

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2024 年 12 月 19 日 (19.12.2024)



(10) 国际公布号
WO 2024/254925 A1

(51) 国际专利分类号:
G16B 20/20 (2019.01)

(21) 国际申请号: PCT/CN2023/105780

(22) 国际申请日: 2023 年 7 月 4 日 (04.07.2023)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
PCT/CN2023/100428
2023年6月15日 (15.06.2023) CN

(71) 申请人: 深圳华大智造科技股份有限公司
(**MGI TECH CO., LTD.**) [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼及 11 栋 2 楼, Guangdong 518083 (CN)。

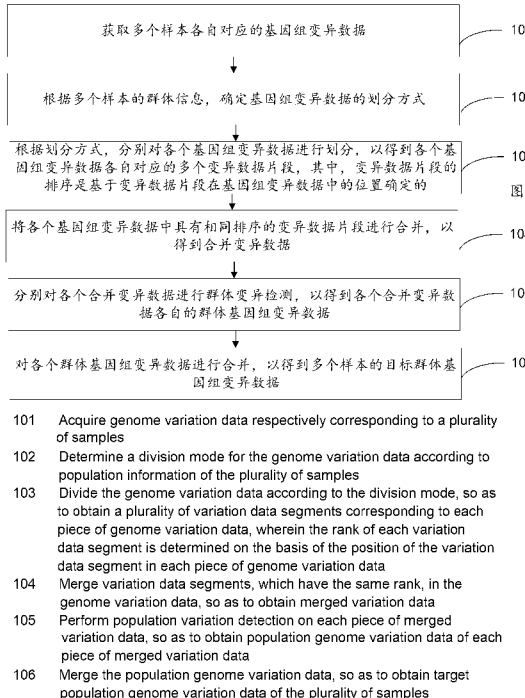
(72) 发明人: 唐静波(**TANG, Jingbo**); 中国广东省深圳市盐田区北山工业区综合楼及 11 栋 2 楼, Guangdong 518083 (CN)。钟子寅(**ZHONG, Ziyin**); 中国广东省深圳市盐田区北山工业区综合楼及 11 栋 2 楼, Guangdong 518083 (CN)。谢寅龙(**XIE, Yinlong**); 中国广东省深圳市盐田区北山工业区综合楼及 11 栋 2 楼, Guangdong 518083 (CN)。

(74) 代理人: 北京法胜知识产权代理有限公司 (**FASHENG INTELLECTUAL PROPERTY COMPANY, LTD.**); 中国北京市海淀区悦秀路 99 号 1 单元 4 层 401, Beijing 100085 (CN)。

(81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI,

(54) **Title:** METHOD AND APPARATUS FOR PERFORMING JOINT-CALLING ACCELERATION ON HIGH-PERFORMANCE SEQUENCING DATA, AND ELECTRONIC DEVICE AND STORAGE MEDIUM

(54) 发明名称: 一种对高性能测序数据Joint-Calling加速的方法、装置、电子设备和存储介质



(57) **Abstract:** Provided in the present disclosure are a method and apparatus for performing joint-calling acceleration on high-performance sequencing data, and an electronic device and a storage medium. The method comprises: determining a division mode for genome variation data according to population information of a plurality of samples; dividing the genome variation data according to

GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。

- (84) 指定国(除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

the division mode, so as to obtain a plurality of variation data segments corresponding to each piece of genome variation data; merging variation data segments, which have the same rank, in the genome variation data, so as to obtain merged variation data; performing population variation detection on each piece of merged variation data, so as to obtain population genome variation data of each piece of merged variation data; and merging the population genome variation data, so as to obtain target population genome variation data of the plurality of samples. Thus, during population variation detection, by means of performing population anomaly detection on a plurality of pieces of local merged data, and merging detection results, the efficiency of acquiring complete population genome variation data can be improved.

