



(12)发明专利

(10)授权公告号 CN 105814574 B

(45)授权公告日 2019.04.26

(21)申请号 201480066706.6

(22)申请日 2014.10.02

(65)同一申请的已公布的文献号
申请公布号 CN 105814574 A

(43)申请公布日 2016.07.27

(30) 优先权数据
61/887,081 2013.10.04 US

(85)PCT国际申请进入国家阶段日
2016.06.06

(86)PCT国际申请的申请数据
PCT/US2014/058885 2014.10.02

(87)PCT国际申请的公布数据
W02015/051163 EN 2015.04.09

(73)专利权人 塞昆纳姆股份有限公司

地址 美国加利福尼亚州

(72)发明人 G·汉纳姆

(74) 专利代理机构 上海专利商标事务所有限公司 31100

代理人 余颖 杨昀

(51) Int. Cl.

审查员 杨静

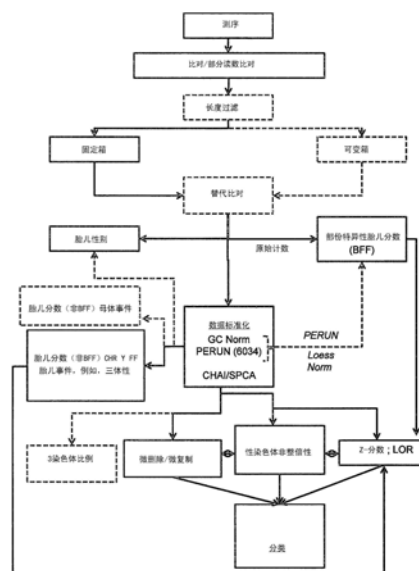
权利要求书5页 说明书139页 附图14页

(54)发明名称

遗传变异的非侵入性评估的方法和过程

(57)摘要

本文提供用于非侵入性评估遗传变异的方法、过程、系统和机器。



1. 一种用于减少样品的序列读数的偏好的方法, 包括:
 - (a) 生成测试样品的序列读数的 (i) 局部基因组偏好评价与 (ii) 偏好频率之间的关系, 从而生成样品偏好关系, 其中,
序列读数是来自测试样品的循环无细胞核酸, 并且
所述序列读数映射至参照基因组;
 - (b) 将所述样品偏好关系与参照偏好关系比较, 从而生成比较, 其中,
所述参照偏好关系是参照的 (i) 局部基因组偏好评价与 (ii) 偏好频率之间的关系; 并且
 - (c) 按照 (b) 中确定的比较将所述样品的序列读数的计数标准化, 从而减少所述样品的序列读数中的偏好。
2. 如权利要求1所述的方法, 其中, 使用微处理器生成所述样品偏好关系。
3. 如权利要求1或2所述的方法, 其中 (c) 中的标准化包括提供标准化的计数。
4. 如权利要求1所述的方法, 其中通过包括使用核心密度评价的过程来确定各所述局部基因组偏好评价。
5. 如权利要求1所述的方法, 其中所述参照偏好关系和所述样品偏好关系的各局部基因组偏好评价代表局部偏好含量。
6. 如权利要求5所述的方法, 其中所述局部偏好含量针对5000bp或更小的多核苷酸区段。
7. 如权利要求1所述的方法, 其中通过包括使用滑动窗分析的过程来确定各所述局部基因组偏好评价。
8. 如权利要求7所述的方法, 其中所述窗是5个连续核苷酸至5000个连续核苷酸, 并且所述窗在滑动窗分析中一次滑动1个碱基至10个碱基。
9. 如权利要求7所述的方法, 其中所述窗是200个连续核苷酸并且所述窗在滑动窗分析中一次滑动1个碱基。
10. 如权利要求1所述的方法, 其中 (b) 包括生成 (i) 和 (ii) 之间的拟合关系: (i) 比例, 各比例包括各所述局部基因组偏好评价的样品偏好关系频率和参照偏好关系频率, 与 (ii) 局部基因组偏好评价。
11. 如权利要求10所述的方法, 其中 (b) 中的拟合关系获自加权的拟合。
12. 如权利要求1所述的方法, 其中, 所述样品的各序列读数表示为二进制格式。
13. 如权利要求12所述的方法, 其中各所述序列读数的二进制格式包括所述读数映射至的染色体和所述读数映射至的染色体位置。
14. 如权利要求13所述的方法, 其中所述二进制格式是5-字节格式, 包括1-字节染色体序数和4-字节染色体位置。
15. 如权利要求12所述的方法, 其中所述二进制格式比序列比对/映射格式小50倍和/或比GZip格式小13%。
16. 如权利要求1所述的方法, 其中 (c) 中的标准化包括因数化除了偏好以外的一个或多个特征, 并且标准化所述序列读数的计数。
17. 如权利要求16所述的方法, 其中所述因数化一个或多个特征是通过包括使用多变量模型的过程。

18. 如权利要求17所述的方法,其中所述包括使用多变量模型的过程通过多变量模型进行。

19. 如权利要求16所述的方法,其中,按照(c)中的标准化和所述一个或多个特征的因数化将所述序列读数的计数标准化。

20. 如权利要求1所述的方法,包括在(c)之后,按照下述过程生成基因组的一个或多个部分或其区段的读数密度,所述过程包括生成含(c)中标准化的序列读数的计数的所述一个或多个部分各自的概率密度评价。

21. 如权利要求20所述的方法,其中,所述概率密度评价是核心密度评价。

22. 如权利要求20所述的方法,包括生成所述基因组或其区段的读数密度概况。

23. 如权利要求22所述的方法,其中,所述读数密度概况包括所述基因组的一个或多个部分,或其区段的读数密度。

24. 如权利要求20所述的方法,包括调整所述一个或多个部份的各读数密度。

25. 如权利要求20所述的方法,其中,所述一个或多个部份经过滤从而提供过滤的部份。

26. 如权利要求20所述的方法,其中,所述一个或多个部份经加权从而提供加权的部份。

27. 如权利要求26所述的方法,其中,所述一个或多个部份经本征函数加权。

28. 如权利要求1所述的方法,其中所述局部基因组偏好评价包括局部GC密度并且所述偏好频率包括GC偏好频率。

29. 一种用于减少样品的序列读数的偏好的方法,包括:

将来自怀有胎儿的妊娠雌性的血液的循环无细胞核酸加载测序设备,或将所述核酸的修饰变体加载测序设备,所述测序设备产生对应于所述核酸的核苷酸碱基的信号;

在任选地将所述信号转移至包括一个或多个计算设备的系统之后,由所述系统从所述核酸的信号生成序列读数,其中所述系统中的一个或多个计算设备包括存储器和一个或多个处理器,并且

其中在所述系统中,一个计算设备,或计算设备的组合设置为将所述序列读数映射至参照基因组并且:

(a) 使用微处理器生成测试样品的序列读数的(i)局部基因组偏好评价与(ii)偏好频率之间的关系,从而生成样品偏好关系,其中,

序列读数是来自测试样品的循环无细胞核酸,并且

所述序列读数映射至参照基因组;

(b) 将所述样品偏好关系与参照偏好关系比较,从而生成比较,其中,

所述参照偏好关系是参照的(i)局部基因组偏好评价与(ii)偏好频率之间的关系;并且

(c) 按照(b)中确定的比较将所述样品的序列读数的计数标准化,从而减少所述样品的序列读数中的偏好。

30. 如权利要求29所述的方法,其中(c)中的标准化包括提供标准化的计数。

31. 如权利要求29所述的方法,其中通过包括使用核心密度评价的过程来确定各所述局部基因组偏好评价。

32. 如权利要求29所述的方法,其中所述参照偏好关系和所述样品偏好关系的各局部基因组偏好评价代表局部偏好含量。

33. 如权利要求32所述的方法,其中所述局部偏好含量针对5000bp或更小的多核苷酸区段。

34. 如权利要求29所述的方法,其中通过包括使用滑动窗分析的过程来确定各所述局部基因组偏好评价。

35. 如权利要求34所述的方法,其中所述窗是5个连续核苷酸至5000个连续核苷酸,并且所述窗在滑动窗分析中一次滑动1个碱基至10个碱基。

36. 如权利要求34所述的方法,其中所述窗是200个连续核苷酸并且所述窗在滑动窗分析中一次滑动1个碱基。

37. 如权利要求29所述的方法,其中 (b) 包括生成 (i) 和 (ii) 之间的拟合关系: (i) 比例,各比例包括各所述局部基因组偏好评价的样品偏好关系频率和参照偏好关系频率,与 (ii) 局部基因组偏好评价。

38. 如权利要求37所述的方法,其中 (b) 中的拟合关系获自加权的拟合。

39. 如权利要求29所述的方法,其中,所述样品的各序列读数表示为二进制格式。

40. 如权利要求39所述的方法,其中各所述序列读数的二进制格式包括所述读数映射至的染色体和所述读数映射至的染色体位置。

41. 如权利要求40所述的方法,其中所述二进制格式是5-字节格式,包括1-字节染色体序数和4-字节染色体位置。

42. 如权利要求39所述的方法,其中所述二进制格式比序列比对/映射格式小50倍和/或比GZip格式小13%。

43. 如权利要求29所述的方法,其中 (c) 中的标准化包括因数化除了偏好以外的一个或多个特征,并且标准化所述序列读数的计数。

44. 如权利要求43所述的方法,其中所述因数化一个或多个特征是通过包括使用多变量模型的过程。

45. 如权利要求44所述的方法,其中所述包括使用多变量模型的过程通过多变量模型进行。

46. 如权利要求43所述的方法,其中,按照 (c) 中的标准化和所述一个或多个特征的因数化将所述序列读数的计数标准化。

47. 如权利要求29所述的方法,包括在 (c) 之后,按照下述过程生成基因组的一个或多个部分或其区段的读数密度,所述过程包括生成含 (c) 中标准化的序列读数的计数的所述一个或多个部分各自的概率密度评价。

48. 如权利要求47所述的方法,其中,所述概率密度评价是核心密度评价。

49. 如权利要求47所述的方法,包括生成所述基因组或其区段的读数密度概况。

50. 如权利要求49所述的方法,其中,所述读数密度概况包括所述基因组的一个或多个部分,或其区段的读数密度。

51. 如权利要求47所述的方法,包括调整所述一个或多个部份的各读数密度。

52. 如权利要求47所述的方法,其中,所述一个或多个部份经过滤从而提供过滤的部份。

53. 如权利要求47所述的方法, 其中, 所述一个或多个部份经加权从而提供加权的部份。

54. 如权利要求53所述的方法, 其中, 所述一个或多个部份经本征函数加权。

55. 如权利要求29所述的方法, 其中所述局部基因组偏好评价包括局部GC密度并且所述偏好频率包括GC偏好频率。

56. 一种用于确定样品是否存在非整倍性的系统, 所述系统包括用于以下的模块:

(a) 生成测试样品的序列读数的 (i) 局部基因组偏好评价与 (ii) 偏好频率之间的关系, 从而生成样品偏好关系, 其中,

序列读数是来自测试样品的循环无细胞核酸, 并且

所述序列读数映射至参照基因组;

(b) 将所述样品偏好关系与参照偏好关系比较, 从而生成比较, 其中,

所述参照偏好关系是参照的 (i) 局部基因组偏好评价与 (ii) 偏好频率之间的关系;

(c) 按照 (b) 中确定的比较将所述样品的序列读数的计数标准化, 从而减少所述样品的序列读数中的偏好;

(d) 按照读数密度分布过滤参照基因组的部份, 从而提供包括过滤的部份的读数密度的测试样品的读数密度概况, 其中,

所述读数密度包括来自妊娠雌性的测试样品的循环无细胞核酸的序列读数, 并且

针对多个样品的部份的读数密度确定所述读数密度分布;

(e) 按照一个或多个主成分调整测试样品的读数密度概况, 所述主成分通过主成分分析获自己知的整倍性样品的组, 从而提供包括调整的读数密度的测试样品概况;

(f) 将所述测试样品概况与参照概况比较, 从而提供比较; 并且

(g) 按照所述比较确定所述测试样品中是否存在染色体非整倍性。

57. 如权利要求56所述的系统, 其中在 (b) 中由1-10个主成分来调整所述读数密度概况。

58. 如权利要求56所述的系统, 其中在 (b) 中由5个主成分来调整所述读数密度概况。

59. 如权利要求56-58中任一项所述的系统, 其中针对读数密度概况中的一个或多个特征调整所述一个或多个主成分, 其特征选自胎儿性别、序列偏好、胎儿分数、与DNA酶I灵敏度相关的偏好、熵、重复序列偏好、染色质结构偏好、聚合酶误差率偏好、回文偏好、插入重复偏好、PCR扩增偏好、和隐藏拷贝数变异。

60. 如权利要求59所述的系统, 其中序列偏好包括鸟嘌呤和胞嘧啶偏好。

61. 如权利要求56-58中任一项所述的系统, 其中所述比较包括确定显著性水平。

62. 如权利要求61所述的系统, 其中, 所述确定显著性水平包括确定p-值。

63. 如权利要求56-58中任一项所述的系统, 其中, 所述参照概况包括获自己知整倍性样品的组的读数密度概况。

64. 如权利要求56-58中任一项所述的系统, 其中, 所述参照概况包括过滤的部份的读数密度。

65. 如权利要求56-58中任一项所述的系统, 其中, 所述参照概况包括按照所述一个或多个主成分调整的读数密度。

66. 如权利要求61所述的系统, 其中, 所述显著性水平表示所述测试样品概况和所述参

照概况之间有统计学显著差异,并且确定存在染色体非整倍性。

67.如权利要求56-58中任一项所述的系统,其中,所述多个样品包括已知整倍性样品的组。

68.如权利要求56-58中任一项所述的系统,其中,所述多个样品的部份的读数密度是中值读数密度。

69.如权利要求56-58中任一项所述的系统,其中,所述测试样品的过滤部份的读数密度是中值读数密度。

70.如权利要求63所述的系统,其中,所述参照概况的读数密度概况包括中值读数密度。

71.如权利要求63所述的系统,其中,按照包括核心密度评价的过程确定所述测试样品概况、所述多个样品和所述参照概况的读数密度。

72.如权利要求69所述的系统,其中,按照所述测试样品的中值读数密度确定所述测试样品概况。

73.如权利要求70所述的系统,其中,按照所述参照的中值读数密度分布确定所述参照概况。

74.如权利要求56-58中任一项所述的系统,包括按照针对所述读数密度分布的不确定性测量对参照基因组的部份进行过滤。

75.如权利要求74所述的系统,其中,所述不确定性测量是MAD。

76.如权利要求56-58中任一项所述的系统,其中,所述测试样品概况代表所述测试样品的染色体剂量。

77.如权利要求76所述的系统,包括比较测试样品概况的染色体剂量与参照概况的染色体剂量,从而生成染色体剂量比较。

78.如权利要求77所述的系统,其中按照所述染色体剂量比较确定所述测试样品是否存在染色体非整倍性。

79.如权利要求56-58中任一项所述的系统,其中确定所述测试样品是否存在染色体非整倍性包括鉴定是否存在1个拷贝的染色体、2个拷贝的染色体、3个拷贝的染色体、4个拷贝的染色体、5个拷贝的染色体、染色体的一个或多个区段的缺失或染色体的一个或多个区段的插入。

遗传变异的非侵入性评估的方法和过程

[0001] 相关专利申请

[0002] 本专利申请要求2013年10月4日提交的名为“遗传变异的非侵入性评估方法和过程(METHODS AND PROCESSES FOR NON-INVASIVE ASSESSMENT OF GENETIC VARIATIONS)”，发明人为Gregory Hannum，档案号为SEQ-6073-PV的美国临时专利申请61/887,081的权利。前述专利申请的全部内容通过引用纳入本文，包括所有文字、表格和附图。

[0003] 领域

[0004] 本文提供的技术部分涉及遗传变异的非侵入性评估方法、过程和机器。

[0005] 背景

[0006] 活体生物(如动物、植物和微生物)的遗传信息和复制遗传信息的其他形式(如病毒)被编码成脱氧核糖核酸(DNA)或核糖核酸(RNA)。遗传信息是代表化学或假定核酸的一级结构的一连串核苷酸或修饰的核苷酸。人的完整基因组包含位于二十四(24)条染色体上的约30,000个基因(见《人类基因组》(The人类基因组), T.Strachan, BIOS科学出版社、1992)。各基因编码特定蛋白质,所述蛋白质在通过转录和翻译表达之后,在活细胞中实现特定的生物化学功能。

[0007] 许多医学病症由一种或多种遗传变异引起。某些遗传变异引起医学病症,包括例如血友病、地中海贫血、杜氏肌营养不良症(DMD)、亨廷顿氏病(HD)、阿尔茨海默病和囊性纤维化(CF)(《人类基因组突变》(Human Genome Mutations), D.N.Cooper和M.Krawczak, BIOS出版社、1993)。这类遗传疾病可能由特定基因DNA中单个核苷酸的加入、取代或缺失所致。某些出生缺陷由染色体异常(也称为非整倍性)造成,例如21三体性(唐氏综合征)、13三体性(帕陶氏综合征)、18三体性(爱德华氏综合征)、16和22三倍体、X单体性(特纳氏综合征)和某些性染色体非整倍性如克氏综合征(XXY)。其他遗传变异是胎儿性别,这通常可基于性染色体X和Y来确定。一些遗传变异使个体倾向于或引起许多疾病中的任一种,例如糖尿病、动脉硬化、肥胖症、各种自体免疫疾病和癌症(如结直肠癌、乳腺癌、卵巢癌、肺癌)。

[0008] 对一种或多种遗传变异或变化的鉴定可有助于诊断特定医学病症,或确定特定医学病症的诱因。鉴定遗传变异能帮助医疗决策和/或使用有益的医疗方案。在某些实施方式中,对一种或多种遗传变异或变化的鉴定涉及分析无细胞DNA。无细胞DNA(CF-DNA)由来自细胞死亡和外周血循环的DNA片段组成。高浓度的CF-DNA能指示某些临床病症,例如癌症、创伤、烧伤、心肌梗塞、中风、败血症、感染和其它疾病。此外,无细胞胎儿DNA(CFF-DNA)能在母本血流中检测,并且用于多种非侵入性产前诊断。

[0009] 概述

[0010] 在某些方面中,本发明提供了一种包括存储器和一个或多个微处理器的系统,一个或多个微处理器设置为按照存储器中的指令进行用于减少样品的序列读数中的偏好的过程,该过程包括(a)生成测试样品的序列读数的(i)局部基因组偏好评价与(ii)偏好频率之间的关系,从而生成样品偏好关系,其中序列读数是来自测试样品的循环无细胞核酸,并且序列读数映射到参照基因组,(b)比较样品偏好关系和参照偏好关系,从而生成比较,其中参照偏好关系是(i)局部基因组偏好评价与(ii)参照偏好频率之间的关系,并且(c)按照

(b) 中确定的比较将样品的序列读数的计数标准化,其中样品的序列读数中的偏好减少。

[0011] 在某些方面中,本发明提供了一种包括存储器和一个或多个微处理器的系统,一个或多个微处理器设置为按照存储器中的指令进行用于减少样品的序列读数中的偏好的过程,该过程包括 (a) 生成测试样品的序列读数的 (i) 鸟嘌呤和胞嘧啶 (GC) 密度与 (ii) GC 密度之间的关系,从而生成样品 GC 密度关系,其中序列读数是来自测试样品的循环无细胞核酸,并且序列读数映射到参照基因组, (b) 比较样品 GC 密度和参照 GC 密度关系,从而生成比较,其中参照 GC 密度关系是 (i) GC 密度与 (ii) 参照的 GC 密度频率之间的关系,并且 (c) 按照 (b) 中确定的比较将样品的序列读数的计数标准化,其中样品的序列读数中的偏好减少。

[0012] 在某些方面中,本发明还提供了一种包括存储器和一个或多个微处理器的系统,一个或多个微处理器设置为按照存储器中的指令进行用于确定样品是否存在非整倍性的过程,该过程包括 (a) 按照读数密度分布过滤参照基因组的部份,从而提供包括过滤的部份的读数密度的测试样品的读数密度概况,其中读数密度包括来自妊娠雌性的测试样品的循环无细胞核酸的序列读数,并且针对多个样品的部份的读数密度确定读数密度分布, (b) 按照一个或多个主成分来调整测试样品的读数密度概况,主成分通过主成分分析获自己知整倍体样品的组,从而提供包括调整的读数密度的测试样品概况, (c) 比较测试样品概况与参照概况,从而提供比较,并且 (d) 根据比较确定测试样品是否存在染色体非整倍性。

[0013] 下述说明、实施例、权利要求和附图中进一步描述某些技术方面。

附图说明

[0014] 附图描述本技术的实施方式但不具限制性。为了说明的清楚和方便,附图未按比例制作,并且在一些情况中,可能夸大或放大多个方面以协助对具体实施方式的理解。

[0015] 图1显示Epanechnikov核提供的GC密度的实施方式(带宽=200bp)。

[0016] 图2显示HTRA1基因的GC密度(y-轴)图,其中GC密度跨越整个基因组进行标准化。基因组位置示于x轴上。

[0017] 图3显示参照基因组(实线)以及样品所获序列读数(虚线)的局部基因组偏好评价(例如GC密度、x轴)。偏移频率(例如密度频率)示于y轴上。GC密度评价跨越整个基因组进行标准化。该实施例中,所述样品相比从参照所预期的有更多高GC含量的读数。

[0018] 图4显示参照基因组的GC密度评价分布和样品序列读数的GC密度评价分布,使用加权第三阶多项式拟合的关系。GC密度评价(x-轴)跨越整个基因组进行标准化。GC密度频率在y轴上用对参照的密度频率除以样品的密度频率的比例取 $\log_2$ 来表示。

[0019] 图5A显示基因组所有部份的中值GC密度(x-轴)的分布。图5B显示根据多种样品的GC密度分布确定的中值绝对偏差(MAD)值(x-轴)。GC密度频率示于y轴上。根据多种参照样品(如训练组)的中值GC密度分布和根据多种样品的GC密度分布确定的MAD值来筛选部份。包括超出既定阈值(例如MAD的四分位间范围的四倍)的GC密度的部份根据筛选方法从考虑中移除。

[0020] 图6A显示基因组的样品的读数密度概况,包括基因组中的中值读数密度(y-轴,例如读数密度/部份)和各基因组部份的相对位置(x-轴,部份的指标)。图6B显示第一主成分(PC1),图6C显示第二主成分(PC2),它们获自500个整倍体的训练组中所获的读数密度概况的主成分分析。

[0021] 图7A-C显示基因组的样品的读数密度概况的示例,所述基因组包括染色体21的三体(例如用两条垂直线括出的)。各基因组部份的相对位置示于x轴上。读数密度示于y轴上。图7A显示原始(例如未校准)读数密度概况。图7B显示7A的包括第一调整(包括扣除中值概况)的概况。图7C显示7B的包括第二调整的概况。第二调整包括扣除8x主成分概况,基于其在该样品中发现的代表进行加权。(例如建立模型)。例如样品概况 = $A*PC1+B*PC2+C*PC3\dots$,而校正概况(例如7C所示) = 样品概况 - $A*PC1+B*PC2+C*PC3\dots$ 。

[0022] 图8显示T21测试的拔靴法(bootstrapped)训练样品的测试p值的QQ图。QQ图通常比较两种分布。图8显示测试样品的ChAI分数(y轴)与均匀分布(即p-值的期待分布,x轴)的比较。各点代表单个测试样品的log-p值的分数。基于均匀分布对样品进行分选并分配“期望”值(x轴)。下部虚线代表对角线,上部的线代表Bonferroni阈值。遵循均匀分布的样品预期落在下部对角线上(下部虚线)。由于部份中的相关性(例如偏移),数值远离对角线,表明样品的分值比预期更高(低p值)。本文所述方法(例如ChAI,例如参见实施例1)能校正这种观察到的偏移。

[0023] 图9A显示读数密度图,显示训练组中男性和女性的PC2系数的差异。图9B显示具有PC2系数的性别判定的接受者操作特征(ROC)曲线。通过测序进行的性别判定用于真参照。

[0024] 图10A-10B显示系统的实施方式。

[0025] 图11显示系统的实施方式。

[0026] 图12显示本文提供的方法的实施方式。

[0027] 发明详述

[0028] 下一代测序允许通过比常规测序方法更快更廉价的方法在全基因组规模上对核酸进行测序。本文提供的方法、系统和产品可利用先进的测序技术来定位并鉴定遗传变异和/或相关的疾病和病症。本发明提供的方法、系统和产品通常可提供使用血液样或其部分的对象基因组(例如,胎儿基因组)的非侵入性评估,并且通常比侵入性更强的技术(例如,羊膜穿刺术,活检)更安全、更快和/或更廉价。在一些实施方式中,本发明提供了部分包括获得样品中存在的核酸的序列读数,该序列读数通常映射到参照序列,处理序列读数的计数并且确定遗传变异是否存在的方法。本发明提供的系统、方法和产品可用于定位和/或鉴定遗传变异并且可用于诊断和治疗与某些遗传变异相关的疾病、病症和障碍。

[0029] 在一些实施方式中,本发明还提供数据操作方法来减少和/或去除由测序技术的多个方面引入的测序偏好。测序偏好通常导致基因组或其片段中读数的不均匀分布,和/或读数质量的变化。测序偏好可破坏基因组测序数据、损害有效数据分析、使结果有污点并且不能进行精确数据解释。有时,可通过增加测序覆盖率来减少测序偏好;然而,该方法通常提高测序成本,并且有非常有限的效果。本发明所述的数据操作方法可减少和/或去除测序偏好,从而改善序列读数数据的质量而不增加测序成本。在一些实施方式中,本文还提供进行本文所述的方法的系统、机器、设备、产品和模块。

[0030] 样品

[0031] 本文提供用于分析核酸的方法和组合物。在一些实施方式中,分析核酸片段混合物中的核酸片段。核酸混合物可包括两种或更多核酸片段种类,所述两种或更多核酸片段种类具有不同核苷酸序列、不同片段长度、不同来源(例如基因组来源、胎儿与母本来源、细胞或组织来源、样品来源、对象来源等)或其组合。

[0032] 本文所述方法、系统、机器和/或设备中使用的核酸或核酸混合物经常从获自对象(例如,测试对象)的样品中分离。在本文中,从中获得试样或样品的对象有时称为测试对象。对象可以是任何活体或非活体生物,包括但不限于人、非人动物、植物、细菌、真菌、病毒或原生生物。能选择任何人或非人动物,包括但不限于哺乳动物、爬行动物、鸟类、两栖类、鱼类、有蹄类动物、反刍动物、牛科动物(如牛)、马科动物(如马)、山羊和绵羊类动物(如绵羊、山羊)、猪科动物(如猪)、羊驼类动物(如骆驼、美洲驼、羊驼)、猴子、猿(如大猩猩、黑猩猩)、熊科动物(如熊)、家禽、犬、猫、小鼠、大鼠、鱼、海豚、鲸鱼和鲨鱼。对象可为雄性或雌性(例如妇女、妊娠妇女、妊娠雌性)。对象可为任何年龄(如胚胎、胎儿、婴儿、儿童、成人)。

[0033] 核酸可以从任何类型的合适生物试样或样品中分离(例如测试样品)。样品或测试样品可为分离或获自对象或其部份(如人对象、妊娠雌性、胎儿)的任何试样。通常从测试对象获得测试样品。测试样品通常获自妊娠雌性(例如,妊娠女性)。试样的非限制性示例包括对象的液体或组织,包括但不限于血液或血液制品(例如,血清、血浆等)、脐带血、绒毛、羊水、脑脊液、脊髓液、洗液(如支气管肺泡、胃、腹膜、导管、耳、关节镜)、活检样品(例如来自移植前胚胎)、膜间液样品、细胞(血液细胞,胎盘细胞、胚胎或胎儿细胞、胎儿有核细胞或胎儿细胞残余)或其部分(例如,线粒体、核、提取物等)、女性生殖道清洗物、尿、粪便、痰、唾液、鼻黏膜、前列腺液、灌洗液、精液、淋巴液、胆汁、眼泪、汗液、母乳、乳腺体液等或其组合。测试样品可包括血液或血液产品(例如,血浆、血清、淋巴细胞、血小板、棕黄层)。测试样品有时包括获自妊娠雌性的血清。测试样品有时包括获自妊娠雌性的血浆。在一些实施方式中,生物样品是来自对象的宫颈擦拭物。在一些实施方式中,生物样品可以是血液,而有时是血浆或血清。本文所用的术语“血液”指来自对象(例如,测试对象,例如,妊娠女性或就可能妊娠而作测试女性)的血液样品或制品。该术语涵盖全血、血液制品或血液的任何部分,例如常规定义的血清和血浆、棕黄层等。血液或其部分常包括核小体(例如母本和/或胎儿核小体)。核小体包括核酸且有时无细胞或为细胞内的。血液还包括棕黄层。棕黄层有时通过菲克(ficoll)梯度来分离。棕黄层可包括白血细胞(例如白细胞、T细胞、B细胞、血小板等)。在一些实施方式中,棕黄层包括母本和/或胎儿核酸。血液血浆指经抗凝剂处理的血液离心所得的全血的部分。血液血清指血液样品凝结后保留的液体水层部分。通常按照医院或临床常规遵循的标准方法来采集液体或组织样品。就血液而言,通常采集适当量的外周血(例如3-40毫升),并且在制备前或后可按标准流程保存。提取核酸所用的液体或组织样品可以是非细胞的(如无细胞)。在一些实施方式中,液体或组织样品可含有细胞要素或细胞残余物。在一些实施方式中,所述样品中可包含胎儿细胞或癌细胞。

[0034] 样品通常是异质性的,即所述样品中存在超过一种类型的核酸物质。例如,异质性核酸能包括但不限于(i) 胎儿源性和母本源性的核酸、(ii) 癌症和非癌症核酸、(iii) 病原体和宿主核酸、和更常见的(iv) 突变的和野生型核酸。样品可以是异质性的原因是,存在超过一种细胞类型,例如胎儿细胞和母本细胞,癌细胞和非癌细胞,或者病原体和宿主细胞。在一些实施方式中,存在少数核酸物质和多数核酸物质。

[0035] 就本文所述技术的产前应用而言,液体或组织样品可采自孕龄适于测试的女性或经测试可能有孕的女性。适当孕龄可能视所进行的产前测试而不同。在某些实施方式中,妊娠雌性对象有时在孕期前三个月,有时在孕中期三个月或有时在孕期末三个月。在某些实施方式中,液体或组织采自胎儿妊娠约1-约45周(如胎儿妊娠1-4、4-8、8-12、12-16、16-20、

20-24、24-28、28-32、32-36、36-40或40-44周)和有时胎儿妊娠约5-约28周(如胎儿妊娠6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26或27周)的妊娠妇女。在某些实施方式中,在分娩(例如阴道或非阴道分娩(如手术分娩))期间或刚分娩后(如0-72小时后)从妊娠雌性收集流体或组织样品。

[0036] 获取血液样品和DNA提取

[0037] 本文方法包括分离、富集和分析母本血液中所发现的胎儿DNA,作为在妊娠期间和有时妊娠后的非侵入性手段来检测是否存在母本和/或胎儿遗传变异和/或监控胎儿和/或妊娠雌性的健康。因此,实施本发明某些方法的第一步包括获取妊娠女性的血液样品和从样品提取DNA。

[0038] 获取血液样品

[0039] 血液样品可获自适合采用本发明所述方法的测试的孕龄妊娠女性。合适的妊娠年龄可根据所测疾病而不同,如下所述。收集妇女血液通常根据医院或诊所一般遵循的标准方案来进行。采集适当量的外周血,例如,通常为5-50毫升,并在进一步制备前按照标准规程保存。可以能使样品中所存在核酸量的降解最小或确保其品质的方式采集、保存或运输所述血液样品。

[0040] 制备血液样品

[0041] 采用例如全血、血清或血浆对母本血液中发现的胎儿DNA进行分析。从母本血液中制备血清或血浆的方法已知。例如,可将妊娠女性的血液置入含有避免血液凝结的EDTA或专用市售产品如Vacutainer SST(新泽西州富兰克林湖市的BD公司(Becton Dickinson))的管内,然后可通过离心从全血获取血浆。血清可通过或不通过血液凝固后的离心来获取。若使用离心,则通常(并不限于)在合适速度(例如1,500-3,000倍g)下进行。血浆或血清可在转移至用于DNA提取的新管之前经过其它离心步骤。

[0042] 除了全血的非细胞部分,DNA还可从细胞组分回收,在棕黄层部分中富集,这可通过从妇女的全血样品离心并去除血浆来获取。

[0043] 提取DNA

[0044] 有多种已知方法用于从包括血液在内的生物样品中提取DNA。可按照DNA制备的常规方法(例如,描述于Sambrook和Russell,Molecular Cloning:A Laboratory Manual(《分子克隆:实验室手册》),第3版,2001);多种市售可得试剂或试剂盒,例如凯杰公司(Qiagen)的QIAamp循环核酸试剂盒,QiaAmp DNA迷你试剂盒或QiaAmp DNA血液迷你试剂盒(德国海尔登的凯杰公司),基因组Prep™血液DNA分离试剂盒(威斯康星州麦迪逊的普洛麦格公司(Promega,Madison,Wis.))和GFX™基因组血液DNA纯化试剂盒(新泽西州皮斯卡特维的安玛西亚公司(Amersham))也可用于从来自妊娠女性的血液样品获取DNA。还可使用这些方法中多于一种的组合。

[0045] 在一些实施方式中,所述样品可首先就胎儿核酸用一种或多种方法富集或相对富集。例如,胎儿和母本DNA的区分可以采用单独的本发明所述组合物和方法进行或与其它区分因子联用。这些因子的示例包括但不限于染色体X和Y中的单核苷酸差异、染色体Y特异序列、基因组中别处的多态性、胎儿和母本DNA之间的大小差异和母本和胎儿组织之间甲基化形式的差异。

[0046] 用于就特定核酸物质富集样品的其它方法描述于2007年5月30日提交的PCT专利

申请号PCT/US07/69991,2007年6月15日提交的PCT专利申请号PCT/US2007/071232,美国临时申请号60/968,876与60/968,878(指定给本申请人)、(PCT专利申请号PCT/EP05/012707,2005年11月28日提交),这些都通过引用纳入本文。在某些实施方式中,从样品中选择性除去(部分、基本、几乎完全或完全)母本核酸。

[0047] 术语“核酸”和“核酸分子”在本文中可互换使用。该术语指任意组合物形式的核酸,来自如:DNA(例如,互补DNA(cDNA),基因组DNA(gDNA)等),RNA(例如,信使RNA(mRNA),短抑制RNA(siRNA),核糖体RNA(rRNA),tRNA,微小RNA,胎儿或胎盘高度表达的RNA等),和/或DNA或RNA类似物(例如,含有碱基类似物,糖类似物和/或非天然主链等),RNA/DNA杂交体和聚酰胺核酸(PNA),所有这些可以是单链或双链形式,且除非另有限定,可涵盖能与天然存在核苷酸相似方式起作用的天然核苷酸的已知类似物。在某些实施方式中,核酸可以是或者可来自:质粒、噬菌体、自主复制序列(ARS)、着丝粒、人工染色体、染色体、或者能够在体外或在宿主细胞、细胞、细胞的细胞核或细胞质中复制或被复制的其它核酸。在一些实施方式中,模板核酸可来自单个染色体(例如核酸样品可来自二倍体生物所得样品的一个染色体)。除非明确限定,该术语涵盖含有结合特性与参比核酸类似且与天然存在核苷酸相似方式代谢的天然核苷酸的已知类似物。除非另有说明,特定核酸序列也包括其保守修饰变体(如,简并密码子取代),等位基因,直向同源物,单核苷酸多态性(SNP)和互补序列,以及明确指出的序列。具体说,可通过产生一个或多个选定(或所有)密码子的第三个位置被混合碱基和/或脱氧肌苷残基取代的序列来获得简并密码子取代。术语核酸与基因座、基因、cDNA、和基因编码的mRNA互换使用。所述术语也可包括从核苷酸类似物、单链(“正义”或“反义”,“正”链或“负”链,“正向”阅读框或“反向”阅读框)和双链多核苷酸合成的RNA或DNA的等价物、衍生物、变体和类似物。术语“基因”指参与产生多肽链的DNA区段;其包括参与基因产物的转录/翻译和所述转录/翻译调节的编码区之前和之后的区域(前导区和尾部区),以及单个编码区段(外显子)之间的插入序列(内含子)。

[0048] 脱氧核糖核苷酸包含脱氧腺苷、脱氧胞苷、脱氧鸟苷和脱氧胸苷。就RNA而言,碱基胞嘧啶替换为尿嘧啶。模板核酸可采用获自对象的核酸作为模板制备。

[0049] 核酸分离和加工

[0050] 可用本领域已知方法从一种或多种样品来源(如细胞、血清、血浆、棕黄层、淋巴液、皮肤、土壤等)中获取核酸。核酸通常分离自测试样品。可采用任何合适的方法从生物样品(例如从血液或血液制品)中分离、提取和/或纯化DNA,非限制性示例包括DNA制备的方法(例如,描述于Sambrook和Russell,Molecular Cloning:A Laboratory Manual(《分子克隆:实验室手册》),第3版,2001);多种市售可得试剂或试剂盒,例如凯杰公司(Qiagen)的QIAamp循环核酸试剂盒,QiaAmp DNA迷你试剂盒或QiaAmp DNA血液迷你试剂盒(德国海尔德的凯杰公司),基因组PrepTM血液DNA分离试剂盒(威斯康星州麦迪逊的普洛麦格公司(Promega, Madison, Wis.))和GFXTM基因组血液DNA纯化试剂盒(新泽西州皮斯卡特维的安玛西亚公司(Amersham))等或其组合。

[0051] 细胞裂解方法和试剂是本领域已知的,且一般可通过化学(例如洗涤剂、低渗溶液、酶促过程等或其组合)、物理(例如法式压滤、超声等)或电解的裂解方法进行。能使用任何合适的裂解过程。例如化学方法通常使用裂解剂破坏细胞并从细胞中提取核酸,然后用离液盐处理。物理方法例如冷冻/解冻然后研磨,使用细胞压滤等也有用。高盐裂解法也是

常用的。例如,可采用碱裂解法。所述后一种方法传统上包括使用苯酚-氯仿溶液,且可采用替代的包括三种溶液的无苯酚-氯仿方法。在后一种方法中,一种溶液可包含15mM Tris,pH 8.0;10mM EDTA和100 μ g/ml RNA酶A;第二溶液可包含0.2N NaOH和1% SDS;以及第三溶液可包含3M KOAc,pH5.5。这些方法可参见纽约约翰韦利森公司(John Wiley&Sons, Inc., New York)的《新编分子生物学实验指南》(Current Protocols in Molecular Biology)的6.3.1-6.3.6(1989),其全文纳入本文。

[0052] 核酸还可以在与另一核酸不同的时间点分离得到,其中各样品来自相同或不同来源。核酸可来自核酸库,例如cDNA或RNA库。核酸可以是样品中核酸分子的核酸纯化或分离和/或扩增的产物。为本文所述方法提供的核酸可包含来自一个样品或来自两个或更多个样品(例如来自1个或更多个、2个或更多个、3个或更多个、4个或更多个、5个或更多个、6个或更多个、7个或更多个、8个或更多个、9个或更多个、10个或更多个、11个或更多个、12个或更多个、13个或更多个、14个或更多个、15个或更多个、16个或更多个、17个或更多个、18个或更多个、19个或更多个、20个或更多个的样品)的核酸。

[0053] 在某些实施方式中,核酸可包括胞外核酸。本文所用术语“胞外核酸”指从基本不具有细胞的来源分离的核酸,还称为“无细胞”核酸和/或“循环无细胞”核酸。胞外核酸可存在于血液中并从中获取(例如从妊娠雌性的血液中)。胞外核酸通常不包括可检测到的细胞且可含有细胞元件或细胞残余物。胞外核酸的无细胞来源的非限制性示例有血液、血浆、血清和尿液。本文所用术语“获取循环无细胞样品核酸”包括直接获取样品(如收集样品例如测试样品)或从已收集样品的人那里获取样品。不受理论限制,胞外核酸可以是细胞凋亡和细胞破裂的产物,这使胞外核酸常具有跨范围的系列长度(例如“梯状带(ladder)”)。

[0054] 在某些实施方式中,胞外核酸可包含不同的核酸物质,因而在本文中称作“异质性”。例如,患有癌症的人的血液血清或血浆可包含来自癌细胞的核酸与来自非癌细胞的核酸。在另一例子中,妊娠雌性的血液血清或血浆可包含母本核酸和胎儿核酸。在一些示例中,胎儿核酸有时占全部核酸的约5%-约50%(例如,总体核酸中约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、45、46、47、48或49%是胎儿核酸)。在一些实施方式中,核酸中的多数胎儿核酸的长度为约500碱基对或更少、约250碱基对或更少、约200碱基对或更少、约150碱基对或更少、约100碱基对或更少、约50碱基对或更少或约25碱基对或更少。

[0055] 在某些实施方式中,可不经对含核酸样品的处理而提供核酸用于进行本文所述方法。在一些实施方式中,在处理含核酸的样品后提供核酸用于进行本文所述方法。例如,可从样品提取、分离、纯化、部分纯化或扩增核酸。如本文所用的术语“分离”指将核酸从其原始环境中取出(例如,天然产生核酸的天然环境或外源表达核酸的宿主细胞),因此核酸从其原始环境通过人的干预(如“人工”)而被改变。本文所用术语“分离的核酸”指从对象(如人类对象)中移出的核酸。与来源样品中具有的成分含量相比,分离的核酸可带有较少的非核酸组分(例如,蛋白质、脂质)。包含分离的核酸的组合物可以是约50%至多于99%不含非核酸组分。包含分离的核酸的组合物可以是约90%、91%、92%、93%、94%、95%、96%、97%、98%、99%或大于99%不含非核酸组分。本文所用术语“纯化”指与将核酸经历纯化程序之前所存在的非核酸组分的含量相比,所提供的核酸带有较少的非核酸组分(例如,蛋白质、脂质、碳水化合物)。包含纯化核酸的组合物可以是约80%、81%、82%、83%、84%、

85%、86%、87%、88%、89%、90%、91%、92%、93%、94%、95%、96%、97%、98%、99%或大于99%不含其他非核酸组分。本文所用术语“纯化”可指提供的核酸与其所衍生自的样品来源相比包含更少的核酸物质。包含纯化核酸的组合物可以是约90%、91%、92%、93%、94%、95%、96%、97%、98%、99%或大于99%不含其他核酸物质。例如，胎儿核酸可从含母本和胎儿核酸的混合物中纯化。在某些示例中，含胎儿核酸的小片段的核小体可从含母本核酸的较大片段的大核小体复合物的混合物中纯化。

[0056] 在一些实施方式中，本发明方法之前、期间或之后对核酸进行片段化或切割。片段化或切割的核酸可具有约5-约10,000个碱基对、约100-约1,000个碱基对、约100-500个碱基对或约10、15、20、25、30、35、40、45、50、55、60、65、70、75、80、85、90、95、100、200、300、400、500、600、700、800、900、1000、2000、3000、4000、5000、6000、7000、8000或9000个碱基对的标称、平均或算术均值长度。可通过本领域已知的合适方法产生片段，且核酸片段的平均、等比中数或标称长度可通过选择适当的片段生成方法而加以控制。

[0057] 核酸片段可含有重叠的核苷酸序列，这样的重叠序列可促进构建未片段化的对应核酸或其区段的核苷酸序列。例如，一个片段可具有亚序列x和y，且其他片段可具有亚序列y和z，其中x、y和z是长度可为5核苷酸或更长的核苷酸序列。在某些实施方式中，重叠核酸y可用于促进从样品的核酸中构建x-y-z核苷酸序列。在某些实施方式中，核酸可以是部分片段化的（例如，来自未完全的或中止的特异性剪切反应）或完全片段化的。

[0058] 在一些实施方式中，核酸可通过合适方法进行片段化或切割，其非限制性示例包括物理方法（例如剪切、例如超声、法式压滤、热、紫外照射等）、酶加工（例如酶切割试剂（例如合适的核酸酶、合适的限制性酶、合适的甲基化敏感的限制性酶））、化学方法（例如烷基化、DMS、哌啶、酸水解、碱水解、热、等或其组合）、美国专利申请公开20050112590中所述方法等，或其组合。

[0059] 本文所用的“片段化”或“剪切”指使核酸分子（如核酸模板基因分子或其扩增产物）可以分成两个或更多较小核酸分子的方法或条件。这种片段化或剪切可以是序列特异性、碱基特异性或非特异性的，并且能通过任意不同方法、试剂或条件（包括例如化学、酶、物理片段化）来完成。

[0060] 本文所用的“片段”、“剪切产物”、“经剪切的产物”或其语法变体指由核酸模板基因分子或其扩增产物的片段化或剪切获得的核酸分子。尽管这种片段或剪切产物可指由剪切反应获得的所有核酸，但是这种片段或剪切产物通常仅指由核酸模板基因分子或其扩增产物区段（包含核酸模板基因分子的相应核苷酸序列）的片段化或剪切获得的核酸分子。如本文所用术语“扩增”是指使处理样品中的靶核酸经过以线性或指数形式产生扩增子核酸的过程，所述扩增子核酸的核苷酸序列与靶核酸或其区段的核苷酸序列相同或基本相同。在某些实施方式中，术语“扩增”指包括聚合酶链式反应（PCR）的方法。例如，扩增产物能含有比核酸模板序列的扩增核苷酸区域多一个或多个的核苷酸（如引物能包含除了与核酸模板基因分子互补的核苷酸以外的“额外”核苷酸例如转录起始序列，生成包含“额外”核苷酸或者与所述核酸模板基因分子的扩增核苷酸区域不对应的核苷酸的扩增产物）。因此，片段能包含来自扩增的核酸分子区段或部分的片段，所述核酸分子至少部分包含来自或基于代表性核酸模板分子的核苷酸序列信息。

[0061] 本文所用的术语“互补剪切反应”是指用不同剪切试剂或者通过改变相同剪切试

剂的剪切特异性在相同核酸上进行的剪切反应,从而产生相同目标或参比核酸或蛋白质的不同剪切模式。在某些实施方式中,可以用一种或多种特异性剪切剂(例如1、2、3、4、5、6、7、8、9、10或更多种特异性剪切剂)在一个或多个反应容器中处理核酸(例如用各种特异性剪切剂在单独的容器内处理核酸)。如本文所用术语“特异性剪切剂”指试剂,有时是可在一个或多个特异性位点处剪切核酸的化学品或酶。

[0062] 在提供核酸用于本文所述方法之前,还可对核酸进行处理修饰核酸中某些核苷酸。例如,可对核酸施用根据核酸中核苷酸的甲基化状态选择性修饰核酸的处理。此外,诸如高温、紫外辐射、x-射线辐射等条件可诱导核酸分子序列中的变异。可以用于进行合适序列分析的任何合适形式提供核酸。

[0063] 核酸可为单链或双链。例如,可通过加热或(例如)用碱处理来变性双链DNA来生成单链DNA。在某些实施方式中,核酸是D环结构,通过双链DNA分子中链入侵有寡核苷酸或DNA样分子例如肽核酸(PNA)来形成。添加大肠杆菌RecA蛋白质和/或改变盐浓度(例如使用本领域已知方法)有助于形成D环。

[0064] 确定胎儿核酸含量

[0065] 在一些实施方式中,确定核酸中胎儿核酸的量(例如,浓度、相对量、绝对量、拷贝数等)。在某些实施方式中,样品中胎儿核酸的量称为“胎儿分数”。在一些实施方式中,“胎儿分数”指获自妊娠雌性的样品(例如血液样品、血清样品、血浆样品)中的循环无细胞核酸中的胎儿核酸分数。在某些实施方式中,根据下述内容确定胎儿核酸的含量:对雄性胎儿特异的标记物(例如Y染色体STR标记物(例如DYS 19、DYS 385、DYS 392标记物);RhD阴性女性中的RhD标记物)、多态性序列的等位基因比例、或对胎儿核酸特异而对母本核酸非特异的一种或多种标记物(例如母本和胎儿之间的差异性表观遗传学生物标记(例如甲基化;如下详述)、或母本血浆中的胎儿RNA标记物(参见例如Lo,2005,Journal of Histochemistry and Cytochemistry 53(3):293-296))。

[0066] 确定胎儿核酸含量(例如胎儿分数)有时用胎儿定量试验(FQA)进行,如美国专利申请公开2010/0105049所述,其通过引用纳入本文。此类试验允许基于样品中核酸的甲基化状态检测和定量母本样品中的胎儿核酸。在某些实施方式中,母本样品中胎儿核酸的含量可相对存在的核酸总量进行确定,从而提供样品中胎儿核酸的百分比。在某些实施方式中,可确定母本样品中胎儿核酸的拷贝数。在某些实施方式中,可以序列特异性(或部份-特异性)方式确定胎儿核酸的量,且有时灵敏度足以进行精确的染色体剂量分析(例如,用以检测胎儿非整倍性的存在与否)。

[0067] 胎儿定量试验(FQA)可与本文所述任何方法联合进行。可通过本领域任何已知方法和/或美国专利申请公开2010/0105049所述进行该试验,例如通过可基于差异性甲基化状态区分母本和胎儿DNA的方法、以及定量胎儿DNA(例如,确定其含量)的方法。基于甲基化状态区分核酸的方法包括但不限于甲基化敏感性捕获(例如使用MBD2-Fc片段,其中MBD2的甲基化结合结构域融合至抗体的Fc片段(MBD-FC)(Gebhard等(2006)Cancer Res.66(12):6118-28));甲基化特异抗体、亚硫酸氢盐转化法,例如MSP(甲基化敏感的PCR)、COBRA、甲基化敏感的单核苷酸引物延伸(Ms-SNuPE)或塞昆纳姆股份有限公司(Sequenom)MassCLEAVE™技术;和甲基化敏感的限制性酶的应用(例如用一种或多种甲基化敏感的限制性酶消化母本样品中的母本DNA,从而富集胎儿DNA)。甲基敏感的酶还可用于基于甲基化状态区分核

酸,例如其DNA识别序列未甲基化时优选或显著切割或消化。因此,未经甲基化的DNA样品会被切成比甲基化样品小的片段,而高甲基化的DNA样品不会被切割。除非明确指出,基于甲基化状态的任何区分核酸的方法均可用于本发明组合物和方法。可通过例如在扩增反应期间引入已知浓度的一种或多种竞争剂来确定胎儿DNA的含量。还可通过例如RT-PCR、引物延伸、测序和/或计数来确定胎儿DNA的含量。在某些示例中,可用美国专利申请公开2007/0065823中所述的BEAMing技术来确定核酸的含量。在某些实施方式中,可确定限制性功效并用该效率比进一步确定胎儿DNA的量。

[0068] 在某些实施方式中,胎儿定量试验(FQA)可用母本样品中的胎儿DNA浓度来确定,例如通过下述方法:a)确定母本样品中存在的DNA总量;b)用一种或多种甲基化敏感的限制性酶选择性消化母本样品中的母本DNA从而富集所述胎儿DNA;c)确定来自步骤b)的胎儿DNA量;以及d)比较步骤c)所得胎儿DNA的量与步骤a)所得DNA的总量,从而确定母本样品中胎儿DNA的浓度。在某些实施方式中,可确定母本样品中胎儿核酸的绝对拷贝数,例如,采用质谱和/或利用针对绝对拷贝数确定的竞争性PCR方法的系统。参见例如Ding和Cantor(2003)PNAS.USA 100:3059-3064,和美国专利申请公开2004/0081993,其均通过引用纳入本文。

[0069] 在某些实施方式中,可基于多肽性序列的等位基因比例(例如单核苷酸多态性(SNP))来确定胎儿分数,例如使用美国专利申请公开2011/0224087中所述的方法,其通过引用纳入本文。该方法中,就母本样品获取核苷酸序列读数,并通过比较映射至第一等位基因的核苷酸序列读数的总数与映射至位于参照基因组中的参照性多态位点(如SNP)的第二等位基因的核苷酸序列读数的总数来确定胎儿分数。在某些实施方式中,通过例如在样品中的胎儿和母本核酸的混合物中,相对于母本核酸对混合物的较大贡献,胎儿等位基因的相对较小贡献来鉴定胎儿等位基因。因此,母本样品中胎儿核酸的相对丰度可作为(就多态位点的两个等位基因中每个来说)映射至参照基因组上靶核酸序列的独特序列读数的总数的参数来确定。

[0070] 在某些实施方式中,可基于一种或多种水平来确定胎儿分数。按照水平确定胎儿分数描述于,例如,国际申请公开号W0 2014/055774,其全部内容通过引用纳入本文,包括所有文字、表格、等式和附图。在一些实施方式中,胎儿分数根据分类为代表母本和/或胎儿拷贝数变异的水平来确定。例如,胎儿分数的确定可包括评估用于确定胎儿分数的母本和/或胎儿拷贝数变异的预期水平。在一些实施方式中,对于如下情况确定胎儿分数:根据确定的相同类型的拷贝数变异的预期水平范围,分类为代表拷贝数变异的水平(例如,第一水平)。可根据落在预期水平范围中的观察到的水平,由此分类为母本和/或胎儿拷贝数变异,来确定胎儿分数。在一些实施方式中,当分类为母本和/或胎儿拷贝数变异的观察到的水平(例如,第一水平)与确定的相同的母本和/或胎儿拷贝数变异的预期水平不同时,确定胎儿分数。胎儿分数可以百分数提供。例如,胎儿分数可除以100,由此提供百分数值。例如,对于代表母本纯合重复且具有155的水平的第一水平和用于母本纯合重复且具有150的水平的预期水平,胎儿分数可确定为10%(例如,(胎儿分数=2x(155-150)))。

[0071] 细胞外核酸中的胎儿核酸的量可定量,并可与本文所述方法联合使用。因此,在某些实施方式中,本文所述技术的方法包括确定胎儿核酸的量的额外步骤。可在加工以制备样品核酸之前或之后确定对象的核酸样品中胎儿核酸的量。在某些实施方式中,在样品核

酸经加工和制备后,确定样品中胎儿核酸的量,并用于进一步评价。在一些实施方式中,结果包括因数化样品核酸中的胎儿核酸分数(如调整计数、除去样品、作出判定或不作出判定)。在某些实施方式中,本文提供的方法可与用于确定胎儿分数的方法联用。例如,包括标准化过程的用于确定胎儿分数的方法可包括本文提供的一个或多个标准化方法(例如,主成分标准化)。

[0072] 所述确定步骤可在本文所述方法之前、期间、本文所述方法之中任何时间点或本文所述某些方法之后(例如非整倍性检测、胎儿性别确定)进行。例如,为了实现有给定灵敏度或特异性的胎儿性别或非整倍性确定方法,胎儿核酸定量方法可以在胎儿性别或非整倍性确定之前、中或后进行,以鉴定有大于约2%、3%、4%、5%、6%、7%、8%、9%、10%、11%、12%、13%、14%、15%、16%、17%、18%、19%、20%、21%、22%、23%、24%、25%或更多胎儿核酸的那些样品。在一些实施方式中,被确定为具有某一胎儿核酸阈值量(如约15%或更多胎儿核酸;如约4%或更多胎儿核酸)的样品进一步用于分析例如胎儿性别或非整倍性确定,或是否存在非整倍性或遗传变异。在某些实施方式中,仅选择(例如,选择并告知患者)具有某一胎儿核酸阈值量(如约15%或更多胎儿核酸;如约4%或更多胎儿核酸)的样品确定例如胎儿性别或是否存在非整倍性。

[0073] 在一些实施方式中,确定胎儿分数或确定胎儿核酸的量并非鉴定是否存在染色体非整倍性所必需。在一些实施方式中,鉴定是否存在染色体非整倍性不需要胎儿与母本DNA的序列区分。在某些实施方式中,这是由于对母本和胎儿序列对具体染色体、染色体部份或其区段的加成贡献进行了分析。在一些实施方式中,鉴定是否存在染色体非整倍性不依赖于区分胎儿DNA和母本DNA的先验序列信息。

[0074] 富集核酸

[0075] 在一些实施方式中,针对核酸亚群或物质对核酸(例如胞外核酸)进行富集或相对富集。核酸亚群可包括例如胎儿核酸、母本核酸、含特定长度或长度范围的片段的核酸、或来自特定基因组区域(例如单一染色体、染色体组、和/或某些染色体区域)的核酸。此类富集的样品可与本文所述方法联合使用。因此,在某些实施方式中,该技术的方法包括富集样品中核酸亚群例如胎儿核酸的额外步骤。在某些实施方式中,上述确定胎儿分数的方法也可用于富集胎儿核酸。在某些实施方式中,从样品中选择性除去(部分、基本、几乎完全或完全)母本核酸。在某些实施方式中,富集特定低拷贝数的核酸(例如胎儿核酸)可提高定量敏感性。富集样品中特定种类核酸的方法例如下述,美国专利号6,927,028、国际申请公开号W02007/140417、国际申请公开号W02007/147063、国际申请公开号W02009/032779、国际申请公开号W02009/032781、国际申请公开号W02010/033639、国际申请公开号W02011/034631、国际申请公开号W02006/056480和国际申请公开号W02011/143659,其都通过引用纳入本文,包括所有文字、表格、等式和附图。

[0076] 在一些实施方式中,针对某些靶片段种类和/或参照片段种类富集核酸。在某些实施方式中,使用下述一种或多种基于长度的分离方法就特定核酸片段长度或片段长度或范围进行核酸富集。在某些实施方式中,使用本文所述和/或本领域已知的一种或多种基于序列的分离方法就选自基因组区域(例如染色体)的片段进行核酸富集。某些富集样品中核酸亚群(如胎儿核酸)的方法如下详述。

[0077] 可与本发明方法一起用的富集核酸亚群(例如胎儿核酸)的方法包括采用母本和

胎儿核酸之间的表观差异的方法。例如可基于甲基化差异区分并分离胎儿核酸和母本核酸。基于甲基化的胎儿核酸富集方法参见美国专利申请公开2010/0105049,其通过引用纳入本文。该方法有时涉及结合样品核酸与甲基化特异的结合试剂(甲基CpG结合蛋白(MBD)、甲基化特异抗体等)并基于不同的甲基化状态分离结合的核酸和未结合的核酸。此类方法还可包括使用甲基化敏感的限制酶(如上所述例如HhaI和HpaII),其通过使用选择性且完全或基本消化母本核酸的酶来选择性消化来自母本样品的核酸从而富集样品中至少一种胎儿核酸区域,这样就能够富集母本样品中的胎儿核酸区域。

[0078] 可与本发明方法一起用的其他富集核酸亚群(例如胎儿核酸)的方法是限制性内切核酸酶增强的多态性序列法,例如美国专利申请公开2009/0317818所述的方法,其通过引用纳入本文。该方法包括用识别含非靶等位基因但不识别靶等位基因的限制性内切核酸酶切割含所述非靶等位基因的核酸;并扩增未切割的核酸但不扩增切割的核酸,其中该未切割的扩增核酸代表相对于非靶核酸(如母本核酸)富集的靶核酸(如胎儿核酸)。在某些实施方式中,可选择核酸从而其包含具有多态性位点的等位基因,所述多态性位点易于被例如切割剂选择性消化。

[0079] 可与本发明方法一起用的富集核酸亚群(例如胎儿核酸)的方法包括选择性酶降解法。该方法涉及保护靶序列免受外切核酸酶的消化,从而有利于消除样品中不需要的序列(例如母本DNA)。例如,在一个方法中,样品核酸变性产生单链核酸,单链核酸在合适的退火条件下接触至少一种靶特异引物对,退火的引物利用核苷酸聚合进行延伸产生双链靶序列,并用消化单链(例如非靶)核酸的核酸酶消化单链核酸。在某些实施方式中,所述方法至少可多重复一循环。在某些实施方式中,可使用相同的靶特异引物对来起始延伸的第一和第二循环,且在某些实施方式中,不同的靶特异引物对用于第一和第二循环。

[0080] 可与本发明方法一起用的富集核酸亚群(例如胎儿核酸)的方法包括大规模平行测序技术(MPSS)。MPSS通常是固相方法,其使用衔接子(例如,标签)连接,然后衔接子解码,并以小的增量读取核酸序列。带标签的PCR产物通常经扩增,从而各核酸产生具有独特标签的PCR产物。标签通常用于接合PCR产物至微珠。例如,基于连接的序列确定进行数轮后,可从各珠鉴定序列签名。分析MPSS数据库中的各签名序列(MPSS标签),比较所有其他签名,并计数所有相同的签名。

[0081] 在一些实施方式中,某些富集方法(如某些基于MPS和/或基于MPSS的富集方法)可包括基于扩增(如PCR)的方法。在一些实施方式中,可使用位点特异的扩增方法(例如使用位点特异的扩增引物)。在某些实施方式中,可使用多重SNP等位基因PCR方法。在某些实施方式中,多重SNP等位基因PCR方法可与单重测序联合使用。例如,该方法可涉及使用多重PCR(MASSARRAY系统)并将捕获探针序列纳入扩增子,然后使用例如亿明达(Illumina)MPSS系统测序。在某些实施方式中,多重SNP等位基因PCR方法可与三引物系统和索引测序联合使用。例如,该方法可涉及使用多重PCR(MASSARRAY系统),所用引物将第一捕获探针纳入某些位点特异的正向PCR引物,并将衔接子序列纳入位点特异的反向PCR引物,从而产生扩增子,然后二级PCR将反向捕获序列和分子索引条码纳入,用于使用例如亿明达MPSS系统的测序。在一些实施方式中,多重SNP等位基因PCR方法可与四引物系统和索引测序联合使用。例如,该方法可涉及使用多重PCR(MASSARRAY系统),所用引物将衔接子序列纳入位点特异的正向和位点特异的反向PCR引物,然后二级PCR将正向和反向捕获序列和分子索引条码纳

入,用于使用例如亿明达MPSS系统的测序。在某些实施方式中,可使用微流体方法。在某些实施方式中,可使用基于阵列的微流体方法。例如,该方法可涉及使用微流体阵列(如Fluidigm)用于低重扩增并纳入索引和捕获探针,然后测序。在某些实施方式中,可使用乳液微流体方法,例如数字液滴PCR。

[0082] 在某些实施方式中,可使用通用扩增方法(例如使用通用或非位点特异的扩增引物)。在一些实施方式中,通用扩增方法可与拉下(pull-down)法联用。在一些实施方式中,方法可包括从通用扩增序列库中拉下生物素化的Ultramer(例如安捷伦或IDT的生物素化拉下试验)。例如,该方法可涉及制备标准库,通过拉下试验富集选择区域,以及二级通用扩增步骤。在某些实施方式中,拉下法可与基于连接的方法联合使用。在某些实施方式中,方法可包括用序列特异的衔接子连接的生物素化的Ultramer拉下(例如HALOPLEX PCR,Halo Genomics公司)。例如,该方法可涉及使用选择器探针来捕获限制性酶-消化片段,然后连接捕获产物和衔接子,和通用扩增然后测序。在某些实施方式中,拉下法可与延伸和基于连接的方法联合使用。在某些实施方式中,方法可包括分子倒置探针(MIP)延伸和连接。例如,该方法可涉及分子倒置探针与序列衔接子组合使用,然后进行通用扩增和测序。在某些实施方式中,互补DNA可合成和测序而无需扩增。

[0083] 在某些实施方式中,可进行延伸和连接方法而无需拉下组件。在某些实施方式中,方法可包括位点特异的正向和反向引物杂交、延伸和连接。该方法还可包括通用扩增或互补DNA合成而无需扩增,然后测序。在某些实施方式中,该方法可在分析期间降低或排除背景序列。

