



(12) 发明专利申请

(10) 申请公布号 CN 115803817 A

(43) 申请公布日 2023. 03. 14

(21) 申请号 202180043728.0

J • S • 维切利

(22) 申请日 2021.10.26

(74) 专利代理机构 北京泛华伟业知识产权代理有限公司 11280

(30) 优先权数据

63/106256 2020.10.27 US

专利代理师 王勇

(85) PCT国际申请进入国家阶段日

2022.12.19

(51) Int. Cl.

G16B 40/10 (2006.01)

G16B 40/30 (2006.01)

(86) PCT国际申请的申请数据

PCT/US2021/056711 2021.10.26

(87) PCT国际申请的公布数据

W02022/093865 EN 2022.05.05

(71) 申请人 因美纳有限公司

地址 美国加利福尼亚州

(72) 发明人 E • J • 奥贾德

A • A • H • 卡格瓦拉 R • 美雄

N • 伍德帕 G • D • 帕纳比

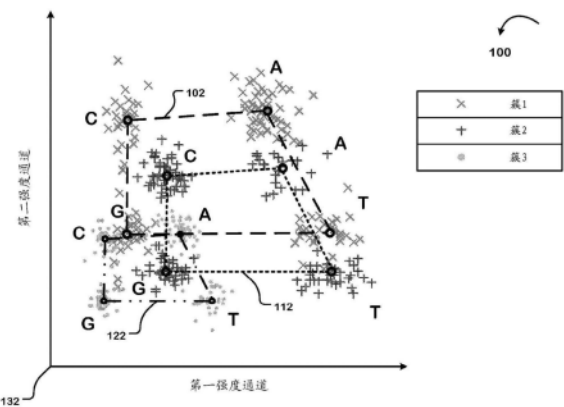
权利要求书4页 说明书30页 附图15页

(54) 发明名称

用于每簇强度校正和碱基检出的系统和方法

(57) 摘要

本发明所公开的技术在逐簇基础上生成变化校正系数,以校正簇间强度分布变化,从而改进碱基检出。放大系数校正缩放变化。通道特异性偏移系数校正沿着相应强度通道的移位变化。基于将在测序运行的先前测序循环处针对目标簇生成的历史强度数据的分析与在该测序运行的当前测序循环处针对该目标簇生成的当前强度数据的分析相结合,生成针对该目标簇的该变化校正系数。然后,使用该变化校正系数来校正正在该测序运行的下一个测序循环处针对该目标簇生成的下一个强度数据。然后,该经过校正的下一个强度数据用于在该下一个测序循环处碱基检出该目标簇。



1. 一种碱基检出目标簇的计算机实现的方法,所述方法包括:
针对目标簇,
从碱基特异性强度分布中读取针对测序运行的当前测序循环记录的当前通道特异性强度,在所述当前测序循环处将所述目标簇碱基检出到所述碱基特异性强度分布,
从所述碱基特异性强度分布的质心读取当前通道特异性分布强度,
基于所述当前通道特异性强度和所述当前通道特异性分布强度,确定用于所述当前测序循环的一组当前强度校正参数,
通过累加用于所述测序运行的先前测序循环的所述一组当前强度校正参数和一组先前累加强度校正参数,确定用于所述当前测序循环的一组当前累加强度校正参数,
基于所述一组当前累加强度校正参数确定用于所述当前测序循环的当前放大系数和当前通道特异性偏移系数,以及
使用所述当前放大系数和所述当前通道特异性偏移系数来校正针对所述测序运行的下一个测序循环记录的下一个通道特异性强度,并针对所述下一个测序循环生成经过校正的下一个通道特异性强度;以及
基于所述经过校正的下一个通道特异性强度,在所述下一个测序循环处碱基检出所述目标簇。
2. 根据权利要求1所述的计算机实现的方法,其中所述一组当前强度校正参数包括当前通道特异性分布强度、当前通道特异性强度误差、当前分布质心到原点的距离和当前分布强度到强度误差相似性度量。
3. 根据权利要求1和2中任一项所述的计算机实现的方法,其中所述当前通道特异性强度误差是所述当前通道特异性强度和所述当前通道特异性分布强度之间的逐通道差异。
4. 根据权利要求1至3中任一项所述的计算机实现的方法,其中所述当前分布质心到原点的距离是多维空间的所述质心和原点之间的欧几里得距离,所述多维空间包含所述碱基特异性强度分布。
5. 根据权利要求1至4中任一项所述的计算机实现的方法,其中所述多维空间是笛卡尔空间、极空间、柱形空间和球形空间中的至少一者。
6. 根据权利要求1至5中任一项所述的计算机实现的方法,其中所述当前分布强度到强度误差相似性度量是所述当前通道特异性分布强度和所述当前通道特异性强度误差之间的逐通道点积的总和。
7. 根据权利要求1至6中任一项所述的计算机实现的方法,其中所述一组当前累加强度校正参数是所述一组当前强度校正参数中的当前强度校正参数和所述一组先前累加强度校正参数中的先前累加强度校正参数的强度校正逐参数总和。
8. 根据权利要求1至7中任一项所述的计算机实现的方法,其中所述一组当前累加强度校正参数是所述当前强度校正参数和所述先前累加强度校正参数的强度校正逐参数平均值。
9. 根据权利要求1至8中任一项所述的计算机实现的方法,其中所述一组先前累加强度校正参数和所述一组当前累加强度校正参数以量化的固定位宽格式存储。
10. 根据权利要求1至9中任一项所述的计算机实现的方法,其中所述当前通道特异性偏移系数被配置为相同。

11. 根据权利要求1至10中任一项所述的计算机实现的方法, 其中所述一组当前累加强度校正参数中的当前累加强度校正参数包括用于所述当前通道特异性分布强度的第一公共当前累加强度校正参数, 以及用于所述当前通道特异性强度误差的第二公共当前累加强度校正参数。

12. 根据权利要求1至11中任一项所述的计算机实现的方法, 其中从所述下一个通道特异性强度中逐通道地减去所述当前通道特异性偏移系数, 以生成下一个通道特异性移位强度, 并且所述下一个通道特异性移位强度除以所述当前放大系数, 以生成所述经过校正的下一个通道特异性强度。

13. 根据权利要求1至12中任一项所述的计算机实现的方法, 所述方法还包括使用加权函数来组合初始放大系数与所述当前放大系数, 以及初始通道特异性偏移系数与所述当前通道特异性偏移系数, 以生成用于所述当前测序循环的加权当前放大系数和加权当前通道特异性偏移系数。

14. 根据权利要求1至13中任一项所述的计算机实现的方法, 其中所述加权函数将最小权重 ($w_{\min}$) 应用于初始放大系数和所述初始通道特异性偏移系数, 并将最大权重 ($w_{\max}$) 应用于所述当前放大系数和所述当前通道特异性偏移系数, 其中 $w_{\min} = (1 - w_{\max})$ 。

15. 根据权利要求1至14中任一项所述的计算机实现的方法, 其中所述最大权重 ($w_{\max}$) 被定义为 $(c-p)/c$, 其中 c 是用于所述当前测序循环的索引, 并且 p 是2至7之间的数。

16. 根据权利要求1至15中任一项所述的计算机实现的方法, 所述方法还包括使用所述加权当前放大系数和所述加权当前通道特异性偏移系数来校正所述下一个通道特异性强度并生成所述经过校正的下一个通道特异性强度。

17. 根据权利要求1至16中任一项所述的计算机实现的方法, 所述方法还包括:

使用最大似然解来针对所述当前测序循环生成用于所述当前放大系数和所述当前通道特异性偏移系数的相应当前最大似然权重;

将所述当前最大似然权重分别应用于所述当前放大系数和所述当前通道特异性偏移系数, 以生成用于所述当前测序循环的最大似然加权当前放大系数和最大似然加权当前通道特异性偏移系数; 以及

使用所述最大似然加权当前放大系数和所述最大似然加权当前通道特异性偏移系数来校正所述下一个通道特异性强度, 并生成所述经过校正的下一个通道特异性强度。

18. 根据权利要求1至17中任一项所述的计算机实现的方法, 所述方法还包括:

将衰减因子应用于所述当前强度校正参数, 以生成用于所述当前测序循环的衰减的当前强度校正参数; 以及

通过强度校正逐参数累加所述衰减的当前强度校正参数和所述先前累加强度校正参数来确定所述当前累加强度校正参数。

19. 根据权利要求1至18中任一项所述的计算机实现的方法, 其中所述衰减因子对于所述测序运行的一定数量的测序循环保持固定, 并且此后基于衰减逻辑以指数方式衰减。

20. 根据权利要求1至19中任一项所述的计算机实现的方法, 其中所述衰减逻辑是 $1 - 1/\tau$, 其中 τ 是预定义的数。

21. 根据权利要求1至20中任一项所述的计算机实现的方法, 所述方法还包括在所述测序运行的连续测序循环处针对所述目标簇迭代所述读取、所述读取、所述确定、所述确定、

所述确定、所述使用和所述碱基检出。

22. 根据权利要求1至21中任一项所述的计算机实现的方法,所述方法还包括针对多个簇并行执行所述读取、所述读取、所述确定、所述确定、所述确定、所述使用和所述碱基检出。

23. 根据权利要求1至22中任一项所述的计算机实现的方法,其中使用最小二乘解来确定用于所述一组当前强度校正参数、所述一组当前累加强度校正参数、所述当前放大系数和所述当前通道特异性偏移系数的闭合形式表达式。

24. 根据权利要求1至23中任一项所述的计算机实现的方法,其中所述当前通道特异性强度分别对应于强度通道。

25. 根据权利要求1至24中任一项所述的计算机实现的方法,其中所述当前通道特异性偏移系数是所述当前通道特异性强度和所述当前通道特异性分布强度之间的所述逐通道差异。

26. 一种碱基检出目标簇的计算机实现的方法,所述方法包括:
针对目标簇,
访问当前强度数据和历史强度数据,
其中所述当前强度数据用于测序运行的当前测序循环,并且
其中所述历史强度数据用于所述测序运行的一个或多个在先前测序循环,
基于所述当前强度数据和所述历史强度数据确定缩放校正系数和通道特异性移位校正系数,

使用所述缩放校正系数和所述通道特异性移位校正系数来校正下一个强度数据并生成经过校正的下一个强度数据,

其中所述下一个强度数据用于所述测序运行的下一个测序循环;以及
基于所述经过校正的下一个强度数据,在所述下一个测序循环处对所述目标簇进行碱基检出。

27. 一种碱基检出目标簇的计算机实现的方法,所述方法包括:
针对目标簇,
访问当前强度数据和历史强度数据,
其中所述当前强度数据用于测序运行的当前测序循环,并且
其中所述历史强度数据用于所述测序运行的一个或多个在先前测序循环,
使用所述当前强度数据和所述历史强度数据来校正下一强度数据并生成经过校正的下一强度数据,
其中所述下一个强度数据用于所述测序运行的下一个测序循环;以及
基于所述经过校正的下一个强度数据,在所述下一个测序循环处对所述目标簇进行碱基检出。

28. 一种系统,所述系统包括存储器以及与之耦合的一个或多个处理器,所述存储器装载有计算机指令以对目标簇进行碱基检出,所述计算机指令在所述处理器上执行时实现根据权利要求1所述的指令。

29. 一种系统,所述系统包括存储器以及与之耦合的一个或多个处理器,所述存储器装载有计算机指令以对目标簇进行碱基检出,所述计算机指令在所述处理器上执行时实现根

据权利要求26所述的指令。

30.一种系统,所述系统包括存储器以及与之耦合的一个或多个处理器,所述存储器装载有计算机指令以对目标簇进行碱基检出,所述计算机指令在所述处理器上执行时实现根据权利要求27所述的指令。

用于每簇强度校正和碱基检出的系统和方法

[0001] 优先权申请

[0002] 本专利申请要求于2020年10月27日提交的名称为“Systems and Methods for Per-Cluster Intensity Correction and Base Calling”的美国临时专利申请第63/106,256号的优先权。

技术领域

[0003] 本发明所公开的技术涉及用于自动分析图像或识别图案的装置和相应方法。本文包括变换图像用于以下目的的系统：(a) 在识别之前增强图像的视觉质量，(b) 相对于传感器或存储的原型定位和记录图像，或通过丢弃无关数据来减少图像数据量，以及(c) 测量图像的显著特征。具体地，所公开的技术涉及生成变化校正系数以校正图像数据中的簇间强度分布变化。

[0004] 文献并入

[0005] 2020年5月5日提交的名称为“EQUALIZATION-BASED IMAGE PROCESSING AND SPATIAL CROSSTALK ATTENUATOR”的美国临时专利申请第63/020,449号(代理人案卷号 ILLM 1032-1/IP-1991-PRV)；

[0006] 2018年3月26日提交的名称为“DETECTION APPARATUS HAVING A MICROFLUOROMETER, A FLUIDIC SYSTEM, AND A FLOW CELL LATCH CLAMP MODULE”的美国非临时专利申请第15/936,365号；

[0007] 2019年9月11日提交的名称为“FLOW CELLS AND METHODS RELATED TO SAME”的美国非临时专利申请第16/567,224号；

[0008] 2019年6月12日提交的名称为“DEVICE FOR LUMINESCENT IMAGING”的美国非临时专利申请第16/439,635号；

[0009] 2017年5月12日提交的名称为“INTEGRATED OPTOELECTRONIC READ HEAD AND FLUIDIC CARTRIDGE USEFUL FOR NUCLEIC ACID SEQUENCING”的美国非临时专利申请第15/594,413号；

[0010] 2019年3月12日提交的名称为“ILLUMINATION FOR FLUORESCENCE IMAGING USING OBJECTIVE LENS”的美国非临时专利申请第16/351,193号；

[0011] 2009年12月15日提交的名称为“DYNAMIC AUTOFOCUS METHOD AND SYSTEM FOR ASSAY IMAGER”的美国非临时专利申请第12/638,770号；

[0012] 2013年3月1日提交的名称为“KINETIC EXCLUSION AMPLIFICATION OF NUCLEIC ACID LIBRARIES”的美国非临时专利申请第13/783,043号；

[0013] 2011年1月13日提交的名称为“DATA PROCESSING SYSTEM AND METHODS”的美国非临时专利申请第13/006,206号；

[0014] 2014年10月31日提交的名称为“IMAGE ANALYSIS USEFUL FOR PATTERNED OBJECTS”的美国非临时专利申请第14/530,299号；

[0015] 2014年12月3日提交的名称为“METHODS AND SYSTEMS FOR ANALYZING IMAGE

DATA”的美国非临时专利申请第15/153,953号；

[0016] 2013年9月6日提交的名称为“CENTROID MARKERS FOR IMAGE ANALYSIS OF HIGH DENSITY CLUSTERS IN COMPLEX POLYNUCLEOTIDE SEQUENCING”的美国非临时专利申请第14/020,570号；

[0017] 2014年10月31日提交的名称为“IMAGE ANALYSIS USEFUL FOR PATTERNED OBJECTS”的美国非临时专利申请第14/530,299号；

[0018] 2009年9月23日提交的名称为“METHOD AND SYSTEM FOR DETERMINING THE ACCURACY OF DNA BASE IDENTIFICATIONS”的美国非临时专利申请第12/565,341号；

[0019] 2007年3月30日提交的名称为“SYSTEMS AND DEVICES FOR SEQUENCE BY SYNTHESIS ANALYSIS”的美国非临时专利申请第12/295,337号；

[0020] 2008年1月28日提交的名称为“IMAGE DATA EFFICIENT GENETIC SEQUENCING METHOD AND SYSTEM”的美国非临时专利申请第12/020,739号；

[0021] 2013年3月15日提交的名称为“BIOSENSORS FOR BIOLOGICAL OR CHEMICAL ANALYSIS AND SYSTEMS AND METHODS FOR SAME”的美国非临时专利申请第13/833,619号(代理人案卷号IP-0626-US)；

[0022] 2016年6月7日提交的名称为“BIOSENSORS FOR BIOLOGICAL OR CHEMICAL ANALYSIS AND METHODS OF MANUFACTURING THE SAME”的美国非临时专利申请第15/175,489号(代理人案卷号IP-0689-US)；

[0023] 2013年4月26日提交的名称为“MICRODEVICES AND BIOSENSOR CARTRIDGES FOR BIOLOGICAL OR CHEMICAL ANALYSIS AND SYSTEMS AND METHODS FOR THE SAME”的美国非临时专利申请第13/882,088号(代理人案卷号IP-0462-US)；

[0024] 2012年9月21日提交的名称为“METHODS AND COMPOSITIONS FOR NUCLEIC ACID SEQUENCING”的美国非临时专利申请第13/624,200号(代理人案卷号IP-0538-US)；

[0025] 2019年3月21日提交的名称为“TRAINING DATA GENERATION FOR ARTIFICIAL INTELLIGENCE-BASED SEQUENCING”的美国临时专利申请第62/821,602号(代理人案卷号ILLM 1008-1/IP-1693-PRV)；

[0026] 2019年3月21日提交的名称为“ARTIFICIAL INTELLIGENCE-BASED GENERATION OF SEQUENCING METADATA”的美国临时专利申请第62/821,618号(代理人案卷号ILLM 1008-3/IP-1741-PRV)；

[0027] 2019年3月21日提交的名称为“ARTIFICIAL INTELLIGENCE-BASED BASE CALLING”的美国临时专利申请第62/821,681号(代理人案卷号ILLM 1008-4/IP-1744-PRV)；

[0028] 2019年3月21日提交的名称为“ARTIFICIAL INTELLIGENCE-BASED QUALITY SCORING”的美国临时专利申请第62/821,724号(代理人案卷号ILLM 1008-7/IP-1747-PRV)；

[0029] 2019年3月21日提交的名称为“ARTIFICIAL INTELLIGENCE-BASED SEQUENCING”的美国临时专利申请第62/821,766号(代理人案卷号:ILLM 1008-9/IP-1752-PRV)；

[0030] 2019年6月14日提交的名称为“TRAINING DATA GENERATION FOR ARTIFICIAL INTELLIGENCE-BASED SEQUENCING”的荷兰申请第2023310号(代理人案卷号ILLM 1008-11/IP-1693-NL)；

[0031] 2019年6月14日提交的名称为“ARTIFICIAL INTELLIGENCE-BASED GENERATION OF SEQUENCING METADATA”的荷兰申请第2023311号(代理人案卷号ILLM 1008-12/IP-1741-NL) ;

[0032] 2019年6月14日提交的名称为“ARTIFICIAL INTELLIGENCE-BASED BASE CALLING”的荷兰申请第2023312号(代理人案卷号ILLM 1008-13/IP-1744-NL) ;

[0033] 2019年6月14日提交的名称为“ARTIFICIAL INTELLIGENCE-BASED QUALITY SCORING”的荷兰申请第2023314号(代理人案卷号ILLM1008-14/IP-1747-NL) ;以及

[0034] 2019年6月14日提交的名称为“ARTIFICIAL INTELLIGENCE-BASED SEQUENCING”的荷兰申请第2023316号(代理人案卷号ILLM 1008-15/IP-1752-NL) 。

[0035] 2020年3月20日提交的名称为“TRAINING DATA GENERATION FOR ARTIFICIAL INTELLIGENCE-BASED SEQUENCING”的美国非临时专利申请第16/825,987号(代理人案卷号ILLM 1008-16/IP-1693-US) ;

[0036] 2020年3月20日提交的名称为“TRAINING DATA GENERATION FOR ARTIFICIAL INTELLIGENCE-BASED SEQUENCING”的美国非临时专利申请第16/825,991号(代理人案卷号ILLM 1008-17/IP-1741-US) ;

[0037] 2020年3月20日提交的名称为“ARTIFICIAL INTELLIGENCE-BASED BASE CALLING”的美国非临时专利申请第16/826,126号(代理人案卷号ILLM 1008-18/IP-1744-US) ;

[0038] 2020年3月20日提交的名称为“ARTIFICIAL INTELLIGENCE-BASED QUALITY SCORING”的美国非临时专利申请第16/826,134号(代理人案卷号ILLM 1008-19/IP-1747-US) ;

[0039] 2020年3月21日提交的名称为“ARTIFICIAL INTELLIGENCE-BASED SEQUENCING”的美国非临时专利申请第16/826,168号(代理人案卷号ILLM 1008-20/IP-1752-PRV) ;

