



(12)发明专利申请

(10)申请公布号 CN 106795568 A

(43)申请公布日 2017. 05. 31

(21)申请号 201580054801.9

(22)申请日 2015.10.09

(30)优先权数据

62/062,636 2014.10.10 US

(85)PCT国际申请进入国家阶段日

2017.04.07

(86)PCT国际申请的申请数据

PCT/IB2015/057716 2015.10.09

(87)PCT国际申请的公布数据

W02016/055971 EN 2016.04.14

(71)申请人 因维蒂公司

地址 美国加利福尼亚州

(72)发明人 K·康维卡 K·雅各布斯

(74)专利代理机构 北京市金杜律师事务所

11256

代理人 王茂华

(51)Int.Cl.

G12Q 1/68(2006.01)

G12N 1/15(2006.01)

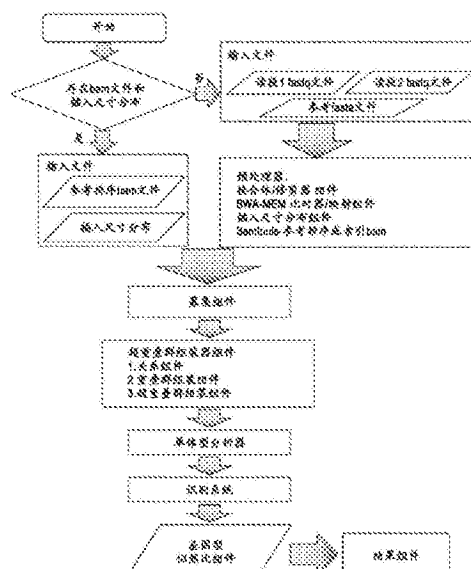
权利要求书13页 说明书41页 附图25页

(54)发明名称

测序读段的DE NOVO组装的方法、系统和过程

(57)摘要

本文提供了映射和组装序列读段的新的方法、系统和过程。本文还提供了标识对象的基因组中遗传变异的存在或不存在的方法、系统和过程。



1. 一种具有存储在其上的可执行程序的非暂时性计算机可读存储介质,所述程序被配置为指示微处理器:

(a) 获得包括多个读段配对物对的双端序列读段的集合,每个对包括两个读段配对物,其中每个对的所述两个读段配对物中的至少一个被映射到参考基因组的至少一部分,所述参考基因组包括预先选择的感兴趣的基因组区域,并且其中所述双端序列读段中的一些未被映射到所述参考基因组的所述至少一部分;

(b) 确定针对序列读段的集合的堆积关系;

(c) 根据(b)中确定的所述堆积关系构建一个或多个重叠群,包括:将至少一个核苷酸迭代地添加到一个或多个起始读段的位置3'或5',其中所述位置包括多数共有核苷酸;

(d) 根据(c)中构建的所述一个或多个重叠群和/或将(c)中构建的所述重叠群中的两个或更多桥接的一个或多个读段配对物对来组装一个或多个超重叠群;

(e) 根据所述一个或多个超重叠群生成基因型似然比;以及

(f) 根据(e)中生成的所述基因型似然比确定基因改变的存在或不存在。

2. 根据权利要求1所述的存储介质,其中所述堆积关系包括所述集合的两个或更多读段之间的多个重叠。

3. 根据权利要求2所述的存储介质,其中根据以下各项选择所述多个重叠中的每一个:

(i) 所述集合中包括与所述集合的第二读段的第一重叠的第一读段, (ii) 所述第一重叠包括大于预定比对得分阈值的比对得分, (iii) 所述第二读段超过所述第一读段的3'端或5'端扩展一个或多个核苷酸,以及(iv) 所述第一重叠包括满足(i)、(ii)和(iii)的所有可能的第一重叠中的最高比对得分。

4. 根据权利要求3所述的存储介质,其中所述第二读段包括与所述集合中的第三读段的第二重叠,其中(i) 所述第二读段包括所述第一重叠, (ii) 所述第二重叠包括大于预定比对得分阈值的比对得分, (iii) 所述第三读段超过所述第二读段的3'端或5'端扩展一个或多个核苷酸,并且所述第二读段和所述第三读段在相同3'或5'方向上扩展所述第一读段,以及(iv) 所述第二重叠包括满足(i)、(ii)和(iii)的所有可能的第二重叠中的最高比对得分。

5. 根据权利要求2至4中任一项所述的存储介质,其中根据针对所述堆积关系确定的所述多个重叠来确定所述多数共有核苷酸。

6. 根据权利要求1至5中任一项所述的存储介质,其中构建所述重叠群包括将至少一个核苷酸迭代地添加到一个或多个中间重叠群的每一个的位置3'或5'。

7. 根据权利要求6所述的存储介质,其中所述位置包括两个不同的多数共有核苷酸,构建所述重叠群包括:生成所述中间重叠群的拷贝,从而提供两个相同的中间重叠群;将所述两个不同的多数共有核苷酸中的一个添加到所述两个相同的中间重叠群中的每一个,其中不同的核苷酸被添加到所述两个相同的中间重叠群中的每一个。

8. 根据权利要求6所述的存储介质,其中所述位置包括三个不同的多数共有核苷酸,构建所述重叠群包括:生成所述中间重叠群的两个拷贝,从而提供三个相同的中间重叠群;将所述三个不同的多数共有核苷酸中的一个添加到所述三个相同的中间重叠群中的每一个,其中不同的核苷酸被添加到所述三个相同的中间重叠群中的每一个。

9. 根据权利要求6所述的存储介质,其中所述位置包括四个不同的多数共有核苷酸,构

建所述重叠群包括：生成所述中间重叠群的三个拷贝，从而提供四个相同的中间重叠群；将所述四个不同的多数共有核苷酸中的一个添加到所述四个相同的中间重叠群中的每一个，其中不同的核苷酸被添加到所述四个相同的中间重叠群中的每一个。

10. 根据权利要求1至9中任一项所述的存储介质，其中所述一个或多个超重叠群包括跨越感兴趣的所述基因组区域的完整长度的重叠群。

11. 根据权利要求1至10中任一项所述的存储介质，其中所述一个或多个超重叠群跨越感兴趣的所述基因组区域的完整长度。

12. 根据权利要求1至11中任一项所述的存储介质，其中所述序列读段从样本获得，所述样本从人类对象获得。

13. 根据权利要求1至12中任一项所述的存储介质，其中所述序列读段中的每一个包括已知的定向。

14. 根据权利要求1至13中任一项所述的存储介质，其中所述读段配对物对中的每一个与插入尺寸分布相关联。

15. 根据权利要求1至13中任一项所述的存储介质，其中根据一个或多个映射权重确定所述基因型假设似然比。

16. 根据权利要求1至15中任一项所述的存储介质，其中根据比对确定所述多数共有核苷酸。

17. 根据权利要求16所述的存储介质，其中根据被比对的5个或更多读段之中的共有序列确定所述多数共有核苷酸。

18. 根据权利要求17所述的存储介质，其中根据被比对的5个或更多读段的10%或更多之中的共有序列确定所述多数共有核苷酸。

19. 根据权利要求2至18中任一项所述的存储介质，其中所述第一读段包括扩展超过所述第一读段的所述5'端的至少两个重叠以及扩展超过所述第一读段的所述3'端的至少两个重叠。

20. 根据权利要求2至19中任一项所述的存储介质，其中所述多个重叠包括针对多个第一读段确定的多个第一重叠。

21. 根据权利要求2至20中任一项所述的存储介质，其中所述多个重叠包括针对多个第二读段确定的多个第二重叠。

22. 根据权利要求2至21中任一项所述的存储介质，其中所述多个重叠中的每一个包括重叠得分。

23. 根据权利要求2至22中任一项所述的存储介质，其中所述多个重叠被过滤。

24. 根据权利要求2至23中任一项所述的存储介质，包括在(c)之前，指示所述微处理器过滤所述多个重叠。

25. 根据权利要求23或24所述的存储介质，其中所述过滤包括剪枝算法。

26. 根据权利要求22至25中任一项所述的存储介质，其中根据所述过滤移除所述多个重叠中的一个或多个。

27. 根据权利要求3至26中任一项所述的存储介质，其中根据预定比对得分阈值移除所述多个重叠中的一个或多个。

28. 根据权利要求2至27中任一项所述的存储介质，其中所述多个重叠不包括插入或缺

失。

29. 根据权利要求1至28中任一项所述的存储介质,包括根据所述堆积关系生成平铺图。

30. 根据权利要求2至29中任一项所述的存储介质,其中根据k-mer哈希策略确定所述多个重叠中的每一个。

31. 根据权利要求1至30中任一项所述的存储介质,其中所述起始读段包括位于所述预先选择的感兴趣的基因组区域的最5'侧的读段。

32. 根据权利要求1至31中任一项所述的存储介质,其中所述起始读段包括位于所述预先选择的感兴趣的基因组区域的最3'侧的读段。

33. 根据权利要求1至32中任一项所述的存储介质,其中两个或更多重叠群由第一读段配对物对中与第一重叠群重叠的第一读段配对物以及所述第一读段配对物对中与第二重叠群重叠的第二读段配对物桥接。

34. 根据权利要求33所述的存储介质,其中根据多个读段配对物对,所述第一重叠群被连接到所述第二重叠群。

35. 根据权利要求1至34中任一项所述的存储介质,其中(d)中的所述两个或更多重叠群中的两个的所述连接包括:构建图,其中第一重叠群和第二重叠群包括顶点,并且读段配对物对中的读段配对物的一个或多个重叠包括被定向的边缘。

36. 根据权利要求34或35所述的存储介质,其中所述第一重叠群和所述第二重叠群由至少5个读段配对物对连接,所述至少5个读段配对物对中的每一个包括与所述第一重叠群重叠的读段配对物以及与所述第二重叠群重叠的读段配对物。

37. 根据权利要求1至36中任一项所述的存储介质,其中所述遗传变异包括短串联重复。

38. 根据权利要求1至37中任一项所述的存储介质,其中所述遗传变异包括一个或多个单核苷酸多态性。

39. 根据权利要求12至38中任一项所述的存储介质,其中所述人类对象是二倍体。

40. 根据权利要求1至39中任一项所述的存储介质,其中所述基因型似然比根据等式(1)确定:

$$\frac{P(G|\{R\})}{P(G_0|\{R\})} = \prod_{\{R\}} \frac{\sum_{\{A_G\}} \frac{1}{N_{A_G}} F_{A_G}(W(R, A_G) + \alpha)}{\sum_{\{A_{G_0}\}} \frac{1}{N_{A_{G_0}}} F_{A_{G_0}}(W(R, A_{G_0}) + \alpha)} \quad (\text{等式 1})$$

其中G是预定倍性的基因型序列,G₀是参考序列,{R}是所述读段配对物对R的集合,N_{AG}是所述基因型序列G中的等位基因AG的数量,N_{AG0}是在所述参考序列G₀中的等位基因AG0的数量,以及F_{AG}是所述等位基因AG在所述基因型序列G中的分数,F_{AG0}是所述等位基因AG0在所述参考序列G₀中的分数,W是读段对映射权重,以及α是映射概率常数。

41. 根据权利要求1至40中任一项所述的存储介质,其中所述遗传变异被包括在从以下各项选择的基因中:AR、ATXN1、ATXN2、ATXN7、ATXN8、ATXN10、DMPK、FXN、JPH3、CACNA1A、PPP2R2B、TBP、ATN1、ARX、PHOX2B、PABPN1、ATT、CFTR和BRCA1。

42. 根据权利要求1至41中任一项所述的存储介质,其中(a)、(b)、(c)和(d)不包括包含误差校正的过程。

43. 根据权利要求1至42中任一项所述的存储介质,其中所述序列读段包括排序的bam文件。

44. 根据权利要求1至43中任一项所述的存储介质,其中所述序列读段包括插入尺寸分布文件。

45. 根据权利要求1至44中任一项所述的存储介质,其中所述序列读段包括fastq文件和fasta文件中的一个或多个。

46. 根据权利要求1至45中任一项所述的存储介质,其中所述序列读段在(b)之前被修剪。

47. 根据权利要求13至46中任一项所述的存储介质,其中插入尺寸似然性被确定。

48. 根据权利要求47所述的存储介质,其中估计的插入尺寸似然性由插入尺寸分布组件确定。

49. 根据权利要求1至48中任一项所述的存储介质,其中所述参考基因组的所述至少一部分是染色体的一部分。

50. 根据权利要求1至49中任一项所述的存储介质,其中所述参考基因组的所述至少一部分是第一染色体的一部分以及第二染色体的一部分。

51. 根据权利要求1至50中任一项所述的存储介质,其中所述读段配对物对包括不一致的读段配对物。

52. 根据权利要求1至51中任一项所述的存储介质,其中所有可能的重叠在(c)之前被标识。

53. 根据权利要求1至52中任一项所述的存储介质,其中(e)的生成所述基因型似然比包括将所述序列读段重新比对到所述一个或多个超重叠群。

54. 根据权利要求40至53中任一项所述的存储介质,其中所述倍性是二倍体。

55. 根据权利要求1至54中任一项所述的存储介质,其中对于(c)中构建的所述一个或多个重叠群中的任何一个的所述构建,所述多个读段配对物对中的每一个不被使用多于一次。

56. 根据权利要求1至55中任一项所述的存储介质,其中根据读段配对物对的所述两个读段配对物中的每一个的定向组装(d)中组装的所述多个超重叠群中的两个或更多。

57. 根据权利要求4至56中任一项所述的存储介质,其中(d)中组装的所述多个超重叠群中的两个或更多与一个或多个插入尺寸似然性相关联。

58. 根据权利要求1至57中任一项所述的存储介质,其中生成所述基因型似然比包括:根据跨越预定倍性的感兴趣的所述基因组区域的所述完整长度的所述一个或多个超重叠群来确定一个或多个可能的基因型。

59. 根据权利要求58所述的存储介质,其中所述一个或多个可能的基因型中的每一个与概率相关联。

60. 根据权利要求1至59中任一项所述的存储介质,其中生成所述基因型似然比包括:根据一个或多个单体型确定一个或多个可能的基因型,其中根据跨越感兴趣的所述基因组区域的所述完整长度的超重叠群确定每个单体型。

61. 根据权利要求1至60中任一项所述的存储介质,其中针对对象确定基因改变的存在或不存在。

62. 一种用于确定对象中基因改变的存在或不存在的系统, 所述系统包括一个或多个硬件处理器, 所述一个或多个硬件处理器被配置为执行计算机程序组件, 所述计算机程序组件包括:

(a) 募集组件, 被配置为获得包括多个读段配对物对的双端序列读段的集合, 每个对包括两个读段配对物, 其中每个对中的所述两个读段配对物中的至少一个被映射到参考基因组的至少一部分, 所述参考基因组包括预先选择的感兴趣的基因组区域, 并且其中所述双端序列读段中的一些未被映射到所述参考基因组的所述至少一部分;

(b) 堆积关系组件, 被配置为针对序列读段的所述集合确定堆积关系;

(c) 重叠群组装组件, 被配置为根据 (b) 中确定的所述堆积关系构建一个或多个重叠群, 所述构建包括将至少一个核苷酸迭代地添加到一个或多个起始读段的位置 3' 或 5', 其中所述位置包括多数共有核苷酸;

(d) 超重叠群组装组件, 被配置为根据 (c) 中构建的所述一个或多个重叠群和/或将 (c) 中构建的所述重叠群中的两个或更多桥接的一个或多个读段配对物对来组装一个或多个超重叠群; 以及

(e) 识别系统组件, 被配置为根据所述一个或多个超重叠群生成基因型似然比。

63. 根据权利要求62所述的系统, 进一步包括单体型组件, 所述单体型组件被配置为从所述一个或多个超重叠群构建单体型。

64. 根据权利要求63所述的系统, 其中所述单体型组件被配置为将所有读段重新映射到所述一个或多个超重叠群。

65. 根据权利要求63或64所述的系统, 其中所述单体型组件被配置为标识所述一个或多个超重叠群中的不规则性。

66. 根据权利要求62至65中任一项所述的系统, 进一步包括插入尺寸分布组件, 所述插入尺寸分布组件被配置为确定由读段对中两个读段的映射暗示的插入尺寸的概率。

67. 根据权利要求66所述的系统, 其中所述插入尺寸分布组件被配置为针对两个或更多读段对确定插入尺寸的两个或更多概率。

68. 根据权利要求62至67中任一项所述的系统, 进一步包括结果组件, 所述结果组件被配置为根据 (e) 中生成的所述基因型似然比确定所述对象中基因改变的存在或不存在。

69. 根据权利要求62至68中任一项所述的系统, 其中所述堆积关系包括所述集合中的两个或更多读段之间的多个重叠。

70. 根据权利要求69所述的系统, 其中根据以下各项选择所述多个重叠中的每一个: (i) 所述集合中包括与所述集合的第二读段的第一重叠的第一读段, (ii) 所述第一重叠包括大于预定比对得分阈值的比对得分, (iii) 所述第二读段超过所述第一读段的 3' 端或 5' 端扩展一个或多个核苷酸, 以及 (iv) 所述第一重叠包括满足 (i)、(ii) 和 (iii) 的所有可能的第一重叠中的最高比对得分。

71. 根据权利要求70所述的系统, 其中所述第二读段包括与所述集合中的第三读段的第二重叠, 其中 (i) 所述第二读段包括所述第一重叠, (ii) 所述第二重叠包括大于预定比对得分阈值的比对得分, (iii) 所述第三读段超过所述第二读段的 3' 端或 5' 端扩展一个或多个核苷酸, 并且所述第二读段和所述第三读段在相同 3' 或 5' 方向上扩展所述第一读段, 以及 (iv) 所述第二重叠包括满足 (i)、(ii) 和 (iii) 的所有可能的第二重叠中的最高比对得

分。

72. 根据权利要求69至71中任一项所述的系统,其中根据针对所述堆积关系确定的所述多个重叠确定所述多数共有核苷酸。

73. 根据权利要求62至72中任一项所述的系统,其中构建所述重叠群包括将至少一个核苷酸迭代地添加到一个或多个中间重叠群中的每一个的位置3' 或5' 。

74. 根据权利要求73所述的系统,其中所述位置包括两个不同的多数共有核苷酸,构建所述重叠群包括:生成所述中间重叠群的拷贝,从而提供两个相同的中间重叠群;将所述两个不同的多数共有核苷酸中的一个添加到所述两个相同的中间重叠群中的每一个,其中不同的核苷酸被添加到所述两个相同的中间重叠群中的每一个。

75. 根据权利要求73所述的系统,其中所述位置包括三个不同的多数共有核苷酸,构建所述重叠群包括:生成所述中间重叠群的两个拷贝,从而提供三个相同的中间重叠群;将所述三个不同的多数共有核苷酸中的一个添加到所述三个相同的中间重叠群中的每一个,其中不同的核苷酸被添加到所述三个相同的中间重叠群中的每一个。

76. 根据权利要求73所述的系统,其中所述位置包括四个不同的多数共有核苷酸,构建所述重叠群包括:生成所述中间重叠群的两个拷贝,从而提供四个相同的中间重叠群;将所述四个不同的多数共有核苷酸中的一个添加到所述四个相同的中间重叠群中的每一个,其中不同的核苷酸被添加到所述四个相同的中间重叠群中的每一个。

77. 根据权利要求62至76中任一项所述的系统,其中所述一个或多个超重叠群包括跨越所述感兴趣的基因组区域的完整长度的重叠群。

78. 根据权利要求1至77中任一项所述的系统,其中所述一个或多个超重叠群跨越所述感兴趣的基因组区域的完整长度。

79. 根据权利要求62至78中任一项所述的系统,其中所述序列读段从样本获得,所述样本从人类对象获得。

80. 根据权利要求62至79中任一项所述的系统,其中所述序列读段中的每一个包括已知的定向。

81. 根据权利要求62至80中任一项所述的系统,其中所述读段配对物对的每一个与插入尺寸分布相关联。

82. 根据权利要求62至80中任一项所述的系统,其中根据一个或多个映射权重确定所述基因型假设似然比。

83. 根据权利要求62至82中任一项所述的系统,其中根据比对确定所述多数共有核苷酸。

84. 根据权利要求83所述的系统,其中根据被比对的5个或更多读段之中的共有序列确定所述多数共有核苷酸。

85. 根据权利要求84所述的系统,其中根据被比对的5个或更多读段中的10%或更多之中的共有序列确定所述多数共有核苷酸。

86. 根据权利要求69至85中任一项所述的系统,其中所述第一读段包括扩展超过所述第一读段的5' 端的至少两个重叠以及扩展超过所述第一读段的3' 端的至少两个重叠。

87. 根据权利要求69至86中任一项所述的系统,其中所述多个重叠包括针对多个第一读段确定的多个第一重叠。

86. 根据权利要求69至87中任一项所述的系统, 其中所述多个重叠包括针对多个第二读段确定的多个第二重叠。

87. 根据权利要求69至86中任一项所述的系统, 其中所述多个重叠中的每一个包括重叠得分。

88. 根据权利要求69至87所述的系统, 其中所述多个重叠被过滤。

89. 根据权利要求69至87中任一项所述的系统, 包括在(c)之前, 指示所述微处理器过滤所述多个重叠。

90. 根据权利要求88或89所述的系统, 其中所述过滤包括剪枝算法。

91. 根据权利要求87至90中任一项所述的系统, 其中根据所述过滤移除所述多个重叠中的一个或多个。

92. 根据权利要求70至91中任一项所述的系统, 其中根据预定比对得分阈值移除所述多个重叠中的一个或多个。

93. 根据权利要求69至92中任一项所述的系统, 其中所述多个重叠不包括插入或缺失。

94. 根据权利要求62至93中任一项所述的系统, 包括根据所述堆积关系生成平铺图。

95. 根据权利要求69至94中任一项所述的系统, 其中根据k-mer哈希策略确定所述多个重叠中的每一个。

96. 根据权利要求62至95中任一项所述的系统, 其中所述起始读段包括位于所述预先选择的感兴趣的基因组区域的最5'侧的读段。

97. 根据权利要求62至96中任一项所述的系统, 其中所述起始读段包括位于所述预先选择的感兴趣的基因组区域的最3'侧的读段。

98. 根据权利要求62至97中任一项所述的系统, 其中所述两个或更多重叠群由第一读段配对物对中与第一重叠群重叠的第一读段配对物以及第一读段配对物对中与第二重叠群重叠的第二读段配对物桥接。

99. 根据权利要求98所述的系统, 其中根据多个读段配对物对, 所述第一重叠群被连接到所述第二重叠群。

100. 根据权利要求62至99中任一项所述的系统, 其中(d)中的所述两个或更多重叠群中的两个的所述连接包括: 构建图, 其中第一重叠群和第二重叠群包括顶点, 并且读段配对物对中的读段配对物的一个或多个重叠包括被定向的边缘。

101. 根据权利要求99或100所述的系统, 其中所述第一重叠群和所述第二重叠群由至少5个读段配对物对连接, 所述至少5个读段配对物对中的每一个包括与所述第一重叠群重叠的读段配对物以及与所述第二重叠群重叠的读段配对物。

102. 根据权利要求62至101中任一项所述的系统, 其中所述遗传变异包括短串联重复。

103. 根据权利要求62至102中任一项所述的系统, 其中所述遗传变异包括一个或多个单核苷酸多态性。

104. 根据权利要求79至103中任一项所述的系统, 其中所述人类对象是二倍体。

105. 根据权利要求62至104中任一项所述的系统, 其中根据等式(1)确定所述基因型似然比:

$$\frac{P(G|\{R\})}{P(G_0|\{R\})} = \prod_{\{R\}} \frac{\sum_{\{A_G\}} \frac{1}{N_{A_G}} F_{A_G}(W(R, A_G) + \alpha)}{\sum_{\{A_{G_0}\}} \frac{1}{N_{A_{G_0}}} F_{A_{G_0}}(W(R, A_{G_0}) + \alpha)} \quad (\text{等式 1})$$

其中G是针对预定倍性的基因型序列, G₀是参考序列, {R} 是所述读段配对物对R的集合, N_{AG}是所述基因型序列G中的等位基因AG的数量, N_{AG₀}是所述参考序列G₀中等位基因AG₀的数量, 以及F_{AG}是所述等位基因AG在所述基因型序列G中的分数, F_{AG₀}是所述等位基因AG₀在所述参考序列G₀中的分数, W是读段对映射权重, 并且α是映射概率常数。

106. 根据权利要求62至105中任一项所述的系统, 其中所述遗传变异被包括在从以下各项选择的基因内: AR、ATXN1、ATXN2、ATXN7、ATXN8、ATXN10、DMPK、FXN、JPH3、CACNA1A、PPP2R2B、TBP、ATN1、ARX、PHOX2B、PABPN1、ATT、CFTR和BRCA1。

107. 根据权利要求62至106中任一项所述的系统, 其中 (a)、(b)、(c) 和 (d) 不包括包含误差校正的过程。

108. 根据权利要求62至107中任一项所述的系统, 其中所述序列读段包括排序的bam文件。

109. 根据权利要求62至108中任一项所述的系统, 其中所述序列读段包括插入尺寸分布文件。

110. 根据权利要求62至109中任一项所述的系统, 其中所述序列读段包括fastq文件和fasta文件中的一个或多个。

111. 根据权利要求62至110中任一项所述的系统, 其中所述序列读段在 (b) 之前被修剪。

112. 根据权利要求81至111中任一项所述的系统, 其中插入尺寸似然性被确定。

113. 根据权利要求112所述的系统, 其中估计的插入尺寸似然性由插入尺寸分布组件确定。

114. 根据权利要求62至113中任一项所述的系统, 其中所述参考基因组的所述至少一部分是染色体的一部分。

115. 根据权利要求62至114中任一项所述的系统, 其中所述参考基因组的所述至少一部分是第一染色体的一部分以及第二染色体的一部分。

116. 根据权利要求62至115中任一项所述的系统, 其中所述读段配对物对包括不一致的读段配对物。

117. 根据权利要求62至116中任一项所述的系统, 其中所有可能的重叠在 (c) 之前被标识。

118. 根据权利要求62至117中任一项所述的系统, 其中 (e) 的生成所述基因型似然比包括将所述序列读段重新比对到所述一个或多个超重叠群。

119. 根据权利要求105至118中任一项所述的系统, 其中所述倍性是二倍体。

120. 根据权利要求62至118中任一项所述的系统, 其中针对 (c) 中构建的所述一个或多个重叠群中的任一个的所述构建, 所述多个读段配对物对中的每一个不被使用多于一次。

121. 根据权利要求62至120中任一项所述的系统, 其中根据读段配对物对的所述两个读段配对物中的每一个的定向组装 (d) 中组装的所述多个超重叠群中的两个或更多。

122. 根据权利要求71至120中任一项所述的系统, 其中 (d) 中组装的所述多个超重叠群

中的两个或更多与一个或多个插入尺寸似然性相关联。

123. 根据权利要求62至122中任一项所述的系统, 其中生成所述基因型似然比包括: 根据跨越针对预定倍性的所述感兴趣的基因组区域的所述完整长度的所述一个或多个超重叠群来确定一个或多个可能的基因型。

124. 根据权利要求123所述的系统, 其中所述一个或多个可能的基因型中的每一个与概率相关联。

125. 根据权利要求62至122中任一项所述的系统, 其中生成所述基因型似然比包括根据一个或多个单体型确定一个或多个可能的基因型, 其中根据跨越所述感兴趣的基因组区域的所述完整长度的超重叠群确定每个单体型。

126. 根据权利要求62至125中任一项所述的系统, 其中针对对象确定基因改变的存在或不存在。

127. 一种用于确定对象中基因改变的存在或不存在的计算机实现的方法, 包括:

(a) 获得包括多个读段配对物对的双端序列读段的集合, 每个对包括两个读段配对物, 其中每个对的所述两个读段配对物中的至少一个被映射到参考基因组的至少一部分, 所述参考基因组包括预先选择的感兴趣的基因组区域, 并且其中所述双端序列读段中的一些未被映射到所述参考基因组的所述至少一部分;

(b) 针对序列读段的所述集合, 确定堆积关系;

(c) 根据(b)中确定的所述堆积关系构建一个或多个重叠群, 包括: 将至少一个核苷酸迭代地添加到一个或多个起始读段的位置3'或5', 其中所述位置包括多数共有核苷酸;

(d) 根据(c)中构建的所述一个或多个重叠群和/或将(c)中构建的所述重叠群中的两个或更多桥接的一个或多个读段配对物对, 组装一个或多个超重叠群;

(e) 根据所述一个或多个超重叠群生成基因型似然比; 以及

(f) 根据(e)中生成的所述基因型似然比确定基因改变的存在或不存在。

128. 根据权利要求127所述的方法, 其中所述堆积关系包括所述集合中的两个或更多读段之间的多个重叠。

129. 根据权利要求128所述的方法, 其中根据以下各项选择所述多个重叠中的每一个:

(i) 所述集合中包括与所述集合的第二读段的第一重叠的第一读段, (ii) 所述第一重叠包括大于预定比对得分阈值的比对得分, (iii) 所述第二读段超过所述第一读段的3'端或5'端扩展一个或多个核苷酸, 以及(iv) 所述第一重叠包括满足(i)、(ii)和(iii)的所有可能的第一重叠中的最高比对得分。

130. 根据权利要求129所述的方法, 其中所述第二读段包括与所述集合中的第三读段的第二重叠, 其中(i) 所述第二读段包括所述第一重叠, (ii) 所述第二重叠包括大于预定比对得分阈值的比对得分, (iii) 所述第三读段超过所述第二读段的3'端或5'端扩展一个或多个核苷酸, 并且所述第二读段在相同3'端或5'段方向上扩展所述第一读段, 以及(iv) 所述第二重叠包括满足(i)、(ii)和(iii)的所有可能的第二重叠中的最高比对得分。

131. 根据权利要求128至130中任一项所述的方法, 其中根据针对所述堆积关系确定的所述多个重叠确定所述多数共有核苷酸。

132. 根据权利要求127至131中任一项所述的方法, 其中构建所述重叠群包括将至少一个核苷酸迭代地添加到一个或多个中间重叠群中的每一个的位置3'或5'。

133. 根据权利要求132所述的方法, 其中所述位置包括两个不同的多数共有核苷酸, 构建所述重叠群包括: 生成所述中间重叠群的拷贝, 从而提供两个相同的中间重叠群; 将所述两个不同的多数共有核苷酸中的一个添加到所述两个相同的中间重叠群中的每一个, 其中不同的核苷酸被添加到所述两个相同的中间重叠群中的每一个。

134. 根据权利要求132所述的方法, 其中所述位置包括三个不同的多数共有核苷酸, 构建所述重叠群包括: 生成所述中间重叠群的两个拷贝, 从而提供三个相同的中间重叠群; 将所述三个不同的多数共有核苷酸中的一个添加到所述三个相同的中间重叠群中的每一个, 其中不同的核苷酸被添加到所述三个相同的中间重叠群中的每一个。

135. 根据权利要求132所述的方法, 其中所述位置包括四个不同的多数共有核苷酸, 构建所述重叠群包括: 生成所述中间重叠群的三个拷贝, 从而提供四个相同的中间重叠群; 将所述四个不同的多数共有核苷酸中的一个添加到所述四个相同的中间重叠群中的每一个, 其中不同的核苷酸被添加到所述四个相同的中间重叠群中的每一个。

136. 根据权利要求127至135中任一项所述的方法, 其中所述一个或多个超重叠群包括跨越所述感兴趣的基因组区域的完整长度的重叠群。

137. 根据权利要求127至136中任一项所述的方法, 其中所述一个或多个超重叠群跨越所述感兴趣的基因组区域的完整长度。

138. 根据权利要求127至137中任一项所述的方法, 其中所述序列读段从样本获得, 所述样本从人类对象获得。

139. 根据权利要求127至138中任一项所述的方法, 其中所述序列读段中的每一个包括已知的定向。

140. 根据权利要求127至139中任一项所述的方法, 其中所述读段配对物对中的每一个与插入尺寸分布相关联。

141. 根据权利要求127至140中任一项所述的方法, 其中根据一个或多个映射权重确定基因型假设似然比。

142. 根据权利要求127至141中任一项所述的方法, 其中根据比对确定所述多数共有核苷酸。

143. 根据权利要求142所述的方法, 其中根据被比对的5个或更多读段之中的共有序列确定所述多数共有核苷酸。

144. 根据权利要求143所述的方法, 其中根据被比对的所述5个或更多读段中的10%或更多之中的共有序列确定所述多数共有核苷酸。

145. 根据权利要求128至144中任一项所述的方法, 其中所述第一读段包括扩展超过所述第一读段的所述5'端的至少两个重叠以及扩展超过所述第一读段的所述3'端的至少两个重叠。