[0084] 在某些实施方式中,拉下法可与任选的扩增组件一起使用或不与扩增组件一起使用。在某些实施方式中,方法可包括修改的拉下试验和连接,其完全纳入捕获探针而无需通用扩增。例如,该方法可涉及使用修改的选择器探针来捕获限制性酶-消化片段,然后连接捕获产物和衔接子,和任选扩增,和测序。在某些实施方式中,方法可包括生物素化的拉下试验,和使用衔接子序列的延伸和连接与环单链连接的组合。例如,该方法可涉及可涉及使用选择器探针捕获感兴趣的区域(例如,靶序列)、延伸探针、衔接子连接、单链环形连接、任选的扩增,和测序。在某些实施方式中,测序结果的分析可分离靶序列和背景。在一些实施方式中,使用本文所述一种或多种基于序列的分离方法就选择基因组区域(例如染色体)的片段进行核酸富集。基于序列的分离通常基于样品中感兴趣的片段中存在而其他片段中基本不存在或其他片段中不存在实质含量(例如5%或更少)的核苷酸序列(例如靶片段和/或参照片段)。在一些实施方式中,基于序列的分离可生成分离的靶片段和/或分离的参照片段。分离的靶片段和/或分离的参照片段通常从核酸样品中的剩余片段分离出来。在某些实施方式中,分离的靶片段和分离的参照片段还可从彼此分离(例如在分别的试验隔室中分离)。在某些实施方式中,分离的靶片段和分离的参照片段可一起分离(例如在相同试验室中分离)。在一些实施方式中,未结合的片段可差异移除或降解或消化。

[0085] 在一些实施方式中,选择性核酸捕获方法用于从核酸样品分离靶片段和/或参照片段。市售可得的核酸捕获系统包括例如,Nimblegen序列捕获系统(威斯康星州麦迪逊的罗氏公司(Roche)的NimbleGen);亿明达BEADARRAY平台(加利福尼亚州圣迭戈的亿明达公司(Illumina));昂飞GENECHIP平台(加利福尼亚州圣克拉拉的昂飞公司(Affymetrix));安捷伦SureSelect目标富集系统(加利福尼亚州圣克拉拉的安捷伦技术公司(Agilent

Technologies))；和相关平台。该方法通常涉及捕获寡核苷酸与靶片段或参照片段的区段或所有核苷酸序列的杂交且可包括使用固相(例如固相阵列)和/或基于溶液的平台。捕获寡核苷酸(有时称为“饵(bait)”)可经选择或设计,从而其优选杂交所选基因组区域或位点的核酸片段(例如染色体21、18、13、x或Y之一,或参照染色体)。在某些实施方式中,可使用基于杂交的方法(例如使用寡核苷酸阵列)从某些染色体(例如可能的非整倍性染色体、参照染色体或其它感兴趣的染色体)或其感兴趣的区段中富集核酸序列。

[0086] 在一些实施方式中,使用一种或多种基于长度的分离方法针对特定核酸片段长度、长度范围、低于或高于特定阈值或截止值的长度来富集核酸。核酸片段长度通常指片段中的核苷酸数量。核酸片段长度有时还指核酸片段大小。在一些实施方式中,基于长度的分离方法不需要测量个体片段的长度。在一些实施方式中,基于长度的分离方法与确定个体片段的长度的方法联合进行。在一些实施方式中,基于长度的分离指大小分级处理,其中所有或部分分级库可经分离(例如保留)和/或分析。大小分级处理为本领域已知(如阵列分离、分子筛分离、凝胶电泳分离、柱色谱分离(例如尺寸排阻柱)和基于微流体的方法)。在某些实施方式中,基于长度的分离法可包括例如片段环化、化学处理(例如甲醛、聚乙二醇(PEG))、质谱和/或大小特异的核酸扩增。

[0087] 可与本发明方法一起使用的某些基于长度的分离方法使用例如选择性序列加标签法。术语“序列加标签”指将可识别的独特序列纳入核酸或核酸群。本文所用术语“序列加标签”与本文后述术语“序列标签”含义不同。在该序列加标签方法中,样品中的片段大小种类(例如短片段)核酸经过选择性序列加标签,所述样品包括长和短核酸。该方法通常涉及使用巢式引物组进行核酸扩增反应,所述引物包括内部引物和外部引物。在某些实施方式中,内部引物之一或二者可加标签从而在靶扩增产物上引入标签。外部引物通常不与载有(内部)靶序列的短片段退火。内部引物可与短片段退火并产生载有标签和靶序列的扩增产物。通常,长片段的加标签受到组合机制的抑制,所述组合机制包括例如,外部引物的先前退火和延伸引起的内部引物延伸受阻。富集带标签的片段可通过多种方法中任一来实现,包括例如,单链核酸的外切核酸酶消化和使用对至少一种标签特异的扩增引物进行的带标签片段的扩增。

[0088] 可与本发明方法一起使用的其他基于长度的分离方法涉及将核酸样品经过聚乙二醇(PEG)沉淀。方法的示例包括国际专利申请公开号W02007/140417和W02010/115016中所述的那些,其各自全部内容通过引用纳入本文,包括所有文字、表格、等式和附图。该方法通常需要在存在一种或多种单价盐时在足以大量沉淀大核酸而不大量沉淀小(例如少于300核苷酸)核酸的条件下将核酸样品与PEG接触。

[0089] 可与本文所述方法一起使用的其他基于大小的富集方法涉及通过连接进行的环化,例如使用环化酶。短核酸片段通常可比长片段更高效的环化。非-环化序列可与环化序列分离,并且富集的短片段可用于进一步分析。

[0090] 核酸库

[0091] 在一些实施方式中,核酸库是针对具体过程进行制备、装配和/或修改的多种多核苷酸分子(例如核酸样品),所述具体过程的非限制性示例包括在固相(例如固体支持物,例如流动池、珠)上固定、富集、扩增、克隆、检测和/或用于核酸测序。在某些实施方式中,在测序过程之前或期间制备核酸库。核酸库(例如测序库)可用本领域已知的合适方法制备。可

通过靶向或非靶向制备过程制备核酸库。

[0092] 在一些实施方式中,核酸库经修饰以包括化学部分(例如功能基团),其配置为用于将核酸固定至固体支持物。在一些实施方式中,核酸库经修饰以包括生物分子(例如功能基团)和/或结合对成员,其配置为用于将库固定至固体支持物,其非限制性示例包括甲状腺素结合球蛋白、类固醇结合蛋白质、抗体、抗原、半抗原、酶、血凝素、核酸、抑制剂、蛋白质A、蛋白质G、抗生物素蛋白、链霉亲和素、生物素、补体组分C1q、核酸-结合蛋白质、受体、碳水化合物、寡核苷酸、多核苷酸、互补核酸序列等及其组合。特异结合对的一些示例包括但不限于:抗生物素蛋白部分和生物素部分;抗原表位和抗体或其免疫活性片段;抗体和半抗原;地高辛配基部分和抗地高辛配基抗体;荧光素部分和抗荧光素抗体;操纵子和抑制剂;核酸酶和核苷;凝集素和多糖;类固醇和类固醇结合蛋白;活性化合物和活性化合物受体;激素及激素受体;酶和底物;免疫球蛋白和蛋白A;寡核苷酸或多多核苷酸和其相应的互补物;等等或其组合。

[0093] 在一些实施方式中,核酸库经修改以包括已知组成的一种或多种多核苷酸,其非限制性示例包括鉴定物(例如标签、索引标签)、捕获序列、标记衔接子、限制性酶位点、启动子、增强子、复制起点、茎环、互补序列(例如引物结合位点、退火位点)、合适的整合位点(例如转座子、病毒整合位点)、修饰的核苷酸等或其组合。已知序列的多核苷酸可加入合适的位置,例如5'末尾、3'末尾或核酸序列内部。已知序列的多核苷酸可为相同或不同序列。在一些实施方式中,已知序列多核苷酸配置为与固定在表面(例如流动池的表面)的一种或多种寡核苷酸杂交。例如,核酸分子的5'已知序列可与第一多种寡核苷酸杂交,而3'已知序列可与第二多种寡核苷酸杂交。在一些实施方式中,核酸库可包括染色体-特异的标签、捕获序列、标记和/或衔接子。在一些实施方式中,核酸库包括一种或多种可检测标记。在一些实施方式中,一种或多种可检测标记可纳入核酸库的5'末端、3'末端和/或库中核酸的任何核苷酸位置。在一些实施方式中核酸库包括杂交的寡核苷酸。在某些实施方式中,杂交的寡核苷酸是标记探针。在一些实施方式中,固定在固相上之前,核酸库包括杂交的寡核苷酸探针。

[0094] 在一些实施方式中,已知序列的多核苷酸包括通用序列。通用序列是整合入两种或更多核酸分子或两种或更多核酸分子子集的特定核苷酸序列,其中所述通用序列就其所整合入的所有分子或分子子集来说相同。通用序列通常设计成使用与通用序列互补的单个通用引物进行杂交和/或扩增多种不同序列。在一些实施方式中,使用两种(例如一对)或更多通用序列和/或通用引物。通用引物通常包括通用序列。在一些实施方式中,衔接子(例如通用衔接子)包括通用序列。在一些实施方式中,一种或多种通用序列用于捕获、鉴定和/或检测多种核酸物质或其子集。

[0095] 在制备核酸库的某些实施方式中,(例如合成程序的某些测序中),核酸的大小经选择和/或经片段化为数百碱基对或更少的长度(例如库生成制备中)。在一些实施方式中,进行库制备无需片段化(例如使用ccfDNA时)。

[0096] 在某些实施方式中,使用基于连接的库制备方法(例如ILLUMINA TRUSEQ,加利福尼亚州圣迭戈亿明达)。基于连接的库制备方法通常使用衔接子(例如甲基化衔接子)设计,其可在初始连接步骤处纳入索引序列并通常可用于制备样品用于单一读数测序、双末端测序和多重测序。例如,有时核酸(例如片段化核酸或ccfDNA)通过填入(fill-in)反应、内切

核酸酶反应或其组合进行末端修复。在一些实施方式中,随后可将所得钝末端修复核酸延伸单个核苷酸,其与衔接子/引物的3'末端的单核苷酸突出互补。任何核苷酸可用于延伸/突出的核苷酸。在一些实施方式中,核酸库制备包括连接衔接子寡核苷酸。衔接子寡核苷酸通常与流动池锚互补,并且有时用于固定核酸库至固体支持物,例如流动池的内表面。在一些实施方式中,衔接子寡核苷酸包括鉴定器、一种或多种测序引物杂交位点(例如与通用测序引物互补的序列、单末端测序引物、双末端测序引物、多重测序引物等)或其组合(例如衔接子/测序、衔接子/鉴定物,衔接子/鉴定器/测序)。

[0097] 鉴定物可为纳入或接合核酸(例如多核苷酸)的合适的可检测标记,其允许检测和/或鉴定包括该鉴定物的核酸。一些实施方式中鉴定物在测序方法期间纳入或接合核酸(例如通过聚合酶)。鉴定物的非限制性示例包括核酸标签、核酸索引或条码、放射标记(例如同位素)、金属标记、化学发光标记、磷光标记、荧光淬灭剂、染料、蛋白质(例如酶、抗体或其部分、连接子、结合对的成员)等或其组合。在一些实施方式中鉴定物(例如核酸索引或条码)是核苷酸或核苷酸类似物的独特、已知和/或可鉴定的序列。在一些实施方式中,鉴定物是六个或更多连续核苷酸。可用许多具有各种不同激发和发射光谱的荧光团。任何合适的类型和/或数量的荧光团可用作鉴定物。在一些实施方式中,1个或更多、2个或更多、3个或更多、4个或更多、5个或更多、6个或更多、7个或更多、8个或更多、9个或更多、10个或更多、20个或更多、30或更多或50或更多个不同鉴定物用于本文所述方法(例如核酸检测和/或测序方法)。在一些实施方式中,一种或两种类型的鉴定物(例如荧光标记)连接至库中各核酸。鉴定物检测和/或定量可通过合适的方法、机器或设备进行,其非限制性示例包括流式细胞术、定量聚合酶链式反应(qPCR)、凝胶电泳、照度计、荧光仪、分光光度计、合适的基因芯片或微阵列分析、Western印迹、质谱、色谱、细胞荧光分析、荧光显微镜、合适的荧光或数字成像方法、共焦激光扫描显微镜、激光扫描流式细胞术、亲和层析、手动批处理模式分离、电场悬浮,合适的核酸测序方法和/或核酸测序装置等及其组合。

[0098] 在一些实施方式中,使用基于转座子的库制备方法(例如EPICENTRE NEXTERA, Epicentre, 威斯康星州麦迪逊)。基于转座子的方法通常在单管反应中使用体外移位至相似片段或标签DNA(通常允许纳入平台特异的标签和任选的条码),并制备测序仪准备库。

[0099] 在一些实施方式中核酸库或其部分经扩增(例如通过基于PCR的方法扩增)。在一些实施方式中测序方法包括扩增核酸库。核酸库可在固定至固体支持物之前或之后进行扩增(例如流动槽中的固体支持物)。核酸扩增包括扩增或增加(例如核酸库中)存在的核酸模板和/或其互补物数量的过程,所述过程通过产生一个或多个模板和/或其互补物的拷贝实现。扩增可通过合适的方法进行。核酸库可通过热循环法或通过等温扩增法进行扩增。在一些实施方式中,使用滚环扩增方法。在一些实施方式中,扩增发生在固体支持物上(例如流动池内)其中核酸库或其部份经固定。某些测序方法中,核酸库加入流动池中并通过合适条件下与锚杂交而固定。此类核酸扩增通常称为固相扩增。在固相扩增的一些实施方式中,所有或部分扩增产物通过从固定化引物开始延伸而合成。固相扩增反应与标准溶液相扩增类似,除了至少一种所述扩增寡核苷酸(例如引物)固定在固体支持物上。

[0100] 在一些实施方式中,固相扩增包括核酸扩增反应,其包括固定在表面上的仅一种寡核苷酸引物。在某些实施方式中,固相扩增包括多种不同固定化寡核苷酸引物物质。在一些实施方式中,固相扩增可包括核酸扩增反应,其包括固定在固体表面的一种寡核苷酸引

物和溶液中的第二种不同寡核苷酸引物。可使用多种不同固定化或溶液引物。固相核酸扩增反应的非限制性示例包括界面扩增、桥式扩增、乳液PCR、WildFire扩增(例如美国专利申请US20130012399)等或其组合。

[0101] 测序

[0102] 在一些实施方式中,核酸(例如核酸片段、样品核酸、无细胞核酸)经测序。在某些实施方式中,获得全序列或基本全序列,有时获得部分序列。

[0103] 在一些实施方式中,在测序之前或期间富集和/或扩增样品中的一些或所有核酸(例如非-特异性,例如通过基于PCR的方法)。在某些实施方式中,在测序之前或期间富集和/或扩增样品中的特定核酸部份或子集。在一些实施方式中,对预选核酸集合的部份或子集进行随机测序。在一些实施方式中,在测序之前或期间不富集和/或扩增样品中的核酸。

[0104] 本文所用的“读数”(例如,“一个读数”、“序列读数”)是由本文所述或本领域已知的任意测序方法生成的短核苷酸序列。读数能从核酸片段的一个末端生成(“单末端读数”),而有时从核酸片段的两个末端生成(例如双末端读数、双末端读数)。

[0105] 序列读数的长度通常与具体测序技术相关。例如高通量方法提供了大小能由数十到数百碱基对(bp)变化的序列读数。例如纳米孔测序提供大小能由数十到数百到数千碱基对变化的序列读数。在一些实施方式中,序列读数是长度约15bp-约900bp长的算术均值、中位数、平均或绝对长度。在某些实施方式中,所述序列读数是长度约1000bp或更长的算术均值、中位数、平均或绝对长度。

[0106] 在一些实施方式中,单末端读数的标称、平均、算术均值或绝对长度有时是约15个连续核苷酸-约50个或更多连续核苷酸,有时是约15个连续核苷酸-约40个或更多连续核苷酸,以及有时是约15个连续核苷酸或约36个或更多连续核苷酸。在某些实施方式中,单末端读数的标称、平均、算术均值或绝对长度是约20-约30个碱基、或约24-约28个碱基。在某些实施方式中,单末端读数的标称、平均、算术均值或绝对长度是约1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、21、22、23、24、25、26、27、28或约29个或更多碱基。

[0107] 在某些实施方式中,配对末端读数的标称、平均、算术均值或绝对长度有时是约10个连续核苷酸-约25个连续核苷酸或更多(如约10、11、12、13、14、15、16、17、18、19、20、21、22、23、24或25个核苷酸长或更多),约15个连续核苷酸-约20个连续核苷酸或更多,以及有时是约17个连续核苷酸或约18个连续核苷酸。

[0108] 读数通常是生理核酸中核苷酸序列的表示。例如,读数中用ATGC描述序列,生理核酸中“A”表示腺嘌呤核苷酸、“T”表示胸腺嘧啶核苷酸、“G”表示鸟嘌呤核苷酸、“C表示胞嘧啶核苷酸。”获自妊娠雌性血液的序列读数可为胎儿和母本核酸混合物的读数。相对较短读数的混合可通过本文所述方法变换为妊娠雌性和/或胎儿中基因组核酸的表示。相对较短读数的混合可变换为表示例如拷贝数变异(例如母本和/或胎儿拷贝数变异)、遗传变异或非整倍性。母本和胎儿核酸的混合物的读数可变换为表示包含母本和胎儿染色体之一或二者的特征的复合染色体或其区段。在某些实施方式中,从对象样品“获得”核酸序列读数,和/或从一个或多个参照个体的生物样品“获得”核酸序列读数能直接涉及测序核酸以获得序列信息。在一些实施方式中,“获得”可涉及接收从其他核酸直接获得的序列信息。

[0109] 在一些实施方式中,基因组的表示组分经测序并且有时称为“覆盖率”或“倍覆盖率”。例如,1-倍覆盖率表明约100%的基因组的核苷酸序列通过读数表示。在一些实施方式

中,“倍覆盖率”是用先前测序运行作为参照的相关术语。例如,第二轮测序可比第一轮测序具有少2-倍的覆盖率。在一些实施方式中,用冗余度对基因组测序,其中所述基因组的给定区域通过两次或更多次读数或者重叠读数来覆盖(如大于1“倍覆盖率”,例如2倍覆盖率)。

[0110] 在一些实施方式中,对来自一个个体的一种核酸样品测序。在某些实施方式中,对两个或更多样品的每个的核酸进行测序,其中样品来自一个个体或来自不同个体。在某些实施方式中,收集来自两个或更多个生物样品的核酸样品(其中各生物样品来自一个个体或两个或更多个个体),并对该集合测序。在后面的实施方式中,常通过一个或多个独特鉴定物来鉴定来自各生物样品的核酸样品。

[0111] 在一些实施方式中,测序方法采用鉴定物,其允许测序过程中的多重序列反应。独特鉴定物的数量越多,检测的样品和/或染色体的数量就越多,例如,可在测序过程多重进行。测序过程可使用任何合适的独特鉴定物数量(例如4、8、12、24、48、96或更多)来进行。

[0112] 测序过程有时使用固相,有时固相包括流动池,其上可接合来自库的核酸且试剂可流动并与所接合的核酸接触。流动池有时包括流动池通道,并且使用鉴定物可便于分析各通道中的样品数量。流动池通常是能被构建以保留和/或可供试剂溶液有序通过结合分析物的固体支持物。流动池通常是平面形状,光学透明,通常在毫米或亚毫米级,并且常具备通道或通路,在所述通道或通路中发生分析物/试剂的相互作用。在一些实施方式中,能在给定流动池通道中分析的样品数目常取决于库制备和/或探针设计中使用的独特鉴定物的数量。单流动池通道。多重使用12种鉴定物,例如可允许在8个通道流动池中同时分析96个样品(如等于96孔微孔板中的孔数)。类似地,多重使用48种鉴定物,例如可允许在8个通道流动池中同时分析384个样品(如等于384孔微孔板中的孔数)。市售可得的多重测序试剂盒的非限制性示例包括亿明达的多重样品制备寡核苷酸试剂盒和多重测序引物和PhiX对照试剂盒(分别例如亿明达的目录号PE-400-1001和PE-400-1002)。

[0113] 可使用测序核酸的任何合适的方法,其非限制性示例包括Maxim和Gilbert,链终止方法、合成测序、连接测序、质谱测序、基于显微镜的技术等或其组合。在一些实施方式中,第一代测序技术例如桑格测序方法包括自动桑格测序方法(包括微流体桑格测序)可用于本发明方法。在一些实施方式中,本文也使用核酸成像技术在内的其它测序技术(如透射电子显微镜(TEM)和原子力显微镜(AFM))。在一些实施方式中,使用高通量测序方法。高通量测序方法通常涉及有时在流动池中以大规模平行方式测序的克隆扩增DNA模板或单个DNA分子。下一代(例如第二和第三代)测序技术(能以大量平行方式测序DNA)可用于本文所述方法并且本文统称为“大量平行测序”(MPS)。在一些实施方式中,MPS测序方法采用靶向方法,其中特异染色体、基因或感兴趣的区域是序列。在某些实施方式中,使用非-靶向方法,其中样品中大多数或所有核酸经测序,扩增和/或随机捕获。

[0114] 在一些实施方式中,使用靶向富集、扩增和/或测序法。靶向方法通常通过序列特异性寡核苷酸分离、选择和/或富集样品中的核酸子集用于进一步加工。在一些实施方式中,采用序列特异性寡核苷酸的库以靶向(例如杂交)样品中的一个或多个核酸组。序列特异性寡核苷酸和/或引物通常就一种或多种感兴趣的染色体、基因、外显子、内含子和/或调控区域中存在的特定序列(例如独特核酸序列)具备选择性。任何合适的方法或方法组合可用于富集、扩增和/或测序一种或多种靶核酸子集。在一些实施方式中,使用一种或多种序列特异性锚通过捕获至固相(例如流动池、珠)分离和/或富集靶序列。在一些实施方式中,

通过基于聚合酶方法(例如基于PCR-方法,通过任何合适的基于聚合酶的延伸)使用序列特异性引物和/或引物组富集和/或扩增靶序列。序列特异性锚通常可用作序列特异性引物。

[0115] MPS测序有时使用通过合成和某些成像方法的测序。本文所述方法中可以使用的核酸测序技术是合成法测序和基于可逆终止子的测序(如亿明达公司(Illumina)的基因组分析仪(基因组分析仪)和基因组分析仪II(基因组分析仪II);HISEQ 2000;HISEQ 2500(亿明达,加利福尼亚州圣迭戈))。采用这种技术能对数百万核酸(如DNA)片段平行测序。在这种测序技术的一个实施例中,使用包含具备8个单独通道的光学透明载玻片的流动池,所述流动池表面结合寡核苷酸锚(如衔接子引物)。流动池通常是能被构建以保留和/或可供试剂溶液有序通过结合分析物的固体支持物。流动池通常是平面形状,光学透明,通常在毫米或亚毫米级,并且常具备通道或通路,在所述通道或通路中发生分析物/试剂的相互作用。

[0116] 在一些实施方式中,合成测序包括以模板引导的方式重复添加(例如通过共价添加)核苷酸至引物或预先存在的核酸链。检测各重复添加的核苷酸并重复所述过程多次直到获得核酸链的序列。所获序列长度部分取决于进行的添加和检测步骤的数量。在一些合成测序的实施方式中,加入一种、两种、三种或更多相同类型的核苷酸(例如A,G,C或T)并在添加核苷酸轮中进行检测。核苷酸可通过任何合适的方法添加(例如酶或化学)。例如,在一些实施方式中,聚合酶或连接酶以模板引导的方式添加核苷酸至引物或预先存在的核酸链。在合成测序的一些实施方式中,使用不同类型核苷酸、核苷酸类似物和/或鉴定物。在一些实施方式中,使用可逆终止子和/或可移除(例如可剪切)鉴定物。在一些实施方式中,使用荧光标记的核苷酸和/或核苷酸类似物。在某些实施方式中,合成测序包括切割(例如切割和移除鉴定物)和/或清洗步骤。在一些实施方式中,通过本文所述或本领域已知合适的方法检测添加一种或多种核苷酸,其非限制性示例包括任何合适的成像设备、合适的照相机、数码照相机、基于CCD(电荷耦合器件)的成像设备(例如CCD照相机)、基于CMOS(互补金属氧化物半导体)的成像设备(例如CMOS照相机)、光二极管(例如光电倍增管)、电子显微镜、场效应晶体管(例如DNA场效应晶体管)、ISFET离子感应器(例如CHEMFET感应器)等或其组合。其他可以用于进行本文所述方法的测序方法包括数字PCR和杂交法测序。

[0117] 其他可以用于进行本文所述方法的测序方法包括数字PCR和杂交法测序。数字聚合酶链式反应(数字PCR或dPCR)能用于直接鉴定和定量样品中的核酸。在一些实施方式中,可在乳液中进行数字PCR。例如,个体核酸在例如微流体设备中分离且各核酸通过PCR单独扩增。分离核酸使得每个孔中不超过一个核酸。在一些实施方式中,能使用不同探针区分多种等位基因(例如胎儿等位基因和母本等位基因)。等位基因可经计数以确定拷贝数。

[0118] 在一些实施方式中,可使用杂交测序。所述方法涉及使多种多核苷酸序列接触多种多核苷酸探针,其中所述多种多核苷酸探针各自可选地连接到底物。在一些实施方式中,所述底物可以是带有已知核苷酸序列阵列的平面。可使用与阵列杂交的模式确定样品中存在的多核苷酸序列。在一些实施方式中,各探针连接到珠(如磁珠等)上。与所述珠的杂交能被鉴定并用于鉴定样品中的多种多核苷酸序列。

[0119] 在一些实施方式中,纳米孔测序能用在本文所述的方法中。纳米孔测序是单分子测序技术,由此当单核酸分子(如DNA)通过纳米孔时被直接测序。

[0120] 进行本文所述方法非人合适的MPS方法、系统或技术平台可用于获取核酸序列读数。MPS平台的非限制性示例包括Illumina/Solex/HiSeq(例如Illumina的基因组分析仪;

基因组分析仪II;HISEQ 2000;HISEQ)、SOLiD、Roche/454、PACBIO和/或SMRT、Helicos True 单个分子测序、基于粒子流(Ion Torrent)和离子半导体(Ion semiconductor)的测序(例如生命技术公司所开发)、基于WildFire、5500、5500xl w和/或5500xl w遗传分析仪的技术(例如Life Technologies公司所开发和销售的、美国专利申请US20130012399);Polony测序、Pyro测序、大量平行签名测序(MPSS)、RNA聚合酶(RNAP)测序、LaserGen系统和方法、基于纳米孔的平台、化学敏感的场效应晶体管(CHEMFET)阵列、基于电子显微镜的测序(例如ZS Genetics公司、Halcyon Molecular公司所开发)、纳米球测序等或其组合。

[0121] 在一些实施方式中,进行染色体特异性测序。在一些实施方式中,染色体-特异性测序采用DANSR(所选区域的数码分析)进行。所选区域的数码分析能同时定量数百个位点,通过两个位置特异性寡核苷酸的cfDNA-依赖性连锁,利用干扰‘桥式’寡核苷酸以形成PCR模板。在一些实施方式中,染色体-特异性测序通过生成富集染色体-特异性序列的库来进行。在一些实施方式中,仅获得所选染色体组的序列读数。在一些实施方式中,仅获得染色体21、18和13的序列读数。在一些实施方式中,获得整个基因组或基因组的区段的序列读数和/或将序列读数映射到整个基因组或基因组的区段。

[0122] 在一些实施方式中,通过序列模块生成、获取、聚集、组装、处理、变换、加工、变换和/或提供序列读数。包括序列模块的机器可为用本领域已知测序技术确定核酸序列的合适机器和/或设备。在一些实施方式中,序列模块可比对、组装、片段化、互补、反向互补、和/或检错(例如,纠错序列读数)。

[0123] 在一些实施方式中获自样品的核苷酸序列读数是部分核苷酸序列读数。本文所用的“部分核苷酸序列读数”是指具有不完整序列信息(也称为序列模糊性)的任何长度的序列读数。部分核苷酸序列读数可能缺少关于核碱基相同性和/或核碱基位置或顺序的信息。部分核苷酸序列读数一般不包括其中仅有的不完整序列信息(或其中不到全部的碱基经测序或确定)来自偶然或无意的测序错误的序列读数。这类测序错误可能对于某些测序过程而言是固有的并且包括,例如,核碱基相同性的不正确读出,和缺失或额外的核碱基。因此,对于本文的部分核苷酸序列读数,通常刻意排除关于序列的某些信息。即,刻意获得不到全部的核碱基的序列信息,或可能另外表征为或成为测序错误。在一些实施方式中,部分核苷酸序列读数可跨核酸片段的一部分。在一些实施方式中,部分核苷酸序列读数可跨核酸片段的整个长度。部分核苷酸序列读数描述于,例如,国际专利申请公开号W02013/052907,其全文通过引用纳入本文,包括所有文字、表格、等式和附图。

[0124] 映射读数

[0125] 可映射序列读数。可使用任何合适的映射方法(例如,过程、算法、程序、软件、模块等或其组合)并且下文中描述了映射过程的某些方面。

[0126] 映射核苷酸序列读数(例如,物理基因组位点未知的片段的序列信息)能以多种方式进行,其通常包括使获得的测序读数与参照基因组中的匹配序列进行比对。所述比对中,序列读数通常与参照序列进行比对,经比对的那些称为经“映射”、“映射的序列读数”或“映射的读数”。

[0127] 本文所用术语“比对”或“对齐”指可鉴定为匹配(例如100%相同性)或部分匹配的两个或更多核酸序列。比对可人工或通过计算机(例如软件、程序、模块或算法)完成,其非限制性示例包括核苷酸数据有效局部比对(ELAND)计算机程序,其为Illumina基因组分析

流程的部分。序列读数的比对可为100%序列匹配。在一些情况中,比对小于100%序列匹配(例如,非完美匹配、部分匹配、部分比对)。在一些实施方式中,比对为约99%、98%、97%、96%、95%、94%、93%、92%、91%、90%、89%、88%、87%、86%、85%、84%、83%、82%、81%、80%、79%、78%、77%、76%或75%匹配。在一些实施方式中,比对包括错配。在一些实施方式中,比对包括1,2,3,4或5个错配。两个或更多序列可使用任一链比对。在某些实施方式中,核酸序列与另一核酸序列的反向互补链比对。

[0128] 使用多种计算机方法将序列读数映射和/或比对至参照基因组。可用于比对序列的计算机算法的非限制性示例包括但不限于BLAST,BLITZ,FASTA,BOWTIE 1,BOWTIE 2,ELAND,MAQ,探针MATCH,SOAP或SEQMAP或其变体或其组合。在一些实施方式中,序列读数可与参照基因组和/或参照基因组中的序列比对。在一些实施方式中,序列读数可从本领域已知核酸数据库获得和/或与其中的序列比对,所述数据库包括例如GenBank,dbEST,dbSTS,EMBL(欧洲分子生物实验室)和DDBJ(日本DNA数据库)。BLAST或相似工具可用于针对序列数据库搜索相同序列。

[0129] 在一些实施方式中,与映射的序列读数关联的映射的序列读数和/或信息在非瞬时计算机可读介质上以合适的计算机可读格式进行储存和/或从其进行评估。本文中“计算机可读格式”有时指一种格式。在一些实施方式中,映射的序列读数以合适的二进制格式、文本格式等或其组合进行储存和/或评估。二进制格式有时是BAM格式。文本格式有时是序列比对/映射(SAM)格式。二进制或文本格式的非限制性示例包括BAM、SAM、SRF、FASTQ、Gzip等或其组合。在一些实施方式中,映射的序列读数在一种格式中储存和/或转化为一种格式,所述一种格式比传统格式(例如SAM格式或BAM格式)需要更少的储存空间(例如更少字节)。在一些实施方式中,第一种格式中的映射的序列读数压缩为第二种格式,后者比第一种需要更少的储存空间。本文所用术语“压缩”指数据压缩、源编码和/或比特率降低的过程,其中计算机可读数据文件大小降低。在一些实施方式中,映射的序列读数从SAM格式压缩成二进制格式。文件压缩后有时损失一些数据。有时压缩过程不损失数据。在一些文件压缩实施方式中,一些数据用另一数据文件的索引和/或参照替代,所述另一数据文件包含涉及映射的序列读数的信息。在一些实施方式中,映射的序列读数以二进制格式储存,包括或由下述组成:读数计数、染色体鉴定物(例如鉴定读数所映射的染色体)和染色体位置鉴定物(例如鉴定读数所映射的染色体上的部份)。在一些实施方式中,二进制格式包括20字节阵列、16字节阵列、8字节阵列、4字节阵列或2字节阵列。在一些实施方式中,映射的读数信息以10字节格式、9字节格式、8字节格式、7字节格式、6字节格式、5字节格式、4字节格式、3字节格式、2字节格式储存在阵列中。有时映射的数据读数储存在4字节阵列中,包括5字节格式。在一些实施方式中,二进制格式包括5-字节格式,包括1-字节染色体序数(ordinal)和4-字节染色体部份。在一些实施方式中,映射的读数以压缩二进制格式储存,所述格式比序列比对/映射(SAM)格式小约100倍、约90倍、约80倍、约70倍、约60倍、约55倍、约50倍、约45倍、约40倍或约30倍。一些实施方式中,映射的读数以压缩二进制格式储存,所述格式比GZip格式小约2倍-约50倍(例如约30、25、20、19、18、17、16、15、14、13、12、11、10、9、8、7、6、或约5倍)。

[0130] 在一些实施方式中,系统包括压缩模块(例如4,图10A)。在一些实施方式中,以计算机可读格式存储在非瞬时计算机可读介质上的映射的序列读数信息通过压缩模块进行

压缩。压缩模块有时将映射的序列读数变换为合适的格式或从合适和格式变换。在一些实施方式中,压缩模块可接受第一格式的映射的序列读数(例如1),将其变换为压缩格式(例如二进制格式,5)和将压缩读数转移至另一模块(例如偏移密度模块6)。压缩模块通常提供二进制格式5(例如BReads格式)的序列读数。压缩模块的非限制性示例包括GZIP、BGZF和BAM等或其变体)。

[0131] 下述提供用java将整数变换为4字节数组的示例:

```
[0132] public static final byte[]  
[0133] convertToByteArray(int value)  
[0134] {  
[0135]     return new byte[] {  
[0136]         (byte) (value>>>24),  
[0137]         (byte) (value>>>16),  
[0138]         (byte) (value>>>8),  
[0139]         (byte) value};  
[0140] }
```

[0141] 在一些实施方式中,读数可唯一或非唯一映射至参照基因组。若读数与参照基因组中的单一序列比对,则其称为“唯一映射”。若读数与参照基因组中的两个或多个序列比对,则其称为“非唯一映射”。在一些实施方式中,非唯一映射的读数从进一步分析(例如定量)中去除。在某些实施方式中,某些小程度的错配(0-1)可说明在所述参照基因组和经映射的来自个体样品的读数之间可能存在单核酸多态性。在一些实施方式中,没有错配可使读数映射到参照序列上。

[0142] 本文所用术语“参照基因组”可指任何生物体或病毒的任何部分或全部都具体已知的、测序的或表征的基因组,其可用于参照鉴定对象序列。参照基因组有时指参照基因组的区段(例如,染色体或其部分,例如,参照基因组的一个或多个部份)。来自任何其它生物体的人基因组、人基因组组件和/或基因组可用作参照基因组。可在国家生物技术信息中心在www.ncbi.nlm.nih.gov上找到一种或多种人基因组、人基因组组件以及其它生物体的基因组。“基因组”指核酸序列中表达的生物体或病毒的完全遗传信息。本文所用参照序列或参照基因组经常是来自个体或多个个体的组装或部分组装的基因组序列。在一些实施方式中,参照基因组是来自一个或多个个人个体的组装或部分组装的基因组序列。在一些实施方式中,参照基因组包括分配至染色体的序列。本文所用术语“参照序列”是指一个或多个参照样品的一个或多个多核苷酸序列。在一些实施方式中,参照序列包括获自参照样品的序列读数。在一些实施方式中,参照序列包括获自一个或多个参照样品的序列读数、读数的组件、共有DNA序列(例如,序列毗连群)、读数密度和/或读数密度概况。获自参照样品的读数密度概况在本文中有时称为参照概况。获自测试样品和/或测试对象的读数密度概况在本文中有时称为测试概况。在一些实施方式中,参照样品获自基本没有遗传变异(例如,所研究的遗传变异)的参照对象。在一些实施方式中,参照样品获自包含已知遗传变异的参照对象。本文所用术语“参照”可以指参照基因组、参照序列、参照样品和/或参照对象。

[0143] 在某些实施方式中,当样品核酸来自妊娠雌性时,有时参照序列不来自胎儿、胎儿母亲或胎儿父亲,从而在本文中被称为“外部参照”。在一些实施方式中可制备和使用母本

参照。基于外部参照制备来自妊娠雌性的参照时(“母本参照序列”),基本不含胎儿DNA的来自妊娠雌性的DNA的读数通常映射至外部参照序列并组装。在某些实施方式中,所述外部参照来自与所述妊娠雌性基本属同一种族的个体的DNA。母本参照序列可以不完全覆盖母本基因组DNA(如可以覆盖母本基因组DNA的约50%、60%、70%、80%、90%或更多),并且所述母本参照可以不与母本基因组DNA序列完美匹配(如所述母本参照序列可以包含多个错配)。

[0144] 在某些实施方式中,可映射能力针对基因组区域(例如部份、基因组部份)进行评估。可映射能力是核苷酸序列读数清楚地比对至参照基因组的部份的能力,通常具有多至特定数量的错配,包括例如,0、1、2或更多错配。在一些实施方式中,以分数或值提供可映射能力,其中通过合适的映射算法后计算机映射软件生成该分数或值。对于给定基因组区域而言,预期的可映射能力可用预定的读数长度的滑动窗口法计算并平均为所得的读数水平可映射能力值。包括独特核苷酸序列的伸展的基因组区域有时具有高可映射能力值。

[0145] 序列读数可通过映射模块或包括映射模块的装置经映射,该映射模块通常映射读数至参照基因组或其区段。映射模块可通过本领域已知合适的方法映射序列读数。在一些实施方式中,需要映射模块或包括映射模块的装置以提供映射的序列读数。

[0146] 计数

[0147] 可对映射的序列读数进行定量以确定映射到参照基因组的区域或部份的读数的数量。在某些实施方式中,映射至参照基因组、或其区域、部份或区段的读数被称为计数。在一些实施方式中,计数包括值。在一些实施方式中,计数值通过数学过程确定。计数可通过合适的方法、操作或数学过程确定。在某些实施方式中,计数经加权、去除、过滤、标准化、调整、平均、加、或减、或通过其组合处理。在某些实施方式中,计数源自通过本领域已知合适的方法、操作或数学过程处理或加工的序列读数。例如,计数通常按照与序列读数相关的一个或多个偏好来标准化和/或加权。在一些实施方式中,计数按照与序列读数相关的GC偏好来标准化和/或加权。在一些实施方式中,计数源自原始序列读数和/或筛选的序列读数。在一些实施方式中,一个或多个计数未经数学处理。本文所用术语“原始计数”是指未经数学处理的一个或多个计数。

[0148] 在一些实施方式中,针对映射至参照基因组或其区域、部份或区段的一些或全部序列读数确定计数。在某些实施方式中,从映射的序列读数的预定义子集确定计数。能利用任意合适的特性或变量来定义或选择映射的序列读数的预定义子集(例如,选择的子集)。在一些实施方式中,映射的测序读数的预定义子集能包含1-n个序列读数,其中n表示与从测试对象或参照对象样品生成的所有序列读数总和相等的数目。

[0149] 计数通常来自获自对象(例如,测试对象)的序列读数。计数有时来自获自怀有胎儿的妊娠雌性的核酸样品的序列读数。核酸序列读数的计数通常是代表胎儿和胎儿母亲(例如,针对妊娠雌性对象)的计数。在某些实施方式中,在对象是妊娠雌性的情况中,一些计数来自胎儿基因组并且一些计数来自母本基因组。

[0150] 读数密度

[0151] 序列读数的计数(例如,加权的计数)通常表示读数密度。通常针对基因组的一个或多个部份确定和/或生成读数密度。在某些实施方式中,针对一个或多个染色体确定和/或生成读数密度。在一些实施方式中,读数密度包括对映射到参照基因组的部份的序列读

数的计数的定量测量。可通过合适的过程来确定读数密度。在一些实施方式中,通过合适的分布和/或合适的分布函数来确定读数密度。分布函数的非限制性示例包括概率函数、概率分布函数、概率密度函数(PDF)、核心(kernel)密度函数(核心密度评价)、累积分布函数、概率质量函数、离散概率分布、绝对连续单变量分布等、任意合适分布、或其组合。在某些实施方式中,PDF包括核心密度函数(核心密度评价)。可用于生成局部基因组偏好评价的核心密度函数的非限制性示例包括均匀核心密度函数(均匀核心)、高斯核心密度函数(高斯核心)、三角核心密度函数(三角核心)、双权核心密度函数(双权核心)、三立方体核心密度函数(三立方体核心)、三权核心密度函数(三权核心)、余弦核心函数(余弦核心)、Epanechnikov核心密度函数(Epanechnikov核心)、正态核心密度函数(正态核心)等或其组合。读数密度通常是来自合适的概率密度函数的密度评价。密度评价是基于观察到的数据的对背后概率密度函数的评价的构建。在一些实施方式中,读数密度包括密度评价(例如,概率密度评价、核心密度评价)。密度评价通常包括核心密度评价。在一些实施方式中,读数密度是按照核心密度函数确定的核心密度评价。通常按照一个过程生成读数密度,包括生成基因组的一个或多个部份各自的密度评价,其中各部份包括序列读数的计数。通常针对映射到部份的标准化和/或加权的计数生成读数密度。在一些实施方式中,映射到部份的各读数通常产生读数密度,其值(例如,计数)等于其获自本文所述的标准化过程的权重。在一些实施方式中,针对一个或多个部份调节读数密度。可通过合适的方法调整读数密度。例如,针对一个或多个部份的读数密度可经加权和/或标准化。

[0152] 在一些实施方式中,系统包括分布模块12。分布模块通常生成和/或提供针对基因组的部份(例如,过滤的部份)的读数密度(例如,22,24)。分布模块可提供针对一个或多个参照样品、训练组(例如,3)和/或测试样品的读数密度、读数密度分布14和/或相关不确定性测量(例如,MAD,分位数)。分布模块可接收、收回、和/或存储序列读数(例如,1、3、5)和/或计数(例如,标准化的计数11,加权的计数)。分布模块通常接收(例如,部份的用户输入和用户参数)、收回、生成和/或存储部份(例如,未过滤或过滤的部份)。有时,分布模块接收和/或收回来自过滤模块18的部份(例如,过滤的部份和/或选择的部份20)。在一些实施方式中,分布模块包括代码和/或源代码形式(例如,标准或自定义脚本集合)的微处理器的指令(例如,算法,脚本)和/或进行分布模块函数的一个或多个软件包(例如,统计学软件包)。在一些实施方式中,分布模块包括以java、采用合适的包(例如,S包或R包)的S或R编写的代码(例如,脚本)。分布模块的非-限制性示例如实施例2所示。

[0153] 在一些实施方式中,确定读数密度概况。在一些实施方式中,读数密度概况包括至少一个读数密度,并且通常包括2个或更多个读数密度(例如,读数密度概况通常包括多个读数密度)。在一些实施方式中,读数密度概况包括合适的定量值(例如,平均、中值、Z-分数等)。读数密度概况通常包括来自一个或多个读数密度的值。读数密度概况有时包括获自基于一个或多个调节(例如,标准化)对读数密度的一个或多个处理的值。在一些实施方式中,读数密度概况包括未处理的读数密度。在一些实施方式中,从包括读数密度的各种数据组或其衍生(例如本领域已知和/或本文所述的一种或多种数学和/或统计学数据处理步骤的结果)中生成一种或多种读数密度概况。在某些实施方式中,读数密度概况包括标准化的读数密度。在一些实施方式中,读数密度概况包括经调整的读数密度。在某些实施方式中,读数密度概况包括原读数密度(例如,未处理、未调整或标准化)、标准化的读数密度、加权的

读数密度、过滤的部份的读数密度, 读数密度的z-分数、读数密度的p-值、读数密度的积分值(例如, 曲线下的面积)、算术平均、平均或中值读数密度、主成分等、或其组合。通常, 读数密度概况的读数密度和/或读数密度概况与不确定性的测量相关(例如, MAD)。在某些实施方式中, 读数密度概况包括中值读数密度的分布。在一些实施方式中, 读数密度概况包括多个读数密度的关系(例如, 拟合的关系、回归等)。例如, 有时, 读数密度概况包括读数密度(例如, 读数密度值)与遗传位置(例如, 部份、部份位置)之间的关系。在一些实施方式中, 使用静态窗过程生成读数密度概况, 和在某些实施方式中, 使用滑动窗过程生成读数密度概况。本文术语“读数密度概况”指读数密度的数学和/或统计学操作的结果, 其可便于在大量序列读数数据中鉴定模式和/或相关性。

[0154] 在一些实施方式中, 读数密度概况有时经印刷和/或显示(例如, 显示为可视表示, 例如, 曲线或图)。

[0155] 读数密度概况通常包括多个数据点, 其中各数据点代表一个或多个读数密度的定量值。任何合适数量的数据点可包括在读数密度概况中, 取决于数据组的性质和/或复杂度。在某些实施方式中, 读数密度概况可包括2或更多数据点、3或更多数据点、5或更多数据点、10或更多数据点、24或更多数据点、25或更多数据点、50或更多数据点、100或更多数据点、500或更多数据点、1000或更多数据点、5000或更多数据点、10,000或更多数据点、100,000或更多数据点、或1,000,000或更多数据点。在一些实施方式中, 数据点是映射到或关联一个或多个部份的序列读数的计数的评价和/或定量值。在一些实施方式中, 读数密度概况中的数据点包括对映射至一个或多个部份的计数的数据处理的结果。在某些实施方式中, 数据点通常是一个或多个读数密度(例如, 平均读数密度)的评价和/或定量值。读数密度概况通常包括关联和/或映射到参照基因组的多个部份的多个读数密度。在一些实施方式中, 读数密度概况包括2至约1,000,000个部份的读数密度。在一些实施方式中, 2-约50,000、2-约40,000、2-约30,000、2-约20,000、2-约10,000、2-约5000、2-约2500、2-约1250、2-约1000、2-约500、2-约250、2-约100或2-约60个部份的读数密度确定读数密度概况。在一些实施方式中, 约10-约50个部份的读数密度确定读数密度概况。

[0156] 在一些实施方式中, 读数密度概况对应于部份的组(例如参照基因组的部份的组、染色体部份的组或染色体区段部份的子集)。在一些实施方式中, 读数密度概况包括与部份的集合(例如, 组, 子集)关联的读数密度和/或计数。在一些实施方式中, 确定连续的部份的读数密度的读数密度概况。在一些实施方式中, 连续部份包括间隙, 其包括密度概况中所不包括的序列读数和/或参照序列的区段(例如, 通过过滤去除的部份)。有时, 邻近部份(例如部份的组)代表基因组的相邻区段或染色体或基因的相邻区段。例如, 通过尾对尾合并部份时, 其的两个或更多连续部份可代表比各部份更长的DNA序列的序列集合。例如, 两个或更多连续部份可代表整个基因组、染色体、基因、内含子、外显子或其区段。有时, 从连续部份和/或非连续部份的集合(例如组, 子集)中确定读数密度概况。在一些情况中, 读数密度概况包括一种或多种部份、所述部份可经加权、移除、过滤、标准化、调整、平均(得出均值)、加、减、或其任意组合的加工或变换。

[0157] 在一些实施方式中, 读数密度概况包括包含遗传变异的基因组的部份的读数密度。在一些实施方式中, 读数密度概况包括不包含遗传变异的基因组的部份(例如, 基本没有遗传变异的基因组的部份)的读数密度。在某些实施方式中, 读数密度概况包括包含遗传

变异的基因组的部份的读数密度和基本上没有遗传变异的基因组的部份的读数密度。

[0158] 通常确定样品和/或参照(例如,参照样品)的读数密度概况。有时生成整个基因组、一个或多个染色体、或者染色体或基因组的区段或一部分的读数密度概况。在一些实施方式中,确定基因组或其区段的一个或多个读数密度概况。在一些实施方式中,读数密度概况表示全部的样品读数密度组,并且在某些实施方式中,读数密度概况表示样品的读数密度的子集或一部分。即,读数密度概况有时包括未过滤移除任何数据的数据的读数密度代表或从其中生成,和有时读数密度概况包括经过过滤移除不想要数据的数据的数据点代表或从其中生成。

[0159] 在一些实施方式中,确定参照(例如,参照样品,训练组)的读数密度概况。参照的读数密度概况在本文中有时称为参照概况。在一些实施方式中,参照概况包括获自一种或多种参照(例如,参照序列、参照样品)的读数密度。在一些实施方式中,参照概况包括一个或多个(例如,一组)已知整倍性样品确定的读数密度。在一些实施方式中,参照概况包括过滤的部份的读数密度。在一些实施方式中,参照概况包括按照一个或多个主成分调整的读数密度。

[0160] 在一些实施方式中,系统包括概况生成模块(例如,26)。概况生成模块通常接收、收回和/或存储读数密度(例如,22、24)。概况生成模块可接收和/或收回来自另一个合适模块(例如,分布模块)的读数密度(例如,调整、加权、标准化、平均、算术平均、中值和/或积分读数密度)。概况生成模块和接收和/或收回来自合适来源(例如,一个或多个参照对象、训练组、一个或多个测试对象等)的读数密度。概况生成模块通常生成和/或提供读数密度概况(例如,32、30、28)至另一个合适的模块(例如,PCA统计模块33,部份加权模块42,打分模块46)和/或至用户(例如,通过描绘、作图和/或印刷)。概况生成模块或其部份的示例提供于实施例2。

[0161] 部份

[0162] 在一些实施方式中,映射的序列读数和/或计数按照多种参数分类在一起并且分配到在本文中称为“部份”的参照基因组的特定区段和/或区域。在一些实施方式中,部份是整个染色体、染色体区段、参照基因组区段、跨多重染色体的区段、多重染色体区段和/或其组合。在一些实施方式中,基于具体参数(例如,预定长度、预定间隔、预定GC含量或任意其它合适的参数)来预定部份。在一些实施方式中,部份基于基因组的划分任意定义(例如根据大小、GC含量、连续区域、任意定义大小的连续区域等的分区)。在一些实施方式中,部份基于一种或多种参数定义,所述参数包括例如序列的长度或具体特征。在一些实施方式中,部份基于基因组序列的具体长度。部份可具有大致相同的长度或部份可具有不同长度。在一些实施方式中,部份长度大约相同。在一些实施方式中,不同长度的部份经调整或加权。部份可为任何合适的长度。在一些实施方式中,部份为约10千碱基(kb)-约100kb、约20kb-约80kb、约30kb-约70kb、约40kb-约60kb,而有时约50kb。在一些实施方式中,部份为约10kb-约20kb。部份不限于连续运行的序列。因此,部份可由连续和/或非-连续序列组成。

[0163] 在一些实施方式中,部份包括包含预选数量的碱基的窗。窗可包括由部份长度确定的任意合适数量的碱基。在一些实施方式中基因组或其区段划分为多个窗。涵盖基因组区域的窗可重叠或不重叠。在一些实施方式中,窗位于彼此距离相等的位置。在一些实施方式中,窗位于彼此距离不等的位置。在某些实施方式中,基因组或其区段划分为多个滑动

窗,其中窗递增地滑过基因组或其区段,其中各增量的各窗代表部份。窗可根据任何数字形式或根据任何数学 (athematic) 定义的序列以任何合适的增量滑过基因组。在一些实施方式中,窗以约100,000bp或更少、约50,000bp或更少、约25,000bp或更少、约10,000bp或更少、约5,000bp或更少、约1,000bp或更少、约500bp或更少、或约100bp或更少的增量滑过基因组或其区段。例如,窗可包括约100,000bp并且可以50,000bp的增量滑过基因组。

[0164] 在一些实施方式中,部份可为感兴趣染色体中的特定染色体区段,感兴趣染色体例如评估遗传变异(例如,染色体13、18和/或21或性染色体的非整倍性)的染色体。部份不限于单一染色体。在一些实施方式中,一个或多个部份包括一条染色体的所有或部份或者两个或更多染色体的所有或部份。在一些实施方式中,一个或多个部份可跨越1个、2个或更多完整染色体。此外,部份可跨越多重染色体的连接或不连接区域。部份可为基因、基因片段、调控序列、内含子、外显子等。

[0165] 在一些实施方式中,基因组的某些区域在将基因组或其区段划分成部份之前经过滤。可使用任意合适的方法选择基因组的区域来从划分过程中排除。通常,去除和/或过滤包含类似区的区(例如,相同或同源区域或序列,例如,重复区)。有时,排除不可映射的区。在一些实施方式中,仅保留独特的区域。划分期间去除的区域可在单一染色体内或可跨越多重染色体。在一些实施方式中,划分的基因组经向下修整并优化以用于快速比对,通常允许关注独特可鉴定序列。在一些实施方式中,可基于分类的上下文中产生的信息将基因组划分成区域(例如,超出染色体的区域)。例如,信息内容可用p值概况定量,测量已证实的正常和异常对象(分别例如整倍体和三倍体对象)的具体基因组位置的显著性。在一些实施方式中,可基于任何其他标准将基因组划分成区域(例如,超出染色体的区域),所述标准例如,比对标签时的速度/便利度、GC含量(例如高或低的GC含量)、GC含量的均匀度、序列含量的其他测量(例如个体核苷酸分数、嘧啶或嘌呤分数、天然与非天然核酸的分数、甲基化核苷酸的分数和CpG含量)、甲基化状态、双重解链温度、测序或PCR的依从度、分配给参照基因组的个体部份的不确定性测量和/或具体特征的靶向搜索。

[0166] 基因组的“区段”有时是包含一个或多个染色体,或染色体的一部分的区域。“区段”一般是与部份不同的染色体的一部分。基因组和/或染色体的“区段”有时与部份在染色体不同区域中,有时不与部份共有多核苷酸,和有时包括部份中的多核苷酸。基因组或染色体区段通常比部份包括更多核苷酸(例如,区段有时包括一个或多个部份),并且有时染色体区段比部份包括更少核苷酸(例如,区段有时在部份内)。

[0167] 过滤部份

[0168] 在某些实施方式中,通过过滤过程将一个或多个部份(例如基因组的部份)从考虑中去除。在某些实施方式中,一个或多个部份经过滤(例如经历过滤过程),从而提供过滤的部份。在一些实施方式中,过滤过程移除某些部份和保留部份(例如部份子集)。过滤过程后,保留的部份通常指本文的过滤的部份。在一些实施方式中,参照基因组的部份被过滤。在一些实施方式中,通过过滤过程去除的参照基因组的部份不包括在确定是否存在遗传变异中(例如染色体非整倍性)。在一些实施方式中,参照基因组中染色体的部份被过滤。在一些实施方式中,关联读数密度(例如读数密度用于部份)的部份通过过滤过程去除并且关联去除的部份的读数密度不包括在确定是否存在遗传变异中(例如染色体非整倍性)。在一些实施方式中,读数密度概况包括过滤的部份的读数密度和/或由其组成。可使用本领域已知

或本文所述的任何合适的标准和/或方法对部份进行选择、过滤和/或从考虑中去除。用于过滤部份的标准的非限制性示例包括冗余数据(例如冗余或重叠映射的读数)、无-信息数据(例如0映射计数的参照基因组的部份)、具有过高频出现或过低频出现序列的参照基因组的部份、GC含量、噪音数据、可映射能力、计数、计数可变性、读数密度、读数密度可变性、不确定性的测量、可重复性测量等或前述的组合。部份有时根据计数分布和/或读数密度分布来进行过滤。在一些实施方式中,部份根据计数分布和/或读数密度分布来进行过滤,其中计数和/或读数密度获自一种或多种参照样品。有时一种或多种参照样品本文指训练组。在一些实施方式中,部份根据计数分布和/或读数密度分布来进行过滤,其中计数和/或读数密度获自一种或多种测试样品。在一些实施方式中,部份根据读数密度分布的不确定性测量进行过滤。在某些实施方式中,显示读数密度中大偏差的部份通过过滤过程去除。例如,可确定读数密度的分布(例如平均、算术平均或中值读数密度的分布,例如图5A),其中分布中的各读数密度映射至相同的部份。不确定性测量(例如MAD)可通过比较多种样品的读数密度的分布来确定,其中基因组各部份与不确定性测量相关联。根据前述示例,部份可根据关联各部份和预定阈值的不确定性测量(例如标准偏差(SD)、MAD)来过滤。图5B显示部份的MAD值的分布,根据多种样品的读数密度分布所确定。预定阈值如垂直虚线所示,所述线围成可接受的MAD值范围。图5B的示例中,通过过滤过程,包括在可接受范围内的MAD值的部份保留并且包括在可接受范围之外的MAD值的部份从考虑中去除。在一些实施方式中,根据前述示例,包括超出预定不确定性测量的读数密度值(例如中值,平均或算术平均读数密度)的部份通常通过过滤过程从考虑中去除。在一些实施方式中,包括超出分布的四分位间范围的读数密度值(例如中值,平均或算术平均读数密度)的部份通常通过过滤过程从考虑中去除。在一些实施方式中,包括超出分布的四分位间范围的2倍、3倍、4倍或5倍的读数密度值的部份通过过滤过程从考虑中去除。在一些实施方式中,包括超出 2σ 、 3σ 、 4σ 、 5σ 、 6σ 、 7σ 或 8σ (例如其中 σ 是标准偏差定义的范围)的读数密度值的部份通过过滤过程从考虑中去除。

[0169] 在一些实施方式中,系统包括过滤模块18。过滤模块通常接收、取回和/或储存部份(例如预定大小和/或重叠的部份,参照基因组内的部份位置)和关联部份的读数密度,通常来自其他合适的模块(例如分布模块12)。在一些实施方式中,所选部份(例如,20,例如过滤的部份)通过过滤模块提供。在一些实施方式中,需要过滤模块来提供过滤的部份和/或从考虑中去除部份。在某些实施方式中,过滤模块从考虑中去除读数密度,其中读数密度关联去除的部份。过滤模块通常提供选择的部份(例如过滤的部份)至其他合适的模块(例如分布模块21)。过滤模块的非-限制性示例如实施例3所示。

[0170] 偏好评价

[0171] 测序技术易受多种来源的偏好所影响。有时测序偏好是局部偏好(例如局部基因组偏好)。局部偏好通常在序列读数水平出现。局部基因组偏好可为任何合适的局部偏好。局部偏好的非限制性示例包括序列偏好(例如GC偏好、AT偏好、等)、关联DNA酶I敏感性的偏好、熵、重复序列偏好、染色质结构偏好、聚合酶误差率偏好、回文偏好、插入重复偏好、PCR相关偏好、等或其组合。在一些实施方式中,局部偏好的来源未确定或未知。

[0172] 在一些实施方式中确定局部基因组偏好评价。局部基因组偏好评价本文中有时指局部基因组偏好评价。局部基因组偏好评价可就参照基因组、其区段或部份确定。在某些实

施方式中,确定参照基因组中一个或多个染色体的局部基因组偏好评价。在一些实施方式中,确定一种或多种序列读数(例如样品的一些或全部序列读数)的局部基因组偏好评价。通常根据参照(例如,参照基因组,参照基因组中的染色体)的对应定位和/或位置的局部基因组偏好评价确定序列读数的局部基因组偏好评价。在一些实施方式中,局部基因组偏好评价包括定量测量序列偏好(例如序列读数、参照基因组序列)。局部基因组偏好评价可通过合适的方法或数学过程确定。在一些实施方式中,通过合适的方法和/或合适的分布函数(例如PDF)确定局部基因组偏好评价。一些实施方式中,局部基因组偏好评价包括PDF的定量表示。在一些实施方式中,局部基因组偏好评价(例如概率密度评价(PDE)、核心密度评价)通过局部偏好含量的概率密度函数(例如PDF,例如核心密度函数)来确定。一些实施方式中密度评价包括核心密度评价。局部基因组偏好评价有时表示为分布的平均、算术平均或中值。有时局部基因组偏好评价表示为加和或积分(例如合适的分布的曲线下面积(AUC))。

[0173] PDF(例如核心密度函数,例如Epanechnikov核心密度函数)通常包括带宽变量(例如带宽)。带宽变量通常定义窗的大小和/或长度,当使用PDF时从中得出概率密度评价(PDE)。得出PDE的窗通常包括多核苷酸的定义长度。在一些实施方式中得出PDE的窗是部份。通常根据带宽变量确定部份(例如部份大小,部份长度)。带宽变量确定用于确定局部基因组偏好评价的窗的长度或大小。从多核苷酸区段(例如核苷酸碱基的连续区段)的长度确定局部基因组偏好评价。PDE(例如读数密度,局部基因组偏好评价(例如GC密度))可使用任何合适的带宽确定,其非限制性示例包括约5碱基-约100,000碱基、约5碱基-约50,000碱基、约5碱基-约25,000碱基、约5碱基-约10,000碱基、约5碱基-约5,000碱基、约5碱基-约2,500碱基、约5碱基-约1000碱基、约5碱基-约500碱基、约5碱基-约250碱基、约20碱基-约250碱基或等的带宽。一些实施方式中,使用约400碱基或更少、约350碱基或更少、约300碱基或更少、约250碱基或更少、约225碱基或更少、约200碱基或更少、约175碱基或更少、约150碱基或更少、约125碱基或更少、约100碱基或更少、约75碱基或更少、约50碱基或更少或约25碱基或更少的带宽来确定局部基因组偏好评价(例如GC密度)。在某些实施方式中,使用带宽确定局部基因组偏好评价(例如GC密度),所述带宽根据给定对象和/或样品所获的序列读数的平均、算术平均、中值或最大读数长度来确定。有时使用带宽确定局部基因组偏好评价(例如GC密度),所述带宽约等于给定对象和/或样品所获的序列读数的平均、算术平均、中值或最大读数长度。在一些实施方式中,使用约250、240、230、220、210、200、190、180、160、150、140、130、120、110、100、90、80、70、60、50、40、30、20或约10碱基的带宽来确定局部基因组偏好评价(例如GC密度)。

[0174] 局部基因组偏好评价可在单碱基分辨率下确定,尽管局部基因组偏好评价(例如局部GC含量)可以更低的分辨率确定。在一些实施方式中就局部偏好含量确定局部基因组偏好评价。通常使用窗确定局部基因组偏好评价(例如使用PDF测定)。在一些实施方式中,局部基因组偏好评价包括使用包括预选数量碱基的窗。有时窗包括连续碱基区段。有时窗包括非-连续碱基的一个或多个部份。有时窗包括一个或多个部份(例如基因组的部份)。窗大小或长度通常通过带宽和根据PDF确定。在一些实施方式中,窗为带宽长度的约10或更多、8或更多、7或更多、6或更多、5或更多、4或更多、3或更多、或约2或更多倍。使用PDF(例如核心密度函数)确定密度评价时,窗有时两倍于所选带宽的长度。窗可包括任何合适数量的