[0040] 2019年5月16日提交的名称为“SYSTEMS AND DEVICES FOR CHARACTERIZATION AND PERFORMANCE ANALYSIS OF PIXEL-BASED SEQUENCING”的美国临时专利申请第62/849,091号(代理人案卷号ILLM 1011-1/IP-1750-PRV) ;

[0041] 2019年5月16日提交的名称为“BASE CALLING USING CONVOLUTIONS”的美国临时专利申请第62/849,132号(代理人案卷号ILLM 1011-2/IP-1750-PR2) ;

[0042] 2019年5月16日提交的名称为“BASE CALLING USING COMPACT CONVOLUTIONS”的美国临时专利申请第62/849,133号(代理人案卷号ILLM 1011-3/IP-1750-PR3) ;

[0043] 2020年2月20日提交的名称为“ARTIFICIAL INTELLIGENCE-BASED BASE CALLING OF INDEX SEQUENCES”的美国临时专利申请第62/979,384号(代理人案卷号ILLM 1015-1/IP-1857-PRV) ;

[0044] 2020年2月20日提交的名称为“ARTIFICIAL INTELLIGENCE-BASED MANY-TO-MANY BASE CALLING”的美国临时专利申请第62/979,414号(代理人案卷号ILLM 1016-1/IP-1858-PRV) ;

[0045] 2020年2月20日提交的名称为“KNOWLEDGE DISTILLATION-BASED COMPRESSION OF ARTIFICIAL INTELLIGENCE-BASED BASE CALLER”的美国临时专利申请第62/979,385号(代理人案卷号ILLM 1017-1/IP-1859-PRV) ;

[0046] 2020年2月20日提交的名称为“MULTI-CYCLE CLUSTER BASED REAL TIME

ANALYSIS SYSTEM”的美国临时专利申请第62/979,412号(代理人案卷号ILLM 1020-1/IP-1866-PRV) ;

[0047] 2020年2月20日提交的名称为“DATA COMPRESSION FOR ARTIFICIAL INTELLIGENCE-BASED BASE CALLING”的美国临时专利申请第62/979,411号(代理人案卷号ILLM 1029-1/IP-1964-PRV) ;以及

[0048] 2020年2月20日提交的名称为“SQUEEZING LAYER FOR ARTIFICIAL INTELLIGENCE-BASED BASE CALLING”的美国临时专利申请第62/979,399号(代理人案卷号ILLM 1030-1/IP-1782-PRV) 。

背景技术

[0049] 本部分中讨论的主题不应仅因为在本部分中有提及就被认为是现有技术。类似地,在本部分中提及的或与作为背景技术提供的主题相关联的问题不应被认为先前在现有技术中已被认识到。本部分中的主题仅表示不同的方法,这些方法本身也可对应于受权利要求书保护的技術的具体实施。

[0050] 本公开涉及在测序运行期间分析图像数据以对簇进行碱基检出。图像数据分析的一个挑战是被称为碱基的簇群体中簇的强度分布的变化。这导致测序运行期间数据吞吐量下降,并且误差率增加。

[0051] 簇间强度分布变化存在许多潜在的原因。其可能是由簇群体中片段长度分布引起的簇亮度差异产生。其可能是由相位误差产生,当簇中的分子在某个测序循环中没有掺入核苷酸并且落后于其他分子时,或当分子在单个测序循环中掺入多于一个的核苷酸时,就会发生相位误差。其可能是由褪色,即随着测序运行的进行,由于过度洗涤和激光照射,簇的信号强度作为测序循环数的函数呈指数衰减产生。其可能是由发育不全的簇菌落,即在图案化流动池上产生空的或部分填充的孔的小的簇尺寸产生。其可能是由非排斥放大引起的簇菌落重叠产生。其可能是由照明不足或照明不均匀产生,例如,由于簇位于流动池的边缘。其可能是由流动池上的杂质产生,这些杂质混淆所发射的信号。其可能是由多克隆簇,即当多个簇沉积在相同的孔中时产生。

[0052] 出现了校正簇间强度分布变化的机会。可能导致测序运行期间碱基检出吞吐量的提高和碱基检出误差率的降低。

附图说明

[0053] 本专利或专利申请文件包含至少一幅彩色附图。具有彩色附图的本专利或本专利申请公布的副本将在提出请求并支付必要费用后由专利局提供。彩色附图也可通过补充内容选项卡成对地获得。

[0054] 在附图中,在所有不同视图中,类似的参考符号通常是指类似的部件。另外,附图未必按比例绘制,而是重点说明所公开的技术的原理。在以下描述中,参考以下附图描述了本发明所公开的技术的各种实施方式,其中:

[0055] 图1例示了由所公开的技术发现和校正的簇间强度分布变化的一个示例。

[0056] 图2例示了实现本文所公开的变化校正逻辑的碱基检出流水线的示例。

[0057] 图3例示了实现本文所公开的最小二乘解的最小二乘解确定器。

- [0058] 图4示出了在当前测序循环处如何测量目标簇的通道特异性分布强度的示例。
- [0059] 图5示出了在当前测序循环处如何针对目标簇计算通道特异性强度误差的示例。
- [0060] 图6示出了在当前测序循环处如何针对目标簇计算分布质心到原点的距离的示例。
- [0061] 图7描绘了实现变化校正逻辑的碱基检出流水线的另一个示例。
- [0062] 图8示出了本文所描述的加权函数的一个具体实施。
- [0063] 图9示出了将最大似然权重直接应用于变化校正系数的一个具体实施。
- [0064] 图10示出了将指数衰减因子应用于变化校正系数的一个具体实施。
- [0065] 图11示出了确定通道特异性偏移系数的另一种具体实施。
- [0066] 图12、图13和图14比较了三种方法的性能,即仅缩放解、仅偏移解(在图11中讨论)和最小二乘解(在图3中讨论)。
- [0067] 图15是可用于实现所公开的技术的计算机系统。

具体实施方式

[0068] 呈现以下讨论以使得本领域的任何技术人员能够实现和使用所公开的技术,并且在特定应用及其要求的上下文中提供以下讨论。对所公开的具体实施的各种修改对于本领域的技术人员而言将是显而易见的,并且在不脱离所公开的技术的实质和范围的情况下,本文所定义的一般原理可应用于其他具体实施和应用。因此,所公开的技术并非旨在限于所示的具体实施,而是要符合与本文所公开的原理和特征一致的最广范围。

[0069] 简介

[0070] 所公开技术的进展始于在测序运行期间对被碱基检出的簇群体中簇的强度分布的分析。分析揭示了簇群体中簇的强度分布采取类似的形式(例如,梯形),但是在比例上不同,并且从多维空间100的原点132移位。我们称之为“簇间强度分布变化”。多维空间100可以是笛卡尔空间、极空间、柱形空间或球形空间。

[0071] 图1例示了由所公开的技术发现和校正的簇间强度分布变化的一个示例。图1分别描绘了簇群体中簇1、2和3的强度分布112、122和132。目标簇的强度分布包括在测序运行的多个测序循环(例如,150个)处捕获由于目标簇中核苷酸掺入而产生的化学发光信号的强度值。

[0072] 在图1所例示的具体实施中,强度值是从由测序器在多个测序循环中的每个测序循环处生成的两个不同的颜色/强度通道测序图像中提取的。测序器的示例包括Illumina的iSeq、HiSeqX、HiSeq 3000、HiSeq 4000、HiSeq2500、NovaSeq 6000、NextSeq 550、NextSeq 1000、NextSeq 2000、NextSeqDx、MiSeq、以及MiSeqDx。

[0073] 在一个具体实施中,测序器使用合成法测序(SBS)来生成测序图像。SBS依赖于生长与具有经荧光标记的核苷酸的簇链互补的新生链,同时跟踪每个新添加的核苷酸的发射信号。经荧光标记的核苷酸具有锚定核苷酸类型的荧光团信号的3'可移除块。SBS在重复的测序循环中进行,每个循环包括三个步骤:(a)通过添加经荧光标记的核苷酸延长新生链;(b)使用测序器的光学系统的一个或多个激光器激发荧光团,并通过光学系统的不同滤波器成像,从而产生测序图像;以及(c)裂解荧光团并移除3'块,以为下一个测序循环做准备。重复掺入和成像直至达到指定数量的测序循环,从而限定读取长度。使用该方法,每个测序

循环询问沿簇链的新位置。

[0074] 在图1中,“×”符号表示簇1的强度值,“+”符号表示簇2的强度值,并且“•”符号表示簇3的强度值。将四种不同核苷酸类型/碱基A、C、T和G的同一性编码为两个彩色图像(即第一强度通道和第二强度通道)中的强度值的组合。例如,可通过提供在第一强度通道(多维空间100的x轴)中检测到的第一核苷酸类型(例如,碱基T)、在第二强度通道(多维空间100的y轴)中检测到的第二核苷酸类型(例如,碱基C)、在第一强度通道和第二强度通道这两者中检测到的第三核苷酸类型(例如,碱基A),以及缺乏标记并且在任一强度通道中未检测到或最低限度检测到的第四核苷酸类型(例如,碱基G),来对核酸进行测序。

[0075] 在一些具体实施中,通过迭代地将四个强度分布(例如,高斯分布)拟合到第一强度通道和第二强度通道中的强度值来生成强度分布。四个强度分布对应于四个碱基A、C、T和G。在强度分布中,相对于第二强度通道中的强度值绘制第一强度通道中的强度值(例如,“作为散点图”),并且强度值分离成四个强度分布。

[0076] 强度分布可采用任何形状(例如,梯形、正方形、矩形、菱形等)。关于如何将四个强度分布拟合到用于碱基检出的强度值的附加细节可在美国专利申请公布第2018/0274023A1中找到,该专利的公开通过引用以其整体并入本文。

[0077] 在一个具体实施中,每个强度通道对应于光学系统使用的多个滤波器波长带中的一个滤波器波长带。在另一种具体实施中,每个强度通道对应于测序循环中的多个成像事件中的一个成像事件。在又一具体实施中,每个强度通道对应于利用特定激光器照明和通过光学系统的特定光学滤波器成像的组合。

[0078] 对于本领域技术人员来说显而易见的是,所公开的技术可类似地应用于使用单通具体实施、四通具体实施等生成的测序图像。例如,在四通具体实施的情况下,确定四个通道特异性偏移系数,以分别校正四个强度通道中的移位变化。

[0079] 变化校正逻辑

[0080] 簇群体中簇的大量(例如,数千个、数百万个、数十亿个等)强度分布中的簇间强度分布变化导致碱基检出吞吐量下降和碱基检出误差率增加。为了校正簇间强度分布变化,我们公开了一种在逐簇基础上生成变化校正系数的变化校正逻辑。

[0081] 在双通道具体实施中,变化校正系数包括说明簇间强度分布变化中的缩放变化的放大系数,以及分别说明簇间强度分布变化中沿第一强度通道和第二强度通道的移位变化的两个通道特异性偏移系数。在另一种具体实施中,通过使用针对不同强度通道(例如,第一强度通道和第二强度通道)的公共偏移系数来解释移位变化。

[0082] 基于将在测序运行的先前测序循环处针对目标簇确定的历史强度统计的分析与在当前测序循环处针对目标簇确定的当前强度统计的分析相结合,在测序运行的当前测序循环处生成目标簇的变化校正系数。变化校正系数用于在测序运行的下一个测序循环处校正针对目标簇记录的下一个强度读数。经过校正的下一个强度读数用于在下一个测序循环处碱基检出目标簇。在测序运行的连续测序循环处,将相应变化校正系数重复地应用于相应簇的相应强度分布的结果是,强度分布变得一致并锚固到原点132(例如,在梯形的底部下角处)。

[0083] 图2例示了实现变化校正逻辑的碱基检出流水线200的示例。

[0084] 当前测序循环

[0085] 在当前测序循环*i*处,测序器生成测序图像202。测序图像202包含在当前测序循环*i*处针对目标簇记录的当前强度数据202t,以及包含针对簇群体中的多个簇记录的当前强度数据202。当前强度数据202t中的“t”是指目标簇。

[0086] 当前强度数据202t被提供给碱基检出器212。碱基检出器212处理当前强度数据202t,并在当前测序循环*i*处针对目标簇生成当前碱基检出222。碱基检出器212的示例包括Illumina的实时分析(RTA)软件、Illumina的基于神经网络的碱基检出器(例如,在美国专利公布第US 2020/0302297 A1中描述的)和Illumina的基于均衡器的碱基检出器(例如,在美国临时专利申请第63/020,449中描述的)

[0087] 在当前测序循环*i*处,目标簇的强度分布包括当前强度数据202t,以及在当前测序循环*i*之前,即在先前测序循环1至*i*-1之前的测序运行的那些测序循环处针对目标簇记录的当前历史强度数据。我们将当前强度数据202t和当前历史强度数据统称为当前可用强度数据。

[0088] 在强度分布中,四个强度分布对应于四个碱基A、C、T和G。在一个具体实施中,通过确定当前强度数据202t属于四个强度分布中的哪一个来进行当前碱基检出222。在一些具体实施中,这是通过使用期望最大化算法来实现的。期望值最大化算法迭代地最大化最拟合当前可用强度数据的观察平均值(质心)和分布(协方差)的似然性。

[0089] 一旦通过使用期望最大化算法在当前测序循环*i*处确定了四个强度分布,就计算属于四个强度分布中的每个强度分布的当前强度数据202t的似然性。最大似然性给出了当前碱基检出222。例如,考虑“m、n”分别是第一强度通道和第二强度通道中的当前强度数据202t的强度值。期望值最大化算法生成四个值,这四个值表示属于四个强度分布中的每个强度分布的“m、n”个强度值的似然性。四个值中的最大值表示被检出的碱基。

[0090] 在其他具体实施中,k-平均值分簇算法、类k-平均值分簇算法、基于直方图的方法等可用于碱基检出。

[0091] 下一个测序循环

[0092] 在下一个测序循环*i*+1处,强度校正参数确定器232基于当前碱基检出222确定用于目标簇的强度校正参数242。在双通道具体实施中,强度校正参数242包括第一强度通道中的分布强度、第二强度通道中的分布强度、第一强度通道中的强度误差、第二强度通道中的强度误差、分布质心到原点的距离以及分布强度到强度误差的相似性度量。

[0093] 我们如下定义强度校正参数242中的每个强度校正参数:

[0094] 1) 第一强度通道中的分布强度是在碱基特异性强度分布的质心处的第一强度通道中的强度值,目标簇在当前测序循环*i*处属于该碱基特异性强度分布。注意,碱基特异性强度分布是用于检出当前碱基检出222的基础。

[0095] 2) 第二强度通道中的分布强度是在碱基特异性强度分布的质心处的第二强度通道中的强度值。

[0096] 3) 第一强度通道中的强度误差是第一强度通道中的当前强度数据202t的测量强度值和第一强度通道中的分布强度之间的差异。

[0097] 4) 第二强度通道中的强度误差是第二强度通道中的当前强度数据202t的测量强度值和第二强度通道中的分布强度之间的差异。

[0098] 5) 分布质心到原点的距离是碱基特异性强度分布的质心和多维空间100的原点

132之间的欧几里得距离,其中碱基特异性强度分布被拟合(例如,通过使用期望最大化算法)。在其他具体实施中,可使用距离度量,诸如马氏距离和最小协方差行列式(MCD)距离,以及其相关联的质心估计器。

[0099] 6) 分布强度到强度误差相似性度量是第一强度通道和第二强度通道中的分布强度和强度误差之间的逐通道点积的总和。

[0100] 累加强度校正参数确定器252将强度校正参数242与来自先前测序循环 $i-1$ 的历史累加强度校正参数254累加,以确定累加强度校正参数262。累加的示例包括求和与求平均。

[0101] 变化校正系数确定器272基于确定的累加强度校正参数262来确定变化校正系数282。

[0102] 在下一个测序循环 $i+1$ 处,测序器生成测序图像294。测序图像294包含在下一个测序循环 $i+1$ 处针对目标簇记录的下一个强度数据294t,以及包含针对簇群体中的多个簇记录的下一个强度数据294。下一个强度数据294t中的“t”是指目标簇。

[0103] 强度校正器292将变化校正系数282应用于下一个强度数据294t,以生成经过校正的下一个强度数据296t。经过校正的下一个强度数据296t中的“t”是指目标簇。

[0104] 在下一个测序循环 $i+1$ 处,目标簇的强度分布包括经过校正的下一个强度数据296t,以及在下一个测序循环 $i+1$ 之前,即在先前测序循环1至 i 之前的测序运行的那些测序循环中针对目标簇记录的下一个历史强度数据。我们将经过校正的下一个强度数据296t和下一个历史强度数据统称为下一个可用强度数据。

[0105] 经过校正的下一个强度数据296t被提供给碱基检出器212。碱基检出器212处理经过校正的下一个强度数据296t,并在下一个测序循环 $i+1$ 处针对目标簇生成下一个碱基检出298。为了生成下一个碱基检出298,期望最大化算法基于经过校正的下一个强度数据296t观察平均值(质心)和分布(协方差),以最佳拟合下一个可用强度数据。

[0106] 一旦通过使用期望最大化算法在下一个测序循环 $i+1$ 处确定了四个强度分布,就计算属于四个强度分布中的每个强度分布的经过校正的下一个强度数据296t的似然性。最大似然性给出了下一个碱基检出298。

[0107] 注意,碱基检出流水线200是在逐簇基础上执行的,并且是针对簇群体中的多个簇并行执行的。另外,对于测序运行的连续测序循环(例如,针对双端测序运行中读取1的连续150个测序循环和读取2的另一个连续150个测序循环),重复执行碱基检出流水线200。

[0108] 最小二乘解

[0109] 图3例示了实现本文所公开的最小二乘解300的最小二乘解确定器。最小二乘解300确定针对累加强度校正参数262和变化校正系数282的闭合形式表达式。最小二乘解确定器302包括强度建模器312和最小化器322。

[0110] 强度建模器312根据以下表达式针对目标簇的测量强度和变化校正系数282之间的关系进行建模:

[0111] $y_{C,i} = ax_{C,i} + d_i + n_{C,i}$ 等式(1)

[0112] 其中

[0113] a 是用于目标簇的放大系数

[0114] d_i 是用于强度通道 i 的通道特异性偏移系数

[0115] $x_{C,i}$ 是在当前测序循环 C 处在强度通道 i 中针对目标簇的分布强度

[0116] $y_{C,i}$ 是在当前测序循环C处在强度通道i中针对目标簇的测量强度

[0117] $n_{C,i}$ 是在当前测序循环C处在强度通道i中针对目标簇的的附加噪声

[0118] 最小化器322使用最小二乘解300来最小化以下表达式：

$$[0119] \quad errorf(\hat{a}, \hat{d}_i) = \sum_{c=1}^C \sum_{i=1}^2 (\hat{a}x_{C,i} + \hat{d}_i - y_{C,i})^2 \quad \text{等式 (2)}$$

[0120] 其中：

[0121] errorf是误差函数

[0122] $\hat{a}$ 是用于目标簇的放大系数

[0123] $\hat{d}_i$ 是强度通道i的通道特异性偏移系数

[0124] C是当前测序循环

[0125] 使用链式法则，最小化器322计算误差函数相对于放大系数 $\hat{a}$ 和通道特异性偏移系数 $\hat{d}_i$ 的两个偏导数。偏导数将等式2设置为零，以最小化误差函数：

$$[0126] \quad \frac{\partial errorf}{\partial \hat{a}} = \frac{\partial}{\partial \hat{a}} \left[\sum_{c=1}^C \sum_{i=1}^2 (\hat{a}x_{C,i} + \hat{d}_i - y_{C,i})^2 \right] = 0 \quad \text{等式 (3)}$$

$$[0127] \quad \frac{\partial errorf}{\partial \hat{d}_i} = \frac{\partial}{\partial \hat{d}_i} \left[\sum_{c=1}^C \sum_{i=1}^2 (\hat{a}x_{C,i} + \hat{d}_i - y_{C,i})^2 \right] = 0 \quad \text{等式 (4)}$$