(57) 摘要: 本公开提出了一种对高性能测序数据Joint-Calling加速的方法、装置、电子设备和存储介质, 该方法包括: 根据多个样本的群体信息, 确定基因组变异数据的划分方式, 并根据划分方式, 划分各个基因组变异数据, 以得到各个基因组变异数据各自对应的多个变异数据片段; 将各个基因组变异数据中具有相同排序的变异数据片段进行合并, 以得到合并变异数据; 分别对各个合并变异数据进行群体变异检测以得到各个合并变异数据的群体基因组变异数据, 并对各个群体基因组变异数据进行合并以得到多个样本的目标群体基因组变异数据。由此, 在群体变异检测过程中, 通过分别对多个局部合并数据进行群体异常检测, 并对检测结果进行合并, 可提高获取完整的群体基因组变异数据的效率。

一种对高性能测序数据 Joint-Calling 加速的方法、装置、电子设备和存储介质 相关申请的交叉引用

5 本公开要求深圳华大智造科技股份有限公司公司于 2023 年 6 月 15 日提交的、发明名称为“群体变异检测方法、装置、电子设备和存储介质”的、国际申请号“PCT/CN2023/100428”的优先权。

技术领域

10 本公开涉及生物信息处理技术领域，尤其涉及一种对高性能测序数据 Joint-Calling 加速的方法、装置、电子设备和存储介质。

背景技术

15 群体基因组学：通过分析 DNA 序列多态数据，推测曾经作用于基因组的各种力量，进而探讨生物演化的过程。群体基因组学的研究内容主要包括群体结构和数量变化、群体突变、自然选择、遗传漂变等，以及这些与基因组中的遗传变异关系。医学研究群体遗传是要探讨遗传病的发病频率，遗传方式及其基因频率和变化的规律，从而了解遗传病在群体中的发生和散布的规律，为预防、监测和治疗遗传病提供重要的信息和措施。

20 在群体基因组学研究中，群体联合变异检测是群体基因组学研究的一个关键环节。在群体联合变异检测的过程中，如何高效对大规模样本进行群体联合变异检测，以得到样本对应的群体变异数据是目前亟需的技术问题。

发明内容

25 本公开提出一种对高性能测序数据 Joint-Calling 加速的方法、装置、电子设备和存储介质。

30 本公开一方面实施例提出一种对高性能测序数据 Joint-Calling 加速的方法，所述方法包括：获取多个样本各自对应的基因组变异数据；根据所述多个样本的群体信息，确定所述基因组变异数据的划分方式；根据所述划分方式，分别对各个所述基因组变异数据进行划分，以得到各个所述基因组变异数据各自对应的多个变异数据片段，其中，所述变异数据片段的排序是基于所述变异数据片段在所述基因组变异数据中的位置确定的；将各个所述基因组变异数据中具有相同排序的变异数据片段进行合并，以得到合并变异数据；分别对各个所述合并变异数据进行群体变异检测，以得到各个所述合并变异数据各自的群体基因组变异数据；对各个所述群体基因组变异数据进行合并，以得到所述多个样本的目标群体基因组变异数据。

本公开另一方面实施例提出一种对高性能测序数据 Joint-Calling 加速的装置，所述装置包括：获取模块，用于获取多个样本各自对应的基因组变异数据；第一确定模块，用于根据所述多个样本的群体信息，确定所述基因组变异数据的划分方式；第一划分模块，用于根据所述划分方式，分别对各个所述基因组变异数据进行划分，以得到各个所述基因组变异数据各自对应的多个变异数据片段，其中，所述变异数据片段的排序是基于所述变异数据片段在所述基因组变异数据中的位置确定的；第一合并处理模块，用于将各个所述基因组变异数据中具有相同排序的变异数据片段进行合并，以得到合并变异数据；第一群体变异检测模块，用于分别对各个所述合并变异数据进行群体变异检测，以得到各个所述合并变异数据各自的群体基因组变异数据；第二合并处理模块，用于对各个所述群体基因组变异数据进行合并，以得到所述多个样本的目标群体基因组变异数据。