碱基。在一些实施方式中,窗包括约5碱基-约100,000碱基、约5碱基-约50,000碱基、约5碱基-约25,000碱基、约5碱基-约10,000碱基、约5碱基-约5,000碱基、约5碱基-约2,500碱基、约5碱基-约1000碱基、约5碱基-约500碱基、约5碱基-约250碱基,或约20碱基-约250碱基。在一些实施方式中基因组或其区段划分为多个窗。涵盖基因组区域的窗可重叠或不重叠。在一些实施方式中,窗位于彼此距离相等的位置。在一些实施方式中,窗位于彼此距离不等的位置。在某些实施方式中,基因组或其区段划分为多个滑动窗,其中窗递增地滑过基因组或其区段,其中各增量的各窗包括局部基因组偏好评价(例如局部GC密度)。窗可根据任何数字形式或根据任何数学(athematic)定义的序列以任何合适的增量滑过基因组。一些实施方式中,对于局部基因组偏好评价确定,窗以下述增量滑过基因组,或其区段:约10,000bp或更多、约5,000bp或更多、约2,500bp或更多、约1,000bp或更多、约750bp或更多、约500bp或更多、约400碱基或更多、约250bp或更多、约100bp或更多、约50bp或更多或约25bp或更多。在一些实施方式中,对于局部基因组偏好评价确定,窗以下述增量滑过基因组或其区段:约25、24、23、22、21、20、19、18、17、16、15、14、13、12、11、10、9、8、7、6、5、4、3、2或约1bp。例如,对于局部基因组偏好评价确定,窗可包括约400bp(例如200bp带宽)且可以1bp增量滑过基因组。在一些实施方式中,确定基因组或其区段中的各碱基的局部基因组偏好评价,使用核心密度函数和约200bp带宽。

[0175] 在一些实施方式中局部基因组偏好评价是局部GC含量和/或表示局部GC含量。本文术语“局部”(例如用于描述局部偏好、局部偏好评价、局部偏好含量、局部基因组偏好、局部GC含量等)指10,000bp或更少的多核苷酸区段。在一些实施方式中,术语“局部”指5000bp或更少、4000bp或更少、3000bp或更少、2000bp或更少、1000bp或更少、500bp或更少、250bp或更少、200bp或更少、175bp或更少、150bp或更少、100bp或更少、75bp或更少或50bp或更少的多核苷酸区段。局部GC含量通常表示(例如数学,定量表示)基因组的局部区段的GC含量、序列读数、序列读数组装(例如毗连群、概况等)。例如,局部GC含量可为局部GC偏好评价或GC密度。

[0176] 通常确定参照或样品(例如测试样品)的多核苷酸的一个或多个GC密度。在一些实施方式中GC密度是表示(例如数学、定量表示)局部GC含量(例如5000bp或更少的多核苷酸区段)。在一些实施方式中,GC密度是局部基因组偏好评价。可使用本文所述和/或本领域已知的合适过程确定GC密度。可使用合适的PDF(例如核心密度函数(例如Epanechnikov核心密度函数,例如参见图1))确定GC密度。在一些实施方式中,GC密度是PDE(例如核心密度评价)。在某些实施方式中,GC密度通过是否存在一个或多个鸟嘌呤(G)和/或胞嘧啶(C)核苷酸来定义。在某些实施方式中,GC密度通过是否存在一个或多个腺嘌呤(A)和/或胸腺嘧啶(T)核苷酸来定义。在一些实施方式中,局部GC含量的GC密度,根据完整基因组或其区段(例如常染色体、染色体组、单染色体、基因,例如参见图2)所确定的GC密度标准化。可确定样品(例如测试样品)或参照样品的多核苷酸的一个或多个GC密度。通常确定参照基因组的GC密度。在一些实施方式中,根据参照基因组确定序列读数的GC密度。通常根据读数映射至的参照基因组的对应位点和/或位置所确定的GC密度确定读数的GC密度。在一些实施方式中,就参照基因组上的定位确定的GC密度分配和/或提供至读数,其中所述读数或其区段映射至参照基因组的相同位点。任何合适的方法可用于确定参照基因组上映射读数的位点用于生成读数的GC密度。在一些实施方式中,映射的读数的中值位置确定参照基因组上的位点,从

其确定读数的GC密度。例如,读数的中值位置映射至参照基因组的染色体12的碱基数 x 处时,读数的GC密度通常以参照基因组的染色体12上碱基数 x 处或附近的位置的核心密度评价所确定的GC密度来提供。在一些实施方式中,根据参照基因组确定读数的一些或所有碱基位置的GC密度。有时读数的GC密度包括参照基因组上多种碱基位置所确定的两个或更多GC密度的平均、加和、中值或积分。

[0177] 在一些实施方式中,局部基因组偏好评价(例如GC密度)经定量和/或提供数值。局部基因组偏好评价(例如GC密度)有时表达为平均、算术平均和/或中值。局部基因组偏好评价(例如GC密度)有时表达为PDE的最大峰高度。有时局部基因组偏好评价(例如GC密度)表示为合适PDE的加和或积分(例如曲线下面积(AUC))。一些实施方式中,GC密度包括核心加权。在某些实施方式中,读数的GC密度包括约等于下述的值:平均、算术平均、加和、中值、最大峰高度或核心加权的积分。

[0178] 偏好频率

[0179] 偏好频率有时根据一种或多种局部基因组偏好评价(例如GC密度)确定。偏好频率有时是样品、参照(例如,参照基因组、参照序列、参照基因组中的染色体)或其部分的局部基因组偏好评价的计数或加和。偏好频率有时是样品、参照或其部分的局部基因组偏好评价(例如各局部基因组偏好评价)的计数或加和。在一些实施方式中,偏好频率是GC密度频率。通常根据一个或多个GC密度确定GC密度频率。例如,GC密度频率可代表值 x 的GC密度相对整个基因组或其区段所代表的倍数。偏好频率通常为局部基因组偏好评价的分布,其中各局部基因组偏好评价的发生数代表偏好频率(例如参见图3)。偏好频率有时经数学处理和/或标准化。偏好频率可通过合适的方法经数学处理和/或标准化。在一些实施方式中,偏好频率根据样品、参照或其部分(例如常染色体、染色体子集、单染色体、或其读数)的各局部基因组偏好评价的表示(例如组分、百分比)进行标准化。可确定样品或参照的一些或全部局部基因组偏好评价的偏好频率。在一些实施方式中,可确定测试样品的一些或全部序列读数的局部基因组偏好评价的偏好频率。

[0180] 在一些实施方式中,系统包括偏好密度模块6。偏好密度模块可以任何合适的格式接受、取回和/或储存映射的序列读数5和参照序列2并生成局部基因组偏好评价、局部基因组偏好分布、偏好频率、GC密度、GC密度分布和/或GC密度频率(盒7统一代表)。在一些实施方式中偏好密度模块转移数据和/或信息(例如7)至其他合适的模块(例如关系模块8)。

[0181] 关系

[0182] 在一些实施方式中,在局部基因组偏好评价和偏好频率之间形成一种或多种关系。本文术语“关系”指两个或更多变量或值之间的数学和/或几何关系。关系可通过合适的数学和/或几何过程生成。关系的非限制性示例包括数学和/或几何过程表示:函数、相关性、分布、线性或非-线性等式、线、回归、拟合回归等或其组合。有时关系包括拟合关系。在一些实施方式中拟合关系包括拟合回归。有时关系包括加权的两个或更多变量或值。在一些实施方式中关系包括拟合回归,其中关系的一种或多种变量或值经加权。有时回归以加权形式拟合。有时回归拟合而不进行加权。在某些实施方式中,生成关系包括作图或图表。

[0183] 在一些实施方式中,在局部基因组偏好评价和偏好频率之间确定合适的关系。在一些实施方式中,在样品的(i)局部基因组偏好评价与(ii)偏好频率之间生成关系提供样品偏好关系。在一些实施方式中,在参照的(i)局部基因组偏好评价与(ii)偏好频率之间生

成关系提供参照偏好关系。在某些实施方式中,在GC密度和GC密度频率之间生成关系。在一些实施方式中,在样品的(i)GC密度与(ii)GC密度频率之间生成关系提供样品GC密度关系。在一些实施方式中,在参照的(i)GC密度与(ii)GC密度频率之间生成的关系提供参照GC密度关系。在一些实施方式中,局部基因组偏好评价是GC密度时,样品偏好关系是样品GC密度关系且参照偏好关系是参照GC密度关系。参照GC密度关系和/或样品GC密度关系的GC密度通常是局部GC含量的表示(例如数学或定量表示)。在一些实施方式中,局部基因组偏好评价和偏好频率之间的关系包括分布。在一些实施方式中,局部基因组偏好评价和偏好频率之间的关系包括拟合关系(例如拟合回归)。在一些实施方式中,局部基因组偏好评价和偏好频率之间的关系包括拟合线性或非-线性回归(例如多项式回归)。在某些实施方式中,局部基因组偏好评价和偏好频率之间的关系包括加权的关系,其中局部基因组偏好评价和/或偏好频率通过合适的过程加权。在一些实施方式中,加权的拟合关系(例如加权的拟合)可通过包括分位数回归、参数化分布或插值的经验分布的过程获得。在某些实施方式中,测试样品、参照或其部分的局部基因组偏好评价和偏好频率之间的关系包括多项式回归,其中局部基因组偏好评价经加权。一些实施方式中,加权的拟合模型包括加权分布的值。可通过合适过程加权分布的值。在一些实施方式中,靠近分布末尾的值比靠近分布中值的值提供更少的权重。例如,对于局部基因组偏好评价(例如GC密度)和偏好频率(例如GC密度频率)之间的分布,根据给定局部基因组偏好评价的偏好频率确定权重,其中包括接近分布算术平均的偏好频率的局部基因组偏好评价比包括较远离算术平均的偏好频率的局部基因组偏好评价提供更多权重。

[0184] 在一些实施方式中,系统包括偏好关系模块8。关系模块可生成关系以及定义关系的函数、系数、常量和变量。关系模块可从合适的模块(例如偏好密度模块6)接受、存储和/或收回数据和/或信息(例如7)和生成关系。关系模块通常生成并比较局部基因组偏好评价的分布。关系模块可比较数据组并有时生成回归和/或拟合关系。在一些实施方式中,关系模块比较一种或多种分布(例如样品和/或参照局部基因组偏好评价的分布)并提供序列读数的计数的加权因子和/或加权分配9至其他合适的模块(例如偏好校正模块)。有时关系模块直接将序列读数的标准化计数提供至分布模块21,其中计数根据关系和/或比较进行标准化。

[0185] 生成比较和其应用

[0186] 在一些实施方式中,降低序列读数中的局部偏好包括标准化序列读数计数。序列读数计数通常根据测试样品与参照的比较进行标准化。例如,有时序列读数计数通过比较测试样品的序列读数的局部基因组偏好评价与参照的局部基因组偏好评价(例如参照基因组或其部分)进行标准化。在一些实施方式中,序列读数计数通过比较测试样品的局部基因组偏好评价的偏好频率与参照的局部基因组偏好评价的偏好频率进行标准化。在一些实施方式中,序列读数计数通过比较样品偏好关系和参照偏好关系进行标准化,从而生成比较。

[0187] 序列读数计数通常根据两个或更多关系的比较标准化。在某些实施方式中,两个或更多关系进行比较,从而提供用于降低序列读数的局部偏好(例如标准化计数)的比较。两个或更多关系可通过合适的方法进行比较。在一些实施方式中,比较包括第二关系加、减、乘和/或除第一关系。在某些实施方式中比较两个或更多关系包括使用合适的线性回归和/或非-线性回归。在某些实施方式中比较两个或更多关系包括合适的多项式回归(例如3

阶多项式回归)。在一些实施方式中,比较包括第二回归加、减、乘和/或除第一回归。在一些实施方式中,两个或更多关系通过包括多种回归的推断框架的过程进行比较。在一些实施方式中,两个或更多关系通过包括合适的多元分析的过程进行比较。在一些实施方式中,两个或更多关系通过包括基函数(例如混合函数、例如多项式基、傅立叶基或等)、样条函数、径向基函数和/或小波的过程进行比较。

[0188] 在某些实施方式中,包括测试样品和参照的偏好频率的局部基因组偏好评价的分布通过包括多项式回归的过程进行比较,其中局部基因组偏好评价经加权。在一些实施方式中,在(i)比率,各比率包括参照的局部基因组偏好评价的偏好频率和样品的局部基因组偏好评价的偏好频率和(ii)局部基因组偏好评价之间生成多项式回归。在一些实施方式中,在(i)参照的局部基因组偏好评价的偏好频率与样品的局部基因组偏好评价的偏好频率的比率和(ii)局部基因组偏好评价之间生成多项式回归。在一些实施方式中,测试样品和参照的读数的局部基因组偏好评价的分布比较包括确定参照和样品的局部基因组偏好评价的偏好频率的对数比率(例如 $\log_2$ 比率)。在一些实施方式中,局部基因组偏好评价的分布的比较包括参照的局部基因组偏好评价的偏好频率的 $\log$ 比率(例如 $\log_2$ 比率)除以样品的局部基因组偏好评价的偏好频率的 $\log$ 比率(例如 $\log_2$ 比率)(例如参见实施例1和图4)。

[0189] 根据比较的标准化计数通常调整一些计数而不调整其他。标准化计数有时调整所有计数而有时不调整任何序列读数计数。序列读数计数有时通过包括确定加权因子的过程进行标准化和有时所述过程不包括直接生成和采用加权因子。根据比较的标准化计数有时包括确定各序列读数计数的加权因子。加权因子通常对序列读数特异且应用于特异性序列读数计数。加权因子通常根据两个或更多偏好关系(例如样品偏好关系比较参照偏好关系)的比较来确定。标准化计数通常通过根据加权因子调整计数值来确定。根据加权因子调整计数有时包括序列读数计数加、减、乘、和/或除加权因子。加权因子和/或标准化计数有时从回归(例如回归线)确定。标准化计数有时直接获自从参照(例如,参照基因组,参照基因组中的染色体)和测试样品的局部基因组偏好评价的偏好频率之间的比较得到的回归线(例如拟合回归线)。在一些实施方式中,样品的读数的各计数根据(i)读数的局部基因组偏好评价的偏好频率相比(ii)参照的局部基因组偏好评价的偏好频率之间的比较提供标准化计数值。在某些实施方式中,所获的样品序列读数计数经标准化且序列读数中的偏好降低。

[0190] 有时系统包括偏好校正模块10。在一些实施方式中,偏好校正模块的函数通过关系模拟模块8进行。偏好校正模块可从合适的模块(例如关系模块8,压缩模块4)接受、收回、和/或存储映射的序列读数和加权因子(例如9)。在一些实施方式中,偏好校正模块将计数提供至映射的读数。在一些实施方式中,偏好校正模块将加权分配和/或偏好校正因子应用至序列读数计数从而提供标准化和/或调整的计数。偏好校正模块通常提供标准化计数至其他合适的模块(例如分布模块21)

[0191] 在某些实施方式中,标准化计数包括因数化GC密度之外的一种或多种特征,并标准化序列读数计数。在某些实施方式中,标准化计数包括因数化一种或多种不同局部基因组偏好评价,并标准化序列读数计数。在某些实施方式中,根据由一种或多种特征(例如一种或多种偏好)所确定的加权对序列读数计数进行加权。在一些实施方式中,根据一种或多

种组合权重标准化计数。有时根据一种或多种组合权重因数化一种或多种特征和/或标准化计数通过包括使用多元模型的过程。任何合适的多元模型可用于标准化计数。多元模型的非限制性示例包括多元线性回归、多元分位数回归、经验数据的多元插值、非-线性多元模型等或其组合。

[0192] 在一些实施方式中,系统包括多元校正模块13。多元校正模块可多次进行偏好密度模块6、关系模块8和/或偏好校正模块10的函数从而调整多种偏好的计数。在一些实施方式中多元校正模块包括一个或多个偏好密度模块6、关系模块8和/或偏好校正模块10。偏好校正模块有时提供标准化计数11至其他合适的模块(例如分布模块21)

[0193] 加权的部份

[0194] 在一些实施方式中,部份经加权。在一些实施方式中,一个或多个部份经加权,从而提供加权的部份。加权部份有时去除部份依赖性。部份可通过合适的过程进行加权。在一些实施方式中,一个或多个部份通过本征函数加权(例如特征函数)。一些实施方式中,本征函数包括用正交的本征部份替代部份。在一些实施方式中系统包括部份加权模块42。在一些实施方式中,加权模块接受、收回和/或存储读数密度、读数密度概况、和/或调整的读数密度概况。在一些实施方式中,加权的部份通过部份加权模块提供。在一些实施方式中,需要加权模块来加权部份。加权模块可通过本领域已知或本文所述一种或多种加权方法加权部份。加权模块通常提供加权的部份至其他合适的模块(例如打分模块46、PCA统计模块33、概况生成模块26等)。

[0195] 主成分分析

[0196] 在一些实施方式中,读数密度概况(例如测试样品的读数密度概况(例如图7A)根据主成分分析(PCA)进行调整。一种或多种参照样品的读数密度概况和/或测试对象的读数密度概况可根据PCA调整。可根据PCA来调整基因组、基因组的一部分、染色体或染色体的区段的读数密度概况。通过PCA相关过程从读数密度概况中去除偏好在本文中有时指调整概况。PCA可通过合适的PCA方法或其变体进行。PCA方法的非限制性示例包括经典相关性分析(CCA)、Karhunen-Loève变换(KLT)、Hotelling变换、合适正交分解(POD)、X的奇异值分解(SVD)、XTX特征值分解(EVD)、因子分析、Eckart-Young定理、Schmidt-Mirsky定理、经验正交函数 EOF)、经验特征函数分解、经验成分分析、似谐波模式、光谱分析,经验模式分析、等、其变体或组合。PCA通常鉴定读数密度概况中的一种或多种偏好。PCA鉴定的偏好在本文中有时指主成分。在一些实施方式中,一种或多种偏好可使用合适的方法根据一种或多种主成分通过调整读数密度概况去除。读数密度概况可通过一种或多种主成分加、减、乘、和/或除读数密度概况而调整。在一些实施方式中,一种或多种偏好可通过读数密度概况减去一种或多种主成分而从读数密度概况中去除。虽然读数密度概况中的偏好通常通过概况的PCA鉴定和/或定量,主成分通常在读数密度水平从概况中减去。通过概况的PCA鉴定和/或定量的读数密度概况中的偏好或特征包括,但不限于,胎儿性别、序列偏好(例如,鸟嘌呤和胞嘧啶(GC)偏好)、胎儿分数、与DNA酶I敏感性相关的偏好、熵、重复序列偏好、染色质结构偏好、聚合酶误差率偏好、回文偏好、插入重复偏好、PCR扩增偏好、和隐藏拷贝数变异。

[0197] PCA通常鉴定一种或多种主成分。在一些实施方式中PCA鉴定第1、第2、第3、第4、第5、第6、第7、第8、第9和第10或更多主成分。在某些实施方式中,1、2、3、4、5、6、7、8、9、10或更多主成分用于调整概况。在某些实施方式中,使用5个主成分来调节概况。通常,主成分用于

以其在PCA中出现的顺序调整概况。例如，三个主成分从读数密度概况减去时，使用第1、第2、第3主成分。有时主成分所鉴定的偏好包括不用于调整概况的概况特征。例如，PCA可鉴定遗传变异（例如非整倍性、缺失、移位、插入）和/或性别差异（例如参见图6C）作为主成分。因此，在一些实施方式中，一种或多种主成分不用于调整概况。例如，有时第1、第2和第4主成分用于调整概况，当第3主成分不用于调整概况时。主成分可获自使用任何合适的样品或参照的PCA。在一些实施方式中，主成分获自测试样品（例如测试对象）。在一些实施方式中，主成分获自一种或多种参照（例如参照样品、参照序列、参照组）。例如，如图6所示，PCA在获自包括多种样品的训练组的中值读数密度概况上进行（图6A），得到第1主成分（图6B）和第2主成分（图6C）的鉴定。在一些实施方式中，主成分获自已知没有在研的遗传变异的对象组。在一些实施方式中，主成分获自已知整倍体组。主成分通常根据使用参照（例如训练组）的一种或多种读数密度概况进行的PCA来鉴定。获自参照的一种或多种主成分通常从测试对象的读数密度概况中减去（例如图7B），从而提供调整的概况（例如图7C）。

[0198] 在一些实施方式中，系统包括PCA统计模块33。PCA统计模块可从其他合适的模块（例如概况生成模块26）接受和/或收回读数密度概况。PCA通常通过PCA统计模块进行。PCA统计模块通常从参照组32、训练组30和/或从一种或多种测试对象28接受、收回和/或存储读数密度概况并加工读数密度概况。PCA统计模块可生成和/或提供主成分和/或调整读数密度概况，根据一种或多种主成分。调整的读数密度概况（例如40, 38）。PCA统计模块可将调整的读数密度概况（例如38, 40）提供和/或转移至其他合适的模块（例如部份加权模块42，打分模块46）。在一些实施方式中PCA统计模块可提供性别判定36。性别判定有时根据PCA和/或根据一种或多种主成分确定胎儿性别。在一些实施方式中PCA统计模块包括下示的R代码的一些、全部或改良。计算主成分的R代码通常从清除数据开始（例如减去中值、过滤部份、和清理极值）：

```
[0199] #清除PCA的数据异常
[0200] dclean<-(dat-m)[mask,]
[0201] for(j in 1:ncol(dclean))
[0202] {
[0203]   q<-quantile(dclean[,j],c(.25,.75))
[0204]   qmin<-q[1]-4*(q[2]-q[1])
[0205]   qmax<-q[2]+4*(q[2]-q[1])
[0206]   dclean[dclean[,j]<qmin,j]<-qmin
[0207]   dclean[dclean[,j]>qmax,j]<-qmax
[0208] }
[0209] 然后计算主成分：
[0210] #计算主成分
[0211] pc<-prcomp(dclean)$x
[0212] 最终，用下述计算各样品的PCA调整的概况：
[0213] #计算残差
[0214] mm<-model.matrix(~pc[,1:numpc])
[0215] for(j in 1:ncol(dclean))
```

[0216] `dclean[,j]<-dclean[,j]-predict(lm(dclean[,j]~mm))`

[0217] 比较概况

[0218] 在一些实施方式中,确定结果包括比较。在某些实施方式中,读数密度概况或其部份用于提供结果。在某些实施方式中,使用基因组、基因组的一部分、染色体、或染色体的区段的读数密度概况来提供结果。在一些实施方式中,确定结果(例如确定是否存在遗传变异)包括比较两个或更多读数密度概况。比较读数密度概况通常包括比较就基因组所选区段生成的读数密度概况。例如,测试概况通常与参照概况比较,其中测试和参照概况就基本相同的基因组区段(例如参照基因组)确定。比较读数密度概况有时包括比较两个或更多读数密度概况部份的子集。读数密度概况的部份子集可代表基因组区段(例如染色体或其区段)。读数密度概况可包括任何量的部份子集。有时读数密度概况包括2或更多、3或更多、4或更多、或5或更多子集。在某些实施方式中,读数密度概况包括两个子集的部份,其中各部份代表毗连的参照基因组区段。在一些实施方式中,测试概况可与参照概况比较,其中测试概况和参照概况都包括部份第一子集和部份第二子集,其中所述第一子集和第二子集代表基因组不同区段。读数密度概况的一些部份子集可包括遗传变异,而部份的其他子集有时基本不含遗传变异。有时概况(例如测试概况)的所有部份子集基本不含遗传变异。有时概况(例如测试概况)的所有部份子集包含遗传变异。在一些实施方式中,测试概况可包括含遗传变异的部份第一子集和基本不含遗传变异的部份第二子集。

[0219] 在一些实施方式中,本文所述方法包括进行比较(例如比较测试概况与参照概况)。两个或更多数据组\两个或更多关系和/或两个或更多概况可通过合适的方法进行比较。适合比较数据组、关系和/或概况的统计学方法的非限制性示例包括Behrens-Fisher法、拔靴法、组合显著独立检验的Fisher方法、Neyman-Pearson检验、确认性数据分析、探测的数据分析、精确检验、F-检验、Z-检验、T-检验、计算和/或比较不确定性测量、原假说、计算空值(counternulls)等、卡方检验、综合检验、显著性(例如统计学显著性)的计算和/或比较水平、荟萃分析、多元分析、回归、简单线性回归、加强线性回归等或前述的组合。在某些实施方式中,比较两个或更多数据组、关系和/或概况包括确定和/或比较不确定性测量。本文所用“不确定性测量”指显著性测量(例如统计学显著性)、误差测量、方差测量、置信度测量等或其组合。不确定性测量可为值(例如阈值)或值的范围(例如区间、置信区间、Bayesian置信区间、阈值范围)。不确定性测量的非限制性示例包括p-值、合适的差异测量(例如标准偏差、 σ 、绝对偏差、算术平均绝对偏差、等),合适的误差测量(例如标准误差、均方误差、根均方误差等),合适的方差测量,合适的标准分数(例如标准偏差、积累百分比、百分比等价、Z-分数、T-分数、R-分数、标准九分(标准九分数)、标准九分数中的百分比等)等或其组合。在一些实施方式中,确定显著性水平包括确定不确定性测量(例如p-值)。在某些实施方式中,两种或更多数据组、关系和/或概况可采用多种(例如2或更多)统计方法进行分析 and/或比较(例如最小二乘回归、主成分分析、线性判别分析、二次判别分析、Bagging、神经网络、支持向量机模型、随机森林、分类树模型、K-最近邻法(k-nearest neighbors),逻辑回归和/或平滑损失(loss smoothing)和/或任何合适的数学和/或统计学操作(例如本文所述操作)。

[0220] 在某些实施方式中,比较两个或更多读数密度概况包括就两个或更多读数密度概况确定和/或比较不确定性测量。读数密度概况和/或关联的不确定性量有时进行比较以便

于阐述数据组的数学和/或统计学处理和/或提供结果。测试对象的读数密度概况生成有时与就一种或多种参照(例如参照样品、参照对象等)生成的读数密度概况比较。在一些实施方式中,通过就染色体、其部份或区段比较测试对象的读数密度概况与参照的读数密度概况来提供结果,其中参照读数密度概况获自己知没有遗传变异的参照对象组(例如参照)。在一些实施方式中,通过就染色体、其部份或区段比较测试对象的读数密度概况与参照的读数密度概况来提供结果,其中参照读数密度概况获自己知含特异遗传变异的参照对象组(例如,染色体非整倍性、三体)。

[0221] 在某些实施方式中,测试对象的读数密度概况与无遗传变异的预定值表示比较,且有时在对应于遗传变异所定位的基因组位点的一种或多种基因组位点(例如部份)偏离预定值。例如,测试对象中(例如患有关联遗传变异的医学病症或有此风险的对象)的读数密度概况预期与含在研遗传变异的测试对象的参照(例如参照序列,参照对象,参照组)的所选部份的读数密度概况显著不同。测试对象读数密度概况通常与不含在研遗传变异的测试对象的参照(例如参照序列、参照对象、参照组)的所选部份的读数密度概况基本相同。读数密度概况通常与预定阈值和/或阈值范围比较(例如参见图8)。本文所用术语“阈值”指用符合要求的数据组计算并作为诊断遗传变异的限制(例如拷贝数变异、非整倍性、染色体异常等)的任何数字。在某些实施方式中,阈值超出本发明方法所获结果,对象诊断为具有遗传变异(例如三体)。在一些实施方式中,阈值或阈值范围通常通过数学和/或统计学处理序列读数数据(例如来自参照和/或对象)来计算。表明是否存在遗传变异的预定阈值或阈值范围可不同,但仍提供可用于确定是否存在遗传变异的结果。在某些实施方式中,生成包括标准化读数密度和/或标准化计数的读数密度概况以便于分类和/或提供结果。结果可基于包括标准化计数(例如使用该读数密度概况图)的读数密度概况图提供。

[0222] 在一些实施方式中,系统包括打分模块46。打分模块可从其他合适的模块(例如概况生成模块26,PCA统计模块33,部份加权模块42等)接受、取回和/或储存读数密度概况(例如调整的,标准化的读数密度概况)。打分模块可接受、收回、存储和/或比较两个或更多读数密度概况(例如测试概况、参照概况、训练组、测试对象)。打分模块可通常提供分数(例如图、概况统计、比较(例如两个或更多概况的差异)、Z-分数、不确定性测量、判定区、样品判定50(例如确定是否存在遗传变异)、和/或结果)。打分模块可提供分数给最终用户和/或至其他合适的模块(例如显示器、打印机等)。在一些实施方式中打分模块包括下述的R代码的一些、所有或改良,其包括计算特异性测试的卡方统计的R函数(例如高-chr21计数)。

[0223] 这三个参数为:

[0224] x=样品读数数据(部份x样品)

[0225] m=部份的中值

[0226] y=测试向量(例如,对所有部份为假,但对chr21为真)

[0227] getChisqP<-function(x,m,y)

[0228] {

[0229] ahigh<-apply(x[!y,],2,function(x) sum((x>m[!y])))

[0230] alow<-sum(!y)-ahigh

[0231] bhigh<-apply(x[y,],2,function(x) sum((x>m[y])))

[0232] blow<-sum(y)-bhigh

```
[0233] p<-sapply(1:length(ahigh),function(i) {  
[0234] p<-chisq.test(matrix(c(ahigh[i],alow[i],bhigh[i],blow[i]),2))$  
p.value/2  
[0235] if(ahigh[i]/alow[i]>bhigh[i]/blow[i])p<-max(p,1-p)  
[0236] else p<=min(p,1-p);p})  
[0237] return(p)
```

[0238] 实验条件

[0239] 在某些实施方式中,可针对与实验条件相关的偏好调整主成分标准化过程。根据实验条件的数据处理描述于,例如,国际专利申请公开号W02013/109981,其全文通过引用纳入本文,包括所有文字、表格、等式和附图。

[0240] 在某些情况中,可通过常见实验条件来影响样品。当与在不同时间和/或在相同时间使用不同条件和/或试剂处理的其它样品相比时,基本同时或使用基本相同的条件和/或试剂处理的样品有时显示出类似的实验条件(例如,常见实验条件)诱导的数据差异(例如,偏好)。通常存在限制可在实验过程期间在给定时间上制备、处理和/或分析的样品数量的实际考虑。在某些实施方式中,处理来自原材料的样品以生成结果的时间框架有时是数天、数周或者甚至数月。由于分离和最后分析之间的时间,分析大量样品的高通量实验有时生成批次效应或者实验条件诱导的数据差异。实验条件诱导的数据差异通常包括作为样品分离、储存、制备和/或分析的结果的任何数据差异。实验条件诱导的的非限制性示例包括基于流动池的差异和/或基于平板的差异,其包括:序列的过高或过低表示;噪音数据;假或异常数据点,试剂影响,个人影响,实验室条件影响等。实验条件诱导的差异有时发生于数据组中样品的亚群(例如,批次效应)。一个批次通常是使用基本相同的试剂处理的样品,在相同样品制备平板(例如,用于样品制备、核酸分离的微孔板)中处理的样品,经分期用于在相同分期平板(例如,用于在加载到流动池之前组织样品的微孔板)中分析的样品,基本同时处理的样品,同一个人处理的样品,和/或在基本相同的实验条件(例如,温度、CO₂水平、臭氧水平等或其组合)下处理的样品。实验条件批次效应有时影响在相同流动池上分析的、在相同试剂平板或微孔板中制备的和/或分期用于在相同试剂平板或微孔板中分析(例如,制备用于测序的核酸文库)的样品。差异的其它来源可包括,分离的核酸的质量、分离的核酸的量、核酸分离后至储存的时间、储存的时间、储存温度等及其组合。批次(例如,同时和/或使用相同试剂和/或实验条件处理的数据组中样品的亚群)中数据点的差异有时大于批次之间看到的数据点的差异。该数据差异有时包括假或异常数据,其幅度可能影响数据组中一些或全部其它数据的解释。可使用本文所述和本领域已知的数据处理步骤,例如针对流动池中分析或微孔板中处理的所有样品标准化至计算的中值绝对偏差来针对实验条件调整数据组的部份或全部。根据实验条件的数据处理描述于,例如,国际专利申请公开号W02013/109981,其全文通过引用纳入本文,包括所有文字、表格、等式和附图。

[0241] 使用比较检测非整倍性

[0242] 在一些实施方式中,主成分标准化过程与用于根据比较确定是否存在非整倍性的方法联用。使用比较检测非整倍性描述于,例如,国际申请公开号W0 2014/116598,其全部内容通过引用纳入本文,包括所有文字、表格、等式和附图。

[0243] 在本节中,比例或比例值、倍性评估和倍性评估值的比较统称为“比较”。在一些实

施方式中,根据一个或多个比较来确定对象中是否存在染色体非整倍性。在一些实施方式中,根据针对三个选择的常染色体(例如,三个选择的常染色体中的一个或多个是测试染色体)的一个或多个比较来确定对象中是否存在染色体非整倍性。在一些实施方式中,根据针对不同的染色体、整倍性区域、非整倍性区域或者整倍性区域或非整倍性区域的组生成的一个或多个比较来确定是否存在染色体非整倍性。在一些实施方式中,根据对象和整倍性区域和/或非整倍性区域(例如,针对参照组确定的整倍性区域和非整倍性区域)获得的比较来确定是否存在染色体非整倍性(例如,胎儿中的染色体非整倍性)。在某些实施方式中,根据对象和整倍性和/或非整倍性区域获得的比较之间的关系来确定是否存在染色体非整倍性。例如,在一些实施方式中,根据比较是在整倍性区域还是非整倍性区域,或者倍性评价距离整倍性区域或非整倍性区域多远来确定是否存在染色体非整倍性。在一些实施方式中,关系是靠近或远离(例如,数学差异和/或图形距离,例如,点和区域之间的距离)。可通过本领域已知或本文所述的合适方法确定关系,其非限制性示例包括概率分布、概率密度函数、累积分布函数、似然函数、Bayesian模型比较、贝叶斯因子、Deviance信息标准、 X^2 检验、欧氏距离、空间分析、马氏距离、曼哈顿距离、Chebyshev距离、Minkowski距离、Bregman散度、Bhattacharyya距离、Hellinger距离、度量空间、Canberra距离、凸包(例如,奇偶环绕法则)、等或其组合。

[0244] 在一些实施方式中,根据比较和整倍性区域确定不存在染色体非整倍性。在一些实施方式中,根据比较与整倍性区域之间的关系确定不存在染色体非整倍性。在一些实施方式中,落在整倍性区域内、中或附近的比较确定整倍性染色体(例如,不存在非整倍性染色体)。在一些实施方式中,在整倍性区域中或附近的比较表明确定比较的各染色体是整倍性。例如,有时,根据映射到ChrA、ChrB和ChrC的计数生成的比较落入整倍性区域(例如,按照映射到ChrA、ChrB和ChrC的计数确定的整倍性区域)中并且确定不存在染色体非整倍性。在一些实施方式中,根据比较确定的不存在染色体非整倍性表明各染色体(例如,衍生出倍性评价值的各染色体)是整倍性(例如,母体和/或胎儿中的整倍性)。

[0245] 在一些实施方式中,落在非整倍性外的比较确定一个或多个整倍性染色体。在一些实施方式中,在整倍性区域外的比较表明比较所确定的一个或多个染色体是整倍性。例如,有时,根据映射到ChrA、ChrB和ChrC的计数生成的比较落到整倍性区域(例如,按照映射到ChrA、ChrB和ChrC的计数确定的整倍性区域)外并且确定不存在染色体非整倍性。在一些实施方式中,在整倍性区域外的比较表明用于比较或评价并且比较所确定的两个或三个染色体是整倍性。

[0246] 在一些实施方式中,比较落入非整倍性区域中并且比较所确定的一个或多个染色体是整倍性。例如,有时,根据映射到ChrA、ChrB和ChrC的计数生成的比较落入非整倍性区域(例如,按照映射到ChrA、ChrB和ChrC的计数确定的非整倍性区域)中并且确定三个染色体中的2个不存在染色体非整倍性。

[0247] 在一些实施方式中,根据比较和整倍性区域确定存在染色体非整倍性。在某些实施方式中,根据比较与整倍性区域之间的关系确定存在染色体非整倍性。在一些实施方式中,落在整倍性区域外的比较确定非整倍性染色体(例如,存在非整倍性染色体)。在一些实施方式中,落在整倍性区域外的比较表明比较所确定的一个或多个染色体是非整倍性。例如,有时,根据映射到ChrA、ChrB和ChrC的计数生成的比较落到整倍性区域(例如,按照映射

到ChrA、ChrB和ChrC的计数确定的整倍性区域)外并且确定存在染色体非整倍性。

[0248] 在一些实施方式中,落在非整倍性区域内、中或附近的比较确定非整倍性染色体(例如,存在非整倍性染色体)。在一些实施方式中,在整倍性区域中或附近的比较表明确定倍性评价值的一个或多个染色体是非整倍性。在一些实施方式中,在非整倍性区域中或附近的比较表明确定比较的1、2、3、4和/或5个染色体是非整倍性。在一些实施方式中,在整倍性区域中或附近的比较表明确定比较的三个染色体中的一个是非整倍性。例如,有时,根据映射到ChrA、ChrB和ChrC的计数生成的比较落入非整倍性区域(例如,按照映射到ChrA、ChrB和ChrC的计数确定的非整倍性区域)中并且染色体中的一个是非整倍性染色体。

[0249] 在一些实施方式中,落在非整倍性区域附近的比较确定非整倍性染色体(例如,存在非整倍性染色体)。在一些实施方式中,在非整倍性区域附近的比较表明比较所确定的一个或多个染色体是非整倍性。在一些实施方式中,参照曲线包括限定的整倍性区域和三个限定的非整倍性区域(例如,针对Chr13、Chr18或Chr21的非整倍性)并且按照落在最接近非整倍性区域之一的比较确定存在非整倍性。例如,比另一个区域(例如,Chr13或Chr18的非整倍性,或整倍性区域)更靠近Chr21的非整倍性区域的比较可表示存在Chr21的非整倍性。

[0250] 在一些实施方式中,根据映射到Chr13、Chr18和Chr21的计数生成的比较落入非整倍性区域(例如,按照映射到Chr13、Chr18和Chr21的计数确定的非整倍性区域)中并且染色体中的一个是非整倍性染色体。在一些实施方式中,根据映射到Chr13、Chr18和Chr21的计数生成的比较落入非整倍性区域(例如,按照映射到Chr13、Chr18和Chr21的计数确定的非整倍性区域)中,Chr18和Chr21确定为整倍性并且Chr13确定为非整倍性。在一些实施方式中,根据映射到Chr13、Chr18和Chr21的计数生成的比较落入非整倍性区域(例如,按照映射到Chr13、Chr18和Chr21的计数确定的非整倍性区域)中,Chr13和Chr21确定为整倍性并且Chr18确定为非整倍性。在一些实施方式中,根据映射到Chr13、Chr18和Chr21的计数生成的比较落入非整倍性区域(例如,按照映射到Chr13、Chr18和Chr21的计数确定的非整倍性区域)中,Chr18和Chr13确定为整倍性并且Chr21确定为非整倍性。

[0251] 在一些实施方式中,根据第一比较和第二比较确定是否存在染色体非整倍性,其中2种比较生成自映射到2个或更多个染色体的相同组的序列读数。在一些实施方式中,根据对象生成的第一比较和第二对象生成的第二比较之间的关系(例如,距离)来确定对象中是否存在染色体非整倍性。在一些实施方式中,第二比较是针对一个或多个对象生成的比较组(例如,区域)。在一些实施方式中,根据对象生成的第一比较和一个或多个对象生成的参照比较组之间的关系(例如,距离)来确定对象中是否存在染色体非整倍性。在一些实施方式中,第一比较是对象的比较并且第二比较是代表一个或多个整倍性胎儿的比较或比较组。在一些实施方式中,第二比较是整倍性胎儿预期的值或一组值(例如,区域)。在一些实施方式中,第二比较是对象(例如,妊娠雌性对象)生成的值或一组值,其中胎儿已知针对比较生成的一个或多个染色体是整倍性。在一些实施方式中,根据不确定性值(例如标准偏差或MAD)确定距离。在一些实施方式中,第一和第二比较之间的距离(例如,第二比较代表一个或多个整倍性对象)是相关不确定性的1、2、3、4、5、6或更多倍并且第一比较确定是非整倍性。在一些实施方式中,第一和第二比较之间的距离(例如,第二比较代表一个或多个整倍性对象)是相关不确定性的3或更多倍并且第一比较确定代表非整倍性染色体。

[0252] 在一些实施方式中,按照根据映射至一个或多个具体染色体和整倍性区域、非整

倍性区域、或整倍性区域和非整倍性区域的计数生成的比较来确定是否存在染色体非整倍性。在一些实施方式中,按照根据映射至一个或多个具体染色体的序列读数生成的比较确定是否存在染色体非整倍性并且该确定不需要映射至其它染色体的序列读数。在一些实施方式中,按照根据映射至2、3、4、5或6个不同染色体的序列读数生成的比较确定是否存在染色体非整倍性并且该确定不需要或得到映射至其它染色体的计数。在一些实施方式中,按照根据3个不同染色体或其区段生成的比较确定是否存在染色体非整倍性并且该确定并不基于除了这3个不同染色体之一以外的染色体。例如,在ChrA、ChrB和ChrC代表3个不同染色体或其片段的情况中,有时按照根据ChrA、ChrB和ChrC生成的比较确定是否存在染色体非整倍性,并且该确定并不基于除ChrA、ChrB和ChrC以外的染色体。在一些实施方式中,ChrA、ChrB和ChrC分别代表Chr13、Chr21和Chr18。

[0253] 性染色体核型

[0254] 在一些实施方式中,主成分标准化过程与用于确定性染色体核型的方法联用。用于确定性染色体核型的方法描述于,例如,国际申请公开号WO 2013/192562,其全部内容通过引用纳入本文,包括所有文字、表格、等式和附图。

[0255] 在一些实施方式中,映射至一个或多个性染色体(即,染色体X,染色体Y)的序列读数计数被标准化。在一些实施方式中,标准化包括主成分标准化。在一些实施方式中,标准化包括确定参照基因组的部份的实验偏好。在一些实施方式中,可从映射至参照基因组的各部份的序列读数的计数和各部份的映射特征(例如,GC含量)之间的各样品的第一拟合关系(例如,拟合的线性关系、拟合的非线性关系)确定多个样品的实验偏好。一般通过线性回归确定拟合的关系(例如,线性关系)的斜率。在一些实施方式中,由实验偏好系数表示各实验偏好。实验偏好系数是例如(i)映射至参照基因组的各部份的序列读数的计数和(ii)各部份的映射特征之间的线性关系的斜率。在一些实施方式中,实验偏好可包括实验偏好曲率评价。

[0256] 在一些实施方式中,方法还包括计算来自实验偏好和映射至各部份的序列读数的计数的第二拟合关系(例如,拟合的线性关系,拟合的非线性关系)的各基因组部份的基因组区段水平(例如,升高,水平),并且可通过线性回归确定关系的斜率。例如,如果第一拟合的关系是线性的并且第二拟合的关系是线性的,则可按照式 α 确定参照基因组的各部份的基因组区段水平 L_i :

$$[0257] \quad L_i = (m_i - G_i S) I^{-1} \quad \text{式}\alpha$$

[0258] 其中 G_i 是实验偏好, I 是第二拟合的关系的截距, S 是第二关系的斜率, m_i 是映射至参照基因组的各部份的测量的计数并且 i 是样品。

[0259] 在一些实施方式中,将二级标准化过程应用于一个或多个计算的基因组区段水平。在一些实施方式中,二级标准化包括GC标准化并且有时包括使用PERUN方法。在一些实施方式中,二级标准化包括主成分标准化。

[0260] 胎儿倍性确定

[0261] 在一些实施方式中,主成分标准化过程与用于确定胎儿倍性的方法联用。用于确定胎儿倍性的方法描述于,例如,美国专利申请公开号US 2013/0288244,其全部内容通过引用纳入本文,包括所有文字、表格、等式和附图。

[0262] 可部分通过胎儿分数的测量确定胎儿倍性,并且胎儿倍性的确定用于确定是否存

在遗传变异(例如,染色体非整倍性、三体性)。胎儿倍性可通过合适的胎儿分数确定的任意方法(包括本文所述的方法)确定的检测的胎儿分数部份确定。在一些实施方式中,该方法需要计算的参照计数 F_i (有时,表示为 f_i),其针对多个样品的基因组的部份(即箱(bin), i)确定,其中,基因组的部份 i 的胎儿的倍性已知是整倍性。在一些实施方式中,不确定性值(例如,标准偏差, σ)被确定用于参照计数 f_i 。在一些实施方式中,参照计数 f_i ,一个不确定性值,测试样品计数和/或测定的胎儿分数(F)被用于确定胎儿倍性。在一些实施方式中,参照计数(例如,平均、算术平均或中值参照计数)通过主成分标准化和/或其它标准化,例如,分批标准化、GC含量标准化、线性和非线性最小二乘方回归、LOESS、GC LOESS、LOWESS、PERUN、RM、GCRM和/或其组合进行标准化。在一些实施方式中,当参照计数用主成分标准化进行标准化时,已知整倍性的基因组的区段的参照计数等于1。在一些实施方式中,测试样品的计数和参照计数(例如,已知为整倍性的胎儿)针对基因组的部份或区段通过主成分标准化进行标准化,并且参照计数等于1。在一些实施方式中,当参照计数用PERUN标准化时,已知整倍性的基因组的区段的参照计数等于1。在一些实施方式中,测试样品的计数和参照计数(例如,已知为整倍性的胎儿)针对基因组的部份或区段通过PERUN进行标准化,并且参照计数等于1。同样地,在一些实施方式中,当所述计数通过参照计数的中值标准化(即计数除以参照计数的中值)时,已知整倍性的基因组的部份或区段的参照计数等于1。例如,在一些实施方式中,基因组的部份或区段的参照计数(例如,已知整倍性的胎儿)和测试样品的计数通过中值参照计数标准化,所述标准化的参照计数等于1,并且所述测试样品计数是经标准化(例如,除以)的中值参照计数。在一些实施方式中,基因组的部份或区段的参照计数(例如,已知整倍性的胎儿)和测试样品的计数通过主成分标准化、GCRM、GC、RM或合适的方法标准化。在一些实施方式中,参照计数是平均、算术平均或中值参照计数。参照计数通常是箱的标准化的计数(例如,标准化的基因组区段水平)。在一些实施方式中,参照计数和测试样品的计数是原始计数。在一些实施方式中,参照计数由平均、算术平均或中值计数概况确定。在一些实施方式中,参照计数是计算的基因组区段水平。在一些实施方式中,参照样品的参照计数和测试样品(例如,患者样品,例如, y_i)的计数通过相同方法或处理标准化。

[0263] 其它数据处理和标准化

[0264] 已经计数的经映射的序列读数在本文中被称为原始数据,因为所述数据表示未处理的计数(如原始计数)。在一些实施方式中,数据组中的序列读数数据能经进一步处理(如数学和/或统计学处理)和/或显示以帮助提供结果。在某些实施方式中,数据组(包括较大数据组)可得益于预处理以帮助进一步的分析。数据组的预处理有时涉及除去冗余的和/或无信息的部份或参照基因组的部份(如具有无信息数据的部份或参照基因组的部份、冗余映射读数、具有0中值计数的部份、过高频出现或过低频出现的序列)。不受理论限制,数据处理和/或预处理可以(i)除去噪音数据、(ii)除去无信息数据、(iii)除去冗余数据、(iv)降低较大数据组的复杂性,和/或(v)帮助所述数据从一种形式变换成一种或多种其它形式。当用于数据或数据组时,术语“预处理”和“处理”在本文中被统称为“处理”。在一些实施方式中,处理能使数据更易于被进一步分析,从而能生成结果。在一些实施方式中,一种或多种或所有处理方法(例如标准化方法、部份筛选、映射、确认等或其组合)通过处理器、微处理器、计算机、与存储器联用来进行和/或通过微处理器控制的设备来进行。

[0265] 本文所用的术语“噪音数据”指(a)分析或作图时在数据点之间有显著差异的

数据、(b) 有显著标准偏差 (例如大于3标准偏差) 的数据、(c) 有显著的标准误差均值的数据等, 及以上组合。由于起始材料 (如核酸样品) 的数量和/或质量, 有时出现噪音数据, 并且噪音数据有时作为制备或复制用于生成序列读数的DNA的方法的部份出现。在某些实施方式中, 噪音来自使用基于PCR的方法制备时的过高频出现的某些序列。本文所述方法能降低或消除噪音数据的基值, 从而降低噪音数据对所提供结果的影响。

[0266] 本文使用术语“无信息数据”、“参照基因组的无信息部份”和“无信息部份”指所具有的数值与预定的阈值显著不同或落在预定截止值范围外的部份或其导出的数据。本文术语“阈值”指用符合要求的数据组计算的任何数字, 作为诊断遗传变异的限制 (例如拷贝数变异、非整倍性、微复制、微缺失、染色体异常等)。在某些实施方式中, 阈值超出本发明方法所获结果, 对象诊断为具有遗传变异 (例如21三体)。在一些实施方式中, 常通过数学和/或统计学处理序列读数数据 (如来自参照和/或对象) 计算阈值或值的范围, 而在某些实施方式中, 经处理以生成阈值或值的范围的序列读数数据是序列读数数据 (如来自参照和/或对象)。在一些实施方式中, 确定不确定值。不确定值通常是方差或误差的量度且可为变异或误差的任何合适量度。在一些实施方式中, 不确定值是标准偏差、标准误差、计算方差、p-值或算术平均绝对偏差 (MAD)。在一些实施方式中, 可根据本文所述的公式计算不确定值。

[0267] 任何合适的程序可用于处理本文所述的数据组。适用于处理数据组的方法的非限定性示例包含过滤、标准化、加权、监测峰高、监测峰面积、监测峰边缘、确定面积比率、数据的数学处理、数据的统计学处理、数学算法的应用、采用固定变量的分析、采用优化变量的分析、对数据作图以鉴定模式或趋势以供其它处理等, 及以上组合。在一些实施方式中, 根据不同特点 (如GC含量、冗余定位读数、着丝粒区域、端粒区域等, 及其组合) 和/或变量 (如胎儿性别、母本年龄、母本倍性、胎儿核酸基值百分比等及其组合) 处理数据组。在某些实施方式中, 处理本文所述的数据组能降低大数据组和/或复杂数据组的复杂性和/或维数。复杂数据组的非限定性示例包括由一个或多个测试对象和不同年龄与种族背景的多种参照对象生成的序列读取数据。在一些实施方式中, 数据组能包含各测试对象和/或参照对象的数千到数百万的序列读数。

[0268] 某些实施方式中, 数据处理可在任何数量的步骤中进行。例如, 在一些实施方式中, 可以仅用单一处理方法调整和/或处理数据, 而在某些实施方式中, 可以使用1个或更多个、5个或更多个、10个或更多个或者20个或更多个处理步骤 (如1个或更多个处理步骤、2个或更多个处理步骤、3个或更多个处理步骤、4个或更多个处理步骤、5个或更多个处理步骤、6个或更多个处理步骤、7个或更多个处理步骤、8个或更多个处理步骤、9个或更多个处理步骤、10个或更多个处理步骤、11个或更多个处理步骤、12个或更多个处理步骤、13个或更多个处理步骤、14个或更多个处理步骤、15个或更多个处理步骤、16个或更多个处理步骤、17个或更多个处理步骤、18个或更多个处理步骤、19个或更多个处理步骤或者20个或更多个处理步骤) 处理数据。在一些实施方式中, 处理步骤可以是重复两次或更多次 (如过滤两次或更多次, 标准化两次或更多次) 的同一处理步骤, 而在某些实施方式中, 处理步骤可以是同时或依次进行的两种或更多种不同的处理步骤 (如过滤、标准化; 标准化、监测峰高和边缘; 过滤、标准化、针对参照标准化、统计学处理以确定p值等)。在一些实施方式中, 可使用相同或不同处理步骤的任意合适次数和/或组合来处理序列读数数据以帮助提供结果。在某些实施方式中, 通过本文所述的标准加工数据组可降低数据组的复杂性和/或维数。

[0269] 在一些实施方式中,一个或多个处理步骤能包括一个或多个过滤步骤。本文所用的术语“过滤”指将部份或参照基因组的部份从考虑中除去。能根据任意合适的标准选择需除去的部份或参照基因组的部份,包括但不限于冗余数据(如冗余或重叠映射读数)、无信息数据(如有0中值计数的部份或参照基因组的部份)、含有过高频出现或过低频出现序列的部份或参照基因组的部份、噪音数据等,及以上组合。过滤方法经常涉及从考虑中除去一个或多个参照基因组的部份,并从考虑的参照基因组、染色体或基因组的计数或总计中减去所选需除去的参照基因组的一个或多个部份中的计数。在一些实施方式中,参照基因组的部份能依次被除去(如一次一个以允许评价各个单独部份的去除影响),而在某些实施方式中,标记为需除去的所有部份能同时被除去。在一些实施方式中,由高于或低于某一水平的差异所表征的参照基因组的部份经移除,其有时称为过滤参照基因组的“噪音”部份。在某些实施方式中,过滤过程包括通过预定的多种概况变化从源自部份、染色体、或染色体区段的平均概况水平的数据组中获取数据点,和在某些实施方式中,过滤过程包括通过预定的多种概况差异从源自部份、染色体、或染色体区段的平均概况水平的数据组中移除数据点。在一些实施方式中,过滤过程用于降低用于分析遗传变异存在或缺失的参照基因组中候选部份的数量。降低用于分析遗传变异(例如微缺失、微复制)存在或缺失的参照基因组中候选部份的数量通常降低数据组的复杂度和/或维度,并且有时提高搜索和/或鉴定遗传变异和/或遗传异常的速度两个或更多数量级。

[0270] 在一些实施方式中,一个或多个处理步骤能包括一个或多个标准化步骤。标准化可通过本文所述或本领域已知的合适方法进行。在某些实施方式中,标准化包括将不同量级的所测数值调整至理论上的共同量级。在某些实施方式中,标准化包括复杂化的数学调整,以在比对中引入调节的数值的概率分布。一些实施方式中,标准化包括将分布与正态分布进行比对。在某些实施方式中,标准化包括数学调整,其允许以消除某些总影响效果(例如误差和异常)的方式针对不同数据组的对应标准化值的比较。在某些实施方式中,标准化包括缩放。标准化有时包括通过预定标量或公式划分一个或多个数据组。标准化有时包括通过预定标量或公式扣除一个或多个数据组。标准化方法的非限制性示例包括逐份标准化、通过GC含量的标准化、中值计数(中值箱计数、中值部份计数)标准化、线性和非线性最小二乘回归、LOESS、GC LOESS、LOWESS(局部加权回归散点平滑法)、PERUN、ChAT、主成分标准化、重复掩蔽(RM)、GC-标准化和重复掩蔽(GCRM)、cQn和/或其组合。在一些实施方式中,确定存在或缺失遗传变异(例如非整倍性、微复制、微缺失)采用标准化方法(例如逐份标准化、通过GC含量的标准化、中值计数(中值箱计数、中值部份计数)标准化、线性和非线性最小二乘回归、LOESS、GC LOESS、LOWESS(局部加权回归散点平滑法)、PERUN、ChAT、主成分标准化、重复掩蔽(RM)、GC-标准化和重复掩蔽(GCRM)、cQn、本领域已知的标准化方法和/或其组合)。在一些实施方式中,确定是否存在遗传变异(例如,非整倍性、微复制、微缺失)采用LOESS、中值计数(中值箱计数、中值部份计数)标准化、和主成分标准化中的一种或多种。在一些实施方式中,确定是否存在遗传变异采用LOESS,之后是中值计数(中值箱计数,中值部份计数)标准化。在一些实施方式中,确定是否存在遗传变异采用LOESS,之后是中值计数(中值箱计数,中值部份计数)标准化,之后是主成分标准化。

[0271] 可使用任何合适的标准化次数。在一些实施方式中,数据组能经标准化1次或更多次、5次或更多次、10次或更多次或甚至20次或更多次。可使数据组针对表示任何合适特点

或变量(如样品数据、参照数据或两者)的值(如标准化值)作标准化。可用的数据标准化类型的非限定性示例包括使一个或多个所选测试或参照部份的原始计数数据针对映射到所选部份或区段的所述染色体或全基因组的映射上的计数总数作标准化;使一个或多个所选的部份的原始计数数据针对映射到所选部份或区段的一个或多个部份或所述染色体的中值参照计数作标准化;使原始计数数据针对前述经标准化的数据或其导出数作标准化;和使前述经标准化的数据针对一个或多个其它预定的标准化变量作标准化。根据所选作为预定标准化变量的特点或属性,使数据组标准化有时具有分离统计误差的作用。通过使数据转为共同量级(scale)(如预定的标准化变量),使数据组标准化有时也使量级不同的数据的数据特征具有可比性。在一些实施方式中,对统计学导出数值的一次或多次标准化可用于使数据差异最小化并减少异常数据(outlying数据)的重要性。涉及标准化值时,对部份或参照基因组的部份标准化有时称为“逐份标准化”。

[0272] 在某些实施方式中,加工步骤包括标准化,包括标准化至静态窗,和在一些实施方式中,加工步骤包括标准化,包括标准化至动态或滑动窗。本文术语“窗”指选择用于分析的一种或多种部份,有时用作比较的参照(如用作标准化和/或其他数学或统计操作)。本文术语“标准化至静态窗”指使用选择用于比较测试对象和参照对象数据组的一种或多种部份的标准化过程。在一些实施方式中,所选部份用于生成概况。静态窗通常包括预定的部份的组,其在操作和/或分析期间不改变。本文术语“标准化至动态窗”或“标准化滑动窗”指对定位于所选测试部份的基因组区域(例如遗传上紧密环绕、毗连的部份或区段)的部份进行标准化,其中一种或多仲所选测试部份标准化至紧密环绕所选测试部份的部份。在某些实施方式中,所选部份用于生成概况。滑动或动态窗标准化通常包括重复移动或滑动至毗邻测试部份,和将新选的测试部份标准化至紧密环绕或毗邻所述新选的测试部份的部份,其中毗邻的窗具有一种或多种共用部份。在某些实施方式中,可通过滑动窗过程分析多种所选测试部份和/或染色体。

[0273] 在一些实施方式中,标准化至滑动或动态窗可产生一个或多个值,其中各值代表针对从基因组不同区域(例如染色体)选择的不同组参照部份的标准化。在某些实施方式中,所得的一个或多个值是累加值(例如就所选部份、结构域(例如染色体的部份)或染色体的标准化计数概况的积分的数字评价)。滑动或动态窗过程所得值可用于产生概况和便于获得结果。在一些实施方式中,一个或多个部份的累加可显示为基因组位置的函数。动态或滑动窗分析有时用于分析基因组是否存在微缺失和/或微插入。在某些实施方式中,显示一个或多个部份的累加用于鉴定是否存在遗传变异区域(例如微缺失、微复制)。在一些实施方式中,动态或滑动窗分析用于鉴定含微缺失的基因组区域和在某些实施方式中,动态或滑动窗分析用于鉴定含微复制的基因组区域。

[0274] 下文详细描述可使用的标准化过程的某些示例,例如,LOESS、PERUN、ChAI和主成分标准化方法。

[0275] 在一些实施方式中,加工步骤包括加权。本文所用的术语“加权的”、“加权”或“加权函数”或其语法衍生形式或等同形式指对数据组的部分或全部的数学处理,所述数据组处理有时用于改变某些数据组特点或变量对其它数据组特点或变量的影响(如根据所选参照基因组的一个或多个部份中数据的质量或实用性,增加或减少参照基因组的一个或多个部份中所含数据的重要性的和/或基值)。在一些实施方式中,加权函数能用于增加测量变量

相对较小的数据的影响,和/或减少测量差异相对较大的数据的影响。例如,能对含有过低频出现或低量序列数据参照基因组的部份“降权重(down weighted)”以最小化对数据组的影响,反之能对参照基因组的所选部份“升权重(up weighted)”以增加对数据组的影响。加权函数的非限定性示例是 $[1/(\text{标准偏差})^2]$ 。加权步骤有时以与标准化步骤基本相似的方式进行。在一些实施方式中,使数据组除以预定变量(如加权变量)。经常选择预定变量(如最小化靶函数, Φ)对数据组的不同部份有区分地加权(如增加某些数据类型的影响,而降低其它数据类型的影响)。

[0276] 在某些实施方式中,处理步骤能包括一种或多种数学和/或统计学处理。任意合适的数学和/或统计学处理可以单一或联合用于分析和/或处理本文所述的数据组。能使用任意合适次数的数学和/或统计学处理。在一些实施方式中,数据组可经数学和/或统计学处理1次或多次、5次或更多次、10次或更多次或者20次或更多次。能使用的数学和统计学处理的非限定性示例包括加法、减法、乘法、除法、代数函数、最小二乘评价、曲线拟合、微分方程、有理多项式、二重多项式、正交多项式、z-分值、p值、x值、 Φ 值、峰水平分析、确定峰边缘位置、计算峰面积比、分析中值染色体水平、计算算术平均绝对偏差、残差平方和、平均、标准偏差、标准误等,或其组合。能对序列读取数据或其经处理的结果的全部或部分进行数学和/或统计学处理。可经统计学处理的数据组变量或特征的非限制性示例包括原始计数、过滤计数、标准化计数、峰高、峰宽、峰面积、峰边缘、侧向公差、P值、中值水平、平均水平、基因组区域内的计数分布、核酸物质的相对值表示等,或其组合。

[0277] 在一些实施方式中,处理步骤能包括使用一种或多种统计学算法。任意合适的统计学算法都可以单一或联合用于分析和/或处理本文所述的数据组。可使用任何合适数量的统计学算法。在一些实施方式中,能使用1种或更多种、5种或更多种、10种或更多种或者20种或更多种统计学算法来分析数据组。合适伴随本文所述方法使用的统计学算法的非限定性示例包括决策树、计数空值(计数ernull)、多重比较、综合检验、贝伦斯-费希尔问题、拔靴法(bootstrapping)、结合显著性独立测试的费舍尔方法、零假设、I型误差、II型误差、精确检验、单样本Z检验、双样本Z检验、单样本t检验、配对t检验、有相等方差的双样本合并t检验、有不相等方差的双样本未合并t检验、单比例z检验、合并的双比例z检验、未合并的双比例z检验、单样本卡方检验、有相等方差的双样本F检验、置信区间、可信区间、显著性、荟萃分析(meta-analysis)、简单线性回归、强线性回归等,或前述组合。能使用统计学算法分析的数据组变量或特征的非限制性示例包括原始计数、过滤计数、标准化计数、峰高、峰宽、峰边缘、侧向公差(lateral tolerance)、P值、中值水平、平均水平、基因组区域内的计数分布、核酸物质的相对值表示等,或其组合。

[0278] 在某些实施方式中,数据组可采用多种(例如2或更多)统计算法进行分析(例如最小二乘回归、主成分分析、线性判别分析、二次判别分析、Bagging、神经网络、支持向量机模型、随机森林、分类树模型、K-最近邻法(k-nearest neighbors),逻辑回归和/或平滑损失(loss smoothing)和/或数学和/或统计学操作(例如本文所述操作)。在一些实施方式中,使用多种操作可产生N维空间,其可用于提供结果。在某些实施方式中,通过采用多种操作的数据组分析可降低数据组的复杂性和/或维度。例如,在参照数据组上使用多种操作可产生N维空间(例如概率图),其可用于代表是否存在遗传变异,这取决于参照样品的遗传状态(例如对所选遗传变异阳性或阴性)。使用基本相似的操作组分析测试样品可用于产生所测