146. 根据权利要求128至145中任一项所述的方法, 其中所述多个重叠包括针对多个第一读段确定的多个第一重叠。

147. 根据权利要求128至146中任一项所述的方法, 其中所述多个重叠包括针对多个第二读段确定的多个第二重叠。

148. 根据权利要求128至147中任一项所述的方法, 其中所述多个重叠中的每一个包括重叠得分。

149. 根据权利要求128至148中任一项所述的方法,其中所述多个重叠被过滤。
150. 根据权利要求128至149中任一项所述的方法,包括在(c)之前,指示所述微处理器过滤所述多个重叠。
151. 根据权利要求149或150所述的方法,其中所述过滤包括剪枝算法。
152. 根据权利要求148至151中任一项所述的方法,其中所述多个重叠中的一个或多个根据所述过滤被移除。
153. 根据权利要求129至152中任一项所述的方法,其中所述多个重叠中的一个或多个根据预定比对得分阈值被移除。
154. 根据权利要求128至153中任一项所述的方法,其中所述多个重叠不包括插入或缺失。
155. 根据权利要求128至154中任一项所述的方法,包括根据所述堆积关系生成平铺图。
156. 根据权利要求128至155中任一项所述的方法,其中根据k-mer哈希策略确定所述多个重叠中的每一个。
157. 根据权利要求127至156中任一项所述的方法,其中所述起始读段包括位于所述预先选择的感兴趣的基因组区域的最5'侧的读段。
158. 根据权利要求127至157中任一项所述的方法,其中所述起始读段包括位于所述预先选择的感兴趣的基因组区域的最3'侧的读段。
159. 根据权利要求127至158中任一项所述的方法,其中所述两个或更多重叠群由第一读段配对物对中与第一重叠群重叠的第一读段配对物以及所述第一读段配对物对中与第二重叠群重叠的第二读段配对物桥接。
160. 根据权利要求159所述的方法,其中根据多个读段配对物对,所述第一重叠群被连接到所述第二重叠群。
161. 根据权利要求127至160中任一项所述的方法,其中(d)中的所述两个或更多重叠群中的两个的所述连接包括:构建图,其中第一重叠群和第二重叠群包括顶点,并且读段配对物对中的读段配对物的一个或多个重叠包括被定向的边缘。
162. 根据权利要求160或161所述的方法,其中所述第一重叠群和所述第二重叠群由至少5个读段配对物对连接,所述至少5个读段配对物对中的每一个包括与所述第一重叠群重叠的读段配对物以及与所述第二重叠群重叠的读段配对物。
163. 根据权利要求127至162中任一项所述的方法,其中所述遗传变异包括短串联重复。
164. 根据权利要求127至163中任一项所述的方法,其中所述遗传变异包括一个或多个单核苷酸多态性。
165. 根据权利要求138至164中任一项所述的方法,其中所述人类对象是二倍体。
166. 根据权利要求127至165中任一项所述的方法,其中根据等式(1)确定所述基因型似然比:

$$\frac{P(G|\{R\})}{P(G_0|\{R\})} = \prod_{\{R\}} \frac{\sum_{\{A_G\}} \frac{1}{N_{A_G}} F_{A_G}(W(R, A_G) + \alpha)}{\sum_{\{A_{G_0}\}} \frac{1}{N_{A_{G_0}}} F_{A_{G_0}}(W(R, A_{G_0}) + \alpha)} \quad (\text{等式 1})$$

其中G是预定倍性的基因型序列,G₀是参考序列,{R}是所述读段配对物对R的集合,N_{AG}是所述基因型序列G中的等位基因AG的数量,N_{AG0}是所述参考序列G₀中等位基因AG₀的数量,以及F_{AG}是所述等位基因AG在所述基因型序列G中的分数,F_{AG0}是所述等位基因AG₀在所述参考序列G₀中的分数,W是读段对映射权重,以及α是映射概率常数。

167. 根据权利要求127至166中任一项所述的方法,其中所述遗传变异被包括在从以下各项选择的基因中:AR、ATXN1、ATXN2、ATXN7、ATXN8、ATXN10、DMPK、FXN、JPH3、CACNA1A、PPP2R2B、TBP、ATN1、ARX、PHOX2B、PABPN1、ATT、CFTR和BRCA1。

168. 根据权利要求127至167中任一项所述的方法,其中(a)、(b)、(c)和(d)不包括包含误差校正的过程。

169. 根据权利要求127至168中任一项所述的方法,其中所述序列读段包括排序的bam文件。

170. 根据权利要求127至169中任一项所述的方法,其中所述序列读段包括插入尺寸分布文件。

171. 根据权利要求127至170中任一项所述的方法,其中所述序列读段包括fastq文件和fasta文件中的一个或多个。

172. 根据权利要求127至171中任一项所述的方法,其中所述序列读段在(b)之前被修剪。

173. 根据权利要求140至172中任一项所述的方法,其中插入尺寸似然性被确定。

174. 根据权利要求173所述的方法,其中估计的插入尺寸似然性由插入尺寸分布组件确定。

175. 根据权利要求127至174中任一项所述的方法,其中所述参考基因组的所述至少一部分是染色体的一部分。

176. 根据权利要求127至175中任一项所述的方法,其中所述参考基因组的所述至少一部分是第一染色体的一部分以及第二染色体的一部分。

177. 根据权利要求127至176中任一项所述的方法,其中所述读段配对物对包括不一致的读段配对物。

178. 根据权利要求127至177中任一项所述的方法,其中所可能的重叠在(c)之前被标识。

179. 根据权利要求127至178中任一项所述的方法,其中(e)的生成所述基因型似然比包括将所述序列读段重新比对到所述一个或多个超重叠群。

180. 根据权利要求166至179中任一项所述的方法,其中所述倍性是二倍体。

181. 根据权利要求127至180中任一项所述的方法,其中针对(c)中构建的所述一个或多个重叠群中的任何一个的所述构建,所述多个读段配对物对中的每一个不被使用多于一次。

182. 根据权利要求127至181中任一项所述的方法,其中根据读段配对物对的所述两个读段配对物中的每一个的定向来组装(d)中组装的所述多个超重叠群中的两个或更多。

183. 根据权利要求130至182中任一项所述的方法,其中(d)中组装的所述多个超重叠群中的两个或更多与一个或多个插入尺寸似然性相关联。

184. 根据权利要求127至183中任一项所述的方法,其中生成所述基因型似然比包括:

根据跨越所述预定倍性的感兴趣的基因组区域的所述完整长度的所述一个或多个超重叠群来确定一个或多个可能的基因型。

185. 根据权利要求184所述的方法, 其中所述一个或多个可能的基因型中的每一个与概率相关联。

186. 根据权利要求127至185中任一项所述的方法, 其中生成所述基因型似然比包括: 根据一个或多个单体型确定一个或多个可能的基因型, 其中根据跨越所述感兴趣的基因组区域的所述完整长度的超重叠群确定每个单体型。

187. 根据权利要求127至186中任一项所述的方法, 其中针对对象确定基因改变的存在或不存在。

188. 一种包括一个或多个硬件处理器的系统, 所述一个或多个硬件处理器被配置为执行权利要求127至187所述的方法。

189. 一种包括权利要求1至61中任一项所述的非暂时性计算机可读存储介质的系统, 所述系统包括一个或多个硬件处理器, 所述一个或多个硬件处理器被配置为执行可执行程序指令, 所述可执行程序被存储在权利要求1至61中任一项所述的非暂时性计算机可读存储介质上。

测序读段的DE NOVO组装的方法、系统和过程

[0001] 相关专利申请

[0002] 本申请要求于2014年10月10日提交的题目为“METHODS, SYSTEMS AND PROCESSES OF DE NOVO ASSEMBLY OF SEQUENCING READS”的临时专利申请号62/062636的权益,其发明人为Karel Knovicka和Kevin Jacobs,并且其由代理档案号055911-0432229指定。包括所有文本、表格和附图的在先专利申请的全部内容通过引用并入本文。

技术领域

[0003] 本技术涉及核酸操作、分析和高通量测序的方法和过程。

背景技术

[0004] 生物体(例如,动物、植物、微生物、病毒)的遗传信息被编码在脱氧核糖核酸(DNA)或核糖核酸(RNA)中。遗传信息是表示核酸的一级结构的核苷酸或修饰的核苷酸序列。生物体的核酸含量(例如,DNA)通常被称为基因组。在人体中,完整的基因组通常包括位于二十四条染色体上的大约30,000个基因。大部分基因编码特定的蛋白质,其经由转录和翻译的表达后实现活细胞内的一个或多个生物化学功能。

[0005] 许多医学病症由基因组内的一个或多个遗传变异引起。一些遗传变异可能使个人容易感染、或导致诸如例如糖尿病、动脉硬化、肥胖症、各种自身免疫性疾病和癌症(例如,结肠直肠癌、乳腺癌、卵巢癌、肺癌)的多种疾病。这种遗传性疾病可以由基因组内的一个或多个核苷酸的添加、取代、插入或缺失引起。

[0006] 通过分析核酸可以标识遗传变异。基因组的核酸可以通过各种方法(包括例如涉及大规模并行测序的方法)来分析。大规模并行测序方法通常生成数千、数百万甚至数十亿的小测序读段。为了确定基因序列,每个读段通常被映射到参考基因组,同时读段集合被组装成个体基因组或其部分的序列表示。读段的映射和组装过程由一个或多个计算机(例如,硬件微处理器(即,微处理器)和存储器)执行,并由人类双手创建的指令(例如,软件指令和/或算法)集合驱动。当对象的基因组中遇到遗传变异时,这样的映射和组装过程通常失败。现有的软件和程序不正确地映射读段、不能映射读段或不能正确地组装包括遗传变异的基因组的区域。本文的方法、系统和过程提供了对当前核酸分析技术的显著进步和改进。

发明内容

[0007] 本文提供的一些方面是分析核酸文库的方法,核酸文库包括具有存储在其上的可执行程序的非暂态计算机可读存储介质,该程序被配置为指示微处理器:(a)获得包括多个读段配对物对的双端序列读段的集合,每个对包括两个读段配对物,其中每个对中的两个读段配对物的至少一个被映射到参考基因组的至少一部分,参考基因组包括预先选择的感兴趣的基因组区域,并且其中双端序列读段中的一些没有被映射到参考基因组的该至少一部分,(b)确定序列读段的集合的堆积关系,(c)根据(b)中确定的堆积关系,构建一个或多个重叠群,包括迭代地将至少一个核苷酸添加到一个或多个起始读段的位置3'或5'中,其

中位置(例如,前进位置)包括多数共有核苷酸,(d)根据桥接两个或多个重叠群的一个或多个读段配对物对,组装一个或多个超重叠群,(e)根据一个或多个超重叠群,生成基因型似然比,以及(f)根据(e)中生成的基因型似然比,确定遗传变异的存在或不存在。

[0008] 在某些方面,堆积关系包括集合中的两个或多个读段之间的多个重叠,其中多个重叠中的每一个根据以下项目选择:(i)包括第一重叠的集合的第一读段,第一重叠具有集合的第二读段,(ii)包括大于预定比对得分阈值的比对得分的第一重叠,(iii)第二读段扩展一个或多个核苷酸并经过第一读段的3'端或5'端,以及(iv)第一重叠包括满足(i)、(ii)和(iii)的、所有可能的第一重叠中的最高比对得分。在一些方面,堆积关系包括第二读段,第二读段包括第二重叠,第二重叠具有集合的第三读段,其中(i)第二读段包括第一重叠,(ii)第二重叠包括大于预定比对得分阈值的比对得分,(iii)第三读段扩展一个或多个核苷酸并经过第二读段的3'端或5'端,以及第二读段和第三读段以同样的3'或5'方向扩展第一读段,以及(iv)第二重叠包括满足(i)、(ii)和(iii)的、所有可能的第二重叠中的最高比对得分。

[0009] 在某些方面,多数共有核苷酸根据针对堆积关系确定的多个重叠来确定的。在某些实施例中,构建重叠群包括将至少一个核苷酸迭代地添加到一个或多个中间重叠群的每一个的位置3'或5'。在某些实施例中,其中位置(例如,前进位置)包括两个不同的多数共有核苷酸,构建重叠群包括:生成中间重叠群的一个拷贝,从而提供两个相同的中间重叠群,将两个不同的多数共有核苷酸中的一个添加到两个相同的中间重叠群的每一个,其中不同的核苷酸被添加到两个相同的中间重叠群中的每一个。在一些示例中,其中位置(例如,前进位置)包括三个不同的多数共有核苷酸,构建重叠群包括:生成中间重叠群的两个拷贝,从而提供三个相同的中间重叠群,将三个不同的多数共有核苷酸中的一个添加到三个相同的中间重叠群的每一个,其中不同的核苷酸被添加到三个相同的中间重叠群中的每一个。在一些示例中,其中位置(例如,前进位置)包括四个不同的多数共有核苷酸,构建重叠群包括:生成中间重叠群的三个拷贝,从而提供四个相同的中间重叠群,将四个不同的多数共有核苷酸中的一个添加到四个相同的中间重叠群的每一个,其中不同的核苷酸被添加到四个相同的中间重叠群中的每一个。

[0010] 在一些方面,从一个或多个个体对象获得样本。

[0011] 在以下的描述、示例、权利要求和附图中进一步描述了某些实施例。

附图说明

[0012] 附图图示了本技术的实施例并且是非限制性的。为了清楚和易于图示,附图不按比例绘制,并在某些情况下,各个方面被夸张或放大,以促进对特定实施例的理解。

[0013] 图1示出了系统流程图(例如,Kragle)的一个实施例。

[0014] 图2示出了重叠的一个实施例,以及具有默认最小扩展读段计数(设为1)的读段-读段连接过滤的一个示例。读段(红色)具有向右扩展它的读段A到G(绿色)。这些扩展读段中的一些也具有向右扩展它们的读段(蓝色)。红色读段将保持到向右扩展它的读段的三个连接。将保留读段A,因为它具有最高得分,但由于读段A不具有向右扩展它的任何读段,所以红色读段还保留读段B和C。这两个读段具有相同的得分(1200)并具有向右扩展它们的读段。不需要附加的读段连接;红色读段在三个连接之间具有其自身被向右扩展的至少一个

读段(读段B和C均可以被另一读段向右扩展)。

[0015] 图3示出了重叠的实施例。读段A必须与读段B(具有多态性碱基A的单体型)和读段C(具有多态性碱基C的单体型)保持连接。将保持到读段B的连接,因为其是向右扩展读段A的最高得分读段(并且如果读段B自身具有将其向右扩展的读段,那么读段A自身不需要附加连接)。然而读段A是向左扩展读段C的最高得分读段,因此读段A和读段C之间的连接由读段C强制。因此读段A将具有两个连接,两个连接均将其向右扩展,然而每一个连接到不同的单体型。

[0016] 图4示出了重叠群组装的一个实施例,示出了“所有募集”边缘和共有序列的扩展。一个碱基扩展的“所有募集”间隔内部的所有读段将在当前迭代中募集。

[0017] 图5示出了重叠群组装的一个实施例,其中多于一个的多数共有核苷酸存在用于前进位置。图5示出了重叠群分裂(复制)的表示。遇到A/C多态位置并使当前重叠群分裂成两个。来自具有碱基A的五个读段(蓝色)的三个读段(或他们的配对物)跨过了先前的分裂位置。来自具有碱基C(绿色)的三个读段对的2个跨过了先前分裂位置。没有碱基获得单倍型调整计数0,因此两个新的重叠群将从当前重叠群生成;除了红色读段外,一个重叠群将采用具有碱基A的读段对(蓝色读段),而另一个重叠群将会采用具有碱基C的读段对(绿色读段)。

[0018] 图6示出了因为具有多态碱基C的读段具有单体型调整计数0(例如,没有具有碱基C的读段对跨过先前的分裂位置)而不支持重叠群分裂的多态读段堆积。因此具有碱基C的读段对将从重叠群中去除并且不生成新的重叠群。

[0019] 图7示出了超重叠群组装的一个示例。a)中红色的重叠群遇到第一A/C多态位置并分裂出新的重叠群(蓝色)。红色重叠群遇到另一G/T多态位置并分裂出另一重叠群(绿色)。当蓝色重叠群被建立时,它遇到与红色重叠群的G/T分裂相同的G/T多态位置(相同的读段组合物和读段中相同的位置)。因此该分裂仅被标记,并且未构建蓝色重叠群的剩余部分。然而,在重叠群整合阶段期间,蓝色重叠群被附加了具有匹配分裂的红色重叠群的两个可能端,产生b)中的四个重叠群中的前两个。

[0020] 图8示出了重叠群如何被连接以形成超重叠群的一个实施例。以下重叠群连接的图产生了5个超重叠群(通过图中的、从开始节点到结束节点的5个路径)。

[0021] 图9示出了读段对被比对到错误插入附近的假设序列。错误插入由读段(蓝色)组成,这些配对物不会跨过相邻的侧翼。这些错误插入通常发生在重复区域中,在重复区域中,在来自真实重复序列的读段和源于该区域外部序列的读段中的重复序列的读段之间,完成一些错误比对。

[0022] 图10示出了包括形成堆积关系、组装重叠群、组装超重叠群和生成基因型概率的一个实施例的概况。

[0023] 图11描述了生成堆积关系(例如,读段-读段比对)的过程的一个示例。

[0024] 图12示出了过滤重叠的一个实施例。

[0025] 图13示出了读段比对图循环的一个实施例。

[0026] 图14示出了组装重叠群和/或超重叠群的一个实施例。

[0027] 图15描述了重叠群组装的一个实施例。

[0028] 图16描述了重叠群组装的另一实施例。

- [0029] 图17描述了包括在重叠群组装过程期间分裂(例如,复制)重叠群的一个实施例。
- [0030] 图18描述了完成重叠群组装的一个实施例。
- [0031] 图19示出了组装超重叠群的一个实施例。
- [0032] 图20示出了二倍体假设的一个示例。
- [0033] 图21示出了包括插入惩罚部分的基因型似然模型的一个示例。
- [0034] 图22示出了分离等位基因表示的推导的一个示例。
- [0035] 图23示出了Kragle方法的一部分的一个实施例。
- [0036] 图24示出了通过应用Kragle获得的结果的一个示例。
- [0037] 图25示出了CFTR识别系统的一个示例。
- [0038] 图26示出了具有挑战性的组装任务的一个示例。
- [0039] 图27示出了针对BRCA1基因的外显子19中的确认的杂合缺失的、两个组装的单体型的映射的一个示例。该图显示了包含缺失的单体型2的3'侧的映射。
- [0040] 图28示出了单体型2的5'侧的映射的、并组装BRCA1基因的外显子19中的实验上确认的缺失的一个示例。

具体实施方式

[0041] 下一代测序(NGS)允许通过比传统测序方法更快速和更便宜的方法对基因组范围内的核酸进行测序。本文的方法和过程提供了可以用于定位和标识遗传变异和/或相关疾病和病症的先进测序技术的改进。在一些实施例中,本文提供了部分地包括操作和分析通常由大规模并行测序方法获得的序列读段的方法。

[0042] 传统的组装器和准直器通常不能正确地组装包括遗传变异(例如,短串联重复(STR)、多态性、插入等)的基因组序列。调用诸如STR的遗传变异对于大多数准直器和映射器来说都是难题。现有的算法和软件包无法正确地映射和比对包括这种遗传变异的基因组区域内的读段。在这方面被测试并失败的组装器的示例包括Lobstr、Repeatseq和诸如GATK Haplotype Caller、AMOS de-novo组装器、Mira de-novo组装器、FERMI、SGA和其他的通用de-novo组装器。非常需要能够正确地和常规地组装包括遗传变异的基因组区域和/或精确地从测序读段的集合中标识遗传变异的新的以及改进的系统和方法(例如,依赖微处理器的方法)。在本文中描述并请求保护这样的方法、系统和过程。

[0043] 对象

[0044] 对象可以是任意生物或非生物体,包括但不限于人类、非人类的动物、植物、细菌、真菌、病毒或原生生物。对象可以是任意年龄(例如,胚胎、胎儿、婴儿、儿童、成人)。对象可以是任意性别(例如,雄性、雌性、或其组合)。对象可以是怀孕的。对象可以是患者(例如,人类患者)。

[0045] 样本

[0046] 本文提供了用于分析样本的方法和组合物。样本(例如,包括核酸的样本)可以从合适的对象获得。样本可以从对象或其部位直接分离或获得。在一些实施例中,样本从个体或医学专业人员处间接获得。样本可以从对象或其部位分离或获得的任意标本。样本可以从多个对象分离或获得的任意标本。标本的非限制性示例包括来自对象的流体或组织,包括但不限于血液或血液制品(例如,血清、血浆、血小板、血沉棕黄层等)、脐带血、绒毛

膜绒毛、羊水、脑脊髓液、脊髓液、灌洗液(例如,肺、胃、腹膜、导管、耳、关节镜)、活组织检查样本、胚胎刺激样本、细胞(血细胞、淋巴细胞、胎盘细胞、干细胞、骨髓来源的细胞、胚胎或胎儿细胞)或其部分(例如,线粒体、细胞核、提取物等)、尿液、粪便、痰液、唾液、鼻粘膜、前列腺液、灌洗液、精液、淋巴液、胆汁、眼泪、汗液、乳汁、乳液等、或其组合。从其中提取核酸的流体或组织样本可以是非细胞组成的(例如,无细胞)。组织的非限制性示例包括器官组织(例如,肝、肾、肺、胸腺、肾上腺、皮肤、膀胱、生殖器官、肠、结肠、脾、脑等)、上皮组织、头发、头发毛囊、人体内的导管、动物体内的管道、骨、眼、鼻、口、喉、耳、指甲等、其部分或其组合。样本可以包括正常的、健康的、患病的(例如,感染)和/或癌性的(例如,癌细胞)细胞或组织。从对象获得的样本可以包括多种微生物(例如,病毒核酸、胎儿核酸、细菌核酸、寄生虫核酸)的细胞或细胞材料(例如,核酸)。

[0047] 在一些实施例中,样本包括核酸或其片段。样本可以包括从一个或多个对象获得的核酸。在一些实施例中,样本包括从单个对象获得的核酸。在一些实施例中,样本包括核酸的混合物。核酸的混合物可以包括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、细胞或组织来源、对象来源等或其组合)或其组合的两个或多个核酸种类。样本可以包括合成核酸。

[0048] 核酸

[0049] 术语“核酸”指来自诸如DNA(例如,互补DNA(cDNA)、基因组DNA(gDNA)等)、RNA(例如,消息RNA(mRNA)、短抑制性RNA(siRNA)、核糖体RNA(rRNA)、tRNA、微RNA)、和/或DNA或RNA类似物(例如,含有碱基类似物、糖类类似物和/或非特异性骨架(non-native backbone)等)、RNA/DNA杂交体和聚酰胺核酸(PNA)的任意组合物的一个或多个核酸(例如,核酸的集合或子集),所有这些可以是单链或双链形式,除非另有限定,可以涵盖可以与天然存在的核苷酸类似的方式作用的已知的天然核苷酸类似物。除非特别限定,该术语涵盖包括脱氧核糖核苷酸、核糖核苷酸和天然核苷酸的已知类似物的核酸。核酸可以包括,作为等同物、衍生物或其变体、由核苷酸类似物、单链(“有义”或“反义”、“加”链或“减”链、“向前”读取框架或“反向”读取框架)和双链核苷酸合成的RNA或DNA的合适的类似物。核酸可以是单链或双链的。核酸可以是2个或更多个、3个或更多个、4个或更多个或5个或更多个连续的核苷酸的任意长度。核酸可以包括本领域已知的从特定5'到3'顺序的核苷酸的序列(例如,核酸序列,例如,序列)。

[0050] 核酸可以是天然存在的和/或可以由人合成、复制或改变。例如,核酸可以是扩增子。核酸可以来自核酸文库(例如,gDNA、cDNA或RNA文库等)。核酸可以是合成的(例如,化学合成)或生成的(例如,通过体外聚合酶扩展、例如通过扩增、例如通过PCR)。在一些实施例中,核酸可以是或可以来自质粒、噬菌体、病毒、自主复制序列(ARS)、着丝粒、人造染色体、染色体或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复制的其他核酸。核酸(例如,核酸文库)可以包含来自一个样本或两个或更多个样本(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、或20个或更多个样本)的核酸。为本文所述的过程或方法提供的核酸可包括来自1至1000个、1至500个、1至200个、1至100个、1至50个、1至20个或1至10个样本的核酸。

[0051] 术语“基因”是指产生多肽链时涉及的DNA片段,并且可以包括涉及基因产品的转录/翻译与转录/翻译的调整以及个体编码片段(外显子)之间的插入序列(内含子)的、编码区域之前和之后的区域(头部和尾部)。由于基因序列的遗传变异(例如,基因的编码和非编码部分中的突变),基因可能不一定产生肽或可能产生截短的蛋白质或非功能性蛋白质。通常可以通过与参考基因组中的基因的同源性来标识无论是功能性还是非功能性的基因。

[0052] 寡核苷酸是相对较短的核酸。寡核苷酸可以为约2至150、2至100、2至50或2至约35个核酸的长度。在一些实施例中,寡核苷酸是单链的。在某些实施例中,寡核苷酸是引物。引物通常被配置为与选定的互补核酸杂交,并且被配置为在杂交后由聚合酶扩展。

[0053] 核酸分离和纯化

[0054] 可以使用本领域已知的合适方法从一个或多个对象、一个或多个样本或一个或多个源衍生、分离、提取、纯化或部分纯化核酸。可以使用任何合适的方法来分离、提取和/或纯化核酸。

[0055] 本文使用的术语“分离”是指从其原始环境(例如,天然存在的天然环境、或外源性表达的宿主细胞)中移除的核酸,因此是通过人的干预(例如,“由人类双手”)从其原始环境改变。本文使用的术语“分离的核酸”可以指从对象(例如,人类对象)移除的核酸。与源样本中存在的核酸分子数目相比,分离的核酸由更少核酸分子(例如,蛋白质、脂质、小化合物、碳水化合物、污染物、颗粒、聚集体、盐、洗涤剂)提供。包含分离的核酸的组合物可以是约50%至大于99%不包含非核酸分子。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%不包含非核酸分子。本文使用的术语“纯化”可以指:与在使核酸进行纯化程序之前存在的非核酸分子的量相比,提供含有较少非核酸分子的核酸。包括纯化的核酸的组合物可以是至少约60%、70%、80%、81%、82%、83%、84%、85%、86%、87%、88%、89%、90%、91%、92%、93%、94%、95%、96%、97%、98%、99%或大于99%不包含其他非核酸分子。包括纯化的核酸的组合物可以是至少约60%、70%、80%、81%、82%、83%、84%、85%、86%、87%、88%、89%、90%、91%、92%、93%、94%、95%、96%、97%、98%、99%或大于99%不包含其他核酸。包括纯化的核酸的组合物可以包括在应用纯化方法之前的样本中存在的总核酸的至少80%、81%、82%、83%、84%、85%、86%、87%、88%、89%、90%、91%、92%、93%、94%、95%、96%、97%、98%、99%或大于99%。

[0056] 核酸测序

[0057] 在某些实施例中,通过包括核酸测序的过程分析核酸(例如,扩增子、文库的核酸、捕获的核酸)。在一些实施例中,核酸可以被测序。在一些实施例中,获得完整的或基本上完整的序列,并且有时获得局部序列。

[0058] 可以使用对核酸测序的合适的方法,方法的非限制性示例包括Maxim&Gilbert、链终止法、合成测序、连接法测序、质谱测序、基于显微镜的技术等或其组合。在一些实施例中,第一代技术(例如诸如Sanger测序方法,其包括自动Sanger测序方法、包括微流体Sanger测序)可以用于本文提供的方法。在一些实施例中,可以使用包括使用核酸成像技术(例如,透射电子显微镜(TEM)和原子力显微镜(AFM))的测序技术。在一些实施例中,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内、以大规模并行方式测序的克隆地扩增的DNA模板或单个DNA分子。能够以大规模并行方式对DNA测序的下一代(例如,第2代

和第3代)测序技术可用于本文所述的方法,并且在本文中统称为“大规模并行测序”(MPS)。用于进行本文描述的方法的任何合适的MPS或下一代测序方法、系统或技术平台可用于获得测序读段,其非限制性示例包括Illumina/Solex/HiSeq(例如,Illumina的基因组分析仪、基因组分析仪II、HISEQ 2000;HISEQ 2500、SOLiD、Roche/454、PACBIO、SMRT、Helicos True Single Molecule Sequencing、Ion Torrent和Ion基于半导体的测序、WildFire、基于5500、555x1 W和/或5500x1 W遗传分析仪(例如,由Life Technologies开发和销售的)的技术、Polony测序、焦磷酸测序、大规模并行签名测序、RNA聚合酶(RNAP)测序、IBS方法、LaserGen系统和方法、化学敏感场效应晶体管(CHEMFET)阵列、基于电子显微镜的测序、纳米球测序、合成测序、连接法测序、杂交测序等或其变体。本文中也考虑包括使用发展中的核酸成像技术(例如,透射电子显微镜(TEM)和原子力显微镜(AFM))的附加测序技术。在一些实施例中,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内、以大规模方式测序的克隆扩增的DNA模板或单个DNA分子。在一些实施例中,MPS测序方法利用靶标方法,其中序列读段从特定染色体、基因或感兴趣区域生成。特定染色体、基因或感兴趣区域在本文中有时被称为靶标基因组区域。在某些实施例中,使用非靶标方法,其中样本中大部分或全部核酸片段被随机测序、扩增和/或捕获。

[0059] 序列读段

[0060] 使核酸通过测序方法通常提供序列读段。如本文中所用,“读段”(例如,“读段”,“序列读段”)是通过本文所述或本领域已知的任何测序过程产生的短核苷酸序列。读段可以从核酸片段的一端生成(“单端读段”),有时从核酸片段的两端生成(例如,双端读段、双端序列读段、两端读段)。双端读段通常包括一个或多个读段对(例如,两个读段、读段配对物对),每个读段对从被测序的核酸片段的每个端获得。读段配对物对的每个读段在本文中有时被称为读段配对物。双端测序方法(例如,其中一个或多个核酸文库被测序)经常产生多个读段配对物对和多个读段配对物。

[0061] 序列读段的长度通常与特定的测序技术相关联。例如,高通量方法和/或下一代序列提供可以从数十到数百个碱基对(bp)的大小变化的序列读段。在一些实施例中,序列读段是约15bp至约900bp长的平均数、中位数、平均值或绝对长度。在某些实施例中,序列读段是约1000bp或更大的平均数、中位数、平均值或绝对长度。

[0062] 单端读段可以是任何合适的长度。在一些实施例中,单端读段的标称值、平均值、平均数或绝对长度有时为约10个核苷酸至约1000个连续核苷酸、约10个核苷酸至约500个连续核苷酸、约10个核苷酸至约250个连续核苷酸、约10个核苷酸至约200个连续核苷酸、约10个核苷酸至约150个连续核苷酸、约15个连续核苷酸至约100个连续核苷酸、约20个连续核苷酸至约75个连续核苷酸、或约30个连续核苷酸或约50个连续核苷酸。在某些实施例中,单端读段的标称值、平均数、平均值或绝对长度为约5个、6个、7个、8个、9个、10个、11个、12个、13个、14个、15个、16个、17个、18个、19个、20个、21个、22个、23个、24个、25个、26个、27个、28个、29个、30个、31个、32个、33个、34个、35个、36个、37个、38个、39个、40个、41个、42个、43个、44个、45个、46个、47个、48个、49个或50个或更多个核苷酸长度。

[0063] 双端读段(例如,读段配对物)可以是任何合适的长度。在某些实施例中,核酸片段的两端以合适的读段长度进行测序,该读段长度足以将每个读段(例如,片段模板的两端的读段)映射到参考基因组。在某些实施例中,双端读段的标称值、平均数、平均值或绝对长度

为约10个连续核苷酸至约500个连续核苷酸、约10个连续核苷酸至约400个连续核苷酸、约10个连续核苷酸至约300个连续核苷酸、约50个连续核苷酸至约200个连续核苷酸、约100个连续核苷酸至约200个连续核苷酸、或约100个连续核苷酸至约150个连续核苷酸。在某些实施例中，双端读段的标称值、平均数、平均值或绝对长度为约125个、126个、127个、128个、129个、130个、131个、132个、133个、134个、135个、136个、137个、138个、139个、140个、141个、142个、143个、144个、145个、146个、147个、148个、149个、150个、151个、152个、153个、154个、155个、156个、157个、158个、159个、160个、165个、166个、167个、168个、169个、170个或更多个核苷酸。