本公开另一方面实施例提出了一种电子设备，包括：存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，所述处理器执行所述程序时实现本公开实施例所公开的一种对高性能测序数据 Joint-Calling 加速的方法。

本公开另一方面实施例提出了一种计算机可读存储介质，其上存储有计算机程序，该程序被处理器执行时实现本公开实施例所公开的一种对高性能测序数据 Joint-Calling 加速的方法。

根据本公开实施例提供的技术方案，在多个样本各自对应的基因组变异数据进行处理的过程中，根据多个样本的群体信息，确定基因组变异数据的划分方式，并根据划分方式，分别对各个基因组变异数据进行划分，以得到各个基因组变异数据各自对应的多个变异数据片段以及将各个基因组变异数据中具有相同排序的变异数据片段进行合并，以得到合并变异数据；分别对各个合并变异数据进行群体变异检测，以得到各个合并变异数据各自的群体基因组变异数据，以及对各个群体基因组变异数据进行合并，以得到多个样本的目标群体基因组变异数据。由此，在群体变异检测的过程中，通过分别对多个局部合并数据进行群体异常检测，并对检测结果进行合并，可提高获取多个样本的完整的群体基因组变异数据的效率。

应当理解，本部分所描述的内容并非旨在标识本公开的实施例的关键或重要特征，也不用于限制本公开的范围。本公开的其它特征将通过以下的说明书而变得容易理解。

附图说明

附图用于更好地理解本方案，不构成对本公开的限定。其中：

图1是根据本公开一个实施例的对高性能测序数据Joint-Calling加速的方法的流程示意图；

图2是根据本公开另一个实施例的对高性能测序数据Joint-Calling加速的方法的流程示意图；

图3是根据本公开另一个实施例的对高性能测序数据Joint-Calling加速的方法的流程示意图；

图4是根据本公开另一个实施例的对高性能测序数据Joint-Calling加速的方法的流程示意图；

5 图5是根据本公开另一个实施例的对高性能测序数据Joint-Calling加速的方法的流程示意图；

图6是根据本公开另一个实施例的对高性能测序数据Joint-Calling加速的方法的流程示意图；

10 图7是根据本公开一个实施例的对高性能测序数据Joint-Calling加速的装置的结构示意图；

图8是根据本公开另一个实施例的对高性能测序数据Joint-Calling加速的装置的结构示意图；

图9是根据本公开一个实施例的电子设备的框图。

15 具体实施方式

下面详细描述本发明的实施例，实施例的示例在附图中示出，其中自始至终相同或类似的标号表示相同或类似的元件或具有相同或类似功能的元件。下面通过参考附图描述的实施例是示例性的，旨在用于解释本发明，而不能理解为对本发明的限制。

如本文公开的，术语“样本”是指受试者的任何样本，其可反应与受试者相关联的生物
20 状态，并且其包括无细胞DNA。样本的实例包括但不限于受试者的血液、全血、血浆、血清、尿液、脑脊液、唾液、汗液、泪液。生物样本可包括来源于活的或死的受试者的任何组织或材料。样本可为无细胞样本。在一些实施方案中，生物组织或流体可以是或包括羊水、房水、腹水、胆汁、骨髓、血液、母乳、脑脊液、耳垢、乳糜、食糜(chyme)、一次射出的精液(ejaculate)、内淋巴、渗出物、粪便、胃酸、胃液、淋巴液、粘液、心包液、外淋
25 巴、腹膜液、胸膜液、脓液、发炎性分泌物、唾液、皮脂、精液、血清、阴垢、痰、滑液、汗液、泪液、尿液、阴道分泌物、玻璃体液、呕吐物，以及/或者它们的组合或一种或多种组分。在一些实施方案中，生物流体可以是或包括细胞内液、细胞外液、血管内液(血浆)、间质液、淋巴液和/或跨细胞液。在一些实施方案中，生物流体可以是或包括植物渗出物。在一些实施方案中，生物组织或样本可以例如通过抽吸、活检(例如，细针或组织活检)、拭
30 子(例如，口腔、鼻、皮肤或阴道拭子)、刮擦、手术、洗涤或灌洗(例如，支气管肺泡、导管、鼻、眼、口腔、子宫、阴道或其他洗涤或灌洗)来获取。在一些实施方案中，生物样本是或包括从个体获得的细胞。在一些实施方案中，样本是通过任何合适的手段直接从感兴趣的来源获得的“初次样本”。在一些实施方案中，如从上下文来看将显而易见的，术语“样本”是指通过处理初次样本(例如，通过去除其一种或多种组分和/或通过向其中添加一种或