样品各自的N维点。测试对象数据组的复杂性和/或维度有时降低至可容易地与参照数据的N维空间进行比较的N维度点或单一值。落入由参照对象数据填充的N维空间内的测试样品数据表示与该参照对象基本相似的遗传状态。落入由参照对象数据填充的N维空间外的测试样品数据表示与该参照对象基本不相似的遗传状态。在一些实施方式中,参照是整倍体或者不具有遗传变异或医学症状。

[0279] 在一些实施方式中,数据组经计算、任选过滤和标准化后,经加工的数据组可用一种或多种过滤和/或标准化程序进一步操作。在某些实施方式中,可用一种或多种过滤和/或标准化程序进一步操作的数据组可用于产生概况。在一些实施方式中,一种或多种过滤和/或标准化程序有时可降低数据组复杂性和/或维度。可基于复杂性和/或维度降低的数据组提供结果。

[0280] 在一些实施方式中,部份可根据误差测量进行过滤(例如标准偏差、标准误差、计算方差、p-值、算术平均绝对误差(MAE)、平均绝对偏差和/或算术平均绝对偏差(MAD)。在某些实施方式中误差测量指计数可变性。在一些实施方式中,部份根据计数可变性进行过滤。在某些实施方式中,计数可变性是针对多种样品(例如获自多种对象的多种样品,例如50或更多、100或更多、500或更多、1000或更多、5000或更多或10,000或更多对象)的映射至参照基因组的部份(即部份)的计数所确定的误差测量。在一些实施方式中,具有高于预定上限范围的计数可变性的部份可被过滤(例如从考虑中排除)。在一些实施方式中,预定上限范围是等于或大于约50,约52,约54,约56,约58,约60,约62,约64,约66,约68,约70,约72,约74或等于或大于约76的MAD值。在一些实施方式中,具有低于预定下限范围的计数可变性的部份可被过滤(例如从考虑中排除)。在一些实施方式中,预定下限范围是等于或小于约40,约35,约30,约25,约20,约15,约10,约5,约1等于或小于约0的MAD值。在一些实施方式中,具有超出预定范围的计数可变性的部份可被过滤(例如从考虑中排除)。在一些实施方式中,预定范围是大于0且小于约76、小于约74、小于约72、小于约71、小于约70、小于约69、小于约68、小于约67、小于约66、小于约65、小于约64、小于约62、小于约60、小于约58、小于约56、小于约54、小于约52、小于约50的MAD值。在一些实施方式中,预定范围是大于0和小于约67.7的MAD值。在一些实施方式中,选择计数可变性在预定范围内的部份(例如用作确定是否存在遗传变异)。

[0281] 在一些实施方式中,部份的计数可变性代表分布(例如正态分布)。在一些实施方式中,可在分布的分位数内选择部份。在一些实施方式中,选择分布的分位数在等于或小于约99.9%、99.8%、99.7%、99.6%、99.5%、99.4%、99.3%、99.2%、99.1%、99.0%、98.9%、98.8%、98.7%、98.6%、98.5%、98.4%、98.3%、98.2%、98.1%、98.0%、97%、96%、95%、94%、93%、92%、91%、90%、85%、80%或等于或小于约75%内的部份。在一些实施方式中,选择计数可变性的分布的分位数在99%内的部份。在一些实施方式中,选择 $MAD > 0$ 和 $MAD < 67.725$,分位数在99%内的部份,得到对参照基因组的稳定部份组的鉴定。

[0282] 涉及PERUN的部份过滤的非限制性示例(例如)本文所述以及国际专利申请号PCT/US12/59123(WO2013/052913),其全文通过引用纳入本文,包括所有文字、表格、等式和图。部份可基于或部分基于误差测量进行过滤。误差测量包括偏差的绝对值,例如R-因子,其在某些实施方式中可用于部份去除或加权。在一些实施方式中,R-因子定义为预测计数值与实测值的绝对偏差除以来自实测值的预测计数值之后的加和。虽然可使用包括偏差绝对值

的误差测量,合适的误差测量也可使用。在某些实施方式中,可使用不包括偏差绝对值的误差测量,例如基于二乘的分散体。在一些实施方式中,根据可映射能力的测量(例如可映射分数)对部份进行过滤或加权。有时根据映射至所述部份相对低数目的序列读数(例如映射至所述部份的0、1、2、3、4、5个读数)对部份进行过滤或加权。可根据进行分析的种类对部份进行过滤或加权。例如,对于染色体13、18和/或21的非整倍性分析,可过滤性染色体,并且仅常染色体或常染色体子集可进行分析。

[0283] 在具体实施方式中,可使用下述过滤过程。选择给定染色体(例如染色体21)内的相同组的部份(例如参照基因组的部份),并比较受影响和未受影响的样品中读数的数量。差距涉及21三体和整倍体样品,其涉及覆盖大多数染色体21的部份组。在整倍体和T21样品之间的所述部份组相同。部份组和单一区段之间的区别并不关键,如部份所定义的那样。比较不同患者中的相同的基因组区域。该过程可用作三体分析,例如T13或T18,除了或代替T21。

[0284] 在一些实施方式中,数据组经计算、任选过滤和标准化后,经加工的数据组可通过加权进行操作。在某些实施方式中,一种或多种部份可选择进行加权以减少所选部份中所含的数据的影响(例如噪音数据、无信息数据),和在一些实施方式中,一种或多种部份可选择进行加权以提高或加强所选部份中所含的数据的影响(例如测量方差小的数据)。在一些实施方式中,采用单一加权函数对数据组进行加权,其降低方差大的数据影响并增加方差小的数据的影响。加权函数有时用于减少方差大的数据影响并增加差异小的数据的影响(例如 $[1/(\text{标准偏差})^2]$)。在一些实施方式中,进一步通过加权进行操作从而产生加工的数据的概况图,以便于分类和/或提供结果。可基于加权数据的概况图提供结果。

[0285] 部份的过滤和加权可在分析中一个或多个合适的点进行。例如,部份可在序列读数映射至参照基因组的部份之前或之后进行过滤或加权。在一些实施方式中,部份可在确定个体基因组部份的实验偏差之前或之后进行过滤或加权。在一些实施方式中,部份可在计算基因组区段水平之前或之后进行过滤或加权。

[0286] 在一些实施方式中,数据组经计算、任选过滤、标准化和任选加权后,经加工的数据组可通过一种或多种数学和/或统计学操作(如统计函数或统计算法)进行操作。在某些实施方式中,处理的数据可通过计算一个或多个所选部份、染色体或染色体的部份的Z分数来进一步操作。在一些实施方式中,处理的数据组可通过计算p值来进一步操作。在某些实施方式中,数学和/或统计学操作包括一种或多种与倍数性和/或胎儿分数有关的假设。在一些实施方式中,进一步通过一种或多种数学和/或统计学操作进行操作从而产生加工的数据的概况图,以便于分类和/或提供结果。可基于数学和/或统计学操作数据的概况图提供结果。基于数学和/或统计学操作数据的概况图所提供的结果通常包括一种或多种与倍数性和/或胎儿分数有关的假设。

[0287] 在某些实施方式中,数据组经计算、任选过滤和标准化后,在处理的数据组上进行多种操作以产生N维空间和/或N维点。可基于N维中分析的数据组的概况图提供结果。

[0288] 在一些实施方式中,采用一种或多种峰水平分析、峰宽度分析、峰边缘位置分析、峰侧向公差等或其衍生分析或前述的组合对数据组进行处理,作为已处理和/或操作的数据组的一部分或在其之后。在一些实施方式中,采用一种或多种峰水平分析、峰宽度分析、峰边缘位置分析、峰侧向公差等或其衍生分析或前述的组合产生处理数据的概况图,以便

于分类和/或提供结果。结果可基于数据的概况图提供,所述数据的概况图已采用一种或多种峰水平分析、峰宽度分析、峰边缘位置分析、峰侧向公差等或其衍生分析或前述的组合进行处理。

[0289] 在一些实施方式中,使用基本不含所研究的遗传变异的一种或多种参照样品可用于生成参照中值计数概况,其可得到表示没有遗传变异的预定值,且通常偏离对应于测试对象中定位有遗传变异的基因组位置的区域中的预定值,若所述测试对象具有该遗传变异。在患有关联遗传变异的病症或有此风险的测试对象中,所选部份或区段的数字值预期与未受影响的基因组位置的预定值显著不同。在某些实施方式中,使用已知携带所研究的遗传变异的一种或多种参照样品可用于生成参照中值计数概况,其可得到表示具有遗传变异的预定值,和通常偏离对应于不具有遗传变异的基因组位置的面积中的预定值,其中测试对象在该基因组位置中不具有遗传变异。在不患有关联遗传变异的病症或没有此风险的测试对象中,所选部份或区段的数字值预期与受影响的基因组位置的预定值显著不同。

[0290] 在一些实施方式中,分析和处理数据能包括使用一个或多个假设。合适数量或类型的假设可用于分析或加工数据组。能用于数据处理和/或分析的假设的非限定性示例包括母本倍性、胎儿基值、参照群中某些序列的普遍性、种族背景、相关家族成员中选择的医学病情的患病率、来自不同患者和/或GC标准化和重复掩蔽(如GCRM)后的运行的原始计数分布之间的对应性、代表PCR人造产物的相同匹配(如相同的碱基位置)、胎儿定量试验(如FQA)中的内在假定、关于双胞胎的假定(例如若有2个双胞胎并且仅有1个受影响,则有效胎儿分数仅是全部测量的胎儿分数的50%(三胞胎、四胞胎等也与之相似))、均一覆盖全基因组的胎儿无细胞DNA(如cfDNA)等,及其组合。

[0291] 在经映射的序列读数的质量和/或深度不能在所需置信水平(如95%或更高的置信水平)预测是否存在遗传变异的结果的那些示例中,根据标准化的计数分布,可使用一种或多种额外数学处理算法和/或统计学预测算法以生成可用于数据分析和/或提供结果的额外数值。本文所用的术语“标准化的计数分布”指使用标准化的计数生成的分布。本文描述了能用于生成标准化计数和标准化计数分布的方法示例。可使所述已经计数的定位的序列读数相对于测试样品计数或参照样品计数作标准化。在一些实施方式中,标准化计数概况能以图示表示。

[0292] LOESS标准化

[0293] LOESS是本领域已知的回归模型,其在基于k-最邻近的元模型中组合多种回归模型。LOESS有时指局部加权的多项式回归。在一些实施方式中,GC LOESS应用LOESS模型至参照基因组的部份的GC组合物和片段计数(例如序列读数、计数)之间的关系。用LOESS通过数据点组描绘平滑曲线有时称为LOESS曲线,特别是通过加权的二次最小二乘回归相对于y轴散点图标准变量的值的跨度给定各平滑值时。对于数据组中的各点,LOESS方法拟合低度多项式至数据组,说明变量值接近经评价的响应的点。用加权最小二乘拟合多项式,使得接近经评价的响应的点具有更多权重而远离的点具有更少的权重。然后使用该数据点的说明变量值通过评估局部多项式来获得点的回归函数值。在回归函数值已经针对各数据点计算后,有时完全考虑LOESS拟合。该方法的许多细节,例如多项式模型的度和权重是灵活的。

[0294] PERUN标准化

[0295] 降低核酸指示剂相关误差的标准化方法本文称为参数化误差去除和无偏标准化

(PERUN), 如本文描述和国际专利申请公开号W02013/052913所述, 其全文通过引用纳入本文, 包括所有文本、表格、等式和图。PERUN法可用于各种核酸指示剂(例如核酸序列读数), 用于降低混淆基于该指示剂的预测的误差影响。

[0296] 在某些实施方式中, PERUN方法包括从下述结果计算参照基因组的部份的基因组区段水平: (a) 测试样品映射至参照基因组的部份的序列读数计数, (b) 测试样品的实验偏好(例如GC偏好) 测试样品, 和(c) 针对(i) 序列读数映射的参照基因组部份的实验偏好和(ii) 映射至所述部份序列读数的计数之间的拟合关系的一种或多种拟合参数(例如拟合评价)。可在多种样品中根据针对(i) 映射至各参照基因组部份的序列读数计数和(ii) 参照基因组各部份的映射特征之间的拟合关系来确定各参照基因组的部份的实验偏好。各样品的这种拟合关系可于三维方向就多种样品进行集合。在某些实施方式中, 该集合可根据实验偏好进行安排, 虽然PERUN方法可实施而无需根据实验偏好安排所述集合。通过本领域已知的合适拟合过程, 各样品的拟合关系和参照基因组的各部份的拟合关系可单独拟合至线性函数或非线性函数。

[0297] 杂交回归标准化

[0298] 在一些实施方式中, 使用杂交标准化。在一些实施方式中, 杂交标准化方法降低偏好(例如GC偏好)。在一些实施方式中, 杂交标准化包括(i) 分析双变量(例如计数和GC含量)的关系和(ii) 根据所述分析选择和应用标准化方法。在某些实施方式中, 杂交标准化包括(i) 回归(例如回归分析) 和(ii) 根据所述回归选择和应用标准化方法。在一些实施方式中, 获自第一样品(例如第一组样品)的计数通过与获自其他样品(例如第二组样品)的计数不同的方法标准化。在一些实施方式中, 获自第一样品(例如第一组样品)的计数通过第一标准化方法进行标准化, 获自第二样品(例如第二组样品)的计数通过第二标准化方法进行标准化。例如, 在某些实施方式中第一标准化方法包括使用线性回归而第二标准化方法包括使用非-线性回归(例如LOESS、GC-LOESS、LOWESS回归、LOESS平滑)。

[0299] 在一些实施方式中, 杂交标准化方法用于标准化映射至基因组的部份或染色体的序列读数(例如计数、映射的计数、映射的读数)。在某些实施方式中, 原始计数经标准化, 和在一些实施方式中, 调整、加权、过滤或先前标准化的计数通过杂交标准化方法标准化。在某些实施方式中, 基因组节段水平或Z-分数经标准化。在一些实施方式中, 映射至所选基因组的部份或染色体的计数通过杂交标准化方法经标准化。计数可指映射至基因组的部份的序列读数的合适测量、其非限制性示例包括原始计数(例如未压缩计数)、标准化计数(例如PERUN、ChAI、主成分标准化或合适的方法的标准化)、部份水平(例如平均水平、算术平均水平、中值水平, 或等)、Z-分数、等或其组合。计数可为一种或多种样品(例如测试样品、来自妊娠雌性的样品)的原始计数或处理计数。在一些实施方式中, 计数获自来自一个或多个对象的一个或多个样品。

[0300] 在一些实施方式中, 标准化方法(例如所述类型的标准化方法) 根据回归(例如回归分析) 和/或相关系数进行选择。回归分析指评价变量间(例如计数和GC含量)关系的统计学技术。在一些实施方式中, 根据参照基因组的多种部份中各部份的计数和GC含量测量生成回归。可使用合适的GC含量测量, 其非限制性示例包括测量鸟嘌呤、胞嘧啶、腺嘌呤、胸腺嘧啶、嘌呤(GC) 或嘧啶(AT或ATU) 的含量、解链温度(T_m) (例如变性温度、退火温度、杂交温度)、测量自由能量、等或其组合。测量鸟嘌呤(G), 胞嘧啶(C)、腺嘌呤(A)、胸腺嘧啶(T)、嘌呤

呤(GC),或嘧啶(AT或ATU)含量的测量可表示为比例或百分比。在一些实施方式中,使用任何合适的比例或百分比、其非限制性示例包括GC/AT、GC/总核苷酸、GC/A、GC/T、AT/总核苷酸、AT/GC、AT/G、AT/C、G/A、C/A、G/T、G/A、G/AT、C/T、等或其组合。在一些实施方式中,测量GC含量是GC与总核苷酸含量的比例或百分比。在一些实施方式中测量GC含量是就映射至参照基因组部份的序列读数而言的GC与总核苷酸含量的比例或百分比。在某些实施方式中,根据映射至参照基因组的各部份的序列读数和/或从其中确定GC含量,且序列读数获自样品(例如样品获自妊娠雌性)。在一些实施方式中,GC含量测量不是根据序列读数和/或从中确定。在某些实施方式中,就获自一个或多个对象的一个或多个样品确定GC含量测量。

[0301] 在一些实施方式中,生成回归包括生成回归分析或相关性分析。可使用合适的回归,其非限制性示例包括回归分析、(例如线性回归分析),拟合优度分析,Pearson相关性分析,分级相关性,未解释的方差组分、Nash-Sutcliffe模型有效性分析、回归模型验证、比例减少损失、均方根差、等或其组合。在一些实施方式中生成回归线。在某些实施方式中生成回归包括生成线性回归。在某些实施方式中生成回归包括生成非-线性回归(例如LOESS回归、LOWESS回归)。

[0302] 在一些实施方式中,回归确定是否存在相关性(例如线性相关性),例如计数和GC含量测量之间的相关性。在一些实施方式中生成回归(例如线性回归)并确定相关系数。在一些实施方式中,确定合适的相关系数,其非限制性示例包括确定系数、R2值、Pearson相关系数等。

[0303] 在一些实施方式中,确定回归(例如回归分析线性回归)的拟合优度。拟合优度有时通过观察或数学分析确定。评估有时包括确定非-线性回归或线性回归的拟合优度是否更大。在一些实施方式中,相关系数是拟合优度的量度。在一些实施方式中评估回归的适合度根据相关系数和/或相关系数截止值确定。在一些实施方式中,拟合优度评估包括比较相关系数与相关系数截止值。在一些实施方式中,评估回归的拟合优度指示线性回归。例如,在某些实施方式中,线性回归的拟合优度大于非-线性回归的拟合优度,且所述拟合优度评估指示线性回归。在一些实施方式中评估指示线性回归且线性回归用于标准化计数。在一些实施方式中,评估回归的拟合优度指示非线性回归。例如,在某些实施方式中,非-线性回归的拟合优度大于线性回归的拟合优度,且所述拟合优度评估指示非线性回归。在一些实施方式中,评估指示非线性回归且非线性回归用于标准化计数。

[0304] 在一些实施方式中,当相关系数等于或大于相关系数截止值,拟合优度评估指示线性回归。在一些实施方式中,当相关系数小于相关系数截止值,拟合优度评估指示非线性回归。在一些实施方式中,相关系数截止值预先确定。在一些实施方式中相关系数截止值为约0.5或更大、约0.55或更大、约0.6或更大、约0.65或更大、约0.7或更大、约0.75或更大、约0.8或更大或约0.85或更大。

[0305] 例如,在某些实施方式中,相关系数等于或大于约0.6时,使用包括线性回归的标准化方法。在某些实施方式中,当相关系数等于或大于相关系数截止值0.6时,样品计数(例如参照基因组的每部份的计数、每部份的计数)根据线性回归经标准化,否则计数根据非-线性回归经标准化(例如当系数小于相关系数截止值0.6时)。在一些实施方式中,标准化过程包括就参照基因组的多个部份的各部份的(i)计数和(ii)GC含量生成线性回归或非-线性回归。在某些实施方式中,当相关系数小于相关系数截止值0.6时,使用包括非-线性回归

(例如LOWESS, LOESS)的标准化方法。在一些实施方式中,当相关系数(例如相关系数)小于约0.7、小于约0.65、小于约0.6、小于约0.55或小于约0.5的相关系数截止值时,使用包括非-线性回归(例如LOWESS)标准化方法。例如在一些实施方式中,当相关系数小于约0.6的相关系数截止值时,使用包括非-线性回归(例如LOWESS, LOESS)的标准化方法。

[0306] 在一些实施方式中,选择特异性类型的回归(例如线性或非-线性回归),并且在生成回归后,通过从计数中减去回归来标准化所述计数。在一些实施方式中,从计数中减去回归提供偏好(例如GC偏好)降低的标准化计数。在一些实施方式中从计数中减去线性回归。在一些实施方式中从计数中减去非-线性回归(例如LOESS、GC-LOESS、LOWESS回归)。任何合适的方法可用于从计数中减去回归线。例如,若计数 x 源自包括0.5的GC含量的部份 i (例如部份 i)且回归线确定GC含量0.5处的计数 y ,则 $x-y$ =部份 i 的标准化计数。在一些实施方式中,减去回归之前和/或之后标准化计数。在一些实施方式中,杂交标准化方法标准化的计数用于生成基因组节段水平, Z -分数、基因组或其区段的水平和/或概况。在某些实施方式中,杂交标准化方法标准化的计数通过本文所述方法分析以确定是否存在遗传变异(例如胎儿中)。

[0307] 在一些实施方式中杂交标准化方法包括标准化之前或之后过滤或加权一个或多个部份。可用本文所述的合适方法过滤部份,包括过滤部份(例如参照基因组的部份)的方法。在一些实施方式中,应用杂交标准化方法之前过滤部份(例如参照基因组的部份)。在一些实施方式中,仅映射至所选部份(例如根据计数可变性选择的部份)的测序读数计数通过杂交标准化进行标准化。在一些实施方式中,映射至参照基因组的过滤部份(例如根据计数可变性过滤的部份)的测序读数计数在使用杂交标准化方法之前去除。在一些实施方式中,杂交标准化方法包括根据合适的方法(例如本文所述方法)选择或过滤部份(例如参照基因组的部份)。在一些实施方式中杂交标准化方法包括根据映射至多种测试样品的各部份的计数的不确定值选择或过滤部份(例如参照基因组的部份)。在一些实施方式中杂交标准化方法包括根据计数可变性选择或过滤部份(例如参照基因组的部份)。在一些实施方式中杂交标准化方法包括根据GC含量、重复元件、重复序列、内含子、外显子、等或其组合选择或过滤部份(例如参照基因组的部份)。

[0308] 例如,在一些实施方式中,分析多个妊娠雌性对象的多个样品且根据计数可变性选择部份子集(例如参照基因组的部份)。在某些实施方式中线性回归用于确定获自妊娠雌性对象的样品的各所选部份的(i)计数和(ii)GC含量的相关系数。在一些实施方式中,确定大于预定相关性截止值(例如约0.6)的相关系数,拟合优度评估指示线性回归并通过从计数中减去线性回归来标准化所述计数。在某些实施方式中,确定小于预定相关性截止值(例如约0.6)的相关系数,拟合优度评估指示非线性回归,生成LOESS回归,并通过从计数中减去LOESS回归来标准化所述计数。

[0309] 概况

[0310] 在一些实施方式中,处理步骤可包括从各种数据组或其衍生(例如本领域已知和/或本文所述的一种或多种数学和/或统计学数据处理步骤的结果)中生成一种或多种概况(例如概况图)。

[0311] 本文术语“概况”指数据的数学和/或统计学操作的结果,其可便于在大量数据中鉴定模式和/或相关性。“概况”通常包括基于一种或多种标准对数据或数据组的一种或多

种操作所得的值。概况通常包括多个数据点。任何合适数量的数据点可包括在概况中,取决于数据组的性质和/或复杂度。在某些实施方式中,概况可包括2或更多数据点、3或更多数据点、5或更多数据点、10或更多数据点、24或更多数据点、25或更多数据点、50或更多数据点、100或更多数据点、500或更多数据点、1000或更多数据点、5000或更多数据点、10,000或更多数据点,或100,000或更多数据点。

[0312] 在一些实施方式中,概况是整个数据组的表示,和在某些实施方式中,概况是数据组部份或子集的表示。即,概况有时包括未过滤移除任何数据的数据的数据点代表或从其中生成,和有时概况包括经过滤移除不想要数据的数据的数据点代表或从其中生成。在一些实施方式中,概况中数据点代表部份的数据操作结果。在某些实施方式中,概况中数据点包括部份组的数据操作结果。在一些实施方式中,部份组可彼此相邻,和在某些实施方式中,部份组可来自染色体或基因组的不同部份。

[0313] 源自数据组的概况中的数据点可表示任何合适的数据分类。数据可分组以生成概况数据点的类别的非限制性示例包括:基于大小的部份、基于序列特征的部份(例如GC含量、AT含量、染色体上的位置(例如短臂、长臂、着丝粒、端粒)等)、表达水平、染色体等或其组合。在一些实施方式中,概况可从获自其他概况(例如再次标准化至不同标准化值的标准化数据概况以生成再标准化数据概况)的数据点生成。在某些实施方式中,获自其他概况的数据点生成的概况减少了数据点数量和/或数据组的复杂性。减少了数据点数量和/或数据组的复杂性通常有利于解释数据和/或有利于提供结果。

[0314] 概况(例如基因组概况、染色体概况、染色体区段概况)通常是两个或更多部份的标准化或非-标准化计数的集合。概况通常包括至少一种水平(例如基因组区段水平),通常包括两个或更多水平(例如概况通常具有多种水平)。水平通常用于具有约相同计数或标准化计数的部份的组。本文详细描述了水平。在某些实施方式中,概况包括一种或多种部份、所述部份可经加权、移除、过滤、标准化、调整、平均(得出均值)、加、减、或其任意组合的加工或变换。概况通常包括映射至定义两个或更多水平的部份的标准化计数,其中计数还通过合适方法根据水平之一进一步标准化。通常概况计数(例如概况水平)关联不确定值。

[0315] 包括一种或多种水平的概况有时经填充(例如孔填充)。填充(例如孔填充)指鉴定和调整概况中源自母本微缺失或母本复制(例如拷贝数变异)的水平的过程。在一些实施方式中,填充源自胎儿微复制或胎儿微缺失的水平。在一些实施方式中,概况中微复制或微缺失可人为升高或降低概况(例如染色体概况)的总体水平,导致染色体非整倍性(例如三体性)确定的假阳性或假阴性。在一些实施方式中,概况中源自微复制和/或缺失的水平通过有时称为填充或孔填充的过程鉴定和调整(例如填充和/或移除)。在某些实施方式中,概况包括明显不同于概况内第二水平的一种或多种第一水平,各所述一种或多种第一水平包括母本拷贝数变异、胎儿拷贝数变异,或母本拷贝数变异和胎儿拷贝数变异,且一种或多种所述第一水平经调整。

[0316] 包括一种或多种水平的概况可包括第一水平和第二水平。在一些实施方式中,第一水平不同于(例如明显不同)第二水平。在一些实施方式中第一水平包括第一组部份、第二水平包括第二组部份且第一组部份不是第二组部份的子集。在某些实施方式中,第一组部份不同于第二组部份,从中确定第一和第二水平。在一些实施方式中,概况可具有不同于(例如明显不同,例如具有显著不同的值)概况内第二水平的多种第一水平。在一些实施方

式中概况包括明显不同于概况内第二水平的一种或多种第一水平且所述一种或多种第一水平经调整。在一些实施方式中,概况包括明显不同于概况内第二水平的一种或多种第一水平,各所述一种或多种第一水平包括母本拷贝数变异、胎儿拷贝数变异,或母本拷贝数变异和胎儿拷贝数变异,且一种或多种所述第一水平经调整。在一些实施方式中概况中的第一水平从概况中去除或经调整(例如填充)。概况可包括多种水平,所述多种水平包括显著不同于一种或多种第二水平的一种或多种第一水平,通常概况中主要水平为第二水平,其中第二水平彼此大约相等。在一些实施方式中,概况中大于50%、大于60%、大于70%、大于80%、大于90%或大于95%的水平为第二水平。

[0317] 概况有时显示为图。例如,表示部份的计数(例如标准化计数)的一种或多种水平可作图并可可视化。可生成的概况图非限制性示例包括原始计数(例如原始计数概况或原始概况)、标准化计数、部份-加权的、Z-分数、p-值、面积比与拟合倍性、中值水平与拟合和测量的胎儿分数之间的比例、主成分、等或其组合。在一些实施方式中,概况图允许观察经操作的数据。某些实施方式中,概况图可用于提供结果(例如面积比与拟合倍性、中值水平与拟合和测量的胎儿分数之间的比例、主成分)。本文所用术语“原始计数概况图”或“原始概况图”指标准化至区域总计数的区域各部份中的计数的图(例如基因组、部份、染色体、参照基因组的染色体部份或染色体区段)。在一些实施方式中,可使用静态窗过程生成概况,和在某些实施方式中,可使用滑动窗过程生成概况。

[0318] 针对测试对象生成的概况有时与一种或多种参照对象生成的概况进行比较,以便于阐述数据组的数学和/或统计学操作和/或提供结果。在一些实施方式中,基于一种或多种起始假设(例如母本核酸贡献(例如母本总积分)、胎儿核酸贡献(例如胎儿分数)、参照样品倍性等或其组合)生成概况。在某些实施方式中,测试概况通常以表示不含遗传变异的预定值为中心、和通常在测试对象中定位有遗传变异(若测试对象具有遗传变异)的基因组位置的对应面积中偏离预定值。在患有关联遗传变异的病症或有此风险的测试对象中,所选部份的数字值预期与未受影响的基因组位置的预定值显著不同。基于起始假设(例如固定的倍性或最佳倍性、固定的胎儿分数或最佳胎儿分数或其组合),表明是否存在遗传变异的预定阈值或截止值或阈值范围可不同,但其仍提供可用于确定是否存在遗传变异的结果。在一些实施方式中,概况指示和/或表示表型。

[0319] 作为非限制性示例,标准化的样品和/或参照计数概况可通过如下方式从原始序列读数数据获得:(a)从已知不携带遗传变异的参照组中计算选定的染色体、部份或其区段的参照中值计数、(b)从所述参照样品原始计数中移除不带信息的部份(例如,滤除);(c)使参照基因组的全部剩余部份针对参照样品选定的染色体或选定的基因组位置的计数的剩余的总数(例如,在移除参照基因组的不带信息的部份之后的余留计数之和)进行标准化,由此产生标准化的参照对象概况;(d)从测试对象样品移除相应的部份;和(e)使一个或多个选定的基因组位置的余留的测试对象计数针对包含所述选定基因组位置的染色体的剩余参照中值计数之和进行标准化,由此产生标准化的测试对象概况。在某些实施方式中,涉及整个基因组的其他标准化步骤(通过(b)中的过滤的部份而减少)可包括在(c)和(d)之间。

[0320] 能通过对计数映射的序列读数数据的一种或多种处理生成数据组概况。一些实施方式包括下述。映射序列读数,并确定映射到各基因组部份上的序列标签数目(如计数)。从

计数的映射的序列读数生成原始计数概况。在某些实施方式中,通过比较测试对象的原始计数概况与已知不含遗传变异的参照对象组的染色体、其部份或区段的参照中值计数概况提供结果。

[0321] 在一些实施方式中,序列读数数据任选经过滤以去除噪音数据或无信息部份。过滤后,通常对剩余计数求和以生成过滤的数据组。在某些实施方式中,从过滤的数据组生成过滤的计数概况。

[0322] 序列读数数据经计数和任选过滤后,可标准化数据组以生成水平或概况。可通过标准化一种或多种所选部份至合适的标准化参照值来标准化数据组。在一些实施方式中,标准化参照值表示从中选择部份的染色体的总计数。在某些实施方式中,标准化参照值表示已知不含遗传变异的参照对象组制备的参照数据组的染色体的一种或多种相应部份。在一些实施方式中,标准化参照值表示分析是否存在遗传变异的测试对象制备的测试对象数据组的染色体的一种或多种相应部份。在某些实施方式中,标准化过程使用静态窗法进行、和在一些实施方式中,标准化过程使用移动或滑动窗法进行。在某些实施方式中,生成包括标准化计数的概况便于分类和/或提供结果。结果可基于包括标准化计数的概况图(例如使用该概况图)提供。

[0323] 水平

[0324] 在一些实施方式中,值(例如数值、定量值)归属为水平。计数可通过合适的方法、操作或数学过程(例如加工的水平)确定。水平通常是或源自部份的组的计数(例如标准化计数)。在一些实施方式中部份的水平基本等于映射至部份的计数总数(例如计数,标准化计数)。通常从本领域已知的合适的方法、操作或数学过程加工、变换或处理的计数来确定水平。在一些实施方式中,水平源自加工的计数,加工的计数的非限制性示例包括加权、去除、过滤、标准化、调整、平均、得出算术平均(例如算术平均水平)、加、减、变换的计数或其组合。在一些实施方式中水平包括标准化的计数(例如部份的标准化计数)。水平可通过合适的过程用于计数标准化,其非限制性示例包括逐份标准化、GC含量的标准化、中值计数标准化、线性和非线性最小二乘回归、LOESS(例如,GC LOESS)、LOWESS、PERUN、ChAI、主成分标准化、RM、GCRM、cQn等和/或其组合。水平可包括标准化计数或计数的相对量。在一些实施方式中,水平用于经平均的两个或更多部份的计数或标准化计数且所述水平指平均水平。在一些实施方式中水平用于具有标准化计数的算术平均的计数或算术平均的部份的组,其称为算术平均水平。在一些实施方式中就包括原始和/或过滤的计数部份得出水平。在一些实施方式中,水平基于原始计数。在一些实施方式中,水平关联不确定值(例如标准偏差, MAD)。在一些实施方式中,水平由Z-分数或p-值代表。本文中一个或多个部份的水平与“基因组区段水平”同义。

[0325] 本文中一个或多个部份的水平与“基因组区段水平”同义。有时,本文中所用的术语“水平”与术语“标高”同义。术语“水平”的含义的确定可由其上下文来确定。例如,术语“水平”,在用于基因组区段、概况、读数和/或计数的内容中时,通常表示标高。术语“水平”,在用于物质或组分(例如, RNA水平、丛水平)的内容中时,通常表示量。术语“水平”,在用于不确定性(例如,错误水平、置信水平、偏差水平、不确定性水平)的内容中时,通常表示量。

[0326] 两个或更多水平(例如概况中两个或更多水平)的标准化或非-标准化计数有时可根据水平经数学操作(例如加、乘、平均、标准化等或其组合)。例如,两个或更多水平的标准

化或非-标准化计数可根据概况中一个、一些或全部水平进行标准化。在一些实施方式中,概况中所有水平的标准化或非-标准化计数根据概况中一个水平进行标准化。在一些实施方式中,概况中第一水平的标准化或非-标准化计数根据概况中第二水平的标准化或非-标准化计数进行标准化。

[0327] 水平(例如第一水平,第二水平)的非限制性示例是包括加工的计数的部份的组水平、包括计数的算术平均、中值或平均的部份的组的水平、包括标准化计数的部份的组的水平等或其任何组合。在一些实施方式中,概况中第一水平和第二水平源自映射至相同染色体的部份的计数。在一些实施方式中,概况中第一水平和第二水平源自映射至不同染色体的部份的计数。

[0328] 在一些实施方式中水平从映射至一个或多个部份的标准化或非-标准化计数确定。在一些实施方式中,水平从映射至两个或更多部份的标准化或非-标准化计数确定,其中各部份的标准化计数通常大约相同。就一水平而言,部份的组中的计数(例如标准化计数)可存在差异。就一水平而言,部份的组中可存在具有与所述组的其他部份(例如峰值和/或倾斜)显著不同的计数的一个或多个部份。关联任何合适数量的部份的任何合适数量的标准化或非-标准化计数可定义水平。

[0329] 在一些实施方式中,可从所有或一些基因组的部份的标准化或非-标准化计数中确定一种或多种水平。通常,可从染色体或其区段的所有或一些标准化或非-标准化计数确定水平。在一些实施方式中,源自两个或更多部份(例如部份的组)的两个或更多计数确定水平。在一些实施方式中,(例如来自两个或更多部份的计数)两个或更多计数确定水平。在一些实施方式中,2-约100,000部份的计数确定水平。在一些实施方式中,2-约50,000、2-约40,000、2-约30,000、2-约20,000、2-约10,000、2-约5000、2-约2500、2-约1250、2-约1000、2-约500、2-约250、2-约100或2-约60部份的计数确定水平。在一些实施方式中,约10-约50部份的计数确定水平。在一些实施方式中,约20-约40或更多部份的计数确定水平。在一些实施方式中,水平包括来自约2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29、30、31、32、33、34、35、36、37、38、39、40、45、50、55、60或更多部份的计数。在一些实施方式中,水平对应于部份的组(例如参照基因组的部份的组、染色体部份的组或染色体区段部份的组)。

[0330] 在一些实施方式中,就邻近部份的标准化或非-标准化计数确定水平。在一些实施方式中,邻近部份(例如部份的组)代表基因组的相邻区段或染色体或基因的相邻区段。例如,通过尾对尾合并部份时,其的两个或更多邻近部份可代表比各部份更长的DNA序列的序列集合。例如,两个或更多邻近部份可代表整个基因组、染色体、基因、内含子、外显子或其区段。在一些实施方式中,从邻近部份和/或非-邻近部份的集合(例如组)中确定水平。

[0331] 结果

[0332] 本文所述方法可提供确定样品是否存在遗传变异(例如胎儿非整倍性),从而提供结果(例如从而提供是否存在遗传变异(例如胎儿非整倍性)的确定性结果)。遗传变异通常包括遗传信息(例如染色体、染色体区段、多态性区域、转位区域、改变的核苷酸序列等或前述的组合)的获得、丧失和/或改变(例如复制、缺失、融合、插入、突变、重组、取代或异常甲基化),其导致测试对象的基因组或遗传信息中相对参照的可检测变化。是否存在遗传变异可通过变换、分析和/或操作映射至部份(例如计数、参照基因组的基因组部份的计数)的序

列读数来确定。在一些实施方式中,确定结果包括分析妊娠雌性的核酸。在某些实施方式中,结果根据获自妊娠雌性的计数(例如标准化计数、读数密度、读数密度概况)确定,其中所述计数获自妊娠雌性的核酸。

[0333] 本文所述方法有时确定怀有胎儿的妊娠雌性的测试样品是否存在胎儿非整倍性(例如全染色体非整倍性、部分染色体非整倍性或区段性染色体异常(例如镶嵌性、缺失和/或插入))。在某些实施方式中本文所述方法测试怀有胎儿的妊娠雌性的测试样品的整倍体性或无整倍体性(非-整倍体性)。本文所述方法有时测试一种或多种染色体(例如染色体13、染色体18、染色体21或其组合)或其区段的三体性。

[0334] 在一些实施方式中,通过本文所述方法、本领域已知方法或其组合确定是否存在遗传变异(例如胎儿非整倍性)。是否存在遗传变异通常测自映射至参照基因组的部份的序列读数计数。

[0335] 来自参照的读数密度有时用于相同妊娠雌性的核酸样品,测试样品获自该妊娠雌性。在某些实施方式中,来自参照的读数密度用于一种或多种妊娠雌性的核酸样品,所述一种或多种妊娠雌性不同于获取测试样品的雌性。在一些实施方式中,来自测试对象的第一组部份的读数密度和/或读数密度概况与来自第二组部份的读数密度和/或读数密度概况比较,其中第二组部份与第一组部份不同。在一些实施方式中,来自测试对象的第一组部份的读数密度和/或读数密度概况与来自第二组部份的读数密度和/或读数密度概况比较,其中第二组部份来自测试对象或来自非测试对象的参照对象。在非限制性示例中,第一组部份在染色体21或其区段中,第二组部份通常在其他染色体中(例如染色体1,染色体13,染色体14,染色体18,染色体19,其区段或其组合)。参照通常位于一般为整倍体的染色体或其区段中。例如,胎儿中染色体1和染色体19通常为整倍体,由于染色体1和染色体19非整倍性与高比例的早期胎儿死亡相关。可生成和/或比较来自测试对象和参照的读数密度和/或读数密度概况之间的不确定性测量。确定是否存在遗传变异(例如胎儿非整倍性)有时无需比较来自测试对象的读数密度和/或读数密度概况与参照。

[0336] 在某些实施方式中,参照包含与测试对象相同的部份的组的读数密度和/或读数概况,其中参照的读数密度来自一个或多个参照样品(例如通常来自多个参照对象的多个参照样品)。参照样品通常来自不同于获得测试样品的雌性的一个或多个妊娠雌性。

[0337] 可生成测试对象和/或参照的读数密度和/或读数概况的不确定性测量。在一些实施方式中,测定测试对象的读数密度和/或读数概况的不确定性测量。在一些实施方式中,测定参照对象的读数密度和/或读数概况的不确定性测量。在一些实施方式中,从整个读数密度概况或读数密度概况中的部份子集确定不确定性测量。

[0338] 在一些实施方式中,参照样品对于基因组的选择区段是整倍性,并且评价选择区段的测试概况和参照概况之间不确定性的测量。在一些实施方式中确定是否存在遗传变异是根据基因组的选择区段(例如,染色体,或其区段)的测试概况和参照概况之间偏差的数量(例如偏差的测量,MAD)。在一些实施方式中,当测试概况和参照概况之间偏差的数量大于约1、大于约1.5、大于约2、大于约2.5、大于约2.6、大于约2.7、大于约2.8、大于约2.9、大于约3、大于约3.1、大于约3.2、大于约3.3、大于约3.4、大于约3.5、大于约4、大于约5或大于约6时,确定存在遗传变异。例如,有时测试概况与参照概况差异大于3量度的偏差时(例如 3σ , 3MAD),确定存在遗传变异。在一些实施方式中获自妊娠雌性的测试概况大于参照概况超

过3量度的偏差时(例如 3σ 、 $3MAD$)，确定存在胎儿染色体非整倍性(例如胎儿三体性)。测试概况和参照概况之间超过3的偏差通常指示基因组的选择区段的非整倍性测试对象(例如，存在遗传变异)。有时，明显大于基因组的选择区段的参照概况的测试概况确定三体性，其参照对于选择区段是整倍性。在一些实施方式中，针对选择区段，获自妊娠雌性的读数密度概况小于参照概况超过3量度的偏差时(例如 3σ 、 $3MAD$)，确定存在胎儿染色体非整倍性(例如胎儿单体性)。测试概况显著低于参照概况(参照概况指示整倍体性)有时决定了单体性。

[0339] 在一些实施方式中，针对基因组的选择区段，当测试概况和参照概况之间偏差的数量小于约3.5、小于约3.4、小于约3.3、小于约3.2、小于约3.1、小于约3.0、小于约2.9、小于约2.8、小于约2.7、小于约2.6、小于约2.5、小于约2.0、小于约1.5或小于约1.0，则确定不存在遗传变异。例如，有时测试概况与参照概况差异小于3量度的偏差时(例如 3σ 、 $3MAD$)，确定不存在遗传变异。在一些实施方式中获自妊娠雌性的测试概况与参照概况的差异小于3量度的偏差时(例如 3σ 、 $3MAD$)，确定不存在胎儿染色体非整倍性(例如胎儿整倍性)。在一些实施方式中，(例如测试概况和参照概况之间偏差小于3(例如标准偏差的 3σ)通常指示基因组的区段是整倍体(例如不存在遗传变异)。测试样品的测试概况和一种或多种参照对象参照概况之间的偏差可作图并可可视化(例如Z-分数作图)。

[0340] 可用测试概况将任何其他合适的参照化为因数，用于确定测试样品的测试区域(例如，测试的基因组的区段)是否存在遗传变异(确定整倍体或非-整倍体)。在一些实施方式中，胎儿分数确定可用序列读数的计数(例如，读数密度)因数化以确定是否存在遗传变异。例如，可根据胎儿分数在比较和/或确定结果之前标准化读数密度和/或读数密度概况。可采用合适的过程定量胎儿分数，其非限制性示例包括质谱法、测序过程或其组合。

[0341] 在一些实施方式中，根据判定区确定是否存在遗传变异(例如，胎儿非整倍性)。在某些实施方式中，当值(例如，读数密度概况和/或不确定性测量)或值的集合落在预定范围内(例如区，判定区)，做出判定(例如确定是否存在遗传变异的判定，例如结果)。在一些实施方式中，根据获自相同患者样品的值(例如，读数密度概况和/或不确定性测量)的集合定义判定区。在某些实施方式中，根据获自相同染色体或其区段的值的集合定义判定区。在一些实施方式中，根据不确定性测量(例如高置信水平、例如低不确定性测量)和/或胎儿分数定义基于遗传变异确定的判定区。

[0342] 在一些实施方式中，根据约2.0%或更多、约2.5%或更多、约3%或更多、约3.25%或更多、约3.5%或更多、约3.75%或更多或约4.0%或更多的遗传变异和胎儿分数的确定定义判定区。例如，在一些实施方式中，基于测试概况与参照概况的比较判定胎儿包含21三体，其中对于获自怀有胎儿的妊娠雌性的测试样品，测试概况所衍生的测试样品包含2%或更大或4%或更大的胎儿分数确定。例如，在一些实施方式中，基于测试概况与参照概况的比较判定胎儿是整倍性，其中对于获自怀有胎儿的妊娠雌性的测试样品，测试概况所衍生的测试样品包含2%或更大或4%或更大的胎儿分数确定。在一些实施方式中，通过约99%或更大、约99.1%或更大、约99.2%或更大、约99.3%或更大、约99.4%或更大、约99.5%或更大、约99.6%或更大、约99.7%或更大、约99.8%或更大或约99.9%或更大的置信水平定义判定区。一些实施方式中不使用判定区进行判定。在一些实施方式中使用判定区和其他数据或信息做出判定。在一些实施方式中，基于比较而不使用判定区做出判定。在一些实施方式中，基于概况的视觉观察(例如，视觉观察读数密度)进行判定。

[0343] 在一些实施方式中非判定区是不进行判定的区。在一些实施方式中,通过表明低精确性、高风险、高误差、低水平置信度、高不确定性测量等或其组合的值或值的集合定义非-判定区。在一些实施方式中,部份通过约5%或更少、约4%或更少、约3%或更少、约2.5%或更少、约2.0%或更少、约1.5%或更少或约1.0%或更少的胎儿分数定义非-判定区。

[0344] 遗传变异有时关联医学病症。遗传变异的确定性结果有时是是否具有病症(例如医学病症)、疾病、症状或异常,或包括检测病症、疾病、症状或异常的确定性结果(非限制性示例列于表1)。在某些实施方式中,诊断包括评估结果。通过本文所述方法对是否存在病症(例如医学病症)、疾病、症状或异常的结果确定有时可通过其他测试(例如染色体组型和/或羊水诊断)单独验证。数据的分析和处理能提供一种或多种结果。本文术语“结果”可指有利于确定是否存在遗传变异(例如非整倍性、拷贝数量差异)的数据处理结果。在一些实施方式中,本文术语“结果”可指预测和/或确定是否存在遗传变异(例如非整倍性、拷贝数量差异)的结论。在一些实施方式中,本文术语“结果”可指预测和/或确定对象(例如胎儿)是否存在遗传变异(例如非整倍性、拷贝数量差异)的风险或可能性的结论。诊断有时包括使用结果。例如,健康医师可分析结果并基于或部份给予该结果提供诊断。在一些实施方式中,确定、检测或诊断病症、症状或异常(例如表1所列)包括使用是否存在遗传变异的确定性结果。在一些实施方式中,基于计数的映射的序列读数或其变换的结果确定是否存在遗传变异。在某些实施方式中,使用本文所述一种或多种数据处理方法生成的结果确定是否存在表1中列举的一种或多种病症、症状或异常。在一些实施方式中,诊断包括确定是否存在病症、症状或异常。通常诊断包括确定作为病症、症状或异常的性质和/或原因的遗传变异。在一些实施方式中,结果不是诊断。结果常包括在关于概率的一种或多种考虑的前提下使用本文所述的处理方法生成的一个或多个数值。风险或概率的考虑可包括但不限于不确定性测量、置信水平、灵敏度、特异性、标准偏差、方差系数(CV)和/或置信水平、Z-分数、Chi值、Phi值、倍数性值、拟合胎儿分数、面积比率、中值水平、等或其组合。关于概率的考虑能帮助确定对象是否有遗传变异的风险或有遗传变异,而是否存在遗传疾病的确定性结果常包括这种考虑。

[0345] 结果有时是表型。结果经常是有相关置信水平的表型(如不确定性测量,胎儿21三体性呈阳性,其置信水平为99%;测试对象为遗传变异相关癌症阴性,其置信水平为95%)。产生结果值的不同方法有时可产生不同类型结果。通常,基于使用本文所述方法生成的结果值有四种可能的评分或判定:真阳性、假阳性、真阴性和假阴性。本文所用的术语“评分”、“分数”、“判定”指就对象/样品中是否存在特定遗传变异的概率的计算。分值可用来确定,例如,可与遗传变异相对应的定位的序列读数的变化、差异或比例。例如,关于参照基因组,对数据组的所选遗传变异或部份计正分能引导鉴定是否存在遗传变异,所述遗传变异有时与医学病症相关联(如癌症、先兆子痫、三体性、单体性等)。在一些实施方式中,结果包括读数密度、读数概况和/或图(如概况图)。在那些结果包括概况的那些实施方式中,可就结果使用合适的概况或概况组合。能就结果使用的概况的非限定性示例包含z-分数概况、p值概况、x值概况、 ϕ 值概况等,及其组合。

[0346] 用于确定是否存在遗传变异而生成的结果有时包括零(null)结果(如两个簇之间的数据点,包括存在和缺失的遗传变异的值的有标准偏差的数值,分布图与有或没有所研

究的遗传变异的对象的分布图不相似的数据组)。在一些实施方式中,指示零结果的结果仍然是确定结果,而所述确定可包含以供确定是否存在遗传变异的对其它信息和/或重复数据生成和/或分析的需要。

[0347] 在一些实施方式中,在进行本文所述一个或多个处理步骤后能生成结果。在某些实施方式中,结果作为本文所述处理步骤之一的结果生成,而在一些实施方式中,在数据组的各统计学和/或数学运算进行之后,能生成结果。关于确定是否存在遗传变异的结果能以任意形式表示,所述形式包括但不限于与对象或样品中遗传变异存在与否相关的概率(如让步比、p值)、可能性、簇内或簇外值、过阈或阈下值、范围内(如阈值范围)的值、有方差或置信测量的值或风险因子。在某些实施方式中,样品间的比较允许确定样品特征(例如允许鉴定重复样品和/或已混合的样品(例如错标的、组合的等))。

[0348] 在一些实施方式中,结果包括高于或低于预定阈值或截止值的值和/或与所述值相关联的置信水平或不确定性测量值。在某些实施方式中,预定阈值或截止值是预期水平或预期水平范围。结果也能描述用于数据处理的假定。在某些实施方式中,结果包含落在值预定范围(如阈值范围)内或外的值,和该值在所述范围内或外的相关不确定性或置信水平。在一些实施方式中,结果包含的值等于预定值(如等于1、等于0)等于预定值范围内的值,以及其等于或者在范围内或外的相关不确定性或置信水平。结果有时以图(如分布图)来图示显示。

[0349] 如上所述,结果能表征为真阳性、真阴性、假阳性或假阴性。本文所用的术语“真阳性”指对象被正确诊断为具有遗传变异。如本文所用术语“假阳性”指对象被错误鉴定为具有遗传变异。如本文所用术语“真阴性”指对象被正确鉴定为不具有遗传变异。如本文所用术语“假阴性”指对象被错误鉴定为不具有遗传变异。可根据发生比例对任何给定方法计算两种性能度量:(i) 灵敏度值,通常是被正确鉴定为阳性的预测阳性部份;和(ii) 特异性值,通常是被正确鉴定为阴性的预测阴性部份。

[0350] 在某些实施方式中,灵敏度、特异性和/或置信水平中的一个或多个表示为百分数。在一些实施方式中,独立地对应各变量的百分数超过约90%(例如,约90、91、92、93、94、95、96、97、98或99%或超过99%(例如,约99.5%或更高,约99.9%或更高、约99.95%或更高、约99.99%或更高))。在一些实施方式中,变异系数(CV)表示为百分数,有时所述百分数为约10%或更低(例如,约10、9、8、7、6、5、4、3、2或1%或低于1%(例如,约0.5%或更低、约0.1%或更低、约0.05%或更低、约0.01%或更低))。在某些实施方式中,概率(如特定结果不是由于偶然)被表示为Z-分值、p值或t检验的结果。在一些实施方式中,可用本文所述的一种或多种数据处理操作生成针对结果的测量的方差、置信区间、灵敏度、特异性等(例如统称为置信参数)。生成结果和相关置信水平的具体示例描述于实施例部分以及国际申请号PCT/US12/59123(WO2013/052913),其全文通过引用纳入本文,包括所有文本、表格、等式和图。

[0351] 本文所用的术语“灵敏度”指真阳性的数量除以真阳性数量与假阴性数量之和,其中灵敏度(sens)可在 $0 \leq \text{sens} \leq 1$ 范围内。本文所用的术语“特异性”指真阴性的数量除以真阴性数量与假阴性数量之和,其中特异性(spec)可在 $0 \leq \text{spec} \leq 1$ 范围内。在一些实施方式中,有时选择灵敏度和特特异性等于1、或100%、或接近1(如约90%-约99%)的方法。在一些实施方式中,选择灵敏度等于1或100%的方法,而在某些实施方式中,选择灵敏度接近1

(如灵敏度约90%、灵敏度约91%、灵敏度约92%、灵敏度约93%、灵敏度约94%、灵敏度约95%、灵敏度约96%、灵敏度约97%、灵敏度约98%或灵敏度约99%)的方法。在一些实施方式中,选择特异性等于1或100%的方法,而在某些实施方式中,选择特异性接近1(如特异性约90%、特异性约91%、特异性约92%、特异性约93%、特异性约94%、特异性约95%、特异性约96%、特异性约97%、特异性约98%或特异性约99%)的方法。

[0352] 在一些实施方式中,确定胎儿是否存在遗传变异(例如染色体非整倍性)。在该实施方式中,确定是否存在胎儿遗传变异(例如胎儿染色体非整倍性)。

[0353] 在某些实施方式中,确定样品是否存在遗传变异(例如染色体非整倍性)。在该实施方式中,确定样品核酸中是否存在遗传变异(例如染色体非整倍性)。在一些实施方式中,检测或未检测的变异在一个来源的样品核酸中残留而在另一来源的样品核酸中不残留。来源的非限制性示例包括胎盘核酸、胎儿核酸、母本核酸、癌细胞核酸、非-癌细胞核酸、等及其组合。在非-限制性示例中,检测或未检测的特定遗传变异(i)在胎盘核酸中残留而在胎儿核酸和母本核酸中不残留;(ii)在胎儿核酸中残留但在母本核酸中不残留;或(iii)在母本核酸中残留但在胎儿核酸中不残留。

[0354] 通常由结果模块提供是否存在遗传变异和/或相关的医学病症(例如,结果)。在一些实施方式中,通过结果模块或包括结果模块的装置鉴定是否存在遗传变异(例如,非整倍性,胎儿非整倍性,拷贝数变异)。结果模块可特异用于确定特异性遗传变异(例如三体、21三体、18三体)。例如,鉴定21三体的结果模块可与鉴定18三体的结果模块不同和/或有差异。在一些实施方式中,需要结果模块或包括结果模块的装置以鉴定遗传变异或遗传变异(例如非整倍性、拷贝数变异)的确定性结果。在某些实施方式中,将结果从结果模块转移至显示模块,其中由显示模块提供结果。

[0355] 通过本文所述方法鉴定的遗传变异或遗传变异的确定性结果可通过其他测试独立验证(例如通过母本和/或胎儿核酸的靶测序)。结果通常提供给健康护理专业人员(如实验室技术人员或管理者;医师或助手)。在某些实施方式中,在合适的可视媒体(例如,机器的组件或周边,例如,打印机或显示器)上提供结果。在一些实施方式中,是否存在遗传变异的确定性结果以报告形式提供给健康护理专业人员,而在某些实施方式中,所述报告包含显示结果值和相关置信参数。通常,结果能以帮助确定是否存在遗传变异和/或医学病症的合适的格式显示。合适用于报告和/或显示数据组或报告结果的格式的非限定性示例包含数字数据、曲线图、2D图、3D图、和4D图、图片(例如,jpg、bitmap(例如,bmp)、pdf、tiff、gif、raw、png、等或合适的格式)、象形图、图表、表格、条线图、饼图、线图、流程图、散点图、图谱、柱状图、密度图、函数图、线路图、框图、起泡图、星座图、轮廓图、统计图、蛛网图、维恩图、列线图,及其组合。结果表示的各种示例如附图所示和实施例所述。

[0356] 在某些实施方式中,生成结果可视为核酸序列读数数据等变换为表示对象细胞核酸。例如,分析对象核酸的序列读数和生成染色体概况和/或结果可视为相对小的序列读数片段变换为相对大的染色体结构。在一些实施方式中,结果获自序列读数从对象(例如妊娠雌性)变换为表示对象(例如母体和/或胎儿核酸)中的现存结构(例如基因组,染色体或其区段)。在一些实施方式中,结果包括序列读数从第一对象(例如妊娠雌性)变换为结构(例如基因组、染色体或其区段)的复合表示,和复合表示的第二变换,其产生对第一对象对象(例如妊娠雌性)和/或第二对象(例如胎儿)中存在结构的表示。

[0357] 涉及性染色体的结果

[0358] 在一些实施方式中,结果涉及性染色体的遗传变异。性染色体的遗传变异描述于,例如,国际申请公开号W0 2013/192562,其全部内容通过引用纳入本文,包括所有文字、表格、等式和附图。在一些实施方式中,结果是性染色体核型的确定,性染色体非整倍性的检测和/或胎儿性别的确定。一些性染色体非整倍性(SCA)条件包括,但不限于,纳氏综合征[45,X]、X三体[47,XXX]、克兰费尔特综合征[47,XXY]、和[47,XYY]综合征(有时称为雅各布斯综合征)。

[0359] 在一些实施方式中,性染色体变异的评价是基于X染色体和Y染色体的序列读数计数转化的分离。序列读数计数转化可包括,例如,X染色体代表和Y染色体代表和/或基于这些代表的Z-分数。对于一组具有各种核型(例如,XX、XY、XXX、X、XXY、XYY)的样品的X染色体对Y染色体的核苷酸序列读数计数转化的二维图(例如,基于PERUN标准化读数计数或主成分标准化读数计数的Z分数)生成了图点的平面场,其可被切成对特定核型特异的区域。例如,给定样品的性染色体核型的确定可通过确定该样品的图点落入平面场的哪个区域来实现。

[0360] 本文所述的某些方法可用于生成具有对具体核型变异限定明确的区域(例如,具有锐利的边界,高分辨率)的图。有助于生成高分辨率图的方法包括序列读数计数标准化、X染色体和Y染色体的信息部份(即,箱)的选择、不可报告(即,“无判定”区)的建立、以及X染色体和Y染色体水平的其它标准化。本文描述了序列读数的标准化和水平的其它标准化并且可包括,例如,映射至X染色体和/或Y染色体和/或X染色体和/或Y染色体的水平(例如,染色体代表)的序列读数的PERUN标准化和/或主成分标准化。X染色体和Y染色体的信息部份的选择描述于,例如,国际专利申请公开号W0 2013/192562,并且可包括,例如,过滤参数的评价如交叉验证的参数、可映射性、可重复性和/或雄性对比雌性分离。

[0361] 结果的应用

[0362] 接收包含确定是否存在遗传变异的一个或多个结果的报告的健康护理专业人员、或其他有资格的人员能使用报告中显示的数据做出关于测试对象或患者的状态判定。在一些实施方式中,健康护理专业人员能根据提供的结果给出建议。在一些实施方式中,健康护理专业人员或有资格的人员能向测试对象或患者提供关于是否存在遗传变异的判定或评分,所述判定或评分基于一个或多个结果值或报告中提供的相关置信参数。在某些实施方式中,通过可见观察提供的报告,由健康护理专业人员或有资格的人员人工做出判定或评分。在某些实施方式中,由自动化程序(有时编入软件)做出评分或判定,并且在由健康护理专业人员或有资格的人员复查准确性之后再向测试对象或患者提供信息。本文所用的术语“接收报告”指通过任意联系方式获得包含结果的书面和或图示表示,其经复查后供健康护理专业人员或其他有资格的人员就关于测试对象或患者中是否存在遗传变异作出决定。所述报告可以通过计算机或人工数据输入生成,并且能使用电子方式(如从一个网络地址向相同或不同物理位点的另一个地址通过因特网、通过计算机、通过传真),或者通过发送或接收数据的任意其他方法(如信件服务、快递服务等)传达。在一些实施方式中,结果以合适介质传送至健康护理专家,包括但不限于口头、档案或文件形式。文件可为例如但不限于声音文件、计算机可读文件、纸质文件、实验室文件或医学报告文件。

[0363] 本文所用的术语“提供结果”及其语法等同形式也能指获得这种信息的任意方法,

包括但不限于从实验室(例如实验室文件)中获得信息。能通过实验室进行一种或多种试验或者一个或多个数据处理步骤生成实验室文件以确定是否存在所述医学病症。所述实验室可以和由所述实验室文件鉴定有或没有所述医学病症的人处于相同地点或不同地点(例如,在另一国家)。例如实验室文件可在一个地点生成并传递至另一地点,在后者处信息将传递至妊娠雌性对象。在某些实施方式中,所述实验室文件可以是有形形式或电子形式(例如计算机可读形式)。

[0364] 在一些实施方式中,结果可提供给健康护理专家、医师或实验室里有资格的个体并且健康护理专家、医师或有资格的个体可基于结果进行诊断。在一些实施方式中,结果可提供给健康护理专家、医师或实验室里有资格的个体并且健康护理专家、医师或有资格的个体可部分基于结果以及其他数据和/或信息和其他结果进行诊断。

[0365] 健康护理专家和有资格的个体可基于本报告中提供的结果提供合适的建议。能根据提供的结果报告提供建议的非限定性示例包括手术、放疗、化疗、遗传咨询、出生后治疗方案(如生命计划、长期辅助治疗、药物、对症治疗)、妊娠终止、器官移植、输血等,或前述组合。在一些实施方式中,建议取决于所提供的基于结果的分类提供(例如唐氏综合征、特纳综合征、关联T13遗传变异的医学病症、关联T18遗传变异的医学病症)。

[0366] 实验室人员(例如实验室管理员)可分析可能确定是否存在遗传变异(或确定测试区域整倍性或非-整倍性)的值(例如测试概况、参照概况、偏差水平)。对于关于是否存在遗传变异的狭窄(close)或可疑的判定,实验室人员可再进行相同测试和/或安排不同测试(例如胎儿非整倍性确定情况中的分核型和/或羊膜穿刺),其使用测试对象的相同或不同样品核酸。

[0367] 遗传变异和医学病症

[0368] 是否存在遗传变异能使用本文所述方法、机器或设备确定。在某些实施方式中,根据本文提供的方法、设备和装置所得到的结果确定是否存在一种或多种遗传变异。遗传变异通常是某些个体中存在的特定遗传表型,并且遗传变异经常是在统计学显著的个体亚群中存在。在一些实施方式中,遗传变异是染色体异常(例如,非整倍性、一个或多个染色体重复、一个或多个染色体缺失)、部份染色体异常或镶嵌(例如,丢失或得到染色体的一个或多个区段)、移位、翻转,其各自在本文中更详细描述。遗传变异的非限制性示例包括一种或多种缺失(例如微缺失)、复制(例如微复制)、插入、突变、多态性(例如单核苷酸多态性)、融合、重复(例如短串联重复)、差异甲基化位点、差异甲基化形式等及其组合。插入、重复(repeat)、缺失、复制(重复)、突变或多态性可以任意长度,并且在一些实施方式中是长度约1个碱基或碱基对(bp)-约250个兆碱基(Mb)。在一些实施方式中,插入、重复(repeat)、缺失、复制(重复)、突变或多态性是长度约1个碱基或碱基对(bp)至约50,000个千碱基(kb)(如长度约为10bp、50bp、100bp、500bp、1kb、5kb、10kb、50kb、100kb、500kb、1000kb、5000kb或10000kb)。

