其包括接收遗传基因座的读段分布。读段分布包含多个潜在等位基因，其中每个潜在等位基因具有等位基因序列和读段计数。读段计数表示分配到所述遗传基因座的来自测序数据的样品读段的数目。方法还包括测定所述读段的数是否超过分析阈值。如果相应的潜在等位基因的读段计数小于所述分析阈值，那么所述方法包括将所述相应的潜在等位基因指定为噪音等位基因。如果相应的潜在等位基因的读段计数通过所述分析阈值，那么所述方法包括将所述潜在等位基因指定为所述遗传基因座的所述等位基因。方法还包括测定所述噪音等位基因的读段计数的总和是否超过噪音阈值。如果所述总和超过所述噪音阈值，那么所述方法包括产生所述遗传基因座具有过度噪音的警报。

[0017] 在一个实施方案中，提供了方法，其包括接收多个遗传基因座的每个遗传基因座的基因座数据。基因座数据包括相应遗传基因座的一个或多个指定等位基因。每个指定等位基因基于从测序数据获得的读段计数。方法还包括针对所述多个遗传基因座的每个遗传基因座，测定所述相应遗传基因座的指定等位基因的数目是否大于所述相应遗传基因座的可允许等位基因的预定最大数目。方法还可以包括如果指定等位基因的数目超过可允许等位基因的预定最大数目，那么产生等位基因数目警报。方法还包括针对所述多个遗传基因座的每个遗传基因座，测定所述指定等位基因的等位基因比例是否不充足。等位基因比例可以基于所述指定等位基因的读段计数。方法还可以包括如果等位基因比例不平衡，那么产生等位基因比例警报。方法还可以包括基于遗传基因座集的等位基因数目警报和等位基因比例警报的数目，确定所述样品包含多个来源的混合物。

[0018] 在一个实施方案中，提供了方法，其包括接收多个Y基因座的基因座数据。基因座数据包括所述Y基因座的指定等位基因。每个指定等位基因基于从测序数据获得的读段计

数。方法还包括将每个Y-基因座的指定等位基因的数目与所述Y-基因座的等位基因的预期数目进行比较。方法还包括基于来自所述比较操作的结果产生所述样品是男性或女性的预测。任选地，遗传基因座包括短串联重复(STR)基因座和单核苷酸多态性(SNP)基因座。

[0019] 附图简述

[0020] 图1是显示根据一个实施方案的方法的流程图。

[0021] 图2是显示为不同分析指定不同类型的样品读段的方法的流程图。

[0022] 图3是显示图2的方法的一部分的示意图。

[0023] 图4是显示根据一个实施方案可以鉴定感兴趣区域(ROI)的示意图。

[0024] 图5是显示如果直接邻近(immediately adjacent to)STR的侧翼区用于接种(seed)对准(alignment)则可以发生的各种错误对准的示意图。

[0025] 图6A是显示基于来自样品混合物的样品输入与理论结果相比实际STR调用(calling)的一组图。

[0026] 图6B是显示基于来自样品混合物的样品输入与理论结果相比实际STR调用的另一组图。

[0027] 图6C是显示基于来自样品混合物的样品输入与理论结果相比实际STR调用的另一组图。

[0028] 图6D是显示基于来自样品混合物的样品输入与理论结果相比实际STR调用的另一组图。

[0029] 图7是显示五个对照DNA样品的已知基因座的等位基因调用的一致性(concordance)的表。

[0030] 图8是显示根据一个实施方案的鉴定样品读段内的打滑产物的方法的流程图。

[0031] 图9包括显示D1S1656基因座的潜在等位基因的读段的表。

[0032] 图10包括基于在图9的表中找到的数据的图。

[0033] 图11是显示分析样品读段以测定一个或多个遗传基因座的基因型的方法的流程图。

[0034] 图12是显示了产生包括多个基因型调用物(genotype call)的样品报告的方法的流程图。

[0035] 图13是显示检测样品是否包括来源的混合物的方法的流程图。

[0036] 图14是显示测定样品的性别的方法的流程图。

[0037] 图15显示了根据可以用于执行本文中阐述的各种方法的实施方案形成的系统。

[0038] 图16A显示了根据一个或多个实施方案的样品报告的一部分。

[0039] 图16B显示了根据一个或多个实施方案的样品报告的另一部分。

[0040] 图17A显示了根据一个或多个实施方案的样品报告的一部分。

[0041] 图17B显示了样品报告的另一部分。

[0042] 图17C显示了样品报告的另一部分。

[0043] 图17D显示了样品报告的另一部分。

[0044] 图17E显示了样品报告的另一部分。

[0045] 图17F显示了样品报告的另一部分。

[0046] 发明详述

[0047] 本申请包括与2013年3月15日提交的题为“METHODS AND SYSTEMS FOR ALIGNING REPETITIVE DNA ELEMENTS”的国际申请No.PCT/US2013/030867(公开号WO 2014/142831)中描述的主题类似的主题,其通过引用整体并入本文。

[0048] 本文中阐述的实施方案可适用于分析核酸序列以鉴定序列变异。实施方案可用于分析遗传基因座的潜在等位基因并测定遗传基因座的基因型,或换言之,提供基因座的基因型调用物。在一些情况下,本文中阐述的方法和系统可产生样品报告或遗传概况,其包含多个此类基因型调用物。实施方案还可适用于监测包括核酸序列(例如包括序列变异的那些)的测序和/或分析的测定法的质量。序列变异可以包括单核苷酸多态性(SNP)或多态性重复元件,如短串联重复(STR)。序列变异可以位于指定的遗传基因座内,如在组合DNA指数系统(Combined DNA Index System,CODIS)数据库中发现的或以其它方式用于遗传分析的遗传基因座。例如,序列变异可以包括选自下组的STR:CODIS常染色体STR基因座、CODIS Y-STR基因座、EU常染色体STR基因座、EU Y-STR基因座等。CODIS是由FBI实验室鉴定的核心STR基因座集,并且包括13种基因座:CSF1PO、FGA、TH01、TPOX、VWA、D3S1358、D5S818、D7S820、D8S1179、D13S317、D16S539、D18S51和D21S11。感兴趣的其它STR可以包括PENTA D和PENTA E,然而,可以通过本文中阐述的实施方案分析其它STR。SNP可以在已知的数据库内,如国家生物技术信息中心(NCBI)dbSNP数据库。也可以在未来的研究中鉴定STR和SNP。

[0049] 如本文中使用,术语“序列”包括或表示彼此偶联的核苷酸链。核苷酸可以基于DNA或RNA。应当理解,一个序列可以包括多个子序列。例如,单个样品读段(例如,PCR扩增子的单个样品读段)可以具有含有350个核苷酸的序列。样品读段可以包含这350个核苷酸内的多个子序列。例如,样品读段可以包括具有例如20-50个核苷酸的第一和第二侧翼子序列。第一和第二侧翼子序列可以位于具有相应子序列(例如,40-100个核苷酸)的重复区段的任一侧。每个侧翼子序列可以包括引物子序列(例如10-30个核苷酸)(或包括其部分)。为了便于阅读,术语“子序列”将被称为“序列”,但是应当理解,两个序列不必在共同链上彼此分开。为了区分本文中所述的各种序列,可以给所述序列不同的标记(例如,靶序列、引物序列、侧翼序列、参考序列等)。可以对其它术语(如“等位基因”)给予不同的标记以区分相似的对象。

[0050] 如本文中使用,术语“感兴趣区域”或“ROI”包括样品读段的包括一个或多个系列重复基序的重复区段。重复基序系列可以是STR。在一些实施方案中,ROI仅是重复区段(例如,STR)。然而,在其它实施方案中,ROI可以包括侧翼区的子序列。例如,ROI可以包括重复区段、从重复区段的一端延伸的约1-5个核苷酸的第一侧翼区、和从重复区段的相反端延伸的约1-5个核苷酸的第二侧翼区。

[0051] 应当理解,不需要重复区段全部具有相同的基序。重复区段可以包括一系列X基序,然后一系列Y基序,然后一系列Z基序(或另一系列X基序)等。[TAGA]11[TAGG]1[TG]5的重复区段是上述的一个具体例子。还应当理解,不需要重复片段全部具有重复基序。如上述例子中显示,重复区段可以包括由非重复基序中断的重复基序。上述例子中的[TAGG]是一种此类非重复基序。

[0052] 如本文中使用的,术语“阈值”指示可以改变分析过程的点和/或可以触发动作的点。不需要阈值为预定数目。相反,阈值可以是例如基于多个因素的函数。换言之,阈值可以适应环境。举例而言,当测定多个样品读段是否构成应当弃去的噪音或应当进一步分析的

数据时，阈值可以是设定数目(例如，10个样品读段)或基于不同因素的函数，如相应遗传基因座的总读段数目和遗传基因座的历史知识。此外，阈值可以指示上限、下限或限制之间的范围。可以触发的动作可以包括例如通知最终用户：样品疑似包括打滑产物，样品含有来源的混合物，测定法具有特定问题区域，样品是质量差的，等等。

[0053] 在一些实施方案中，可以将基于测序数据的度量或得分与阈值进行比较。如本文中使用的，术语“度量”或“得分”可包括从测序数据测定的值或结果，或可包括基于从测序数据测定的值或结果的功能。像阈值一样，度量或得分可以适应环境。例如，度量或得分可以是标准化值。

[0054] 作为得分或度量的例子，一个或多个实施方案可以在分析数据时使用计数得分。计数得分可以基于样品读段的数目。样品读段可能已经经历了一个或多个过滤阶段，使得样品读段具有至少一个共同的特性或质量。例如，用于测定计数得分的每个样品读段可能已与参考序列对准或可归为潜在等位基因。可以对具有共同特性的样品读段的数目进行计数以测定读段计数。计数得分可以基于读段计数。在一些实施方案中，计数得分可以是等于读段计数的值。在其它实施方案中，计数得分可以基于读段计数和其它信息。例如，计数得分可以基于遗传基因座的特定等位基因的读段计数和遗传基因座的读段总数。在一些实施方案中，计数得分可以基于读段计数和遗传基因座的先前获得的数据。在一些实施方案中，计数得分可以是预定值之间的标准化得分。计数得分还可以是来自样品的其它基因座的读段计数的函数或来自与感兴趣样品同时运行的其它样品的读段计数的函数。例如，计数得分可以是特定等位基因的读段计数和样品中其它基因座的读段计数和/或来自其它样品的读段计数的函数。作为一个例子，来自其它基因座的读段计数和/或来自其它样品的读段计数可用于标准化特定等位基因的计数得分。

[0055] 通常从测序数据测定读段计数。读段计数可以是例如已经测定为具有包括ROI的相同ROI的样品读段的数目。读段计数(例如，350个样品读段)可以用于计算然后与指定阈值进行比较的打滑度量。例如，可以通过将读段计数乘以基于历史知识、样品知识、基因座知识等的指定因素来测定打滑度量。打滑度量可以是读段计数的标准化值。

[0056] 当结合附图阅读时，将更好地理解各种实施方案的上述和以下详细描述。就附图显示各种实施方案的功能块的图而言，功能块不一定指示硬件电路之间的划分。因此，例如，可以在单片硬件(例如，通用信号处理器或随机存取存储器，硬盘等的块)或多片硬件中执行功能块(例如，模块、处理器或存储器)中的一种或多种。类似地，程序可以是单机程序，可以作为子例程并入操作系统中，可以是所安装的软件包中的功能等。应当理解，各种实施方案不限于附图中显示的排列(arrangements)和手段(instrumentality)。

[0057] 本申请描述了各种方法和用于执行方法的系统。在图中将至少一些方法显示为多个步骤。然而，应当理解，实施方案不限于图中显示的步骤。可以省略步骤，可以修改步骤，和/或可以添加其它步骤。举例而言，尽管本文中所述的一些实施方案可以包括制备样品并对其测序以获得测序数据，但是其它实施方案可以包括直接接收测序数据，而不制备样品和/或对样品测序。此外，可以组合本文中所述的步骤，可以同时执行步骤，可以并行执行步骤，可以将步骤分成多个子步骤，可以以不同的顺序执行步骤，或者可以以迭代方式重新执行步骤(或一系列步骤)。此外，尽管在本文中阐述不同的方法，但是应当理解，可以在其它实施方案中组合不同的方法(或不同方法的步骤)。

[0058] 图1显示了根据一个实施方案的方法100。方法100包括在102处接收包括或疑似包括核酸(如DNA)的生物样品。生物样品可以来自已知或未知来源,如动物(例如人)、植物、细菌或真菌。可以直接从来源采集生物样品。例如,可以直接从个体采集血液或唾液。或者,可以不直接从来源获得样品。例如,生物样品可以从犯罪现场获得,从挖掘中,或正在研究的其它区域(例如,历史场所)保留。如本文中使用的,术语"生物样品"包括生物样品具有来自不同来源的多个生物样品的可能性。例如,通过犯罪现场获得的生物样品可以包括来自不同个体的DNA的混合物。

[0059] 方法100还可包括在104处制备用于测序的样品。制备104可包括除去外来物质和/或分离某些物质(例如DNA)。可以制备生物样品以包括特定测定所需的特征。例如,可以制备生物样品以用于合成测序(SBS)。在某些实施方案中,制备可以包括扩增基因组的某些区域。例如,在104处的制备可以包括扩增已知包含STR和/或SNP的预定遗传基因座。可以使用预定的引物序列扩增遗传基因座。

[0060] 在106处,可以对样品进行测序。可以通过多种已知的测序方案进行测序。在具体实施方案中,测序包括SBS。在SBS中,多个荧光标记的核苷酸用于对存在于光学基底的表面(例如,至少部分限定流中的通道的表面)上的多个扩增DNA簇(可能数百万个簇)进行测序。流动池可含有用于测序的核酸样品,其中流动池置于合适的流动池保持器内。用于测序的样品可以采取彼此分离的单个核酸分子的形式以便个别可分辨,簇或其它特征形式的核酸分子的扩增群体,或附着到一种或多种核酸分子的珠。

[0061] 可以制备核酸,使得它们包含与未知靶序列相邻的已知寡核苷酸引物,其可以称为引物序列。为了启动第一SBS测序循环,可以通过流体流子系统(未显示)将一种或多种不同标记的核苷酸和DNA聚合酶等流过/通过流动池。可以一次添加单一类型的核苷酸,或者可以特殊设计用于测序程序的核苷酸以拥有可逆的终止性质,从而允许在存在几种类型的标记的核苷酸(例如A、C、T、G)的情况下同时发生测序反应的每个循环。核苷酸可以包括可检测的标记物部分,如荧光团。当将四个核苷酸混合在一起时,聚合酶能够选择正确的碱基并且将每个序列延伸单个碱基。可以通过使清洗溶液流过流动池洗去非掺入的核苷酸。一个或多个激光器可以激发核酸并诱导荧光。从核酸发射的荧光基于掺入的碱基的荧光团,并且不同的荧光团可以发射不同波长的发射光。可以将去封闭试剂添加到流动池,以从延伸和检测的DNA链中除去可逆终止剂基团。然后,可以通过使清洗溶液流过流动池洗去去封闭试剂。然后,流动池准备好进行进一步的测序循环,以引入标记的核苷酸开始,如上文阐述。可以将流体和检测步骤重复几次以完成测序运行。示例性测序方法描述于Bentley et al.,Nature 456:53-59(2008),国际公开号WO 04/018497;美国专利号7,057,026;国际公开号WO 91/06678;国际公开号WO 07/123744;美国专利号7,329,492;美国专利号7,211,414;美国专利号7,315,019;美国专利号7,405,281,以及美国公开号2008/0108082,其各自通过引用并入本文。

[0062] 在一些实施方案中,可以在测序之前或期间将核酸附着于表面并扩增。例如,可以使用桥式扩增进行扩增以在表面上形成核酸簇。有用的桥式扩增方法描述于例如美国专利号5,641,658;美国专利公开号2002/0055100;美国专利号7,115,400;美国专利公开号2004/009685;美国专利公开号2004/0002090;美国专利公开号2007/0128624;和美国专利公开号2008/0009420,其全部内容各自通过引用并入本文。用于扩增表面上的核酸的另一

种有用的方法是滚环扩增(RCA),例如,如Lizardi et al.,Nat.Genet.19:225-232(1998)和美国专利公开号2007/0099208A1中描述,其各自通过引用并入本文。

[0063] 特别有用的SBS方案利用具有可除去的3'区块(3'blocks)的经修饰的核苷酸,例如,如记载于国际公开号WO 04/018497、美国专利公开号2007/0166705A1和美国专利号7,057,026,其各自通过引用并入本文。例如,作为桥式扩增方案的结果,可以将SBS试剂的重复循环递送至具有对其附着的靶核酸的流动池。可以使用线性化溶液将核酸簇转化为单链形式。线性化溶液可以含有例如能够切割每个簇的一条链的限制性内切核酸酶。其它切割方法可以用作限制酶或切口酶的备选,特别包括化学切割(例如用高碘酸盐切割二醇连接)、通过用内切核酸酶切割来切割无碱基位点(例如"USER",如由NEB,Ipswich,MA,USA提供,部件号M5505S),通过暴露于热或碱,切割掺入以其它方式由脱氧核糖核苷酸构成的扩增产物中的核糖核苷酸,光化学切割或肽接头的切割。在线性化步骤之后,可以在用于将测序引物与待测序的靶核酸杂交的条件下将测序引物递送至流动池。

[0064] 然后,可以在通过单核苷酸添加延伸与每个靶核酸杂交的引物的条件下使流动池与具有带有可除去的3'区块和荧光标记物的经修饰的核苷酸的SBS延伸试剂接触。仅将单核苷酸添加到每个引物,因为一旦已经将经修饰的核苷酸掺入与测序的模板区域互补的生长的多核苷酸链中,没有游离的3'-OH基团可用于指导进一步的序列延伸,因此聚合酶不能添加进一步的核苷酸。可以除去SBS延伸试剂,并用含有在照射激发下保护样品的组分的扫描试剂替换。用于扫描试剂的示例性组分描述于美国公开US 2008/0280773A1和美国流水号13/018,255,每篇通过引用并入本文。然后,可以在扫描试剂的存在下荧光检测延伸的核酸。一旦已经检测到荧光,可以使用适合于所使用的封闭基团的去封闭试剂除去3'区块。可用于各自封闭基团的示例性去封闭试剂描述于WO04018497、US 2007/0166705A1和US7057026中,其各自通过引用并入本文。可以洗去去封闭试剂,留下与具有3'OH基团的延伸引物杂交的靶核酸,所述引物现在能够添加另外的核苷酸。因此,可以重复添加延伸试剂、扫描试剂和去封闭试剂的循环,以及在一个或多个步骤之间任选的清洗,直到获得期望的序列。当每个经修饰的核苷酸具有已知对应于特定碱基的与其附着的不同标记物时,可以使用每个循环的单个延伸试剂递送步骤进行上述循环。不同的标记物有利于区分每个掺入步骤期间添加的核苷酸。或者,每个循环可以包括延伸试剂递送的分开的步骤,然后是扫描试剂递送和检测的分开的步骤,在这种情况下,两种或更多种核苷酸可以具有相同的标记,并且可以基于已知的递送顺序来区分。

[0065] 继续流动池中核酸簇的实例,可以进一步处理核酸以在称为"配对末端测序"的方法中从相对端获得第二读段。配对末端测序允许用户对目标片段的两端进行测序。配对末端测序可以促进基因组重排和重复区段,以及基因融合和新转录物的检测。配对末端测序的方法描述于PCT公开WO07010252、PCT申请流水号PCTGB2007/003798和美国专利申请公开US 2009/0088327,其各自通过引用并入本文。在一个实例中,可以如下实施一系列步骤:(a)产生核酸簇;(b)使核酸线性化;(c)杂交第一测序引物并且实施延伸、扫描和去封闭的重复循环;(d)通过合成互补拷贝"转化"流动池表面上的靶核酸;(e)使再合成链线性化;并且(f)杂交第二测序引物并实施延伸、扫描和去封闭的重复循环,如上文阐述。可以通过递送如上文对桥式扩增的单个循环阐述的试剂实施转化步骤。

[0066] 虽然上文就特定的SBS方案而言例示了在106处的测序操作,但是应当理解,可以

根据需要实施用于对多种其它分子分析中的任一种进行测序的其它方案。例如，也可以使用珠上的乳液PCR，例如如Dressman et al.,Proc.Natl.Acad.Sci.USA 100:8817-8822(2003)、WO 05/010145、或美国专利公开号2005/0130173或2005/0064460中描述，其各自通过引用整体并入本文。适用于本文中阐述的方法和系统的用途的其它测序技术是焦磷酸测序、纳米孔测序和通过连接的测序。示例性焦磷酸测序技术和特别有用的样品描述于US 6,210,891;US 6,258,568;US 6,274,320以及Ronaghi,Genome Research 11:3-11(2001)，其各自通过引用并入本文。示例性的纳米孔技术和也有用的样品描述于Deamer et al.,Acc.Chem.Res.35:817-825(2002);Li et al.,Nat.Mater.2:611-615(2003);Soni et al.,Clin Chem.53:1996-2001(2007)Healy et al.,Nanomed.2:459-481(2007)和Cockroft et al.,J.am.Chem.Soc.130:818-820;以及US 7,001,792，其各自通过引用并入本文。特别地，这些方法利用试剂递送的重复步骤。本文中阐述的仪器或方法可以配置有储器、阀、流体线和其它流体组件以及那些部件的控制系统，以便根据期望的方案，如上文引用的参考文献中阐述的那些方案引入试剂和检测光信号。多种样品中的任一种可用于这些系统中，如具有通过乳液PCR产生的珠的基底、具有零模式波导的基底、具有集成CMOS检测器的基底、在脂质双层中具有生物纳米孔的基底、具有合成纳米孔的固态基底、和本领域已知的其它基底。在上文引用的参考文献中和进一步在US 2005/0042648;US 2005/0079510;US 2005/0130173;和WO 05/010145(其各自通过引用并入本文)中的各种测序技术的背景下描述此类样品。

[0067] 能够实施上文描述的一种或多种SBS方案的系统包括由Illumina,Inc.开发的系统，诸如MiSeq、HiSeq 2500、HiSeq X Ten、NeoPrep、HiScan和iScan系统。能够实施上文描述的一种或多种SBS方案的系统描述于美国申请号13/273,666和13/905,633，其各自通过引用整体并入本文。

[0068] 在108处，可以接收测序数据用于随后在110处的分析。测序数据可以包括例如多个样品读段。每个样品读段可以包括核苷酸序列，其可以称为样品序列或靶序列。样品序列可以包括例如引物序列、侧翼序列和靶序列。样品序列内的核苷酸数目可以包括30、40、50、60、70、80、90、100或更多个。在一些实施方案中，一个或多个样品读段(或样品序列)包括至少150个核苷酸、200个核苷酸、300个核苷酸、400个核苷酸、500个核苷酸或更多。在一些实施方案中，样品读段可以包括超过1000个核苷酸、2000个核苷酸或更多。样品读段(或样品序列)可以在一端或两端包括引物序列。在某些实施方案中，每个样品读段可以与沿着模板的相反方向上的另一个读段相关联。例如，在106，测序可以包括配对末端测序，其中进行第一读段(读段1)，接着是在相反方向上的第二读段(读段2)。第一和第二读段中的每个可以包括靶序列的全部或几乎整个靶序列。然而，在其它实施方案中，可以使用“不对称”配对末端测序，其中第二读段仅包括可以获得的事物的一部分。例如，第二读段可以仅包括有限数目的核苷酸以确认位于接近用于第二读段的序列的开始的引物序列的鉴定。举例来说，第一读段可以包括300-500个核苷酸，但是第二读段可以仅包括20-50个核苷酸。

[0069] 下面更详细地描述了在110处的分析。在110处的分析可以包括单个方案或以指定方式分析样品读段以获得期望信息的方案的组合。在110处的分析的非限制性实例可以包括分析样品读段以将样品读段分配到某些遗传基因座(或指定所述某些遗传基因座的样品读段)；分析样品读段以鉴定样品读段的长度和/或序列；分析样品读段以分选与某个基因

座的靶等位基因相关联的ROI;分析不同靶等位基因的样品读段(或ROI),以测定一种靶等位基因的ROI是否是另一种靶等位基因的ROI的疑似打滑产物;鉴定遗传基因座的基因型;和/或监测测定法的良好或质量控制。

[0070] 方法100还可以包括在112处产生或提供样品报告。样品报告可以包括例如关于就样品而言的多个遗传基因座的信息。例如,对于预定的遗传基因座集的每个遗传基因座,样品报告可以下列至少一项:提供基因型调用物;指示不能产生基因型调用物;提供关于基因型调用物的确定性的置信得分;或指示关于一种或多种遗传基因座的测定法的潜在问题。样品报告还可以指示提供样品和/或指示样品包括多个来源的个体的性别。如本文中使用的,"样品报告"可以包括遗传基因座或预定的遗传基因座集的数目数据(例如,数据文件)和/或遗传基因座或遗传基因座集的印刷报告。因此,在112处的产生或提供可以包括创建数据文件和/或打印样品报告,或显示样品报告。

[0071] 图2是显示分析具有序列变异的样品读段的测序数据的方法150的流程图。下面参照图3描述图2,图3进一步显示了图1的不同步骤。方法150包括在152处接收来自一个或多个来源的测序数据。测序数据可以包括具有核苷酸的相应样品序列的多个样品读段。图3显示了样品读段180的实例。术语"鉴定序列"和"序列变异"表示样品序列的部分。虽然仅显示了一个样品读段180,但是应当理解,测序数据可以包括例如数百、数千、数十万或数百万个样品个读段。不同的样品读段可以具有不同数目的核苷酸。例如,样品读段的范围可以在10个核苷酸至约500个核苷酸或更多之间。然而,在其它实施方案中,样品读段可以包括更多的核苷酸。样品读段可以跨越来源的整个基因组。在具体实施方案中,样品读段针对预定的遗传基因座,如具有疑似STR或疑似SNP的那些遗传基因座。可以基于与感兴趣的遗传基因座相关的已知引物序列来选择样品读段。例如,样品读段可以包括使用与感兴趣的遗传基因座相关的引物序列获得的PCR扩增子。

[0072] 在154处,可以将每个样品读段分配到相应的遗传基因座。基于样品读段的核苷酸序列,或者换言之,样品读段内的核苷酸(例如,A、C、G、T)的顺序,可以将样品读段分配到相应的遗传基因座。基于该分析,样品读段可以指定为包括特定遗传基因座的可能等位基因。样品读段可以与已经指定为包括遗传基因座的可能等位基因的其它样品读段一起收集(或聚集或框并(binned))。不同的遗传基因座在图3中表示为箱(bin)182。遗传基因座可以是用于特定测定法的预定的遗传基因座集。例如,联邦调查局已经测定了十三(13)种STR位点,其可用于产生犯罪中可能的嫌疑人的遗传概况。使用FBI标准作为实例,如果可能,方法150可以将每个样品读段分配到13种箱之一。