多种剂)获得的制品。例如, 使用半透膜过滤。此类“经处理的样本”可包括例如从样本中提取或通过使初次样本经受一种或多种技术(诸如核酸的扩增或逆转录, 某些组分的分离和/或纯化等)而获得的核酸(包括 DNA、RNA)、蛋白质或染色质等。

下面参考附图描述本公开实施例的对高性能测序数据 Joint-Calling 加速的方法、装置、
5 电子设备和存储介质。

图 1 是根据本公开一个实施例的对高性能测序数据 Joint-Calling 加速的方法的流程示意图。其中, 需要说明的是, 本实施例提供的对高性能测序数据 Joint-Calling 加速的方法由对高性能测序数据 Joint-Calling 加速的装置执行, 其中, 本实施例中的对高性能测序数据 Joint-Calling 加速的装置可以由软件和/或硬件的方式实现, 该对高性能测序数据
10 Joint-Calling 加速的装置可以为电子设备, 或者可以配置在电子设备中。

其中, 该电子设备可以包括但不限于终端设备、服务器、计算机集群等, 该实施例对此不作具体限定。

其中, 终端设备可以包括但不限于个人计算机(Personal Computer, PC)、移动设备、平板电脑等, 该实施例对此不作具体限定。

15 如图 1 所示, 该对高性能测序数据 Joint-Calling 加速的方法可以包括:

步骤 101, 获取多个样本各自对应的基因组变异数据。

在一些示例性的实施方式中, 可通过分别对多个样本进行变异检测, 以得到多个样本各自对应的基因组变异检测格式(Genome Variant Call Format)GVCF 文件, 其中, GVCF 文件可以包括基因组变异数据。

20 步骤 102, 根据多个样本的群体信息, 确定基因组变异数据的划分方式。

其中, 群体信息用于表示样本所属于群体的信息, 例如, 群体信息可以包括但不限于群体标识信息、群体名称信息等。

在一些示例性的实施方式中, 根据多个样本的群体信息, 确定基因组变异数据的划分方式的一种可能实现方式可以为: 比较多个样本的群体信息是否相同, 以得到比较结果;
25 根据比较结果, 确定基因组变异数据的划分方式。也就是说, 根据多个样本各自对应的群体信息来确定多个样本是否来自于同一个群体, 并根据多个样本是否来自于同一个群体, 确定对基因组变异数据进行划分的划分方式。

作为一种示例中, 在比较结果为多个样本的群体信息相同的情况下, 确定基因组变异数据的划分方式为第一划分方式。也就是说, 在根据多个样本各自对应的群体信息确定多个样本来自同一个群体, 即, 多个样本所属于的群体是相同的, 由于来自同一个群体上的
30 多个样本有着共同的起源, 因此, 可采用第一划分方式对基因组变异数据进行划分, 其中, 第一划分方式用于指示基于变异位点密度进行划分的方式。

