(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 119731731 A
(43)申请公布日 2025.03.28

(21)申请号 202380059358.9

(22)申请日 2023.08.15

(30)优先权数据
63/398,512 2022.08.16 US
63/445,541 2023.02.14 US

(85)PCT国际申请进入国家阶段日
2025.02.12

(86)PCT国际申请的申请数据
PCT/US2023/072246 2023.08.15

(87)PCT国际申请的公布数据
WO2024/040078 EN 2024.02.22

(71)申请人 诺杰有限公司
地址 美国

(72)发明人 琼·斯内德克尔
凯瑟琳·M·斯蒂芬斯
萨拉·M·拉德克
戈塔米·帕德马班杜
约安娜·亚历山大·佩雷拉·安特内斯

(74)专利代理机构 中原信达知识产权代理有限责任公司 11219
专利代理师 韩峰 孙志湧

(51)Int.Cl.
G16B 10/00(2006.01)
C12Q 1/6876(2006.01)
G16B 20/20(2006.01)
G16B 20/40(2006.01)
G16B 40/00(2006.01)

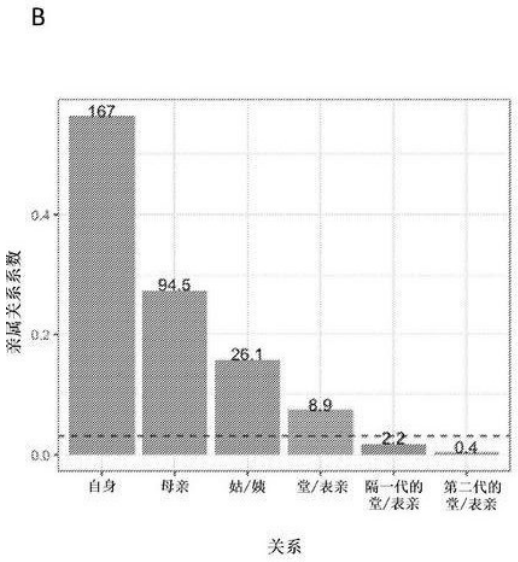
权利要求书17页 说明书76页 附图34页

(54)发明名称

用于失踪人员和灾难/冲突受害者的亲属关系评估的方法和系统

(57)摘要

在某些方面,本公开涉及进行基于DNA的亲属关系分析,其涉及分析2,000至50,000个之间的SNP,包括样本制备和测序技术和方法,能够用于计算来自感兴趣人员(诸如失踪人员或冲突或灾难的受害者)的DNA谱与参考DNA谱集中的一个或多个参考DNA谱的关系程度,该参考DNA谱集包括来自感兴趣人员的亲属的至少一个DNA谱。



CN 119731731 A

1.一种执行基于DNA的亲属关系分析的方法,包括:

提供来自感兴趣人员的核酸样本,

使用与多个靶序列特异性杂交的多个引物来扩增所述核酸样本,所述多个靶序列共同包括多个至少在2,000至50,000个之间或约2,000至50,000个之间的单核苷酸多态性(SNP),从而生成扩增产物,其中,所述扩增在一个或多个多重PCR反应中进行,

从所述扩增产物生成核酸库,

对从所述扩增产物生成的所述核酸库进行测序,

分析所述扩增产物的序列,

确定所述多个SNP的基因型,从而生成DNA谱,以及

计算所述DNA谱与一个或多个参考DNA谱的关系程度,其中,所述一个或多个参考DNA谱被包含在参考DNA谱集内,所述参考DNA谱集包括来自所述感兴趣人员的亲属的一个或多个参考DNA谱。

2.一种执行基于DNA的亲属关系分析的方法,包括:

提供来自感兴趣人员的核酸样本,

使用与多个靶序列特异性杂交的多个引物来扩增所述核酸样本,所述多个靶序列共同包括多个至少在2,000至50,000个之间或约2,000至50,000个之间的单核苷酸多态性(SNP),从而生成扩增产物,其中,所述扩增在一个或多个多重PCR反应中进行,

从所述扩增产物生成核酸库,

对从所述扩增产物生成的所述核酸库进行测序,

确定所述多个SNP的基因型,从而生成DNA谱,以及

计算所述DNA谱与一个或多个参考DNA谱的关系程度,其中,所述一个或多个参考DNA谱被包含在参考DNA谱集内,所述参考DNA谱集包括来自所述感兴趣人员的亲属的一个或多个参考DNA谱。

3.根据权利要求1或2所述的方法,其中,使用大规模并行测序(MPS)来进行所述测序。

4.根据权利要求1至3中任一项所述的方法,其中,所述测序不包括全基因组测序(WGS)。

5.根据权利要求1至4中任一项所述的方法,还包括生成包含与一个或多个DNA谱相关的所述DNA谱的家谱。

6.一种构建感兴趣人员的核酸库的方法,包括:

提供来自感兴趣人员的核酸样本,

使用与多个靶序列特异性杂交的多个引物来扩增所述核酸样本,所述多个靶序列共同包括多个至少在2,000至50,000个之间或约2,000至50,000个之间的单核苷酸多态性(SNP),从而生成包括扩增产物的核酸库,其中,所述扩增在一个或多个多重PCR反应中进行。

7.根据权利要求6所述的方法,还包括对所述扩增产物进行测序以产生所述感兴趣人员的DNA谱的步骤。

8.一种构建参考DNA样本的核酸库的方法,包括:

提供来自感兴趣人员的亲属的核酸样本,

使用与多个靶序列特异性杂交的多个引物来扩增所述核酸样本,所述多个靶序列共同

包括多个至少在2,000至50,000个之间或约2,000至50,000个之间的单核苷酸多态性(SNP),从而生成包括扩增产物的核酸库,其中,所述扩增在一个或多个多重PCR反应中进行。

9.根据权利要求8所述的方法,其中,所述亲属是所述感兴趣人员的一级、二级、三级、四级或五级亲属。

10.根据权利要求8或9所述的方法,其中,所述亲属是所述感兴趣人员的一级、二级或三级亲属。

11.根据权利要求1至10中任一项所述的方法,其中,所述核酸样本包括基因组DNA。

12.根据权利要求1至11中任一项所述的方法,其中,所述核酸样本包括一种或多种酶抑制剂。

13.根据权利要求12所述的方法,其中,所述一种或多种酶抑制剂包括选自由血红素、亚铁血红素、腐殖酸、靛蓝、单宁酸、胶原蛋白、钙和羟基磷灰石组成的组中的一种或多种抑制剂。

14.根据权利要求1至13中任一项所述的方法,其中,所述核酸样本包括低质量核酸分子和/或低数量核酸分子。

15.根据权利要求14所述的方法,其中,所述低质量核酸分子是退化的基因组DNA和/或片段化的基因组DNA。

16.根据权利要求14或权利要求15所述的方法,其中,所述低质量核酸分子具有为1、2、3、4、5、6、7、8、9、10、15、20、25、30、35、40、45、50、55、60、65、70、75、80、85、90、95、100、105、110、115、120、125、130、135、140、145、150、155、160、165、170、175、180、185、190、195或200的退化指数(DI)或至少为1、2、3、4、5、6、7、8、9、10、15、20、25、30、35、40、45、50、55、60、65、70、75、80、85、90、95、100、105、110、115、120、125、130、135、140、145、150、155、160、165、170、175、180、185、190、195或200的退化指数(DI)。

17.根据权利要求14或权利要求15所述的方法,其中,所述低质量核酸分子具有至少1且多达或小于158.3的DI。

18.根据权利要求1至13中任一项所述的方法,其中,所述核酸样本包括高质量核酸分子。

19.根据权利要求18所述的方法,其中,所述高质量核酸分子具有小于1的DI。

20.根据权利要求1至19中任一项所述的方法,其中,所述感兴趣人员是失踪人员。

21.根据权利要求1至19中任一项所述的方法,其中,所述感兴趣人员是灾难或冲突的受害者。

22.根据权利要求1至21中任一项所述的方法,其中,所述核酸样本源自唾液、血液、精液、毛发、牙齿、骨骼或皮肤。

23.根据权利要求22所述的方法,其中,所述核酸样本源自唾液、血液或精液。

24.根据权利要求22所述的方法,其中,所述核酸样本源自骨骼或毛发。

25.根据权利要求1至21中任一项所述的方法,其中,所述核酸样本源自口腔拭子、纸、织物,或浸渍有唾液、血液、精液或其他体液、或者包含毛发或皮肤细胞的其他基质或物体。

26.根据权利要求1至25中任一项所述的方法,其中,所述核酸样本包括在3pg至100ng之间或在约3pg至100ng之间的基因组DNA。

27.根据权利要求1至26中任一项所述的方法,其中,所述核酸样本包括在100pg至5ng之间或在约100pg至5ng之间的基因组DNA,在50pg至5ng之间或在约50pg至5ng之间的基因组DNA,或者在3pg至5ng之间或在约3pg至5ng之间的基因组DNA。

28.根据权利要求26或27所述的方法,其中,所述核酸样本包括1ng或约1ng的基因组DNA。

29.根据权利要求1至28中任一项所述的方法,其中,所述多个SNP包括亲属关系SNP(kiSNP)。

30.根据权利要求1至29中任一项所述的方法,其中,所述多个SNP包括Y染色体SNP(Y-SNP)。

31.根据权利要求1至30中任一项所述的方法,其中,所述多个SNP包括kiSNP和Y-SNP。

32.根据权利要求1至31中任一项所述的方法,其中,所述多个SNP包括kiSNP、生物地理祖先SNP(aiSNP)、身份SNP(iiSNP)、表型SNP(piSNP)、X-染色体SNP(X-SNP)和Y-染色体SNP(Y-SNP)。

33.根据权利要求1至28中任一项所述的方法,其中,所述多个SNP包括选自由kiSNP、aiSNP、iiSNP、piSNP、X-SNP和Y-SNP组成的组中的一个或多个的SNP。

34.根据权利要求1至33中任一项所述的方法,其中,所述多个SNP中的至少80%、81%、82%、83%、84%、85%、86%、87%、88%、89%、90%、91%、92%、93%、94%、95%、96%、97%、98%或99%是亲属关系SNP,或至少约80%、81%、82%、83%、84%、85%、86%、87%、88%、89%、90%、91%、92%、93%、94%、95%、96%、97%、98%或99%是亲属关系SNP。

35.根据权利要求1至34中任一项所述的方法,其中,所述参考DNA谱集包括多达5、10、15、20、25、30、35、40、45、50、55、60、65、70、75、80、85、90、95、100、125、150、175、200、300、400、500、600、700、800、900或1000个参考DNA谱。

36.根据权利要求1至35中任一项所述的方法,其中,所述参考DNA谱集中至少5%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%、65%、70%、75%、80%、85%、90%或95%的所述参考DNA谱来自所述感兴趣人员的亲属。

37.根据权利要求1至36中任一项所述的方法,其中,所述参考DNA谱集中至少5%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%、65%、70%、75%、80%、85%、90%或95%的所述参考DNA谱来自所述感兴趣人员的亲属,并且其中,所述参考DNA谱集中至少5%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%、65%、70%、75%、80%、85%、90%或95%的所述参考DNA谱中的每一个为一级亲属、二级亲属、三级亲属、四级亲属或五级亲属。

38.根据权利要求1至37中任一项所述的方法,其中,所述参考DNA谱集中至少50%的所述参考DNA谱来自所述感兴趣人员的亲属。

39.根据权利要求36至38中任一项所述的方法,其中,所述参考DNA谱集中的所述感兴趣人员的每个亲属分别是所述感兴趣人员的一级亲属、二级亲属、三级亲属、四级亲属或五级亲属。

40.根据权利要求39所述的方法,其中,所述参考DNA谱集中的所述感兴趣人员的每位亲属分别是所述感兴趣人员的一级亲属、二级亲属或三级亲属。

41.根据权利要求1至40中任一项所述的方法,其中,所述参考DNA谱集中的所述感兴趣

人员的每个亲属的身份都是已知的。

42.根据权利要求1至41中任一项所述的方法,其中,所述参考DNA谱集中的所述一个或多个参考DNA谱中的每一个的身份都是已知的。

43.根据权利要求1至42中任一项所述的方法,其中,所述参考DNA谱集在数据库中。

44.根据权利要求43所述的方法,其中,所述数据库不可公开访问。

45.根据权利要求1至44中任一权利要求所述的方法,其中,所述测序包括多达40重的测序复杂性。

46.根据权利要求1至44中任一项所述的方法,其中,所述测序包括多达32重的测序复杂性。

47.根据权利要求1至44中任一项所述的方法,其中,所述测序包括12重至32重的测序复杂性。

48.根据权利要求1至44中任一项所述的方法,其中,所述测序包括24重至32重的测序复杂性。

49.根据权利要求1至44中任一项所述的方法,其中,所述测序包括10重、11重、12重、13重、14重、15重、16重、17重、18重、19重、20重、21重、22重、23重、24重、25重、26重、27重、28重、29重、30重、31重、32重、33重、34重或35重的测序复杂性,或者约10重、11重、12重、13重、14重、15重、16重、17重、18重、19重、20重、21重、22重、23重、24重、25重、26重、27重、28重、29重、30重、31重、32重、33重、34重或35重的测序复杂性。

50.根据权利要求1至49中任一项所述的方法,其中,所述测序包括对于死后样本为8至16重或约8至16重的测序复杂性,和/或所述测序包括对于生前样本为24至40重或约24至40重的测序复杂性。

51.根据权利要求1至50中任一项所述的方法,其中,所述测序包括对于死后样本为12重或约12重的测序复杂性,和/或所述测序包括对于生前样本为32重或约32重的测序复杂性。

52.根据权利要求1至51中任一项所述的方法,其中,所述测序包括为30重、31重或32重的测序复杂性或者约为30重、31重或32重的测序复杂性。

53.根据权利要求1至52中任一项所述的方法,还包括识别所述感兴趣人员。

54.一种计算关联度的方法,包括:

获得包括至少在2,000至50,000个SNP之间或在约2,000至50,000个SNP之间的基因型的DNA谱,其中,所述DNA谱来自所述感兴趣人员;以及

计算所述DNA谱与一个或多个参考DNA谱的关系程度,其中,所述一个或多个参考DNA谱包括在参考DNA谱集内,所述参考DNA谱集包括来自所述感兴趣人员的亲属的一个或多个参考DNA谱。

55.一种计算关联度的方法,包括:

生成包括至少在2,000至50,000个SNP之间或在约2,000至50,000个SNP之间的基因型的DNA谱,其中,所述DNA谱来自所述感兴趣人员;以及

计算所述DNA谱与一个或多个参考DNA谱的关系程度,其中,所述一个或多个参考DNA谱包括在参考DNA谱集内,所述参考DNA谱集包括来自所述感兴趣人员的亲属的一个或多个参考DNA谱。

56.根据权利要求1至55中任一项所述的方法,其中,所述关系程度是使用亲属关系模型来计算的。

57.根据权利要求1至56中任一项所述的方法,其中,所述关系程度是使用利用PCA方法训练的亲属关系模型来计算的。

58.根据权利要求57所述的方法,其中,用于训练所述亲属关系模型的所述PCA方法是PCA或者涉及PCA。

59.根据权利要求57或权利要求58所述的方法,其中,所述PCA方法是PC-AiR。

60.根据权利要求59所述的方法,其中,所述PC-AiR包括以下步骤:

(1)估计训练数据库的所有样本对之间的亲属关系系数,所述样本对可选地训练DNA谱,其中,亲属关系系数>0.025的配对被识别为相关的,而亲属关系系数<-0.025的配对被识别为祖先分化的;

(2)初始化包括所有样本的无关样本集;以及

(3)迭代地:(i)识别所述无关样本集中具有所述无关样本集中最相关样本的集合,从而指定为X,(ii)识别X中与所述无关样本集中的样本相比具有最少祖先分化配对的样本集,从而指定为Y;以及(iii)如果Y具有零个样本,则终止过程,或者,如果Y具有至少一个样本,则从Y中随机选择一个样本以从U中移除,并且在步骤(3)(i)处开始重复。

61.根据权利要求57或权利要求58所述的方法,其中,所述PCA方法是改进的PC-Air。

62.根据权利要求61所述的方法,其中,所述改进的PC-AiR包括以下步骤:

(1)估计训练数据库的所有样本对之间的亲属关系系数,所述样本对可选地训练DNA谱,其中,亲属关系系数>0.01的配对被识别为相关的,而亲属关系系数<-0.025的配对被识别为祖先分化的;

(2)移除所有具有≥5%失踪数据的DNA谱;以及

(3)通过用排序值识别每个DNA谱来对所有DNA谱进行排序。在一些实施例中,所述排序值是基于完整数据库中从最少到最多排序的相关DNA谱的数量来确定的,并且通过所述完整数据库中从最多到最少排序的祖先分化DNA谱的数量来打破纽带。在一些实施例中,步骤(3)涉及迭代地遍历排序后的DNA谱,并且对于每个DNA谱:(i)如果所述DNA谱尚未出现在相关样本集中,则将其添加到所述无关样本集并且将所有相关DNA谱添加到所述相关样本集,以及(ii)如果所述DNA谱已经在所述相关样本集中,则跳至下一个DNA谱,并且在步骤(3(i)处开始重复。

63.根据权利要求1至62中任意一项所述的方法,其中,计算所述关系程度包括使用PC-Relate来计算亲属关系系数。

64.根据权利要求63所述的方法,其中,通过将所述感兴趣人员的所述DNA谱作为PC-Relate的输入来计算所述关系程度。

65.根据权利要求57至64中任一项所述的方法,其中,通过提供亲属关系模型和所述感兴趣人员的所述DNA谱作为PC-Relate的输入来计算所述关系程度。

66.根据权利要求63至65中任一项所述的方法,其中,所述一个或多个参考DNA谱还被提供作为PC-Relate的输入。

67.根据权利要求1至66中任一项的方法,其中,计算所述关系程度包括使用全基因组亲属关系算法来计算亲属关系系数,如下所示:

$$\widehat{\phi_{ij}} = \frac{\sum_{s\in S}(g_{is} - 2\hat{u}_{is})(g_{js} - 2\hat{u}_{js})}{4\sum_{s\in S}[\hat{u}_{is}(1 - \hat{u}_{is}) \cdot \hat{u}_{js}(1 - \hat{u}_{js})]^{1/2}}$$

其中,所述感兴趣人员和所述一个或多个参考DNA谱中的参考DNA谱是i和j,$\widehat{\phi_{ij}}$是所述亲属关系系数,$\hat{u}$是估计等位基因频率,s是在两个个体中都被分型的S个SNP中的SNP,g_{is}和g_{js}分别是在SNP s处的i和j中的参考等位基因的数量,并且$\hat{u}_{is}$和$\hat{u}_{js}$分别是由在SNP s处的i和j的PC-AiR计算出的预期等位基因频率。

68.根据权利要求1至67中任一项所述的方法,其中,计算所述关系程度包括计算似然比。

69.根据权利要求68所述的方法,其中,计算所述似然比包括比较所述DNA谱和所述一个或多个参考DNA谱之间的所述多个SNP。

70.根据权利要求68所述的方法,其中,计算所述似然比包括比较所述DNA谱和所述一个或多个参考DNA谱之间的SNP集合,所述SNP集合包括来自所述多个SNP内的亲属关系SNP。

71.根据权利要求68至70中任一项所述的方法,其中,计算所述似然比包括基于所述多个SNP的所述基因型,将所述DNA谱与所述一个或多个参考DNA谱中的参考DNA谱相关的概率除以所述DNA谱与所述参考DNA谱不相关的概率。

72.根据权利要求68至71中任一项所述的方法,其中,所述似然比(LR)计算如下:

$$LR = \frac{P(D \mid H_r)}{P(D \mid H_u)}$$

其中,D表示所述基因型,H_r表示个体是相关的假设,并且H_u表示个体是无关的假设。

73.根据权利要求68至71中任一项所述的方法,其中,所述LR计算如下:

$$LR = \frac{0.001}{p^2q^2}$$

其中,0.001表示基因分型错误率,p是等位基因1的等位基因频率,并且q是等位基因2的等位基因频率。

74.根据权利要求1至73中任一项所述的方法,其中,所述感兴趣人员在生物学上是男性,并且所述方法还包括计算所述DNA谱与所述一个或多个参考DNA谱之间共享Y染色体的似然比。

75.根据权利要求74所述的方法,其中,计算共享Y染色体的似然比包括比较所述DNA谱和所述一个或多个参考DNA谱之间的SNP集合,所述SNP集合包括一个或多个Y-SNP。

76.根据权利要求75所述的方法,其中,所述一个或多个Y-SNP包含在所述多个SNP内。

77.根据权利要求75所述的方法,其中,所述一个或多个Y-SNP包括至少20、25、30、35、40、45、50、55、60、65、70、75、80、81、82、83、84或85个Y-SNP。

78.根据权利要求75至77中任一项所述的方法,其中,所述一个或多个Y-SNP包括85个Y-SNP。

79.根据权利要求75至78中任一项所述的方法,其中,计算共享Y染色体的所述似然比包括基于所述一个或多个Y-SNP的所述基因型,将所述DNA谱与所述一个或多个参考DNA谱中的参考DNA谱共享Y染色体的概率除以所述DNA谱与所述参考DNA谱不共享Y染色体的概

率。

80.根据权利要求1至79中任一项所述的方法,其中,所述参考DNA谱集中至少90%、91%、92%、93%、94%、95%、96%、97%、98%或99%的所述DNA谱来自失踪人员的或灾难或冲突的受害者的亲属。

81.根据权利要求1至80中任一项所述的方法,其中,所述参考DNA谱集中的每个所述DNA谱来自失踪人员的或灾难或冲突的受害者的亲属。

82.根据权利要求1至81中任一项所述的方法,其中,所述参考DNA谱集包括多达5、10、15、20、25、30、35、40、45、50、55、60、65、70、75、80、85、90、95、100、125、150、175、200、300、400、500、600、700、800、900或1000个参考DNA谱。

83.根据权利要求1至82中任一项所述的方法,其中,所述参考DNA谱集包括多达100个参考DNA谱。

84.根据权利要求1至83中任一项所述的方法,其中,所述参考DNA谱集中至少5%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%、65%、70%、75%、80%、85%、90%或95%的所述参考DNA谱来自所述感兴趣人员的亲属。

85.根据权利要求1至84中任一项所述的方法,其中,所述参考DNA谱集中至少50%的所述参考DNA谱来自所述感兴趣人员的亲属。

86.根据权利要求84或权利要求85所述的方法,其中,所述参考DNA谱集中的所述感兴趣人员的每个亲属分别是所述感兴趣人员的一级亲属、二级亲属、三级亲属、四级亲属或五级亲属。

87.根据权利要求1至86中任一项所述的方法,其中,所述参考DNA谱集中的所述感兴趣人员的每个亲属的身份都是已知的。

88.根据权利要求1至87中任一项所述的方法,其中,所述DNA参考谱集中的所述感兴趣人员的每个亲属的所述身份都是已知的。

89.根据权利要求1至88中任一项所述的方法,其中,所述参考DNA谱集中的所述一个或多个参考DNA谱中的每一个的身份都是已知的。

90.根据权利要求1至89中任一项所述的方法,其中,所述参考DNA谱集在数据库中。

91.根据权利要求90所述的方法,其中,所述数据库不可公开访问。

92.根据权利要求90或权利要求91所述的方法,其中,所述数据库不能被第三方遗传学服务访问。

93.一种核酸库,所述核酸库是使用根据权利要求6至92中任一项所述的方法构建的。

94.多个引物,所述多个引物与多个靶序列特异性杂交,所述多个靶序列包括来自感兴趣人员的核酸样本中的至少在2,000至50,000个之间的单核苷酸多态性(SNP)或约2,000至50,000个之间的所述SNP,其中,在一个或多个多重PCR反应中使用所述多个引物来扩增所述核酸样本以产生扩增产物。

95.多个引物,所述多个引物与多个靶序列特异性杂交,所述多个靶序列包括来自感兴趣人员的核酸样本和来自一个或多个参考样本的核酸样本中的至少在2,000至50,000个之间的单核苷酸多态性(SNP)或约2,000至50,000个之间的所述SNP,

其中,所述一个或多个参考样本包括来自所述感兴趣人员的亲属的样本,以及

其中,在一个或多个多重PCR反应中使用所述多个引物来扩增来自所述感兴趣人员的

所述核酸样本和来自一个或多个参考样本的所述核酸样本以产生扩增产物。

96.根据权利要求94或权利要求95所述的多个引物,其中,来自所述感兴趣人员的所述核酸样本包括基因组DNA。

97.根据权利要求94至96中任一项所述的多个引物,其中,来自所述感兴趣人员的所述核酸样本包括一种或多种酶抑制剂。

98.根据权利要求97所述的多种引物,其中,所述一种或多种酶抑制剂包括选自由血红素、亚铁血红素、腐殖酸、靛蓝、单宁酸、胶原蛋白、钙和羟基磷灰石组成的组中的一种或多种抑制剂。

99.根据权利要求94至98中任一项所述的多个引物,其中,来自所述感兴趣人员的所述核酸样本包括低质量核酸分子和/或低数量核酸分子。

100.根据权利要求99所述的多种引物,其中,所述低质量核酸分子是退化的基因组DNA和/或片段化的基因组DNA。

101.根据权利要求99或权利要求100所述的多个引物,其中,所述低质量核酸分子具有为1、2、3、4、5、6、7、8、9、10、15、20、25、30、35、40、45、50、55、60、65、70、75、80、85、90、95、100、105、110、115、120、125、130、135、140、145、150、155、160、165、170、175、180、185、190、195或200的退化指数(DI),或至少为1、2、3、4、5、6、7、8、9、10、15、20、25、30、35、40、45、50、55、60、65、70、75、80、85、90、95、100、105、110、115、120、125、130、135、140、145、150、155、160、165、170、175、180、185、190、195或200的退化指数(DI)。

102.根据权利要求99至101中任一项所述的多个引物,其中,所述低质量核酸分子具有至少1且多达或小于158.3的DI。

103.根据权利要求94至102中任一项的所述的多个引物,其中,来自所述感兴趣人员的所述核酸样本和/或来自一个或多个参考样本的所述核酸样本包括高质量核酸分子。

104.根据权利要求103所述的多个引物,其中,所述高质量核酸分子具有小于1的DI。

105.根据权利要求94至104中任一项所述的多个引物,其中,所述感兴趣人员是失踪人员。

106.根据权利要求94至104中任一项所述的多个引物,其中,所述感兴趣人员是灾难或冲突的受害者。

107.根据权利要求94至106中任一项所述的多个引物,其中,来自所述感兴趣人员的所述核酸样本源自口腔拭子、纸、织物,或浸渍有唾液、血液或其他体液、或者包含毛发或皮肤细胞的其他基质或物体。

108.根据权利要求94至107中任一项所述的多个引物,其中,来自所述感兴趣人员的所述核酸样本包括在3pg至100ng之间或在约3pg至100ng之间的基因组DNA。

109.根据权利要求94至108中任一项所述的多个引物,其中,来自所述感兴趣人员的所述核酸样本包括在100pg至5ng之间或在约100pg至5ng之间的基因组DNA,在50pg至5ng之间或在约50pg至5ng之间的基因组DNA,或者在3pg至5ng之间或在约3pg至5ng之间的基因组DNA。

110.根据权利要求108或权利要求109所述的多个引物,其中,来自所述感兴趣人员的所述核酸样本包括1ng或约1ng的基因组DNA。

111.根据权利要求94至110中任一项所述的多个引物,其中,所述多个SNP包括亲属关

系SNP(kiSNP)。

112.根据权利要求94至111中任一项所述的方法,其中,所述多个SNP包括Y染色体SNP(Y-SNP)。

113.根据权利要求94至112中任一项所述的方法,其中,所述多个SNP包括kiSNP和Y-SNP。

114.根据权利要求94至113中任一项所述的多个引物,其中,所述多个SNP包括kiSNP、生物地理祖先SNP(aiSNP)、身份SNP(iiSNP)、表型SNP(piSNP)、X-染色体SNP(X-SNP)和Y-染色体SNP(Y-SNP)。

115.根据权利要求94至111中任一项所述的多个引物,其中,所述多个SNP包括选自由kiSNP、aiSNP、iiSNP、piSNP、X-SNP和Y-SNP组成的组中的一个或多个的SNP。

116.根据权利要求94至115中任一项所述的多个引物,其中,所述多个SNP中的至少80%、81%、82%、83%、84%、85%、86%、87%、88%、89%、90%、91%、92%、93%、94%、95%、96%、97%、98%或99%是亲属关系SNP,或至少约80%、81%、82%、83%、84%、85%、86%、87%、88%、89%、90%、91%、92%、93%、94%、95%、96%、97%、98%或99%是亲属关系SNP。

117.根据权利要求94至116中任一项所述的多个引物,其中,所述参考DNA谱集中至少90%、91%、92%、93%、94%、95%、96%、97%、98%或99%的所述DNA谱来自失踪人员或灾难或冲突的受害者的亲属。

118.根据权利要求94至116中任一项所述的多个引物,其中,所述一个或多个参考样本中的每一个来自失踪人员或灾难或冲突的受害者的亲属。

119.根据权利要求94至118中任一项所述的多种引物,其中,所述一个或多个参考样本包括多达5、10、15、20、25、30、35、40、45、50、55、60、65、70、75、80、85、90、95、100、125、150、175、200、300、400、500、600、700、800、900或1000个参考样本。

120.根据权利要求94至119中任一项所述的多个引物,其中,所述一个或多个参考样本中的至少5%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%、65%、70%、75%、80%、85%、90%或95%来自所述感兴趣人员的亲属。

121.根据权利要求94至120中任一项所述的多个引物,其中,所述参考DNA谱集中至少5%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%、65%、70%、75%、80%、85%、90%或95%的所述参考DNA谱来自所述感兴趣人员的亲属,并且其中,所述参考DNA谱集中至少5%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%、65%、70%、75%、80%、85%、90%或95%的所述参考DNA谱中的每一个为一级亲属、二级亲属、三级亲属、四级亲属或五级亲属。

122.根据权利要求94至121中任一项所述的多个引物,其中,至少50%的所述一个或多个参考样本来自所述感兴趣人员的亲属。

123.根据权利要求120至122中任一项所述的多个引物,其中,所述一个或多个参考样本中的所述感兴趣人员的每个亲属分别是所述感兴趣人员的一级亲属、二级亲属、三级亲属、四级亲属或五级亲属。

124.根据权利要求123所述的多个引物,其中,所述一个或多个参考样本中的所述感兴趣人员的每个亲属分别是所述感兴趣人员的一级亲属、二级亲属或三级亲属。

125.根据权利要求94至124中任一项所述的多个引物,其中,所述一个或多个参考样本中的所述感兴趣人员的每个亲属的身份是已知的。

126.根据权利要求94至125中任一项所述的多个引物,其中,所述一个或多个参考样本中的每一个的身份是已知的。

127.一种构建DNA谱的方法,包括:

提供来自感兴趣人员的核酸样本,

使用与多个靶序列特异性杂交的多个引物来扩增所述核酸样本,所述多个靶序列共同包括多个至少在2,000至50,000个之间或约2,000至50,000个之间的单核苷酸多态性(SNP),从而生成扩增产物,其中,所述扩增在一个或多个多重PCR反应中进行,

对所述扩增产物进行测序,

确定所述多个SNP的基因型,从而生成DNA谱。

128.一种构建DNA谱的方法,包括:

提供来自感兴趣人员的核酸样本,

提供来自所述感兴趣人员的亲属的核酸样本,

使用与多个靶序列特异性杂交的多个引物来扩增来自所述感兴趣人员的所述核酸样本和来自所述亲属的所述核酸样本,所述多个靶序列共同包括多个至少在2,000至50,000个之间或约2,000至50,000个之间的单核苷酸多态性(SNP),从而生成扩增产物,其中,所述扩增在一个或多个多重PCR反应中进行,

对所述扩增产物进行测序,

确定所述多个SNP的基因型,从而生成所述感兴趣人员和所述感兴趣人员的所述亲属的DNA谱。

129.根据权利要求127或权利要求128所述的方法,其中,所述测序不包括全基因组测序(WGS)。

130.根据权利要求127或权利要求129所述的方法,其中,所述核酸样本包括基因组DNA。

131.根据权利要求128或权利要求129所述的方法,其中,所述感兴趣人员的所述核酸样本和/或所述感兴趣人员的所述亲属的所述核酸样本包括基因组DNA。

132.根据权利要求127至131中任一项所述的方法,其中,所述核酸样本、所述感兴趣人员的所述核酸样本和/或所述亲属的所述核酸样本包括一种或多种酶抑制剂。

133.根据权利要求132所述的方法,其中,所述一种或多种酶抑制剂包括选自由血红素、亚铁血红素、腐殖酸、靛蓝、单宁酸、胶原蛋白、钙和羟基磷灰石组成的组中的一种或多种抑制剂。

134.根据权利要求127至133中任一项所述的方法,其中,所述核酸样本、所述感兴趣人员的所述核酸样本和/或所述亲属的所述核酸样本包括低质量核酸分子和/或低数量核酸分子。

135.根据权利要求134所述的方法,其中,所述低质量核酸分子是退化的基因组DNA和/或片段化的基因组DNA。

136.根据权利要求134或权利要求135所述的方法,其中,所述低质量核酸分子具有为1、2、3、4、5、6、7、8、9、10、15、20、25、30、35、40、45、50、55、60、65、70、75、80、85、90、95、100、

105、110、115、120、125、130、135、140、145、150、155、160、165、170、175、180、185、190、195或200的退化指数(DI),或至少为1、2、3、4、5、6、7、8、9、10、15、20、25、30、35、40、45、50、55、60、65、70、75、80、85、90、95、100、105、110、115、120、125、130、135、140、145、150、155、160、165、170、175、180、185、190、195或200的退化指数(DI)。

137.根据权利要求134至136中任一项所述的方法,其中,所述低质量核酸分子具有至少1且多达或小于158.3的DI。

138.根据权利要求127至133中任一项所述的方法,其中,所述核酸样本、所述感兴趣人员的所述核酸样本和/或所述亲属的所述核酸样本包括高质量核酸分子。

139.根据权利要求138所述的方法,其中,所述高质量核酸分子具有小于1的DI。

140.根据权利要求127至139中任一项所述的方法,其中,所述感兴趣人员是失踪人员。

141.根据权利要求127至139中任一项所述的方法,其中,所述感兴趣人员是灾难或冲突的受害者。

142.根据权利要求128、129和131至141中任一项所述的方法,其中,所述感兴趣人员的所述亲属是一级、二级、三级、四级或五级亲属。

143.根据权利要求128、129和131-141中任一项所述的方法,其中,所述感兴趣人员的所述亲属是一级、二级或三级亲属。

144.根据权利要求127至143中任一项所述的方法,其中,所述核酸样本、所述感兴趣人员的所述核酸样本和/或所述亲属的所述核酸样本源自口腔拭子、纸、织物,或浸渍有唾液、血液或其他体液、或者包含毛发或皮肤细胞的其他基质或物体。

145.根据权利要求127至144中任一项所述的方法,其中,所述核酸样本、所述感兴趣人员的所述核酸样本和/或所述亲属的所述核酸样本包括在3pg至100ng之间或在约3pg至100ng之间的基因组DNA。

146.根据权利要求127至145中任一项所述的方法,其中,所述核酸样本、所述感兴趣人员的所述核酸样本和/或所述亲属的所述核酸样本包括在100pg至5ng之间或在约100pg至5ng之间的基因组DNA,在50pg至5ng之间或在约50pg至5ng之间的基因组DNA,或者在3pg至5ng之间或在约3pg至5ng之间的基因组DNA。

147.根据权利要求145或权利要求146所述的方法,其中,所述核酸样本、所述感兴趣人员的核酸样本和/或所述亲属的所述核酸样本包括1ng或约1ng的基因组DNA。

148.根据权利要求127至147中任一项所述的方法,其中,所述多个SNP包括亲属关系SNP。

149.根据权利要求127至148中任一项所述的方法,其中,所述多个SNP包括Y染色体SNP(Y-SNP)。

150.根据权利要求127至149中任一项所述的方法,其中,所述多个SNP包括kiSNP和Y-SNP。

151.根据权利要求127至150中任一项所述的方法,其中,所述多个SNP包括kiSNP、生物地理祖先SNP(aiSNP)、身份SNP(iiSNP)、表型SNP(piSNP)、X-染色体SNP(X-SNP)和Y-染色体SNP(Y-SNP)。

152.根据权利要求127至151中任一项所述的方法,其中,所述多个SNP包括选自由kiSNP、aiSNP、iiSNP、piSNP、X-SNP和Y-SNP组成的组中的一个或多个的SNP。

153.根据权利要求127至152中任一项所述的方法,其中,所述多个SNP中的至少80%、81%、82%、83%、84%、85%、86%、87%、88%、89%、90%、91%、92%、93%、94%、95%、96%、97%、98%或99%是亲属关系SNP,或至少约80%、81%、82%、83%、84%、85%、86%、87%、88%、89%、90%、91%、92%、93%、94%、95%、96%、97%、98%或99%是亲属关系SNP。

154.根据权利要求1至92和127至153中任一项所述的方法,其中,所述测序包括多达40重的测序复杂性。

155.根据权利要求1至92和127至153中任一项所述的方法,其中,所述测序包括多达32重的测序复杂性。

156.根据权利要求1至92和127-至153中任一项所述的方法,其中,所述测序包括12重至32重的测序复杂性。

157.根据权利要求1至92和127至153中任一项所述的方法,其中,所述测序包括24重至32重的测序复杂性。

158.权利要求1-92和127-153中任一项的方法,其中:

所述测序包括4重、5重、6重、7重、8重、9重、10重、11重、12重、13重、14重、15重、16重、17重、18重、19重、20重、21重、22重、23重、24重、25重、26重、27重、28重、29重、30重、31重、32重、33重、34重、35重、36重、37重、38重、39重、40重、41重、42重、43重、44重、45重的测序复杂性,或者约4重、5重、6重、7重、8重、9重、10重、11重、12重、13重、14重、15重、16重、17重、18重、19重、20重、21重、22重、23重、24重、25重、26重、27重、28重、29重、30重、31重、32重、33重、34重、35重、36重、37重、38重、39重、40重、41重、42重、43重、44重、45重的测序复杂性;或者

所述测序包括10重、11重、12重、13重、14重、15重、16重、17重、18重、19重、20重、21重、22重、23重、24重、25重、26重、27重、28重、29重、30重、31重、32重、33重、34重或35重的测序复杂性,或者约10重、11重、12重、13重、14重、15重、16重、17重、18重、19重、20重、21重、22重、23重、24重、25重、26重、27重、28重、29重、30重、31重、32重、33重、34重或35重的测序复杂性。

159.根据权利要求1至92和127至158中任一项所述的方法,其中,所述测序包括对于死后样本为8至16重或约8至16重的测序复杂性,和/或所述测序包括对于生前样本为24至40重或约24至40重的测序复杂性。

160.根据权利要求1至92和127至158中任一项所述的方法,其中,所述测序包括对于死后样本为12重或约12重的测序复杂性,和/或所述测序包括对于生前样本为32重或约32重的测序复杂性。

161.根据权利要求1至92和127至153中任一项所述的方法,其中,所述测序包括为30重、31重或32重的测序复杂性,或者约为30重、31重或32重的测序复杂性。

162.一种识别DNA谱的遗传亲属的方法,包括:

计算根据权利要求127至161中任一项所述的DNA谱与一个或多个参考DNA谱的关系程度,其中,所述一个或多个参考DNA谱被包括在参考DNA谱集内,所述参考DNA谱集包括来自所述感兴趣人员的亲属的一个或多个参考DNA谱;以及

生成家谱,所述家谱包括与所述一个或多个参考DNA谱相关的所述DNA谱。

163.根据权利要求162所述的方法,其中,所述一个或多个参考DNA谱是数据库的一部分。

164.根据权利要求162或权利要求163所述的方法,其中,所述参考DNA谱集包括多达5、10、15、20、25、30、35、40、45、50、55、60、65、70、75、80、85、90、95、100、125、150、175、200、300、400、500、600、700、800、900或1000个参考DNA谱。

165.根据权利要求162至164中任一项所述的方法,其中,在所述参考DNA谱集中的至少5%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%、65%、70%、75%、80%、85%、90%或95%的所述参考DNA谱来自所述感兴趣人员的亲属。

166.根据权利要求162至165中任一项所述的方法,其中,在所述参考DNA谱集中至少50%的所述参考DNA谱来自所述感兴趣人员的亲属。

167.根据权利要求165或权利要求166所述的方法,其中,所述感兴趣人员的每个亲属分别是所述感兴趣人员的一级亲属、二级亲属、三级亲属、四级亲属或五级亲属。

168.根据权利要求162至167中任一项所述的方法,其中,在所述参考DNA谱集中的所述感兴趣人员的每个亲属的身份都是已知的。

169.根据权利要求162至168中任一项所述的方法,其中,在所述参考DNA谱集中的所述一个或多个参考DNA谱中的每一个的身份都是已知的。

170.根据权利要求162至169中任一项所述的方法,其中,所述参考DNA谱集在数据库中。

171.根据权利要求170所述的方法,其中,所述数据库不可公开访问。

172.根据权利要求170或权利要求171所述的方法,其中,所述数据库不能被第三方遗传学服务访问。

173.一种识别DNA谱身份的方法,包括:

计算包括至少在2,000至50,000个SNP之间或在约2,000至50,000个SNP之间的基因型的DNA谱与一个或多个参考DNA谱的关系程度,其中,所述DNA谱来自感兴趣人员,并且其中,所述一个或多个参考DNA谱被包括在参考DNA谱集内,所述参考DNA谱集包括来自所述感兴趣人员的亲属的一个或多个参考DNA谱;以及

生成家谱,所述家谱包括与所述一个或多个参考DNA谱相关的所述DNA谱。

174.根据权利要求173所述的方法,其中,所述DNA谱是通过根据权利要求127至161中任一项所述的方法生成的。

175.根据权利要求173或权利要求174所述的方法,其中,所述关系程度是使用亲属关系模型来计算的。

176.根据权利要求173至175中任一项所述的方法,其中,所述关系程度是使用利用PCA方法训练的亲属关系模型来计算的。

177.根据权利要求176所述的方法,其中,用于训练所述亲属关系模型的所述PCA方法是PCA或涉及PCA。

178.根据权利要求176或权利要求177所述的方法,其中,所述PCA方法是PC-AiR。

179.根据权利要求178所述的方法,其中,所述PC-AiR包括以下步骤:

(1)估计训练数据库的所有样本对之间的亲属关系系数,所述样本对可选地训练DNA谱,其中,亲属关系系数>0.025的配对被识别为相关的,而亲属关系系数<-0.025的配对被

识别为祖先分化的;

(2)初始化包括所有样本的无关样本集;以及

(3)迭代地:(i)识别所述无关样本集中具有所述无关样本集中最相关样本的集合,从而指定为X,(ii)识别X中与所述无关样本集中的样本相比具有最少祖先分化配对的样本集,从而指定为Y;以及(iii)如果Y具有零个样本,则终止过程,或者,如果Y具有至少一个样本,则从Y中随机选择一个样本以从U中移除,并且在步骤(3)(i)处开始重复。

180.根据权利要求176或权利要求177所述的方法,其中,所述PCA方法是改进的PC-Air。

181.根据权利要求180所述的方法,其中,所述改进的PC-AiR包括以下步骤:

(1)估计训练数据库的所有样本对之间的亲属关系系数,所述样本对可选地训练DNA谱,其中,亲属关系系数>0.01的配对被识别为相关的,而亲属关系系数<-0.025的配对被识别为祖先分化的;

(2)移除所有具有≥5%失踪数据的DNA谱;以及

(3)通过用排序值识别每个DNA谱来对所有DNA谱进行排序。在一些实施例中,所述排序值是基于完整数据库中从最少到最多排序的相关DNA谱的数量来确定的,并且通过所述完整数据库中从最多到最少排序的祖先分化DNA谱的数量来打破纽带。在一些实施例中,步骤(3)涉及迭代地遍历排序后的DNA谱,并且对于每个DNA谱:(i)如果所述DNA谱尚未出现在相关样本集中,则将其添加到所述无关样本集并且将所有相关DNA谱添加到所述相关样本集,以及(ii)如果所述DNA谱已经在所述相关样本集中,则跳至下一个DNA谱,并且在步骤(3)(i)处开始重复。

182.根据权利要求173至181中任一项所述的方法,其中,计算所述关系程度包括使用PC-Relate来计算亲属关系系数。

183.根据权利要求182所述的方法,其中,通过提供所述感兴趣人员的所述DNA谱作为PC-Relate的输入来计算所述关系程度。

184.根据权利要求182或权利要求183所述的方法,其中,通过提供所述亲属关系模型和所述感兴趣人员的所述DNA谱作为PC-Relate的输入来计算所述关系程度。

185.根据权利要求182至184中任一项所述的方法,其中,还提供所述一个或多个参考DNA谱作为PC-Relate的输入。

186.根据权利要求173至185中任一项所述的方法,其中,计算所述关系程度包括使用全基因组亲属关系算法来计算亲属关系系数,如下所示:

$$\widehat{\phi_{ij}} = \frac{\sum_{s\in S}(g_{is} - 2\hat{u}_{is})(g_{js} - 2\hat{u}_{js})}{4\sum_{s\in S}[\hat{u}_{is}(1 - \hat{u}_{is}) \cdot \hat{u}_{js}(1 - \hat{u}_{js})]^{1/2}}$$

其中,所述感兴趣人员和所述一个或多个参考DNA谱中的参考DNA谱是i和j,$\widehat{\phi_{ij}}$是所述亲属关系系数,$\hat{u}$是估计等位基因频率,s是在两个个体中都被分型的S个SNP中的SNP,g_{is}和g_{js}分别是在SNP s处的i和j中的参考等位基因的数量,并且$\hat{u}_{is}$和$\hat{u}_{js}$分别是由在SNP s处的i和j的PC-AiR计算出的预期等位基因频率。

187.根据权利要求173至186中任一项所述的方法,其中,计算所述关系程度包括计算似然比。

188.根据权利要求187所述的方法,其中,计算所述似然比包括比较所述DNA谱和所述一个或多个参考DNA谱之间的所述多个SNP。

189.根据权利要求187所述的方法,其中,计算所述似然比包括比较所述DNA谱和所述一个或多个参考DNA谱之间的SNP集合,所述SNP集合包括来自所述多个SNP内的亲属关系SNP。

190.根据权利要求187至189中任一项所述的方法,其中,计算所述似然比包括基于所述多个SNP的所述基因型,将所述DNA谱与所述一个或多个参考DNA谱中的参考DNA谱相关的概率除以所述DNA谱与所述参考DNA谱不相关的概率。

191.根据权利要求187至190中任一项所述的方法,其中,所述似然比(LR)计算如下:

$$LR = \frac{P(D \mid H_r)}{P(D \mid H_u)}$$

其中,D表示所述基因型,H_r表示个体是相关的假设,并且H_u表示个体是无关的假设。

192.根据权利要求187至190中任一项所述的方法,其中,所述LR计算如下:

$$LR = \frac{0.001}{p^2q^2}$$

其中,0.001表示基因分型错误率,p是等位基因1的等位基因频率,并且q是等位基因2的等位基因频率。

193.根据权利要求173至192中任一项所述的方法,其中,所述感兴趣人员在生物学上是男性,并且所述方法还包括计算所述DNA谱与所述一个或多个参考DNA谱之间共享Y染色体的似然比。

194.根据权利要求193所述的方法,其中,计算共享Y染色体的似然比包括比较所述DNA谱和所述一个或多个参考DNA谱之间的SNP集合,所述SNP集合包括一个或多个Y-SNP。

195.根据权利要求194所述的方法,其中,所述一个或多个Y-SNP被包含在所述多个SNP内。

196.根据权利要求194或权利要求195所述的方法,其中,所述一个或多个Y-SNP包括至少20、25、30、35、40、45、50、55、60、65、70、75、80、81、82、83、84或85个Y-SNP。

197.根据权利要求194至196中任一项所述的方法,其中,所述一个或多个Y-SNP包括85个Y-SNP。

198.根据权利要求193至197中任一项所述的方法,其中,计算共享Y染色体的所述似然比包括基于所述一个或多个Y-SNP的所述基因型,将所述DNA谱与所述一个或多个参考DNA谱中的参考DNA谱共享Y染色体的概率除以所述DNA谱与所述参考DNA谱不共享Y染色体的概率。

199.根据权利要求173至198中任一项所述的方法,其中,所述参考DNA谱集中至少90%、91%、92%、93%、94%、95%、96%、97%、98%或99%的所述DNA谱来自失踪人员或灾难或冲突的受害者的亲属。

200.根据权利要求173至199中任一项所述的方法,其中,所述一个或多个参考样本中的每一个来自失踪人员或灾难或冲突的受害者的亲属。

201.根据权利要求173-200中任一项所述的方法,其中,所述一个或多个参考样本包括多达5、10、15、20、25、30、35、40、45、50、55、60、65、70、75、80、85、90、95、100、125、150、175、

200、300、400、500、600、700、800、900或1000个参考样本。

202.根据权利要求173至201中任一项所述的方法,其中,所述一个或多个参考样本的至少5%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%、65%、70%、75%、80%、85%、90%或95%来自所述感兴趣人员的亲属。

203.根据权利要求173至202中任一项所述的方法,其中,所述参考DNA谱集中至少5%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%、65%、70%、75%、80%、85%、90%或95%的所述参考DNA谱来自所述感兴趣人员的亲属,并且其中,所述参考DNA谱集中至少5%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%、65%、70%、75%、80%、85%、90%或95%的所述参考DNA谱中的每一个为一级亲属、二级亲属、三级亲属、四级亲属或五级亲属。

204.根据权利要求173至203中任一项所述的方法,其中,所述一个或多个参考样本的至少50%来自所述感兴趣人员的亲属。

205.根据权利要求199至204中任一项所述的方法,其中,所述一个或多个参考样本中的所述感兴趣人员的每个亲属分别是所述感兴趣人员的一级亲属、二级亲属、三级亲属、四级亲属或五级亲属。

206.根据权利要求205所述的方法,其中,所述一个或多个参考样本中的所述感兴趣人员的每个亲属分别是所述感兴趣人员的一级亲属、二级亲属或三级亲属。

207.根据权利要求173至206中任一项所述的方法,其中,所述一个或多个参考样本中的所述感兴趣人员的每个亲属的身份是已知的。

208.根据权利要求173至207中任一项所述的方法,其中,所述一个或多个参考样本中的每一个的所述身份都是已知的。

209.一种试剂盒,包括至少一个容器装置,其中,所述至少一个容器装置包括根据权利要求94至126中任一项所述的多个引物。

210.根据权利要求1至92和127至208中任一项所述的方法,其中,所述多个SNP包括介于2,000至11,000个SNP、3,000至11,000个SNP、4,000至11,000个SNP、5,000至11,000个SNP、5,500至11,000个SNP、6,000至11,000个SNP、7,000至15,000个SNP、7,000至14,000个SNP、7,000至13,000个SNP、7,000至12,000个SNP、7,000至11,000个SNP、8,000至15,000个SNP、8,000至14,000个SNP、8,000至13,000个SNP、8,000至12,000个SNP、8,000至11,000个SNP、9,000至15,000个SNP、9,000至14,000个SNP、9,000至13,000个SNP、9,000至12,000个SNP或9,000至11,000个SNP,或介于约2,000至11,000个SNP、3,000至11,000个SNP、4,000至11,000个SNP、5,000至11,000个SNP、5,500至11,000个SNP、6,000至11,000个SNP、7,000至15,000个SNP、7,000至14,000个SNP、7,000至13,000个SNP、7,000至12,000个SNP、7,000至11,000个SNP、8,000至15,000个SNP、8,000至14,000个SNP、8,000至13,000个SNP、8,000至12,000个SNP、8,000至11,000个SNP、9,000至15,000个SNP、9,000至14,000个SNP、9,000至13,000个SNP、9,000至12,000个SNP或9,000至11,000个SNP。

211.根据权利要求1至92和127至208中任一项所述的方法,其中,所述多个SNP包括10,230个SNP。

212.根据权利要求94至126中任一项所述的多个引物,其中,所述多个SNP包括介于2,000至11,000个SNP、3,000至11,000个SNP、4,000至11,000个SNP、5,000至11,000个SNP、5,

500至11,000个SNP、6,000至11,000个SNP、7,000至15,000个SNP、7,000至14,000个SNP、7,000至13,000个SNP、7,000至12,000个SNP、7,000至11,000个SNP、8,000至15,000个SNP、8,000至14,000个SNP、8,000至13,000个SNP、8,000至12,000个SNP、8,000至11,000个SNP、9,000至15,000个SNP、9,000至14,000个SNP、9,000至13,000个SNP、9,000至12,000个SNP或9,000至11,000个SNP,或介于约2,000至11,000个SNP、3,000至11,000个SNP、4,000至11,000个SNP、5,000至11,000个SNP、5,500至11,000个SNP、6,000至11,000个SNP、7,000至15,000个SNP、7,000至14,000个SNP、7,000至13,000个SNP、7,000至12,000个SNP、7,000至11,000个SNP、8,000至15,000个SNP、8,000至14,000个SNP、8,000至13,000个SNP、8,000至12,000个SNP、8,000至11,000个SNP、9,000至15,000个SNP、9,000至14,000个SNP、9,000至13,000个SNP、9,000至12,000个SNP或9,000至11,000个SNP。

213.根据权利要求94至126中任一项所述的多个引物,其中,所述多个SNP包括10,230个SNP。

214.根据权利要求1至5、11至93、127至161和210至213中任一项所述的方法,还包括生成家谱,所述家谱包括与所述参考DNA谱集中包含的一个或多个DNA谱相关的所述DNA谱。

215.根据权利要求214所述的方法,其中,所述家谱包括与来自所述感兴趣人员的亲属的一个或多个DNA谱相关的所述DNA谱。

用于失踪人员和灾难/冲突受害者的亲属关系评估的方法和系统

[0001] 相关申请的交叉引用

[0002] 本申请要求于2022年8月16日提交的标题为“METHODS AND SYSTEMS FOR KINSHIP EVALUATION FOR MISSING PERSONS ANDDISASTER/CONFLICT VICTIMS(用于失踪人员和灾难/冲突受害者的亲属关系评估的方法和系统)”的美国临时申请No.63/398,512,以及于2023年2月14日提交的标题为“METHODS AND SYSTEMS FOR KINSHIP EVALUATION FOR MISSING PERSONS AND DISASTER/CONFLICT VICTIMS(用于失踪人员和灾难/冲突受害者的亲属关系评估的方法和系统)”的美国临时申请No.63/445,541的优先权,其内容通过引用整体并入本文。

技术领域

[0003] 本公开在一些方面涉及针对感兴趣人员——诸如失踪人员以及冲突和灾难的受害者——进行基于DNA的亲属关系评估的方法和系统。

背景技术

[0004] 目前,生成用于基因数据库比较的DNA谱的方法包括使用密集SNP微阵列和全基因组测序(WGS)进行基因分型,然后将证据样本与数据库中的远亲关联,这需要大量和高质量的DNA样本,并且不适用于家族搜索或法医目的。法医案例样本通常是数量少和质量低的样本,并且通常需要查询可公开访问的基因数据库来识别家族关系。然而,在失踪人员或灾难或冲突受害者的情况下,如果家庭成员的基因样本被上传到可公开访问的数据库中,他们可能会犹豫是否提供有助于识别这样的人员的基因样本。因此,需要一种新的、改进的方法来生成基于DNA的谱分析,以允许识别感兴趣人员——诸如失踪人员或灾难或冲突的受害者——的家族关系,而无需使用可公开访问的基因数据库。

发明内容

[0005] 本文提供了一种执行基于DNA的亲属关系(kinship)分析的方法,包括:提供来自感兴趣人员的核酸样本,使用与多个靶序列特异性杂交的多个引物来扩增核酸样本,所述多个靶序列共同包括多个至少在2,000至50,000个之间或约2,000至50,000个之间的单核苷酸多态性(SNP),从而生成扩增产物,其中,扩增在一个或多个多重PCR反应中进行,从扩增产物生成核酸库,对从扩增产物生成的核酸库进行测序,分析扩增产物的序列,确定多个SNP的基因型,从而生成DNA谱,以及计算DNA谱与一个或多个参考DNA谱的关系程度(degree of relationship),其中,一个或多个参考DNA谱包含在参考DNA谱集内,所述参考DNA谱集包括来自感兴趣人员的亲属(relative)的一个或多个参考DNA谱。

[0006] 本文还提供了一种执行基于DNA的亲属关系分析的方法,包括:提供来自感兴趣人员的核酸样本,使用与多个靶序列特异性杂交的多个引物来扩增核酸样本,所述多个靶序列共同包括多个至少在2,000至50,000个之间或约2,000至50,000个之间的单核苷酸多态

性(SNP),从而生成扩增产物,其中,扩增在一个或多个多重PCR反应中进行,从扩增产物生成核酸库,对从扩增产物生成的核酸库进行测序,确定多个SNP的基因型,从而生成DNA谱,以及计算DNA谱与一个或多个参考DNA谱的关系程度,其中,一个或多个参考DNA谱包含在参考DNA谱集内,所述参考DNA谱集包括来自感兴趣人员的亲属的一个或多个参考DNA谱。

[0007] 在任何一些这样的实施例中,使用大规模并行测序(MPS)来进行测序。在任何一些这样的实施例中,测序不包括全基因组测序(WGS)。

[0008] 在任何一些这样的实施例中,方法还包括生成包含与一个或多个DNA谱相关的DNA谱的家谱。

[0009] 本文还提供了一种构建感兴趣人员的核酸库的方法,包括:提供来自感兴趣人员的核酸样本,使用与多个靶序列特异性杂交的多个引物来扩增核酸样本,所述多个靶序列共同包括多个至少在2,000至50,000个之间或约2,000至50,000个之间的单核苷酸多态性(SNP),从而生成包括扩增产物的核酸库,其中,扩增在一个或多个多重PCR反应中进行。在一些实施例中,方法还包括对扩增产物进行测序以产生感兴趣人员的DNA谱的步骤。