[0128] 通道特异性强度误差 $e_{C,i}$ 定义如下：

$$[0129] \quad e_{C,i} = y_{C,i} - x_{C,i} \quad \text{等式 (5)}$$

[0130] 闭合形式表达式

[0131] 第一偏导数确定针对放大系数 $\hat{a}$ 的闭合形式表达式如下：

$$\frac{\partial errorf}{\partial \hat{a}} = \sum_{c=1}^C \sum_{i=1}^2 2x_{C,i} (\hat{a}x_{C,i} + \hat{d}_i - y_{C,i}) = 0 \quad \text{等式 (6)}$$

$$= \sum_{c=1}^C \sum_{i=1}^2 2x_{C,i} (\hat{a}x_{C,i} - x_{C,i} + \hat{d}_i - y_{C,i} + x_{C,i}) = 0 \quad \text{等式 (7)}$$

$$[0132] \quad = \sum_{c=1}^C \sum_{i=1}^2 x_{C,i}^2 (\hat{a} - 1) + \hat{d}_i x_{C,i} - e_{C,i} x_{C,i} = 0 \quad \text{等式 (8)}$$

$$\begin{aligned} &= (\hat{a} - 1) \sum_{c=1}^C (x_{C,1}^2 + x_{C,2}^2) + \hat{d}_1 \sum_{c=1}^C x_{C,1} + \hat{d}_2 \sum_{c=1}^C x_{C,2} - \sum_{c=1}^C e_{C,1} x_{C,1} \\ &\quad - \sum_{c=1}^C e_{C,2} x_{C,2} = 0 \quad \text{等式 (9)} \end{aligned}$$

[0133] 累加强度校正参数262的闭合形式表达式 $\bar{x}_1$ 、 $\bar{x}_2$ 、 $\bar{e}_1$ 、 $\bar{e}_2$ 、 $\overline{xx}$ 和 $\overline{xe}$ 将等式9重新表征如下：

$$[0134] \quad = C(\hat{a} - 1)\overline{xx} + C\bar{x}_1\hat{d}_1 + C\bar{x}_2\hat{d}_2 - C\overline{xe} = 0 \quad \text{等式 (10)}$$

$$[0135] \quad = (\hat{a} - 1)\overline{xx} + \bar{x}_1\hat{d}_1 + \bar{x}_2\hat{d}_2 - \overline{xe} = 0 \quad \text{等式 (11)}$$

[0136] 其中：

$$[0137] \quad \bar{x}_1 = \sum_{c=1}^C x_{c,1} \quad \text{中间术语(1)}$$

$$[0138] \quad \bar{x}_2 = \sum_{c=1}^C x_{c,2} \quad \text{中间术语(2)}$$

$$[0139] \quad \bar{e}_1 = \sum_{c=1}^C e_{c,1} \quad \text{中间术语(3)}$$

$$[0140] \quad \bar{e}_2 = \sum_{c=1}^C e_{c,2} \quad \text{中间术语(4)}$$

$$[0141] \quad \overline{xx} = \sum_{c=1}^C (x_{c,1}^2 + x_{c,2}^2) \quad \text{中间术语(5)}$$

$$[0142] \quad \overline{xe} = \sum_{c=1}^C (x_{c,1}e_{c,1} + x_{c,2}e_{c,2}) \quad \text{中间术语(6)}$$

[0143] 我们如下定义累加强度校正参数262中的每个累加强度校正参数：

[0144] 1) 第一累加强度校正参数 $\bar{x}_1$ 是在先前测序循环1至i-1和当前测序循环i的每一者处针对目标簇测量的第一强度通道中的分布强度的总和。

[0145] 2) 第二累加强度校正参数 $\bar{x}_2$ 是在先前测序循环1至i-1和当前测序循环i的每一者处针对目标簇测量的第二强度通道中的分布强度的总和。

[0146] 3) 第三累加强度校正参数 $\bar{e}_1$ 是在先前测序循环1至i-1和当前测序循环i的每一者处针对目标簇计算的第一强度通道中的强度误差的总和。

[0147] 4) 第四累加强度校正参数 $\bar{e}_2$ 是在先前测序循环1至i-1和当前测序循环i的每一者处针对目标簇计算的第二强度通道中的强度误差的总和。

[0148] 5) 第五累加强度校正参数 $\overline{xx}$ 是在先前测序循环1至i-1和当前测序循环i的每一者处针对目标簇计算的分布质心到原点的距离的总和。

[0149] 6) 第六累加强度校正参数 $\overline{xe}$ 是在先前测序循环1至i-1和当前测序循环i的每一者处针对目标簇计算的分布强度-强度误差相似性度量的总和。

[0150] 第二偏导数确定针对偏移系数 $\hat{d}_i$ 的闭合形式表达式如下：

$$\frac{\partial errorf}{\partial \hat{d}_i} = \sum_{c=1}^C \sum_{i=1}^2 2 (\hat{a}x_{C,i} + \hat{d}_i - y_{C,i}) = 0 \quad \text{等式 (12)}$$

$$[0151] \quad = \sum_{c=1}^C \sum_{i=1}^2 (\hat{a}x_{C,i} - x_{C,i} + \hat{d}_i - y_{C,i} + x_{C,i}) = 0 \quad \text{等式 (13)}$$

$$= \sum_{c=1}^C \sum_{i=1}^2 x_{C,i}^2 \left((\hat{a} - 1) + \sum_{i=1}^2 \hat{d}_i + \sum_{C=1}^C x_{C,i} \right) = 0 \quad \text{等式 (14)}$$

[0152] 然后,针对每个强度通道:

$$[0153] \quad = C(\hat{a} - 1)\bar{x}_i + C\hat{d}_i - C\bar{e}_i = 0 \quad \text{等式 (15)}$$

$$[0154] \quad = (\hat{a} - 1)\bar{x}_i + \hat{d}_i - \bar{e}_i = 0 \quad \text{等式 (16)}$$

[0155] 针对第一强度通道,即*i*=1:

$$[0156] \quad \hat{d}_1 = \bar{e}_1 + (1 - \hat{a})\bar{x}_1 \quad \text{等式 (17)}$$

[0157] 其中:

[0158] $\hat{d}_1$ 是针对第一强度通道的偏移系数

[0159] 针对第二强度通道,即*i*=2:

$$[0160] \quad \hat{d}_2 = \bar{e}_2 + (1 - \hat{a})\bar{x}_2 \quad \text{等式 (18)}$$

[0161] 其中:

[0162] $\hat{d}_2$ 是针对第二强度通道的偏移系数

[0163] 将等式17和等式18代入等式11:

$$\frac{\partial errorf}{\partial \hat{a}} = (\hat{a} - 1)\overline{xx} + \bar{x}_1[\bar{e}_1 + (1 - \hat{a})\bar{x}_1] + \bar{x}_2[\bar{e}_2 + (1 - \hat{a})\bar{x}_2] - \overline{xe} = 0 \quad \text{等式 (19)}$$

$$[0164] \quad = (\hat{a} - 1)\overline{xx} + (1 - \hat{a})\bar{x}_1^2 + (1 - \hat{a})\bar{x}_2^2 + \bar{x}_1\bar{e}_1 + \bar{x}_2\bar{e}_2 - \overline{xe} = 0 \quad \text{等式 (20)}$$

$$= (\hat{a} - 1)(\overline{xx} - \bar{x}_1^2 - \bar{x}_2^2) + \bar{x}_1\bar{e}_1 + \bar{x}_2\bar{e}_2 - \overline{xe} = 0 \quad \text{等式 (21)}$$

$$[0165] \quad \hat{a} - 1 = \frac{\bar{x}_1\bar{e}_1 + \bar{x}_2\bar{e}_2 - \overline{xe}}{\overline{xx} - \bar{x}_1^2 - \bar{x}_2^2} \quad \text{等式 (22)}$$

$$[0166] \quad \hat{a} = 1 + \frac{\bar{x}_1\bar{e}_1 + \bar{x}_2\bar{e}_2 - \overline{xe}}{\overline{xx} - \bar{x}_1^2 - \bar{x}_2^2} \quad \text{等式 (23)}$$

[0167] 其中:

[0168] $\hat{a}$ 是用于目标簇的放大系数

[0169] 在另一种具体实施中,为了减少每簇的存储器需求,通过引入约束 $\hat{d}_1 = \hat{d}_2$,如下确定不同强度通道(例如,第一强度通道和第二强度通道)的公共偏移系数:

$$[0170] \quad \bar{x} = \frac{1}{2C} \sum_{c=1}^C (x_{C,1} + x_{C,2}) \quad \text{中间术语 (1.1)}$$

$$[0171] \quad \overline{xx} = \frac{1}{C} \sum_{c=1}^C (x_{c,1}^2 + x_{c,2}^2) \quad \text{中间术语 (2.1)}$$

$$[0172] \quad \bar{e} = \frac{1}{2C} \sum_{c=1}^C (e_{c,1} + e_{c,2}) \quad \text{中间术语 (3.1)}$$

$$[0173] \quad \overline{xe} = \frac{1}{C} \sum_{c=1}^C (x_{c,1}e_{c,1} + x_{c,2}e_{c,2}) \quad \text{中间术语 (4.1)}$$

$$[0174] \quad \hat{a} = 1 + \frac{\overline{xe} - 2\bar{x}\bar{e}}{\overline{xx} - 2\bar{x}^2} \quad \text{等式 (24)}$$

$$[0175] \quad \hat{d}_1 = \bar{e} + \bar{x}(1 - \hat{a}) \quad \text{等式 (25)}$$

$$[0176] \quad \hat{d}_2 = \bar{e} + \bar{x}(1 - \hat{a}) \quad \text{等式 (26)}$$

[0177] 对本领域技术人员来说显而易见的是,在测序运行之前执行最小二乘解300以确定闭合形式表达式。一旦确定,闭合形式表达式被应用于在逐簇测序运行期间生成的强度值,并在测序运行的每个测序循环处迭代。

[0178] 强度校正参数

[0179] 以下讨论集中在如何在当前测序循环处针对目标簇确定六个强度校正参数,即第一强度通道中的分布强度、第二强度通道中的分布强度、第一强度通道中的强度误差、第二强度通道中的强度误差、分布质心到原点的距离以及分布强度到强度误差相似性度量。

[0180] 对本领域技术人员来说显而易见的是,强度校正参数的数量将根据强度通道的数量而变化。例如,在四通道路体实施的情况下,将分别针对四个强度通道计算四个通道特异性分布强度和四个通道特异性强度误差。

[0181] 图4示出了如何在当前测序循环*i*处针对目标簇测量通道特异性分布强度的示例400。在图4中,符号“**×**”表示在当前测序循环*i*和先前测序循环1至*i*-1处针对簇1记录的第一强度通道和第二强度通道中的强度值。

[0182] 在图4中,四个强度分布C 402、A 406、G 462和T 466被连接以形成针对簇1的星座102。在图4中,符号“*****”表示在当前测序循环*i*处针对簇1记录的第一强度通道和第二强度通道中的测量强度“m、n”422。由于测量强度“m、n”422最接近强度分布C 402的质心414,簇1属于强度分布C 402,因此在当前测序循环*i*处被分配了碱基检出C。

[0183] 此外,由于簇1属于C强度分布402,所以质心414处的强度值“a、b”是簇1在当前测序循环*i*处的分布强度。另外,“a”是第一强度通道的通道特异性分布强度,并且“b”是第二强度通道的通道特异性分布强度。

[0184] 图5示出了如何在当前测序循环*i*处针对目标簇计算通道特异性强度误差的示例500。在当前测序循环*i*处针对簇1计算第一强度通道中的强度误差($\overline{er}_1$)532,以作为第一强度通道中的通道特异性测量强度(m)和第一强度通道中的通道特异性分布强度(a)之间的差异:

$$[0185] \quad \overline{er}_1 = m - a \quad \text{等式 (27)}$$

[0186] 第二强度通道中的强度误差($\overline{er}_2$) 502在当前测序循环*i*处针对簇1计算,以作为

第二强度通道中的通道特异性测量强度 (n) 和第二强度通道中的通道特异性分布强度 (b) 之间的差异:

$$[0187] \quad \overline{er}_2 = n - b \quad \text{等式 (28)}$$

[0188] 图6示出了如何在当前测序循环i处针对目标簇计算分布质心到原点的距离的示例600。簇1属于C强度分布402,并且质心414处的强度值“a、b”是在当前测序循环i处的针对簇1的分布强度。

[0189] 在当前测序循环i处,将簇1的分布质心到原点的距离计算为质心414和原点132“x,y”之间的欧几里得距离 (d) 652:

$$[0190] \quad d = \sqrt{(a - x)^2 + (b - y)^2} \quad \text{等式 (29)}$$

[0191] 在当前测序循环i处,针对簇1计算分布强度-强度误差相似性度量,以作为通道特异性分布强度和通道特异性强度误差之间的逐通道点积的总和:

$$[0192] \quad \text{相似性度量} = \overline{er}_1 \bullet a + \overline{er}_2 \bullet b \quad \text{等式 (30)}$$

[0193] 其中:

[0194] • 是点积运算符

[0195] 碱基检出流水线

[0196] 图7描绘了实现变化校正逻辑的碱基检出流水线700的另一个示例。考虑当前测序循环i是测序运行的第25个测序循环,即i=25。先前测序循环i-1是测序运行的第24个测序循环,即i-1=24。下一个测序循环i+1是测序运行的第26个测序循环,即i+1=26。随后测序循环i+2是测序运行的第27个测序循环,即i+2=27。

[0197] 先前测序循环

[0198] 在第1个测序循环至第24个测序循环中的每一者中,从相应组的强度校正参数中确定相应组的累加强度校正参数。目标簇的先前累加强度校正参数702是24组强度校正参数的强度校正逐参数累加。在双通道具体实施中,24组强度校正参数中的每组强度校正参数包括六个强度校正参数,即,第一强度通道中的分布强度、第二强度通道中的分布强度、第一强度通道中的强度误差、第二强度通道中的强度误差、分布质心到原点的距离以及分布强度到强度误差的相似度。24组累加强度校正参数中的每组累加强度校正参数包括6个累加强度校正参数 $\bar{x}_1$ 、 $\bar{x}_2$ 、 $\bar{e}_1$ 、 $\bar{e}_2$ 、 $\overline{xx}$ 和 $\overline{xe}$ 。

[0199] 先前累加强度校正参数702是关于底层先前强度值和先前强度校正参数的元数据(或统计数据),从这些参数对其进行计算。结果,与底层先前强度值和先前强度校正参数相比,先前累加强度校正参数702具有小得多的存储器印迹。先前累加强度校正参数702在测序运行期间被高速缓存在存储器中,并且与针对目标簇的当前强度校正参数732累加,以生成针对目标簇的当前累加强度校正参数742,如三角形734所描绘的。

[0200] 在一个具体实施中,先前累加强度校正参数702以量化的固定位宽格式存储。例如,一个或两个字节可用于将每个先前累加强度校正参数存储在先前累加强度校正参数702中。

[0201] 当前测序循环

[0202] 针对目标簇的当前测量强度712包括在第25个测序循环处针对目标簇记录的强度值。基于当前测量强度712,在第25个测序循环处针对目标簇对当前碱基检出722进行检出

(例如,通过使用期望最大化算法)。

[0203] 下一个测序循环

[0204] 基于当前碱基检出722,针对目标簇确定当前强度校正参数732。当前累加强度校正参数742是基于累加先前累加强度校正参数702和当前强度校正参数732来针对目标簇计算的,如三角形734所描绘的。累加的一个示例是求和。在求和具体实施中,当前累加强度校正参数742是通过基于强度校正参数对先前累加强度校正参数702和当前强度校正参数732求和来计算的(如以上中间术语1至6所示)。

[0205] 累加的另一个示例是求平均:

$$[0206] \quad \bar{x}_1 = \frac{1}{C} \sum_{c=1}^C x_{c,1} \quad \text{中间术语 (1.2)}$$

$$[0207] \quad \bar{x}_2 = \frac{1}{C} \sum_{c=1}^C x_{c,2} \quad \text{中间术语 (2.2)}$$

$$[0208] \quad \bar{e}_1 = \frac{1}{C} \sum_{c=1}^C e_{c,1} \quad \text{中间术语 (3.2)}$$

$$[0209] \quad \bar{e}_2 = \frac{1}{C} \sum_{c=1}^C e_{c,2} \quad \text{中间术语 (4.2)}$$

$$[0210] \quad \overline{xx} = \frac{1}{C} \sum_{c=1}^C (x_{c,1}^2 + x_{c,2}^2) \quad \text{中间术语 (5.2)}$$

$$[0211] \quad \overline{xe} = \frac{1}{C} \sum_{c=1}^C (x_{c,1}e_{c,1} + x_{c,2}e_{c,2}) \quad \text{中间术语 (6.2)}$$

[0212] 其中C是当前测序循环i的索引,即在本文所讨论的示例中C=25

[0213] 基于中间术语1.2至6.2,我们将累加强度校正参数中的每个累加强度校正参数定义如下:

[0214] 1) 第一累加强度校正参数 $\bar{x}_1$ 是在先前测序循环1至i-1和当前测序循环i的每一者处针对目标簇测量的第一强度通道中的分布强度的平均值。

[0215] 2) 第二累加强度校正参数 $\bar{x}_2$ 是在先前测序循环1至i-1和当前测序循环i的每一者处针对目标簇测量的第二强度通道中的分布强度的平均值。

[0216] 3) 第三累加强度校正参数 $\bar{e}_1$ 是在先前测序循环1至i-1和当前测序循环i的每一者处针对目标簇计算的第一强度通道中的强度误差的平均值。

[0217] 4) 第四累加强度校正参数 $\bar{e}_2$ 是在先前测序循环1至i-1和当前测序循环i的每一者处针对目标簇计算的第二强度通道中的强度误差的平均值。

[0218] 5) 第五累加强度校正参数 $\overline{xx}$ 是在先前测序循环1至i-1和当前测序循环i的每一者处针对目标簇计算的分布质心到原点的距离的平均值。

[0219] 6) 第六累加强度校正参数 $\overline{xe}$ 是在先前测序循环1至i-1和当前测序循环i的每一者处针对目标簇计算的分布强度-强度误差相似性度量的平均值。

[0220] 紧凑表示

[0221] 在一个具体实施中, 先前累加强度校正参数702存储在紧凑表示中(例如, 求和表示或求平均表示)。在求平均具体实施中, 先前累加强度校正参数702被存储在其求平均表示中, 并且首先乘以其被累加的测序循环的数量, 以检索预平均表示, 即, 24是本文所讨论的示例中的乘数。

[0222] 然后, 在强度校正参数的基础上, 将相乘的结果, 即预平均表示与当前强度校正参数732进行求和。然后, 求和的结果除以当前测序循环i的索引C (C=25), 以确定当前累加强度校正参数742。

[0223] 考虑 $\bar{x}_1^{24}$ 是第24个测序循环的第一累加强度校正参数。考虑 $x_{c,1}^{25}$ 是针对第25个测序循环的第一强度通道中的分布强度。考虑 $\bar{x}_1^{25}$ 是针对第25个测序循环的第一累加强度校正参数, 并用于校正第26个测序循环的测量强度。然后,

$$[0224] \quad \bar{x}_1^{25} = \frac{[(\bar{x}_1^{24} * 24) + x_{c,1}^{25}]}{25}$$

[0225] 考虑 $\bar{x}_2^{24}$ 是第24个测序循环的第二累加强度校正参数。考虑 $x_{c,2}^{25}$ 是针对第25个测序循环的第二强度通道中的分布强度。考虑 $\bar{x}_2^{25}$ 是针对第25个测序循环的第二累加强度校正参数, 并用于校正第26个测序循环的测量强度。然后,

$$[0226] \quad \bar{x}_2^{25} = \frac{[(\bar{x}_2^{24} * 24) + x_{c,2}^{25}]}{25}$$

[0227] 考虑 $\bar{e}_1^{24}$ 是第24个测序循环的第三累加强度校正参数。考虑 $e_{c,1}^{25}$ 是针对第25个测序循环的第一强度通道中的强度误差。考虑 $\bar{e}_1^{25}$ 是针对第25个测序循环的第三累加强度校正参数, 并用于校正第26个测序循环的测量强度。然后,

$$[0228] \quad \bar{e}_1^{25} = \frac{[(\bar{e}_1^{24} * 24) + e_{c,1}^{25}]}{25}$$