[0073] 不同箱的样品读段可以随后经历不同的分析。例如,可以将样品读段分配到包括STR的遗传基因座。此类基因座可以称为STR基因座。然而,可以将样品读段分配到包括SNP的遗传基因座。此类基因座可以称为SNP基因座。对于典型的样品读段,仅将样品读段分配到一种遗传基因座(或一种箱)。在这些情况下,样品读段然后将经历针对遗传基因座类型配置的分析。更具体地,分配到STR基因座的样品读段将经历STR分析,而分配到SNP基因座的样品读段将经历SNP分析。然而,在一些情况下,可能将样品读段分配到超过一种遗传基因座,因此,样品读段可能经历超过一种类型的分析。

[0074] 在154处的分配操作也可以称为基因座调用,其中将样品读段鉴定为可能与特定遗传基因座相关联。可以分析样品读段以定位区分样品读段与其它样品读段的核苷酸的一

种或多种鉴定序列(例如,引物序列)。更具体地,鉴定序列可以将样品读段与其它样品读段鉴定为与特定遗传基因座相关联。鉴定序列可以包括样品读段的任一端或位于样品读段的任一端附近(例如,在10-30个核苷酸内)。在具体实施方案中,样品读段的鉴定序列基于用于从一种或多种来源选择性地扩增序列的引物序列。然而,在其它实施方案中,鉴定序列可以不位于样品读段的末端附近。

[0075] 在一些实施方案中,将鉴定序列与多个预定序列进行比较,以测定任何鉴定序列是否与预定序列之一相同或几乎相同。例如,每个鉴定序列可以与数据库184内的预定序列的列表(例如,查找表)进行比较。预定序列可以与某些遗传基因座相关联。以下将数据库的预定序列称为选择序列。每个选择序列代表核苷酸序列。如果鉴定序列与任何选择序列有效匹配,那么可以将具有鉴定序列的读段样品分配到与选择序列相关联的遗传基因座。有可能的是,样品读段有效地匹配超过一种选择序列。在这种情况下,可以将样品读段分配到那些选择序列的每个遗传基因座,并进行进一步分析以测定应当调用样品读段的哪一种遗传基因座。

[0076] 可以存在有分析期间使用的预定数目的选择序列。例如,由本文中阐述的实施方案产生的遗传概况可包括约5至约300种遗传基因座的分析。在具体实施方案中,遗传基因座的数目可以是约5至约100种遗传基因座,或更具体地,约10至约30种遗传基因座。然而,可以使用其它数目的遗传基因座。每种遗传基因座可以具有与遗传基因座相关的有限数目的选择序列。利用有限数目的遗传基因座和与每种遗传基因座相关的有限数目的选择序列,可以在不过度使用计算资源的情况下对遗传基因座调用样品读段。在一些实施方案中,选择序列基于用于选择性扩增预定DNA序列的引物序列。

[0077] 尽管每种选择序列可以基于遗传基因座的鉴定序列(例如引物序列),但选择序列可以不包括鉴定序列的每种核苷酸。作为实例,选择序列可以包括样品读段之一的鉴定序列的n个核苷酸的系列。在具体实施方案中,选择序列可以包括鉴定序列的前n个核苷酸。数目n可足以将一种遗传基因座的等位基因与另一种靶基因座的等位基因区分开。在一些实施方案中,数目n在10和30之间。

[0078] 在154处的分配操作可以包括分析鉴定序列的n个核苷酸的系列以测定鉴定序列的n个核苷酸系列是否与一种或多种选择序列有效地匹配。在具体实施方案中,在154处的分配操作可以包括分析样品序列的前n个核苷酸以测定样品序列的前n个核苷酸是否与一种或多种选择序列有效地匹配。数目n可以具有多种值,其可以编程到方案中或由用户输入。例如,数目n可以定义为数据库内最短选择序列的核苷酸数目。数目n可以是预定数目。预定数目可以是例如10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29、或30个核苷酸。然而,在其它实施方案中可以使用更少或更多的核苷酸。数目n也可以由个人(诸如系统的用户)选择。数目n可以基于一种或多种条件。例如,数目n可以定义为数据库内最短引物序列的核苷酸数目或指定数目,其甚至是更小的数目。在一些实施方案中,可以使用n的最小值,如15,使得小于15个核苷酸的任何引物序列可以称为例外。

[0079] 在一些情况下,鉴定序列的n个核苷酸的系列可能不精确匹配选择序列的核苷酸。尽管如此,如果鉴定序列与选择序列几乎相同,那么鉴定序列可以有效地匹配选择序列。例如,如果鉴定序列的n个核苷酸的系列(例如,前n个核苷酸)与具有不超过指定数目的错配(例如,3)和/或指定数目的移位(例如,2)的选择序列匹配,那么可以对遗传基因座调用样

品读段。可以建立规则，使得每个错配或移位可以计数为样品读段和引物序列之间的差异。如果差异数目小于指定数目，那么可以对相应的遗传基因座调用样品读段(即，归入相应的遗传基因座)。在一些实施方案中，可以测定匹配得分，其基于读段样品的鉴定序列和与遗传基因座相关的选择序列之间的差异的数目。如果匹配得分通过指定的匹配阈值，那么对应于选择序列的遗传基因座可以指定为样品读段的潜在基因座。在一些实施方案中，可以进行后续分析以测定是否对遗传基因座调用样品读段。

[0080] 鉴定序列和选择序列之间的差异的指定数目可以是例如小于相应选择序列内核苷酸总数的20％的数目，或者更具体地，小于相应选择序列内核苷酸总数的15％的数目。差异的指定数目可以是预定值，如6、5、4、3或2。因此，短语“有效匹配”包括具有n个核苷酸的系列的样品序列，其与选择序列精确匹配，或者与在选择序列和n个核苷酸的系列之间具有有限数目的差异的选择序列几乎匹配。

[0081] 如果样品读段有效地匹配数据库中的选择序列之一(即，如上所述完全匹配或几乎匹配)，那么将样品读段分配或指定到与选择序列相关的遗传基因座。这可以称为基因座调用或临时基因座调用，其中对与选择序列相关的遗传基因座调用样品读段。然而，如上文讨论，可以对超过一种遗传基因座调用样品读段。在此类实施方案中，可以进行进一步分析以仅对潜在遗传基因座之一调用或分配样品读段。

[0082] 在一些实施方案中，与数据库比较的样品读段是来自配对末端测序的第一读段。对于更具体的实施方案，可以分析与样品读段相关的第二读段，以确认第二读段内的鉴定序列有效地匹配来自数据库的选择序列。用于第二读段的数据库中的选择序列可以不同于用于第一读段的选择序列。在一些实施方案中，仅在确认第二读段也与数据库中的选择序列有效匹配后才对遗传基因座调用样品读段。测定第二读段是否有效地匹配选择序列可以以与上文描述类似的方式进行。通过确认第二读段有效地匹配选择序列，可以从进一步分析中过滤脱靶样品读段(例如，脱靶扩增子)。

[0083] 已经为特定遗传基因座调用的样品读段可以称为特定遗传基因座的“分配读段”。在此阶段时，虽然已经将分配读段鉴定为可能与特定遗传基因座相关，但是分配读段不会适合于进一步分析。更具体地，随后，可以基于其它因素从进一步分析中除去分配读段(或多个读段)。

[0084] 在154处将分配读段分配到相应的遗传基因座后，可以进一步分析样品读段。用分配读段进行的随后的分析可以基于已经为分配读段调用的遗传基因座的类型。例如，如果遗传基因座就包括SNP而言是已知的，那么在156处已经为遗传基因座调用的分配读段可以经历分析，以鉴定分配读段的SNP。如果遗传基因座就包括多态性重复DNA元件而言是已知的，那么可以在158处分析分配读段，以鉴定或表征样品读段内的多态性重复DNA元件。在一些实施方案中，如果分配读段有效地与STR基因座和SNP基因座匹配，那么可以将警告或标志(flag)分配到样品读段。样品读段可以指定为经历例如在156的分析和在158的分析的STR基因座和SNP基因座两者。

[0085] 在一些实施方案中，可以使用下面就图4-7而言描述的方案执行STR分析。在158处的分析可以包括分析样品读段以鉴定ROI，这可以包括测定ROI的序列和/或ROI的长度。ROI可以是样品读段的序列(例如，样品序列的子序列)。ROI可以包括重复区段。ROI可以是仅包括一个或多个系列的重复基序(即，重复区段)的核苷酸序列，或者除了指定数目的从重复

区段的一端或两端延伸的核苷酸之外还包括一个或多个系列的重复基序。更具体地，每个ROI可以包括一个或多个系列的重复基序，其中每个重复基序包括核苷酸的相同核苷酸(例如，2、3、4、5、6个核苷酸或更多)集。常用的重复基序包括四核苷酸，但可使用其它基序，如单核苷酸、二核苷酸、三核苷酸、五核苷酸或六核苷酸。在具体实施方案中，重复基序包括四核苷酸。

[0086] 在158处的分析可以包括分析每个指定基因座的分配读段，以鉴定分配读段内的相应ROI。更具体地，可以测定ROI的长度和/或序列。在158处的分析可以包括根据对准方案对准分配读段以测定分配读段的序列和/或长度。对准方案可以包括在2013年3月15日提交的国际申请号PCT/US2013/030867(公开号WO 2014/142831)中描述的方法，其通过引用整体并入本文。

[0087] 然而，可以使用其它对准方案。例如，一个已知的对准方案将样品读段与参考序列对准。另一种现有方法将样品读段与参考梯对准。在该实例中，通过构建所有已知STR等位基因的梯并将读段与参考基因组对准来创建“参考基因组”，如通常用NGS全基因组序列数据或非重复DNA区域的靶向测序所完成的。可以与本文中阐述的实施方案一起使用的另一种方法被称为lobSTR。在没有STR的先验知识的情况下，lobSTR方法从单个样品的测序数据从头感测，然后调用所有现有的STR(参见Gymrek et al.2012Genome Research 22:1154-62)，其通过引用整体并入本文。

[0088] 现在针对包括ROI的遗传基因座描述国际申请号PCT/US2013/030867(公开号WO 2014/142831)中阐述的对准方法。为了便于阅读，此类遗传基因座可以称为STR基因座。在一些实施方案中，STR基因座的保守侧翼用于有效地测定重复区段的序列。在154处将样品读段到相应的STR基因座之后，实施方案可以在相应的重复片段的每侧上对准侧翼序列的部分以测定重复区段的长度和序列。可以使用k聚体策略接种对准。种子区可以例如在侧翼序列的选择的高复杂性区中，接近重复区段，但避免具有与重复区段具有同源性的低复杂性序列。此类方法可以避免接近重复区段的低复杂性侧翼序列的错误对准(misalignment)。

[0089] 实施方案可以利用STR自身侧翼中的已知序列，其先前已经基于人群中已知的现有变异定义。有利地，进行短跨距的侧翼区的对准在计算上比其它方法更快。例如，整个读段的动态编程对准(Smith-Waterman类型)可以是CPU密集型的、耗时的，特别是在要将多个样品序列对准的情况下。此外，对准整个序列(对于该序列，参考可能甚至不存在)花费的时间占用了宝贵的计算资源。

[0090] 实施方案可利用侧翼序列的现有知识来确保STR等位基因的正确调用。相比之下，依赖于每种等位基因的完整参考序列的现有方法在存在不完全参考的情况下面临显著的失败率。存在许多不知道序列的等位基因和可能一些尚未知的等位基因。作为例示，假设具有简单重复基序[TCTA]的重复区段，具有以序列TCAGCTA开始的3’侧翼。因此，参考可以包括诸如[侧翼1][TCTA]$_n$TCAGCTA[rest_of_flank2]等序列，其中“n”是等位基因中的重复数目。9.3等位基因与10等位基因的相差之处将在于在沿着序列的某处具有缺失。这些可能包括在参考中，虽然情况可以不完全如此。[TCTA]$_7$TCA[TCTA]$_2$是此类等位基因的实例。根据现有的对准方案，在[TCTA]$_7$之后和最终[TCTA]前结束的任何读段将与[侧翼1][TCTA]$_7$TCAGCTA对准，从而产生不适当的调用物。

[0091] 本文中提供的实施方案允许测定位于第一保守侧翼区和第二保守侧翼区之间的多态性重复DNA元件或重复区段的长度。在一个实施方案中，方法包括提供包含多态性重复DNA元件的至少一个样品读段的数据集；提供包含第一保守侧翼区和第二保守侧翼区的参考序列；将参考序列的第一侧翼区的一部分与样品读段对准；将参考序列的第二侧翼区的一部分与样品读段对准；以及测定重复区段的长度和/或序列。在典型的实施方案中，使用适当编程的计算机进行方法中的一个或多个步骤。

[0092] 如本文中使用的，术语“样品读段”是指要测定重复元件的长度和/或同一性的序列数据。样品读段可以基于DNA或RNA。样品读段可以包括所有重复元件或其一部分。样品读段可以进一步包含在重复元件的一端上的保守侧翼区(例如，5’侧翼区)。样品读段可以进一步包含在重复元件的另一端上的另外的保守侧翼区(例如，3’侧翼区)。在典型的实施方案中，样品读段包括来自具有正向和反向引物序列的PCR扩增子的序列数据。序列数据可以从任何选择序列方法获得。样品读段可以是例如来自合成测序(SBS)反应、连接测序反应、或任何其它合适的测序方法，对此期望测定重复元件的长度和/或同一性。样品读段可以是从多个样品读段衍生的共有(例如，平均或加权)序列。在某些实施方案中，提供参考序列包括基于PCR扩增子的引物序列鉴定感兴趣基因座。

[0093] 如本文中使用的，术语“多态性重复DNA元件”是指任何重复DNA序列，其可以称为重复区段。本文中提供的方法可用于对准任何此类重复DNA序列的相应侧翼区。本文中呈现的方法可用于难以对准的任何区域，而不管重复类别。本文中呈现的方法对于具有保守侧翼区的区域特别有用。另外/或者，本文中呈现的方法对于跨越包括每个侧翼区的至少一部分的整个重复区段的样品读段特别有用。在典型的实施方案中，重复DNA元件是可变数目串联重复(VNTR)。VNTR是多态性，其中特定序列在该基因座处重复多次。一些VNTR包括小卫星(minisatellites)和微卫星(microsatellites)，也称为简单序列重复(SSR)或短串联重复(STR)。在一些实施方案中，重复区段小于100个核苷酸，尽管可以对准更大的重复区段。重复部分的重复单元(例如，重复基序)可以是2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20或更多个核苷酸，并且可以重复直至2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29、30、40、45、50、55、60、65、70、75、80、85、90、95或直至至少100，或者更多。

[0094] 在某些实施方案中，多态性重复DNA元件是STR。在一些实施方案中，STR用于法医目的。在用于法医应用的典型实施方案中，例如，多态性重复DNA元件包含四核苷酸或五核苷酸重复基序，然而，本文中呈现的方法适合于任何长度的重复基序。在某些实施方案中，重复区段是短串联重复(STR)，如例如选自CODIS常染色体STR基因座、CODIS Y-STR基因座、EU常染色体STR基因座、EU Y-STR基因座等的STR。作为实例，CODIS(组合DNA指数系统)数据库是由FBI实验室鉴定的核心STR基因座集，包括13种基因座：CSF1PO、FGA、TH01、TPOX、VWA、D3S1358、D5S818、D7S820、D8S1179、D13S317、D16S539、D18S51和D21S11。法医界感兴趣并且可以使用本文中呈现的方法和系统对准的其它STR包括PENTA D和PENTA E。本文中呈现的方法和系统可以应用于任何重复DNA元件，并且不限于上文描述的STR。

[0095] 如本文中使用的，术语“参考序列”是指充当可以与样品序列对准的支架的已知序列。在本文中提供的方法和系统的典型实施方案中，参考序列包含至少第一保守侧翼区和第二保守侧翼区。术语“保守侧翼区”是指重复区段外的序列区(例如，STR)。该区域通常在

许多等位基因间是保守的，即使重复区段可以是多态性的。如本文中使用的保守侧翼区通常将具有比重复区段更高的复杂性。在典型的实施方案中，单个参考序列可用于对准遗传基因座内的所有等位基因。在一些实施方案中，由于侧翼区内的变异，超过一种参考序列用于对准遗传基因座的样品序列。例如，用于釉原蛋白(Amelogenin)的重复区段在X和Y之间的侧翼中具有差异，但是如果参考中包括较长区域，那么单个参考可以表示重复区段。

[0096] 在本文中呈现的实施方案中，参考序列的侧翼区的一部分与样品序列对准。通过测定保守侧翼区的位置，然后进行侧翼区的所述部分与样品读段的相应部分的序列对准来进行对准。根据已知的对准方法进行侧翼区的一部分的对准。在某些实施方案中，侧翼区(例如，第一或第二侧翼区)的一部分的对准包括：(i)通过使用与重复区段重叠或相邻的接种区的精确k聚体匹配来测定样品读段上保守侧翼区的位置；并且(ii)将侧翼区与样品读段对准。在一些实施方案中，对准可以进一步包括将侧翼序列和包括重复区段的一部分的短相邻区域进行对准。

[0097] 图4中显示了该方法的实例。也可称为样品读段的扩增子(“模板”)显示于图4中，其具有未知长度和/或同一性的STR。如上文就图2而言描述，可以分析样品读段以将样品读段分配到遗传基因座，其在这种情况下已知包括STR。在测定样品读段的遗传基因座后，对准方案可以包括将样品读段的预定序列与参考序列的预定序列进行对准。例如，引物显示为p1和p2，其基于用于产生扩增子的引物序列。在图4所示的实施方案中，在初始对准步骤期间使用单独的p1。在一些实施方案中，单独的p2用于引物对准。在其它实施方案中，p1和p2都用于引物对准。然而，在其它实施方案中，其它序列可以用于初始对准步骤。

[0098] 在初始对准之后，将侧翼1对准，在图4中指定为“$f1_{al}$”。侧翼1对准之前可以接种侧翼1，在图4中称为“$f1_{种子}$”。侧翼1接种是为了校正在样品序列的开始和STR之间的少数(表示为“e”)插入缺失(indel)。接种区可以直接邻近STR的开始，或者可以是偏移的(如图中一样)，以避免低复杂性区。可以通过精确的k聚体匹配来完成接种。继续侧翼1对准以测定STR序列的起始位置。如果STR模式足够保守以预测前几个核苷酸(s1)，那么将这些添加到对准以实现改善的精确度。

[0099] 由于STR的长度是未知的，因此如下对侧翼2进行对准。进行侧翼2接种以快速找出STR的可能的末端位置。作为侧翼1的接种，可以偏移接种以避免低复杂性区和错误对准。弃去任何不能对准的侧翼2种子。一旦侧翼2正确对准，可以测定STR的末端位置(s2)。利用在s1处已知的STR序列的开始和在s2处已知的STR序列的末端，可以计算STR的长度。

[0100] 接种区可以直接与重复段(例如，STR)相邻和/或包括重复段的一部分。在一些实施方案中，接种区的位置将取决于与重复区段直接相邻的区域的复杂性。STR的起始或末端可以由包含额外重复或具有低复杂性的序列限定。因此，可以有利的是偏移侧翼区的接种，以避免低复杂性的区域。如本文中使用的，术语“低复杂性”是指具有与重复基序和/或重复区段的序列类似的序列的区域。另外/或者，低复杂性区掺入低多样性的核苷酸。例如，在一些实施方案中，低复杂性区包含与重复序列具有超过30％、40％、50％、60％、70％或超过80％序列同一性的序列。在典型的实施方案中，低复杂性区以区域中所有核苷酸的小于20％、15％、10％或小于5％的频率掺入四种核苷酸中的每种。可以利用任何合适的方法来测定低复杂性的区域。测定低复杂性区域的方法是本领域中已知的，如Morgulis et al., (2006)Bioinformatics.22(2):134-41中公开的方法例示，其通过引用整体并入本文。例

如，如Morgulis等人的并入材料中所述，诸如DUST的算法可用于鉴定给定核苷酸序列内具有低复杂性的区域。

[0101] 在一些实施方案中，接种与STR的开始偏移至少1、2、3、4、5、6、7、8、9、10、15、20、25、30、35、40个或更多个核苷酸。在一些实施方案中，评估侧翼区以鉴定高复杂性的区域。如本文中使用的，术语“高复杂性区”是指具有与重复基序和/或重复区段的序列足够不同的序列，使得其降低错误对准的可能性的区域。另外/或者，高复杂性区掺入多种核苷酸。例如，在一些实施方案中，高复杂性区包含与重复序列具有小于80％、70％、60％、50％、40％、30％、20％或小于10％同一性的序列。在典型的实施方案中，高复杂性区以区域中所有核苷酸的至少10％、15％、20％或至少25％的频率掺入四种核苷酸中的每种。

[0102] 如本文中使用的，术语“精确k-匹配”是指通过使用字方法找到最佳对准的方法，其中字长度定义为具有值k。在一些实施方案中，k值的长度是3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29、30、31、32、33、34、35、36、37、38、39、40或更多个核苷酸。在典型的实施方案中，k具有5至30个核苷酸的长度值。在一些实施方案中，k具有5至16个核苷酸的长度值。在某些实施方案中，k由系统或用户基于一个或多个因素来选择。例如，如果侧翼区短，如当引物序列位于相对接近STR序列时，可以适当地减少k。在典型的实施方案中，选择k以便在+/-e的距离内找到所有匹配。

[0103] 字方法鉴定查询序列中的一系列短的、非重叠的子序列(“字”)，然后将其与候选数据库序列匹配。减去在所比较的两种序列中的字的相对位置以获得偏移；如果多个不同的字产生相同的偏移，那么这将指示对准区域。只有当检测到此区域时，这些方法的确应用更灵敏的对准标准；因此，消除了与没有明显相似性的序列的许多不必要的比较。进行k聚体匹配(包括精确的k聚体匹配)的方法是本领域中已知的，如Lipman,et al.,(1985)Science227:1435-41，以及Altschul,et al.,(1990)Journal of Molecular Biology215:403-410的公开例示，其各自的内容通过引用整体并入。

[0104] 如本文中使用的，术语“扩增子”是指获得序列的任何合适的扩增产物。通常，扩增产物是使用靶物特异性引物如PCR引物的选择性扩增方法的产物。在某些实施方案中，序列数据来自具有正向和反向引物序列的PCR扩增子。在一些实施方案中，选择性扩增方法可包括一个或多个非选择性扩增步骤。例如，使用随机或简并引物的扩增过程之后可以是使用靶物特异性引物的一个或多个扩增循环。用于选择性扩增的合适方法包括但不限于聚合酶链式反应(PCR)、链置换扩增(SDA)、转录介导的扩增(TMA)和基于核酸序列的扩增(NASBA)，如美国专利号8,003,354中描述，其通过引用整体并入本文。上述扩增方法可用于选择性扩增一种或多种感兴趣的核酸。例如，包括多重PCR、SDA、TMA、NASBA等的PCR可用于选择性扩增一种或多种感兴趣的核酸。在此类实施方案中，在扩增反应中包括特异性针对感兴趣的核酸的引物。用于扩增核酸的其它合适方法可包括寡核苷酸延伸和连接、滚环扩增(RCA)(Lizardi et al.,Nat.Genet.19:225-232(1998)，其通过引用并入本文)和寡核苷酸连接测定法(OLA)(通常参见美国专利号7,582,420、5,185,243、5,679,524和5,573,907;EP 0 320 308 B1;EP 0 336 731 B1;EP 0 439 182 B1;WO 90/01069;WO 89/12696;和WO 89/09835，所有这些文献通过引用并入本文)技术。

[0105] 应当理解，这些扩增方法可以设计为选择性扩增感兴趣的靶核酸。例如，在一些实施方案中，选择性扩增方法可以包括连接探针扩增或寡核苷酸连接测定(OLA)反应，其包含

特异性针对感兴趣的核酸的引物。在一些实施方案中，选择性扩增方法可包括引物延伸-连接反应，其含有特异性针对感兴趣的核酸的引物。作为可以特异性设计为扩增感兴趣的核酸的引物延伸和连接引物的非限制性实例，扩增可以包括用于GoldenGate™测定(Illumina,Inc.,San Diego,CA)的引物，如美国专利号7,582,420，其通过引用整体并入本文。本方法不限于任何特定的扩增技术，并且就本公开的方法和实施方案而言，本文中所述的扩增技术仅仅是示例性的。

[0106] 用于扩增重复DNA元件的引物通常与侧翼区的独特序列杂交。可以根据任何合适的方法设计和产生引物。用于重复区段的侧翼区的引物的设计是本领域中公知的，如Zhi,et al.(2006)Genome Biol,7(1):R例示，其通过引用整体并入本文。例如，可以手动设计引物。这涉及搜索基因组DNA序列的微卫星重复，这可以通过眼或通过使用自动化工具，如RepeatMasker软件完成。一旦测定了重复区段和相应的侧翼区，侧翼序列可以用于设计将在PCR反应中扩增特定重复的寡核苷酸引物。

[0107] 以下描述根据上述描述进行的实施例。

[0108] 实施例1:基因座D18S51的对准

[0109] 本实施例描述了根据一个实施方案的基因座D18S51的对准。一些基因座具有低复杂性且类似于STR重复序列的侧翼序列。这可引起侧翼序列错误对准(有时与STR序列本身)，因此可以错误调用等位基因。麻烦的基因座的一个例子是D18S51。重复基序是[AGAA]n AAAG AGAGAG。下文显示了侧翼序列，低复杂性“问题”序列加下划线：

[0110] GAGACCTTGTCTC**(STR)GAAAGAAAGAGAAAAAGAAAAGAAA**TAGTAGCAACTGTTAT

[0111] 如果使用直接邻近STR的侧翼区接种对准，那么将产生k-聚体，如GAAAG、AAAGAA、AGAGAAA，其定位到STR序列。这阻碍了性能，因为从接种中获得了许多可能性，但是最重要的是，该方法产生错误对准，如图5中显示的那些错误对照。在图5中显示的序列中，突出显示了真正的STR序列，源自错误对准的STR是加下划线的，并且读段错误以粗体显示。

[0112] 对于这些低复杂性侧翼，通过将它们进一步推离STR序列，确保接种区不在低复杂性区中。虽然这需要更长的读段来调用STR，但是其确保高准确性并且防止侧翼区与STR序列(或侧翼的其它部分)的错误对准。仍然将低复杂性侧翼与读段对准以找到STR的结束位置，但是因为对准是用高复杂性序列接种的，所以它应该在正确的位置中。