[0010] 本文还提供了一种构建参考DNA样本的核酸库的方法,包括:提供来自感兴趣人员的亲属的核酸样本,使用与多个靶序列特异性杂交的多个引物来扩增所述核酸样本,所述多个靶序列共同包括多个至少在2,000至50,000个之间或约2,000至50,000个之间的单核苷酸多态性(SNP),从而生成包括扩增产物的核酸库,其中,扩增在一个或多个多重PCR反应中进行。

[0011] 在一些实施例中,亲属是感兴趣人员的一级、二级、三级、四级或五级亲属。在任何一些这样的实施例中,亲属是感兴趣人员物的一级、二级或三级亲属。

[0012] 在任何一些这样的实施例中,核酸样本包括基因组DNA。在任何一些这样的实施例中,核酸样本包括一种或多种酶抑制剂。在任何一些这样的实施例中,一种或多种酶抑制剂包括选自由血红素、亚铁血红素、腐殖酸、靛蓝、单宁酸、胶原蛋白、钙和羟基磷灰石组成的组中的一种或多种抑制剂。在任何一些这样的实施例中,核酸样本包括低质量核酸分子和/或低数量核酸分子。在任何一些这样的实施例中,低质量核酸分子是退化的基因组DNA和/或片段化的基因组DNA。在任何一些这样的实施例中,低质量核酸分子具有为1、2、3、4、5、6、7、8、9、10、15、20、25、30、35、40、45、50、55、60、65、70、75、80、85、90、95、100、105、110、115、120、125、130、135、140、145、150、155、160、165、170、175、180、185、190、195或200的退化指数(DI)或至少为1、2、3、4、5、6、7、8、9、10、15、20、25、30、35、40、45、50、55、60、65、70、75、80、85、90、95、100、105、110、115、120、125、130、135、140、145、150、155、160、165、170、175、180、185、190、195或200的退化指数(DI)。在任何一些这样的实施例中,低质量核酸分子具有至少1且多达或小于158.3的DI。在任何一些这样的实施例中,核酸样本包含高质量核酸分子。在任何一些这样的实施例中,高质量核酸分子具有小于1的DI。

[0013] 在任何一些这样的实施例中,感兴趣人员是失踪人员。在任何一些这样的实施例中,感兴趣人员是灾难或冲突的受害者。

[0014] 在任何一些这样的实施例中,核酸样本源自唾液、血液、精液、毛发、牙齿、骨骼或皮肤。在任何一些这样的实施例中,核酸样本源自唾液、血液或精液。在任何一些这样的实施例中,核酸样本来自骨骼或毛发。在任何一些这样的实施例中,核酸样本源自口腔拭子、纸、织物,或浸渍有唾液、血液、精液或其他体液、或者包含毛发或皮肤细胞的其他基质或物

体。

[0015] 在任何一些这样的实施例中,核酸样本包括在3pg至100ng之间或在约3pg至100ng之间的基因组DNA。在任何一些这样的实施例中,核酸样本包括在100pg至5ng之间或在约100pg至5ng之间的基因组DNA,在50pg至5ng之间或在约50pg至5ng之间的基因组DNA,或者在3pg至5ng之间或在约3pg至5ng之间的基因组DNA。

[0016] 在任何一些这样的实施例中,多个SNP包括亲属关系SNP(kiSNP)。在任何一些这样的实施例中,多个SNP包括Y染色体SNP(Y-SNP)。在任何一些这样的实施例中,多个SNP包括kiSNP和Y-SNP。在任何一些这样的实施例中,多个SNP包括kiSNP、生物地理祖先SNP(aiSNP)、身份SNP(iiSNP)、表型SNP(piSNP)、X-染色体SNP(X-SNP)和Y-染色体SNP(Y-SNP)。在一些这样的实施例中,多个SNP包括选自由kiSNP、aiSNP、iiSNP、piSNP、X-SNP和Y-SNP组成的组中的一个或多个的SNP。在一些这样的实施例中,多个SNP中的至少或至少约80%、81%、82%、83%、84%、85%、86%、87%、88%、89%、90%、91%、92%、93%、94%、95%、96%、97%、98%或99%是亲属关系SNP。

[0017] 在任何一些这样的实施例中,参考DNA谱集包括多达5、10、15、20、25、30、35、40、45、50、55、60、65、70、75、80、85、90、95、100、125、150、175、200、300、400、500、600、700、800、900或1000个参考DNA谱。在任何一些这样的实施例中,参考DNA谱集中至少5%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%、65%、70%、75%、80%、85%、90%或95%的所述参考DNA谱来自所述感兴趣人员的亲属。在任何一些这样的实施例中,参考DNA谱集中至少5%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%、65%、70%、75%、80%、85%、90%或95%的所述参考DNA谱来自所述感兴趣人员的亲属,并且其中,所述参考DNA谱集中至少5%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%、65%、70%、75%、80%、85%、90%或95%的所述参考DNA谱中的每一个为一级亲属、二级亲属、三级亲属、四级亲属或五级亲属。在任何一些这样的实施例中,参考DNA谱集中至少50%的所述参考DNA谱来自所述感兴趣人员的亲属。

[0018] 在任何一些这样的实施例中,参考DNA谱集中的感兴趣人员的每个亲属分别是感兴趣人员的一级亲属、二级亲属、三级亲属、四级亲属或五级亲属。在任何一些这样的实施例中,参考DNA谱集中的感兴趣人员的每位亲属分别是感兴趣人员的一级亲属、二级亲属或三级亲属。在任何一些这样的实施例中,参考DNA谱集中的感兴趣人员的每个亲属的身份都是已知的。在任何一些这样的实施例中,参考DNA谱集中的一个或多个参考DNA谱中的每一个的身份都是已知的。在任何一些这样的实施例中,参考DNA谱集在数据库中。在一些实施例中,数据库不可公开访问。

[0019] 在任何一些这样的实施例中,测序包括多达40重的测序复杂性。在任何一些这样的实施例中,测序包括多达32重的测序复杂性。在任何一些这样的实施例中,测序包括多达12重至32重的测序复杂性。在任何一些这样的实施例中,测序包括24重至32重的测序复杂性。在任何一些这样的实施例中,测序包括10重、11重、12重、13重、14重、15重、16重、17重、18重、19重、20重、21重、22重、23重、24重、25重、26重、27重、28重、29重、30重、31重、32重、33重、34重或35重的测序复杂性或者约10重、11重、12重、13重、14重、15重、16重、17重、18重、19重、20重、21重、22重、23重、24重、25重、26重、27重、28重、29重、30重、31重、32重、33重、34重或35重的测序复杂性。在任何一些这样的实施例中,所述测序包括对于死后样本为8至16

重或约8至16重的测序复杂性,和/或所述测序包括对于生前样本为24至40重或约24至40重的测序复杂性对于死后样本,测序包括8到16重或约8到16重的测序复杂性,和/或对于生前样本,测序包括24到40重或约24到40重的测序复杂性。在任何一些这样的实施例中,所述测序包括对于死后样本为12重或约12重的测序复杂性,和/或所述测序包括对于生前样本为32重或约32重的测序复杂性对于死后样本,测序包括12重或约12重的测序复杂性,和/或对于生前样本,测序包括32重或约32重的测序复杂性。在任何一些这样的实施例中,所述测序包括为30重、31重或32重的测序复杂性或者约为30重、31重或32重的测序复杂性测序包括约30重、31重或32重的测序复杂性。

[0020] 在任何一些这样的实施例中,方法还包括识别感兴趣人员。

[0021] 本文还提供了一种计算关联度的方法,包括:获得包括至少在2,000至50,000个SNP之间或在约2,000至50,000个SNP之间的基因型的DNA谱,其中,所述DNA谱来自感兴趣人员;以及计算DNA谱与一个或多个参考DNA谱的关系程度,其中,一个或多个参考DNA谱包括在参考DNA谱集内,所述参考DNA谱集包括来自感兴趣人员的亲属的一个或多个参考DNA谱。

[0022] 本文还提供了一种计算关联度的方法,包括:生成包括至少在2,000至50,000个SNP之间或在约2,000至50,000个SNP之间的基因型的DNA谱,其中,所述DNA谱来自感兴趣人员;以及计算DNA谱与一个或多个参考DNA谱的关系程度,其中,一个或多个参考DNA谱包括在参考DNA谱集内,所述参考DNA谱集包括来自感兴趣人员的亲属的一个或多个参考DNA谱。

[0023] 在任何一些这样的实施例中,使用亲属关系模型来计算关系程度。在任何一些这样的实施例中,使用利用PCA方法训练的亲属关系模型来计算关系程度。在一些实施例中,用于训练亲属关系模型的PCA方法是PCA或涉及PCA。在任何一些这样的实施例中,PCA方法是PC-AiR。在一些实施例中,PC-AiR包括以下步骤:(1)估计训练数据库的所有样本对——可选地训练DNA谱——之间的亲属关系系数,其中,亲属关系系数>0.025的配对被识别为相关的,而亲属关系系数<-0.025的配对被识别为祖先分化的;(2)初始化包括所有样本的无关样本集;以及(3)迭代地:(i)识别所述无关样本集中具有所述无关样本集中最相关样本的集合,从而指定为X,(ii)识别X中与所述无关样本集中的样本相比具有最少祖先分化配对的样本集,从而指定为Y;以及(iii)如果Y具有零个样本,则终止过程,或者,如果Y具有至少一个样本,则从Y中随机选择一个样本以从U中移除,并且在步骤(3)(i)处开始重复。

[0024] 在任何一些这样的实施例中,PCA方法是改进的PC-Air。在任何一些这样的实施例中,改进的PC-AiR包括以下步骤:(1)估计训练数据库的所有样本对(可选地训练DNA谱)之间的亲属关系系数,其中,亲属关系系数>0.01的配对被识别为相关的,而亲属关系系数<-0.025的配对被识别为祖先分化的;(2)移除所有具有≥5%失踪数据的DNA谱;以及(3)通过用排序值识别每个DNA谱来对所有DNA谱进行排序。在一些实施例中,排序值是基于完整数据库中从最少到最多排序的相关DNA谱的数量来确定的,并且通过完整数据库中从最多到最少排序的祖先分化DNA谱的数量来打破纽带(tie)。在一些实施例中,步骤(3)涉及迭代地遍历排序后的DNA谱,并且对于每个DNA谱:(i)如果DNA谱尚未出现在相关样本集中,则将其添加到无关样本集并且将所有相关DNA谱添加到相关样本集,以及(ii)如果DNA谱已经在相关样本集中,则跳至下一个DNA谱,并且在步骤任何一些这样的(3)(i)处开始重复。

[0025] 在任何一些这样的实施例中,计算关系程度包括使用PC-Relate来计算亲属关系系数。在任何一些这样的实施例中,通过将感兴趣人员的DNA谱作为PC-Relate的输入来计

算关系程度。在任何一些这样的实施例中,通过将亲属关系模型和感兴趣人员的DNA谱作为PC-Relate的输入来计算关系程度。在任何一些这样的实施例中,一个或多个参考DNA谱还被提供作为PC-Relate的输入。

[0026] 在任何一些这样的实施例中,计算关系程度包括使用全基因组亲属关系算法来计算亲属关系系数,如下所示:

[0027] $$\widehat{\phi_{ij}} = \frac{\sum_{s\in S}(g_{is} - 2\hat{u}_{is})(g_{js} - 2\hat{u}_{js})}{4\sum_{s\in S}[\hat{u}_{is}(1 - \hat{u}_{is}) \cdot \hat{u}_{js}(1 - \hat{u}_{js})]^{1/2}}$$

[0028] 其中,感兴趣人员和一个或多个参考DNA谱中的参考DNA谱是i和j,$\widehat{\phi_{ij}}$是亲属关系系数,$\hat{u}$是估计等位基因频率,s是在两个个体中都被分型的S个SNP中的SNP,g_{is}和g_{js}分别是在SNP s处的i和j中的参考等位基因的数量,并且$\hat{u}_{is}$和$\hat{u}_{js}$分别是由在SNP s处的i和j的PC-AiR计算出的预期等位基因频率。

[0029] 在任何一些这样的实施例中,计算关系程度包括计算似然比。在一些实施例中,计算似然比包括比较DNA谱和一个或多个参考DNA谱之间的多个SNP。在一些实施例中,计算似然比包括比较DNA谱和一个或多个参考DNA谱之间的SNP集合,所述SNP集合包括来自所述多个SNP内的亲属关系SNP。

[0030] 在任何一些这样的实施例中,计算似然比包括基于多个SNP的基因型,将DNA谱与一个或多个参考DNA谱中的参考DNA谱相关的概率除以DNA谱与参考DNA谱不相关的概率。

[0031] 在任何一些这样的实施例中,似然比(LR)计算如下:

[0032] $$LR = \frac{P(D \mid H_r)}{P(D \mid H_u)}$$

[0033] 其中,D表示所述基因型,H_r表示个体是相关的假设,并且H_u表示个体是无关的假设。

[0034] 在任何一些这样的实施例中,LR计算如下:

[0035] $$LR = \frac{0.001}{p^2q^2}$$

[0036] 其中,0.001表示基因分型错误率,p是等位基因1的等位基因频率,并且q是等位基因2的等位基因频率。

[0037] 在任何一些这样的实施例中,感兴趣人员在生物学上是男性,并且方法还包括计算DNA谱与一个或多个参考DNA谱之间共享Y染色体的似然比。在一些实施例中,计算共享Y染色体的似然比包括比较DNA谱和一个或多个参考DNA谱之间的SNP集合,所述SNP集合包括一个或多个Y-SNP。

[0038] 在一些这样的实施例中,一个或多个Y-SNP包含在多个SNP内。在一些实施例中,一个或多个Y-SNP包括至少20、25、30、35、40、45、50、55、60、65、70、75、80、81、82、83、84或85个Y-SNP。在一些这样的实施例中,一个或多个Y-SNP包括85个Y-SNP。

[0039] 在任何一些这样的实施例中,计算共享Y染色体的似然比包括基于一个或多个Y-SNP的基因型,将DNA谱与一个或多个参考DNA谱中的参考DNA谱共享Y染色体的概率除以DNA谱与参考DNA谱不共享Y染色体的概率。

[0040] 在任何一些这样的实施例中,参考DNA谱集中至少90%、91%、92%、93%、94%、

95％、96％、97％、98％或99％的DNA谱来自失踪人员或灾难或冲突的受害者的亲属。在任何一些这样的实施例中,参考DNA谱集中的每个DNA谱来自失踪人员或灾难或冲突的受害者的亲属。

[0041] 在任何一些这样的实施例中,参考DNA谱集包括多达5、10、15、20、25、30、35、40、45、50、55、60、65、70、75、80、85、90、95、100、125、150、175、200、300、400、500、600、700、800、900或1000个参考DNA谱。在任何一些这样的实施例中,参考DNA谱集包括多达100个参考DNA谱。

[0042] 在任何一些这样的实施例中,参考DNA谱集中至少5％、10％、15％、20％、25％、30％、35％、40％、45％、50％、55％、60％、65％、70％、75％、80％、85％、90％或95％的参考DNA谱来自感兴趣人员的亲属。在任何一些这样的实施例中,参考DNA谱集中至少50％的参考DNA谱来自感兴趣人员的亲属。

[0043] 在任何一些这样的实施例中,参考DNA谱集中的感兴趣人员的每个亲属分别是感兴趣人员的一级亲属、二级亲属、三级亲属、四级亲属或五级亲属。

[0044] 在任何一些这样的实施例中,参考DNA谱集中的感兴趣人员的每个亲属的身份都是已知的。在任何一些这样的实施例中,DNA参考谱集中的感兴趣人员的每个亲属的身份都是已知的。在任何一些这样的实施例中,参考DNA谱集中的一个或多个参考DNA谱中的每一个的身份都是已知的。

[0045] 在任何一些这样的实施例中,参考DNA谱集在数据库中。在一些实施例中,数据库不可公开访问。在任何一些这样的实施例中,所述数据库不能被第三方遗传学(geneaological)服务访问。

[0046] 本文还提供了使用本文所述任何方法来构建的核酸库。

[0047] 本文还提供了多个引物,所述多个引物与多个靶序列特异性杂交,所述多个靶序列包括来自感兴趣人员的核酸样本中的至少在2,000至50,000个之间的单核苷酸多态性(SNP)或约2,000至50,000个之间的所述SNP,其中,在一个或多个多重PCR反应中使用所述多个引物来扩增所述核酸样本以产生扩增产物。

[0048] 本文还提供了多个引物,所述多个引物与多个靶序列特异性杂交,所述多个靶序列包括来自感兴趣人员的核酸样本和来自一个或多个参考样本的核酸样本中的至少在2,000至50,000个之间的单核苷酸多态性(SNP)或约2,000至50,000个之间的所述SNP,其中,一个或多个参考样本包括来自所述感兴趣人员的亲属的样本,并且其中,在一个或多个多重PCR反应中使用多个引物来扩增来自感兴趣人员的核酸样本和来自一个或多个参考样本的核酸样本以产生扩增产物。

[0049] 在任何一些这样的实施例中,来自感兴趣人员的核酸样本包括基因组DNA。

[0050] 在任何一些这样的实施例中,来自感兴趣人员的核酸样本包括一种或多种酶抑制剂。在任何一些这样的实施例中,一种或多种酶抑制剂包括选自由血红素、亚铁血红素、腐殖酸、靛蓝、单宁酸、胶原蛋白、钙和羟基磷灰石组成的组中的一种或多种抑制剂。在任何一些这样的实施例中,来自感兴趣人员的核酸样本包括低质量核酸分子和/或低数量核酸分子。在一些实施例中,低质量核酸分子是退化的基因组DNA和/或片段化的基因组DNA。在任何一些这样的实施例中,低质量核酸分子具有为1、2、3、4、5、6、7、8、9、10、15、20、25、30、35、40、45、50、55、60、65、70、75、80、85、90、95、100、105、110、115、120、125、130、135、140、145、

150、155、160、165、170、175、180、185、190、195或200的退化指数(DI)或至少为1、2、3、4、5、6、7、8、9、10、15、20、25、30、35、40、45、50、55、60、65、70、75、80、85、90、95、100、105、110、115、120、125、130、135、140、145、150、155、160、165、170、175、180、185、190、195或200的退化指数(DI)。在任何一些这样的实施例中,低质量核酸分子具有至少1且多达或小于158.3的DI。在任何一些这样的实施例中,来自感兴趣人员的核酸样本和/或来自一个或多个参考样本的核酸样本包括高质量核酸分子。在任何一些这样的实施例中,高质量核酸分子具有小于1的DI。

[0051] 在任何一些这样的实施例中,感兴趣人员是失踪人员。在任何一些这样的实施例中,感兴趣人员是灾难或冲突的受害者。

[0052] 在任何一些这样的实施例中,来自感兴趣人员的核酸样本源自口腔拭子、纸、织物,或浸渍有唾液、血液或其他体液、或者包含毛发或皮肤细胞的其他基质或物体。

[0053] 在任何一些这样的实施例中,来自感兴趣人员的核酸样本包括在3pg至100ng之间或在约3pg至100ng之间的基因组DNA。在任何一些这样的实施例中,来自感兴趣人员的核酸样本包括在100pg至5ng之间或在约100pg至5ng之间的基因组DNA,在50pg至5ng之间或在约50pg至5ng之间的基因组DNA,或者在3pg至5ng之间或在约3pg至5ng之间的基因组DNA。在任何一些这样的实施例中,来自感兴趣人员的核酸样本包括1ng或约1ng的基因组DNA。

[0054] 在任何一些这样的实施例中,多个SNP包括亲属关系SNP(kiSNP)。在任何一些这样的实施例中,多个SNP包括Y染色体SNP(Y-SNP)。在任何一些这样的实施例中,多个SNP包括kiSNP和Y-SNP。在任何一些这样的实施例中,多个SNP包括kiSNP、生物地理祖先SNP(aiSNP)、身份SNP(iiSNP)、表型SNP(piSNP)、X染色体SNP(X-SNP)和Y染色体SNP(Y-SNP)。在任何一些这样的实施例中,多个SNP包括选自由kiSNP、aiSNP、iiSNP、piSNP、X-SNP和Y-SNP组成的组中的一个或多个的SNP。

[0055] 在任何一些这样的实施例中,多个SNP中的至少或至少约80%、81%、82%、83%、84%、85%、86%、87%、88%、89%、90%、91%、92%、93%、94%、95%、96%、97%、98%或99%是亲属关系SNP。在任何一些这样的实施例中,参考DNA谱集中至少90%、91%、92%、93%、94%、95%、96%、97%、98%或99%的DNA谱来自失踪人员或灾难或冲突的受害者的亲属。

[0056] 在任何一些这样的实施例中,一个或多个参考样本中的每一个来自失踪人员或灾难或冲突的受害者的亲属。在任何一些这样的实施例中,一个或多个参考样本包括多达5、10、15、20、25、30、35、40、45、50、55、60、65、70、75、80、85、90、95、100、125、150、175、200、300、400、500、600、700、800、900或1000个参考样本。在任何一些这样的实施例中,一个或多个参考样本中的至少5%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%、65%、70%、75%、80%、85%、90%或95%来自所述感兴趣人员的亲属。在任何一些这样的实施例中,参考DNA谱集中至少5%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%、65%、70%、75%、80%、85%、90%或95%的参考DNA谱来自所述感兴趣人员的亲属,并且其中,参考DNA谱集中至少5%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%、65%、70%、75%、80%、85%、90%或95%的参考DNA谱中的每一个为一级亲属、二级亲属、三级亲属、四级亲属或五级亲属。在任何一些这样的实施例中,至少50%的一个或多个参考样本来自感兴趣人员的亲属。在任何一些这样的实施例中,一个或多个参

考样本中的感兴趣人员的每个亲属分别是所述感兴趣人员的一级亲属、二级亲属、三级亲属、四级亲属或五级亲属。在任何一些这样的实施例中,一个或多个参考样本中的感兴趣人员的每个亲属分别是感兴趣人员的一级亲属、二级亲属或三级亲属。在任何一些这样的实施例中,一个或多个参考样本中的所述感兴趣人员的每个亲属的身份是已知的。在任何一些这样的实施例中,一个或多个参考样本中的每一个的身份是已知的。

[0057] 本文还提供了一种构建DNA谱的方法,包括:提供来自感兴趣人员的核酸样本,使用与多个靶序列特异性杂交的多个引物来扩增所述核酸样本,所述多个靶序列共同包括多个至少在2,000至50,000个之间或约2,000至50,000个之间的单核苷酸多态性(SNP),从而生成扩增产物,其中,扩增在一个或多个多重PCR反应中进行,对所述扩增产物进行测序,确定多个SNP的基因型,从而生成DNA谱。

[0058] 本文还提供了一种构建DNA谱的方法,包括:提供来自感兴趣人员的核酸样本,提供来自感兴趣人员的亲属的核酸样本,使用与多个靶序列特异性杂交的多个引物来扩增来自感兴趣人员的核酸样本和来自亲属的核酸样本,所述多个靶序列共同包括多个至少在2,000至50,000个之间或约2,000至50,000个之间的单核苷酸多态性(SNP),从而生成扩增产物,其中,扩增在一个或多个多重PCR反应中进行,对所述扩增产物进行测序,确定多个SNP的基因型,从而生成感兴趣人员和感兴趣人员的亲属的DNA谱。

[0059] 在任何一些这样的实施例中,测序不包括全基因组测序(WGS)。在任何一些这样的实施例中,核酸样本包括基因组DNA。在任何一些这样的实施例中,感兴趣人员的核酸样本和/或感兴趣人员的亲属的核酸样本包括基因组DNA。

[0060] 在任何一些这样的实施例中,核酸样本、感兴趣人员的核酸样本和/或亲属的核酸样本包含一种或多种酶抑制剂。在一些实施例中,一种或多种酶抑制剂包括选自由血红素、亚铁血红素、腐殖酸、靛蓝、单宁酸、胶原蛋白、钙和羟基磷灰石组成的组中的一种或多种抑制剂。

[0061] 在任何一些这样的实施例中,核酸样本、感兴趣人员的核酸样本和/或亲属的核酸样本包括低质量核酸分子和/或低数量核酸分子。在一些实施例中,低质量核酸分子是退化的基因组DNA和/或片段化的基因组DNA。在任何一些这样的实施例中,低质量核酸分子具有为1、2、3、4、5、6、7、8、9、10、15、20、25、30、35、40、45、50、55、60、65、70、75、80、85、90、95、100、105、110、115、120、125、130、135、140、145、150、155、160、165、170、175、180、185、190、195或200的退化指数(DI)或至少为1、2、3、4、5、6、7、8、9、10、15、20、25、30、35、40、45、50、55、60、65、70、75、80、85、90、95、100、105、110、115、120、125、130、135、140、145、150、155、160、165、170、175、180、185、190、195或200的退化指数(DI)。在任何一些这样的实施例中,低质量核酸分子具有至少1且多达或小于158.3的DI。在任何一些这样的实施例中,核酸样本、感兴趣人员的核酸样本和/或亲属的核酸样本包含高质量核酸分子。在一些实施例中,高质量核酸分子具有小于1的DI。

[0062] 在任何一些这样的实施例中,感兴趣人员是失踪人员。在任何一些这样的实施例中,感兴趣人员是灾难或冲突的受害者。

[0063] 在任何一些这样的实施例中,感兴趣人员的亲属是一级、二级、三级、四级或五级亲属。在任何一些这样的实施例中,感兴趣人员的亲属是一级、二级或三级亲属。

[0064] 在任何一些这样的实施例中,核酸样本、感兴趣人员员的核酸样本和/或亲属的核

酸样本源自口腔拭子、纸、织物,或浸渍有唾液、血液或其他体液、或者包含毛发或皮肤细胞的其他基质或物体。

[0065] 在任何一些这样的实施例中,核酸样本、感兴趣人员的核酸样本和/或亲属的核酸样本包括在3pg至100ng之间或在约3pg至100ng之间的基因组DNA。在任何一些这样的实施例中,核酸样本、感兴趣人员的核酸样本和/或亲属的核酸样本包括在100pg至5ng之间或在约100pg至5ng之间的基因组DNA,在50pg至5ng之间或在约50pg至5ng之间的基因组DNA,或者在3pg至5ng之间或在约3pg至5ng之间的基因组DNA。在任何一些这样的实施例中,核酸样本、感兴趣人员的核酸样本和/或亲属的核酸样本包括1ng或约1ng的基因组DNA。

[0066] 在任何一些这样的实施例中,多个SNP包括亲属关系SNP。在任何一些这样的实施例中,多个SNP包括Y染色体SNP(Y-SNP)。在任何一些这样的实施例中,多个SNP包括kiSNP和Y-SNP。在任何一些这样的实施例中,多个SNP包括kiSNP、生物地理祖先SNP(aiSNP)、身份SNP(iiSNP)、表型SNP(piSNP)、X-染色体SNP(X-SNP)和Y-染色体SNP(Y-SNP)。在一些这样的实施例中,多个SNP包括选自由kiSNP、aiSNP、iiSNP、piSNP、X-SNP和Y-SNP组成的组中的一个或多个的SNP。在一些这样的实施例中,多个SNP中的至少或至少约80%、81%、82%、83%、84%、85%、86%、87%、88%、89%、90%、91%、92%、93%、94%、95%、96%、97%、98%或99%是亲属关系SNP。

[0067] 在任何一些这样的实施例中,测序包括多达40重的测序复杂性。在任何一些这样的实施例中,测序包括多达32重的测序复杂性。在任何一些这样的实施例中,测序包括12重至32重的测序复杂性。在任何一些这样的实施例中,测序包括24重至32重的测序复杂性。在任何一些这样的实施例中,测序包括4重、5重、6重、7重、8重、9重、10重、11重、12重、13重、14重、15重、16重、17重、18重、19重、20重、21重、22重、23重、24重、25重、26重、27重、28重、29重、30重、31重、32重、33重、34重、35重、36重、37重、38重、39重、40重、41重、42重、43重、44重、45重的测序复杂性或者约4重、5重、6重、7重、8重、9重、10重、11重、12重、13重、14重、15重、16重、17重、18重、19重、20重、21重、22重、23重、24重、25重、26重、27重、28重、29重、30重、31重、32重、33重、34重、35重、36重、37重、38重、39重、40重、41重、42重、43重、44重、45重的测序复杂性。在任何一些这样的实施例中,测序包括10重、11重、12重、13重、14重、15重、16重、17重、18重、19重、20重、21重、22重、23重、24重、25重、26重、27重、28重、29重、30重、31重、32重、33重、34重或35重的测序复杂性或者约10重、11重、12重、13重、14重、15重、16重、17重、18重、19重、20重、21重、22重、23重、24重、25重、26重、27重、28重、29重、30重、31重、32重、33重、34重或35重的测序复杂性。在任何一些这样的实施例中,测序包括对于死后样本为8至16重或约8至16重的测序复杂性,和/或所述测序包括对于生前样本为24至40重或约24至40重的测序复杂性。在任何一些这样的实施例中,测序包括对于死后样本为12重或约12重的测序复杂性,和/或所述测序包括对于生前样本为32重或约32重的测序复杂性。在任何一些这样的实施例中,测序包括为30重、31重或32重的测序复杂性或者约为30重、31重或32重的测序复杂性。

[0068] 本文还提供了一种识别DNA谱的遗传亲属的方法,包括:计算根据权利要求127至161中任一项所述的DNA谱与一个或多个参考DNA谱的关系程度,其中,一个或多个参考DNA谱包括在参考DNA谱集内,所述参考DNA谱集包括来自感兴趣人员的亲属的一个或多个参考DNA谱;以及生成家谱,所述家谱包括与所述一个或多个参考DNA谱相关的DNA谱。

[0069] 在一些实施例中,一个或多个参考DNA谱是数据库的一部分。

[0070] 在任何一些这样的实施例中,参考DNA谱集包括多达5、10、15、20、25、30、35、40、45、50、55、60、65、70、75、80、85、90、95、100、125、150、175、200、300、400、500、600、700、800、900或1000个参考DNA谱。在任何一些这样的实施例中,参考DNA谱集中的至少5%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%、65%、70%、75%、80%、85%、90%或95%的所述参考DNA谱来自感兴趣人员的亲属。在任何一些这样的实施例中,参考DNA谱集中至少50%的参考DNA谱来自感兴趣人员的亲属。在任何一些这样的实施例中,感兴趣人员的每个亲属分别是感兴趣人员的一级亲属、二级亲属、三级亲属、四级亲属或五级亲属。

[0071] 在任何一些这样的实施例中,参考DNA谱集中的所述感兴趣人员的每个亲属的身份都是已知的。在任何一些这样的实施例中,参考DNA谱集中的所述一个或多个参考DNA谱中的每一个的身份都是已知的。

[0072] 在任何一些这样的实施例中,参考DNA谱集在数据库中。在一些实施例中,数据库不可公开访问。在任何一些这样的实施例中,数据库不能被第三方遗传学服务访问。

[0073] 本文还提供了一种识别DNA谱身份的方法,包括:计算包括至少在2,000至50,000个SNP之间或在约2,000至50,000个SNP之间的基因型的DNA谱与一个或多个参考DNA谱的关系程度,其中,DNA谱来自感兴趣人员,并且其中,一个或多个参考DNA谱包括在参考DNA谱集内,所述参考DNA谱集包括来自感兴趣人员的亲属的一个或多个参考DNA谱;并且生成家谱,所述家谱包括与所述一个或多个参考DNA谱相关的所述DNA谱。

[0074] 在一些实施例中,DNA谱是通过本文描述的任何一种生成DNA谱的方法来生成的。

[0075] 在任何一些这样的实施例中,使用亲属关系模型来计算关系程度。在任何一些这样的实施例中,使用利用PCA方法训练的亲属关系模型来计算关系程度。在一些实施例中,用于训练亲属关系模型的PCA方法是PCA或涉及PCA。在任何一些这样的实施例中,PCA方法是PC-AiR。在任何一些这样的实施例中,PC-AiR包括以下步骤:(1)估计训练数据库的所有样本对——可选地训练DNA谱——之间的亲属关系系数,其中,亲属关系系数>0.025的配对被识别为相关的,而亲属关系系数<-0.025的配对被识别为祖先分化的;(2)初始化包括所有样本的无关样本集;以及(3)迭代地:(i)识别所述无关样本集中具有所述无关样本集中最相关样本的集合,从而指定为X,(ii)识别X中与所述无关样本集中的样本相比具有最少祖先分化配对的样本集,从而指定为Y;以及(iii)如果Y具有零个样本,则终止过程,或者,如果Y具有至少一个样本,则从Y中随机选择一个样本以从U中移除,并且在步骤(3)(i)处开始重复。

[0076] 在任何一些这样的实施例中,PCA方法是改进的PC-Air。在一些实施例中,改进的PC-AiR包括以下步骤:(1)估计训练数据库的所有样本对(可选地训练DNA谱)之间的亲属关系系数,其中,亲属关系系数>0.01的配对被识别为相关的,而亲属关系系数<-0.025的配对被识别为祖先分化的;(2)移除所有具有≥5%失踪数据的DNA谱;以及(3)通过用排序值识别每个DNA谱来对所有DNA谱进行排序。在一些实施例中,排序值是基于完整数据库中从最少到最多排序的相关DNA谱的数量来确定的,并且通过所述完整数据库中从最多到最少排序的祖先分化DNA谱的数量来打破纽带(tie)。在一些实施例中,步骤(3)涉及迭代地遍历排序后的DNA谱,并且对于每个DNA谱:(i)如果DNA谱尚未出现在相关样本集中,则将其添加到

无关样本集并且将所有相关DNA谱添加到相关样本集,以及(ii)如果DNA谱已经在相关样本集中,则跳至下一个DNA谱,并且在步骤(3)(i)处开始重复。

[0077] 在任何一些这样的实施例中,计算关系程度包括使用PC-Relate来计算亲属关系系数。在一些实施例中,通过提供感兴趣人员的DNA谱作为PC-Relate的输入来计算关系程度。在任何一些这样的实施例中,通过提供亲属关系模型和感兴趣人员的DNA谱作为PC-Relate的输入来计算关系程度。

[0078] 在任何一些这样的实施例中,还提供一个或多个参考DNA谱作为PC-Relate的输入。

[0079] 在任何一些这样的实施例中,计算关系程度包括使用全基因组亲属关系算法来计算亲属关系系数,如下所示:

[0080] $$\widehat{\phi_{ij}} = \frac{\sum_{s\in S}(g_{is} - 2\hat{u}_{is})(g_{js} - 2\hat{u}_{js})}{4\sum_{s\in S}[\hat{u}_{is}(1 - \hat{u}_{is}) \cdot \hat{u}_{js}(1 - \hat{u}_{js})]^{1/2}}$$

[0081] 其中,感兴趣人员和一个或多个参考DNA谱中的参考DNA谱是i和j,φ_{ij}是亲属关系系数,$\hat{u}$是估计等位基因频率,s是在两个个体中都被分型的S个SNP中的SNP,g_{is}和g_{js}分别是在SNP s处的i和j中的参考等位基因的数量,并且$\hat{u}_{is}$和$\hat{u}_{js}$分别是由在SNP s处的i和j的PC-AiR计算出的预期等位基因频率。

[0082] 在任何一些这样的实施例中,计算关系程度包括计算似然比。在一些实施例中,计算似然比包括比较所述DNA谱和所述一个或多个参考DNA谱之间的多个SNP。在一些实施例中,计算似然比包括比较DNA谱和一个或多个参考DNA谱之间的SNP集合,所述SNP集合包括来自多个SNP内的亲属关系SNP。

[0083] 在任何一些这样的实施例中,计算似然比包括基于多个SNP的基因型,将DNA谱与一个或多个参考DNA谱中的参考DNA谱相关的概率除以DNA谱与参考DNA谱不相关的概率。

[0084] 在任何一些这样的实施例中,似然比(LR)计算如下:

[0085] $$LR = \frac{P(D \mid H_r)}{P(D \mid H_u)}$$

[0086] 其中,D表示所述基因型,H_r表示个体是相关的假设,并且H_u表示个体是无关的假设。

[0087] 在任何一些这样的实施例中,LR计算如下:

[0088] $$LR = \frac{0.001}{p^2q^2}$$

[0089] 其中,0.001表示基因分型错误率,p是等位基因1的等位基因频率,并且q是等位基因2的等位基因频率。

[0090] 在任何一些这样的实施例中,感兴趣人员在生物学上是男性,并且所述方法还包括计算DNA谱与一个或多个参考DNA谱之间共享Y染色体的似然比。在一些实施例中,计算共享Y染色体的似然比包括比较DNA谱和一个或多个参考DNA谱之间的SNP集合,所述SNP集合包括一个或多个Y-SNP。在一些实施例中,一个或多个Y-SNP包含在多个SNP内。在任何一些这样的实施例中,一个或多个Y-SNP包括至少20、25、30、35、40、45、50、55、60、65、70、75、80、81、82、83、84或85个Y-SNP。在任何一些这样的实施例中,一个或多个Y-SNP包括85个Y-SNP。

[0091] 在任何一些这样的实施例中,计算共享Y染色体的似然比包括基于一个或多个Y-SNP的基因型,将DNA谱与一个或多个参考DNA谱中的参考DNA谱共享Y染色体的概率除以DNA谱与参考DNA谱不共享Y染色体的概率。

[0092] 在任何一些这样的实施例中,参考DNA谱集中至少90%、91%、92%、93%、94%、95%、96%、97%、98%或99%的所述DNA谱来自失踪人员或灾难或冲突的受害者的亲属。

[0093] 在任何一些这样的实施例中,一个或多个参考样本中的每一个来自失踪人员或灾难或冲突的受害者的亲属。在任何一些这样的实施例中,一个或多个参考样本包括多达5、10、15、20、25、30、35、40、45、50、55、60、65、70、75、80、85、90、95、100、125、150、175、200、300、400、500、600、700、800、900或1000个参考样本。

[0094] 在任何一些这样的实施例中,一个或多个参考样本的至少5%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%、65%、70%、75%、80%、85%、90%或95%来自所述感兴趣人员的亲属。在任何一些这样的实施例中,参考DNA谱集中至少5%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%、65%、70%、75%、80%、85%、90%或95%的参考DNA谱来自感兴趣人员的亲属,并且其中,参考DNA谱集中至少5%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%、65%、70%、75%、80%、85%、90%或95%的参考DNA谱中的每一个为一级亲属、二级亲属、三级亲属、四级亲属或五级亲属。在任何一些这样的实施例中,一个或多个参考样本的至少50%来自感兴趣人员的亲属。在任何一些这样的实施例中,一个或多个参考样本中的感兴趣人员的每个亲属分别是感兴趣人员的一级亲属、二级亲属、三级亲属、四级亲属或五级亲属。在任何一些这样的实施例中,一个或多个参考样本中的感兴趣人员的每个亲属分别是感兴趣人员的一级亲属、二级亲属或三级亲属。

[0095] 在任何一些这样的实施例中,一个或多个参考样本中的感兴趣人员的每个亲属的身份是已知的。在任何一些这样的实施例中,一个或多个参考样本中的每一个的身份都是已知的。

[0096] 本文还提供了包含至少一个容器装置的试剂盒,其中,至少一个容器装置包括本文描述的多个引物中的任一个。在任何一些这样的实施例中,多个SNP包括介于或介于约2,000至11,000个SNP、3,000至11,000个SNP、4,000至11,000个SNP、5,000至11,000个SNP、5,500至11,000个SNP、6,000至11,000个SNP、7,000至15,000个SNP、7,000至14,000个SNP、7,000至13,000个SNP、7,000至12,000个SNP、7,000至11,000个SNP、8,000至15,000个SNP、8,000至14,000个SNP、8,000至13,000个SNP、8,000至12,000个SNP、8,000至11,000个SNP、9,000至15,000个SNP、9,000至14,000个SNP、9,000至13,000个SNP、9,000至12,000个SNP或9,000至11,000个SNP。在任何一些这样的实施例中,多个SNP包括10,230个SNP。在任何一些这样的实施例中,多个SNP包括介于或介于约2,000至11,000个SNP、3,000至11,000个SNP、4,000至11,000个SNP、5,000至11,000个SNP、5,500至11,000个SNP、6,000至11,000个SNP、7,000至15,000个SNP、7,000至14,000个SNP、7,000至13,000个SNP、7,000至12,000个SNP、7,000至11,000个SNP、8,000至15,000个SNP、8,000至14,000个SNP、8,000至13,000个SNP、8,000至12,000个SNP、8,000至11,000个SNP、9,000至15,000个SNP、9,000至14,000个SNP、9,000至13,000个SNP、9,000至12,000个SNP或9,000至11,000个SNP。在任何一些这样的实施例中,多个SNP包括10,230个SNP。

[0097] 在任何一些这样的实施例中,方法还包括生成家谱,所述家谱包括与参考DNA谱集中包含的一个或多个DNA谱相关的DNA谱。在一些实施例中,家谱包括与来自感兴趣人员的亲属的一个或多个DNA谱相关的DNA谱。

附图说明

[0098] 图1描述了生成能够被测序的库的方法的示例性示意图。

[0099] 图2显示了使用包括5ng、2.5ng、1ng、500pg、250pg、100pg和50pg的不同输入滴定度的基因组DNA所识别的基因座(loci)数量的结果。

[0100] 图3显示了使用本文所述测定法检测到的具有退化DNA的基因座的百分比(调用率)与微阵列(GSA)调用率的比较。

[0101] 图4显示了与参考对照相比,在存在抑制剂血红素、腐殖酸、靛蓝和单宁酸的情况下检测到的基因座的数量。

[0102] 图5示出了通过本文所述方法生成的示例性家谱。

[0103] 图6显示了使用本文所述算法计算出的预期和观察到的亲属关系系数。

[0104] 图7显示了在示例案例研究中的1:many搜索算法的结果。

[0105] 图8描绘了根据1:many搜索算法的结果生成的示例性家谱。

[0106] 图9是总结使用不同输入滴定度的基因组DNA(包括5ng、2.5ng、1ng、500pg、250pg、100pg和50pg)检测到的基因座的数量和类型的表格。

[0107] 图10是总结了在存在抑制剂血红素、腐殖酸、单宁酸和靛蓝的情况下使用DNA检测到的基因座的数量和类型,与在不存在抑制剂的情况下阳性扩增对照进行比较的表格。

[0108] 图11是总结了模拟性侵犯后9小时和22小时获取的两个DNA样本中检测到的基因座的数量和类型的表格。DNA是从差异提取方法的精子级分中分离出来的,并且输入量为500pg DNA。

[0109] 图12显示了通过苯酚-氯仿-异戊醇(PCIA)提取方法,在唾液样本中随着苯酚(已知的PCR扩增抑制剂)的含量增加而检测到的基因座的数量。

[0110] 图13显示了在法医实验室中通常执行的不同基质或方法分离的血液样本中检测到的基因座的数量,包括带铁锈的血液、牛仔布中的血液、拭子上的血液和从Chelex™提取中残留不同水平的血红素(已知的PCR扩增抑制剂)的血液。

[0111] 图14描绘了评估感兴趣个体——例如,失踪人员或者冲突或灾难的受害者——的亲属关系的方法的示例性示意图,该方法包括使用上传到本地服务器的DNA谱——例如SNP报告——来分析亲属关系,该本地服务器包括来自亲属的至少一个DNA谱。

[0112] 图15显示了在一组真实的死后样本中针对个体样本检测到的SNP总数。

[0113] 图16显示了在示例性相关模拟死后样本集内针对个体样本检测到的SNP总数,该样本集包括通过煮沸人工退化的样本或具有ODI且输入DNA少于1ng的低输入样本。

[0114] 图17显示了在示例性相关生前私人家族样本集内针对个体样本检测到的SNP总数,该样本集包括来自包括经Coriell验证的多达二级关系程度的CEPH/Utah的样本,以及三个无关样本。

[0115] 图18A-E描绘了关于一级、二级和三级亲属的2,000个SNP(图18A)、4,000个SNP(图18B)、6,000个SNP(图18C)、8,000个SNP(图18D)和10,000个SNP(图18E)的结果的受试者工

作特征(ROC)曲线。

[0116] 图19A-E描绘了关于四级或五级亲属的2,000个SNP(图19A)、4,000个SNP(图19B)、6,000个SNP(图19C)、8,000个SNP(图19D)和10,000个SNP(图19E)的结果的ROC曲线。

[0117] 图20A描绘了在16重和30重被测序的库的下采样数据集中分型的SNP数量分布。图20B和20C描绘了:在用于模拟生前(AM)样本的3(3plex)、12(12plex)、16(16plex),24(24plex)和32(32plex)样本运行中测序的序列库样本中键入的SNP数量,该生前(AM)样本的库是用1ng完整DNA生成的(图20B),以及在用于模拟死后(PM)样本的3(3plex)和12(12plex)样本运行中测序的序列库样本中键入的SNP数量,该死后(PM)样本的库由火化、埋葬和焚烧的骨骼、牙科遗骸、全血退化的DNA样本以及50、100、250和500pg输入DNA的低输入DNA样本生成(图20C)。

[0118] 图21A和21B描绘了来自同时期牙齿、血液、埋藏骨骼、同时期骨骼或低输入DNA的模拟生前(模拟AM)样本(图21A)和模拟死后(模拟PM)(图21B)样本的等位基因一致性和杂合性。

[0119] 图22A-D描绘了匿名GEDMatch样本的整个亲属关系系数的灵敏度和特异性的图形表示,针对一级、二级、三级、四级和五级关系,其中分型的SNP数量为2,000(图22A)、4,000(图22B)、6,000(图23C)或8,000(图22D)。

[0120] 图23描绘了Utah/CEPH 1463家族的谱系,该家族由祖父母、父母和兄弟姐妹组成,从而代表一级和二级关系程度。在12、16、24和32个样本合并库中对样本进行了测序。

[0121] 图24描绘了在以每次运行四种不同数量的样本进行测序时Utah/CEPH 1463家族中样本间SNP分型数量的分布:每次运行12个样本(12重)、每次运行16个样本(16重)、每次运行24个样本(24重)和每次运行32个样本(32重)。

[0122] 图25A和25B描绘了针对从Utah/CEPH 1463家族中获取的所有配对组合以及从1000基因组计划中随机选择的100个样本(其表示无关对照组)的亲属关系系数(图25A)和以10为底的对数似然比(LogLR)(图25B)的分布。这些样本包括祖父母(G)、父母(P)、兄弟姐妹(S)、无关对照组(U)、无关祖父母(GU)、无关父母(PU)和无关兄弟姐妹(SU)的样本。

[0123] 图26描绘了由父母、姑/姨、堂/表亲(Cousin)、隔一代的堂/表亲(1C1R)和第二代的堂/表亲组成的私人相关家族(RF)的谱系。

[0124] 图27A描绘了使用以12重退化的/低输入DNA样本和以30重完整样本,针对私人相关家族(RF)进行分型的SNP的数量分布,这些SNP数量引发了一级、二级、三级、四级和五级关系程度。图27B描绘了来自私人相关家族(RF)的个体对的亲属关系系数,其中在每条柱的顶部指示相应的对数似然比。

具体实施方式

[0125] 除非另有说明,本文所述技术的实施可以采用分子生物学、细胞生物学、生物化学和测序技术的常规技术和描述,这些技术和描述属于本领域技术人员的技能范围。能够参考本文的示例来具体说明合适的技术。

[0126] 本申请中提及的所有出版物(其包括专利文献、科学论文和数据库)均以引用的方式全文并入本申请中,用于所有目的,其程度与每篇单独的出版物以引用的方式单独并入本申请中相同。如果本申请中所述的定义与本申请中以引用的方式并入的专利、申请、已公

布的申请和其他出版物中所述的定义相反或不一致,则本申请中所述的定义优先于以引用的方式并入本申请中的定义。

[0127] 本文使用的章节标题仅用于组织目的,不应被视为限制所描述的主题。

[0128] 概述

[0129] 失踪人员或灾难或冲突受害者的样本可能高度退化,并且可能不适合全基因组测序(WGS)、微阵列或短串联重复序列(STR)分析。线粒体分析具有高灵敏度并且可能适用于某些情况,但仅考虑母系遗传,因此在亲属关系分析中使用存在缺点。此外,当前生成用于基因数据库中比较的DNA谱的方法包括使用密集SNP微阵列和WGS进行基因分型,然后将证据样本与数据库中的远亲关联,这需要大量和高质量的DNA样本,并且不适用于家族搜索或用于识别失踪人员或灾难或冲突的受害者。这些样本可能是低数量和低质量的样本,例如,包括退化的DNA,并且当前方法的数据需要大量归纳才能生成能够上传到搜索数据库的结果。最后,失踪人员或灾难或冲突受害者的亲属通常不希望将其基因数据上传到公共数据库。本文提供的新方法和改进方法克服了这些限制,允许使用低数量和低质量——例如退化的——DNA来生成核酸谱,从而比WGS或SNP微阵列等替代方法更有效地进行遗传分析,并且无需将遗传数据上传到可公开访问的基因数据库中。此外,本文提供的新方法和改进方法还包括改进的执行亲属关系分析的方法,该方法需要更少的计算来计算准确的亲属关系。

[0130] 出于人道主义和法律原因,需要对死亡事件的受害者进行身份识别:为寻找失踪家庭成员的家庭提供解决方案,并且在民事和刑事案件中,当死亡不是意外时,需要为死者家属伸张正义。由于遇难者人数众多,而且灾难对遇难者尸体完整性的影响很大,因此进行大规模死亡事件(MFI)的遇难者身份识别可能具有挑战性。MFI是由诸如疾病/饥荒、地震/海啸/飓风、飞机/火车/汽车相撞或火灾等事故或诸如战争、恐怖袭击或侵犯人权/种族灭绝等人为行为造成的。最近发生的事件,诸如2001年9月11日纽约世界贸易中心遭受的恐怖袭击或2004年12月26日苏门答腊北部西海岸地震引起的圣诞节海啸,都造成了难以想象的人员伤亡,也表明需要有效的程序和方法来发掘和编目遇难者遗体,存储与遗体相关的信息,并且进行身份识别。

[0131] 最常见的身份识别方法(称为灾难受害者身份识别或DVI)是指纹识别、牙齿对比(牙科检查或放射学)、个人物品明细、尸检以寻找手术疤痕/程序和纹身的证据、以及DNA分析。诸如指纹识别和牙齿对比等传统方法通常被用作第一步行动,因为它们所需的劳动力和成本低,而且程序速度快,但这些方法需要生前指纹和牙齿图像记录才能进行身份识别。个人物品可能会产生误导,因为受害者可能有类似的珠宝或其他个人物品项。手术程序和纹身的对比也需要生前医疗记录或纹身或其他改动的记录。最后,这些方法要求受害者的遗体相对完整。一些MFI可能导致遗体碎裂以及混合。在碎裂的情况下,对死后(PM)样本进行DNA分析不仅可以帮助识别失踪人员,还可以帮助将多个遗体或身体部位分配给特定个体。DNA分析需要失踪人员的生前(AM)样本,诸如剃须刀、刮胡刀、牙刷或发刷以供比较。当无法获得失踪人员的生前(AM)样本时,近亲家庭成员捐赠的样本将有助于身份识别。DNA分析比传统方法更耗时,并且对实验室清洁度和待分析样本的保管链跟踪有特定要求,这在现场情况下可能具有挑战性。

[0132] 如果传统方法无法成功或遗骸不完整,DNA分析的成功将取决于样本的采集、采集

时间、储存条件和所获样本的数量(与腐烂和分解有关)。DNA身份识别依赖于法医基因组学中使用的非编码DNA标记,其包括短串联重复序列(STR),诸如组合DNA索引系统或CODIS中包含的20个常染色体核心基因座、母系的线粒体DNA或父系的Y染色体上的STR(Y-STR)。由于大规模灾难中可能有多个具有相同母系或父系的家庭成员失踪的复杂性,因此常染色体标记是首选。

[0133] 多年来,STR分析已成功用于刑事、失踪人员和亲子关系案件中的人类身份识别。这种类型的分析的成功得益于这些标记的高度多态性以及能够同时用于一次分析的标记数量。这些标记也已成功用于DVI。在MFI的情况下,软件解决方案有助于协助对PM(受害者)和AM(自身或亲属)样本的大量成对比较以及关联度(degree of relatedness)的统计计算。有几种软件包可用于分析常染色体和Y-STR以及线粒体DNA数据。并非所有MFI样本都适合进行STR分析。高度退化的DNA样本不会扩增常用的毛细管电泳(CE)试剂盒中的较大STR标记,从而限制了进行身份识别所分型的标记数量。利用较小扩增子的CE试剂盒的开发有助于退化的DNA的CE分析(Butler论文)。下一代测序(NGS或大规模并行测序)STR检测允许更小的扩增子尺寸以及在一次检测中分析更多的标记,这两者都可以改进从退化的DNA样本中恢复信息的能力。

[0134] 利用STR数据尤其适用于有失踪人员或非常近的家庭成员(诸如一级亲属(父母、子女或兄弟姐妹))的AM DNA样本的情况。如果只有较远的家庭成员可用于比较,STR分析的成功率会较低,因为无关个体可能会出现大量假阳性身份识别。在MFI发生在多年前且失踪人员的非常近的家庭成员已去世的情况下,情况尤其如此。在有较远阶的亲属(二级和三级亲属,诸如侄女、侄子、孙子或曾孙)的DNA可用于比较的情况下,利用大量标记(诸如单核苷酸多态性(SNP))能够帮助身份识别。

[0135] 本文所述方法的开发是为了查询10,230个法医相关SNP以用于解决悬案,包括失踪人员身份识别等目的。以前的方法涉及利用每次测序运行多达3个样本的NGS来分析扩增的DNA。这种方法最初设计用于在搜索DNA微阵列数据库——诸如GEDmatch PRO——时对足够的SNP进行分型,以检测多达五度的关系,从而协助执法部门解决悬案。为了促进DVI,需要具有成本效益的更高通量解决方案,开发了一种新的改进方法,其中能够对扩增的DNA进行更高复杂度的测序,以分型具有显著重叠的足够SNP以确定三级关系。如示例13所示,模拟PM样本以每次运行12个样本的多重性进行测序,并且模拟AM样本以每次运行32个样本的多重性进行测序,生成了足够的分型基因座数据,可以使用本文所述方法和亲属关系算法以识别三级关系,且没有假阳性身份识别。

[0136] 该亲属关系算法的靶有三方面:在没有家谱或关系信息的情况下来识别关系,这在试图将多具遗骸分配给单一个体或在无法找到或不了解受害者亲属时非常重要;降低识别关系的假阳性率,当使用STR来计算似然比时,该假阳性率通常会升高(Alonso等人,Croat Med J2005,46,540-548,其内容在此通过引用整体并入本文);以及,维护MFI的受害者的隐私,这需要了解诸如R之类的编程语言、如Familias的似然比软件的本地构建(Kling等人,Forensic Sci Int Genet2014,13,121-127,doi:10.1016/j.fsigen.2014.07.004,其内容在此通过引用整体并入本文),或Bonaparte上的私人账户(Slooten等人,Forensic Science International:Genetics 2011,5,308-315,doi:10.1016/j.fsigen.2010.06.005,其内容在此通过引用整体并入本文)(只能识别多达二级关系)。由

于MFI的受害者和受害者家属经常要求在遗体识别过程中保密,因此需要在私人服务器上维护基因型信息。因此,在一些实施例中,本文所述的亲属关系算法位于私人服务器上以识别按本文所述制备的样本之间的关系。本地软件不会将结果上传到执法数据库;而是在私人服务器上维护结果以供审查。此外,在一些实施例中,针对以12的复杂度测序的退化/低输入模拟PM样本和以32的复杂度测序的模拟参考或AM样本,本文所述的亲属关系算法能够有效地识别多达和包括三级的关系,且具有完美的灵敏度和特异性。

[0137] 因此,本文公开了识别感兴趣人员的方法,即通过使用来自感兴趣人员——例如,失踪人员或者灾难或冲突的受害者——的DNA谱执行基于DNA的亲属关系分析,并且确定DNA谱与包括感兴趣人员的已知亲属的一个或多个参考DNA谱之间的关系程度,从而识别感兴趣人员。

[0138] 本文提供了一种执行基于DNA的亲属关系分析的方法,包括:提供来自感兴趣人员的核酸样本,使用与多个靶序列特异性杂交的多个引物来扩增核酸样本,所述多个靶序列共同包括多个至少在2,000至50,000个之间或约2,000至50,000个之间的单核苷酸多态性(SNP),从而生成扩增产物,其中,扩增在一个或多个多重PCR反应中进行,从扩增产物生成核酸库,对从扩增产物生成的核酸库进行测序,分析扩增产物的序列,确定多个SNP的基因型,从而生成DNA谱,以及计算DNA谱与一个或多个参考DNA谱的关系程度,其中,一个或多个参考DNA谱包含在参考DNA谱集内,所述参考DNA谱集包括来自感兴趣人员的亲属的一个或多个参考DNA谱。

[0139] 本文还提供了一种执行基于DNA的亲属关系分析的方法,包括:提供来自感兴趣人员的核酸样本,使用与多个靶序列特异性杂交的多个引物来扩增核酸样本,所述多个靶序列共同包括多个至少在2,000至50,000个之间或约2,000至50,000个之间的单核苷酸多态性(SNP),从而生成扩增产物,其中,扩增在一个或多个多重PCR反应中进行,从扩增产物生成核酸库,对从扩增产物生成的核酸库进行测序,确定多个SNP的基因型,从而生成DNA谱,以及计算DNA谱与一个或多个参考DNA谱的关系程度,其中,一个或多个参考DNA谱包含在参考DNA谱集内,所述参考DNA谱集包括来自感兴趣人员的亲属的一个或多个参考DNA谱。