其中, 变异位点密度是指变异位点在单位长度样本上变异位点的数量。

其中, 单位长度是根据需求设定的一个参考标准, 单位长度就是可供参考的标准, 它

没有固定值，依设定而变动的。

在另一些示例中，在比较结果为多个样本的群体信息不相同的情况下，确定基因组变异数据的划分方式为第二划分方式。也就是说，在基于多个样本的群体信息确定多个样本来自于不同于群体的情况下，由于多个样本来自于不同群体，群体来源广泛，群体杂合性高，同一个位点等位基因数量较多，因此，可确定对基因组变异数据进行划分的方式为第二划分方式，其中，第二划分方式为用于指示基于等位基因密度进行划分的方式。

其中，杂合性：杂合性又称群体的平均杂合性或杂合度，它是群体遗传变异的另一个度量参数，是指某一基因座上的等位基因是杂合体的频率。

其中，等位基因密度是指等位基因在单位长度样本上等位基因的数量。

其中，等位基因是指位于一对同源染色体相同位置上控制同一性状不同形态的基因，不同的等位基因产生例如发色或血型等遗传特征的变化。在一个群体内，同源染色体的某个相同座位上的等位基因超过 2 个以上时，就称作复等位基因。例如，人类 ABO 血型基因座位是在 9 号染色体长臂的末端，在这个座位上的等位基因，有 A、B、O 三个基因，因此人类的 ABO 血型是由 3 个复等位基因决定的。群体基因组联合变异检测 Joint-Calling 时，同一个位点的等位基因数量可达几十个。

步骤 103，根据划分方式，分别对各个基因组变异数据进行划分，以得到各个基因组变异数据各自对应的多个变异数据片段，其中，变异数据片段的排序是基于变异数据片段在基因组变异数据中的位置确定的。

在一些示例性的实施方式中，在划分方式为第一划分方式的情况下，根据划分方式，分别对各个基因组变异数据进行划分，以得到各个基因组变异数据各自对应的多个变异数据片段的一种可能实现方式为：针对每个基因组变异数据，将基因组变异数据按照变异位点密度进行划分，以得到基因组变异数据对应的多个变异数据片段。

在本示例中，基因组变异数据中的变异位点并不是均匀分布的，为了充分利用计算资源和降低计算时间，因此，针对每个基因组变异数据，本示例中对该基因组变异数据中的变异位点的分布情况进行统计，以得到该基因组变异数据的变异位点统计结果，并根据变异位点统计结果，基于变异位点密度对该基因组变异数据进行划分，以得到该基因组变异数据对应的多个变异数据片段，其中，多个变异数据片段所对应的长度是不同的，但是多个变异数据片段上的变异位点数量之间的数量差在第一预设数量之内，例如，多个变异数据片段上的变异位点数量基本一致。

其中，第一预设数量可以是预先设置的。

在一些示例中，可根据预设片段数量和该基因组变异数据中的变异位点的总数，确定出分配给每个变异数据片段的变异位点的平均值，并根据平均值，确定出基因组变异数据中每个变异数据片段的起止位置，并根据每个变异数据片段的起止位置，对该基因组变异数据进行划分，以得到预设片段数量个变异数据片段。

在另一些示例性的实施方式中，在划分方式为第二划分方式的情况下，根据划分方式，分别对各个基因组变异数据进行划分，以得到各个基因组变异数据各自对应的多个变异数据片段的一种可能实现方式为：针对每个基因组变异数据，将基因组变异数据按照等位基因密度进行划分，以得到基因组变异数据对应的多个变异数据片段。

5 具体而言，针对每个基因组变异数据，可确定出该基因组变异数据中每个变异位点上所对应的等位基因信息，并对该基因组变异数据中的等位基因信息进行统计，以得到该基因组变异数据的等位基因信息的统计结果，并根据统计结果，基于等位基因密度对该基因组变异数据进行划分，以得到该基因组变异数据对应的多个变异数据片段，其中，变异数据片段所对应的长度是不同的，并且多个变异数据片段中等位基因的数量之间的差值在第二预设数量内例如，多个变异数据片段中等位基因的数量可以是基本相同的。