[0229] 考虑 $\bar{e}_2^{24}$ 是针对第24个测序循环的第四累加强度校正参数。考虑 $e_{c,2}^{25}$ 是针对第25个测序循环的第二强度通道中的强度误差。考虑 $\bar{e}_2^{25}$ 是针对第25个测序循环的第四累加强度校正参数, 并用于校正第26个测序循环的测量强度。然后,

$$[0230] \quad \bar{e}_2^{25} = \frac{[(\bar{e}_2^{24} * 24) + e_{c,2}^{25}]}{25}$$

[0231] 考虑 $\overline{xx}^{24}$ 是针对第24个测序循环的第五累加强度校正参数。考虑 $[(x_{c,1}^{25})^2 + (x_{c,2}^{25})^2]$ 是针对第25个测序循环的分布质心到原点的距离。考虑 $\overline{xx}^{25}$ 是针对第25个测序循环的第五累加强度校正参数, 并用于校正第26个测序循环的测量强度。然后,

$$[0232] \quad \overline{xx}^{25} = \frac{[(\overline{xx}^{24} * 24) + [(x_{c,1}^{25})^2 + (x_{c,2}^{25})^2]]}{25}$$

[0233] 考虑 $\overline{xe}^{24}$ 是针对第24个测序循环的第六累加强度校正参数。考虑 $x_{c,1}^{25}e_{c,1}^{25} + (x_{c,2}^{25}e_{c,2}^{25})$ 是第25个测序循环的分布强度-强度误差相似性度量。考虑 $\overline{xe}^{25}$ 是针对第25个测序循环的第六累加强度校正参数, 并用于校正第26个测序循环的测量强度。然后,

$$[0234] \quad \overline{xe}^{25} = \frac{[(\overline{xe}^{24} * 24) + [(x_{c,1}^{25}e_{c,1}^{25}) + (x_{c,2}^{25}e_{c,2}^{25})]]}{25}$$

[0235] 当前累加强度校正参数742是关于当前测量强度712和当前强度校正参数732的元数据(或统计)。结果,与当前测量强度712和当前强度校正参数732相比,当前累加强度校正参数742具有小得多的存储器印迹。当前累加强度校正参数742在测序运行期间被高速缓存在存储器中,并且与针对目标簇的下一个强度校正参数794累加,以生成针对目标簇的下一个累加强度校正参数796,如三角形784所描绘的。

[0236] 在一个具体实施中,当前累加强度校正参数742以量化的固定位宽格式存储。例如,一个或两个字节可用于将每个先前累加强度校正参数存储在当前累加强度校正参数742中。

[0237] 当前累加强度校正参数742用于确定针对目标簇的当前放大系数752。这包括根据当前累加强度校正参数742执行等式23中的闭合形式表达式。

[0238] 当前累加强度校正参数742和当前放大系数752用于确定针对目标簇的当前通道特异性偏移系数762。这包括根据当前累加强度校正参数742和放大系数752执行等式17和18中的闭合形式表达式。

[0239] 当前放大系数752和当前通道特异性偏移系数762用于校正第26个测序循环处针对目标簇测量的下一个测量强度772。在一个具体实施中,校正包括从下一个测量强度772中逐通道减去当前通道特异性偏移系数762,以生成下一个移位强度,并将下一个移位强度除以当前放大系数752,以生成针对目标簇的下一个经过校正的测量强度782。

[0240] 然后,在第26个测序循环处,使用下一个经过校正的测量强度782针对目标簇对下一个碱基检出792进行检出。这是通过将下一个经过校正的测量强度782作为输入提供给碱基检出器212(例如,通过使用期望最大化算法)来实现的。

[0241] 后续测序循环

[0242] 如操作794、796、798和799所例示的,控制器(未示出)迭代针对测序运行的连续测序循环的碱基检出流水线700。例如,针对第27个测序循环,当前累加强度校正参数742用作先前累加强度校正参数702,如三角形784所描绘的。注意,碱基检出流水线700是在逐簇基础上执行的,并且是针对簇群体中的多个簇并行执行的。

[0243] 加权最小二乘解

[0244] 图8示出了本文所描述的加权函数800的一个具体实施。由于最小二乘解300可能需要若干测序循环才能收敛,加权函数800用于衰减测序运行的初始测序循环中的变化校正系数,并用于放大测序运行的稍后测序循环中的变化校正系数。

[0245] 加权函数800工作如下。首先,初始化初始放大系数802和初始偏移系数822和832。在一个具体实施中,初始放大系数802在测序运行的第1个测序循环处被初始化为预定值(例如,“1”),并且初始偏移系数822和832在第1个测序循环处被初始化为预定值(例如,“0”)。加权函数800将初始放大系数802与放大系数806(由最小二乘解300确定)相结合(例如求和),并将初始第一偏移系数822和初始第二偏移系数832与第一偏移系数826和第二偏移系数836(由最小二乘解300确定)相结合,使得放大系数806以及第一偏移系数826和第二偏移系数836在初始测序循环中被衰减,并在稍后测序循环中被放大。

[0246] 在一个具体实施中,加权函数800将初始最小权重(inimin权重)804应用于(例如,

乘以) 初始放大系数802以及初始第一偏移系数822和初始第二偏移系数832,并将最小二乘最大权重 (lsqmax权重) 808应用于放大系数806以及第一偏移系数826和第二偏移系数836,使得:

$$[0247] \quad \text{lsqmax 权重} = \left(0, \frac{c-p}{c}\right) \quad \text{等式 (31)}$$

$$[0248] \quad \text{inimin 权重} = \left(1, 1 - \left[\frac{c-p}{c}\right]\right) \quad \text{等式 (32)}$$

[0249] 其中:

[0250] c是当前测序循环的索引

[0251] p是2到7之间的数

[0252] 针对第1个测序循环,即 $c=1$,并且针对 $p=2$,表达式 $\left(\frac{c-p}{c}\right)$ 等于“-1”。然后,在“0”和“-1”之间,lsqmax权重808选择两个值中的最大值,即0。表达式 $\left(1 - \left[\frac{c-p}{c}\right]\right)$ 等于“2”。然后,在“1”和“2”之间,inimin权重804选择两个值中的最小值,即1。

[0253] 继续向前,来自lsqmax权重808的0乘以放大系数806以及第一偏移系数826和第二偏移系数836,并且来自inimin权重804的1乘以初始放大系数802以及初始第一偏移系数822和初始第二偏移系数832。将两次相乘的结果进行求和以生成加权放大系数810以及加权第一偏移系数820和加权第二偏移系数830。

[0254] 随着测序运行的进行,并且索引“c”的值增加,并且lsqmax权重808和inimin权重804的值也类似地改变和应用,使得放大系数806以及第一偏移系数826和第二偏移系数836(从最小二乘解300获知)在每个连续测序循环处被逐渐放大。

[0255] 加权函数800生成加权放大系数810和加权第一偏移系数820和加权第二偏移系数830,其用于在下一个测序循环 $i+1$ 处校正针对目标簇的测量强度,并在下一个测序循环 $i+1$ 处生成对目标簇进行碱基检出的经过校正的测量强度。

$$[0256] \quad \hat{a} = 1 + \frac{W\bar{x}\bar{e} - \bar{x}_1\bar{e}_1 - \bar{x}_2\bar{e}_2}{W\bar{x}\bar{x} - \bar{x}_1^2 - \bar{x}_2^2} \quad \text{等式(33)}$$

$$[0257] \quad \hat{d}_1 = \frac{-\bar{x}_1 W\bar{x}\bar{e} + (W\bar{x}\bar{x} - \bar{x}_2^2)\bar{e}_1 + \bar{x}_1\bar{x}_2\bar{e}_2}{W(W\bar{x}\bar{x} - \bar{x}_1^2 - \bar{x}_2^2)} \quad \text{等式 (34)}$$

$$[0258] \quad \hat{d}_2 = \frac{-\bar{x}_2 W\bar{x}\bar{e} + (W\bar{x}\bar{x} - \bar{x}_1^2)\bar{e}_2 + \bar{x}_1\bar{x}_2\bar{e}_1}{W(W\bar{x}\bar{x} - \bar{x}_1^2 - \bar{x}_2^2)} \quad \text{等式 (35)}$$

[0259] 其中:

[0260] W是权重

[0261] 最大似然解

[0262] 图9示出了将最大似然权重906、908、和910直接应用于变化校正系数的一个具体实施。最大似然权重906、908和910是通过将最大似然解900应用于在先前测序运行中针对变化校正系数观察到的历史值的概率分布902而生成的。图9还例示了累加强度校正参数904。

[0263] 最大似然权重906、908和910是当前测序循环的函数,如索引“C”所表示的。最大似然权重906、908和910根据索引C在测序循环的基础上变化。最大似然权重906、908和910也

是加性噪声的函数,如字符“n”所表示的。西格玛术语“ σ ”表示针对相应变化校正系数观察到的历史值的变化范围,即方差(σ^2)。在一些具体实施中,可使用最大似然解900来估计针对加性噪声的西格玛术语,或可由用户指定。针对放大系数、通道特异性偏移系数和加性噪声的西格玛术语是在测序运行的基础上确定的,并在测序运行的所有测序循环内保持固定。西格玛术语包含了关于变化校正系数中观察到的不确定性的先验知识。

[0264] 西格玛术语的一些示例值如下:

[0265] 'ml_chanest_西格玛_a',0.15

[0266] 'ml_chanest_西格玛_d1',0.1

[0267] 'ml_chanest_西格玛_d2',0.02

[0268] 'ml_chanest_西格玛_n',0.14

[0269] 在一个具体实施中,针对放大系数的概率分布的中心/初始/平均值被设置为“1”,并且针对通道特异性偏移系数的概率分布的中心/初始/平均值被设置为“0”。

[0270] 针对最大似然权重906、908和910中的放大系数和通道特异性偏移系数的西格玛术语的较小值指示其相应历史值的低变化。这导致最大似然权重906、908和910的更高值。这进而导致加权放大系数920,其被加权以有利于中心值1,以及加权通道特异性偏移系数930和940,其被加权以有利于中心值0,特别是在早期测序循环中。

[0271] 相反,针对最大似然权重906、908和910中的放大系数和通道特异性偏移系数的西格玛术语的较大值指示其相应历史值的高变化。这导致最大似然权重906、908和910的较低值。这进而导致加权放大系数920,其被加权以有利于最小二乘解300的输出(例如,等式23),以及加权通道特异性偏移系数930和940,其被加权以有利于最小二乘解300的输出(例如,等式17和18),特别是在稍后测序循环中。

[0272] 最大似然权重906、908和910被直接合并以分别计算加权放大系数920和加权通道特异性偏移系数930和940。

[0273] 指数衰减因子解

[0274] 图10示出了将指数衰减因子应用于变化校正系数的一个具体实施。指数衰减逻辑1000基于所谓的“tau”和“统计.循环”。术语“统计.循环”是指当前测序循环。

[0275] 根据在强度校正参数中观察到的时间变化程度,Tau被设置为预定值。如果强度校正参数是时间不变的,则tau可被设置为无穷大。如果强度校正参数是快速时变的,则tau可被设置为小值。在一个具体实施中,tau被设置为32。

[0276] 考虑tau是32。然后,根据语句1002、1004和1006,针对第1个测序循环至第31个测序循环,衰减因子是“1”,并且这导致累加强度校正参数没有衰减。根据语句1008,针对32个及以上的测序循环,衰减因子为31/32。指数衰减特征来自于以下事实:在语句1010、1012、1014、1016、1018和1020中,累加强度校正参数中的每个累加强度校正参数在每个连续测序循环处乘以衰减因子,如下所示:

[0277] 针对测序循环32,衰减因子 $=\frac{31}{32}$

[0278] $\bar{x}_1^D = \frac{31}{32} * \bar{x}_1$ 中间术语 (1.3)

[0279] $\bar{x}_2^D = \frac{31}{32} * \bar{x}_2$ 中间术语 (2.3)

$$[0280] \quad \bar{e}_1^D = \frac{31}{32} * \bar{e}_1 \quad \text{中间术语 (3.3)}$$

$$[0281] \quad \bar{e}_2^D = \frac{31}{32} * \bar{e}_2 \quad \text{中间术语 (4.3)}$$

$$[0282] \quad \overline{xx^D} = \frac{31}{32} * \overline{xx} \quad \text{中间术语 (5.3)}$$

$$[0283] \quad \overline{xe^D} = \frac{31}{32} * \overline{xe} \quad \text{中间术语 (6.3)}$$

$$[0284] \quad \text{针对测序循环33, 衰减因子} = \left[\frac{31}{32} \right]^2$$

$$[0285] \quad \bar{x}_1^D = \left[\frac{31}{32} \right]^2 * \bar{x}_1 \quad \text{中间术语 (1.4)}$$

$$[0286] \quad \bar{x}_2^D = \left[\frac{31}{32} \right]^2 * \bar{x}_2 \quad \text{中间术语 (2.4)}$$

$$[0287] \quad \bar{e}_1^D = \left[\frac{31}{32} \right]^2 * \bar{e}_1 \quad \text{中间术语 (3.4)}$$

$$[0288] \quad \bar{e}_2^D = \left[\frac{31}{32} \right]^2 * \bar{e}_2 \quad \text{中间术语 (4.4)}$$

$$[0289] \quad \overline{xx^D} = \left[\frac{31}{32} \right]^2 * \overline{xx} \quad \text{中间术语 (5.4)}$$

$$[0290] \quad \overline{xe^D} = \left[\frac{31}{32} \right]^2 * \overline{xe} \quad \text{中间术语 (6.4)}$$

[0291] 在图10中,累加强度校正参数使用求和运算来累加。在指数衰减因子的求平均具体实施中,使用平均运算来累加累加强度校正参数。在指数衰减因子的求平均具体实施中,中间术语1.2至6.2中的除数“C”在测序运行的tau个测序循环之后,即在测序运行的第32个测序循环之后保持固定。

[0292] 在一些具体实施中,加权最小二乘解(图8)、最大似然解(图9)和指数衰减因子解(图10)相结合,在测序运行的每个测序循环处生成加权变化校正系数。

[0293] 通道特异性偏移系数

[0294] 图11示出了确定通道特异性偏移系数的另一种具体实施。在双通道具体实施中,在当前测序循环处,针对第一强度通道并针对目标簇的第一通道特异性偏移系数(“ Δx ”)被计算为在当前测序循环处针对目标簇的第一强度通道中的测量强度(“p”)与在当前测序循环处目标簇所属的碱基特异性强度分布A 1102的质心1104处的第一强度通道中的强度值(“u”)之间的差异(例如,由期望最大化算法确定)。

[0295] 在双通道具体实施中,在当前测序循环处,针对第二强度通道并针对目标簇的第二通道特异性偏移系数(“ Δy ”)被计算为在当前测序循环处针对目标簇的第二强度通道中的测量强度(“q”)与在当前测序循环处目标簇所属的碱基特异性强度分布A 1102的质心1104处的第二强度通道中的强度值(“v”)之间的差异。

[0296] 在一个具体实施中,在测序运行的每个测序循环处确定第一通道特异性偏移系数(“ Δx ”)和第二通道特异性偏移系数(“ Δy ”)。在一些具体实施中,在可配置数量的测序循

环(例如,十个测序循环或二十个测序循环)之后,第一通道特异性偏移系数(“ Δx ”)和第二通道特异性偏移系数(“ Δy ”)被初始化为预定值(例如,“0”)。

[0297] 在移动平均值具体实施中,在可配置数量的测序循环之后,针对第一偏移通道特异性系数(“ Δx ”)和第二通道特异性偏移系数(“ Δy ”)计算平均值。然后,该平均值被用作针对第一偏移通道特异性系数(“ Δx ”)和第二通道特异性偏移系数(“ Δy ”)的替代,直到针对下一组可配置数量的测序循环计算出下一个平均值。

[0298] 在一些具体实施中,仅当目标簇属于A碱基特异性强度分布、C碱基特异性强度分布和T碱基特异性强度分布时,而不是当目标簇属于G(暗)碱基特异性强度分布时,才计算第一通道特异性偏移系数(“ Δx ”)和第二通道特异性偏移系数(“ Δy ”)。在具体实施中,当测序运行涉及双端读取时,第一通道特异性偏移系数(“ Δx ”)和第二通道特异性偏移系数(“ Δy ”)被初始化用于第二次读取,具有在第一次读取结束时可用的值,但是此后在每组可配置数量的测序循环中被更新。

[0299] 性能结果

[0300] 图12、图13和图14比较了三种方法的性能,即仅缩放解、仅偏移解(在图11中讨论)和最小二乘解300。这三种方法应用于使用Illumina的实时分析(RTA)软件从Illumina的测序器NextSeq 2000的二十个数据集生成的强度数据。

[0301] 在图12中,仅缩放解(蓝色)、仅偏移解(橙色)和最小二乘解(灰色)的性能相对于通过RTA的双通道chastity滤波器的簇百分比进行绘制。比较是在二十个数据集上完成的(示出为x轴)。虽然所有三种方法都实现了>65%的簇通过双通道chastity滤波器,但是20个案例中的16个(80%)得分高于75%的通过率,并且8/20(或20%)得分>80%的通过率。注意,最小二乘解300表现最佳。

[0302] 在图13中,针对低多样性样品中的误差率,比较绘制了仅缩放解(蓝色)、仅偏移解(橙色)和最小二乘解(灰色)的性能。该比较是在掺有已知噬菌体基因组(PhiX)(示出为x轴)的二十个低多样性数据集上完成的。二十个中有十七个(17/20或68%)实现了<35%的误差率,而大多数享有低于25%的误差率。注意,最小二乘解300表现最佳。

[0303] 在图14中,针对具有高于Q30(即碱基检出误差 $\leq 10^{-3}$ 或0.1%)的质量得分的测序数据百分比,比较绘制了仅缩放解(蓝色)、仅偏移解(橙色)和最小二乘解(灰色)的性能。比较是在相同的二十个数据集上完成的(示出为x轴)。虽然所有三种方法都享有较高的Q30质量得分(例如>80%),但20个案例中有16个(80%)得分高于75%的通过率,并且8/20(或20%)得分>80%的通过率。注意,最小二乘解300优于其他解2%以上。

[0304] 计算机系统

[0305] 图15是可以用于实现本发明所公开的技术的计算机系统1500。计算机系统1500包括经由总线子系统1555与多个外围设备通信的至少一个中央处理单元(CPU) 1572。这些外围设备可以包括存储子系统1510,该存储子系统包括例如存储器设备和文件存储子系统1536、用户界面输入设备1538、用户界面输出设备1576和网络接口子系统1574。输入设备和输出设备允许用户与计算机系统1500进行交互。网络接口子系统1574提供通向外部网络的接口,该接口包括通向其他计算机系统1500中的对应接口设备的接口。

[0306] 在一个具体实施中,最小二乘解确定器302可通信地链接到存储子系统1510和用户界面输入设备1538。

[0307] 用户界面输入设备1538可包括键盘;指向设备,诸如鼠标、轨迹球、触摸板或图形输入板;扫描仪;结合到显示器中的触摸屏;音频输入设备,诸如语音识别系统和麦克风;以及其他类型的输入设备。一般来讲,使用术语“输入设备”旨在包括将信息输入到计算机系统1500中的所有可能类型的设备和方式。

[0308] 用户界面输出设备1576可以包括显示子系统、打印机、传真机或非视觉显示器(诸如音频输出设备)。显示子系统可包括LED显示器、阴极射线管(CRT)、平板设备诸如液晶显示器(LCD)、投影设备或用于产生可见图像的一些其他机构。显示子系统还可提供非视觉显示器,诸如音频输出设备。一般来讲,使用术语“输出设备”旨在包括将信息从计算机系统1500输出到用户或者输出到另一机器或计算机系统的所有可能类型的设备和方式。