[0140] 本文还提供了一种构建感兴趣人员员的核酸库的方法,包括:提供来自感兴趣人员的核酸样本,使用与多个靶序列特异性杂交的多个引物来扩增核酸样本,所述多个靶序列共同包括多个至少在2,000至50,000个之间或约2,000至50,000个之间的单核苷酸多态性(SNP),从而生成包括扩增产物的核酸库,其中,扩增在一个或多个多重PCR反应中进行。

[0141] 本文还提供了一种构建参考DNA样本的核酸库的方法,包括:提供来自感兴趣人员的亲属的核酸样本,使用与多个靶序列特异性杂交的多个引物来扩增所述核酸样本,所述多个靶序列共同包括多个至少在2,000至50,000个之间或约2,000至50,000个之间的单核苷酸多态性(SNP),从而生成包括扩增产物的核酸库,其中,扩增在一个或多个多重PCR反应中进行。在一些实施例中,亲属是感兴趣人员的一级、二级、三级、四级或五级亲属。在一些实施例中,亲属是感兴趣人员的一级、二级或三级亲属。

[0142] 本文还提供了一种计算关联度的方法,包括:获得包括至少在2,000至50,000个SNP之间或在约2,000至50,000个SNP之间的基因型的DNA谱,其中,所述DNA谱来自感兴趣人员;以及计算DNA谱与一个或多个参考DNA谱的关系程度,其中,一个或多个参考DNA谱包括在参考DNA谱集内,所述参考DNA谱集包括来自感兴趣人员的亲属的一个或多个参考DNA谱。

[0143] 本文还提供了一种计算关联度的方法,包括:生成包括至少在2,000至50,000个SNP之间或在约2,000至50,000个SNP之间的基因型的DNA谱,其中,所述DNA谱来自感兴趣人员;以及计算DNA谱与一个或多个参考DNA谱的关系程度,其中,一个或多个参考DNA谱包括在参考DNA谱集内,所述参考DNA谱集包括来自感兴趣人员的亲属的一个或多个参考DNA谱。

[0144] 本文还提供了使用本文所述的任何方法构建的核酸库,例如,如本文所述的任何构建核酸库的方法。

[0145] 本文还提供了多个引物,所述多个引物与多个靶序列特异性杂交,所述多个靶序列包括来自感兴趣人员的核酸样本中的至少在2,000至50,000个之间的单核苷酸多态性(SNP)或约2,000至50,000个之间的所述SNP,其中,在一个或多个多重PCR反应中使用所述多个引物来扩增所述核酸样本以产生扩增产物。

[0146] 本文还提供了多个引物,所述多个引物与多个靶序列特异性杂交,所述多个靶序列包括来自感兴趣人员的核酸样本和来自一个或多个参考样本的核酸样本中的至少在2,000至50,000个之间的单核苷酸多态性(SNP)或约2,000至50,000个之间的所述SNP,其中,一个或多个参考样本包括来自所述感兴趣人员的亲属的样本,并且其中,在一个或多个多重PCR反应中使用多个引物来扩增来自感兴趣人员的核酸样本和来自一个或多个参考样本的核酸样本以产生扩增产物。

[0147] 本文还提供了一种构建DNA谱的方法,包括:提供来自感兴趣人员的核酸样本,使用与多个靶序列特异性杂交的多个引物来扩增所述核酸样本,所述多个靶序列共同包括多个至少在2,000至50,000个之间或约2,000至50,000个之间的单核苷酸多态性(SNP),从而生成扩增产物,其中,扩增在一个或多个多重PCR反应中进行,对所述扩增产物进行测序,确定多个SNP的基因型,从而生成DNA谱。

[0148] 本文还提供了一种构建DNA谱的方法,包括:提供来自感兴趣人员的核酸样本,提供来自感兴趣人员的亲属的核酸样本,使用与多个靶序列特异性杂交的多个引物来扩增来自感兴趣人员的核酸样本和来自亲属的核酸样本,所述多个靶序列共同包括多个至少在2,000至50,000个之间或约2,000至50,000个之间的单核苷酸多态性(SNP),从而生成扩增产物,其中,扩增在一个或多个多重PCR反应中进行,对所述扩增产物进行测序,确定多个SNP的基因型,从而生成感兴趣人员和感兴趣人员的亲属的DNA谱。

[0149] 本文还提供了使用本文所述的任何方法构建的DNA谱,例如,使用本文所述的任何构建DNA谱的方法。

[0150] 本文还提供了一种识别DNA谱的遗传亲属的方法,包括:计算本文所述的任何DNA谱与一个或多个参考DNA谱的关系程度,其中,一个或多个参考DNA谱包括在参考DNA谱集内,所述参考DNA谱集包括来自感兴趣人员的亲属的一个或多个参考DNA谱;以及生成家谱,所述家谱包括与所述一个或多个参考DNA谱相关的DNA谱。

[0151] 本文还提供了一种识别DNA谱身份的方法,包括:计算包括至少在2,000至50,000个SNP之间或在约2,000至50,000个SNP之间的基因型的DNA谱与一个或多个参考DNA谱的关系程度,其中,DNA谱来自感兴趣人员,并且其中,一个或多个参考DNA谱包括在参考DNA谱集内,所述参考DNA谱集包括来自感兴趣人员的亲属的一个或多个参考DNA谱;并且生成家谱,所述家谱包括与所述一个或多个参考DNA谱相关的所述DNA谱。

[0152] 本文还提供了包括至少一个容器装置的试剂盒,其中至少一个容器装置包括本文

描述的多个引物中的任一个。

[0153] 在任何一些这样的实施例中,使用亲属关系模型来计算亲属关系程度。在任何一些这样的实施例中,使用通过利用PCA方法训练的亲属关系模型来计算亲属关系程度。在任何一些这样的实施例中,用于训练亲属关系模型的PCA方法是PCA或涉及PCA。

[0154] 在任何一些这样的实施例中,PCA方法是PC-AiR。在一些实施例中,PC-AiR包括以下步骤:(1)估计训练数据库的所有样本对——可选地训练DNA谱——之间的亲属关系系数,其中,亲属关系系数>0.025的配对被识别为相关的,而亲属关系系数<-0.025的配对被识别为祖先分化的;(2)初始化包括所有样本的无关样本集;以及(3)迭代地:(i)识别所述无关样本集中具有所述无关样本集中最相关样本的集合,从而指定为X,(ii)识别X中与所述无关样本集中的样本相比具有最少祖先分化配对的样本集,从而指定为Y;以及(iii)如果Y具有零个样本,则终止过程,或者,如果Y具有至少一个样本,则从Y中随机选择一个样本以从U中移除,并且在步骤(3)(i)处开始重复。

[0155] 在任何一些这样的实施例中,PCA方法是改进的PC-Air。在一些实施例中,改进的PC-AiR包括以下步骤:(1)估计训练数据库的所有样本对——可选地训练DNA谱——之间的亲属关系系数,其中,亲属关系系数>0.01的配对被识别为相关的,而亲属关系系数<-0.025的配对被识别为祖先分化的;(2)移除所有具有≥5%失踪数据的DNA谱;以及(3)通过用排序值识别每个DNA谱来对所有DNA谱进行排序。在一些实施例中,排序值是基于完整数据库中从最少到最多排序的相关DNA谱的数量来确定的,并且通过所述完整数据库中从最多到最少排序的祖先分化DNA谱的数量来打破纽带(tie)。在一些实施例中,步骤(3)涉及迭代地遍历排序后的DNA谱,并且对于每个DNA谱:(i)如果DNA谱尚未出现在相关样本集中,则将其添加到无关样本集并且将所有相关DNA谱添加到相关样本集,以及(ii)如果DNA谱已经在相关样本集中,则跳至下一个DNA谱,并且在步骤(3)(i)处开始重复。

[0156] 在任何一些这样的实施例中,计算关系程度包括使用PC-Relate来计算亲属关系系数。在一些实施例中,通过提供感兴趣人员的DNA谱作为PC-Relate的输入来计算关系程度。在任何一些这样的实施例中,通过提供亲属关系模型和感兴趣人员的DNA谱作为PC-Relate的输入来计算关系程度。在任何一些这样的实施例中,还提供一个或多个参考DNA谱作为PC-Relate的输入。在任何一些这样的实施例中,计算关系程度包括计算似然比。在一些实施例中,计算似然比包括比较所述DNA谱和所述一个或多个参考DNA谱之间的多个SNP。在一些实施例中,计算似然比包括比较DNA谱和一个或多个参考DNA谱之间的SNP集合,所述SNP集合包括来自多个SNP内的亲属关系SNP。在一些实施例中,感兴趣人员在生物学上是男性,并且所述方法还包括计算DNA谱与一个或多个参考DNA谱之间共享Y染色体的似然比。在一些实施例中,计算共享Y染色体的似然比包括比较DNA谱和一个或多个参考DNA谱之间的SNP集合,所述SNP集合包括一个或多个Y-SNP。在一些实施例中,一个或多个Y-SNP包含在多个SNP内。在一些实施例中,一个或多个Y-SNP包括至少20、25、30、35、40、45、50、55、60、65、70、75、80、81、82、83、84或85个Y-SNP。在一些实施例中,一个或多个Y-SNP包括85个Y-SNP。在一些实施例中,计算共享Y染色体的似然比包括基于一个或多个Y-SNP的基因型,将DNA谱与一个或多个参考DNA谱中的参考DNA谱共享Y染色体的概率除以DNA谱与参考DNA谱不共享Y染色体的概率。

[0157] 在一些实施例中,计算似然比包括基于多个SNP的基因型,将DNA谱与一个或多个

参考DNA谱中的参考DNA谱相关的概率除以DNA谱与参考DNA谱不相关的概率。

[0158] 在任何一些这样的实施例中,计算共享染色体Y的似然率包括通过提供感兴趣人员的DNA谱和匹配谱作为输入来计算对数似然率以识别匹配染色体。

[0159] 样本和样本处理

[0160] 在一些方面,本文公开的样本能够是或包括任何合适的生物样本,或由其衍生的样本。在一些方面,本文描述的样本使用任何已知的合适方法处理和扩增以补充本文描述的方法。下文描述了示例性样本、样本处理方法和样本扩增方法。

[0161] A.核酸样本

[0162] 本文公开的核酸样本能够源自任何生物样本,例如,来自感兴趣人员的任何生物样本。生物样本可以源自血液、口腔拭子、毛发、牙齿、骨骼、皮肤、组织和/或精液,或用于获得感兴趣人员的DNA的任何其他来源。在一些实施例中,核酸样本源自生物样本,该生物样本是或包括血液、毛发、牙齿、骨骼、精液、皮肤或精子。在一些实施例中,核酸样本源自组织样本。在一些实施例中,生物样本是DNA样本。在一些实施例中,核酸样本包括DNA。在一些实施例中,DNA是基因组DNA(gDNA)。在一些实施例中,来自感兴趣人员的核酸样本包括基因组DNA和/或来自参考DNA样本——例如感兴趣人员的亲属——的核酸样本包括,基因组DNA。在一些实施例中,感兴趣人员的核酸样本和/或感兴趣人员的亲属的核酸样本包括基因组DNA。可从中获得核酸样本的DNA可能是完整的或部分退化的。可从中获得核酸样本的DNA可能由于(但不限于)源材料年龄、可变提取、存储程序或环境暴露而受损、退化或抑制。在一些实施例中,DNA因钙抑制、火化、焚烧和防腐而受损。在一些实施例中,本文所述方法包括提供来自感兴趣人员的核酸样本。

[0163] 在一些实施例中,从中获得核酸样本的DNA是低数量和/或低质量的DNA样本。在一些实施例中,从中获得核酸样本的DNA是低数量和低质量的DNA样本。在一些实施例中,低质量DNA样本包括低质量的核酸分子。在一些实施例中,低质量核酸分子是退化的DNA,例如,基因组DNA,和/或片段化的DNA,例如,基因组DNA。

[0164] 能够通过计算退化指数(DI)来确定核酸(例如,DNA)样本的质量。通过将小DNA靶标的浓度除以大DNA靶标的浓度来计算DI(DI=小DNA靶标的浓度/大DNA靶标的浓度)。一般而言,小于1的DI值通常指示核酸(例如,DNA)未退化、不是低质量样本和/或为高质量样本;1至10的DI值通常指示核酸(例如,DNA)具有轻微至中等的退化度;大于10的DI值通常表示核酸(例如,DNA)高度退化。

[0165] 在一些实施例中,低质量核酸分子具有为1、2、3、4、5、6、7、8、9、10、15、20、25、30、35、40、45、50、55、60、65、70、75、80、85、90、95、100、105、110、115、120、125、130、135、140、145、150、155、160、165、170、175、180、185、190、195或200或更高的DI,或者至少为1、2、3、4、5、6、7、8、9、10、15、20、25、30、35、40、45、50、55、60、65、70、75、80、85、90、95、100、105、110、115、120、125、130、135、140、145、150、155、160、165、170、175、180、185、190、195或200或更高的DI。在一些实施例中,低质量核酸分子具有至少为1的DI,并且等于或小于2、3、4、5、6、7、8、9、10、15、20、25、30、35、40、45、50、55、60、65、70、75、80、85、90、95、100、105、110、115、120、125、130、135、140、145、150、155、160、165、170、175、180、185、190、195或200或更高。在一些实施例中,低质量核酸分子具有至少为2的DI,并且等于或小于3、4、5、6、7、8、9、10、15、20、25、30、35、40、45、50、55、60、65、70、75、80、85、90、95、100、105、110、115、120、125、130、

135、140、145、150、155、160、165、170、175、180、185、190、195或200或更高。在一些实施例中,低质量核酸分子的DI为2或至少为2或更高。在一些实施例中,低质量核酸分子的DI为5或至少为5或更高。在一些实施例中,低质量核酸分子的DI为10或至少为10或更高。在一些实施例中,低质量核酸分子的DI为20或至少为20或更高。在一些实施例中,低质量核酸分子的DI为1至200之间。在一些实施例中,低质量核酸分子的DI为1至175之间。在一些实施例中,低质量核酸分子的DI至少为1且为158.3或更低。在一些实施例中,低质量核酸分子的DI为2至200之间。在一些实施例中,低质量核酸分子的DI为2至175之间。在一些实施例中,低质量核酸分子的DI至少为2且等于或小于158.3。在一些实施例中,低质量核酸分子的DI为5至200之间。在一些实施例中,低质量核酸分子的DI为5至175之间。在一些实施例中,低质量核酸分子的DI至少为5且等于或小于158.3。在一些实施例中,低质量核酸分子的DI为10至200之间。在一些实施例中,低质量核酸分子的DI为10至175之间。在一些实施例中,低质量核酸分子的DI至少为10且等于或小于158.3。

[0166] 在一些实施例中,低质量核酸分子具有的DI为1至10之间或约1至10之间、1至50之间或约1至50之间、1至100之间或约1至100之间、1至200之间或约1至200之间、2至10之间或约2至10之间、2至50之间或约2至50之间、2至100之间或约2至100之间、200之间或约200之间、5至10之间或约5至10之间、5至50之间或约5至50之间、5至100之间或约5至100之间、5至200之间或约5至200之间。

[0167] 在一些实施例中,获取核酸样本的DNA是高质量核酸样本。在一些实施例中,高质量核酸样本的DI小于1。

[0168] 在一些实施例中,核酸样本包含一种或多种酶抑制剂。在一些实施例中,一种或多种酶抑制剂包含由血红素、腐殖酸(例如,亚铁血红素)、腐殖酸、靛蓝、单宁酸、胶原蛋白、钙和羟基磷灰石组成的组中的一种或多种抑制剂。在一些实施例中,一种或多种酶抑制剂包含亚铁血红素。

[0169] 在一些实施例中,核酸样本来自感兴趣人员,例如,失踪人员或灾难或冲突的受害者。在一些实施例中,感兴趣人员是失踪人员。失踪人员可能因任何原因失踪,可能是自愿或非自愿失踪。例如,在一些实施例中,失踪人员是非自愿失踪,并且已经被劫持或绑架。在一些实施例中,失踪人员是自愿失踪,并且已经逃跑、正在逃避侦查或以其他方式躲藏。

[0170] 在一些实施例中,核酸样本来自参考DNA样本。在一些实施例中,参考DNA样本来自感兴趣人员的亲属。因此,在一些实施例中,核酸样本来自感兴趣人员的亲属,诸如感兴趣人员的一级、二级、三级、四级或五级亲属。在一些实施例中,一个或多个参考DNA谱中的一个或多个来自参考DNA样本,例如,来自感兴趣人员的亲属的参考DNA样本。

[0171] 在一些实施例中,感兴趣人员是灾难或冲突的受害者。灾难或冲突的受害者可以是任何类型的灾难或冲突的受害者。例如,在一些实施例中,灾难或冲突的受害者是灾难的受害者,诸如飓风、龙卷风、风暴、火灾——包括野火/森林火灾、海啸、地震、洪水、火山爆发、雪崩等。在一些实施例中,灾难是自然灾害。如本文所用,“自然灾害”是指由地球自然过程——诸如与天气和/或地质事件相关的过程——引起的任何灾难,例如,飓风、洪水、风暴、海啸、地震、火山爆发等。在一些实施例中,灾难是非自然灾害。本文所用的“非自然灾害”是指除自然灾害以外的任何灾害,包括人为因素造成的灾害,包括涉及汽车、飞机、轮船和火车的灾害,涉及建筑物、道路、矿井和桥梁倒塌的灾害,涉及建筑物燃烧的灾害,以及其

他人为因素造成的灾害。在一些实施例中,灾害或冲突的受害者是冲突的受害者,诸如战争或人群之间的其他冲突。本文所用的“冲突”是指不同国家或州之间或国家或州内不同群体之间的任何冲突,例如武装冲突,例如战争或恐怖袭击,或导致人员死亡和/或受伤的群体之间的任何其他冲突。

[0172] 在一些实施例中,感兴趣人员物在生物学上是女性。在一些实施例中,感兴趣人员物在生物学上是男性。

[0173] 在一些实施例中,核酸样本源自口腔拭子、纸、织物——例如牛仔布或其他基质或浸渍有唾液、血液、精子或其他体液或含有毛发或皮肤细胞的物体。在一些实施例中,浸渍有唾液、血液、精子或其他体液或含有毛发或皮肤细胞的物体是个人物品,例如牙刷或发刷。在一些实施例中,核酸样本源自包含毛发或皮肤细胞的物体,例如发刷或牙刷。在一些实施例中,核酸样本源自个人物品,例如牙刷或发刷。在一些实施例中,核酸样本源自牙刷或发刷。在一些实施例中,个人物品是被核酸样本来源人使用的和/或与其关联的物品,使得该人的核酸存在于物品上或物品之中。

[0174] 在一些实施例中,核酸样本来自犯罪现场,诸如凶杀案、袭击案——诸如性侵犯案、入室盗窃案或任何其他需要参与者的身份识别的犯罪案。在一些实施例中,核酸样本来自性侵犯案。

[0175] 在一些实施例中,在包含核酸样本的样本由其来源(例如,人类受试者)存放后30或约30分钟、1小时或约1小时、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23或24或更长小时或约2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23或24或更长小时来获得核酸样本。在一些实施例中,在包括核酸样本的样本由其来源(例如,人类受试者)存放后3小时、9小时、12小时、15小时、18小时、21小时、22小时、24小时、36小时、48小时、3天、4天、5天、6天、7天、2周、3周、4周、1个月、2个月、3个月、4个月、5个月、6个月、7个月、8个月、9个月、10个月、11个月、1年、2年、3年或4年或更长时间,或者约3小时、9小时、12小时、15小时、18小时、21小时、22小时、24小时、36小时、48小时、3天、4天、5天、6天、7天、2周、3周、4周、1个月、2个月、3个月、4个月、5个月、6个月、7个月、8个月、9个月、10个月、11个月、1年、2年、3年或4年或更长时间来获得核酸样本。在一些实施例中,在包含核酸样本的样本由其来源(例如,人类受试者)存放后24小时或小于24小时(例如,22小时或小于22小时)来获得核酸样本。

[0176] 在一些实施例中,核酸样本包括在3pg至100ng之间的DNA或在约3pg至100ng之间的DNA,例如基因组DNA,或在50pg至100ng之间的DNA或在约50pg至100ng之间的DNA,例如基因组DNA。在一些实施例中,核酸样本包括在100pg至5ng之间的DNA或在约100pg至5ng之间的DNA,例如基因组DNA。在一些实施例中,核酸样本包括1ng的DNA或约1ng的DNA,例如基因组DNA。在一些实施例中,核酸样本包括在3pg至100ng之间的DNA或约3pg至100ng之间的DNA,例如基因组DNA。

[0177] 在一些实施例中,核酸样本包括3pg或约3pg至100ng或约100ng的DNA,例如基因组DNA。在一些实施例中,核酸样本包括10pg或约10pg至100pg或约100ng的DNA,例如基因组DNA,或包括10pg或约10pg至5ng或约5ng的DNA,例如基因组DNA。在一些实施例中,核酸样本包括10pg或约10pg至10ng、10pg或约10pg至5ng,25pg或约25pg至10ng、25pg或约25pg至5ng、50pg或约50pg至10ng、或50pg或约50pg至5ng的DNA,例如基因组DNA。

[0178] 在一些实施例中,核酸样本包括3pg或约3pg至5ng或约5ng的DNA,例如基因组DNA。在一些实施例中,核酸样本包括50pg或约50pg至5ng或约5ng的DNA,例如基因组DNA。在一些实施例中,核酸样本包括为或约2.5ng、3pg、4pg、5pg、6pg、7pg、8pg、9pg、10pg、15pg、20pg、25pg、30pg、35pg、40pg、45pg、50pg、55pg、60pg、70pg、75pg、80pg、85pg、90pg、95pg、100pg、125pg、150pg、175pg、200pg、225pg、250pg、275pg、300pg、325pg、350pg、375pg、400pg、425pg、450pg、475pg、500pg、600pg、700pg、800pg、900pg、1ng、1.1ng、1.2ng、1.3ng、1.4ng、1.5ng、1.6ng、1.7ng、1.8ng、1.9ng、2ng、2.1ng、2.2ng、2.3ng、2.4ng、2.5ng、2.6ng、2.7ng、2.8ng、2.9ng、3ng、3.25ng、3.5ng、3.75ng、4ng、4.25ng、4.5ng、4.75ng或5ng的DNA,例如基因组DNA,或在任何两个前述值之间。在一些实施例中,核酸样本包括介于或介于约3pg与10ng之间、介于或介于约3pg与5ng之间、介于或介于约3pg与4ng之间、介于或介于约3pg与3ng之间、介于或介于约3pg与2ng之间、介于或介于约10pg与10ng之间、介于或介于约10pg与5ng之间、介于或介于约10pg与4ng之间、介于或介于约10pg与3ng之间、介于或介于约10pg与2ng之间、介于或介于约25pg与10ng之间、介于或介于约25pg与5ng之间、介于或介于约25pg与4ng之间、介于或介于约25pg与3ng之间、介于或介于约25pg与2ng之间、介于或介于约40pg与10ng之间、介于或介于约40pg与5ng之间、介于或介于约40pg与4ng之间、介于或介于约40pg与3ng之间、介于或介于约40pg与2ng之间、介于或介于约50pg与10ng之间、介于或介于约50pg与5ng之间、介于或介于约50pg与4ng之间、介于或介于约50pg与3ng之间、介于或介于约50pg与2ng之间、介于或介于约10pg与2ng之间、介于或介于约10pg与1.5ng之间、介于或介于约10pg与1ng之间、介于或介于约20pg至2ng之间、介于或介于约20pg与1.5ng之间、介于或介于约20pg与1ng之间、介于或介于约25pg与2ng之间、介于或介于约25pg与1.5ng之间、介于或介于约25pg与1ng之间、介于或介于约30pg与2ng之间、介于或介于约30pg与1.5ng之间、介于或介于约30pg与1ng之间、介于或介于约35pg与2ng之间、介于或介于约35pg与1.5ng之间、介于或介于约35pg与1ng之间、介于或介于约40pg与2ng之间、介于或介于约40pg与1.5ng之间、介于或介于约40pg与1ng之间、介于或介于约45pg与2ng之间、介于或介于约45pg与1.5ng之间、介于或介于约45pg与1ng之间、介于或介于约50pg与2ng之间、介于或介于约50pg与1.5ng之间、或介于或介于约50pg与1ng之间。

[0179] B.样本处理和扩增

[0180] 在一些实施例中,本文提供的方法包括使用与多个靶序列特异性杂交的多个引物来扩增核酸样本的步骤,所述多个靶序列共同包括多个至少在或在约2,000至50,000个之间的单核苷酸多态性(SNP),从而生成扩增产物,其中,所述扩增在一个或多个多重PCR反应中进行。

[0181] 能够执行多种步骤来制备或处理用于测定(assay)和/或测定期间的核酸样本。除非另有说明,下文描述的制备或处理步骤通常可以以任何方式和任何顺序组合,以适当地制备或处理本文公开的特定样本以供分析和/或测序。

[0182] 在一些实施例中,所提供的核酸样本的量为、约为或小于1ng的基因组DNA。在一些实施例中,本文公开的方法包括基因组DNA的扩增。在一些实施例中,基因组DNA的扩增包括一个或多个包括多个引物的多重聚合酶链式反应(PCR),从而生成扩增产物。在一些实施例中,基因组DNA的扩增包括单个多重PCR反应。在一些实施例中,基因组DNA的扩增包括两个多重PCR反应。在一些实施例中,基因组DNA的扩增包括三个多重PCR反应。在一些实施例中,

基因组DNA的扩增包括四个多重PCR反应。

[0183] 在一些实施例中,多个引物中的一个或多个引物是根据WO 2015/126766A1中所述的非典型设计策略设计的,其在此通过引用整体并入本文。在一些实施例中,多个引物中的一个或多个引物的长度至少为24个核苷酸,和/或熔解温度小于60摄氏度,和/或富含AT,AT含量至少为60%。在一些实施例中,多个引物中的一个或多个引物的长度至少为与靶序列杂交的24个核苷酸,和/或熔解温度在50摄氏度和60摄氏度之间,和/或富含AT,AT含量至少为60%。在一些实施例中,多个引物中的一个或多个引物的熔解温度小于58摄氏度,或小于54摄氏度。

[0184] 在一些实施例中,可以使用多个引物对基因组DNA进行多个循环扩增,所述多个引物与多个靶序列杂交和/或标记多个靶序列,所述多个靶序列共同包括多个至少在或在约2,000至50,000个之间的单核苷酸多态性(SNP),或至少在或在约5,000至50,000个之间的单核苷酸多态性(SNP)。在一些实施例中,可以使用多个引物对基因组DNA进行多个循环扩增,所述多个引物与多个靶序列杂交和/或标记多个靶序列,所述多个靶序列共同包括多个至少在或在约2,000至15,000、20,000、25,000、30,000、35,000、40,000、45,000或50,000个之间的SNP。在一些实施例中,可以使用多个引物对基因组DNA进行多个循环扩增,所述多个引物与多个靶序列杂交和/或标记多个靶序列,所述多个靶序列共同包括至少在或在约5,000至15,000、20,000、25,000、30,000、35,000、40,000、45,000或50,000个SNP之间。在一些实施例中,可以使用多个引物对基因组DNA进行多个循环扩增,所述多个引物与多个靶序列杂交和/或标记多个靶序列,所述多个靶序列共同包括至少在或在约10,000至11,000个SNP之间。在一些实施例中,可以使用多个引物对基因组DNA进行多个循环扩增,所述多个引物与多个靶序列杂交和/或标记多个靶序列,所述多个靶序列共同包括至少在或在约2,000至11,000个SNP之间、3,000至11,000个SNP之间、4,000至11,000个SNP之间、5,000至11,000个SNP之间、5,500至11,000个SNP之间、6,000至11,000个SNP之间、7,000至15,000个SNP之间、7,000至14,000个SNP之间、7,000至13,000个SNP之间、7,000至12,000个SNP之间、7,000至11,000个SNP之间、8,000至15,000个SNP之间、8,000至14,000个SNP之间、8,000至13,000个SNP之间、8,000至12,000个SNP之间、8,000至11,000个SNP之间、9,000至15,000个SNP之间、9,000至14,000个SNP之间、9,000至13,000个SNP之间、9,000至12,000个SNP之间、或9,000至11,000个SNP之间。在一些实施例中,可以使用多个引物对基因组DNA进行多个循环扩增,所述多个引物与多个靶序列杂交和/或标记多个靶序列,所述多个靶序列共同包括至少在或在约6,000至11,000个SNP之间。在一些实施例中,多个SNP包括2,639个或约2,639个SNP。在一些实施例中,可以使用多个引物对基因组DNA进行多个循环扩增,所述多个引物与多个靶序列杂交和/或标记多个靶序列,所述多个靶序列共同包括10,230个或约10,230个SNP。

[0185] 在一些实施例中,多个SNP包括至少在或在约2,000至15,000、20,000、25,000、30,000、35,000、40,000、45,000或50,000个SNP之间。在一些实施例中,多个SNP包括至少在或在约5,000至15,000、20,000、25,000、30,000、35,000、40,000、45,000或50,000个SNP之间。在一些实施例中,多个SNP包括至少在或在约6,000至15,000、20,000、25,000、30,000、35,000、40,000、45,000或50,000个SNP之间。在一些实施例中,多SNP包括至少在或在约2,000至11,000个SNP之间、3,000至11,000个SNP之间、4,000至11,000个SNP之间、5,000至11,000个SNP之间、5,500至11,000个SNP之间、6,000至11,000个SNP之间、7,000至15,000个SNP之

间、7,000至14,000个SNP之间、7,000至13,000个SNP之间、7,000至12,000个SNP之间、7,000至11,000个SNP之间、8,000至15,000个SNP之间、8,000至14,000个SNP之间、8,000至13,000个SNP之间、8,000至12,000个SNP之间、8,000至11,000个SNP之间、9,000至15,000个SNP之间、9,000至14,000个SNP之间、9,000至13,000个SNP之间、9,000至12,000个SNP之间、或9,000至11,000个SNP之间。在一些实施例中,多个SNP包括2,639个或约2,639个SNP。在一些实施例中,多个SNP包括10,230个或约10,230个SNP。在一些实施例中,多个SNP包括至少在或在约2,000至50,000个SNP之间、5,000至50,000个SNP之间、5,000至45,000个SNP之间、5,000至40,000个SNP之间、5,000至35,000个SNP之间、5,000至30,000个SNP之间、5,000至25,000个SNP之间、5,000至20,000个SNP之间、6,000至50,000个SNP之间、6,000至45,000个SNP之间、6,000至40,000个SNP之间、6,000至35,000个SNP之间、6,000至30,000SNP之间、6,000至25,000个SNP之间、6,000至20,000个SNP之间、7,000至50,000个SNP之间、7,000至45,000个SNP之间、7,000至40,000个SNP之间、7,000至35,000个SNP之间、7,000至30,000个SNP之间、7,000至25,000个SNP之间、7,000至20,000个SNP之间、8,000至50,000个SNP之间、8,000至45,000个SNP之间、8,000至40,000个SNP之间、8,000至35,000个SNP之间、8,000至30,000个SNP之间、8,000至25,000个SNP之间、8,000至20,000个SNP之间、9,000至50,000个SNP之间、9,000至45,000个SNP之间、9,000至40,000个SNP之间、9,000至35,000个SNP之间、9,000至30,000个SNP之间、9,000至25,000个SNP之间、或9,000至20,000个SNP之间。

[0186] 在一些实施例中,多个SNP包括至少在或在约2,000至11,000个SNP之间、2,500至11,000个SNP之间、3,000至11,000个SNP之间、3,500至11,000个SNP之间、4,000至11,000个SNP之间、4,500至11,000个SNP之间、5,000至11,000个SNP之间、5,550至11,000个SNP之间、6,000至11,000个SNP之间、6,500至11,000个SNP之间、7,000至11,000个SNP之间、7,500至11,000个SNP之间、8,000至11,000个SNP之间、8,500至11,000个SNP之间、9,000至11,000个SNP之间、9,500至11,000个SNP之间、或10,000至11,000个SNP之间。

[0187] 在一些实施例中,多个SNP包括选自由亲属关系SNP、祖先SNP、身份SNP、表型SNP、X-SNP和Y-SNP组成的组中的一个或多个SNP中的SNP。在一些实施例中,多个SNP包括亲属关系SNP、祖先SNP、身份SNP、表型SNP、X-SNP和Y-SNP。在一些实施例中,多个SNP包括亲属关系SNP。在一些实施例中,多个SNP包括Y-SNP。在一些实施例中,多个SNP包括亲属关系SNP和Y-SNP。

[0188] 在任何一些这样的实施例中,多个SNP包括一个或多个微单倍型(microhaplotype)。因此,在一些实施例中,微单倍型是包含在多个SNP中的SNP类型。在一些实施例中,每个微单倍型包括在单个扩增子上共享或在基因组上彼此接近的一个或多个SNP。一般而言,微单倍型是长度通常小于300个核苷酸的生物标记,其显示多个等位基因组合,例如,多个基于SNP的等位基因标记。

[0189] 在一些实施例中,SNP不包括具有已知医学关联——例如与已知医学状况相关——或低次要等位基因频率的SNP。通过排除具有已知医学关联——例如与已知医学状况相关——或低次要等位基因频率的SNP,隐私关注受到限制并且遗传健康数据得到保护。

[0190] 在一些实施例中,SNP包括已经用多个基因型样本过滤的SNP。在一些实施例中,SNP选自包括祖先SNP、身份SNP、亲属SNP、表型SNP、X-SNP和Y-SNP的类别。在一些实施例中,祖先SNP包括在或在约10-100个SNP之间。在一些实施例中,身份SNP包括在或在约10-200个

SNP之间。在一些实施例中,亲属关系SNP包括在或在约7,000-12,000个SNP之间。在一些实施例中,表型SNP包括在或在约1-50个SNP之间。在一些实施例中,X-SNP包括在或在约10-200个SNP之间。在一些实施例中,Y-SNP包括在或在约10-200个SNP之间。在一些实施例中,祖先SNP包括在SNP总数的0-10%或约0-10%之间。在一些实施例中,身份SNP包括在SNP总数的0-10%或约0-10%之间。在一些实施例中,亲属关系SNP包括在SNP总数的80-100%或约80-100%之间。在一些实施例中,多个SNP中至少或至少约80%、81%、82%、83%、84%、85%、86%、87%、88%、89%、90%、91%、92%、93%、94%、95%、96%、97%、98%或99%是亲属关系SNP。在一些实施例中,多个SNP中至少或至少约85%是亲属关系SNP。在一些实施例中,多个SNP中至少或至少约90%是亲属关系SNP。在一些实施例中,多个SNP中至少或至少约95%是亲属关系SNP。在一些实施例中,多个SNP中至少或至少约99%是亲属关系SNP。在一些实施例中,多个SNP中100%是亲属关系SNP。在一些实施例中,表型SNP包括在SNP总数的0-5%或约0-5%之间。在一些实施例中,X-SNP包括在SNP总数的0-5%或约0-5%之间。在一些实施例中,Y-SNP包括在SNP总数的0-5%或约0-5%之间。在一些实施例中,SNP不包括医学信息性或次要等位基因频率SNP。标签区能够是任意序列,诸如通用标签区、捕获标签区、扩增标签区、测序标签区、UMI标签区等。

[0191] 在一些实施例中,靶序列被纯化和富集,并且生成原始DNA样本的库,也被称为核酸库。在一些实施例中,纯化将纯化珠与酶相结合以从其他反应组分中纯化扩增的靶。在一些实施例中,通过扩增DNA并添加UDI接头和簇生成所需的序列来富集纯化的靶序列。UDI接头能够使用识别每个样本进行分析的独特的序列组合来标记DNA。

[0192] 在一些实施例中,从扩增产物——其包括通过本文所述方法或实施例产生的扩增产物——生成核酸库。因此,在一些实施例中,核酸库包括通过使用与多个靶序列特异性杂交的多个引物来扩增核酸样本而生成的扩增产物,所述靶序列共同包括多个至少在5,000至50,000个之间或约5,000至50,000个之间的SNP,或多个至少在2,000至50,000个之间或约2,000至50,000个之间的SNP。

[0193] 在一些实施例中,对核酸库或DNA库进行归一化以进行量化和质量检查,并通过组合等量的归一化库进行池化以创建能够在同一流动池上一起测序的库池。在一些实施例中,量化包括使用荧光法。在一些实施例中,量化包括定量PCR方法。池化DNA库后,能够使用基于氢氧化钠(NaOH)的方法对其进行变性和稀释,并能够添加测序对照。

[0194] 在一些实施例中,核酸库按照Forenseq Kintelligence套件用户指南(Verogen PN:V16000120,其内容在此通过引用整体并入)中的说明进行定量、归一化、变性和稀释。

[0195] 在一些实施例中,使用任何已知合适的方法通过大规模并行测序来制备DNA库的核酸库以进行测序,以补充本文所述的方法。

[0196] 在一些实施例中,本文还提供了使用本文所述任何方法构建的核酸库。

[0197] 在一些实施例中,本文提供的方法包括从扩增产物生成核酸库的步骤。

[0198] 测序与分析

[0199] 在一些方面,本文所述的核酸库或DNA库能够使用任何已知合适的方法测序以补充本文所述的方法,并且不限于任何特定的测序平台。在一些方面,本文公开的样本能够使用任何已知合适的方法分析以补充本文所述的方法。下文描述了示例性的测序方法和方法分析。

[0200] A.测序

[0201] 在一些实施例中,本文提供的方法包括对由扩增产物生成的核酸库进行测序的步骤。

[0202] 在一些实施例中,对通过实践本文所述方法创建的核酸库或DNA库进行测序的技术包括使用基于聚合酶的合成测序、基于连接(ligation based)、焦磷酸测序或基于聚合酶的测序方法。

[0203] 在一些实施例中,核酸库按照MiSeq FGx测序系统参考指南(例如,文档#VD2018006,其内容在此通过引用整体并入)上的说明进行测序。在一些实施例中,按照MiSeq FGx测序系统参考指南(例如,文档#VD2018006)上的说明进行测序的核酸库已变性。

[0204] 在一些方面,本文公开的测序方法包括使用大规模并行测序(MPS)。在一些方面,本文公开的测序方法不包括使用全基因组测序(WGS)。在一些方面,本文公开的测序方法不包括使用微阵列。

[0205] 在一些实施例中,本文公开的测序方法检测了90%或约90%的SNP的基因座。

[0206] 在一些实施例中,本文公开的测序方法生成输出报告,该输出报告包括含有多个SNP的扩增产物的测序结果。

[0207] 在一些实施例中,测序包括多达40重的测序复杂性。在一些实施例中,测序包括2重至40重的测序复杂性。在一些实施例中,测序包括12重至40重的测序复杂性。在一些实施例中,测序包括12重至32重的测序复杂性。在一些实施例中,测序包括12重至30重的测序复杂性。在一些实施例中,测序包括24重至40重的测序复杂性。在一些实施例中,测序包括24重至32重的测序复杂性。在一些实施例中,测序包括28重至32重的测序复杂性。在一些实施例中,测序包括2重、3重、4重、5重、6重、7重、8重、9重、10重、11重、12重、13重、14重、15重、16重、17重、18重、19重、20重、21重、22重、23重、24重、25重、26重、27重、28重、29重、30重、31重或32重的测序复杂性。在一些实施例中,测序包括30重或约30重的测序复杂性。在一些实施例中,测序包括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或40重的测序复杂性。在一些实施例中,测序包括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、41、42、43、44或45重的测序复杂性。测序复杂性是指例如在流动细胞(cell)上一起测序的个体样本的数量。

[0208] 在一些实施例中,测序包括以在或在约6重和16重之间的测序复杂性对死后样本进行测序。在一些实施例中,测序包括以在或在约8重和14重之间的测序复杂性对死后样本进行测序。在一些实施例中,测序包括以在或在约10重和14重之间的测序复杂性对死后样本进行测序。在一些实施例中,测序包括以在或在约10重、11重、12重、13重或14重的测序复杂性对死后样本进行测序。在一些实施例中,测序包括以在或在约12重的测序复杂性对死后样本进行测序。

[0209] 在一些实施例中,测序包括以在或在约24重和40重之间的测序复杂性对生前样本进行测序。在一些实施例中,测序包括以在或在约26重和38重之间的测序复杂性对死后样本进行测序。在一些实施例中,测序包括以在或在约28重和36重之间的测序复杂性对死后样本进行测序。在一些实施例中,测序包括以在或在约28重、29重、30重、31重、32重、33重或34重的测序复杂性对死后样本进行测序。在一些实施例中,测序包括以在或在约32重的测

序复杂性对死后样本进行测序。

[0210] B.分析

[0211] 在一些实施例中,本文提供的方法包括分析扩增产物的序列的步骤。

[0212] 在一些方面,本文公开的方法涉及使用在样本(即扩增产物)测序完成后自动启动分析的分析模块。在一些实施例中,分析模块包括通用分析软件(UAS)。

[0213] 在一些实施例中,本文公开的分析方法生成包括含有多个SNP的扩增产物的测序结果的输出报告。

[0214] 在一些实施例中,使用本领域中可用的任何合适的序列分析软件来分析测序结果。

[0215] 在一些实施例中,使用Forenseq通用分析软件——诸如版本2.1或2.2或更高版本(Verogen,加利福尼亚州圣地亚哥)——按照Forenseq通用分析软件参考指南(诸如版本2.2或更高版本)中概述的说明进行分析,并且例如在参考指南文档#VD2019002中提供,其内容通过引用整体并入本文。

[0216] 基因型和DNA谱确定

[0217] 在一些实施例中,本文提供的方法包括确定多个SNP的基因型的步骤,从而生成DNA谱。

[0218] 在一些实施例中,通过确定多个SNP的基因型来生成DNA谱。

[0219] 在一些方面,包括由本文所述任何方法生成的包含多个SNP的扩增产物的测序结果的输出报告能够用于使用任何已知合适的方法对样本进行基因分型,以补充本文所述方法。在一些方面,包括由本文所述任何方法生成的包含多个SNP的扩增产物的测序结果的输出报告能够用于使用任何已知合适的方法来生成DNA谱,以补充本文所述方法。

[0220] 在一些实施例中,DNA谱包括多个SNP中的每一个的基因型。在一些实施例中,DNA谱包括针对多个SNP中的至少或至少约80%、85%、90%、91%、92%、93%、94%、95%、96%、97%、98%或99%的基因型。在一些实施例中,DNA谱包括针对多个SNP中的至少或至少约85%的基因型。在一些实施例中,DNA谱包括针对多个SNP中的至少或至少约90%的基因型。在一些实施例中,DNA谱包括针对多个SNP中的至少或至少约95%的基因型。在一些实施例中,DNA谱包括针对多个SNP中的至少或至少约99%或约100%的基因型。

[0221] 在一些实施例中,本文公开的方法包括确定毛发颜色、眼睛颜色和生物地理祖先。

[0222] 关系程度确定

[0223] 在一些实施例中,本文提供的方法包括计算DNA谱与一个或多个参考DNA谱的关系程度的步骤,其中,一个或多个参考DNA谱包含在参考DNA谱集内,该参考DNA谱集包括来自感兴趣人员的亲属的一个或多个参考DNA谱。

[0224] 在一些方面,能够使用任何已知的合适方法参考一个或多个参考DNA谱来计算本文所述的DNA谱的关系程度,以补充本文所述的方法。

[0225] 在一些实施例中,一个或多个参考DNA谱包含在参考DNA谱集中,该参考DNA谱集包括来自感兴趣人员的亲属的一个或多个参考DNA谱。

[0226] 在一些实施例中,参考DNA谱集包括多达5、10、15、20、25、30、35、40、45、50、55、60、65、70、75、80、85、90、95、100、125、150、175、200、300、400、500、600、700、800、900或1000个参考DNA谱。在一些实施例中,参考DNA谱集包括多达1000个参考DNA谱。在一些实施例中,参

考DNA谱集包括多达500个参考DNA谱。在一些实施例中,参考DNA谱集包括多达250个参考DNA谱。在一些实施例中,参考DNA谱集包括多达150个参考DNA谱。在一些实施例中,参考DNA谱集包括多达100个参考DNA谱。在一些实施例中,参考DNA谱集包括多达75个参考DNA谱。在一些实施例中,参考DNA谱集包括多达50个参考DNA谱。在一些实施例中,参考DNA谱集包括多达25个参考DNA谱。在一些实施例中,参考DNA谱集包括多达15个参考DNA谱。在一些实施例中,参考DNA谱集包括1和1,000个之间的参考DNA谱、1和500个之间的参考DNA谱、1和400个之间的参考DNA谱、1和300个之间的参考DNA谱、1和250个之间的参考DNA谱、1和200个之间的参考DNA谱、1和150个之间的参考DNA谱、1和100个之间的参考DNA谱、1和75个之间的参考DNA谱、1和50个之间的参考DNA谱、1和25个之间的参考DNA谱、1和20个之间的参考DNA谱、1和15个之间的参考DNA谱、1和10个之间的参考DNA谱或1和5个之间的参考DNA谱。

[0227] 在一些实施例中,参考DNA谱集包括至少1、2、3、4、5、6、7、8、9、10、15、20或25个参考DNA谱,并且包括多达30、35、40、45、50、55、60、65、70、75、80、85、90、95、100、125、150、175、200、300、400、500、600、700、800、900或1000个参考DNA谱。

[0228] 在一些实施例中,参考DNA谱集包括1、2、3、4、5、6、7、8、9、10、11、12、12、14、15、16、17、18、19或20个参考DNA谱。

[0229] 在一些实施例中,参考DNA谱集包括来自感兴趣人员的亲属的DNA谱。在一些实施例中,参考DNA谱集中至少5%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%、65%、70%、75%、80%、85%、90%或95%的参考DNA谱来自感兴趣人员的亲属。在一些实施例中,参考DNA谱集中的70%、71%、72%、73%、74%、75%、76%、77%、78%、79%、80%、81%、82%、83%、84%、85%、86%、87%、88%、89%、90%、91%、92%、93%、94%、95%、96%、97%、98%、99%或100%或约70%、71%、72%、73%、74%、75%、76%、77%、78%、79%、80%、81%、82%、83%、84%、85%、86%、87%、88%、89%、90%、91%、92%、93%、94%、95%、96%、97%、98%、99%或100%的参考DNA谱来自感兴趣人员的亲属。在一些实施例中,参考DNA谱集中100%的参考DNA谱来自感兴趣人员的亲属。在一些实施例中,参考DNA谱集中至少50%的参考DNA谱来自感兴趣人员的亲属。

[0230] 在一些实施例中,来自感兴趣人员的亲属的一个或多个参考DNA谱中的每一个都是生前样本。在一些实施例中,来自感兴趣人员的亲属的一个或多个参考DNA谱中的一个或多个是生前样本。在一些实施例中,来自感兴趣人员的亲属的一个或多个参考DNA谱中的一个或多个是死后样本。在一些实施例中,来自感兴趣人员的亲属的一个或多个参考DNA谱包括死后样本和生前样本。

[0231] 在一些实施例中,参考DNA谱集中的感兴趣人员的每位亲属分别为感兴趣人员的一级亲属、二级亲属、三级亲属、四级亲属或五级亲属。例如,在包括参考DNA谱集的实施例中,该参考DNA谱集包括来自感兴趣人员的亲属的三个参考DNA谱,三个参考DNA谱中的每一个能够独立地来自一级亲属、二级亲属、三级亲属、四级亲属或五级亲属,例如,第一参考DNA谱可以来自一级亲属,第二参考DNA谱可以来自三级亲属,并且第三参考DNA谱可以来自一级亲属。

[0232] 在一些实施例中,参考DNA谱集中的一个或多个参考DNA谱中至少50%、55%、60%、65%、70%、75%、80%、85%、90%或95%是相关的,其中,一个或多个参考DNA谱中的每一个独立地来自参考DNA谱集中的其他一个或多个参考DNA谱中的每一个的一级亲属、二

级亲属、三级亲属、四级亲属或五级亲属。

[0233] 在一些实施例中,参考DNA谱集中至少5%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%、65%、70%、75%、80%、85%、90%或95%的参考DNA谱来自感兴趣人员的亲属,并且参考DNA谱集中至少5%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%、65%、70%、75%、80%、85%、90%或95%的参考DNA谱中的每一个是一级亲属、二级亲属、三级亲属、四级亲属或五级亲属。

[0234] 在一些实施例中,参考DNA谱集中的感兴趣人员的每个亲属的身份是已知的。在一些实施例中,参考DNA谱集中一个或多个参考DNA谱中的每一个的身份是已知的。

[0235] 在一些实施例中,参考DNA谱集包括源自感兴趣人员的样本的DNA谱。例如,在一些实施例中,参考DNA谱集包括源自感兴趣人员消失前——例如,在感兴趣人员是失踪人员的情况——或感兴趣人员成为灾难或冲突的受害者前的感兴趣人员的样本的DNA谱。在一些实施例中,源自感兴趣人员消失前——例如,在感兴趣人员是失踪人员的情况——或感兴趣人员成为灾难或冲突的受害者前的感兴趣人员的样本的DNA谱被用作感兴趣人员的阳性对照,因为该样本是在感兴趣人员生前或失踪或受害之前获得的,并且是被确认为源自感兴趣人员的样本的样本。因此,在一些实施例中,参考DNA谱集包括源自感兴趣人员消失前——例如,在感兴趣人员是失踪人员的情况——或感兴趣人员成为灾难或冲突的受害者前的感兴趣人员的样本的DNA谱,并且在被扩增和/或测序之前,已知是源自感兴趣人员的样本。

[0236] 在一些实施例中,参考DNA谱集在数据库中,例如,基因数据库。在一些实施例中,数据库不可公开访问,即,公众无法访问。在一些实施例中,数据库不是公共数据库,诸如执法机构或第三方遗传学服务可访问的公共数据库。在一些实施例中,数据库不是通过订阅服务公开可访问的。在一些实施例中,数据库不是第三方遗传学服务可访问的。

[0237] 在一些实施例中,计算DNA谱与一个或多个参考DNA谱的关系程度不包括访问可公开访问的数据库,例如,可公开访问的基因数据库。在一些实施例中,计算DNA谱与一个或多个参考DNA谱的关系程度不需要互联网访问来访问包括参考DNA谱集的数据库。在一些实施例中,计算DNA谱与一个或多个参考DNA谱的关系程度包括使用包含参考DNA谱集的本地数据库。如本文所用,“本地数据库”是指仅在本地存储和可访问的数据库,尝试查询该数据库的公众——例如第三方——无法访问该数据库。

[0238] 在一些实施例中,参考DNA谱集包括来自两个或多个无关家庭的DNA谱,例如,两个或多个无关家庭(即,彼此不相关的家庭),每个家庭包括失踪人员和/或灾难或冲突的受害者的一个或多个亲属。例如,在具有来自彼此无关的多个家庭的多个受害者的灾难或冲突事件中,来自每个家庭的一个或多个家庭成员可以在参考DNA谱集中贡献参考DNA谱。然后,该本地参考DNA谱集可以在本地用于从多个无关家庭中识别灾难或冲突的受害者。

[0239] 在一些实施例中,参考DNA谱集或数据库(例如,基因数据库或本地数据库)包括来自具有感兴趣的种族的个体的一个或多个DNA谱。在一些实施例中,参考DNA谱集中至少5%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%、65%、70%、75%、80%、85%、90%或95%的参考DNA谱来自感兴趣的种族。在一些实施例中,参考DNA谱集中至少25%、30%、35%、40%、45%、50%、55%、60%、65%、70%、75%、80%、85%、90%或95%的参考DNA谱来自感兴趣的种族。在一些实施例中,参考DNA谱集中至少50%、55%、

60％、65％、70％、75％、80％、85％、90％或95％的参考DNA谱来自感兴趣的种族。在一些实施例中，参考DNA谱集中至少75％、80％、85％、90％或95％的参考DNA谱来自感兴趣的种族。在一些实施例中，参考DNA谱集中至少95％、95％、97％、98％或99％的参考DNA谱来自感兴趣的种族。在一些实施例中，参考DNA谱集中100％的参考DNA谱来自感兴趣的种族。

[0240] 在一些实施例中，感兴趣人员具有感兴趣的种族。

[0241] 在一些实施例中，感兴趣的种族可以是任何种族，例如，来自任何地方的任何种族。

[0242] 在一些实施例中，感兴趣的种族是稀有种族。在一些实施例中，稀有种族占感兴趣的国家或全世界人口的0.01％、0.05％、0.1％、0.2％、0.3％、0.4％、0.5％、0.6％、0.7％、0.8％、0.9％、1％、2％、3％、4％或5％或更少。在一些实施例中，感兴趣的种族是感兴趣的国家中的任何种族。在一些实施例中，感兴趣的种族是感兴趣的国家中的主导种族。在一些实施例中，感兴趣的种族是感兴趣的国家中的少数种族。在一些实施例中，感兴趣人员来自感兴趣的国家。

[0243] 感兴趣的国家可以是任何感兴趣的国家。在一些实施例中，感兴趣的国家选自阿富汗、阿尔巴尼亚、阿尔及利亚、安道尔、安哥拉、安提瓜和巴布达、阿根廷、亚美尼亚、澳大利亚、奥地利、阿塞拜疆、巴哈马、巴林、孟加拉国、巴巴多斯、白俄罗斯、比利时、伯利兹、贝宁、不丹、玻利维亚、波斯尼亚和黑塞哥维那、博茨瓦纳、巴西、文莱、保加利亚、布基纳法索、布隆迪、科特迪瓦、佛得角、柬埔寨、喀麦隆、加拿大、中非共和国、乍得、智利、中国、哥伦比亚、科摩罗、刚果(刚果-布拉柴维尔)、哥斯达黎加、克罗地亚、古巴、塞浦路斯、捷克共和国、刚果民主共和国、丹麦、吉布提、多米尼加、多米尼加共和国、厄瓜多尔、埃及、萨尔瓦多、赤道几内亚、厄立特里亚、爱沙尼亚、斯威士兰(以前称为“斯威士兰”)、埃塞俄比亚、斐济、芬兰、法国、加蓬、冈比亚、格鲁吉亚、德国、加纳、希腊、格林纳达、危地马拉、几内亚、几内亚比绍、圭亚那、海地、罗马教廷、洪都拉斯、匈牙利、冰岛、印度、印度尼西亚、伊朗、伊拉克、爱尔兰、以色列、意大利、牙买加、日本、约旦、哈萨克斯坦、肯尼亚、基里巴斯、科威特、吉尔吉斯斯坦、老挝、拉脱维亚、黎巴嫩、莱索托、利比里亚、利比亚、列支敦士登、立陶宛、卢森堡、马达加斯加、马拉维、马来西亚、马尔代夫、马里、马耳他、马绍尔群岛、毛里塔尼亚、毛里求斯、墨西哥、密克罗尼西亚、摩尔多瓦、摩纳哥、蒙古、黑山、摩洛哥、莫桑比克、缅甸(前缅甸)、纳米比亚、瑙鲁、尼泊尔、荷兰、新西兰、尼加拉瓜、尼日尔、尼日利亚、朝鲜、北马其顿、挪威、阿曼、巴基斯坦、帕劳、巴勒斯坦国、巴拿马、巴布亚新几内亚、巴拉圭、秘鲁、菲律宾、波兰、葡萄牙、卡塔尔、罗马尼亚、俄罗斯、卢旺达、圣基茨和尼维斯、圣圣卢西亚、圣文森特和格林纳丁斯、萨摩亚、圣马力诺、圣多美和普林西比、沙特阿拉伯、塞内加尔、塞尔维亚、塞舌尔、塞拉利昂、新加坡、斯洛伐克、斯洛文尼亚、所罗门群岛、索马里、南非、韩国、南苏丹、西班牙、斯里兰卡、苏丹、苏里南、瑞典、瑞士、叙利亚、塔吉克斯坦、坦桑尼亚、泰国、东帝汶、多哥、汤加、特立尼达和多巴哥、突尼斯、土耳其、土库曼斯坦、图瓦卢、乌干达、乌克兰、阿拉伯联合酋长国、英国、美利坚合众国、乌拉圭、乌兹别克斯坦、瓦努阿图、委内瑞拉、越南、也门、赞比亚、津巴布韦。在一些实施例中，感兴趣的国家是美利坚合众国。

[0244] 在一些实施例中，本文所述的基于DNA的亲属关系分析包括使用本地数据库。在一些实施例中，本文所述的基于DNA的亲属关系分析允许以最少的用户输入生成报告。在一些实施例中，本文所述的基于DNA的亲属关系分析包括使用算法来计算亲属关系系数。在一些

实施例中,亲属关系系数确定样本或DNA谱与数据库中的参考DNA谱的关系状态。例如,在一些实施例中,亲属关系系数基于亲属关系系数的相对值指示一个或多个已识别的遗传亲属中的每一个是否可能是曾曾祖母、曾曾祖父、曾祖父、曾祖母、祖母、祖父、堂/表亲、隔一代的堂/表亲或第二代的堂/表亲。在一些实施例中,参考DNA谱是遗传学数据库的一部分。

[0245] 在一些实施例中,本文所述的基于DNA的亲属关系分析包括识别为1级、2级、3级、4级或5级的遗传亲属或为约1级、2级、3级、4级或5级的遗传亲属。在一些实施例中,本文所述的基于DNA的亲属关系分析包括识别多于1级、2级、3级、4级或5级的遗传亲属。在一些实施例中,本文所述的基于DNA的亲属关系分析包括识别感兴趣人员与参考DNA谱集中的一个或多个参考DNA谱之间的关系程度。例如,在一些实施例中,该方法包括识别感兴趣人员独立地是一个或多个参考DNA谱中的一个或多个的一级亲属、二级亲属、三级亲属、四级亲属或五级亲属。一个人的一级亲属是该人的父母(例如,父亲或母亲)、同胞兄弟姐妹(例如,姐妹或兄弟)或孩子(例如,儿子或女儿)。一个人的二级亲属是与该人共享约25%基因的人,诸如该人的祖父母、姑/姨(aunt)、叔/舅(uncle)、侄女、侄子、孙子或半亲的兄弟姐妹。一个人的三级亲属是与该人共享约12.5%基因的人,诸如曾祖父母、堂/表亲和曾孙。四级亲属包括,例如,隔一代的堂/表亲、半亲的叔祖父、半亲的姑祖母、半亲的侄孙女、半亲的侄孙和半亲的堂/表亲。五级亲属包括,例如,第二代的堂/表亲、半亲的隔一代的堂/表亲和隔两代的堂/表亲。