[0369] 遗传变异有时是缺失。在某些实施方式中,缺失是部分染色体或DNA序列丢失的突变(例如遗传失常)。缺失通常是遗传材料丢失。可缺失任何数量的核苷酸。缺失可包括缺失一种或多种全染色体、染色体区段、等位基因、基因、内含子、外显子、任何非-编码区域、任何编码区域、其区段或其组合。缺失可包括微缺失。缺失可包括缺失单个碱基。

[0370] 遗传变异有时是遗传复制。在某些实施方式中,复制是部分染色体或DNA序列复制

并插入基因组的突变(例如遗传失常)。在某些实施方式中,遗传复制(例如,复制)是DNA区域的任何复制。在一些实施方式中复制是基因组或染色体内的重复(通常串联)核酸序列。在一些实施方式中,复制可包括下述的拷贝:一种或多种全染色体、染色体区段、等位基因、基因、内含子、外显子、任何非-编码区域、任何编码区域、其区段或其组合。复制可包括微复制。复制有时包括复制核酸的一个或多个拷贝。有时复制表征为重复一次或多次(如重复1、2、3、4、5、6、7、8、9或10次)的遗传区域。在一些示例中,复制的范围可以是小区域(数千碱基对)到全部染色体。复制经常由于同源重组的错误或由于逆转座子事件造成。复制关联某些类型的增殖疾病。复制能使用基因组微阵列或比较基因组杂交(CGH)来表征。

[0371] 遗传变异有时是插入。插入有时是添加一种或多种核苷酸碱基对至核酸序列中。插入有时是微插入。在某些实施方式中插入包括添加染色体区段至基因组、染色体或其区段。在某些实施方式中插入包括添加等位基因、基因、内含子、外显子、任何非-编码区域、任何编码区域、其区段或其组合至基因组或其区段。在某些实施方式中插入包括添加(例如,插入)未知来源的核酸至基因组、染色体或其区段。在某些实施方式中插入包括添加(例如,插入)单个碱基。

[0372] 本文所用“拷贝数变异”通常是一类或一种遗传变异或染色体失常。拷贝数变异可为缺失(例如微缺失)、复制(例如微复制)或插入(例如微插入)。通常,本文所用前缀“微”有时是长度小于5Mb的核酸区段。拷贝数变异可包括染色体区段的一种或多种缺失(例如微缺失)、复制和/或插入(例如微复制、微插入)。在某些实施方式中复制包括插入。在某些实施方式中插入是复制。在某些实施方式中插入不是复制。

[0373] 在一些实施方式中拷贝数变异是胎儿拷贝数变异。通常,胎儿拷贝数变异是胎儿基因组中的拷贝数变异。在一些实施方式中拷贝数变异是母本和/或胎儿拷贝数变异。在某些实施方式中母本和/或胎儿拷贝数变异是妊娠雌性(例如怀有胎儿的雌性对象)、生育胎儿的雌性对象或能生育胎儿的雌性对象的基因组内的拷贝数变异。拷贝数变异可为杂合拷贝数变异,其中所述变异(例如复制或缺失)存在于基因组的一个等位基因上。拷贝数变异可为纯合拷贝数变异,其中所述变异存在于基因组的两个等位基因上。在一些实施方式中拷贝数变异是杂合或纯合拷贝数变异。在一些实施方式中拷贝数变异是杂合或纯合的母本和/或胎儿拷贝数变异。拷贝数变异有时存在于母本基因组和胎儿基因组中、母本基因组中存在但胎儿基因组中不存在、或胎儿基因组中存在但母本基因组不存在。

[0374] “倍性”是胎儿或母本中存在的染色体数量的参照。在某些实施方式中“倍性”与“染色体倍性”相同。例如人类中,常染色体通常成对存在。例如,不含遗传变异时,多数人具有各常染色体的两份(例如染色体1-22)。人中存在2个常染色体的正常互补物通常称为整倍体或二倍体。“微倍性”与倍性含义相似。“微倍性”通常指染色体区段的倍性。术语“微倍性”有时是染色体内是否存在拷贝数变异(例如缺失、复制和/或插入)的参照(例如纯合或杂合缺失、复制,或插入、等或或无变异)。

[0375] 在某些实施方式中胎儿微倍性与胎儿母本(例如,妊娠雌性对象)微倍性相匹配。在某些实施方式中胎儿微倍性与胎儿母本微倍性相匹配且母本和胎儿带有相同杂合拷贝数变异、纯合拷贝数变异或均为整倍体。在某些实施方式中胎儿微倍性不同于胎儿母本微倍性。例如,有时胎儿微倍性的拷贝数变异为杂合、而母本的拷贝数变异为纯合,针对特定拷贝数变异的胎儿微倍性与母本微倍性不匹配(例如不相等)。

[0376] 在某些实施方式中,鉴定对象是否存在遗传变异与医学病症相关联。因此,本文所述技术能用于鉴定是否存在与医疗病情或医疗状态相关联的一种或多种遗传变异。医学病症的非限定性示例包括与智力残疾(如唐氏综合征)、异常细胞增殖(如癌症)、存在微生物(如病毒、细菌、真菌、酵母)核酸和先兆子痫相关联的那些。

[0377] 遗传变异、医学病症和状态的非限制性示例如下所述。

[0378] 胎儿性别

[0379] 在一些实施方式中,可通过本文所述方法、机器和/或设备确定对胎儿性别或性别相关疾病(例如性染色体非整倍性)的预测。性别决定通常基于性染色体。人有两个性染色体,X和Y染色体。Y染色体含有启动胚胎发育为雄性的基因SRY。人和其他哺乳动物的Y染色体还含有产生正常精子所需的其他基因。有XX的个体是雌性,有XY的个体是雄性,而非限定性变异通常指性染色体非整倍性,包括X0、XYY、XXX和XXY。在某些实施方式中,雄性具有两个X染色体和一个Y染色体(XXY;克氏综合征)、或一个X染色体和两个Y染色体(XYY综合征;雅各布斯综合征),和一些雌性具有三个X染色体(XXX;三X综合征)或单一X染色体而非两个(X0;特纳综合征)。在某些实施方式中,仅部分个体细胞受性染色体非整倍性的影响,这称为镶嵌型(例如特纳镶嵌型)。其他情况包括SRY受损(产生XY雌性),或X复制(产生XX雄性)。

[0380] 在某些情况中,确定子宫中胎儿性别有益处。例如,具有一种或多种性连锁疾病家族史的亲本(例如妊娠雌性)可能希望确定所怀胎儿性别以评估胎儿遗传该疾病的风险。性连锁疾病包括,但不限于X连锁和Y连锁疾病。X连锁疾病包括X连锁隐性和X连锁显性疾病。X连锁隐性疾病的示例包括但不限于免疫疾病(例如慢性肉芽肿性疾病(CYBB)、Wiskott-Aldrich综合征、X连锁重症联合免疫缺陷、X连锁无丙种球蛋白血症缺乏症、1型高-IgM综合征、IPEX、X连锁淋巴增殖性疾病、备解素缺乏症)、血液病(例如血友病A、血友病B、X连锁铁粒幼细胞贫血)、内分泌障碍(例如雄激素不敏感综合征/Kennedy病、KAL1Kallmann综合征、X连锁先天性肾上腺发育不良)、代谢疾病(例如鸟氨酸转氨甲酰酶、眼脑肾综合征、肾上腺白质营养不良、葡萄糖六磷酸去氧酶缺乏症、丙酮酸盐脱氢酶缺乏症、Danon病/IIb型糖原贮积病、Fabry's病、Hunter综合征、Lesch-Nyhan综合征、Menkes病/后头角综合征)、神经系统疾病(例如Coffin-Lowry综合征、MASA综合征、X连锁 α 地中海贫血症精神发育迟滞综合征、SideriusX连锁智力低下综合征、色盲、眼白化、Norrie病、无脉络膜、Charcot-Marie-Tooth病(CMTX2-3)、Pelizaeus-Merzbacher病、SMA2)、皮肤和皮肤相关疾病(例如先天性角化不良、少汗性外胚层增生不良症(EDA)、X连锁鱼鳞病、X连锁角膜内皮失养症)、神经肌肉疾病(例如Becker's肌营养不良/Duchenne、中央核肌病(MTM1)、Conradi-Hünemann综合征、Emery-Dreifuss肌营养不良1)、泌尿疾病(例如Alport综合征、Dent's病、X连锁肾性尿崩症)、骨/牙疾病(例如AMELX釉质形成缺陷症)、和其他疾病(例如Barth综合征、McLeod综合征、Smith-Fineman-Myers综合征、Simpson-Golabi-Behmel综合征、Mohr-Tranebjærg综合征、Nasodigitoacoustic综合征)。X连锁显性疾病的示例包括但不限于X连锁低磷血症、灶性皮肤发育不全、脆性X综合征、Aicardi综合征、色素失调症、Rett综合征、CHILD综合征、Lujan-Fryns综合征、和Orofacio数码综合征1。Y连锁疾病的示例包括,但不限于男性不育、视网膜色素变性和无精子症。

[0381] 染色体异常

[0382] 在一些实施方式中,胎儿染色体异常可以使用本文所述方法、机器和/或设备确

定。染色体异常包括但不限于整个染色体或者包含一个或多个基因的染色体区域的获得或丢失。染色体异常包含单体性、三体性、多体性、杂合性的丢失、一个或多个核苷酸序列(如一个或多个基因)的缺失和/或重复,包含非平衡移位造成的缺失和重复。本文所用术语“染色体异常”或“非整倍性”指对象染色体与正常同源染色体的结构之间的偏差。术语“正常”指具体物种的健康个体中的主要染色体组型或结合形式,例如,整倍体基因组(例如,人中为二倍体,例如,46,XX或46,XY)。由于不同生物有广泛不同的染色体套数(complement),所述术语“非整倍性”并非指染色体的特定数目,而是指给定生物体细胞中染色体含量异常的情况。在一些实施方式中,本文术语“非整倍性”指由于缺失或获得全染色体,或部分染色体引起的遗传材料的不平衡。“非整倍性”可指染色体区段的一种或多种缺失和/或插入。在一些实施方式中,术语“整倍性”指染色体的正常套数。

[0383] 本文使用术语“单体性”指缺乏正常套数(complement)的一个染色体。部分单体性可在非平衡易位或缺失中发生,其中仅有一部分染色体以单个拷贝存在。例如性染色体(45,X)的单体性造成特纳氏综合征。术语“二体性”指存在染色体的两个拷贝。就各染色体有两个拷贝的生物体(二倍体或“整倍体”的那些)(例如人)而言,二体性是正常情况。就各染色体通常有三个或更多个拷贝的生物体(三倍体或更多倍体的那些)而言,二体性是非整倍染色体套数的状态。在单亲二体性中,染色体的两个拷贝来自相同亲本(另一亲本没有贡献)。

[0384] 本文所用术语“三体性”指存在特定染色体的三个拷贝,而不是两个拷贝。人唐氏综合征中发现额外一条染色体21的存在,称为“21三体性”。18三体性和13三体性是其它两种人常染色体三体性。性染色体的三体性可出现在雌性中(如47,三X症状中的XXX)或雄性中(如克氏综合征中的47,XXY;或47,雅各布斯症状中的XYY)。在一些实施方式中,三体性是大多数或所有常染色体的复制。在一些实施方式中,三体性是全染色体非整倍性,导致具体种类染色体的三种情况(如三拷贝)(而非二倍体的具体种类染色体的两种情况(即一对))。

[0385] 本文使用“四体性”和“五体性”指分别存在染色体的四个或五个拷贝。尽管对常染色体罕见,但已报道了人的性染色体的四体性和五体性,包括XXXX、XXXY、XXYY、YYYY、XXXXX、XXXXY、XXXYY、XXYYY和YYYYY。

[0386] 染色体异常可由各种机制引起。机制包括但不限于(i)有丝分裂检查点弱化导致的不分离、(ii)有丝分裂检查点失活造成多个染色体处的不分离、(iii)当一个着丝粒连接两个有丝分裂纺锤体极时发生单极向型(merotelic)连接、(iv)当形成多于两个纺锤体极时形成多极性纺锤体、(v)当形成仅一个纺锤体极时形成单极性纺锤体,和(vi)单极性纺锤体机制最终导致出现四倍体中间型。

[0387] 本文使用术语“部分单体性”和“部分三体性”指部分染色体的丧失或获得造成的遗传材料的不平衡。非平衡易位可导致部分单体性或部分三体性,此时个体载有通过两条不同染色体的破裂和融合形成的衍生染色体。在这种情况下,所述个体可以有一条染色体部分的三个拷贝(两个正常拷贝和所述衍生染色体上存在的部分),和所述衍生染色体中所带的其它染色体部分的仅仅一个拷贝。

[0388] 本文所用术语“镶嵌性”指生物体的一些细胞但不是全部细胞中的非整倍性。某些染色体异常能以镶嵌性(mosaic)和非镶嵌性(非mosaic)染色体异常形式存在。例如,某些21三体性个体有镶嵌性唐氏综合征而一些有非镶嵌性唐氏综合征。不同机制可导致镶嵌

性。例如 (i) 起始受精卵可以有三条21号染色体,正常情况下会导致简单的21三体性,但是在细胞分裂中一个或多个细胞系丢失了所述21号染色体中的一条;和(ii)起始受精卵可以有三条21号染色体,但是在细胞分裂中所述21号染色体中的一条复制。体细胞镶嵌性似乎通过不同于通常与涉及完全或镶嵌性非整倍性的遗传症状相关联的那些的机制发生。例如已在某些类型的癌症和神经元中鉴定了体细胞镶嵌性。在某些示例中,在慢性淋巴细胞性白血病 (CLL) 中鉴定了12三体性,而在急性骨髓性白血病 (AML) 中鉴定了8三体性。同样,有染色体破裂倾向 (染色体不稳定性综合征) 的个体的遗传症状常与多种类型癌症的风险增加相关联,从而突出了癌发生中体细胞非整倍性的作用。本文所述方法和试方案能鉴定是否存在非镶嵌性和镶嵌性的染色体异常。

[0389] 表1A和1B为可通过本文所述的方法、机器和/或设备可能鉴定的染色体病症、综合征和/或异常的非限定性列表。表1B来自2011年10月6日的DECIPHER数据库(如版本5.1,根据定位到GRCh37的位置;统一资源定位符 (URL) dechipher.sanger.ac.uk上可用)。

[0390] 表1A

[0391]

染色体	异常	相关疾病
X	XO	特纳氏综合征
Y	XXY	克氏综合征
Y	XYY	双 Y 综合征
Y	XXX	三体性 X 综合征
Y	XXXX	四 X 综合征

[0392]

染色体	异常	相关疾病
Y	Xp21 缺失	Duchenne/Becker 综合征、先天性肾上腺发育不良、慢性肉芽肿病
Y	Xp22 缺失	类固醇硫酸酯酶缺乏症
Y	Xq26 缺失	X-连锁淋巴细胞异常增生症
1	1p(体细胞)单体性三体性	成神经细胞瘤
2	单体性三体性 2q	发育迟缓、发育和智力延迟以及微小体细胞异常
3	单体性三体性 (体细胞)	非霍奇金淋巴瘤
4	单体性三体性 (体细胞)	急性非淋巴细胞性白血病(ANLL)
5	5p	猫叫综合征；累若纳(Lejeune)综合征
5	5q (体细胞) 单体性三体性	脊髓发育不良综合征
6	单体性三体性 (体细胞)	透明细胞性瘤
7	7q11.23 缺失	威廉姆斯综合征
7	单体性三体性	儿童单体性 7 综合征；体细胞：肾皮质腺瘤；骨髓增生异常综合征
8	8q24.1 缺失	Langer-Giedon 综合征

[0393]

染色体	异常	相关疾病
8	单体性三体性	骨髓增生异常综合征; Warkany 综合征; 体: 慢性髓细胞性白血病
9	9p 单体性	阿尔菲综合征
9	单体性 9p 部分 三体性	Rethore 综合征
9	三体性	完全 9 三体性综合征; 镶嵌性 9 三体性综合征
10	单体性三体性 (体细胞)	ALL 或 ANLL
11	11p-	先天性无虹膜症; 维尔姆斯瘤
11	11q-	雅各布森综合征(雅各布斯综合征)
11	单体性(体细胞) 三体性	受影响的骨髓系(ANLL、MDS)
12	单体性三体性 (体细胞)	CLL、幼年型粒层细胞瘤(JGCT)
13	13q-	13q-综合征; 欧拜里综合征
13	13q14 缺失	成视网膜细胞瘤
13	单体性三体性	帕套氏综合征
14	单体性三体性 (体细胞)	骨髓疾病(MDS、ANLL、非典型 CML)
15	15q11-q13 缺 失单体性	帕-魏二氏、安琪曼(Angelman)综合征

[0394]

染色体	异常	相关疾病
15	三体性(体细胞)	受影响的骨髓和淋巴系，如 MDS、ANLL、ALL、CLL)
16	三体性	完全 16 三体性 镶嵌 16 三体性
16	16q13.3 缺失	鲁宾斯坦-泰比(Rubenstein-Taybi)
3	单体性三体性 (体细胞)	乳头状肾细胞癌(恶性)
17	17p-(体细胞)	骨髓性恶性肿瘤中的 17p 综合征
17	17q11.2 缺失	史密斯-马吉利(Smith-Magenis)综合征
17	17q13.3	敏-迪(Miller-Dieker)综合征
17	单体性三体性 (体细胞)	肾皮质腺瘤
17	17p11.2-12 三体性	1 型恰克-马利-杜斯氏(Charcot-Marie-Tooth)综合征；HNPP
18	18p-	18p 部分单体性综合征或 Grouchy Lamy Thieffry 综合征
18	18q-	Grouchy Lamy Salmon Landry 综合征
18	单体性三体性	爱德华氏综合征
19	单体性三体性	
20	20p-	三体性 20p 综合征
20	20p11.2-12 缺失	阿拉吉耶(Alagille)综合征

[0395]

染色体	异常	相关疾病
20	20q-	体:体细胞: MDS、ANLL、真性红细胞增多症、慢性嗜中性粒细胞白血病
20	单体性三体性 (体细胞)	乳头状肾细胞癌(恶性)
21	单体性三体性	唐氏综合征
22	22q11.2 缺失	迪格奥尔格(DiGeorge)综合征, 软腭-心-面综合征、先天性异常面容综合征、常染色体显性 Opitz G/BBB 综合征、Cayler 心面综合征
22	单体性三体性	完全 22 三体性综合征

[0396] 表1B

[0397]

综合征	染色体	起始	终止	间隔(Mb)	级别
12q14 微缺失综合征	12	65,071,919	68,645,525	3.57	
15q13.3 微缺失综合征	15	30,769,995	32,701,482	1.93	
15q24 复发微缺失综合征	15	74,377,174	76,162,277	1.79	
15q26 过度生长综合征	15	99,357,970	102,521,392	3.16	
16p11.2 微重复综合征	16	29,501,198	30,202,572	0.70	
16p11.2-p12.2 微缺失综合征	16	21,613,956	29,042,192	7.43	
16p13.11 复发性微复制(微重复)	16	15,504,454	16,284,248	0.78	

[0398]

综合征	染色体	起始	终止	间隔(Mb)	级别
(神经认知疾病的易感位点)					
16p13.11 复发性微复制(微重复)(神经认知疾病的易感位点)	16	15,504,454	16,284,248	0.78	
17q21.3 复发微缺失综合征	17	43,632,466	44,210,205	0.58	1
1p36 微缺失综合征	1	10,001	5,408,761	5.40	1
1q21.1 复发性微缺失(神经发育疾病的易感位点)	1	146,512,930	147,737,500	1.22	3
1q21.1 复发性微复制(神经发育疾病的可能的易感位点)	1	146,512,930	147,737,500	1.22	3
1q21.1 血小板减少症-桡骨缺失(TAR)综合征易感位点	1	145,401,253	145,928,123	0.53	3
22q11 缺失综合征(软腭心面/迪格奥尔格(DiGeorge)综合征)	22	18,546,349	22,336,469	3.79	1
22q11 重复综合	22	18,546,349	22,336,469	3.79	3

[0399]

综合征	染色体	起始	终止	间隔(Mb)	级别
征					
22q11.2 远端缺失综合征	22	22,115,848	23,696,229	1.58	
22q13 缺失综合征 (Phelan-Mcdermid 综合征)	22	51,045,516	51,187,844	0.14	1
2p15-16.1 微缺失综合征	2	57,741,796	61,738,334	4.00	
2q33.1 缺失综合征	2	196,925,089	205,206,940	8.28	1
2q37 单体性	2	239,954,693	243,102,476	3.15	1
3q29 微缺失综合征	3	195,672,229	197,497,869	1.83	
3q29 微重复综合征	3	195,672,229	197,497,869	1.83	
7q11.23 重复综合征	7	72,332,743	74,616,901	2.28	
8p23.1 缺失综合征	8	8,119,295	11,765,719	3.65	
9q 亚端粒缺失综合征	9	140,403,363	141,153,431	0.75	1
成年型常染色体 主导性脑白质病 变(ADLD)	5	126,063,045	126,204,952	0.14	
安格尔曼综合征 (1 型)	15	22,876,632	28,557,186	5.68	1

[0400]

综合征	染色体	起始	终止	间隔(Mb)	级别
安格尔曼综合征 (2 型)	15	23,758,390	28,557,186	4.80	1
ATR-16 综合征	16	60,001	834,372	0.77	1
AZFa	Y	14,352,761	15,154,862	0.80	
AZFb	Y	20,118,045	26,065,197	5.95	
AZFb+AZFc	Y	19,964,826	27,793,830	7.83	
AZFc	Y	24,977,425	28,033,929	3.06	
猫眼综合征(I 型)	22	1	16,971,860	16.97	
1A 型恰克-马利- 杜斯氏综合征 (CMT1A)	17	13,968,607	15,434,038	1.47	1
猫叫综合征(5p 缺失)	5	10,001	11,723,854	11.71	1
伴随大脑淀粉样 血管病的早发型 阿兹海默症	21	27,037,956	27,548,479	0.51	
家族性腺瘤性息 肉病	5	112,101,59 6	112,221,37 7	0.12	
遗传性压力易感 性神经病(HNPP)	17	13,968,607	15,434,038	1.47	1
Leri-Weill 不良 软骨发生症 (LWD) - SHOX 缺失	X	751,878	867,875	0.12	
Leri-Weill 不良 软骨发生症 (LWD) - SHOX 缺失	X	460,558	753,877	0.29	

[0401]

综合征	染色体	起始	终止	间隔(Mb)	级别
敏 - 迪 (Miller-Dieker)综 合征(MDS)	17	1	2,545,429	2.55	1
NF1-微缺失综合 征	17	29,162,822	30,218,667	1.06	1
佩 - 梅 二 氏 (Pelizaeus-Merzb acher)病	X	102,642,05 1	103,131,76 7	0.49	
Potocki-Lupski 综 合征(17p11.2 复 制综合征)	17	16,706,021	20,482,061	3.78	
Potocki-Shaffer 综合征	11	43,985,277	46,064,560	2.08	1
Prader-Willi 综合 征(1 型)	15	22,876,632	28,557,186	5.68	1
Prader-Willi 综合 征(2 型)	15	23,758,390	28,557,186	4.80	1
RCAD(肾囊肿和 糖尿病)	17	34,907,366	36,076,803	1.17	
Rubinstein-Taybi 综合征	16	3,781,464	3,861,246	0.08	1
史密斯 - 马 吉 利 (Smith-Magenis) 综合征	17	16,706,021	20,482,061	3.78	1
索托氏(Sotos)综 合征	5	175,130,40 2	177,456,54 5	2.33	1
裂手 / 足 畸 形 1(SHFM1)	7	95,533,860	96,779,486	1.25	

[0402]

综合征	染色体	起始	终止	间隔(Mb)	级别
类固醇硫酸酯酶缺陷(STS)	X	6,441,957	8,167,697	1.73	
WAGR 11p13 缺失综合征	11	31,803,509	32,510,988	0.71	
Williams-Beuren 综合征(WBS)	7	72,332,743	74,616,901	2.28	1
Wolf-Hirschhorn 综合征	4	10,001	2,073,670	2.06	1
Xq28(MECF2) 复制 (重复)	X	152,749,900	153,390,999	0.64	

[0403] 1级病症常具有以下特点中的一种或多种:致病性异常;遗传学家之间的一致认同;高度外显;仍然可以有可变的表型,但有一些共同特性;文献中的所有示例有临床表型;没有含所述异常的健康个体示例;在DVG数据库上没有报导或健康人群中没有发现;证明单个基因或多个基因剂量效应的函数型数据;经证明的候选基因或强候选基因;明确的临床管理意义;监测意义伴随的已知的癌症风险;多个信息来源 (OMIM、GeneReviews、Orphanet、Unique、Wikipedia);和/或可以用于诊断应用(生殖咨询(counseling))。

[0404] 2级病症常具有以下特点中的一种或多种:可能的致病性异常;高度外显;除了DD以外没有持续特性的可变表型;文献中少量示例/报导;所有报导的示例都有临床表型;没有函数型数据或经证明的病原基因;多个信息来源 (OMIM、GeneReviews、Orphanet、Unique、Wikipedia);和/或可以用于诊断目的和生殖咨询。

[0405] 3级病症常具有以下特点中的一种或多种:易感位点;健康个体或所述先证者的未受影响的父母;在对照人群中存在;非显性;表型温和且无特异性;特征较不一致;无函数型数据或经证明的病原基因;更有限的数据来源;针对偏离大部分或有新临床发现设想提出的情况仍可能有二次诊断的可能性;和/或用于诊断目的时需谨慎且就生殖咨询给出有保留的建议。

[0406] 先兆子痫

[0407] 在一些实施方式中,使用本文所述的方法、机器或设备确定是否存在先兆子痫。先兆子痫是妊娠中出现高血压(例如,妊娠诱导的高血压)且与尿中高蛋白含量相关联的病症。在一些实施方式中,先兆子痫也与升高的细胞外核酸和/或甲基化形式改变有关。例如,已经观察到了胞外胎儿源性高甲基化RASSF1A水平和先兆子痫的严重性正相关。在某些实施例中,对比正常对照,在先兆子痫胎盘中观察到了H19基因DNA甲基化的增加。

[0408] 先兆子痫是世界范围内母本和胎儿/新生儿死亡率和发病率的主要原因之一。血浆和血清中的循环无细胞核酸是在包括产前诊断在内的不同医学领域中具有临床应用前景的新型生物标志物。不同研究中已报道了将母本血浆中无细胞胎儿(cff) DNA的量变作为

即将发生先兆子痫的指示物,例如针对雄性特异性SRY或DYS 14位点使用实时定量PCR。在早发型先兆子痫的示例中,在头三个月内可以观察到水平提高。症状发作前cffDNA水平的增加可归因为绒毛间空隙中的缺氧/复氧,导致了组织氧化应激和胎盘凋亡及坏死的增加。除有证据证明先兆子痫中排入母本循环的cffDNA增多以外,也有证据证明在先兆子痫中cffDNA的肾清除率降低。由于当前通过定量Y-染色体特异性序列确定胎儿DNA的量,替代性方法例如测量总无细胞DNA或使用性别无关的胎儿表观遗传学标记物(如DNA甲基化)提供了其它选择。胎盘来源的无细胞RNA是可以在临床实践中用于筛选和诊断先兆子痫的另一个备选标记。胎儿RNA与保护其免于降解的亚细胞胎盘颗粒相关联。有先兆子痫的妊娠雌性中的胎儿RNA水平有时相较于对照高出10倍,因此所述胎儿RNA水平是可以在临床实践中用于筛选和诊断先兆子痫的备选生物标记物。

[0409] 病原体

[0410] 在一些实施方式中,通过本文所述的方法、机器或设备确定是否存在病原性病症。病原性病症能通过病原体(包括但不限于细菌、病毒或真菌)感染宿主而产生。由于病原体通常具有能与宿主核酸区分的核酸(如基因组DNA、基因组RNA、mRNA),本文提供的方法、机器和设备能用于确定是否存在病原体。通常,病原体具有特定病原体独有特性的核酸,例如表观遗传状态和/或一种或多种序列变异、复制和/或缺失。因此,本方法可用于鉴定具体病原体或病原体变体(例如株系)。

[0411] 癌症

[0412] 在一些实施方式中,能使用本文所述方法、装置或设备确定是否存在细胞增殖疾病(如癌症)。例如,相较于健康患者,多种类型癌症患者内血清中的无细胞核酸水平会升高。例如,患有转移性疾病的患者的血清DNA水平有时能比无转移患者高出大约两倍。患有转移性疾病的患者还可以通过例如癌症特异性标志物和/或某些单核苷酸多态性或短串联重复来鉴定。可以与循环DNA水平提高正相关的癌症类型的非限定性示例包括乳腺癌、结直肠癌、胃肠癌、肝细胞癌、肺癌、黑素瘤、非霍奇金淋巴瘤、白血病、多发性骨髓瘤、膀胱癌、肝细胞瘤、宫颈癌、食道癌、胰腺癌和前列腺癌。多种癌症能具有与非癌健康细胞核酸特性(例如表观遗传状态和/或序列变异、复制和/或缺失)不同的核酸,并且有时释放其进入血液。例如此类特性可对特定类型的癌症有特异性。因此,还考虑本文所提供的方法能用于鉴定特定类型的癌症。

[0413] 软件能用于进行本文所述方法的一个或多个步骤,包括但不限于:计数、数据处理、生成结果和/或根据生成的结果提供一个或多个建议,如下文详细描述。

[0414] 机器、软件和接口

[0415] 在没有计算机、处理器、软件、模块或其它设备的情况下,本文所述的某些过程和方法通常不可进行。本文所述方法通常是计算机-执行方法,且方法的一或多部分有时通过一种或多种处理器(例如微处理器)、计算机、或微处理器控制的设备进行。在一些实施方式中,本文所述或已知的一个或多个或全部处理方法(例如,映射、数据压缩、局部基因组偏好评价确定、关系确定、关系比较、计数标准化、读数密度和/或读数密度概况生成、PCA、概况调整、部份过滤、部份加权、概况比较、概况评分、结果确定等或其组合)由处理器、微处理器、计算机与存储器联用和/或微处理器控制的设备进行。本文相关方法的实施方式通常可应用于本文所述系统、装置和计算机程序产物中的指令执行的相同或相关过程。在一些实

施方式中,本文所述过程和方法(例如定量、计数和/或确定序列读数、计数、水平和/或概况)通过自动化方法进行。在一些实施方式中,本文所述一种或多种步骤和方法通过处理器和/或计算机进行、和/或联合存储器进行。在一些实施方式中,自动化方法物化在软件、模块、处理器、外围设备和/或装置等中,其确定序列读数、计数、映射、映射的序列标签、水平、概况、标准化、比较、范围设定、归类、调整、作图、结果、变换和鉴定。如本文所用,软件指在由处理器执行时进行计算机操作的计算机可读程序指令。

[0416] 源自测试对象(例如患者,妊娠雌性)和/或参照对象的序列读数、计数、读数密度、和读数密度概况可进一步和加工以确定是否存在遗传变异。测序读数、计数、水平和/或概况有时称为“数据”或“数据组”。在一些实施方式中,数据或数据组能表征为一种或多种特性或变量(如基于序列的[如GC含量、特异性核苷酸序列等]、功能特异性[如表达的基因、癌基因等]、基于定位的[基因组特异性、染色体特异性、部份特异性]等及其组合)。在某些实施方式中,能根据一种或多种特性或变量将数据或数据组组织成有两个或多个维度的矩阵。能使用任何合适的特性或变量对组织成矩阵的数据分级。矩阵中的数据的非限定性示例包含通过母本年龄、母本倍性和胎儿基值组织的数据。在某些实施方式中,根据一种或多种特征或变量分类的数据组有时在计数后加工。

[0417] 可使用设备(多个设备,本文中也称为设备的复数)、软件和接口来进行本文所述的方法。使用设备、软件和接口,用户可以进入、请求、查询或确定用于使用特定信息、程序或方法的选项(如定位序列读数、处理定位的数据和/或提供结果),例如,所述信息、程序或方法可涉及实现统计学分析算法、统计学显著性算法、统计学变化算法、比较、重复步骤、验证算法和图示显示。在一些实施方式中,数据组可以作为输入信息由用户输入,用户可以通过任意合适的硬件介质(如闪存)下载一个或多个数据组,并且/或者用户可以从一个系统向另一个系统发送数据组以供后续处理和/或提供结果(如从一个测序器向计算机系统发送序列读取数据以定位序列读数;向计算机系统发送定位的序列数据以处理和生成结果和/或报告)。

[0418] 系统通常包括一种或多种设备。在一些实施方式中,设备是机器。在一些实施方式中,设备包括机器。设备可包括一种或多种存储器、一种或多种处理器和/或指令。系统包括两个或更多设备时,一些或全部设备可位于相同位置、一些或全部设备可位于不同位置、所有设备可位于一个位置和/或所有设备可位于不同位置。系统包括两个或更多设备时,一些或全部设备可位于用户的相同位置、一些或全部设备可位于用户的不同位置、所有设备可位于用户的相同位置和/或所有设备可位于用户的一个或多个不同位置。本文所述的系统的设备可通过合适的方法对接一个或多个远程计算服务器和/或计算机(例如,云、云计算服务器)。本文所用的术语“云”部份是指通过实时可进行集中化功能(例如,本文所述的方法)的通信网络(例如,英特网)连接的2台或更多台计算机(例如,通常多台计算机),其中功能的部份由网络中的多台计算机共享。“云”通常可同时在多个连接的计算机上运行一个或多个程序(例如,软件程序,模块)。在一些实施方式中,本文所述的系统和/或设备包括云(例如,云服务器、云计算、云计算服务)。本文所述的系统和/或设备的一个或多个功能可通过云来进行。可使用合适的方法从设备和/云中转入和/或转出数据和/或信息。本文所用的术语“计算机”是指电子人造设备,其包括可进行数学和逻辑运算的微处理器。计算机有时包括指令、软件(例如,模块)、存储器、显示器、一个或多个外设和/或存储介质。在一些实

施方式中,机器包括计算机。在一些实施方式中,机器是计算机。计算机通常对接和/或连接其它计算机(例如,英特网、网络、云)。

[0419] 系统有时包括计算设备或测序设备,或计算设备和测序设备(即,测序机器和/或计算机器)。测序设备通常设置成接收生理核酸并生成对应于核酸的核苷酸碱基的信号。通常用包含核酸的样品“加载”测序设备并且加载到测序设备的样品的核酸一般经过核酸测序过程。本文所用术语“加载测序设备”是指使测序设备的部份(例如,流动池)接触核酸样品,测序设备的部份设置为接收样品用于进行核酸测序过程。在一些实施方式中,用多种样品核酸来加载测序设备。有时通过修饰样品核酸至适于对核酸测序的形式来产生变体(例如,通过连接(例如,通过连接在样品核酸的末端添加衔接子)、扩增、限制性消化等或其组合)。测序设备通常部份设置为进行合适的DNA测序方法,其生成对应于加载的核酸的核苷酸碱基的信号(例如,电子信号、检测器信号、图像等或其组合)。

[0420] 对应于DNA序列的各碱基的一个或多个信号通常通过合适的过程加工和/或转化成碱基判定(例如,具体核苷酸碱基,例如,鸟嘌呤、胞嘧啶、胸腺嘧啶、尿嘧啶、腺嘌呤等)。来自加载的核酸的碱基判定的集合通常经处理和/或组装成一个或多个序列读数。在同一时间对多个样品核酸进行测序(即,多重化)的实施方式中,可使用合适的去多重化方法来使特定读数与它们来源的样品核酸相关联。可通过合适方法将序列读数与参照基因组比对,并且可对与参照基因组的部份比对的读数进行计数,如本文所述。

[0421] 在系统中,测序设备有时关联和/或包括一个或多个计算设备。该一个或多个计算设备有时设置成进行以下过程中的一个或多个:从测序设备信号中生成碱基判定、组装读数(例如,生成读数)、去多重化读数、将读数与参照基因组比对、对与参照基因组中的基因组部份对齐的读数进行计数等。该一个或多个计算设备有时设置成进行以下额外过程中的一个或多个:标准化读数计数(例如,减少或去除偏好)、生成一个或多个确定(例如,确定胎儿分数、胎儿倍性、胎儿性别、胎儿染色体计数、结果、是否存在遗传变异(例如,是否存在胎儿染色体非整倍性(例如,染色体13、18和/或21三体性)))等。

[0422] 在一些实施方式中,一个计算设备关联测序设备,并且在某些实施方式中,这一个计算设备进行以下过程中的大多数或全部:从测序设备信号生成碱基判定、组装读数、去多重化读数、比对读数和与参照基因组的基因组部份对齐的读数进行计数、标准化读数计数和生成一个或多个结果(例如,胎儿分数,是否存在特定遗传变异)。在后一个实施方式中,其中一个计算设备关联测序设备,该计算设备通常包括一个或多个处理器(例如,微处理器)和具有由一个或多个处理器进行的指令的存储器以进行该过程。在一些实施方式中,这一个计算设备可以是测序设备本地的单核或多核计算设备(例如,位于同一位置(例如,相同地址、相同构造、相同层、相同室等))。在一些实施方式中,这一个计算设备与测序设备集成。

[0423] 在一些实施方式中,系统中的多个计算设备关联测序设备,并且由系统进行的总过程的子集可在系统中的特定计算设备之间分配或划分。过程的总数的子集可在2个或更多个计算设备或其组之间以任意合适的组合划分。在某些实施方式中,由第一计算设备或其组进行从测序设备信号生成碱基判定、组装读数和去多重化,由第二计算设备或其组进行对映射至参照基因组的部份的读数的比对和计数,并且由第三计算设备或其组进行对读数计数的标准化并提供一个或多个结果。在包括2个或更多个计算设备或其组的系

统中,各特定计算设备可包括存储器、一个或多个处理器或其组合。多计算设备系统有时包括测序设备本地的一个或多个合适的服务器,并且有时包括非测序设备本地的一个或多个合适的服务器(例如,网页服务器、即时服务器、应用服务器、远端文件服务器、云服务器(例如,云环境、云计算))。

[0424] 不同系统构造中的设备可生成不同类型的输出数据。例如,测序设备可输出基础信号并且该基础信号输出数据可转移至将基础信号数据转化为基础判定的计算设备。在一些实施方式中,基础判定是来自一个计算设备的输出数据并且转移至另一个计算设备用于生成序列读数。在某些实施方式中,基础判定不是来自特定设备的输出数据,并且相反,用于接收测序设备基础信号以生成序列读数的同一设备中。在一些实施方式中,一个设备接收测序设备计数信号,生成基础判定、序列读数和去多重化序列读数,并且输出样品的去多重化序列读数,其可转移至将测序读数与参照基因组比对的另一个设备或其组。在一些实施方式中,一个设备或其组可输出映射至参照基因组的部份(例如,SAM或BAM文件)的对齐的序列读数,并且这种输出数据可转移至将序列读数标准化(例如,将序列读数的计数标准化)并生成结果(例如,胎儿分数和/或是否存在胎儿三体性)的第二计算设备或其组。来自一个设备的输出数据可以任意合适的方式转移至第二设备。例如,来自一个设备的输出处理有时置于物理存储装置上并且将该存储装置运输并连接至输出数据转移至的第二设备。有时由数据库中的一个设备来存储输出数据,并且第二设备评估来自相同数据库的输出数据。

[0425] 系统有时包括偏好减少机器。偏好减少机器有时包括一台或多台计算机。在一些实施方式中,偏好减少机器映射序列读数和/或压缩读数(例如,映射的序列读数)。偏好减少机器有时将序列读数压缩成合适的压缩格式(例如,BReads格式)。在一些实施方式中,偏好减少机器生成读数密度、密度概况、调整的读数密度概况和/或结果。可通过网络和/或云(例如,云计算网络)来进行偏好减少机器的一个或多个功能。偏好减少机器可对接包括微处理器、存储器和存储介质、模块、数据和/或信息(例如,参照、参照序列读数、参照读数密度、参照密度概况等)和/或软件的多个服务器(例如,云服务器)。偏好减少机器可将数据和/或信息转移至云,其中进行偏好减少机器的一个或多个功能。处理的数据和/或信息可从云转移至偏好减少机器。

[0426] 系统有时包括测序机器和偏好减少机器,其中测序机器从样品核酸生成序列读数,有时映射序列读数,并且提供和/或转移未映射或映射的序列读数至偏好减少机器。测序机器可通过任意合适的方法提供或转移读数至偏好减少机器。测序机器和偏好减少机器有时通过合适的硬件接口连接在一起。在一些实施方式中,测序机器和偏好减少机器连接至网络和/或云。在一些实施方式中,测序机器和偏好减少机器由网络和/或云连接在一起。可由云进行测序机器和/或偏好减少机器的一些或全部方法和/或功能。测序机器可通过使用瞬时和/或非瞬时计算机可读介质将读数转移至偏好减少机器。例如,可通过有线电缆传输的数字或模拟信号和/或无线信号来转移序列读数。在一些实施方式中,使用非-瞬时计算机可读储存介质将序列读数从测序机器转移至偏好减少机器。

[0427] 偏好减少机器可包括本文所述的一个或多个模块,其可进行偏好减少机器的一些或全部功能。在一些实施方式中,偏好减少机器包括压缩模块并且进行压缩模块的功能。在一些实施方式中,偏好减少机器包括偏好密度模块、关系模块、偏好校正模块和/或多变量

校正模块中的一个或多个。偏好校正机器可使用一个或多个模块来从读数中去除偏好(例如,GC偏好)和/或提供样品读数的标准化计数。在一些实施方式中,偏好校正机器包括分布模块、过滤模块和/或概况生成模块中的一个或多个。偏好校正机器可通常处理来自训练组或参照的序列读数以及来自测试样品的序列读数。在一些实施方式中,偏好校正机器包括PCA统计模块和/或部份加权模块中的一个或多个。偏好校正机器通常采用映射的读数和多个模块并且向打分模块、终端用户、计算机周边(例如,显示器,打印机)或向结果生成机器提供读数密度、密度概况和/或调整的读数密度概况。在一些实施方式中,偏好减少机器提供结果。有时,偏好减少机器不提供结果。在一些实施方式中,偏好减少机器包括结果生成机器。有时,偏好减少机器将标准化的读数、读数密度、密度概况和/或调整的读数密度概况转移至结果生成机器。偏好减少机器可通过任意合适的方法将数据和/或信息(例如,读数密度概况)转移至结果生成机器。在一些实施方式中,系统包括测序机器、偏好减少机器和/或结果生成机器中的一个或多个。结果生成机器可接收来自偏好校正机器的读数的标准化的计数、读数密度、密度概况和/或调整的读数密度概况。结果生成机器通常提供判定或结果(例如,确定是否存在遗传变异)。结果生成机器通常向终端用户和/或计算机外设(例如,显示器,印刷机)提供判定或结果。结果生成机器有时包括过滤模块、分布模块、概况生成模块、PCA统计模块、部份加权模块、打分模块和/或一个或多个其它合适的模块中的一个或多个。

[0428] 在一些实施方式中,用于与设备(例如,计算设备,测序设备)交互。在一些实施方式中,用户可以向系统、计算机或模块设置查询,所述系统、计算机或模块随后可以通过因特网入口(例如,云)获得数据组,而在某些实施方式中,可指示可编程的处理器根据给定参数获得合适的数据组。可编程的处理器也可以提示用户选择由处理器在给定参数基础上所选的一个或多个数据组选项。可编程的处理器可以提示用户选择由所述处理器基于通过因特网、其它内部或外部信息等发现的信息所选的一个或多个数据组选项。可以选定选项以选择方法、设备或计算机程序的一个或多个数据特性选择、一种或多种统计学算法、一种或多种统计学分析算法、一种或多种统计学显著性算法、重复步骤、一种或多种确证算法和一种或多种图示显示。

[0429] 本文所述的系统可以包括计算机系统的通用组件,例如网络服务器、笔记本系统、台式系统、手持系统、个人数字助理、计算机自助服务终端等。计算机系统可包括一种或多种输入方式例如键盘、触摸屏、鼠标、语音识别或其他方式以允许用户向系统输入数据系统。系统还可以包含一个或多个输出,包括但不限于显示屏(如CRT或LCD)、扬声器、传真机、打印机(如激光、喷墨、击打式、黑白或彩色打印机)或用于提供信息的视觉、听觉和/或硬拷贝输出(如结果和/或报告)的其它方式。在一些实施方式中,显示模块处理、变换和/或转移数据和/或信息至合适的可视媒体用于在合适的显示器(例如监视器、LED、LCD、CRT等或其组合)、打印机、合适的外围设备或装置上呈现。在某些实施方式中,显示模块提供对关系、概况或结果的可视显示。合适的可视媒体和/或显示器的非限制性示例包括表、图、照片等或其组合。在一些实施方式中,显示模块处理、变换数据和/或信息为胎儿或母本基因组,或其区段(例如,染色体或其部分)的可视表示。在一些实施方式中,需要显示模块或包括显示模块的机器以提供合适的可视显示。

[0430] 系统中,可使所述输入和输出方式连接中央处理单元,该单元可含有运行程序指

令的微处理器和存储程序编码与数据的存储器和其它组件。在一些实施方式中,处理可作为位于单一地理位置的单用户系统实施。在某些实施方式中,处理可作为多用户系统实施。在多用户执行的情况中,多中央处理单元可通过网络连通。所述网络可位于本地,涵盖建筑物的部分中的单一隔室、整个建筑物、跨多个建筑物、跨区域、跨国家或全球。所述网络可为私人的,由提供者所有并控制,或其可作为基于网络的服务而执行,用户进入网页来输入或找回信息。因此,在某些实施方式中,系统包含可由用户定位或遥控的一种或多种机器。用户可以访问在一个或多个位置的多于一台机器,并且数据可以以系列和/或平行方式作图和/或处理。因此,可利用任意合适的结构和控制来使用多机器绘图和/或处理数据,所述机器例如局部网络、远程网络和/或“云”计算机平台。

[0431] 在一些实施方式中,系统能包含通信接口。通信接口使软件和数据能在计算机系统和一种或多种外部设备之间转移。通信接口的非限定性示例可包括调制解调器、网络接口(例如以太网卡)、通信端口、PCMCIA槽和卡等。经通信接口转移的软件和数据通常为信号形式,其可以是能被通信接口接收的电子、电磁、光学和/或其它信号。信号经常通过通道提供给通信接口。通道经常携带信号,并能采用导线或线缆、光纤、电话线、手机连接、RF连接和其它通信通道实现。因此,在一个实施例中,可采用通信接口接收能由信号检测模块确定的信号信息。

[0432] 数据能由任意合适的设备和/或方法输入,所述设备和/或方法包括但不限于人工输入设备或直接数据输入设备(DDE)。人工设备的非限定性示例包括键盘、概念键盘、触敏屏、光笔、鼠标、轨迹球、操纵杆、图形平板、扫描仪、数码相较、视频数字化仪和语音识别设备。DDE的非限定性示例包括条形码扫描仪、磁条编码、智能卡、磁墨字符识别、光学字符识别、光学标记识别、和周转文件。

[0433] 在一些实施方式中,测序设备的输出可以作为能通过输入设备输入的数据。在某些实施方式中,定位的序列读数可以作为能通过输入设备输入的数据。在某些实施方式中,模拟数据通过计算机虚拟(in silico)方法生成,并且所述模拟数据作为能通过输入设备输入的数据。术语“计算机虚拟(in silico)”指采用计算机进行的研究和实验。计算机虚拟方法包括但不限于根据本文所述方法的定位的序列读数和处理定位的序列读数。

[0434] 系统可包含用于运行本文所述方法的软件,并且软件能包含用于运行这种方法的一种或多种模块(如测序模块、偏好校正模块、显示模块)。如本文所用,软件指在由计算机执行时进行计算机操作的计算机可读程序指令。一个或多个处理器可执行的指令有时作为可执行代码提供,其在运行时可使一个或多个处理器执行本发明方法。本文所述模块可以软件形式存在,且软件中内置的指令(例如过程、例程、子例程)可通过处理器执行或进行。例如,模块(例如软件模块)是执行特定方法和任务的程序的部分。术语“模块”指能用于较大设备或软件系统的独立功能单元。模块可包括指令组以通过一个或多个微处理器进行模块的功能。可通过使用合适的编程语言、合适的软件、和/或写成合适语言的代码(例如,本领域已知的计算机编程语言)和/或操作系统来执行模块的指令,其非限制性示例包括UNIX、Linux、oracle、windows、Ubuntu、ActionScript、C、C++、C#、Haskell、Java、JavaScript、Objective-C、Perl、Python、Ruby、Smalltalk、SQL、Visual Basic、COBOL、Fortran、UML、HTML(例如,PHP)、PGP、G、R、S等或其组合。在一些实施方式中,本文所述的模块包括以采用合适的包(例如,S包或R包)的S或R编写的代码(例如,脚本)。R、R源代码、R程

序、R包和R存档可从镜像网站下载获得(R综合档案网络(CRAN) [在线], [2013-04-24上线], 从英特网<URL:*>http://cran.us.r-project.org/<>上线)。CRAN是全世界ftp和web服务器的网络,其储存R的存档和代码的相同最新版本。

[0435] 模块可变换数据和/或信息。数据和/或信息可为合适的形式。例如数据和/或信息可为数字或模拟的。在一些实施方式中,数据和/或信息可为包裹、字节、字符或比特。在一些实施方式中,数据和/或信息可为任何收集的、集合的或有用的数据或信息。数据和/或信息的非限制性示例包括合适的媒介、图片、视频、声音(例如听得见或听不见的频率)、数字、常量、值、物体、时间、文字、函数、指令、计算机代码、图谱、参照、序列、读数、映射的读数、读数密度、读数密度概况、范围、阈值、显示、表示、变换等或其组合。模块可接受或接收数据和/或信息、变换数据和/或信息为第二形式、和提供或转移该第二形式至机器、外围设备、组件或其他模块。模块可进行一种或多种下述非-限制性功能:例如映射序列读数、压缩文件(例如,映射的读数数据)、过滤部份、选择部份、进行PCA、提供主成分、调整读数密度和/或读数密度概况、加权部份、打分、提供计数、集合部份、标准化计数、提供局部基因组偏好评价、提供偏好频率、提供读数密度概况、提供判定区和/或无判定区、提供不确定性测量、提供或确定预期范围(例如,阈值范围和阈值水平)、作图、和/或确定结果。在某些实施方式中,处理器可进行模块中的指令。在一些实施方式中,需要一种或多种处理器进行模块或模块组中的指令。模块可提供数据和/或信息至其他模块、设备或源并可接收其他模块、设备或源的数据和/或信息。

[0436] 非-瞬时计算机可读存储介质有时包括其上面的可执行程序并且有时该程序对微处理器下达指令进行功能(例如,本文所述的方法)。计算机程序产品有时物化在实体计算机-可读介质上、和有时实体物化在非-瞬时计算机-可读介质上。模块有时储存在计算机可读介质(例如磁盘、驱动器)或存储器中(例如随机存取存储器)。能执行来自模块的指令的模块和处理器可位于机器内或不同设备内。能执行模块的指令的模块和/或处理器可位于用户的相同位置(例如局部网络)或用户的不同位置(例如远程网络、云系统)。方法与两个或更多模块联合进行的实施方式中,模块可位于相同设备、一种或多种模块可位于相同物理位置中的不同设备中、和一种或多种模块可位于不同物理位置中的不同设备中。

[0437] 在一些实施方式中,装置包括至少一个处理器用于进行模块中的指令。映射至参照基因组的部份的序列读数计数有时通过处理器存取,所述处理器运行进行本文所述方法的指令。通过处理器存取的计数可在系统的存储器内,且可在获取计数后对其进行存取并将其置于和系统的存储器内。在一些实施方式中,机器包括处理器(例如一种或多种处理器),该处理器可进行和/或执行模块的一种或多种指令(例如过程、例程和/或子例程)。在一些实施方式中,机器包括多种处理器,例如协同和并行工作的处理器。在一些实施方式中,装置用一种或多种外部处理器操作(例如内部或外部网络、服务器、储存设备和/或储存网络(例如云))。在一些实施方式中,装置包括模块。在某些实施方式中,装置包括一个或多个模块。装置包括模块通常可从其他模块接受和传递一种或多种数据和/或信息和至其他模块。在某些实施方式中,装置包括外围设备和/或组件。在某些实施方式中装置可包括一种或多种外围设备或组件,其可将数据和/或信息传递至其他模块、外围设备和/或组件和从其传递。在某些实施方式中,装置与提供数据和/或信息的外围设备和/或组件交互。在某些实施方式中外围设备和组件协助装置行使功能或直接与模块交互。外设和/或组件的非

限制性示例包括合适的计算机外设、I/O或储存方法或设备,包括但不限于扫描仪、打印机、显示器(例如监视器、LED、LCT或CRT)、相机、麦克风、平板电脑(例如、书写板)、触摸屏、智能手机、移动电话、USB I/O设备、USB存储器、键盘、计算机鼠标、数码笔、调制解调器、硬盘、跳跃引擎、闪存驱动器、处理器、服务器、CD、DVD、图形卡、专用I/O设备(例如序列发生器、光电电池、光电放大管、光阅读器、感应器、等)、一种或多种流动槽、液体处理部件、网络交互控制器、ROM、RAM、无线传递方法和设备(蓝牙、WiFi等)、万维网(www)、网络、计算机和/或其他模块。

[0438] 软件常在含有程序指令的程序产品上提供,所述指令记录在计算机可读介质(例如,非-瞬时计算机可读介质)上,包括但不限于,磁性介质包括软盘、硬盘和磁带;和光学介质包括CD-ROM盘、DVD盘、磁光盘、固态驱动器、闪存驱动器、RAM、ROM、BUS等和其它可记录所述程序指令的此类介质。在线执行中,由组织维持的服务器和网站能被设置成向远端用户提供软件下载,或者远端用户可以使用由组织维持的远端系统来远端获取软件。软件可获得或接收输入信息。软件可以包含特定获得或接收数据的模块(如接收序列读取数据和/或定位的读取数据的数据接收模块)并且可以包含特定处理数据的模块(如处理数据的处理模块(如过滤器、标准化、提供结果和/或报告)。术语“获得”和“接收”输入信息指通过计算机通信方式从本地或远端位点、人工数据输入或任意其它接收数据的方法来接收数据(如序列读数、定位读数)。输入信息可在其接收的相同位置生成,或其可在不同位置生成并传送至接收位置。在一些实施方式中,输入信息在处理前经修饰(例如置于利于处理的形式中(例如表格))。

[0439] 在某些实施方式中,软件能包含一种或多种算法。算法可以用于根据有限的序列指令来处理数据和/或提供结果或报告。算法经常是用于完成任务的定义指令表。从起始状态开始,所述指令可以描述通过定义的一系列连续的状况进行并且以最终结束状态终止的计算。从一个状态转换成下一个状态不必是确定性的(例如,一些算法会纳入任意性)。作为非限定性示例,算法能是搜索算法、分类算法、归并算法、数值算法、图解算法、字符串搜索算法、建模算法、计算几何(几何学)算法、组合算法、机器学习算法、密码术算法、数据压缩算法、分析算法等。算法能包含一种算法或者两种或更多种算法的组合应用。算法可以是任意合适的复杂性分类和/或参数化的复杂性。算法能用于计算和/或数据处理,而在一些实施方式中能用在确定性的或盖然论/预测方法中。算法能通过使用合适的程序语言(非限定性示例是C、C++、Java、Perl、R、S、Python、Fortran等)植入到计算机环境中。在一些实施方式中,算法能构建或改进成包含误差容限、统计学分析、统计学显著性、不确定性测量和/或其它信息或数据组的比较(如当使用神经网络或簇算法时的应用)。

[0440] 在某些实施方式中,可将若干算法植入软件以便于使用。在一些实施方式中,这些算法能采用原始数据训练。对各种新的原始数据样品而言,所述经训练的算法可以生成代表性的处理的数据组或结果。相较经处理的父数据组(亲本数据集)而言,经处理的数据组有时复杂性降低。在一些实施方式中,基于经处理的数据组,能根据灵敏度和特异性来评价经训练的算法的实现。在某些实施方式中,可以鉴定并利用有最高灵敏度和/或特异性的算法。

[0441] 在某些实施方式中,模拟数据可协助数据处理,例如通过算法的训练或算法的测试。在一些实施方式中,模拟数据包含不同组序列读数的多种假定取样。模拟数据可基于真

实群体中可能的预期情况或可被歪曲以测试算法和/或分配正确的分类。模拟数据本文中还可指“实质”数据。在某些实施方式中模拟可通过计算机程序进行。使用模拟数据组中的一个可能步骤是评价经鉴定结果的置信度,如随机取样匹配或最佳代表原始数据的良好程度。一种方法是计算概率值(p值),该值评估随机样品比选定样品更好的概率。在一些实施方式中,可评估经验模型,其中假设至少一个样品匹配参照样品(具有或没有分辨出(resolved)的变异)。在一些实施方式中,其他分布例如泊松分布可用于定义概率分布。

[0442] 在某些实施方式中,系统可以包括一个或多个处理器。处理器可连接通信总线。计算机系统可包括主存储器(经常为随机读取存储器(RAM)),也可包括第二存储器。在一些实施方式中,存储器包括非-瞬时计算机-可读储存介质。二级存储器可包括例如,硬盘设备和/或可移除储存设备、代表软盘设备、磁带设备、光盘设备、存储卡等。移动存储驱动器经常读取和/或写入可移动的储存单元。可移动储存单元的非限定性示例包括能读取或写入例如移动存储驱动器的软盘、磁带、光盘等。可移动储存单元可包括其内储存有计算机软件和/或数据的计算机-可用的储存介质。

[0443] 处理器可以执行系统中的软件。在一些实施方式中,可以对处理器编程以自动运行本文所述用户可以进行的任务。因此,处理器或者由这种处理器执行的算法能几乎不需要监控至没有监控或者来自用户的输入(如可以编写软件以自动化实施功能)。在一些实施方式中,所述处理具有很大复杂性以至于单个个人或一组人不能在足够短的时间范围内实行所述处理以确定是否存在遗传变异。

[0444] 在一些实施方式中,第二存储器可包括允许计算机程序的其它相似方式或装载到计算机系统的其它指令。例如系统可包括可移除储存单元和交互设备。这种系统的非限定性示例可包括程序模块和模块接口(例如视频游戏设备中发现的那种)、可移动存储芯片(例如EPROM或PROM)以及关联插座和允许软件和数据从可移动储存单元转移到计算机系统的其它可移动储存单元和接口。

[0445] 在一些实施方式中,一个实体可生成序列读数计数、映射序列读数至部份、计数映射读数、并在本文所述方法、系统、机器或计算机程序产物中使用计数的映射读数。在某些实施方式中,本文所述方法、系统、机器或计算机程序产品中,映射至部份的序列读数计数有时从一个实体转移至第二实体由其使用。

[0446] 在一些实施方式中,一个实体生成序列读数和第二实体映射那些序列读数至参照基因组中的部份。第二实体有时计数映射的读数和在本文所述方法、系统、装置或计算机程序产品中采用该计数的映射的读数。在某些实施方式中第二实体转移映射的读数至第三实体、和第三实体计数该映射的读数并在本文所述方法、系统、装置或计算机程序产品中采用该映射的读数。在某些实施方式中第二实体计数映射的读数并将该计数的映射的读数转移至第三实体、和第三实体计数在本文所述方法、系统、装置或计算机程序产品中采用该计数的映射的读数。在涉及第三实体的实施方式中,第三实体有时与第一实体相同。即,第一实体有时转移序列读数至第二实体,第二实体可映射序列读数至参照基因组的部份和/或计数该映射的读数,且和第二实体可转移该映射的和/或计数的读数至第三实体。第三实体有时可在本文所述方法、系统、装置或计算机程序产品中采用该映射的和/或计数的读数,其中所述第三实体有时与第一实体相同,而有时该第三实体与第一或第二实体不同。

[0447] 在一些实施方式中,一个实体从妊娠雌性获得血液、任选从血液(例如从血浆或血

清)分离核酸血液、和转移所述血液或核酸至第二实体,所述第二实体从所述核酸生成序列读数。

[0448] 图11显示计算环境510的非限制性示例,其中可执行本文所述各种系统、方法、算法、和数据结构。计算环境510仅为合适的计算环境的一个实施例,并非旨在限制本文所述系统、方法、和数据结构的使用或功能范围。计算环境510也不应解释为对计算环境510中所示的任何一种组件或其组合的任何依赖或需求。在某些实施方式中,可使用图11中所示的系统、方法、和数据结构的子集。本文所述系统、方法、和数据结构可用其他通常目的或特定目的大量计算系统环境或配置进行操作。已知的合适计算系统、环境、和/或配置的示例包括但不限于个人计算机、服务器计算机、瘦客户机、厚客户机、手持-或膝式设备、多处理器系统、基于微处理器的系统、顶置盒、可编程客户电子装置、网络PC、迷你计算机、主机计算机、包括任何上述系统或设备分布的计算环境等。

[0449] 图11的操作环境510包括常规目的的计算设备,其为计算机520形式,包括处理单元521、系统存储器522、和可操作地偶联各种系统组件(包括系统存储器522)至处理单元521的系统总线523。可仅有一个或可有多于一个处理单元521,从而计算机520的处理器包括单中央-处理单元(CPU)或多种处理单元,通常称为平行处理环境。计算机520可为传统计算机、分布计算机,或任何其他类型计算机。

[0450] 系统总线523可为任何数种总线结构,包括存储器总线或存储器控制器、外围设备总线、和局部总线,使用各种总线构架中的任何。系统存储器还可简单称为存储器,仅包括读数存储器(ROM) 524和随机存取存储器(RAM)。基本输入/输出系统(BIOS) 526存于ROM 524, BIOS包含例如在启动时协助在计算机520元件之间转移信息的基本例程。计算机520还可包括硬盘驱动器接口527用于从其读取和写入硬盘(未显示),磁盘驱动器528用于从其读取或写入可移除磁盘529,和光盘驱动器530用于从其读取或写入可移除光盘531,例如CD ROM或其他光学介质。

[0451] 硬盘驱动器527、磁盘驱动器528、和光盘驱动器530分别通过硬盘驱动器接口532,磁盘驱动器接口533、和光盘驱动器接口534连接至系统总线523。驱动器和其关联的计算机-可读介质提供计算机520的计算机-可读指令、数据结构、程序模块和其他数据的固定存储器。任何类型的计算机可存取并可存储数据的计算机-可读介质例如磁盒、闪存卡、数码视频盘、贝努利筒、随机存取存储器(RAM)、只读存储器(ROM)等可用于该操作环境。

[0452] 许多程序模块可存于硬盘、磁盘529、光盘531、ROM 524或RAM上,包括操作系统535、一种或多种应用程序536、其他程序模块537、和程序数据538。用户可通过输入设备例如540和设备542键入命令和信息至个人计算机520。其他输入设备(未显示)可包括麦克风、操纵杆、游戏板、圆盘式卫星电视天线、扫描仪或等。这些和其他输入设备通常通过偶联至系统总线的串行端口接口546连接至处理单元521,但可不通过其他结构连接,例如并行端口、游戏端口或通用串行总线(USB)。监视器547或其他类型的显示设备还通过接口例如视频适配器548连接至系统总线523。除监视器外,计算机通常包括其他外围设备输出设备(未显示)例如扬声器和打印机。