[0309] 存储子系统1510存储提供本文描述的模块和方法中的一些或全部模块和方法的功能的编程结构和数据结构。这些软件模块通常由处理器1578执行。

[0310] 处理器1578可以是图形处理单元(GPU)、现场可编程门阵列(FPGA)、专用集成电路(ASIC)和/或粗粒度可重构架构(CGRA)。处理器1578可由深度学习云平台(诸如Google Cloud Platform™、Xilinx™和Cirrascle™)托管。处理器1578的示例包括Google的Tensor Processing Unit(TPU)™、Rackmount解决方案(如GX4 Rackmount Series™、GX15Rackmount Series™)、NVIDIA DGX-1™、Microsoft的Stratix V FPGA™、Graphcore的Intelligent Processor Unit(IPU)™、Qualcomm的具有Snapdragon processors™的Zeroth Platform™、NVIDIA的Volta™、NVIDIA的DRIVE PX™、NVIDIA的JETSON TX1/TX2 MODULE™、Intel的Nirvana™、Movidius VPU™、Fujitsu DPI™、ARM的DynamicIQ™、IBM TrueNorth™、具有Testa V100s™的Lambda GPU服务器,等等。

[0311] 在存储子系统1510中使用的存储器子系统1522可包括多个存储器,包括用于在程序执行期间存储指令和数据的主随机存取存储器(RAM) 1532和其中存储固定指令的只读存储器(ROM) 1534。文件存储子系统1536可以为程序文件和数据文件提供持久性存储,并且可以包括硬盘驱动器、软盘驱动器以及相关联的可移动介质、CD-ROM驱动器、光盘驱动器或可移动介质磁盘盒。实现某些具体实施的功能的模块可由文件存储子系统1536存储在存储子系统1510中,或存储在处理器可访问的其他机器中。

[0312] 总线子系统1555提供用于使计算机系统1500的各种部件和子系统按照预期彼此通信的机构。尽管总线子系统1555被示意性地示出为单条总线,但是总线子系统的替代性具体实施可使用多条总线。

[0313] 计算机系统1500本身可具有不同类型,包括个人计算机、便携式计算机、工作站、计算机终端、网络计算机、电视机、主机、服务器群、一组广泛分布的松散联网的计算机,或者任何其他数据处理系统或用户设备。由于计算机和网络的不断变化的性质,对图15中描绘的计算机系统1500的描述仅旨在作为用于展示本发明的优选实施方式的具体示例。计算机系统1500的许多其他配置也是可能的,其具有比图15中描绘的计算机系统更多或更少的部件。

[0314] 本文所讨论的处理器或模块中的每一者可包括用于执行特定过程的算法(例如,存储在有形和/或非暂态计算机可读存储介质上的指令)或子算法。变化校正器232在概念上被例示为模块的集合,但可利用专用硬件板、DSP、处理器等的任何组合来实现。另选地,变化校正器232可利用具有单个处理器或多个处理器的现成PC来实现,其中功能操作分布

在处理器之间。作为进一步的选择,下文所述的模块可利用混合配置来实现,其中某些模块化功能利用专用硬件来执行,而其余模块化功能利用现成PC等来执行。模块还可被实现为处理单元内的软件模块。

[0315] 本文所阐述的方法的各种过程和步骤(例如,图9)可使用计算机来执行。计算机可包括处理器,该处理器是检测设备的一部分,与用于获得由计算机处理的数据的检测设备联网,或与检测设备分离。在一些具体实施中,信息(例如,图像数据)可直接或经由计算机网络在本文所公开的系统的部件之间传输。局域网(LAN)或广域网(WAN)可以是公司计算网络,包括对因特网的访问,包括该系统的计算机和计算设备连接到该网络。在一个具体实施中,LAN符合传输控制协议/因特网协议(TCP/IP)行业标准。在一些情况下,信息(例如,图像数据)经由输入设备(例如,磁盘驱动器、光盘播放器、USB端口等)输入到本文所公开的系系统)。在一些情况下,通过例如从存储设备,诸如磁盘或闪存驱动器加载信息来接收信息。

[0316] 用于运行本文所阐述的算法或其他过程的处理器可包括微处理器。微处理器可以是任何常规的通用单芯片或多芯片微处理器,诸如由Intel Corporation制造的奔腾™处理器。特别有用的计算机可利用英特尔艾维布里奇双12核处理器、LSI raid控制器、128GB的RAM和2TB固态硬盘。另外,处理器可包括任何常规的专用处理器,诸如数字信号处理器或图形处理器。处理器通常具有常规的地址线、常规的数据线以及一个或多个常规的控制线。

[0317] 本文所公开的具体实施可被实现为使用标准编程或工程技术来产生软件、固件、硬件或它们的任何组合的方法、装置、系统或制品(例如,图2和图7)。如本文所用,术语“制品”是指在硬件或计算机可读介质,诸如光学存储设备和易失性或非易失性存储器设备中实现的代码或逻辑。此类硬件可包括但不限于现场可编程门阵列(FPGA)、专用集成电路(ASIC)、复杂可编程逻辑器件(CPLD)、可编程逻辑阵列(PLA)、微处理器或其他类似的处理设备。在特定具体实施中,本文阐述的信息或算法存在于非暂态存储介质中。

[0318] 在特定具体实施中,本文所阐述的计算机实现的方法(例如,(在图11中讨论))可在获得对象的多个图像时实时发生。此类实时分析尤其可用于核酸测序应用,其中核酸阵列经受流体和检测步骤的重复循环。测序数据的分析通常可能是计算密集型的,使得在进行其他数据采集或分析算法时实时或在后台执行本文所述的方法可能是有益的。可与本发明方法一起使用的示例性实时分析方法是用于可从Illumina公司(San Diego, Calif.)商购获得和/或描述于美国专利申请公布2012/0020537A1中的MiSeq和HiSeq测序设备的那些方法,该专利申请公布以引用方式并入本文。

[0319] 在本专利申请中,术语“簇”、“孔”、“样品”、“分析物”和“荧光样品”是可互换的,因为孔包含对应的簇/样品/分析物/荧光样品。如本文所定义,“样品”及其衍生物以其最广泛的意义使用,包括怀疑包含目标的任何标本、培养物等。在一些实施方式中,样品包括DNA、RNA、PNA、LNA、嵌合或杂交形式的核酸。样品可包括含有一种或多种核酸的任何基于生物、临床、外科、农业、大气或水生动植物的标本。该术语还包括任何分离的核酸样品,诸如基因组DNA、新鲜冷冻或福尔马林固定石蜡包埋的核酸标本。还设想样品的来源可以是:单个个体、来自遗传相关成员的核酸样品的集合、来自遗传不相关成员的核酸样品、来自单个个体的(与之匹配的)核酸样品(诸如肿瘤样品和正常组织样品),或者来自含有两种不同形式的遗传物质(诸如从母体受试者获得的母体DNA和胎儿DNA)的单个来源的样品,或者在含有植物或动物DNA的样品中存在污染性细菌DNA。在一些实施方式中,核酸材料的来源可以包括

从新生儿获得的核酸,例如通常用于新生儿筛检的核酸。

[0320] 该核酸样品可以包括高分子量物质,诸如基因组DNA(gDNA)。该样品可以包括低分子量物质,诸如从FFPE样品或存档的DNA样品获得的核酸分子。在另一种实施方式中,低分子量物质包括酶促片段化或机械片段化的DNA。该样品可以包含无细胞循环DNA。在一些实施方式中,该样品可以包括从活检组织、肿瘤、刮取物、拭子、血液、黏液、尿液、血浆、精液、毛发、激光捕获显微解剖、手术切除和其他临床或实验室获得的样品获得的核酸分子。在一些具体实施中,该样品可以是流行病学样品、农业样品、法医学样品或病原性样品。在一些实施方式中,该样品可以包括从动物(诸如人或哺乳动物来源)获得的核酸分子。在另一种具体实施中,该样品可包括从非哺乳动物来源(诸如植物、细菌、病毒或真菌)获得的核酸分子。在一些实施方式中,核酸分子的来源可以是存档或灭绝的样品或物种。

[0321] 另外,本文所公开的方法和组合物可以用于扩增具有低质量核酸分子的核酸样品,诸如来自法医学样品的降解的和/或片段化的基因组DNA。在一种实施方式中,法医学样品可以包括从犯罪现场获得的核酸、从失踪人员DNA数据库获得的核酸、从与法医调查相关联的实验室获得的核酸,或者包括由执法机关、一种或多种军事服务或任何此类人员获得的法医学样品。核酸样品可以是经纯化的样品或含有粗DNA的溶胞产物,例如来源于口腔拭子、纸、织物或者其他可以用唾液、血液或其他体液浸渍的基材。因此,在一些实施方式中,核酸样品可以包含少量DNA(诸如基因组DNA),或者DNA的片段化部分。在一些具体实施中,目标测序可存在于一种或多种体液中,其中体液包括但不限于血液、痰、血浆、精液、尿液和血清。在一些具体实施中,目标测序可获自受害者的毛发、皮肤、组织样品、尸体解剖或遗骸。在一些实施方式中,包括一种或多种目标序列的核酸可以获自死亡的动物或人。在一些实施方式中,目标序列可以包括从非人DNA(诸如微生物、植物或昆虫DNA)获得的核酸。在一些实施方式中,目标序列或扩增的目标序列涉及人类身份识别的目的。在一些实施方式中,本公开整体涉及用于识别法医学样品的特征的方法。在一些实施方式中,本公开整体涉及使用本文所公开的一种或多种目标特异性引物或者用本文概述的引物设计标准设计的一种或多种目标特异性引物的人类身份识别方法。在一种实施方式中,含有至少一种目标序列的法医学样品或人类身份识别样品可以使用本文所公开的任何一种或多种目标特异性引物或者使用本文概述的引物标准进行扩增。

[0322] 所公开的技术生成变化校正系数以校正图像数据中的簇间强度分布变化。本发明所公开的技术可以作为系统、方法或制品来实践。具体实施的一个或多个特征可与基本具体实施组合。不互相排斥的具体实施被教导为可组合的。具体实施的一个或多个特征可与其他具体实施组合。本公开周期性地提醒用户这些选项。从一些具体实施中省略重复这些选项的表述不应被视为限制前述部分中教导的组合,这些表述将据此以引用方式并入以下具体实施中的每个具体实施中。

[0323] 在本部分中描述的方法的其他具体实施可包括存储指令的非暂态计算机可读存储介质,这些指令可由处理器执行以执行上述方法中的任一种方法。在本部分中描述的方法的又一具体实施可包括一种系统,该系统包括存储器和一个或多个处理器,该一个或多个处理器可操作以执行存储在存储器中的指令,以执行上述方法中的任一种方法。

[0324] 在另一种具体实施中,对非强度数据进行变化校正,诸如在分子延伸期间由氢离子释放所诱导的pH变化。检测pH变化并将其转化为与引入的碱基的数量成比例的电压变化

(例如,就Ion Torrent而言)。

[0325] 在又另一种具体实施中,非强度数据根据纳米孔感测构造而成,该纳米孔感测使用生物传感器来测量当分析物穿过纳米孔或靠近其孔口时电流的中断,同时确定碱基的种类。例如,牛津纳米孔技术(ONT)测序基于以下概念:使单链DNA(或RNA)经由纳米孔穿过膜,并且跨膜施加电压差。孔中存在的核苷酸将影响孔的电阻,因此随时间推移的电流测量结果可指示DNA碱基穿过孔的序列。该电流信号(由于其在绘制时的外观而被称为“波形曲线(squiggle)”)是由ONT测序器收集的原始数据。这些测量结果被存储为在(例如)4kHz频率下获取的16位整数数据采集(DAC)值。在DNA链速度为约450碱基对/秒的情况下,这给出了平均每种碱基大约九个原始观察结果。然后处理该信号以识别对应于各个读数的开孔信号的中断。对原始信号的这些最大限度的利用是进行碱基检出,即将DAC值转换成DNA碱基序列的过程。在一些具体实施中,非强度数据包括归一化或缩放的DAC值。

[0326] 可以以计算机产品的形式实现所公开的技术的一个或多个具体实施或其元素,该计算机产品包括具有用于执行所指示的方法步骤的计算机可用程序代码的非暂态计算机可读存储介质。此外,所公开的技术的一个或多个具体实施或其元素可以以包括存储器的装置和耦合到存储器的至少一个处理器的形式来实现,并且操作以执行示例性方法步骤。进一步地,在另一方面,可以以用于执行本文描述的方法步骤中的一个或多个方法步骤的装置的形式来实现所公开的技术的一个或多个具体实施或其元素;该装置可包括(i)硬件模块,(ii)在一个或多个硬件处理器上执行的软件模块,或(iii)硬件和软件模块的组合;(i)至(iii)中的任何一个实施本文所阐述的特定技术,并且软件模块被存储在计算机可读存储介质(或多个此类介质)中。

[0327] 本专利申请可互换地使用术语“累加强度校正参数”和“中间术语”。

[0328] 本专利申请可互换地使用术语“一个或多个纯强度”和“一个或多个分布强度”。

[0329] 本专利申请可互换地使用术语“强度分布”和“星座”。

[0330] 本专利申请可互换地使用术语“变化校正系数”和“强度校正系数”。

[0331] 本专利申请可互换地使用术语“放大系数”和“缩放因子”。

[0332] 本专利申请可互换地使用术语“偏移系数”和“偏移”。

[0333] 本专利申请可互换地使用术语“目标簇”和“特定簇”。

[0334] 本专利申请可互换地使用术语“下一个”、“随后”和“连续”。

[0335] 本专利申请可互换地使用术语“核苷酸”和“碱基”。

[0336] 本专利申请可互换地使用术语“累加强度校正参数确定器”和“累加器”。

[0337] 本专利申请可互换地使用术语“累加强度校正参数确定器”和“累加器”。

[0338] 条款

[0339] 1. 一种碱基检出目标簇的计算机实现的方法,所述方法包括:

[0340] 针对目标簇,

[0341] 从碱基特异性强度分布中读取针对测序运行的当前测序循环记录的当前通道特异性强度,在所述当前测序循环处将所述目标簇碱基检出到所述碱基特异性强度分布,

[0342] 从所述碱基特异性强度分布的质心读取当前通道特异性分布强度,

[0343] 基于所述当前通道特异性强度和所述当前通道特异性分布强度,确定用于所述当前测序循环的一组当前强度校正参数,通过累加用于所述测序运行的先前测序循环的所述

一组当前强度校正参数和一组先前累加强度校正参数,确定用于所述当前测序循环的一组当前累加强度校正参数,

[0344] 基于所述一组当前累加强度校正参数确定用于所述当前测序循环的当前放大系数和当前通道特异性偏移系数,以及使用所述当前放大系数和所述当前通道特异性偏移系数来校正针对所述测序运行的下一个测序循环记录的下一个通道特异性强度,并针对所述下一个测序循环生成经过校正的下一个通道特异性强度;以及

[0345] 基于所述经过校正的下一个通道特异性强度,在所述下一个测序循环处碱基检出所述目标簇。

[0346] 2.根据条款1所述的计算机实现的方法,其中所述一组当前强度校正参数包括当前通道特异性分布强度、当前通道特异性强度误差、当前分布质心到原点的距离和当前分布强度到强度误差相似性度量。

[0347] 3.根据条款2所述的计算机实现的方法,其中所述当前通道特异性强度误差是所述当前通道特异性强度和所述当前通道特异性分布强度之间的逐通道差异。

[0348] 4.根据条款2所述的计算机实现的方法,其中所述当前分布质心到原点的距离是多维空间的所述质心和原点之间的欧几里得距离,所述多维空间包含所述碱基特异性强度分布。

[0349] 5.根据条款4所述的计算机实现的方法,其中所述多维空间是笛卡尔空间、极空间、柱形空间和球形空间中的至少一者。

[0350] 6.根据条款2所述的计算机实现的方法,其中所述当前分布强度到强度误差相似性度量是所述当前通道特异性分布强度和所述当前通道特异性强度误差之间的通道式点积的总和。

[0351] 7.根据条款1所述的计算机实现的方法,其中所述一组当前累加强度校正参数是所述一组当前强度校正参数中的当前强度校正参数和所述一组先前累加强度校正参数中的先前累加强度校正参数的强度校正参数式总和。

[0352] 8.根据条款7所述的计算机实现的方法,其中所述一组当前累加强度校正参数是所述当前强度校正参数和所述先前累加强度校正参数的强度校正逐参数平均值。

[0353] 9.根据条款1所述的计算机实现的方法,其中所述一组先前累加强度校正参数和所述一组当前累加强度校正参数以量化的固定位宽格式存储。

[0354] 10.根据条款1所述的计算机实现的方法,其中所述当前通道特异性偏移系数被配置为相同。

[0355] 11.根据条款10所述的计算机实现的方法,其中所述一组当前累加强度校正参数中的当前累加强度校正参数包括用于所述当前通道特异性分布强度的第一公共当前累加强度校正参数,以及用于所述当前通道特异性强度误差的第二公共当前累加强度校正参数。

[0356] 12.根据条款1所述的计算机实现的方法,其中从所述下一个通道特异性强度中逐通道地减去所述当前通道特异性偏移系数,以生成下一个通道特异性移位强度,并且所述下一个通道特异性移位强度除以所述当前放大系数,以生成所述经过校正的下一个通道特异性强度。

[0357] 13.根据条款1所述的计算机实现的方法,所述方法还包括使用加权函数来组合初

始放大系数与所述当前放大系数,以及初始通道特异性偏移系数与所述当前通道特异性偏移系数,以生成用于所述当前测序循环的加权当前放大系数和加权当前通道特异性偏移系数。

[0358] 14. 根据条款13所述的计算机实现的方法,其中所述加权函数将最小权重($w_{\min}$)应用于初始放大系数和所述初始通道特异性偏移系数,并将最大权重($w_{\max}$)应用于所述当前放大系数和所述当前通道特异性偏移系数,其中 $w_{\min} = (1 - w_{\max})$ 。

[0359] 15. 根据条款14所述的计算机实现的方法,其中所述最大权重($w_{\max}$)被定义为 $(c - p) / c$,其中 c 是用于所述当前测序循环的索引,并且 p 是2至7之间的数。

[0360] 16. 根据条款15所述的计算机实现的方法,所述方法还包括使用所述加权当前放大系数和所述加权当前通道特异性偏移系数来校正所述下一个通道特异性强度并生成所述经过校正的下一个通道特异性强度。

[0361] 17. 根据条款1所述的计算机实现的方法,还包括:

[0362] 使用最大似然解来针对所述当前测序循环生成用于所述当前放大系数和所述当前通道特异性偏移系数的相应当前最大似然权重;

[0363] 将所述当前最大似然权重分别应用于所述当前放大系数和所述当前通道特异性偏移系数,以生成用于所述当前测序循环的最大似然加权当前放大系数和最大似然加权当前通道特异性偏移系数;以及

[0364] 使用所述最大似然加权当前放大系数和所述最大似然加权当前通道特异性偏移系数来校正所述下一个通道特异性强度,并生成所述经过校正的下一个通道特异性强度。

[0365] 18. 根据条款1所述的计算机实现的方法,还包括:

[0366] 将衰减因子应用于所述当前强度校正参数,以生成用于所述当前测序循环的衰减的当前强度校正参数;以及

[0367] 通过强度校正逐参数累加所述衰减的当前强度校正参数和所述先前累加强度校正参数来确定所述当前累加强度校正参数。

[0368] 19. 根据条款18所述的计算机实现的方法,其中所述衰减因子对于所述测序运行的一定数量的测序循环保持固定,并且此后基于衰减逻辑以指数方式衰减。

[0369] 20. 根据条款19所述的计算机实现的方法,其中所述衰减逻辑是 $1 - 1/\tau$,其中 τ 是预定义的数。

[0370] 21. 根据条款1所述的计算机实现的方法,所述方法还包括在所述测序运行的连续测序循环处针对所述目标簇迭代所述读取、所述读取、所述确定、所述确定、所述确定、所述使用和所述碱基检出。

[0371] 22. 根据条款1所述的计算机实现的方法,所述方法还包括针对多个簇并行执行所述读取、所述读取、所述确定、所述确定、所述确定、所述使用和所述碱基检出。

[0372] 23. 根据条款1所述的计算机实现的方法,其中使用最小二乘解来确定用于所述一组当前强度校正参数、所述一组当前累加强度校正参数、所述当前放大系数和所述当前通道特异性偏移系数的闭合形式表达式。