[0246] 在一些实施例中,本文所述的基于DNA的亲属关系分析包括生成包含与一个或多个DNA谱相关的DNA谱的家谱。能够使用任何可用的手段或方法来生成家谱。

[0247] 在一些实施例中,本文描述的基于DNA的亲属关系分析包括通过共同祖先来识别嫌疑人。

[0248] 在一些实施例中,计算关系程度包括计算DNA谱——即来自感兴趣人员的DNA谱——与参考DNA谱集——例如,包括来自感兴趣人员的亲属的一个或多个参考DNA谱的参考DNA谱集——中包含的一个或多个参考DNA谱之间的关系程度。

[0249] 在一些实施例中,计算关系程度包括通过比较SNP集合——其为或包括一个或多个Y-SNP——来计算DNA谱(即,来自感兴趣人员的DNA谱)与参考DNA谱集(例如,包括来自感兴趣人员的亲属的一个或多个参考DNA谱的参考DNA谱集)中包含的一个或多个参考DNA谱之间的关系程度。在一些实施例中,一个或多个Y-SNP包括至少5、10、15、20、25、30、35、40、45、50、55、60、65、70、75、80、81、82、83、84或85个Y-SNP。在一些实施例中,一个或多个Y-SNP是或包括85个Y-SNP。通过比较包括来自生物学男性的DNA谱与来自生物学男性的一个或多个参考DNA谱之间的SNP集合,能够确定男性世系的似然比。

[0250] 似然比(LR)和亲属关系值能够使用本领域已知的任何方法或算法来计算。在一些实施例中,似然比是使用算法pedprobr(Brustad等人,Int.J.Legal Med.,2021,135:117-129,其内容在此通过引用整体并入本文)和dvir(Vigeland等人,Scientific Reports,2021,11:13661,其内容在此通过引用整体并入本文)来计算的。在一些实施例中,在LR计算中使用来自基因组聚合数据库(gnoMAD)(Karczewski等人,Nature,2020,581:434-443,其内容在此通过引用整体并入本文)v3.0的群体频率的平均值。在一些实施例中,不使用突变模型,并且当选择用于分析的SNP具有低连锁不平衡时,将theta设置为0(Karczewski等人,同上,其内容通过引用整体并入本文)。在一些实施例中,LR计算如下:

[0251] $$LR = \frac{P(D \mid H_r)}{P(D \mid H_u)}$$

[0252] 其中,D表示所述基因型,H_r表示个体是相关的假设,并且H_u表示个体是无关的假设。相关假设由谱系表示,其中未识别个体被测试为亲属。无关假设由Hardy-Weinberg平衡计算表示。在一些实施例中,计算每个基因座的LR值,然后将其乘以基因座,从而得到关系的最终LR。为了提高计算效率,能够将每个基因座LR转换为对数,并将基因座LR相加。然而,由于该平台的复杂性高,并且在大规模死亡事件(MFI)案例中,PM样本的退化程度通常很高和/或输入量很低,PCR期间的随机效应会导致等位基因脱扣(dropout),并可能导致两个个体在某个基因座的等位基因组合似乎不可能的情况。这通常发生在亲子关系中,其中关系中的每个个体都被基因分型为不同等位基因的纯合子,但其中一个或两个实际上是杂合子(例如,当父母或父母和孩子这两者实际上是AB时,父母的基因型为AA而孩子为基因型BB)。基因分型错误和新生突变也是造成这种情况的可能原因。这种情况导致LR为0,并且可能导致总体LR为0,因为最终LR是基因座LR的乘积。为了确保基因座LR不为0,在一些实施例中对pedprobr进行了修改。因此,在一些实施例中,这些情况下基因座的似然比计算如下:

[0253] $$LR = \frac{0.001}{p^2 q^2}$$

[0254] 其中,0.001表示基因分型错误率,p是等位基因1的等位基因频率,并且q是等位基因2的等位基因频率(Galván-Femenía等,Heredity,2021,126:537-547,其内容在此通过引用整体并入)。等位基因1和等位基因2,也称为第一等位基因和第二等位基因或基因座,能够是特定基因座的任何两个感兴趣的等位基因。

[0255] 在一些实施例中,LR是按照Galván-Femenía等人,Heredity,2021,126:537-547中所述来计算的,其内容在此通过引用整体并入。

[0256] 在一些实施例中,计算关系程度包括计算DNA谱(即,来自感兴趣人员的DNA谱)与参考DNA谱集(例如,包括来自感兴趣人员的亲属的一个或多个参考DNA谱的参考DNA谱集)中包含的一个或多个参考DNA谱之间共享Y染色体的似然比。在一些实施例中,计算共享Y染色体的似然比包括比较SNP集合,该SNP集合是一个或多个Y-SNP或包括一个或多个Y-SNP。在一些实施例中,一个或多个Y-SNP包括5、10、15、20、25、30、35、40、45、50、55、60、65、70、75、80、81、82、83、84或85个Y-SNP或至少5、10、15、20、25、30、35、40、45、50、55、60、65、70、75、80、81、82、83、84或85个Y-SNP。在一些实施例中,一个或多个Y-SNP是或包括85个Y-SNP。

[0257] 在一些实施例中,计算共享Y染色体的似然比包括基于来自多个SNP中的一个或多个Y-SNP来计算亲属关系系数。在一些实施例中,计算共享Y染色体的似然比包括通过提供DNA谱和包含在参考DNA谱集内的一个或多个参考DNA谱作为输入(例如,作为PC-Relate的输入)来计算对数似然比。在一些实施例中,计算共享Y染色体的似然比能够允许识别DNA谱(例如,感兴趣人员的DNA谱)与一个或多个参考DNA谱中的一个或多个之间的匹配Y染色体,然后能够使用其来确定感兴趣人员的男性世系的似然比。

[0258] 在一些实施例中,这包括使用计算两个样本之间(例如,DNA谱与一个或多个参考DNA谱之间)的关系测量值的亲属关系系数。在一些实施例中,本文提供的方法涉及使用从例如公共遗传学数据库构建的亲属关系模型(例如使用PC-AiR或改进的PC-AiR方法)计算亲属关系系数,并且确定本地靶样本集的亲属关系(例如,使用PC-Relate),而不是使用大

量可公开访问的样本集在公共数据库上确定亲属关系。在一些实施例中,这还包括计算每次比较的似然比(LR)。例如,似然比(LR)是法医领域中关联度的标准度量。

[0259] 在一些实施例中,使用PC-Relate(Conomos等人,American Journal of Human Genetics,2016,98:127-148,其内容在此通过引用整体并入)和PC-AiR(Conomos等人,Genetic Epidemiology,2015,39:276-293,其内容在此通过引用整体并入)来计算全基因组亲属系数、共享cMs和最长片段cM,诸如先前在Snedecor等人,Forensic Sci.Int.Genet.,2022,61:102769中所述,其内容在此通过引用整体并入。当两个个体之间的关系未知时,此方法很有用,因此不需要谱系,这适用于诸如具有失踪人员或冲突受害者的情况。

[0260] 在一些实施例中,PC-AiR方法首先获取基因型个体集合,并将它们分成两个不重叠的子集:一个集合包含代表所有个体祖先的无关个体(无关子集),另一个集合包含在第一个子集中至少有一个亲属的个体(相关子集)。为了构建无关子集,对原始PC-AiR方法进行了修改以提高构建模型的计算效率。没有亲属或亲属最少的样本被添加到无关子集中,而亲属较多的样本则被排除在无关子集之外。这是通过计算每对亲属关系值来执行的,如Conomos等人在Genetic Epidemiology,2015,39:276-293中所述,其内容在此通过引用整体并入本文中,并基于严格的阈值将每个个体分类为有亲属或没有亲属关系:如果亲属关系值大于0.01,则认为是亲属,如果亲属关系值小于-0.025,则认为是无亲属关系。排除具有小于5%的失踪SNP数据的样本。接下来,对无关子集进行主组分分析(PCA),然后基于与无关子集中个体的遗传相似性,预测相关子集中所有个体的变异组分值。所得组分代表能够用于代替静态种群频率来识别未知个体集合中的匹配的模型。

[0261] 在一些实施例中,PC-Relate方法使用来自PC-AiR的主成分并且将遗传相关性分为两个组分:一个组分用于共享由最近的共同祖先传下来的相同的等位基因,并且另一个组分用于共享由于更远的共同祖先而产生的等位基因。来自PC-AiR的组分用于基于个体的祖先背景估计等位基因频率,使用线性回归而不是静态种群频率(诸如来自gnoMAD的频率)。对于两个个体i和j,然后使用来自PC-AiR模型的估计等位基因频率$\hat{u}$来计算亲属关系系数$\widehat{\phi_{ij}}$,如下所示

[0262] $$\widehat{\phi_{ij}} = \frac{\sum_{s\in S}(g_{is} - 2\hat{u}_{is})(g_{js} - 2\hat{u}_{js})}{4\sum_{s\in S}[\hat{u}_{is}(1 - \hat{u}_{is}) \cdot \hat{u}_{js}(1 - \hat{u}_{js})]^{1/2}}$$

[0263] 其中,s是在两个个体中都被分型的S个SNP中的SNP,g_{is}和g_{js}分别是在SNP s处的i和j中的参考等位基因的数量,并且$\hat{u}_{is}$和$\hat{u}_{js}$分别是由在SNP s处的i和j的PC-AiR计算出的预期等位基因频率。该算法被称为“全基因组亲属关系”,因为它将整个基因组视为一个关联度的片段,其用全基因组亲属关系系数表示。在引用全基因组亲属关系算法时,会使用全基因组亲属关系系数来识别关系。当在两个个体之间共享的SNP数量少于6000时,本研究中需要两个个体之间的全基因组亲属关系系数超过0.031才能被视为亲属。

[0264] 因此,在一些实施例中,计算关系程度包括使用全基因组亲属关系算法来计算亲属关系系数,如下所示:

[0265] $$\widehat{\phi_{ij}} = \frac{\sum_{s\in S}(g_{is} - 2\hat{u}_{is})(g_{js} - 2\hat{u}_{js})}{4\sum_{s\in S}[\hat{u}_{is}(1 - \hat{u}_{is}) \cdot \hat{u}_{js}(1 - \hat{u}_{js})]^{1/2}}$$

[0266] 其中,感兴趣人员和参考DNA谱是i和j,$\widehat{\phi_{ij}}$是亲属关系系数,$\hat{u}$是估计等位基因频率,s是在两个个体中都被分型的S个SNP中的SNP,g_{is}和g_{js}分别是在SNP s处的i和j中的参考等位基因的数量,并且$\hat{u}_{is}$和$\hat{u}_{js}$分别是由在SNP s处的i和j的PC-AiR计算出的预期等位基因频率。

[0267] 在一些实施例中,计算关系程度包括使用"窗口亲属关系"方法来计算亲属关系系数。参见Snedecor等人,Forensic Sci Int Genet 2022,61,102769,doi:10.1016/j.fsigen.2022.102769,其内容在此通过引用整体并入本文。窗口亲属关系涉及计算跨基因组的亲属关系的窗口以找到共享的亲属关系片段。这是通过枚举每个染色体内的所有可能窗口并计算所有窗口的亲属关系系数来执行的。然后通过最小亲属关系系数阈值过滤这些窗口并将其包含在共享cMs计算中。然后迭代过滤后的片段,并分别对共享至少一个等位基因和两个等位基因的SNP延伸进行分类。然后跨所有片段来计算总的共享cM。在参考窗口亲属关系算法时,使用总的共享cM和最长cM片段来识别关系。当两个个体之间共享的SNP数量在6000和8000之间时,共享cM值必须高于180且最长cM片段必须高于30才被视为存在关系。当两个个体之间共享的SNP数量在8000和9000之间时,共享cM值必须高于150且最长cM片段必须高于30才被视为存在关系。当两个个体之间共享的SNP数量为9000或更多时,共享cM值必须高于140且最长cM片段必须高于30才被视为存在关系。全基因组亲属关系系数能够用于过滤任意数量的共享SNP。然而,Snedecor等人(上文)观察到,当SNP重叠大于6000时,对共享cM和最长片段cM进行过滤(例如,使用窗口亲属关系)时具有更高的特异性,特别是对于更高的关系程度。

[0268] 更简单地说,两个个体之间分型的SNP的数量(SNP重叠)能够用于决定何时使用全基因组亲属关系算法(<6000个SNP重叠)以及何时使用窗口亲属关系算法(>6000个SNP重叠)。并且,当基于该SNP重叠来决定一种算法时,将使用一个值或值的集合来过滤数据以识别关系,具体取决于所选的算法。全基因组亲属关系和窗口亲属关系这两者的截止值都经过选择以确保高灵敏度,但更重要的是,高特异性,如Snedecor等人(上文)所示。降低这些阈值可能会捕获更多的关系(即增加灵敏度),但预计会引入更多的假阳性结果,特别是对于更远的关系(例如,四级和五级)。

[0269] 在一些实施例中,计算关系程度包括计算DNA谱(例如,来自感兴趣人员的DNA谱)和一个或多个参考DNA谱中的一个的亲属关系系数。在一些实施例中,计算DNA谱和一个或多个参考DNA谱中的每一个的关系程度(例如,亲属关系系数)。在一些实施例中,基于在两个样本中观察到的基因型,通过将查询(例如,感兴趣人员的DNA谱)和靶(例如,一个或多个参考DNA谱中的一个)相关的概率除以查询和靶不相关的概率来计算似然比。然后,能够基于亲属关系系数和LR来过滤结果以识别最可能的关系并且从参考DNA谱集中包含的一个或多个参考DNA谱中消除错误匹配。

[0270] 在一些实施例中,计算似然比包括比较DNA谱和一个或多个参考DNA谱之间的多个SNP。在一些实施例中,计算关系程度包括基于来自多个SNP中的亲属关系SNP来计算亲属关系系数。在一些实施例中,计算关系程度包括基于来自多个SNP中的亲属关系SNP来计算亲属关系系数以及基于来自多个SNP中的Y-SNP来计算亲属关系系数。在一些实施例中,计算关系程度包括基于来自多个SNP中的Y-SNP来计算亲属关系系数。在一些实施例中,计算似

然比包括比较SNP集合,该SNP集合包括来自DNA谱和一个或多个参考DNA谱之间的多个SNP中的亲属关系SNP。

[0271] 在一些实施例中,感兴趣人员在生物学上是男性,并且方法还包括计算DNA谱与一个或多个参考DNA谱之间共享Y染色体的似然比。在一些实施例中,计算共享Y染色体的似然比包括比较DNA谱和一个或多个参考DNA谱之间的SNP集合,所述SNP集合包括一个或多个Y-SNP。在一些实施例中,一个或多个Y-SNP包含在多个SNP内。在一些实施例中,一个或多个Y-SNP包括至少20、25、30、35、40、45、50、55、60、65、70、75、80、81、82、83、84或85个Y-SNP。在一些这样的实施例中,一个或多个Y-SNP包括85个Y-SNP。在一些实施例中,计算共享Y染色体的似然比包括基于一个或多个Y-SNP的基因型,将DNA谱与一个或多个参考DNA谱中的参考DNA谱共享Y染色体的概率除以DNA谱与参考DNA谱不共享Y染色体的概率。

[0272] 在一些实施例中,计算关系程度,例如亲属关系系数,包括使用主组分分析(PCA)方法。在一些实施例中,使用亲属关系模型来计算关系程度。在一些实施例中,使用利用PCA方法训练的亲属关系模型来计算关系程度。在一些实施例中,用于训练亲属关系模型的PCA方法是PCA。在一些实施例中,用于训练亲属关系模型的PCA方法涉及PCA。在一些实施例中,用于训练亲属关系模型的PCA方法能够解释样本关联度,例如能够由跨样本的家庭结构引起的已知或隐性关联度。在一些实施例中,PCA方法是PC-AiR,其能够允许在存在已知或隐性关联度的情况下用于祖先确定。参见,例如,Conomos等人,稳健的人口结构推断用于祖先预测和在存在关联度的情况下对分层的校正,Genet Epidemiol.,2015,39(4):276-293,其内容通过引用并入本文。在一些实施例中,PCA方法是改进的PC-AiR方法,例如本文所述。

[0273] 在一些实施例中,使用训练数据库来构建亲属关系模型。在一些实施例中,训练数据库是基因数据库。在一些实施例中,训练数据库是遗传学数据库。在一些实施例中,训练数据库是可公开访问的数据库。在一些实施例中,训练数据库包括1到1000万或更多个训练DNA谱。在一些实施例中,训练数据库包括为或约为或至少为或至少约为1、5、25、50、75、100、500、1,000、1,500、2,000、3,000、4,000、5,000、10,000、20,000、30,000、40,000、50,000、75,000、100,000、125,000、150,000、175,000、200,000、225,000、250,000、275,000、300,000、400,000、500,000、600,000、700,000、800,000、900,000、1,000,000、1,250,000、1,500,000、1,750,000、2,000,000、3,000,000、4,000,000、5,000,000或10,000,000个训练DNA谱,或者前述任意两个值之间的范围。在一些实施例中,训练数据库包括多达或多达约100、500、1,000、1,500、2,000、3,000、4,000、5,000、10,000、20,000、30,000、40,000、50,000、75,000、100,000、125,000、150,000、175,000、200,000、225,000、250,000、275,000、300,000、400,000、500,000、600,000、700,000、800,000、900,000、1,000,000、1,250,000、1,500,000、1,750,000、2,000,000、3,000,000、4,000,000、5,000,000或10,000,000个训练DNA谱。在一些实施例中,训练数据库包括5,000至500,000个之间、或10,000至500,000个之间、或15,000至500,000个之间、或20,000至500,000个之间、或25,000至500,000个之间、或25,000至400,000个之间、或25,000至300,000个之间、或25,000至250,000个之间、或50,000至500,000个之间、或50,000至400,000个之间、或50,000至300,000个之间、或50,000至250,000个之间的训练DNA谱。

[0274] 在一些实施例中,PCA方法是PC-AiR,并且训练数据库包括至少1并且多达100、200、300、400、500、600、700、800、900、1,000、1,100、1,200、1,300、1,400、1,500、1,600、1,

700、1,800、1,900、2,000、2,100、2,200、2,300、2,400、2,500、2,600、2,700、2,800、2,900、3,000、3,500、4,000、4,500或5,000个训练DNA谱,或上述任意两个值之间的范围。

[0275] 在一些实施例中,PCA方法是改进的PC-Air方法,并且训练数据库包括为或约为或至少为或至少约为3,000、4,000、5,000、10,000、20,000、30,000、40,000、50,000、75,000、100,000、125,000、150,000、175,000、200,000、225,000、250,000、275,000、300,000、400,000、500,000、600,000、700,000、800,000、900,000、1,000,000、1,250,000、1,500,000、1,750,000、2,000,000、3,000,000、4,000,000、5,000,000或10,000,000个训练DNA谱,或者上述任意两个值之间的范围。

[0276] 在一些实施例中,访问训练数据库不需要互联网访问。在一些实施例中,训练亲属关系模型不需要互联网访问。在一些实施例中,训练数据库可在本地访问。

[0277] 在一些实施例中,通过将PCA方法应用于训练数据库来训练亲属关系模型。在一些实施例中,训练DNA谱包括多个SNP的基因型。在一些实施例中,亲属关系模型包括使用PCA方法为训练数据库获得的主组分(PC)。

[0278] 在一些方面,给定训练DNA谱的训练数据库,PC-AiR和改进的PC-AiR方法都能够从训练数据库中识别足够可接受的训练DNA谱的无关样本集合,该无关样本集合尽可能接近训练数据库中存在的所有祖先背景,同时还可以从训练数据库中存在的所有祖先背景中进行良好的采样。在一些实施例中,PC-AiR和改进的PC-AiR方法都能够在训练数据库中识别无关样本集合,例如训练DNA谱。在一些实施例中,无关样本集合是对训练数据库中存在的所有或几乎所有祖先背景进行采样的样本集合。

[0279] 在一些实施例中,PC-AiR和改进的PC-AiR方法都包括估计训练数据库中所有样本对之间的亲属关系的初始步骤。在一些实施例中,估计亲属关系系数。在一些实施例中,使用称为“KING-Robust”的简化亲属关系估计方法来估计亲属关系系数。

[0280] 在一些实施例中,PC-AiR然后进入后续步骤,包括:(1)使用来自训练数据库的所有样本来初始化集合“U”;(2)针对每个样本,扫描该集合以计算该样本在U中与多少个样本相关(称为“R”),以及该样本在U中与多少个样本“祖先分化”(称为“D”);(3)选择具有最高R的样本,如果有多个样本具有最高R,则选择具有最高R和最低D的样本;(4)从U中移除选定的样本;以及(5)从步骤(2)重复。例如,使用PC-AiR,如果有50,000个样本,则该过程将在第一次迭代中查看$50,000^2$个数据点,在第二次迭代中查看$49,999^2$个数据点,以此类推,直到集合中不再有相关样本,这可能会继续下降到例如$20,000^2$或$10,000^2$个数据点。在一些实施例中,该过程持续进行,直到U仅包含无关样本为止。

[0281] 在一些实施例中,PC-AiR基于估计亲属关系将样本视为相关样本。在一些实施例中,估计亲属关系系数≥0.025的样本被视为相关样本。

[0282] 在一些实施例中,PC-AiR基于估计亲属关系将样本视为祖先分化样本。在一些实施例中,估计亲属关系系数<0.025的样本被视为祖先分化样本。

[0283] 在一些实施例中,PC-AiR包括以下步骤:估计训练数据库的所有样本对(例如,训练DNA谱)之间的亲属关系系数,其中,亲属关系系数>0.025的配对被识别为相关的,而亲属关系系数<-0.025的配对被识别为祖先分化的;(2)初始化包括所有样本的无关样本集;以及(3)迭代地:(i)识别所述无关样本集中具有所述无关样本集中最相关样本的集合,从而指定为X,(ii)识别X中与所述无关样本集中的样本相比具有最少祖先分化配对的样本集,

从而指定为Y;以及(iii)如果Y具有零个样本,则终止过程,或者,如果Y具有至少一个样本,则从Y中随机选择一个样本以从U中移除,并且在步骤(3)(i)处开始重复。

[0284] 在一些实施例中,改进的PC-AiR方法与PC-AiR相比包括一个或多个调整。在一些实施例中,改进的PC-AiR方法对样本是否相关有更严格的定义。在一些实施例中,如果估计亲属关系系数≥0.01,则改进的PC-AiR方法认为样本相关。

[0285] 在一些实施例中,改进的PC-AiR方法基于估计亲属关系系数将样本视为祖先分化。在一些实施例中,估计亲属关系系数<0.025的样本被视为祖先分化。

[0286] 在一些实施例中,改进的PC-AiR方法包括去除所有具有>5%失踪基因型的样本(例如,DNA谱中超过5%的SNP),以确保每个样本都具有足够的信息量。

[0287] 在一些实施例中,改进的PC-AiR方法包括以下步骤:(1)对于每个样本,计算:“R”是训练数据库中相关样本的总数,“D”是数据库中祖先分化样本的数量,并且“S”是相关样本的集合;(2)按R(升序)和D(降序)对所有样本进行排序;(3)迭代地遍历已排序的样本列表,并且:(i)如果样本不在“相关”集合中,则将其添加到无关集合中,并将来自S的所有样本(即,与样本相关的DNA谱)添加到相关集合中;或者(ii)如果样本在“相关”集合中,则忽略该样本并移动到下一个样本。在某些方面,这种改进的PC-AiR方法允许实现很大程度上线性复杂度(即,运行时间随样本数量线性扩展)而不是指数级的过程

[0288] 在一些实施例中,改进的PC-AiR方法包括以下步骤:(1)估计训练数据库的所有样本对——例如DNA谱——之间的亲属关系系数,其中,亲属关系系数>0.01的配对被识别为相关的,而亲属关系系数<-0.025的配对被识别为祖先分化的;(2)移除所有具有≥5%失踪数据的DNA谱;以及(3)通过用排序值识别每个DNA谱来对所有DNA谱进行排序。在一些实施例中,排序值是基于完整数据库中从最少到最多排序的相关DNA谱的数量来确定的,并且通过所述完整数据库中从最多到最少排序的祖先分化DNA谱的数量来打破纽带(tie)。在一些实施例中,步骤(3)涉及迭代地遍历排序后的DNA谱,并且对于每个DNA谱:(i)如果DNA谱尚未出现在相关样本集中,则将其添加到无关样本集并且将所有相关DNA谱添加到相关样本集,以及(ii)如果DNA谱已经在相关样本集中,则跳至下一个DNA谱,并且在步骤(3)(i)处开始重复。

[0289] 在一些实施例中,使用AiR或改进的PC-AiR方法确定无关样本集之后,将PCA应用于无关样本集以训练亲属关系模型。在一些实施例中,亲属关系模型还包括针对相关样本集计算的PC值。在一些实施例中,相关样本集的PC值是基于针对无关样本集获得的PC而确定的。

[0290] 在一些实施例中,PCA被应用于整个训练数据库以构建亲属关系模型。

[0291] 在一些实施例中,所提供的方法涉及训练亲属关系模型。

[0292] 在一些实施例中,所提供的方法不涉及训练亲属关系模型。在一些实施例中,在计算关系程度(例如,亲属关系系数)之前训练亲属关系模型。

[0293] 在一些实施例中,访问亲属关系模型不需要互联网访问。在一些实施例中,亲属关系模型可在本地访问。

[0294] 在一些实施例中,使用亲属关系模型来计算关系程度,例如,亲属关系系数。在一些实施例中,使用亲属关系模型的PC来计算关系程度。在一些实施例中,计算关系程度涉及获取DNA谱的PC值,例如,感兴趣人员的DNA谱。在一些实施例中,计算关系程度涉及获取参

考DNA谱或多个参考DNA谱的PC值。在一些实施例中,使用DNA谱的PC值来计算关系程度。在一些实施例中,使用DNA谱和参考DNA谱或多个参考DNA谱的PC值来计算关系程度。

[0295] 在一些实施例中,使用PC-Relate来计算关系程度(例如,亲属关系系数)。参见,例如,Conomos等人,近期遗传关联度的无模型估计,Am.J.Hum.Genet.,98(1):127-148(2016),其内容在此通过引用整体并入本文。在一些实施例中,通过将DNA谱(例如,感兴趣人员的DNA谱)提供作为PC-Relate的输入来计算关系程度。在一些实施例中,通过将亲属关系模型(例如,PC)和DNA谱提供作为PC-Relate的输入来计算关系程度。在一些实施例中,参考DNA谱或多个参考DNA谱还被提供作为PC-Relate的输入。

[0296] 在一些实施例中,关系程度(例如,亲属关系系数)是在本地计算的。在一些实施例中,计算关系程度不需要互联网访问。

[0297] 在一些实施例中,本文所述方法还包括识别感兴趣人员。在一些实施例中,识别感兴趣人员包括通过感兴趣人员的法定姓名来识别感兴趣人员。在一些实施例中,识别感兴趣人员包括通过感兴趣人员与参考DNA谱集中的一个或多个已知人的家族关系来识别感兴趣人员。例如,在一些实施例中,识别感兴趣人员包括识别感兴趣人员是特定已知人的儿子或女儿,和/或特定已知人的全体兄弟姐妹。

[0298] 试剂盒

[0299] 本文提供了包括本文所述任何引物、试剂或组合物的试剂盒,其还可以包括关于使用试剂盒的方法(诸如本文所述用途)的说明。本文所述试剂盒还可以包括从商业和用户角度来看理想的其他材料,其包括其他缓冲液、稀释剂、过滤器和带有用于执行本文所述任何方法的说明的包装说明书。

[0300] 在一些实施例中,本文提供了包括至少一个容器装置的试剂盒,其中,至少一个容器装置包括如本文所述的多个引物中的任一个。

[0301] 示例性实施例

[0302] 本文提供的示例性实施例包括:

[0303] 1.一种执行基于DNA的亲属关系分析的方法,包括:

[0304] 提供来自感兴趣人员的核酸样本,

[0305] 使用与多个靶序列特异性杂交的多个引物来扩增所述核酸样本,所述多个靶序列共同包括多个至少在2,000至50,000个之间或约2,000至50,000个之间的单核苷酸多态性(SNP),从而生成扩增产物,其中,所述扩增在一个或多个多重PCR反应中进行,

[0306] 从所述扩增产物生成核酸库,

[0307] 对从所述扩增产物生成的所述核酸库进行测序,

[0308] 分析所述扩增产物的序列,

[0309] 确定所述多个SNP的基因型,从而生成DNA谱,以及

[0310] 计算所述DNA谱与一个或多个参考DNA谱的关系程度,其中,所述一个或多个参考DNA谱包含在参考DNA谱集内,所述参考DNA谱集包括来自所述感兴趣人员的亲属的一个或多个参考DNA谱。

[0311] 2.一种执行基于DNA的亲属关系分析的方法,包括:

[0312] 提供来自感兴趣人员的核酸样本,

[0313] 使用与多个靶序列特异性杂交的多个引物来扩增所述核酸样本,所述多个靶序列

共同包括多个至少在2,000至50,000个之间或约2,000至50,000个之间的单核苷酸多态性(SNP),从而生成扩增产物,其中,所述扩增在一个或多个多重PCR反应中进行,

[0314] 从所述扩增产物生成核酸库,

[0315] 对从所述扩增产物生成的所述核酸库进行测序,

[0316] 确定所述多个SNP的基因型,从而生成DNA谱,以及

[0317] 计算所述DNA谱与一个或多个参考DNA谱的关系程度,其中,所述一个或多个参考DNA谱包含在参考DNA谱集内,所述参考DNA谱集包括来自所述感兴趣人员的亲属的一个或多个参考DNA谱。

[0318] 3.根据实施例1或2所述的方法,其中,使用大规模并行测序(MPS)来进行所述测序。

[0319] 4.根据实施例1至3中任一项所述的方法,其中,所述测序不包括全基因组测序(WGS)。

[0320] 5.根据实施例1至4中任一项所述的方法,还包括生成包含与一个或多个DNA谱相关的所述DNA谱的家谱。

[0321] 6.一种构建感兴趣人员的核酸库的方法,包括:

[0322] 提供来自感兴趣人员的核酸样本,

[0323] 使用与多个靶序列特异性杂交的多个引物来扩增所述核酸样本,所述多个靶序列共同包括多个至少在2,000至50,000个之间或约2,000至50,000个之间的单核苷酸多态性(SNP),从而生成包括扩增产物的核酸库,其中,所述扩增在一个或多个多重PCR反应中进行。

[0324] 7.根据实施例6所述的方法,还包括对所述扩增产物进行测序以产生所述感兴趣人员的DNA谱的步骤。

[0325] 8.一种构建参考DNA样本的核酸库的方法,包括:

[0326] 提供来自感兴趣人员的亲属的核酸样本,

[0327] 使用与多个靶序列特异性杂交的多个引物来扩增所述核酸样本,所述多个靶序列共同包括多个至少在2,000至50,000个之间或约2,000至50,000个之间的单核苷酸多态性(SNP),从而生成包括扩增产物的核酸库,其中,所述扩增在一个或多个多重PCR反应中进行。

[0328] 9.根据实施例8所述的方法,其中,所述亲属是所述感兴趣人员的一级、二级、三级、四级或五级亲属。

[0329] 10.根据实施例8或9所述的方法,其中,所述亲属是所述感兴趣人员的一级、二级或三级亲属。

[0330] 11.根据实施例1至10中任一项所述的方法,其中,所述核酸样本包括基因组DNA。

[0331] 12.根据实施例1至11中任一项所述的方法,其中,所述核酸样本包括一种或多种酶抑制剂。

[0332] 13.根据实施例12所述的方法,其中,所述一种或多种酶抑制剂包括选自由血红素、亚铁血红素、腐殖酸、靛蓝、单宁酸、胶原蛋白、钙和羟基磷灰石组成的组中的一种或多种抑制剂。

[0333] 14.根据实施例1至13中任一项所述的方法,其中,所述核酸样本包括低质量核酸

分子和/或低数量核酸分子。

[0334] 15.根据实施例14所述的方法,其中,所述低质量核酸分子是退化的基因组DNA和/或片段化的基因组DNA。

[0335] 16.根据实施例14或实施例15所述的方法,其中,所述低质量核酸分子具有为1、2、3、4、5、6、7、8、9、10、15、20、25、30、35、40、45、50、55、60、65、70、75、80、85、90、95、100、105、110、115、120、125、130、135、140、145、150、155、160、165、170、175、180、185、190、195或200的退化指数(DI)或至少为1、2、3、4、5、6、7、8、9、10、15、20、25、30、35、40、45、50、55、60、65、70、75、80、85、90、95、100、105、110、115、120、125、130、135、140、145、150、155、160、165、170、175、180、185、190、195或200的退化指数(DI)。

[0336] 17.根据实施例14或实施例15所述的方法,其中,所述低质量核酸分子具有至少1且多达或小于158.3的DI。

[0337] 18.根据实施例1至13中任一项所述的方法,其中,所述核酸样本包括高质量核酸分子。

[0338] 19.根据实施例18所述的方法,其中,所述高质量核酸分子具有小于1的DI。

[0339] 20.根据实施例1至19中任一项所述的方法,其中,所述感兴趣人员是失踪人员。

[0340] 21.根据实施例1至19中任一项所述的方法,其中,所述感兴趣人员是灾难或冲突的受害者。

[0341] 22.根据实施例1至21中任一项所述的方法,其中,所述核酸样本源自唾液、血液、精液、毛发、牙齿、骨骼或皮肤。

[0342] 23.根据实施例22所述的方法,其中,所述核酸样本源自唾液、血液或精液。

[0343] 24.根据实施例22所述的方法,其中,所述核酸样本源自骨骼或毛发。

[0344] 25.根据实施例1至21中任一项所述的方法,其中,所述核酸样本源自口腔拭子、纸、织物,或浸渍有唾液、血液、精液或其他体液、或者包含毛发或皮肤细胞的其他基质或物体。

[0345] 26.根据实施例1至25中任一项所述的方法,其中,所述核酸样本包括在3pg至100ng之间或在约3pg至100ng之间的基因组DNA。

[0346] 27.根据实施例1至26中任一项所述的方法,其中,所述核酸样本包括在100pg至5ng之间或在约100pg至5ng之间的基因组DNA,在50pg至5ng之间或在约50pg至5ng之间的基因组DNA,或者在3pg至5ng之间或在约3pg至5ng之间的基因组DNA。

[0347] 28.根据实施例26或27所述的方法,其中,所述核酸样本包括1ng或约1ng的基因组DNA。

[0348] 29.根据实施例1至28中任一项所述的方法,其中,所述多个SNP包括亲属关系SNP(kiSNP)。

[0349] 30.根据实施例1至29中任一项所述的方法,其中,所述多个SNP包括Y染色体SNP(Y-SNP)。

[0350] 31.根据实施例1至30中任一项所述的方法,其中,所述多个SNP包括kiSNP和Y-SNP。

[0351] 32.根据实施例1至31中任一项所述的方法,其中,所述多个SNP包括kiSNP、生物地理祖先SNP(aiSNP)、身份SNP(iiSNP)、表型SNP(piSNP)、X-染色体SNP(X-SNP)和Y-染色体

SNP(Y-SNP)。

[0352] 33.根据实施例1至28中任一项所述的方法,其中,所述多个SNP包括选自由kiSNP、aiSNP、iiSNP、piSNP、X-SNP和Y-SNP组成的组中的一个或多个的SNP。

[0353] 34.根据实施例1至33中任一项所述的方法,其中,所述多个SNP中的至少或至少约80%、81%、82%、83%、84%、85%、86%、87%、88%、89%、90%、91%、92%、93%、94%、95%、96%、97%、98%或99%是亲属关系SNP。

[0354] 35.根据实施例1至34中任一项所述的方法,其中,所述参考DNA谱集包括多达5、10、15、20、25、30、35、40、45、50、55、60、65、70、75、80、85、90、95、100、125、150、175、200、300、400、500、600、700、800、900或1000个参考DNA谱。

[0355] 36.根据实施例1至35中任一项所述的方法,其中,所述参考DNA谱集中至少5%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%、65%、70%、75%、80%、85%、90%或95%的所述参考DNA谱来自所述感兴趣人员的亲属。

[0356] 37.根据实施例1至36中任一项所述的方法,其中,所述参考DNA谱集中至少5%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%、65%、70%、75%、80%、85%、90%或95%的所述参考DNA谱来自所述感兴趣人员的亲属,并且其中,所述参考DNA谱集中至少5%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%、65%、70%、75%、80%、85%、90%或95%的所述参考DNA谱中的每一个为一级亲属、二级亲属、三级亲属、四级亲属或五级亲属。

[0357] 38.根据实施例1至37中任一项所述的方法,其中,所述参考DNA谱集中至少50%的所述参考DNA谱来自所述感兴趣人员的亲属。

[0358] 39.根据实施例36至38中任一项所述的方法,其中,所述参考DNA谱集中的所述感兴趣人员的每个亲属分别是所述感兴趣人员的一级亲属、二级亲属、三级亲属、四级亲属或五级亲属。

[0359] 40.根据实施例39所述的方法,其中,所述参考DNA谱集中的所述感兴趣人员的每位亲属分别是所述感兴趣人员的一级亲属、二级亲属或三级亲属。

[0360] 41.根据实施例1至40中任一项所述的方法,其中,所述参考DNA谱集中的所述感兴趣人员的每个亲属的身份都是已知的。

[0361] 42.根据实施例1至41中任一项所述的方法,其中,所述参考DNA谱集中的所述一个或多个参考DNA谱中的每一个的身份都是已知的。

[0362] 43.根据实施例1至42中任一项所述的方法,其中,所述参考DNA谱集在数据库中。

[0363] 44.根据实施例43所述的方法,其中,所述数据库不可公开访问。

[0364] 45.根据实施例1至44中任一实施例所述的方法,其中,所述测序包括多达40重的测序复杂性。

[0365] 46.根据实施例1至44中任一项所述的方法,其中,所述测序包括多达32重的测序复杂性。

[0366] 47.根据实施例1至44中任一项所述的方法,其中,所述测序包括12重至32重的测序复杂性。

[0367] 48.根据实施例1至44中任一项所述的方法,其中,所述测序包括24重至32重的测序复杂性。

[0368] 49.根据实施例1至44中任一项所述的方法,其中,所述测序包括10重、11重、12重、13重、14重、15重、16重、17重、18重、19重、20重、21重、22重、23重、24重、25重、26重、27重、28重、29重、30重、31重、32重、33重、34重或35重的测序复杂性或者约10重、11重、12重、13重、14重、15重、16重、17重、18重、19重、20重、21重、22重、23重、24重、25重、26重、27重、28重、29重、30重、31重、32重、33重、34重或35重的测序复杂性。

[0369] 50.根据实施例1至49中任一项所述的方法,其中,所述测序包括对于死后样本为8至16重或约8至16重的测序复杂性,和/或所述测序包括对于生前样本为24至40重或约24至40重的测序复杂性。

[0370] 51.根据实施例1至50中任一项所述的方法,其中,所述测序包括对于死后样本为12重或约12重的测序复杂性,和/或所述测序包括对于生前样本为32重或约32重的测序复杂性。

[0371] 52.根据实施例1至51中任一项所述的方法,其中,所述测序包括为30重、31重或32重的测序复杂性或者约为30重、31重或32重的测序复杂性。

[0372] 53.根据实施例1至52中任一项所述的方法,还包括识别所述感兴趣人员。

[0373] 54.一种计算关联度的方法,包括:

[0374] 获得包括至少在2,000至50,000个SNP之间或在约2,000至50,000个SNP之间的基因型的DNA谱,其中,所述DNA谱来自所述感兴趣人员;以及

[0375] 计算所述DNA谱与一个或多个参考DNA谱的关系程度,其中,所述一个或多个参考DNA谱包括在参考DNA谱集内,所述参考DNA谱集包括来自所述感兴趣人员的亲属的一个或多个参考DNA谱。

[0376] 55.一种计算关联度的方法,包括:

[0377] 生成包括至少在2,000至50,000个SNP之间或在约2,000至50,000个SNP之间的基因型的DNA谱,其中,所述DNA谱来自所述感兴趣人员;以及

[0378] 计算所述DNA谱与一个或多个参考DNA谱的关系程度,其中,所述一个或多个参考DNA谱包括在参考DNA谱集内,所述参考DNA谱集包括来自所述感兴趣人员的亲属的一个或多个参考DNA谱。

[0379] 56.根据实施例1至55中任一项所述的方法,其中,所述关系程度是使用亲属关系模型来计算的。

[0380] 57.根据实施例1至56中任一项所述的方法,其中,所述关系程度是使用利用PCA方法训练的亲属关系模型来计算的。

[0381] 58.根据实施例57所述的方法,其中,用于训练所述亲属关系模型的所述PCA方法是PCA或者涉及PCA。

[0382] 59.根据实施例57或实施例58所述的方法,其中,所述PCA方法是PC-AiR。

[0383] 60.根据实施例59所述的方法,其中,所述PC-AiR包括以下步骤:(1)估计训练数据库的所有样本对(可选地训练DNA谱)之间的亲属关系系数,其中,亲属关系系数>0.025的配对被识别为相关的,而亲属关系系数<-0.025的配对被识别为祖先分化的;(2)初始化包括所有样本的无关样本集;以及(3)迭代地:(i)识别所述无关样本集中具有所述无关样本集中最相关样本的集合,从而指定为X,(ii)识别X中与所述无关样本集中的样本相比具有最少祖先分化配对的样本集,从而指定为Y;以及(iii)如果Y具有零个样本,则终止过程,或

者,如果Y具有至少一个样本,则从Y中随机选择一个样本以从U中移除,并且在步骤(3)(i)处开始重复。

[0384] 61.根据实施例57或实施例58所述的方法,其中,所述PCA方法是改进的PC-Air。

[0385] 62.根据实施例61所述的方法,其中,所述改进的PC-AiR包括以下步骤:(1)估计训练数据库的所有样本对(可选地训练DNA谱)之间的亲属关系系数,其中,亲属关系系数>0.01的配对被识别为相关的,而亲属关系系数<-0.025的配对被识别为祖先分化的;(2)移除所有具有≥5%失踪数据的DNA谱;以及(3)通过用排序值识别每个DNA谱来对所有DNA谱进行排序。在一些实施例中,所述排序值是基于完整数据库中从最少到最多排序的相关DNA谱的数量来确定的,并且通过所述完整数据库中从最多到最少排序的祖先分化DNA谱的数量来打破纽带(tie)。在一些实施例中,步骤(3)涉及迭代地遍历排序后的DNA谱,并且对于每个DNA谱:(i)如果所述DNA谱尚未出现在相关样本集中,则将其添加到所述无关样本集并且将所有相关DNA谱添加到所述相关样本集,以及(ii)如果所述DNA谱已经在所述相关样本集中,则跳至下一个DNA谱,并且在步骤(3)(i)处开始重复。

[0386] 63.根据实施例1至62中任意一项所述的方法,其中,计算所述关系程度包括使用PC-Relate来计算亲属关系系数。

[0387] 64.根据实施例63所述的方法,其中,通过将所述感兴趣人员的所述DNA谱作为PC-Relate的输入来计算所述关系程度。

[0388] 65.根据实施例57至64中任一项所述的方法,其中,通过提供亲属关系模型和所述感兴趣人员的所述DNA谱作为PC-Relate的输入来计算所述关系程度。

[0389] 66.根据实施例63至65中任一项所述的方法,其中,所述一个或多个参考DNA谱还被提供作为PC-Relate的输入。

[0390] 67.根据实施例1至66中任一项的方法,其中,计算所述关系程度包括使用全基因组亲属关系算法来计算亲属关系系数,如下所示:

[0391] $$\widehat{\phi_{ij}} = \frac{\sum_{s\in S}(g_{is} - 2\hat{u}_{is})(g_{js} - 2\hat{u}_{js})}{4\sum_{s\in S}[\hat{u}_{is}(1 - \hat{u}_{is}) \cdot \hat{u}_{js}(1 - \hat{u}_{js})]^{1/2}}$$

[0392] 其中,所述感兴趣人员和所述一个或多个参考DNA谱中的参考DNA谱是i和j,$\widehat{\phi_{ij}}$是所述亲属关系系数,$\hat{u}$是估计等位基因频率,s是在两个个体中都被分型的S个SNP中的SNP,g_{is}和g_{js}分别是在SNP s处的i和j中的参考等位基因的数量,并且$\hat{u}_{is}$和$\hat{u}_{js}$分别是由在SNP s处的i和j的PC-AiR计算出的预期等位基因频率。

[0393] 68.根据实施例1至67中任一项所述的方法,其中,计算所述关系程度包括计算似然比。

[0394] 69.根据实施例68所述的方法,其中,计算所述似然比包括比较所述DNA谱和所述一个或多个参考DNA谱之间的所述多个SNP。

[0395] 70.根据实施例68所述的方法,其中,计算所述似然比包括比较所述DNA谱和所述一个或多个参考DNA谱之间的SNP集合,所述SNP集合包括来自所述多个SNP内的亲属关系SNP。

[0396] 71.根据实施例68至70中任一项所述的方法,其中,计算所述似然比包括基于所述多个SNP的所述基因型,将所述DNA谱与所述一个或多个参考DNA谱中的参考DNA谱相关的概

率除以所述DNA谱与所述参考DNA谱不相关的概率。

[0397] 72.根据实施例68至71中任一项所述的方法,其中,所述似然比(LR)计算如下:

[0398] $$LR = \frac{P(D \mid H_r)}{P(D \mid H_u)}$$

[0399] 其中,D表示所述基因型,H_r表示个体是相关的假设,并且H_u表示个体是无关的假设。

[0400] 73.根据实施例68至71中任一项所述的方法,其中,所述LR计算如下:

[0401] $$LR = \frac{0.001}{p^2 q^2}$$

[0402] 其中,0.001表示基因分型错误率,p是等位基因1的等位基因频率,并且q是等位基因2的等位基因频率。

[0403] 74.根据实施例1至73中任一项所述的方法,其中,所述感兴趣人员在生物学上是男性,并且所述方法还包括计算所述DNA谱与所述一个或多个参考DNA谱之间共享Y染色体的似然比。

[0404] 75.根据实施例74所述的方法,其中,计算共享Y染色体的似然比包括比较所述DNA谱和所述一个或多个参考DNA谱之间的SNP集合,所述SNP集合包括一个或多个Y-SNP。

[0405] 76.根据实施例75所述的方法,其中,所述一个或多个Y-SNP包含在所述多个SNP内。

[0406] 77.根据实施例75所述的方法,其中,所述一个或多个Y-SNP包括至少20、25、30、35、40、45、50、55、60、65、70、75、80、81、82、83、84或85个Y-SNP。

[0407] 78.根据实施例75至77中任一项所述的方法,其中,所述一个或多个Y-SNP包括85个Y-SNP。

[0408] 79.根据实施例75至78中任一项所述的方法,其中,计算共享Y染色体的所述似然比包括基于所述一个或多个Y-SNP的所述基因型,将所述DNA谱与所述一个或多个参考DNA谱中的参考DNA谱共享Y染色体的概率除以所述DNA谱与所述参考DNA谱不共享Y染色体的概率。

[0409] 80.根据实施例1至79中任一项所述的方法,其中,所述参考DNA谱集中至少90%、91%、92%、93%、94%、95%、96%、97%、98%或99%的所述DNA谱来自失踪人员或灾难或冲突的受害者的亲属。

[0410] 81.根据实施例1至80中任一项所述的方法,其中,所述参考DNA谱集中的每个所述DNA谱来自失踪人员或灾难或冲突的受害者的亲属。

[0411] 82.根据实施例1至81中任一项所述的方法,其中,所述参考DNA谱集包括多达5、10、15、20、25、30、35、40、45、50、55、60、65、70、75、80、85、90、95、100、125、150、175、200、300、400、500、600、700、800、900或1000个参考DNA谱。

[0412] 83.根据实施例1至82中任一项所述的方法,其中,所述参考DNA谱集包括多达100个参考DNA谱。

[0413] 84.根据实施例1至83中任一项所述的方法,其中,所述参考DNA谱集中至少5%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%、65%、70%、75%、80%、85%、90%或95%的所述参考DNA谱来自所述感兴趣人员的亲属。

[0414] 85.根据实施例1至84中任一项所述的方法,其中,所述参考DNA谱集中至少50%的所述参考DNA谱来自所述感兴趣人员的亲属。

[0415] 86.根据实施例84或实施例85所述的方法,其中,所述参考DNA谱集中的所述感兴趣人员的每个亲属分别是所述感兴趣人员的一级亲属、二级亲属、三级亲属、四级亲属或五级亲属。

[0416] 87.根据实施例1至86中任一项所述的方法,其中,所述参考DNA谱集中的所述感兴趣人员的每个亲属的身份都是已知的。

[0417] 88.根据实施例1至87中任一项所述的方法,其中,所述DNA参考谱集中的所述感兴趣人员的每个亲属的所述身份都是已知的。

[0418] 89.根据实施例1至88中任一项所述的方法,其中,所述参考DNA谱集中的所述一个或多个参考DNA谱中的每一个的身份都是已知的。

[0419] 90.根据实施例1至89中任一项所述的方法,其中,所述参考DNA谱集在数据库中。

[0420] 91.根据实施例90所述的方法,其中,所述数据库不可公开访问。

[0421] 92.根据实施例90或实施例91所述的方法,其中,所述数据库不能被第三方遗传学服务访问。

[0422] 93.一种核酸库,所述核酸库是使用根据实施例6至92中任一项所述的方法构建的。

[0423] 94.多个引物,所述多个引物与多个靶序列特异性杂交,所述多个靶序列包括来自感兴趣人员的核酸样本中的至少在2,000至50,000个之间的单核苷酸多态性(SNP)或约2,000至50,000个之间的所述SNP,其中,在一个或多个多重PCR反应中使用所述多个引物来扩增所述核酸样本以产生扩增产物。

[0424] 95.多个引物,所述多个引物与多个靶序列特异性杂交,所述多个靶序列包括来自感兴趣人员的核酸样本和来自一个或多个参考样本的核酸样本中的至少在2,000至50,000个之间的单核苷酸多态性(SNP)或约2,000至50,000个之间的所述SNP,

[0425] 其中,所述一个或多个参考样本包括来自所述感兴趣人员的亲属的样本,以及

[0426] 其中,在一个或多个多重PCR反应中使用所述多个引物来扩增来自所述感兴趣人员的所述核酸样本和来自一个或多个参考样本的所述核酸样本以产生扩增产物。

[0427] 96.根据实施例94或实施例95所述的多个引物,其中,来自所述感兴趣人员的所述核酸样本包括基因组DNA。

[0428] 97.根据实施例94至96中任一项所述的多个引物,其中,来自所述感兴趣人员的所述核酸样本包括一种或多种酶抑制剂。

[0429] 98.根据实施例97所述的多种引物,其中,所述一种或多种酶抑制剂包括选自由血红素、亚铁血红素、腐殖酸、靛蓝、单宁酸、胶原蛋白、钙和羟基磷灰石组成的组中的一种或多种抑制剂。

[0430] 99.根据实施例94至98中任一项所述的多个引物,其中,来自所述感兴趣人员的所述核酸样本包括低质量核酸分子和/或低数量核酸分子。

[0431] 100.根据实施例99所述的多种引物,其中,所述低质量核酸分子是退化的基因组DNA和/或片段化的基因组DNA。

[0432] 101.根据实施例99或实施例100所述的多个引物,其中,所述低质量核酸分子具有

为1、2、3、4、5、6、7、8、9、10、15、20、25、30、35、40、45、50、55、60、65、70、75、80、85、90、95、100、105、110、115、120、125、130、135、140、145、150、155、160、165、170、175、180、185、190、195或200的退化指数(DI)或至少为1、2、3、4、5、6、7、8、9、10、15、20、25、30、35、40、45、50、55、60、65、70、75、80、85、90、95、100、105、110、115、120、125、130、135、140、145、150、155、160、165、170、175、180、185、190、195或200的退化指数(DI)。

[0433] 102.根据实施例99至101中任一项所述的多个引物,其中,所述低质量核酸分子具有至少1且多达或小于158.3的DI。

[0434] 103.根据实施例94至102中任一项的所述的多个引物,其中,来自所述感兴趣人员的所述核酸样本和/或来自一个或多个参考样本的所述核酸样本包括高质量核酸分子。

[0435] 104.根据实施例103所述的多个引物,其中,所述高质量核酸分子具有小于1的DI。

[0436] 105.根据实施例94至104中任一项所述的多个引物,其中,所述感兴趣人员是失踪人员。

[0437] 106.根据实施例94至104中任一项所述的多个引物,其中,所述感兴趣人员是灾难或冲突的受害者。

[0438] 107.根据实施例94至106中任一项所述的多个引物,其中,来自所述感兴趣人员的所述核酸样本源自口腔拭子、纸、织物,或浸渍有唾液、血液或其他体液、或者包含毛发或皮肤细胞的其他基质或物体。

[0439] 108.根据实施例94至107中任一项所述的多个引物,其中,来自所述感兴趣人员的所述核酸样本包括在3pg至100ng之间或在约3pg至100ng之间的基因组DNA。

[0440] 109.根据实施例94至108中任一项所述的多个引物,其中,来自所述感兴趣人员的所述核酸样本包括在100pg至5ng之间或在约100pg至5ng之间的基因组DNA,在50pg至5ng之间或在约50pg至5ng之间的基因组DNA,或者在3pg至5ng之间或在约3pg至5ng之间的基因组DNA。

[0441] 110.根据实施例108或实施例109所述的多个引物,其中,来自所述感兴趣人员的所述核酸样本包括1ng或约1ng的基因组DNA。

[0442] 111.根据实施例94至110中任一项所述的多个引物,其中,所述多个SNP包括亲属关系SNP(kiSNP)。

[0443] 112.根据实施例94至111中任一项所述的方法,其中,所述多个SNP包括Y染色体SNP(Y-SNP)。

[0444] 113.根据实施例94至112中任一项所述的方法,其中,所述多个SNP包括kiSNP和Y-SNP。

[0445] 114.根据实施例94至113中任一项所述的多个引物,其中,所述多个SNP包括kiSNP、生物地理祖先SNP(aiSNP)、身份SNP(iiSNP)、表型SNP(piSNP)、X-染色体SNP(X-SNP)和Y-染色体SNP(Y-SNP)。

[0446] 115.根据实施例94至111中任一项所述的多个引物,其中,所述多个SNP包括选自由kiSNP、aiSNP、iiSNP、piSNP、X-SNP和Y-SNP组成的组中的一个或多个的SNP。

[0447] 116.根据实施例94至115中任一项所述的多个引物,其中,所述多个SNP中的至少或至少约80％、81％、82％、83％、84％、85％、86％、87％、88％、89％、90％、91％、92％、93％、94％、95％、96％、97％、98％或99％是亲属关系SNP。

[0448] 117.根据实施例94至116中任一项所述的多个引物,其中,所述参考DNA谱集中至少90%、91%、92%、93%、94%、95%、96%、97%、98%或99%的所述DNA谱来自失踪人员或灾难或冲突的受害者的亲属。

[0449] 118.根据实施例94至116中任一项所述的多个引物,其中,所述一个或多个参考样本中的每一个来自失踪人员或灾难或冲突的受害者的亲属。

[0450] 119.根据实施例94至118中任一项所述的多种引物,其中,所述一个或多个参考样本包括多达5、10、15、20、25、30、35、40、45、50、55、60、65、70、75、80、85、90、95、100、125、150、175、200、300、400、500、600、700、800、900或1000个参考样本。

[0451] 120.根据实施例94至119中任一项所述的多个引物,其中,所述一个或多个参考样本中的至少5%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%、65%、70%、75%、80%、85%、90%或95%来自所述感兴趣人员的亲属。

[0452] 121.根据实施例94至120中任一项所述的多个引物,其中,所述参考DNA谱集中至少5%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%、65%、70%、75%、80%、85%、90%或95%的所述参考DNA谱来自所述感兴趣人员的亲属,并且其中,所述参考DNA谱集中至少5%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%、65%、70%、75%、80%、85%、90%或95%的所述参考DNA谱中的每一个为一级亲属、二级亲属、三级亲属、四级亲属或五级亲属。

[0453] 122.根据实施例94至121中任一项所述的多个引物,其中,至少50%的所述一个或多个参考样本来自所述感兴趣人员的亲属。

[0454] 123.根据实施例120至122中任一项所述的多个引物,其中,所述一个或多个参考样本中的所述感兴趣人员的每个亲属分别是所述感兴趣人员的一级亲属、二级亲属、三级亲属、四级亲属或五级亲属。

[0455] 124.根据实施例123所述的多个引物,其中,所述一个或多个参考样本中的所述感兴趣人员的每个亲属分别是所述感兴趣人员的一级亲属、二级亲属或三级亲属。

[0456] 125.根据实施例94至124中任一项所述的多个引物,其中,所述一个或多个参考样本中的所述感兴趣人员的每个亲属的身份是已知的。

[0457] 126.根据实施例94至125中任一项所述的多个引物,其中,所述一个或多个参考样本中的每一个的身份是已知的。

[0458] 127.一种构建DNA谱的方法,包括:

[0459] 提供来自感兴趣人员的核酸样本,

[0460] 使用与多个靶序列特异性杂交的多个引物来扩增所述核酸样本,所述多个靶序列共同包括多个至少在2,000至50,000个之间或约2,000至50,000个之间的单核苷酸多态性(SNP),从而生成扩增产物,其中,所述扩增在一个或多个多重PCR反应中进行,