[0453] 计算机520可在联网环境中操作,使用逻辑连接至一种或多种远程计算机、例如远程计算机549。这些逻辑连接可通过偶联至或部分的计算机520的通信设备或以其他方式实现。远程计算机549可以是其他计算机、服务器、路由器、网络PC、对等设备或其它常见的网

络节点,且一般包括以上关于计算机520描述的多个或所有的元件,虽然图11仅显示存储器储存设备550。图11中所描绘的逻辑连接包括局域网 (LAN) 551和广域网 (WAN) 552。这些联网环境在办公室、企业范围的计算机网络、内联网和因特网中是常见的。

[0454] 当在LAN联网环境中使用时,计算机520通过网络接口或适配器553连接到局域网551,其为一种通信设备。当在WAN联网环境中使用时,计算机520通常包括调制解调器554,一种通信设备,或用于通过广域网552建立通信的其它类型通信设备。调制解调器554可以是内置的或外置的,可经由串行端口接口546连接到系统总线523。在联网环境中,相关于计算机520所示的程序模块或其部分可被存储在远程存储器储存设备中。应当理解,所示的网络连接是示例性的,并且可使用在计算机之间建立通信链路的其它手段。

[0455] 在一些实施方式中,系统包括存储器和一个或多个微处理器,存储器包括可由一个或多个微处理器执行的指令并且可由一个或多个微处理器执行的质量设置为(a)生成测试样品的序列读数的(i)局部基因组偏好评价与(ii)偏好频率之间的关系,从而生成样品偏好关系,其中序列读数是来自测试样品的循环无细胞核酸,并且序列读数映射到参照基因组,(b)比较样品偏好关系和参照偏好关系,从而生成比较,其中参照偏好关系是(i)局部基因组偏好评价与(ii)参照的偏好频率之间的关系,并且(c)按照(b)中确定的比较将样品的序列读数的计数标准化,从而减少样品的序列读数中的偏好。

[0456] 在一些实施方式中,系统包括存储器和一个或多个微处理器,存储器包括可由一个或多个微处理器执行的指令并且可由一个或多个微处理器执行的指令设置为(a)生成测试样品的序列读数的(i)鸟嘌呤和胞嘧啶(GC)密度与(ii)GC密度频率之间的关系,从而生成样品GC密度关系,其中,

[0457] 序列读数是来自测试样品的循环无细胞核酸,并且

[0458] 序列读数映射至参照基因组,(b)比较样品GC密度关系和参照GC密度关系,从而生成比较,其中,

[0459] 参照GC密度关系是(i)GC密度与(ii)参照的GC密度频率之间的关系,并且(c)按照(b)中确定的比较将样品的序列读数的计数标准化,从而减少样品的序列读数中的偏好。

[0460] 在一些实施方式中,系统包括存储器和一个或多个微处理器,该存储器包括可由一个或多个微处理器执行的指令并且可由一个或多个微处理器执行的指令设置为(a)按照序列密度分布,过滤参照基因组的部份,从而提供包括过滤的部份的读数密度的测试样品的读数密度概况,其中使用映射至参照基因组的来自妊娠雌性的测试样品的循环无细胞核酸的序列读数确定读数密度并且确定多个样品的部份的读数密度的读数密度分布,(b)使用微处理器按照一个或多个主成分调整测试样品的读数密度概况,该主成分通过主成分分析获自己知的整倍性样品的组,从而提供包括调整的读数密度的测试样品概况,(c)比较测试样品概况与参照概况,从而提供比较,并且(d)按照比较确定测试样品是否存在染色体非整倍性。

[0461] 在一些实施方式中,本文提供了包括其上储存可执行程序的非-瞬时计算机可读存储介质。在一些实施方式中,其上包括可执行程序的非-瞬时计算机可读存储介质包括计算机程序产品。在一些实施方式中,其上包括可执行程序的非-瞬时计算机可读存储介质是指软件。计算机程序产品通常是软件。在一些实施方式中,本文提供一种非瞬时计算机可读存储介质,其上储存有可执行程序,其中所述程序对微处理器下指令,以进行如下操作:(a)

生成测试样品的序列读数的 (i) 鸟嘌呤和胞嘧啶 (GC) 密度与 (ii) GC 密度之间的关系, 从而生成样品 GC 密度关系, 其中序列读数是来自测试样品的循环无细胞核酸, 并且序列读数映射到参照基因组, (b) 比较样品 GC 密度关系和参照 GC 密度关系, 从而生成比较, 其中参照 GC 密度关系是 (i) GC 密度与 (ii) 参照的 GC 密度频率之间的关系, 并且 (c) 按照 (b) 中确定的比较将样品的序列读数的计数标准化, 从而减少样品的序列读数中的偏好。

[0462] 在一些实施方式中, 本文还提供一种非瞬时计算机可读存储介质, 其上储存有可执行的程序, 其中所述程序对微处理器下指令, 以进行如下操作: (a) 按照序列密度分布, 过滤参照基因组的部份, 从而提供包括过滤的部份的读数密度的测试样品的读数密度概况, 其中读数密度包括来自妊娠雌性的测试样品的循环无细胞核酸的序列读数并且确定多个样品的部份的读数密度的读数密度分布, (b) 按照一个或多个主成分调整测试样品的读数密度概况, 该主成分通过主成分分析获自己知的整倍性样品的组, 从而提供包括调整的读数密度的测试样品概况, (c) 比较测试样品概况与参照概况, 从而提供比较, 并且 (d) 按照比较确定测试样品是否存在染色体非整倍性。

[0463] 模块

[0464] 可在本文所述的方法中使用一个或多个模块, 其非限制性示例包括压缩模块、测序模块、映射模块、过滤模块、偏好密度模块、关系模块、偏好校正模块、多变量校正模块、分布模块、概况生成模块、PCA 统计模块、部份加权模块、打分模块、结果模块、显示模块等或其组合。在一些实施方式中, 模块是包括一组指令的非-瞬时计算机可读介质 (例如, 计算机程序产品, 例如, 软件, 程序), 其中这组指令指导一个或多个微处理器来实施功能。在一些实施方式中, 模块包括合适的计算机代码 (例如, 源代码) 形式的指令。源代码有时包括程序。计算机代码有时包括一个或多个文件 (例如, 文本文件)。可在合适的非-瞬时存储介质上储存计算机代码 (例如, 在存储器中, 例如, 在计算机硬盘上)。通常将计算机代码文件排列成目录树 (例如, 源树)。可以合适的编程语言编写模块的计算机代码, 其非限制性示例包括 C 编程语言、basic、R、R++、S、java、HTML 等或其组合。在一些实施方式中, 合适的主程序用作计算机代码的翻译器。在一些实施方式中, 模块包括和/或访问存储器。模块有时受微处理器控制。在某些实施方式中模块或包括一种或多种模块的装置从其他模块、装置、组件、装置的外围设备或操作器聚集、集合、接受、获取、存取、回收提供和/或转移数据和/或信息或至其他模块、装置、组件、装置的外围设备或操作器。在一些实施方式中, 通过包含一种或多种下述部件的装置将数据和/或信息 (例如, 序列读数, 计数等) 提供至模块: 一个或多个流动池、照相机、检测器 (例如, 光检测器、光电池、电子检测器 (例如, 正交调幅检测器、频率和相调节检测器、相锁定环检测器)、计数器、感应器 (例如压力、温度、体积、流动、重量的感应器)、流体操作装置、打印机、显示器 (例如 LED、LCT 或 CRT) 等或其组合。有时装置的操作器提供常量、阈值、公式或预定值至模块。模块通常配置为从其他模块或装置转移数据和/或信息或至其他模块或装置。模块可接收来自另一个模块的数据和/或信息, 其非限制性示例包括压缩模块、测序模块、映射模块、过滤模块、偏好密度模块、关系模块、偏好校正模块、多变量校正模块、分布模块、概况生成模块、PCA 统计模块、部份加权模块、打分模块、结果模块、显示模块等或其组合。模块可操作和/或变换数据和/或信息。来自模块或由模块变换的数据和/或信息可转移至另一个合适的机器和/或模块, 其非限制性示例包括压缩模块、测序模块、映射模块、过滤模块、偏好密度模块、关系模块、偏好校正模块、多变量校正模块、分布

模块、概况生成模块、PCA统计模块、部份加权模块、打分模块、结果模块、显示模块等或其组合。包括模块的机器可包括至少一个处理器。在一些实施方式中,数据和/或信息由包括模块的装置接受和/或提供。包括模块的机器包括处理器(例如一种或多种处理器),该处理器可进行和/或执行模块的一种或多种指令(例如过程、例程和/或子例程)。在一些实施方式中,模块用一种或多种外部处理器操作(例如内部或外部网络、服务器、储存设备和/或储存网络(例如云))。在一些实施方式中,系统(图10所示的系统的实施方式)包括压缩模块、测序模块、映射模块、过滤模块、偏好密度模块、关系模块、偏好校正模块、多变量校正模块、分布模块、概况生成模块、PCA统计模块、部份加权模块、打分模块、结果模块、显示模块等或其组合中的一个或多个。

[0465] 变换

[0466] 如上所述,数据有时从一种形式变换成另一种形式。本文使用术语“变换的”、“变换”和其语法衍生形式或等同形式指从物理起始材料(如测试对象和/或参照对象样品核酸)变成物理起始材料的数字表现(如序列读数数据)的数据变化,而在一些实施方式中,其包括将所述数字表示进一步变换成能用于提供结果的一种或多种数值或图示。在某些实施方式中,所述数字形式表示的数据的一种或多种数值和/或图示能用于表示测试对象的物理基因组的显示(如虚拟表示和/或直观表示是否存在基因组插入、复制或删除;表示是否存在与医学病症相关联的序列的物理量变化)。有时将虚拟表示进一步变换成所述起始材料的数字表示的一种或多种数值或图示。这些方法能把物理起始材料变换成数值或图示,或者表示测试对象基因组的物理存在。

[0467] 在一些实施方式中,本发明的方法和系统将在妊娠雌性的血液中发现的大量多核苷酸片段的混合物变换成胎儿、母体或胎盘细胞中存在的具体显微和/或亚显微结构(例如,染色体,或其区段)的一个或多个示意图。这些多核苷酸片段一般源自不同的细胞和组织(例如,母体、胎盘、胎儿,例如,肌肉、心脏、肝脏、淋巴细胞、肿瘤)、不同的染色体、和不同的遗传元件和/或位置(例如,着丝粒区、重复元件、富GC区、高变区、不同基因、不同调控元件、内含子、外显子等)。在一些实施方式中,本文所述的系统通过使用测序机器将多核苷酸片段变换成序列读数。在一些实施方式中,本文所述的系统将序列读数(其序列读数包括偏好)变换成标准化的序列计数、读数密度和/或概况。序列读数通常变换成标准化的序列计数、读数密度和/或概况,其中通常通过使用偏好减少机器和/或一种或多种合适的过程和/或模块(例如,映射模块、偏好密度模块、关系模块、偏好校正模块、和/或多变量校正模块)来显著减少偏好。具有减少的偏好的标准化的序列读数和从标准化的序列读数生成的读数密度和/或读数密度概况可用于生成置信度更高的结果。通常通过变换来改变序列读数,其改变具体序列读数参数并减少偏好,从而提供有时变换成概况和结果的标准化的序列读数。

[0468] 在一些实施方式中,变换数据组通过降低数据复杂性和/或数据维数来帮助提供结果。有时在将物理起始材料变换成所述起始材料的虚拟表示(如物理起始材料的序列读数表示)的处理过程中降低数据组复杂性。合适的特性或变量都能用于降低数据组的复杂性和/或维数。能选择用作数据处理的靶标特征的特征的非限定性示例包括GC含量、胎儿性别预测、染色体非整倍性鉴定、特定基因或蛋白质鉴定、癌症鉴定、疾病、遗传基因/性状、染色体异常、生物学分类、化学分类、生物化学分类、基因或蛋白质分类、基因本体学(gene

ontology)、蛋白质本体学 (protein ontology)、共调基因、细胞信号转导基因、细胞循环基因、与前述基因有关的蛋白质、基因变异、蛋白质变异、共调基因、共调蛋白、氨基酸序列、核苷酸序列、蛋白质结构数据等,及以上组合。降低数据组复杂性和/或维数的非限定性示例包括:使多重序列读数简化为分布图,使多重序列读数简化为数值(如标准化的值、Z-分值、p值);使多种分析方法简化为概率图或单点;对导出的量进行主成分分析等或其组合。

实施例

[0469] 仅以说明的形式而非限制的形式提供以下实施例。下述实施例说明某些实施方式但不限制本技术。本领域技术人员不难了解,可改变或调整各种非关键参数而获得基本相同或相似的结果。

[0470] -实施例1:ChAI

[0471] ChAI是用于从获自测试对象(例如,妊娠雌性)的序列读数中确定胎儿中是否存在染色体非整倍性的示例性系统。ChAI系统流程图的示例示于图10A和10B。序列读数获自妊娠雌性测试对象和一种或多种参照对象,本文有时指训练组。训练组妊娠雌性对象怀有其他测试方法证实为整倍体的胎儿。

[0472] 序列读数先从SAM或BAM格式压缩为二元读数格式(B读数格式),这可使ChAI运行更快。BReads格式存储各读数的基因组位置,包括根据参照基因组测定的染色体和碱基对位置,并丢弃其他信息。Breads文件从所含读数的计数开始。这样通过消除存储器重新配置的需要而改善装载时间。值以4字节阵列储存在盘上。然后使用5-字节格式储存读数,1个用于染色体序数(1-22,X,Y,M的0-索引),和4个用于染色体位置。通过先从前4比特读取序列读数计数来装载Breads文件。然后一次装载5字节各序列读数,第一字节表示染色体序数和后4字节转换为整数位置。通过使用针对特异读数索引的盘跳(disk-skip)命令可快速进行读数的随机取样。

[0473] 例如,表I显示17,673,732映射的读数的不同格式的磁盘用量相比BReads格式的磁盘用量。

[0474] 表1:基于具有17,673,732读数的样品的不同格式的磁盘用量。

[0475]	格式	空间用量
	SAM	4.0 GB
[0476]	映射的读数位置	247 MB
	GZip 读数位置	97 MB
	BReads	85 MB

[0477] BRead格式大约比原始SAM文件小50倍,比GZip格式使用约12%更少的空间。BRead还具有在一次存储器储存配置的开头储存读数数量的优点,并可快速取样,用于读数不必按顺序读取。其他格式不可能实现这些特征。

[0478] GC偏好建模

[0479] 然后各样品学习GC偏好模型。分配用于训练的样品部分用来产生部份的过滤器并

用来学习单独通过GC偏好无法良好解决的其他基因组偏好。最终,训练的统计数字用于过滤器和评分测试样品。

[0480] ChAI使用局部GC含量的密度评价来建模GC偏好。使用核函数例如Epanechnikov核(图1)从参照基因组评价GC密度。其他核同样适用,包括Gaussian或三加权(triweight)核。带宽选为200bp,然而带宽参数可灵活选用。

[0481] 使用核,GC密度以参照基因组上的碱基对分辨率进行评价(例如图2所示)。使用参照的GC密度评价,确定样品的各读数的局部GC含量。然后样品的GC密度评价的分布与跨全参照基因组的分布进行比较以确定GC偏好(图3)。映射至富含AT的区域的读数和参照值(GC密度=0)舍弃。

[0482] 样品的GC-密度分布和参照之间的差异使用多项式建模,在参照分布密度除以样品分布的密度的对数比上拟合(图4)。模型以加权形式拟合,就给定的GC-密度值的样品的分布-密度值取各权重。这保证分布的尾部不过分趋于所述拟合。其他拟合模型,例如分位数回归模型或参数化分布可用作偏好分布的合适模型。

[0483] 使用拟合GC模型,样品的各序列读数计数经加权以针对其相对参照的过度或不足表示进行调整。通过将这些权重纳入读数密度的评价,ChAI算法可就GC偏好进行校正。

[0484] 多维度偏好校正

[0485] GC偏好仅是基因组中数种偏好影响读数模式之一。其他偏好有时建模并校正以使用普遍的多元模型来评价读数权重。该校正如下进行:

[0486] 1.就测试样品和参照基因组在基因组位置的各子集处评价N偏好值。

[0487] 2.偏好值密度使用N-维平滑核或合适的参数函数进行建模。

[0488] 3.就从参照和测试密度所取的密度值组计算对数比。

[0489] 4.使用已选点和多元模型对密度对数比建模(例如就各维的加权的3阶多项式)。

[0490] 5.模型用于评价给定读数相比参照的频率比,并分配合适的权重。

[0491] 部份过滤

[0492] 就染色体异常基于基因组上的序列读数表示(例如计数)评分样品。该表示使用密度函数确定,类似于用于局部GC评价的那个。读数-密度核通常具有大得多的带宽,默认50,000bp。读数各计数贡献密度的值等于其GC-偏好模型的权重。读数-密度可在任何或所有碱基对处评估,但用于计算机性能仅使用某些位置。这些位置称为“部份”。部份可位于对评价读数-密度来说最终要的位置。为了分类染色体非整倍性,部份最初(例如过滤前)在基因组上平均分开。各部份包括50,000bp窗且在过滤前重叠下一邻近部份的25,000bp。

[0493] 一些部份包括弱映射的基因组区域,这导致样品间读数-密度的极端波动。ChAI使用训练组通过过滤过程鉴定和去除这些部份。显示较大中值偏差的部份(例如图5A)和/或MAD值(例如图5B)从考虑中去除。这些偏差的阈值取超出训练群四分位大于四倍四分位间范围的任何值(图5A)。该阈值可精细调节以就ChAI参数的特定组达到最大化测试性能。

[0494] 训练和评分

[0495] 仅使用映射至过滤的部份的读数,计算各样品的基因组读数-密度概况。然后作为训练组部分的样品用于评价训练统计数字,其用于打分测试组。这些统计数字由部份中值、主成分和打分测试统计的空分布组成。部份中值和主成分用于建模基因组-宽的读数偏好,其可从任何数量的生物和技术产品呈现(图6A-C)。为了最小化极端部份值对剩余样品的影

响,样品中其他部份中超出 $4 \times \text{IQR}$ 的各值修剪为 $4 \times \text{IQR}$ 。

[0496] 测试样品针对隐藏的偏好进行校正,通过先从测试部份值中减去训练的中值。与顶部训练主成分关联的样品值的成分也去除。这通过使用多元线性回归基于主成分项目对部份值建模(图7)。通过所述模型预测的值从样品值中减去,仅剩下无偏好的残差。所用主成分的数量任选,默认为8。

[0497] 校正后,样品使用Fisher-精确测试评分。该测试比较值大于或小于感兴趣染色体区域中的训练中值的部份的数量。这些计数针对基因组中的剩余进行评估。评分的统计数字取 $\log_{10} p$ -值的负数。其他评分统计数字可用在该步骤。例如Wilcoxon符号-评级测试或F-测试。

[0498] 由于部份之间的残差相关性,训练和测试样品中测试的统计数字均经夸大。这种夸大从训练组的拔靴法中评价(图8)。

[0499] 使用该空分布作为经验背景对测试样品的分数进行校正。远大于经验分布中分数的那些分数使用空分布尾部的Pareto外推法进行校正。

[0500] 判定性别

[0501] 从样品的主成分概况确定性别。在训练数据组中,第二主成分(如PC2)与性别高度关联。使用该成分的回归系数作为测试的统计数字是高度精确的性别测试(图9A-9B)。

[0502] 移除部份依赖

[0503] ChAI运行中采取其他步骤以改善方法的预期功效。这涉及降低部份-样品矩阵中的相关性结构的量,这更好地支持变量独立性的测试假设以及降低空置换中显著性分数的频率。该方法涉及用正交的本征部份替代部份,其含有几乎所有的相同信息,但没有相关性结构。

[0504] 第一步是就训练位置M的组学习变换矩阵 Meig :

[0505] 1. SVD分解: $M = U * D * V^T$

[0506] 2. 选择独立本征部份数量N: (例如,使得D的N对角元素的累积分数(cumulative fraction)大于95%)

[0507] 3. 计算伪逆矩阵: $\text{Meig} = \text{pinv}(U[\dots, 1:N] * D[1:N, 1:N])$

[0508] 任何部份矩阵M的子集左乘其对应 Meig 得到该子集的维度降低的无相关性表示。以此方法,在训练数据组上得到 Meig 并且将其应用于测试样品而无需进一步修饰。

[0509] Meig 还有用于变换测试变量。测试变量表示为由所有的0组成的向量,其中一些位于预期误差的位置(例如Chr 21部份)。该向量用 Meig 变换,通过左乘以合适匹配变换的部份的数据。

[0510] 该方法仅可产生训练组中样品数一样的独立本征部份。例如,对于50,000部份的训练组和1,000样品,变换的数据至多含1,000部份。这类似过度校正,显著降低部份的数量。该方法可通过就部份数据的更小子集计算分开的 Meig 转换并分开应用它们来更宽松地进行。这对于从邻近部份去除局部校正结构特别有用。

[0511] 还可用其他方法降低部份相关性结构。例如,许多簇方法可用于分组部份并用更小组的聚集部份替代(例如基于组平均或图心)。

[0512] 实施例2: 分布/概况生成模块

[0513] 在Java中写脚本用于从序列读数数据(如BReads)生成读数密度概况。分配下述代

码以收集各序列读数的读数数据并在合适的读书密度窗升级密度概况(如部份的个体读数密度),根据读数与部份中值或中点的距离并按照样品的GC偏好校正进行加权(参见实施例4)。下述脚本可判定或用于关系模块或偏好校正模块生成的加权的和/或标准化的计数(实施例4)。在一些实施方式中,分布模块可包括一些或所有下述java脚本或其变体。在一些实施方式中,概况生成模块可包括一些或所有下述java脚本或其变体:

```
[0514] package utilities.genome;  
import java.util.Iterator;
```

[0515]

```
import utilities.data.VectorUtil;
import utilities.text.DataFormatter;

public class ChromDensScaleRunnable implements Runnable{

    private GenomeScaleBoolean mask;
    private GenomeScaleFloat density;
    private final String modelPath;
    private final String brPath;
    private final int bandwidth;
    private final GenomeFloat gcdens;
    private final int stepSize;
    private final int sampleSize;
    private final int shift;
    private final String report;

    public ChromDensScaleRunnable(String modelPath, String brPath, int
bandwidth, GenomeFloat gcdens, int stepSize, GenomeScaleBoolean mask,
String report, int sampleSize, int shift)
    {
        this.modelPath = modelPath;
        this.brPath = brPath;
        this.bandwidth = bandwidth;
        this.gcdens = gcdens;
        this.stepSize = stepSize;
        this.mask = mask;
        this.report = report;
        this.sampleSize = sampleSize;
```

[0516]

```
        this.shift = shift;
    }

    public void run()
    {
        double[]    mdat    =    (gcdens==null)    ?    null    :
VectorUtil.loadDoubleFromFile(modelPath, 6);

        //Build density
        density = new GenomeScaleFloat(stepSize);

        double correction = 0;

        try
        {
            Iterator<GenomicPosition>  readIterator  = (sampleSize==-1) ?
GenomeIO.scanBReads(brPath)    :    new    BReadsSampler(brPath,
sampleSize,true);

            while (readIterator.hasNext())
            {
                GenomicPosition gp = readIterator.next().shift(shift);
                int pos = gp.pos;
                int start = Math.max(0, pos-bandwidth);
                int    end    =    Math.min(pos+bandwidth,
GenomeUtil.chromosomeSize(gp.chr)-1);

                int cindex = gp.chr.ordinal();
```

[0517]

```
double weight;
if (gcdens!=null)
{
    float gc = gcdens.values[cindex][pos-1];
    if (gc==0) continue;

    weight = modelWeight(mdat, gc);
} else weight = 1;

int[][] gpoints = density.getScalePoints(cindex, start, end,
mask);

if (gpoints[0].length==0 || Double.isNaN(weight)) continue;
if (weight>2) weight = 2;
if (weight<.5) weight = .5;

correction += weight;

for (int i=0;i<gpoints[0].length;i++)
    density.values[cindex][gpoints[0][i]] +=
kernel((gpoints[1][i]-pos)/(double)bandwidth) * weight;
}
} catch (Exception e)
{
    System.out.println("THROW!");
    e.printStackTrace();
    System.exit(0);
}
```

[0518]

```
//System.out.println(GenomeIO.countReadsFromBReads(brPath));
//System.out.println(correction);

//Normalize intensity
for (int i=0;i<density.values.length;i++)
    for (int j=0;j<density.values[i].length;j++)
    {
        float blah = density.values[i][j];

        density.values[i][j] /= correction;

        if (Double.isNaN(density.values[i][j]) ||
Double.isInfinite(density.values[i][j]))
        {
            System.out.println("NA    val2:    "+modelPath+",
"+density.values[i][j]+",    "+blah+",    "+correction);
            System.exit(0);
        }
    }

    if (report!=null) System.out.println(report);
}

public GenomeScaleFloat density()
{
    return density;
```

[0519]

```
    }

    private static double kernel(double x)
    {
        return .75 * (1.0 - x*x);
    }

    public static double modelWeight(double[] mdat, double gcdens)
    {
        if (mdat[5]==1) gcdens = Math.log(gcdens);

        double x2 = gcdens * gcdens;

        return Math.pow(2, mdat[0] + mdat[1] * gcdens + mdat[2] * x2 +
mdat[3] * x2 * gcdens);
    }

}
```

[0520] 实施例3:过滤模块

[0521] 在R中写脚本用于过滤读数密度概况的部份。基于四分位间范围,该代码在所有样品中检测读数密度概况并鉴定保留的部份和/或丢弃的部份(如从分析去移除)。在一些实施方式中,过滤模块可包括一些或所有下述R脚本或其变体:

[0522]

```
rcodepath <- "I:/ghannum/Projects/Binless/RCode"
mdistpath                                     <-
"I:/ghannum/Projects/Binless/Reference/MarkerDistribution_LDTv2_200_5000
0_50000.txt"
```

[0523]

```
      outpath                                     <-  
      "I:/ghannum/Projects/Binless/Reference/LDTv2_200_50000_50000_MarkerMa  
      sk.txt"
```

```
      args <- commandArgs(trailingOnly = TRUE)
```

```
      rcodepath <- args[1]
```

```
      mdistpath <- args[2]
```

```
      outpath <- args[3]
```

```
      source(paste(rcodepath,"/src/utilities/scanmatrix.R",sep=""))
```

```
      dat <- scanMatrix(mdistpath,rownames=FALSE,colnames=TRUE)
```

```
      m <- apply(dat,1,median)
```

```
      v <- apply(dat,1,mad)
```

```
      qm <- quantile(m,c(.25,.75))
```

```
      qv <- quantile(v,c(.25,.75))
```

```
      scalem <- qm[2]-qm[1]
```

```
      scalev <- qv[2]-qv[1]
```

```
      ok <- m > qm[1]-4*scalem & m < qm[2]+4*scalem & v > qv[1]-4*scalev  
& v < qv[2]+4*scalev
```

```
      write.table(matrix(as.integer(ok),1),row.names=F,col.names=F,quote=F,fi  
      le=outpath,sep="")
```

[0524] 实施例4: 偏好密度模块、关系模块、偏好校正模块和作图模块

[0525] R中编写脚本用于生产偏好密度、生成并比较关系并用于校正序列读数中的偏好。该代码通常涉及微处理器以分析一个或多个样品并基于局部基因组偏好评价(例如GC密

度)就各样品和参照建立偏好模型(例如关系和/或关系比较)。下述脚本部分涉及一种或多种处理器,以生成下述之间的关系:测试样品的序列读数的(i)鸟嘌呤和胞嘧啶(GC)密度与(ii)GC密度频率,从而生成样品GC密度关系,(b)比较样品GC密度关系和参照GC密度关系,从而生成比较,其中参照GC密度关系是(i)GC密度与(ii)参照的GC密度频率之间的关系,以及通过适当修改脚本而(c)根据(b)中确定的比较标准化样品的序列读数计数,其中降低样品的序列读数中的偏好。在一些实施方式中,偏好密度模块、关系模块、偏好校正模块和/或作图模块包括一些或所有下述R脚本或其变体。

[0526]

```
gcpath <-  
"I:/ghannum/Projects/Binless/Reference/BiasMaps/DnaseDensity_200_dist.txt"  
inpath <- "I:/ghannum/Projects/Binless/Models/LDTv2_DNase-200"  
outpath <- "I:/ghannum/Projects/Binless/Models/LDTv2_DNase-200"  
makePlots <- TRUE  
logTransform <- TRUE  
  
args <- commandArgs(trailingOnly = TRUE)  
gcpath <- args[1]  
inpath <- args[2]  
outpath <- args[3]  
makePlots <- args[4]  
logTransform <- as.logical(args[5])  
paths <- dir(inpath)  
paths <- paths[grepl("_BiasDistr.txt$",paths)]  
  
gcref <- scan(gcpath,0)
```

```
gcref <- gcref[gcref!=0]

if (logTransform) gcref <- log(gcref)

from <- quantile(gcref,.005)
to <- quantile(gcref,.995)

x <- seq(from,to,length.out=100);

d1y <- predict(smooth.spline(density(gcref,from=from,to=to)),x)$y
if (!logTransform) d1y <- sapply(d1y,function(x){max(x,0)})

print(paste("Processing",length(paths),"models."))

[0527] for (f in paths)
{
  distr <- scan(paste(inpath,"/",f,sep=""),0)
  distr <- distr[distr!=0]
  if (logTransform) distr <- log(distr)
  d2y <- predict(smooth.spline(density(distr,from=from,to=to)),x)$y
  if (!logTransform) d2y <- sapply(d2y,function(x){max(x,0)})
  pp <- log2(d1y / d2y)
  pp[pp > 2] <- 2; pp[pp < -2] <- -2
  ok <- !is.na(pp)
  mod <- lm(pp[ok]~x+I(x^2)+I(x^3), data=list(x=x[ok]), w=d2y[ok])
  w <- 2^predict(mod,list(x=distr))
  fname <- substr(f,1,nchar(f)-14)
  out <- c(mod$coefficients,mean(w))
  out[out==Inf] <- "Infinity"
```

```
[0528]
    out[out==-Inf] <- "-Infinity"
    write.table(matrix(c(out,as.integer(logTransform)),ncol=1),file=paste(outpath,"/",fname,"_BiasMod.txt",sep=""),row.names=F,col.names=F,quote=F)

    if (makePlots)
    {

        png(units="in",height=4,width=4,res=300,file=paste(outpath,"/",fname,"_BiasMod.png",sep=""))
        if (logTransform)
        {
            plot(x[ok],pp[ok],ylim=c(-4,4),xlab="Bias
Density",ylab="Log2 Ratio (Reference / Sample)")
        } else
            plot(x[ok],pp[ok],ylim=c(-4,4),xlab="Log-Bias
Density",ylab="Log2 Ratio (Reference / Sample)")
        abline(h=0,lty=2)
        lines(x[ok],predict(mod),col=3)
        dev.off()
    }
}

##Demo transformation
#load("I:/ghannum/Projects/Binless/2012_11_13_cewi_PERUN_19FCs_A
ltGCbias_chrFractions.RData")
#
#d <- dir("I:/ghannum/Projects/Binless/GCDistribution/LDTv2/")
#d <- d[grep("_GCDistr.txt",d)]
#v <- as.numeric(df.cewi.GCbiasTable[, "gcBiasRobust"])[1:length(d)]
#
```

```

[0529]
      #a <-
scan(paste("I:/ghannum/Projects/Binless/GCDistribution/LDTv2/",d[which.min
(v)],sep=""),0); a <- sort(a)
      #b <-
scan(paste("I:/ghannum/Projects/Binless/GCDistribution/LDTv2/",d[which.ma
x(v)],sep=""),0); d <- sort(d)
      #
      #r <-
scan("I:/ghannum/Projects/Binless/Reference/GCDensity_200_density.txt",0)
      #
      #plot(density(r),ylim=c(0,1e10),xlab="GC Density");
lines(density(a),col=3); lines(density(b),col=2)
      #
      #a <- a[a!=0]
      #b <- b[b!=0]
      #r <- r[r!=0]
      #
      #plot(density(r),ylim=c(0,1e10),xlab="GC Density");
lines(density(a),col=3); lines(density(b),col=2)
      #
      #modA <-
as.numeric(scan(paste("I:/ghannum/Projects/Binless/GCDistribution/LDTv2/",s
ubstr(d[which.min(v)],1,nchar(d[which.min(v)])-12),"_GCMod.txt",sep=""),""))
      #modB <-
as.numeric(scan(paste("I:/ghannum/Projects/Binless/GCDistribution/LDTv2/",s
ubstr(d[which.max(v)],1,nchar(d[which.max(v)])-12),"_GCMod.txt",sep=""),""
))
      #

```

[0530]

```
#wa <- sapply(a,function(x){2^sum(c(1,x,x^2,x^3)*modA[1:4])})  
#wb <- sapply(b,function(x){2^sum(c(1,x,x^2,x^3)*modB[1:4])})  
#  
#wa <- wa/(length(wa)*modA[5])  
#wb <- wb/(length(wb)*modB[5])  
#  
#plot(density(r),ylim=c(0,1e10),xlab="GC Density");  
lines(density(a,weights=wa),col=3); lines(density(b,weights=wb),col=2)
```

[0531] 实施例5:实施方式示例

[0532] 下述实施例说明某些实施方式但不限制本技术。

[0533] A1.一种包括存储器和一个或多个微处理器的系统,所述一个或多个微处理器设置为按照所述存储器中的指令进行用于减少样品的序列读数中的偏好的过程,所述过程包括:

[0534] (a)生成下述(i)和(ii)之间的关系:测试样品的序列读数的(i)鸟嘌呤和胞嘧啶(GC)密度与(ii)GC密度频率,从而生成样品GC密度关系,其中

[0535] 所述序列读数是来自测试样品的循环无细胞核酸,并且

[0536] 所述序列读数映射至参照基因组;

[0537] (b)将所述样品GC密度关系与参照GC密度关系比较,从而生成比较,其中,

[0538] 所述参照GC密度关系是参照的(i)GC密度与(ii)GC密度频率之间的关系;并且

[0539] (c)按照(b)中确定的比较将所述样品的序列读数的计数标准化,从而减少所述样品的序列读数中的偏好。

[0540] A1.1.一种包括测序设备和一个或多个计算设备的系统,

[0541] 所述测序设备设置成产生对应于加载到所述测序设备中的核酸的核苷酸碱基的信号,所述核酸是来自怀有胎儿的妊娠雌性的血液的循环无细胞核酸,或者加载到所述测序设备中的核酸是所述循环无细胞核酸的修饰变体;并且

[0542] 所述一个或多个计算设备包括存储器和一个或多个处理器,所述存储器包括可由所述一个或多个处理器执行的指令并且可由所述一个或多个处理器执行的指令设置为:

[0543] 从所述信号产生序列读数并且将所述序列读数映射至参照基因组;

[0544] (a)生成测试样品的序列读数的(i)鸟嘌呤和胞嘧啶(GC)密度与(ii)GC密度频率之间的关系,从而生成样品GC密度关系;

[0545] (b)将所述样品GC密度关系与参照GC密度关系比较,从而生成比较,其中,

[0546] 所述参照GC密度关系是参照的(i)GC密度与(ii)GC密度频率之间的关系;并且

[0547] (c)按照(b)中确定的比较将所述样品的序列读数的计数标准化,从而减少所述样品的序列读数中的偏好。

[0548] A1.2.如实施方式A1或A1.1所述的系统,其中(c)中的标准化包括提供标准化的计数。

[0549] A2.如实施方式A1至A1.2中任一项所述的系统,其中通过使用核心密度评价的过程来确定各所述GC密度。

[0550] A2.1.如实施方式A1至A2中任一项所述的系统,其中所述参照GC密度关系和所述样品GC密度关系的各GC密度代表局部GC密度。

[0551] A2.2.如实施方式A2.1所述的系统,其中所述局部GC含量针对5000bp或更小的多核苷酸区段。

[0552] A3.如实施方式A1至A2.2中任一项所述的系统,其中通过使用滑动窗分析的过程来确定各所述GC密度。

[0553] A4.如实施方式A3所述的系统,其中所述窗是约5个连续核苷酸至约5000个连续核苷酸,并且所述窗在滑动窗分析中一次滑动约1个碱基至约10个碱基。

[0554] A5.如实施方式A3所述的系统,其中所述窗是约200个连续核苷酸并且所述窗在滑动窗分析中一次滑动约1个碱基。

[0555] A6.如实施方式A1-A5中任一项所述的系统,其中(b)包括生成(i)和(ii)之间的拟合关系:(i)比例,各比例包括各所述GC密度的样品GC密度关系频率和参照GC密度关系频率,与(ii)GC密度。

[0556] A7.如实施方式A6所述的系统,其中(a)中的拟合关系获自加权的拟合。

[0557] A8.如实施方式A1-A7中任一项所述的系统,其中,所述样品的各序列读数表示为二进制格式和/或文本格式。

[0558] A9.如实施方式A8所述的系统,其中各所述序列读数的二进制格式包括所述读数映射至的染色体和所述读数映射至的染色体位置。

[0559] A10.如实施方式A9所述的系统,其中所述二进制格式是5-字节格式,包括1-字节染色体序数和4-字节染色体位置。

[0560] A11.如实施方式A8-A10中任一项所述的系统,其中所述二进制格式比序列比对/映射(SAM)格式小50倍和/或比GZip格式小13%。

[0561] A12.如实施方式A1-A11中任一项所述的系统,其中(c)中的标准化包括因数化除了GC密度以外的一个或多个特征,并且标准化所述序列读数。

[0562] A13.如实施方式A12中任一项所述的系统,其中所述因数化一个或多个特征是通过包括使用多变量模型的过程进行。

[0563] A14.如实施方式A13所述的系统,其中所述包括使用多变量模型的过程通过多变量模型进行。

[0564] A14.1.如实施方式A12-A14中任一项所述的系统,其中,按照(c)中的标准化和所述一个或多个特征的因数化将所述序列读数的计数标准化。

[0565] A15.如实施方式A1-A14.1中任一项所述的系统,包括在(c)之后,按照下述过程生成基因组的一个或多个部份或其区段的读数密度,所述过程包括生成含(c)中标准化的序列读数的计数的所述一个或多个部份各自的概率密度评价。

[0566] A16.如实施方式A15所述的系统,其中,所述概率密度评价是核心密度评价。

[0567] A17.如实施方式A16或A16所述的系统,包括生成所述基因组或其区段的读数密度概况。

[0568] A18.如实施方式A17所述的系统,其中,所述读数密度概况包括所述基因组的一个

或多个部份,或其区段的读数密度。

[0569] A19.如实施方式A15-A18中任一项所述的系统,包括调整所述一个或多个部份的各读数密度。

[0570] A20.如实施方式A15-A19中任一项所述的系统,其中,所述一个或多个部份经过滤从而提供过滤的部份。

[0571] A21.如实施方式A15-A20中任一项所述的系统,其中,所述一个或多个部份经加权从而提供加权的部份。

[0572] A22.如实施方式A21所述的系统,其中,所述一个或多个部份经本征函数加权。

[0573] A23.如实施方式A1-A22中任一项所述的系统,包括,在(a)之前获得所述序列读数。

[0574] A24.如实施方式A23所述的系统,其中,所述序列读数通过大规模平行测序(MPS)生成。

[0575] A25.如实施方式A1-A24中任一项所述的系统,包括获得映射至完整参照基因组或基因组区段的序列读数。

[0576] A26.如实施方式A25所述的系统,其中,所述基因组的区段包括染色体或其区段。

[0577] A27.如实施方式A25或A26所述的系统,其中,在(a)之前将映射至所述参照基因组的序列读数的计数标准化。

[0578] A28.如实施方式A27所述的系统,其中,由GC含量、箱式标准化、GC LOESS、PERUN、GCRM或其组合将映射至所述参照基因组的序列读数的计数标准化。

[0579] A29.如实施方式A27或A28中任一项所述的系统,其中,所述映射至所述参照基因组的序列读数的计数是原始计数。

[0580] A30.如实施方式A15-A29中任一项所述的系统,其中,所述参照基因组的各部份包括大约相等长度的连续核苷酸。

[0581] A31.如实施方式A15或A30中任一项所述的系统,其中,所述参照基因组的各部份包括约50kb。

[0582] A32.如实施方式A15-A31中任一项所述的系统,其中,所述参照基因组的各部份包括约100kb。

[0583] A33.如实施方式A15-A32中任一项所述的系统,其中,所述参照基因组的各部份包括与所述参照基因组的相邻部份一样的连续核苷酸的区段。

[0584] A34.如实施方式A1-A33中任一项所述的系统,其中,所述测试样品获自妊娠雌性。

[0585] A35.如实施方式A1-A34中任一项所述的系统,其中,所述测试样品包括来自妊娠雌性的血液。

[0586] A36.如实施方式A1-A35中任一项所述的系统,其中,所述测试样品包括来自妊娠雌性的血浆。

[0587] A37.如实施方式A1-A36中任一项所述的系统,其中,所述测试样品包括来自妊娠雌性的血清。

[0588] A38.如实施方式A1-A37中任一项所述的系统,其中,从所述测试样品分离核酸。

[0589] A39.如实施方式A8-A38中任一项所述的系统,包括将(a)中映射至参照基因组的序列读数从序列比对格式压缩成二进制格式。

- [0590] A40.如实施方式A39所述的系统,其中,所述压缩通过压缩模块进行。
- [0591] A41.如实施方式A1-A40中任一项所述的系统,其中通过偏好密度模块提供所述测试样品的序列读数和所述参照的GC密度和GC密度频率。
- [0592] A42.如实施方式A1-A41中任一项所述的系统,其中,由关系模块生成(b)中的比较。
- [0593] A43.如实施方式A1-A42中任一项所述的系统,其中,由偏好校正模块进行(c)中的标准化。
- [0594] A44.如实施方式A15-A43中任一项所述的系统,其中,由分布模块提供所述读数密度。
- [0595] A45.如实施方式A20-A44中任一项所述的系统,其中,由过滤模块提供过滤的部份。
- [0596] A46.如实施方式A21-A45中任一项所述的系统,其中,由读数密度调整模块提供调整的读数密度。
- [0597] A46.1.如实施方式A21-A46中任一项所述的系统,其中,由部份加权模块提供加权的部份。
- [0598] A47.如实施方式A46.1所述的系统,包括一个或多个压缩模块、偏好密度模块、关系模块、偏好校正模块、分布模块、过滤模块、读数密度调整模块和部份加权模块。
- [0599] A48.如实施方式A1-A47中任一项所述的系统,其中,所述系统的存储器包括来自所述测试样品的映射至所述参照基因组的循环无细胞核酸的序列读数。
- [0600] B1.一种包括存储器和一个或多个微处理器的系统,所述一个或多个微处理器设置为按照所述存储器中的指令进行用于确定样品是否存在非整倍性的过程,所述过程包括:
- [0601] (a)按照读数密度分布过滤参照基因组的部份,从而提供包括过滤的部份的读数密度的测试样品的读数密度概况,其中,
- [0602] 所述读数密度包括来自妊娠雌性的测试样品的循环无细胞核酸的序列读数,并且
- [0603] 针对多个样品的部份的读数密度确定所述读数密度分布;
- [0604] (b)按照一个或多个主成分调整测试样品的读数密度概况,所述主成分通过主成分分析获自己知的整倍性样品的组,从而提供包括调整的读数密度的测试样品概况;
- [0605] (c)将所述测试样品概况与参照概况比较,从而提供比较;并且
- [0606] (d)按照所述比较确定所述测试样品中是否存在染色体非整倍性。
- [0607] B1.1.一种包括测序设备和一个或多个计算设备的系统,
- [0608] 所述测序设备设置成产生对应于加载到所述测序设备中的核酸的核苷酸碱基的信号,所述核酸是来自怀有胎儿的妊娠雌性的血液的循环无细胞核酸,或者加载到所述测序设备中的核酸是所述循环无细胞核酸的修饰变体;并且
- [0609] 所述一个或多个计算设备包括存储器和一个或多个处理器,所述存储器包括可由所述一个或多个处理器执行的指令并且可由所述一个或多个处理器执行的指令设置为:
- [0610] 从所述信号产生序列读数并且将所述序列读数映射至参照基因组;
- [0611] (a)按照读数密度分布过滤参照基因组的部份,从而提供包括过滤的部份的读数密度的测试样品的读数密度概况,其中,

- [0612] 所述读数密度包括来自妊娠雌性的测试样品的循环无细胞核酸的序列读数,并且
- [0613] 针对多个样品的部份的读数密度确定所述读数密度分布;
- [0614] (b) 按照一个或多个主成分调整测试样品的读数密度概况,所述主成分通过主成分分析获自己知的整倍性样品的组,从而提供包括调整的读数密度的测试样品概况;
- [0615] (c) 将所述测试样品概况与参照概况比较,从而提供比较;并且
- [0616] (d) 按照所述比较确定所述测试样品中是否存在染色体非整倍性。
- [0617] B2.如实施方式B1或B1.1所述的系统,其中,所述比较包括确定显著性水平。
- [0618] B3.如实施方式B1-B2中任一项所述的系统,其中,确定所述显著性水平包括确定p-值。
- [0619] B4.如实施方式B1-B3中任一项所述的系统,其中,所述参照概况包括获自己知整倍性样品的组的读数密度概况。
- [0620] B5.如实施方式B1-B4中任一项所述的系统,其中,所述参照概况包括过滤的部份的读数密度。
- [0621] B6.如实施方式B1-B5中任一项所述的系统,其中,所述参照概况包括按照所述一个或多个主成分调整的读数密度。
- [0622] B7.如实施方式B2-B6中任一项所述的系统,其中,所述显著性水平表示所述测试样品概况和所述参照概况之间有统计学显著差异,并且确定存在染色体非整倍性。
- [0623] B8.如实施方式B1-B7中任一项所述的系统,其中,所述多个样品包括已知整倍性样品的组。
- [0624] B9.如实施方式B1-B8中任一项所述的系统,其中,所述多个样品的部份的读数密度是中值读数密度。
- [0625] B10.如实施方式B1-B9中任一项所述的系统,其中,所述测试样品的过滤的部份的读数密度是中值读数密度。
- [0626] B11.如实施方式B4-B10中任一项所述的系统,其中,所述参照概况的读数密度概况包括中值读数密度。
- [0627] B12.如实施方式B4-B11中任一项所述的系统,其中,按照包括核心密度评价的过程确定所述测试样品概况、所述多个样品和所述参照概况的读数密度。
- [0628] B13.如实施方式B10-B12中任一项所述的系统,其中,按照所述测试样品的中值读数密度确定所述测试样品概况。
- [0629] B14.如实施方式B11-B13中任一项所述的系统,其中,按照所述参照的中值读数密度分布确定所述参照概况。
- [0630] B15.如实施方式B1-B14中任一项所述的系统,包括按照针对所述读数密度分布的不确定性测量对参照基因组的部份进行过滤。
- [0631] B16.如实施方式B15所述的系统,其中,所述不确定性测量是MAD。
- [0632] B17.如实施方式B1-B16中任一项所述的系统,其中,映射至所述测试样品的过滤的部份的序列读数的计数通过在(a)之前进行的过程来标准化,包括:
- [0633] (I) 生成测试样品的序列读数的(i)局部基因组偏好评价与(ii)偏好频率之间的关系,从而生成样品偏好关系,其中,
- [0634] 序列读数是来自测试样品的循环无细胞核酸,并且

- [0635] 所述序列读数映射至参照基因组；
- [0636] (II) 将所述样品偏好关系与参照偏好关系比较，从而生成比较，其中，
- [0637] 所述参照偏好关系是参照的 (i) 局部基因组偏好评价与 (ii) 偏好频率之间的关系；并且
- [0638] (III) 按照 (II) 中确定的比较将所述样品的序列读数的计数标准化，从而减少所述样品的序列读数中的偏好。
- [0639] B18. 如实施方式B17所述的系统，其中 (III) 中的标准化包括提供标准化的计数。
- [0640] B19. 如实施方式B17或B18所述的系统，其中通过包括核心密度评价的过程确定各局部基因组偏好评价。
- [0641] B20. 如实施方式B17-B19中任一项所述的系统，其中通过包括使用滑动窗分析的过程来确定各所述局部基因组偏好评价。
- [0642] B21. 如实施方式B20所述的系统，其中所述窗是约5个连续核苷酸至约5000个连续核苷酸，并且所述窗在滑动窗分析中一次滑动约1个碱基至约10个碱基。
- [0643] B22. 如实施方式B20所述的系统，其中所述窗是约200个连续核苷酸并且所述窗在滑动窗分析中一次滑动约1个碱基。
- [0644] B23. 如实施方式B17-B22中任一项所述的系统，其中 (II) 包括生成 (i) 和 (ii) 之间的拟合关系：(i) 比例，各比例包括各所述局部基因组偏好评价的样品偏好关系频率和参照偏好关系频率，与 (ii) 局部基因组偏好评价。
- [0645] B24. 如实施方式B23所述的系统，其中 (I) 中的拟合关系获自加权的拟合。
- [0646] B25. 如实施方式B17-B24中任一项所述的系统，其中，所述样品的各序列读数表示为二进制格式。
- [0647] B26. 如实施方式B25所述的系统，其中各所述序列读数的二进制格式包括所述读数映射至的染色体和所述读数映射至的染色体位置。
- [0648] B27. 如实施方式B26所述的系统，其中所述二进制格式是5-字节格式，包括1-字节染色体序数和4-字节染色体位置。
- [0649] B28. 如实施方式B25-B27中任一项所述的系统，其中所述二进制格式比序列比对/映射 (SAM) 格式小50倍和/或比GZip格式小13%。
- [0650] B29. 如实施方式B17-B28中任一项所述的系统，其中 (III) 中的标准化包括因数化除了偏好以外的一个或多个特征，并且标准化所述序列读数的计数。
- [0651] B30. 如实施方式B29中任一项所述的系统，其中所述因数化一个或多个特征是通过包括使用多变量模型的过程进行。
- [0652] B31. 如实施方式B30所述的系统，其中所述包括使用多变量模型的过程通过多变量模型进行。
- [0653] B32. 如实施方式B29-B31中任一项所述的系统，其中，按照 (III) 中的标准化和所述一个或多个特征的因数化将所述序列读数的计数标准化。
- [0654] B33. 如实施方式B17-B32中任一项所述的系统，包括在 (III) 之后，按照包括生成基因组的一个或多个部份各自的概率密度评价的过程，生成所述一个或多个部份或其区段的读数密度，所述概率密度评价包括 (III) 中标准化的序列读数的计数。
- [0655] B34. 如实施方式B33所述的系统，其中，所述概率密度评价是核心密度评价。

[0656] B35.如实施方式B33或B34所述的系统,包括生成所述基因组或其区段的读数密度概况。

[0657] B36.如实施方式B35所述的系统,其中,所述读数密度概况包括所述基因组的一个或多个部份,或其区段的读数密度。

[0658] B37.如实施方式B33-B36中任一项所述的系统,包括调整所述一个或多个部份的各读数密度。

[0659] B38.如实施方式B33-B37中任一项所述的系统,其中,所述一个或多个部份经过滤从而提供过滤的部份。

[0660] B39.如实施方式B33-B38中任一项所述的系统,其中,所述一个或多个部份经加权从而提供加权的部份。

[0661] B40.如实施方式B39所述的系统,其中,所述一个或多个部份经本征函数加权。

[0662] B41.如实施方式B17-B40中任一项所述的系统,其中所述局部基因组偏好评价是局部GC密度并且所述偏好频率是GC偏好频率。

[0663] B42.如实施方式B1-B16中任一项所述的系统,其中,映射至所述测试样品的过滤的部份的序列读数的计数通过在(a)之前进行的过程来标准化,包括:

[0664] (1)生成所述测试样品的序列读数的(i)鸟嘌呤和胞嘧啶(GC)密度与(ii)GC密度频率之间的关系,从而生成样品GC密度关系,其中所述序列读数映射至所述参照基因组;

[0665] (2)将所述样品GC密度关系与参照GC密度关系比较,从而生成比较,其中

[0666] 所述参照GC密度关系是参照的(i)GC密度与(ii)GC密度频率之间的关系;并且

[0667] (3)按照(b)中确定的比较将所述样品的序列读数的计数标准化,从而减少所述样品的序列读数中的偏好。

[0668] B43.如实施方式B42所述的系统,其中(3)中的标准化包括提供标准化的计数。

[0669] B44.如实施方式B42或B43所述的系统,其中通过包括使用核心密度评价的过程确定各所述GC密度。

[0670] B44.1.如实施方式B42-B44中任一项所述的系统,其中所述参照GC密度关系和所述样品GC密度关系的各GC密度代表局部GC密度。

[0671] B44.2.如实施方式B44.1所述的系统,其中所述局部GC含量针对5000bp或更小的多核苷酸区段。

[0672] B45.如实施方式B42至B44.2中任一项所述的系统,其中通过使用滑动窗分析的过程来确定各所述GC密度。

[0673] B46.如实施方式B45所述的系统,其中所述窗是约5个连续核苷酸至约5000个连续核苷酸,并且所述窗在滑动窗分析中一次滑动约1个碱基至约10个碱基。

[0674] B47.如实施方式B46所述的系统,其中所述窗是约200个连续核苷酸并且所述窗在滑动窗分析中一次滑动约1个碱基。

[0675] B48.如实施方式B42-B47中任一项所述的系统,其中(2)包括生成(i)和(ii)之间的拟合关系:(i)比例,各比例包括各所述GC密度的样品GC密度关系频率和参照GC密度关系频率,与(ii)GC密度。

[0676] B49.如实施方式B48所述的系统,其中(1)中的拟合关系获自加权的拟合。

[0677] B50.如实施方式B42-B49中任一项所述的系统,其中,所述样品的各序列读数表示

为二进制格式。

[0678] B51.如实施方式B50所述的系统,其中各所述序列读数的二进制格式包括所述读数映射至的染色体和所述读数映射至的染色体位置。

[0679] B52.如实施方式B51所述的系统,其中所述二进制格式是5-字节格式,包括1-字节染色体序数和4-字节染色体位置。

[0680] B53.如实施方式B50-B52中任一项所述的系统,其中所述二进制格式比序列比对/映射(SAM)格式小50倍和/或比GZip格式小13%。

[0681] B54.如实施方式B42-B53中任一项所述的系统,其中(c)中的标准化包括因数化除了GC密度以外的一个或多个特征,并且标准化所述序列读数。

[0682] B55.如实施方式B54中任一项所述的系统,其中所述因数化一个或多个特征是通过包括使用多变量模型的过程进行。

[0683] B56.如实施方式B55所述的系统,其中所述包括使用多变量模型的过程通过多变量模型进行。

[0684] B57.如实施方式B42-B56中任一项所述的系统,其中对所述测试样品的过滤的部份加权。

[0685] B58.如实施方式B57所述的系统,其中所述测试样品的过滤的部份通过包括本征函数的过程加权。

[0686] B59.如实施方式B1-B58中任一项所述的系统,包括,在(a)之前获得所述序列读数。

[0687] B60.如实施方式B59所述的系统,其中,所述序列读数通过大规模平行测序(MPS)生成。

[0688] B61.如实施方式B1-B60中任一项所述的系统,包括获得映射至完整参照基因组或基因组区段的序列读数。

[0689] B62.如实施方式B61所述的系统,其中,所述基因组的区段包括染色体或其区段。

[0690] B63.如实施方式B61或B62所述的系统,其中,在(1)之前将映射至所述参照基因组的序列读数的计数标准化。

[0691] B64.如实施方式B63所述的系统,其中,由GC含量、箱式标准化、GC LOESS、PERUN、GCRM或其组合将映射至所述参照基因组的序列读数的计数标准化。

[0692] B65.如实施方式B61或B62所述的系统,其中,映射至所述参照基因组的序列读数的计数是原始计数。

[0693] B66.如实施方式B1-B65中任一项所述的系统,其中,所述参照基因组的各部份包括大约相等长度的连续核苷酸。

[0694] B67.如实施方式B1-B66中任一项所述的系统,其中,所述参照基因组的各部份包括约50kb。

[0695] B68.如实施方式B1-B67中任一项所述的系统,其中,所述参照基因组的各部份包括约100kb。

[0696] B69.如实施方式B1-B68中任一项所述的系统,其中,所述参照基因组的各部份包括与所述参照基因组的相邻部份一样的连续核苷酸的区段。

[0697] B70.如实施方式B1-B69中任一项所述的系统,其中,所述测试样品包括来自妊娠

雌性的血液。

[0698] B71.如实施方式B1-B70中任一项所述的系统,其中,所述测试样品包括来自妊娠雌性的血浆。

[0699] B72.如实施方式B1-B71中任一项所述的系统,其中,所述测试样品包括来自妊娠雌性的血清。

[0700] B73.如实施方式B1-B72中任一项所述的系统,其中,从所述测试样品分离核酸。

[0701] B74.如实施方式B50-B73中任一项所述的系统,包括将(1)中映射至参照基因组的序列读数从序列比对格式压缩成二进制格式。

[0702] B75.如实施方式B74所述的系统,其中,所述压缩通过压缩模块进行。

[0703] B76.如实施方式B42-B75中任一项所述的系统,其中通过偏好密度模块提供所述测试样品的序列读数和所述参照的GC密度和GC密度频率。

[0704] B77.如实施方式B42-B76中任一项所述的系统,其中,由关系模块生成(2)中的比较。

[0705] B78.如实施方式B44-B77中任一项所述的系统,其中,由偏好校正模块进行(3)中的标准化。

[0706] B79.如实施方式B1-B78中任一项所述的系统,其中,由分布模块提供所述读数密度。

[0707] B80.如实施方式B1-B79中任一项所述的系统,其中,由过滤模块提供过滤的部份。

[0708] B81.如实施方式B57-B80中任一项所述的系统,其中通过部份加权模块对所述测试样品的过滤的部份加权。

[0709] B81.1.如实施方式B57-B81中任一项所述的系统,其中,由读数密度调整模块调整所述读数密度。

[0710] B82.如实施方式B81.1所述的系统,其中,设备包括一个或多个压缩模块、偏好密度模块、关系模块、偏好校正模块、分布模块、过滤模块、读数密度调整模块和部份加权模块中。

[0711] B83.如实施方式B1-B82中任一项所述的系统,其中所述测试样品概况包括染色体或其区段的概况。

[0712] B84.如实施方式B1-B83中任一项所述的系统,其中所述参照概况包括染色体或其区段的概况。

[0713] B85.如实施方式B1-B84中任一项所述的系统,其中所述(d)中的确定具有等于或大于90%的特异性和等于或大于90%的灵敏度。

[0714] B86.如实施方式B1-B85中任一项所述的系统,其中非整倍性是三体性。

[0715] B87.如实施方式B86所述的系统,其中三体性是21三体性、18三体性、或13三体性。

[0716] B88.如实施方式B17-B87中任一项所述的系统,其中,所述系统的存储器包括来自所述测试样品的映射至所述参照基因组的循环无细胞核酸的序列读数。

[0717] C1.如实施方式A1-A48和B1-B88中任一项所述的系统,其体现在一个或多个机器中。

[0718] C2.如实施方式C1所述的系统,其体现在一个机器中。

[0719] C3.如实施方式C1或C2所述的系统,其包括设置为对核酸测序并生成序列读数的

机器。

[0720] D1.一种用于减少样品的序列读数的偏好的方法,包括:

[0721] (a) 使用微处理器生成测试样品的序列读数的(i)鸟嘌呤和胞嘧啶(GC)密度与(ii)GC密度频率之间的关系,从而生成样品GC密度关系,其中,

[0722] 序列读数是来自测试样品的循环无细胞核酸,并且

[0723] 所述序列读数映射至参照基因组;

[0724] (b) 将所述样品GC密度关系与参照GC密度关系比较,从而生成比较,其中

[0725] 所述参照GC密度关系是参照的(i)GC密度与(ii)GC密度频率之间的关系;并且

[0726] (c) 按照(b)中确定的比较将所述样品的序列读数的计数标准化,从而减少所述样品的序列读数中的偏好。

[0727] D1.1.一种用于减少样品的序列读数的偏好的方法,包括:

[0728] 用来自怀有胎儿的妊娠雌性的血液的循环无细胞核酸加载测序设备,或用所述核酸的修饰变体加载测序设备,所述测序设备产生对应于所述核酸的核苷酸碱基的信号;

[0729] 在任选地将所述信号转移至包括一个或多个计算设备的系统之后,由所述系统从所述核酸的信号生成序列读数,其中所述系统中的一个或多个计算设备包括存储器和一个或多个处理器,并且

[0730] 其中所述系统中一个计算设备,或计算设备的组合设置为:

[0731] 将所述序列读数映射至参照基因组;

[0732] (a) 生成测试样品的序列读数的(i)鸟嘌呤和胞嘧啶(GC)密度与(ii)GC密度频率之间的关系,从而生成样品GC密度关系,其中,

[0733] 序列读数是来自测试样品的循环无细胞核酸,并且

[0734] 所述序列读数映射至参照基因组;

[0735] (b) 将所述样品GC密度关系与参照GC密度关系比较,从而生成比较,其中

[0736] 所述参照GC密度关系参照的(i)GC密度与(ii)GC密度频率之间的关系;并且

[0737] (c) 按照(b)中确定的比较将所述样品的序列读数的计数标准化,从而减少所述样品的序列读数中的偏好。

[0738] D1.2.如实施方式D1或D1.1所述的方法,其中(c)中的标准化包括提供标准化的计数。

[0739] D2.如实施方式D1至D1.2中任一项所述的方法,其中通过使用核心密度评价的过程来确定各所述GC密度。

[0740] D2.1.如实施方式D1至D2中任一项所述的方法,其中所述参照GC密度关系和所述样品GC密度关系的各GC密度代表局部GC密度。

[0741] D2.2.如实施方式D2.1所述的方法,其中所述局部GC含量针对5000bp或更小的多核苷酸区段。

[0742] D3.如实施方式D1至D2.2中任一项所述的方法,其中通过使用滑动窗分析的过程来确定各所述GC密度。

[0743] D4.如实施方式D3所述的方法,其中所述窗是约5个连续核苷酸至约5000个连续核苷酸,并且所述窗在滑动窗分析中一次滑动约1个碱基至约10个碱基。

[0744] D5.如实施方式D3所述的方法,其中所述窗是约200个连续核苷酸并且所述窗在滑

动窗分析中一次滑动约1个碱基。

[0745] D6.如实施方式D1-D5中任一项所述的方法,其中(b)包括生成(i)和(ii)之间的拟合关系:(i)比例,各比例包括各所述GC密度的样品GC密度关系频率和参照GC密度关系频率,与(ii)GC密度。

[0746] D7.如实施方式D6所述的方法,其中(a)中的拟合关系获自加权的拟合。

[0747] D8.如实施方式D1-D7中任一项所述的方法,其中,所述样品的各序列读数表示为二进制格式。

[0748] D9.如实施方式D8所述的方法,其中各所述序列读数的二进制格式包括所述读数映射至的染色体和所述读数映射至的染色体位置。

[0749] D10.如实施方式D9所述的方法,其中所述二进制格式是5-字节格式,包括1-字节染色体序数和4-字节染色体位置。

[0750] D11.如实施方式D8-D10中任一项所述的方法,其中所述二进制格式比序列比对/映射(SAM)格式小50倍和/或比GZip格式小13%。

[0751] D12.如实施方式D1-D11中任一项所述的方法,其中(c)中的标准化包括因数化除了GC密度以外的一个或多个特征,并且标准化所述序列读数。

[0752] D13.如实施方式D12所述的方法,其中所述因数化一个或多个特征是通过包括使用多变量模型的过程进行。

[0753] D14.如实施方式D13所述的方法,其中所述包括使用多变量模型的过程通过多变量模型进行。

[0754] D14.1.如实施方式D12-D14中任一项所述的方法,其中,按照(c)中的标准化和所述一个或多个特征的因数化将所述序列读数的计数标准化。

[0755] D15.如实施方式D1-D14.1中任一项所述的方法,包括在(c)之后,按照下述过程生成基因组的一个或多个部份或其区段的读数密度,所述过程包括生成含(c)中标准化的序列读数的计数的所述一个或多个部份各自的概率密度评价。