[0113] 实施例2:通过短STR序列添加的基因座Penta-D的对准

[0114] 一组Penta-D序列倾向于具有比预期短1nt的STR。进一步检查后，发现两个侧翼都含有聚A区段并且测序/扩增错误经常除去那些区段中的A之一。如下述序列中显示，在两侧均发现同多聚体A区段。

[0115]

```
...CAAGAAAGAAAAAAAAG [AAAGA]nAAAAACGAAGGGGAAAAAAAGAGAAT...
```

[0116] 引起第一侧翼中缺失的读段错误将产生两个相等可行的对准：

[0117]

读段： ...CAAGAAAG**AAAAAAA**-GA...
侧翼： ...CAAGAAAG**AAAAAAAA**G- （2个插入缺失）

读段： ...CAAGAAAG**AAAAAAA**GA... （2个错配）
侧翼： ...CAAGAAAG**AAAAAAAA**G

[0118] 将最接近STR的碱基强制为匹配不起作用，因为STR之一中的侧翼之一以其中具有SNP结束。发现仅添加STR序列的2个核苷酸解决了问题：

[0119]

读段： ...CAAGAAAG**AAAAAAA**-GAA
侧翼： ...CAAGAAAG**AAAAAAAA**GAA （1个插入缺失） ✔

读段： ...CAAGAAAG**AAAAAAA**G-AA （1个插入缺失+ 1个错配）
侧翼： ...CAAGAAAG**AAAAAAAA**GAA

[0120] 实施例3:DNA样品的混合物的分析

[0121] 使用本文中提供的方法分析样品的混合物以对一组1STR中的每种基因座产生调用物。对于每种基因座，对与每种等位基因和与所述等位基因的每种不同序列对应的读段数目进行计数。

[0122] 典型的结果示于图6A-6D中。如显示，每对右侧的柱形表示获得的实际数据，指示每种等位基因的读段的比例。不同的阴影代表不同的序列。省略了具有小于0.1％的基因座读段计数的等位基因和具有小于1％的等位基因计数的序列。每对左侧的柱形代表理论比例(无打滑)。不同的阴影在输入中代表不同的对照DNA，如图例中指示。在图6A-6D中，x轴为等位基因次序，并且Y轴指示具有指定等位基因的读段的比例。

[0123] 如图中显示，使用本文中呈现的方法的STR调用方法实现了组中每种等位基因的准确得令人惊讶的调用。

[0124] 实施例4:法医STR组的分析

[0125] 在5份不同样品中分析了15种不同基因座的组。样品获自Promega Corp，并且包括样品9947A、K562、2800M、NIST:A和B(SRM 2391c)。基因座从CODIS STR法医学标志物中选择，并且包括CSF1PO、D3S1358、D7S820、D16S539、D18S51、FGA、PentaE、TH01、vWA、D5S818、D8S1179、D13S317、D21S11、PentaD和TPOX，使用本文中呈现的对准方法。简言之，使用标准引物扩增标志物，如Krenke,et al.(2002)J.Forensic Sci.47(4):773-785中阐述，其通过引用整体并入本文。将扩增子合并，并在MiSeq测序仪(Illumina,San Diego,CA)上使用1x460循环获得测序数据。

[0126] 根据本文中提供的方法进行对准。如图7中阐述，与对照数据相比，显示了这些对照样品的100％一致性。此外，该方法鉴定了标志物D8S1179的样品之一中的先前未知的SNP，进一步证明当与本文中提供的对准方法组合时基于序列的STR分析的强大工具。

[0127] 图8显示了鉴定打滑产物的方法160。在已经鉴定了分配读段内的ROI之后，本文中阐述的实施方案可以基于ROI的序列在162处对ROI(或分配读段)进行分类。如上文描述，在

某些情况下，对准方案可以分析除了重复区段的序列之外的侧翼区之一或两者的一部分。因此，在某些实施方案中，在162处分选可以包括基于重复段的序列和侧翼区之一或两者的子序列分类。作为实例，分类可以包括分析重复区段和从重复区段延伸的每个侧翼区的几个核苷酸。在其它实施方案中，在162处的分选可以包括基于仅包括重复区段的序列的ROI的分类。

[0128] 可以对ROI(或重复区段)进行分类，使得具有不同序列的ROI(或重复区段)指定为潜在(或疑似)等位基因。例如，每种潜在的等位基因可以具有独特的样品序列和/或独特的长度。更具体地，每种潜在等位基因可以具有ROI或重复区段的独特序列和/或ROI或重复片段的独特长度。如下文描述，在一些实施方案中，可以基于CE等位基因名称对重复区段进行排序。

[0129] 可以对每种指定基因座进行在162处的分类。在将样品读段分配到相应的遗传基因座后，每种遗传基因座可以具有与遗传基因座相关的多个分配读段。例如，在一些实施方案中，一种或多种遗传基因座可以具有数百个分配读段，它们彼此分组或框并。如已知的，相应的遗传基因座，如已知的STR基因座，可以具有多种等位基因，其中每种等位基因包括不同的序列。通过共同分析已经鉴定为来自共同遗传基因座的多个分配读段，可以分析多个分配读段，以提供个体或多个个体的基因型调用物。

[0130] 方法160还可以包括在164处对具有共同序列的共同遗传基因座的分配读段进行计数(或求和)。在164处的计数可以包括测定计数得分，如本文中描述。作为实例，图9包括包含D1S1656基因座的潜在等位基因的表190，并且图10包括显示CE等位基因的分布的图192。根据惯例命名CE等位基因，如图10中显示，可能包括打滑产物。在该实施例中，在从单一来源测序核酸后，分析样品读段以鉴定D1S1656基因座的ROI(例如重复片段)。对ROI进行分类和计数以鉴定D1S1656基因座内的多个潜在等位基因。在该实施例中，没有考虑具有计数低于D1S1656基因座的分配读段的总数的1％的等位基因。如图9中所示，过滤的分配读段包括总共四种独特序列，其可以认为是D1S1656基因座的潜在等位基因。分析后，如下所述，基因座的基因型调用物是杂合的12/13。

[0131] 在一些实施方案中，基于遗传基因座的潜在等位基因的计数得分，可以对遗传基因座产生基因型调用物。然而，在一些实施方案中，可以进行序列的进一步分析。例如，方法160可以包括在166处分析潜在等位基因的序列以确定第一等位基因是否是第二等位基因的疑似打滑产物。打滑是在核酸扩增期间可能发生的现象，特别是包括一个或多个系列的重复基序(如在STR等位基因内发现的那些)的核酸。打滑产物具有通常比真实等位基因大小更小(或大小更大)的一个或多个重复基序的序列。在核酸序列的复制过程中，两条链可沿着STR分开。因为每个重复基序是相同的，所以两条链可能不适当地重新退火，使得两条链偏移一个或多个重复基序。因此，可以进一步扩增的所得产物与真实序列相差一个或多个重复基序。

[0132] 因为打滑产物与真正的等位基因具有几乎相同的大小，所以测定打滑产物是遗传基因座的真正等位基因还是相邻等位基因的打滑产物是具有挑战性的。因此，打滑产物可以降低基因型调用的置信度。在某些情况下，打滑产物可以阻止提供基因型调用物或潜在地引起不正确的基因型调用物。打滑产物可以使包括多个来源的样品的基因型调用物变得特别具有挑战性。

[0133] 在166处的分析可以确定第一等位基因是否是第二等位基因的疑似打滑产物。在一些实施方案中，分析包括将一种或多种规则或条件应用于第一和第二等位基因的序列。例如，如果在171处测定已经在第一和第二等位基因之间添加或丢失了k个重复基序，那么第一等位基因可以是第二等位基因的疑似打滑产物。数目k是整数。在具体的实施方案中，数目k是1或2。尽管打滑产物通常少包含一个重复基序，但打滑产物也可少包括两个重复基序或包含一个添加的重复基序。打滑产物可能还包括重复基序中的其它差异。在166处的分析可以包括将与遗传基因座相关的每种潜在等位基因与相同遗传基因座的每种其它潜在等位基因进行比较。

[0134] 在一些实施方案中，在166处的分析可以包括在172处鉴定已经添加或丢失的重复基序。在172处鉴定已添加或丢失的重复基序可包括沿着ROI或重复区段对准两种等位基因的两种序列以测定何时丢失或添加重复基序。例如，序列可以在一端彼此对准，以测定何时添加或丢失重复基序。

[0135] 或者或除上述之外，分析可以包括在173处比较第一和第二等位基因的重复片段的长度，以确定第一和第二等位基因之间的重复区段的长度是否相差一个重复基序或多个重复基序的长度。例如，在图9中所示的实例中，重复基序是TAGA，其是具有四种核苷酸的四核苷酸。靶等位基因的序列长度显示在图9中。等位基因1和等位基因2中的每种具有62个核苷酸，并且等位基因3和等位基因4中的每种具有58个核苷酸。因此，等位基因1的序列长度与等位基因3的序列和等位基因4的序列差异四个核苷酸，或换言之，重复基序的长度。同样地，等位基因2的序列长度与等位基因3的序列和等位基因4的序列相差重复基序的长度。

[0136] 在一些实施方案中，在166处的分析可以包括在174处测定添加的或丢失的重复基序是否与相同序列中的相邻重复基序相同。如上所述，可以通过对准等位基因序列来鉴定已经添加或丢失的重复基序来测定添加或丢失的重复基序。在对准序列后，可以测定添加/删除的重复基序与其相邻的重复基序相同。在一些实施方案中，可以通过使用贪婪算法(greedy algorithm)来实现对准。

[0137] 第一等位基因(或疑似为打滑产物的等位基因)通常包括小于第二等位基因的读段计数(或计数得分)的读段计数(或计数得分)。在某些情况下，如当样品包括次要贡献者时，这可能不是真的。在一些情况下，等位基因的打滑产物可以小于指定的打滑阈值或落入基因座和/或等位基因的预定范围内。打滑阈值可以基于例如第二等位基因的读段计数的数目、相应基因座和/或等位基因的历史数据、和/或在测定期间对相应基因座和/或等位基因的观察。为了提供关于等位基因的历史数据或观察的实例，可以通过关于指定测定的经验测定等位基因提供大于或小于通常预期的预定量的打滑。该数据和/或观察可以用于修改阈值。作为等位基因的知识可以影响打滑阈值的另一个实例，更长的等位基因平均可以提供更大百分比的打滑产物。因此，可以基于等位基因的长度来改变打滑阈值。

[0138] 在一些实施方案中，在166处的分析可以包括在175测定第一等位基因的计数得分是否落入第二等位基因的计数得分的预定范围内。例如，如果第一等位基因的计数得分(例如，读段计数)在第二等位基因的计数得分(例如，读段计数)的预定百分位数范围内，那么第一等位基因可以是疑似性打滑产物。预定百分位数范围可以在约5％和约40％之间。在具体实施方案中，预定百分位数范围可以在约10％和约30％之间或在约10％和约25％之间。预定百分位数范围可以使用历史数据或在测定期间相应的STR基因座的观察来计算或获

得。同样地，如果第一等位基因的计数得分小于基于第二等位基因的计数得分的指定的打滑阈值，那么第一等位基因可以是疑似打滑产物。作为实例，指定的打滑阈值可以基于第二等位基因的计数得分的预定百分比。例如，预定百分比可以是大约20％、25％、30％、35％或40％。可以使用相应STR的历史数据或在测定期间相应STR基因座的观察来测定或获得预定百分比。

[0139] 在一些实施方案中，潜在等位基因的计数得分可用于测定打滑度量(或打滑得分)。打滑度量可以是基于第一等位基因的计数得分的值或函数。打滑度量也可以基于第二等位基因的计数得分。可以将打滑度量与指定的打滑阈值进行比较，以测定相应的潜在等位基因是否是疑似打滑产物。如果打滑度量小于指定的打滑阈值，那么第一等位基因可以是第二等位基因的疑似打滑产物。如果打滑度量不小于指定的打滑阈值，那么第一等位基因可以认为是潜在的等位基因。在这种情况下，第一等位基因和第二等位基因可以各自是基因座的真正等位基因。

[0140] 可以应用另外的条件来测定一种等位基因是否是另一种等位基因的打滑产物。例如，在166处的分析可以包括在176处测定在第一和第二等位基因的序列之间不存在其它错配。可以分析ROI，或者更具体地，可以分析重复段以鉴定各个序列之间的任何错配。例如，如果一种序列的核苷酸与另一种序列的核苷酸(除了添加的/丢失的重复基序)不匹配，那么该序列可能不是打滑产物。

[0141] 在其它实施方案中，可以测定疑似打滑产物不是第二等位基因的打滑产物。相反，疑似打滑产物可能来自另一个贡献者或者可能由测序错误引起。例如，如果第一等位基因的打滑度量(例如，基于计数得分的计数得分或其它函数)大于指定的打滑阈值，那么一个或多个实施方案可以测定疑似打滑产物来自另一个贡献者。指定的阈值可以基于第二等位基因的计数得分和预定的打滑函数，其可以基于历史数据和/或感兴趣测定内的数据。一个或多个实施方案可以测定如果第一等位基因的打滑度量度小于基线值，那么疑似打滑产物是测序错误。基线值可以基于第二等位基因的计数得分和预定的打滑函数，其可以基于历史数据和/或感兴趣测定内的数据。作为实例，某个基因座可以历史上具有10-30％的基因座范围。如果某一基因座的第二等位基因的读段为100，那么如果读段小于10，那么第一等位基因可能是测序错误。如果读段大于30，那么第一等位基因可能来自另一个贡献者。

[0142] 在具体实施方案中，如果：(A)第一和第二等位基因的等位基因序列的长度相差k个重复基序，那么第一等位基因认为是第二等位基因的打滑产物；(B)丢失或添加的重复基序与相邻重复基序相同；(C)在两种等位基因(例如，ROI或重复片段)之间没有其它错配；和任选地，(D)第一等位基因的打滑度量在第二等位基因的打滑度量的预定的打滑范围(或小于指定的打滑阈值)内。

[0143] 返回到图9中显示的实例，D1S1656基因座的两种真等位基因的序列是等位基因12的[TAGA]11[TAGG]1[TG]5和等位基因13的[TAGA]13[TG]5。等位基因12在最后的“TAGA”重复单元中具有SNP。由此，我们可以测定等位基因12序列[TAGA]12[TG]5事实上是等位基因13的-1打滑，并且等位基因13序列[TAGA]12[TAGG]1[TG]5是等位基因12的+1打滑。如可以看出，本文阐述的实施方案可以优于CE系统。更具体地，CE系统将不能测定等位基因12序列[TAGA]12[TG]5是等位基因13的-1打滑，并且等位基因13序列[TAGA]12[TAGG]1[TG]5是等位基因12的+1打滑。

[0144] 图11显示了根据实施方案的分析测序数据的方法200。方法200可以与本文中阐述的其它实施方案结合。方法200包括在202处接收包括配置为对应于遗传基因座集的多个样品读段的测序数据。遗传基因座集可以配置用于预定的遗传应用，如法医学(forensics)或亲子鉴定。样品读段可以形成相应扩增子的读段对，其中每个读段对包括相应扩增子的第一读段和第二读段。例如，第一和第二读段对可以从对末端测序获得，在具体实施方案中，可以从不对称配对末端测序获得。第一和第二读段中的每个可以具有相应的序列，以下称为读段序列。每个读段序列可以包括例如鉴定序列(例如引物序列)和包括序列变异，如SNP或STR的序列。

[0145] 方法200可以包括在204处鉴定样品读段的一个或多个潜在遗传基因座。鉴定操作可类似于上文就图2而言描述的在154处的分配。例如，在204处，可临时鉴定读段对的第一读段的一个或多个遗传基因座。可以将每个读段对的第一读段与数据库(例如，查找表)的选择序列进行比较。数据库的每个选择序列可以对应于遗传基因座集的指定遗传基因座。如果第一读段的读段序列有效地匹配一个或多个选择序列，那么可以暂时调用对应于选择序列的遗传基因座的第一读段。例如，如果来自第一读段的鉴定序列的一系列n个核苷酸(例如，前n个核苷酸)有效地匹配一个或多个选择序列，那么可以暂时调用那些相应的遗传基因座的第一读段。一种或多种相应的遗传基因座可以称为一种或多种临时指定基因座。

[0146] 如果第一读段没有有效地与任何选择序列匹配，那么可以弃去未分配的读段。任选地，可以将可以是第一读段和/或相应的第二读段的未分配的读段与其它未分配的读段收集或聚合在一起。在206处，可以分析未分配的读段以进行质量控制。例如，可以分析第一读段的读段序列以测定为何不分配第一读段。

[0147] 方法200还可包括在208处针对具有潜在遗传基因座的每个第一读段来测定第一读段是否与潜在遗传基因座的一个或多个参考序列对准。可以使用一个或多个对准方案来进行测定208。例如，在208处的测定可以包括将第一读段与潜在遗传基因座的相应参考序列进行对准，如上文就图3-7而言描述。如果第一读段与仅潜在基因座之一的参考序列对准，那么可以将第一读段暂时指定为一个遗传基因座的有效读段，并且该方法可以进行到步骤210。在其它实施方案中，可以将第一读段指定为一个遗传基因座的有效读段，并且该方法可以进行到步骤212。

[0148] 然而，如果第一读段有效地与超过一个参考序列对准，那么在208处的测定可以包括鉴定与第一读段最佳对准或最对准的参考序列。更具体地，尽管第一读段可以有效地与多个参考序列对准，一个对准可能优于其它对准。作为一个简单的实例，对准分析可以分析第一读段并将第一读段与三种参考序列Ref Seq A、Ref Seq B和Ref Seq C进行对准，所述参考序列是与204处鉴定的三种潜在遗传基因座有关的参考序列。对准分析可以测定第一读段有效与Ref Seq A对准，在Ref Seq A和第一读段之间具有总共三个差异。对准分析可以测定第一读段有效与Ref Seq B对准，在Ref Seq B和第一读段之间具有总共四个差异。对准分析可以测定第一读段和参考序列C不彼此对准。例如，在第一读段和Ref Seq C之间可以存在过多数目的差异(例如，高于10)。作为另一个实例，差异的过多比例或百分比(例如，相对于读段或参考序列中的核苷酸总数的差异数目)可以存在于第一读段和Ref Seq C之间。基于该数据，该方法可以测定第一读段与Ref Seq A比与Ref Seq B更好地对准。因此，可以将第一读段暂时指定为对应于Ref Seq A的遗传基因座的有效读段。

[0149] 在一些实施方案中，测定哪个参照序列与第一读段最佳匹配可以包括计算每个参考序列的对准得分，其中对准得分基于差异的数目。如上文描述，对准得分可以是原始数目(例如，差异数目)。在其它实施方案中，对准得分可以是差异的数目和/或类型的函数。例如，可以不同地对插入缺失和不匹配得分。

[0150] 任选地，方法200可以包括在210处分析第二读段，以确认应该为暂时指定的遗传基因座调用第一读段。可以以与相应读段对的第一读段类似的方式来分析第二读段。可以分析第二读段以测定第二读段的鉴定序列是否有效地匹配数据库的一个或多个选择序列。如果第二读段的鉴定序列有效地仅匹配一个选择序列，那么该方法可以包括鉴定对应于一个选择序列的遗传基因座。如果遗传基因座是与将第一读段临时指定的相同的遗传基因座，那么遗传基因座可以称为第一读段的遗传基因座，并且可以在212处将第一读段指定为遗传基因座的有效读段。

[0151] 然而，如果第二读段的鉴定序列有效地匹配多个选择序列，那么该方法可以包括鉴定对应于多个选择序列的遗传基因座。如果这些遗传基因座之一是与将第一读段临时指定的相同的遗传基因座，那么遗传基因座可以称为第一读段的遗传基因座，并且可以在212处将第一读段指定为遗传基因座的有效读段。

[0152] 如果在210处的分析不确认第二读段对应于第一读段的临时指定的基因座，那么方法200可以包括将相应的第一读段指定为未确认的读段。在214处，可以收集和可选地进一步分析未确认的读段，用于质量控制。例如，与临时指定基因座的第一选择序列有效匹配，但没有与临时指定基因座的第二选择序列有效匹配的读段对可以指示分析中的关注。未确认的读段可以指示一个或多个脱靶扩增子。在214处可以分析读段对，以测定例如质量控制问题是否存在于测定内或指示等位基因退出。

[0153] 然而，如果在208处，第一读段不与潜在遗传基因座的参考序列对准，那么该方法可以包括在216将第一读段指定为未对准的前导。未对准读段可以表示通过一个过滤阶段但不能与参考序列对准的第一读段。具体地，未对准读段可以是已经确认具有与一个或多个选择序列有效匹配但不能与参考序列对准的鉴定序列的第一读段。

[0154] 任选地，方法200可以包括在218处分析每个未对准读段以测定相应的未对准读段的最佳拟合遗传基因座。如上文描述，鉴定序列可以有效地与超过一个选择序列匹配。在218处的分析可以包括将未对准读段的鉴定序列与先前在204处鉴定的选择序列进行比较。最佳拟合遗传基因座可以是对应于与未对准读段的鉴定序列最佳匹配或最匹配的选择序列的遗传基因座。因此，在218处，该方法可以测定多个选择序列中的哪个选择序列与鉴定序列最佳匹配。例如，最佳拟合遗传基因座可以是对应于与鉴定序列具有最少差异的选择序列的遗传基因座。在一些实施方案中，在218处的分析可以包括测定每个选择序列相对于鉴定序列的匹配得分。可以将对应于具有最大匹配得分的选择序列的遗传基因座指定为最佳拟合遗传基因座。

[0155] 在220处，可分析与未对准读段(即，第一读段)相关联的第二读段以测定第二读段是否确认在218处鉴定的最佳拟合基因座。可分析第二读段以测定是否第二读段的鉴定序列有效地与一个或多个选择序列匹配。如果第二读段的鉴定序列有效地与选择序列匹配并且该选择序列对应于最佳拟合遗传基因座，那么可以在222将未对准读段指定为双中靶非对准读段(也称为对中靶未对准读段)。双中靶非对准读段可以表示具有与未对准读段的两

端接近的序列的未对准读段，其与来自数据库的选择序列有效匹配。尽管与两个选择序列有效匹配，未对准读段的ROI不能与参考序列对准。

[0156] 然而，如果第二读段的鉴定序列不与对应于最佳拟合遗传基因座的选择序列有效匹配，那么可以在224处将未对准读段指定为一中靶未对准读段。一中靶未对准读段可以表示仅具有一个有效地与来自数据库的选择序列匹配的鉴定序列的未对准读段。

[0157] 出于质量控制的目的，可分别在226和228处分析双中靶未对准读段和一中靶未对准读段两者。在226或228处的分析可以包括分析未对准读段的总数(或相当的得分)和/或分析未对准读段的ROI的序列。例如，可以在228处分析一中靶未对准读段，以测定测定法的良好。更具体地，可以分析一中靶非对准读段以测定是否存在嵌合体和/或是否存在引物二聚体。过多数目的嵌合体和/或引物二聚体可以指示测定法是差的(例如，扩增问题)或样品DNA具有低质量。任选地，在228处的分析可以包括分析214的未确认的读段以测定测定法的良好。在228处的分析可以包括共同分析未确认的读段和一中靶未对准读段。或者，在228处的分析可以包括单独分析未确认的读段和一中靶未对准读段。

[0158] 就双中靶非对准读段而言，过多数目的此类读段可以指示潜在等位基因退出。在一些实施方案中，在226处的分析可以包括测定双中靶非对准读段的数目是否超过指定基因座的总读段的百分比，然后可以测定与指定基因座存在问题。指定基因座的“总读段”可以是在212处指定的有效读段和在216处指定的未对准读段的函数。例如，总读段可以等于有效读段和未对准读段的总和。在其它实施方案中，总读段也可以是未确认的读段的函数。在226处，可以将双中靶未对准读段(或相当得分)的数目与阈值进行比较，以测定是否存在与指定基因座有关的问题(例如，等位基因退出)。

[0159] 在230处，可以提供关于测定的质量和/或遗传概况的置信度的通知。例如，通知可以对用户通知未对准读段的数目。在具体实施方案中，通知可以对用户通知一中靶未对准读段的数目和/或双中靶目标未对准读段的数目。在一些情况下，方法可以比较未对准读段的数目(或相当的得分)、一中靶未对准读段的数目(或相当的得分)和/或双中靶未对准读段的数目(相当的得分)与指定阈值。如果数目或得分超过阈值，那么通知可以包括针对用户的特定警告或特定指导。例如，该通知可以通知用户证据指示样品是质量差的和/或是少量的。该通知可以作为整体指向测定法，或者可以是对特定基因座特异的。更具体地，过多数目的一对靶未对准读段可以指示测定的问题，而过多数目的两对靶未对准读段可以指示等位基因退出。

[0160] 在232处，可以对有效读段进行分类以形成指定基因座的读段分布。读段分布通常包括已经通过多个过滤阶段并分配到指定基因座的许多样品读段。例如，读段分布可以包括已经分配到指定基因座的数十、数百或数千个第一读段。读段分布可以收集在文件(例如，“分布文件”)中，并且包括关于样品读段的分布的信息，例如不同的潜在等位基因、等位基因的序列和计数得分(例如读段计数或基于读段计数的其它值/功能)。例如，当针对读段分布对有效读段进行分类时，可以基于序列对有效读段进行分类。有效读段可以具有许多不同的序列，尽管不同但已经分配到指定基因座。每个不同的序列代表指定基因座的潜在等位基因。一个或多个序列可以是噪音(例如，测序错误)，一个或多个序列可以是打滑产物，并且一个或多个序列可以是真正的等位基因。

[0161] 可以将有效读段与具有相同序列的其它有效读段聚合在一起。可以对特定序列计

数具有相同序列的有效读段的数目。例如，假设具有对其分配的1000个有效读段的遗传基因座，读段分布可以指示存在8种不同的序列。可以在8种不同的序列间分布有效读段。例如，等位基因1可以具有10个有效读段；等位基因2可以具有20个有效读段；等位基因3可以具有10个有效读段；等位基因4可以具有400个有效读段；等位基因5可以具有15个有效读段；等位基因6可以具有500个有效读段；等位基因7可以具有25个有效读段；并且等位基因8可以具有20个有效读段。如下所述，进一步分析可以测定一些等位基因是噪音和/或打滑产物。

[0162] 在一些实施方案中，潜在等位基因可以提供基于CE中的常规命名实践的CE等位基因名称。潜在等位基因的CE等位基因名称可以部分地基于序列内的重复基序的数目。CE等位基因命名也可以基于历史使用。在一些实施方案中，基于CE等位基因名称在读段分布内对潜在等位基因进行排序。例如，CE等位基因名称通常包括数值。可以基于数值对潜在等位基因进行排序。作为一个实例，图10中所示的图192显示了一个读段分布。如显示，潜在的等位基因包括11、11.2、12、13和14。图192中所示的遗传基因座的读段分布可以是排序的11、11.2、12和13。