[0461] 对所述扩增产物进行测序,

[0462] 确定所述多个SNP的基因型,从而生成DNA谱。

[0463] 128.一种构建DNA谱的方法,包括:

[0464] 提供来自感兴趣人员的核酸样本,

[0465] 提供来自所述感兴趣人员的亲属的核酸样本,

[0466] 使用与多个靶序列特异性杂交的多个引物来扩增来自所述感兴趣人员的所述核

酸样本和来自所述亲属的所述核酸样本,所述多个靶序列共同包括多个至少在2,000至50,000个之间或约2,000至50,000个之间的单核苷酸多态性(SNP),从而生成扩增产物,其中,所述扩增在一个或多个多重PCR反应中进行,

[0467] 对所述扩增产物进行测序,

[0468] 确定所述多个SNP的基因型,从而生成所述感兴趣人员和所述感兴趣人员的所述亲属的DNA谱。

[0469] 129.根据实施例127或实施例128所述的方法,其中,所述测序不包括全基因组测序(WGS)。

[0470] 130.根据实施例127或实施例129所述的方法,其中,所述核酸样本包括基因组DNA。

[0471] 131.根据实施例128或实施例129所述的方法,其中,所述感兴趣人员的所述核酸样本和/或所述感兴趣人员的所述亲属的所述核酸样本包括基因组DNA。

[0472] 132.根据实施例127至131中任一项所述的方法,其中,所述核酸样本、所述感兴趣人员的所述核酸样本和/或所述亲属的所述核酸样本包括一种或多种酶抑制剂。

[0473] 133.根据实施例132所述的方法,其中,所述一种或多种酶抑制剂包括选自由血红素、亚铁血红素、腐殖酸、靛蓝、单宁酸、胶原蛋白、钙和羟基磷灰石组成的组中的一种或多种抑制剂。

[0474] 134.根据实施例127至133中任一项所述的方法,其中,所述核酸样本、所述感兴趣人员的所述核酸样本和/或所述亲属的所述核酸样本包括低质量核酸分子和/或低数量核酸分子。

[0475] 135.根据实施例134所述的方法,其中,所述低质量核酸分子是退化的基因组DNA和/或片段化的基因组DNA。

[0476] 136.根据实施例134或实施例135所述的方法,其中,所述低质量核酸分子具有为1、2、3、4、5、6、7、8、9、10、15、20、25、30、35、40、45、50、55、60、65、70、75、80、85、90、95、100、105、110、115、120、125、130、135、140、145、150、155、160、165、170、175、180、185、190、195或200的退化指数(DI)或至少为1、2、3、4、5、6、7、8、9、10、15、20、25、30、35、40、45、50、55、60、65、70、75、80、85、90、95、100、105、110、115、120、125、130、135、140、145、150、155、160、165、170、175、180、185、190、195或200的退化指数(DI)。

[0477] 137.根据实施例134至136中任一项所述的方法,其中,所述低质量核酸分子具有至少1且多达或小于158.3的DI。

[0478] 138.根据实施例127至133中任一项所述的方法,其中,所述核酸样本、所述感兴趣人员的所述核酸样本和/或所述亲属的所述核酸样本包括高质量核酸分子。

[0479] 139.根据实施例138所述的方法,其中,所述高质量核酸分子具有小于1的DI。

[0480] 140.根据实施例127至139中任一项所述的方法,其中,所述感兴趣人员是失踪人员。

[0481] 141.根据实施例127至139中任一项所述的方法,其中,所述感兴趣人员是灾难或冲突的受害者。

[0482] 142.根据实施例128、129和131至141中任一项所述的方法,其中,所述感兴趣人员的所述亲属是一级、二级、三级、四级或五级亲属。

[0483] 143.根据实施例128、129和131-141中任一项所述的方法,其中,所述感兴趣人员的所述亲属是一级、二级或三级亲属。

[0484] 144.根据实施例127至143中任一项所述的方法,其中,所述核酸样本、所述感兴趣人员的所述核酸样本和/或所述亲属的所述核酸样本源自口腔拭子、纸、织物,或浸渍有唾液、血液或其他体液、或者包含毛发或皮肤细胞的其他基质或物体。

[0485] 145.根据实施例127至144中任一项所述的方法,其中,所述核酸样本、所述感兴趣人员的所述核酸样本和/或所述亲属的所述核酸样本包括在3pg至100ng之间或在约3pg至100ng之间的基因组DNA。

[0486] 146.根据实施例127至145中任一项所述的方法,其中,所述核酸样本、所述感兴趣人员的所述核酸样本和/或所述亲属的所述核酸样本包括在100pg至5ng之间或在约100pg至5ng之间的基因组DNA,在50pg至5ng之间或在约50pg至5ng之间的基因组DNA,或者在3pg至5ng之间或在约3pg至5ng之间的基因组DNA。

[0487] 147.根据实施例145或实施例146所述的方法,其中,所述核酸样本、所述感兴趣人员的核酸样本和/或所述亲属的所述核酸样本包括1ng或约1ng的基因组DNA。

[0488] 148.根据实施例127至147中任一项所述的方法,其中,所述多个SNP包括亲属关系SNP。

[0489] 149.根据实施例127至148中任一项所述的方法,其中,所述多个SNP包括Y染色体SNP(Y-SNP)。

[0490] 150.根据实施例127至149中任一项所述的方法,其中,所述多个SNP包括kiSNP和Y-SNP。

[0491] 151.根据实施例127至150中任一项所述的方法,其中,所述多个SNP包括kiSNP、生物地理祖先SNP(aiSNP)、身份SNP(iiSNP)、表型SNP(piSNP)、X-染色体SNP(X-SNP)和Y-染色体SNP(Y-SNP)。

[0492] 152.根据实施例127至151中任一项所述的方法,其中,所述多个SNP包括选自由kiSNP、aiSNP、iiSNP、piSNP、X-SNP和Y-SNP组成的组中的一个或多个的SNP。

[0493] 153.根据实施例127至152中任一项所述的方法,其中,所述多个SNP中的至少或至少约80%、81%、82%、83%、84%、85%、86%、87%、88%、89%、90%、91%、92%、93%、94%、95%、96%、97%、98%或99%是亲属关系SNP。

[0494] 154.根据实施例1至92和127至153中任一项所述的方法,其中,所述测序包括多达40重的测序复杂性。

[0495] 155.根据实施例1至92和127至153中任一项所述的方法,其中,所述测序包括多达32重的测序复杂性。

[0496] 156.根据实施例1至92和127-至53中任一项所述的方法,其中,所述测序包括12重至32重的测序复杂性。

[0497] 157.根据实施例1至92和127至153中任一项所述的方法,其中,所述测序包括24重至32重的测序复杂性。

[0498] 158.实施例1-92和127-153中任一项的方法,其中:(a)所述测序包括4重、5重、6重、7重、8重、9重、10重、11重、12重、13重、14重、15重、16重、17重、18重、19重、20重、21重、22重、23重、24重、25重、26重、27重、28重、29重、30重、31重、32重、33重、34重、35重、36重、37

重、38重、39重、40重、41重、42重、43重、44重、45重的测序复杂性或者约4重、5重、6重、7重、8重、9重、10重、11重、12重、13重、14重、15重、16重、17重、18重、19重、20重、21重、22重、23重、24重、25重、26重、27重、28重、29重、30重、31重、32重、33重、34重、35重、36重、37重、38重、39重、40重、41重、42重、43重、44重、45重的测序复杂性;或者(b)所述测序包括10重、11重、12重、13重、14重、15重、16重、17重、18重、19重、20重、21重、22重、23重、24重、25重、26重、27重、28重、29重、30重、31重、32重、33重、34重或35重的测序复杂性或者约10重、11重、12重、13重、14重、15重、16重、17重、18重、19重、20重、21重、22重、23重、24重、25重、26重、27重、28重、29重、30重、31重、32重、33重、34重或35重的测序复杂性。

[0499] 159.根据实施例1至92和127至158中任一项所述的方法,其中,所述测序包括对于死后样本为8至16重或约8至16重的测序复杂性,和/或所述测序包括对于生前样本为24至40重或约24至40重的测序复杂性。

[0500] 160.根据实施例1至92和127至158中任一项所述的方法,其中,所述测序包括对于死后样本为12重或约12重的测序复杂性,和/或所述测序包括对于生前样本为32重或约32重的测序复杂性。

[0501] 161.根据实施例1至92和127至153中任一项所述的方法,其中,所述测序包括为30重、31重或32重的测序复杂性或者约为30重、31重或32重的测序复杂性。

[0502] 162.一种识别DNA谱的遗传亲属的方法,包括:

[0503] 计算根据实施例127至161中任一项所述的DNA谱与一个或多个参考DNA谱的关系程度,其中,所述一个或多个参考DNA谱包括在参考DNA谱集内,所述参考DNA谱集包括来自所述感兴趣人员的亲属的一个或多个参考DNA谱;以及

[0504] 生成家谱,所述家谱包括与所述一个或多个参考DNA谱相关的所述DNA谱。

[0505] 163.根据实施例162所述的方法,其中,所述一个或多个参考DNA谱是数据库的一部分。

[0506] 164.根据实施例162或实施例163所述的方法,其中,所述参考DNA谱集包括多达5、10、15、20、25、30、35、40、45、50、55、60、65、70、75、80、85、90、95、100、125、150、175、200、300、400、500、600、700、800、900或1000个参考DNA谱。

[0507] 165.根据实施例162至164中任一项所述的方法,其中,在所述参考DNA谱集中的至少5%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%、65%、70%、75%、80%、85%、90%或95%的所述参考DNA谱来自所述感兴趣人员的亲属。

[0508] 166.根据实施例162至165中任一项所述的方法,其中,在所述参考DNA谱集中至少50%的所述参考DNA谱来自所述感兴趣人员的亲属。

[0509] 167.根据实施例165或实施例166所述的方法,其中,所述感兴趣人员的每个亲属分别是所述感兴趣人员的一级亲属、二级亲属、三级亲属、四级亲属或五级亲属。

[0510] 168.根据实施例162至167中任一项所述的方法,其中,在所述参考DNA谱集中的所述感兴趣人员的每个亲属的身份都是已知的。

[0511] 169.根据实施例162至168中任一项所述的方法,其中,在所述参考DNA谱集中的所述一个或多个参考DNA谱中的每一个的身份都是已知的。

[0512] 170.根据实施例162至169中任一项所述的方法,其中,所述参考DNA谱集在数据库中。

[0513] 171.根据实施例170所述的方法,其中,所述数据库不可公开访问。

[0514] 172.根据实施例170或实施例171所述的方法,其中,所述数据库不能被第三方遗传学服务访问。

[0515] 173.一种识别DNA谱身份的方法,包括:

[0516] 计算包括至少在2,000至50,000个SNP之间或在约2,000至50,000个SNP之间的基因型的DNA谱与一个或多个参考DNA谱的关系程度,其中,所述DNA谱来自感兴趣人员,并且其中,所述一个或多个参考DNA谱包括在参考DNA谱集内,所述参考DNA谱集包括来自所述感兴趣人员的亲属的一个或多个参考DNA谱;以及

[0517] 生成家谱,所述家谱包括与所述一个或多个参考DNA谱相关的所述DNA谱。

[0518] 174.根据实施例173所述的方法,其中,所述DNA谱是通过根据实施例127至161中任一项所述的方法生成的。

[0519] 175.根据实施例173或实施例174所述的方法,其中,所述关系程度是使用亲属关系模型来计算的。

[0520] 176.根据实施例173至175中任一项所述的方法,其中,所述关系程度是使用利用PCA方法训练的亲属关系模型来计算的。

[0521] 177.根据实施例176所述的方法,其中,用于训练所述亲属关系模型的所述PCA方法是PCA或涉及PCA。

[0522] 178.根据实施例176或实施例177所述的方法,其中,所述PCA方法是PC-AiR。

[0523] 179.根据实施例178所述的方法,其中,所述PC-AiR包括以下步骤:(1)估计训练数据库的所有样本对——可选地训练DNA谱——之间的亲属关系系数,其中,亲属关系系数>0.025的配对被识别为相关的,而亲属关系系数<-0.025的配对被识别为祖先分化的;(2)初始化包括所有样本的无关样本集;以及(3)迭代地:(i)识别所述无关样本集中具有所述无关样本集中最相关样本的集合,从而指定为X,(ii)识别X中与所述无关样本集中的样本相比具有最少祖先分化配对的样本集,从而指定为Y;以及(iii)如果Y具有零个样本,则终止过程,或者,如果Y具有至少一个样本,则从Y中随机选择一个样本以从U中移除,并且在步骤(3)(i)处开始重复。

[0524] 180.根据实施例176或实施例177所述的方法,其中,所述PCA方法是改进的PC-Air。

[0525] 181.根据实施例180所述的方法,其中,所述改进的PC-AiR包括以下步骤:(1)估计训练数据库的所有样本对——可选地训练DNA谱——之间的亲属关系系数,其中,亲属关系系数>0.01的配对被识别为相关的,而亲属关系系数<-0.025的配对被识别为祖先分化的;(2)移除所有具有≥5%失踪数据的DNA谱;以及(3)通过用排序值识别每个DNA谱来对所有DNA谱进行排序。在一些实施例中,所述排序值是基于完整数据库中从最少到最多排序的相关DNA谱的数量来确定的,并且通过所述完整数据库中从最多到最少排序的祖先分化DNA谱的数量来打破纽带。在一些实施例中,步骤(3)涉及迭代地遍历排序后的DNA谱,并且对于每个DNA谱:(i)如果所述DNA谱尚未出现在相关样本集中,则将其添加到所述无关样本集并且将所有相关DNA谱添加到所述相关样本集,以及(ii)如果所述DNA谱已经在所述相关样本集中,则跳至下一个DNA谱,并且在步骤(3)(i)处开始重复。

[0526] 182.根据实施例173至181中任一项所述的方法,其中,计算所述关系程度包括使

用PC-Relate来计算亲属关系系数。

[0527] 183.根据实施例182所述的方法,其中,通过提供所述感兴趣人员的所述DNA谱作为PC-Relate的输入来计算所述关系程度。

[0528] 184.根据实施例182或实施例183所述的方法,其中,通过提供所述亲属关系模型和所述感兴趣人员的所述DNA谱作为PC-Relate的输入来计算所述关系程度。

[0529] 185.根据实施例182至184中任一项所述的方法,其中,还提供所述一个或多个参考DNA谱作为PC-Relate的输入。

[0530] 186.根据实施例173至185中任一项所述的方法,其中,计算所述关系程度包括使用全基因组亲属关系算法来计算亲属关系系数,如下所示:

[0531] $$\widehat{\phi_{ij}} = \frac{\sum_{s\in S}(g_{is} - 2\hat{u}_{is})(g_{js} - 2\hat{u}_{js})}{4\sum_{s\in S}[\hat{u}_{is}(1 - \hat{u}_{is}) \cdot \hat{u}_{js}(1 - \hat{u}_{js})]^{1/2}}$$

[0532] 其中,所述感兴趣人员和所述一个或多个参考DNA谱中的参考DNA谱是i和j,$\widehat{\phi_{ij}}$是所述亲属关系系数,$\hat{u}$是估计等位基因频率,s是在两个个体中都被分型的S个SNP中的SNP,g_{is}和g_{js}分别是在SNP s处的i和j中的参考等位基因的数量,并且$\hat{u}_{is}$和$\hat{u}_{js}$分别是由在SNP s处的i和j的PC-AiR计算出的预期等位基因频率。

[0533] 187.根据实施例173至186中任一项所述的方法,其中,计算所述关系程度包括计算似然比。

[0534] 188.根据实施例187所述的方法,其中,计算所述似然比包括比较所述DNA谱和所述一个或多个参考DNA谱之间的所述多个SNP。

[0535] 189.根据实施例187所述的方法,其中,计算所述似然比包括比较所述DNA谱和所述一个或多个参考DNA谱之间的SNP集合,所述SNP集合包括来自所述多个SNP内的亲属关系SNP。

[0536] 190.根据实施例187至189中任一项所述的方法,其中,计算所述似然比包括基于所述多个SNP的所述基因型,将所述DNA谱与所述一个或多个参考DNA谱中的参考DNA谱相关的概率除以所述DNA谱与所述参考DNA谱不相关的概率。

[0537] 191.根据实施例187至190中任一项所述的方法,其中,所述似然比(LR)计算如下:

[0538] $$LR = \frac{P(D \mid H_r)}{P(D \mid H_u)}$$

[0539] 其中,D表示所述基因型,H_r表示个体是相关的假设,并且H_u表示个体是无关的假设。

[0540] 192.根据实施例187至190中任一项所述的方法,其中,所述LR计算如下:

[0541] $$LR = \frac{0.001}{p^2q^2}$$

[0542] 其中,0.001表示基因分型错误率,p是等位基因1的等位基因频率,并且q是等位基因2的等位基因频率。

[0543] 193.根据实施例173至192中任一项所述的方法,其中,所述感兴趣人员在生物学上是男性,并且所述方法还包括计算所述DNA谱与所述一个或多个参考DNA谱之间共享Y染色体的似然比。

[0544] 194.根据实施例193所述的方法,其中,计算共享Y染色体的似然比包括比较所述DNA谱和所述一个或多个参考DNA谱之间的SNP集合,所述SNP集合包括一个或多个Y-SNP。

[0545] 195.根据实施例194所述的方法,其中,所述一个或多个Y-SNP包含在所述多个SNP内。

[0546] 196.根据实施例194或实施例195所述的方法,其中,所述一个或多个Y-SNP包括至少20、25、30、35、40、45、50、55、60、65、70、75、80、81、82、83、84或85个Y-SNP。

[0547] 197.根据实施例194至196中任一项所述的方法,其中,所述一个或多个Y-SNP包括85个Y-SNP。

[0548] 198.根据实施例193至197中任一项所述的方法,其中,计算共享Y染色体的所述似然比包括基于所述一个或多个Y-SNP的所述基因型,将所述DNA谱与所述一个或多个参考DNA谱中的参考DNA谱共享Y染色体的概率除以所述DNA谱与所述参考DNA谱不共享Y染色体的概率。

[0549] 199.根据实施例173至198中任一项所述的方法,其中,所述参考DNA谱集中至少90%、91%、92%、93%、94%、95%、96%、97%、98%或99%的所述DNA谱来自失踪人员或灾难或冲突的受害者的亲属。

[0550] 200.根据实施例173至199中任一项所述的方法,其中,所述一个或多个参考样本中的每一个来自失踪人员或灾难或冲突的受害者的亲属。

[0551] 201.根据实施例173-200中任一项所述的方法,其中,所述一个或多个参考样本包括多达5、10、15、20、25、30、35、40、45、50、55、60、65、70、75、80、85、90、95、100、125、150、175、200、300、400、500、600、700、800、900或1000个参考样本。

[0552] 202.根据实施例173至201中任一项所述的方法,其中,所述一个或多个参考样本的至少5%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%、65%、70%、75%、80%、85%、90%或95%来自所述感兴趣人员的亲属。

[0553] 203.根据实施例173至202中任一项所述的方法,其中,所述参考DNA谱集中至少5%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%、65%、70%、75%、80%、85%、90%或95%的所述参考DNA谱来自所述感兴趣人员的亲属,并且其中,所述参考DNA谱集中至少5%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%、65%、70%、75%、80%、85%、90%或95%的所述参考DNA谱中的每一个为一级亲属、二级亲属、三级亲属、四级亲属或五级亲属。

[0554] 204.根据实施例173至203中任一项所述的方法,其中,所述一个或多个参考样本的至少50%来自所述感兴趣人员的亲属。

[0555] 205.根据实施例199至204中任一项所述的方法,其中,所述一个或多个参考样本中的所述感兴趣人员的每个亲属分别是所述感兴趣人员的一级亲属、二级亲属、三级亲属、四级亲属或五级亲属。

[0556] 206.根据实施例205所述的方法,其中,所述一个或多个参考样本中的所述感兴趣人员的每个亲属分别是所述感兴趣人员的一级亲属、二级亲属或三级亲属。

[0557] 207.根据实施例173至206中任一项所述的方法,其中,所述一个或多个参考样本中的所述感兴趣人员的每个亲属的身份是已知的。

[0558] 208.根据实施例173至207中任一项所述的方法,其中,所述一个或多个参考样本

中的每一个的所述身份都是已知的。

[0559] 209.一种试剂盒,包括至少一个容器装置,其中,所述至少一个容器装置包括根据实施例94至126中任一项所述的多个引物。

[0560] 210.根据实施例1至92和127至208中任一项所述的方法,其中,所述多个SNP包括介于或介于约2,000至11,000个SNP、3,000至11,000个SNP、4,000至11,000个SNP、5,000至11,000个SNP、5,500至11,000个SNP、6,000至11,000个SNP、7,000至15,000个SNP、7,000至14,000个SNP、7,000至13,000个SNP、7,000至12,000个SNP、7,000至11,000个SNP、8,000至15,000个SNP、8,000至14,000个SNP、8,000至13,000个SNP、8,000至12,000个SNP、8,000至11,000个SNP、9,000至15,000个SNP、9,000至14,000个SNP、9,000至13,000个SNP、9,000至12,000个SNP或9,000至11,000个SNP。

[0561] 211.根据实施例1至92和127至208中任一项所述的方法,其中,所述多个SNP包括10,230个SNP。

[0562] 212.根据实施例94至126中任一项所述的多个引物,其中,所述多个SNP包括介于或介于约2,000至11,000个SNP、3,000至11,000个SNP、4,000至11,000个SNP、5,000至11,000个SNP、5,500至11,000个SNP、6,000至11,000个SNP、7,000至15,000个SNP、7,000至14,000个SNP、7,000至13,000个SNP、7,000至12,000个SNP、7,000至11,000个SNP、8,000至15,000个SNP、8,000至14,000个SNP、8,000至13,000个SNP、8,000至12,000个SNP、8,000至11,000个SNP、9,000至15,000个SNP、9,000至14,000个SNP、9,000至13,000个SNP、9,000至12,000个SNP或9,000至11,000个SNP。

[0563] 213.根据实施例94至126中任一项所述的多个引物,其中,所述多个SNP包括10,230个SNP。

[0564] 214.根据实施例1至5、11至93、127至161和210至213中任一项所述的方法,还包括生成家谱,所述家谱包括与所述参考DNA谱集中包含的一个或多个DNA谱相关的所述DNA谱。

[0565] 215.根据实施例214所述的方法,其中,所述家谱包括与来自所述感兴趣人员的亲属的一个或多个DNA谱相关的所述DNA谱。

[0566] 示例

[0567] 下列示例仅用于说明目的,并不旨在限制本发明的范围。

[0568] 示例1:序列库的生成和灵敏度的确定

[0569] 本实施例描述了一种确定本文所述多重聚合酶链式反应的灵敏度以生成能够被测序的库的方法。图1描绘了本示例中所述的生成能够被测序的库的方法的示例性示意图。

[0570] A.基因组DNA靶的PCR扩增

[0571] 进行多重聚合酶链式反应以扩增基因组DNA样本中的10,230个个体扩增子。每个引物对都被设计为选择性杂交并促进基因组DNA样本的特定单核苷酸多态性(SNP)的扩增。测试的输入基因组DNA范围从50ng到50pg,更具体地,5ng、2.5ng、1ng、500pg、250pg、100pg和50pg)。简而言之,将含有足够缓冲液、dNTP、MgC12、盐和PCR添加剂(诸如甘油)的18.5ml PCR主混合物添加到96孔PCR板的单个孔中。还添加了5微升引物池(其包含10,530个引物对)、2-4单位DNA聚合酶(诸如Phusion热启动DNA聚合酶(Thermo Fisher,cat#F549L或任何其他热稳定DNA聚合酶)),还添加50ng至50pg的基因组DNA。

[0572] 将PCR板密封并装入热循环仪(Veriti 96孔热循环仪,Thermo Fisher

Scientific,4413964)并按照下面描述的温度曲线运行以生成扩增子库。

[0573] 98℃持续3分钟

[0574] 18个周期:

[0575] 96℃,持续45秒

[0576] 80℃,持续10秒

[0577] 采用适用的升温模式,54℃持续4分钟

[0578] 采用适用的升温模式,66℃持续90秒

[0579] 68℃,持续10分钟

[0580] 保持在4℃

[0581] 循环后,扩增子库保持在2-8℃直至进行到下面概述的纯化步骤为止。

[0582] B.从输入DNA和引物中纯化扩增子

[0583] 使用MagBind Total Pure NGS珠子(Omega Biotek,M1378-02)以1.6X和0.6X体积比结合、清洗和洗脱进行两轮清理,发现可以去除基因组DNA和未结合或过量的引物。本文概述的扩增和纯化步骤产生长度约为150-350bp的扩增子。然后,纯化的扩增子用于第二轮PCR以添加用于测序的衔接物(adapter)。

[0584] C.富集纯化的扩增子以生成能够被测序的库

[0585] 第二轮PCR扩增是通过将来自上述步骤的25ml纯化扩增子与在Forenseq Kintelligence试剂盒(Verogen PN:V16000120)中提供的5ml衔接物和在Forenseq Kintelligence试剂盒(VerogenPN:V160000120)中提供的20ml KPCR主混合物在96孔PCR板中混合来进行的。将PCR板密封并装入热循环仪(Veriti 96孔热循环仪,Thermo Fisher Scientific,4413964)中,并且按照下面描述的温度曲线运行以生成扩增子库。

[0586] 98℃,30秒

[0587] 15个周期:

[0588] 98℃,持续20秒

[0589] 66℃,持续30秒

[0590] 72℃,持续30秒

[0591] 72℃,持续1分钟

[0592] 保持在4℃

[0593] 使用MagBind Total Pure NGS珠子(Omega Biotek,M1378-02)以1X结合、清洗和洗脱来纯化库。根据在Forenseq Kintelligence试剂盒用户指南(Verogen PN:V16000120,其内容在此通过引用整体并入)中的说明对纯化的库进行定量、归一化、变性和稀释。

[0594] 变性库按照MiSeq FGx测序系统参考指南(文档#VD2018006,其内容通过引用整体并入本文)上的说明进行测序。如图2所示,跨一系列输入基因组DNA滴定中检测到的基因座数量相似。

[0595] 使用Forenseq通用分析软件2.1(Verogen,加利福尼亚州圣地亚哥)来分析结果,其遵循Forenseq通用分析软件2.1中概述的说明并在参考指南文档#VD2019002中提供,其内容在此通过引用整体并入本文。

[0596] 示例2:使用退化的DNA来生成序列库

[0597] 本示例描述了从低量和高度退化的样本中对DNA进行测序。从Innogenomics(新奥

尔良,路易斯安那州)获得了退化的血液DNA的退化的DNA A系列。如示例1所述,DNA样本用于生成测序库,除了本示例中使用了10,327个基因座的引物对。图3中显示了使用本文所述测定法检测到的具有退化DNA的基因座的百分比(调用率)与微阵列(GSA)调用率的比较。在x轴上显示退化指数(DI),并且在Y轴上显示检测到的基因座数量。这些结果表明,即使退化程度较高的DNA(DI为158.3),该测定法也能检测到9167个基因座,其足以上传到遗传学数据库以搜索亲属。诸如微阵列的替代技术无法在退化指数高的样本中检测到任何基因座。

[0598] 示例3:抑制剂对库制备的活性评估

[0599] 本示例描述了PCR抑制剂对本文公开的库制备的影响的评估。来自犯罪现场的DNA样本通常含有抑制PCR的共纯化杂质。当存在足够的DNA副本时,PCR抑制是PCR失败的最常见原因。腐殖质化合物(腐烂过程中产生的一系列物质),被认为是土壤、天然水和近期沉积物中污染DNA的物质。其他常见抑制剂包括血红素(来自血液)、靛蓝(来自牛仔裤)和单宁酸。

[0600] 为了评估法医样本中常见的抑制剂的影响,按照示例1中所述进行库制备,但在上述“扩增和标记靶”步骤中加入了200uM血红素、50ng/uL腐殖酸、133uM靛蓝和16uM单宁酸,并使用了10380个基因座的引物对。结果显示在图4中,其中不含任何抑制剂的PCR反应被标记为对照。

[0601] 示例4:关系程度的确定

[0602] 本示例描述了通常按照上述示例1所述制备的样本的示例性结果。

[0603] 使用Utah CEPH家族1463DNA(Coriell Institute)样本中的200ng样本来运行Illumina Global Screening Array(GSA)2.0。将SNP调用上传到GEDmatch数据库(Verogen)。在图5中显示了示例性家谱。样本之一,NA12889(祖父)按照示例1中所述的库制备方案在ForenSeq UAS2.1模块上运行。生成的报告上传到数据库,并使用1:many搜索关系工具进行搜索。将根据数据库中算法的亲属关系系数与预期亲属关系系数进行比较。预期亲属关系系数和观察到的亲属关系系数如图6所示。

[0604] 示例5:示例性案例研究中的亲属关系系数确定

[0605] 本示例描述了使用样本SNP谱确定亲属关系系数的示例性案例研究的结果。使用GEDmatch数据库中的10个已建立的谱系(具有12-28个家庭成员)测试了1:many搜索算法检测潜在亲属的能力。本文公开的来自测定法的样本SNP谱被认为是X先生=POI(感兴趣人员/未知犯罪现场谱)。图7显示了候选查询结果(hit)、亲属关系系数和亲属状态。

[0606] 然后使用根据搜索算法生成的结果为X先生生成家谱,如图8所示。如家谱中所示,在前11个候选查询结果中返回了X先生的表兄弟姐妹(1C)和曾祖父(G GF),其是三级关系。在前15个候选查询结果中返回了X先生的曾曾祖母(GG GM)、曾曾叔父(GG uncle)和隔一代的堂/表亲(1C1R),其是四级关系。X先生的第二代的堂/表亲(2C),其是五级关系并且是第12个候选查询结果。

[0607] 示例6:序列库的生成和灵敏度的确定,包括按基因座类型进行评估

[0608] 本示例涉及一种确定本文描述的多重聚合酶链式反应的灵敏度的方法以生成能够被测序的库,并且包括由基因座类型进行的评估。

[0609] 序列库(测序的核酸库),也称为DNA谱,以与示例1中描述的方式相同的方式生成,区别是使用Forenseq通用分析软件版本2.2来分析结果。

[0610] 如图9所示,该表总结了基于每种不同类型基因座的输入DNA量(ng)检测到的基因座数量(三次重复的平均值),例如,Y染色体SNP(Y-SNP)、X染色体SNP(X-SNP)、表型SNP(piSNP)、亲属关系SNP(kiSNP)、身份SNP(iiSNP)和生物地理祖先SNP(aiSNP),总共分析了10,230个基因座。测试的基因组DNA输入滴定包括5ng、2.5ng、1ng、0.5ng(500pg)、0.25ng(250pg)、0.10ng(100pg)、0.05ng(50pg)的输入基因组DNA。如图9所示,检测出的总SNPs中,每个输入DNA的量在0.05ng至5ng之间时,至少98.9%(10,117)的基因座被检测到,而输入DNA量为0.10ng及以上时,至少99.5%(10,179)的基因座被检测到。

[0611] 该数据表明,使用不同类型的SNP和0.05g(50pg)至5ng的输入DNA量,能够高效和高灵敏度地检测到多于10,000个基因座。

[0612] 示例7:抑制剂对序列库制备的活性评估,包括根据基因座类型进行评估

[0613] 本示例描述了某些抑制剂对本文公开的序列库(被测序的核酸库)制备的影响的评估,序列库也称为DNA谱,包括按被检测和测序的基因座类型。来自犯罪现场的DNA样本通常含有抑制扩增的共纯化杂质。常见的抑制剂包括血红素、腐殖酸和靛蓝。

[0614] 为了评估法医样本中常见的抑制剂的影响,库制备如示例1所述进行,但使用Forenseq通用分析软件版本2.2来分析结果,某些抑制剂对扩增的影响评估如示例3所述进行,但测试的抑制剂如下:200μM血红素、100μM血红素、50ng/μL腐殖酸、25ng/μL腐殖酸、16μM单宁酸、8μM单宁酸、133μM靛蓝和66.5μM靛蓝如示例1所述包含在扩增步骤中,并使用10230个基因座的引物对。还进行了不含任何抑制剂的阳性对照反应。使用1ng输入DNA。

[0615] 结果如图10所示,其表明根据本文所述方法,能够以高效率和高检测率来扩增和检测包括kiSNP、Y-SNP、X-SNP、piSNP、iiSNP和aiSNP的各种SNP并相互组合,如通过例如即使在存在抑制剂的情况下也可以检测到所有或几乎所有每种类型的SNP所表明的。例如,检测到的kiSNP、Y-SNP、X-SNP、piSNP、iiSNP和aiSNP的数量与在缺乏抑制剂的阳性对照中检测到的数量相似(图10)。该数据表明,样本中存在常见抑制剂不会对使用本文所述方法在PCR反应中扩增多于10,000个SNP的能力产生不利影响。

[0616] 示例8:使用模拟性侵犯后获得的DNA样本对序列库制备进行评估

[0617] 本示例描述了使用来自模拟性侵犯样本的DNA来生成序列库(被测序的核酸库),也称为DNA谱,以便确认是否可以使用少量输入DNA(例如,少于推荐量1ng,诸如500pg)成功生成序列库,例如,被测序的核酸库。

[0618] 模拟性侵犯DNA是从模拟性侵犯发生后9小时和22小时收集的样本中获取的。使用差异提取法从小部分精子中分离DNA,收集和保存两个时间点的小部分精子以供分析。小部分精子中可用作分析输入的DNA量(用于生成序列库)仅为500pg,其是建议量1ng的一半。

[0619] DNA样本用于生成序列库(被测序的核酸库),如示例1所述,但是使用Forenseq通用分析软件版本2.2来分析结果。图11显示了检测出的基因座百分比(调用率)以及测定法中存在的每种SNP的数量。结果表明,即使只有500pg的输入DNA,大多数SNP都能被检测到,在9小时时间点处检测到所有SNP的99.99%(10,230个SNP中的10,229个),并且在22小时时间点处检测到所有SNP的99.93%(10,230个SNP中的10,223个)。具体来说,所有aiSNP、iiSNP、piSNP、X-SNP和Y-SNP在模拟性侵犯发生后的9小时和22小时时间点处均被检测到。在9小时时间点处,9,867个kiSNP中只有一个未被检测到,并且在22小时时间点处,9,867个kiSNP中只有7个未被检测到。检测到的基因座数量足以上传到遗传学数据库以搜索亲属。

[0620] 该数据表明,本文所述方法能够用于检测多于10,000个SNP,包括各种kiSNP、Y-SNP、X-SNP、piSNP、iiSNP和aiSNP并且创建序列库,在模拟性侵犯发生后9小时和22小时处仅使用500pg的DNA,可检测到超过99.9%的所有SNP。因此,本文所述方法适用于在犯罪事件(包括性侵犯)后、或从灾难或冲突的受害者、或从失踪人员留下的样本中,使用少于推荐量——例如,500pg——的DNA来创建序列库。

[0621] 示例9:关于根据唾液样本生成序列库的PCIA残留的评估

[0622] 本示例描述了根据来自唾液样本的DNA对核酸文库进行测序(例如,生成DNA谱),该唾液样本是使用苯酚-氯仿-异戊醇(PCIA)提取法进行有机提取而提取的。

[0623] 唾液DNA是从唾液样本中提取的,其中有意将越来越多的提取试剂PCIA(例如,无PCIA、轻度PCIA、中度PCIA和重度PCIA)留在提取的DNA中作为残留,其模拟不完美的提取。PCIA(包括其成分苯酚)是已知的PCR扩增抑制剂。

[0624] 使用具有无PCIA、轻度PCIA、中度PCIA或重度PCIA的DNA样本生成序列库(被测序的核酸库),如示例1中所述,但使用Forenseq通用分析软件版本2.2来分析结果。确定针对每个样本检测到的SNP总数,如图12所示。结果表明,即使PCIA残留很高且是重度PCIA残留,但PCIA残留并不影响测定法检测SNP的能力,因为每个样本中检测到的SNP超过10,170个。

[0625] 示例10:评估在各种基质上从血液样本生成序列库以及血红素的影响

[0626] 本示例描述了对DNA进行核酸库测序(例如,以生成DNA谱),该DNA来源于犯罪现场常见的不同基质(包括铁锈和牛仔布)中留存的血液样本,以及对其中只有420pg的DNA可用的拭子上的血液样本,以及使用其中血红素水平随着DNA的增加而增加的CheleX™提取的血液样本。血红素是已知的PCR扩增抑制剂。牛仔布含有靛蓝染料,该靛蓝染料是已知的PCR扩增抑制剂。

[0627] 每份DNA样本均用于生成序列库(被测序的核酸库),如示例1所述,但使用Forenseq通用分析软件2.2版来分析结果,其包括含有血液和铁锈的样本、两份牛仔布中的血液样本、拭子上的420pg血液样本、作为对照的含有少量或中等量血红素残留或无血红素的血液样本,以及阳性对照血液样本。确定了针对每份样本和参考对照检测到的SNP总数,并显示在图13中。结果表明,留存在不同基质中的血液样本仍然可以检测到10,230个总SNP中的10,114个或更多个SNP。具有仅420pg的血液样本就检测到了9,563个SNP,而含有血红素的样本检测到了超过10,000个SNP,并且检测到的SNP数量不受样本中存在的血红素量的影响。这表明,根据本文提供的方法,从在犯罪现场常见的各种基质上留存的血液样本中提取的DNA能够用于检测超过10,000个SNP以用于法医应用。

[0628] 示例11:使用相关样本、相关生前样本、无关死后样本和相关模拟死后样本进行亲属关系分析

[0629] 本示例描述了执行本文所述的亲属关系分析以在四个不同的样本集(包括非退化、高度退化和低输入样本)中识别多达三级的关系。具体地,本示例的目标是根据以高复杂度测序的退化样本来确定多达三级的家族关系,同时仍允许足够的SNP以准确预测这样的家庭关系,潜在匹配位于本地的私有数据库(而不是可公开访问的数据库)中。图14中描绘了所涉及方法的示意图,其包括以下步骤:(a)整理法医相关SNP靶的列表,并选择>10,000个SNP,诸如10,230个SNP;(b)通过标记和复制靶、富集靶、纯化靶和归一化靶量,从死后和生前类型的DNA样本中制备测序库;(c)以更高的复杂度(例如,12重或更高)执行下一代

测序;(d)生成SNP报告(也称为DNA谱);(e)将SNP报告上传到本地服务器;(f)进行成对比较;以及(g)计算亲属关系系数和似然比,并过滤出最有可能的家族关系。在一些示例中,步骤(a)中的整理是在先前的工作流程中进行的,并且在本工作流程中利用相同的选定SNP靶,例如,特定的10,230个SNP靶的集合。

[0630] 在四个样本集合中的每个集合中,选择了10,230个SNP靶的集合进行检测。对四个不同的样本集合进行测序以生成序列库,如示例1所述。这四个不同的样本集合包括:(1)来自包括在Coriell验证的多达二级关系的CEPH/Utah的相关生前样本的集合(本文称为"相关生前CEPH/Utah样本");(2)来自包括多达五级关系的私人家族的相关生前样本的集合(本文称为"相关生前私人家族样本");(3)包括骨骼(火化、防腐、烧毁和埋葬)、牙齿遗骸/牙齿和不同退化指数(DI)水平的退化血液的无关死后样本的集合(本文称为"真实死后样本");以及(4)包括来自集合(2)的相同样本的相关模拟死后样本的集合,但是包括以下DNA:(a)通过将DNA煮沸24小时而人工退化(DI范围为2.1-20)或(b)在低DNA输入(50pg)下测序(本文称为"相关模拟死后样本")。

[0631] 使用MiSeq FGx测序系统对真实死后样本进行12重测序并且结果如图15所示,其显示了该真实死后样本的集合中针对个体样本检测到的SNP数量。这包括标记为"牙齿"、"退化血液"、"埋葬骨骼"、"低输入样本"、"其他退化样本"和"其他真实死后样本"的真实死后样本。如图15所示,对于真实死后样本,4个埋葬骨骼样本中的3个具有检测到(或调用)的最低的SNP数量,范围从248到1319个SNP,并且具有最高DI(DI 158和DI 56)的退化血液样本具有检测到的次低的SNP数量,分别检测到4,603和5,069个SNP。所有剩余的真实死后样本检测到的SNP最少为6,737个并且最多为9,903个(图15)。"总通过"计数反映在全体10,230个SNP的集合中针对每个样本检测到的SNP总数,而"计数通过"计数反映在跨样本中一致调用的2,639个SNP子集中检测到的SNP总数。如图15所示,跨样本存在被一致调用(即,被检测到)的核心SNP集合,因为在2,639个SNP子集中检测到的SNP数量("计数通过"SNP)与总体被调用的SNP总数相比变化较小。

[0632] 使用MiSeq FGx测序系统对模拟死后样本进行12重测序,并且结果如图16所示。如图16所示,通过煮沸人工退化的相关模拟死后样本,即,DI大于0的样本,检测到的SNP数量范围在1,470到8,999之间,针对来自相关父母和女儿的样本平均检测到6,462个SNP。低输入DNA样本具有0的DI和0.05ng的DNA输入(图16)。

[0633] 使用MiSeq FGx测序系统对相关生前CEPH/Utah样本进行12重、16重、24重和32重测序,以确定最高复杂度,这将产生足够多的检测到的SNP数量(即,SNP调用率)以用于亲属关系分析。24重测序运行平均检测到9,691个SNP,范围从检测到的8,297个SNP到检测到的9,982个SNP,具体取决于样本;并且32重测序运行平均检测到9,048个SNP,范围从检测到的6,894个SNP到检测到的9,827个SNP(数据未显示)。这表明30重运行将允许足够高的检测到的SNP通量,而不会显著影响检测到的SNP数量和亲属关系分析的置信度。

[0634] 使用MiSeq FGx测序系统利用30重测序运行对相关生前私人家族样本以及三个无关样本进行测序,并且结果如图17所示。如图17所示,除了"自身"样本的一个重复样本(标记为"rep1*")外,每个样本中均检测到超过7,000个SNP,该重复样本可能是由于该特定样本的库制备错误造成的。检测到的超过7,000个SNP的样本包括"自身"样本和"自身"个体的几个亲属,包括来自隔一代的堂/表亲、女儿、姐妹、侄子、堂/表亲和丈夫的样本。

[0635] 然后使用测序运行后生成的DNA谱进行亲属关系分析。比较了相关生前私人家族样本(以30重测序)和模拟死后样本(以12重测序),这些样本源自相同的原始相关样本,但相关模拟死后样本已被人工退化或以低输入使用。使用最小亲属关系系数值0.031,所有预期关系被匹配到多达三级(例如,堂/表亲),并且没有获得假匹配(例如,没有假阳性),从而获得100%的特异性和100%的灵敏度(数据未显示)。获得了一些但不是全部的预期四级匹配(例如,隔一代的堂/表亲)(数据未显示)。这些数据证实,该方法能够准确识别多达三级关系,同时排除所有无关关系,即使在高度退化和低输入样本(诸如失踪人员和灾难/冲突受害者的情况下可能存在的样本)的情况下也是如此。

[0636] 接下来,为了评估最合适的亲属关系系数阈值,将相关模拟死后样本(以12重测序)与GEDMatch数据库进行比较,假设相关模拟死后样本所代表的相关个体在GEDMatch数据库中没有任何近亲。确定要实现100%的特异性(即,无假阳性),亲属关系系数需要为0.062。将此亲属关系系数值阈值(0.062)应用于相关模拟死后样本(以12重测序)并与相关生前私人家族样本(以30重测序)进行比较,确定这会通过排除一些已知的三级关系来降低灵敏度。

[0637] 示例12:评估最少数量的SNP以准确确定亲属关系

[0638] 为了识别检测到的SNP的最小数量(即,SNP调用率)以准确确定亲属关系,比较了GEDMatch数据库中模拟一系列检测到的SNP(即,调用的SNP)的已知关系。这包括测试2,000个SNP、4,000个SNP、6,000个SNP、8,000个SNP和10,000个SNP的调用率,以识别一级、二级和三级家族关系的灵敏度和特异性。结果如图18A-E所示,其描绘了2,000个SNP(图18A)、4,000个SNP(图18B)、6,000个SNP(图18C)、8,000个SNP(图18D)和10,000个SNP(图18E)的结果的接收者操作特性(ROC)曲线。最低SNP调用率(n=2,000)导致一级、二级和三级关系的特异性降低(图18A),这表明检测到的2,000个SNP的SNP调用率是准确识别关系的SNP调用率的绝对底线。

[0639] 对以3重未退化样本进行测序的能力为识别四级和五级关系提供了可能性。使用GEDMatch数据库进行了类似的分析,但用于识别四级和五级关系。这包括测试2,000个SNP、4,000个SNP、6,000个SNP、8,000个SNP和10,000个SNP的调用率,以识别四级和五级家族关系的灵敏度和特异性。结果显示在图19A-E中,其描绘了2,000个SNP(图19A)、4,000个SNP(图19B)、6,000个SNP(图19C)、8,000个SNP(图19D)和10,000个SNP(图19E)的结果的ROC曲线。如图19A-E所示,需要更高的调用的SNP的最小数量(~6,000)以准确识别真实四级和五级关系。

[0640] 示例13:使用相关样本、模拟生前样本和模拟死后样本进行高复杂性SNP测序以用于亲属关系分析

[0641] 本示例描述了执行本文所述的亲属关系分析以在不同的样本集(包括非退化、高度退化和低输入死后(PM)和生前(AM)样本)中识别关系程度。具体地,本示例的目标是根据以高复杂度测序的退化样本来确定多达三级的家族关系,同时仍允许足够的SNP以准确预测这样的家庭关系,潜在匹配位于本地的私有数据库(而不是可公开访问的数据库)中。图14中描绘了所涉及方法的示意图,其包括以下步骤:(a)整理法医相关SNP靶的列表,并选择>10,000个SNP,诸如10,230个SNP;(b)通过标记和复制靶、富集靶、纯化靶和归一化靶量,从死后和生前类型的DNA样本中制备测序库;(c)以更高的复杂度(例如,12重或更高)执行下

一代测序;(d)生成SNP报告(也称为DNA谱);(e)将SNP报告上传到本地服务器;(f)进行成对比较;以及(g)计算亲属关系系数和似然比,并过滤出最有可能的家族关系。在一些实施例中,步骤(a)中的整理是在先前的工作流程中进行的,并且在本工作流程中利用相同的选定SNP靶,例如,特定的10,230个SNP靶的集合。在一些实施例中,使用窗口亲属关系算法。

[0642] A.方法论

[0643] 选择了10,230个SNP靶的集合以用于在每个样本集合中检测,包括模拟生前样本和模拟死后样本。

[0644] 如示例1所述来生成库,以1ng的NA24385 DNA作为这些研究中的阳性对照,并且使用预期的质量控制度量在ForenSeq通用软件版本2.3(UAS)中自动进行分析。

[0645] 这些研究购买了商业上获得的完整DNA,其用作模拟生前(AM)样本,包括来自1000基因组计划、CEPH收集或来自Coriell医学研究所(美国新泽西州卡姆登)的个人基因组计划DNA样本的81个DNA样本,以及来自Innogenomics Inc.(美国路易斯安那州新奥尔良)的全血中提取的4个DNA样本。低输入DNA样本包括样本NA24385,其量分别为0.05ng、0.1ng、0.25ng和0.5ng。

[0646] 模拟死后(PM)样本DNA提取物由五个同时期牙齿(contemporary tooth)(CT)样本(指定为CT1、CT2、CT3、CT4和CT5)、七个同时期骨骼样本和一个来自东欧古代骨骼的DNA提取物组成。针对样本CB 1、CB 3、CB 4、CB 6和CB 7使用PrepFiler™法医DNA提取试剂盒(Thermo Fisher,美国马萨诸塞州沃尔瑟姆),或针对骨骼样本CB 2和CB 5使用脱矿方案来提取七个同时期骨骼(CB)样本的DNA。使用Quantifiler™Trio DNA定量试剂盒(Thermo Fisher,美国马萨诸塞州沃尔瑟姆)来确定CB骨骼DNA样本的退化指数和DNA浓度。CB样本的DI分别为针对CB1、CB2、CB3、CB4、CB5、CB6和CB7的13.6、4.3、5.6、1.1、1.8、2.5和6.5。

[0647] 此外,故意退化的DNA被用作模拟AM或模拟PM样本。从Innogenomics Inc.(美国路易斯安那州新奥尔良)购买了两组退化的DNA:使用有机方法从两名不同男性供体的全血中提取的DNA,以及在50℃下时间范围为0至16小时(样本1231和3551)通过超声剪切的DNA。还确定了人类DNA的量和退化状态。1231样本的DI分别为26.3、33.6、48.6、160.3和459.8,其中1231样本被指定为7、8、10、11和12。3551样本的DI分别为56和158,其中3551样本被指定为56和158。

[0648] 口腔样本收集自具有已知谱系的家族(RF004、RF016-021)的志愿者,本文称为相关家族(RF),其谱系如图26所示。从口腔拭子中提取和纯化DNA。使用高温处理对来自相关家族(RF004和RF016)的两个DNA样本进行人工退化,具体如下:将来自每个个体的五个纯化口腔DNA重复样本进行21个周期的加热和冷却,在98℃下持续1小时,然后在4℃下持续10分钟。将DNA级水加入随后干燥的DNA中以使DNA溶解。确定针对所有相关家族DNA样本的退化指数和DNA浓度。退化指数随重复样本的不同而变化,样本RF004的值为1、2.1、2.6、5.1和20,样本RF016的值为1、1.5、2.0、2.2和2.9。

[0649] 使用ForenSeq Kintelligence试剂盒(Verogen,美国加利福尼亚州圣地亚哥)并按照制造商的说明来制备DNA序列库,并使用QuantiFluor ONE dsDNA系统(Promega,美国威斯康星州麦迪逊)对库进行定量。在使用更高复杂度对库进行测序时,利用了独特的双索引接头(UDI)。在库制备之前,对完整的DNA样本进行定量用于到库制备的输入。利用qPCR方法对模拟PM DNA样本进行定量。除非另有说明,针对被添加到库制备反应中的1ng的总DNA,

DNA被稀释到40pg/μL。针对500、250、100和50pg的总DNA输入,阳性对照DNA NA24385被连续稀释到20、10、4和2pg/μL以模拟低输入PM样本。购买的人工退化样本的DNA浓度足以将1ng的DNA添加到库制备反应中。并非所有退化的相关家族样本都具有足够的DNA浓度以将1ng DNA输入到库制备反应中。对于样本RF004,具有2.1、20、5.1和2.6的DI和600pg、600pg、700pg和250pg的退化的重复样本分别被添加到库制备反应中。对于样本RF016,DI为2.0、2.2和2.9的退化重复样本具有足够的DNA浓度以将1ng添加到库制备反应中。为了模拟低输入样本,分别将50pg的RF004(DI为1)和50和250pg的RF016(DI为1和1.5)添加到库制备反应中。基于~1400mtDNA副本/μL的线粒体DNA定量来估计古代骨骼DNA浓度为390pg。每个库制备的集合包括一个1ng NA24385 DNA的阳性扩增对照和一个阴性模板对照(NTC)。

[0650] 在靶的扩增和纯化后,库被归一化为0.75ng/μL。如果库产量低于0.75ng/μL,则库被整齐池化而无需稀释。从市售获得的完整DNA生成的模拟AM库显示库产量>0.75ng/μL,但有一个例外是0.67ng/μL。从模拟PM、低DNA输入和市售退化的DNA样本生成的一些库的产量也>0.75ng/μL。通过将8μL每个被归一化或整齐的库移液到1.7ml微量离心管中,以不同的复杂度来池化库。从模拟AM样本生成的库以3、12、16、24、30或32个总库的样本复杂度被池化以进行变性和测序。从模拟PM DNA样本生成的库以3或12个总库的样本复杂度被池化以进行变性和测序。被池化的库在室温下用新鲜稀释的NaOH(HP3)孵育5分钟使其变性,然后用HT1稀释。人类测序对照(HSC),用作阳性测序对照的由33个STR组成的库,也类似地变性并且用HT1稀释。然后将变性的库池与变性的HSC结合,在带有遵循制造商的建议的MiSeq FGx试剂盒的MiSeq FGx仪器上测序。在可能的情况下,测序运行是在ForenSeq通用分析软件v2.2中创建的。测序对所有库都使用了151个周期双端读取。测序运行包括需要利用UDI接头中存在的索引对库进行解复用的两个八周期索引读取。

[0651] 为了生成模拟研究中每个样本具有非常高读取的模拟AM文库的测序结果,遵循制造商的建议(参考NextSeq和变性/池化指南),使用NextSeq 500/550高输出试剂盒v2.5(300周期)试剂盒(Illumina,美国加利福尼亚州圣地亚哥)在NextSeq 500仪器上对来自模拟AM样本的30个库进行测序。

[0652] 然后使用如下的二级和三级数据分析来分析测序数据。针对MiSeq FGx™运行质量以UAS来设置度量。这些度量包括簇密度、簇通过过滤器、定相、预定相和Q-分数阈值。簇密度是运行中每平方毫米的簇数(K)并且该度量被设置为400-1650K/mm^2以获得最佳测序结果。簇通过过滤器度量经由通过Illumina贞洁过滤器(ref)的簇百分比来衡量碱基调用的质量,并且该度量被设置为≥80%。当此度量失败时,可用读取的数量会受到影响,但不会影响那些通过读取的质量。定相度量表示簇中在读取中落后于当前周期的DNA链百分比,并且≤0.25%的值通过。或者,预定相表示簇中在读取中领先于当前周期的分子,并且≤0.15%的值通过。如果定相或预定相不符合规范,则测序错误的百分比可能会更高。在使用运行数据之前,确定HSC是否通过其度量非常重要。

[0653] 使用在MiSeq FGx上测序的ForenSeq Kintelligence试剂盒中提供的六个UDI接头制备的所有库样本,均使用ForenSeq通用分析软件(UAS)v2.3(Verogen,加利福尼亚州圣地亚哥)进行等位基因和基因型调用分析,如前所述(Jager等人,Forensic Sci.Int.Genet.,2017,28:52-70,其全部内容通过引用并入本文)。对于使用额外UDI接头制备以实现更高复杂性的库样本,使用与UAS中使用的相同生物信息学流水线来分析测序

运行,但通过命令行工具在单独的服务器上运行。该流水线(在UAS内或通过命令行工具运行)具有与用于ForenSeq DNA特征分析(Jager ref)(Verogen,San Diego,CA,USA)的UAS v1.3中相同的SNP基因型调用基本算法。首先,通过对二进制碱基调用(BCL)文件进行解复用并生成FASTQ文件,基于在UDI接头上提供的索引序列对样本进行解复用。使用Smith-Waterman-Gotoh算法(Gotoh,O.,J.Mol.Biol.,1982,162:705-708)将读取1和2与引物序列对齐。与特定引物对对齐的读取被分配到与这些对相对应的基因座。然后以BAM格式写入对齐。在每个SNP的位置处,计算与参考碱基调用的匹配和与替代碱基调用的匹配,对30的最低碱基质量进行过滤。然后对每个基因座处每种类型的调用(参考或替代)的读取数量进行求和,要求最小覆盖大于10个读取。然后通过过滤分析(AT)和解释(IT)阈值来确定SNP基因型,这两个阈值均被设置为3%。AT和IT阈值是通过将3%乘以在该基因座处计数的读取总和来确定的。当出现低覆盖时,使用最少650个读取来计算阈值。然后将得到的AT和IT值与每个基因座的参考和替代等位基因的总读取计数进行比较。如果调用通过了AT和IT阈值,则确定每个基因座的基因型。然后将基因分型结果写入变体调用格式(VCF)。

[0654] 在一些研究中,执行了GEDMatch数据模拟以进行全基因组亲属关系算法测试。对于这些研究,下载并分析了GEDMatch数据库配置文件,如Snedecor等人,Forensic Sci.Int.Genet.,2022,61,102769,doi:10.1016/j.fsigen.2022.102769(其内容在此以引用的方式全部并入)中所述。从GEDMatch中随机选择了1000个匿名样本(称为“查询样本”)的集合。然后在GEDMatch数据库中查询这些样本的亲属,并且基于由GEDMatch一对多工具计算出的共享厘摩(cM)值来选择任何查询结果(称为“靶样本”)。此搜索产生了2954个靶样本以及匹配的查询样本。这些结果包括具有零共享cM的查询-靶对,代表真实无关对。因此,由于包含相关和无关的靶样本,靶样本的数量大于查询样本的数量。关系程度是通过将从一对多工具生成的所得共享cM值与由DNA Painter提供的每个程度的预期共享cM范围进行比较来确定的,DNA Painter可在以下网址访问:https://dnapainter.com/tools/sharedcmv4。首先对查询和靶集中包含的样本分型的基因座针对面板中的10,230个SNP进行过滤,然后随机过滤到80%、60%、40%和20%的调用率,从而在每个查询-靶对中分别调用8000、6000、4000和2000个基因座。使用亲属关系算法来计算针对每个降低的基因座调用率级别的每个查询-靶样本对的全基因组亲属关系系数。全基因组亲属关系系数大于0.031的配对被认为是相关的;全基因组亲属关系系数小于或等于0.031的配对被认为是无关的。通过将这些结果与一对多工具查询结果进行比较来计算灵敏度和特异性。换句话说,一对多查询结果被视为真值集,而由亲属关系算法生成的结果被视为测试集。

[0655] 然后按如下方式来计算似然比(LR)和亲属关系值。使用算法pedprobr(Brustad等人,Int.J.Legal Med.,2021,135:117-129,其内容在此通过引用整体并入本文)和dvir(Vigeland等人,Scientific Reports,2021,11:13661,其内容在此通过引用整体并入本文)来计算LR。在LR计算中使用来自基因组聚合数据库(gnoMAD)(Karczewski等人,Nature,2020,581:434-443,其内容在此通过引用整体并入本文)v3.0的群体频率平均值。未使用突变模型,并且将theta设置为0,因为选择用于分析的SNP具有较低的连锁不平衡(Karczewski等人,上文)。LR计算如下:

[0656] $$LR = \frac{P(D \mid H_r)}{P(D \mid H_u)}$$

[0657] 其中,D表示基因型,H_r表示个体是相关的假设,并且H_u表示个体是无关的假设。相关假设通过谱系表示,其中未被识别的个体被测试为亲属。无关假设通过Hardy-Weinberg平衡计算表示。计算每个基因座的LR值,然后将其乘以基因座,从而得出关系的最终LR。为了提高计算效率,将每个基因座LR转换为对数,并将基因座LR相加。然而,由于该平台的高复杂性,并且在大规模死亡事件(MFI)案例中,PM样本通常是高退化程度和/或低输入量,PCR期间的随机效应会导致等位基因脱扣,并可能导致两个个体在某个基因座上的等位基因组合似乎不可能的情况。这通常发生在亲子关系中,其中关系中的每个个体都被基因分型为不同等位基因的纯合子,但实际上一个或两个都是杂合子(例如,父母的基因型为AA,孩子的基因型为BB,而父母或父母和孩子实际上是AB)。基因分型错误和新生突变也是造成这种情况的可能原因。这种情况导致LR为0,并且可能导致总LR为0,因为最终LR是基因座LR的乘积。为了确保基因座LR不为0,对pedprobr进行了修改。这些情况下基因座的似然比计算如下:

[0658] $$LR = \frac{0.001}{p^2q^2}$$

[0659] 其中,0.001表示基因分型错误率,p是等位基因1的等位基因频率,并且q是等位基因2的等位基因频率(Galván-Femenía等人,Heredity,2021,126:537-547,其内容在此通过引用整体并入本文)。

[0660] 全基因组亲属系数、共享cMs和最长片段cM是使用PC-Relate(Conomos等人,American Journal of Human Genetics,2016,98:127-148,其内容在此通过引用整体并入本文中)和PC-AiR(Conomos等人,Genetic Epidemiology,2015,39:276-293,其内容在此通过引用整体并入本文中)计算的,如先前在Snedecor等人,Forensic Sci.Int.Genet.,2022,61:102769(其内容在此通过引用整体并入本文中)中所述。当两个个体之间的关系未知时,此方法很有用,因此不需要谱系。

[0661] PC-AiR方法首先获取基因型个体的集合,并将它们分成两个不重叠的子集:一个集合包含代表所有个体祖先的无关个体(无关子集),另一个集合包含在第一个子集中至少有一个亲属的个体(相关子集)。为了构建无关子集,对原始PC-AiR方法进行了修改以提高构建模型的计算效率。没有亲属或亲属最少的样本被添加到无关子集中,而亲属较多的样本则被排除在无关子集之外。这是通过计算每对亲属关系值来执行的,如Conomos等人在Genetic Epidemiology,2015,39:276-293中所述,其内容在此通过引用整体并入本文中,并基于严格的阈值将每个个体分类为有亲属或没有亲属关系:如果亲属关系值大于0.01,则认为是亲属,如果亲属关系值小于-0.025,则认为是无亲属关系。排除具有小于5%的失踪SNP数据的样本。接下来,对无关子集进行主组分分析(PCA),然后基于与无关子集中个体的遗传相似性,预测相关子集中所有个体的变异组分值。所得组分代表能够用于代替静态种群频率来识别未知个体集合中的匹配的模型。本研究中使用的模型是使用GEDMatch数据库构建的。

[0662] PC-Relate方法使用来自PC-AiR的主成分并且将遗传相关性分为两个组分:一个组分用于共享由最近的共同祖先传下来的相同的等位基因,并且另一个组分用于共享由于更远的共同祖先而产生的等位基因。来自PC-AiR的组分用于基于个体的祖先背景估计等位基因频率,使用线性回归而不是静态种群频率(诸如来自gnoMAD的频率)。对于两个个体i和

j,然后使用来自PC-AiR模型的估计等位基因频率$\hat{u}$来计算亲属关系系数$\widehat{\phi_{ij}}$,如下所示

[0663] $$\widehat{\phi_{ij}} = \frac{\sum_{s\in S}(g_{is} - 2\hat{u}_{is})(g_{js} - 2\hat{u}_{js})}{4\sum_{s\in S}[\hat{u}_{is}(1 - \hat{u}_{is}) \cdot \hat{u}_{js}(1 - \hat{u}_{js})]^{1/2}}$$

[0664] 其中,s是在两个个体中都被分型的S个SNP中的SNP,g_{is}和g_{js}分别是在SNP s处的i和j中的参考等位基因的数量,并且$\hat{u}_{is}$和$\hat{u}_{js}$分别是由在SNP s处的i和j的PC-AiR计算出的预期等位基因频率。该算法被称为“全基因组亲属关系”,因为它将整个基因组视为一个关联度的片段,其用全基因组亲属关系系数表示。在引用全基因组亲属关系算法时,会使用全基因组亲属关系系数来识别关系。当在两个个体之间共享的SNP数量少于6000时,本研究中需要两个个体之间的全基因组亲属关系系数超过0.031才能被视为亲属。

[0665] Snedecor等人(上文)引入了额外的步骤,称为“窗口亲属关系”,以更准确地识别远亲关系。窗口亲属关系包括计算跨基因组的亲属关系的窗口以找到共享的亲属关系片段。这是通过枚举每个染色体内的所有可能窗口并计算所有窗口的亲属关系系数来执行的。然后通过最小亲属关系系数阈值过滤这些窗口并将其包含在共享cMs计算中。然后迭代过滤后的片段,并分别对共享至少一个等位基因和两个等位基因的SNP延伸进行分类。然后跨所有片段来计算总的共享cM。在参考窗口亲属关系算法时,使用总的共享cM和最长cM片段来识别关系。当两个个体之间共享的SNP数量在6000和8000之间时,共享cM值必须高于180且最长cM片段必须高于30才被视为存在关系。当两个个体之间共享的SNP数量在8000和9000之间时,共享cM值必须高于150且最长cM片段必须高于30才被视为存在关系。当两个个体之间共享的SNP数量为9000或更多时,共享cM值必须高于140且最长cM片段必须高于30才被视为存在关系。全基因组亲属关系系数能够用于过滤任意数量的共享SNP。然而,Snedecor等人(上文)观察到,当SNP重叠大于6000时,对共享cM和最长片段cM进行过滤(例如,使用窗口亲属关系)时具有更高的特异性,特别是对于更高的关系程度。

[0666] 更简单地说,两个个体之间分型的SNP的数量(SNP重叠)能够用于决定何时使用全基因组亲属关系算法(<6000个SNP重叠)以及何时使用窗口亲属关系算法(>6000个SNP重叠)。并且,当基于该SNP重叠来决定一种算法时,将使用一个值或值的集合来过滤数据以识别关系,具体取决于所选的算法。全基因组亲属关系和窗口亲属关系这两者的截止值都经过选择以确保高灵敏度,但更重要的是,高特异性,如Snedecor等人(上文)所示。降低这些阈值可能会捕获更多的关系(即增加灵敏度),但预计会引入更多的假阳性结果,特别是对于更远的关系(例如,四级和五级)。

[0667] 利用上述方法,进行了一项研究以更高复杂度对库进行测序,从而减少总样本读数,旨在生成稳定但更小的分型SNP的集合,从而实现一种经济有效的亲属关系分析方法。

[0668] B.结果

[0669] 为了显示使用序列库实现高复杂度(plexity)的可行性,在从模拟AM样本中的高质量DNA样本生成的30个库的集合上模拟了高复杂度(在一次测序运行中>3个样本)。如上所述,制备了30个模拟AM样本并一起测序,以生成具有大量读取的样本数据。使用ForenSeq UAS二级分析流水线的自定义本地构建(build)对BCL文件进行解复用,由于NextSeq生成的原始数据不同,该流水线仅在FASTQ文件的生成上有所不同。为了确定模拟16个序列库的读取数,将25,000,000(总估计读取数)除以16(样本数)并乘以平均对齐百分比(96%),得出

此复杂度下每个样本的总读取数为1,500,000。然后执行相同的计算以模拟30个序列库,但不是除以16,而是除以30,结果为800,000个读数。模拟读数较少的测序运行被称为“下采样”。诸如seqtk(https://github.com/lh3/seqtk)的生物信息学工具能够用于模拟读数较少的运行或下采样测序运行。Seqtk用于将每个样本下采样至150万个读数以模拟16个样本/运行的复杂度,或800,000个读数以模拟30个样本/运行的复杂度。seqtk从FASTQ文件中随机选择读数并输出具有所需读取数量的新FASTQ。随后的下采样FASTQ通过本地构建的ForenSeq UAS流水线进行处理,该流水线如上所述分析了FASTQ。

[0670] 每个样本产生的总读取数范围为8,086,090到32,707,490,平均为23,186,251个读取数。为了模拟高复杂度,从每个样本的FASTQ文件中随机选择读取数,直到满足所需的读取数为止,这称为下采样。下采样被定义为从FASTQ文件中随机选择读取数,直到达到所需的读取数为止。随后将随机选择的读取数输出到新的FASTQ文件。使用如上所述的生物信息学算法分析所得的FASTQ文件。通过将数据分别下采样为针对每个样本的1.5M读取数和800,000个读取数,模拟了16和30的测序复杂度。减少每个样本的读取数会导致分型SNP的数量预期减少。

[0671] 为了确定跨样本的SNP调用率,在每个模拟复杂度(16和30)下针对每个样本来确定分型SNP的数量。通过将FASTQ文件中的原始读取下采样到每个样本150万个读取,生成16个样本运行(16重)数据,而通过将原始读取下采样到每个样本80万个读取,生成30个样本运行(30重)数据。随后进行基因分型,并汇总每个样本分型SNP的数量。对于16个样本运行,最小值为7781,第一四分位数为8472,中值为8630,第三四分位数为8708,并且最大值为8848。对于30重,最小值为6375,第一四分位数为7179,中值为7299,第三四分位数为7382,最大值为7516。图20A显示了30个库的两个模拟复杂度的分型SNP分布。复杂度为16时的平均恢复率为8586个SNP(范围从7781至8848个SNP),中位数为8630个SNP,第一四分位数为8472个SNP,并且第三四分位数为8708个SNP;复杂度为30时的平均恢复率为7234个SNP(范围从6375至7516个SNP),中位数为7299,第一四分位数为7179,第三四分位数为7382个SNP(图20A)。

[0672] 使用上面描述的亲属关系算法和似然比来确定关系在很大程度上依赖于一对一比较中每个样本中分型SNP的数量,也称为SNP重叠。

[0673] 两个样本中分型SNP的数量越多,亲属关系和似然比值的可信度就越高。因此,计算了模拟的16和30复杂度测序运行这两者中样本之间的平均SNP重叠,以查看样本之间是否有足够的分型SNP基因座可以被分享以识别具有亲属关系和似然比值的真实关系。在模拟的16和30复杂度测序运行这两者中跨所有样本组合的平均常见分型SNP是6998个基因座,最小重叠为6058并且最大重叠为7322。模拟表明,以针对每个样本获得的总读取较少的方式对库进行测序将允许针对每个样本分型较小的稳定的SNP集合,并且有足够的重叠来确定亲属关系。常见的分型SNP的数量少于8000,这可能不足以识别更高阶的关系(例如,四级和五级),但可能足以识别到三级的关系。

[0674] 预期以增加的复杂度对库进行测序将导致分型SNP的数量减少,这可能导致等位基因脱扣增加和随后的杂合性降低,特别是对于DNA退化和/或低数量的死后(PM)样本。为了评估更高复杂度下的基因座和等位基因脱扣以及杂合性的水平,对如上所述的从模拟生前样本(AM)和模拟死后(PM)生成的库进行测序,推荐的复杂度为每次测序运行3个样本,然

后以每次测序运行12、16、24和32个样本的四个更高复杂度进行测序。所得的分型SNP的数量如图20B所示。如图20B所示,最小值、第一四分位数、中值、第三四分位数和最大值如下:3重的为9853、9976、10009、10059和10135;12重的为9332、9394、9419、9520和9945;16重的为8881、9091、9303、9419和9901;32重的为7653、8348、8515、8706和9753。如图20C所示,最小值、第一四分位数、中位数、第三四分位数和最大值如下:3重的为7215、8677、9724、9923和9991;12重的为4603、8261、9360、9664和9903。如图20B所示,对于参考样本,随着复杂度的增加,具有读数低于可接受阈值(AT)的基因座的数量增加。此外,随着测序复杂性的增加,下降到低于AT的基因座的分布范围扩大,对于每次运行3个样本的测序样本,平均为10,111个(最小9,853个,最大10,135个)分型SNP,对于每次运行32个样本的测序样本,平均为8,528个(最小7,653个,最大9,753个)分型SNP。在测试的每次测序运行32个样本的最高复杂度下,分型SNP的最小数量仍保持高于7,000,这表明与上述模拟结果和图20A所示相比,以高复杂度对这些库进行测序会产生更高数量的分型SNP基因座。

[0675] 接下来,通过比较以推荐的每次运行3个样本的复杂度对样本库进行测序所确定的SNP基因型与以每次运行32个样本进行测序的相同样本库所确定的SNP基因型,评估了更高复杂度测序对杂合基因座的姊妹等位基因丢失的影响。对于都以推荐的复杂度测序(3重)和更高的复杂度测序(32重)的每个样本和每个基因座,如果两次运行都分型相同的等位基因,则基因型被认为是一致的,否则不一致。3重测序和32重测序之间的平均重叠为8,610个SNP,最小值为7,808个SNP,最大值为9,667个SNP。对于常染色体基因座,两个等位基因必须匹配才被认为是一致的。通过将一致等位基因的数量除以在两次测序运行中分型的基因座中的等位基因总数来计算每个样本的等位基因一致性。对于模拟生前样本,与32的复杂度相比,在3的复杂度下测序的库之间的等位基因不一致(等位基因下降到低于AT)平均增加了1.9%,最小值为0.50%,最大值为2.8%(图21A,左侧y轴)。通过将每个样本的杂合基因座的数量相加并且将该值除以调用的基因座总数来确定杂合性。使用每次运行32个样本(32重)测序的样本与每次运行3个样本(3重)的标准复杂度相比表明杂合性差异最大,平均差异为6.8%,最小差异为2.0%,最大差异为10.2%,而3重测序表明比这些值更大的杂合性。(图21A,右侧y轴)。

[0676] 来自MFI的受害者的样本的特性是它们经常退化,并且可能含有低水平的基因组DNA。然而,在成本效益和整个过程中,增加每次测序运行的样本数量也是有利的,尽管这可能导致分型SNP的数量减少和杂合性降低,从而使这些问题更加复杂。针对退化和低输入DNA样本,为了评估与每次测序运行3个样本的标准方案相比,增加测序复杂度对模拟死后样本的影响,从具有不同退化程度、低输入DNA样本、牙齿遗骸以及火化、防腐、烧毁和埋葬骨骼的30个模拟死后(PM)样本生成了库。这些库以标准3重和12重进行测序。在两次测序运行之间,每个样本和每次运行中高于AT的分型SNP的数量没有显著差异,尽管以更高复杂度测序的一些样本具有显著更低的分型SNP的数量。例如,如图20C所示,针对模拟PM样本以3重测序的结果是平均SNP为9,313个,最小SNP为7,215个,最大SNP为9,991个,而针对模拟PM样本以12重测序的结果是平均SNP为8,796个,最小SNP为4,603个,最大SNP为9,903个。这些结果表明,针对退化和/或低DNA输入样本(例如,死后样本)增加每次运行的样本数量将增加低于AT的基因座数量,但对分型基因座数量的最大影响是DNA的退化状态和/或数量。

[0677] 如针对上述模拟AM样本所述,对以3重或12重测序的库的模拟PM样本进行基因型

调用一致性测试。这包括同时期骨骼样本和同时期牙齿样本,此外,还从Innogenomics Inc.(美国路易斯安那州新奥尔良)购买了两组退化DNA(样本1231和3551)。通过将32重或12重与相应的3重运行进行比较并且将在每个基因座处的相同等位基因调用归类为一致,将不同的等位基因调用归类为不一致(低于AT),来计算出百分比一致性。对于常染色体基因座,两个等位基因必须匹配才能被视为一致。通过将匹配基因座的数量除以针对每个样本调用的两次运行的基因座总数来计算百分比。图21A中针对模拟AM样本调用的基因座范围为7808至9667,平均值为8610;并且针对图21B中的模拟PM样本,调用的基因座范围为4580至9795,平均值为8627。通过将杂合基因座的数量相加,然后将该值除以所调用的基因座的总数来确定杂合度百分比。模拟PM样本是从牙齿残骸(Tooth)、不同退化程度的血液(样本名称以1231和3551开头)、埋葬骨骼(BB1)、火化骨骼(CB1)、防腐骨骼(CB2和CB3)、烧焦骨骼(CB5、CB6、CB7)提取的DNA,以及量为0.05、0.1和0.25ng(分别为50pg、100pg和250pg)的Coriell样本NA24385的一系列低DNA输入。

[0678] 两次运行中分型的基因座数量范围为4580至9795,平均值为8627个SNP基因座。等位基因不一致(等位基因下降到低于AT)平均增加1.2%,最低为0.1%并且最高为2.6%(图21B,右轴)。牙齿遗骸和火化和烧焦的骨骼表现出最低的等位基因不一致,平均不一致率为0.6%,最小不一致为0.1%,并且最大不一致率为1.3%。

[0679] 如上所述针对模拟AM样本来计算杂合性。杂合性水平随输入DNA的退化水平和量而变化(图21B,右轴)。退化指数为158和56的两个退化DNA样本(分别为3551_158和3551_56)和埋葬骨骼样本(BB1)以及0.05ng(50pg)NA24385样本在测序复杂度之间表现出最高的杂合性差异(分别为9.0%、13.5%、26.9%和10.8%)。1231库7-8和10-12(图21B)由退化量不断增加的相同DNA样本生成(对于被指定为7、8、10、11和12的1231样本的DI分别为26.3、33.6、48.6、160.3和459.8)。杂合性随DNA退化增加而降低,对于在MiSeq FGx上以12的复杂度测序的这些库,DI为26.3的样本1231_7的杂合性为37%,并且DI为459.8的样本1231_12的杂合性为31%。BB1在以3重测序时具有最高水平的杂合性。对该样本(BB1)的基因座内平衡(ILB)的调查表明,针对每次运行以3个样本的复杂度测序的样本和针对每次运行以12个样本的复杂度测序的样本的ILB分别为53%和28%。这些结果表明,在每次运行以3个样本的复杂度测序的BB1样本中观察到的杂合性增加的原因不是由于测序错误,因为预期是ILB会非常低。

[0680] 具有CB标记的骨骼样本(烧毁、防腐或火化骨骼DNA提取物)显示出最小的杂合性差异(七个样本中平均差异为1.7%;最小差异为0.1%;最大差异为3.9%)。此外,使用此方法的高通量在牙齿遗骸中表现良好,牙齿遗骸显示出2.5%的平均杂合性差异,最小差异为1.8%,最大差异为4.5%。

[0681] 为了评估MFI受害者的亲属关系,将AM样本与PM样本进行比较以测试关联度,从而实现1对1比较。两个个体中SNP分型(SNP重叠)的数量对确定关系有显著影响,尤其是远亲关系,例如四级和五级。因此,在以12重测序的模拟PM样本和以32重测序的模拟AM样本之间的所有样本组合中计算SNP重叠(共同的SNP基因座)。这是通过将每个模拟PM样本与每个模拟AM样本配对,识别两个样本中基因分型的SNP基因座的数量,并将该值相加,从而得出针对每对的SNP重叠。模拟PM样本和模拟AM样本组合之间的平均重叠为8,020,最小值为4,071,最大值为9,727。这些数据表明,平均约有8,000个SNP,使用此方法以更高复杂度(每次

运行针对PM样本为12个,针对AM样本为32)对这些库进行测序将能够识别一级、二级、三级、大多数四级关系以及一些五级关系。

[0682] 接下来研究了全基因组亲属关系算法在SNP调用率小于9000个共同基因座(2000-8000个SNP分型)时识别关系的能力。Snedecor等人提供的数据(上文)表明,当SNP调用率约为9,000个基因座时,使用窗口亲属关系算法识别所有关系程度的准确性很高。窗口亲属关系通过计算跨基因组的亲属关系窗口来查找两个个体之间的共享片段。参见Snedecor等人(上文)。全基因组亲属关系算法将整个基因组视为亲属关系的一个片段,而不是两个个体之间的窗口。Snedecor等人表明,为了识别更高阶的关系(例如,四级和五级),当在两个相应样本中分型SNP(SNP重叠)的数量至少为9000个SNP时,窗口亲属关系更准确。如图20B和20C所示,以高复杂度测序的模拟AM样本和模拟PM样本不太可能展现出窗口亲属关系算法所需的>/=9000个SNP调用率。如Snedecor等人所述(上文),当SNP重叠在6000-8000个共同基因座的范围时,窗口亲属关系对于一级、二级、三级和大多数四级关系表现良好。但是,当SNP重叠范围为2000-4000个共同基因座时,窗口亲属关系的性能下降。

[0683] 使用上述的模拟相关谱的方法,接下来研究了当SNP调用率少于9,000个共同基因座(2,000到8,000个SNP分型)时全基因组亲属关系算法识别关系的能力。

[0684] 窗口亲属关系算法报告具有共享厘摩(cM)和最长片段cM的度量的关联度,而全基因组亲属关系方法使用全基因组亲属关系系数来报告关联度,其中亲属关系系数阈值为>0.031才被视为相关。

[0685] 从GEDMatch数据库中随机选择了1000个匿名查询谱,并在整个数据库中搜索以选出针对涵盖一级到五级关系以及具有零个共享cM的无关样本的2954个“相关”靶谱。简而言之,从GEDMatch中随机选择了1000个匿名样本的集合,称为“查询样本”。然后在GEDMatch数据库中查询这些样本的亲属,并基于由GEDMatch一对多工具计算的共享厘摩(cM)值来选择任何查询结果(称为“靶样本”)。此搜索产生了2954个靶样本以及匹配的查询样本。这些结果包括具有零个共享cM的查询-靶对,代表真实无关对。因此,由于包含相关和不相关的靶样本,靶样本的数量大于查询样本的数量。通过将从一对多工具生成的所得的共享cM值与由DNA Painter提供的每级共享cM的预期范围进行比较来确定关系程度。首先针对Kintelligence多重中的10,230个SNP来过滤谱。随后,将谱随机过滤为80%、60%、40%和20%的调用率,从而在针对每个查询-靶对的每个谱中分别产生8000、6000、4000和2000个基因座。然后计算全基因组亲属关系算法以在所有SNP数量上测试每个查询-靶对的一对一关系,总共进行446,182次比较。

[0686] 如图22A所示,即使对于具有2000个分型SNP的谱(平均灵敏度为70.6%,平均特异性为100%),一级和二级关系也表现出高特异性和灵敏度(图22A-D)。对于三级关系,灵敏度开始下降,但对于具有4,000至8,000个分型SNP的谱,全基因组亲属关系算法在关系检测方面的表现优于窗口亲属关系(平均灵敏度为25.1%,平均特异性为99.2%)(图22B、C和D)。对于更少的分型SNP,四级和五级匹配表现出较低的灵敏度和特异性(图22A-D)。窗口亲属关系算法利用这些具有大量分型SNP的库谱,以更高的灵敏度和更高的特异性识别了这些高阶关系(Snedecor等人,Forensic Sci Int Genet 2022,61,102769,doi:10.1016/j.fsigen.2022.102769,其内容通过引用整体并入本文)。重要的是,所有程度的特异性均保持在99.98%以上,其中6,000和8,000个SNP的范围保持在99.9975%以上(图22C和图

22D)。这些结果表明全基因组亲属关系具有非常低的假阳性率。总体而言,这些数据表示以更高复杂度对如本文制备的库进行测序将产生足够的分型SNP数据以自信地识别一级、二级和三级关系。这些模拟结果还表示,将这种生成序列库的方法与本文描述的全基因组算法相结合应该对于灾难受害者身份识别(DVI)有用,并且能够在本地服务器上实现(而不是公众和/或执法机构可以访问的服务器),以允许跨生前样本的小型数据库进行亲属关系确定。

[0687] 接下来,使用如上所述的相关生前CEPH/Utah样本和如图23A所示,测试了本文所述的全基因组算法在如通过上述方法制备的库中识别真实关系的能力。之所以选择这些样本,是因为它们包括两组祖父母、一组父母和十一个兄弟姐妹,从而代表一级和二级关系(图23A)。这些库以每次测序运行12、16、24或32个样本的复杂度进行测序,以确定在排除已知无关对的同时能够维持识别真实关系的准确性的最高复杂度。

[0688] 确定每次运行每个样本中分型SNP的数量,以确保足够的SNP被分型以进行亲属关系和似然比计算(图24)。SNP调用率随每次运行样本数量的增加而降低(图24),并且结果与在图20B和20C中的初始高通量量测试相当,从而表明这些库的高复杂度测序的可重复性。如图24所示,最小值、第一四分位数、中位数、第三四分位数和最大值如下:12重为9831、9859、9907、9987和10079;16重为9718、9816、9862、9957和10024;24重为9428、9570、9618、9798和9907;32重为8979、9294、9387、9588和9743。

[0689] 接下来计算亲属关系度量以使用这些高复杂度谱来确认本文所述算法的敏感性和特异性。计算的亲属关系度量包括全基因组亲属关系系数、共享cM、最长峰值cM和谱系中所有成员组合的似然比。这个家族中的大多数成员都是相关的,只有五对个体是无关的。为了增加无关对照组的数量并最好地模拟用于失踪人员案件的数据库(其主要由从受害者家族成员的定义模拟生前样本组成),包括随机选择的100个1000基因组计划样本。基因型从国际基因组样本资源数据库(Fairley等人,Nucleic Acids Res 2020,48,D941-D947,doi:10.1093/nar/gkz836,其内容在此通过引用整体并入本文)下载并过滤以仅包括ForenSeq Kintelligence试剂盒(Verogen公司)的10,230个SNP面板中的基因座。

[0690] 如上所示,确定每次运行12个样本的复杂度足以从已退化和/或DNA含量低的模拟死后(PM)样本中调用足够的SNP。为了模拟对DVI进行的PM与生前(AM)比较,12个样本运行的样本被视为PM样本。然后将每个PM样本谱与来自所有测序运行(每次运行12、16、24和32个样本)的其他样本谱中的每一个以及来自1000基因组计划的100个无关谱进行配对,以计算每对的全基因组亲属关系系数、共享cM、最长片段cM和似然比,从而识别相关对。Utah/CEPH 1463家族包括祖父母、父母和兄弟姐妹。每个独特的图表代表12个样本运行与其自身(12_vs_12)、16个样本(12_vs_16)、24个样本(12_vs_24)和32个样本(12_vs_32)运行的比较,以模拟死后与生前的比较,并且确定能够保持对生前样本密切关系的准确身份识别的最高复杂度。针对每个对来配对样本和计算亲属关系系数和logLR。对于Utah/CEPH 1463家族的亲属关系分析,从1000基因组计划中随机选择100个以HG前缀开头的样本(Fairley等人,Nucleic Acids Res 2020,48,D941-D947,doi:10.1093/nar/gkz836,其内容在此以引用方式整体并入本文)作为无关对照组。从国际基因组样本资源数据库下载基因型并进行过滤以仅包括在10,230Kintelligence SNP面板中的基因座。如果亲属关系系数大于0.031,则认为该对相关,否则无关,这由图25A中的黑色垂直线表示。如果logLR大于0,则认

为该对相关,否则无关,这由图25B中的黑色垂直线表示。样本包括来自祖父母(G)、父母(P)、兄弟姐妹(S)、无关对照组(U)、无关祖父母(GU)、无关父母(PU)和无关兄弟姐妹(SU)的样本。如图25A和25B所示,亲属关系系数和logLR阈值足以区分相关对和无关对。

[0691] 由于以所有复杂度测序的样本中最小SNP重叠>8,900,因此使用如上所述的窗口亲属关系算法来确定共享cM和最长片段cM。使用如上所述的全基因组算法确定关系,对于所有预期相关对的亲属关系系数远高于0.031,对于所有预期无关对的亲属关系系数低于0.031。此外,由于该家族具有最大的二级关系,因此窗口亲属关系和全基因组亲属关系表现同样出色。

[0692] 全基因组亲属关系系数清楚地区分了每种预期关系而没有检测到假阳性关系(图25A)。如图25A所示,所有无关对的亲属关系系数均低于0.031的阈值,而所有相关对的亲属关系系数均远高于0.031的阈值,从而清楚地区分了相关对和无关对。所有亲属关系的度量都能够准确地识别来自Utah/CEPH 1463家族的所有关系,即使在将使用12个样本运行所生成的谱与使用32个样本运行所生成的谱进行比较时也是如此,因为所有比较都是在小型数据库(诸如DVI案例中使用的数据库)中进行的,其包含很少的无关个体。似然比显示出类似的趋势,所有预期的相关对都通过具有高于阈值0的正值而被正确识别为相关,并且所有无关对通过具有低于阈值0的负值而被归类为无关(图25B)。这些数据表明,对于模拟AM库,在每次运行的最大复杂度为32个样本时,谱包含足够的SNP数据以准确识别多达二级的关系并且排除无关对,其包括在对只有少数无关个体的小型数据库进行分析时,诸如失踪人员和DVI案例中使用的数据库。

[0693] 接下来,如上文所述并如图26所示,相关生前私人家族("相关家族"或RF)样本用于测试更高阶关系,包括一级、二级、三级、四级和五级关系,如图26所示。该相关私人家族包括自身样本和与自身样本具有特定关系的个体,包括父母(Father和Mother)、姑/姨(aunt)、堂/表亲(Cousin)、隔一代的堂/表亲(1C1R)和第二代的堂/表亲(2Cousin)(图26)。这些在图26中被标记有关系的个体被测序并包括在以下亲属关系分析中。

[0694] 针对8个家族成员中的每一个成员都生成了三个重复样本库,具有多个阳性和阴性扩增对照,并且在运行中对总共30个样本进行了测序。

[0695] 来自自身(RF016)和母亲(RF004)的重复样本通过温度经受人工退化以模拟PM样本,具体来说,这些样本经受21个加热和冷却周期,在98℃持续1小时,然后在4℃持续10分钟。测量退化指数后,得到母亲(RF004)的四个重复样本的DI分别为2.1、2.6、5.1和20,自身(RF016)的四个重复样本的DI分别为1.5、2.0、2.2和2.9。从这些退化样本以及完整的自身和母亲DNA样本(50pg)生成库。退化和低输入样本以每次运行12个样本(12重)进行测序,完整样本以30的复杂度进行测序,以及与以12的复杂度测序的模拟PM样本(其中分型SNP的数量明显更低且分布范围很广)相比,完整样本的分型SNP的数量更多且分布紧密(图27A)。这些样本的最小值、第一四分位数、中位数、第三四分位数和最大值如下:对于退化/低输入的12重为3256、8106、8391、8580和8896;对于完整的30重为1470、5057、7807、8782和9898(图27A)。

[0696] 此外,模拟AM样本谱(从完整DNA生成的库),与从以相似复杂度测序的市售DNA样本生成的模拟AM样本谱相比,其分型SNP较少(图27A与图20B和24相比)。以30的复杂度测序的完整模拟AM母亲(RF004)样本与以12的复杂度测序的对应的退化/低DNA输入模拟PM母亲

(RF004)样本之间的杂合性平均百分比差异为14.6％,最小差异为1.6％,最大差异为25.5％。

[0697] 通过将每个退化/低输入模拟PM样本与每个完整模拟AM样本配对进行一对一比较,以计算全基因组亲属关系系数、共享cM、最长片段cM和似然比。当使用本文所述的两种亲属关系算法时,所有多达和包括三级的关系均以100％的灵敏度和100％的特异性确定。使用本文所述的窗口亲属关系算法识别出可能的三十个四级关系中的八个,但使用本文所述的全基因组算法只能识别出可能的三十个四级关系中的四个,但使用这两种算法都没有通过亲属关系阈值的五级关系。在该数据集中,总共可能的样本对中的三十九个是无关的,并且未检测到假阳性关系。19个样本对具有少于2,000个SNP重叠,其中18对为DI为2.6并且杂合度为7.2％的模拟PM样本母亲(RF004),其中一对为输入DNA为50pg并且杂合度为15.4％的模拟PM样本自身(RF0016)。在这19对中,七对为一级亲属,三对为二级亲属,并且三对为三级亲属,所有的亲属关系系数均超过0.031。三对为具有较高亲属关系系数的自身到自身。这些结果证实,使用本文所述方法能够针对以高复杂性和高特异性测序的模拟PM和模拟AM样本库来确定一级至三级关系。图27B显示了来自一个DI为2.9,与完整自身样本配对的退化模拟PM样本(自身,RF016),以及包括已知的母亲(一级)、已知的姑/姨(二级)、已知的堂/表亲(三级)、已知的隔一代的堂/表亲(四级)和已知的第二代的堂/表亲(五级)的模拟AM样本的全基因组亲属关系系数和对应的对数似然比的示例。使用全基因组算法来确定自身(RF016)(退化指数为2.9)模拟死后样本与完整模拟生前(AM)样本的亲属关系。图27B中描绘的水平虚线表示亲属关系系数阈值>0.031。跨6对的SNP重叠平均为7400,最小值为7061并且最大值为7755(图27B)。自身、母亲、姑/姨、堂/表亲、隔一代的堂/表亲和第二代的堂/表亲的对数似然比分别为167、94.5、26.1、8.9、2.2和0.4。除了隔一代的堂/表亲和第二代的堂/表亲(四级和五级)外,所有关系均通过了使用本文所述方法论的窗口亲属关系和全基因组算法所需的0.031阈值,从而证实了针对使用本文所述方法识别一级、二级和三级亲属使用这些算法的有效性。

[0698] C.讨论

[0699] 利用DNA进行灾难受害者身份识别(DVI)通常包括询问基因组中高度多态性的区域,特别是常染色体短串联重复序列(STR)、Y染色体上的STR和线粒体DNA(mtDNA)(Alonso等人,Croat Med J2005,46,540-548;Ambers等人,Int J Legal Med 2018,132,1545-1553;Alvarez-Cubero等人,Pathobiology 2012,79,228-238;Watherston等人,Forensic Sci Int Genet 2018,37,270-282;Prinz等人,Forensic Sci Int Genet 2007,1,3-12,其内容在此以引用方式整体并入本文)。STR标记已经用于识别多起大规模死亡事件(MFI)中的遗骸,但只能提供足够的信息来识别一级和二级关系,并可能导致较高的假阳性率(Alonso等人,Croat Med J 2005,46,540-548;Alvarez-Cubero等人,Pathobiology 2012,79,228-238;Birus等人,Croat Med J2003,44,322-326;Brenner等人,Theor Popul Biol 2003,63,173-178;Graham等人,Forensic Sci Med Pathol 2006,2,203-207,其内容在此以引用方式整体并入本文)。用于STR分析的下一代测序(NGS)测定法的开发已经克服了利用STR进行DVI的一些障碍(Alvarez-Cubero等人,Ann Hum Biol 2017,44,581-592;Zavala等人,DNA退化对基于大规模并行测序的常染色体STR、iiSNP和线粒体DNA分型系统的影响,2019;Senst等人,Forensic Science International:Genetics 2023,62;Senst等人,J

Forensic Sci 2022,67,1382-1398;Ambers等人,BMC Genomics 2016,17,750,其内容在此以引用方式整体并入本文)。mtDNA控制区域或高变区域或整个线粒体基因组(mtGenome)也已用于识别遗骸(Holland等人,线粒体DNA序列分析-法医案例的验证和使用。法医科学评论1999;Holland等人,Croat Med J 2003,44,264-272,其内容在此以引用方式整体并入本文)。mtGenome具有高拷贝数并且是环状基因组;因此,与片段化、老化遗骸中的核DNA相比,mtDNA的恢复几率更高(Amorim等人,PeerJ 2019,7,e7314,doi:10.7717/peerj.7314,其内容在此以引用方式整体并入本文)。然而,mtDNA对母系进行分型,如果多个家族成员受到MFI影响,或者只有父系相关亲属可用于亲属关系分析,则可能会出现问题(Zavala等人,DNA退化对基于大规模并行测序的常染色体STR、iiSNP和线粒体DNA分型系统的影响,2019;Holland等人,线粒体DNA序列分析-法医案例的验证和使用。法医科学评论1999;Holland等人,Croat Med J 2003,44,264-272,Amorim等人,Peer J,7,e7314,doi:10.7717/peerj.7314)。相反,单核苷酸多态性SNP)能够用于确定父系或母系关系,或两者兼而有之(Watherston等人,Forensic Sci Int Genet 2018,37,270-282;Amorim等人,Forensic Sci Int 2005,150,17-21;Gorden等人,Forensic Sci Int Genet 2022,57,102636)。此外,对于基于PCR的测定法,SNP扩增子往往比STR的扩增子短,并且更有可能在受损样本中被扩增(Watherston等人,Forensic Sci Int Genet 2018,37,270-282;Zavala等人,DNA退化对基于大规模并行测序的常染色体STR、iiSNP和线粒体DNA分型系统的影响,2019;Senst等人,J Forensic Sci 2022,67,1382-1398;Ambers等人,BMC Genomics 2016,17,750,其内容在此以引用方式整体并入本文)。并且,由于要询问的SNP数量很多,因此亲属关系以及尤其是似然比值将比STR更具鉴别性(Turner等人,Front Genet 2022,13,882268;Cho等人,Transfus Med Hemother 2016,43,429-432;Tillmar等人,Genes 2021,12,1968)。尽管STR比SNP更具多态性(Devesse等人,Forensic Sci Int Genet 2018,34,57-61;Phillips等人,ELECTROPHORESIS2018,39,2708-2724;Gettings等人,Forensic Science International:Genetics 2016,21,15-21;Novroski等人,Forensic Sci Int Genet 2016,25,214-226;Wendt等人,Forensic Sci Int Genet 2017,28,146-154;Delest等人,Forensic Sci Int Genet,2020,,47,102304,其内容在此以引用方式整体并入本文),SNP测定法需要询问数百到数千个数据点,而STR只需要询问数十个数据点。SNP数据的高密度可以降低假阳性率(Snedecor等人,Forensic Sci Int Genet 2022,61,102769,其内容在此以引用方式整体并入本文)。最后,SNP以前曾在DNA退化的情况下使用,并被证明能够提供足够的鉴别能力来识别遗骸(Snedecor等人,Forensic Sci Int Genet 2022,61,102769;Gorden等人,Forensic Sci Int Genet 2022,57,102636;Marshall等人,Genes(Basel)2020,11,doi:10.3390/genes11080938,其内容在此以引用方式整体并入本文)。

[0700] ForenSeq Kintelligence Library PrepKit®包含10,230个法医相关的SNP集合,其能够用于利用法医遗传学来解决暴力犯罪和失踪人员案件(Kling等人,Forensic Sci Int Genet 2021,52,102474;Snecedor等人,Forensic Science International:Genetics 2022,61,102769;Peck等人,ForenSeq Kintelligence Kit在法医遗传学中的应用的内部验证。bioRxiv 2022,2022.2010.2028.514056,doi:10.1101/2022.10.28.514056;Verogen.ForenSeq Kintelligence Kit数据表:文档#VD2020054.Rev.A.2021.在线获取:https://verogen.com/wp-content/uploads/2021/03/

forenseq-kintelligence-datasheet-vd2020054-a.pdf;Verogen.使用ForenSeq Kintellingence从低质量样本中获得高质量结果应用说明:文档#VD2021002 Rev.B.2021.2021,其内容在此以引用方式整体并入本文)。ForenSeq Kintelligence已用于解决俄勒冈州一起身份不明遗骸案件(International,DL.DNA.Labs International是第一家使用ForenSeq Kintelligence系统协助识别身份不明遗骸的认证实验室。2022)。Kintelligence结果可以上传到数据库,诸如GEDMatch PRO或FamilyTreeDNA(Verogen.Verogen和Gene by Gene建立开创性合作伙伴关系以加速法医调查基因遗传学的应用。可在线获取:https://www.businesswire.com/news/home/20220815005116/en/Verogen-and-Gene-by-Gene-Form-Groundbreaking-Partnership-to-Accelerate-Adoption-of-Forensic-Investigative-Genetic-Genealogy),以搜索数据库中可用的未知亲属。GEDMatch PRO中使用的算法专门设计用于处理Kintelligence试剂盒中的10,230个SNP,但需要上传到公共数据库才能搜索亲属,并且上传时样本中至少需要输入6000个SNP。此外,Kintelligence试剂盒的当前配置允许在Miseq FGx上一次最多对3个样本进行测序,确保为GEDMatch PRO上传分型足够的SNP,但也降低了MFI案例的成本效益,因为这些案例不希望有远亲关系。此外,如果亲属已知但不存在于两个数据库中的一个,则匹配死后(PM)样本将需要已知亲属将其谱上传到数据库。

[0701] 协助利用库制备试剂盒和类似的方法论来生成核酸库,诸如ForenSeq Kintelligence Library PrepKit®,以更高的通量和成本效益的方式用于DVI,以超过每次测序运行3个库的推荐复杂度对库进行测序。目标是最大化生前(AM)和死后(PM)样本这两者的复杂度,同时维持识别关系的准确性多达三级。

[0702] 基于模拟(图20A)和在MiSeq FGx上对模拟AM和PM样本进行测序(图20B、20C、24和27A),多达三级关系确定的最佳测序复杂度为每次测序运行12个模拟PM库或32个模拟AM库。尽管与每次测序3个样本的推荐复杂度相比,针对以更高复杂度的文库测序SNP等位基因和基因座降到低于AT(图21A和21B),但在模拟AM和PM样本之间有足够的被共同分型的基因座,即使对于高度退化和低DNA输入样本,也允许确定多达三级的关系(图22、25A、25B和27B)。

[0703] 对相关家族(RF)进行的分析证实,除了识别一些四级关系外,还识别出了三级关系(图27B)。具有四级关系的样本之间的SNP重叠范围为1356-7872个SNP。在30个可能的配对中,重叠超过7000个SNP的8对没有通过亲属关系确定的阈值。如果PM样本的质量足够好以维持高SNP调用率,则可以确定四级关系,但可能不会通过阈值(图22A-D)。

[0704] 如关于图21A和21B的研究所述,在模拟PM分析中观察到SNP调用率降低和等位基因脱扣增加,以及如图27A的研究所示,对退化的相关家族样本进行分析,如图26所示的谱系,PM样本和AM样本之间调用的基因座的重叠将随着PM DNA质量的降低而减少。这种重叠显著影响了亲属关系算法的性能,降低了关系程度较高时的灵敏度(图22A-D)。因此,必须最大化在AM样本中分型SNP的数量,以防止基因座的进一步丢失。如果在寻找MFI时预期有更远的关系,并且PM样本显示出低SNP调用率,则能够以更低的复杂度对这些库进行重新测序以确保调用足够的SNP。

[0705] 在图20A的上下文中进行的高通量模拟和在图20B和20C的上下文中执行的实际高通量分析表明,复杂度的增加将减少分型SNP的数量。PM样本的退化水平越高和DNA量越低

将进一步影响分型SNP的数量。正如从GEDMatch数据库采样匿名亲属和通过随机将SNP数量减少到8000、6000、4000和2000来确定亲属关系的研究所表明的,全基因组亲属关系(使用亲属关系系数而不是共享cM来识别关系)能够以高水平的特异性和敏感性来识别多达三级亲属关系(图22A-D)。降低亲属关系系数阈值将提高灵敏度(可能匹配的数量);然而,特异性会受到影响,并且会出现更多的假阳性(将被称为真实匹配的假匹配的数量)。当前研究中使用的阈值导致了更高的特异性(近100%),有利于丢失真阳性以避免假阳性的增加。此外,在这些较低的SNP调用率下,识别更远的关系(诸如四级和五级关系)变得困难。全基因组亲属关系和窗口亲属关系算法都对被测试的两个个体之间的分型SNP的数量敏感。在相关家族分析中,其包括退化和低输入样本这两者,我们能够识别出所有关系,多达和包括三级关系。然而,由于分型SNP的平均数量太低,无法指代四级和五级关系的窗口亲属关系(少于6000个),我们无法利用窗口亲属关系算法,该算法更适用于这些更远的关系。全基因组亲属关系算法在识别四级和五级关系方面不如窗口亲属关系算法具体或敏感。

[0706] 杂合性是数据脱扣的另一种度量,如果太低,其可能会影响似然比和亲属关系值的准确性。杂合性的丢失是由于样本的高度退化或DNA的低含量造成的。即使在相关家族样本中观察到展现7.2%杂合性的退化样本,也捕获了所有一级、二级和三级关系。用展现22.6%杂合性的模拟PM样本来确定四级关系。结合在退化和低输入样本中观察到的基因座脱扣水平,预期将能够同时测序的样本数量增加到针对退化和/或低DNA输入样本的12个和针对AM样本的32个,将能够捕获多达和包括三级关系的所有关系。

[0707] 如果预期PM样本是四级或五级亲属,建议通过减少同时测序的样本数量来最大化SNP重叠,特别是在输入低和/或样本高度退化的情况下。或者,如果预期PM样本是一级、二级或三级亲属,那么即使针对退化和/或低输入样本,同时测序的样本数量更高(多达12个)也不应影响识别一级至三级关系的亲属关系的能力。或者,针对AM样本,如上所述,必须最大化分型SNP的数量;然而,据观察,使用所述亲属关系算法和似然比这两者,多达32个同时测序的样本能够识别多达和包括所有三级关系。

[0708] 总之,根据本文所述的方法论,以较高多路复杂度对库进行测序,通过将针对PM样本的多路复杂度从3提高到12以及针对AM样本的多路复杂度提高到32,改进了关系身份识别的成本效益。算法中设置的高度调整的阈值和大面板的SNP将识别关系的假阳性率降至零,同时以完美的灵敏度和特异性识别多达三级的所有关系(例如,堂/表亲或曾祖父母)。这些亲属关系算法(其安装在具有用于似然比计算的额外方法的私人服务器上)可以作为私人数据库,能够找到亲属并正确确定MFI的关系,再加上更高的多路复杂度测序的成本效益,是DVI的解决方案。

[0709] 本发明并不旨在将范围限制在特定公开的示例中,这些示例例如用于说明本发明的各个方面。从本文的描述和教导中,对所述组合物和方法的各种修改将变得显而易见。可以在不背离本公开内容的真实范围和精神的情况下实施这样的变化,并且旨在落入本公开内容的范围内。

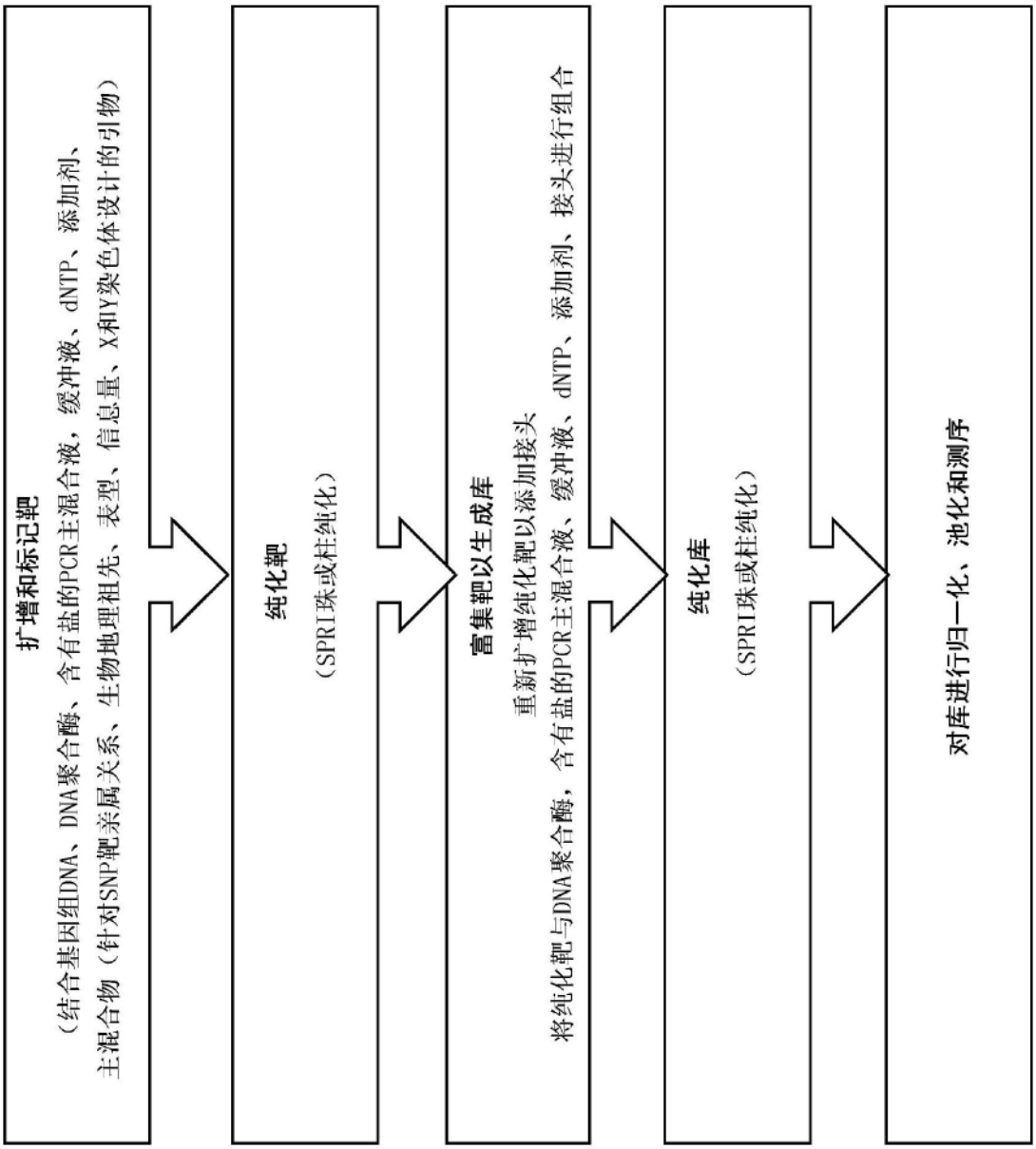

扩增和标记靶
（结合基因组DNA、DNA聚合酶、含有盐的PCR主混合液，缓冲液、dNTP、添加剂、
主混合物（针对SNP靶亲属关系、生物地理祖先、表型、信息量、X和Y染色体设计的引物）
纯化靶
（SPRI珠或柱纯化）
富集靶以生成库
重新扩增纯化靶以添加接头
将纯化靶与DNA聚合酶，含有盐的PCR主混合液、缓冲液、dNTP、添加剂、接头进行组合
纯化库
（SPRI珠或柱纯化）
对库进行归一化、池化和测序

图1

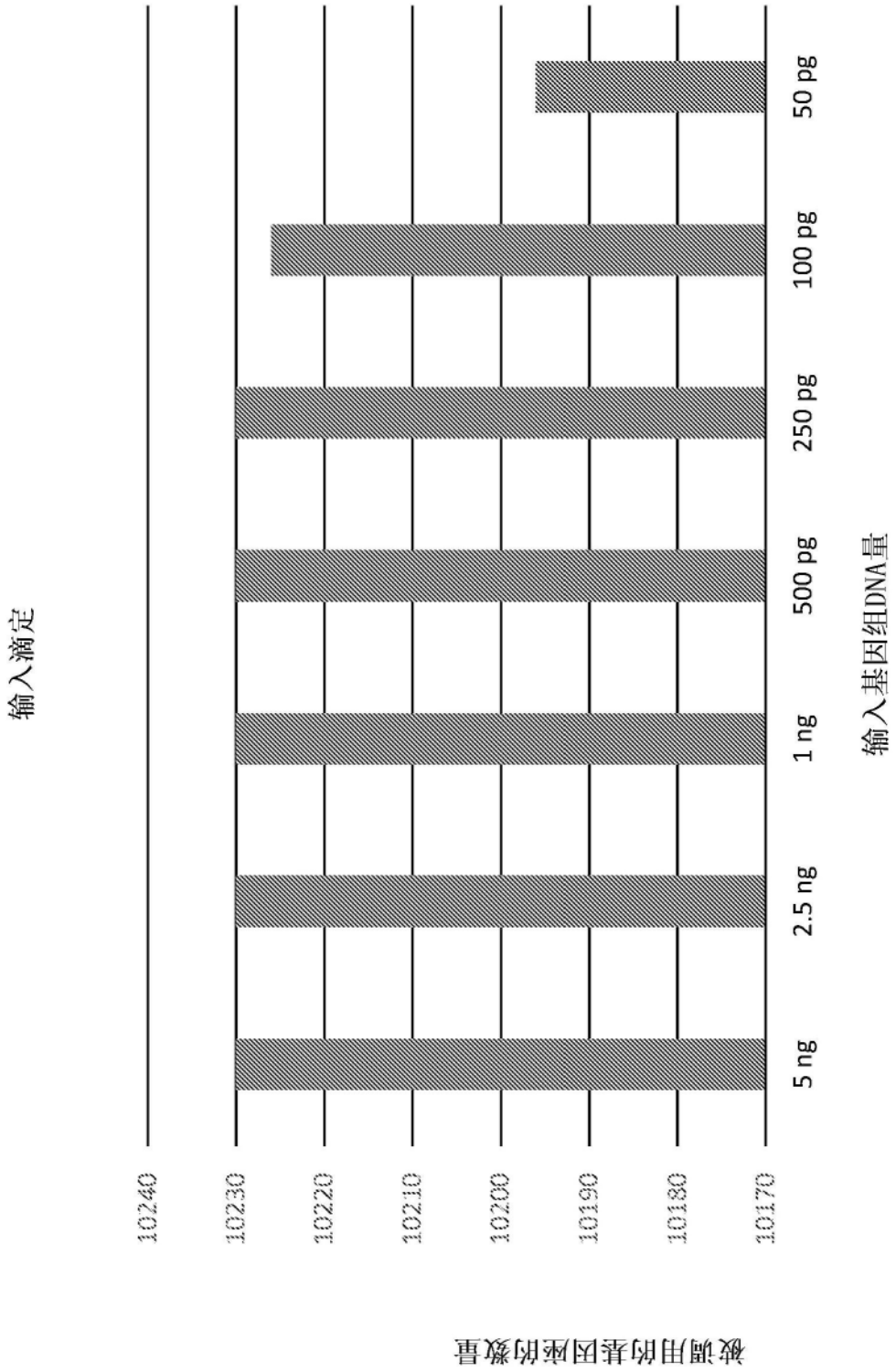

输入滴定
10240
10230
10220
10210
10200
10190
10180
10170
被调用的基因座的数量
5 ng
2.5 ng
1 ng
500 pg
250 pg
100 pg
50 pg
输入基因组DNA量

图2

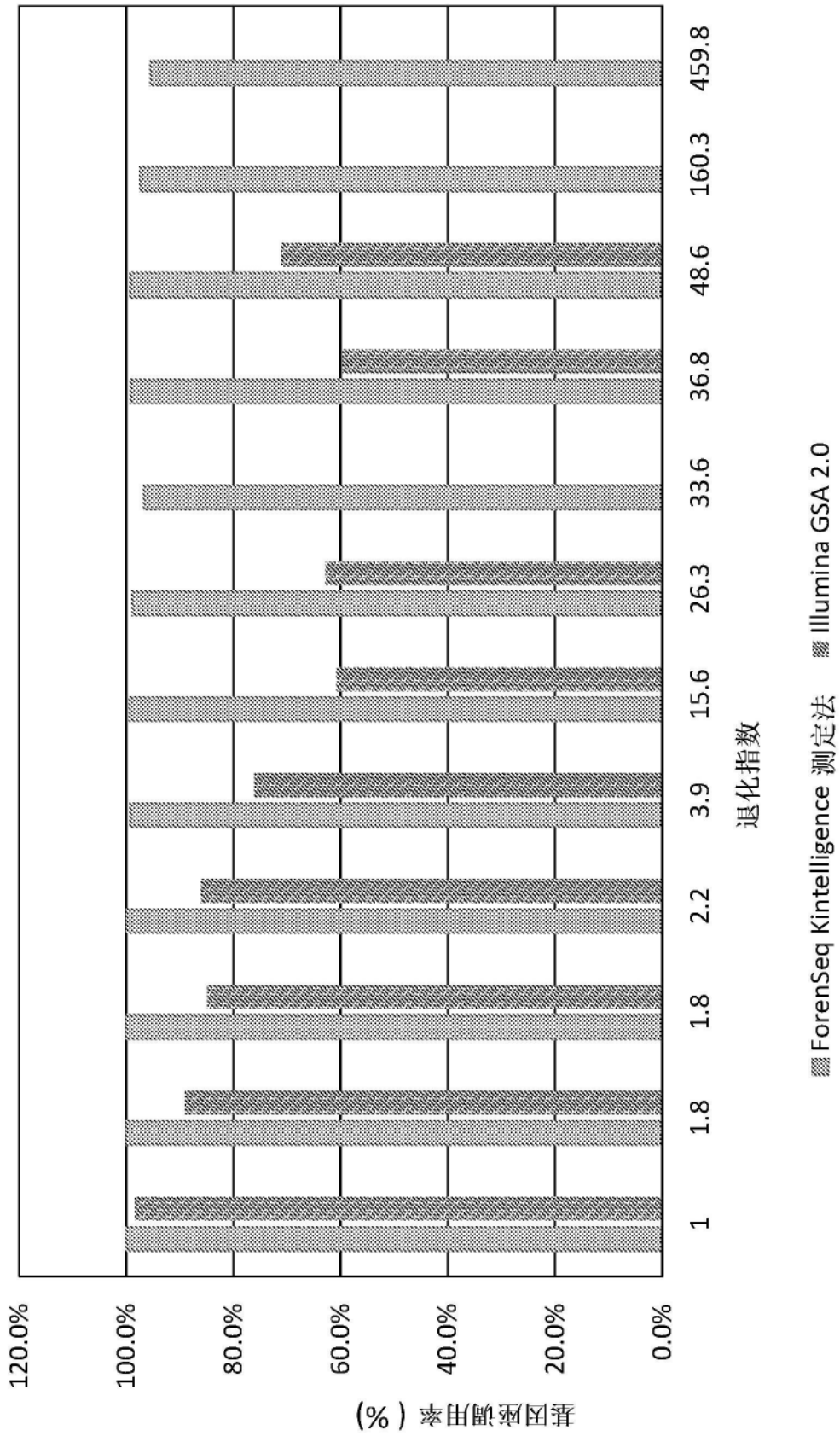

基因座调用率（%）
120.0%
100.0%
80.0%
60.0%
40.0%
20.0%
0.0%
1
1.8
1.8
2.2
3.9
15.6
26.3
33.6
36.8
48.6
160.3
459.8
退化指数
ForenSeq Kintelligence 测定法
Illumina GSA 2.0