[0373] 24. 根据条款1所述的计算机实现的方法,其中所述当前通道特异性强度分别对应于强度通道。

[0374] 25. 根据条款1所述的计算机实现的方法,其中所述当前通道特异性偏移系数是所

述当前通道特异性强度和所述当前通道特异性分布强度之间的逐通道差异。

[0375] 26. 一种印有碱基检出目标簇的计算机程序指令的非暂态计算机可读存储介质, 该指令在处理器上执行时实现包括以下各项的方法:

[0376] 针对目标簇,

[0377] 从碱基特异性强度分布中读取针对测序运行的当前测序循环记录的当前通道特异性强度, 在所述当前测序循环处将所述目标簇碱基检出到所述碱基特异性强度分布,

[0378] 从所述碱基特异性强度分布的质心读取当前通道特异性分布强度,

[0379] 基于所述当前通道特异性强度和所述当前通道特异性分布强度, 确定用于所述当前测序循环的一组当前强度校正参数,

[0380] 通过累加用于所述测序运行的先前测序循环的所述一组当前强度校正参数和一组先前累加强度校正参数, 确定用于所述当前测序循环的一组当前累加强度校正参数,

[0381] 基于所述一组当前累加强度校正参数确定用于所述当前测序循环的当前放大系数和当前通道特异性偏移系数, 以及使用所述当前放大系数和所述当前通道特异性偏移系数来校正针对所述测序运行的下一个测序循环记录的下一个通道特异性强度, 并针对所述下一个测序循环生成经过校正的下一个通道特异性强度; 以及

[0382] 基于所述经过校正的下一个通道特异性强度, 在所述下一个测序循环处碱基检出所述目标簇。

[0383] 27. 根据条款26所述的非暂态计算机可读存储介质, 该非暂态计算机可读存储介质实现最终取决于条款1的每个条款。

[0384] 28. 一种包括耦合到存储器的一个或多个处理器的系统, 该存储器装载有碱基检出目标簇的计算机指令, 该指令在该处理器上执行时实现包括以下各项的动作:

[0385] 针对目标簇,

[0386] 从碱基特异性强度分布中读取针对测序运行的当前测序循环记录的当前通道特异性强度, 在所述当前测序循环处将所述目标簇碱基检出到所述碱基特异性强度分布,

[0387] 从所述碱基特异性强度分布的质心读取当前通道特异性分布强度,

[0388] 基于所述当前通道特异性强度和所述当前通道特异性分布强度, 确定用于所述当前测序循环的一组当前强度校正参数,

[0389] 通过累加用于所述测序运行的先前测序循环的所述一组当前强度校正参数和一组先前累加强度校正参数, 确定用于所述当前测序循环的一组当前累加强度校正参数,

[0390] 基于所述一组当前累加强度校正参数确定用于所述当前测序循环的当前放大系数和当前通道特异性偏移系数, 以及使用所述当前放大系数和所述当前通道特异性偏移系数来校正针对所述测序运行的下一个测序循环记录的下一个通道特异性强度, 并针对所述下一个测序循环生成经过校正的下一个通道特异性强度; 以及

[0391] 基于所述经过校正的下一个通道特异性强度, 在所述下一个测序循环处碱基检出所述目标簇。

[0392] 29. 根据条款28所述的系统, 所述系统实现最终取决于条款1的每个条款。

[0393] 30. 一种碱基检出目标簇的计算机实现的方法, 所述方法包括:

[0394] 针对目标簇,

[0395] 访问当前强度数据和历史强度数据,

- [0396] 其中所述当前强度数据用于测序运行的当前测序循环,并且
- [0397] 其中所述历史强度数据用于所述测序运行的一个或多个在先前测序循环,
- [0398] 基于所述当前强度数据和所述历史强度数据确定缩放校正系数和通道特异性移位校正系数,
- [0399] 使用所述缩放校正系数和所述通道特异性移位校正系数来校正下一个强度数据并生成经过校正的下一个强度数据,
- [0400] 其中所述下一个强度数据用于所述测序运行的下一个测序循环;以及
- [0401] 基于所述经过校正的下一个强度数据,在所述下一个测序循环处对所述目标簇进行碱基检出。
- [0402] 31.根据条款30所述的计算机实现的方法,所述计算机实现的方法实现最终取决于条款1的每个条款。
- [0403] 32.一种印有碱基检出目标簇的计算机程序指令的非暂态计算机可读存储介质,该指令在处理器上执行时实现包括以下各项的方法:
- [0404] 针对目标簇,
- [0405] 访问当前强度数据和历史强度数据,
- [0406] 其中所述当前强度数据用于测序运行的当前测序循环,并且
- [0407] 其中所述历史强度数据用于所述测序运行的一个或多个在先前测序循环,
- [0408] 基于所述当前强度数据和所述历史强度数据确定缩放校正系数和通道特异性移位校正系数,
- [0409] 使用所述缩放校正系数和所述通道特异性移位校正系数来校正下一个强度数据并生成经过校正的下一个强度数据,
- [0410] 其中所述下一个强度数据用于所述测序运行的下一个测序循环;以及
- [0411] 基于所述经过校正的下一个强度数据,在所述下一个测序循环处对所述目标簇进行碱基检出。
- [0412] 33.根据条款32所述的非暂态计算机可读存储介质,该非暂态计算机可读存储介质实现最终取决于条款1的每个条款。
- [0413] 34.一种包括耦合到存储器的一个或多个处理器的系统,该存储器装载有碱基检出目标簇的计算机指令,该指令在该处理器上执行时实现包括以下各项的动作:
- [0414] 针对目标簇,
- [0415] 访问当前强度数据和历史强度数据,
- [0416] 其中所述当前强度数据用于测序运行的当前测序循环,并且
- [0417] 其中所述历史强度数据用于所述测序运行的一个或多个在先前测序循环,
- [0418] 基于所述当前强度数据和所述历史强度数据确定缩放校正系数和通道特异性移位校正系数,
- [0419] 使用所述缩放校正系数和所述通道特异性移位校正系数来校正下一个强度数据并生成经过校正的下一个强度数据,
- [0420] 其中所述下一个强度数据用于所述测序运行的下一个测序循环;以及
- [0421] 基于所述经过校正的下一个强度数据,在所述下一个测序循环处对所述目标簇进行碱基检出。

- [0422] 35. 根据条款34所述的系统,所述系统实现最终取决于条款1的每个条款。
- [0423] 36. 一种碱基检出目标簇的计算机实现的方法,所述方法包括:
- [0424] 针对目标簇,
- [0425] 访问当前强度数据和历史强度数据,
- [0426] 其中所述当前强度数据用于测序运行的当前测序循环,并且
- [0427] 其中所述历史强度数据用于所述测序运行的一个或多个在先前测序循环,
- [0428] 使用所述当前强度数据和所述历史强度数据来校正下一强度数据并生成经过校正的下一强度数据,
- [0429] 其中所述下一个强度数据用于所述测序运行的下一个测序循环;以及
- [0430] 基于所述经过校正的下一个强度数据,在所述下一个测序循环处对所述目标簇进行碱基检出。
- [0431] 37. 根据条款36所述的计算机实现的方法,所述计算机实现的方法实现最终取决于条款1的每个条款。
- [0432] 38. 一种印有碱基检出目标簇的计算机程序指令的非暂态计算机可读存储介质,该指令在处理器上执行时实现包括以下各项的方法:
- [0433] 针对目标簇,
- [0434] 访问当前强度数据和历史强度数据,
- [0435] 其中所述当前强度数据用于测序运行的当前测序循环,并且
- [0436] 其中所述历史强度数据用于所述测序运行的一个或多个在先前测序循环,
- [0437] 使用所述当前强度数据和所述历史强度数据来校正下一强度数据并生成经过校正的下一强度数据,
- [0438] 其中所述下一个强度数据用于所述测序运行的下一个测序循环;以及
- [0439] 基于所述经过校正的下一个强度数据,在所述下一个测序循环处对所述目标簇进行碱基检出。
- [0440] 39. 根据条款38所述的非暂态计算机可读存储介质,该非暂态计算机可读存储介质实现最终取决于条款1的每个条款。
- [0441] 40. 一种包括耦合到存储器的一个或多个处理器的系统,该存储器装载有碱基检出目标簇的计算机指令,该指令在该处理器上执行时实现包括以下各项的动作:
- [0442] 针对目标簇,
- [0443] 访问当前强度数据和历史强度数据,
- [0444] 其中所述当前强度数据用于测序运行的当前测序循环,并且
- [0445] 其中所述历史强度数据用于所述测序运行的一个或多个在先前测序循环,
- [0446] 使用所述当前强度数据和所述历史强度数据来校正下一强度数据并生成经过校正的下一强度数据,
- [0447] 其中所述下一个强度数据用于所述测序运行的下一个测序循环;以及
- [0448] 基于所述经过校正的下一个强度数据,在所述下一个测序循环处对所述目标簇进行碱基检出。
- [0449] 41. 根据条款40所述的系统,所述系统实现最终取决于条款1的每个条款。
- [0450] 虽然通过参考上文详细描述的首选实施方案和示例公开了本发明,但是应当理