[0064] 读段通常是物理核酸中核苷酸序列的表示。例如，在含有序列的ATGC描绘读段中，在物理核酸中，“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。相对短的读段的混合物可以通过本文所述的过程转化成对象中存在的基因组核酸的表示。例如，相对较短的读段的混合物可以被转化为复制数变异（例如，复制数变异）、遗传变异或非整倍体的表示。来自多个对象的核酸混合物的读段可以转化为多个对象中的每一个的基因组或其部分的表示。在某些实施例中，“获得”来自对象的样本的核酸序列读段和/或“获得”来自一个或多个参考者的生物标本的核酸序列读段可以涉及对核酸直接测序以获得序列信息。在一些实施例中，“获得”可以涉及接收由他人从核酸直接获得的序列信息。

[0065] 映射读段

[0066] 序列读段可以被映射。在一些实施例中，可以使用合适的映射方法、过程或算法。在某些实施例中，在本文中使用修改的映射方法和过程。下面描述了映射过程的某些方面。

[0067] 映射核苷酸序列读段（例如，来自其物理基因组位置未知的片段的序列信息）可以以多种方式进行，并且通常包括将获得的序列读段或其部分与参考基因组中的匹配序列比对。在这种比对中，序列读段通常比对到参考序列，比对的序列读段被命名为“映射的”、“映射的序列读段”或“映射的读段”。

[0068] 如本文中所用，术语“比对的”、“比对”或“比对中”是指可以被标识为匹配（例如，100%相同）或部分匹配的两个或多个核酸序列。比对可以手动地或通过计算机（例如，软件、程序、计算机程序组件或算法）进行，其非限制性示例包括作为Illumina基因组分析管道的部分分布的核苷酸数据（ELAND）高效局部比对计算机程序。序列读段的比对可以是100%的序列匹配。在一些情况下，比对小于100%的序列匹配（例如，非完美匹配、部分匹配、部分比对）。在某些实施例中，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%的匹配。在一些实施例中，比对包括错配。在一些实施例中，比对包括1个、2个、3个、4个、5个或更多个错配。可以使用两个链中任意一个来比对两个或多个序列。在某些实施例中，核酸序列与另一个核酸序列的反向互补序列比对。

[0069] 可以使用各种计算方法将序列读段映射和/或比对到参考基因组。序列读段可以由映射组件或包括映射组件（例如，合适的映射和/或比对程序）的机器或计算机映射，该映射组件通常将读段映射到参考基因组或其片段。通常通过使用合适的映射和/或比对程序来将序列读段和/或双端读段映射到参考基因组，程序的非限制性示例包括BWA (Li H. and Durbin R. (2009) *Bioinformatics* 25,1754-60)、Novoalign [Novocraft (2010)]、Bowtie

(Langmead B, et al., (2009) Genome Biol. 10:R25)、SOAP2 (Li R, et al., (2009) Bioinformatics 25,1966-67)、BFAST (Homer N, et al., (2009) PLoS ONE 4,e7767)、GASSST (Rizk, G. 和 Lavenier, D. (2010) Bioinformatics 26,2534-2540)、和 MPscan (Rivals E., et al. (2009) Lecture Notes in Computer Science 5724,246-260) 等。可以使用合适的短读段比对程序来映射和/或比对序列读段和/或双端读段。短读段比对程序的非限制性示例是 BarraCUDA、BFAST、BLASTN、BLAST、BLAT、BLITZ、Bowtie (例如, BOWTIE 1、BOWTIE 2)、BWA、CASHX、CUDA-EC、CUSHAW、CUSHAW2、drFAST、FASTA、ELAND、ERNE、GNUMAP、GEM、GensearchNGS、GMAP、Geneious Assembler、iSAAC、LAST、MAQ、mrFAST、mrsFAST、MOSAIC、MPscan、Novoalign、NovoalignCS、Novocraft、NextGENe、Omixon、PALMapper、Partek、PASS、PerM、PROBEMATCH、QPalma、RazerS、REAL、cREAL、RMAP、rNA、RTG、Segemehl、SeqMap、Shrec、SHRIMP、SLIDER、SOAP、SOAP2、SOAP3、SOCS、SSAHA、SSAHA2、Stampy、STORM、Subread、Subjunc、Taipan、UGENE、VelociMapper、TimeLogic、XpressAlign、ZOOM 等、其变体或其组合。映射组件可以通过本领域公知或本文描述的合适方法对测序读段进行映射。在某些实施例中,需要映射组件或包括映射组件的机器或计算机来提供映射的序列读段。映射组件通常包括合适的映射和/或比对程序或算法。

[0070] 在某些实施例中,与序列读段相关联的一个或多个序列读段和/或信息以合适的计算机可读格式存储在非暂态计算机可读存储介质上和/或从非暂态计算机可读存储介质访问。存储在非暂态计算机可读存储介质上的信息有时被称为文件或数据文件。读段(例如,单独读段、双端读段、读段配对物、读段配对物对)、所选择的读段、读段的集合或子集和/或与一个或多个读段相关联的信息通常存储在文件或数据文件中。文件通常包括格式。例如,序列读段有时以包括关于一个或多个序列读段的信息的格式存储,这种信息的非限制性示例包括完整的或部分的核酸序列、可映射性、可映射性得分,映射位置、与其他映射或未映射读段的相对位置或距离(例如,读段配对物之间的估计距离)、相对于参考基因组或其他读段(例如,相对于读段配对物)的定向、读段配对物的估计或精确位置、G/C 内容等或其组合。“计算机可读格式”通常在本文中被称作格式。在一些实施例中,以合适的二进制格式、文本格式等或其组合存储和/或访问序列读段。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。二进制和/或文本格式的非限制性示例包括 BAM、排序的 BAM、SAM、SRF、FASTA、FASTQ、Gzip 等或其组合。

[0071] 在某些实施例中,本文的程序被配置为指示微处理器获得或获取一个或多个文件(例如,排序的 bam 文件)。在一些实施例中,本文的程序被配置为指示微处理器获得或获取一个或多个 FASTQ 文件(例如,针对第一读段和第二读段的 FASTQ 文件)和/或一个或多个参考文件(例如,FASTA 或 FASTQ 文件)。在一些实施例中,本文中的程序指示微处理器调用计算机程序组件和/或向一个或多个计算机程序组件(例如,接合体(adapter)修剪器组件、BWA-MEM 准直器、插入尺寸分布组件、samtools 等)传递数据和/或信息(例如,文件)或传递来自一个或多个计算机程序组件的数据和/或信息(例如,文件)。在一些实施例中,程序指示处理器调用计算机程序组件,该计算机程序组件创建用于输入到另一处理步骤的新文件和格式(参见示例 1 和图 1)。在一些实施例中,第一格式的序列读段被压缩成第二格式,第二格式需要比第一格式少的存储空间。本文使用的术语“压缩”是指数据压缩、源编码和/或比特率降低的过程,其中计算机可读数据文件的大小减小。压缩组件的非限制性示例包括 GZIP 和

BGZF等或其变体。

[0072] 在一些实施例中,读段可以唯一地或非唯一地映射到参考基因组。如果读段与参考基因组中的单个序列比对,则将其视为“唯一映射”。如果读段与参考基因组中的两个或多个序列比对,则将其视为“非唯一映射”。在一些实施例中,从进一步分析(例如,量化)消除非唯一映射的读段。在某些实施例中,可以允许一定程度的错配(0-1)来考虑可能存在于参考基因组和来自被映射的个体样本的读段之间的单核苷酸多态性。在一些实施例中,被映射到参考序列的读段不允许任何程度的错配。

[0073] 如本文中所用,术语“参考基因组”可以指可用于参考来自对象的经标识的序列的、任何生物或病毒的、任何特定已知的、经测序或表征(无论是部分地或完全地)的基因组。参考基因组有时指参考基因组的片段(例如,染色体或其部分,例如参考基因组的一个或多个部分)。人类基因组、人类基因组组装和/或来自任何其他生物体的基因组可用作参考基因组。在国家生物技术信息中心www.ncbi.nlm.nih.gov可以找到一个或多个人类基因组、人类基因组组装以及其他生物的基因组。“基因组”是指以核酸序列表达的、生物或病毒的完整遗传信息。如本文中所用,参考序列或参考基因组通常是来自个体或多个个体的组装的或部分组装的基因组序列。在一些实施例中,参考基因组是来自一个或多个人类个体的组装的或部分组装的基因组序列。在一些实施例中,参考基因组包括分配给染色体的序列。本文所用的术语“参考序列”是指一个或多个参考样本的一个或多个多核苷酸序列。在一些实施例中,参考序列包括从参考样本获得的序列读段。在一些实施例中,参考序列包括序列读段、读段的组装和/或共有DNA序列(例如,序列重叠群)。在一些实施例中,从基本上不包含遗传变异(例如,所讨论的遗传变异)的参考对象获得参考样本。在一些实施例中,从包括已知的遗传变异的参考对象获得参考样本。本文所用的术语“参考”可以指参考基因组、参考序列、参考样本和/或参考对象。在一些实施例中,可以在本领域已知的核酸数据库中找到序列读段和/或将序列读段与本领域已知的核酸数据库中的序列比对,本领域已知的核酸数据库包括例如GenBank、dbEST、dbSTS、EMBL(欧洲分子生物学实验室)和DDBJ(日本DNA数据库)。BLAST或类似工具可用于根据序列数据库搜索所标识的序列。

[0074] 在一些实施例中,针对基因组区域(例如,部分、基因组部分)评估可映射性。可映射性是将核苷酸序列读段明确地比对到参考基因组的一部分的能力,通常存在多达指定数量的错配,包括例如0个、1个、2个或多个错配。在一些实施例中,可映射性被提供作为得分或值,其中得分或值由合适的映射算法或计算机映射软件生成。与包括唯一核苷酸序列的延伸的基因组区域比对的高质量序列读段有时具有高的可映射性值。

[0075] 根据参考基因组,双端读段有时被映射到相同多核苷酸片段的相对端。在一些实施例中,仅读段配对物对的一个读段被映射到参考基因组。在一些实施例中,读段配对物对的读段配对物被独立映射。在一些实施例中,在映射过程中考虑来自读段配对物对的所有读段配对物的信息(例如,定向、估计插入尺寸、读段之间的估计距离)。参考基因组通常用于确定和/或推断位于读段配对物对之间的核酸序列。位于两个双端读段之间的核酸在本文中通常被称为插入。在一些实施例中,通过将读段配对物对的所有读段配对物映射到参考序列来确定或估计插入尺寸。在一些实施例中,根据分布来估计或确定插入尺寸(例如,长度)。在某些实施例中,从插入尺寸分布确定包括可行插入的插入尺寸的概率。在一些实施例中,由合适的分布和/或合适的分布函数确定插入尺寸。在一些实施例中,由通常包括

分布函数的插入尺寸分布组件来确定插入尺寸或估计的插入尺寸。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)、核密度函数 (核密度估计)、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何合适的分布或其组合。有时由平均的、归一化的和/或加权的插入长度生成插入尺寸。有时根据估计的和/或已知的从被测序的核酸文库的片段衍生的核酸片段长度来估计插入尺寸分布。在一些实施例中,合适的存储介质包括存储的估计插入长度、插入长度分布等。在某些实施例中,序列读段包括插入尺寸分布、估计插入长度、读段配对物之间的估计距离等,或其组合。

[0076] 读段募集

[0077] 在一些实施例中,本文的方法、过程或系统包括读段募集过程。读段募集过程通常由读段募集组件进行。在某些实施例中,读段募集过程包括获得和/或选择如本文所述的序列读段。在一些实施例中,读段募集过程包括从多个读段中获取和/或选择读段子集的方法。

[0078] 在一些实施例中,读段配对物对 (例如,从双端测序方法获得) 的一个读取配对物映射到参考基因组,并且读段配对物对的另一个读段配对物被错误地映射到参考基因组、不能映射到参考基因组或包括低的可映射性得分。这样的读段配对物对有时被称为不一致的读段配对物对。在一些实施例中,不一致的读段配对物对包括映射到感兴趣的参考基因组的区域 (例如,感兴趣的基因组区域) 的一个读段配对物,而另一个读段配对物不能映射到感兴趣的参考基因组的一部分。在一些实施例中,不一致的读段配对物对包括映射到感兴趣的参考基因组的一部分 (例如,感兴趣的基因组区域的一部分) 的第一读段配对物和映射到参考基因组的不期望位置的第二读段配对物。参考基因组的不期望位置的非限制性示例包括 (i) 与第一读段映射的染色体不同的染色体, (ii) 与第一读段配对物分离超过预定距离的基因组位置,预定距离的非限制性示例包括从估计的插入尺寸预测的距离;超过 300bp、超过 500bp、超过 1000bp、超过 5000bp、或超过 10000bp 的距离,以及 (iii) 与第一读段不一致的定向 (例如,相反定向) 等或其组合。在一些实施例中,不一致的读段配对物对包括映射到参考基因组或其部分的第一片段的第一读段配对物,以及不可映射的和/或包括低可映射性 (例如,低可映射性得分) 的第二读段配对物。在一些实施例中,不一致的读段配对物对包括映射到参考基因组或其部分的第一片段的第一读段配对物,以及第二读段配对物,其中第二读段配对物或其部分的可映射性未被确定。可以通过合适的不一致读段标识组件或包括不一致读段标识组件的机器标识不一致的读段配对物对,该不一致读段标识组件通常标识不一致的读段配对物对。不一致读段标识组件的非限制性示例包括 SVDetect、Lumpy、BreakDancer、BreakDancerMax、CREST、DELLY 等或其组合。在一些实施例中,不一致读段配对物对不被算法或组件标识。在某些实施例中,通过标识双端读段配对物的算法来标识不一致读段对,其中读段配对物对的一个读段配对物映射到参考基因组,并且读段配对物对的另一个读段配对物被错误地映射到参考基因组、不能映射到参考基因组或包括低可映射性得分。

[0079] 在一些实施例中,读段募集过程 (例如,从多个读段中) 选择和/或获得映射到感兴趣的基因组区域中的参考基因组的所有双端读段。在一些实施例中,获得和/或使用感兴趣的基因组区域中的所有双端读段用于本文中的分析,其中每个读段配对物对中的至少一个完全或部分地映射到参考基因组。在一些实施例中,获得和/或使用所有双端读段用于本文

中的分析,其中每个读段配对物对中的至少一个完全或部分地映射到感兴趣的基因组区域的参考基因组。在一些实施例中,获得和/或使用所有不一致的读段配对物对用于本文中的分析,其中每个不一致的读段配对物对的读段中的至少一个映射到感兴趣的基因组区域中的参考基因组。

[0080] 在一些实施例中,本文的方法或系统包括获得包含多个读段配对物对的双端序列读段的集合。在一些实施例中,本文的方法或系统包括获得由多个读段配对物对组成的双端序列读段的集合。在某些实施例中,从双端测序方法获得读段配对物对的测序读段的每个对。在某些实施例中,读段配对物对的测序读段的每个对由两个读段配对物组成。读段配对物通常是测序读段。在一些实施例中,本文的方法或系统包括获得包含多个读段配对物对的双端序列读段的集合,其中每个对的读段配对物中的至少一个或其部分被映射到参考基因组的至少一部分,参考基因组包括预先选择的感兴趣的基因组区域,并且其中双端序列读段中的一些未被映射到包括预先选择的感兴趣的基因组区域的参考基因组的至少一部分。

[0081] 在一些实施例中,本文中的方法和系统通过募集从双端序列读段获得的读段配对物对的所有读段配对物来绕过包括扩充的STR、序列结点和大的复杂变异的区域中的读段映射问题,其中读段配对物对的第一读段映射到感兴趣的基因组区域,而不管读段配对物对的第二读段的可映射性。在某些实施例中,本文中的方法和系统利用映射的读段配对物的位置、读段配对物对的所有读段配对物的定向和/或读段配对物之间的估计距离(例如,估计的插入尺寸)来组装从可能包括遗传变异的对象获得的基因组核酸的区域。

[0082] 在某些实施例中,本文中的方法和系统使用读段被映射到的、一个感兴趣的基因组区域。在某些实施例中,本文的方法和系统使用两个感兴趣的基因组区域(其可能已经使用分裂读段信号或不一致的配对物信号来标识)以募集和/或获取位于遗传变异处或接近遗传变异处的读段,遗传变异包括转运和/或结点。在一些实施例中,感兴趣的基因组区域被预先选择(例如,在获得读段之前、在募集读段之前、在分析、映射和/或组装读段之前)。感兴趣的基因组区域可以是基因组的任何合适的部分。感兴趣的基因组区域可以包括一个或多个染色体、基因、外显子、内含子、非编译区(例如,调节区、启动子/增强子区)、甲基化区、非甲基化区或其部分。在一些实施例中,感兴趣的基因组区域包括怀疑具有遗传变异的区域或可能含有已知遗传变异(例如,先前在另一对象或亚群体中标识的遗传变异)的区域。在一些实施例中,感兴趣的基因组区域包括遗传变异。在一些实施例中,感兴趣的基因组区域不包括遗传变异。

[0083] 序列读段(例如,读段配对物)通常包括已知的定向。例如,存储介质通常包括包含读段配对物的已知定向的文件。在一些实施例中,读段配对物的定向和/或估计的插入尺寸被用于确定在堆积、重叠群和/或超重叠群内的映射的、未映射的、不良映射或不一致读段配对物的位置。

[0084] 在一些实施例中,序列读段被修剪。在某些实施例中,修剪是指从序列读段中标识和/或移除合成和/或异源核酸或核酸的部分,该合成和/或异源核酸用于构建文库和/或用于测序方法。异源核酸通常是异源的或对象基因组外的。通常被修剪的合成和/或异源核酸的非限制性示例包括接合体、质粒、载体、引物结合位点、索引标签(例如,核酸条形码序列)、核酸捕获序列等或其组合。在一些实施例中,修剪包括指示处理器删除和/或忽略测序

读段的合成和/或异源的那些部分。合成核酸、异源核酸和/或修剪的核酸通常不包括在本文的方法或过程中。在一些实施例中,在获得双端序列读段的集合之前或期间修剪序列读段。在一些实施例中,在确定堆积关系、过滤、构建一个或多个重叠群、组装一个或多个超重叠群和/或生成基因型似然比之前或期间修剪序列读段。在某些实施例中,修剪由修剪器组件执行。

[0085] 堆积关系

[0086] 在一些实施例中,本文的方法或过程包括确定用于序列读段的集合或子集的堆积关系。在一些实施例中,堆积关系包括集合中的多个读段之间的一个或多个重叠(例如,多个重叠),其中一些读段映射到感兴趣的参考基因组的区域。在一些实施例中,堆积关系包括构建平铺图。在一些实施例中,堆积关系包括双端序列读段的集合中的所有读段。在一些实施例中,堆积关系包括双端序列读段的集合中的选择的读段。在一些实施例中,重叠包括两个或多个读段的比对。在一些实施例中,重叠包括比对得分。在某些实施例中,根据k-mer 哈希策略来确定重叠。

[0087] 在一些实施例中,堆积关系包括多个重叠。在某些实施例中,堆积关系包括被选择和/或存储(例如,存储到存储器中)的一个或多个重叠。有时确定堆积关系包括确定和/或评估多个读段之间的所有可能的重叠。在某些实施例中,仅选择和/或存储所有可能的重叠中的一些重叠。在某些实施例中,所选择的所有重叠被存储并用于堆积关系。

[0088] 用于堆积关系的重叠通常符合一个或多个标准。例如,在一些实施例中,第一标准包括高于比对阈值得分的、第一读段和第二读段之间的重叠。在一些实施例中,在集合的第一读段包括与集合的第二读段的重叠并且重叠包括高于预定比对得分阈值或截点的比对得分的情况下,选择和/或存储重叠。在一些实施例中,在集合的第一读段包括与集合的一个、两个、三个或多个其他读段的重叠并且重叠包括高于预定比对得分阈值的比对得分的情况下,选择和/或存储重叠。在一些实施例中,在集合的第一读段包括与集合的一个、两个、三个或多个其他读段的重叠、每个重叠包括高于预定比对得分阈值的比对得分、并且每个重叠包括相同的比对得分的情况下选择和/或存储重叠。可以通过任何合适的方法或算法来确定比对得分,合适的方法或算法的非限制性示例包括Smith和Waterman的方法(Smith TF, Waterman MS., (1981) J. Theor. Biol. 91 (2): 379-80; 和Smith TF, Waterman MS., (1981) J. Mol. Biol. 147 (1): 195-7) 以及Needleman的方法(Needleman, S.B. and Wunsch, C.D. (1970) J. Mol. Biol. 48 (3): 443-53)。例如,在一些实施例中,使用具有比对得分截点为500的Smith-Waterman算法,其中匹配得分为10,并且错配惩罚为-500。在某些实施例中,在读段-读段比对中禁止和/或排除插入和缺失(in/del)。在一些实施例中,用于引发或扩展插入或缺失的惩罚被设置为足够高,以排除全部或大部分in/del。在一些实施例中,不允许间隙。在一些实施例中,在读段-读段的比对中可以允许或包括一些in/del。

[0089] 在某些实施例中,第二标准需要作为所有可能重叠中的最高比对得分的重叠。在一些实施例中,被选择和/或存储的重叠包括在读段集合中的第一读段和任何其他读段之间的所有可能的重叠(例如,所有可能的比对)的最高比对得分。有时,被选择和/或存储的重叠包括在第一读段和多个其他读段之间确定的多个重叠(例如,多个比对)的最高比对得分。

[0090] 在一些实施例中,重叠超过读段的3'端或5'端扩展一个或多个核苷酸。在一些实

施例中,第三标准需要重叠超过第一读段的5'或3'端扩展第一读段。在某些实施例中,第一读段包括在5'或3'方向上扩展第一读段并超过第一读段的端部的重叠。第一读段和第二读段之间的、扩展第一读段的重叠通常包括超过第一读段的3'端或5'端扩展第二读段的一个或多个核苷酸。有时,当第一读段和第二读段重叠、并且重叠超过第一读段的3'或5'端扩展第一读段时,重叠被选择或存储。在一些实施例中,重叠超过读段的3'端或5'端扩展至少1个、2个、3个、4个、5个、6个、7个、8个、9个、10个、15个、20个、50个、100个或至少150个核苷酸。在某些实施例中,第一读段包括与在3'方向上扩展第一读段的第二读段的第一重叠,并且第一读段包括与在5'方向上扩展第一读段的第三读段的第二重叠。在某些实施例中,堆积关系包括第一读段和第二读段之间的重叠以及第一读段和第三读段之间的重叠,其中重叠在3'和5'方向上扩展第一读段。

[0091] 在一些实施例中,堆积关系包括针对第一读段、第二读段和例如第三读段的附加选择的重叠。例如,第一读段通常包括与第二读段的第一重叠,其中当第二读段包括与扩展第二读段的第三读段的重叠时,第一重叠被选择和/或存储。在上述示例中,重叠将在相同的3'或5'方向上扩展第一读段和第二读段。此外,第三读段可以与或不与第一读段重叠。在一些实施例中,第一读段包括具有多个读段的多个重叠,多个重叠在5'和/或3'方向上扩展第一读段,其中每个重叠符合标准中的一个或多个。在一些实施例中,第一读段包括扩展超过第一读段的5'端的至少两个重叠以及扩展超过第一读段的3'端的至少两个重叠。

[0092] 在一些实施例中,堆积关系包括针对多个读段的多个被选择和/或存储的重叠,其中可以从多个重叠中选择每个重叠。在一些实施例中,堆积关系包括针对读段的集合的被选择和/或存储的多个重叠,其中每个重叠满足以下条件:(i)重叠必须包括第一读段和第二读段之间的比对,其中比对得分高于预定的比对得分阈值,(ii)第一读段和第二读段之间的重叠必须超过第一读段的3'端或5'端扩展第一读段,并且(iii)第一读段和第二读段之间的重叠包括满足(i)和(ii)的、在第一读段和读段集合中的任何其他读段之间的所有可能重叠的最高比对得分。在某些实施例中,除了上述(i)、(ii)和(iii)之外,第二读段包括(iv)高于预定比对得分阈值的重叠,(v)在相同的3'或5'方向上扩展第一读段和第二读段并且超过第二读段的端部的重叠,(vi)是第二读段和满足上述(iv)和(v)的任何其他读段之间的最高比对得分的重叠。在某些实施例中,一种方法或过程包括确定堆积关系,确定堆积关系包括选择和/或存储其中每个重叠都满足上述(i)、(ii)和(iii)的集合中的多个读段的重叠。在某些实施例中,集合的每个读段包括在5'和/或3'方向上扩展读段的1个、2个、3个、4个、5个、6个、7个、8个、9个、10个、11个、12个、13个、14个或15个或更多个重叠。堆积关系通常包括多个读段,每个读段包括多个重叠。

[0093] 在一些实施例中,重叠包括得分或索引。例如,在某些实施例中,针对读段集合的所有可能的重叠被确定,并且有时每个重叠与得分或值相关联。与重叠相关联的得分或值(例如,点值)有时是从以上的条件(i)、(ii)、(iii)、(iv)和/或(v)中的一个或多个确定的和、平均数或平均值。在一些实施例中,重叠与比对得分相关联。在某些实施例中,重叠被过滤。被过滤的重叠通常从堆积关系中被移除或删除。对于重叠群或超重叠群的de novo组装,通常不考虑被删除或过滤的重叠。在一些实施例中,根据得分或预定的截点得分过滤重叠。在一些实施例中,根据预定的比对得分阈值过滤重叠。在一些实施例中,不满足(i)、(ii)、(iii)、(iv)和(v)中的一些或全部的要求的重叠被过滤。过滤算法是已知的,并且可

以修改任何合适的过滤器以过滤堆积关系的重叠。在一些实施例中,过滤器包括剪枝算法,其遍历集合中的所有读段,并维持(例如,根据(i)、(ii)、(iii)、(iv)和/或(v))被选择和/或存储的、针对每个读段的重叠列表。在某些实施例中,程序指示微处理器针对读段的集合过滤多个重叠。

[0094] 在某些实施例中,确定堆积关系不包括包含纠错的过程。在一些实施例中,堆积关系不包括包含插入或缺失的重叠。在某些实施例中,堆积关系包括包含一个或多个错配的重叠。

[0095] 重叠群

[0096] 在一些实施例中,针对读段的集合,组装和/或构建一个或多个重叠群。在某些实施例中,根据针对读段的集合被选择/或存储的多个重叠构建一个或多个重叠群。在某些实施例中,根据包括针对读段集合的多个重叠的堆积关系构建一个或多个重叠群。在某些实施方案中,从一个或多个起始读段构建重叠群。在某些实施例中,从1个、2个、3个、4个、5个、6个、7个、8个、9个或10个或更多个起始读段构建一个或多个重叠群。起始读段可以是集合中的任何合适的读段。有时起始读段包括读段集合的最5'方向的读段和/或最3'方向的读段。最5'方向的读段通常是被映射到感兴趣的基因组区域的最5'方向的区域的读段,其中集合的一些或所有序列读段被映射到该感兴趣的基因组区域。同样,最3'方向的读段通常是被映射到感兴趣的基因组区域的最3'方向的区域的读段,其中集合的一些或所有序列读段被映射到该感兴趣的基因组区域。在某些实施例中,从不是集合中最3'方向或最5'方向的读段的起始读段组装重叠群。

[0097] 在一些实施例中,从起始读段组装重叠群,并且该过程包括将至少一个核苷酸迭代地添加到起始读段的位置3'或5'。起始读段的位置3'或5'可以是起始读段的任何合适的核苷酸的位置3'或5'。在某些实施例中,起始读段的位置3'或5'是起始读段的端部(例如,3'端或5'端)的位置3'或5'。在某些实施例中,起始读段的位置3'或5'是起始读段的中值或中间核苷酸的位置3'或5'。通常,向起始读段的3'或5'位置迭代地添加至少一个核苷酸的过程包括:首先选择起始读段内的合适位置(例如,位于合适位置的核苷酸);根据堆积关系确定针对所选择的位置的多数共有核苷酸(例如,参见下文确定多数共有核苷酸);以及将一个或多个核苷酸迭代地添加到根据堆积关系确定的多数共有核苷酸的3'和/或5'位置,从而启动了重叠群的组装。在某些实施例中,起始读段是开始重叠群组装过程的第一读段,并且募集的读段的堆积关系确定了起始读段的每个核苷酸位置的多数共有核苷酸。例如,在某些实施例中,通过与用于组装重叠群或超重叠群的过程类似的过程重新组装起始读段。

[0098] 在某些实施例中,从起始读段组装重叠群,并且该方法包括将至少一个核苷酸迭代地添加中间重叠群的位置3'或5'。在某些实施例中,中间重叠群包括起始读段(例如,起始读取的至少一些核苷酸)和添加到起始读段的3'和/或5'侧的一个或多个核苷酸。在某些实施例中,中间重叠群包括起始读段的部分或全部核苷酸。起始读段或中间重叠群的位置3'或5'通常是紧邻并超过起始读段或中间重叠群的in silico组装核酸序列中的3'或5'端的核苷酸位置。在某些实施例中,位于紧邻并超过起始读段或中间重叠群的3'或5'端的核苷酸位置,在尚未添加多数共有核苷酸的情况下(例如,在in silico重叠群组装过程期间尚未添加),在本文中被称作前进位置(例如,参见图4)。在某些实施例中,起始读段的位置

3' 或5' ,在起始读段的位置3' 或5' (例如,起始读段内的核苷酸的3' 或5')尚未被多数共有核苷酸填充的情况下,被称为前进位置。在某些实施例中,中间重叠群包括起始读段和添加到起始读段的3' 或5' 位置的一个或多个核苷酸。核苷酸通常被添加到起始读段或中间重叠群的位置3' 或5' ,其中该位置(例如,前进位置)包括多数共有核苷酸。

[0099] 在某些实施例中,根据多个重叠或比对来确定多数共有核苷酸,根据堆积关系确定多个重叠或比对。有时,根据被选择和/或存储的重叠来将一个或多个核酸读段与起始读段、中间重叠群或其部分比对。在某些实施例中,被选择和/或存储的重叠(例如,重叠的读段)被募集到包括起始读段或中间重叠群的比对,其中读段或重叠的一些或全部包括与前进位置重叠或比对的核苷酸。在某些实施例中,根据与前进位置重叠或比对的核苷酸确定多数共有核苷酸。在一些实施例中,多数共有核苷酸是位于前进位置处或与前进位置比对的核苷酸(例如,A、T、C、G或U),其中重叠的读段中的至少3个、至少4个、至少5个、至少6个、至少7个、至少8个、至少9个、至少10个、至少15个、至少20个、至少30个、至少50个、至少100个或至少200个包括前进位置处的相同的核苷酸(例如A、T、G、C或U)。在一些实施例中,多数共有核苷酸是位于前进位置处或与前进位置比对的核苷酸(例如,A、T、C、G或U),其中重叠读段的至少5%、至少6%、至少7%、至少8%、至少9%、至少10%、至少15%、至少20%、至少25%、至少30%、或至少50%包括前进位置处的相同的核苷酸(例如A、T、G、C或U)。

[0100] 在重叠群组装的某些实施例中,前进位置包括单个多数共有核苷酸,多数共有核苷酸被添加到起始读段或中间重叠群的3' 或5' 位置,并且针对前进位置重复in silico过程。在一些实施例中,前进位置包括多态碱基位置,例如其中多于一个多数共有核苷酸存在于前进位置(例如,多态碱基位置)。在针对多态碱基位置标识两个多数共有核苷酸的情况下,通常由产生两个相同的中间重叠群拷贝的中间重叠群进行复制。在这种情况下,被标识的两个多数共有核苷酸中的一个被添加到两个拷贝中的一个的前进位置,而被标识的另一个多数共有核苷酸被添加到另一拷贝的前进位置。这个过程有时被称为分裂或分裂重叠群。在某些实施例中,本文的系统、方法、过程或算法包括分裂一个或多个重叠群的方法。在一些实施例中,计算机程序组件(即,组件)向微处理器提供分裂一个或多个重叠群的指令。

[0101] 在其中针对前进位置(例如,多态碱基位置),三个多数共有核苷酸被标识的某些实施例中,生成中间重叠群的两个拷贝,从而产生3个相同的重叠群,并且三个多数共有核苷酸中的每一个中的一个被添加到三个相同的重叠群中的每一个的前进位置。在这种情况下,不同的核苷酸被添加到三个相同的重叠群中的每一个。换言之,重叠群被分裂为三个重叠群。同样地,在针对前进位置(例如,多态碱基位置)标识四个多数共有核苷酸的情况下,通常生成中间重叠群的两个拷贝,从而产生4个相同的重叠群,并且四个多数共有核苷酸中的每一个中的一个被添加到四个相同的重叠群中的每一个的前进位置。换言之,重叠群被分裂为四个重叠群。在某些实施例中,中间重叠群包括分裂重叠群(例如,由将重叠群分裂导致的重叠群)。