图3

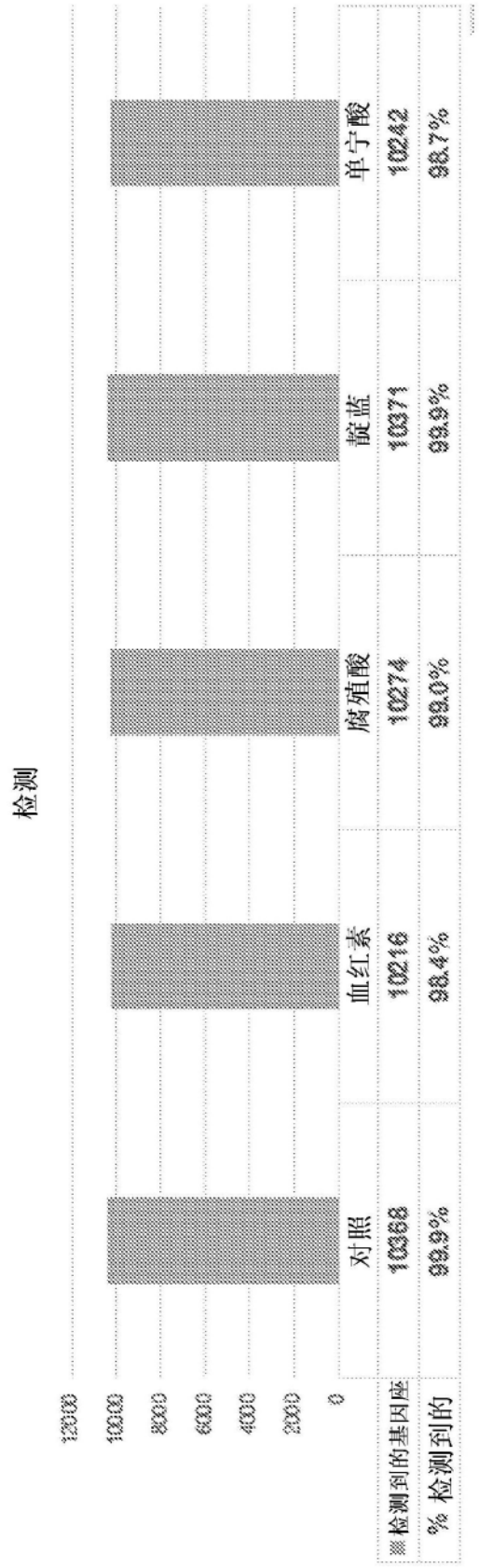

检测
12000
10000
8000
6000
4000
2000
0
对照
血红素
腐殖酸
靛蓝
单宁酸
※检测到的基因座
10368
10216
10274
10371
10242
% 检测到的
99.9%
98.4%
99.0%
99.9%
98.7%

图4

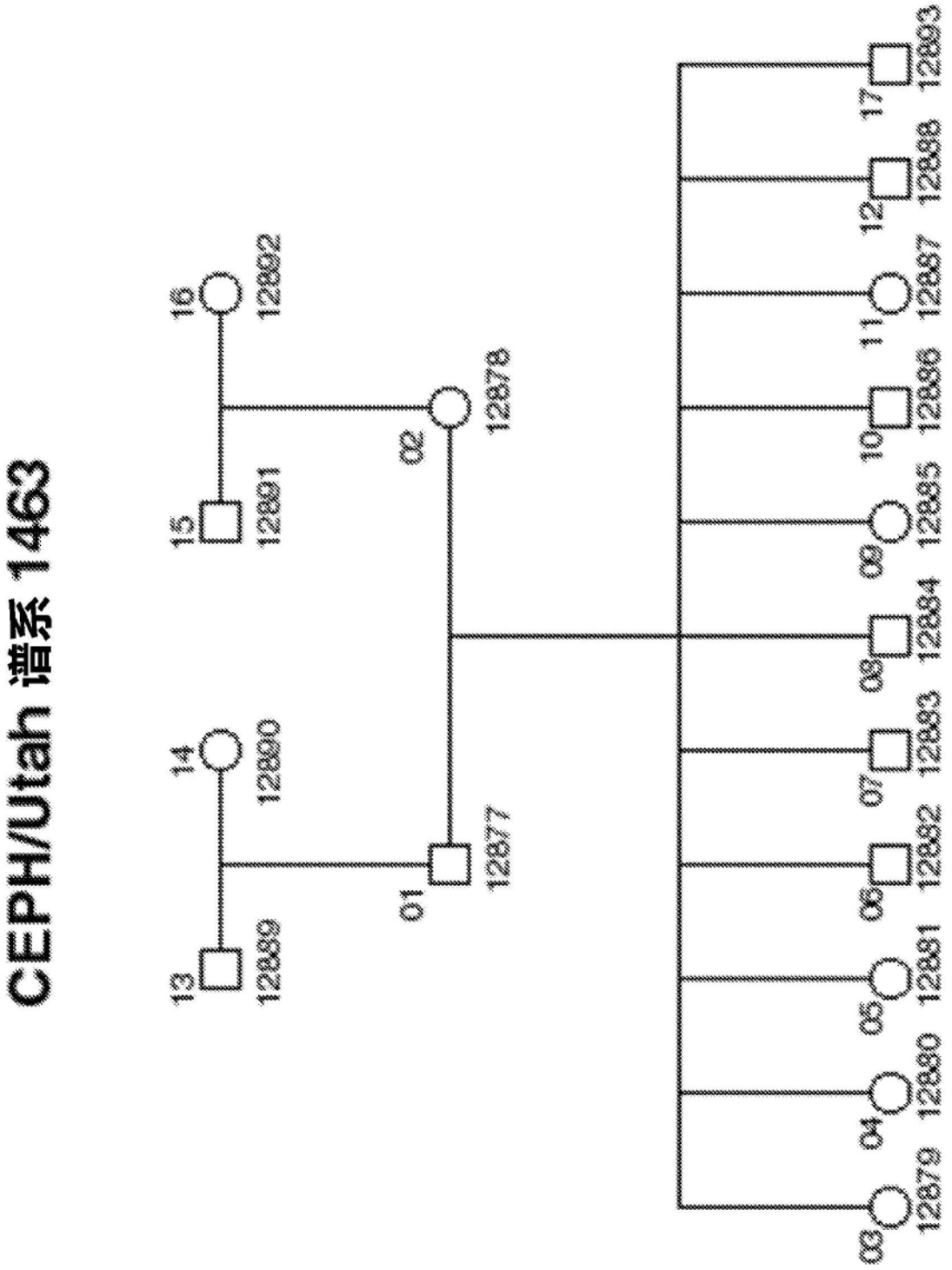

CEPH/Utah 谱系 1463
01
12877
02
12878
03
12879
04
12880
05
12881
06
12882
07
12883
08
12884
09
12885
10
12886
11
12887
12
12888
13
12889
14
12890
15
12891
16
12892
17
12893

图5

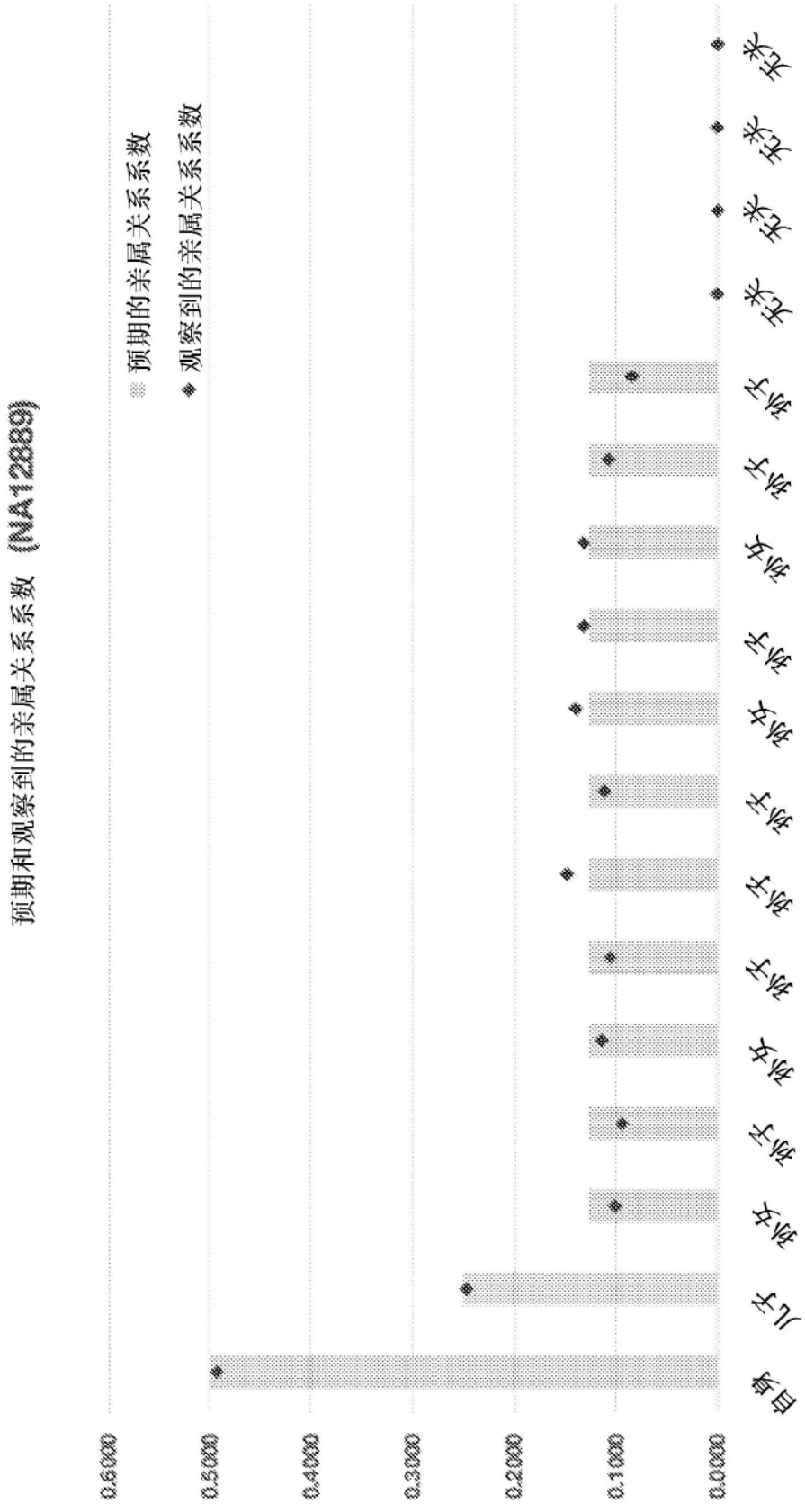
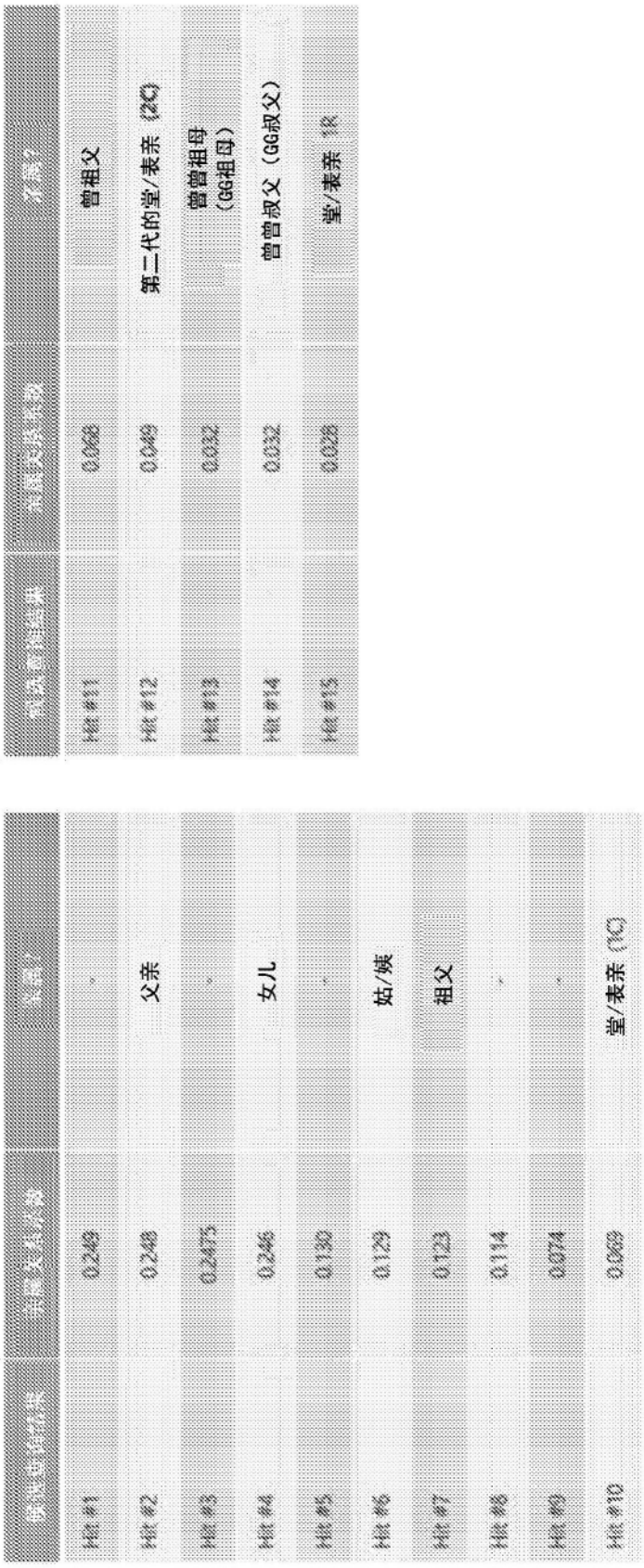

预期和观察到的亲属关系系数 (NA12889)
预期的亲属关系系数
观察到的亲属关系系数
0.6000
0.5000
0.4000
0.3000
0.2000
0.1000
0.0000
自身
儿子
孙女
孙子
孙女
孙子
孙子
孙子
孙女
孙子
孙女
孙子
孙子
无关
无关
无关
无关

图6



例性查询结果	余属支持系数	实属?
Hit #1	0.249	-
Hit #2	0.248	父亲
Hit #3	0.2475	-
Hit #4	0.246	女儿
Hit #5	0.130	-
Hit #6	0.129	姑/姨
Hit #7	0.123	祖父
Hit #8	0.114	-
Hit #9	0.074	-
Hit #10	0.069	堂/表亲 (1C)

例性查询结果	余属支持系数	实属?
Hit #11	0.068	曾祖父
Hit #12	0.049	第二代的堂/表亲 (2C)
Hit #13	0.032	曾曾祖母 (GG祖母)
Hit #14	0.032	曾曾叔父 (GG叔父)
Hit #15	0.028	堂/表亲 1R

图7

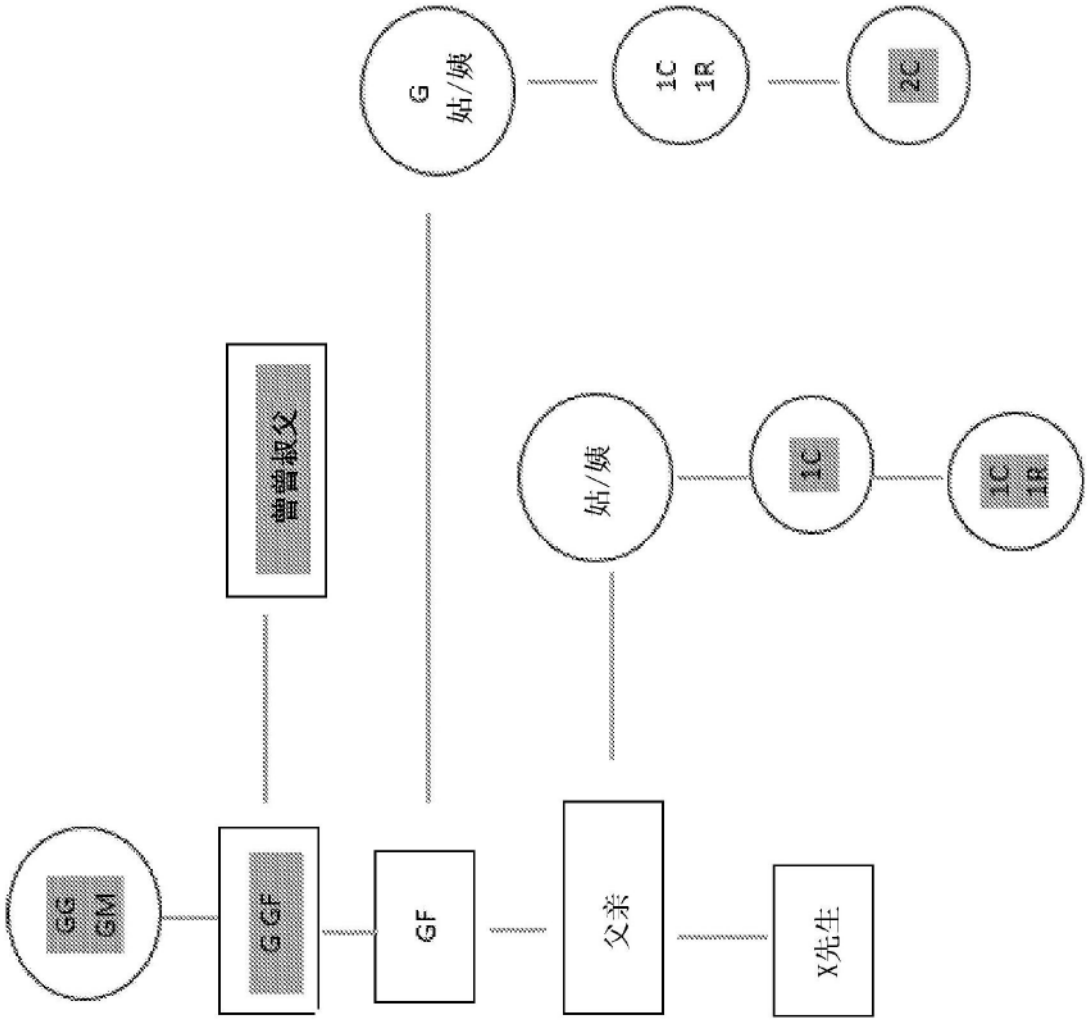
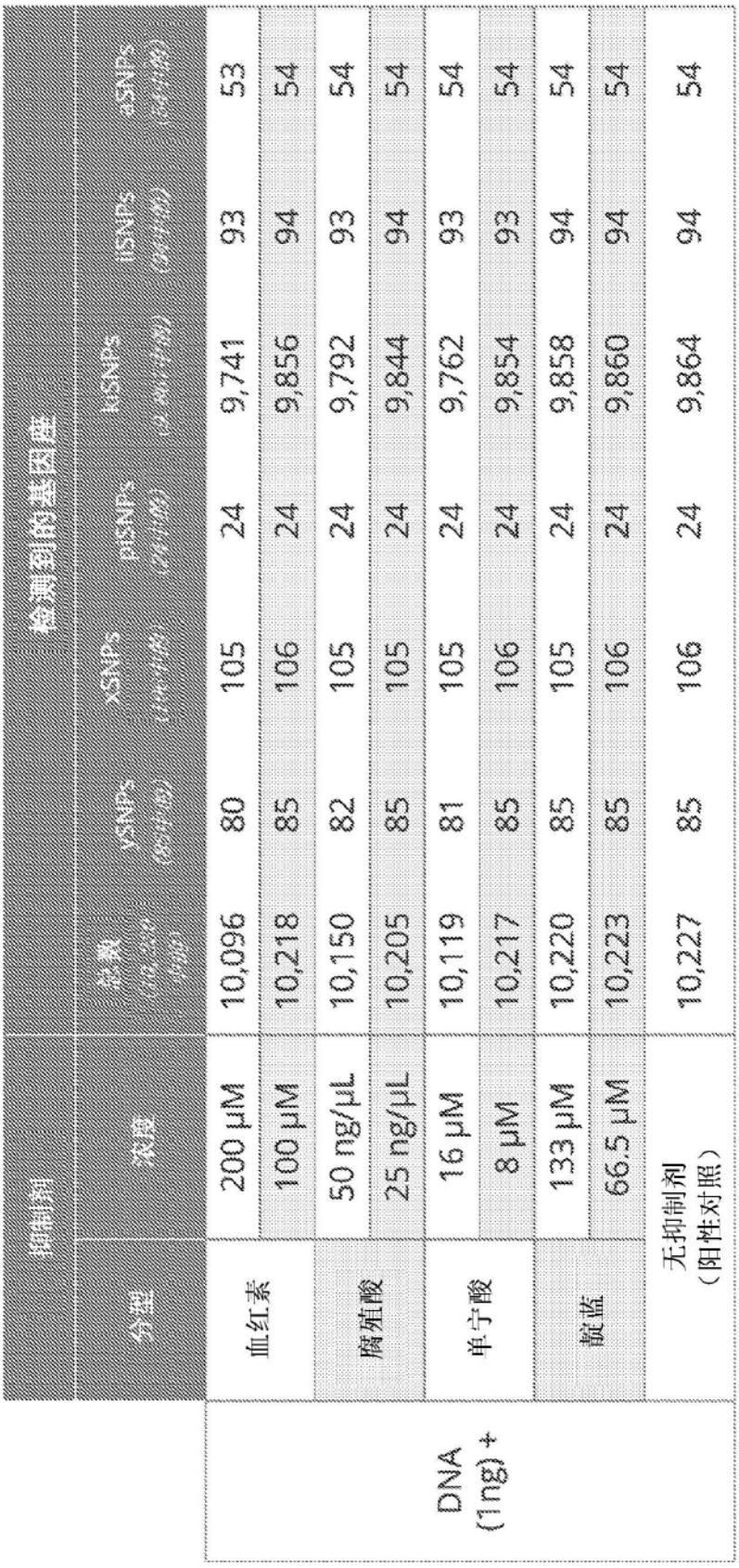


图8

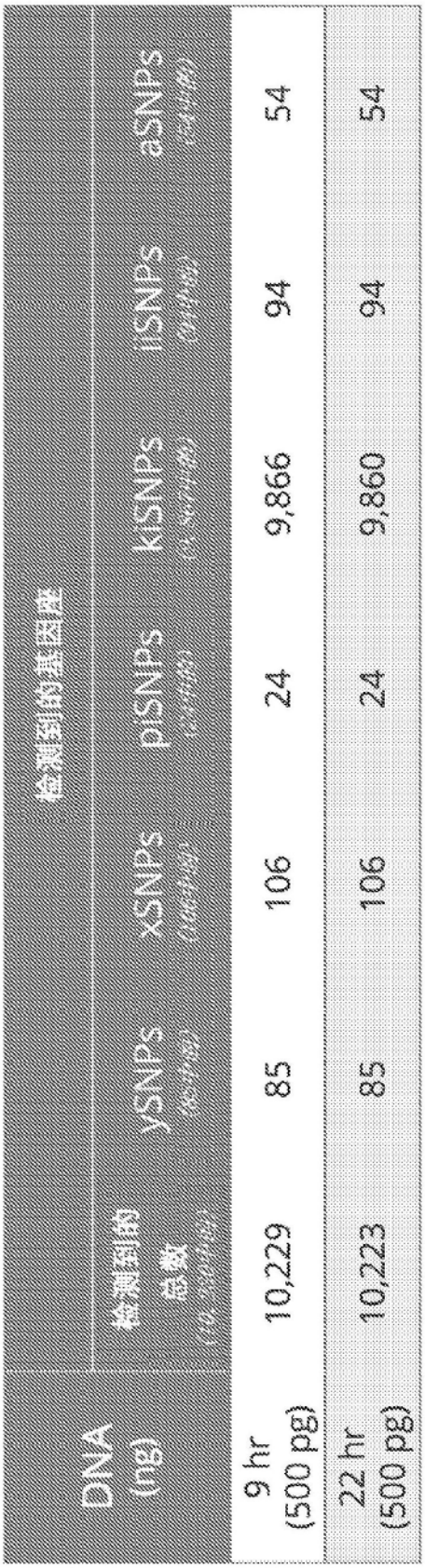
输入DNA (ng)	检测到的基因座总数	YSNPs	XSNPs	piSNPs	kiSNPs	iiSNPs	eSNPs
	(二次重复的平均值)						
5	10,225	85	106	24	9,863	93	54
2.5	10,223	85	106	24	9,861	93	54
1	10,221	85	106	24	9,859	93	54
0.5	10,221	85	106	24	9,859	93	54
0.25	10,210	85	105	24	9,849	93	54
0.10	10,181	85	104	24	9,821	93	54
0.05	10,119	82	104	24	9,762	93	54

图9



	抑制剂		检测到的基因座						
	分型	浓度	总数 (10,230中的)	ySNPs (85中的)	xSNPs (106中的)	piSNPs (24中的)	kiSNPs (9,867中的)	iiSNPs (94中的)	aSNPs (54中的)
DNA (1ng) +	血红素	200 μM	10,096	80	105	24	9,741	93	53
		100 μM	10,218	85	106	24	9,856	94	54
	腐殖酸	50 ng/μL	10,150	82	105	24	9,792	93	54
		25 ng/μL	10,205	85	105	24	9,844	94	54
	单宁酸	16 μM	10,119	81	105	24	9,762	93	54
		8 μM	10,217	85	106	24	9,854	93	54
	靛蓝	133 μM	10,220	85	105	24	9,858	94	54
		66.5 μM	10,223	85	106	24	9,860	94	54
	无抑制剂（阳性对照）		10,227	85	106	24	9,864	94	54

图10



DNA (ng)	检测到的总数 (10,230个数)	检测到的基因座					
		ySNPs (85个数)	xSNPs (106个数)	piSNPs (24个数)	kiSNPs (9,867个数)	iiSNPs (94个数)	aSNPs (54个数)
9 hr (500 pg)	10,229	85	106	24	9,866	94	54
22 hr (500 pg)	10,223	85	106	24	9,860	94	54

图11

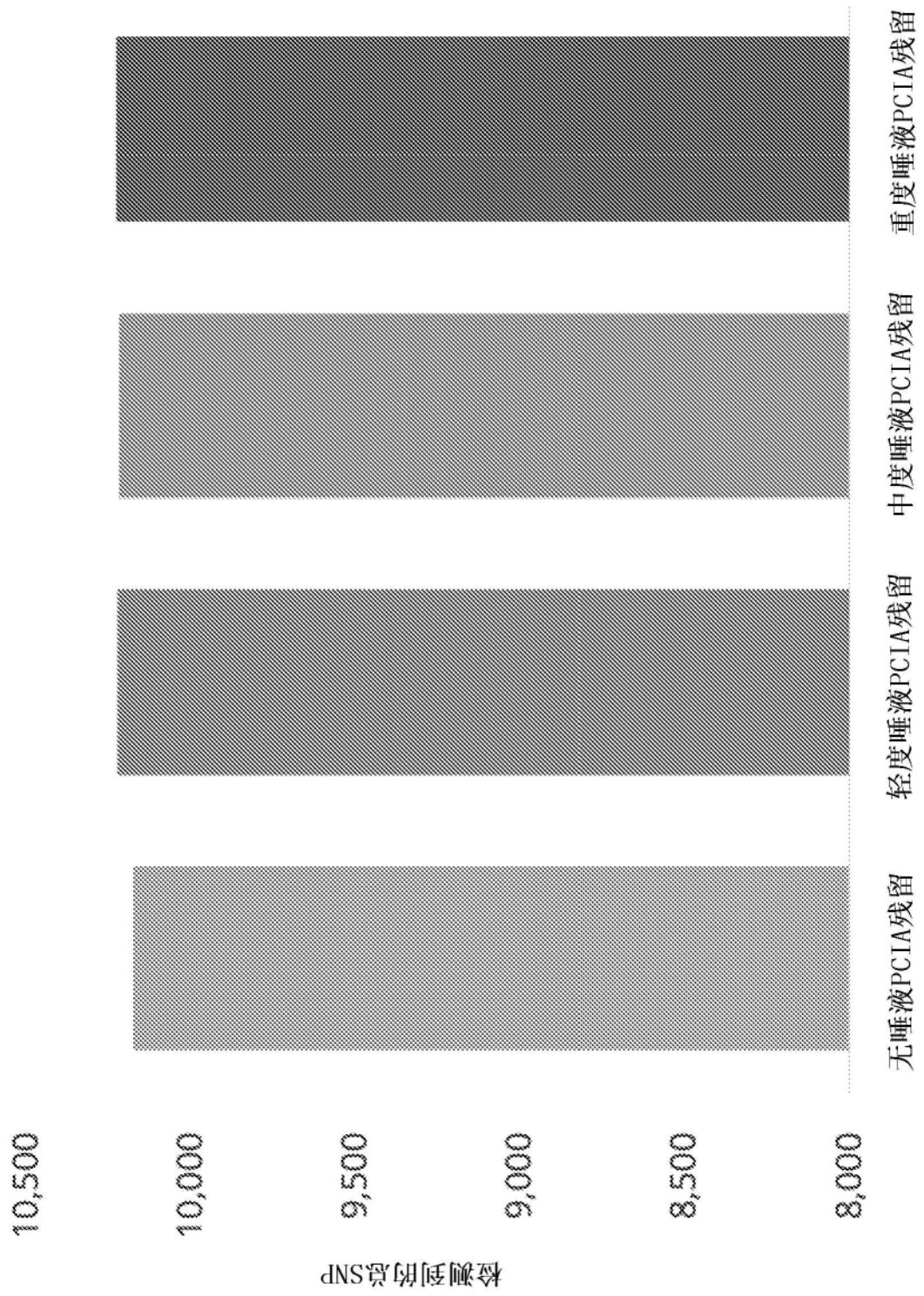

检测到的总SNP
10,500
10,000
9,500
9,000
8,500
8,000
无唾液PCIA残留
轻度唾液PCIA残留
中度唾液PCIA残留
重度唾液PCIA残留

图12

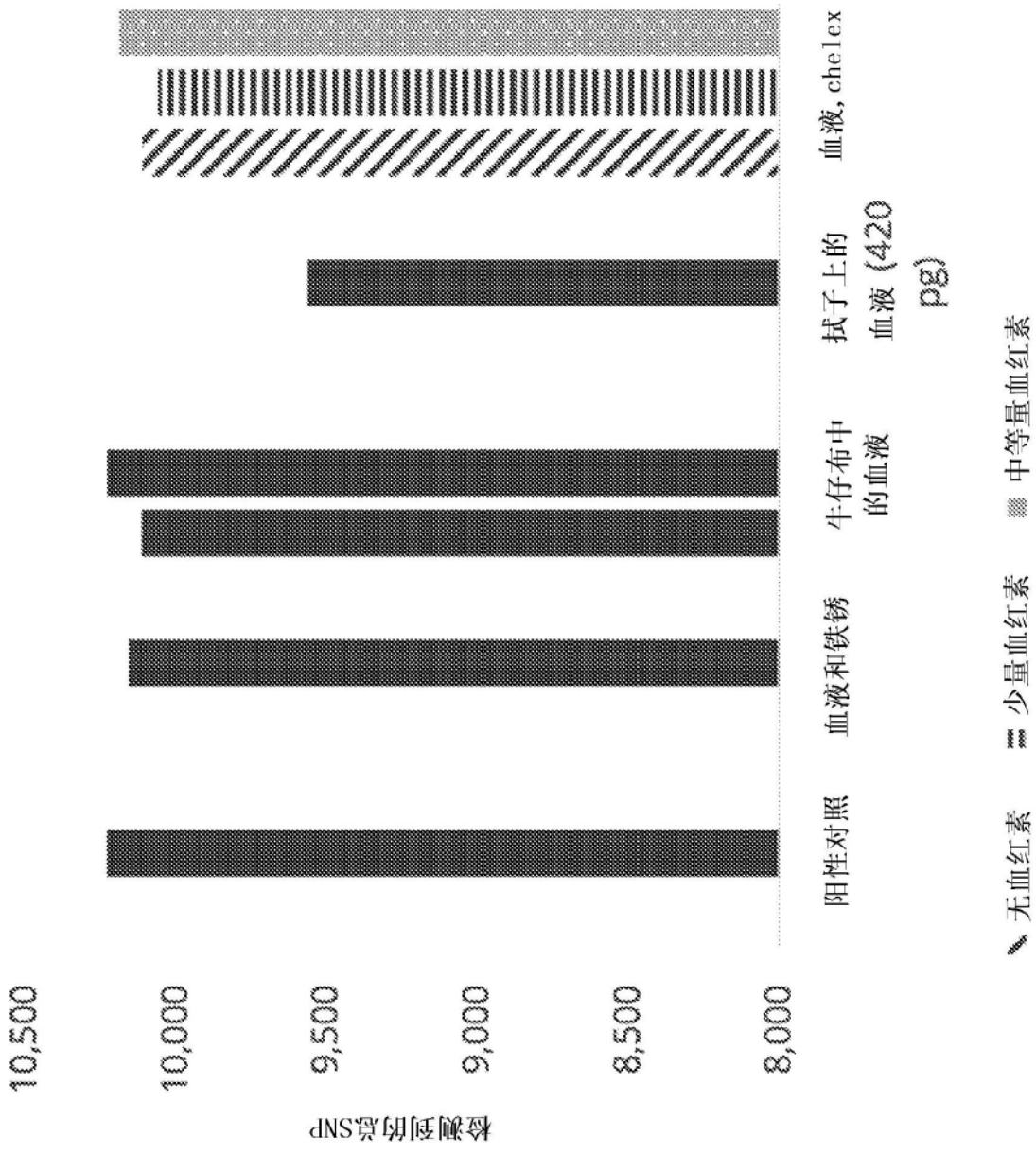

检测到的总SNP
10,500
10,000
9,500
9,000
8,500
8,000
阳性对照
血液和铁锈
牛仔布中的血液
拭子上的血液(420 pg)
血液, chelex
无血红素
少量血红素
中等量血红素

图13

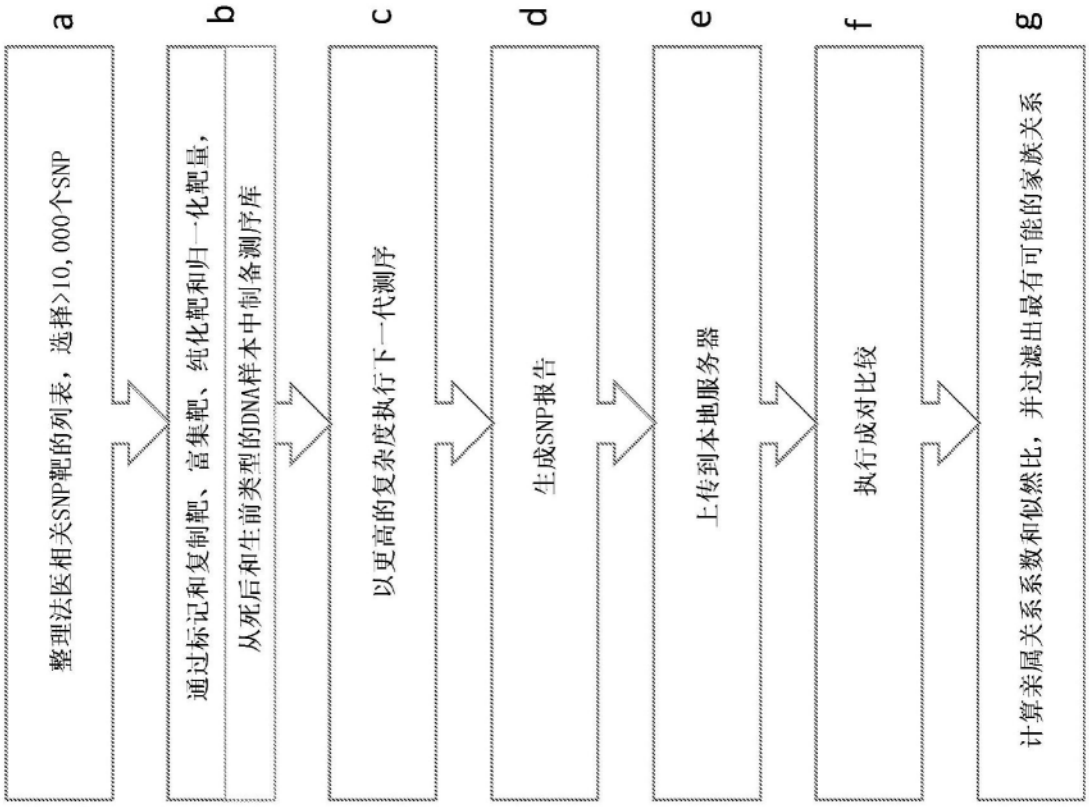

a
整理法医相关SNP靶的列表，选择>10,000个SNP
b
通过标记和复制靶、富集靶、纯化靶和归一化靶量，
从死后和生前类型的DNA样本中制备测序库
c
以更高的复杂度执行下一代测序
d
生成SNP报告
e
上传到本地服务器
f
执行成对比较
g
计算亲属关系系数和似然比，并过滤出最有可能的家族关系

图14

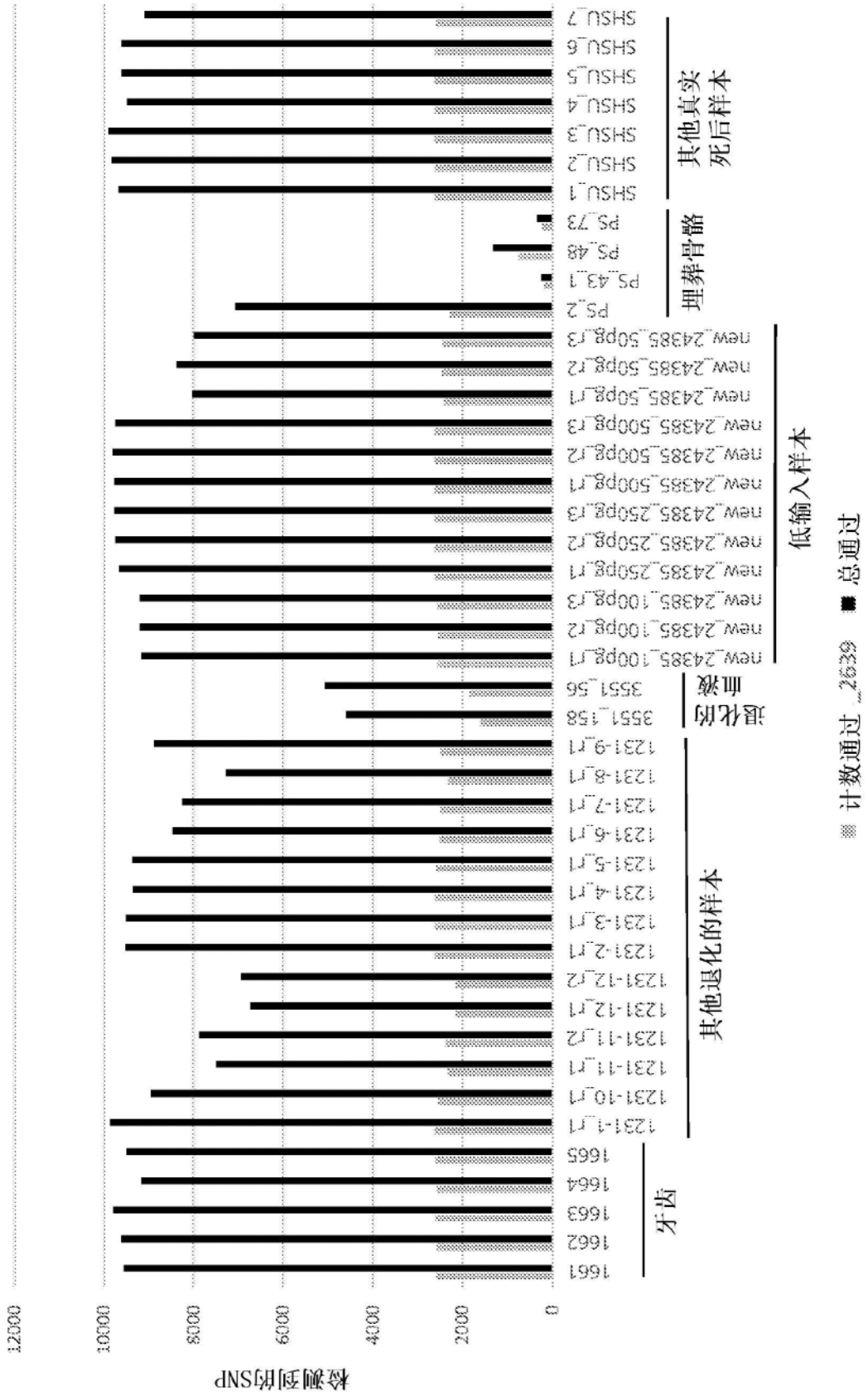

检测到的SNP
0
2000
4000
6000
8000
10000
12000
计数通过_26339
总通过
牙齿
其他退化的样本
退化的血液
低输入样本
埋葬骨骼
其他真实死后样本
1661
1662
1663
1664
1665
1231-1_r1
1231-10_r1
1231-11_r1
1231-11_r2
1231-12_r1
1231-12_r2
1231-2_r1
1231-3_r1
1231-4_r1
1231-5_r1
1231-6_r1
1231-7_r1
1231-8_r1
1231-9_r1
3551_158
3551_56
new_24385_100pg_r1
new_24385_100pg_r2
new_24385_100pg_r3
new_24385_250pg_r1
new_24385_250pg_r2
new_24385_250pg_r3
new_24385_500pg_r1
new_24385_500pg_r2
new_24385_500pg_r3
new_24385_50pg_r1
new_24385_50pg_r2
new_24385_50pg_r3
PS_2
PS_43_1
PS_48
PS_73
SHSU_1
SHSU_2
SHSU_3
SHSU_4
SHSU_5
SHSU_6
SHSU_7

图15

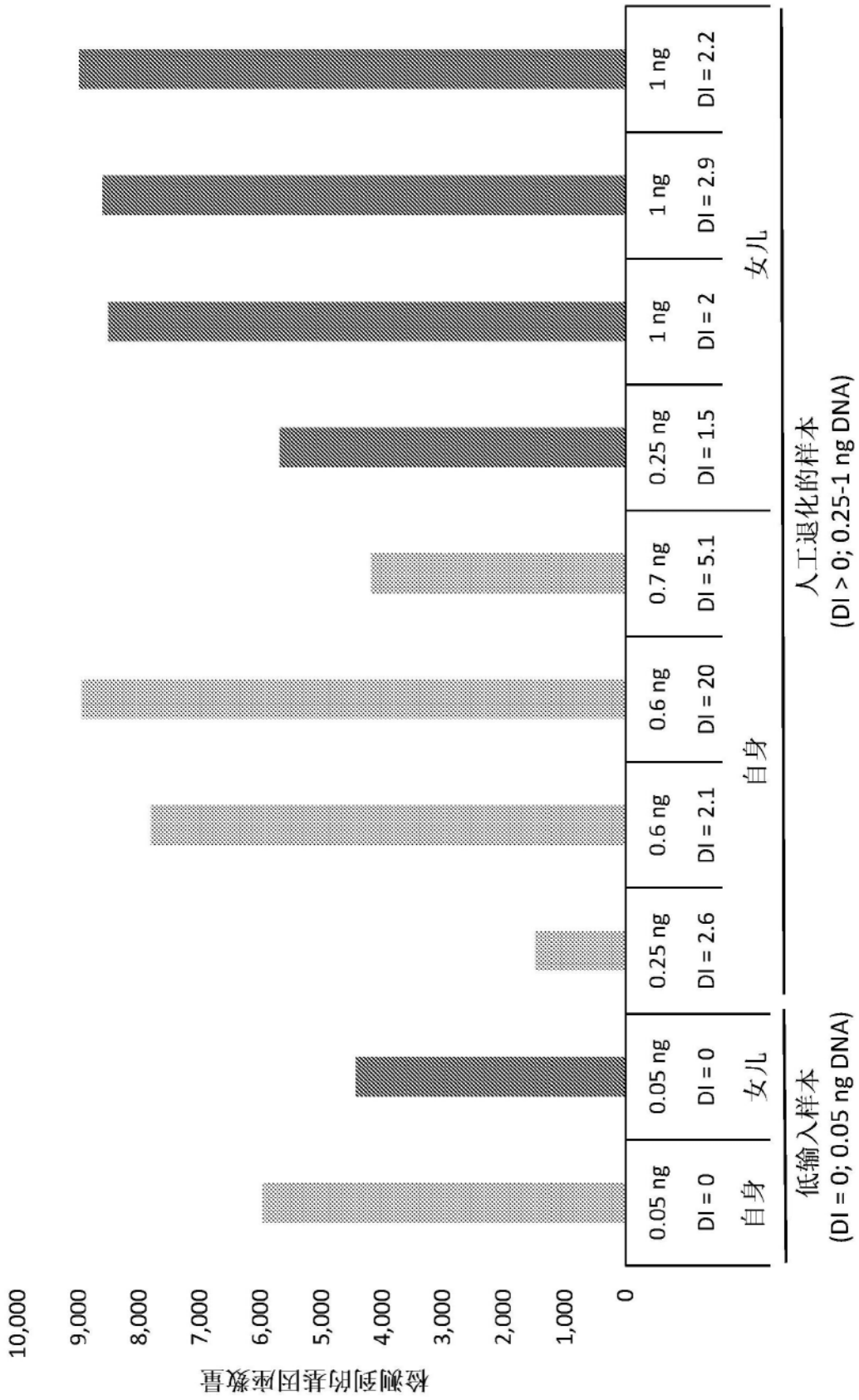

检测到的基因座数量
10,000
9,000
8,000
7,000
6,000
5,000
4,000
3,000
2,000
1,000
0
0.05 ng
DI = 0
自身
0.05 ng
DI = 0
女儿
0.25 ng
DI = 2.6
0.6 ng
DI = 2.1
0.6 ng
DI = 20
0.7 ng
DI = 5.1
0.25 ng
DI = 1.5
1 ng
DI = 2
1 ng
DI = 2.9
1 ng
DI = 2.2
自身
女儿
低输入样本
(DI = 0; 0.05 ng DNA)
人工退化的样本
(DI > 0; 0.25-1 ng DNA)

图16

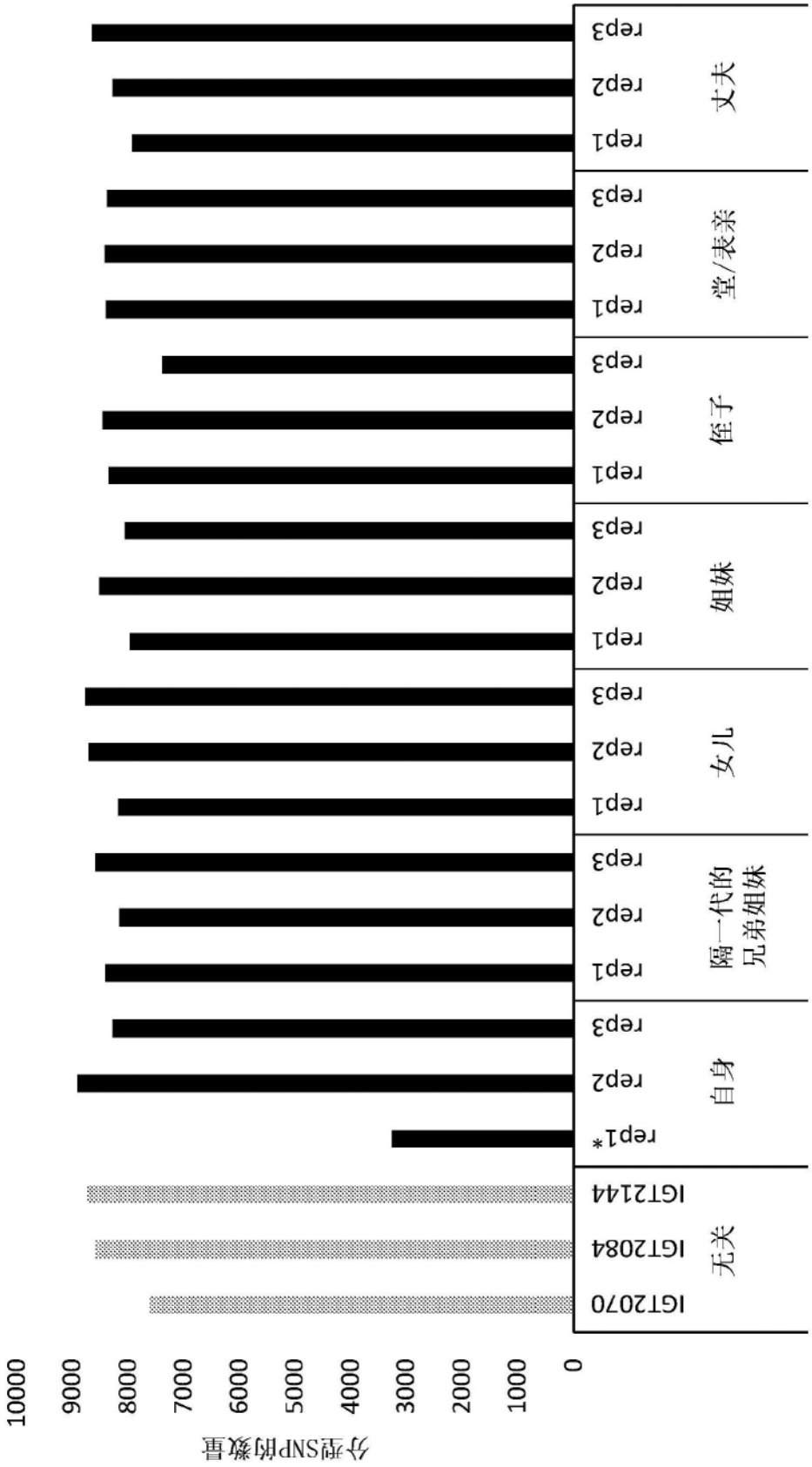

分型SNP的数量
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
无关
IGT2070
IGT2084
IGT2144
自身
rep1*
rep2
rep3
隔一代的兄弟姐妹
rep1
rep2
rep3
女儿
rep1
rep2
rep3
姐妹
rep1
rep2
rep3
侄子
rep1
rep2
rep3
堂/表亲
rep1
rep2
rep3
丈夫
rep1
rep2
rep3

图17

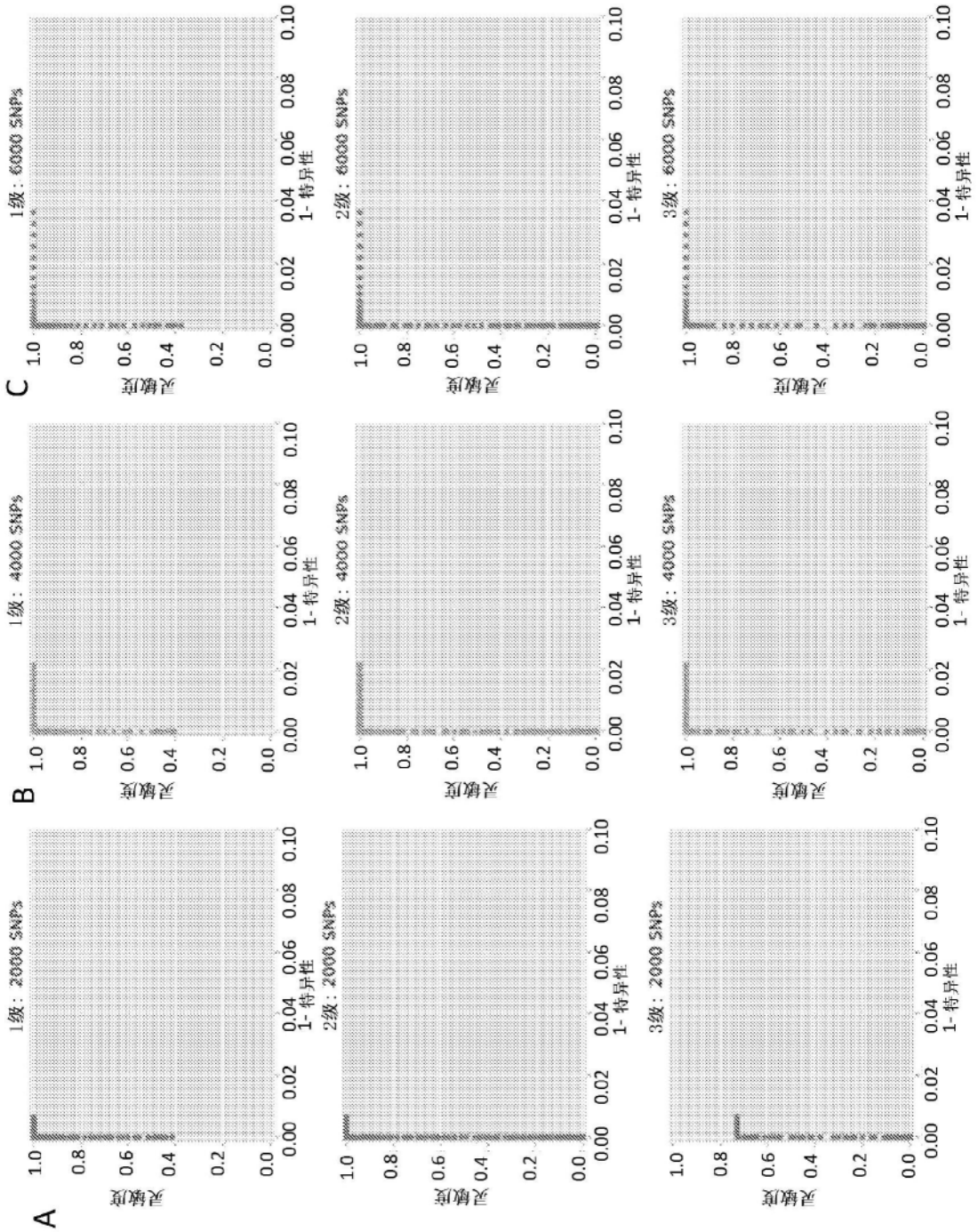

A
B
C
1级：2000 SNPs
2级：2000 SNPs
3级：2000 SNPs
1级：4000 SNPs
2级：4000 SNPs
3级：4000 SNPs
1级：6000 SNPs
2级：6000 SNPs
3级：6000 SNPs
灵敏度
1-特异性
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.02
0.04
0.06
0.08
0.10

图18

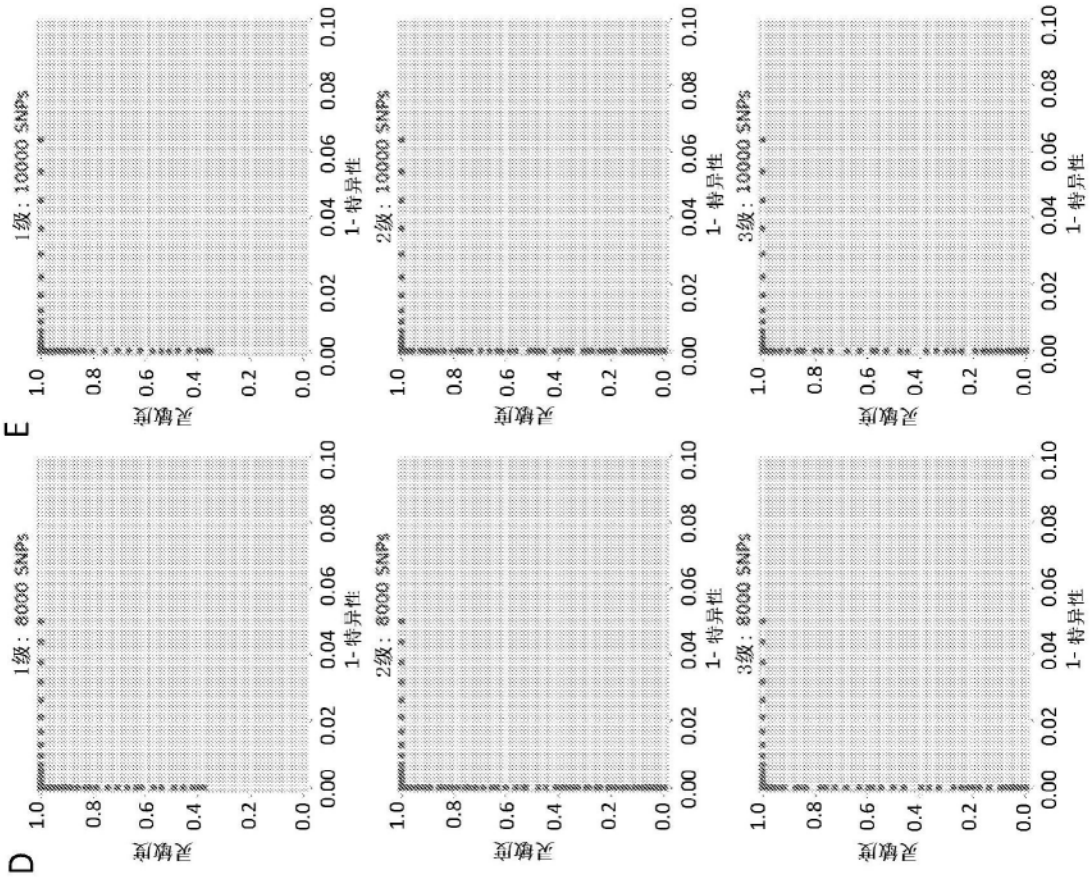

D
E
1级：8000 SNPs
2级：8000 SNPs
3级：8000 SNPs
1级：10000 SNPs
2级：10000 SNPs
3级：10000 SNPs
灵敏度
1-特异性
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.02
0.04
0.06
0.08
0.10