[0756] D16.如实施方式D15所述的方法,其中,所述概率密度评价是核心密度评价。

[0757] D17.如实施方式D15或D16所述的方法,包括生成所述基因组或其区段的读数密度概况。

[0758] D18.如实施方式D17所述的方法,其中,所述读数密度概况包括所述基因组的一个或多个部份,或其区段的读数密度。

[0759] D19.如实施方式D15-D18中任一项所述的方法,包括调整所述一个或多个部份的各读数密度。

[0760] D20.如实施方式D15-D19中任一项所述的方法,其中,所述一个或多个部份经过滤从而提供过滤的部份。

[0761] D21.如实施方式D15-D20中任一项所述的方法,其中,所述一个或多个部份经加权从而提供加权的部份。

[0762] D22.如实施方式D21所述的方法,其中,所述一个或多个部份经本征函数加权。

[0763] D23.如实施方式D1-D22中任一项所述的方法,包括,在(a)之前获得所述序列读数。

[0764] D24.如实施方式D23所述的方法,其中,所述序列读数通过大规模平行测序(MPS)

生成。

[0765] D25.如实施方式D1-D24中任一项所述的方法,包括获得映射至完整参照基因组或基因组区段的序列读数。

[0766] D26.如实施方式D25所述的方法,其中,所述基因组的区段包括染色体或其区段。

[0767] D27.如实施方式D25或D26所述的方法,其中,在(a)之前将映射至所述参照基因组的序列读数的计数标准化。

[0768] D28.如实施方式D27所述的方法,其中,由GC含量、箱式标准化、GC LOESS、PERUN、GCRM或其组合将映射至所述参照基因组的序列读数的计数标准化。

[0769] D29.如实施方式D27或D28中任一项所述的方法,其中,所述映射至所述参照基因组的序列读数的计数是原始计数。

[0770] D30.如实施方式D15-D29中任一项所述的方法,其中,所述参照基因组的各部份包括大约相等长度的连续核苷酸。

[0771] D31.如实施方式D15或D30中任一项所述的方法,其中,所述参照基因组的各部份包括约50kb。

[0772] D32.如实施方式D15-D30中任一项所述的方法,其中,所述参照基因组的各部份包括约100kb。

[0773] D33.如实施方式D15-D32中任一项所述的方法,其中,所述参照基因组的各部份包括与所述参照基因组的相邻部份一样的连续核苷酸的区段。

[0774] D34.如实施方式D1-D33中任一项所述的方法,其中,所述测试样品获自妊娠雌性。

[0775] D35.如实施方式D1-D34中任一项所述的方法,其中,所述测试样品包括来自妊娠雌性的血液。

[0776] D36.如实施方式D1-D35中任一项所述的方法,其中,所述测试样品包括来自妊娠雌性的血浆。

[0777] D37.如实施方式D1-D36中任一项所述的方法,其中,所述测试样品包括来自妊娠雌性的血清。

[0778] D38.如实施方式D1-D37中任一项所述的方法,其中,从所述测试样品分离核酸。

[0779] D39.如实施方式D8-D38中任一项所述的方法,包括将(a)中映射至参照基因组的序列读数从序列比对格式压缩成二进制格式。

[0780] D40.如实施方式D39所述的方法,其中,所述压缩通过压缩模块进行。

[0781] D41.如实施方式D1-D40中任一项所述的方法,其中通过偏好密度模块提供所述测试样品的序列读数和所述参照的GC密度和GC密度频率。

[0782] D42.如实施方式D1-D41中任一项所述的方法,其中,由关系模块生成(b)中的比较。

[0783] D43.如实施方式D1-D42中任一项所述的方法,其中,由偏好校正模块进行(c)中的标准化。

[0784] D44.如实施方式D15-D43中任一项所述的方法,其中,由分布模块提供所述读数密度。

[0785] D45.如实施方式D20-D44中任一项所述的方法,其中,由过滤模块提供过滤的部份。

[0786] D46.如实施方式D21-D45中任一项所述的方法,其中,由部份加权模块提供加权的部份。

[0787] D46.1.如实施方式D21-D46中任一项所述的方法,其中,由读数密度调整模块调整读数密度。

[0788] D47.如实施方式D46.1所述的方法,包括一个或多个压缩模块、偏好密度模块、关系模块、偏好校正模块、分布模块、过滤模块、读数密度调整模块和部份加权模块。

[0789] E0.一种用于确定样品是否存在非整倍性的方法,包括:(a)按照读数密度分布过滤参照基因组的部份,从而提供包括过滤的部份的读数密度的测试样品的读数密度概况,其中,

[0790] 所述读数密度包括来自妊娠雌性的测试样品的循环无细胞核酸的序列读数,并且

[0791] 针对多个样品的部份的读数密度确定所述读数密度分布;

[0792] (b)按照一个或多个主成分调整测试样品的读数密度概况,所述主成分通过主成分分析获自己知的整倍性样品的组,从而提供包括调整的读数密度的测试样品概况;

[0793] (c)将所述测试样品概况与参照概况比较,从而提供比较;并且

[0794] (d)按照所述比较确定所述测试样品中是否存在染色体非整倍性。

[0795] E0.1.一种用于确定样品是否存在非整倍性的方法,包括:

[0796] (a)按照读数密度分布过滤参照基因组中染色体的部份,从而提供包括过滤的部份的读数密度的测试样品的读数密度概况,其中,

[0797] 所述读数密度包括来自妊娠雌性的测试样品的循环无细胞核酸的序列读数,并且

[0798] 针对多个样品的部份的读数密度确定所述读数密度分布;

[0799] (b)按照一个或多个主成分调整测试样品的染色体的读数密度概况,所述主成分通过主成分分析获自己知的整倍性样品的组,从而提供包括调整的读数密度的测试样品染色体概况;

[0800] (c)将所述测试样品染色体概况与参照概况比较,从而提供比较;并且

[0801] (d)按照所述比较确定所述测试样品中是否存在染色体非整倍性。

[0802] E1.一种用于确定样品是否存在非整倍性的方法,包括:

[0803] (a)按照读数密度分布过滤参照基因组的部份,从而提供包括过滤的部份的读数密度的测试样品的读数密度概况,其中,

[0804] 所述读数密度包括来自妊娠雌性的测试样品的循环无细胞核酸的序列读数,并且

[0805] 针对多个样品的部份的读数密度确定所述读数密度分布;

[0806] (b)使用微处理器按照一个或多个主成分调整测试样品的读数密度概况,所述主成分通过主成分分析获自己知的整倍性样品的组,从而提供包括调整的读数密度的测试样品概况;

[0807] (c)将所述测试样品概况与参照概况比较,从而提供比较;并且

[0808] (d)按照所述比较确定所述测试样品中是否存在染色体非整倍性。

[0809] E1.1.一种用于确定样品是否存在非整倍性的方法,包括:

[0810] 用来自怀有胎儿的妊娠雌性的血液的循环无细胞核酸加载测序设备,或用所述核酸的修饰变体加载测序设备,所述测序设备产生对应于所述核酸的核苷酸碱基的信号;

[0811] 在任选地将所述信号转移至包括一个或多个计算设备的系统之后,由所述系统从

所述核酸的信号生成序列读数,其中所述系统中的一个或多个计算设备包括存储器和一个或多个处理器,并且

[0812] 其中所述系统一个计算设备,或计算设备的组合设置为:

[0813] 将所述序列读数映射至参照基因组;

[0814] (a) 按照读数密度分布过滤参照基因组的部份,从而提供包括过滤的部份的读数密度的测试样品的读数密度概况,其中,

[0815] 所述读数密度包括来自妊娠雌性的测试样品的循环无细胞核酸的序列读数,并且

[0816] 针对多个样品的部份的读数密度确定所述读数密度分布;

[0817] (b) 使用微处理器按照一个或多个主成分调整测试样品的读数密度概况,所述主成分通过主成分分析获自己知的整倍性样品的组,从而提供包括调整的读数密度的测试样品概况;

[0818] (c) 将所述测试样品概况与参照概况比较,从而提供比较;并且

[0819] (d) 按照所述比较确定所述测试样品中是否存在染色体非整倍性。

[0820] E1.2. 一种用于减少样品的序列读数的偏好的方法,包括:

[0821] 用来自怀有胎儿的妊娠雌性的血液的循环无细胞核酸加载测序设备,或用所述核酸的修饰变体加载测序设备,所述测序设备产生对应于所述核酸的核苷酸碱基的信号;

[0822] 在任选地将所述信号转移至包括一个或多个计算设备的系统之后,由所述系统从所述核酸的信号生成序列读数,其中所述系统中的一个或多个计算设备包括存储器和一个或多个处理器,并且

[0823] 其中所述系统一个计算设备,或计算设备的组合设置为:

[0824] 将所述序列读数映射至参照基因组;

[0825] (a) 按照读数密度分布过滤参照基因组的部份,从而提供包括过滤的部份的读数密度的测试样品的读数密度概况,其中,

[0826] 所述读数密度包括来自妊娠雌性的测试样品的循环无细胞核酸的序列读数,并且

[0827] 针对多个样品的部份的读数密度确定所述读数密度分布;

[0828] (b) 使用微处理器按照一个或多个主成分调整测试样品的读数密度概况,所述主成分通过主成分分析获自己知的整倍性样品的组,从而提供包括调整的读数密度的测试样品概况;

[0829] (c) 将所述测试样品概况与参照概况比较,从而提供比较;并且

[0830] (d) 按照所述比较确定所述测试样品中是否存在染色体非整倍性。

[0831] E1.3. 如实施方式E0-E1.2中任一项所述的方法,其中在(b)中由1-10个主成分来调整所述读数密度概况。

[0832] E1.4. 如实施方式E0-E1.3中任一项所述的方法,其中在(b)中由5个主成分来调整所述读数密度概况。

[0833] E1.5. 如实施方式E0-E1.4中任一项所述的方法,其中针对读数密度概况中的一个或多个特征调整所述一个或多个主成分,其中特征选自胎儿性别、序列偏好、胎儿分数、与DNA酶I灵敏度相关的偏好、熵、重复序列偏好、染色质结构偏好、聚合酶误差率偏好、回文偏好、插入重复偏好、PCR扩增偏好、和隐藏拷贝数变异。

[0834] E1.6. 如实施方式E1.5所述的方法,其中序列偏好包括鸟嘌呤和胞嘧啶(GC)偏好。

- [0835] E2.如实施方式E0-E1.6中任一项所述的方法,其中所述比较包括确定显著性水平。
- [0836] E3.如实施方式E0-E2中任一项所述的方法,其中,确定所述显著性水平包括确定p-值。
- [0837] E4.如实施方式E0-E3中任一项所述的方法,其中,所述参照概况包括获自己知整倍性样品的组的读数密度概况。
- [0838] E5.如实施方式E0-E4中任一项所述的方法,其中,所述参照概况包括过滤的部份的读数密度。
- [0839] E6.如实施方式E0-E5中任一项所述的方法,其中,所述参照概况包括按照所述一个或多个主成分调整的读数密度。
- [0840] E7.如实施方式E2-E6中任一项所述的方法,其中,所述显著性水平表示所述测试样品概况和所述参照概况之间有统计学显著差异,并且确定存在染色体非整倍性。
- [0841] E8.如实施方式E1-E7中任一项所述的方法,其中,所述多个样品包括已知整倍性样品的组。
- [0842] E9.如实施方式E0-E8中任一项所述的方法,其中,所述多个样品的部份的读数密度是中值读数密度。
- [0843] E10.如实施方式E0-E9中任一项所述的方法,其中,所述测试样品的过滤的部份的读数密度是中值读数密度。
- [0844] E11.如实施方式E4-E10中任一项所述的方法,其中,所述参照概况的读数密度概况包括中值读数密度。
- [0845] E12.如实施方式E4-E11中任一项所述的方法,其中,按照包括核心密度评价的过程确定所述测试样品概况、所述多个样品和所述参照概况的读数密度。
- [0846] E13.如实施方式E10-E12中任一项所述的方法,其中,按照所述测试样品的中值读数密度确定所述测试样品概况。
- [0847] E14.如实施方式E11-E13中任一项所述的方法,其中,按照所述参照的中值读数密度分布确定所述参照概况。
- [0848] E15.如实施方式E0-E14中任一项所述的方法,包括按照针对所述读数密度分布的不确定性测量对参照基因组的部份进行过滤。
- [0849] E16.如实施方式E15所述的方法,其中,所述不确定性测量是MAD。
- [0850] E16.1.如实施方式E0-E16中任一项所述的方法,其中,所述测试样品概况代表所述测试样品的染色体剂量。
- [0851] E16.2.如实施方式E16.1所述的方法,包括比较测试样品概况的染色体剂量与参照概况的染色体剂量,从而生成染色体剂量比较。
- [0852] E16.3.如实施方式E16.2所述的方法,其中按照所述染色体剂量比较确定所述测试样品是否存在染色体非整倍性。
- [0853] E16.4.如实施方式E0-E16.3中任一项所述的方法,其中确定所述测试样品是否存在染色体非整倍性包括鉴定是否存在1个拷贝的染色体、2个拷贝的染色体、3个拷贝的染色体、4个拷贝的染色体、5个拷贝的染色体、染色体的一个或多个区段的缺失或染色体的一个或多个区段的插入。

[0854] E17.如实施方式E0-E16.4中任一项所述的方法,其中,映射至所述测试样品的过滤的部份的序列读数的计数通过在(a)之前进行的过程来标准化,包括:

[0855] (I)生成测试样品的序列读数的(i)局部基因组偏好评价与(ii)偏好频率之间的关系,从而生成样品偏好关系,其中,

[0856] 序列读数是来自测试样品的循环无细胞核酸,并且

[0857] 所述序列读数映射至参照基因组;

[0858] (II)将所述样品偏好关系与参照偏好关系比较,从而生成比较,其中,

[0859] 所述参照偏好关系是参照的(i)局部基因组偏好评价与(ii)偏好频率之间的关系;并且

[0860] (III)按照(II)中确定的比较将所述样品的序列读数的计数标准化,从而减少所述样品的序列读数中的偏好。

[0861] E18.如实施方式E17所述的方法,其中(III)中的标准化包括提供标准化的计数。

[0862] E19.如实施方式E17或E18所述的方法,其中通过包括核心密度评价的过程确定各局部基因组偏好评价。

[0863] E19.1.如实施方式E17-E19中任一项所述的方法,其中所述参照偏好关系和所述样品偏好关系的各局部基因组偏好评价代表局部偏好含量。

[0864] E19.2.如实施方式E19.1所述的方法,其中所述局部偏好含量针对5000bp或更小的多核苷酸区段。

[0865] E20.如实施方式E17-E19.2中任一项所述的方法,其中通过包括使用滑动窗分析的过程来确定各所述局部基因组偏好评价。

[0866] E21.如实施方式E20所述的方法,其中所述窗是约5个连续核苷酸至约5000个连续核苷酸,并且所述窗在滑动窗分析中一次滑动约1个碱基至约10个碱基。

[0867] E22.如实施方式E20所述的方法,其中所述窗是约200个连续核苷酸并且所述窗在滑动窗分析中一次滑动约1个碱基。

[0868] E23.如实施方式E17-E22中任一项所述的方法,其中(II)包括生成(i)和(ii)之间的拟合关系:(i)比例,各比例包括各所述局部基因组偏好评价的样品偏好关系频率和参照偏好关系频率,与(ii)局部基因组偏好评价。

[0869] E24.如实施方式E23所述的方法,其中(I)中的拟合关系获自加权的拟合。

[0870] E25.如实施方式E17-E24中任一项所述的方法,其中,所述样品的各序列读数表示为二进制格式。

[0871] E26.如实施方式E25所述的方法,其中各所述序列读数的二进制格式包括所述读数映射至的染色体和所述读数映射至的染色体位置。

[0872] E27.如实施方式E26所述的方法,其中所述二进制格式是5-字节格式,包括1-字节染色体序数和4-字节染色体位置。

[0873] E28.如实施方式E25-E27中任一项所述的方法,其中所述二进制格式比序列比对/映射(SAM)格式小50倍和/或比GZip格式小13%。

[0874] E29.如实施方式E17-E28中任一项所述的方法,其中(III)中的标准化包括因数化除了偏好以外的一个或多个特征,并且标准化所述序列读数的计数。

[0875] E30.如实施方式E29所述的方法,其中所述因数化一个或多个特征是通过包括使

用多变量模型的过程进行。

[0876] E31.如实施方式E30所述的方法,其中所述包括使用多变量模型的过程通过多变量模型进行。

[0877] E32.如实施方式E29-E31中任一项所述的方法,其中,按照(III)中的标准化和所述一个或多个特征的因数化将所述序列读数的计数标准化。

[0878] E33.如实施方式E17-E32中任一项所述的方法,包括在(III)之后,按照包括生成基因组的一个或多个部份各自的概率密度评价的过程,生成所述一个或多个部份或其区段的读数密度,所述概率密度评价包括(III)中标准化的序列读数的一个或多个计数。

[0879] E34.如实施方式E33所述的方法,其中,所述概率密度评价是核心密度评价。

[0880] E35.如实施方式E33或E34所述的方法,包括生成所述基因组或其区段的读数密度概况。

[0881] E36.如实施方式E35所述的方法,其中,所述读数密度概况包括所述基因组的一个或多个部份,或其区段的读数密度。

[0882] E37.如实施方式E33-E36中任一项所述的方法,包括调整所述一个或多个部份的各读数密度。

[0883] E38.如实施方式E33-E37中任一项所述的方法,其中,所述一个或多个部份经过滤从而提供过滤的部份。

[0884] E39.如实施方式E33-E38中任一项所述的方法,其中,所述一个或多个部份经加权从而提供加权的部份。

[0885] E40.如实施方式E39所述的方法,其中,所述一个或多个部份经本征函数加权。

[0886] E41.如实施方式E17-E40中任一项所述的方法,其中所述局部基因组偏好评价是局部GC密度并且所述偏好频率是GC偏好频率。

[0887] E42.如实施方式E0-E16中任一项所述的方法,其中,映射至所述测试样品的过滤的部份的序列读数的计数通过在(a)之前进行的过程来标准化,包括:

[0888] (1)生成所述测试样品的序列读数的(i)鸟嘌呤和胞嘧啶(GC)密度与(ii)GC密度频率之间的关系,从而生成样品GC密度关系,其中所述序列读数映射至所述参照基因组;

[0889] (2)将所述样品GC密度关系与参照GC密度关系比较,从而生成比较,其中

[0890] 所述参照GC密度关系是参照的(i)GC密度与(ii)GC密度频率之间的关系;并且

[0891] (3)按照(b)中确定的比较将所述样品的序列读数的计数标准化,从而减少所述样品的序列读数中的偏好。

[0892] E43.如实施方式E42所述的方法,其中(3)中的标准化包括提供标准化的计数。

[0893] E44.如实施方式E42或E43所述的方法,其中通过包括使用核心密度评价的过程确定各所述GC密度。

[0894] E45.如实施方式E42至E44中任一项所述的方法,其中通过使用滑动窗分析的过程来确定各所述GC密度。

[0895] E46.如实施方式E45所述的方法,其中所述窗是约5个连续核苷酸至约5000个连续核苷酸,并且所述窗在滑动窗分析中一次滑动约1个碱基至约10个碱基。

[0896] E47.如实施方式E46所述的方法,其中所述窗是约200个连续核苷酸并且所述窗在滑动窗分析中一次滑动约1个碱基。

[0897] E48.如实施方式E42-E47中任一项所述的方法,其中(2)包括生成(i)和(ii)之间的拟合关系:(i)比例,各比例包括各所述GC密度的样品GC密度关系频率和参照GC密度关系频率,与(ii)GC密度。

[0898] E49.如实施方式E48所述的方法,其中(1)中的拟合关系获自加权的拟合。

[0899] E50.如实施方式E42-E49中任一项所述的方法,其中,所述样品的各序列读数表示为二进制格式。

[0900] E51.如实施方式E50所述的方法,其中各所述序列读数的二进制格式包括所述读数映射至的染色体和所述读数映射至的染色体位置。

[0901] E52.如实施方式E51所述的方法,其中所述二进制格式是5-字节格式,包括1-字节染色体序数和4-字节染色体位置。

[0902] E53.如实施方式E50-E52中任一项所述的方法,其中所述二进制格式比序列比对/映射(SAM)格式小50倍和/或比GZip格式小13%。

[0903] E54.如实施方式E42-E53中任一项所述的方法,其中(c)中的标准化包括因数化除了GC密度以外的一个或多个特征,并且标准化所述序列读数。

[0904] E55.如实施方式E54所述的方法,其中所述因数化一个或多个特征是通过包括使用多变量模型的过程进行。

[0905] E56.如实施方式E55所述的方法,其中所述包括使用多变量模型的过程通过多变量模型进行。

[0906] E57.如实施方式E42-E56中任一项所述的方法,其中对所述测试样品的过滤的部份加权。

[0907] E58.如实施方式E57所述的方法,其中所述测试样品的过滤的部份通过包括本征函数的过程加权。

[0908] E59.如实施方式E0-E58中任一项所述的方法,包括,在(a)之前获得所述序列读数。

[0909] E60.如实施方式E59所述的方法,其中,所述序列读数通过大规模平行测序(MPS)生成。

[0910] E61.如实施方式E0-E60中任一项所述的方法,包括获得映射至完整参照基因组或基因组区段的序列读数。

[0911] E62.如实施方式E61所述的方法,其中,所述基因组的区段包括染色体或其区段。

[0912] E63.如实施方式E61或E62所述的方法,其中,在(1)之前将映射至所述参照基因组的序列读数的计数标准化。

[0913] E64.如实施方式E63所述的方法,其中,由GC含量、箱式标准化、GC LOESS、PERUN、GCRM或其组合将映射至所述参照基因组的序列读数的计数标准化。

[0914] E65.如实施方式E61或E62所述的方法,其中,映射至所述参照基因组的序列读数的计数是原始计数。

[0915] E66.如实施方式E0-E65中任一项所述的方法,其中,所述参照基因组的各部份包括大约相等长度的连续核苷酸。

[0916] E67.如实施方式E0-E66中任一项所述的方法,其中,所述参照基因组的各部份包括约50kb。

[0917] E68.如实施方式E0-E67中任一项所述的方法,其中,所述参照基因组的各部份包括约100kb。

[0918] E69.如实施方式E0-E68中任一项所述的方法,其中,所述参照基因组的各部份包括与所述参照基因组的相邻部份一样的连续核苷酸的区段。

[0919] E70.如实施方式E0-E69中任一项所述的方法,其中,所述测试样品包括来自妊娠雌性的血液。

[0920] E71.如实施方式E0-E70中任一项所述的方法,其中,所述测试样品包括来自妊娠雌性的血浆。

[0921] E72.如实施方式E0-E71中任一项所述的方法,其中,所述测试样品包括来自妊娠雌性的血清。

[0922] E73.如实施方式E0-E72中任一项所述的方法,其中,从所述测试样品分离核酸。

[0923] E74.如实施方式E50-E73中任一项所述的方法,包括将(1)中映射至参照基因组的序列读数从序列比对格式压缩成二进制格式。

[0924] E75.如实施方式E74所述的方法,其中,所述压缩通过压缩模块进行。

[0925] E76.如实施方式E42-E75中任一项所述的方法,其中通过偏好密度模块提供所述测试样品的序列读数和所述参照的GC密度和GC密度频率。

[0926] E77.如实施方式E42-E76中任一项所述的方法,其中,由关系模块生成(2)中的比较。

[0927] E78.如实施方式E44-E77中任一项所述的方法,其中,由偏好校正模块进行(3)中的标准化。

[0928] E79.如实施方式E0-E78中任一项所述的方法,其中,由分布模块提供所述读数密度。

[0929] E80.如实施方式E0-E79中任一项所述的方法,其中,由过滤模块提供过滤的部份。

[0930] E81.如实施方式E57-E80中任一项所述的方法,其中通过部份加权模块对所述测试样品的过滤的部份加权。

[0931] E81.1.如实施方式E57-E81中任一项所述的方法,其中,由读数密度调整模块调整所述读数密度。

[0932] E82.如实施方式E81.1所述的方法,其中,设备包括一个或多个压缩模块、偏好密度模块、关系模块、偏好校正模块、分布模块、过滤模块、读数密度调整模块和部份加权模块。

[0933] E83.如实施方式E0-E82中任一项所述的方法,其中所述测试样品概况包括染色体或其区段的概况。

[0934] E84.如实施方式E0-E83中任一项所述的方法,其中所述参照概况包括染色体或其区段的概况。

[0935] E85.如实施方式E0-E84中任一项所述的方法,其中所述(d)中的确定具有等于或大于90%的特异性和等于或大于90%的灵敏度。

[0936] E86.如实施方式E0-E85中任一项所述的方法,其中非整倍性是三体性。

[0937] E87.如实施方式E86所述的方法,其中所述三体性是21三体性、18三体性、或13三体性。

[0938] F1.一种非瞬时计算机可读存储介质,包括其上储存有可执行的程序,其中所述程序对微处理器下指令,以进行如下操作:

[0939] (a)生成测试样品的序列读数的(i)鸟嘌呤和胞嘧啶(GC)密度与(ii)GC密度频率之间的关系,从而生成样品GC密度关系,其中:

[0940] 序列读数是来自测试样品的循环无细胞核酸,并且

[0941] 所述序列读数映射至参照基因组;

[0942] (b)将所述样品GC密度关系与参照GC密度关系比较,从而生成比较,其中

[0943] 所述参照GC密度关系是参照的(i)GC密度与(ii)GC密度频率之间的关系;并且

[0944] (c)按照(b)中确定的比较将所述样品的序列读数的计数标准化,从而减少所述样品的序列读数中的偏好。

[0945] F1.1.如实施方式F1所述的存储介质,其中(c)中的标准化包括提供读数的标准化计数。

[0946] F2.如实施方式F1或F1.1所述的存储介质,其中通过包括使用核心密度评价的过程确定各所述GC密度。

[0947] F2.1.如实施方式F1-F2中任一项所述的存储介质,其中所述参照GC密度关系和所述样品GC密度关系的各GC密度代表局部GC密度。

[0948] F2.2.如实施方式F2.1所述的存储介质,其中所述局部GC含量针对5000bp或更小的多核苷酸区段。

[0949] F3.如实施方式F1-F2.2中任一项所述的存储介质,其中通过使用滑动窗分析的过程来确定各所述GC密度。

[0950] F4.如实施方式F3所述的存储介质,其中所述窗是约5个连续核苷酸至约5000个连续核苷酸,并且所述窗在滑动窗分析中一次滑动约1个碱基至约10个碱基。

[0951] F5.如实施方式F3所述的存储介质,其中所述窗是约200个连续核苷酸并且所述窗在滑动窗分析中一次滑动约1个碱基。

[0952] F6.如实施方式F1-F5中任一项所述的存储介质,其中(b)包括生成(i)和(ii)之间的拟合关系:(i)比例,各比例包括各所述GC密度的样品GC密度关系频率和参照GC密度关系频率,与(ii)GC密度。

[0953] F7.如实施方式F6所述的存储介质,其中(a)中的拟合关系获自加权的拟合。

[0954] F8.如实施方式F1-F7中任一项所述的存储介质,其中,所述样品的各序列读数表示为二进制格式。

[0955] F9.如实施方式F8所述的存储介质,其中各所述序列读数的二进制格式包括所述读数映射至的染色体和所述读数映射至的染色体位置。

[0956] F10.如实施方式F9所述的存储介质,其中所述二进制格式是5-字节格式,包括1-字节染色体序数和4-字节染色体位置。

[0957] F11.如实施方式F8-F10中任一项所述的存储介质,其中所述二进制格式比序列比对/映射(SAM)格式小50倍和/或比GZip格式小13%。

[0958] F12.如实施方式F1-F11中任一项所述的存储介质,其中(c)中的标准化包括因数化除了GC密度以外的一个或多个特征,并且标准化所述序列读数。

[0959] F13.如实施方式F12中任一项所述的存储介质,其中所述因数化一个或多个特征

是通过包括使用多变量模型的过程进行。

[0960] F14.如实施方式F13所述的存储介质,其中所述包括使用多变量模型的过程通过多变量模型进行。

[0961] F14.1.如实施方式F12-F14中任一项所述的存储介质,其中,按照(c)中的标准化和所述一个或多个特征的因数化将所述序列读数的计数标准化。

[0962] F15.如实施方式F1-F14.1中任一项所述的存储介质,其中所述程序对微处理器下指令,在(c)之后,按照下述过程生成基因组的一个或多个部份或其区段的读数密度,所述过程包括生成含(c)中标准化的序列读数的计数的所述一个或多个部份各自的概率密度评价。

[0963] F16.如实施方式F15所述的存储介质,其中,所述概率密度评价是核心密度评价。

[0964] F17.如实施方式F15或F16所述的存储介质,其中,所述程序向微处理器下指令以生成所述基因组或其区段的读数密度概况。

[0965] F18.如实施方式F17所述的存储介质,其中,所述读数密度概况包括所述基因组的一个或多个部份,或其区段的读数密度。

[0966] F19.如实施方式F15-F18中任一项所述的存储介质,其中,所述程序向所述微处理器下指令以调整所述一个或多个部份的各读数密度。

[0967] F20.如实施方式F15-F19中任一项所述的存储介质,其中,所述一个或多个部份经过滤从而提供过滤的部份。

[0968] F21.如实施方式F15-F20中任一项所述的存储介质,其中,所述程序向所述微处理器下指令以对所述一个或多个部份加权,从而提供加权的部份。

[0969] F22.如实施方式F21所述的存储介质,其中,所述一个或多个部份经本征函数加权。

[0970] F23.如实施方式F1-F22中任一项所述的存储介质,其中所述程序向所述微处理器下指令以在(a)之前获得所述序列读数。

[0971] F24.如实施方式F23所述的存储介质,其中,所述序列读数通过大规模平行测序(MPS)生成。

[0972] F25.如实施方式F23或F24所述的存储介质,其中,所述获得的序列读数是映射至整个参照基因组或基因组的区段的序列读数。

[0973] F26.如实施方式F25所述的存储介质,其中,所述基因组的区段包括染色体或其区段。

[0974] F27.如实施方式F25或F26所述的存储介质,其中,映射至所述参照基因组的序列读数的计数是序列读数的标准化计数。

[0975] F28.如实施方式F27所述的存储介质,其中,由GC含量、箱式标准化、GC LOESS、PERUN、GCRM或其组合将映射至所述参照基因组的序列读数的计数标准化。

[0976] F29.如实施方式F25或F26所述的存储介质,其中,映射至所述参照基因组的序列读数的计数是原始计数。

[0977] F30.如实施方式F15-F29中任一项所述的存储介质,其中,所述参照基因组的各部份包括大约相等长度的连续核苷酸。

[0978] F31.如实施方式F15或F30中任一项所述的存储介质,其中,所述参照基因组的各

部份包括约50kb。

[0979] F32.如实施方式F15-F31中任一项所述的存储介质,其中,所述参照基因组的各部份包括约100kb。

[0980] F33.如实施方式F15-F32中任一项所述的存储介质,其中,所述参照基因组的各部份包括与所述参照基因组的相邻部份一样的连续核苷酸的区段。

[0981] F34.如实施方式F1-F33中任一项所述的存储介质,其中,所述测试样品获自妊娠雌性。

[0982] F35.如实施方式F1-F34中任一项所述的存储介质,其中,所述测试样品包括来自妊娠雌性的血液。

[0983] F36.如实施方式F1-F35中任一项所述的存储介质,其中,所述测试样品包括来自妊娠雌性的血浆。

[0984] F37.如实施方式F1-F36中任一项所述的存储介质,其中,所述测试样品包括来自妊娠雌性的血清。

[0985] F38.如实施方式F1-F37中任一项所述的存储介质,其中,所述测试样品包括分离的核酸。

[0986] F39.如实施方式F8-F38中任一项所述的存储介质,其中所述程序向所述微处理器下指令以将(a)中映射至参照基因组的序列读数从序列比对格式压缩成二进制格式。

[0987] F40.如实施方式F39所述的存储介质,其中,所述压缩通过压缩模块进行。

[0988] F41.如实施方式F1-F40中任一项所述的存储介质,其中通过偏好密度模块提供所述测试样品的序列读数和所述参照的GC密度和GC密度频率。

[0989] F42.如实施方式F1-F41中任一项所述的存储介质,其中,由关系模块生成(b)中的比较。

[0990] F43.如实施方式F1-F42中任一项所述的存储介质,其中,由偏好校正模块进行(c)中的标准化。

[0991] F44.如实施方式F15-F43中任一项所述的存储介质,其中,由分布模块提供所述读数密度。

[0992] F45.如实施方式F20-F44中任一项所述的存储介质,其中,由过滤模块提供过滤的部份。

[0993] F46.如实施方式F21-F45中任一项所述的存储介质,其中,由部份加权模块提供加权的部份。

[0994] F46.1.如实施方式F21-F45中任一项所述的存储介质,其中,由读数密度调整模块提供调整的读数密度。

[0995] F47.如实施方式F46所述的存储介质,包括一个或多个压缩模块、偏好密度模块、关系模块、偏好校正模块、分布模块、过滤模块、读数密度调整模块和部份加权模块。

[0996] G1.一种非瞬时计算机可读存储介质,包括其上储存有可执行的程序,其中所述程序对微处理器下指令,以进行如下操作:

[0997] (a)按照读数密度分布过滤参照基因组的部份,从而提供包括过滤的部份的读数密度的测试样品的读数密度概况,其中:

[0998] 所述读数密度包括来自妊娠雌性的测试样品的循环无细胞核酸的序列读数,并且

- [0999] 针对多个样品的部份的读数密度确定所述读数密度分布；
- [1000] (b) 按照一个或多个主成分调整测试样品的读数密度概况，所述主成分通过主成分分析获自己知的整倍性样品的组，从而提供包括调整的读数密度的测试样品概况；
- [1001] (c) 将所述测试样品概况与参照概况比较，从而提供比较；并且
- [1002] (d) 按照所述比较确定所述测试样品中是否存在染色体非整倍性。
- [1003] G2. 如实施方式G1所述的存储介质，其中，所述比较包括确定显著性水平。
- [1004] G3. 如实施方式G2所述的存储介质，其中，所述确定显著性水平包括确定p-值。
- [1005] G4. 如实施方式G1-G3中任一项所述的存储介质，其中，所述参照概况包括获自己知整倍性样品的组的读数密度概况。
- [1006] G5. 如实施方式G1-G4中任一项所述的存储介质，其中，所述参照概况包括过滤的部份的读数密度。
- [1007] G6. 如实施方式G1-G5中任一项所述的存储介质，其中，所述参照概况包括按照所述一个或多个主成分调整的读数密度。
- [1008] G7. 如实施方式G2-G6中任一项所述的存储介质，其中，所述显著性水平表示所述测试样品概况和所述参照概况之间有统计学显著差异，并且确定存在染色体非整倍性。
- [1009] G8. 如实施方式G1-G7中任一项所述的存储介质，其中，所述多个样品包括已知整倍性样品的组。
- [1010] G9. 如实施方式G1-G8中任一项所述的存储介质，其中，所述多个样品的部份的读数密度是中值读数密度。
- [1011] G10. 如实施方式G1-G9中任一项所述的存储介质，其中，所述测试样品的过滤的部份的读数密度是中值读数密度。
- [1012] G11. 如实施方式G4-G10中任一项所述的存储介质，其中，所述参照概况的读数密度概况包括中值读数密度。
- [1013] G12. 如实施方式G4-G11中任一项所述的存储介质，其中，按照包括核心密度评价的过程确定所述测试样品概况、所述多个样品和所述参照概况的读数密度。
- [1014] G13. 如实施方式G10-G12中任一项所述的存储介质，其中，按照所述测试样品的中值读数密度确定所述测试样品概况。
- [1015] G14. 如实施方式G11-G13中任一项所述的存储介质，其中，按照所述参照的中值读数密度分布确定所述参照概况。
- [1016] G15. 如实施方式G1-G14中任一项所述的存储介质，其中所述程序向所述微处理器下指令以按照针对所述读数密度分布的不确定性测量对参照基因组的部份进行过滤。
- [1017] G15.1. 如实施方式G14.1所述的存储介质，其中，所述不确定性测量是MAD。
- [1018] G16. 如实施方式G1-G15.1中任一项所述的存储介质，其中所述程序向所述微处理器下指令以通过在(a)之前进行的过程对映射至所述测试样品的过滤部份的序列读数的计数加权，包括：
- [1019] (1) 生成所述测试样品的序列读数的(i) 鸟嘌呤和胞嘧啶(GC) 密度与(ii) GC密度频率之间的关系，从而生成样品GC密度关系，其中所述序列读数映射至所述参照基因组；
- [1020] (2) 将所述样品GC密度关系与参照GC密度关系比较，从而生成比较，其中，
- [1021] 所述参照GC密度关系是参照的(i) GC密度与(ii) GC密度频率之间的关系；并且

[1022] (3) 按照 (b) 中确定的比较将所述样品的序列读数的计数标准化,从而减少所述样品的序列读数中的偏好。

[1023] G16.1.如实施方式G16所述的存储介质,其中(3)中的标准化包括提供标准化的计数。

[1024] G17.如实施方式G16或G16.1所述的存储介质,其中通过包括使用核心密度评价的过程确定各所述GC密度。

[1025] G17.1.如实施方式G16-G17中任一项所述的存储介质,其中所述参照GC密度关系和所述样品GC密度关系的各GC密度代表局部GC密度。

[1026] G17.2.如实施方式G17.1所述的存储介质,其中所述局部GC含量针对5000bp或更小的多核苷酸区段。

[1027] G18.如实施方式G16-G17.2中任一项所述的存储介质,其中通过使用滑动窗分析的过程来确定各所述GC密度。

[1028] G19.如实施方式G18所述的存储介质,其中所述窗是约5个连续核苷酸至约5000个连续核苷酸,并且所述窗在滑动窗分析中一次滑动约1个碱基至约10个碱基。

[1029] G20.如实施方式G19所述的存储介质,其中所述窗是约200个连续核苷酸并且所述窗在滑动窗分析中一次滑动约1个碱基。

[1030] G21.如实施方式G16-G20中任一项所述的存储介质,其中(2)包括生成(i)和(ii)之间的拟合关系:(i)比例,各比例包括各所述GC密度的样品GC密度关系频率和参照GC密度关系频率,与(ii)GC密度。

[1031] G22.如实施方式G21所述的存储介质,其中(1)中的拟合关系获自加权的拟合。

[1032] G23.如实施方式G16-G22中任一项所述的存储介质,其中,所述样品的各序列读数表示为二进制格式。

[1033] G24.如实施方式G23所述的存储介质,其中各所述序列读数的二进制格式包括所述读数映射至的染色体和所述读数映射至的染色体位置。

[1034] G25.如实施方式G24所述的存储介质,其中所述二进制格式是5-字节格式,包括1-字节染色体序数和4-字节染色体位置。

[1035] G26.如实施方式G23-G25中任一项所述的存储介质,其中所述二进制格式比序列比对/映射(SAM)格式小50倍和/或比GZip格式小13%。

[1036] G27.如实施方式G16-G26中任一项所述的存储介质,其中(c)中的标准化包括因数化除了GC密度以外的一个或多个特征,并且标准化所述序列读数的计数。

[1037] G28.如实施方式G27所述的存储介质,其中所述因数化一个或多个特征是通过包括使用多变量模型的过程进行。

[1038] G29.如实施方式G28所述的存储介质,其中所述包括使用多变量模型的过程通过多变量模型进行。

[1039] G29.1.如实施方式G16-G29中任一项所述的存储介质,其中,所述程序向所述微处理器下指令以对所述测试样品的过滤部份加权。

[1040] G29.2.如实施方式G29.1所述的存储介质,其中所述测试样品的过滤的部份通过包括本征函数的过程加权。

[1041] G30.如实施方式G1-G29.2中任一项所述的存储介质,其中所述程序向所述微处理

器下指令以在 (a) 之前获得所述序列读数。

[1042] G31. 如实施方式G30所述的存储介质, 其中, 所述序列读数通过大规模平行测序 (MPS) 生成。

[1043] G32. 如实施方式G1-G31中任一项所述的存储介质, 包括获得映射至完整参照基因组或基因组区段的序列读数。

[1044] G33. 如实施方式G32所述的存储介质, 其中, 所述基因组的区段包括染色体或其区段。

[1045] G34. 如实施方式G32或G33所述的存储介质, 其中, 在 (1) 之前将映射至所述参照基因组的序列读数的计数标准化。

[1046] G35. 如实施方式G34所述的存储介质, 其中, 由GC含量、箱式标准化、GC LOESS、PERUN、GCRM或其组合将映射至所述参照基因组的序列读数的计数标准化。

[1047] G36. 如实施方式G32或G33所述的存储介质, 其中, 映射至所述参照基因组的序列读数的计数是原始计数。

[1048] G37. 如实施方式G1-G36中任一项所述的存储介质, 其中, 所述参照基因组的各部份包括大约相等长度的连续核苷酸。

[1049] G38. 如实施方式G1-G37中任一项所述的存储介质, 其中, 所述参照基因组的各部份包括约50kb。

[1050] G39. 如实施方式G1-G38中任一项所述的存储介质, 其中, 所述参照基因组的各部份包括约100kb。

[1051] G40. 如实施方式G1-G39中任一项所述的存储介质, 其中, 所述参照基因组的各部份包括与所述参照基因组的相邻部份一样的连续核苷酸的区段。

[1052] G41. 如实施方式G1-G40中任一项所述的存储介质, 其中, 所述测试样品包括来自妊娠雌性的血液。

[1053] G42. 如实施方式G1-G41中任一项所述的存储介质, 其中, 所述测试样品包括来自妊娠雌性的血浆。

[1054] G43. 如实施方式G1-G42中任一项所述的存储介质, 其中, 所述测试样品包括来自妊娠雌性的血清。

[1055] G44. 如实施方式G1-G43中任一项所述的存储介质, 其中, 从所述测试样品分离核酸。

[1056] G45. 如实施方式G23-G44中任一项所述的存储介质, 其中所述程序向微处理器下指令以将 (1) 中映射至所述参照基因组的序列读数从序列比对格式压缩成二进制格式。

[1057] G46. 如实施方式G45所述的存储介质, 其中, 所述压缩通过压缩模块进行。

[1058] G47. 如实施方式G16-G46中任一项所述的存储介质, 其中通过偏好密度模块提供所述测试样品的序列读数和所述参照的GC密度和GC密度频率。

[1059] G48. 如实施方式G16-G47中任一项所述的存储介质, 其中, 由关系模块生成 (2) 中的比较。

[1060] G49. 如实施方式G17-G48中任一项所述的存储介质, 其中, 由偏好校正模块进行 (3) 中的标准化。

[1061] G50. 如实施方式G1-G49中任一项所述的存储介质, 其中, 由分布模块提供所述读

数密度。

[1062] G51.如实施方式G1-G50中任一项所述的存储介质,其中,由过滤模块提供过滤的部份。

[1063] G51.1.如实施方式G29.1-G51中任一项所述的存储介质,其中通过部份加权模块对所述测试样品的过滤的部份加权。

[1064] G51.1.如实施方式G29.1-G51中任一项所述的存储介质,其中,由读数密度调整模块提供调整的读数密度。

[1065] G52.如实施方式G51.1所述的存储介质,其中,设备包括一个或多个压缩模块、偏好密度模块、关系模块、偏好校正模块、分布模块、过滤模块、读数密度调整模块和部份加权模块。

[1066] G53.如实施方式G1-G52中任一项所述的存储介质,其中所述测试样品概况包括染色体或其区段的概况。

[1067] G54.如实施方式G1-G53中任一项所述的存储介质,其中所述参照概况包括染色体或其区段的概况。

[1068] G55.如实施方式G1-G54中任一项所述的存储介质,其中所述(d)中的确定具有等于或大于90%的特异性和等于或大于90%的灵敏度。

[1069] G56.如实施方式G1-G55中任一项所述的存储介质,其中非整倍性是三体性。

[1070] G57.如实施方式G56所述的存储介质,其中所述三体性是21三体性、18三体性、或13三体性。

[1071] H1.一种包括存储器和一个或多个微处理器的系统,所述一个或多个微处理器设置为按照所述存储器中的指令进行用于减少样品的序列读数中的偏好的过程,所述过程包括:

[1072] (a)生成测试样品的序列读数的(i)局部基因组偏好评价与(ii)偏好频率之间的关系,从而生成样品偏好关系,其中,

[1073] 序列读数是来自测试样品的循环无细胞核酸,并且

[1074] 所述序列读数映射至参照基因组;

[1075] (b)将所述样品偏好关系与参照偏好关系比较,从而生成比较,其中,

[1076] 所述参照偏好关系是参照的(i)局部基因组偏好评价与(ii)偏好频率之间的关系;并且

[1077] (c)按照(b)中确定的比较将所述样品的序列读数的计数标准化,从而减少所述样品的序列读数中的偏好。

[1078] H1.1.一种包括测序设备和一个或多个计算设备的系统,

[1079] 所述测序设备设置成产生对应于加载到所述测序设备中的核酸的核苷酸碱基的信号,所述核酸是来自怀有胎儿的妊娠雌性的血液的循环无细胞核酸,或者加载到所述测序设备中的核酸是所述循环无细胞核酸的修饰变体;并且

[1080] 所述一个或多个计算设备包括存储器和一个或多个处理器,所述存储器包括可由所述一个或多个处理器执行的指令并且可由所述一个或多个处理器执行的指令设置为:

[1081] 产生来自所述信号的序列读数并且映射所述序列读数;

[1082] (a)生成测试样品的序列读数的(i)局部基因组偏好评价与(ii)偏好频率之间的

关系,从而生成样品偏好关系;

[1083] (b) 将所述样品偏好关系与参照偏好关系比较,从而生成比较,其中,

[1084] 所述参照偏好关系是参照的 (i) 局部基因组偏好评价与 (ii) 偏好频率之间的关系;并且

[1085] (c) 按照 (b) 中确定的比较将所述样品的序列读数的计数标准化,从而减少所述样品的序列读数中的偏好。

[1086] H1.2.如实施方式H1或H1.1所述的系统,其中(c)中的标准化包括提供标准化的计数。

[1087] H2.如实施方式H1-H1.2中任一项所述的系统,其中通过包括使用核心密度评价的过程来确定各所述局部基因组偏好评价。

[1088] H2.1.如实施方式H1-H2中任一项所述的系统,其中所述参照偏好关系和所述样品偏好关系的各局部基因组偏好评价代表局部偏好含量。

[1089] H2.2.如实施方式H2.1所述的系统,其中所述局部偏好含量针对5000bp或更小的多核苷酸区段。

[1090] H3.如实施方式H1-H2.2中任一项所述的系统,其中通过包括使用滑动窗分析的过程来确定各所述局部基因组偏好评价。

[1091] H4.如实施方式H3所述的系统,其中所述窗是约5个连续核苷酸至约5000个连续核苷酸,并且所述窗在滑动窗分析中一次滑动约1个碱基至约10个碱基。

[1092] H5.如实施方式H3所述的系统,其中所述窗是约200个连续核苷酸并且所述窗在滑动窗分析中一次滑动约1个碱基。

[1093] H6.如实施方式H1-H5中任一项所述的系统,其中(b)包括生成(i)和(ii)之间的拟合关系:(i)比例,各比例包括各所述局部基因组偏好评价的样品偏好关系频率和参照偏好关系频率,与(ii)局部基因组偏好评价。

[1094] H7.如实施方式H6所述的系统,其中(a)中的拟合关系获自加权的拟合。

[1095] H8.如实施方式H1-H7中任一项所述的系统,其中,所述样品的各序列读数表示为二进制格式。

[1096] H9.如实施方式H8所述的系统,其中各所述序列读数的二进制格式包括所述读数映射至的染色体和所述读数映射至的染色体位置。

[1097] H10.如实施方式H9所述的系统,其中所述二进制格式是5-字节格式,包括1-字节染色体序数和4-字节染色体位置。

[1098] H11.如实施方式H8-H10中任一项所述的系统,其中所述二进制格式比序列比对/映射(SAM)格式小50倍和/或比GZip格式小13%。

[1099] H12.如实施方式H1-H11中任一项所述的系统,其中(c)中的标准化包括因数化除了偏好以外的一个或多个特征,并且标准化所述序列读数的计数。

[1100] H13.如实施方式H12中任一项所述的系统,其中所述因数化一个或多个特征是通过包括使用多变量模型的过程进行。

[1101] H14.如实施方式H13所述的系统,其中所述包括使用多变量模型的过程通过多变量模型进行。

[1102] H14.1.如实施方式H12-H14中任一项所述的系统,其中,按照(c)中的标准化和所

述一个或多个特征的因数化将所述序列读数的计数标准化。

[1103] H15.如实施方式H1-H14.1中任一项所述的系统,包括在(c)之后,按照下述过程生成基因组的一个或多个部份或其区段的读数密度,所述过程包括生成含(c)中标准化的序列读数的计数的所述一个或多个部份各自的概率密度评价。

[1104] H16.如实施方式H15所述的系统,其中,所述概率密度评价是核心密度评价。

[1105] H17.如实施方式H15或H16所述的系统,包括生成所述基因组或其区段的读数密度概况。

[1106] H18.如实施方式H17所述的系统,其中,所述读数密度概况包括所述基因组的一个或多个部份,或其区段的读数密度。

[1107] H19.如实施方式H15-H18中任一项所述的系统,包括调整所述一个或多个部份的各读数密度。

[1108] H20.如实施方式H15-H19中任一项所述的系统,其中,所述一个或多个部份经过滤从而提供过滤的部份。

[1109] H21.如实施方式H15-H20中任一项所述的系统,其中,所述一个或多个部份经加权从而提供加权的部份。

[1110] H22.如实施方式H21所述的系统,其中,所述一个或多个部份经本征函数加权。

[1111] H23.如实施方式H1-H22中任一项所述的系统,其中所述局部基因组偏好评价包括局部GC密度并且所述偏好频率包括GC偏好频率。

[1112] H24.如实施方式H1-H23中任一项所述的系统,包括:

[1113] (a)按照读数密度分布过滤参照基因组的部份,从而提供包括过滤的部份的读数密度的测试样品的读数密度概况,其中,

[1114] 所述读数密度包括来自妊娠雌性的测试样品的循环无细胞核酸的序列读数,并且

[1115] 针对多个样品的部份的读数密度确定所述读数密度分布;

[1116] (b)按照一个或多个主成分调整测试样品的读数密度概况,所述主成分通过主成分分析获自己知的整倍性样品的组,从而提供包括调整的读数密度的测试样品概况;

[1117] (c)将所述测试样品概况与参照概况比较,从而提供比较;并且

[1118] (d)按照所述比较确定所述测试样品中是否存在染色体非整倍性。

[1119] H25.如实施方式H24所述的系统,其中,所述比较包括确定显著性水平。

[1120] H26.如实施方式H25所述的系统,其中,所述确定显著性水平包括确定p-值。

[1121] H27.如实施方式H24-H26中任一项所述的系统,其中,所述参照概况包括获自己知整倍性样品的组的读数密度概况。

[1122] H28.如实施方式H24-H27中任一项所述的系统,其中,所述参照概况包括过滤的部份的读数密度。

[1123] H29.如实施方式H24-H28中任一项所述的系统,其中,所述参照概况包括按照所述一个或多个主成分调整的读数密度。

[1124] H30.如实施方式H25-H29中任一项所述的系统,其中,所述显著性水平表示所述测试样品概况和所述参照概况之间有统计学显著差异,并且确定存在染色体非整倍性。

[1125] H31.如实施方式H24-H30中任一项所述的系统,其中,所述多个样品包括已知整倍性样品的组。

[1126] H32.如实施方式H24-H31中任一项所述的系统,其中,所述多个样品的部份的读数密度是中值读数密度。

[1127] H33.如实施方式H24-H32中任一项所述的系统,其中,所述测试样品的过滤的部份的读数密度是中值读数密度。

[1128] H34.如实施方式H27-H33中任一项所述的系统,其中,所述参照概况的读数密度概况包括中值读数密度。

[1129] H35.如实施方式H27-H34中任一项所述的系统,其中,按照包括核心密度评价的过程确定所述测试样品概况、所述多个样品和所述参照概况的读数密度。

[1130] H36.如实施方式H33-H35中任一项所述的系统,其中,按照所述测试样品的中值读数密度确定所述测试样品概况。

[1131] H37.如实施方式H34-H36中任一项所述的系统,其中,按照所述参照的中值读数密度分布确定所述参照概况。

[1132] H38.如实施方式H24-H37中任一项所述的系统,包括按照针对所述读数密度分布的不确定性测量对参照基因组的部份进行过滤。

[1133] H39.如实施方式H38所述的系统,其中,所述不确定性测量是MAD。

[1134] H40.如实施方式H1-H39中任一项所述的系统,其中,所述系统的存储器包括来自所述测试样品的映射至所述参照基因组的循环无细胞核酸的序列读数。

[1135] I1.一种用于减少样品的序列读数的偏好的方法,包括:

[1136] (a) 使用微处理器生成测试样品的序列读数的(i)局部基因组偏好评价与(ii)偏好频率之间的关系,从而生成样品偏好关系,其中,

[1137] 序列读数是来自测试样品的循环无细胞核酸,并且

[1138] 所述序列读数映射至参照基因组;

[1139] (b) 将所述样品偏好关系与参照偏好关系比较,从而生成比较,其中,

[1140] 所述参照偏好关系是参照的(i)局部基因组偏好评价与(ii)偏好频率之间的关系;并且

[1141] (c) 按照(b)中确定的比较将所述样品的序列读数的计数标准化,从而减少所述样品的序列读数中的偏好。

[1142] I1.1.一种用于减少样品的序列读数的偏好的方法,包括:

[1143] 用来自怀有胎儿的妊娠雌性的血液的循环无细胞核酸加载测序设备,或用所述核酸的修饰变体加载测序设备,所述测序设备产生对应于所述核酸的核苷酸碱基的信号;

[1144] 在任选地将所述信号转移至包括一个或多个计算设备的系统之后,由所述系统从所述核酸的信号生成序列读数,其中所述系统中的一个或多个计算设备包括存储器和一个或多个处理器,并且

[1145] 其中所述系统中一个计算设备,或计算设备的组合设置为:

[1146] 将所述序列读数映射至参照基因组;

[1147] (a) 使用微处理器生成测试样品的序列读数的(i)局部基因组偏好评价与(ii)偏好频率之间的关系,从而生成样品偏好关系,其中,

[1148] 序列读数是来自测试样品的循环无细胞核酸,并且

[1149] 所述序列读数映射至参照基因组;

- [1150] (b) 将所述样品偏好关系与参照偏好关系比较,从而生成比较,其中,
- [1151] 所述参照偏好关系是参照的 (i) 局部基因组偏好评价与 (ii) 偏好频率之间的关系;并且
- [1152] (c) 按照 (b) 中确定的比较将所述样品的序列读数的计数标准化,从而减少所述样品的序列读数中的偏好。
- [1153] I1.2.如实施方式I1或I1.1所述的方法,其中 (c) 中的标准化包括提供标准化的计数。
- [1154] I2.如实施方式I1、I1.1或I1.2所述的方法,其中通过包括核心密度评价的过程确定各局部基因组偏好评价。
- [1155] I2.1.如实施方式I1-I2中任一项所述的方法,其中所述参照偏好关系和所述样品偏好关系的各局部基因组偏好评价代表局部偏好含量。
- [1156] I2.2.如实施方式I2.1所述的方法,其中所述局部偏好含量针对5000bp或更小的多核苷酸区段。
- [1157] I3.如实施方式I1-I2.2中任一项所述的方法,其中通过包括使用滑动窗分析的过程来确定各所述局部基因组偏好评价。
- [1158] I4.如实施方式I3所述的方法,其中所述窗是约5个连续核苷酸至约5000个连续核苷酸,并且所述窗在滑动窗分析中一次滑动约1个碱基至约10个碱基。
- [1159] I5.如实施方式I3所述的方法,其中所述窗是约200个连续核苷酸并且所述窗在滑动窗分析中一次滑动约1个碱基。
- [1160] I6.如实施方式I1-I5中任一项所述的方法,其中 (b) 包括生成 (i) 和 (ii) 之间的拟合关系: (i) 比例,各比例包括各所述局部基因组偏好评价的样品偏好关系频率和参照偏好关系频率,与 (ii) 局部基因组偏好评价。
- [1161] I7.如实施方式I6所述的方法,其中 (a) 中的拟合关系获自加权的拟合。
- [1162] I8.如实施方式I1-I7中任一项所述的方法,其中,所述样品的各序列读数表示为二进制格式。
- [1163] I9.如实施方式I8所述的方法,其中各所述序列读数的二进制格式包括所述读数映射至的染色体和所述读数映射至的染色体位置。
- [1164] I10.如实施方式I9所述的方法,其中所述二进制格式是5-字节格式,包括1-字节染色体序数和4-字节染色体位置。
- [1165] I11.如实施方式I8-I10中任一项所述的方法,其中所述二进制格式比序列比对/映射 (SAM) 格式小50倍和/或比GZip格式小13%。
- [1166] I12.如实施方式I1-I11中任一项所述的方法,其中 (c) 中的标准化包括因数化除了偏好以外的一个或多个特征,并且标准化所述序列读数的计数。
- [1167] I13.如实施方式I12所述的方法,其中所述因数化一个或多个特征是通过包括使用多变量模型的过程进行。
- [1168] I14.如实施方式I13所述的方法,其中所述包括使用多变量模型的过程通过多变量模型进行。
- [1169] I14.1.如实施方式I12-I14中任一项所述的方法,其中,按照 (c) 中的标准化和所述一个或多个特征的因数化将所述序列读数的计数标准化。

[1170] I15.如实施方式I1-I14.1中任一项所述的方法,包括在(c)之后,按照下述过程生成基因组的一个或多个部份或其区段的读数密度,所述过程包括生成含(c)中标准化的序列读数的计数的所述一个或多个部份各自的概率密度评价。

[1171] I16.如实施方式I15所述的方法,其中,所述概率密度评价是核心密度评价。

[1172] I17.如实施方式I15或I16所述的方法,包括生成所述基因组或其区段的读数密度概况。

[1173] I18.如实施方式I17所述的方法,其中,所述读数密度概况包括所述基因组的一个或多个部份,或其区段的读数密度。

[1174] I19.如实施方式I15-I18中任一项所述的方法,包括调整所述一个或多个部份的各读数密度。

[1175] I20.如实施方式I15-I19中任一项所述的方法,其中,所述一个或多个部份经过滤从而提供过滤的部份。

[1176] I21.如实施方式I15-I20中任一项所述的方法,其中,所述一个或多个部份经加权从而提供加权的部份。

[1177] I22.如实施方式I21所述的方法,其中,所述一个或多个部份经本征函数加权。

[1178] I23.如实施方式I1-I22中任一项所述的方法,其中所述局部基因组偏好评价包括局部GC密度并且所述偏好频率包括GC偏好频率。

[1179] I23.1.如实施方式I1-I23中任一项所述的方法,包括:

[1180] (a)按照读数密度分布过滤参照基因组中染色体的部份,从而提供包括过滤的部份的读数密度的测试样品的读数密度概况,其中,

[1181] 所述读数密度包括来自妊娠雌性的测试样品的循环无细胞核酸的序列读数,并且

[1182] 针对多个样品的部份的读数密度确定所述读数密度分布;

[1183] (b)按照一个或多个主成分调整测试样品的染色体的读数密度概况,所述主成分通过主成分分析获自己知的整倍性样品的组,从而提供包括调整的读数密度的测试样品染色体概况;

[1184] (c)将所述测试样品染色体概况与参照概况比较,从而提供比较;并且

[1185] (d)按照所述比较确定所述测试样品中是否存在染色体非整倍性。

[1186] I24.如实施方式I1-I23中任一项所述的方法,包括:

[1187] (a)按照读数密度分布过滤参照基因组的部份,从而提供包括过滤的部份的读数密度的测试样品的读数密度概况,其中,

[1188] 所述读数密度包括来自妊娠雌性的测试样品的循环无细胞核酸的序列读数,并且

[1189] 针对多个样品的部份的读数密度确定所述读数密度分布;

[1190] (b)按照一个或多个主成分调整测试样品的读数密度概况,所述主成分通过主成分分析获自己知的整倍性样品的组,从而提供包括调整的读数密度的测试样品概况;

[1191] (c)将所述测试样品概况与参照概况比较,从而提供比较;并且

[1192] (d)按照所述比较确定所述测试样品中是否存在染色体非整倍性。

[1193] I24.1.如实施方式I23.1或I24所述的方法,其中在(b)中由1-10个主成分来调整所述读数密度概况。

[1194] I24.2.如实施方式I23.1、I24或I24.1所述的方法,其中在(b)中由5个主成分来调

整所述读数密度概况。

[1195] I24.3.如实施方式I23.1-I24.2中任一项所述的方法,其中针对读数密度概况中的一个或多个特征调整所述一个或多个主成分,其特征选自胎儿性别、序列偏好、胎儿分数、与DNA酶I灵敏度相关的偏好、熵、重复序列偏好、染色质结构偏好、聚合酶误差率偏好、回文偏好、插入重复偏好、PCR扩增偏好、和隐藏拷贝数变异。

[1196] I24.4.如实施方式I24.3所述的方法,其中序列偏好包括鸟嘌呤和胞嘧啶(GC)偏好。

[1197] I25.如实施方式I23.1-I24.4中任一项所述的方法,其中所述比较包括确定显著性水平。

[1198] I26.如实施方式I25所述的方法,其中,所述确定显著性水平包括确定p-值。

[1199] I27.如实施方式I23.1-I26中任一项所述的方法,其中,所述参照概况包括获自己知整倍性样品的组的读数密度概况。

[1200] I28.如实施方式I23.1-I27中任一项所述的方法,其中,所述参照概况包括过滤的部份的读数密度。

[1201] I29.如实施方式I23.1-I28中任一项所述的方法,其中,所述参照概况包括按照所述一个或多个主成分调整的读数密度。

[1202] I30.如实施方式I25-I29中任一项所述的方法,其中,所述显著性水平表示所述测试样品概况和所述参照概况之间有统计学显著差异,并且确定存在染色体非整倍性。