[0102] 在某些实施例中,在组装重叠群的过程期间,重叠群或中间重叠群被分裂多次。例如,在重叠群或中间重叠群的组装期间,可遇到第一多态碱基位置和第二多态碱基位置,其中第一多态位置导致重叠群的第一分裂,而第二多态碱基位置可导致重叠群的第二分裂。例如,中间重叠群可以分裂1次或多次、5次或多次、或50次或多次。在某些实施例中,中间重叠群被分裂1至500次、1至100次、1至50次、1至25次或1至10次。在某些实施例中,中间重叠

群不分裂。在某些实施例中，在由第一分裂产生的中间重叠群（例如，由遇到第一多态碱基位置的先前分裂产生的中间重叠群）的组装期间，遇到第二多态碱基位置。在这种情况下，中间重叠群可以被再次分裂，或者重叠群可以不被分裂。如果重叠群先前在某个位置被分裂（例如，第一多态位置），则分裂过程确定读段对或读段对的集合是否与第一多态位置和当前遇到的第二个多态位置（例如，其中两个或多个多数共有核苷酸比对的前进位置）重叠。在某些实施例中，如果存在这样的重叠读段对的集合，并且读段对的集合包括 (i) 在第一多态位置中添加的第一多态碱基和 (ii) 第二多态碱基位置处的单个多数共有核苷酸（例如，相同核苷酸），则针对第二多态碱基位置的多数共有核苷酸在前进位置处被添加到中间重叠群链，并且重叠群不被分裂。此外，满足条件 (i) 和 (ii) 的上述读段的集合不用于分裂任何其他重叠群，也不用于组装另一个重叠群。在某些实施例中，如果存在这样的重叠的读段对的集合，并且读段对的集合包括 (i) 在第一多态位置中添加的第一多态碱基和 (iii) 第二多态碱基位置处的两个或多个多数共有核苷酸，那么中间重叠体再次被分裂。在某些实施例中，不满足条件 (i) 但是为第二多态位置提供多数共有核苷酸的重叠的读段对的集合不被用于分裂上述示例中的中间重叠群，并且这样的读段对被排除在以上示例中的组装中间重叠群之外。这样的设计背后的理由是防止多态碱基处的分裂，其中包括该多态碱基的单倍型已经被包含在另一重叠群的组装中。在上述示例中，如果满足条件 (i)，并且与第一多态碱基位置重叠的读段集合包括针对第二多态碱基位置的两个或多个多数共有核苷酸，则在某些实施例中，重叠群将被相应地分裂。类似地，在某些实施例中，如果不满足条件 (i)，则重叠群将被分裂。在示例1中描述针对分裂的其他细节。

[0103] 在某些实施例中，通过复制已经在重叠群中产生的分裂来检测图形循环。在某些实施例中，如果检测到重复分裂，则重叠群被标记为“坏”，并且终止“坏”重叠群的组装。在某些实施例中，标记为“坏”的重叠群不用于超重叠群构建。

[0104] 在某些实施例中，读段集合中的读段仅被使用一次来构建重叠群。在某些实施例中，包括被并入重叠群的前进位置中的多数共有核苷酸的读段不用于向另一个重叠群添加附加的核苷酸。在某些实施例中，在重叠群被复制的情况下，由于存在两个或多个多数共有核苷酸，读段将仅被用于继续构建重叠群拷贝中的一个。在某些实施例中，读段可以在不同的重叠群中被重新使用。

[0105] 在某些实施例中，如果遇到已经遇到过并在某些其他重叠群中分裂的多态位置，则该重叠群中的重叠群分裂不被执行，而是仅在该重叠群中称为“重复”分裂。在这样的实施例中，重复分裂包含共有碱基的同一集合以及在读段的同一位置处支撑他们的读段的同一集合。在这样的实施例中，一旦所有重叠群被组装，则通过添加共有序列（来自具有“复制”分裂的重叠群和来自“复制分裂”位置后的重叠群所分裂的其他重叠群）的所有可能端来重新引入这些被跳过的分裂。在某些实施例中，假设一旦在重叠群建立过程期间遇到以相同方式堆积的读段的同一集合，则该位置之后的共有序列将是相同的，因为这些读段将在之后募集相同的读段集合。在某些实施例中，这种“重复”分裂检测不应该改变组装的重叠群，而只是加速重叠群组装计算过程。

[0106] 在某些实施例中，如果分裂的重叠群不能募集任何新的读段来扩展重叠群，而由该分裂位置产生的其他重叠群中的一些可以募集新的读段，则将该重叠群标记为“死胡同”。这些重叠群通常是跟随由于系统测序错误而不是真正的多态性而导致的共有碱基的

结果。在某些实施例中,这些“死胡同”重叠群被丢弃。

[0107] 在某些实施例中,从小于预定数量的读段或从包含小于预定数量的读段的堆积关系组装的重叠群被丢弃或删除。在某些实施例中,预定数量的读段是大约200读段或更少、100读段或更少、50读段或更少、25读段或更少或10读段或更少。在某些实施例中,通过滤波器丢弃、删除和/或删除从小于预定数量的读段组装的重叠群。在某些实施例中,被丢弃、删除和/或删除的重叠群不用于组装超重叠群。

[0108] 超重叠群

[0109] 在先前步骤中组装的重叠群可以跨越感兴趣的整个基因组区域,或者可以例如在覆盖度下降或者高读段错误率(例如,通常系统的错误)禁止高得分重叠的地方终止。在某些实施例中,跨越感兴趣的整个基因组区域的重叠群是超重叠群并且不需要附加的组装。超重叠群通常跨越感兴趣的整个基因组区域。没有跨越感兴趣的整个基因组区域的重叠群可以被组装成超重叠群。在某些实施例中,一个或多个超重叠群由两个或多个重叠群组装。在某些实施例中,使用读段配对物(例如,读段配对物对)来将重叠群链接在一起以形成超重叠群。例如,在某些实施例中,两个相邻重叠群之间的覆盖间隙可以由读段配对物对的读段配对物桥接,其中该对的第一读段配对物提供与第一个重叠群的重叠,并且该对的第二个读段配对物提供与另一重叠群的重叠。桥接或连接两个相邻重叠群的读的读段配对物可以提供关于重叠群之间的估计距离、重叠群的顺序和定向的信息。例如,读段配对物之间的估计插入长度可以提供两个桥接重叠群之间的估计距离。有时,桥接两个重叠群的读段配对物的定向提供了两个桥接的重叠群相对彼此的定向和顺序。在某些实施例中,根据多个读段配对物对,第一重叠群被连接到第二重叠群。在某些实施例中,根据至少3个、至少4个、至少5个、至少6个、至少7个、至少8个、至少9个、至少10个、至少15个、至少20个、至少30个、或至少50个读段配对物对,第一重叠群被连接到第二重叠群。

[0110] 在某些实施例中,一旦两个重叠群由一个或多个读段配对物链接,则可以募集和/或比对附加的读段、重叠(例如,根据堆积关系确定的)和/或重叠群,以组装被桥接的相邻重叠群之间的插入序列。

[0111] 在某些实施例中,超重叠群构建涉及构建以重叠群作为顶点并以标识的链路(例如,链接两个重叠群的读段配对物)作为定向边缘的图形。在某些实施例中,在两个相邻的重叠群被最少数量的读段配对物对桥接的情况下,定向边缘被记录,其中最小数量的读段配对物对是至少3个、至少4个、至少5个、至少6个、至少7个、至少8个、至少9个、至少10个、至少15个、至少20个、至少30个、或至少50个读段配对物对。在某些实施例中,桥接两个相邻重叠群所需的读段配对物对的最小数目为平均重叠群覆盖度的至少5%、至少6%、至少7%、至少8%、至少9%、至少10%、至少15%、至少20%、至少25%、至少30%、或至少50%。平均重叠群覆盖度是指跨越重叠群或中间重叠群的每个核苷酸位置的读段的平均数。例如,与重叠群中的每个核苷酸位置重叠的读段的数目通常被计算为位置覆盖度,并且在重叠群中的所有位置上的位置覆盖度的平均值是平均重叠群覆盖度。在某些实施例中,桥接两个重叠群的读段配对物对共享相同的定向。在某些实施例中,通过遍历图形中所有路径,同时避免循环,从具有0入度的所有顶点开始并以具有0出度的顶点结束(例如,参见图8),多于两个的重叠群被桥接,从而形成超重叠群。在某些实施例中,未被连接到任何其他重叠群(例如,同时具有0入度和0出度)的重叠群创建仅具有一个重叠群的超重叠群。

[0112] 单体型分析

[0113] 在某些实施例中,通过上述过程组装的超重叠群表示所有可能的序列排列,因此表示所有可能的单体型序列(即,单体型)。在某些实施例中,根据预定的倍性,单体型由识别系统直接组合,从而产生所有可能的基因型(例如,基因型假设、基因型似然性或基因型似然比)。在某些实施例中,所有单体型在被识别系统处理之前进行单体型分析过程。在某些实施例中,单体型分析过程启动与每个单体型相关联的对象(例如,单体型对象)。单体型对象可以包括映射权重、标识的错误结点和/或标识的错误插入。例如,在某些实施例中,单体型分析过程包括一些或所有读段到单体型序列(例如,超重叠群)的重新映射。在某些实施例中,该重新映射包括在示例I的“识别系统”部分中描述的映射权重的预计算,其中映射权重与每个单体型相关联。在某些实施例中,单体型分析器(haplotyper)过程还执行单体型序列中的错误结点和错误插入的标识(参见下文)。单体型分析器过程通常与识别系统的功能分开进行,以允许识别系统组件在它们被组合成基因型序列假设前,有机会基于单体型分析过程的输出(例如,单体型对象,例如映射权重、错误结点和/或错误插入的标识)来过滤单体型。在某些实施例中,在每个单体型对象中列出被标识的错误结点和错误插入,其中必要信息确定其读段的支持。然后,识别系统组件可以基于单体型对象的属性(例如,错误插入)使用截点来过滤和/或移除单体型。可以使用任何合适的截点。

[0114] 错误结点标识

[0115] 在某些实施例中,单体型分析器过程包括标识错误结点的方法。在一些实施例中,错误结点由于假阳性比对而创建。在某些实施例中,错误结点由序列读段组成,该序列读段源自由于某些序列相似性而被募集(例如,获得)并被包括在重叠群组装中的基因组的不同部分(例如,感兴趣的基因组区域外部的基因组的部分)。这样的序列相似性有时允许一些读段来加入某些序列,然而结点将被稀疏覆盖。单体型分析器过程可以标识单体型序列中的位置,其中跨接结点位置的读段对的数量远低于预期。在某些实施例中,单体型分析器过程通过计算一些距离外的读段配对物的预期数量(例如,从插入尺寸分布估计)并将它们与观察到的计数进行比较来找到可能的错误结点。低观察/预期比的位置可以被标记为可能的错误结点。在某些实施例中,统计拟合的合适估计(例如,卡方检验)可用于确定观察-预期差异的显著性。在某些实施例中,针对错误结点,通过使用插入尺寸分布的平均值附近的中心或非中心范围测试来标识错误结点。有时使用插入尺寸分布平均值附近范围的-20%和+80%(例如,对于Illumina读段对文库,当前尺寸为50)的间隔来计算预期计数并用于搜索观察到的计数。在某些实施例中,单体型分析器过程单独地计算向前读段和反向读段(以反方向的)的观察和预期计数,然后找到比率中的局部最小值。单体型分析器过程有时会报告超过规定比率截点的所有局部最小值。在某些实施例中,在向前和反向两个方向中执行搜索可以给出结点的算法确认。

[0116] 错误插入标识

[0117] 在某些实施例中,单体型分析器过程包括错误插入检测过程。在某些实施例中,错误插入是在in silico组装的超重叠群内的外来或错位的核酸序列的不期望插入或错误插入。在某些实施例中,错误插入检测过程确定单体型中存在或不存在错误插入。在某些实施例中,错误插入检测过程确定在单体型中存在或不存在错误插入的可能性或概率。在某些实施例中,错误插入检测过程对潜在的错误插入进行标记、设置权重或评分,并将这些对象

与单体型相关联。在某些实施例中,可以使用错误结点的配对(例如,如上所述)来标识错误插入。然而,为了标识错误插入,专用的错误插入检测过程通常比错误结点算法更灵敏和明确。

[0118] 在某些实施例中,错误插入检测过程1)标识映射到单体型的读段配对物对,其中读段配对物之间的距离大于估计的插入长度(例如,由插入尺寸分布组件确定),从而定义读段配对物之间假定的错误插入,以及2)确定假定的插入是否仅由完全包含在假定的插入内的读段配对物对占据。占据假定的插入区域的读段配对物是对特定区域的in silico组装有贡献的读段。可以使用任何合适的方法来确定读段配对物是否被完全包括在假定的插入内。例如,位于假定的插入侧面的读段配对物的中点或末端可用于定义假定的插入的开始和结束。在某些实施例中,可以使用与位于假定的插入侧面的读段配对物重叠(例如,根据堆积关系)的读段的集合的中点或端部来定义假定的插入区域。有时使用方法的组合。例如,错误插入过程可以首先标识映射到单体型的读段配对物对,其中读段配对物的中点之间的距离大于估计的插入长度(例如,由插入尺寸分布组件确定),从而根据侧面读段配对物的中点的位置来定义假定的插入的开始和结束。如果算法确定假定的插入仅被完全包含在假定的插入内的读段配对物对占据,则在某些实施例中,算法可以根据包括该插入的读段配对物的边缘或端部来重新定义假定的插入边缘。

[0119] 在某些实施例中,错误插入过程针对假定的插入内的每个碱基位置重新计算插入纯度的测量。如果在插入内存在大部分由插入读段(完全包含在建议的错误插入区域内的读段对)组成的、并且不被跨插入边界的读段(或具有跨插入边界或在插入边界外部的配对物的读段)污染的位置,则这种插入被识别为错误插入。可以使用任何合适的方法来计算对于假定的插入内的碱基位置的插入纯度,和/或定义、重新定义和/或确认错误插入区域的长度和/或边缘。例如,在某些实施例中,根据重叠和根据堆积关系,由重叠群组装过程重新计算每个碱基位置,其中从组装过程中排除完全包含在假定的插入内的读段配对物对。当使用这种方法时,不可以被多数共有核苷酸占据的碱基位置被用于定义和报告错误插入区域。可以使用任何类似的过程来定义、重新定义和/或确认错误插入区域。

[0120] 在某些实施例中,单体型分析器过程对被确定为包含错误插入的单体型进行标记、设置权重、惩罚或评分。在某些实施例中,单体型分析器过程对被确定不包含错误插入的单体型进行标记、设置权重或评分。在某些实施例中,识别系统使用单体型分析器过程分配给单体型的对象来确定单体型是否将被包括在基因型假设中。

[0121] 识别系统和单体型似然比

[0122] 在某些实施例中,识别系统过程组装基因型并确定基因型似然比。识别系统组件通常执行识别系统过程。识别系统(例如,识别系统组件)可以从超重叠群组装组件和/或单体型分析器(例如,单体型组件)接收单体型。在某些实施例中,识别系统过程将单体型组合,以生成针对给定倍性的所有可能的基因型。在某些实施例中,针对给定倍性的所有可能的基因型由识别系统组件(例如,“识别系统”)组装。在某些实施例中,针对给定倍性确定的每个可能的基因型被称为基因型假设。针对单倍体、二倍体、三倍体对象或任何倍性的对象,可以以所有可能的排列组合单体型。例如,对于二倍体序列假设,任何两个单体型的所有可能的配对(包括由相同单体型的两个拷贝组成的纯合排列)可由识别系统组装,其中的每个配对被称为基因型假设。

[0123] 在这种二倍体基因型中,针对每个单体型单体型贡献为0.5。在某些实施例中,单体型可以以任何比例组合,导致对基因型的分数单体型贡献。这样的分数基因型可用于基因型嵌合体(mosaic)个体样本或可反映正常组织污染和/或肿瘤异质性的肿瘤样品。在某些实施例中,由识别系统组装的每个基因型各自是基因型假设。因此,在某些实施例中,本文的方法和/或过程根据一个或多个单体型生成基因型似然比。在某些实施例中,本文的方法和/或过程根据一个或多个单体型及其对基因型的分数贡献生成基因型似然比。在某些实施例中,本文的方法和/或过程根据一个或多个基因型假设生成基因型似然比。因此,在某些实施例中,识别系统过程根据一个或多个单体型生成基因型似然比。在某些实施例中,识别系统过程根据一个或多个基因型假设(例如,一个被选择的基因型假设)生成基因型似然比。在某些实施例中,识别系统过程根据包括纯合参考基因组排列的基因型假设生成基因型似然比。

[0124] 在某些实施例中,例如根据错误结点、错误插入的存在或不存在和/或通过映射权重,由识别系统过程过滤(例如,排除)由识别系统从单体型分析器获得的单体型。经过滤的单体型通常不被识别系统用来组装基因型或确定基因型似然比。在某些实施例中,单体型不被识别系统过程过滤。

[0125] 在某些实施例中,针对感兴趣的基因组区域组装的基因型的数量表示针对给定倍性的该区域的所有可能的单体型序列排列。任何合适数量的基因型可以被组装用于感兴趣的基因组区域。有时多个基因型被组装。有时1个或多个基因型被组装。在某些实施例中,1至100,000,000个、1至1,000,000个、1至100,000个、1至10,000个、1至1000个、1至500个、1至200个、1至50个或1至20个基因型被组装以用于感兴趣的基因组区域。在某些实施例中,至少5个、至少10个、至少20个、至少30个、至少50个、至少100个、至少500个或至少1000个基因型被组装以用于感兴趣的基因组区域。

[0126] 在一些实施例中,识别系统过程根据一个或多个基因型似然比确定感兴趣的基因组区域(例如,针对对象)的基因型。在某些实施例中,识别系统过程根据一个或多个基因型似然比确定多个可能的基因型假设中最可能和/或最可信的基因型。在某些实施例中,识别系统过程可以向医疗保健专业人员或结果组件提供基因型假设的列表,其中列表包括概率、似然性、统计置信度量、误差度量、排名等,或其与每个基因型假设相关联的组合。在一些实施例中,识别系统过程根据一个或多个基因型假设确定基因型似然比。在一些实施例中,识别系统过程根据一个或多个基因型假设来确定一个或多个基因型似然比。

[0127] 在一些实施例中,基因似然比根据等式(1)确定

$$[0128] \quad \frac{P(G|\{R\})}{P(G_0|\{R\})} = \prod_{\{R\}} \frac{\sum_{\{A_G\}} \frac{1}{N_{A_G}} F_{A_G}(W(R, A_G) + \alpha)}{\sum_{\{A_{G_0}\}} \frac{1}{N_{A_{G_0}}} F_{A_{G_0}}(W(R, A_{G_0}) + \alpha)} \quad (\text{等式 1})$$

[0129] 其中G是预定倍性的基因型序列,G₀是参考序列,{R}是读段配对物对R的集合,N_{AG}是基因型序列G中的等位基因AG的数量,N_{AG₀}是在参考序列G₀中等位基因AG₀的数量,以及F_{AG}是基因型序列G中等位基因AG的分数,F_{AG₀}是参考序列G₀中等位基因AG₀的分数,W是读段对映射权重,α是映射概率常数。在一些实施例中,根据等式(1)的推导或等式(1)的变形来确定基因型似然比。在示例1中进一步描述等式(1)中的项及其推导。

[0130] 在一些实施例中,对象的倍性是已知的、预先确定的或假设的。在一些实施例中,

本文的方法或过程不确定对象的倍性。在一些实施例中，本文的方法或过程可以确定对象的估计的倍性，其中估计的倍性与概率相关联。在一些实施例中，本文的方法或过程可以确定对象的估计的倍性，其中估计的倍性与最大似然性相关联。在一些实施例中倍性是二倍体。在一些实施例中，针对人类对象（二倍体）确定基因型概率。例如，对于二倍体基因组，等位基因 F_{AG} 和 F_{AGO} 的分数各自等于0.5的值。

[0131] 在一些实施例中，等式(1)的 α 的值取决于读段对（例如，读段对的映射或可映射性）。例如，如果读段对在重叠群组装区域外部或感兴趣的基因组区域外部具有第二映射，则 α 值较大（例如，与 W 的值相当）。在读段对的可映射性差的一些实施例中， α 可以对应于 W 值。在一些实施例中，对于不具有第二映射（例如，良好可映射性）的读段的 α 的默认值可以为约 $1e-5$ 或更小、约 $1e-10$ 或更小、 $1e-20$ 或更小、约 $1e-25$ 或更小、约 $1e-30$ 或更小、约 $1e-40$ 或更小、约 $1e-50$ 或更小、约 $1e-60$ 或更小或约 $1e-70$ 或更小。在一些实施例中，对于不具有第二映射（例如，良好可映射性）的读段的 α 的默认值为约 $1e-50$ 或更少。在示例1中提供关于 α 和 W 的附加细节。

[0132] 在一些实施例中，生成基因型似然比包括将获得的或募集的一些或所有读段重新比对和/或映射。在一些实施例中，读段被识别系统组件（例如，“识别系统”）重新比对和/或映射到参考（例如，参考单体型或参考基因型假设）。在一些实施例中，读段被单体型组件重新比对和/或映射到参考（例如，参考单体型或参考基因型假设）。在一些实施例中，生成基因型似然比包括将所有读段重新对准和/或映射到参考基因组。在一些实施例中，生成基因型似然比包括将所有读段重新比对和/或映射到一个或多个单体型。在一些实施例中，生成基因型似然比包括将所有读段重新比对和/或映射到一个或多个单体型。在一些实施例中，生成基因型似然比包括将所有读段重新比对和/或映射到被指定为参考的一个或多个单体型（例如，基因型假设）。任何合适的单体型或基因型假设都可以作为参考。

[0133] 在一些实施例中，根据等式(1)确定多个基因型似然比。在一些实施例中，针对多个基因型假设（例如，可能的基因型）确定基因型似然比。在一些实施例中，根据一个或多个单体型或跨越感兴趣的基因组的整个长度的单体型对，针对多种基因型假设（例如，可能的基因型）确定基因型似然比。在某些实施例中，每个基因型假设与概率（例如，通过其总和归一化的基因型似然比）相关联。

[0134] 在一些实施例中，根据基因型似然比确定对象中遗传变异的存在或不存在。在某些实施例中，包括所有可能基因型中针对感兴趣的基因组区域的最高概率（例如，最高基因型似然比）的基因型假设是针对给定感兴趣的基因组区域的最可能的基因型。在一些实施例中，最可能的基因型表示针对感兴趣的基因组区域中的一个或多个单体型的核酸序列。在一些实施例中，根据最可能的基因型确定遗传变异的存在或不存在。

[0135] 在一些实施例中，具有最高似然比的基因型假设被用于产生响应(call)或确定结果。在一些实施例中，具有最高似然比的基因型假设被用于确定对象中遗传变异的存在或不存在。在一些实施例中，根据预定的截点确定最高似然比。在某些实施例中，将两个或多个似然比确定为最高似然比，并且其他参数或数据被用于确定结果或基因型。在一些实施例中，最高似然比的值包括约800至10,000的对数似然比。在一些实施例中，最高似然比包括约1000的对数似然比。

[0136] 在一些实施例中，两个顶部基因型假设之间的似然比可以用于估计遗传变异存在

或不存在的置信度。在一些实施例中,可以评估基因型假设的整个集合来确定遗传变异的存在或不存在,并且具有变异和不具有变异的假设的集合可以用于确定样本中存在变异的置信度。

[0137] 系统、机器、存储介质和接口

[0138] 在没有计算机、微处理器、软件、计算机程序组件或其他机器的情况下,通常不可以执行本文描述的某些过程和方法。本文描述的方法通常是计算机实现的方法,并且方法的一个或多个部分有时由一个或多个硬件处理器(例如,微处理器)、计算机或微处理器控制的机器执行。属于本文中描述的方法的实施例通常适用于通过本文所述的系统、机器和计算机程序产品中的指令实现的相同或相关过程。属于本文中描述的方法的实施例通常适用于由其上存储有可执行程序的非暂时性计算机可读存储介质实现的相同或相关过程,其中程序指示微处理器执行该方法或其部分。本文所使用的描述性术语“非暂时性”明确地是限制性的并且排除了暂时性的传播信号(例如,传输信号、电子传输、波(例如,载波))。本文使用的术语“非暂时性计算机可读媒介”和/或“非暂时性计算机可读介质”包括除了暂时性的传播信号之外的所有计算机可读介质。在一些实施例中,本文所述的方法和过程通过自动化方法执行。在一些实施例中,本文所述的一个或多个步骤和方法由微处理器和/或计算机执行和/或与存储器结合执行。在一些实施例中,自动化方法在软件、计算机程序组件、微处理器、外围设备和/或包括类似的机器中实现,其(i)获得包括多个读段配对物对的双端序列读段的集合,每个对包括两个读段配对物,其中每个对的两个读段配对物中的至少一个被映射到参考基因组的至少一部分,参考基因组包括预先选择的感兴趣的基因组区域,并且其中双端序列读段中的一些未被映射到参考基因组的所述至少一部分,(ii)确定序列读段的集合的堆积关系,(iii)根据堆积关系构建一个或多个重叠群,(iv)组装一个或多个超重叠群,(v)生成基因型似然比,(vi)确定遗传变异的存在或不存在,或(vii)执行其组合。机器、软件和接口可以用于实施本文描述的方法。使用机器、软件和接口,用户可以输入、请求、查询或确定针对使用特定信息、程序或过程(例如,获得读段、募集读段、映射读段、生成堆积关系、构建重叠群、组装单体型、生成基因型似然比、确定遗传变异的存在或不存在等,或其组合)的选项,这可以涉及例如实现统计分析算法、统计显著性算法、统计误差算法、统计概率算法、迭代步骤、验证算法和图形表示。在一些实施例中,数据文件可以由用户输入作为输入信息,用户可以通过合适的硬件介质(例如,闪存驱动器)下载一个或多个数据文件,和/或用户可以将来自一个系统的数据集发送到另一个系统,用于后续处理和/或提供结果(例如,将测序器的序列读段数据发送到计算机系统,用于序列读段映射;将被映射的序列数据发送到计算机系统,用于处理和产生一个或多个基因型似然比)。

[0139] 系统通常包括一个或多个机器。每个机器包括一个或多个存储器、一个或多个微处理器以及指令。在系统包括两个或多个机器的情况下,部分或全部机器可以位于相同的位置处,部分或全部机器可以位于不同的位置处,全部机器可以位于一个位置处和/或全部机器可以位于不同的位置处。在系统包括两个或多个机器的情况下,部分或全部机器可以位于与用户相同的位置处,部分或全部机器可以位于与用户不同的位置处,全部机器可以位于与用户相同的位置处,和/或全部机器可以位于与用户不同的一个或多个位置处。

[0140] 系统有时包括计算装置或测序装置,或者计算装置和测序装置(即,测序机器和/或计算机器)。本文所述的装置有时是机器。测序装置通常被配置为接收物理核酸并生成对

应于核酸的核苷酸碱基的信号。测序装置通常被“加载”有包括核酸的样本,并且加载在测序装置中的样本的核酸通常被施加核酸测序过程。如本文所使用的术语“加载测序装置”是指将测序装置(例如,流动池)的一部分与核酸样本接触,测序装置的该部分被配置为接收用于进行核酸测序过程的样本。在一些实施例中,测序装置被加载有样本核酸的变体。有时通过将样本核酸修改为适合于对核酸测序的形式过程(例如,通过连接;例如通过连接、扩增、限制性消化等或其组合将接合体添加到样本核酸的端部)产生变体。测序装置通常被配置为部分地执行合适的DNA测序方法,DNA测序方法生成对应于被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、数据文件、图像等或其组合)。

[0141] 对应于DNA序列的每个碱基的一个或多个信号通常通过合适的过程被处理和/或转化成碱基响应(base call)(例如,特定核苷酸碱基,例如鸟嘌呤、胞嘧啶、胸腺嘧啶、尿嘧啶、腺嘌呤等)。源自加载的核酸的碱基响应的集合经常被处理和/或组装成一个或多个序列读段。在多个样本核酸同时被测序(即,复用)的实施例中,可以利用合适的去复用过程将特定的读段与其源自的样本核酸相关联。序列读段可以通过合适的过程被比对到参考基因组,并且读段(被比对到参考基因组的部分)以及读段配对物(可以不与参考基因组比对)(例如,具有低可映射性得分的读段配对物或不可映射的读段配对物)可以如本文所述被存储和处理。

[0142] 测序装置有时与系统中的一个或多个计算装置相关联和/或包括系统中的一个或多个计算装置。一个或多个计算装置有时被配置为执行以下过程中的一个或多个:获得读段、募集读段、过滤读段、确定针对序列读段集合的堆积关系、构建一个或多个重叠群(例如,重叠群或中间重叠群)、组装一个或多个超重叠群、过滤重叠群、过滤单体型、执行单体型分析器的一个或多个功能、执行识别系统的一个或多个功能、组装一个或多个基因型、生成一个或多个基因型假设、生成一个或多个基因型似然比、确定基因改变的存在或不存在等或其组合。一个或多个计算装置有时被配置为执行以下附加过程中的一个或多个:从测序装置信号生成碱基响应、生成读段、修剪读段、将读段去复用、将读段比对或映射到参考基因组等。

[0143] 在一些实施例中,由多个计算装置执行方法或过程,并且由系统执行的全部过程的子集可以被分配或划分在系统中的特定计算装置中。可以以任何合适的组合在两个或多个计算装置或其组中划分全部数目的过程的子集。多计算装置系统有时包括测序装置本地的一个或多个适当的服务器,并且有时包括一个或多个非测序装置本地的一个或多个适当的服务器(例如,web服务器、在线服务器、应用服务器、远程文件服务器、云服务器(如云环境,云计算))。

[0144] 不同系统配置中的装置可以生成不同类型的输出数据。例如,测序装置可以输出碱基信号,并且碱基信号输出数据可以被传送到将碱基信号数据转换为碱基响应的计算装置。在一些实施例中,碱基响应是来自一个计算装置的输出数据,并被传送到另一个计算装置,以产生序列读段。在某些实施例中,碱基响应不是来自特定装置的输出数据,而是在接收到测序装置碱基信号以生成序列读段的同一装置中被使用。在一些实施例中,一个装置接收测序装置碱基信号,生成碱基响应、序列读段并将序列读段去复用,并且输出用于样本的去复用序列读段,该样本可以被传送到将序列读段比对到参考基因组的另一装置或其组。例如,来自一个装置的输出数据有时被置于物理存储设备上,并且存储设备被输送并连

接到第二装置,输出数据被传送到该第二装置。输出数据有时被一个装置存储在数据库中,并且第二装置从相同的数据库访问输出数据。

[0145] 在一些实施例中,用户与装置(例如,计算装置,测序装置)交互。例如,用户可以向软件发出查询,然后软件可以经由因特网访问获取数据集,并且在某些实施例中,可以提示可编程微处理器基于给定参数获取合适的数据集。可编程微处理器还可以提示用户选择由微处理器基于给定的参数选择的一个或多个数据集选项。可编程微处理器可以提示用户选择由微处理器基于经由互联网找到的信息、其他内部或外部信息等选择的一个或多个数据集选项。可以选择选项来选择一个或多个数据特征选择、一个或多个统计算法、一个或多个统计分析算法、一个或多个统计显著性算法、迭代步骤、一个或多个验证算法、以及方法、机器、装置(多个装置,本文中其复数也被称为(多个)装置)、计算机程序或其上存储有可执行程序的非暂时性计算机可读存储介质的一个或多个图形表示。

[0146] 本文所述的系统可以包括设备、外围设备、接口、存储介质、传感器和典型计算机系统(例如,网络服务器、膝上型计算机系统、桌面系统、手持系统、个人数字助理、手机、计算亭等)的部件。计算机系统可以包括一个或多个输入装置(例如,键盘、触摸屏、鼠标、语音识别或其他装置),以允许用户将数据输入到系统中。系统还可以包括一个或多个输出,输出包括但不限于显示器(例如CRT、LED或LCD)、扬声器、传真机、打印机(例如,激光式、喷墨式、击打式、黑白或彩色打印机)或用于提供信息(例如,结果和/或报告)的视觉、听觉和/或硬拷贝输出的任何其他合适的输出。