图18

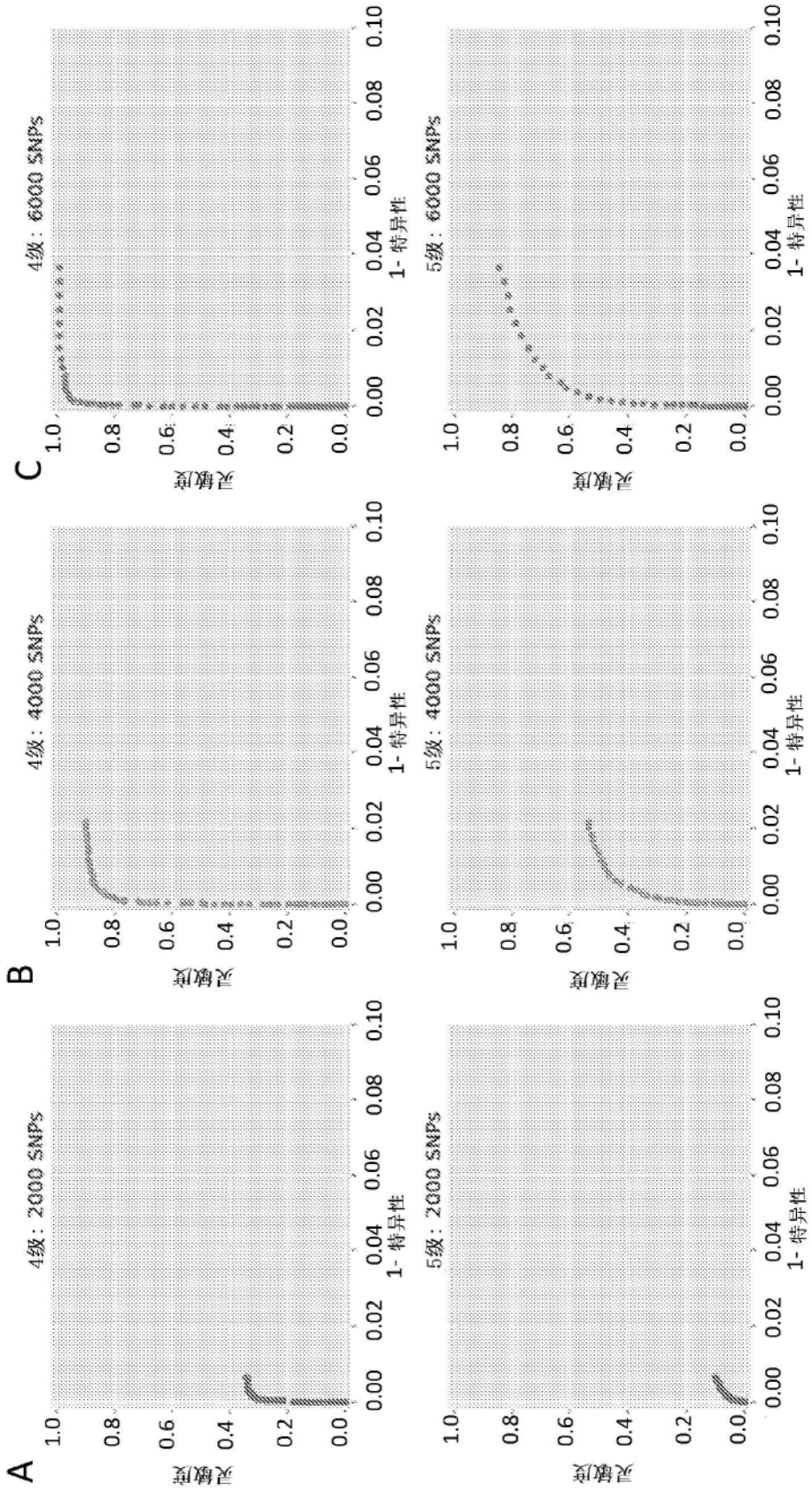

A
4级：2000 SNPs
5级：2000 SNPs
B
4级：4000 SNPs
5级：4000 SNPs
C
4级：6000 SNPs
5级：6000 SNPs
灵敏度
1- 特异性
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.02
0.04
0.06
0.08
0.10

图19

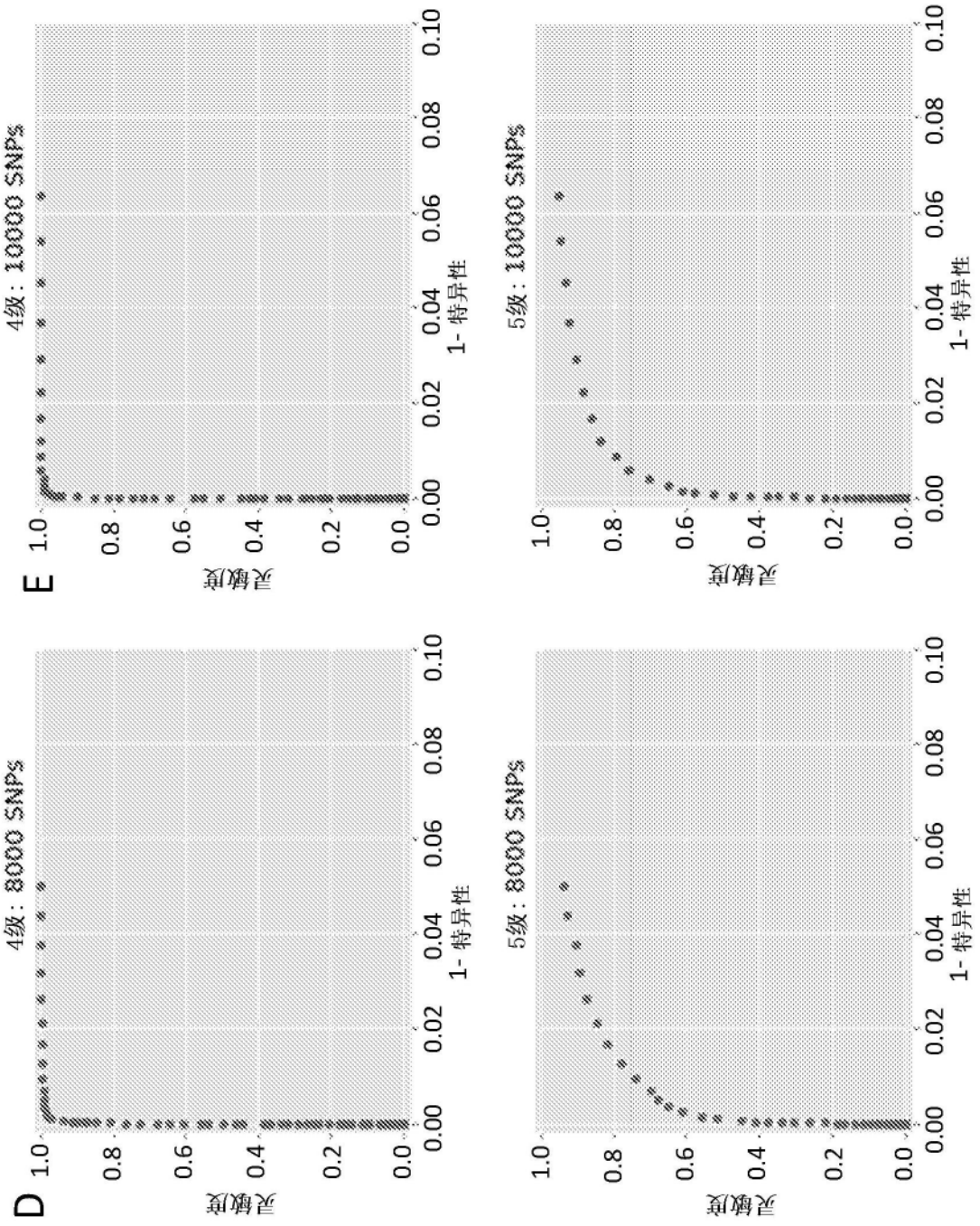

D
4级：8000 SNPs
灵敏度
1- 特异性
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.02
0.04
0.06
0.08
0.10
E
4级：10000 SNPs
灵敏度
1- 特异性
5级：8000 SNPs
灵敏度
1- 特异性
5级：10000 SNPs
灵敏度
1- 特异性

图19

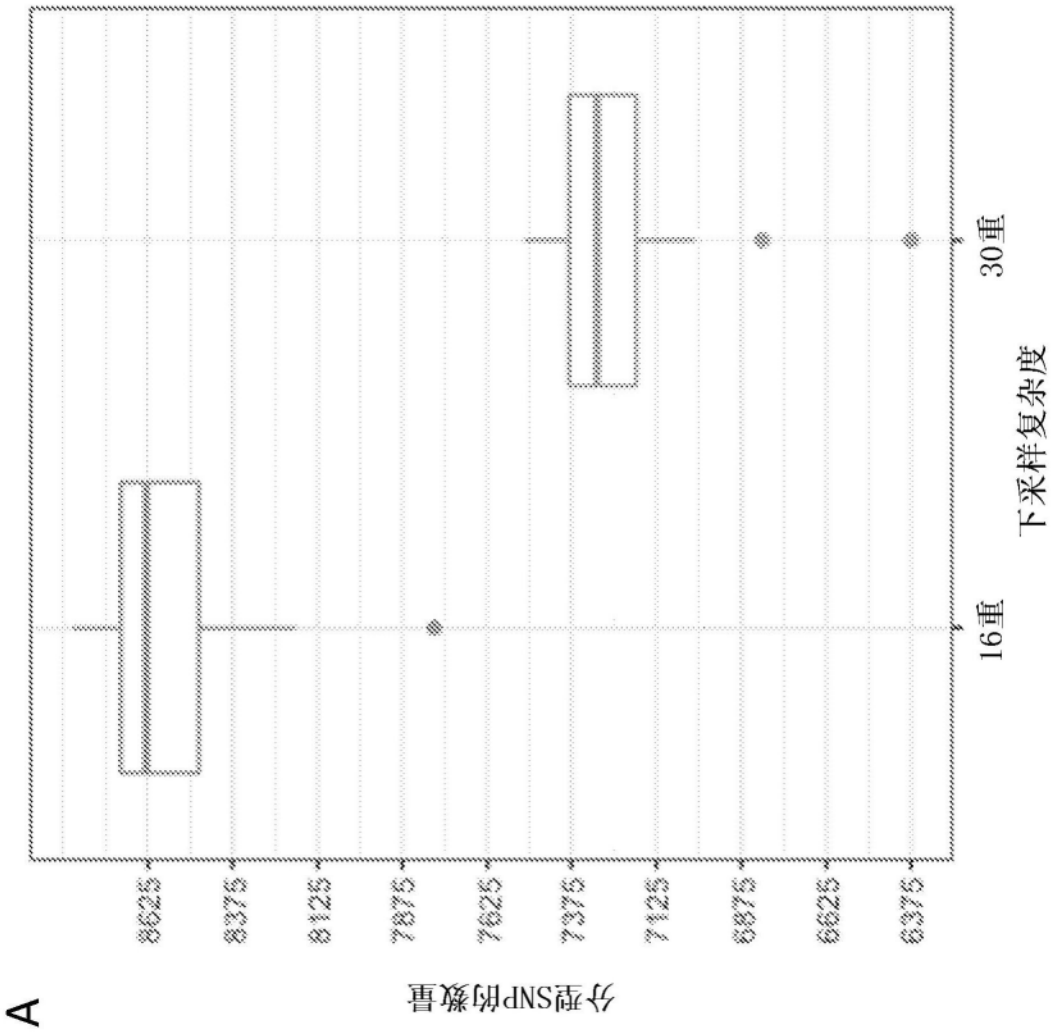

A
分型SNP的数量
8625
8375
8125
7875
7625
7375
7125
6875
6625
6375
16重
30重
下采样复杂度

图20

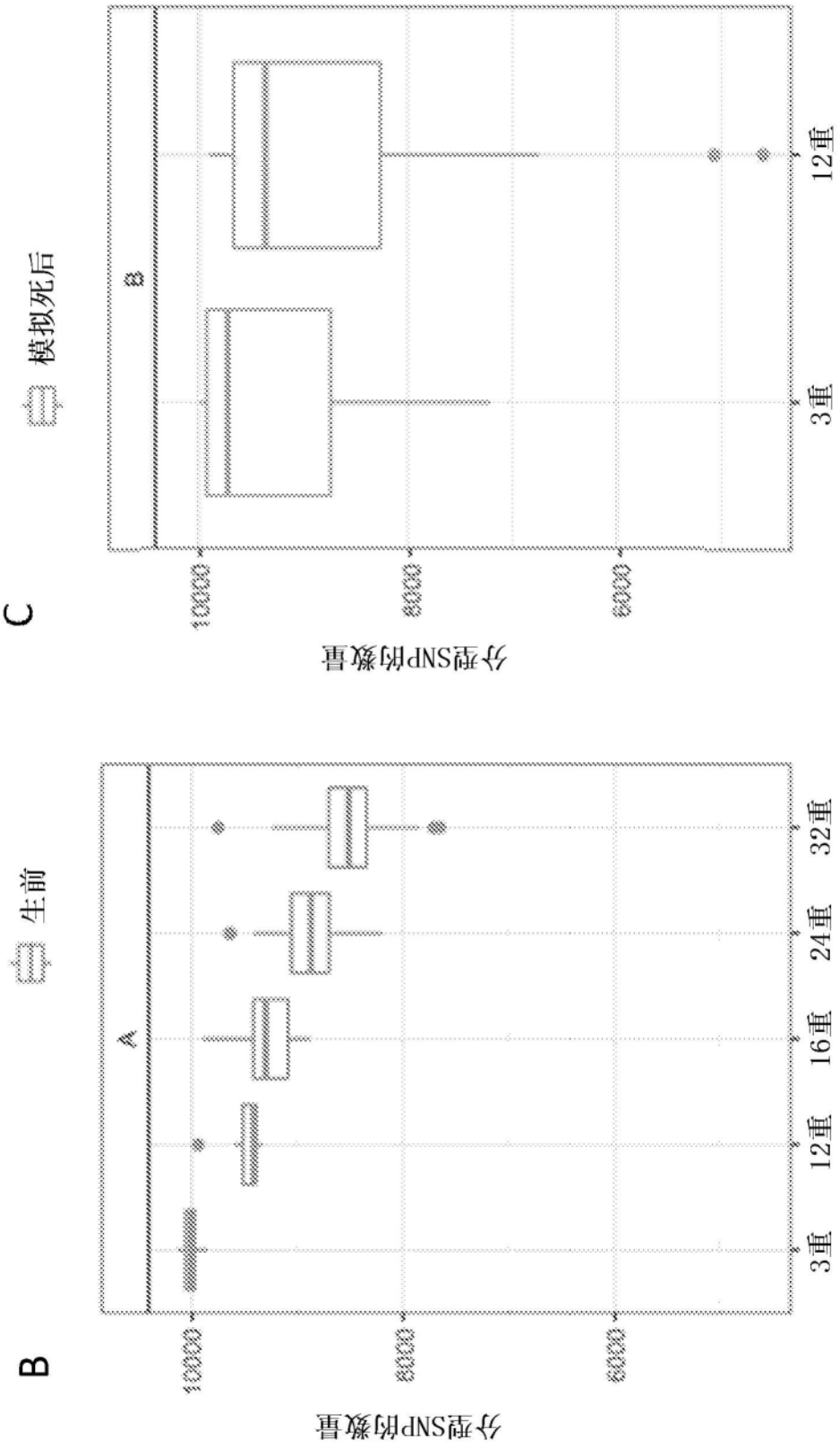

C
模拟死后
B
分型SNP的数量
10000
8000
6000
3重
12重
B
生前
A
分型SNP的数量
10000
8000
6000
3重
12重
16重
24重
32重

图20

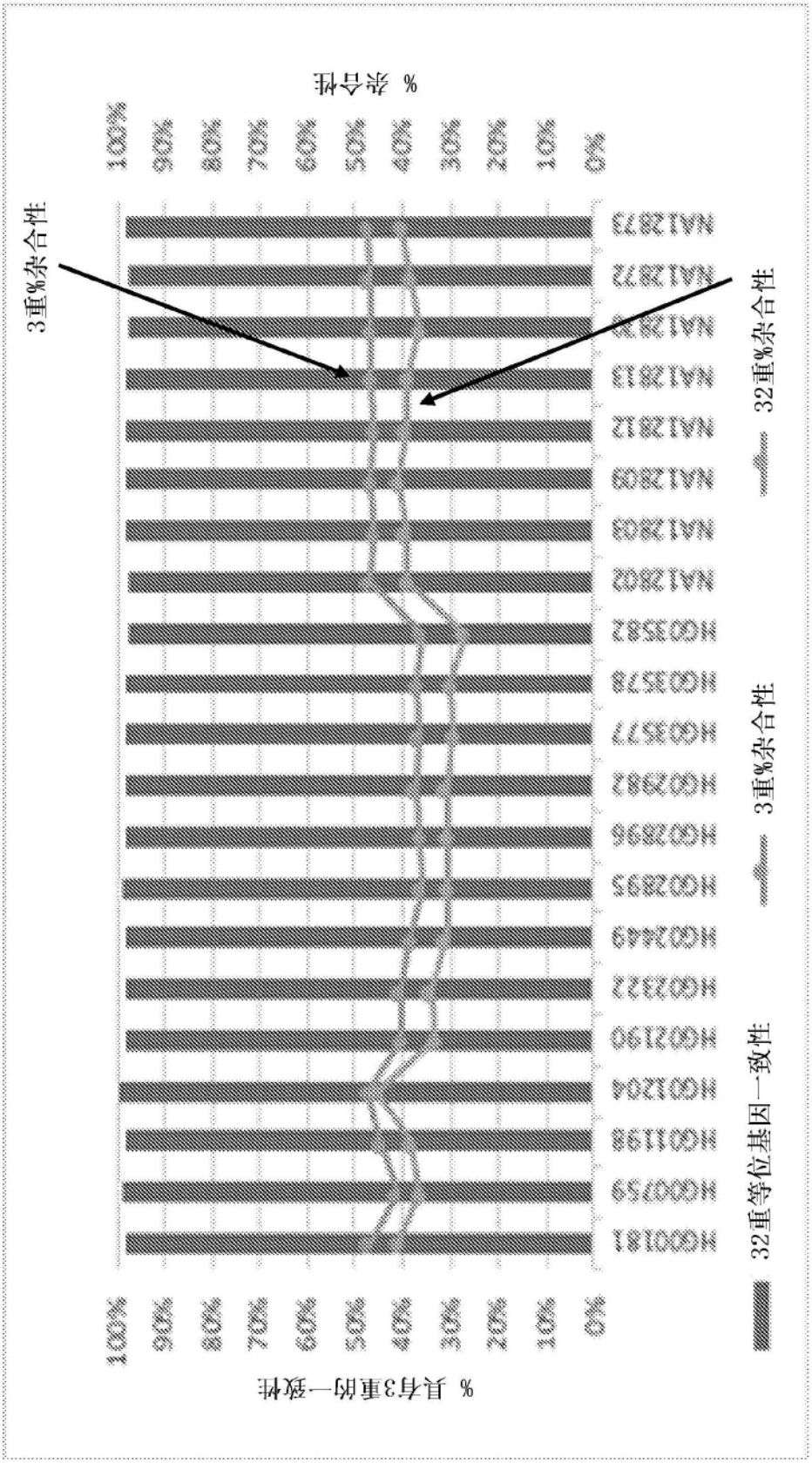

% 杂合性
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
3重%杂合性
NA12873
NA12872
NA12871
NA12813
NA12812
NA12809
NA12803
NA12802
HG03582
HG03578
HG03577
HG02982
HG02896
HG02895
HG02449
HG02322
HG02190
HG01204
HG01198
HG00759
HG00181
32重%杂合性
3重%杂合性
32重等位基因一致性
% 具有3重的一致性

图21A

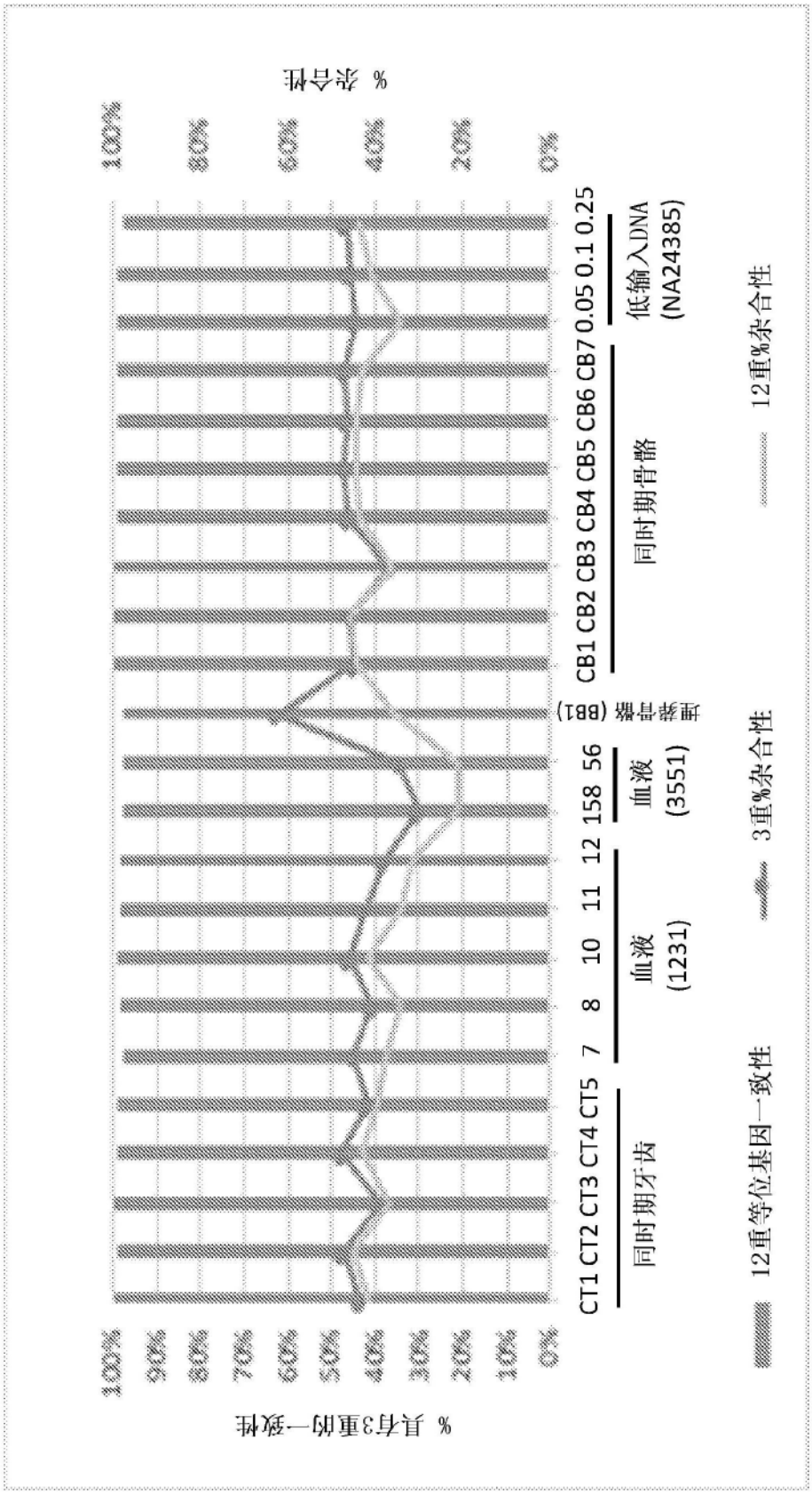

% 具有3重的一致性
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
% 杂合性
100%
80%
60%
40%
20%
0%
CT1 CT2 CT3 CT4 CT5 7 8 10 11 12 158 56 CB1 CB2 CB3 CB4 CB5 CB6 CB7 0.05 0.1 0.25
同时期牙齿
血液 (1231)
埋葬骨骼 (BB1)
血液 (3551)
同时期骨骼
低输入DNA (NA24385)
12重等位基因一致性
3重%杂合性
12重%杂合性

图21B

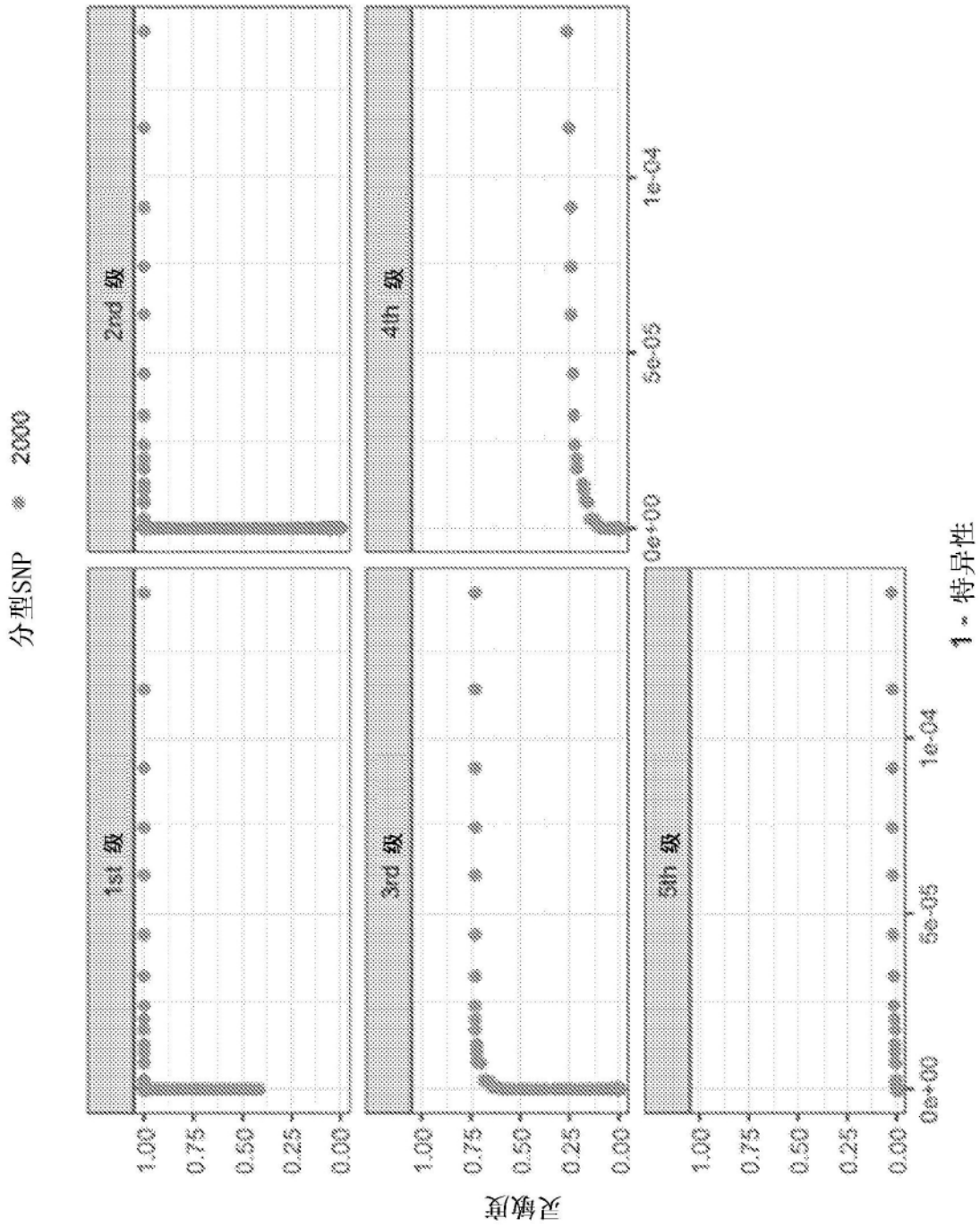

分型SNP * 2000
1st 级
2nd 级
3rd 级
4th 级
5th 级
灵敏度
1 - 特异性
1.00
0.75
0.50
0.25
0.00
0e+00
5e-05
1e-04

图22A

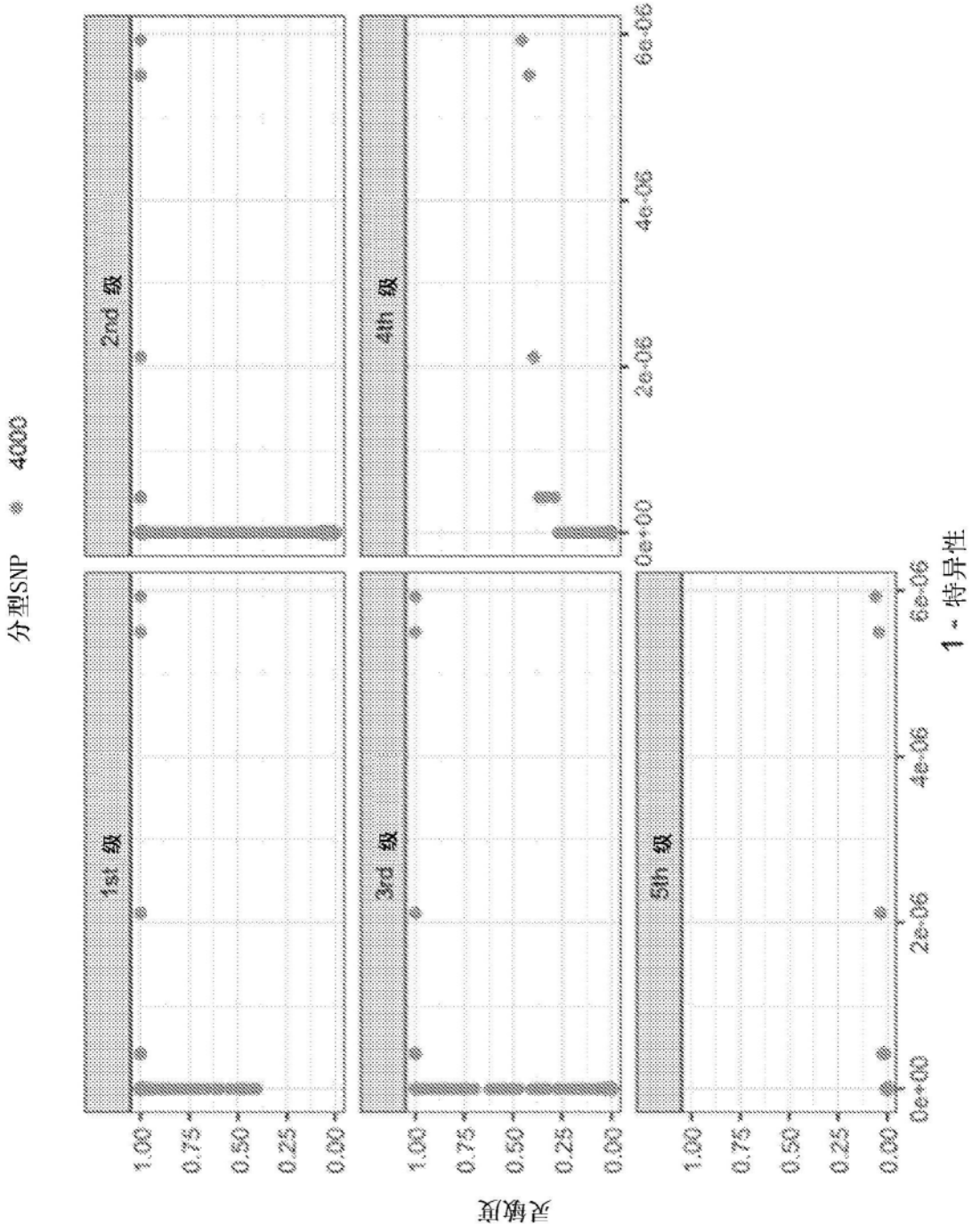

分型SNP
4003
1st 级
2nd 级
3rd 级
4th 级
5th 级
灵敏度
1 - 特异性
0.00
0.25
0.50
0.75
1.00
0e+00
2e-06
4e-06
6e-06

图22B

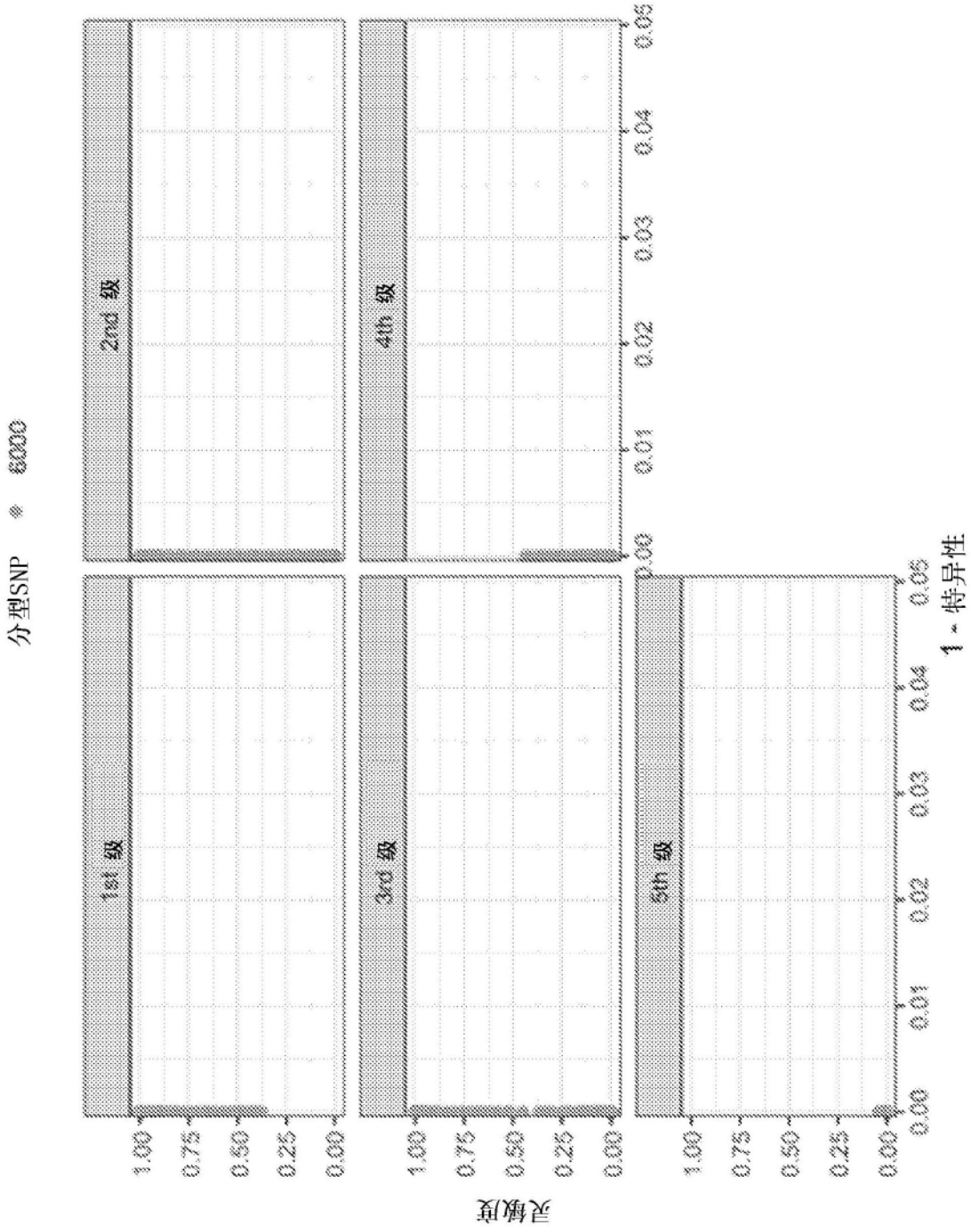

分型SNP * 6030
1st 级
2nd 级
3rd 级
4th 级
5th 级
灵敏度
1 - 特异性
0.00
0.25
0.50
0.75
1.00
0.01
0.02
0.03
0.04
0.05

图22C

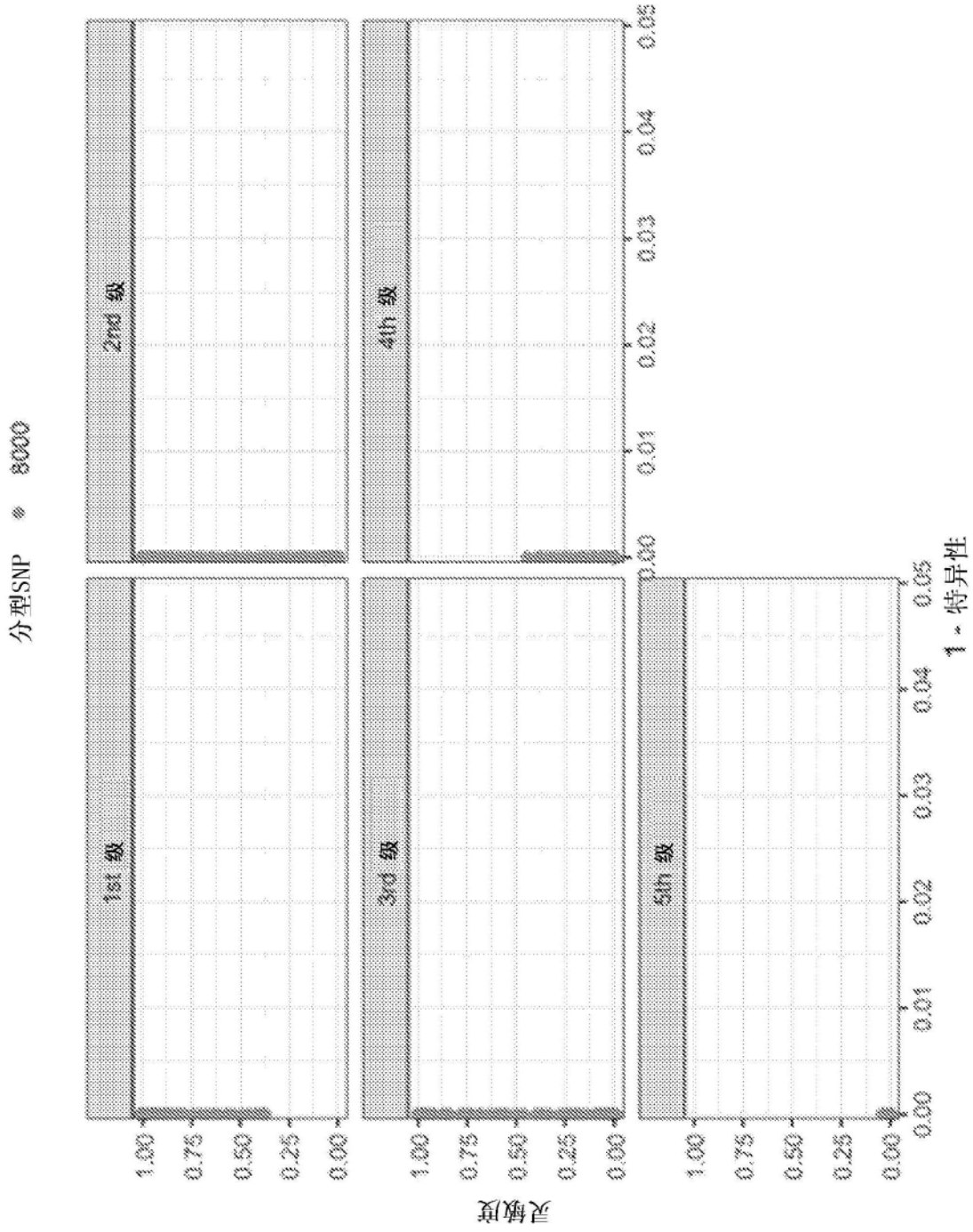

分型SNP * 8000
1st 级
2nd 级
3rd 级
4th 级
5th 级
灵敏度
1 - 特异性
1.00
0.75
0.50
0.25
0.00
0.00
0.01
0.02
0.03
0.04
0.05

图22D

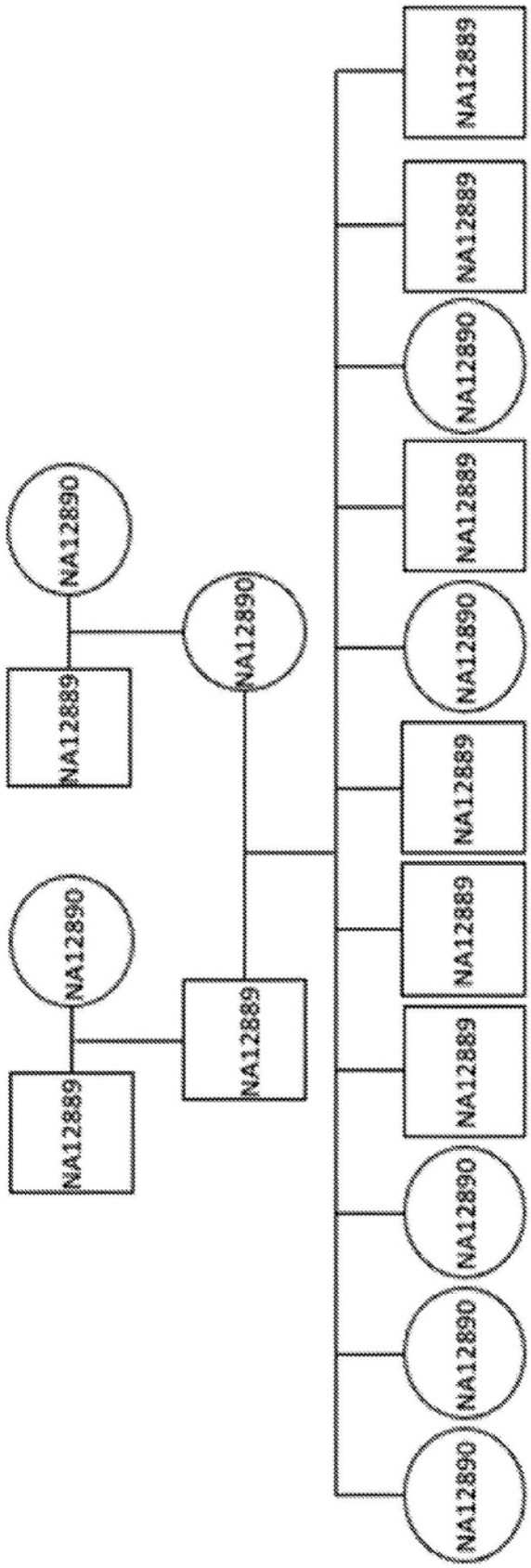

NA12889
NA12889
NA12890
NA12889
NA12890
NA12889
NA12889
NA12889
NA12890
NA12890
NA12890
NA12890
NA12889
NA12889
NA12890
NA12889

图23

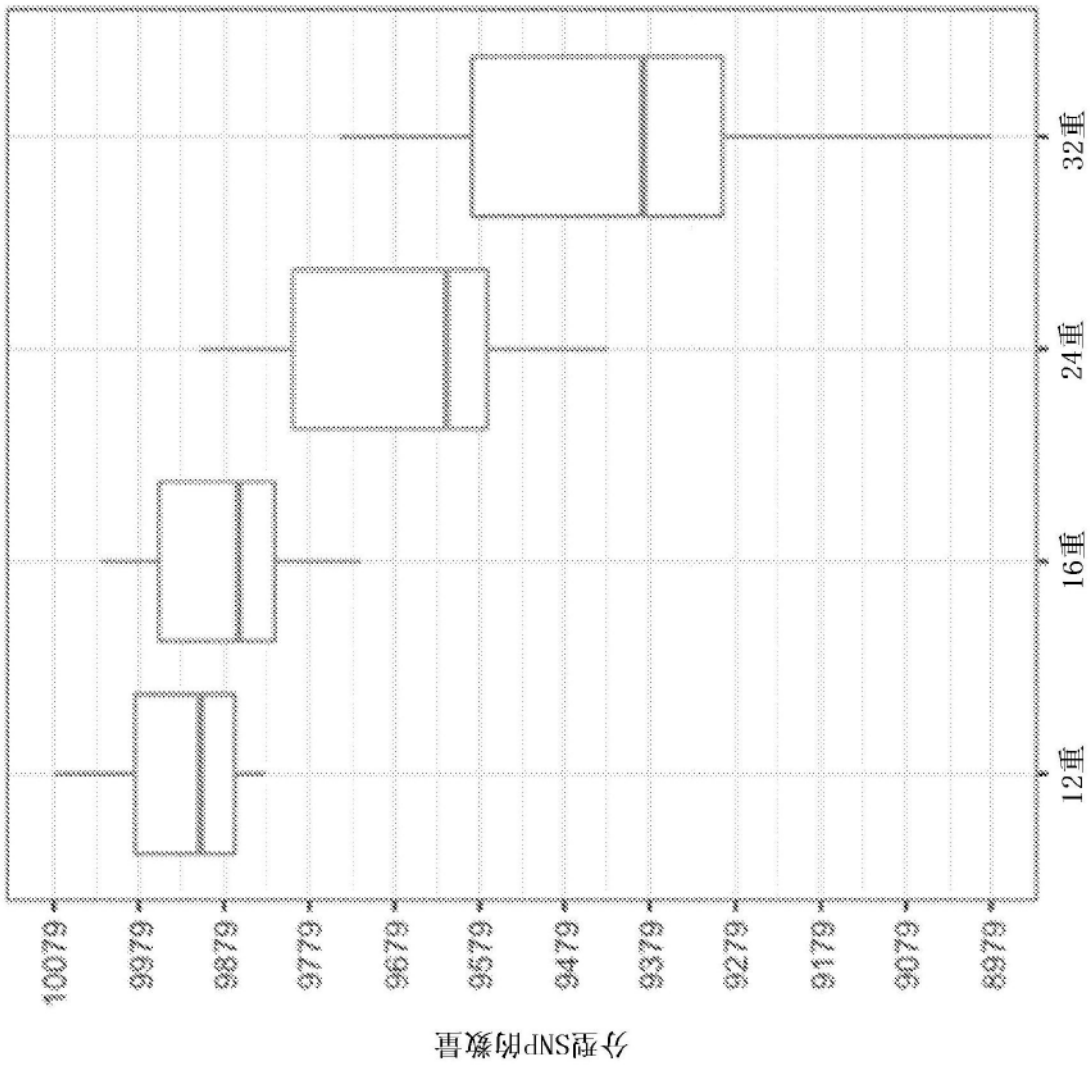

32重
24重
16重
12重
10079
9979
9879
9779
9679
9579
9479
9379
9279
9179
9079
8979
分型SNP的数量

图24

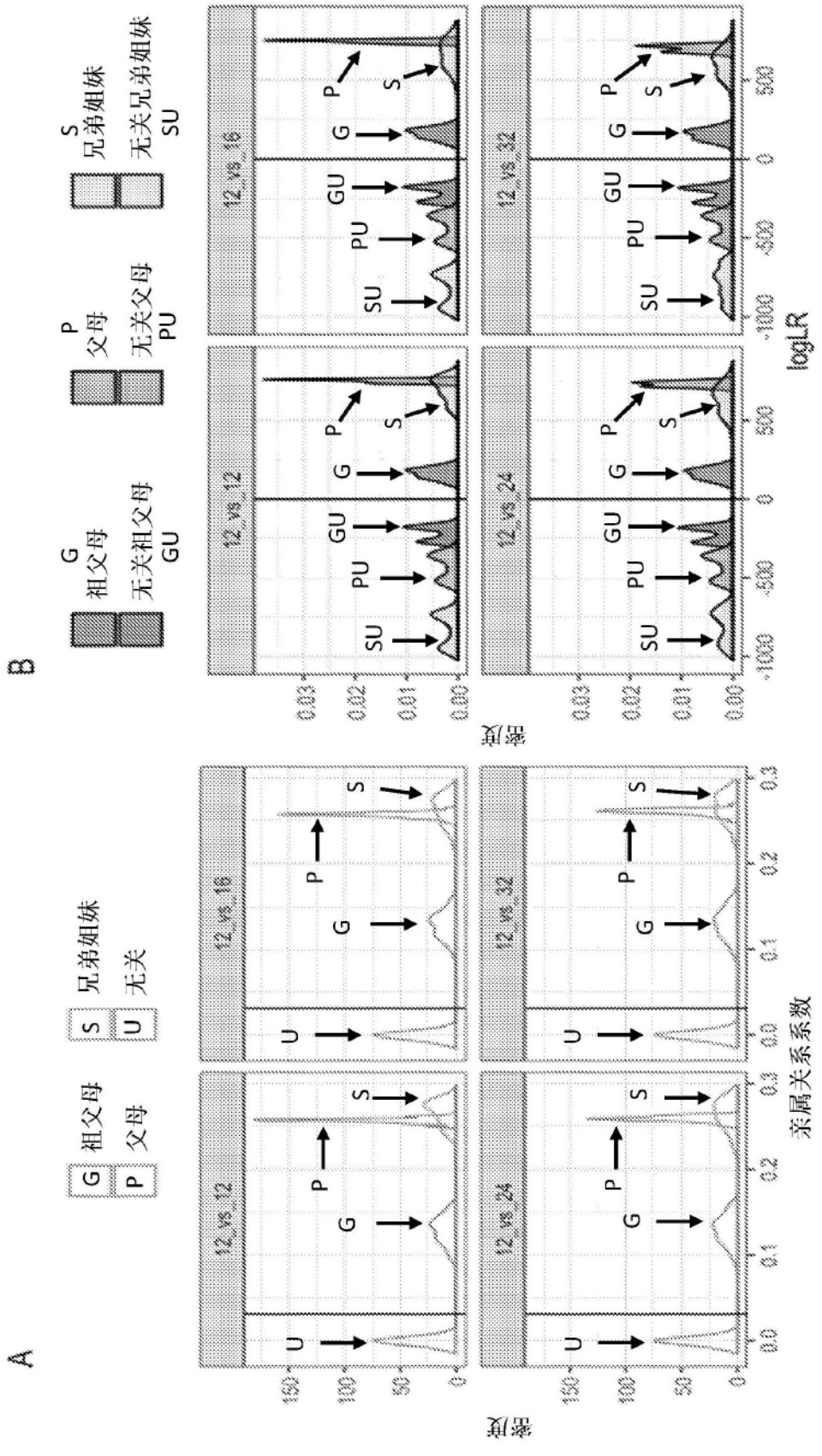

A
B
G 祖父母
P 父母
S 兄弟姐妹
U 无关
GU 无关祖父母
PU 无关父母
SU 无关兄弟姐妹
12_vs_12
12_vs_16
12_vs_24
12_vs_32
密度
亲属关系系数
logLR

图25

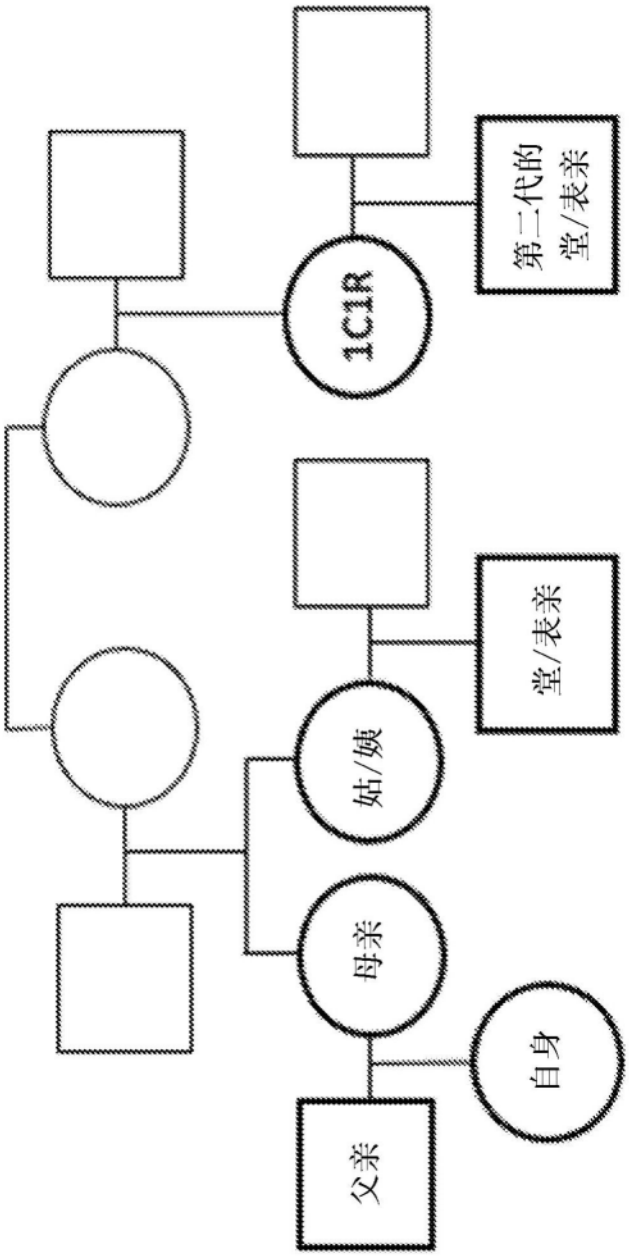


图26

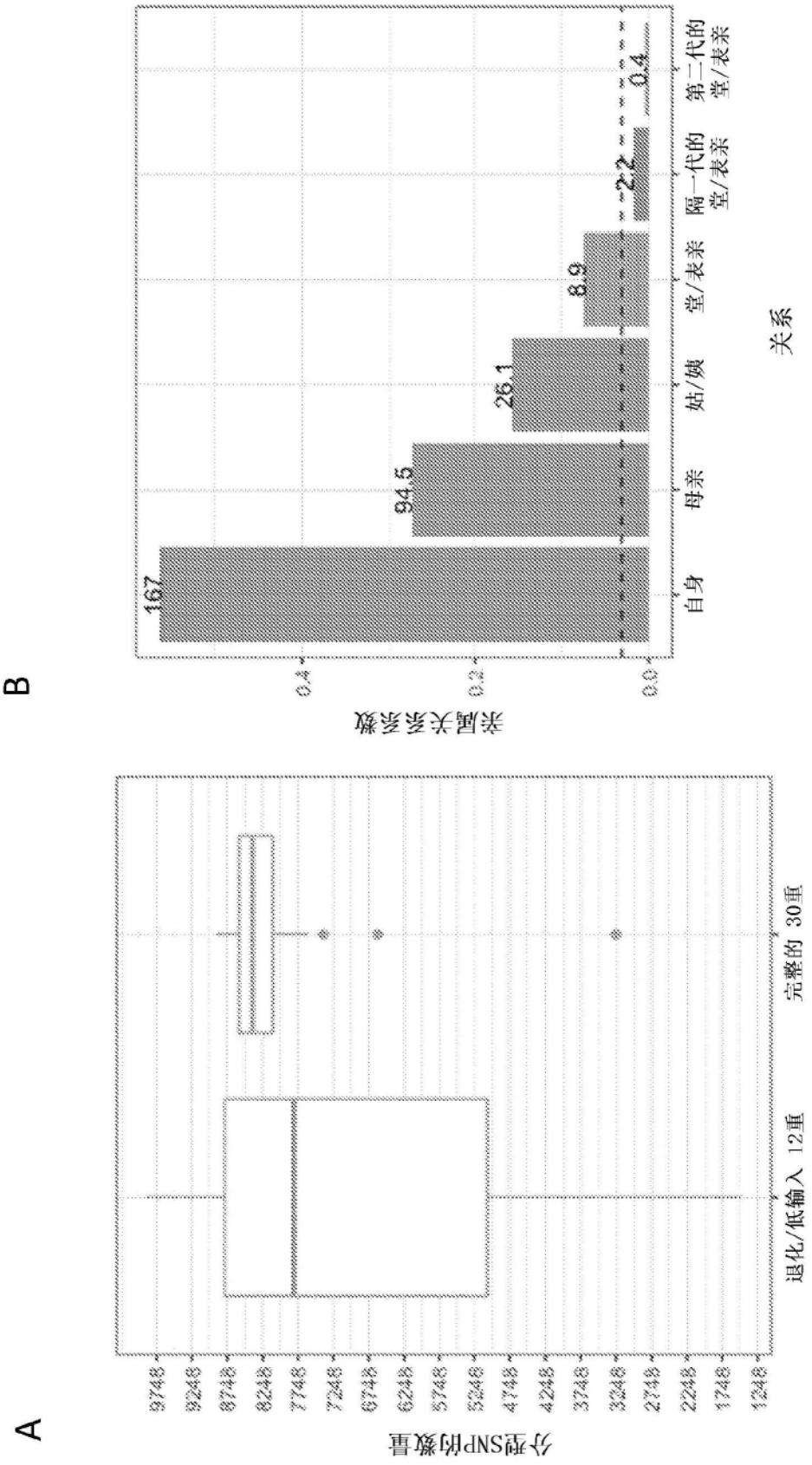

A
分型SNP的数量
9748
9248
8748
8248
7748
7248
6748
6248
5748
5248
4748
4248
3748
3248
2748
2248
1748
1248
退化/低输入 12重
完整的 30重
B
亲属关系系数
0.4
0.2
0.0
167
94.5
26.1
8.9
2.2
0.4
自身
母亲
姑/姨
堂/表亲
隔一代的堂/表亲
第二代的堂/表亲
关系

图27