[0163] 在一些情况下，两种不同的潜在等位基因可以具有相同的CE等位基因名称。例如，基于常规命名实践，可以对潜在等位基因给予相同的CE等位基因名称。在一些实施方案中，读段分布可以指示两种不同序列具有相同的CE等位基因名称。例如，读段分布可以指示CE等位基因名称(例如，13)，然后列出对应于相同CE等位基因名称的不同序列。

[0164] 在对读段进行分类以形成读段分布之后，然后可以传送读段分布用于不同的分析。例如，可以通过SNP分析指导就包括SNP而言已知的遗传基因座。可以通过STR分析指导就STR而言已知的遗传基因座。虽然SNP和STR分析可以包括不同的步骤，但是分析还可以包括类似的步骤。

[0165] 图12显示了根据实施方案的分析测序数据的方法240。具体地，方法240包括分析指定基因座的读段分布。读段分布可以是STR基因座、SNP基因座或与序列变异相关的其它基因座。方法240包括在242处接收指定基因座的读段分布。就以下步骤而言，每个步骤可以至少部分地基于指定的基因座。例如，可以应用各种功能(例如，阈值)，其中那些功能基于指定的基因座。更具体地，一个遗传基因座的功能可能不是另一个遗传基因座的相同功能。

[0166] 任选地，方法240包括在244处测定指定的遗传基因座的每个潜在等位基因的计数得分。计数得分可以基于潜在等位基因的读段计数。读段计数表示包括共同序列的有效读段的数目。在一些实施方案中，计数得分是等于潜在等位基因的读段的值。例如，如果读段为300，那么计数得分可以为300。在其它实施方案中，潜在等位基因的计数得分可以基于读段计数和遗传基因座的读段的总数。读段的总数可以是例如所有潜在等位基因的读段分布内的读段的总数。在一些实施方案中，潜在等位基因的计数得分可以基于读段计数和先前获得的遗传基因座的数据。在具体实施方案中，计数得分可以是预定数目(例如，0和1)之间的标准化得分。标准化得分可以基于遗传基因座的读段的总数。任选地，标准化得分是来自其它基因座的读段计数和/或来自其它样品的读段计数的函数。计数得分还可以是来自样品的其它基因座的读段计数的函数或来自与感兴趣样品同时运行的其它样品的读段计数的函数。计数得分还可以是历史数据的函数。例如，可以运行不同类型的测定以获得读段。在一些实施方案中，计数得分基于关于特定测定的历史数据。

[0167] 方法240还可以包括在245处测定潜在等位基因的计数得分中的一个或多个是否通过解读阈值。解读阈值可以是预定值，或者可以是基于多个因素的函数。例如，解读阈值可以基于对应于指定基因座的总读段的数目。总读段的数目可以包括基因座内所有潜在等位基因的有效读段。在一些实施方案中，总读段的数目可以包括基因座的有效读段和基因座的未对准读段。在具体实施方案中，总读段的数目可以包括基因座的有效读段、未对准读段和未确认读段。如果在245处计数得分之一通过解读阈值，那么方法240可以进行到步骤246或另一个后续步骤。在一些实施方案中，解读阈值可以基于样品中读段的总数。

[0168] 如果在245处计数得分无一通过解读阈值，那么方法240可以在248处提供关于指定的基因座的警报或其它通知。例如，警报可以通知用户指定的基因座具有低覆盖。更具体地，警报可以通知用户关于指定基因座的数据量可能不充足以提供基因型调用物。

[0169] 在一个具体实施方案中，方法240包括鉴定在读段分布内具有最大读段(或等位基因计数)的潜在等位基因。读段计数表示包括共同序列的有效读段的数目。就STR而言，读段计数可以表示包括ROI或重复区段的共同序列的有效读段的数目。方法240还可以包括将最大读段计数与解读阈值进行比较。如果在245处，最大读段计数通过解读阈值，那么方法240可以进行到步骤246或另一个后续步骤。如果最大等位基因计数未通过解读阈值，那么方法240可以在248处提供关于指定基因座的警报或其它通知，如上文描述。

[0170] 在其它实施方案中，可将计数得分与另一阈值(如下文所述的分析阈值)进行比较。分析阈值通常比解读阈值更容易通过。如果潜在等位基因无一具有通过分析阈值的计数得分，那么可以测定遗传基因座具有低覆盖。作为用于测定遗传基因座是否具有足够覆盖的另一个实例，可以将遗传基因座的读段(例如，有效读段)的总数与读段阈值进行比较。读段阈值可以基于样品中的读段总数和/或历史数据。如果遗传基因座的读段总数没有通过读段阈值，那么可以测定遗传基因座具有低覆盖。在其它实施方案中，可以使用一个或多个步骤(如上文所述的那些)的组合来测定遗传基因座是否具有低覆盖。

[0171] 任选地，在246处，可以将读段分布内的每个计数得分或相应的读段计数与分析阈值进行比较。与解读阈值一样，分析阈值可以是预定值或基于多个因素的函数，如基因座的读段总数(例如，有效读段的总数)和/或指定基因座的历史知识。分析阈值可以比解读阈值更不严格(例如，更容易通过)。更具体地，解读阈值可能需要比分析阈值通过更大的读段计数。

[0172] 在246处通过分析阈值之后，方法240可以包括在247处测定潜在等位基因是否是疑似打滑产物。可以应用各种规则或条件来测定潜在等位基因是否是疑似打滑产物。例如，可以应用上文就图8而言描述的因素171-175中的一种或多种。在具体实施方案中，在247处的测定包括测定第一等位基因是否具有已经相对于第二等位基因添加或丢失的重复基序。

[0173] 如果怀疑潜在等位基因不是打滑产物，那么在250处将潜在等位基因指定为基因座的指定或调用的等位基因。如果疑似潜在等位基因是打滑产物，那么方法240包括在249处测定第一等位基因的计数得分是否小于指定阈值。计数得分可以是读段计数或基于读段计数的函数。指定的阈值可以基于第二等位基因的计数得分。在具体实施方案中，在249处的测定可以包括测定第一等位基因的计数得分是否在第二等位基因的计数得分的预定范围(例如，10%-30%)内。

[0174] 尽管在图12中未指示，但是如果潜在等位基因小于指定阈值或在预定范围内，那

么可以将潜在等位基因指定为第二等位基因的打滑产物。可以用基因座的基因型调用物记录打滑产物。例如，样品报告可以包括基因座的基因型，以及存在打滑产物的指示。可以在样品报告中提供关于打滑产物的信息(例如，第二等位基因的序列和百分比)。然而，如果计数得分或读段计数通过指定的阈值(或在预定范围内)，那么在250处可以将潜在等位基因指定为遗传基因座的指定等位基因。

[0175] 在一些实施方案中，在252处收集噪音等位基因的计数得分。在246处，噪音等位基因可包括未通过分析阈值的潜在等位基因。在一些实施方案中，噪音等位基因还可以包括来自未对准读段和任选地上文描述的未确认读段的计数得分。可以在252处收集噪音等位基因的计数得分，并在254处分析，以测定过多数目的读段是否指示与相应基因座的潜在问题。例如，可以将所有噪音等位基因的计数得分相加并与预定噪音阈值进行比较。噪音阈值可以基于读段的总数和/或历史数据。如果噪音阈值在254处通过，那么可以在256处提供警报，即该基因座具有过多的噪音。

[0176] 在一些实施方案中，可以在258处分析噪音等位基因用于质量控制。在具体实施方案中，STR基因座的噪音等位基因可以细分成具有与调用的等位基因相同长度的序列的噪音等位基因和具有与调用等位基因不同长度的序列的噪音等位基因。噪音等位基因的分离可以提供关于为何与相应基因座存在过多噪音的附加信息。

[0177] 在测定指定等位基因之后，在250处，方法240可以包括在对指定基因座产生基因型调用物之前进一步分析指定等位基因。基因型调用物通常包括杂合调用物(即两种不同的等位基因)或纯合调用物(即，一个观察到的等位基因)。对于杂合调用物，数据通常支持读段基本上均匀地成比例。如果两种等位基因在数据中没有基本上相等呈现，那么可能存在基因座的问题。因此，在一些实施方案中，方法240可以包括在260处分析调用的等位基因以测定调用的等位基因是否平衡或成比例。例如，可以计算调用的等位基因的比率以测定比率是否满足平衡阈值。仅作为实例，如果一个等位基因的计数得分(例如，读段计数)是另一个等位基因的计数得分(例如，读段计数)的小于50％或小于75％，那么等位基因可称为不平衡的。因此，可以在262处提供等位基因比例警报，指示等位基因不平衡。如下文讨论，可以用其它证据(例如，其它警报)分析等位基因比例警报，以测定样品是否包括多个来源。

[0178] 在一些实施方案中，方法240可以包括在264处测定基因座的拷贝数是否超过拷贝阈值。对于常染色体位点，拷贝数通常最多为2。对于非常染色体基因座，如X位点或Y位点，拷贝数可以不同。例如，Y-基因座的拷贝数可以为至多一个。X基因座的拷贝数可以为至多2。如下文描述，在一些情况下，可以预测样品的性别，然后当查询样品内是否存在多个来源时使用。

[0179] 因此，在264处测定可以包括获得指定基因座的拷贝数(例如0、1或2)，并将指定基因座的调用的等位基因的数目与拷贝数进行比较。如果调用的等位基因的数目超过拷贝数，那么可以在266处提供等位基因数目警报，即该基因座包括过多数目的等位基因。如下文讨论，可以用其它证据(例如，其它警报)分析等位基因数目警报，以测定样品是否包括多个来源。

[0180] 在268处，可以调用基因座的基因型。在250处，基因型调用物基于指定等位基因，并且通常将是一个或两个等位基因。然而，在一些实施方案中，基因型调用物将包括超过两个等位基因。具有超过两个等位基因的基因型调用物可以包括通知，其指示问题可以存在

于基因座处或一般与样品有关。在270处，可以产生样品报告，其包括对于预定集的遗传基因座的基因型调用物(若可能的话)。样品报告还可以包括已经由方法240或方法200(图11)鉴定的多个通知(例如，警报)。在一些实施方案中，可以提供基因座的基因型调用物以及指示符(indicator)，该指示符对阅读者通知关于基因座的潜在问题(例如，覆盖、噪音，等位基因退出、打滑等)。在其它实施方案中，如果鉴定了遗传基因座的某些警报(例如，覆盖或噪音)，那么不为遗传基因座提供基因型调用物。在一些实施方案中，样品报告可以包括调用的等位基因的序列，和任选地打滑产物和/或其它鉴定的潜在等位基因的序列。在一些实施方案中，样品报告可以包括相对于样品作为整体的置信得分。例如，如果存在大量一中靶未对准读段，那么样品报告可以指示样品可能质量较差。

[0181] 图13是显示了预测样品来源的性别的方法300的流程图。方法300假定样品来自单个源。如果随后测定样品来自多个来源，如下所述，那么可以除去性别预测。在一些实施方案中，在测定样品包括多个来源之后，该方法可以预测样品的所有来源是单个性别，如男性。

[0182] 方法300可以与方法240(图12)结合。方法300可以在从遗传基因座集测定每个遗传基因座的指定等位基因后执行。例如，方法300可以在图12中的步骤250针对遗传基因座集(或对于该集内的所有遗传基因座)内的多个遗传基因座的所有潜在等位基因发生之后执行。方法300包括在302处接收多个遗传基因座的基因座数据。基因座数据可以包括相应遗传基因座的一个或多个指定(或调用)等位基因。多个遗传基因座可以是基于样品的性别预期具有不同数目的等位基因的基因座。换言之，基因座数据可以对应于X和Y基因座。X基因座可以包括X染色体上的已知SNP或STR基因座。Y基因座可以包括Y染色体上的已知SNP或STR基因座。

[0183] 方法300可以包括在304处将每个Y-基因座的指定等位基因的数目在与样品是男性时的预期数目和/或在与样品是女性时的预期数目进行比较。预期数目可以是基于历史数据的预设数目。男性的样品的指定等位基因的预期数目可以基于一个或多个基因座出现在Y染色体上的次数。虽然这通常是一个，但是它可以是超过一个(例如，两个)。Y-基因座内女性的样品的指定等位基因的预期数目为零。

[0184] 任选地，方法300可以包括在306处将每个X-基因座的指定等位基因的数目在与样品是男性时的预期数目和/或在与样品是女性时的预期数目进行比较。X基因座内男性的样品的指定等位基因的预期数目通常为1，但如果基因座或等位基因在X染色体上出现超过一次，那么可以超过1。X基因座内的女性的样品的指定等位基因的预期数目通常为2，但如果基因座/等位基因在X染色体上出现超过一次，那么可以更多。

[0185] 方法300还包括在308处基于来自在304处比较的结果和/或来自在306处比较的结果预测样品的性别。理想地，每个Y-基因座在样品是男性时将包括1个指定等位基因，并且在样品是女性时会包括0个指定等位基因。同样地，每个X-基因座理想地在样品是男性时将包括1个指定等位基因，并且在样品是女性时会包括1或2个指定等位基因。然而，由于测序错误、污染、不适当的分析等，可能X-基因座和Y-基因座在预测样品的性别方面不一致。在一些情况下，分析可以考虑许多遗传基因座。例如，可以有约五(5)至十(10)个Y基因座和约二十(20)至三十(30)个X基因座。因此，尽管样品可以是男性，但是一个或多个Y-基因座可能具有0个指定等位基因。同样，尽管样品可以是女性，但是可能的是一个或多个Y-基因座

可能具有指定等位基因。

[0186] 因此，用于预测样品的性别的分析可以包括分析全部的证据以预测样品的性别。例如，分析可以包括计数下列至少一项：(i)与作为男性的样品一致的Y-基因座的数目；(ii)与作为女性的样品一致的Y-基因座的数目；(iii)与作为男性的样品一致的X基因座的数目；(iv)或作为男性的样品一致的X基因座的数目。在一些实施方案中，在308处，可以在分析中仅考虑Y-基因座的数目，或者，备选，可以仅考虑X-基因座的数目。在一些实施方案中，在308处，在分析中可以考虑X和Y基因座两者的数目。在一些实施方案中，可以对一个或多个X基因座和/或一个或多个Y-基因座给予比其它基因座更大的权重。

[0187] 作为一个实例，分析可以审阅10个Y-基因座。如果10个Y-基因座中的9个包括与作为男性的样品一致的指定等位基因，那么可以预测样品的性别是男性。如果10个Y-基因座之一包括指定等位基因，那么可以预测样品的性别是女性。在一些实施方案中，分析可以测定样品包括混合物。例如，如果在308处的分析测定Y-基因座的数目和X-基因座的数目支持男性和女性样品，那么可以预测来源的混合物。

[0188] 图14是显示检测样品是否包括来源的混合物的方法320的流程图。方法320可以与方法240(图12)合并，并且任选地，可以在预测样品的性别之后执行。方法300包括在322处接收遗传基因座集的每个遗传基因座的基因座数据。基因座数据可以包括相应遗传基因座的一个或多个指定或调用的等位基因。基因座数据还可以包括指定等位基因的计数得分(例如，读段计数)、噪音等位基因的计数得分和打滑产物的计数得分。可以如本文所述获得计数得分。

[0189] 对于每个遗传基因座，方法320可以包括在324处测定遗传基因座的拷贝数是否超过最大可允许的等位基因数(在下文中称为“最大等位基因数”)。如上所述，常染色体基因座的最大等位基因数通常为2。X基因座或Y基因座的最大等位基因数是基于样品(假设单一来源样品)是男性还是女性。如果样品是男性，Y-基因座的最大等位基因数为1，并且X基因座的最大等位基因数为1。如果样品是女性，Y-基因座的最大等位基因数为零，并且X-基因座的最大等位基因数为2。基于上述方法300，样品可以预测为男性或预测为女性。

[0190] 因此，在324处的测定可以包括获得遗传基因座的最大等位基因数目(例如，0、1、2)，并比较每个遗传基因座的拷贝数(即指定等位基因的数目)与相应的最大等位基因数。如果拷贝数超过最大等位基因数，可以为遗传基因座提供等位基因数目警报或标记。

[0191] 对于每个遗传基因座，方法300还可以包括在326处测定指定等位基因的等位基因比例是否不平衡。如上所述，遗传基因座的等位基因比例可以基于第一指定等位基因的计数得分(例如读段计数)和第二指定等位基因的计数得分(例如读段计数)。可以预期单个来源样品在遗传基因座是纯合的或在遗传基因座是杂合的。如果是杂合的，那么可以预期等位基因比例将为约1:1的比率。基本上不成比例的比率可以指示遗传基因座不是杂合的，或者样品包括超过一个来源。更具体地，计算的比率越大偏离1:1，遗传基因座是纯合的或者样品作为整体包括来源的混合物的可能性越大。如下文描述，测定样品包括来源的混合物是基于分析多个遗传基因座(例如所有调用的遗传基因座)。

[0192] 在一些实施方案中，在326处的测定可以包括计算基于遗传基因座的两个指定等位基因之间的计数得分的比率的平衡得分。如果平衡得分不在指定范围内，例如0.8:1.0至约1.2:1.0，那么平衡得分可以指示等位基因比例不平衡。如果测定遗传基因座具有不平衡

的等位基因比例，那么可以为遗传基因座产生等位基因比例警报。在一些实施方案中，可以将平衡得分与指定阈值进行比较，以测定等位基因比例是否不平衡。

[0193] 方法320还可以包括在328处分析在324处测定和在326处测定的结果以测定样品内是否存在多个来源。在328处的分析可以基于等位基因数目警报的数目和遗传基因座集的等位基因比例警报的数目。在一个实施方案中，可以计算警报的总数。如果警报的总数超过混合阈值，那么可以在具有多个来源方面标记样品。混合物阈值可以基于所分析的遗传基因座的数目(即，遗传基因座集中的遗传基因座的数目)。在具体实施方案中，混合物阈值可以基于调用的遗传基因座的数目。在一些实施方案中，混合物阈值基于历史数据或就特定测定法而言的知识。

[0194] 在一些实施方案中，遗传基因座集可以包括例如10、20、30、40、50、60、70、80、90、100个遗传基因座或更多。在具体实施方案中，遗传基因座集可以包括120、140、160、180、200个遗传基因座或更多。在更具体的实施方案中，遗传基因座集可以包括250、300、350个遗传基因座或更多。

[0195] 在一些实施方案中，混合物阈值是等于集内遗传基因座的预定百分比的预定值。预定百分比可以是至少例如5％、10％、15％、20％、25％、30％、35％、40％、50％、60％、70％或更多。

[0196] 在一些实施方案中，等位基因数目警报可以包括基于指定等位基因的数目的等位基因数目得分。更具体地，包含混合物的样品的可能性可以随着超出遗传基因座的可允许等位基因的最大数目的指定等位基因的数目增加而增加。为了例示，如果第一遗传基因座的指定等位基因的数目是三(3)，并且第二遗传基因座的指定等位基因的数目是(4)，那么可以对第二遗传基因座的等位基因数目得分分配比测定混合物是否存在时第一遗传基因座的等位基因数目得分更大的值(或更大的权重)。

[0197] 在一些实施方案中，等位基因比例警报可以包括基于遗传基因座的指定等位基因的比例的等位基因比例得分。更具体地，包含混合物的样品的可能性可以随着指定等位基因的比例变得更不成比例而增加。例如，如果第一遗传基因座的等位基因比例为1.3:1.0，并且第二遗传基因座的等位基因比例为2.0:1.0，那么可以对第二遗传基因座的等位基因数目得分分配比测定混合物是否存在时第一遗传基因座的等位基因比例得分更大的值(或更大重量)。

[0198] 在一些实施方案中，样品报告可以包括通知用户样品疑似包含多个来源的混合警报。在一些实施方案中，混合警报可伴随有通知用户混合警报中的置信度水平的置信度得分。置信度得分可以基于下列至少一项：等位基因数目警报的数目、与等位基因数目警报相关的等位基因数目得分、等位基因比例警报的数目和与等位基因比例警报相关的等位基因比例得分。

[0199] 图15显示了根据可以用于实施本文中阐述的各种方法的实施方案形成的系统400。例如，系统400可以用于实施方法100(图1)、150(图1)、160(图8)、200(图11)、240(图12)、300(图13)和340(图14)中的一种或多种。各种步骤可以由系统400自动化，如测序，而一个或多个步骤可以手动进行或者另外需要用户交互。在具体实施方案中，用户可以提供样品(例如，血液、唾液、毛发精液等)，并且系统400可以自动地制备、测序和分析样品，并提供样品来源的遗传概况。在一些实施方案中，系统400是位于一个地点的集成单机系统。在

其它实施方案中，系统的一个或多个部件相对于彼此远程定位。

[0200] 如显示，系统400包括样品发生器402、测序仪404和样品分析仪406。样品发生器402可以为指定的测序方案制备样品。例如，样品发生器可以制备用于SBS的样品。测序仪404可以进行测序以产生测序数据。如上所述，测序数据可以包括多个样品读段。每个样品读段可以包括样品序列。在具体实施方案中，样品读段形成从配对末端测序或更具体地不对称配对末端测序产生的读段对。

[0201] 样品分析仪406可以从测序仪404接收测序数据。图15包括根据一个实施方案形成的样品分析仪406的框图。样品分析仪406可用于例如分析测序数据以提供对特定基因座的基因型调用物或产生样品的遗传概况。样品分析仪406包括系统控制器412和用户接口414。系统控制器412以通信方式偶联到用户接口414，并且还可以以通信方式偶联到测序仪404和/或样品产生器402。

[0202] 在示例性实施方案中，系统控制器412包括一个或多个处理器/模块，其配置为根据本文中阐述的一个或多个方法处理和可选地分析测序数据。例如，系统控制器412可以包括一个或多个模块，配置为执行存储在一个或多个存储元件中的指令集(例如，存储在有形和/或非暂时性计算机可读存储介质上的指令，排除信号)来处理测序数据。指令集可以包括指示系统控制器412作为处理机器来执行诸如本文中描述的工作流、过程和方法的特定操作的各种命令。作为实例，样品分析仪406可以是或者包括台式计算机、膝上型计算机(laptop)、笔记本计算机(notebook)、平板计算机或智能电话。用户接口414可以包括使得个人(例如，用户)能够直接或间接地控制系统控制器412及其各种组件的操作的硬件、固件、软件或其组合。如显示，用户接口414包括操作人员显示器410。

[0203] 在例示的实施方案中，系统控制器412包括控制系统控制器412的操作的多个模块或子模块。例如，系统控制器412可以包括模块421-426和存储系统426，其与模块421-426中的至少一些通信。模块包括第一滤器模块421、对准器模块422、第二滤器模块423、打滑模块(stutter module)424、检测器模块425和分析模块426、系统400可以包括其它模块或模块的子模块，其配置为执行本文中描述的操作。第一滤器模块421配置成分析样品读段以测定样品读段是否是如本文中阐述的指定基因座的确认读段。对准器模块422配置成分析确认的读段并测定确认的读段是否是如本文中阐述的指定基因座的对准读段。第二滤器模块423配置为接收有效读段并测定有效读段是否代表如本文中阐述的相应基因座的潜在等位基因。打滑模块424配置为测定有效读段是否是如本文中阐述的另一等位基因的打滑产物。检测器模块425配置为测定是否应当针对如本文中阐述的相应基因座指示任何错误或警报。例如，检测器模块425可以测定基因座具有过多数目的未对准读段、低覆盖、过多数目的噪音等位基因、不平衡的等位基因、和/或来自不同来源的等位基因的混合物。分析模块426配置为测定如本文中阐述的遗传基因座的基因型。

[0204] 如本文中使用的，术语“模块”、“系统”或“系统控制器”可以包括操作以执行一个或多个功能的硬件和/或软件系统和电路。例如，模块、系统或系统控制器可以包括基于存储在有形和非暂时性计算机可读存储介质(诸如计算机存储器)上的指令来执行操作的计算机处理器、控制器或其它基于逻辑的设备。或者，模块、系统或系统控制器可以包括基于硬连线逻辑和电路执行操作的硬连线设备。附图中所示的模块、系统或系统控制器可以表示基于软件或硬连线指令操作的硬件和电路、引导硬件执行操作的软件、或其组合。模块、

系统或系统控制器可以包括或表示包括一个或多个处理器(诸如一个或计算机微处理器)和/或与一个或多个处理器(诸如一个或计算机微处理器)连接的硬件电路或电路。

[0205] 如本文中使用的,术语“软件”和“固件”是可互换的,并且包括存储在存储器中以由计算机执行的任何计算机程序,所述存储器包括RAM存储器、ROM存储器、EPROM存储器、EEPROM存储器和非易失性RAM(NVRAM)存储器。上述存储器类型仅是示例性的,并且因此不限于可用于存储计算机程序的存储器的类型。

[0206] 在一些实施方案中,“配置为”执行任务或操作的处理单元、处理器、模块或计算系统可以理解为特别地构造为执行任务或操作(例如,具有在其上存储或与其结合使用的一个或多个程序或指令,改编或意图用于执行任务或操作,和/或具有定制或意图用于执行任务或操作的处理电路的布置)。为了清楚和避免疑问的目的,通用计算机(其可以变得“配置为”执行任务或操作,若适当编程的话)未“配置为”执行任务或操作,除非或直到特定编程或结构上修改为执行任务或操作。

[0207] 图16A、16B和17A-17F显示了可以由本文中所述的实施方案产生的样品报告500、520的实例。可以将样品报告500、520存储在一个或多个文件中并通过通信网络传送。样品报告500、520可以例如显示在屏幕上或打印在纸上。图16A和16B仅显示了整个样品报告500的一部分。如显示,样品报告500可以包括对最初认为是单个来源样品的概述或概要分析。样品报告500包括用于STR分析的第一部分511和用于SNP分析的第二部分512。样品报告500可以利用标志或指示符510来确认样品是单一来源。

[0208] 样品报告500包括调用盒504的阵列502。每个调用盒504可以与指定的遗传基因座相关。例如,调用盒504A对应于遗传基因座釉原蛋白,并且调用盒504B对应于遗传基因座TPOX。每个调用盒504包括遗传基因座的基因型调用物506。釉原蛋白的基因型调用物506是X、Y,并且TPOX的基因型调用物是等位基因11,11。等位基因的名称可以基于常规命名或可以通过其它命名方案(例如,专有方案)决定。

[0209] 每个调用盒504可以指示标志或通知是否与遗传基因座相关联。例如,在图16中,颜色编码调用盒504以指示是否存在标志或通知。调用盒504A是灰色的,并且调用盒504C是橙色或红色。在备选实施方案中可以使用其它指示方法。在图16中,颜色编码的每个调用盒504包括标志508。标志508在上文在定义标志508的图例516中引用。例如,样品报告500包括用于“打滑”、“等位基因计数”、“不平衡”、“低覆盖”、“解读阈值”和“用户修改”的标志508。在例如本文描述的分析之后,可以将这些标志508分配到调用盒504。