[0147] 计算机系统通常包括用户输入组件。用户输入组件便于对象和/或其他用户输入和/或选择信息。用户输入组件通常便于经由用户界面和/或其他接口设备输入和/或选择信息。例如,用户输入组件可以使得户界面向用户显示一个或多个视图图形视图,这便于用户输入和/或选择信息。在一些实施例中,用户输入组件被配置为便于经由与一个或多个用户相关联的一个或多个用户界面来输入和/或选择信息。在一些实施例中,用户输入组件被配置为便于通过网站、移动应用、利用其发送文本消息和/或电子邮件的智能程序和/或经由其他方法来输入和/或选择信息。在一些实施例中,输入和/或选择的信息包括与核酸序列、用户、样本和向微处理器提供附加指令的选项参数相关的信息。在一些实施例中,用户输入组件被配置为提示对象或用户和/或其他用户回答特定问题和/或提供其他信息。在一些实施例中,用户输入组件被配置为将时刻、持续时间和/或其他时间相关信息与其他被输入的、被选择的、被存储的、被提取的和/或被处理的信息相关联。

[0148] 在系统中,输入和输出装置可以被连接到中央处理单元,该中央处理单元可以包括用于执行程序指令的微处理器和用于存储程序代码和数据的存储器。在一些实施例中,过程可以被实现为位于单个地理站点中的单个用户系统。在某些实施例中,过程可以被实施为多用户系统。在多用户实现的情况下,可以通过网络连接多个中央处理单元。网络可以是本地的,涵盖建筑物的一部分中的单个部门、整个建筑物、跨越多个建筑物、跨越一个区域、跨越整个国家或全世界。网络可以是私有的,由提供方拥有和控制,或者可以被实现为基于因特网的服务,其中用户访问网页以输入和取回信息。因此,在某些实施例中,系统包括一个或多个机器,该机器可以位于用户本地也可以相对用户处于远程。一个位置或多个位置中的多个机器可以由用户访问,并且数据可以被串行地和/或并行地映射和/或处理。因此,可以使用合适的配置和控制来使用多个机器(例如,在本地网络、远程网络和/或“云”

计算平台中)对数据进行映射和/或处理。

[0149] 在一些实施例中,系统可以包括通信接口。通信接口允许在计算机系统和一个或多个外部设备之间传送软件和数据。通信接口的非限制性示例包括调制解调器、网络接口(以太网/WiFi)、通信端口(例如,USB端口、HDMI端口)、蓝牙、PCMCIA插槽和/或卡等。数据可以由合适的通信接口、设备和/或方法(包括但不限于手动输入设备和/或直接数据输入设备(DDE))输入。手动设备的非限制性示例包括键盘、概念键盘、触敏屏幕、光笔、鼠标、跟踪球、操纵杆、图形平板电脑、扫描仪、数码相机、视频数字化仪和语音识别设备。DDE的非限制性示例包括条形码读取器、磁条代码、智能卡、磁性墨水字符识别、光学字符识别、光学标记识别和周转文件。

[0150] 在某些实施例中,通过in silico过程生成模拟数据,并且模拟数据用作可以经由输入设备输入的数据。术语“in silico”是指数据(例如,重叠群、中间重叠群、超重叠群等)和/或使用计算机、一个或多个计算机程序组件或其组合执行的数据的操纵或变换。在某些实施例中,本文中的方法和过程以in silico方式执行。in silico过程包括但不限于映射读段、比对读段、重叠读段、生成堆积关系、迭代过程(例如,迭代组装或构建重叠群、中间重叠群和/或超重叠群或其部分)、组装单体型、组装基因型和/或基因型假设。

[0151] 系统可以包括用于执行本文所述的过程的软件,并且软件可以包括用于执行这些过程的一个或多个计算机程序组件。术语“软件”是指包括程序指令(例如,可执行程序)的计算机可读存储介质,当程序指令由计算机执行时,执行计算机操作。由一个或多个微处理器可执行的指令有时被提供为可执行代码,当可执行代码被执行时,可以使得一个或多个微处理器实现本文描述的方法。

[0152] 本文描述的计算机程序组件(即,组件)可以作为软件、和/或在可以由处理器或微处理器实现或执行的软件中实现的指令(例如,过程、例程、子例程)存在。例如,计算机程序组件可以是执行特定过程或任务的程序的一部分。术语“计算机程序组件”和“组件”在本文中被同义使用,并且是指可用于较大机器或软件系统的独立功能单元。组件可以包括用于通过一个或多个微处理器执行计算机程序组件的功能的指令的集合。计算机程序组件的指令可以在计算环境中、通过使用合适的编程语言、合适的软件和/或以合适的语言编写(例如,本领域已知的计算机编程语言)的代码和/或操作系统来实现,操作系统的非限制性示例包括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,与PHP一同)、PGP、G、R、S等或其组合。

[0153] 在一些实施例中,计算机程序组件包括一个或多个数据文件,并且可以将数据文件传送到另一个计算机程序组件和/或从另一个计算机程序组件接收数据文件。在一些实施例中,组件将数据和/或信息例如转换成有形的打印物、给用户的指令、结果、显示、基因型等或其组合。例如,本文所述的一个或多个组件和/或微处理器(例如,装置或机器)可以获得测序读段(其表示对象基因组的随机的、无序的核酸片段),并将这些读段转换成对象主体的特定部分(例如,对象主体的一部分(例如,感兴趣的基因组区域的基因型))的精确表示(例如,显示)。该过程可以相当于将数百万个拼图转换成图片的过程或将X射线数据的比特转换成对象主体的一部分的显示(例如,骨骼、器官和其他身体组织的显示)的过程。

[0154] 一个或多个组件可以用于本文所述的方法,其非限制性示例包括测序组件、募集

组件、堆积关系组件、超重叠群组组装器组件、重叠群组组装组件、超重叠群组组装组件、插入尺寸分布组件、接合体 (adaptor/adapters) 修剪器组件、读段-读段比对器、单体型组件、识别系统、结果组件等或其组合。组件有时由微处理器控制。在某些实施例中, 组件或包括一个或多个组件的机器收集、组装、接收、获得、访问、恢复来自另一组件、机器、接口、外围设备或机器的操作者 (用户) 的数据和/或信息, 和/或向其他组件、机器、接口、外围设备或机器的操作者 (用户) 提供和/或传送数据和/或信息。在一些实施例中, 由包括以下项目中的一个或多个的机器向组件提供数据和/或信息 (例如, 序列读段): 一个或多个流动池、相机、检测器 (例如, 光电检测器、光电池、电检测器 (例如, 幅度调制检测器、频率和相位调制检测器、锁相环检测器))、计数器、传感器 (例如, 压力、温度、体积、流量、重量的传感器)、液体处理设备、数据输入设备 (例如, 键盘、鼠标、扫描仪、语音识别软件和麦克风、手写笔等)、打印机、显示器 (例如, LED、LCT 或 CRT) 等或其组合。例如, 机器或装置的操作者有时向组件提供常数、阈值、公式或预定值。计算机程序组件通常被配置为向微处理器、存储介质和/或存储器传送数据和/或信息、或传送来自微处理器、存储介质和/或存储器的数据和/或信息。组件通常被配置为将数据和/或信息传送到另一合适的组件或机器, 或从另一合适的组件或机器接收数据和/或信息。组件可以操纵和/或转换数据和/或信息。从组件导出或转换的数据和/或信息可以被传送到另一个合适的机器和/或组件。包括计算机程序组件的机器可以包括至少一个微处理器。包括组件的机器可以包括微处理器 (例如, 一个或多个微处理器), 该微处理器可以执行和/或实现组件的一个或多个指令 (例如, 过程、例程和/或子例程)。在一些实施例中, 组件与一个或多个外部微处理器 (例如, 内部或外部网络、服务器、存储设备和/或存储网络 (例如, 云)) 一起操作。

[0155] 数据和/或信息可以是合适的形式。例如, 数据和/或信息可以是数字的或模拟的。在某些实施例中, 数据和/或信息有时可以是分组、字节、字符或比特。在一些实施例中, 数据和/或信息可以是任何被收集、组装的或可用的数据或信息。数据和/或信息的非限制性示例包括合适的媒体、图片、视频、声音 (例如, 频率、可听见的或不可听见的)、数字、常数、数据文件、值、对象、时间、函数、指令、地图、参考、序列、读段、被映射的读取、级别、范围、阈值、信号、显示、表示或其变换。计算机程序组件可以接受或接收数据和/或信息, 将数据和/或信息转换成第二形式, 并将第二形式的信息提供或传送到机器、外围设备、设备、微处理器、存储设备、接口或到另一个计算机程序组件。在某些实施例中, 微处理器可以执行组件中的指令。在一些实施例中, 需要一个或多个处理器来执行计算机程序组件或计算机程序组件组中的指令。计算机程序组件可以向另一个计算机程序组件、机器或源提供数据和/或信息, 并且可以从另一个计算机程序组件、机器或源接收数据和/或信息。

[0156] 计算机程序产品有时在非暂时性计算机可读介质上实现, 并且有时在非暂时性计算机可读介质上有形地实现。在某些实施例中, 计算机可读存储介质包括存储在其上的可执行程序。计算机程序组件有时被存储在非暂时性计算机可读介质 (例如, 磁盘、驱动器) 或存储器 (例如, 随机访问存储器) 中。能够实现来自计算机程序组件的指令的计算机程序组件和微处理器可以位于机器或不同的机器中。能够实现计算机程序组件的指令的计算机程序组件和/或微处理器可以位于与用户相同的位置 (例如, 本地网络) 中, 或位于与用户不同的位置 (例如, 远程网络、云系统) 中。在其中结合两个或多个计算机程序组件执行方法的实施例中, 计算机程序组件可以位于同一机器中, 一个或多个计算机程序组件可以位于相同

物理位置中的不同机器中,并且一个或多个计算机程序组件可以位于不同物理位置中的不同机器中。

[0157] 在某些实施例中,机器、装置或计算机包括一个或多个计算机组件部分、外围设备和/或接口。外围设备和/或计算机组件部分有时可以将数据和/或信息传送到计算机程序组件、接口、显示器、外围设备和/或其他计算机组件部分,并且传送来自计算机程序组件、接口、显示器、外围设备和/或其他计算机组件部分的数据和/或信息。在某些实施例中,机器与提供数据和/或信息的外围设备和/或计算机组件部分交互。在某些实施例中,外围设备和计算机组件部分帮助机器执行功能或直接与计算机程序组件交互。外围设备和/或计算机组件部分的非限制性示例包括合适的计算机外围设备、I/O或存储方法或设备,该存储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(例如,iPad、平板电脑)、触摸屏、智能电话、移动电话、USB I/O设备、电子存储装置(USB大容量存储设备、光学可读存储介质(例如,光盘等)、磁可读存储介质(例如,磁带、磁性硬盘驱动器、软盘驱动器等)、基于电荷的存储介质(例如,EPROM、RAM等)、固态存储介质(例如,闪存驱动器等)、和/或其他电子可读存储介质)、键盘、计算机鼠标、数字笔、调制解调器、硬盘驱动器、跳转驱动器、闪存驱动器、微处理器、服务器、CD、DVD、图形卡、专用I/O设备(例如,定序器、光电池、光电倍增管、光学读取器、传感器等)、网络接口控制器、只读存储器(ROM)、随机访问存储器(RAM)、无线传输设备(蓝牙设备、WiFi设备等)、万维网(www)、因特网、计算机和/或另一个计算机程序组件。

[0158] 计算机程序组件和计算机实现

[0159] 在一些实施例中,系统包括被配置为生成序列读段的序列组件。序列组件可以包含核酸测序仪(例如,被设计和配置为生成核酸库的序列读段的机器或装置)和/或被配置为生成、组装、映射和修剪序列读段的软件和指令。序列组件通常以数据文件(例如,bam文件、fasta文件等)的形式提供序列读段。序列组件可以以任何合适的文件格式提供序列读段。

[0160] 在一些实施例中,系统包括募集组件。在一些实施例中,募集组件被配置为从合适的源和/或输入装置获得数据文件的形式的序列读段(例如,双端序列读段)。例如,募集组件可以从被配置为产生读段的装置(例如,被配置用于核酸排序的装置)和/或从被配置为组装和/或映射读段的计算机获得或接收读段。在一些实施例中,募集组件从序列组件获得和/或募集读段。在一些实施例中,募集组件从合适的非暂时性或暂时性存储介质获得读段。例如,人可以通过以任何合适的方式(例如,经由跳转驱动器、光盘、电子邮件、因特网等)向募集组件提供数据文件来向募集组件提供序列读段。在某些实施例中,募集组件获得和/或募集被映射到参考的读段和/或未被映射、不一致地或不良地映射到参考的读段(例如,具有低可映射性的读段)。在某些实施例中,募集组件获得被映射的双端序列读段的读段配对物以及它们对应的读段配对物(不管它们是被映射的、未映射的、不一致还是不良映射的)。在某些实施例中,募集组件获得包括多个读段配对物对的双端序列读段的集合,每个对包括两个读段配对物,其中每个对的两个读段配对物中的至少一个被映射到包括预先选择的感兴趣的基因组区域的参考基因组的至少一部分,并且其中双端序列读段的一些未被映射到参考基因组的该至少一部分。在一些实施例中,募集组件获得和/或存储与读段(例如,读段长度、读段配对物对的定向和读段配对物对的估计插入长度)相关联的信息。在

一些实施例中,募集组件被配置为将所选择的读段(例如,募集的读段、获得的读段、所选择的读段集合)传送到另一个计算机程序组件。例如在一些实施例中,募集组件将选择的读段传送到过滤器组件、修剪器组件、映射组件或堆积关系组件、重叠群组装组件、超重叠群组装组件和/或识别系统组件。

[0161] 在一些实施例中,系统或存储介质包括插入尺寸分布组件。在一些实施例中,插入尺寸分布组件从募集组件、超重叠群组装器、超重叠群组装组件、单体型组件或基因型似然比组件传送和/或接收数据和/或信息。在一些实施例中,插入尺寸分布组件通常被配置为确定读段对或双端读段子集的插入尺寸分布、估计插入尺寸、估计插入长度和/或插入尺寸的估计似然性。在一些实施例中,插入尺寸分布组件生成估计的、计算的或测量的插入片段长度的分布,并确定针对双端读段的子集的估计插入尺寸。插入尺寸分布组件有时将插入尺寸的估计插入尺寸和/或估计的似然性并入或索引到数据文件中。在一些实施例中,插入尺寸的估计似然性与读段配对物对相关,并且用于确定给定读段配对物对被映射或比对到重叠群或超重叠群的可能性。在一些实施例中,插入尺寸分布组件根据估计的插入长度确定与读段配对物到另一读段配对物对、参考序列、重叠群或超重叠群的映射或比对相关联的似然性或概率。

[0162] 在一些实施例中,插入尺寸分布组件将似然性(例如,插入尺寸似然性)、似然性得分或惩罚分配到超重叠群或基因型假设。在一些实施例中,插入尺寸分布组件将似然性、似然性得分或惩罚分配到建议的插入尺寸。例如,在一些实施例中,单体型分析器将读段对与基因型假设的超重叠群或单体型进行重新比对。在某些实施例中,插入尺寸分布组件由单体型分析器募集以确定被重新比对的读段对之间的距离(该距离由超重叠群或单体型定义)是否与针对读段对的所估计的插入尺寸或插入尺寸分布一致。在一些实施例中,插入尺寸分布组件比较(i)被重新比对的读段对之间的距离(该距离由读段对比对到的超重叠群或单体型定义)与(ii)针对读段对确定的插入尺寸分布或估计的插入尺寸,并返回在一些实施例中与读段对比对到的超重叠群和单体型相关联的似然性、似然性得分或概率。例如,在一些实施例中,当读段配对物对的读段配对物映射到假设序列(例如,超重叠群、基因型假设的单体型)并且被映射成彼此间隔比估计的插入尺寸更长的距离时,这么长的插入尺寸的似然性将会低,并且低似然性被用于对该假设的似然比进行惩罚。在一些实施例中,插入分布组件根据被重新比对到超重叠群或基因型假设的多个读段对确定似然性、似然性得分或概率。在一些实施例中,插入分布组件根据被重新比对到超重叠群或基因型假设的多个读段对确定似然性、似然性得分或概率,并将似然性、似然性得分或概率与读段被重新比对到的超重叠群或基因型假设相关联。在某些实施例中,似然性或似然性得分包括与超重叠群、单体型或基因型假设相关联的惩罚或惩罚得分。在某些实施例中,数据和/或信息(例如,插入尺寸分布、估计的插入尺寸、似然性、插入尺寸似然性、似然性得分、得分或概率)被传送到单体型组件和/或由单体型组件处理,以确定或预计算每个读段对到每个单体型的映射权重。在一些实施例中,映射权重部分地根据由插入尺寸分布组件确定的一个或多个插入尺寸似然性或惩罚来确定。在一些实施例中,映射权重包括从插入尺寸分布组件导出的插入尺寸似然性。

[0163] 在一些实施例中,根据堆积关系(例如,由堆积关系组件确定)确定多个重叠或比对,根据多个重叠或比对确定多数共有核苷酸。在某些实施例中,根据所映射的读段配对物

之间的估计或隐含的插入长度并针对其对应的读段配对物的重叠和/或比对来检查读段的重叠和/或比对。这种功能通常由插入尺寸分布组件执行。例如,其中读段配对物对的两个读段配对物与重叠群的部分比对并且两个读段配对物端部之间的距离暗示某个插入尺寸,则比对将接收插入尺寸似然性,该插入尺寸似然性与正由实验室DNA片段协议生成的这样的插入尺寸的频率对应。在一些实施例中,其中读段配对物对的两个读段配对物与重叠群的部分重叠或比对,并且两个读段配对物端部之间的距离(这暗示着其插入长度)在给定DNA分段过程中非常频繁,则这样的读段对比对被分配相对较高的似然性。在另一方面,低的似然性将被分配给给定DNA分段协议中暗示着太短或太长插入尺寸的读段对比对。在重叠群组装期间或在超重叠群组装期间,有时会评估重叠和比对惩罚。在一些实施例中,包括惩罚的重叠和/或比对不包括在重叠群组装、中间重叠群组装或超重叠群组装中,或用于重叠群组装、中间重叠群组装或超重叠群组装。包括读段配对物对的两个读段配对物的比对和重叠通常由插入尺寸分布组件检查。在某些实施例中,插入尺寸分布组件根据由读段配对物端部的位置确定的插入长度来评估读段配对物到重叠群、中间重叠群、超重叠群和单体型的重叠和比对。在某些实施例中,插入尺寸分布组件将权重和/或惩罚或似然性分配给读段配对物对的某些重叠和比对(例如,读段配对物到重叠群、中间重叠群、超重叠群和单体型的比对)。在一些实施例中,插入尺寸分布组件确定读段配对物对(例如,两个读段配对物)是由in silico生成序列(例如,重叠群、超重叠群、单体型或单体型假设)生成的似然性,其中根据其到序列的比对所暗示的读段对的插入长度确定似然性。在一些实施例中,插入分布组件将似然性与每个读段对到in silico生成的序列(例如,重叠群、超重叠群、单体型或单体型假设)的比对相关联,其中根据插入尺寸确定似然性并且似然性被包括为如等式(3)和(6)中所示的概率 $P(I_M)$,其中 I_M 是由读段对的映射 M 暗示的插入尺寸。在一些实施例中,可以从经验插入尺寸分布中获得概率 $P(I_M)$ 。插入尺寸分布组件通常将数据和/或信息发送到堆积关系组件、重叠群组装组件、超重叠群组装组件、识别系统和/或识别系统组件。

[0164] 在一些实施例中,系统包括堆积关系组件(即,关系组件)。在一些实施例中,堆积关系组件确定一个或多个堆积关系。在一些实施例中,堆积关系组件被配置为执行比对、生成重叠并且确定关系或将关系(例如,堆积关系)分配到读段和/或读段配对物。在一些实施例中,堆积关系组件被配置为针对读段的集合生成一个或多个堆积关系。堆积关系组件通常从募集组件获得和/或接收读段,并根据所接收的读段生成一个或多个堆积关系。在某些实施例中,堆积关系组件针对读段的集合或子集生成所有可能的重叠。在某些实施例中,堆积关系组件根据合适的k-mer哈希策略生成针对读段集合或子集的重叠。在某些实施例中,堆积关系组件过滤、移除和/或修剪重叠。在某些实施例中,堆积关系组件选择和/或存储重叠。在一些实施例中,堆积关系组件生成堆积图形和/或平铺图。堆积关系组件经常将针对读段集合选择的重叠和/或读段-读段比对传送到重叠群组装组件。

[0165] 在一些实施例中,系统包括重叠群组装组件。在某些实施例中,重叠群组装组件从募集组件或堆积关系组件接收数据和/或信息(例如,数据文件)。重叠群组装组件通常被配置为根据堆积关系通过将核苷酸(例如,in silico)迭代地添加到起始读段或中间重叠群来组装重叠群。重叠群组装组件通常确定读段、读段配对物和/或读段配对物对到起始读段、重叠群或中间重叠群的部分的重叠和/或比对。在一些实施例中,重叠群组装组件通常根据堆积关系来确定读段、读段配对物和/或读段配对物对的重叠和/或比对。在一些实施

例中,重叠群组装组件通常根据针对读段的某些重叠和/或比对确定的惩罚和/或权重来确定读段、读段配对物和/或读段配对物对的重叠和/或比对。惩罚、权重和/或其不存在通常由插入尺寸分布组件确定,并被发送到重叠群组装组件,其中信息被用于在重叠群组装期间包括或排除某些读段的重叠或比对。由重叠群组装组件生成的重叠群(例如,重叠群和中间重叠群)通常被传送到超重叠群组装组件。

[0166] 在一些实施例中,系统包括超重叠群组装组件。在某些实施例中,超重叠群组装组件从重叠群组装组件、关系组件、插入尺寸分布组件和/或从募集组件接收数据和/或信息(例如,数据文件)。超重叠群组装组件通常被配置为通过将重叠群与一个或多个读段配对物对桥接来构建和组装超重叠群。超重叠群组装组件通常确定读段、读段配对物和/或读段配对物对到一个或多个重叠群或中间重叠群的部分的重叠和/或比对。在一些实施例中,超重叠群组装组件通常确定连接两个或多个重叠群的读段、读段配对物和/或读段配对物对的重叠和/或比对。在一些实施例中,超重叠群组装组件通常根据针对读段和读段配对物对的某些重叠和/或比对确定的惩罚和/或权重来确定读段、读段配对物和/或读段配对物对的重叠和/或比对。惩罚、权重和/或其不存在通常由插入尺寸分布组件确定并被发送到超重叠群组装组件,其中该信息用于在超重叠群组装期间包括或排除某些读段重叠或比对。在一些实施例中,由超重叠群组装组件生成的超重叠群被传送到识别系统或识别系统组件。在一些实施例中,由超重叠群组装组件生成的超重叠群被传送到单体型组件。

[0167] 在一些实施例中,系统包括执行一个或多个单体型分析器过程的单体型分析器(例如,单体型组件)。一个或多个单体型分析器过程通常由单体型组件执行。单体型组件可以利用超重叠群组装器组件、超重叠群组装组件、堆积关系组件、插入尺寸分布组件或募集组件中的一个或多个来接收和/或交换对象、数据和/或信息。单体型组件可以将对象、数据和/或信息发送到识别系统或结果组件。在一些实施例中,系统不包括单体型组件。

[0168] 在一些实施例中,系统包括识别系统(例如,识别系统组件)。在某些实施例中,识别系统组件从超重叠群组装组件、单体型组件、关系组件、插入尺寸分布组件和/或从募集组件接收数据和/或信息(例如,数据文件)。在某些实施例中,识别系统组件针对给定倍性的所有可能的基因型。在一些实施例中,识别系统执行等式6(等式6)和/或等式1(等式1)的功能。在一些实施例中,识别系统预先计算针对每个读段对和每个等位基因(超重叠群)的读段对权重,并且在假设似然性计算期间调用该值,该过程可由识别系统组件执行。在一些实施例中,为了便于计算针对所有读段的读段权重,识别系统将所有读段重新比对到所有超重叠群。在一些实施例中,识别系统将所有读段映射到参考基因组或选择单体型之一(例如,序列假设)作为参考。在某些实施例中,确定的第一假设成为参考,并且计算关于第一假设的所有似然性。

[0169] 在一些实施例中,识别系统组件根据倍性(例如,输入倍性、默认倍性)从一个或多个超重叠群生成一个或多个基因型序列。识别系统可以根据任何合适的倍性组装基因型序列(例如,基因型可能性、基因型假设)。在一些实施例中,基因型序列可以是超重叠群对,其中倍性是二倍体。基因型序列可以是单个超重叠群,其中倍性是单倍体。基因型序列可以由三个超重叠群组成,其中倍性是三倍体。在一些实施例中,识别系统从表示针对给定倍性的每个可能的基因型假设中组合多个基因型序列。通常从识别系统向结果组件传送基因型序列(例如,基因型假设)。

[0170] 在某些实施例中,识别系统从超重叠群组装组件、堆积关系组件、插入尺寸分布组件和/或从募集组件接收数据和/或信息(例如,数据文件)。在一些实施例中,识别系统针对一个或多个超重叠群生成一个或多个基因型序列似然比。在一些实施例中,识别系统组件生成多个基因型似然比,其中针对单体型的组装(例如,基因型假设)生成每个似然比。由识别系统组件生成的基因型序列似然比通常被传送到结果组件。

[0171] 在一些实施例中,系统包括结果组件。结果组件通常从识别系统组件接收数据和/或信息(例如,基因型概率)。在一些实施例中,结果组件通常从识别系统组件获得一个或多个基因型似然比。结果通常由结果组件提供。有时从结果组件向健康护理专业人员(例如,实验室技术人员或管理者;医师或助理)提供结果。结果组件可以包括合适的统计软件包。在某些实施例中,结果组件生成绘图、表格、图表或图形。在一些实施例中,结果组件生成和/或比较标准得分(例如,Z得分)。遗传变异的存在或不存在和/或相关的身体状况(例如,结果)通常由结果组件确定和/或由结果组件提供。在一些实施例中,对象中遗传变异的存在或不存在由包括结果组件的机器标识。结果组件可以专门用于确定特定的遗传变异(例如,STR、易位、多态性、插入)。例如,标识STR的结果组件可以不同于和/或区别于标识易位的结果组件。在一些实施例中,需要结果组件或包括结果组件的机器通过将基因型序列比对到参考序列来标识遗传变异或遗传变异决定的结果。在某些实施例中,结果从结果组件传送到显示组件,其中结果由显示组件提供(例如,合适的显示器,例如,LED等)。在一些实施例中,结果组件向显示器提供基因型(例如,基因型序列、基因型图像)的表示。

[0172] 遗传变异和身体状况

[0173] 在一些实施例中,本文所述的系统、方法或过程确定对象中遗传变异的存在或不存在。在一些实施例中,根据基因型似然比和/或结果组件确定对象中遗传变异的存在或不存在。遗传变异通常是存在于某些个体中的特定遗传表型。在一些实施例中,遗传变异是染色体异常(例如,染色体的一个或多个部分的缺失或增加)。遗传变异的非限制性示例包括一个或多个缺失、复制、插入、微插入、添加、易位、突变、多态性(例如,单核苷酸多态性、多核苷酸多态性)、融合、重复(例如,短串联重复(即,STR))等及其组合。插入、重复、缺失、复制、突变或多态性可以是任何长度,并且在一些实施例中,其长度为大约1碱基或碱基对(bp)至约250兆碱基(Mb)。在一些实施例中,插入、重复、STR、缺失、复制、突变或多态性长度为约1个核苷酸(nt)至约50,000nt(例如,约1个至约10,000个核苷酸、约1个至约10,000个核苷酸、约1个至约1,000个核苷酸、约1个至约500个核苷酸、约1个至约400个核苷酸、约1个至约300个核苷酸、约1个至约200个核苷酸、约1个至约100个核苷酸、或约1个至约100个核苷酸、或约1个至约50个核苷酸)。在一些实施例中,由本文所述的方法、系统或过程确定的遗传变异包括以下长度:约2个至约500个核苷酸、约2个至约400个核苷酸、约2个至约300个核苷酸、约2个至约200个核苷酸、约2个至约100个核苷酸、约2个至约50个核苷酸、10个至约500个核苷酸、约10个至约400个核苷酸、约10个至约300个核苷酸、约10个至约200个核苷酸、约10个至约100个核苷酸、约10个至约50个核苷酸、20个至约500个核苷酸、约20个至约400个核苷酸、约20个至约300个核苷酸、约20个至约200个核苷酸、约20个至约100个核苷酸、或约20个至约50个核苷酸。

[0174] 遗传变异可以被包括在基因内。包括遗传变异的基因可以包括基因中或基因附近的遗传变异,该遗传变异可以在基因的内含子、外显子、基因的非翻译区或其组合中。任何

基因可以包括通过本文所述的方法或过程确定的遗传变异。例如,遗传变异可以包括AR、ATXN1、ATXNNX2、ATXN3、ATXN7、ATXN8、ATXN10、DMPK、FXN、JPH3、CACNA1A、PPP2R2B、TBP、ATN1、ARX、PHOX2B、PABPN1、ATT、CFTR和BRCA1基因。

[0175] 在某些实施例中,针对对象所标识存在或不存在的遗传变异有时与身体状况相关联。身体状况的非限制性示例包括与以下相关联的那些:智力障碍(例如,唐氏综合症)、异常细胞增殖(例如,癌症)、非霍奇金淋巴瘤、骨髓增生异常综合症、威廉综合症、兰格-吉登综合症、阿尔菲综合症、雷波尔综合症、雅各布综合症、视网膜母细胞瘤、史密斯-马格尼斯、爱德华综合症、乳头状肾细胞癌、迪格奥尔格综合症、天使人综合症、猫眼综合症、家族性腺瘤性息肉病、米勒-德克尔综合症、微生物核酸的存在(例如,病毒、细菌、真菌、酵母)和先兆子痫。

[0176] 示例

[0177] 以下描述的示例图示某些实施例并且不限制本技术。

[0178] 示例1:Kragle:用于短串联重复序列和其他染色体组位点(基因座)的局部de Novo组装器和基因型识别系统

[0179] Kragle被设计为局部de-novo序列组装和基因分型包。Kragle被设计为从双端读段组装任何倍性的序列。Kragle被专门设计于处理延伸到读段长度的重复序列,但也可以识别(call)由序列反转、易位、复制或缺失引起的序列结点。Kragle被成功应用于识别基因状况阵列中涉及的AR、ATXN1、ATXN2、ATXN3、ATXN7、DMPK、FXN和HTT基因中的短串联重复(STR)的二倍体基因型。Kragle也用于确认人类BRCA1基因中大量缺失导致的假设结点,以及识别涉及人类CFTR基因中均聚物和相邻二核苷酸重复的复杂变异。

[0180] Kragle的功能分为四个主要组件(图1):读段募集组件(即,募集器、募集组件)、超重叠群组装器、单体型创建器(单体型分析器)和基因型假设识别系统(识别系统)。超重叠群组装器包括三个附加组件(堆叠关系组件、超重叠群组装组件和超重叠群组装组件)。前两个主要组件(募集器组件和超重叠群组装器)被设计为充分利用双端读段来组装低复杂度序列,低复杂度序列可以包括与读段长度相同的不中断的重复内容,并且如果重复包含中断,则有时重复内容甚至长于读段长度。单体型分析器从超重叠群序列构建单体型,将所有读段重新映射到它们,并且尝试标识组装序列中的不规则性。基因型概率组件从超重叠群组装组件来组装单体型序列(例如,超重叠群)并生成基因型概率并分配置信度值。识别系统从组装的单体型序列中识别二倍体基因型,并基于似然比将置信度分配给备选的基因型假设。本示例中的Kragle实施例不利用任何读段误差校正,因为读段误差校正算法干扰重复(例如,STR)的存在或不存在的确定。

[0181] 在一些实施例中,Kragle的输入是参考排序的bam文件和插入尺寸分布文件。备选地,Kragle接受两个fastq文件(针对读段1和读段2)和一个参考fasta文件来调用接合体修剪器组件、BWA-MEM比对器、插入尺寸分配组件和samtools(参考排序和索引bam文件),以创建在预处理步骤(参见图1)中必要的输入。

[0182] 募集组件

[0183] 从与参考序列显著不同的样本的基因组区域获得的序列读段对标准读段比对器呈现了显著的挑战。例如,源自基因组改变部分的读段通常被映射到错误的基因组位置或保持未被映射。然而,在这种情况下,来自读段配对物对的读段配对物通常包含未改变(或