解,这些示例旨在进行说明而非进行限制。可以预期,本领域的技术人员将容易想到修改和组合,这些修改和组合将在本发明的实质和以下权利要求书的范围之内。

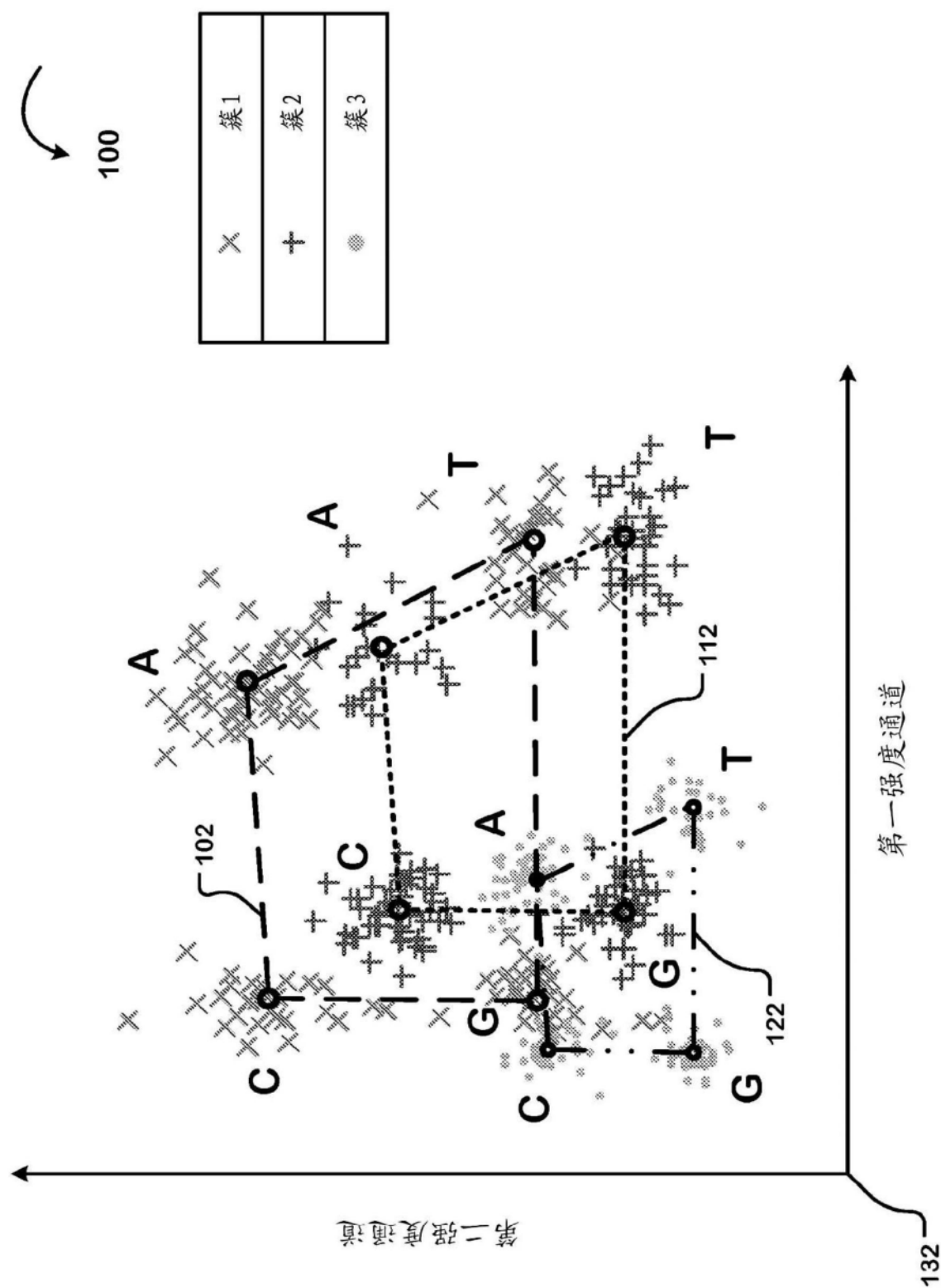


图1

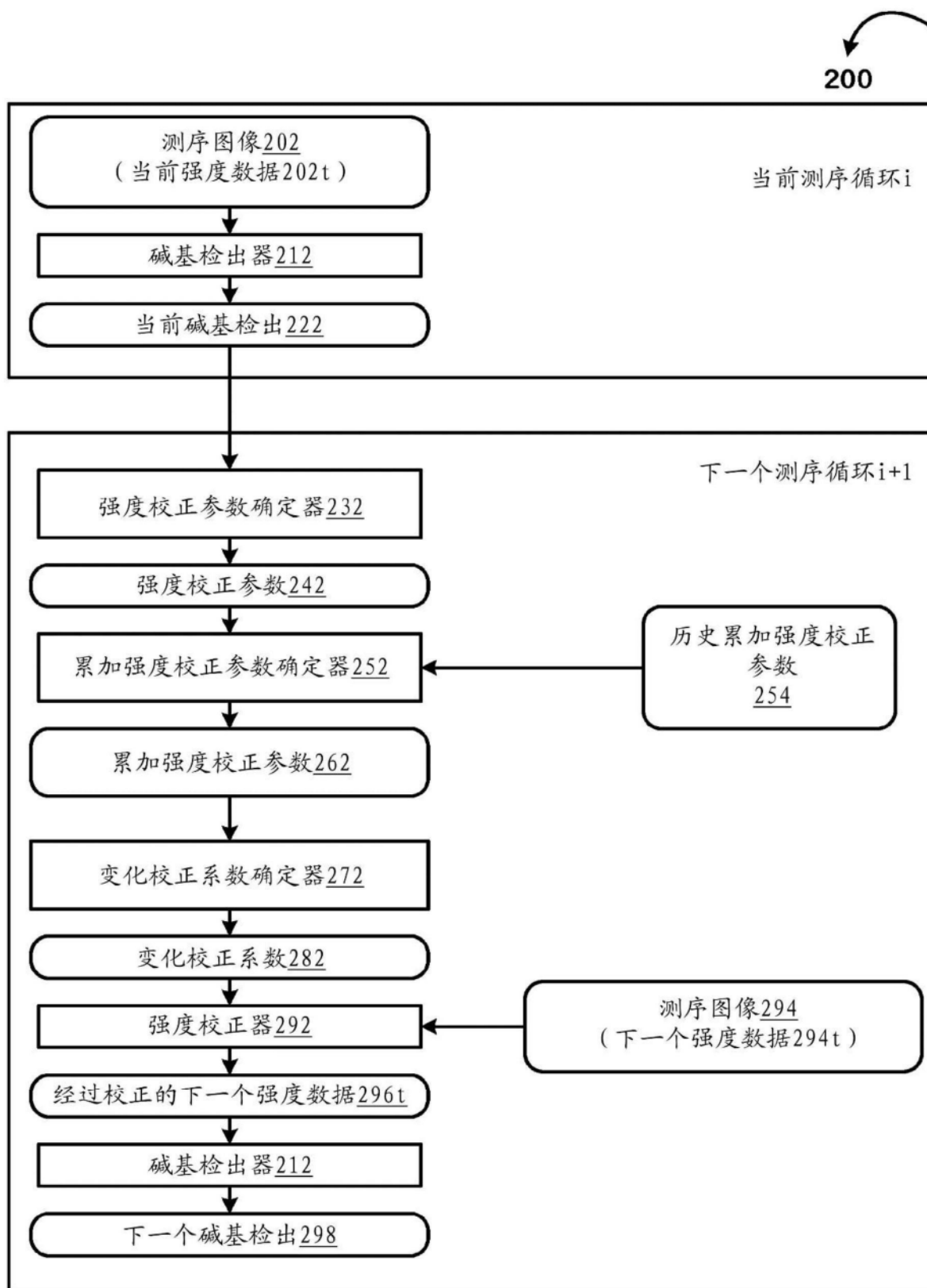


图2

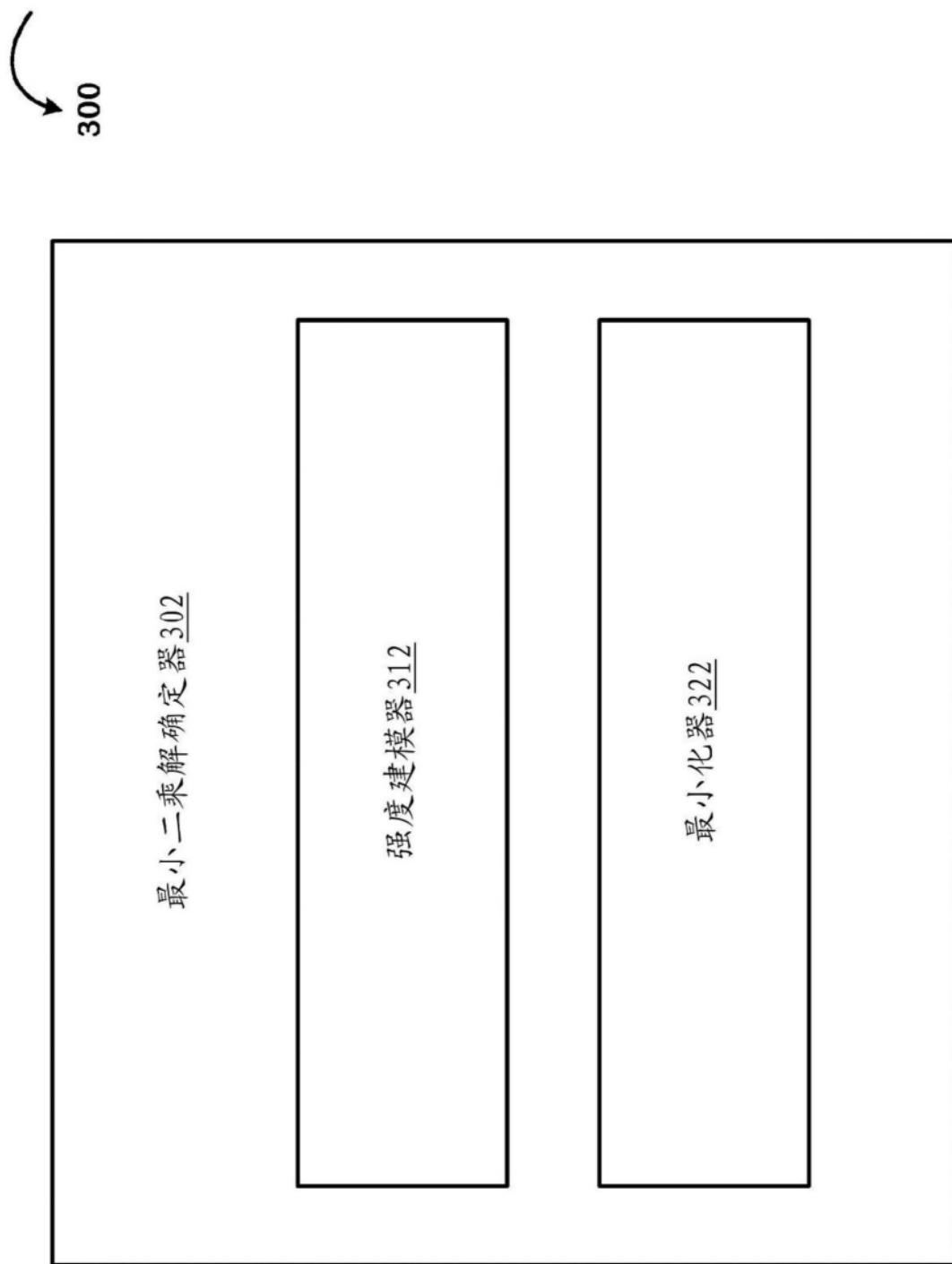


图3

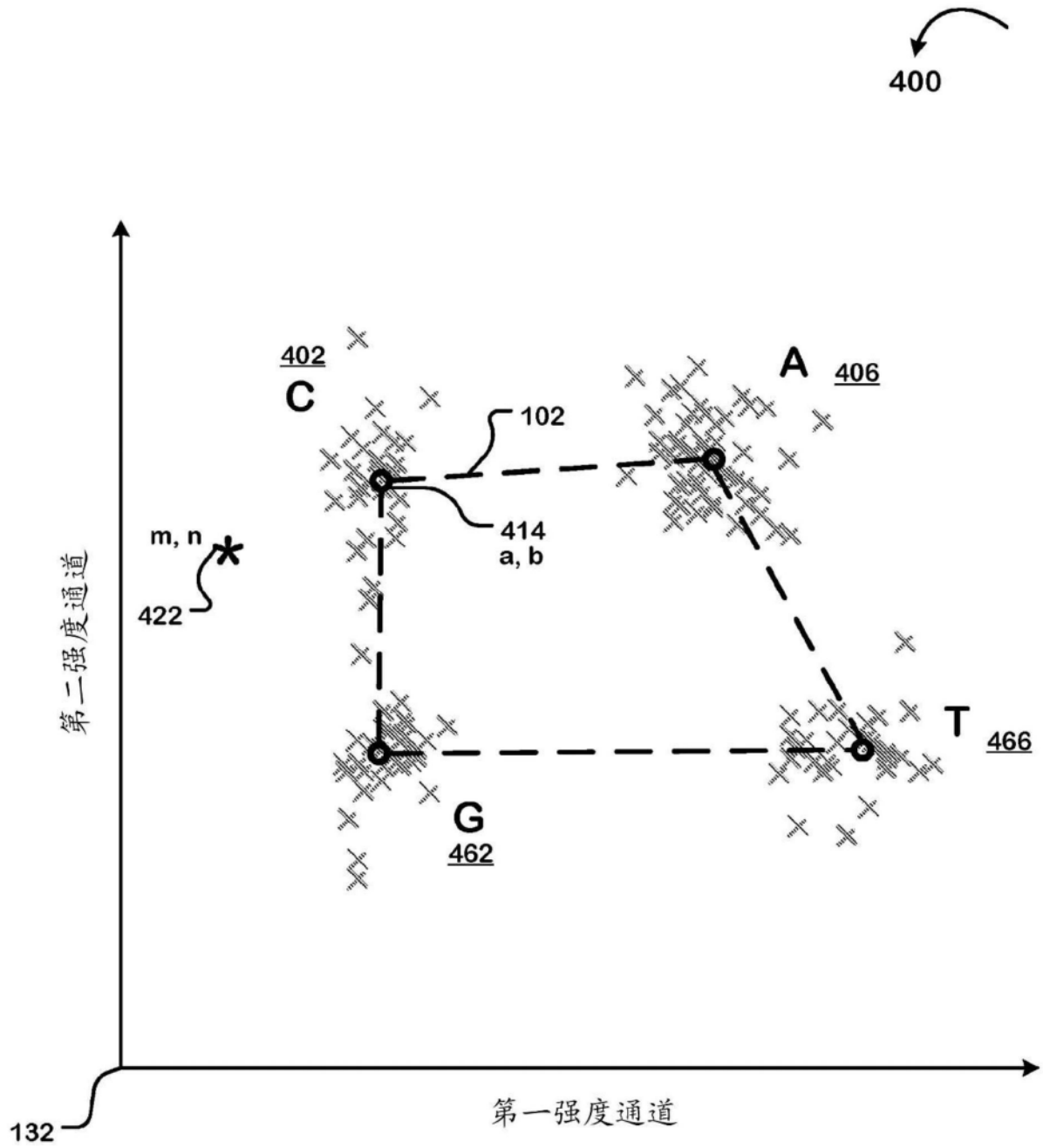


图4

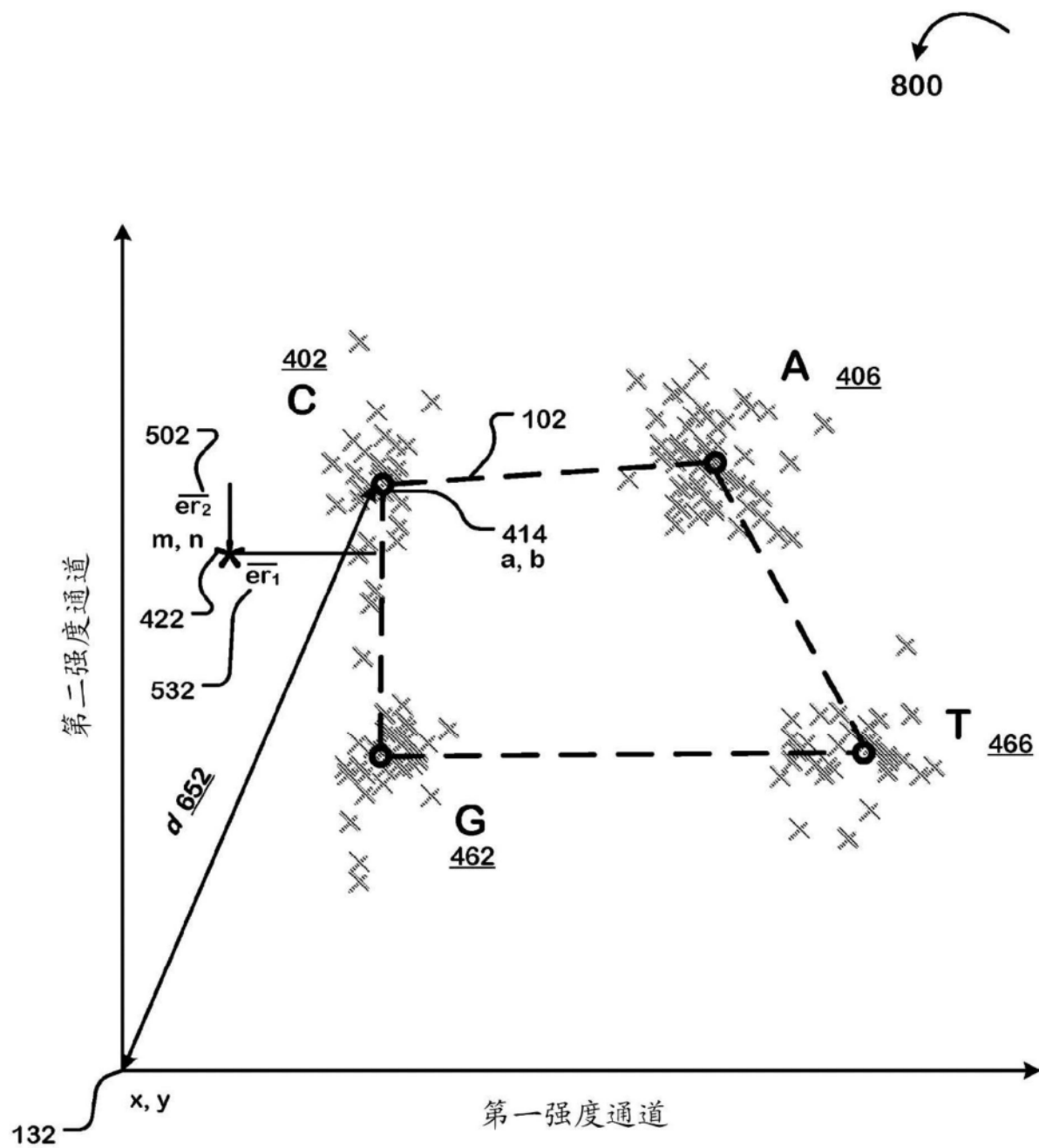


图6

先前测序循环 $i-1$

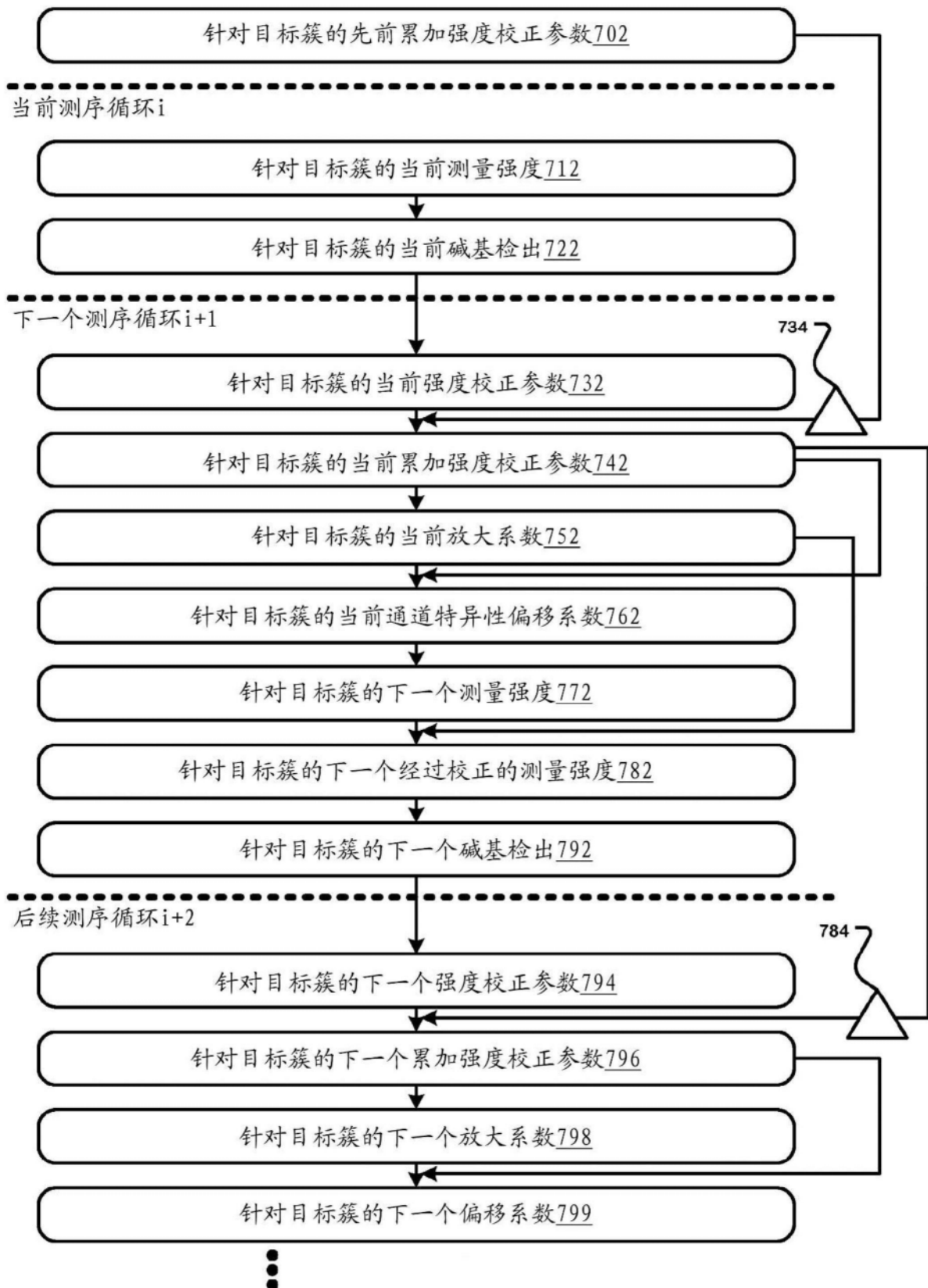


图7

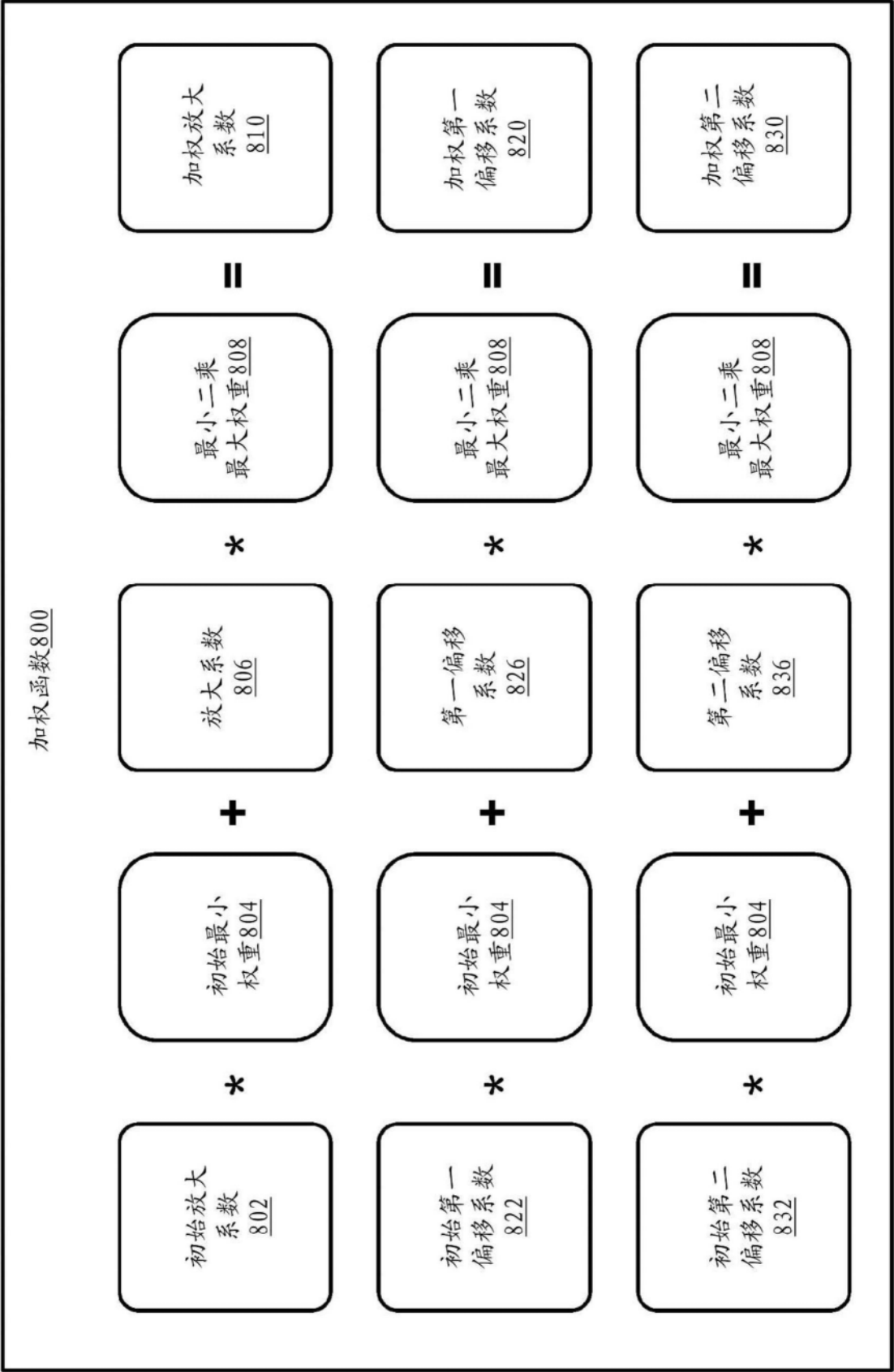


图8

如果 α 、 d_1 、 d_2 和 $n_{c,i}$ 具有先验概率分布 $\alpha \sim N(1, \sigma_\alpha^2)$, $d_1 \sim N(0, \sigma_{d_1}^2)$, $d_2 \sim N(0, \sigma_{d_2}^2)$, $n_{c,i} \sim N(0, \sigma_n^2)$,

可找到最大似然解:

设 $e_{c,i} = y_{c,i} - x_{c,i}$

$$\begin{aligned} \bar{x}_1 &\triangleq \frac{1}{C} \sum_{c=1}^C x_{c,1} & \bar{x}_2 &\triangleq \frac{1}{C} \sum_{c=1}^C x_{c,2} & \overline{xx} &\triangleq \frac{1}{C} \sum_{c=1}^C (x_{c,1}^2 + x_{c,2}^2) \\ \bar{e}_1 &\triangleq \frac{1}{C} \sum_{c=1}^C e_{c,1} & \bar{e}_2 &\triangleq \frac{1}{C} \sum_{c=1}^C e_{c,2} & \overline{xe} &\triangleq \frac{1}{C} \sum_{c=1}^C (x_{c,1}e_{c,1} + x_{c,2}e_{c,2}) \end{aligned}$$

$$\begin{aligned} k_\alpha(C) &= \frac{\sigma_n^2}{C\sigma_\alpha^2}, & k_{d_1}(C) &= \frac{C\sigma_{d_1}^2}{C\sigma_{d_1}^2 + \sigma_n^2}, & k_{d_2}(C) &= \frac{C\sigma_{d_2}^2}{C\sigma_{d_2}^2 + \sigma_n^2} \end{aligned}$$

$$\hat{\alpha} = 1 + \frac{\overline{xe} - \bar{x}_1 \bar{e}_1 k_{d_1}(C) - \bar{x}_2 \bar{e}_2 k_{d_2}(C)}{\overline{xx} + k_\alpha(C) - \bar{x}_1^2 k_{d_1}(C) - \bar{x}_2^2 k_{d_1}(C)}$$