[0210] 图17A-17F提供了遗传基因座的更详细的分析。在一些实施方案中,样品报告520可以是样品报告500的一部分(图16)。如显示,对每个遗传基因座分配图522,其在视觉上表示相应遗传基因座的数据。在例示的实施方案中,图522是条形图,但是其它图可以用于可视地表示数据。图522具体地显示了相对于不同等位基因的读段强度。读段强度可以是计数得分或基于如上文描述的计数得分。在一些实施方案中,读段强度/计数得分是读段计数。

[0211] 图522可以指示相对于读段强度(或计数得分)的解读阈值和分析阈值。例如,D2S441基因座具有解读阈值530和分析阈值532。解读和分析阈值530、532可以类似于上文描述的解读和分析阈值。如图17中显示,解读和分析阈值对于不同的基因座可以是不同的。例如,D21S11基因座具有大于PentaE基因座的解读阈值551的解读阈值550。如上文描述,解读阈值和/或分析阈值可以基于对应于指定基因座的读段总数(即,函数)。任选地,解读阈

值和/或分析阈值可以是特定基因座的读段计数的函数以及还可以是来自其它基因座的读段计数和/或来自其它样品的读段计数的函数。

[0212] 在一些实施方案中，图522还可以指示打滑产物。图522可以在视觉上区分打滑产物与真实等位基因。例如，D1S1656基因座包括分别与D1S1656的CE等位基因11、12和13相关的柱形541-543。柱形541-543可以指示相应等位基因的读段强度(或计数得分)。图17中显示的D1S1656基因座的等位基因历史上基于CE数据并且已经按照惯例标记为11、12和13。如图17中不同颜色(例如，蓝色和棕色)所示，D1S1656基因座的等位基因可能包括打滑产物。更具体地，条形541是打滑产物，并且不超过D1S1656基因座的解读阈值555。条形542包括条形部分546、547。条形部分546、547中的每个在视觉上表示读段强度。尽管对应于条形部分546、547的读段具有相同的序列长度，但是对应于条形部分546、547的读段具有不同的序列。条形部分546表示打滑产物。然而，如上文所述，由条形部分546表示的打滑产物可以是另一等位基因的如CE等位基因13的打滑产物。因此，颜色编码(或区分打滑产物和真实等位基因的其它指示符)可以通知或警告用户分析CE等位基因11、12、13的不同序列以提供遗传调用物的更确信的测定。在图17中，D1S1656基因座的遗传调用物是12/13。然而，在其它情况下，分析打滑产物的序列可以改变遗传调用物。更具体地，在一些情况下，使用已知CE过程的遗传调用物将是不正确的。本文中阐述的实施方案可以能够提供正确的遗传调用物。

[0213] 样品报告520还提供不同遗传基因座的标志或通知。图例524定义通知。作为一个实例，D21S11基因座具有“不平衡”和“等位基因计数”的标志。换言之，样品报告520向观察者指示等位基因的数目不是预期的，并且等位基因的平衡不是预期。观察者可能希望进一步调查关于D21S11基因座的数据。

[0214] 在一个实施方案中，提供了一种方法。方法包括接收包含多个样品读段的测序数据，所述样品读段具有相应的核苷酸序列。方法还包括基于所述核苷酸序列将所述样品读段分配到指定基因座，其中分配到相应的指定基因座的所述样品读段是所述相应的指定基因座的分配读段。方法还包括分析每个指定基因座的分配读段以鉴定所述分配读段内的相应的感兴趣区域(ROI)。每个所述ROI具有一个或多个系列的重复基序，其中相应系列的每个重复基序包含相同的核苷酸集。方法还包括基于所述ROI的序列，对具有多个指定读段的指定基因座分选所述分配读段，使得具有不同序列的ROI归为不同的潜在等位基因。每个潜在等位基因具有与所述指定基因座内的其它潜在等位基因的序列不同的序列。方法还包括针对具有多个潜在等位基因的指定基因座分析所述潜在等位基因的序列以确定所述潜在等位基因的第一等位基因是否是所述潜在等位基因的第二等位基因的疑似打滑产物(stutter product)。如果已经在所述第一和第二等位基因之间添加或丢失了相应序列内的k个重复基序，其中k是整数，那么所述第一等位基因是所述第二等位基因的疑似打滑产物。任选地，k等于1或2。

[0215] 在一个方面，其中针对所述具有多个潜在等位基因的指定基因座分析所述潜在等位基因的序列以确定所述第一等位基因是否是所述第二等位基因的疑似打滑产物，包括比较所述第一和第二等位基因的ROI的长度以确定所述第一和第二等位基因的ROI的长度是否相差一个重复基序或多个重复基序。

[0216] 在另一方面，针对所述具有多个潜在等位基因的指定基因座分析所述潜在等位基因的序列以确定所述第一等位基因是否是所述第二等位基因的疑似打滑产物，可以包括鉴

定已经添加或丢失的重复基序，并测定添加的或丢失的重复基序是否与所述相应序列中的相邻重复基序相同。

[0217] 在另一方面，如果在所述第一和第二等位基因的ROI的序列之间不存在其它错配，那么所述第一等位基因可以是所述第二等位基因的打滑产物。

[0218] 在另一方面，所述方法还可以包括产生基因型序型，所述基因型序型调用至少多个所述指定基因座的基因型，其中具有疑似打滑产物的所述指定基因座指示为具有所述疑似打滑产物。

[0219] 在另一方面，所述方法还可包括提供针对至少多个所述指定基因座的基因型调用物，其中所述基因型调用物中的至少一个指示对于所述至少一个基因型调用物的所述指定基因座存在疑似打滑产物。

[0220] 在另一方面，所述方法还可包括针对具有多个潜在等位基因的每个指定基因座计数对所述潜在等位基因调用的所述样品读段的总数。如果所述第一等位基因的样品读段小于所述第二等位基因的样品读段的指定阈值，那么所述第一等位基因可以是所述第二等位基因的打滑产物。任选地，指定阈值为第二等位基因的样品读段的约40％。任选地，如果所述第一等位基因的样品读段超过所述第二等位基因的样品读段的预定百分比，那么将所述疑似打滑产物指定为来自另一贡献者。任选地，如果第一等位基因的样品读段小于第二等位基因的样品读段的预定百分比，那么将疑似打滑产物指定为噪音。

[0221] 在另一方面，分配读段包括具有位于其间的相应重复区段的第一和第二保守侧翼区。对于每个分配读段，所述方法可以包括：(a)提供包含所述第一保守侧翼区和所述第二保守侧翼区的参考序列；(b)将所述参考序列的第一侧翼区的一部分与所述相应的指定读段对准；(c)将所述参考序列的第二侧翼区的一部分与所述相应的指定读段对准；并且(d)测定所述重复区段的长度和/或序列。

[0222] 任选地，在步骤(b)和(c)之一或两者中对准所述侧翼区的一部分包括：(i)通过使用与所述重复区段重叠或相邻的接种区的精确k聚体匹配来测定所述分配读段上的相应保守侧翼区的位置；并且(ii)将所述侧翼区与所述分配读段对准。

[0223] 任选地，所述接种区包含所述保守侧翼区的高复杂性区。例如，所述高复杂性区可以包括与所述重复区段充分不同以避免错误对准的序列。作为另一个例子，所述高复杂性区可以包含具有多种多样的核苷酸混合物的序列。

[0224] 任选地，所述接种区避免所述相应的保守侧翼区的低复杂性区。例如，所述低复杂性区可以具有基本上类似于多个所述重复基序的序列。

[0225] 任选地，所述接种区直接邻近所述重复区段；所述接种区可以包括所述重复区段的一部分；或者所述接种区与所述重复区段偏移。

[0226] 在另一方面，所述样品读段可以是具有正向和反向引物序列的PCR扩增子。

[0227] 在另一方面，将所述样品读段分配到所述指定基因座可以包括鉴定与PCR扩增子的引物序列对应的所述样品读段的序列。

[0228] 在另一方面，测序数据可来自合成测序(SBS)测定。

[0229] 在另一方面，ROI包括短串联重复(STR)。任选地，所述STR选自下列中的至少一种：CODIS常染色体STR基因座、CODIS Y-STR基因座、EU常染色体STR基因座或EU Y-STR基因座。

[0230] 在一个实施方案中，提供了一种方法，其包括接收测序数据，所述测序数据包含与

遗传基因座集对应的扩增子的多个样品读段。所述样品读段包含读段对，其中相应扩增子的每个读段对包含所述相应扩增子的第一和第二读段。所述第一和第二读段中的每个具有相应的读段序列。所述方法还包括基于对所述第一读段的读段序列的分析，鉴定所述第一读段的潜在遗传基因座。所述潜在遗传基因座来自所述遗传基因座集。方法还包括针对具有至少一个潜在基因座的每个所述第一读段，确定所述第一读段是否与每个所述潜在遗传基因座的参考序列对准。如果所述第一读段与仅一个遗传基因座的参考序列对准，那么所述方法包括确定所述第一读段包括所述一个遗传基因座的潜在等位基因。如果所述第一读段与超过一个参考序列对准，那么所述方法包括确定第一读段包括具有与所述第一读段最佳对准的参考序列的遗传基因座的潜在等位基因。如果所述第一读段不与参考序列对准，那么所述方法包括将所述第一读段指定为未对准读段，并分析所述未对准读段以从与所述未对准读段最佳拟合的潜在遗传基因座鉴定遗传基因座。方法还包括产生遗传概况，其包含至少多个所述遗传基因座的调用的基因型，其中所述调用的基因型基于所述相应遗传基因座的潜在等位基因。遗传概况还包含具有未对准读段的遗传基因座的一个或多个通知。

[0231] 在一个方面，所述通知中的至少一个包括与所述相应遗传基因座有关的置信度得分。置信度得分可以基于与所述相应遗传基因座最佳拟合的非对准读段的数目，其中未对准读段数目越大指示所述调用的基因型越不可信。

[0232] 在另一方面，分析所述未对准读段以从与所述未对准读段最佳拟合的所述潜在遗传基因座鉴定遗传基因座可以包括分析所述未对准读段的鉴定子序列以鉴定与所述鉴定子序列最佳拟合的所述遗传基因座。

[0233] 在另一方面，鉴定子序列包括引物序列的至少一部分。

[0234] 在另一方面，鉴定所述第一读段的潜在遗传基因座包括确定所述第一读段的引物序列有效匹配与所述潜在遗传基因座相关的序列。

[0235] 在另一方面，通过不对称配对末端测序产生所述测序数据。

[0236] 在另一方面，所述方法还可包括分析所述未对准读段以确定是否存在潜在的等位基因退出。

[0237] 在另一方面，所述方法还可包括分析所述未对准读段以确定测定法的良好。

[0238] 在另一方面，所述方法还可以包括分析所述未对准读段以确定所述未对准读段是否指示嵌合体。

[0239] 在另一方面，所述方法还可包括分析所述未对准读段以确定引物二聚体的数目。

[0240] 在另一方面，确定所述第一读段包含所述遗传基因座的潜在等位基因可以包括确认与所述第一读段对应的所述第二读段也与所述遗传基因座相关。

[0241] 在另一方面，所述方法还可以包括分析所述未对准读段以确定所述未对准读段是一中靶读段还是对中靶读段。所述对中靶读段可以具有第一和第二鉴定子序列，其与数据库的第一和第二选择序列有效匹配。所述一中靶读段可以仅具有与数据库的第一选择序列有效匹配的第一鉴定子序列。

[0242] 在一个实施方案中，提供了一种方法，其包括接收测序数据，所述测序数据具有与遗传基因座集对应的扩增子的多个样品读段。所述样品读段包含读段对，其中相应扩增子的每个读段对包含所述相应扩增子的第一和第二读段。所述第一和第二读段中的每个具有相应的读段序列。方法还包括基于对所述第一读段的读段序列的分析，鉴定所述第一读段

的潜在遗传基因座。所述潜在遗传基因座来自所述遗传基因座集。方法还包括针对具有至少一个潜在基因座的每个所述第一读段，确定所述第一读段是否与每个所述潜在遗传基因座的参考序列对准。方法还包括将不与参考序列对准的所述第一读段指定为未对准读段。方法还包括分析所述未对准读段以从与所述未对准读段最佳拟合的所述潜在遗传基因座鉴定遗传基因座。方法还包括分析未对准读段以确定对于所述最佳拟合遗传基因座是否存在潜在的等位基因退出。

[0243] 在一个方面，所述方法还可以包括分析所述未对准读段以确定所述未对准读段是一中靶读段还是对中靶读段。所述对中靶读段可以具有第一和第二鉴定子序列，其与数据库的第一和第二选择序列有效匹配。所述一中靶读段可以仅具有与数据库的第一选择序列有效匹配的第一鉴定子序列。分析所述未对准读段以确定对于所述最佳拟合遗传基因座是否存在所述潜在等位基因退出可以基于多个对中靶读段。

[0244] 在一个实施方案中，提供了一种方法，其包括接收多个遗传基因座的每个遗传基因座的读段分布。所述读段分布包含多个潜在等位基因，其中每个潜在等位基因具有等位基因序列和读段计数。所述读段计数表示测定为包括所述潜在等位基因的来自测序数据的样品读段的数目。所述方法还可以包括针对所述多个遗传基因座的每个遗传基因座，鉴定具有最大读段计数的所述读段分布的潜在等位基因之一。方法还可以包括针对所述多个遗传基因座的每个遗传基因座，确定所述最大读段计数是否超过解读阈值。如果所述最大读段超过所述解读阈值，那么所述方法包括分析对应遗传基因座的潜在等位基因以调用所述遗传基因座的基因型。如果所述最大读段小于所述解读阈值，那么所述方法包括产生所述遗传基因座具有低覆盖的警报。方法还包括产生遗传概况和具有低覆盖的遗传基因座的警报，所述遗传概况具有调用基因型的每个所述遗传基因座的基因型。

[0245] 在一个方面，分析所述相应遗传基因座的潜在等位基因以调用所述遗传基因座的基因型还可以包括如果所述遗传基因座具有相对于彼此不充足的比例的多个潜在等位基因，那么产生所述遗传基因座不平衡的警报。

[0246] 在另一个方面，方法还可以包括针对所述多个遗传基因座中的每个遗传基因座，确定所述潜在等位基因的读段计数是否通过分析阈值。所述分析阈值可以比所述解读阈值更容易通过。

[0247] 在另一方面，具有未通过所述解读阈值的读段计数的所述潜在等位基因指定为噪音等位基因，所述方法还包括将所述噪音等位基因的读段计数的总和与噪音阈值进行比较，并且如果所述总和超过所述噪音阈值，那么产生所述遗传基因座包括过多噪音的警报。

[0248] 任选地，遗传基因座包括短串联重复(STR)基因座和单核苷酸多态性(SNP)基因座。

[0249] 在一个实施方案中，提供了一种方法，其包括：(a)接收遗传基因座的读段分布。所述读段分布包含多个潜在等位基因，其中每个潜在等位基因具有等位基因序列和计数得分。所述计数得分基于测定为包含所述潜在等位基因的来自测序数据的样品读段的数目。方法还包括(b)基于一个或多个所述潜在等位基因的计数得分确定所述遗传基因座是否具有低覆盖。如果所述遗传基因座具有低覆盖，那么所述方法包括产生所述遗传基因座具有低覆盖的通知。如果所述遗传基因座不具有低覆盖，那么所述方法包括分析所述潜在等位基因的计数得分以测定所述遗传基因座的基因型。方法还包括(d)产生包括遗传基因座的

基因型的遗传概况或遗传基因座具有低覆盖的警报。

[0250] 在一个方面，确定遗传基因座是否具有低覆盖可以包括确定潜在等位基因的计数得分中的一个或多个是否通过解读阈值。如果计数得分中的至少一个超过解读阈值，那么该方法还可以包括分析相应遗传基因座的潜在等位基因以调用遗传基因座的基因型。如果计数得分无一通过解读阈值，那么该方法可以包括产生遗传基因座具有低覆盖的通知。

[0251] 在另一方面，确定遗传基因座是否具有低覆盖包括确定潜在等位基因的计数得分中的一个或多个是否通过分析阈值。如果计数得分中的至少一个超过分析阈值，那么该方法还可以包括分析相应遗传基因座的潜在等位基因以调用遗传基因座的基因型。如果计数得分无一通过分析阈值，那么该方法还可以包括产生遗传基因座具有低覆盖的通知。

[0252] 在另一方面，确定遗传基因座是否具有低覆盖包括将遗传基因座的对准读段的总数与读段阈值进行比较。如果对准读段的总数通过读段阈值，那么该方法可包括分析相应遗传基因座的潜在等位基因以调用遗传基因座的基因型。如果对准读段的总数未通过读段阈值，那么该方法可以包括产生遗传基因座具有低覆盖的通知。

[0253] 在另一方面，每个计数得分是等于相应的潜在等位基因的读段的值。

[0254] 在另一方面，每个计数得分是基于读段计数和遗传基因座的读段总数的函数。

[0255] 在另一方面，每个计数得分是基于读段计数和先前获得的遗传基因座的数据的函数。

[0256] 在另一方面，每个计数得分是基于来自样品的其它遗传基因座的读段的函数。

[0257] 在另一方面，每个计数得分是基于来自其它样品的遗传基因座的读段计数的函数。

[0258] 在另一方面，分析遗传基因座的潜在等位基因以调用遗传基因座的基因型还包括将遗传基因座的多个潜在等位基因与遗传基因座的可允许等位基因的预定最大数目进行比较，并且如果潜在等位基因的数目超过允许的等位基因的预定最大数目，那么警告遗传基因座具有过多数目的等位基因。

[0259] 在另一方面，分析所述遗传基因座的潜在等位基因以调用所述遗传基因座的基因型还可以包括如果所述遗传基因座具有相对于彼此不充足的比例的多个潜在等位基因，那么产生所述遗传基因座不平衡的通知。

[0260] 在另一方面，所述方法还可以包括确定潜在等位基因的计数得分是否通过分析阈值。分析阈值可以比解读阈值更容易通过。任选地，具有未通过分析阈值的计数得分的潜在等位基因指定为噪音等位基因。该方法还可以包括将噪音得分与噪音阈值进行比较，并且如果噪音得分超过噪音阈值，那么产生所述遗传基因座包括过多噪音的警报。噪音得分可以基于噪音等位基因的计数得分。

[0261] 任选地，遗传基因座是短串联重复(STR)基因座或单核苷酸多态性(SNP)基因座之一。

[0262] 在另一方面，所述方法包括对多个遗传基因座重复(a)-(c)，其中产生遗传概况包括调用每个遗传基因座的基因型或为具有低覆盖的每个遗传基因座提供通知。

[0263] 在一个实施方案中，提供了一种方法，其包括接收遗传基因座的读段分布。读段分布包括多个潜在等位基因，其中每个潜在等位基因具有等位基因序列和读段计数。读段计数表示来自分配到遗传基因座的测序数据的样品读段的数目。该方法还可以包括测定每个

潜在等位基因的计数得分。计数得分可以基于潜在等位基因的读段计数。该方法还可以包括确定潜在等位基因的计数得分是否通过分析阈值。如果相应的潜在等位基因的计数得分未通过分析阈值，那么该方法包括弃去相应的潜在等位基因。如果相应的潜在等位基因的计数得分通过分析阈值，那么该方法包括将潜在等位基因指定为遗传基因座的指定等位基因。

[0264] 在一个方面，弃去相应的潜在等位基因包括将潜在等位基因指定为噪音等位基因。该方法还可以包括确定噪音等位基因的计数得分是否共同通过噪音阈值。如果计数得分共同通过噪音阈值，那么该方法可以包括产生遗传基因座具有过多噪音的警报。

[0265] 在另一方面，每个计数得分是等于相应潜在等位基因的读段计数的值。

[0266] 在另一方面，每个计数得分是基于读段和基因座的读段总数的函数。

[0267] 在另一方面，每个计数得分是基于读段计数和先前获得的遗传基因座的数据的函数。

[0268] 在另一方面，所述方法还可以包括将指定等位基因的数目与所述遗传基因座的可允许的等位基因的预定最大数目进行比较，并且如果指定等位基因的数目超过可允许等位基因的所述预定最大数目，那么产生所述遗传基因座具有过多数目的等位基因的警报。

[0269] 在另一方面，所述方法还包括如果所述遗传基因座具有相对于彼此不充足的比例的多个指定等位基因，那么产生所述遗传基因座不平衡的警报。

[0270] 任选地，遗传基因座包括短串联重复(STR)基因座和单核苷酸多态性(SNP)基因座。

[0271] 在一个实施方案中，提供了一种方法，其包括接收遗传基因座的读段分布。读段分布包括多个潜在等位基因，其中每个潜在等位基因具有等位基因序列和读段计数。读段计数表示来自分配到遗传基因座的测序数据的样品读段的数目。该方法还包括确定读段计数是否超过分析阈值。如果相应的潜在等位基因的读段计数小于分析阈值，那么该方法包括将相应的潜在等位基因指定为噪音等位基因。如果相应的潜在等位基因的读段计数通过分析阈值，那么该方法包括将潜在等位基因指定为遗传基因座的等位基因。该方法还包括确定噪音等位基因的读段计数的总和是否超过噪音阈值。如果总和超过噪音阈值，那么该方法包括产生遗传基因座具有过多噪音的警报。

[0272] 在一个方面，所述方法还可以包括将指定等位基因的数目与所述遗传基因座的可允许的等位基因的预定最大数目进行比较，并且如果指定等位基因的数目超过可允许等位基因的所述预定最大数目，那么产生所述遗传基因座具有过量数目的等位基因的警报。

[0273] 在另一方面，所述方法还可以包括如果所述遗传基因座具有相对于彼此不充足的比例的多个指定等位基因，那么产生所述遗传基因座不平衡的警报。

[0274] 任选地，遗传基因座包括短串联重复(STR)基因座和单核苷酸多态性(SNP)基因座。

[0275] 在一个实施方案中，提供了一种方法，其包括接收多个遗传基因座的每个遗传基因座的基因座数据。基因座数据包括相应遗传基因座的一个或多个指定等位基因。每个指定等位基因基于从测序数据获得的读段。该方法还包括针对所述多个遗传基因座的每个遗传基因座，确定所述相应遗传基因座的指定等位基因的数目是否大于所述相应遗传基因座的可允许等位基因的预定最大数目。所述方法可以包括如果指定等位基因的数目超过可允

许等位基因的预定最大数目，那么产生等位基因数目警报。方法还包括针对所述多个遗传基因座的每个遗传基因座，确定所述指定等位基因的等位基因比例是否不充足。等位基因比例可以基于所述指定等位基因的读段计数。所述方法还可以包括如果等位基因比例不平衡，那么产生等位基因比例警报，方法还可以包括基于遗传基因座集的等位基因数目警报和等位基因比例警报的数目，确定所述样品包含多个来源的混合物。

[0276] 在一个方面，确定样品包括多个来源的混合物包括测定警报的总数通过混合物阈值。任选地，混合物阈值基于遗传基因座集中的遗传基因座的数目。任选地，混合物阈值是等于集内遗传基因座的预定百分比的预定值。

[0277] 在另一方面，产生等位基因数目警报包括提供基于指定等位基因的数目的等位基因数目得分。测定样品包括多个来源的混合物可以包括分析等位基因数目得分。任选地，所述样品包含混合物的可能性随着超出可允许等位基因的最大数目的指定等位基因的数目增加而增加。

[0278] 在另一方面，产生等位基因比例警报包括提供基于等位基因比例的等位基因比例得分。测定样品包括多个来源的混合物包括分析等位基因比例得分。任选地，所述样品包含混合物的可能性随着所述等位基因之间的不成比例增加而增加。

[0279] 任选地，遗传基因座包括短串联重复(STR)基因座和单核苷酸多态性(SNP)基因座。

[0280] 在一个实施方案中，提供了一种方法，其包括接收多个Y基因座的基因座数据。基因座数据包括Y基因座的指定等位基因。每个指定等位基因基于从测序数据获得的读段。该方法还包括将每个Y-基因座的指定等位基因数目与Y-基因座的等位基因的预期数目进行比较。该方法还包括基于来自比较操作的结果产生样品是男性或女性的预测。任选地，遗传基因座包括短串联重复(STR)基因座和单核苷酸多态性(SNP)基因座。

[0281] 在一个或多个实施方案中，提供了一种系统，其包括配置为实施本文中所阐述的一个或多个权利要求的样品分析仪。

[0282] 在本申请中，已经参考了各种出版物。专利和/或专利申请。这些出版物的公开内容在此通过引用整体并入。

[0283] 如本文中使用的，术语“包含”、“包括”和“具有”等旨在是开放式的，不仅包括所列举的要素，而且可能包括额外的元素。

[0284] 应当理解，上述描述旨在是示例性的，而不是限制性的。例如，上述实施方案(和/或其方面)可以彼此组合使用。另外，可以进行许多修改以使特定情况或材料适应本发明的教导而不脱离其范围。本文中描述的尺寸、材料类型、各种部件的取向以及各种部件的数目和位置旨在限定某些实施方案的参数，并且决不是限制性的，而仅仅是示例性实施方案。在审阅上述描述之后，在权利要求书的精神和范围内的许多其它实施方案和修改对于本领域技术人员将是显而易见的。因此，本发明的范围应当参考所附权利要求书以及这些权利要求所赋予的等同物的全部范围来确定。

[0285] 如本说明书中所使用的，短语“在示例性实施方案中”、“在一些实施方案中”、“在具体实施方案中”等是指所描述的实施方案是可以根据本申请形成或执行的实施方案的实例。该短语不旨在将发明主题限于该实施方案。更具体地，本发明主题的其它实施方案可以不包括用具体实施方案描述的叙述的特征或结构。

[0286] 在所附权利要求书中,术语“包括(including)”和“其中(in which)”用作相应术语“包含(comprising)”和“其中(wherein)”的简单英语等同物。此外,在所附权利要求书中,术语“第一”、“第二”和“第三”等仅仅用作标记,并且不旨在对其对象强加数目要求。此外,所附权利要求书的限制不是以装置加功能格式撰写的,并且不意在基于35U.S.C.§112(f)的解释,除非和直到此类权利要求限制明确地使用短语“装置”,随后是没有进一步结构的功能声明。