少量改变)侧翼区域的序列,并且因此可以被正确映射。为了规避延伸的STR、序列结点和大的复杂变异区中的读段映射问题,募集组件使用来自读段配对物对的被映射的读段配对物的位置来标识可以提供Kragle尝试组装的区域(例如,特定的感兴趣的基因组区域)的信息的读段。

[0184] 在一些实施例中,Kragle被配置为使用参考排序的bam文件来快速索引到感兴趣的区域(一个或多个)中。在结点的情况下,Kragle使用两个基因组区域(其可能使用分裂读段信号或不一致的配对物信号被标识)来获取提供信息的读段。然后,募集组件收集具有与到感兴趣的区域(一个或多个)的最初比对映射的配对物中的至少一个的读段。

[0185] 超重叠群组装器(supercontiger)

[0186] 超重叠群组装器由三个组件组成:堆积关系组件(例如,读段-读段比对器)、重叠群组装组件和超重叠群组装组件。三个组件从募集组件获得的募集读段对的集合开始,并产生单体型序列(超重叠群)的集合。超重叠群由以正确定向链接在一起的一个或多个重叠群组成,并被排序以产生组装的感兴趣的基因组区域的单体型序列(例如,可能被中断的单体型序列)。

[0187] 使用两个一般范式进行重叠群组装:基于de Bruin图(Idury RM,et al.,Pevzner PA,et al.)和依赖于重叠图的重叠布局共有序列(Myers EW,et al.,(2005))。Kragle使用了读段平铺策略,该读段平铺策略类似于但不同于Celera组装器(Myers EW,et al.,(2000))中构建“unitigs”的过程,并且精神上类似于但不同于重叠布局共有序列(overlap-layout-consensus)策略。在该示例中,读段之间的所有可能的重叠在重叠群组装开始之前被标识。

[0188] 堆积关系组件

[0189] 堆积关系组件被配置为执行标识这样可能的读段-读段重叠的功能,并且还可以消除所构建的图中的一些冗余。与重叠图(Myers EW,et al.,(2005)) (其消除了完全包含在另一读段的序列中的每个读段)相反,由堆积关系组件产生的读段平铺图包含作为顶点的所有读段,并且边缘表示读段-读段的重叠。仅由映射的配对物执行局部de-novo组装和募集读段对的优点是每个读段的定向(链)是已知的并且读段平铺图不需要表示针对每个读段的两个链的可能性。因为未探索备选的读段定向,所以这样简化了读段平铺图和组装任务。

[0190] 读段平铺图构建分两步执行。第一步标识超过得分阈值的所有读段-读段重叠。第二步修剪平铺图中的边缘,以仅保留用于构建重叠群的完整平铺所需的重叠的最小集合。

[0191] k-mer哈希策略被用于加快读段-读段重叠的标识。每个读段被分解为所有可能的k-mer的集合(例如,默认k-mer大小为50),并且所有读段-读段对被快速筛选用于k-mer的匹配集合。如果找到匹配,如果比对得分超过预定的得分截点,并且如果对应的k-mer位置并不暗示读段-读段比对中的任何插入或缺失(in/del),则存储读段平铺图中的两个读段之间的推断的重叠。比对得分被计算为比对软件(Smith et al.,(1981)、*Smith et al.,(1981)、Needleman et al.,(1970))中常规的匹配和错配得分的和。只有在对应的k-mer位置暗示比对中的插入或缺失时,才计算完整的读段-读段比对。然而,针对启动或延伸插入和缺失的惩罚被设置得足够高以禁止它们,因此所得到的读段比对总是插入/缺失(in/del)自由的。再次,如果得到的全局比对得分超过截点,则重叠被存储在读段平铺图中。

[0192] 作为读段-读段重叠比对中不允许插入/缺失(in/del)的结果,通常从组装中排除具有in/del序列错误的读段。这样的错误是相当罕见的,并不会显著降低读段覆盖度。然而,插入/缺失自由的(in/del-free)重叠简化并显著加速了重叠群组装。注意,真实插入/缺失(in/del)变体(与参考序列相关)在读段之间正确比对,因为源自这种单体型的所有读段都包含变异,因此它们的重叠被正确地比对。

[0193] 一旦完整的读段平铺图被构建,就会修剪它以消除不必要的和可能的错误重叠。在修剪之后,对于每个读段只保留(例如,存储)最佳得分重叠(一个或多个)和为了保持图中的连接性所必需的重叠(一个或多个)。为了保持图的连接性,每个读段保留最小的数量(默认一个)的最佳得分重叠(其在3'和5'侧的每一个上延伸读段),然而这些连接的读段也必须具有在相同侧上延伸读段的、自己的最小数量的重叠。例如,在具有默认最小数量的一个连接的情况下,如果读段具有与另一读段(在3'侧上扩展读段)的至少一个重叠,并且该扩展读段本身具有与另一读段(在3'侧上扩展该扩展读段)的至少一个重叠,则认为在3'侧上满足读段的连接(例如,参见图2)。然后如果在过程开始时具有这样的连接,则读段维持其在图形中的连接性,并且可以从3'和5'侧的重叠边缘到达。因此,由不具有覆盖间隙的连续读段堆积支持的任何序列都可以利用读段连接维持不中断的路径。

[0194] 剪枝算法遍历所有读段并维持针对每个读段所保留的重叠列表。在每次迭代中,对于每个读段,除非读段已经满足在每一侧上所需的最小数量的连接(例如,在一侧上连接到读段,其中读段在同一侧上也具有(多个)连接),否则过程选择在3'和5'侧上延伸读段的最佳的得分重叠(或多个重叠,如果存在具有相同得分的多个重叠)。当每个读段在每一侧上具有所需的最少数量的连接时,剪枝迭代终止。如果读段耗尽所有重叠,则其不需要满足所需数量的连接。这样的读段很可能是位于重叠群端部处的读段,或具有太多的测序错误、与其他读段没有任何重叠的读段。在终止迭代之后,算法将删除不在保留边缘列表中的所有重叠边缘。应该注意,如果重叠读段中的任何一个认为需要该连接,则该重叠边缘被保留。例如,如果存在跨越多态性的两个读段(每一个在相对的单体型上),并且正好在多态位置外部的读段具有与两个多态读段的重叠,则外部的读段可以保留与多态读段中的仅一个的重叠来满足其连接,然而另一个多态读段可能需要与该读段的连接以维持其自身的连接性,因此在多态性外部的读段保留两个重叠(图3)。这确保了从所有单体型特定的读段到达多个单体型所共有的序列延伸的读段。

[0195] 重叠群组装组件

[0196] 重叠群组装组件使用读段平铺图来收集重叠读段并扩展图形中重叠读段的路径。每个重叠群组装从单个读段开始。重叠群组装首先组装从组装器试图识别的区域(例如,感兴趣的基因组区域)的3'侧和5'侧挑选的读段启动的两个重叠群。每个重叠群仅使用每个读段一次,但不同的重叠群可以共享读段。重叠群建立过程通过在遇到多态位置时分裂现有的重叠群来创建新的重叠群。一旦完成两个初始的重叠群和它们的分裂重叠群,重叠群组装器检查未在任何重叠群中使用的读段的集合。如果在未使用的读段中发现大于截点(计算为已组装的重叠群的平均覆盖深度的百分比,默认值为10%)的读段的连接集群,则从集群中的一个读段开始建立新的重叠组。开始新的重叠群直到不存在大于截点的未使用的读段集群。

[0197] 重叠群建立器保留在重叠群中使用的所有读段的踪迹并将它们分成两组:(在读

段平铺图中使用重叠)已募集了其重叠读段的读段以及尚未被募集的读段。建立器还保留3'和5'的“所有募集的”边界(即,边缘)。在尚未被募集的第一个读段的起始之前,“所有募集的”边界被设置在3'侧上(在5'侧上等同)。因此,这些边界在不断增加的重叠群中设置了不会被新募集的读段改变其读段构成的间隔(例如,前进位置)。因此,从“所有募集的”间隔内部的读段堆积计算共有序列是安全的。

[0198] 通过重复以下三个步骤进行重叠群的建立:

[0199] (1) 使用尚未被募集的、并且与当前“所有募集的”3'或5'边界外部的位置1碱基重叠的读段募集新读段(见图4)。被募集的读段必须被包含在募集读段中,或者将读段延伸到外部-远离“所有募集的”边缘。

[0200] (2) 重新计算“所有募集的”边界。如果边界的该侧上的任何读段募集了当其被放置在重叠群中时以相同的位置开始的读段,则3'边界可能没有改变。类似地,5'边界可能没有改变。然而,如果新放置的读段留下较大的间隙,则两个“所有募集的”边界中的任何一个可以移位多个碱基。

[0201] (3) 使用“所有募集的”间隔内部的读段堆积来计算新的“所有募集的”间隔边界的共有序列。

[0202] 当不存在未被募集的读段时,重叠群建立迭代终止。如果重叠群建立到达被募集的读段覆盖的区域的端部(或到达覆盖中的间隙),并且边缘上的读段不具有将重叠群扩展到任一侧的任何重叠读段,则会发生这种情况。

[0203] 当共有序列的计算遇到多态位置(即,读段堆积包含两个或更多不同碱基的显著的计数的位置)时,当前的重叠群被分裂成表示两个(或多个)可能的单体型的两个(或多个)。然而,仅当碱基的计数超过该位置处的覆盖度的10%或至少5个读段时,碱基数才被认为是显著的。因此,读段中的随机碱基响应错误将可能不会触发重叠群分裂。此时,还针对单体型依从性检查多态读段。如果重叠群先前已经在一些位置处被分裂,则分裂过程检查读段或其配对物是否与先前的分裂位置重叠,并收集确实重叠的读段的计数(见图5)。如果针对一个多态性碱基的读段(或其配对物)确实与具有显著计数(默认为至少5个)的位置重叠,但是一些其他多态碱基读段和其配对物没有跨过该位置(计数0),则该多态碱基不会用于分裂另一个重叠群,并且这些多态读段及其配对物将从重叠群中被移除(参见图6)。这种设计决策背后的理由是为了防止其单体型已在另一重叠群中分离的多态位置上的分裂。如果多态位置与已将重叠群分裂为两个(或多个)的先前多态位置间隔足够远,并且读段重叠允许从已分裂的单体型再次募集一些读段,则这些读段(其自身或其读段配对物)将不具有与先前的分裂位置的重叠,因为该位置已经被“纯化”以包含仅属于其单体型的读段。因此,这些读段(及其读段配对物)将从该重叠群中被删除,但是将很可能被用于已在先前的分裂位置处分裂的重叠群中。

[0204] 重叠群分裂的过程涉及复制两个重叠群中的读段成员(除了跨多态位置的读段),其中每个重叠群将采用具有其被分配的碱基的读段。将多态读段及其配对物一起分配给其相应的重叠群。由于配对物重新分配可能会消除“所有募集的”间隔的某些部分中的读段覆盖度(这是罕见的),我们需要重新识别每个分裂重叠群中的共有序列,将N分配给覆盖度完全丢失的区域。

[0205] 因为分裂出新的重叠群并建立它们在计算上是昂贵的,所以我们采用了几种检查

来减少几何的重叠群增加:

[0206] 1、如果在其他重叠群中遇到了相同的分裂(利用多态读段的相同集合并且在读段的相同位置处分裂),则分裂不生成新的重叠群(一个或多个)。这些分裂位置对应于“闭合单体型气泡”,并且具有其对应“重复”的这些“未采用”的分裂被保存。随后在“未采用”的分裂之后,通过将所有可能的重叠群序列末端连接到重叠群,这些“未采用”路径被添加。这些末端取自“重复的”分裂重叠群和在此位置后分裂的所有重叠群(见图7)。该标准假设遇到“重复的”分裂之前的潜在不同的读段成员不会因读段重用的约束而导致不同的末端。假设一旦在重叠群中遇到相同的读段堆积,可能的末端相同。

[0207] 2、如果分裂多态读段不能募集任何新的读段以扩展分裂的重叠群,则这该重叠群被标记为“死胡同”,并且不会用于超重叠群的构建。我们观察到,当重叠群遇到具有系统的碱基响应错误的读段的集合时会发生这些分裂。特别是在具有挑战性的领域(如STR区域),这些可能是非常普遍的。

[0208] 3、如果通过重复已在重叠群中采用的分裂检测到图形循环(非常罕见),则该重叠群被标记为“坏的”并终止。这些重叠群也不用于超重叠群的构建。

[0209] 如上所述组装的重叠群被过滤掉重复。这些重复可能来自于从两个读段开始组装相同区域的重复工作—一个从3'侧,另一个从5'侧。然而,两个起始点给了重叠群组装器抵抗贪婪的读段募集过程的缺点(其中读段由重叠群中与其在读段平铺图中有重叠的第一读段募集)的更好的鲁棒性。从两个方向探索图中的路径可能会导致在某种程度上不同的组装序列中的一些情况。

[0210] 超重叠群组装组件

[0211] 先前步骤中组装的重叠群可以跨越Kragle试图识别的整个区域,或者可以在覆盖度下降的地方或者在高读段错误率(通常是系统错误)禁止高得分读段—读段重叠的地方终止。在这种情况下,读段对可用于将重叠群链接在一起以形成超重叠群。如果读段对中的读段之间的间隙允许将它们放置在两个相邻的重叠群中,横跨重叠群之间的覆盖间隙,则这样的读段对可以通知重叠群链接及其定向。

[0212] 超重叠群构建涉及构建将重叠群作为顶点并将所标识的链接作为定向边缘的图。经定向的边缘被记录,其中重叠群对共享具有相同定向的至少最小数量的读段对(默认设置的最小数量为平均重叠覆盖的10%或至少5个读段对)。通过遍历图中所有路径,同时避免循环,从具有0入度的所有顶点开始并以具有0出度的顶点结束(例如,参见图8),重叠群然后被连接到超重叠群。未连接到任何其他重叠群的重叠群(例如,具有0入度和0出度)创建仅具有一个重叠群的超重叠群。

[0213] 单体型分析器

[0214] 由于重叠群和超重叠群构建创建了所有可能的序列排列,所以超重叠群表示可能的单体型序列,其将被组合以在识别系统中生成正确的倍性(即,二倍体)序列假设。单体型分析器启动单体型对象,并执行所有读段到所有单体型序列的重新映射。该重新映射还包括在“识别系统”部分中描述的映射权重的预计算。单体型分析器执行对单体型序列中的错误结点和错误插入的标识。单体型分析器与识别系统分离的主要原因是:在将单体型组合成序列假设之前,允许识别程序(calling program)有机会基于错误结点和错误插入的标识的结果来过滤单体型。在每个单体型对象中列出了所标识的错误结点和错误插入,并确

定了读段支持的必要信息。识别程序可以使用关于支持信息的截点来应用其自身的单体型过滤的严格性。

[0215] 错误结点标识

[0216] 该算法目标在于标识属于基因组不同部分的、并且由于某些序列相似性而参与组装过程的序列之间的结点。序列相似性将允许一些读段加入序列,然而结点将被稀疏覆盖。因此,这些结点可以被标识为序列中的点,其中横跨结点位置的读段对的数量远低于预期。该算法通过计算一些距离(例如,插入尺寸分布范围)外的读段配对物的预期数量并将它们与观察到的计数进行比较来找到可能的错误结点。低观察/预期比的位置然后被标记为可能的错误结点。可以使用卡方检验来确定观察-预期差的显著性,然而,随着对插入尺寸分布中具有合理变异性的位置的覆盖度的增加,这样的p值将变得越来越敏感。

[0217] 该算法使用插入尺寸分布的平均值附近的非中心带来测试错误结点。使用插入尺寸分布平均值附近带的-20%和+80%(例如,对于Illumina读段对文库,当前尺寸为50)的间隔来计算预期计数并用于搜索观察到的计数。该算法单独地处理针对向前读段和反向读段(以反方向的)的观察和预期计数,然后找到比率中的局部最小值。该算法报告报告超过规定比率截点的所有局部最小值。在向前和反向两个方向中执行搜索可以给出结点的算法确认。

[0218] 然而,该算法在某些情况下可能会产生假阳性,因此必须谨慎进行具有错误插入的单体型的过滤。该算法针对被压缩成两个等位基因的读段长度的大纯合重复产生假阳性。在这种情况下,将不会存在跨越压缩的重复序列的任何读段,但预期的计数仍然可能很大。该算法还可能错误地标记高度错误的位置(Illumina测序过程中的系统和相关错误)。高度错误的序列位置将具有与映射读段低的覆盖度(具有太多错误的读段将不映射),从而可能导致预期计数和观察计数之间的差异。

[0219] 错误插入标识

[0220] 错误插入检测算法旨在使用结点处的一些序列相似性来检测在侧翼处与亲代序列连接的外来(或错位)序列的插入。这样的插入可以通过使用错误结点的配对(如上所述)来潜在地被标识,然而,下面描述的专用算法比错误结点算法更加敏感和特异。

[0221] 对于读段对意外地跳过序列的片段的情况,可以通过检查读段配对物对的位置来标识错误插入,其中被插入的序列具有只映射到其中而不映射到其外部的读段对(参见图9)。该算法在前进和反向两个方向上进行搜索。首先找到不重叠的配对物对被分隔开大于插入尺寸分布的平均值的距离的区域。读段中间的中点标记了假设插入的起始和结束。然后,该算法尝试定位完全包含在假设插入内的读段对。如果找到这样的读对,则该算法将插入边缘细化为这些读段的最左边和最右边的碱基。然后,在该间隔内,算法在每个位置处计算未被完全包含的读段对对被完全包含的读段对的污染。如果在任何位置处,外部读段对的污染低于截点,则该区域被报告为可能的错误插入。

[0222] 识别系统

[0223] 由单体型分析器产生的单体型可以被过滤掉错误结点和错误插入,然后被组合以创建单倍体、二倍体或一般而言的任何倍性的序列假设。对于二倍体序列假设,识别系统探索了所有可能的单体型对,包括由相同单体型两个拷贝组成的纯合排列。因此,二倍体假设评估过程的计算复杂度与单体型数量的平方成比例。因此,为了避免专设的单体型和假

设过滤,假设似然性评估必须具有计算效率。Carnevali (Carnevali et al.2012) 描述的统计框架在某些方面被认为是失败的。例如Carnevali的框架未能适应任何等位基因比率(对于嵌合体和癌症基因组)。等式(2)的统计框架被扩展和修改以适应任何等位基因比率(嵌合体和癌症基因组)。这个新框架(例如,参见等式1)也允许针对每个单体型对似然性计算的许多分量进行预先计算,因此将单体型与二倍体假设配对,并且计算它们的似然性仅需要相当微不足道的计算量。

$$[0224] \quad \frac{P(G|\{R\})}{P(G_0|\{R\})} = \left(\frac{N_{G_0}}{N_G}\right)^{N_R} \prod_{\{R\}} \frac{\sum_{\{M\}} P(R|G,M)}{\sum_{\{M\}} P(R|G_0,M)} \quad (\text{等式 } 2)$$

[0225] 其中G表示任何倍性的基因组序列,G₀表示参考基因组序列,或任何其他固定序列假设。{R}表示读段对的集合,并且读段对的数目为N_R。N_G和N_{G0}分别表示基因组G和G₀中的碱基数。M表示读段对中两个读段的映射位置,并且读段对R的所有可能的映射位置的集合由{M}表示。

[0226] P(R|G,M)可以被计算为基因组中给定映射位置的匹配和错配碱基概率(Carnevali et al.2012)与由读段对的两个读段的映射M暗示的插入尺寸的概率的乘积。匹配和错配概率可以从测序平台分配的碱基响应错误率推导出。

$$[0227] \quad P(R|G,M) = P(I_M) * \prod_i P(b_i | G[M(i)]) \quad (\text{等式 } 3)$$

[0228] 其中I_M是由读段对的映射M暗示的插入尺寸,并且概率P(I_M)可以从经验插入尺寸分布获得。乘积 $\prod_i$ 接管读段对中的所有位置i,并且P(b_i|G[M(i)])是基因组G中被映射的参考碱基在被映射的位置i处生成读段对中的被映射碱基b_i的概率。这些可以使用测序仪提供的估计碱基响应错误概率来计算。

[0229] $P(b_i | G[M(i)]) = (1-\epsilon_i) \delta[b_i, G[M(i)]] + (\epsilon_i/3) (1-\delta[b_i, G[M(i)]])$ (等式4) 在该等式中,如果Kronecker符号δ的两个参数相同,则其定义为1,否则为0,并且ε_i是读段对中位置i处针对碱基响应的错误概率。G[M(i)]是基因组G中、读段对中的碱基b_i使用映射M被映射到的碱基。

[0230] 即使对于小基因组,可能映射的整个集合上的读段概率P(R|G,M)的和也可能是计算上难以处理的。两个读段的所有可能位置的组合可能太多而无法枚举。然而,仅存在到基因组的很少映射将对和作出显著贡献。因此,所有可能的映射之上的和可以被划分成“良好”映射之和(称为映射权重W(R,G)),而对和的剩余的小的贡献可以被聚合到一个小项α:

$$[0231] \quad \sum_{\{M\}} P(R|G,M) = \sum_{\{M_{good}\}} P(R|G,M) + \alpha = W(R,G) + \alpha \quad (\text{等式 } 5)$$

[0232] 在局部de novo测序中,可以使用α捕获在被组装区域之外的映射权重的概率和。一些读段,特别是包含低复杂度序列的读段,可以具有到参考基因组的多个映射,因此它们在感兴趣区域的位置是不确定的,并且应该伴随较大的α,这有效地降低了它们对似然比的贡献。

[0233] 上述等式2因此可以被扩展和修改为适应具有不同比率的等位基因的以下等式1:

[0234]

$$\frac{P(G|\{R\})}{P(G_0|\{R\})} = \prod_{\{R\}} \frac{\sum_{\{A_G\}} \frac{1}{N_{A_G}} F_{A_G} (W(R, A_G) + \alpha)}{\sum_{\{A_{G_0}\}} \frac{1}{N_{A_{G_0}}} F_{A_{G_0}} (W(R, A_{G_0}) + \alpha)} \quad (\text{等式 } 1)$$

[0235] 其中G是针对预定倍性的基因型序列,G₀是参考序列,{R}是读段配对物对R的集合,N_{AG}是基因型序列G中的等位基因A_G的数量,N_{AG0}是在参考序列G₀中等位基因A_{G0}的数量,以及F_{AG}是基因型序列G中等位基因A_G的分数,F_{AG0}是参考序列G₀中等位基因A_{G0}的分数,W是读段对映射权重,α是映射概率常数。在一些实施例中,根据等式(1)的推导或等式(1)的变形来确定基因型似然比。在二倍体基因组中,两个等位基因分数各自为0.5。

[0236] 使用上面的等式,W(R,A_G)可以表示为以下等式6:

$$W(R, A_G) = \sum_{\{M_{good}\}} \left(P(I_M) * \prod_{i \in M} \left((1 - \varepsilon_i) \delta[b_i, G[M(i)]] + \right. \right. \\ \left. \left. \left(\varepsilon_i / 3 \right) (1 - \delta[b_i, G[M(i)]]) \right) \right) \quad [0237]$$

[0238] 如上所述,α捕获在被组装的区域中的映射位置M_{good}的集合之外的可能映射的映射权重的和。每个映射程序提供映射质量值(mapQ),映射质量值是错误的比对的phred变换概率(mapQ=-10*log₁₀(P_{wrongAlignment}))。这个概率不是所有可能的备选映射位置处的映射权重的和,因此开发了缩放,将mapQ的转化近似为映射权重(W_{mapQ})的和。缩放的目的在于使得对于高mapQ值(即,BWA中为60),W_{mapQ}的贡献可忽略,并且对于mapQ=0,W_{mapQ}具有贡献1(大的值)。因此,等式5中的α可以进一步扩展为:

$$[0239] \quad \alpha = W_{mapQ} + \alpha_R \quad (\text{等式7})$$

[0240] 其中对于非常高的mapQ值(当W_{mapQ}可忽略时),α_R是捕获位置外部的剩余映射权重的非常小的常数。

[0241] W_{mapQ}使用以下等式获得:

$$[0242] \quad W_{mapQ} = 10^{mapQ * f / -10} \quad (\text{等式8})$$

$$[0243] \quad f = 1 \text{ IF } mapQ \leq 30 \text{ ELSE } \left(1 + \frac{mapQ - 30}{30} * \left(\frac{p_{max}}{60} - 1 \right) \right) \quad (\text{等式9})$$

$$[0244] \quad p_{max} = -10 * \log_{10}(\alpha_R) + 100 \quad (\text{等式10})$$

[0245] 由于W_{mapQ}是针对读段对计算的值,所以上述等式中的mapQ值是读段对中两个读段的最大值。因此,如果读段对中的读段中的一个在组装区域中具有令人信服的唯一映射,则假设第二读段也唯一地属于该组装区域。

[0246] 从上面的等式可以看出,mapQ变换的操作范围是在30和60的mapQ值之间。对于小于30的值(低质量读段),W_{mapQ}的值可以与良好映射的映射权重相当。这有效地降低了等式1中读段对对总体概率的贡献。另一方面,对于值接近60的mapQ(高质量映射),W_{mapQ}小于α_R,因此可忽略,并且α≈α_R。

[0247] 上述等式6-等式10允许针对每个读段对和每个等位基因(超重叠群)预计算读段对权重,并且可以在假设似然性计算期间重新调用该值。这种预计算显著地加快了似然比的计算,从而允许对从数千个超重叠群构建的所有二倍体假设进行实时评估。在一些实施例中,为了便于计算针对所有读段的读段权重,识别系统将所有读段重新比对到所有超重叠群。

[0248] 似然比计算要求识别系统将所有读段映射到参考基因组或选择序列假设中的一

个作为参考。默认第一假设为参考,然后相对于第一假设计算所有似然性。任何两个假设的对数似然比可以相减来获得它们的相对似然比。这允许计算最靠前假设与任何其他假设的似然比,以获得置信度测量。

[0249] 本文参考的每个专利、专利申请、出版物和文献的整体内容通过引用并入本文。上述专利、专利申请、出版物和文件的引用不是承认上述任何内容是相关的现有技术,也不构成对这些出版物或文献的内容或日期的任何承认。

[0250] 在不脱离本技术的基本方面的情况下,可以对前述内容进行修改。虽然已经参考一个或多个具体实施例大致详细地描述了本技术,但是本领域普通技术人员将认识到可以对本申请中具体公开的实施例进行改变,但是这些修改和改进在本技术的范围和精神内。

[0251] 本文示例性地描述的技术适当地可以在没有本文具体公开的任何元件(一个或多个)的情况下实践。因此,例如,在本文的每个实例中,任何术语“包括”、“基本上由...组成”和“由...组成”可以使用其他两个术语中的任一个替代。已经使用的术语和表达用作描述的非限制性术语,并且这些术语和表达的使用不排除所示出和描述的特征或其部分的任何等同物,并且在所要求保护的技术的范围内,可以进行各种修改。除非上下文清楚描述了元素中的任一个或元素的多个,否则术语“一”或“一个”可以指其修饰的元素中的一个或多个(例如,“一个试剂”可以是一个或多个试剂)。本文所使用的术语“约”是指基础参数的10%内的值(即,加或减10%),并且在一串值的开头使用的术语“约”修饰每个值(即,“约1、2和3”是指约1、约2和约3)。例如,“约100克”的重量可以包括90克和110克之间的重量。此外,当本文描述值列表(例如,约50%、60%、70%、80%、85%或86%)时,该列表包括其所有中间值和分数值(例如,54%、85.4%)。因此,应当理解,虽然已由代表性实施例和可选特征具体公开了本技术,但是本领域技术人员可以采用本文所公开的概念的修改和变化,并且这些修改和变化被认为在本技术的范围内。

[0252] 本技术的某些实施例在所附权利要求书中阐述。

[0253] 参考文献

[0254] Carnevali, P., et al. 2012. Computational Techniques for Human Genome Resequencing Using Mated Gapped Reads. *J. Comput. Biol.* 19, 279-292. Idury RM, Waterman MS (1995) *J. Comput. Biol.* 2 (2) : 291-306.

[0255] Pevzner PA, Tang H, Waterman MS (2001) *Proc. Natl. Acad. Sci. U S A.* 98 (17) : 9748-53) .

[0256] Myers EW (2005) *Bioinformatics* 21:Suppl 2:ii79-85) .

[0257] Myers EW, et al., (2000) *Science* 287 (5461) : 2196-204) .

[0258] Smith TF, Waterman MS (1981) *J. Theor. Biol.* 91 (2) : 379-80.

[0259] *Smith TF, Waterman MS (1981) *J. Mol. Biol.* 147 (1) : 195-7.

[0260] Needleman SB, Wunsch CD (1970) *J. Mol. Biol.* 48 (3) : 443-53) .