$$\hat{d}_1 = (\bar{e}_1 + \bar{x}_1(1 - \hat{\alpha}))k_{d_1}(C)$$

$$\hat{d}_2 = (\bar{e}_2 + \bar{x}_2(1 - \hat{\alpha}))k_{d_2}(C)$$

当 $C \rightarrow \infty$, 这接近最小二乘解

图9

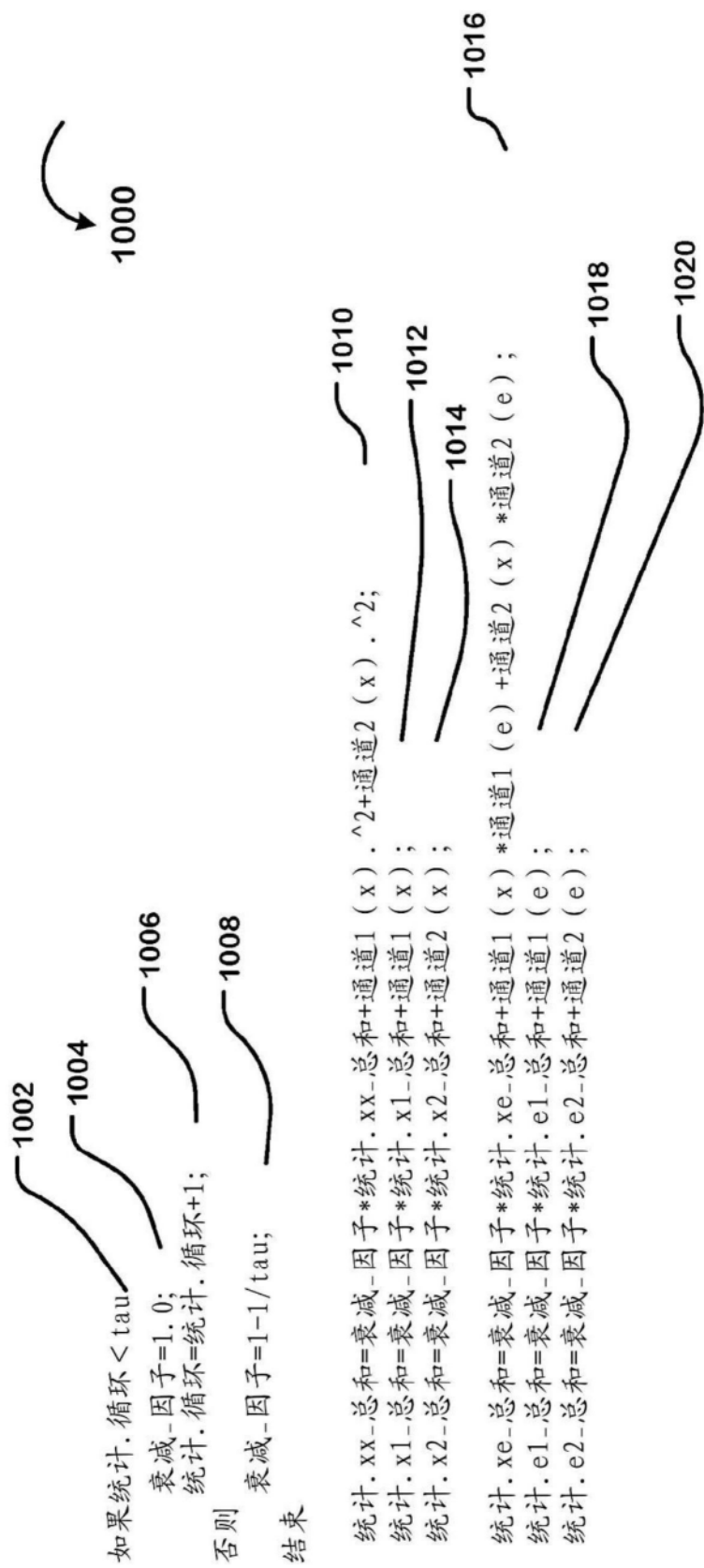


图10

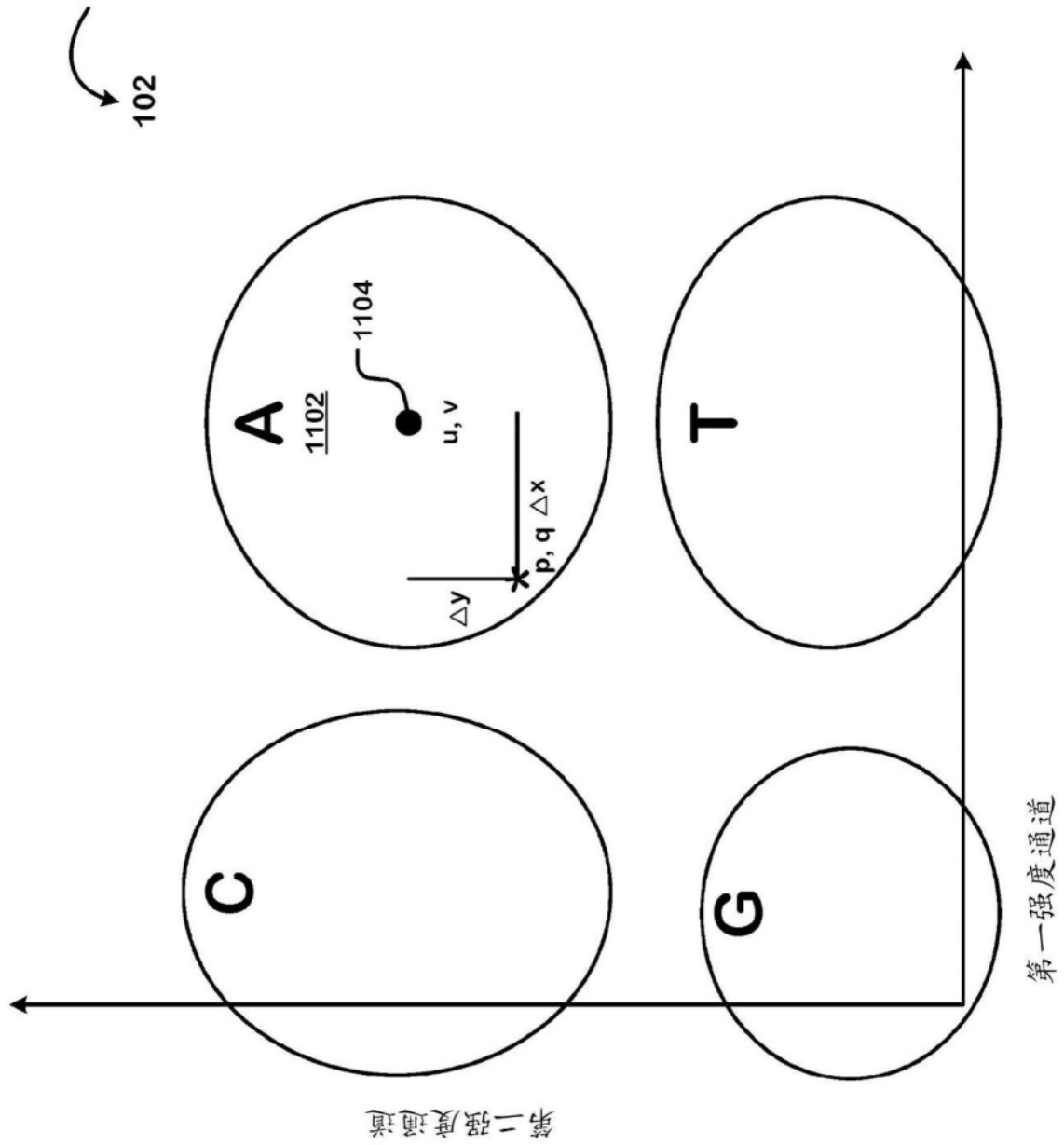


图11

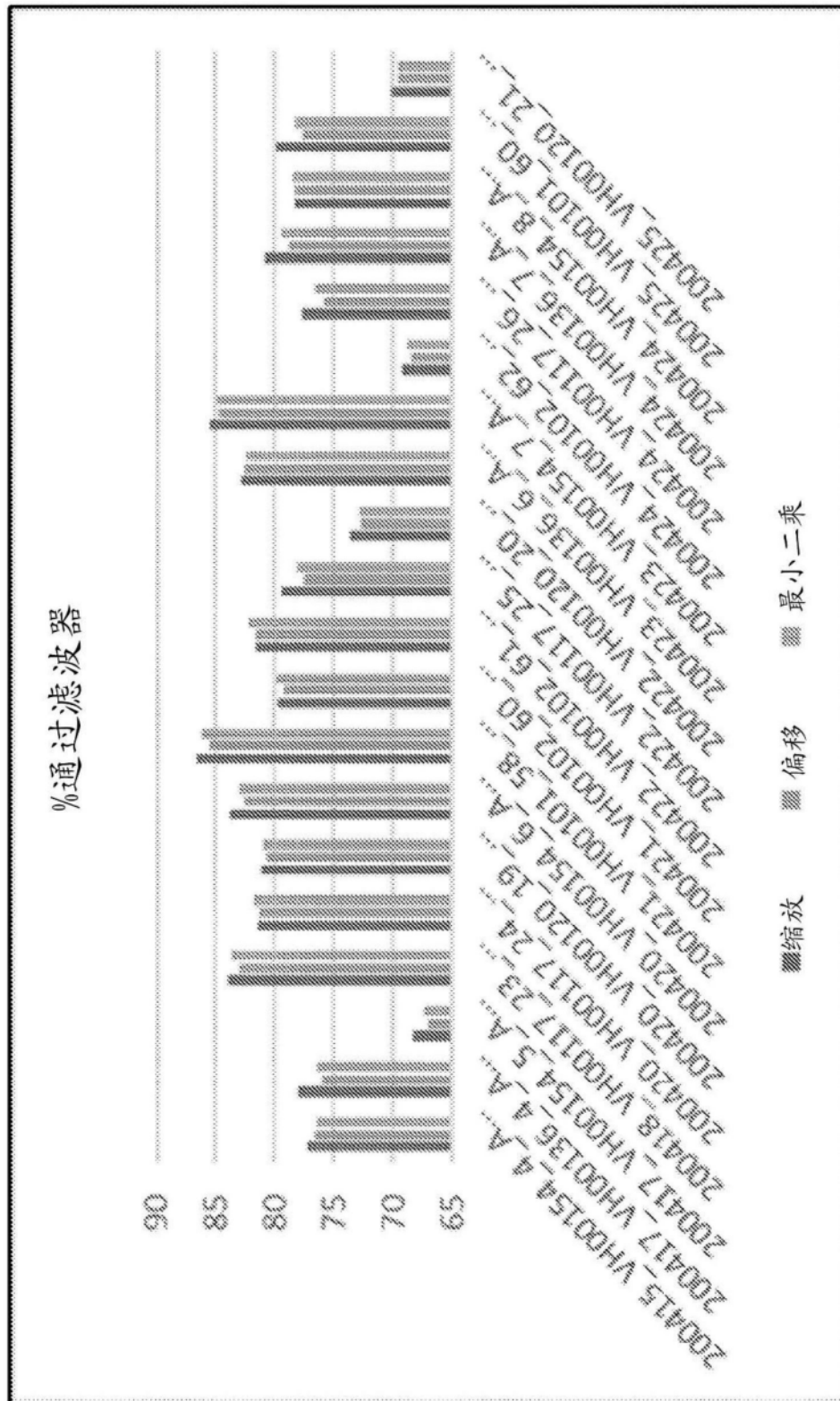


图12

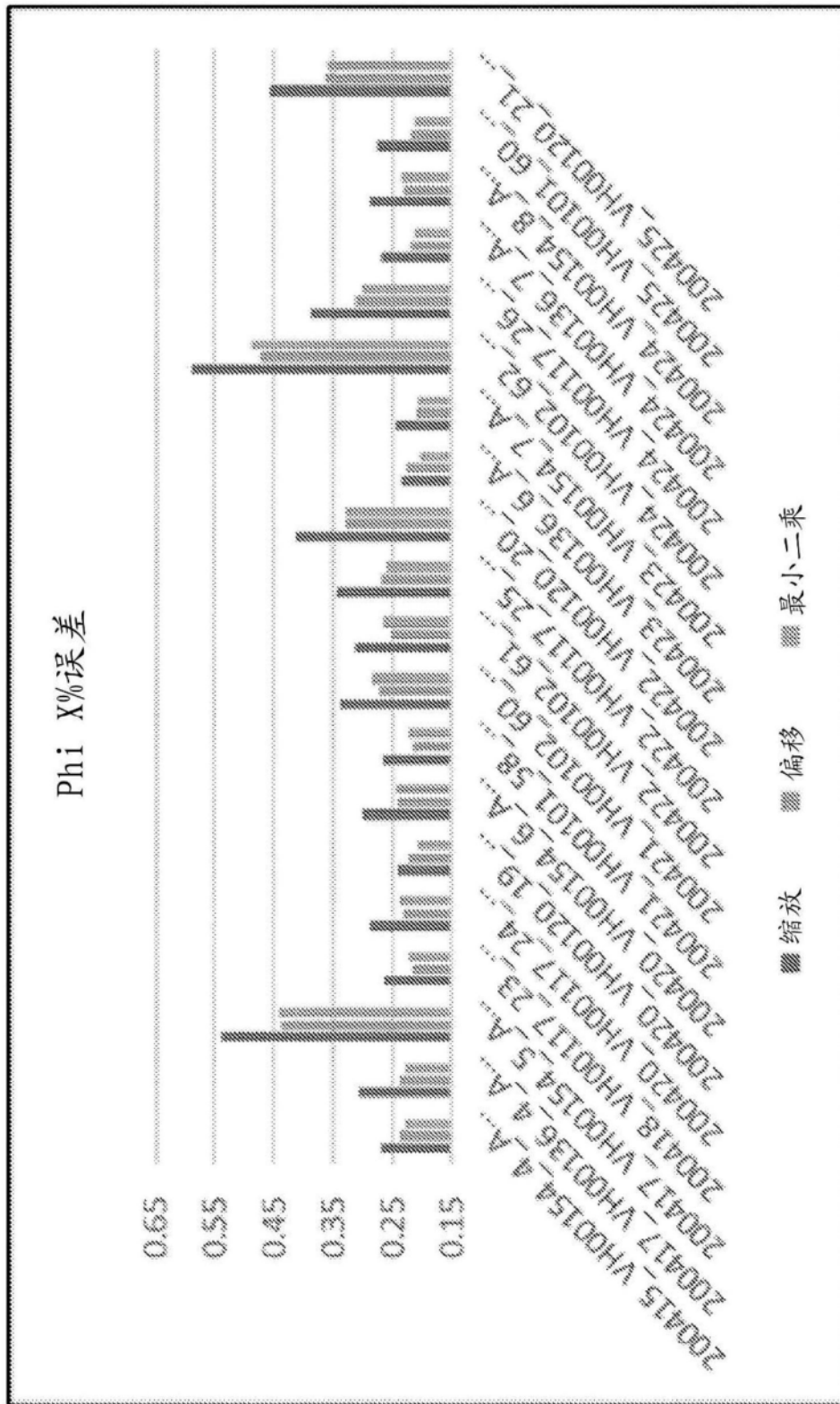


图13

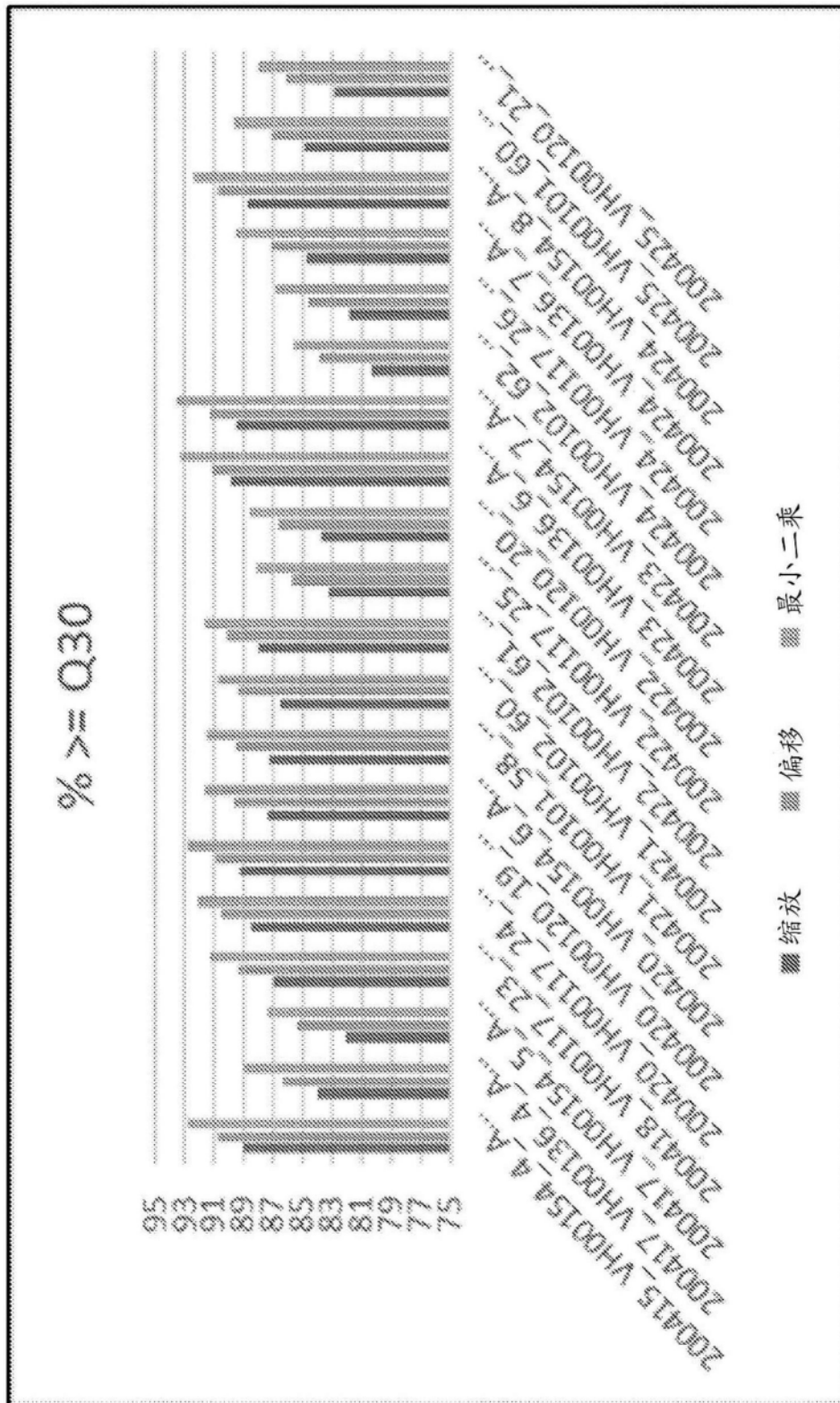


图14

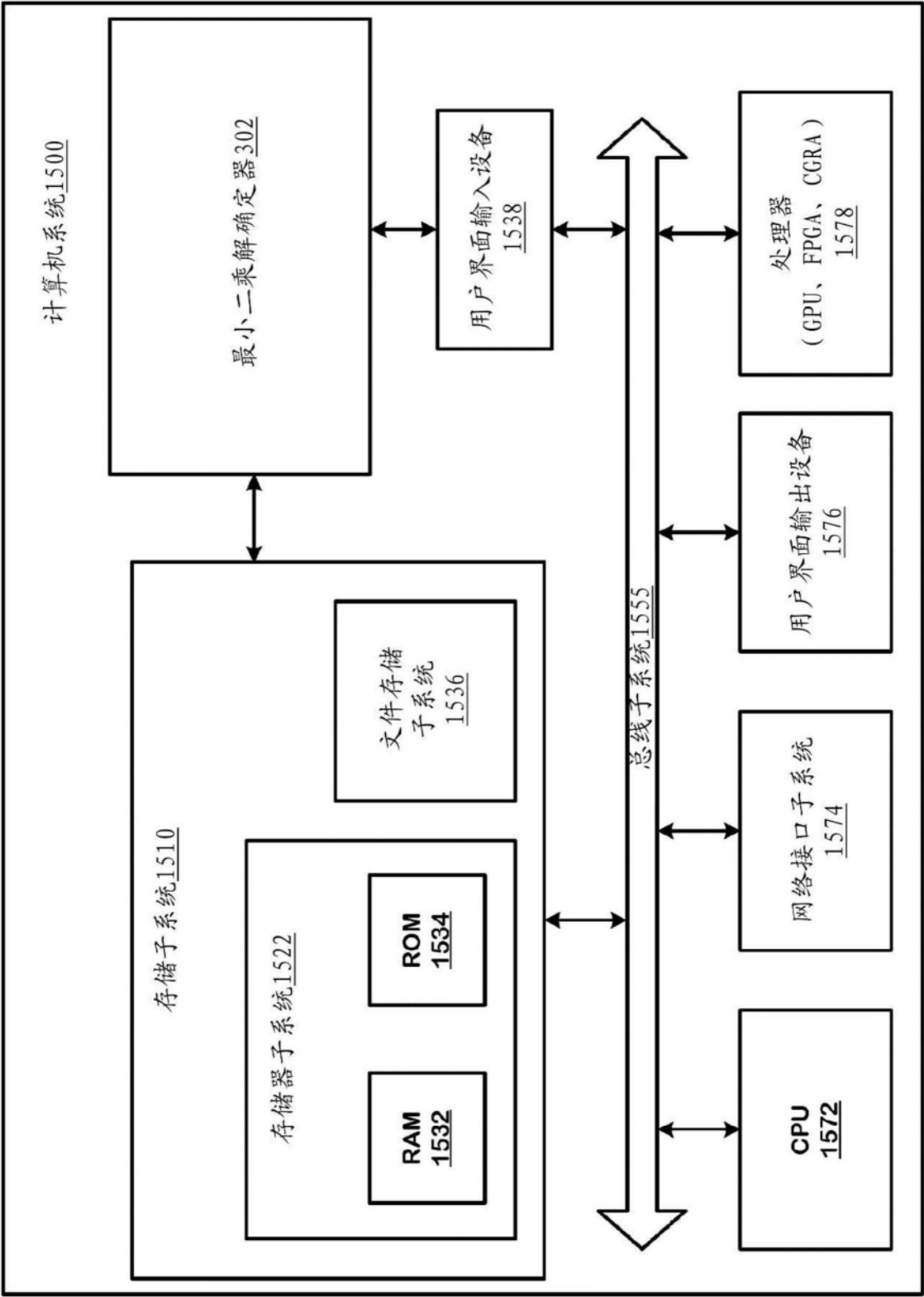


图15