[0287] 所附权利要求书记载了本申请的一个或多个实施方案,并且因此并入本申请的描述中。

序列表

<110> 亿明达股份有限公司

<120> 用于分析核酸测序数据的方法和系统

<130> IP-1270-PCT

<140> PCT/US2015/050129
<141> 2015-09-15

<150> 62/052,189
<151> 2014-09-18

<160> 25

<170> PatentIn version 3.5

<210> 1
<211> 58
<212> DNA
<213> 人工序列

<220>
<223> 人工序列的描述: 合成
寡核苷酸

<400> 1
tagatagata gatagataga tagatagata gatagataga tagataggtg tgtgtgtg 58

<210> 2
<211> 11
<212> DNA
<213> 人工序列

[0001]

<220>
<223> 人工序列的描述: 合成
寡核苷酸

<400> 2
tctatcagct a 11

<210> 3
<211> 39
<212> DNA
<213> 人工序列

<220>
<223> 人工序列的描述: 合成
寡核苷酸

<400> 3
tctatctatc tatctatcta tctatctatc atctatcta 39

<210> 4
<211> 28
<212> DNA
<213> 人工序列

<220>
<223> 人工序列的描述: 合成
寡核苷酸

<400> 4
tctatctatc tatctatcta tctatcta 28

<210> 5
<211> 35

<212> DNA
<213> 人工序列

<220>
<223> 人工序列的描述: 合成
寡核苷酸

<400> 5
tctatctatc tatctatcta tctatctatc agcta 35

<210> 6
<211> 14
<212> DNA
<213> 人工序列

<220>
<223> 人工序列的描述: 合成
寡核苷酸

<400> 6
agaaaaagag agag 14

<210> 7
<211> 13
<212> DNA
<213> 人工序列

<220>
<223> 人工序列的描述: 合成
寡核苷酸

<400> 7
gagaccttgt ctc 13

[0002]

<210> 8
<211> 41
<212> DNA
<213> 人工序列

<220>
<223> 人工序列的描述: 合成
寡核苷酸

<400> 8
gaaagaaaga gaaaaagaaa agaaatagta gcaactgtta t 41

<210> 9
<211> 48
<212> DNA
<213> 人工序列

<220>
<223> 人工序列的描述: 合成
寡核苷酸

<400> 9
caagaaagaa aaaaaagaaa gaaaaaacga aggggaaaaa aagagaat 48

<210> 10
<211> 17
<212> DNA
<213> 人工序列

<220>
<223> 人工序列的描述: 合成
寡核苷酸

<400> 10
caagaaagaa aaaaaga 17

<210> 11
<211> 17
<212> DNA
<213> 人工序列

<220>
<223> 人工序列的描述: 合成
寡核苷酸

<400> 11
caagaaagaa aaaaaag 17

<210> 12
<211> 17
<212> DNA
<213> 人工序列

<220>
<223> 人工序列的描述: 合成
寡核苷酸

<400> 12
caagaaagaa aaaaaga 17

<210> 13
<211> 18
<212> DNA
<213> 人工序列

[0003]

<220>
<223> 人工序列的描述: 合成
寡核苷酸

<400> 13
caagaaagaa aaaaagaa 18

<210> 14
<211> 19
<212> DNA
<213> 人工序列

<220>
<223> 人工序列的描述: 合成
寡核苷酸

<400> 14
caagaaagaa aaaaaagaa 19

<210> 15
<211> 18
<212> DNA
<213> 人工序列

<220>
<223> 人工序列的描述: 合成
寡核苷酸

<400> 15
caagaaagaa aaaaagaa 18

<210> 16
<211> 62
<212> DNA

<213> 人工序列

<220>
<223> 人工序列的描述: 合成
寡核苷酸

<400> 16
tagatagata gatagataga tagatagata gatagataga tagatagata gatgtgtgtg 60

tg 62

<210> 17
<211> 58
<212> DNA
<213> 人工序列

<220>
<223> 人工序列的描述: 合成
寡核苷酸

<400> 17
tagatagata gatagataga tagatagata gatagataga tagatagatg tgtgtgtg 58

<210> 18
<211> 62
<212> DNA
<213> 人工序列

<220>
<223> 人工序列的描述: 合成
寡核苷酸

<400> 18
[0004] tagatagata gatagataga tagatagata gatagataga tagatagata ggtgtgtgtg 60

tg 62

<210> 19
<211> 21
<212> DNA
<213> 人工序列

<220>
<223> 人工序列的描述: 合成
寡核苷酸

<400> 19
agaaagaaag aaagaaagaa a 21

<210> 20
<211> 61
<212> DNA
<213> 人工序列

<220>
<223> 人工序列的描述: 合成
寡核苷酸

<400> 20
agaaagaaag aaagaaagag aaaaagagag gaaagaaaga gaaaaagaaa agaaatagta 60

g 61

<210> 21
<211> 35
<212> DNA
<213> 人工序列

<220>
<223> 人工序列的描述: 合成
寡核苷酸

<400> 21
gaaagaaaga gaaaaagaaa agaaatagta gcaac 35

<210> 22
<211> 49
<212> DNA
<213> 人工序列

<220>
<223> 人工序列的描述: 合成
寡核苷酸

<400> 22
agaaagaaaa agagagagga aagaaagaaa aaaagaaaaa aaatagtag 49

<210> 23
<211> 27
<212> DNA
<213> 人工序列

<220>
<223> 人工序列的描述: 合成
寡核苷酸

[0005]

<400> 23
gaaagaaaga gaaaaagaaa agaaata 27

<210> 24
<211> 52
<212> DNA
<213> 人工序列

<220>
<223> 人工序列的描述: 合成
寡核苷酸

<400> 24
tctatctatc tatctatcta tctatctatc tatctatcta tctatctatc ta 52

<210> 25
<211> 52
<212> DNA
<213> 人工序列

<220>
<223> 人工序列的描述: 合成
寡核苷酸

<400> 25
tctatctgtc tatctatcta tctatctatc tatctatcta tctatctatc ta 52

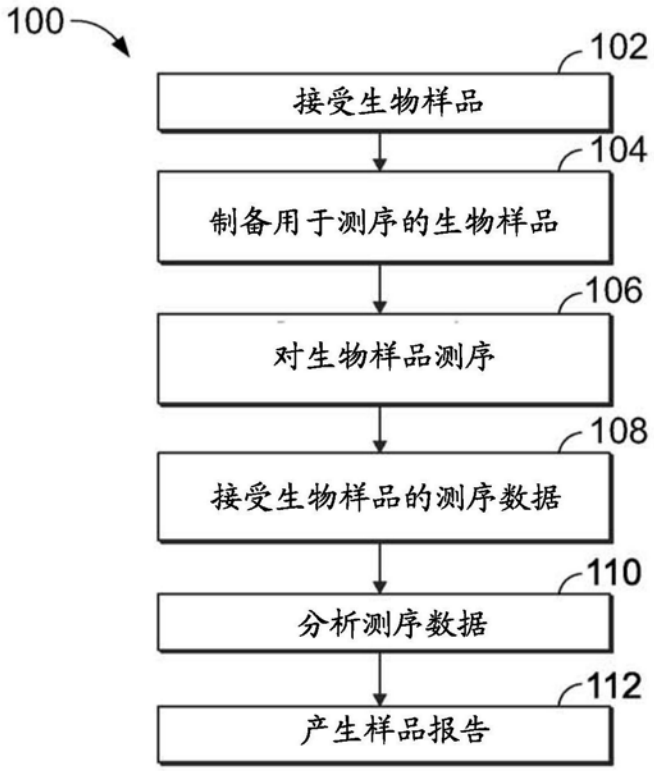

100
102
接受生物样品
104
制备用于测序的生物样品
106
对生物样品测序
108
接受生物样品的测序数据
110
分析测序数据
112
产生样品报告

图1

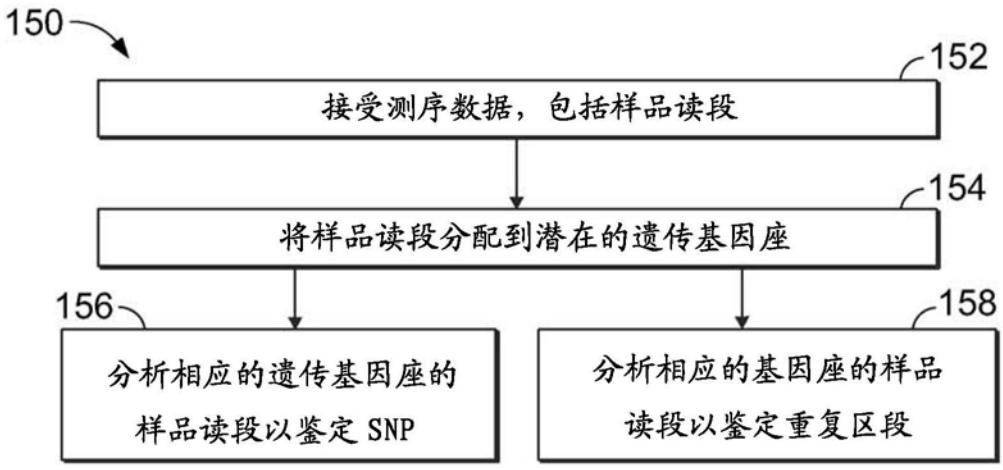

150
152
接受测序数据，包括样品读段
154
将样品读段分配到潜在的遗传基因座
156
分析相应的遗传基因座的样品读段以鉴定 SNP
158
分析相应的基因座的样品读段以鉴定重复区段

图2

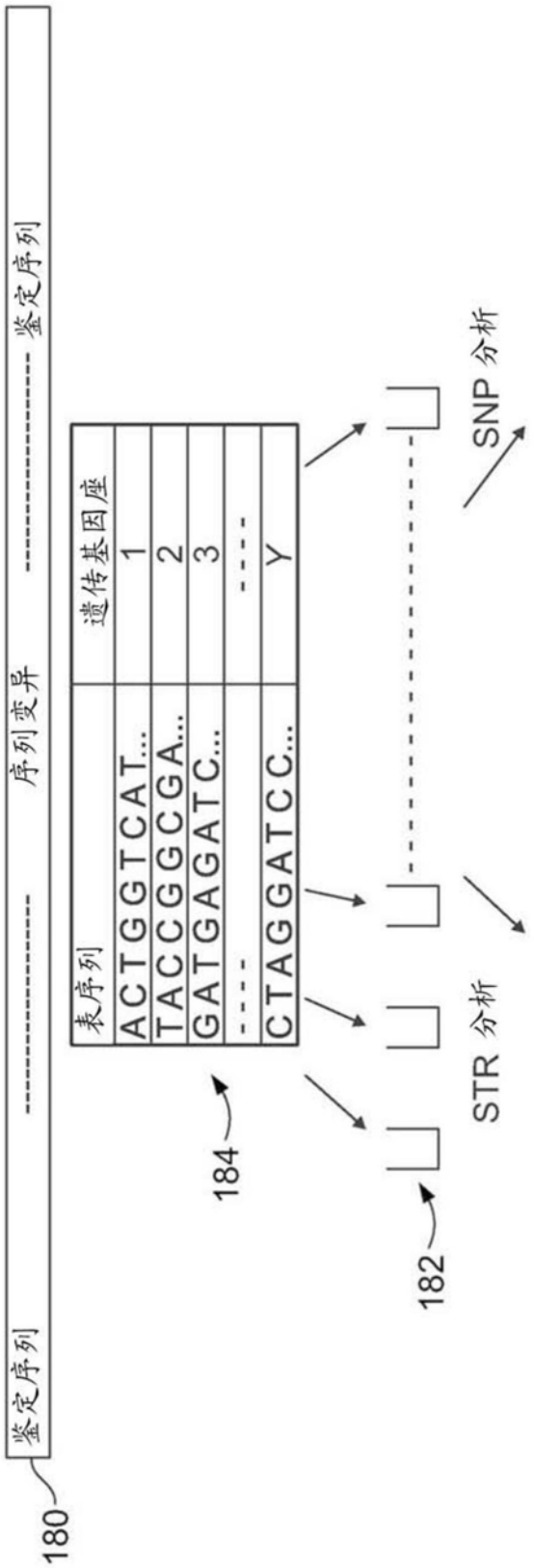

鉴定序列
序列变异
鉴定序列
180
表序列
遗传基因座
ACTGGTCAT...
1
TACCGGGCGA...
2
GATGAGATC...
3
- - - -
- - - -
CTAGGATCC...
Y
184
182
STR 分析
SNP 分析

图3

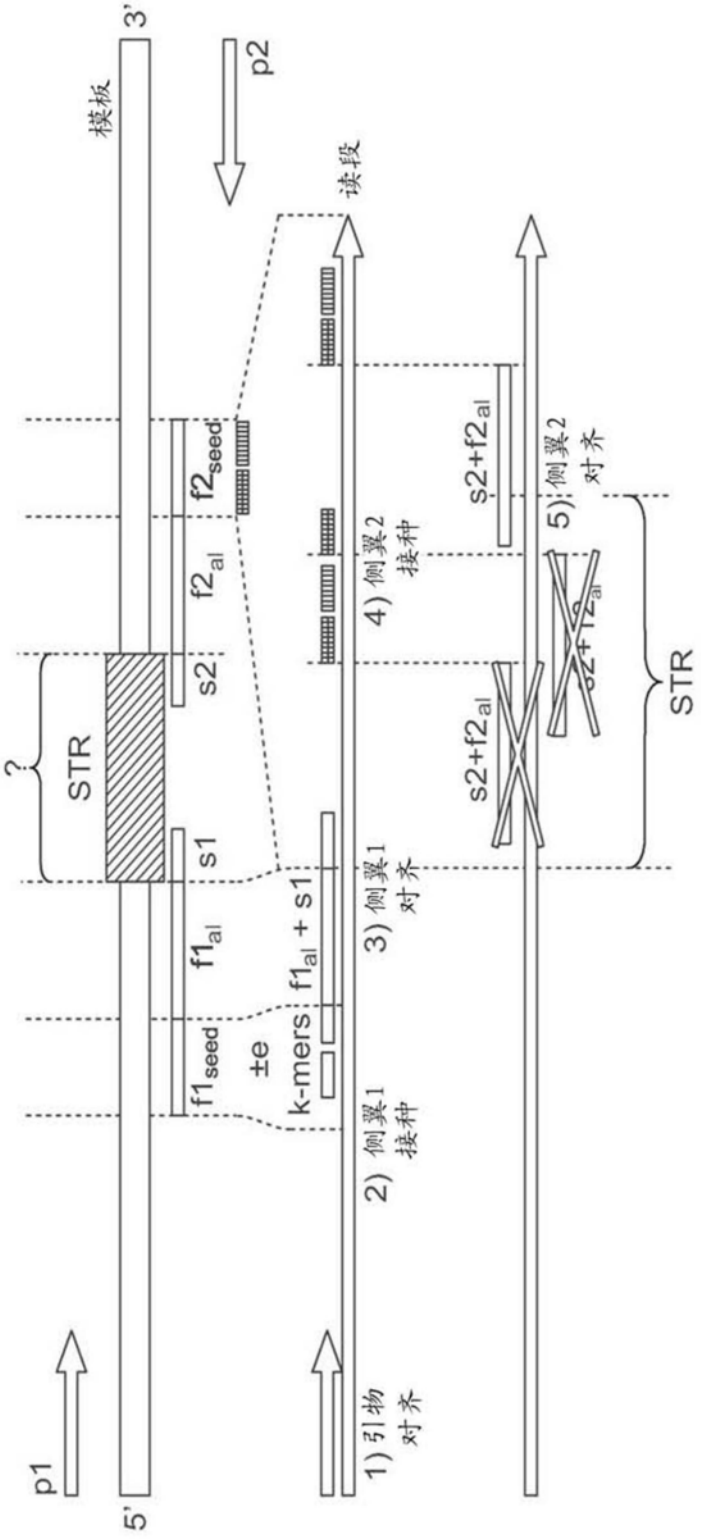

3'
模板
p2
读段
f2seed
f2al
s2
?
STR
s1
f1al
f1seed
±e
k-mers
f1al + s1
4) 侧翼2 接种
s2+f2al
5) 侧翼2 对齐
3) 侧翼1 对齐
2) 侧翼1 接种
1) 引物 对齐
p1
5'

图4

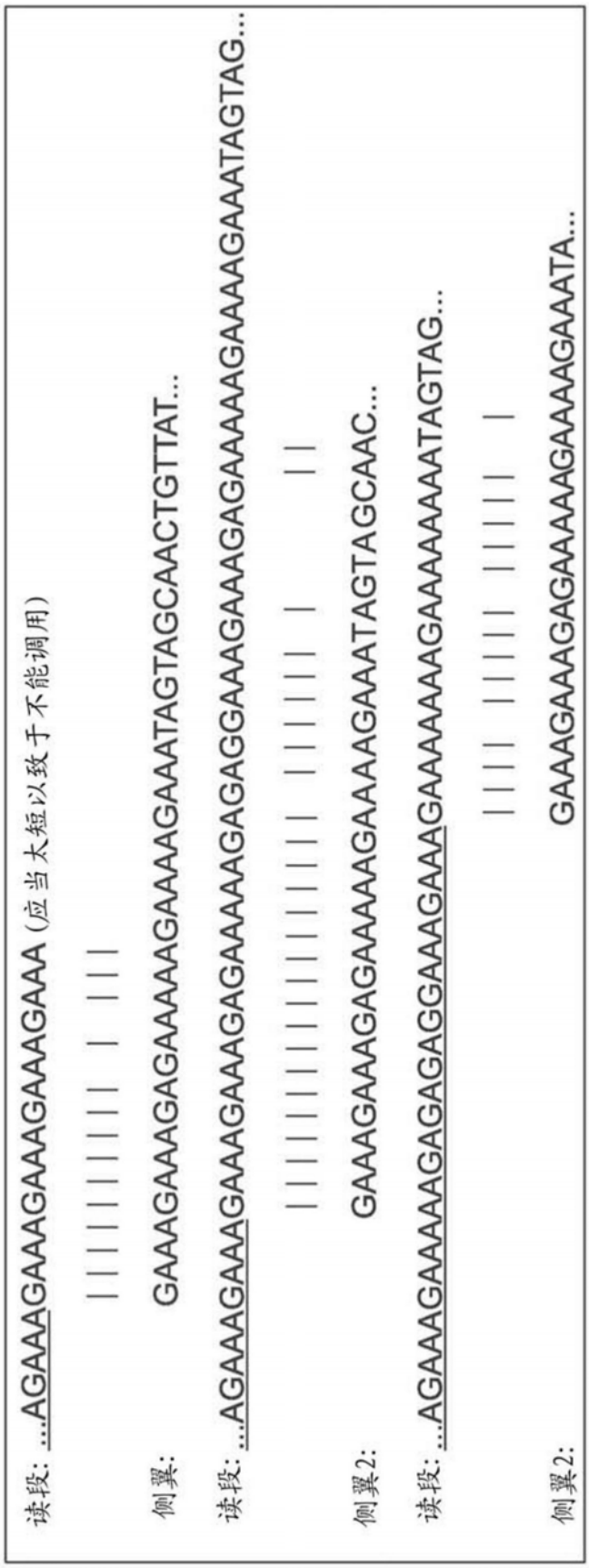

读段: ...AGAAAGAAAGAAAGAAA (应当太短以致于不能调用)
侧翼: GAAAGAAAGAGAAAAAGAAATAGTAGCAACTGTTAT...
读段: ...AGAAAGAAAGAAAGAGAGGAAAGAAAGAGAAAAGAAAATAGTAG...
侧翼2: GAAAGAAAGAGAAAAAGAAATAGTAGCAAC...
读段: ...AGAAAGAAAAAGAGAGGAAAGAAAAAAAAAGAAAAAAAATAGTAG...
侧翼2: GAAAGAAAGAGAAAAAGAAATA...