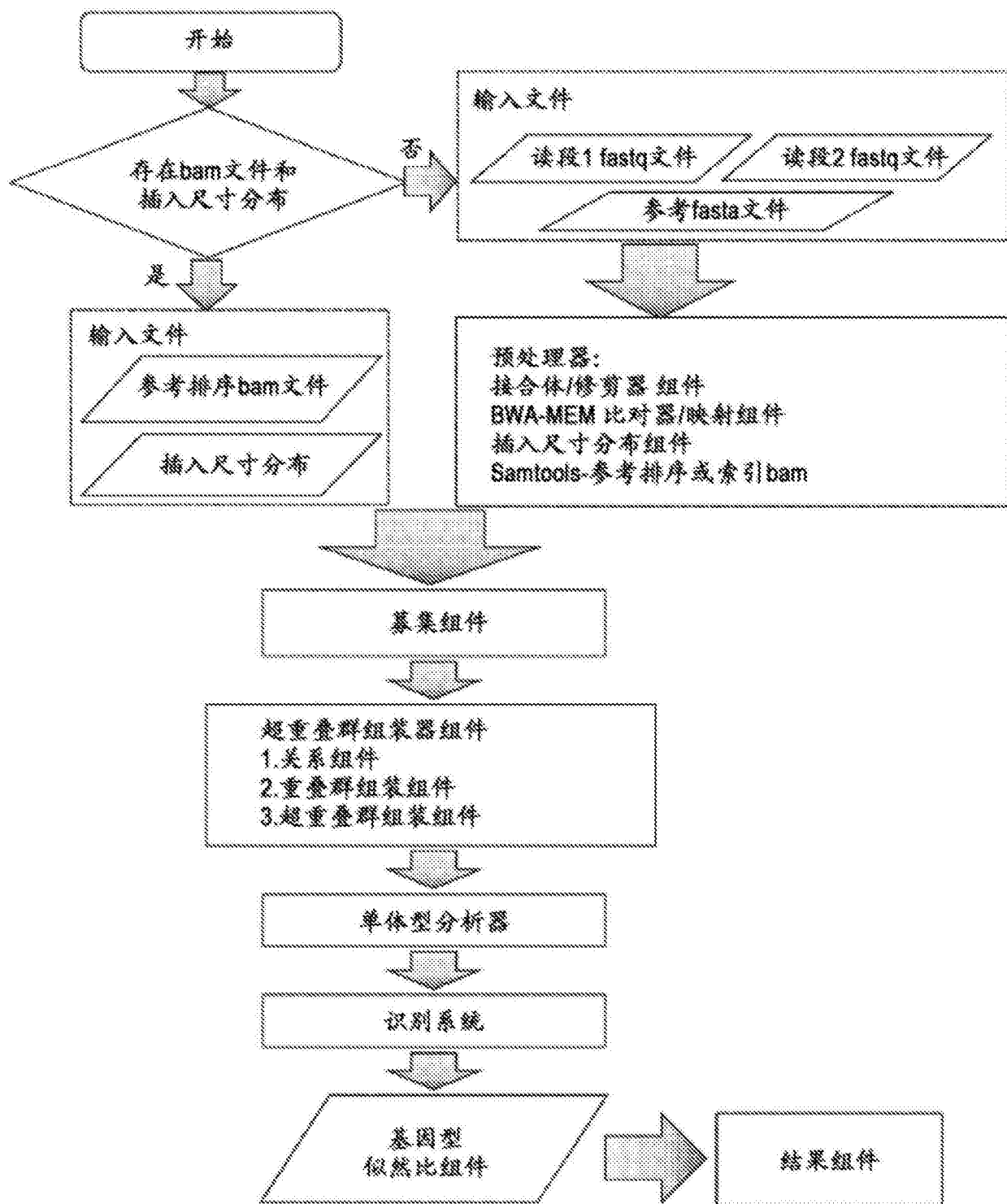


图1

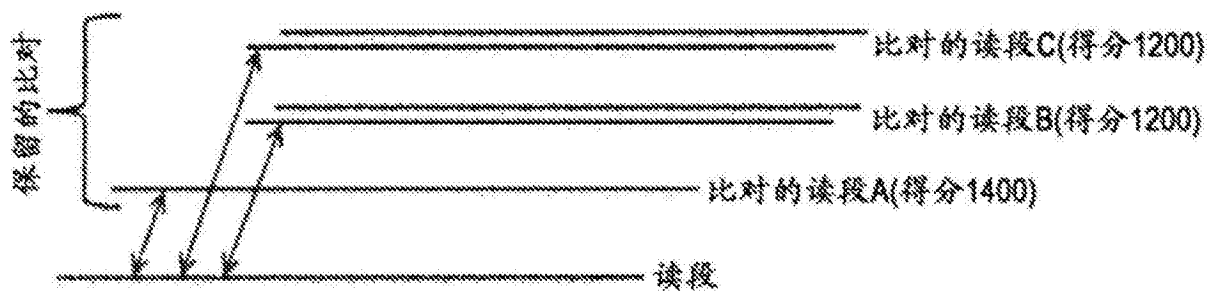
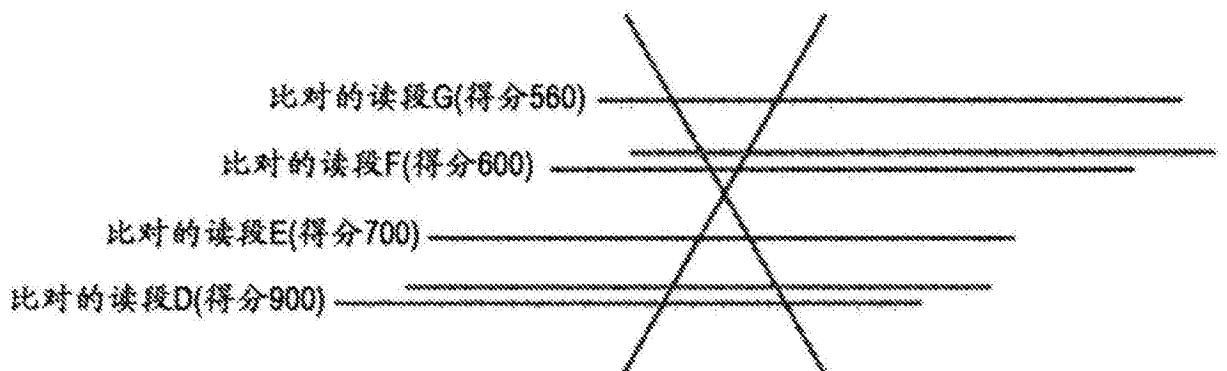


图2

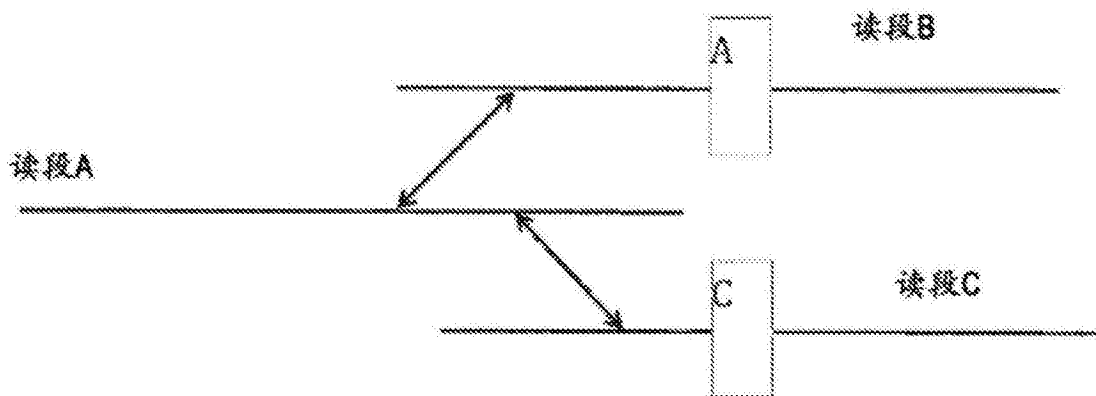


图3

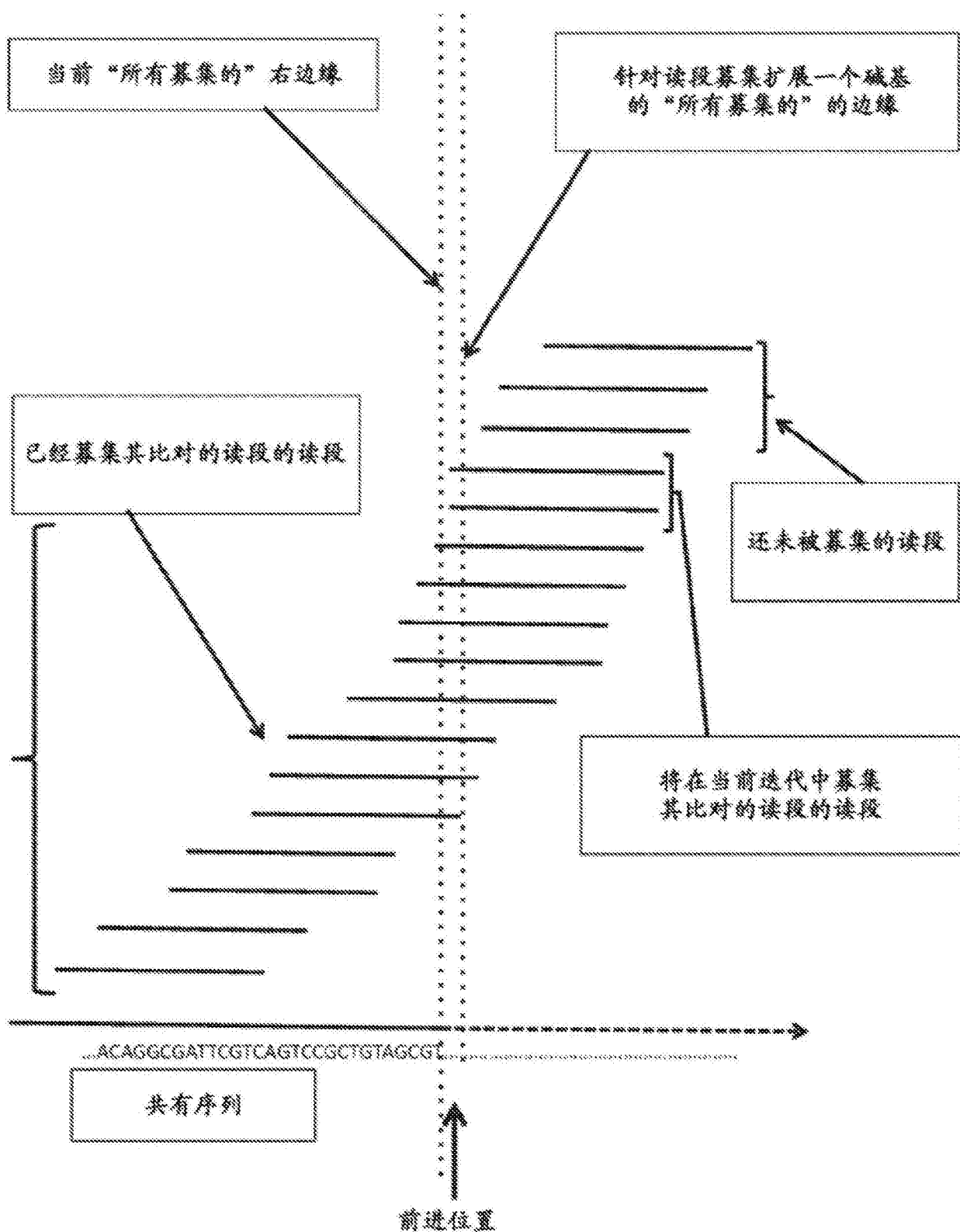


图4

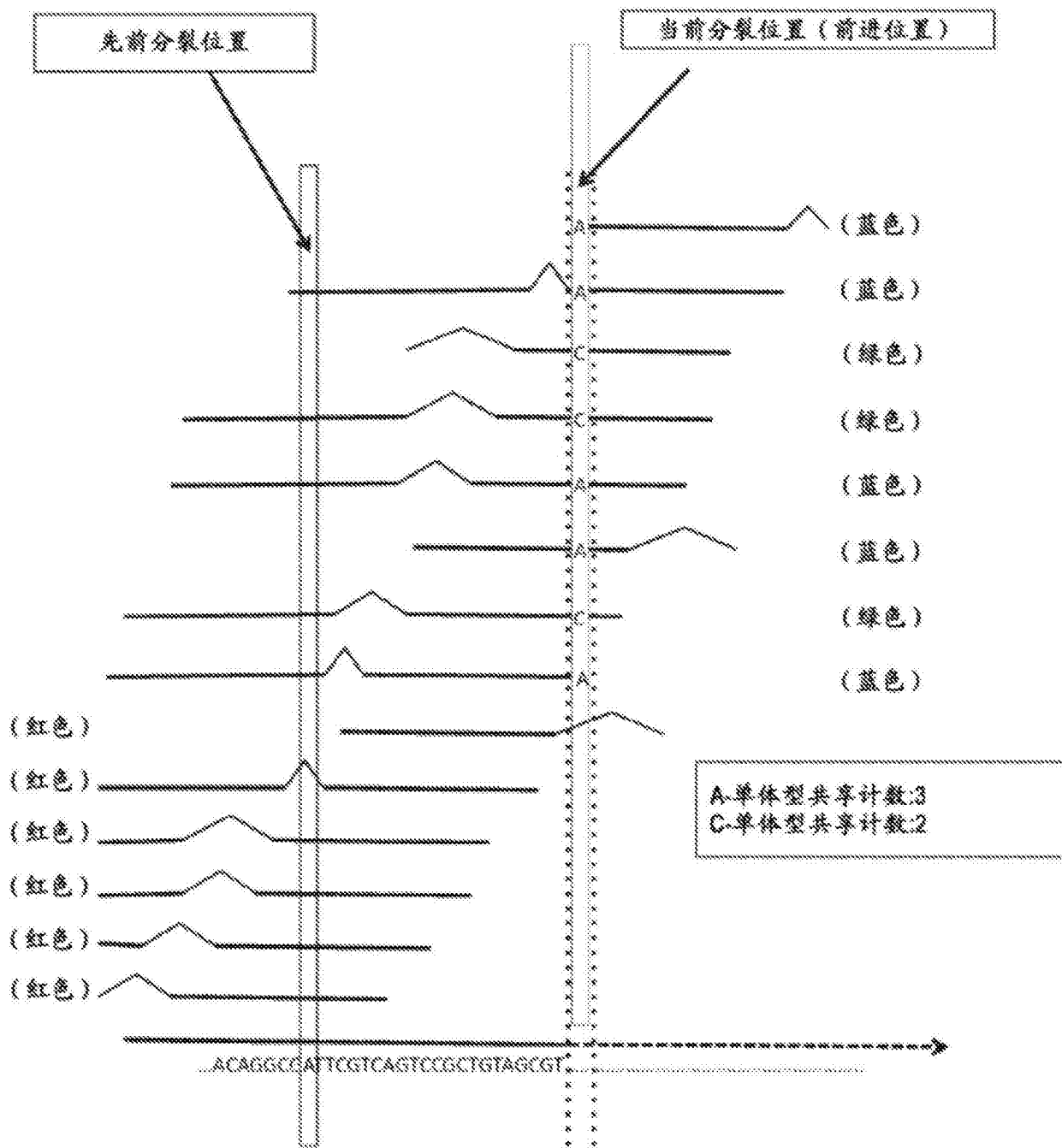


图5

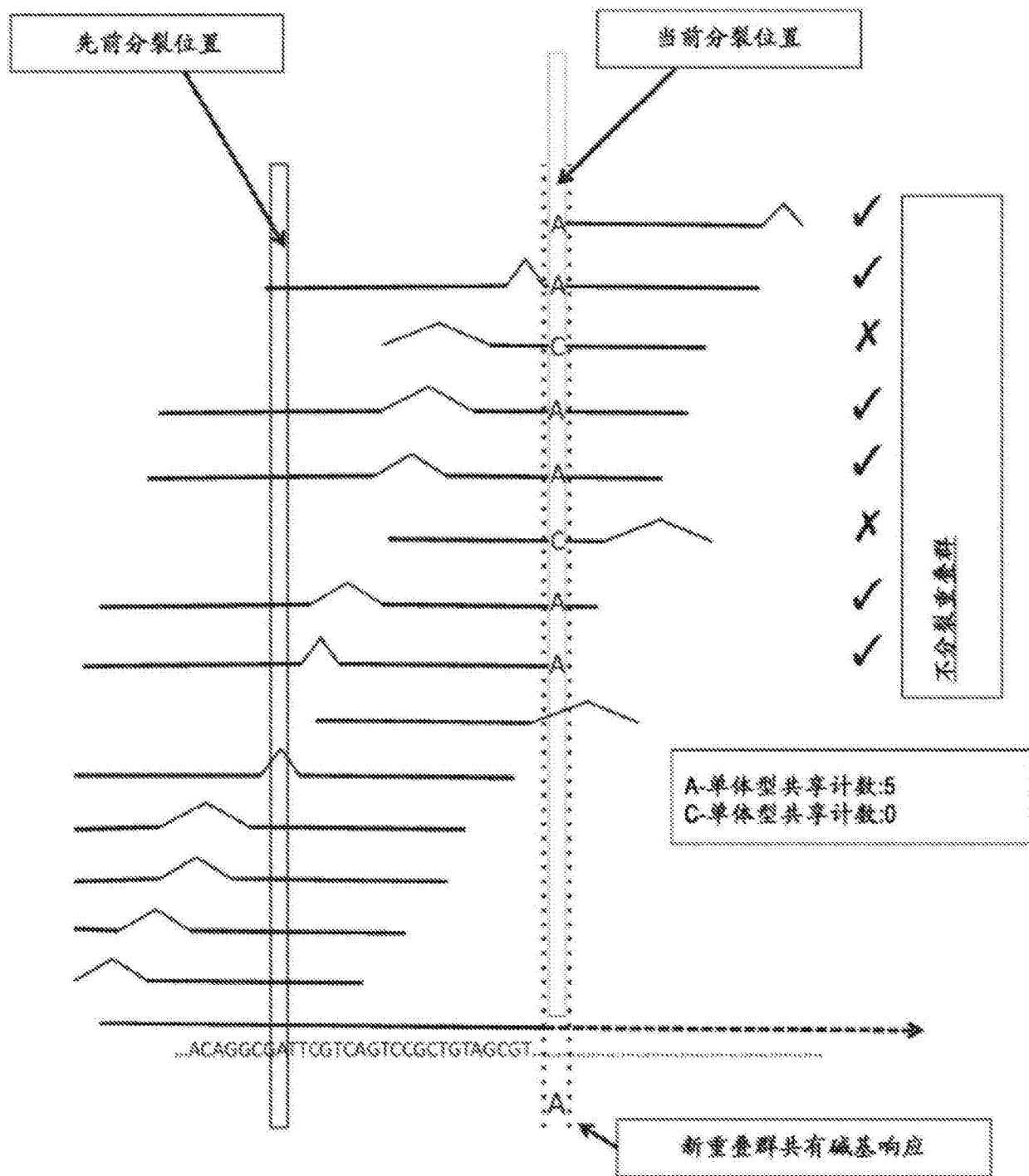


图6

“**新地**”

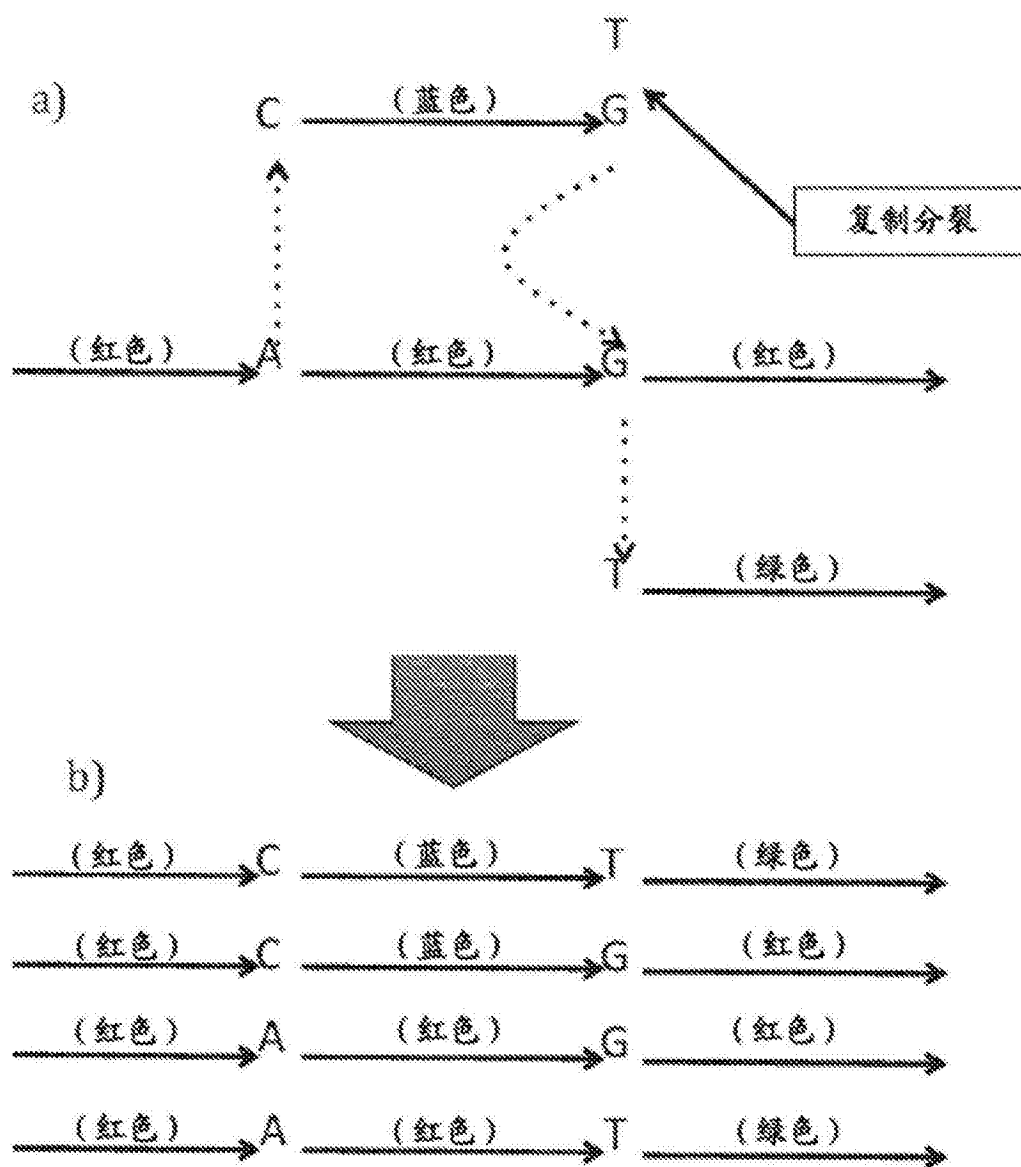


图7

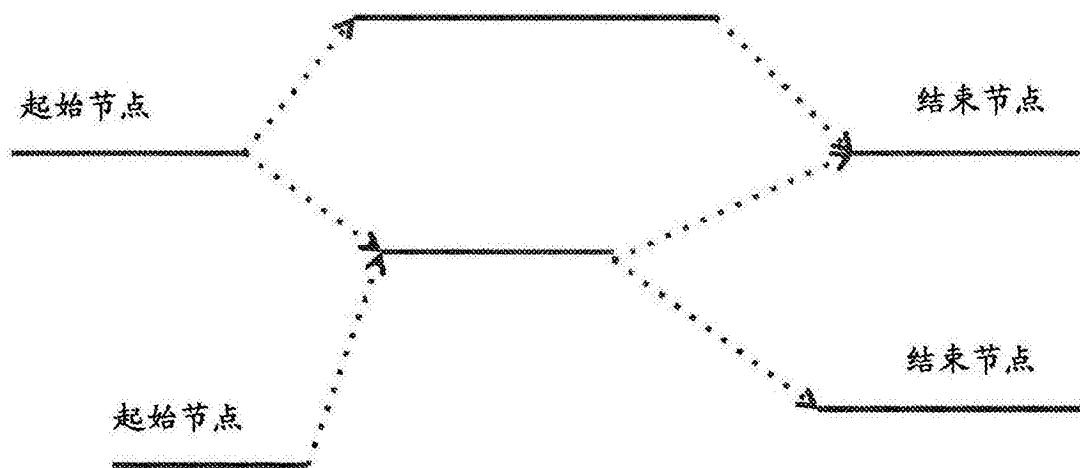


图8

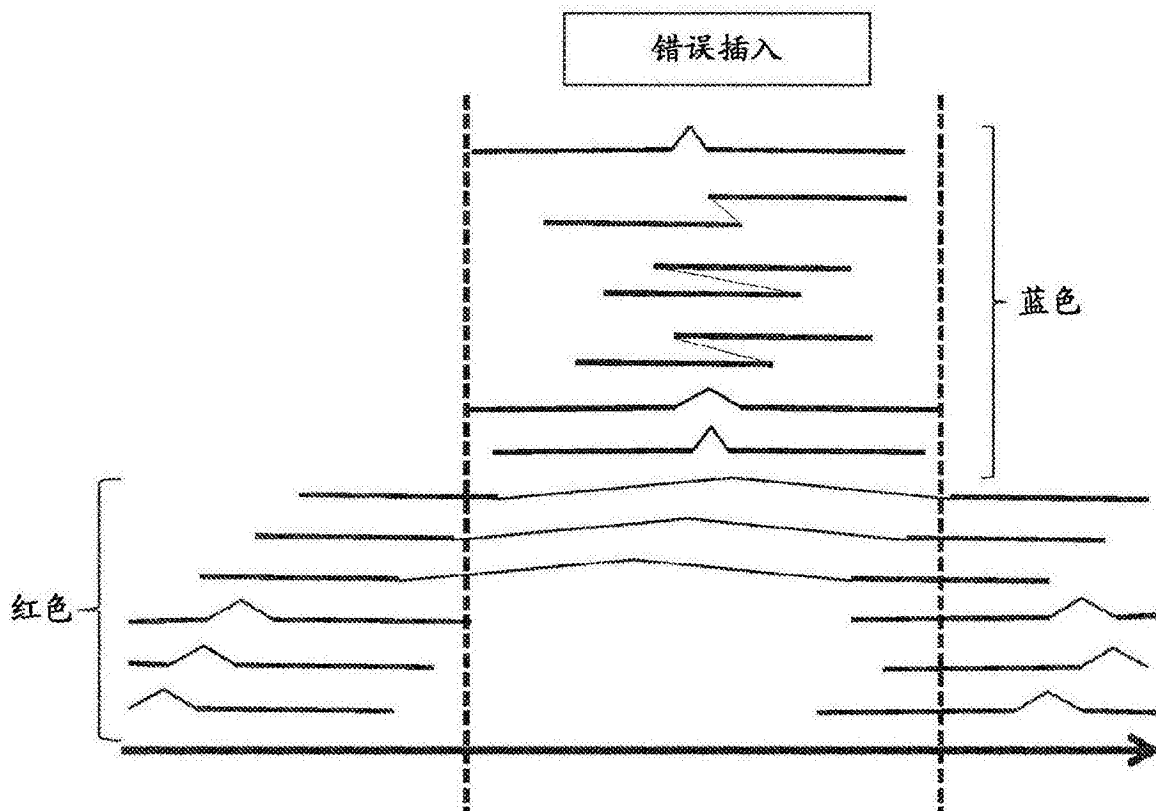
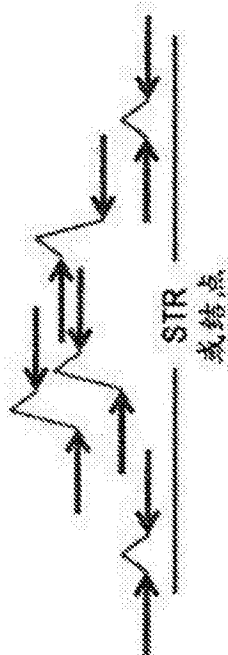


图9

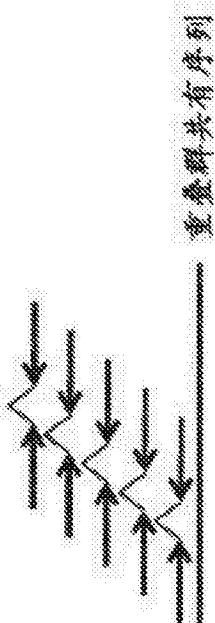
Krage=通用de novo组装机和基因型识别系统

工作流程概述:

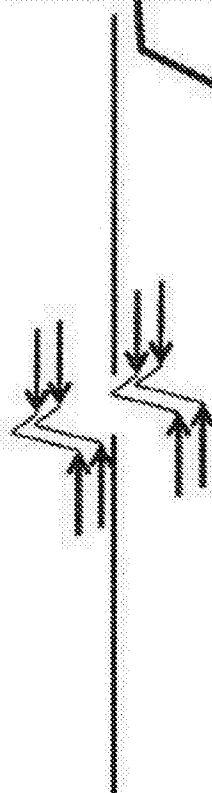
由映射的配对
物募集读段对



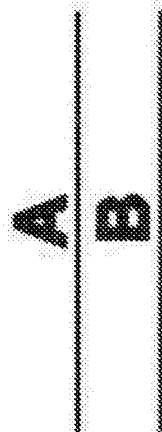
平铺重叠群



重叠群被到
超重叠群



评估基因型假设



二倍体基因型识别

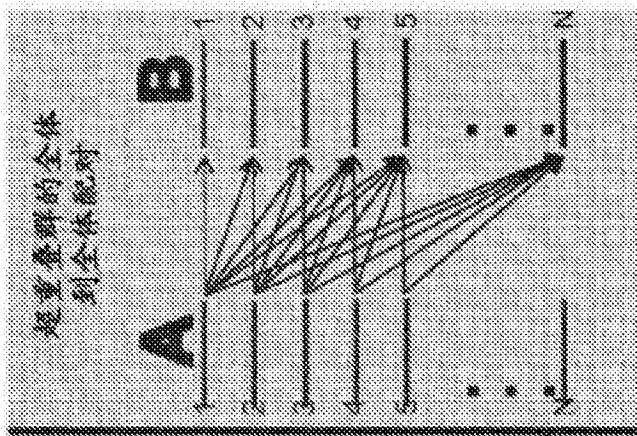


图10

重叠群：读段-读段比对

不允许间隙！！

如果可能避免运行比对算法（算法复杂度 $O(N^2)$ ）

1.针对每个读段，聚集所有子序列及其位置：

- 1.子序列（长度50）被分配唯一的ID
- 2.针对读段中的每个子序列ID，创建位置列表

2.针对读段的每个对，接受具有得分>阈值的比对：

- 1.如果其不具有任何共有子序列，则跳过
- 2.如果共有子序列暗示单个比对，计算得分和cigar
- 3.否则调用`global_align()`来创建完全的读段-读段比对（具有得分和cigar）

3.过滤比对，同时确保读段-读段图形的连接性：

- 1.如果读段完美重叠，保留比对
 - 2.对于每个读段，保留向左向右延伸其的最高得分比对：
 - 1.如果最佳比对的读段的当前集合自身不具有在同一方向上延伸他们的比对，重复添加下一个最佳比对
 - 2.读段-读段连接被对称建立（ $A \rightarrow B \Leftrightarrow A \leftarrow B$ ）
- =>读段-读段比对创建最小连接的有向图（不确保无循环）