[1203] I31.如实施方式I23.1-I30中任一项所述的方法,其中,所述多个样品包括已知整倍性样品的组。

[1204] I32.如实施方式I23.1-I31中任一项所述的方法,其中,所述多个样品的部份的读数密度是中值读数密度。

[1205] I33.如实施方式I23.1-I32中任一项所述的方法,其中,所述测试样品的过滤的部份的读数密度是中值读数密度。

[1206] I34.如实施方式I27-I33中任一项所述的方法,其中,所述参照概况的读数密度概况包括中值读数密度。

[1207] I35.如实施方式I27-I34中任一项所述的方法,其中,按照包括核心密度评价的过程确定所述测试样品概况、所述多个样品和所述参照概况的读数密度。

[1208] I36.如实施方式I33-I35中任一项所述的方法,其中,按照所述测试样品的中值读数密度确定所述测试样品概况。

[1209] I37.如实施方式I34-I36中任一项所述的方法,其中,按照所述参照的中值读数密度分布确定所述参照概况。

[1210] I38.如实施方式I23.1-I37中任一项所述的方法,包括按照针对所述读数密度分布的不确定性测量对参照基因组的部份进行过滤。

[1211] I39.如实施方式I38所述的方法,其中,所述不确定性测量是MAD。

[1212] I40.如实施方式I23.1-I39中任一项所述的方法,其中,所述测试样品概况代表所述测试样品的染色体剂量。

[1213] I41.如实施方式I40所述的方法,包括比较测试样品概况的染色体剂量与参照概况的染色体剂量,从而生成染色体剂量比较。

[1214] I42.如实施方式I41所述的方法,包括按照所述染色体剂量比较确定所述测试样品是否存在染色体非整倍性。

[1215] I43.如实施方式I42所述的方法,其中确定所述测试样品是否存在染色体非整倍性包括鉴定是否存在1个拷贝的染色体、2个拷贝的染色体、3个拷贝的染色体、4个拷贝的染色体、5个拷贝的染色体、染色体的一个或多个区段的缺失或染色体的一个或多个区段的插入。

[1216] J1.一种用于确定是否存在非整倍性的方法,包括:

[1217] (a) 获得映射至参照基因组的基因组部份的部分核苷酸序列读数的计数,所述部分核苷酸序列读数是来自妊娠雌性的测试样品的循环无细胞核酸的读数,其中所述部分核苷酸序列读数的至少一些包括:

[1218] i) 鉴定的核碱基之间的多个核碱基间隙,或

[1219] ii) 一个或多个核碱基类别,其中各核碱基类别包括样品核酸中存在的核碱基子集,或

[1220] (i) 和(ii)的组合,

[1221] (b) 按照读数密度分布过滤参照基因组的部份,从而提供包括过滤的部份的读数密度的所述测试样品的读数密度概况,其中,

[1222] 所述读数密度包括来自所述测试样品的部分核苷酸序列读数,并且

[1223] 针对多个样品的部份的读数密度确定所述读数密度分布;

[1224] (c) 按照一个或多个主成分调整测试样品的读数密度概况,所述主成分通过主成分分析获自己知的整倍性样品的组,从而提供包括调整的读数密度的测试样品概况;

[1225] (d) 将所述测试样品概况与参照概况比较,从而提供比较;并且

[1226] (e) 按照所述比较确定所述测试样品中是否存在非整倍性。

[1227] J2.一种用于基于拷贝数变异确定胎儿分数的方法,包括:

[1228] (a) 获取映射至参照基因组的基因组部份的核酸序列读数的计数,其中序列读数是来自妊娠雌性的测试样品的循环无细胞核酸的读数;

[1229] (b) 标准化所述映射至参照基因组的基因组部份的计数,从而提供所述基因组部份的标准化计数;其中所述标准化包括:

[1230] (i) 按照读数密度分布过滤所述参照基因组的部份,从而提供包括过滤的部份的读数密度的所述测试样品的读数密度概况,其中,

[1231] 所述读数密度包括来自所述测试样品的核苷酸序列读数,并且

[1232] 针对多个样品的部份的读数密度确定所述读数密度分布;并且

[1233] (ii) 按照一个或多个主成分调整测试样品的读数密度概况,所述主成分通过主成分分析获自己知的整倍性样品的组,从而提供包括调整的读数密度的测试样品概况;

[1234] (c) 鉴定与所述标准化计数的第二水平明显不同的所述标准化计数的第一水平,所述第一水平针对第一组基因组部份,并且所述第二水平针对第二组基因组部份;

[1235] (d) 将拷贝数变异分配到所述第一水平,从而提供归类;并且

[1236] (e) 按照所述归类确定所述循环无细胞核酸的胎儿分数,从而从所述核酸序列读数生成所述胎儿分数。

[1237] J3.一种用于确定来自妊娠雌性的血液的循环无细胞核酸中胎儿核酸的分数的方

法,包括:

[1238] (a) 获取映射至参照基因组的基因组部份的核酸序列读数的计数,其中序列读数是来自怀有雄性胎儿的妊娠雌性的测试样品的循环无细胞核酸的读数;

[1239] (b) 生成实验X染色体代表,所述X染色体代表是(i)映射至X染色体中参照基因组的基因组部份的序列读数的计数与(ii)映射至基因组或其区段中参照基因组的基因组部份的序列读数的计数的比例;并且

[1240] (c) 按照所述实验X染色体代表和预期的X染色体代表,从所述实验X染色体代表确定所述妊娠雌性的血液中胎儿核酸的分数,所述预期的X染色体代表是(i)X染色体中参照基因组的基因组部份的数量与(ii)基因组或其区段中参照基因组的基因组部份的数量的比例,其中(b)中的计数通过包括以下的方法标准化:

[1241] (1) 按照读数密度分布过滤所述参照基因组的部份,从而提供包括过滤的部份的读数密度的所述测试样品的读数密度概况,其中,

[1242] 所述读数密度包括来自所述测试样品的核苷酸序列读数,并且

[1243] 针对多个样品的部份的读数密度确定所述读数密度分布;并且

[1244] (2) 按照一个或多个主成分调整测试样品的读数密度概况,所述主成分通过主成分分析获自己知的整倍性样品的组,从而提供包括调整的读数密度的测试样品概况。

[1245] J4.一种用于按照核酸序列读数确定胎儿倍性的方法,包括:

[1246] (a) 确定测试样品中胎儿核酸的分数,所述测试样品包括来自妊娠雌性的循环无细胞核酸;

[1247] (b) 获取映射至参照基因组的部份的序列读数的计数,所述序列读数来自所述样品中的核酸;

[1248] (c) 计算各所述参照基因组的部份的基因组区段水平,从而提供计算的基因组区段水平;并且

[1249] (d) 按照(i)和(ii)之间的关系确定胎儿倍性:(i)参照基因组的部份的子集的计算的基因组区段水平与(ii)(a)中确定的胎儿核酸的分数,其中(b)中的计数通过包括以下的过程来标准化:

[1250] (1) 按照读数密度分布过滤所述参照基因组的部份,从而提供包括过滤的部份的读数密度的所述测试样品的读数密度概况,其中,

[1251] 所述读数密度包括来自所述测试样品的核苷酸序列读数,并且

[1252] 针对多个样品的部份的读数密度确定所述读数密度分布;并且

[1253] (2) 按照一个或多个主成分调整测试样品的读数密度概况,所述主成分通过主成分分析获自己知的整倍性样品的组,从而提供包括调整的读数密度的测试样品概况。

[1254] J5.一种用于确定是否存在胎儿非整倍性的方法,包括:

[1255] (a) 获取映射至参照基因组部份的核苷酸序列读数的计数,其中所述核苷酸序列读数是来自妊娠雌性的测试样品的循环无细胞核酸的读数;

[1256] (b) 通过包括从第一基因组部份的计数中减去预期计数的过程来标准化所述第一基因组部份的计数,从而生成减值,并且将所述减值除以计数变异的评价,或者使用微处理器标准化所述第一基因组部份的计数衍生,从而获取标准化的样品计数,

[1257] 获得针对包括接触一个或多个共同实验条件的样品、参照、或样品和参照的组的

所述预期计数,或所述预期计数衍生;并且

[1258] (c) 基于标准化的样品计数确定是否存在胎儿非整倍性,其中标准化(b)中计数还包括:

[1259] (1) 按照读数密度分布过滤所述参照基因组的部份,从而提供包括过滤的部份的读数密度的所述测试样品的读数密度概况,其中,

[1260] 所述读数密度包括来自所述测试样品的核苷酸序列读数,并且

[1261] 针对多个样品的部份的读数密度确定所述读数密度分布;并且

[1262] (2) 按照一个或多个主成分调整测试样品的读数密度概况,所述主成分通过主成分分析获自己知的整倍性样品的组,从而提供包括调整的读数密度的测试样品概况。

[1263] J6. 一种用于确定胎儿中性染色体核型的方法,包括:

[1264] (a) 获取映射至参照基因组的部份的核苷酸序列读数的计数,其中序列读数是妊娠雌性的测试样品的循环无细胞核酸的读数;

[1265] (b) 从各样品的(i)映射至各参照基因组的部份的序列读数的计数与(ii)各部份的映射特征之间的拟合关系确定多个样品的参照基因组的各部份的实验偏好;

[1266] (c) 从所述实验偏好与映射至各所述参照基因组的部份的序列读数的计数之间的拟合关系确定各所述参照基因组的部份的基因组区段水平,从而提供计算的基因组区段水平;并且

[1267] (d) 按照所述计算的基因组区段水平确定所述胎儿的性染色体核型,其中(b)中确定实验偏好还包括:

[1268] (1) 按照读数密度分布过滤所述参照基因组的部份,从而提供包括过滤的部份的读数密度的所述测试样品的读数密度概况,其中,

[1269] 所述读数密度包括来自所述测试样品的核苷酸序列读数,并且

[1270] 针对多个样品的部份的读数密度确定所述读数密度分布;并且

[1271] (2) 按照一个或多个主成分调整测试样品的读数密度概况,所述主成分通过主成分分析获自己知的整倍性样品的组,从而提供包括调整的读数密度的测试样品概况。

[1272] J7. 一种用于确定是否存在非整倍性的方法,包括:

[1273] (a) 获取参照基因组中映射至染色体13、18或21或其区段的序列读数的计数,所述序列读数是来自妊娠雌性的测试样品的循环无细胞核酸的读数;

[1274] (b) 确定3个比例或比例值,所述3个比例各自是(i)映射至各染色体13、18和21或其区段的计数与(ii)映射至各其它染色体13、18和21或其区段的计数的比例;

[1275] (c) 比较所述3个比例或比例值,从而生成比较;并且

[1276] (d) 基于(c)中生成的比较确定受否存在染色体非整倍性,前提是(c)中生成的比较和(d)中的确定并不基于除了染色体13、18和21以外的基因组的区段;从而从所述序列读数中生成确定是否存在所述染色体非整倍性,其中映射至染色体13、18和21或其区段的序列读数的计数通过包括以下的过程来标准化:

[1277] (1) 按照读数密度分布过滤所述参照基因组的部份,从而提供包括过滤的部份的读数密度的所述测试样品的读数密度概况,其中,

[1278] 所述读数密度包括来自所述测试样品的核苷酸序列读数,并且

[1279] 针对多个样品的部份的读数密度确定所述读数密度分布;并且

[1280] (2) 按照一个或多个主成分调整测试样品的读数密度概况,所述主成分通过主成分分析获自己知的整倍性样品的组,从而提供包括调整的读数密度的测试样品概况。

[1281] * * *

[1282] 本文中引用的各专利、专利申请、出版物和文献的全部内容均通过引用纳入本文。对上述专利、专利申请、出版物和文献的引用并不表示承认上述任何内容是相关的现有技术,也并不表示承认这些出版物或文献的内容或日期。

[1283] 可以对上述内容进行改变而不背离本技术的基本方面。尽管参照一个或多个具体实施方式充分详细描述了本技术,但是本领域普通技术人员应认识到可对本申请中具体公开的实施方案进行改变,而这些改良和改进在本技术的范围和精神内。

[1284] 本文中适当地说明性描述的技术可在没有任何本文未具体公开的元素的情况下实施。因此,例如,在本文的各个例子中,术语“包括”、“基本由……组成”和“由……组成”中的任何一个都可用其它两个中的任意一个代替。已经使用的术语和表达用作说明而非限制性的术语,此类术语和表达的使用并不排除对所显示和所描述的特征或其部分的任何等价物,以及在要求权利的本技术范围内可进行各种改良。术语“一个”或“一种”表示一种或多种其修饰的元素(例如“一种试剂”可表示一种或多种试剂),除非上下文清楚表示所描述的是元素之一或是一种以上的元素。本文所使用的术语“约”表示在基础参数的10%范围内的数值(例如,±10%),在一列数值的开头处使用的术语“约”表示修饰该列数值中的每个数值(例如,“约1、2和3”指约1、约2和约3)。例如,“约100克”的重量能包含90克-110克的重量。此外,当本文描述数值列表(例如,约50%、60%、70%、80%、85%或86%)时,该列表包含其所有中间值和分数值(例如,54%、85.4%)。因此,应理解,尽管通过代表性实施方案和任选的特征具体公开了本技术,但是本领域技术人员能对本文所公开内容进行改良和变化,应认为此类改良和变化落在本技术的范围内。

[1285] 本技术的某些实施方式在所附的权利要求中列出。

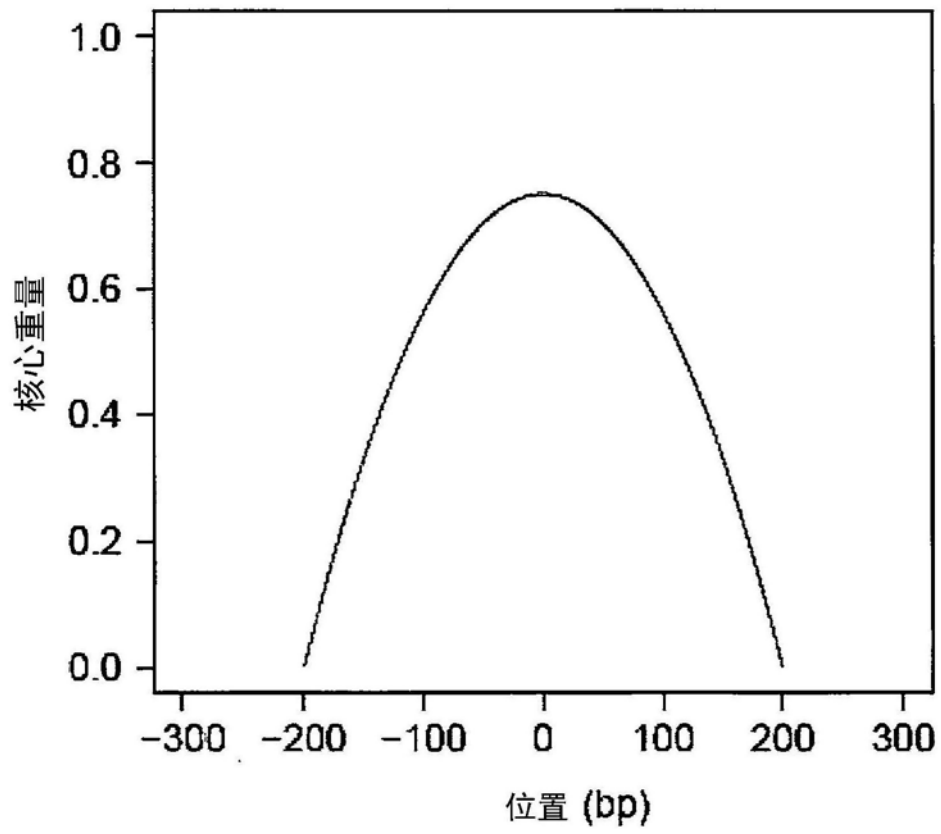


图1

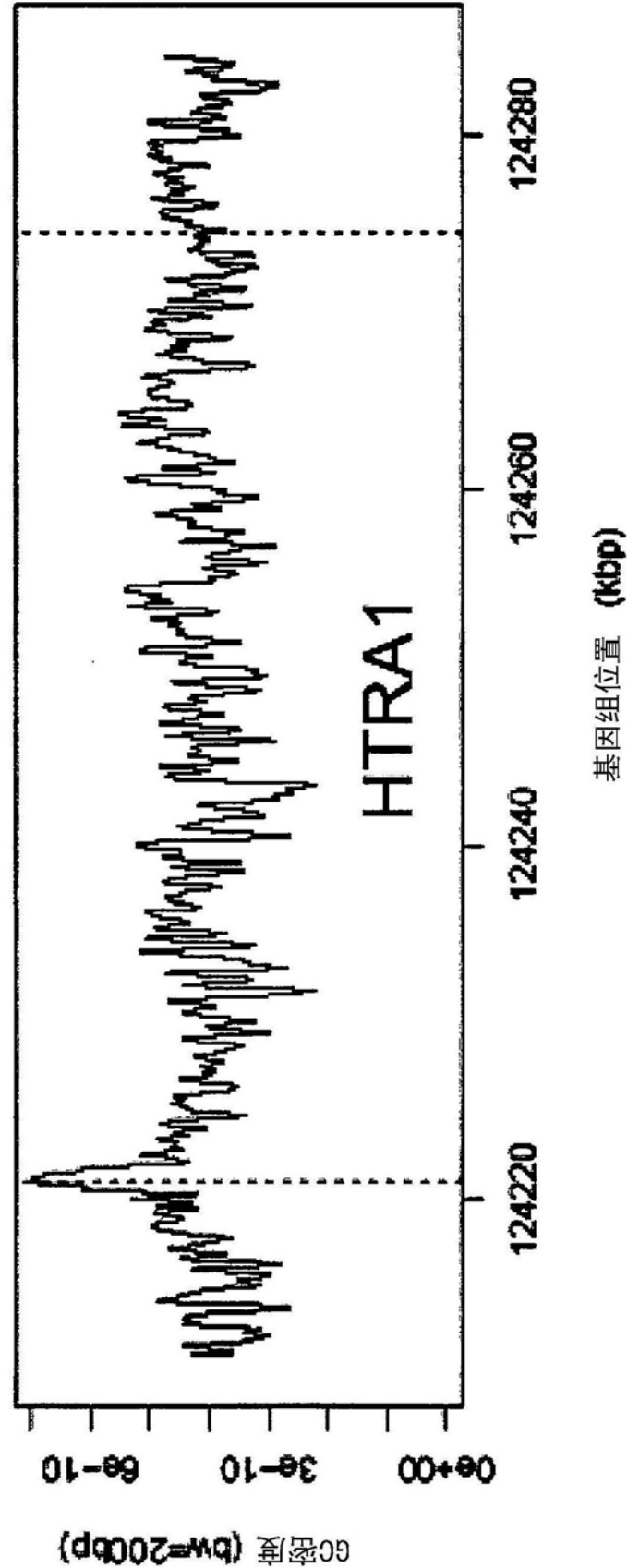


图2

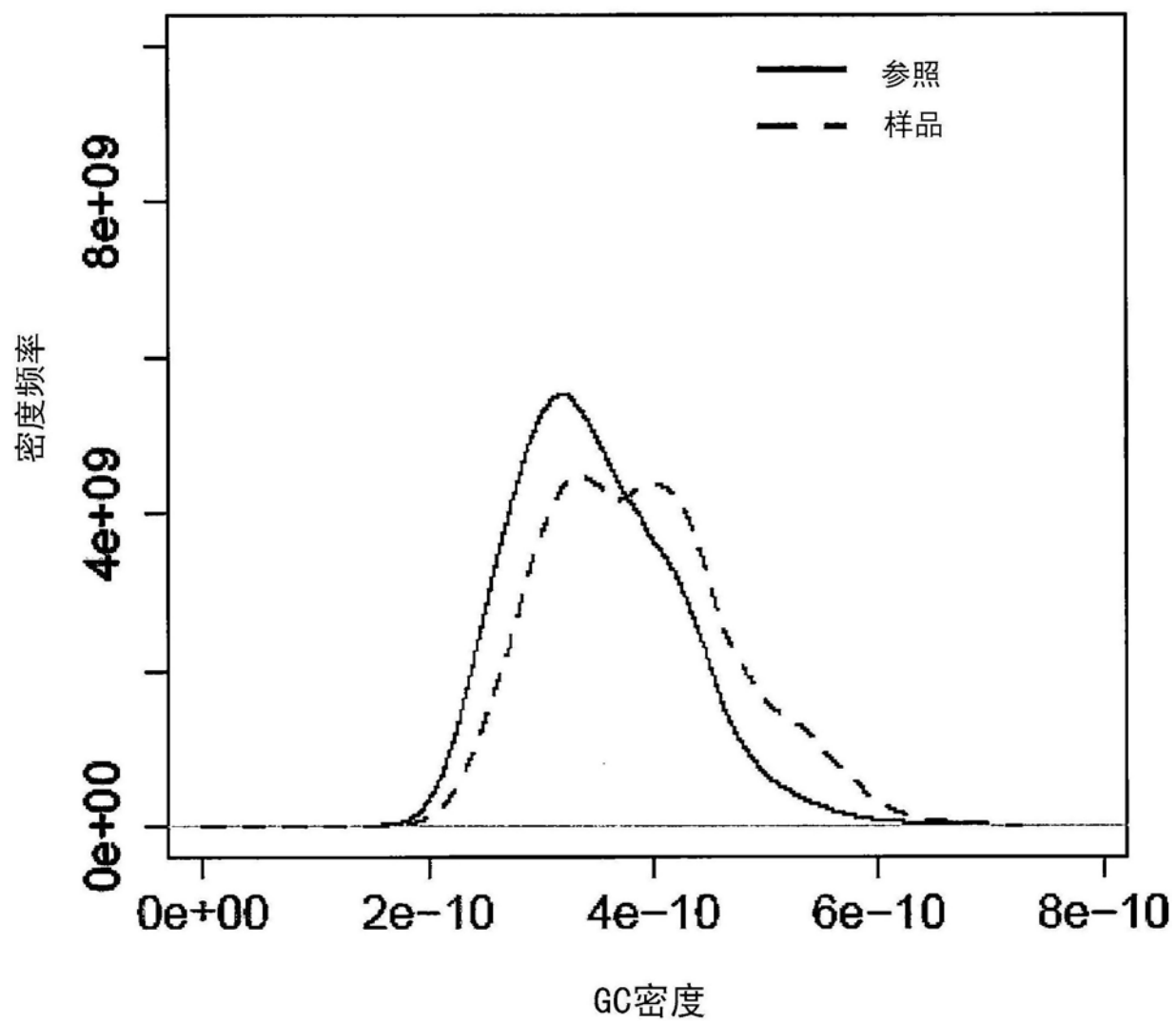


图3

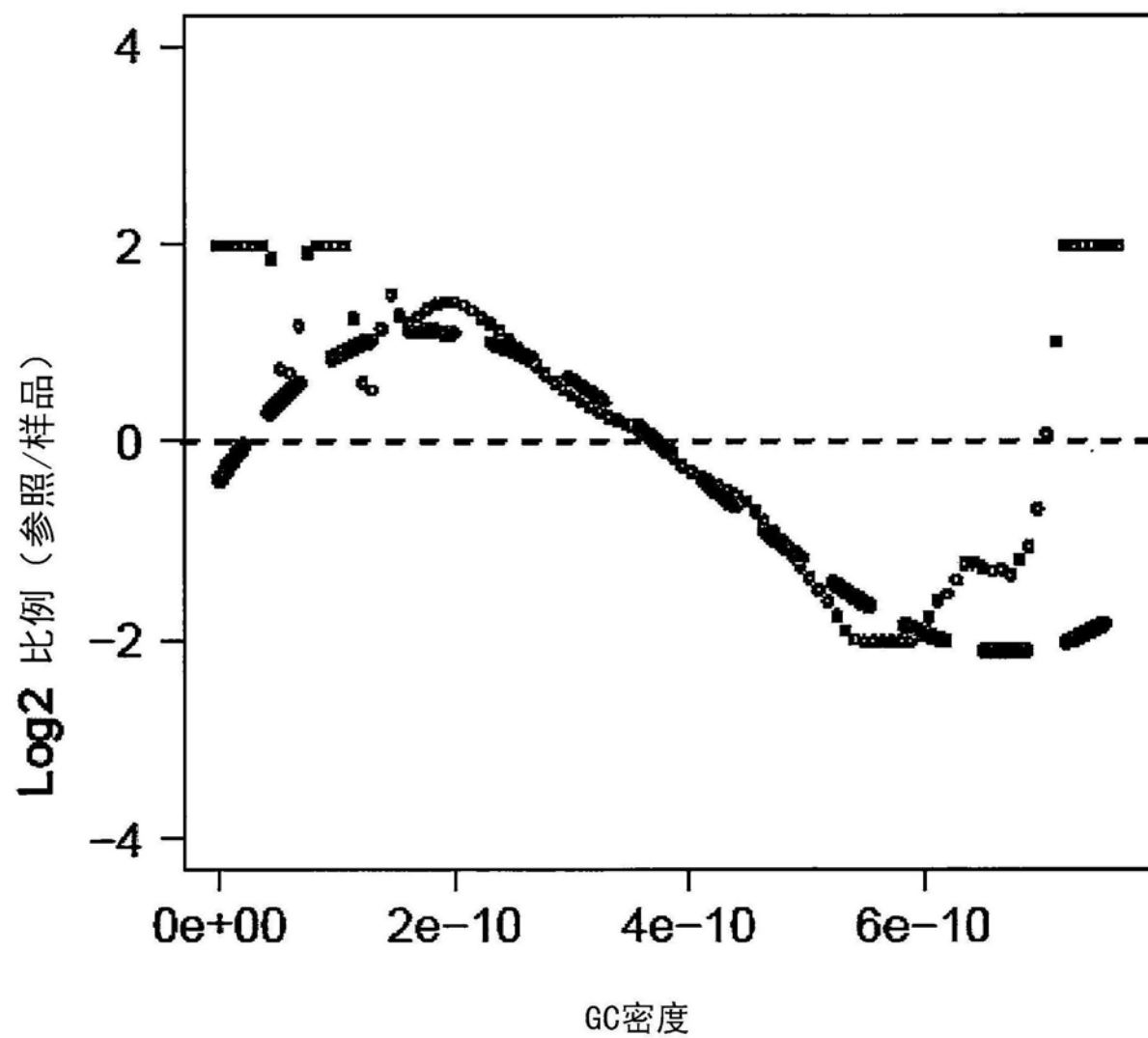


图4

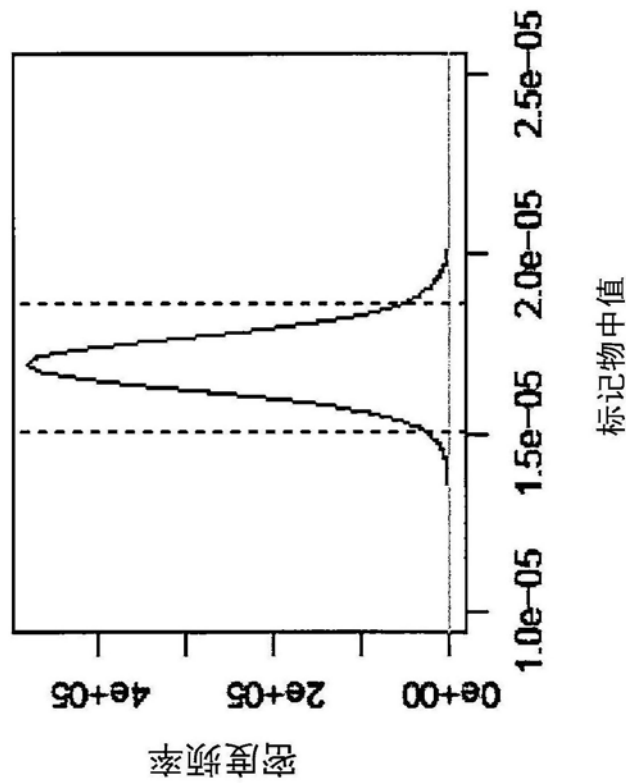


图5A

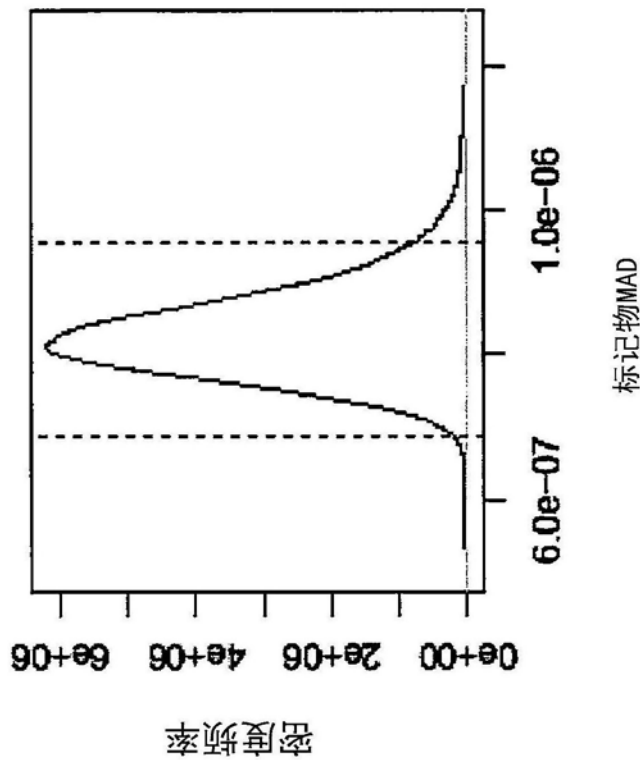


图5B

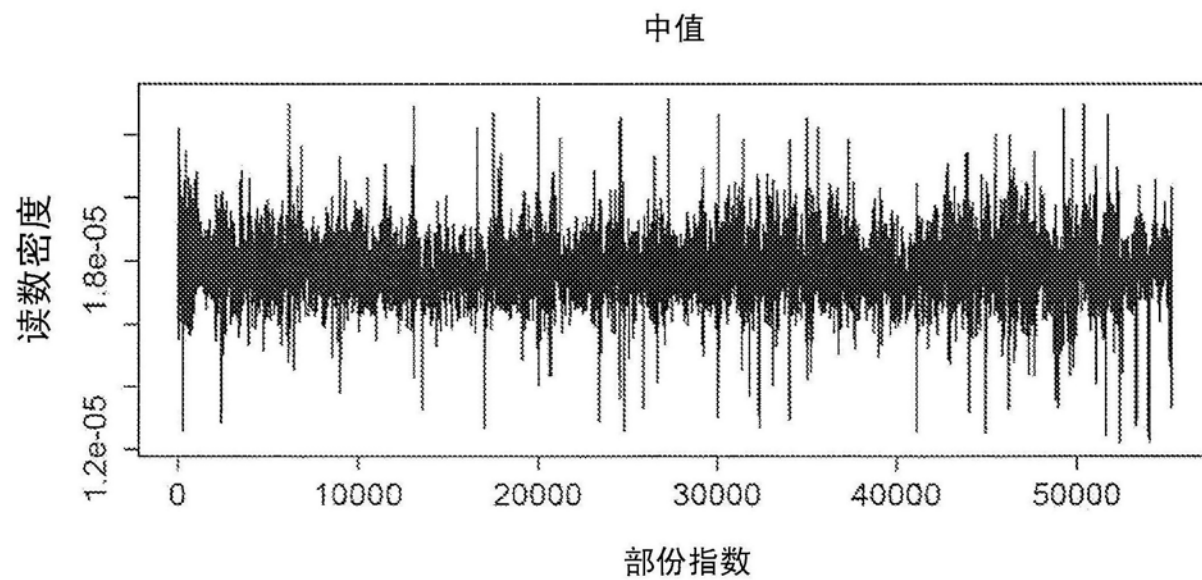


图6A

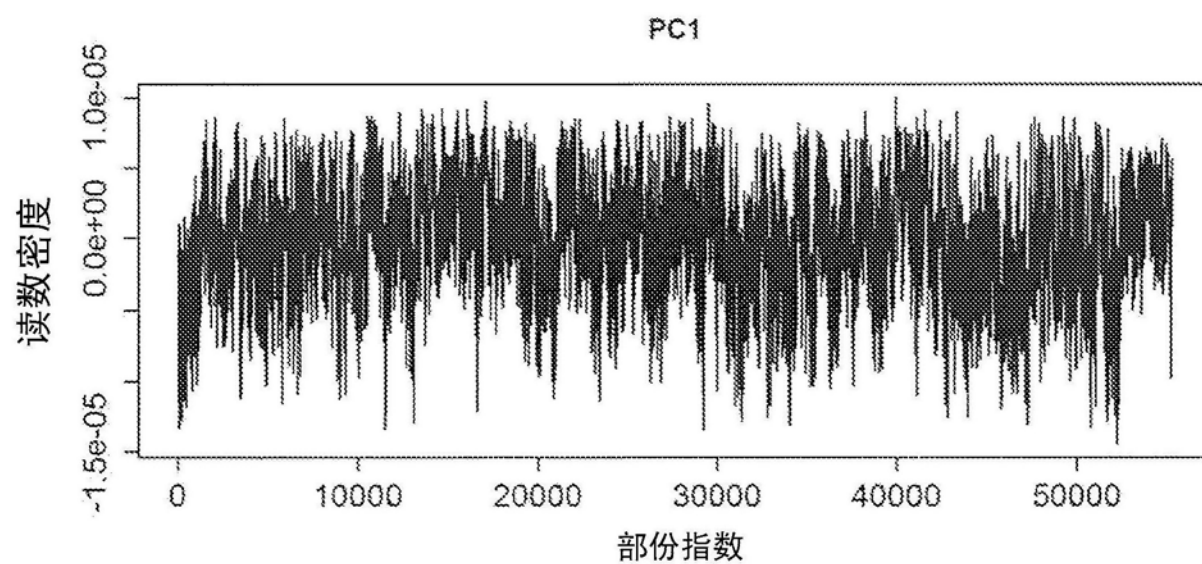


图6B

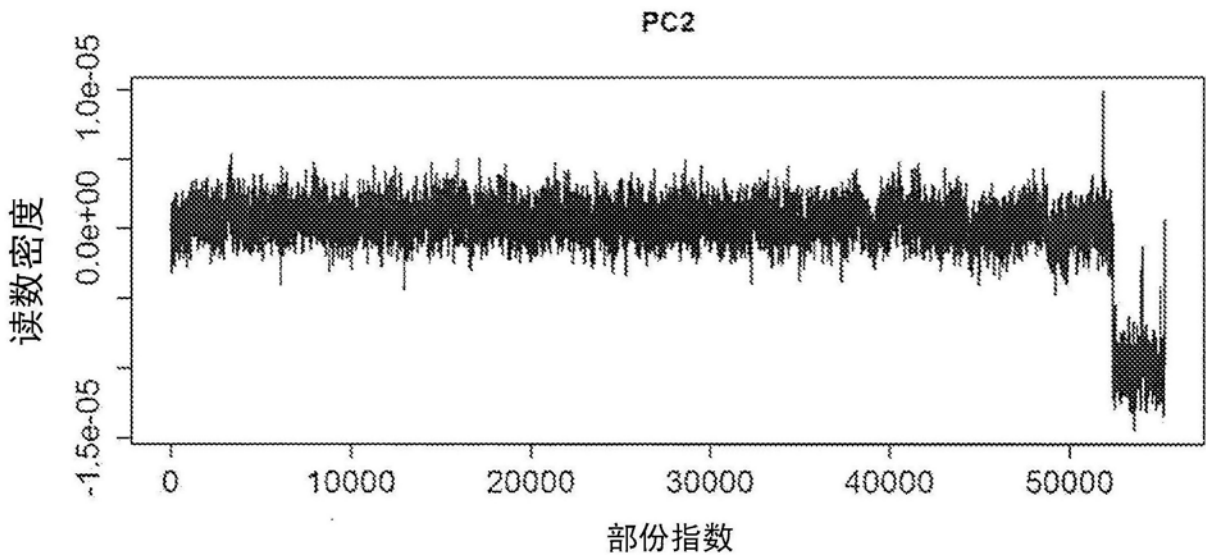


图6C

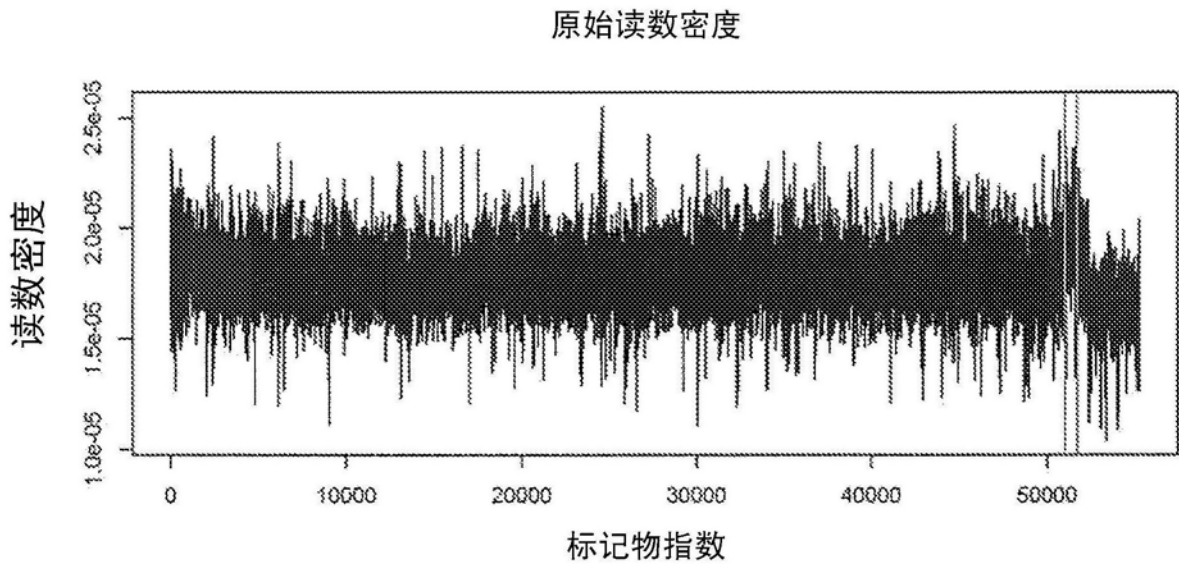


图7A

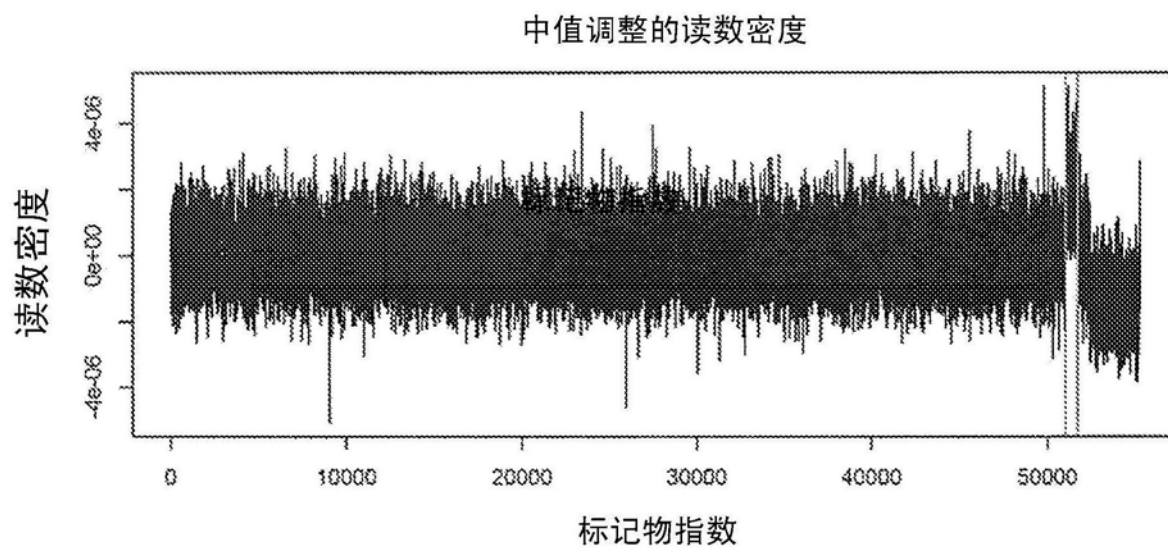


图7B

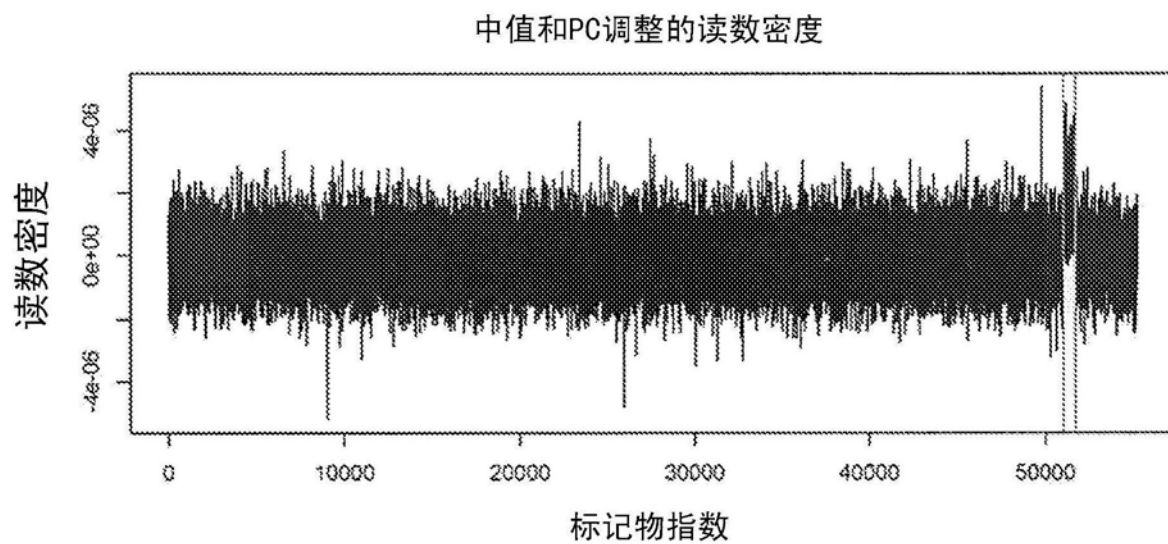


图7C

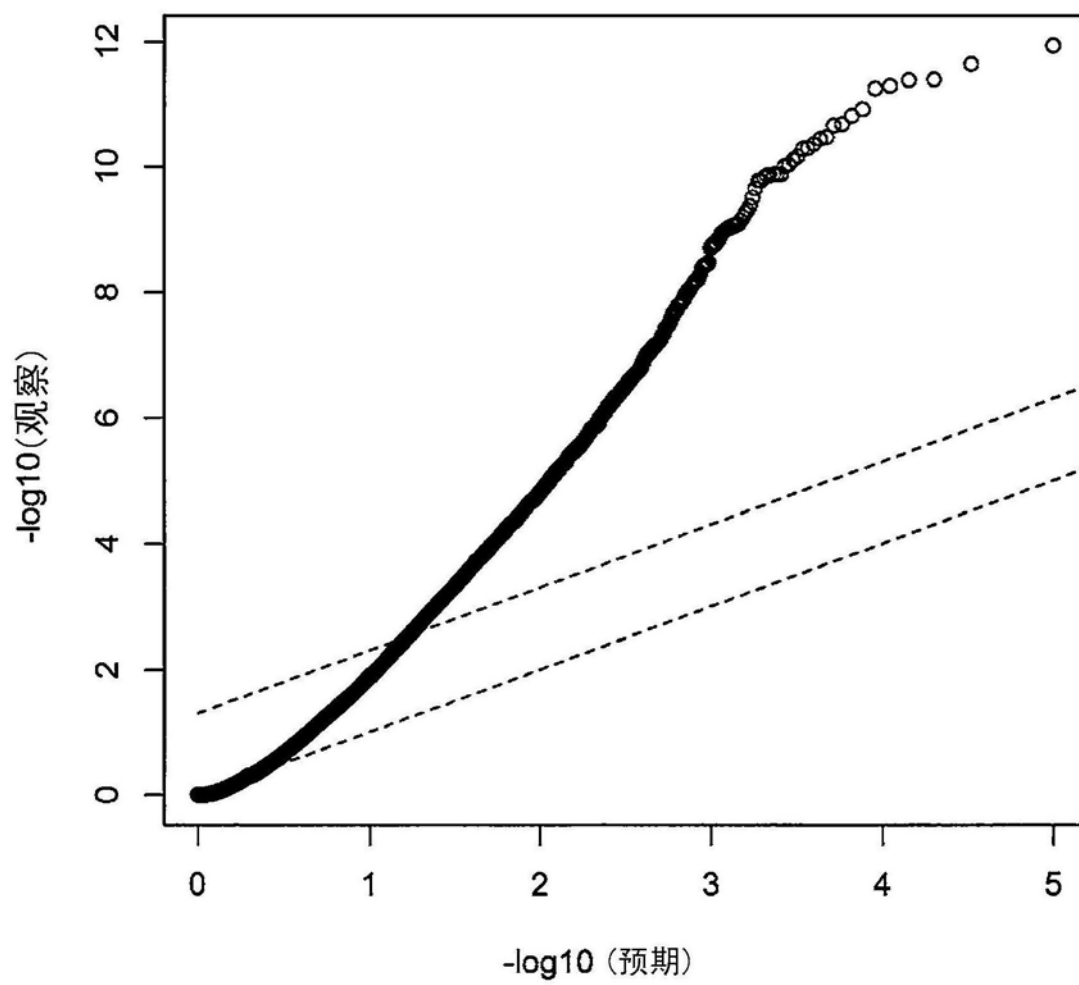


图8

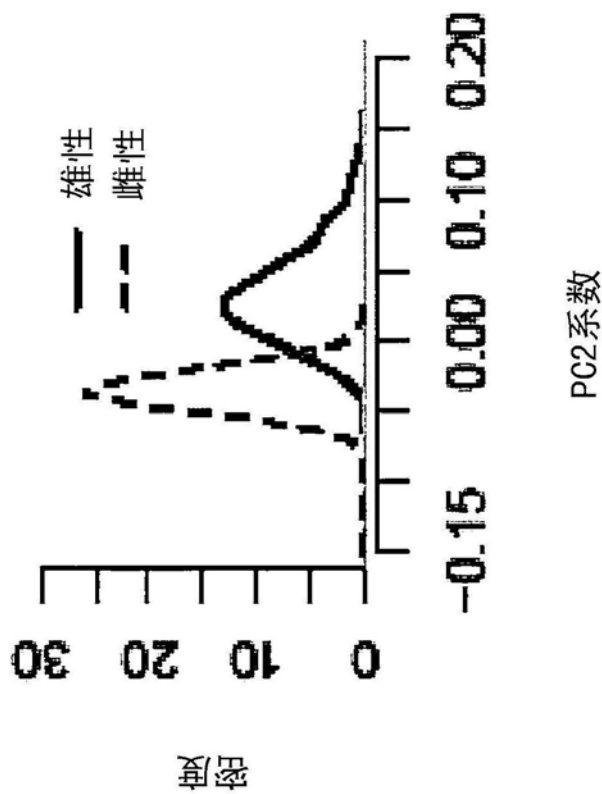


图9A

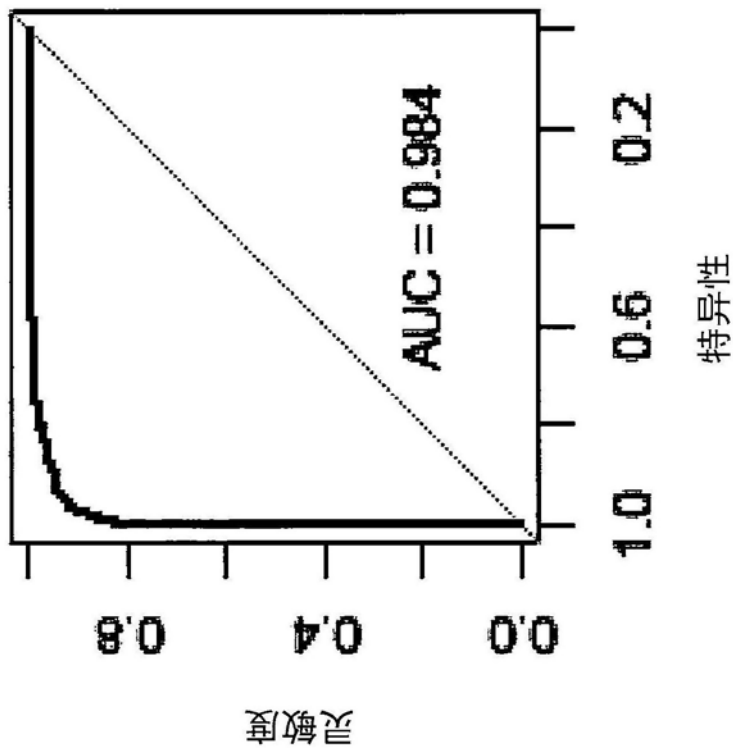


图9B

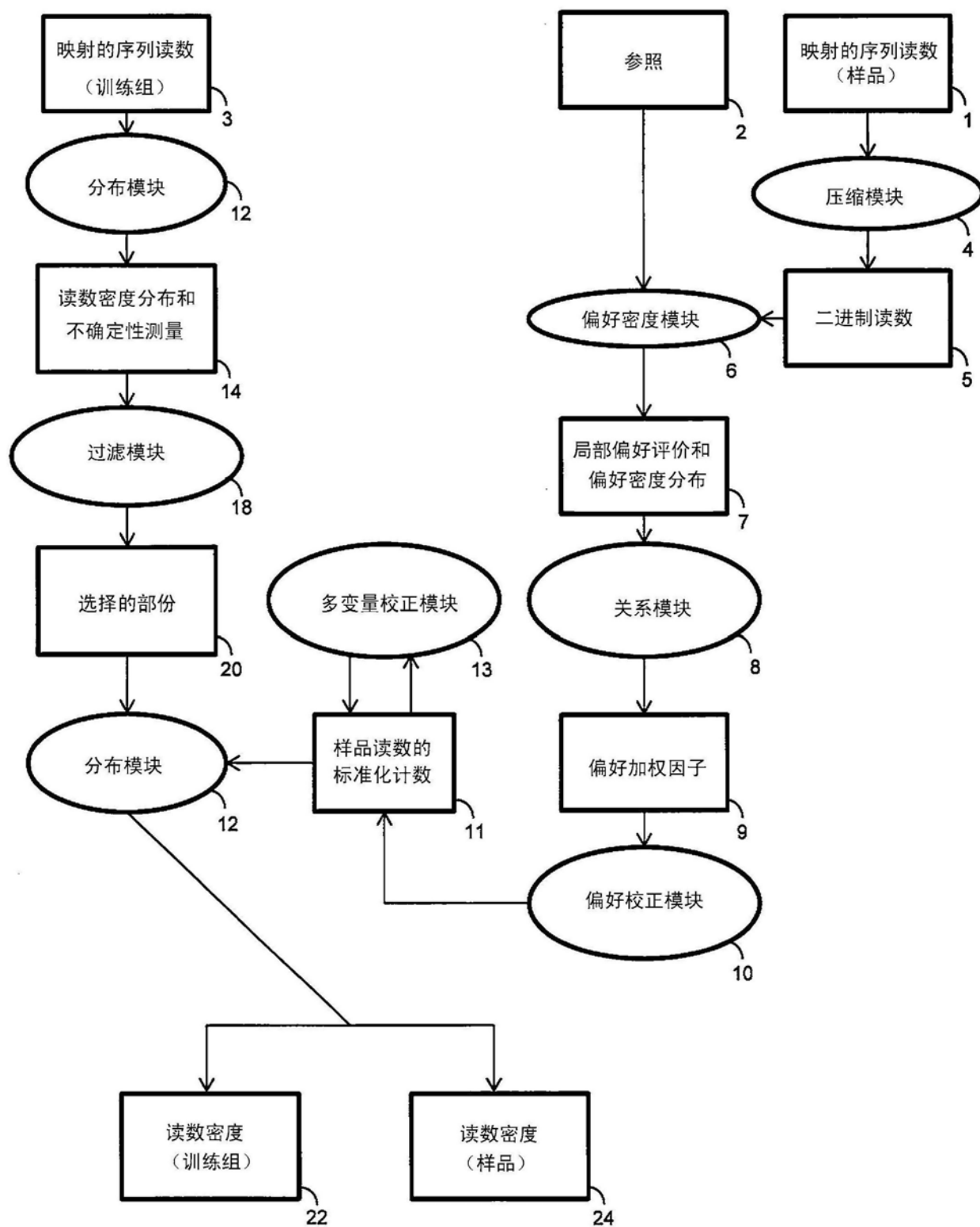


图10A

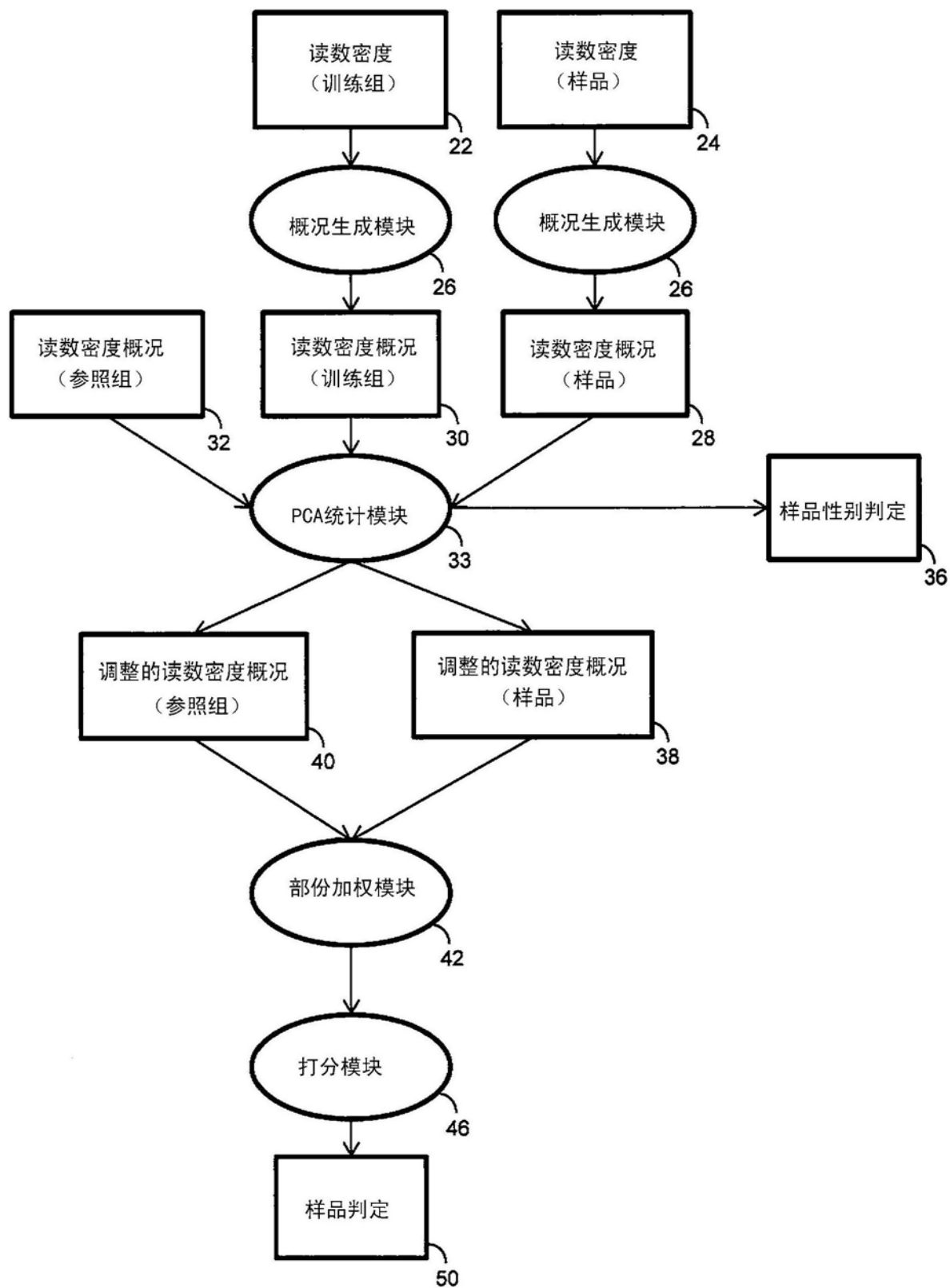


图10B

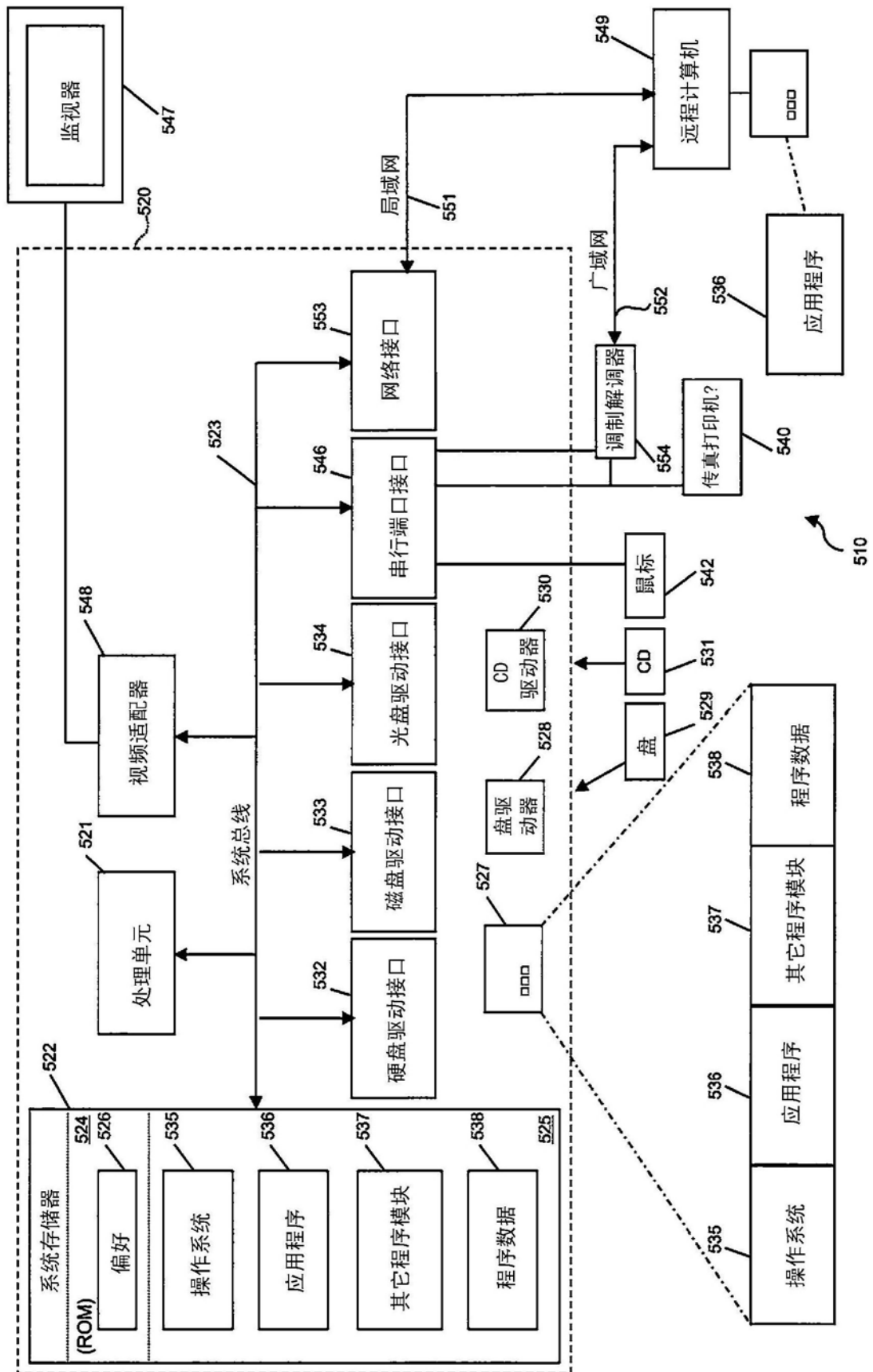


图11

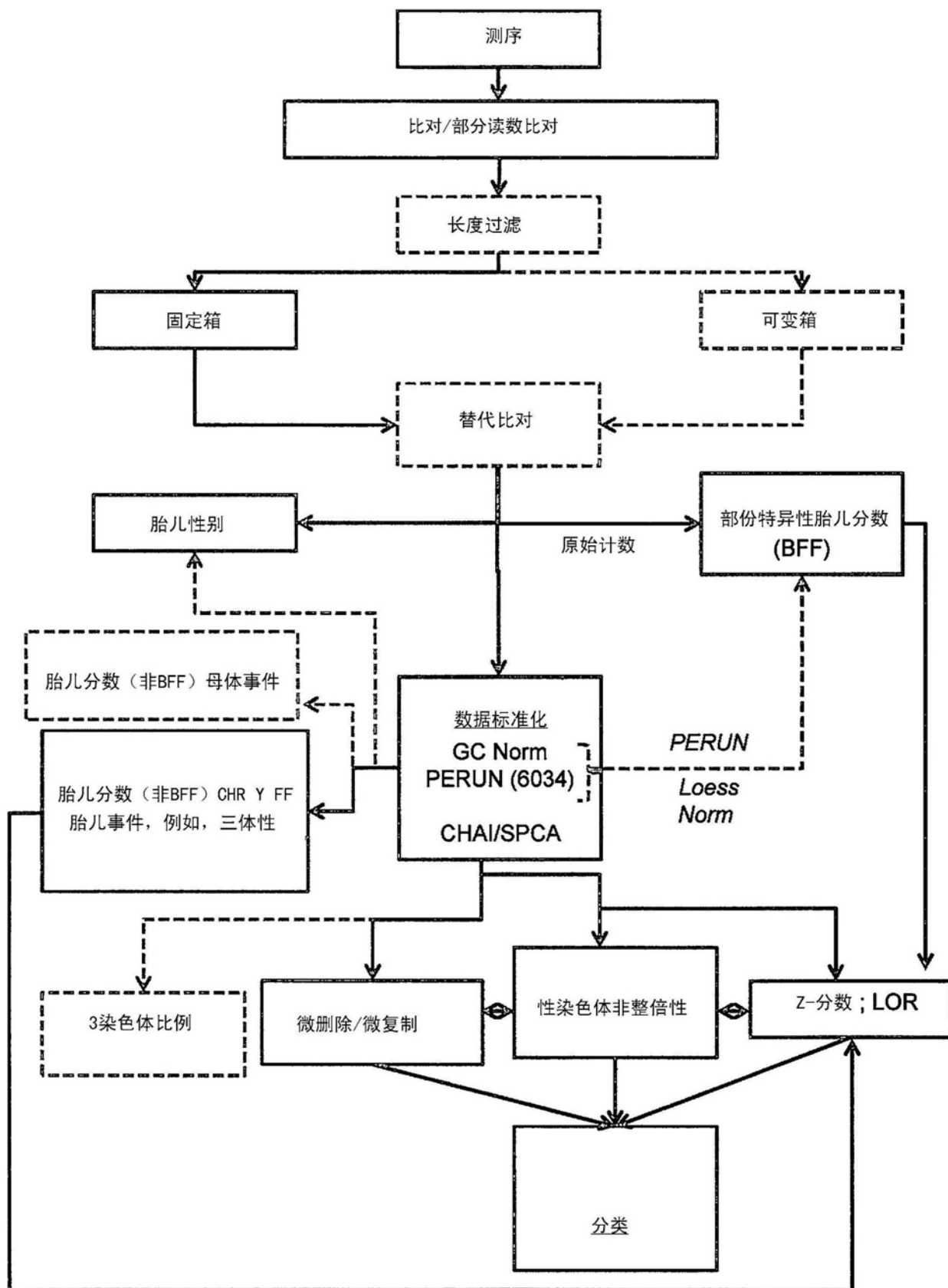


图12