图5

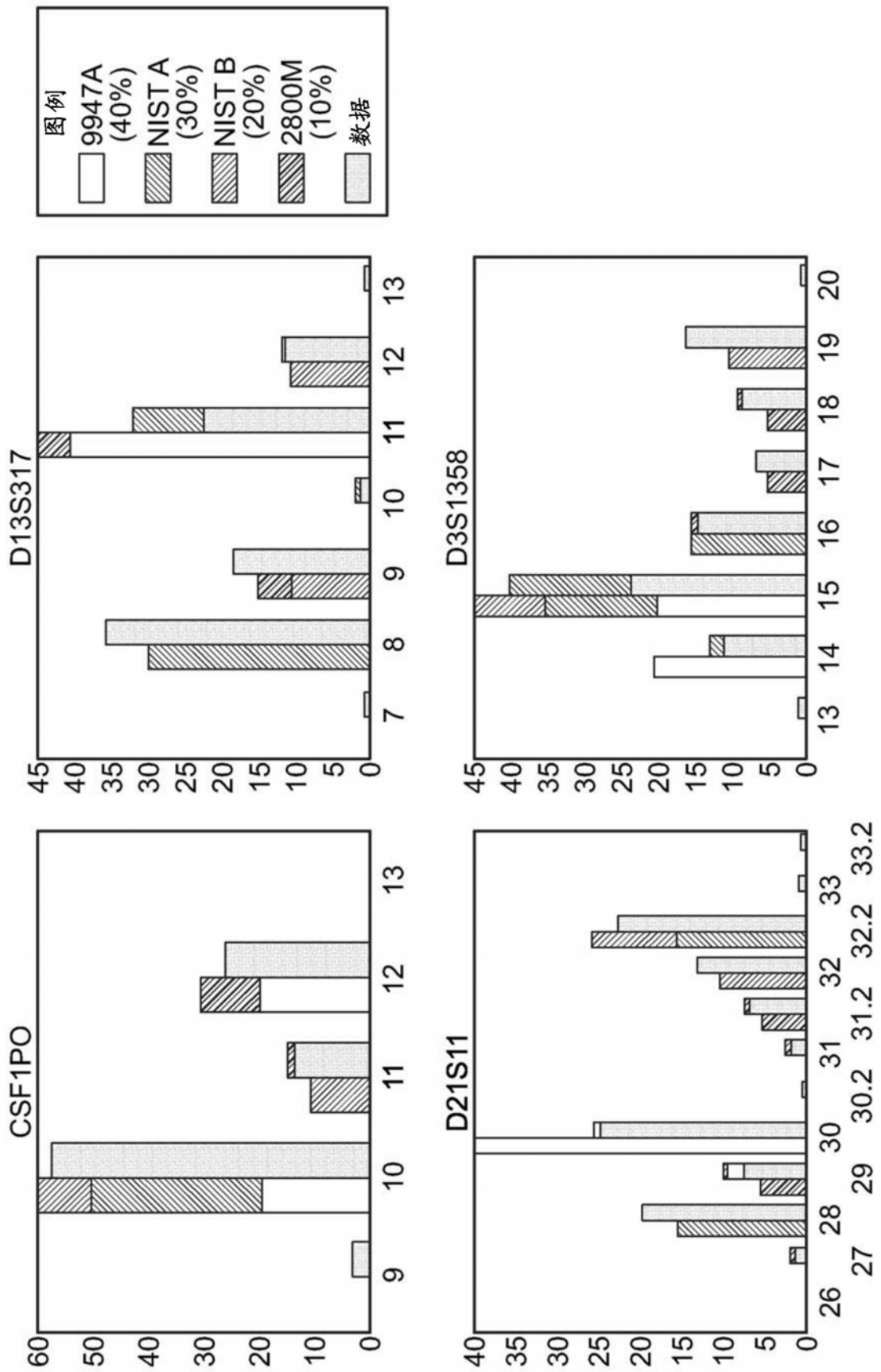

图例
9947A (40%)
NIST A (30%)
NIST B (20%)
2800M (10%)
数据
D13S317
D3S1358
CSF1PO
D21S11

图6A

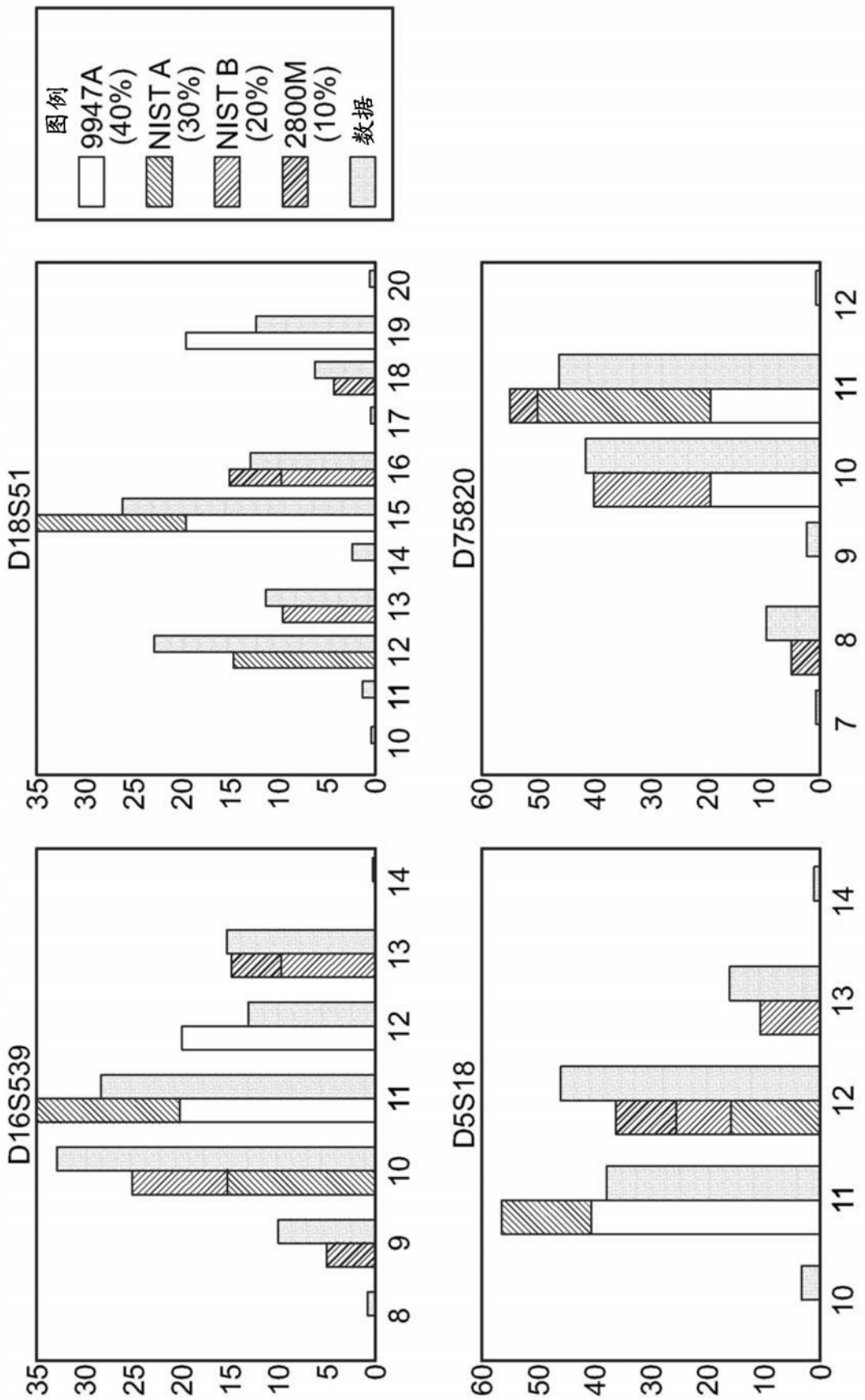

图例
9947A (40%)
NIST A (30%)
NIST B (20%)
2800M (10%)
数据
D18S51
D75820
D16S539
D5S18

图6B

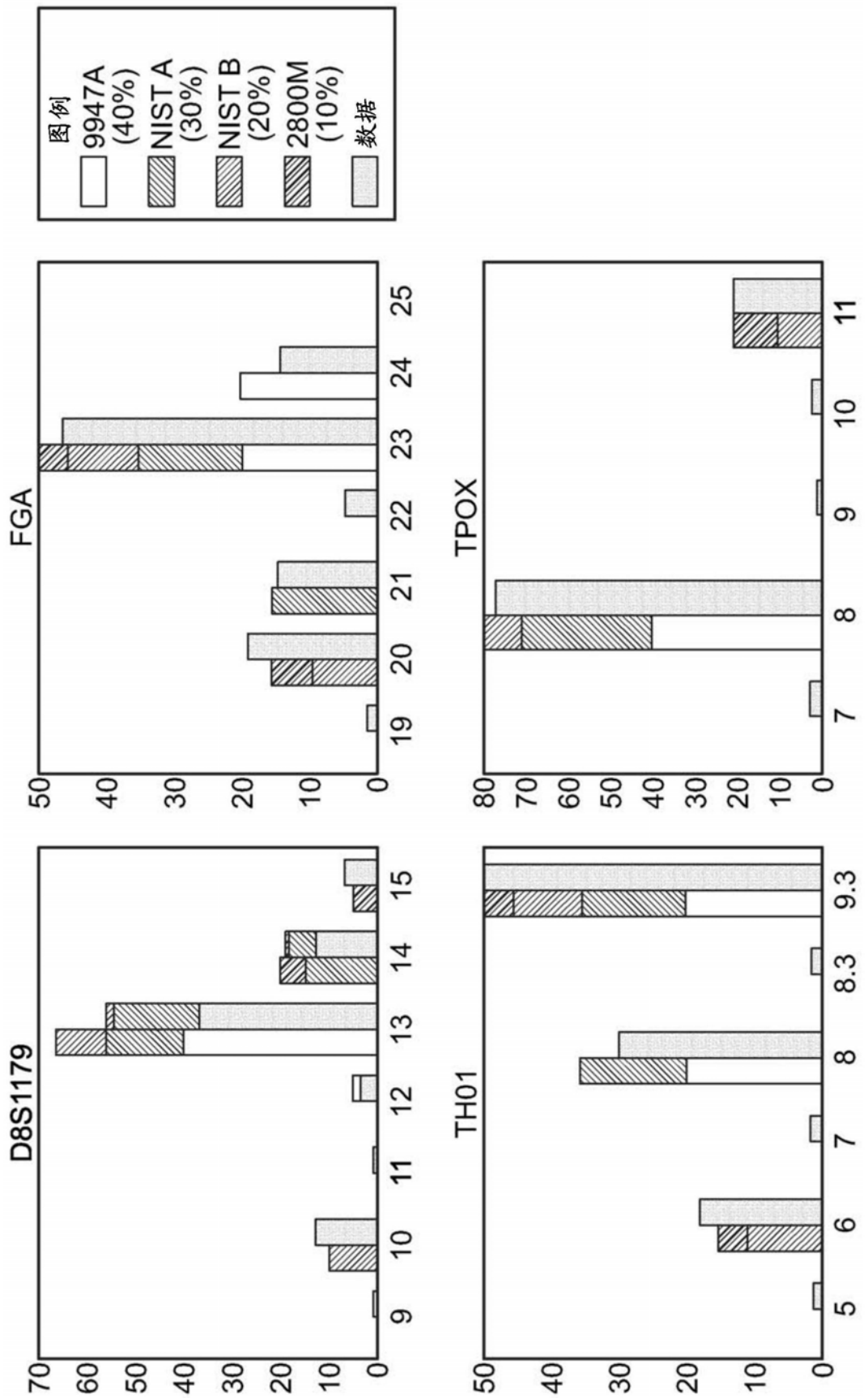

图例
9947A (40%)
NIST A (30%)
NIST B (20%)
2800M (10%)
数据
FGA
D8S1179
TPOX
TH01

图6C

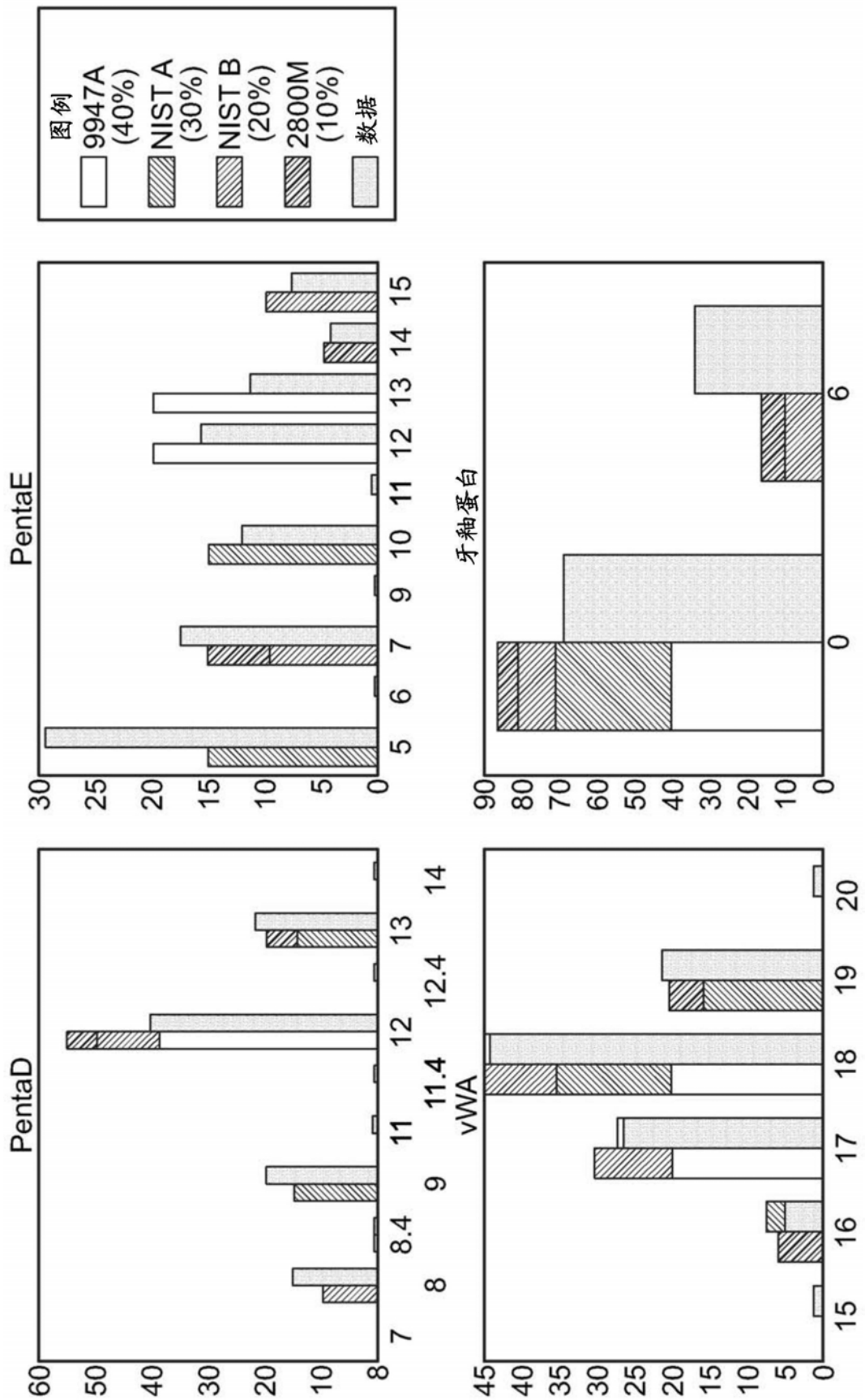

图例
9947A (40%)
NIST A (30%)
NIST B (20%)
2800M (10%)
数据
PentaE
PentaD
牙釉蛋白
vWA

图6D

DNA 样品	CSF1PO	D351358	D7S820	D16S539	D18S51	FGA	PentaE	TH01
9947A	10,12	14,15	10,11	11,12	15,19	23,24	12,13	8,9.3
2800M	12	17,18	8,11	9,13	16,18	20,23	7,14	6,9.3
NIST A	10	15,16	11	10,11	12,15	21,23	5,10	8,9.3
NIST B	10,11	15,19	10	10,13	13,16	20,23	7,15	6,9.3
NIST C	10,12	16,18	10,12	10	16,19	24,26	12,13	6,8

A

vWA	D5S818	D8S1179	D13S317	D21S11	PentaD	TPOX
19,20	11	13,13[1]	11	30	12	8
16,19	12	14,15	9,11	29,31.2	12,13	11
18,19	11,12	13,14	8	28,32.2	9,13	8
17,18	12,13	10,13	9,12	32,32.2	8,12	8,11
16,18	10,11	10,17	11	29,30	10,11	11

A

[1]半数重复中发现的SNP: 46% [TCTA]13, 56% [TCTA]1[TCTG]1[TCTA]11

图7

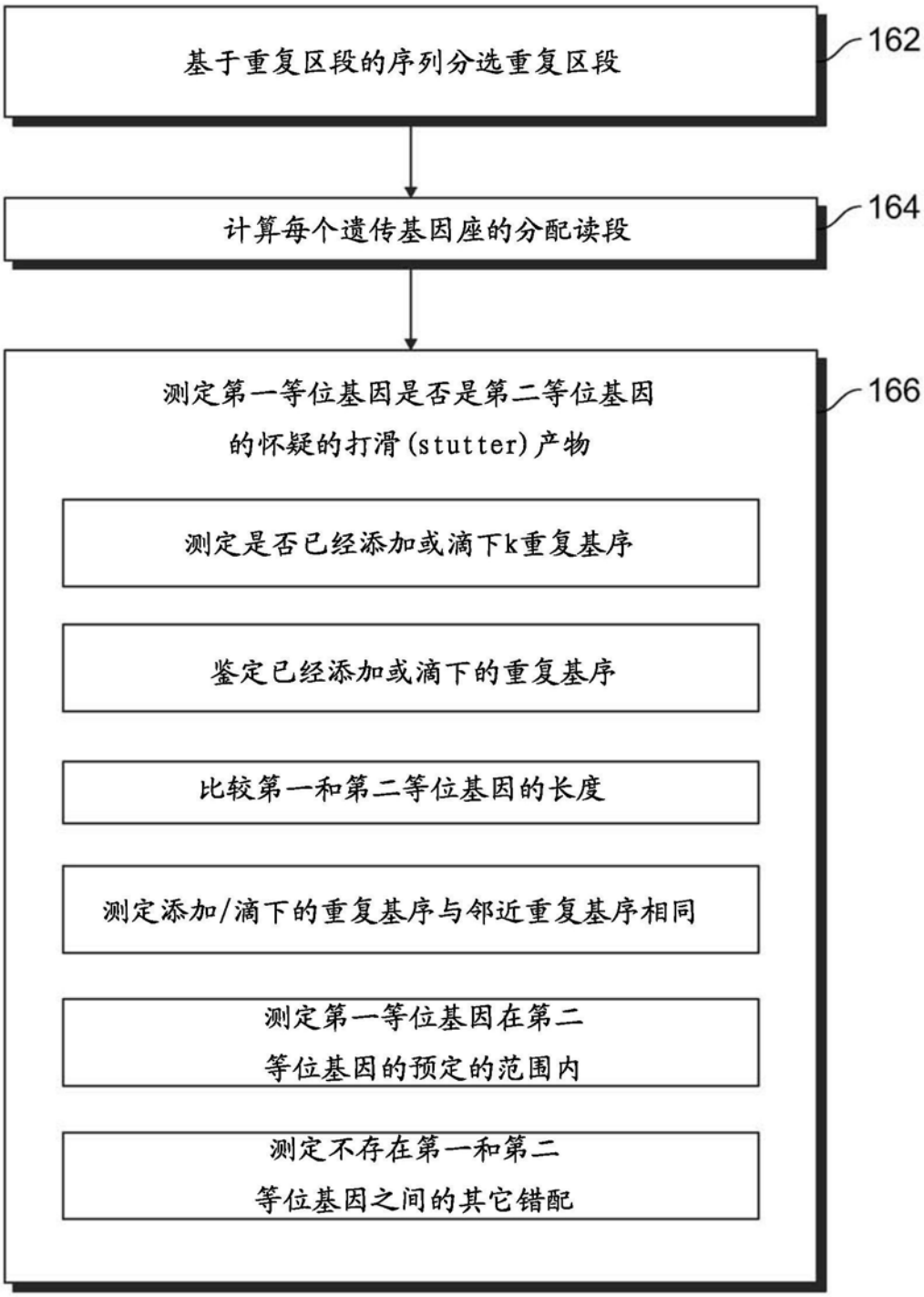
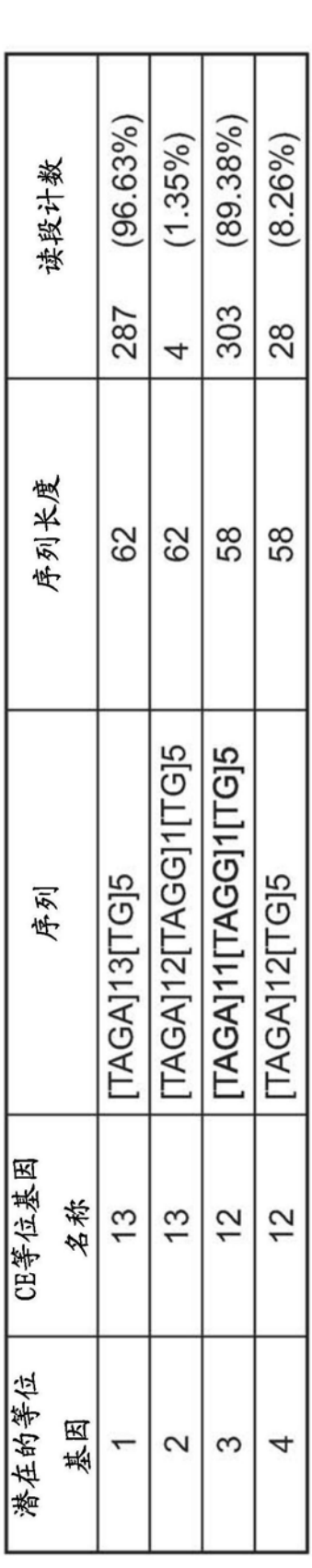

基于重复区段的序列分选重复区段
162
计算每个遗传基因座的分配读段
164
测定第一等位基因是否是第二等位基因
的怀疑的打滑(stutter)产物
166
测定是否已经添加或滴下k重复基序
鉴定已经添加或滴下的重复基序
比较第一和第二等位基因的长度
测定添加/滴下的重复基序与邻近重复基序相同
测定第一等位基因在第二
等位基因的预定的范围内
测定不存在第一和第二
等位基因之间的其它错配

图8



潜在的等位基因	CE等位基因名称	序列	序列长度	读段计数
1	13	[TAGA]13[TG]5	62	287 (96.63%)
2	13	[TAGA]12[TAGG]1[TG]5	62	4 (1.35%)
3	12	[TAGA]11[TAGG]1[TG]5	58	303 (89.38%)
4	12	[TAGA]12[TG]5	58	28 (8.26%)

图9

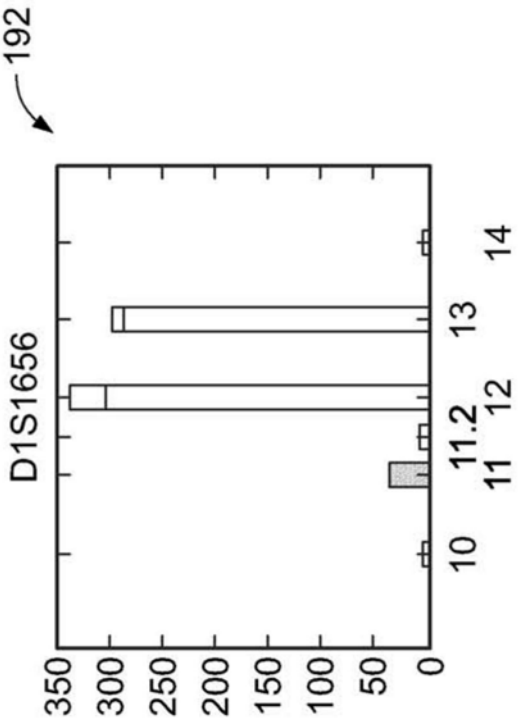


图10

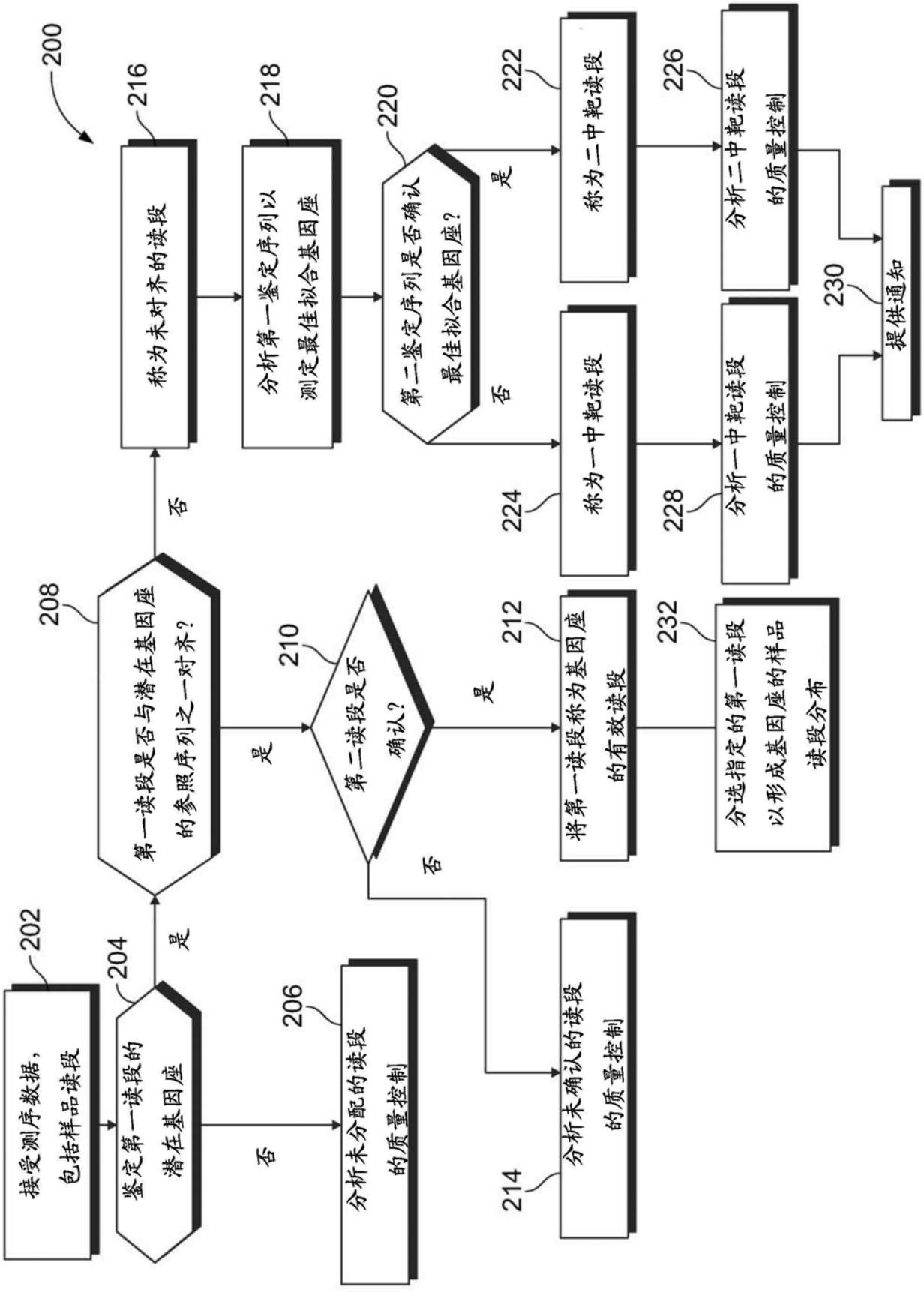

200
接受测序数据，包括样品读段
202
鉴定第一读段的潜在基因座
204
是
否
分析未分配的读段的质量控制
206
第一读段是否与潜在基因座的参照序列之一对齐?
208
是
否
第二读段是否确认?
210
是
否
将第一读段称为基因座的有效读段
212
分选指定的第一读段以形成基因座的样品读段分布
232
分析未确认的读段的质量控制
214
称为未对齐的读段
216
分析第一鉴定序列以测定最佳拟合基因座
218
第二鉴定序列是否确认最佳拟合基因座?
220
是
否
称为二中靶读段
222
称为一中靶读段
224
分析二中靶读段的质量控制
226
分析一中靶读段的质量控制
228
提供通知
230

图11

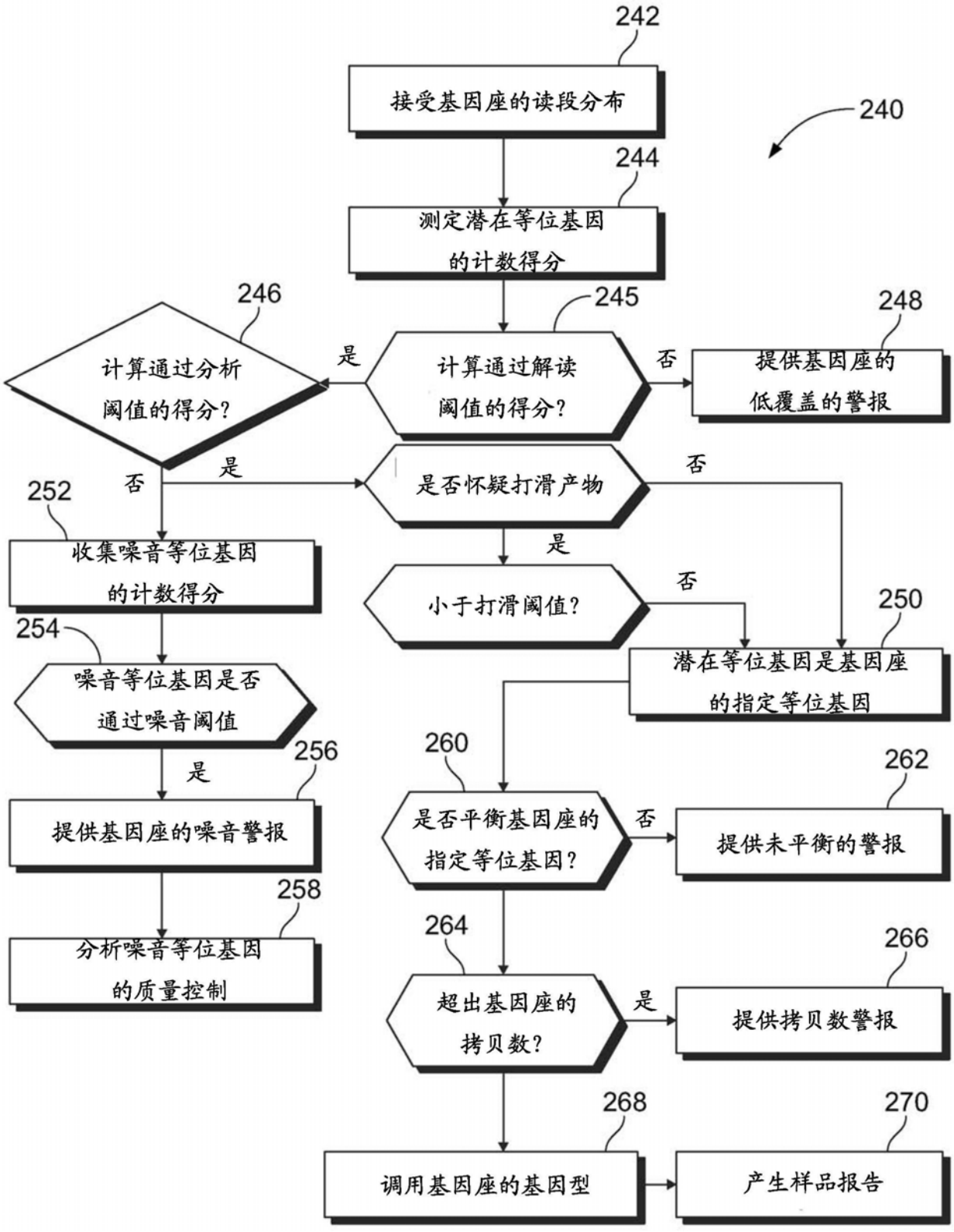

242
接受基因座的读段分布
240
244
测定潜在等位基因的计数得分
246
计算通过分析阈值的得分?
是
245
计算通过解读阈值的得分?
否
248
提供基因座的低覆盖的警报
否
是
是否怀疑打滑产物
否
252
收集噪音等位基因的计数得分
是
小于打滑阈值?
否
250
潜在等位基因是基因座的指定等位基因
254
噪音等位基因是否通过噪音阈值
256
是
提供基因座的噪音警报
260
是否平衡基因座的指定等位基因?
否
262
提供未平衡的警报
258
分析噪音等位基因的质量控制
264
超出基因座的拷贝数?
是
266
提供拷贝数警报
268
调用基因座的基因型
270
产生样品报告