图11

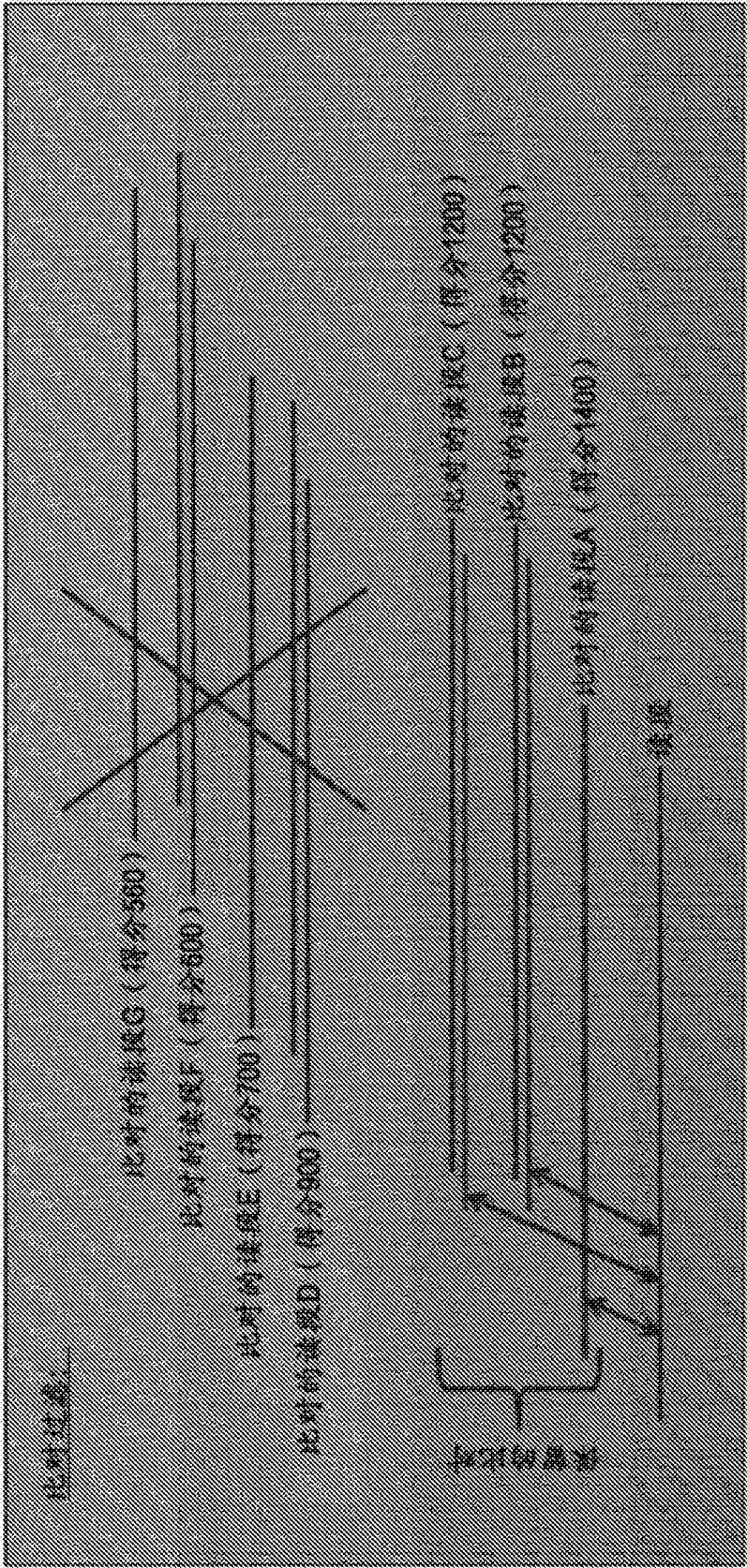


图12

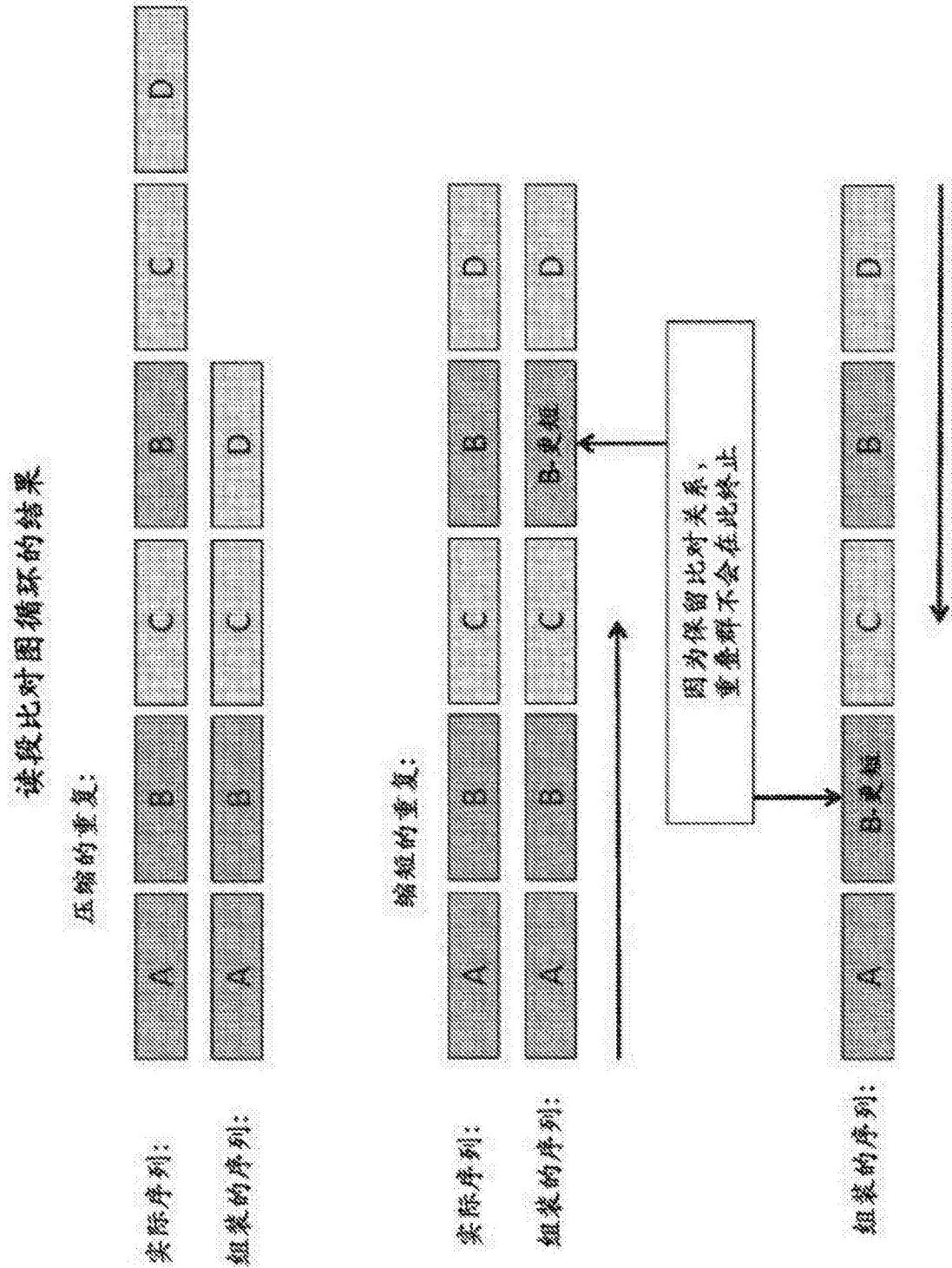


图13

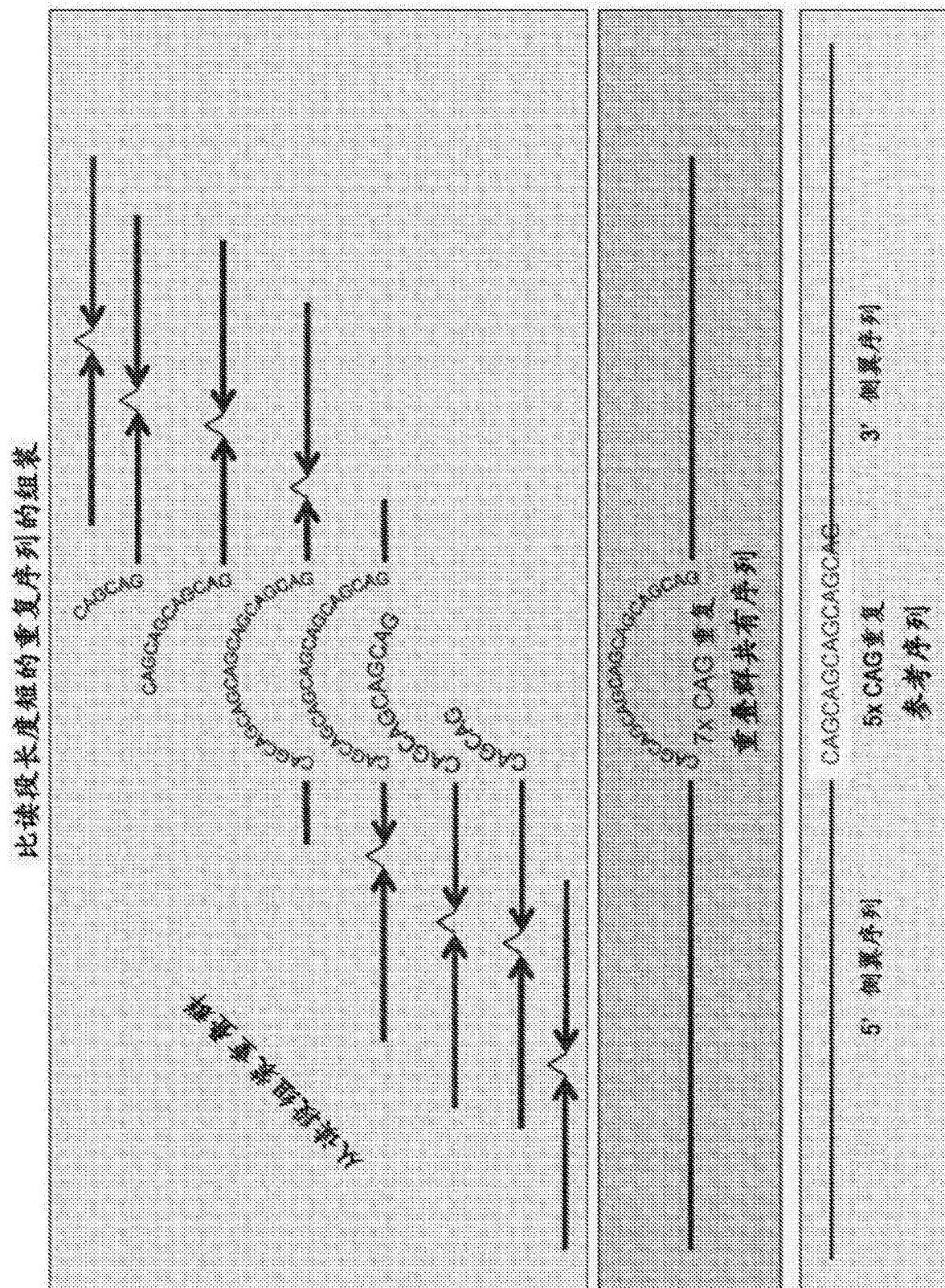


图14

平铺重叠群

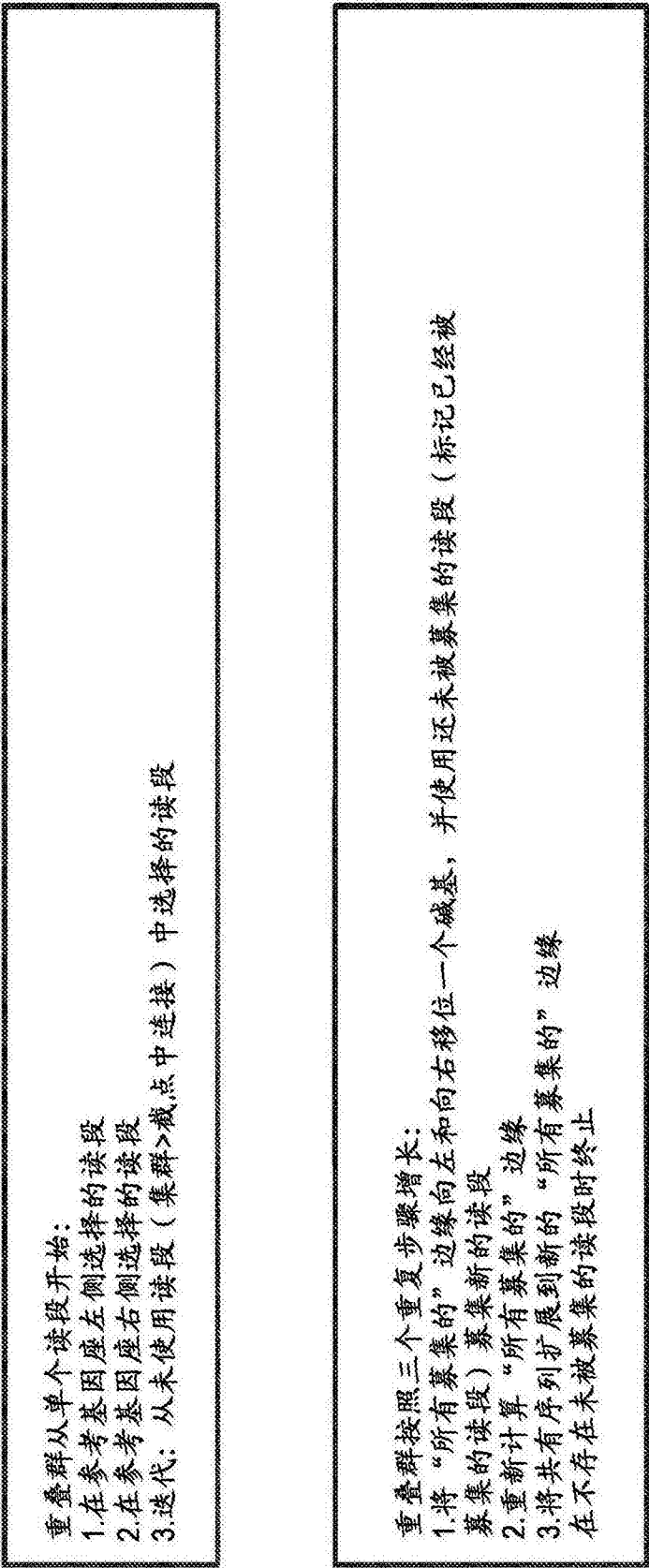


图15

扩展重叠群

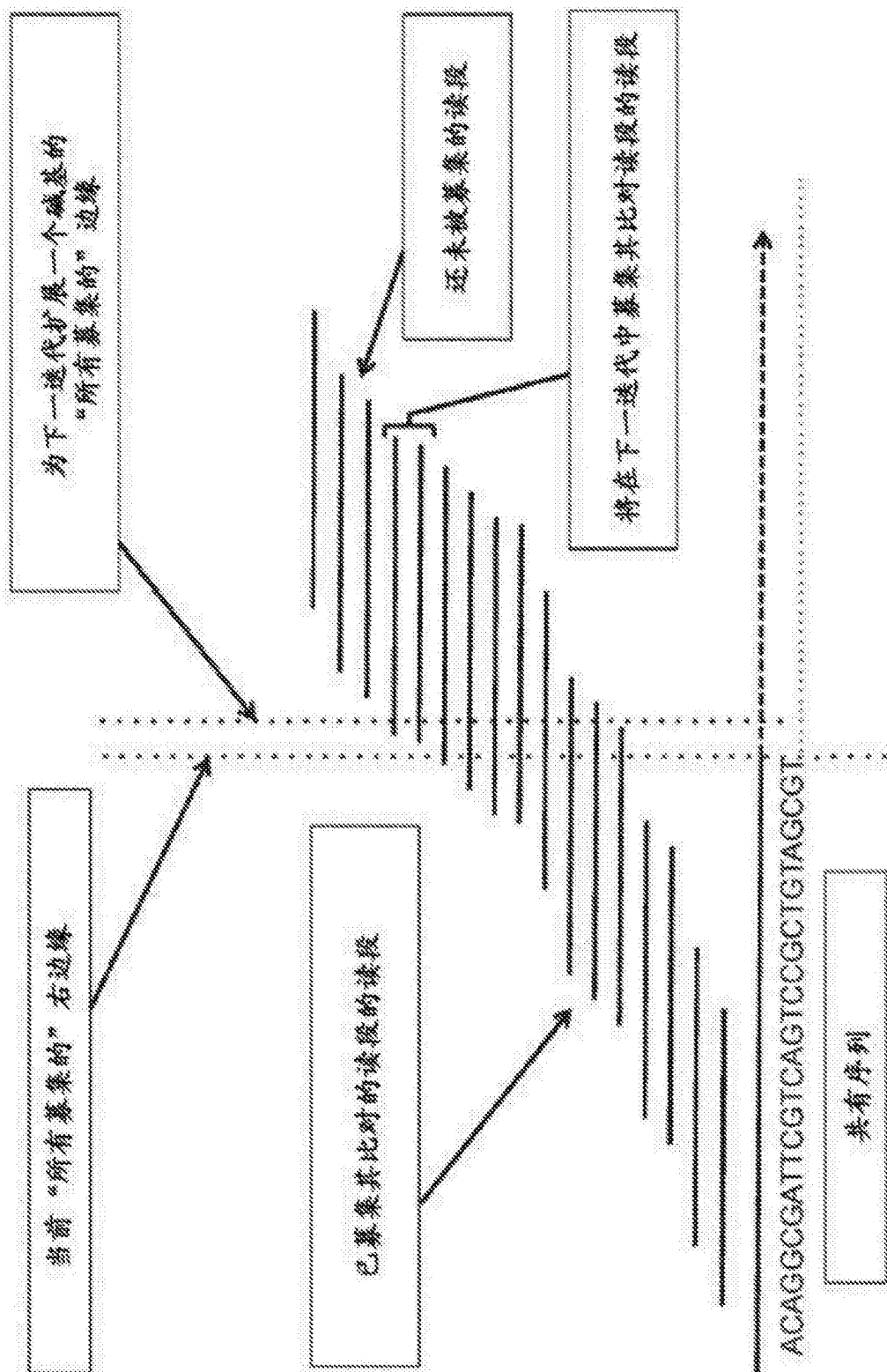


图16

平铺重叠群：分裂

当共有序列遇到多态性（具有不同碱基的N个读段的位置； $N = \max(5, \text{读段的}10\%)$ ）时，单个重叠群可以分裂成若干重叠群；

1、将共有碱基确定为多数单体型纠正的碱基计数

2、通过使备选碱基读段与共有碱基读段分离而将重叠群分裂为两个或多个

- 如果在其他重叠群中遇到相同的分裂，则该分裂不生成新的重叠群（在读段的相同位置处分裂读段的相同集合）。
- 这些分裂位置被追踪并且“未采取”的路径之后被添加。
- 如果检测到图循环，则终止重叠群，并将其标记为“坏”
- 如果分裂出的读段不能募集任何新的读段，则该重叠群被标记为“死胡同”

图17

完成重叠群

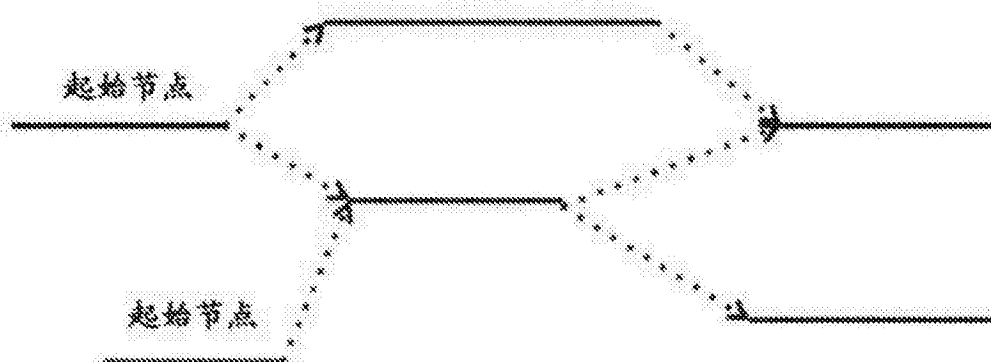
- 
1. 创建重叠群的最后集合:
 1. 跟随复制分裂并创建新的重叠群
 2. 删除复制重叠群序列
 3. 删除遇到循环的“坏”的重叠群
 4. 删除“死胡同”重叠群

图18

超重叠群

创建超重叠群:

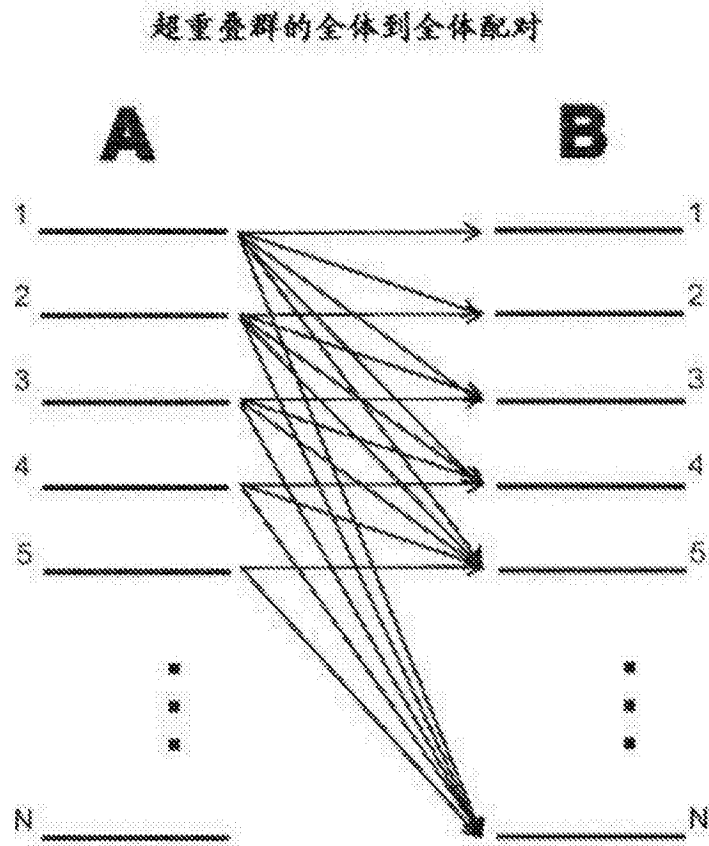
1. 创建重叠群连接的有向图
2. 发现图的起始节点 (具有0个内连接的重叠群)
3. 从起始节点跟随图中的所有路径
 - 如果检测到循环, 终止路径并删除该超重叠群
4. 添加在图中未使用的重叠群作为“单个重叠群的超重叠群”



图产生5个超重叠群

图19

二倍体假设



二倍体假设的数量: $(N^2 - N) / 2$

$\Rightarrow$ 需要非常高效的假设评估

图20

基因组模型

$$\frac{P(G|\{R\})}{P(G_0|\{R\})} = \frac{P(\{R\}|G) * P(G)}{P(\{R\}|G_0) * P(G_0)} = \frac{P(G)}{P(G_0)} * \prod_{\{R\}} \frac{P(R|G)}{P(R|G_0)}$$

插入惩罚

$$= \left(\frac{N_{G_0}}{N_G} \right)^{N_R} * \prod_{\{R\}} \frac{\sum_{\{M\}} P(R|G, M)}{\sum_{\{M\}} P(R|G_0, M)}$$

$$P(R|G) = \sum_{\{M\}} P(M) * P(R|G, M)$$

$$P(M(x, s, a)) = \frac{1}{2N_G}$$

$$\sum_{\{M\}} P(R|G, M) = W(R, G) + \alpha$$

 α 表示剩余的低可能性映射

它是基因组宽的常数

图21

分离等位基因表示的推导

$$\frac{P(G|\{R\})}{P(G_0|\{R\})} = \frac{P(\{R\}|G)*P(G)}{P(\{R\}|G_0)*P(G_0)} = \frac{P(G)}{P(G_0)} * \prod_{\{R\}} \frac{P(R|G)}{P(R|G_0)}$$

$G = \{A\}$ (具有任何可能的比率 F_A 的单体的集合)

$$P(R|G) = P(R|\{A\}) = \frac{P(R, \{A\})}{P(\{A\})} = \frac{\sum_{\{A\}} P(R, A)}{P(\{A\})} = \frac{\sum_{\{A\}} P(R|A)*P(A)}{P(\{A\})} = \sum_{\{A\}} P(R|A)*F_A$$

$$P(R|A) = \sum_{Pos} P(R|A, Pos)*P(Pos) = \sum_{Pos} P(R|A, Pos)*\frac{1}{2N_A} = \frac{1}{2N_A} (\sum_M P(R|A, M) + \alpha)$$

$$\frac{P(G|\{R\})}{P(G_0|\{R\})} = \frac{P(G)}{P(G_0)} * \prod_{\{R\}} \frac{\sum_{\{A\}} \frac{1}{N_{AG}} * F_{AG} \left(\alpha + \sum_M P(R|A_G, M) \right)}{\sum_{\{A\}} \frac{1}{N_{AG0}} * F_{AG0} \left(\alpha + \sum_M P(R|A_{G0}, M) \right)}$$

(对于简单的SNP, 该比率为二项式似然比)

图22

STR识别系统Kragle封装器

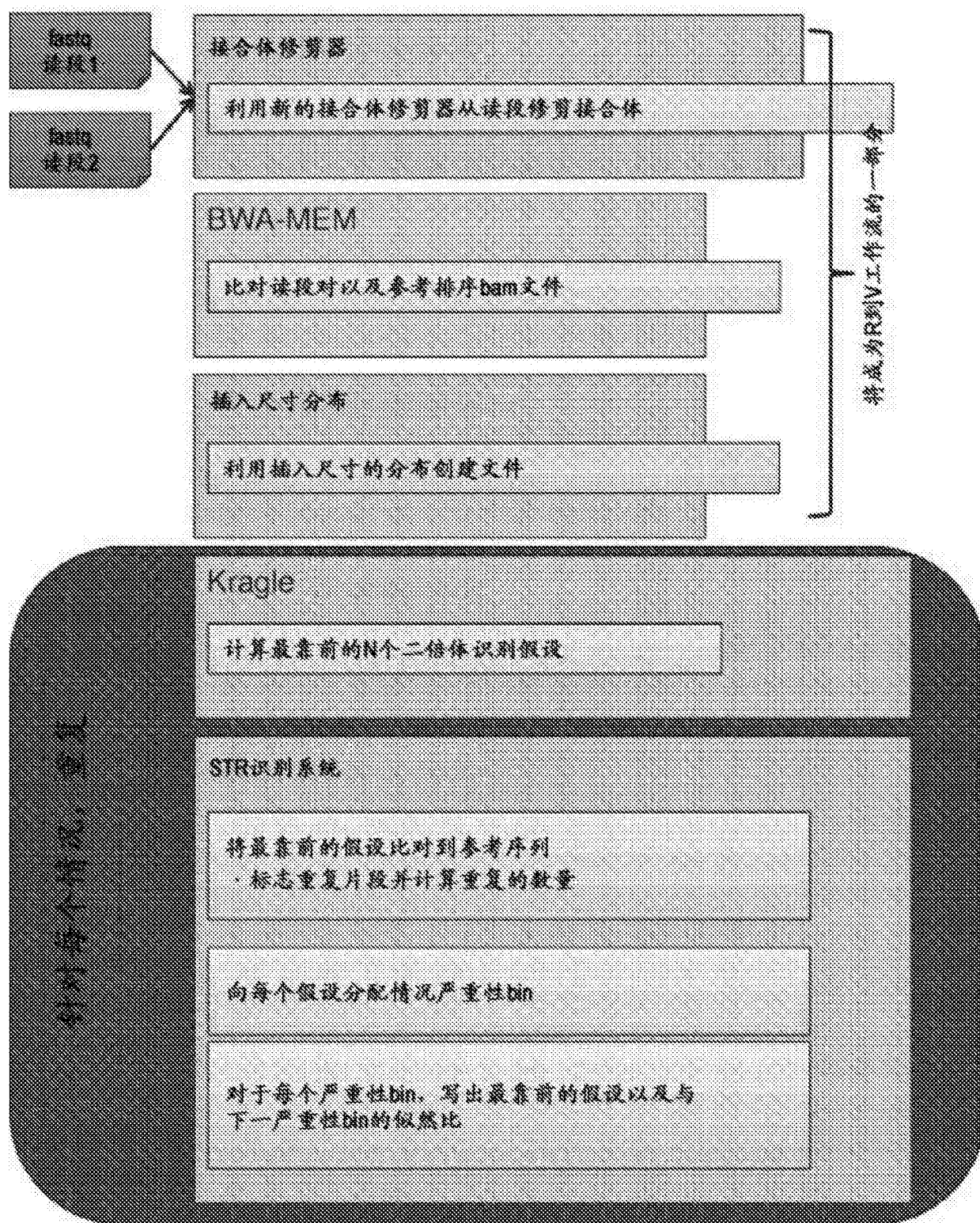


图23

结果

针对20个状况重复完成32个样本
20个状况中的19个可识别 (DM2试验失败)
一个样本由于实验室错误重复失败

特异性及敏感性 (与外部实验室结果相比较):

- 假阳性率 = $1 / ((32 + 31) \times 19) = 0.08\% \Rightarrow$
特异性 = 99.2%
- 假阴性率 = 0% $\Rightarrow$ 敏感性 = 100%

最靠前假设的再现性 (31个样本的技术性复制):

- 准确重复计数的再现性: 587对中64个不同 = 89%
- ± 1 重复再现性: 587对中17个不同 = 97%

图24

针对Kragle的定制的CFTR识别系统前端

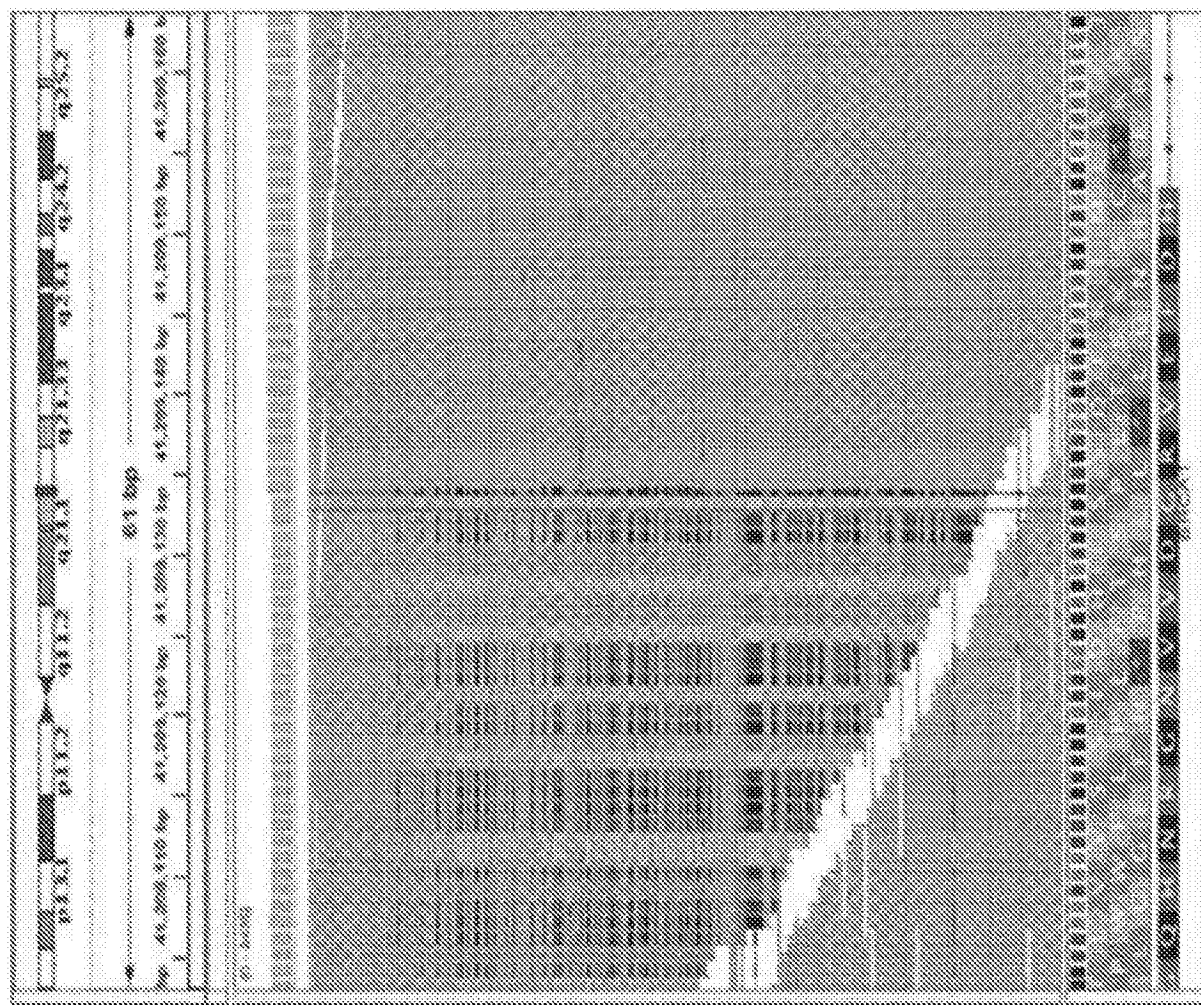
- 类似于通用STR识别系统, 但执行一些特殊过程:

- 稍微不同的侧翼映射以及重复计数
- 该区域中募集的读段包括来自分段复制的畸形的读段
 - 需要一些代码来丢弃这些单体型

- 测试的100个“随机的”临床样本:

- 93个具有与CoalGen相同的识别
- 7个不同, 不过看起来似乎合理

图25



更加具有挑战性的 组装任务

- 针对该真正的BRCA1分裂读段信号，Kragle能够正确地组装结点？

• 输入

- 参考比对读段
- 该区域的坐标
- 被剪掉的区域的坐标

- 从被剪掉的碱基到参考基因组的映射中获得

图26

这儿发生了什么？

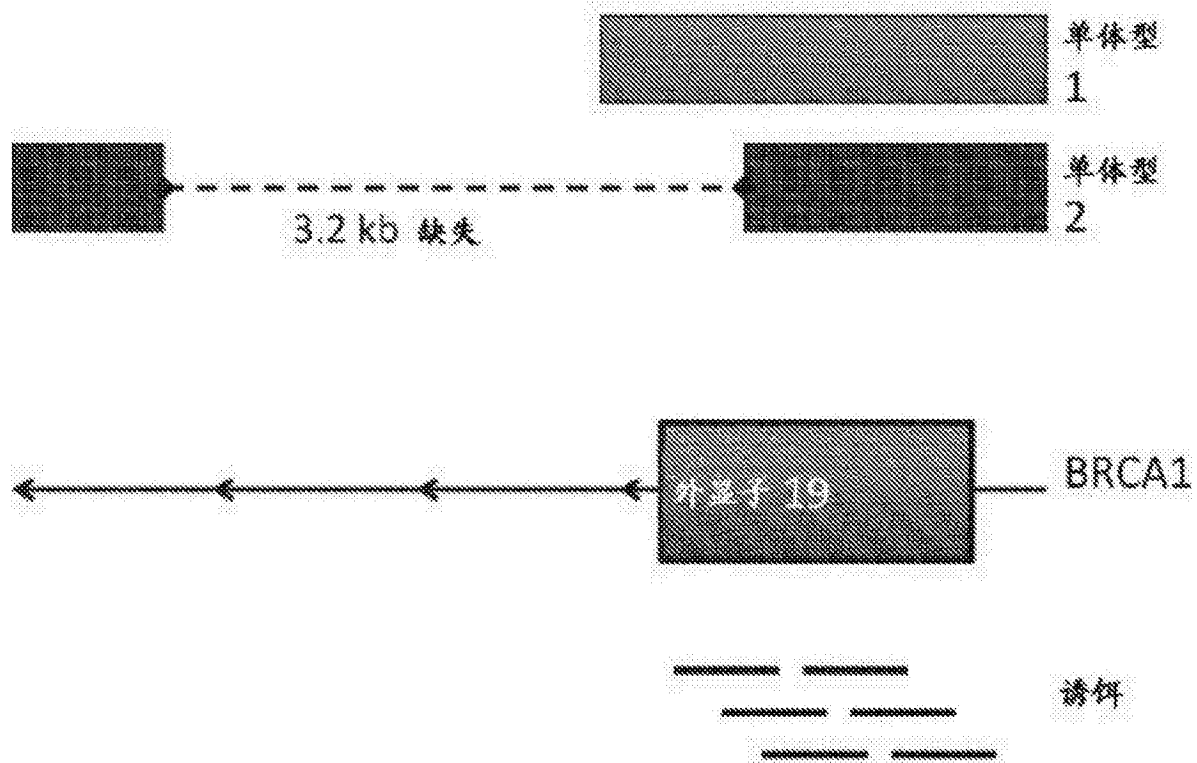


图27

Kragle结果

● 仅组装两个重叠群

— 单体型 1: 813 bp

— 单体型 2: 701 bp

接合性	等位基因	LLR
Hom	1/1	-2,703
Het	1/2	0
Hom	2/2 ✓	-11,443

图28