图12

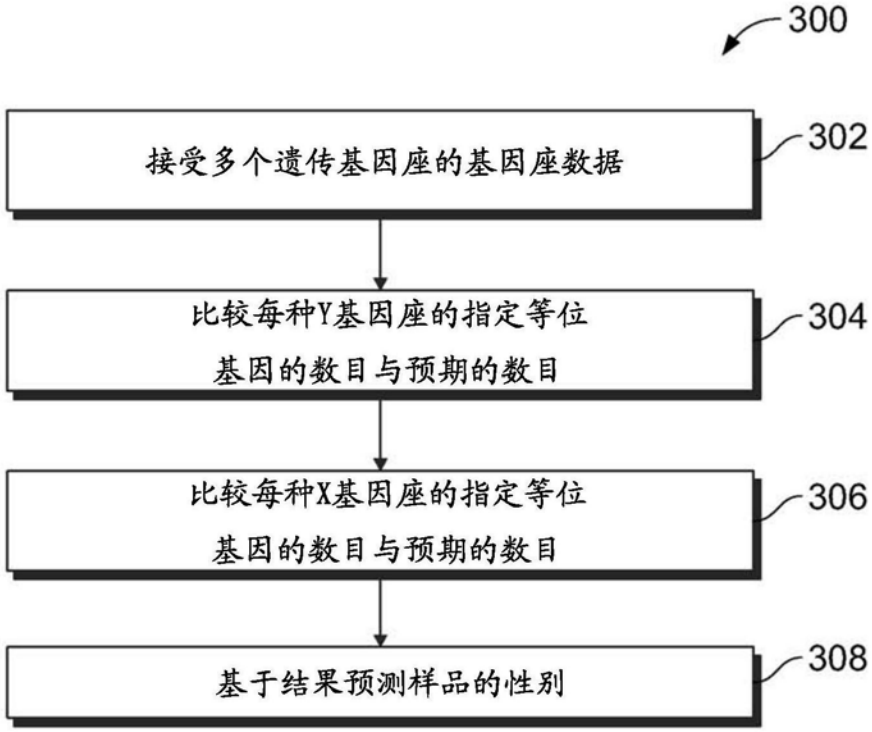

300
接受多个遗传基因座的基因座数据
302
比较每种Y基因座的指定等位基因的数目与预期的数目
304
比较每种X基因座的指定等位基因的数目与预期的数目
306
基于结果预测样品的性别
308

图13

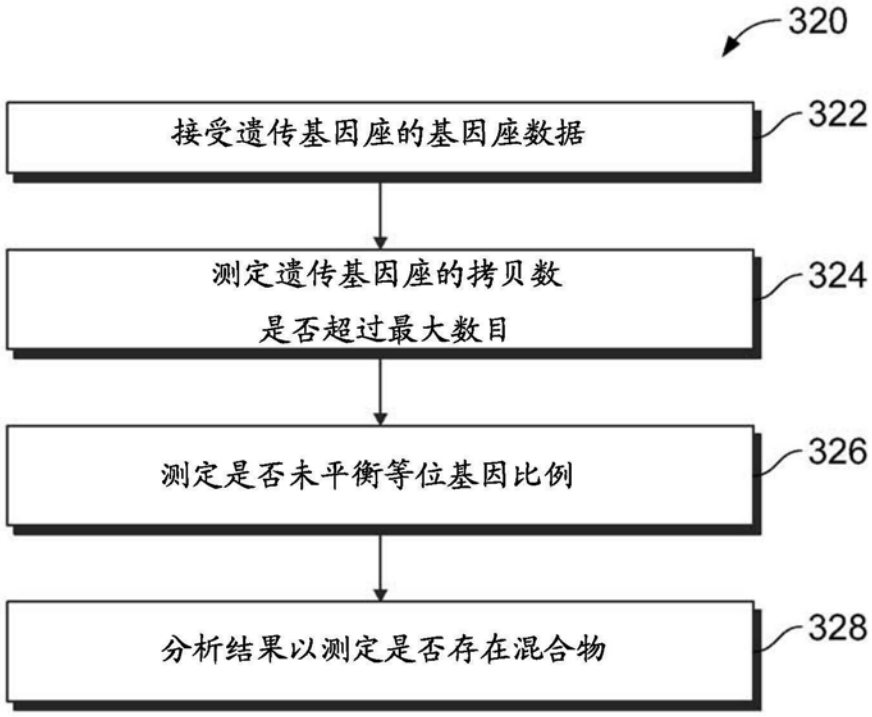

320
接受遗传基因座的基因座数据
322
测定遗传基因座的拷贝数是否超过最大数目
324
测定是否未平衡等位基因比例
326
分析结果以测定是否存在混合物
328

图14

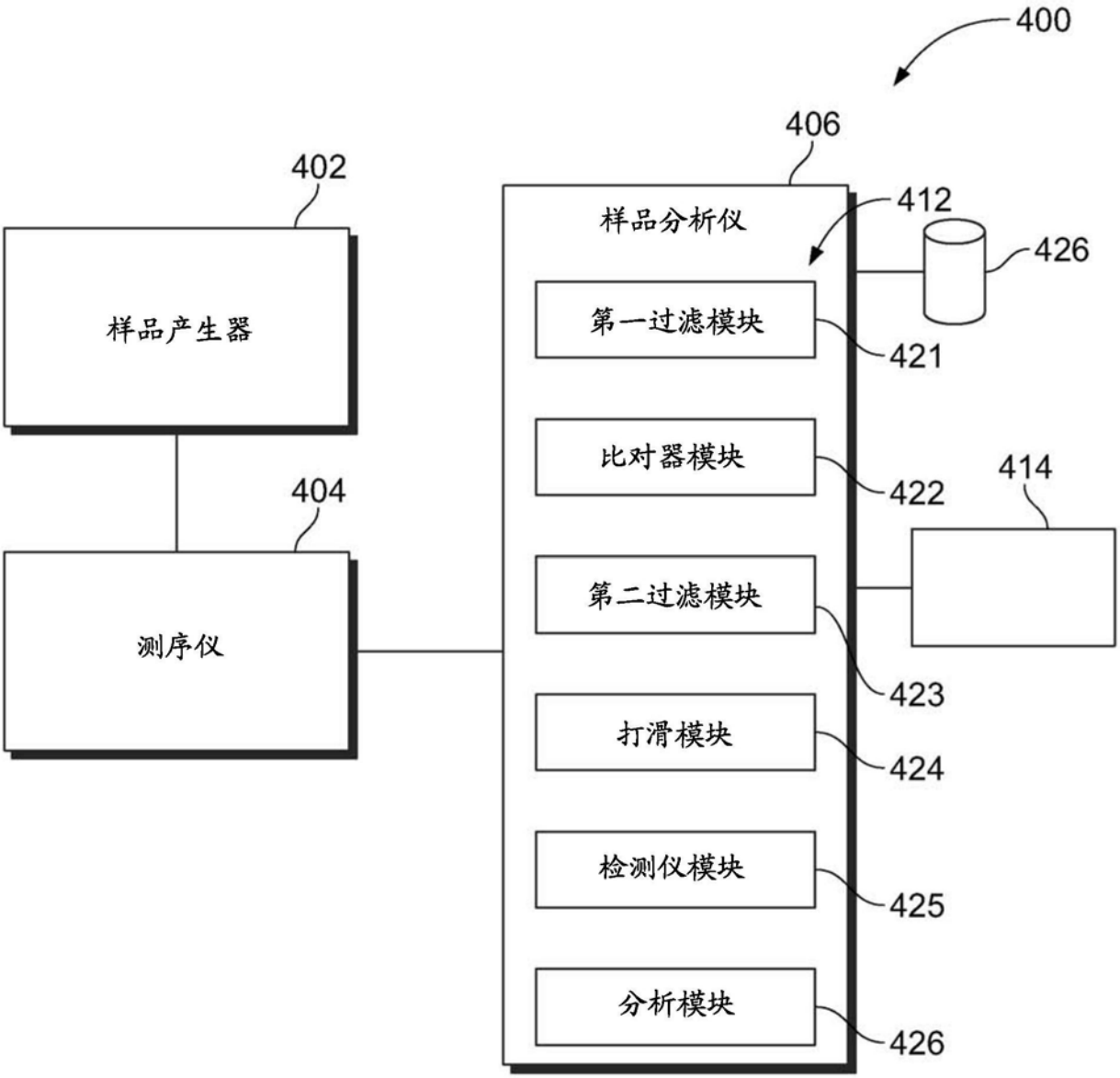

400
406
402
样品分析仪
412
426
样品产生器
第一过滤模块
421
比对器模块
414
422
404
第二过滤模块
测序仪
423
打滑模块
424
检测仪模块
425
分析模块
426

图15

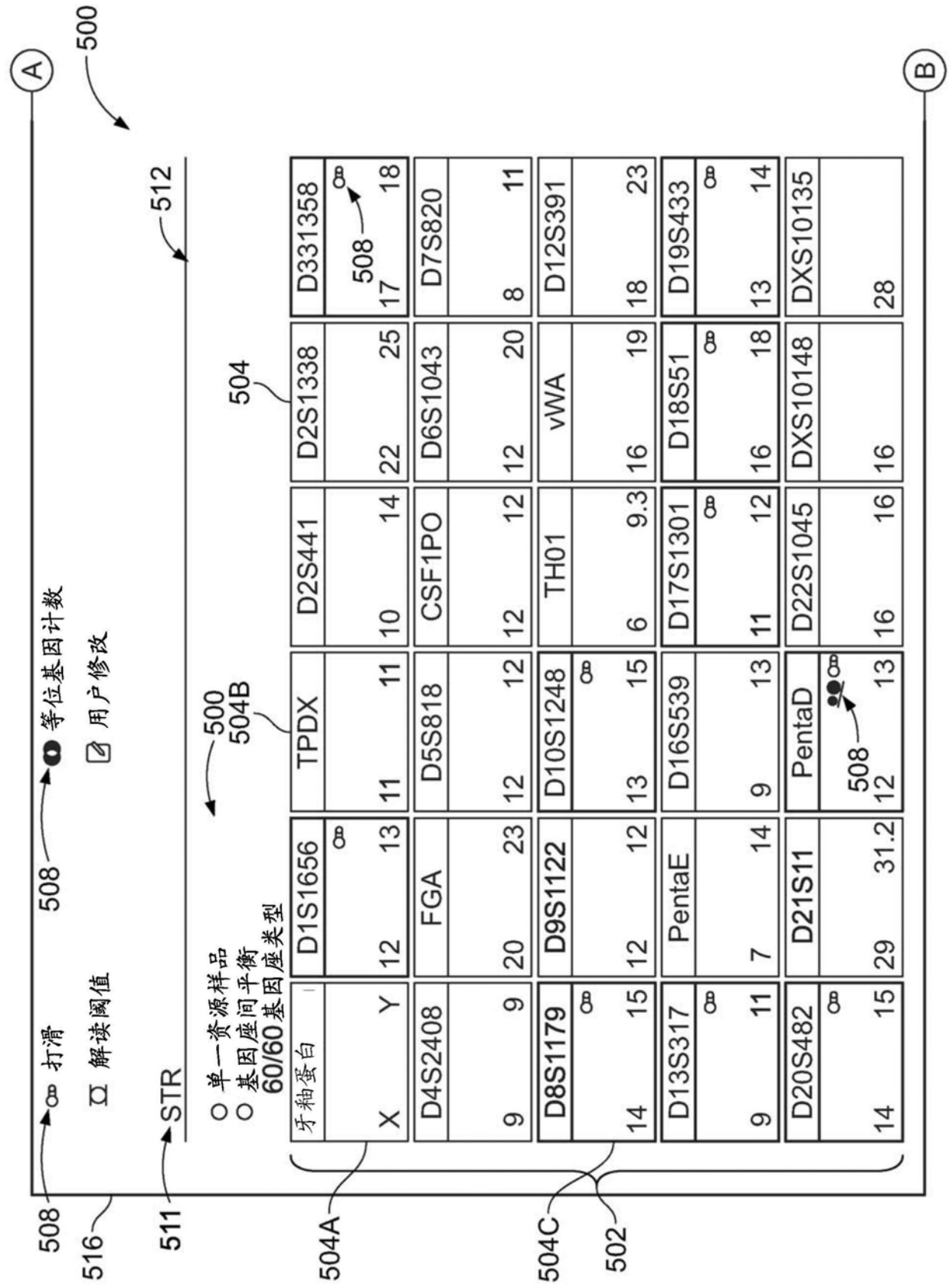

A
B
500
512
504
504B
504A
504C
502
508
516
511
打滑
等位基因计数
解读阈值
用户修改
STR
单一资源样品
基因座间平衡
60/60基因座类型
牙釉蛋白
X Y
D1S1656
12 13
D4S2408
9 9
D8S1179
14 15
D13S317
9 11
D20S482
14 15
TPDX
11 11
D5S818
12 12
D10S1248
13 15
D16S539
9 13
D21S11
29 31.2
D2S441
10 14
CSF1PO
12 12
TH01
6 9.3
D17S1301
11 12
D22S1045
16 16
D2S1338
22 25
D6S1043
12 20
vWA
16 19
D18S51
16 18
DXS10148
16 16
D331358
17 18
D7S820
8 11
D12S391
18 23
D19S433
13 14
DXS10135
28
PentaD
12 13

图16A

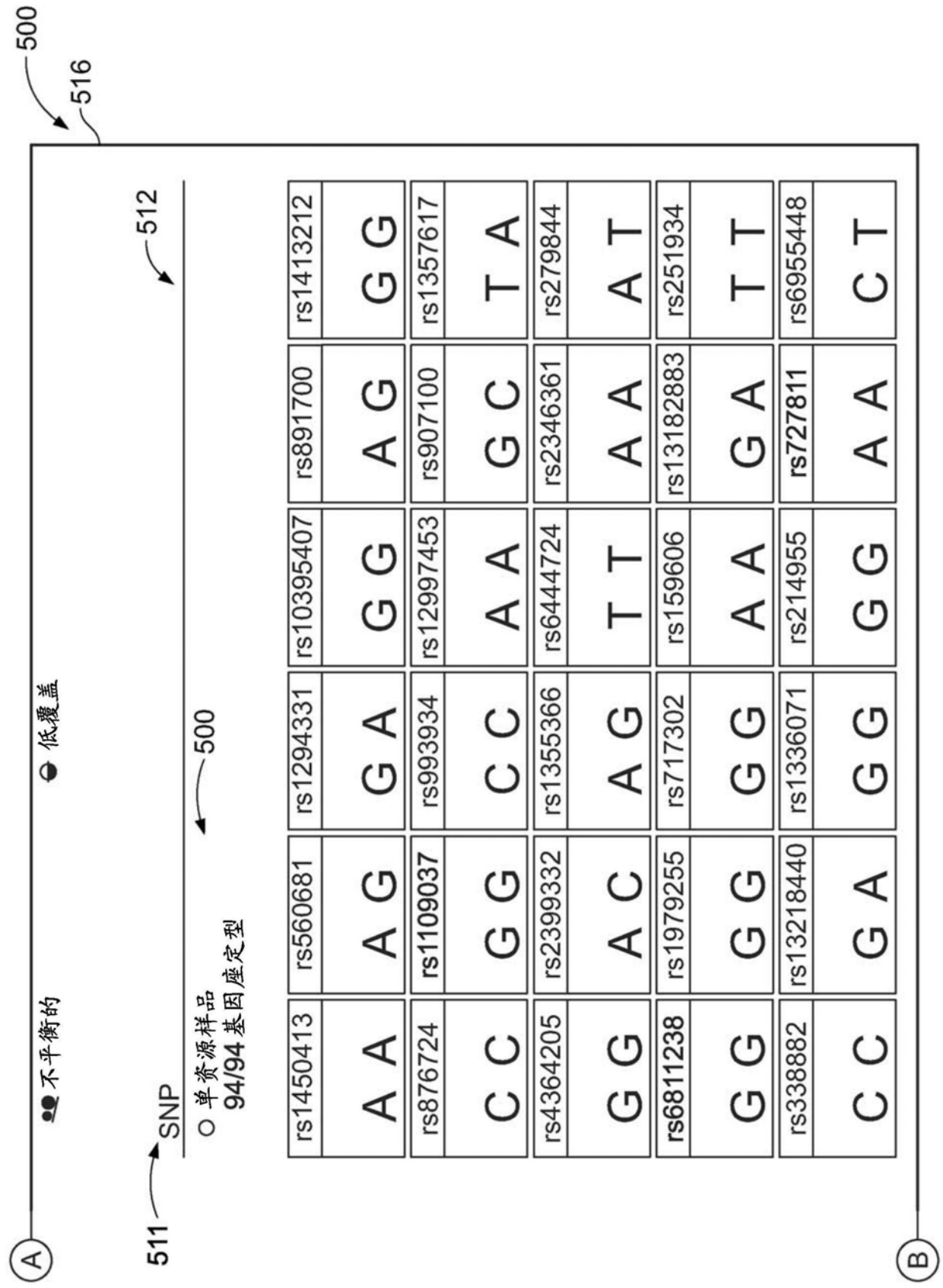

500
516
512
500
511
SNP
单资源样品
94/94 基因座定型
不平衡的
低覆盖
A
B
rs1450413 rs560681 rs1294331 rs10395407 rs891700 rs1413212
A A A G G A G G A G G G
rs876724 rs1109037 rs993934 rs12997453 rs907100 rs1357617
C C G G C C A A G C T A
rs4364205 rs2399332 rs1355366 rs6444724 rs2346361 rs279844
G G A C A G T T A A A T
rs6811238 rs1979255 rs717302 rs159606 rs13182883 rs251934
G G G G G G A A G A T T
rs338882 rs13218440 rs1336071 rs214955 rs727811 rs6955448
C C G A G G G G A A C T

图16B

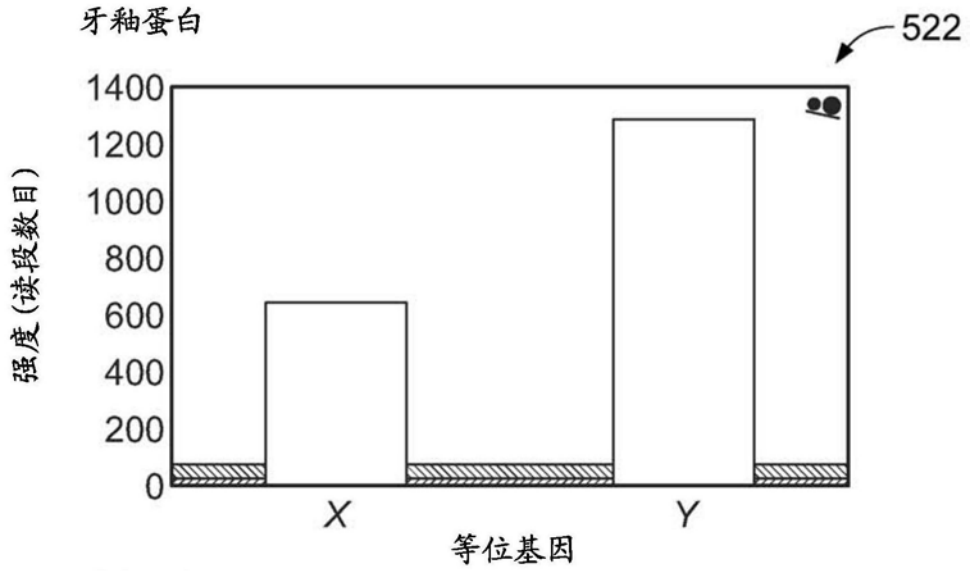

牙釉蛋白
522
1400
1200
1000
800
600
400
200
0
强度(读段数目)
X
Y
等位基因

图17A

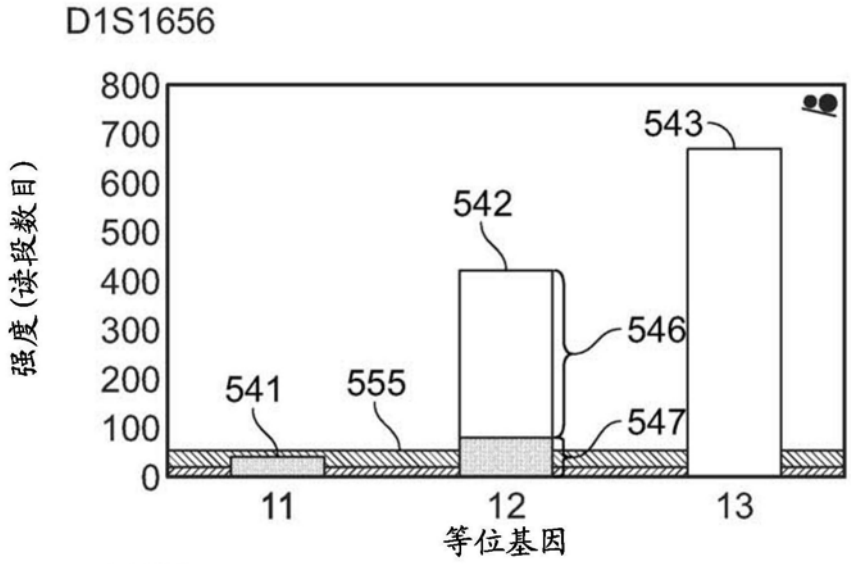

D1S1656
800
700
600
500
400
300
200
100
0
强度(读段数目)
543
542
546
541
555
547
11
12
13
等位基因

图17B

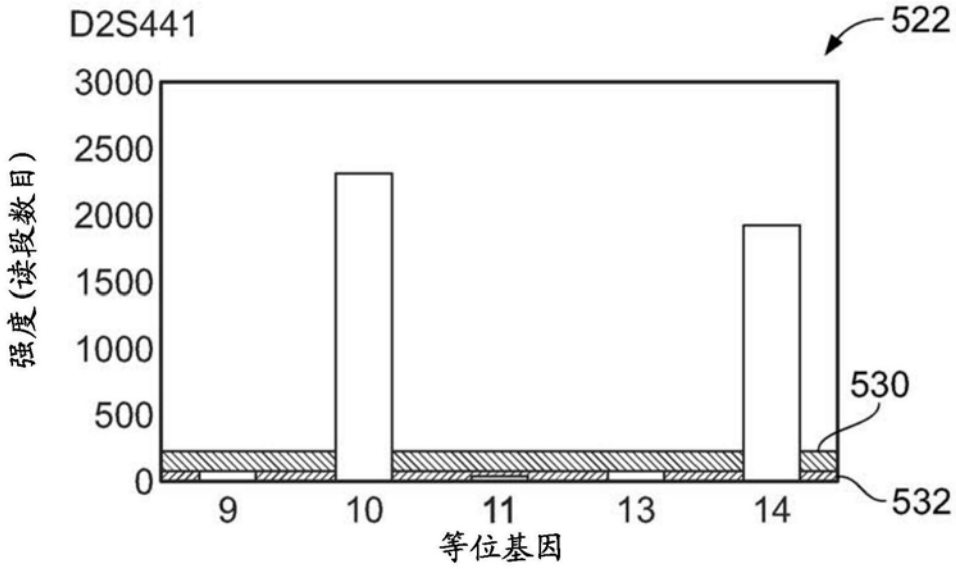

D2S441
522
3000
2500
2000
1500
1000
500
0
强度(读段数目)
530
532
9
10
11
13
14
等位基因

图17C

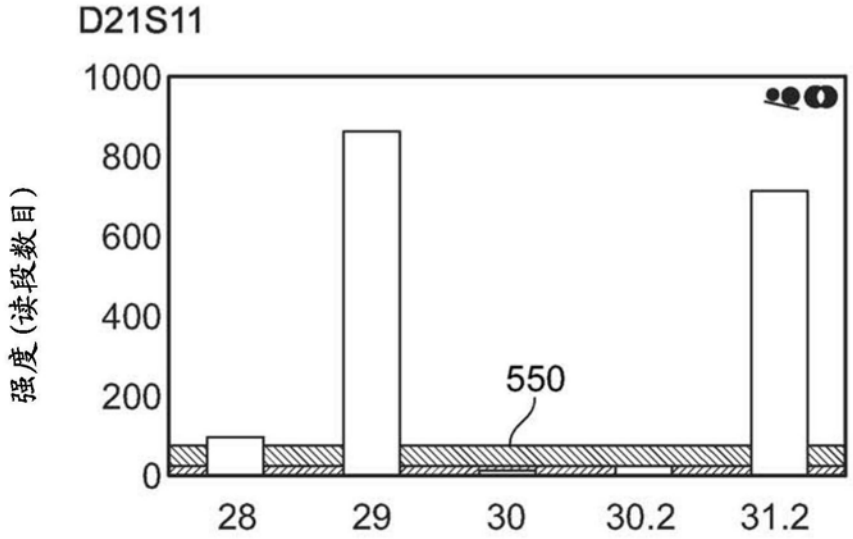

D21S11
1000
800
600
400
200
0
强度(读段数目)
550
28
29
30
30.2
31.2

图17D

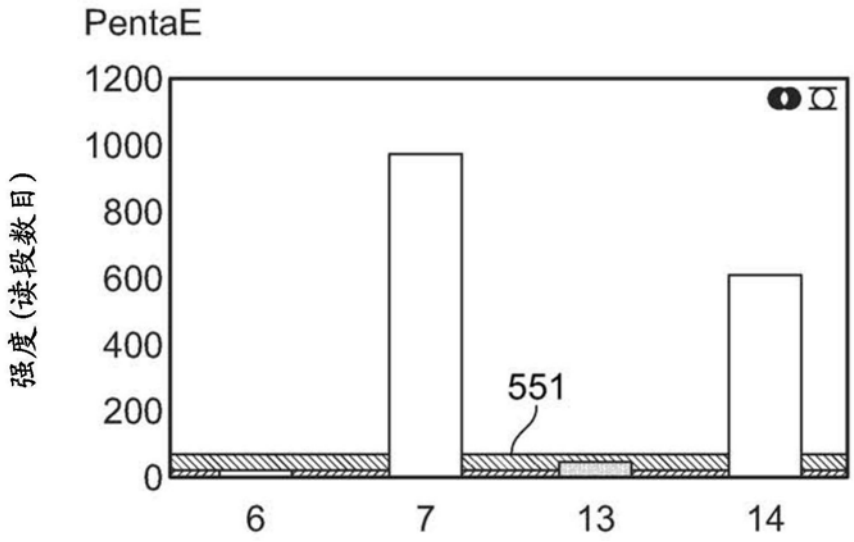

PentaE
强度(读段数目)
1200
1000
800
600
400
200
0
551
6
7
13
14

图17E

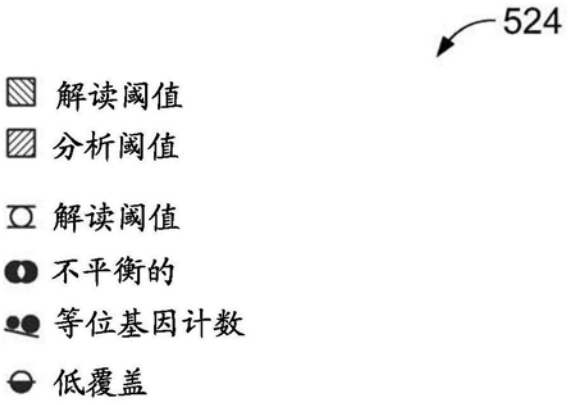

524
解读阈值
分析阈值
解读阈值
不平衡的
等位基因计数
低覆盖

图17F