10 在一些示例中，可根据预设片段数量和该基因组变异数据中的等位基因的总数，确定出分配给每个变异数据片段的等位基因的平均值，并根据平均值，确定出基因组变异数据中每个变异数据片段的起止位置，并根据每个变异数据片段的起止位置，对该基因组变异数据进行划分，以得到预设片段数量个变异数据片段。

15 步骤 104，将各个基因组变异数据中具有相同排序的变异数据片段进行合并，以得到合并变异数据。

其中，可以理解的是，本示例中的合并变异数据有多个。其中，合并变异数据的数量与对一个基因组变异数据进行划分所得到的基因变异数据片段的数量是相同的。

20 步骤 105，分别对各个合并变异数据进行群体变异检测，以得到各个合并变异数据各自的群体基因组变异数据。

25 在一些示例性的实施方式中，可调用多个处于空闲状态的计算节点，并且每一个计算节点对一个合并变异数据进行群体变异检测（Joint-Calling），其中，不同节点所处理的合并变异数据是不同的。也就是说，在一些示例中，可通过多个计算节点，分别对多个合并变异数据进行群体变异检测，其中，不同计算节点所处理的合并变异数据是不同的。由此，通过多个计算节点，分别对多个合并变异数据进行群体变异检测，可加速联合变异检测过程，提高获取多个合并变异数据各自对应的群体基因组变异数据的效率。

30 在另一些示例性的实施方式中，在通过多个计算节点对多个合并变异数据进行群体变异检测的情况下，分别对各个合并变异数据进行群体变异检测，以得到各个合并变异数据各自的群体基因组变异数据的一种可能实现方式为：针对每个计算节点：从多个合并变异数据中获取一个未处理的目标合并变异数据；调用计算节点对目标合并变异数据进行群体变异检测，以得到目标合并变异数据所对应的群体基因组变异数据；从多个合并变异数据获取下一个未处理的目标合并变异数据，直至多个合并变异数据均完成群体变异检测。

步骤 106，对各个群体基因组变异数据进行合并，以得到多个样本的目标群体基因组变异数据。

本公开实施例的对高性能测序数据 Joint-Calling 加速的方法, 在多个样本各自对应的基因组变异数据进行处理的过程中, 根据多个样本的群体信息, 确定基因组变异数据的划分方式, 并根据划分方式, 分别对各个基因组变异数据进行划分, 以得到各个基因组变异数据各自对应的多个变异数据片段以及将各个基因组变异数据中具有相同排序的变异数据片段进行合并, 以得到合并变异数据; 分别对各个合并变异数据进行群体变异检测, 以得到各个合并变异数据各自的群体基因组变异数据, 以及对各个群体基因组变异数据进行合并, 以得到多个样本的目标群体基因组变异数据。由此, 在群体变异检测的过程中, 通过分别对多个局部合并数据进行群体异常检测, 并对检测结果进行合并, 可提高获取多个样本的完整的群体基因组变异数据的效率。

图2是根据本公开另一个实施例的对高性能测序数据 Joint-Calling 加速的方法的流程示意图。其中, 需要说明的是, 本实施例是对上述实施例的进一步细化。

如图2所示, 该对高性能测序数据 Joint-Calling 加速的方法, 可以包括:

步骤 201, 获取多个样本各自对应的基因组变异数据。

步骤 202, 根据多个样本的群体信息, 确定基因组变异数据的划分方式。

步骤 203, 根据划分方式, 分别对各个基因组变异数据进行划分, 以得到各个基因组变异数据各自对应的多个变异数据片段, 其中, 变异数据片段的排序是基于变异数据片段在基因组变异数据中的位置确定的。

步骤 204, 将各个基因组变异数据中具有相同排序的变异数据片段进行合并, 以得到合并变异数据。

步骤 205, 分别对各个合并变异数据进行群体变异检测, 以得到各个合并变异数据各自的群体基因组变异数据。

其中, 需要说明的是, 关于步骤 201 至步骤 205 的具体实现方式, 可参见本公开实施例的相关描述, 此处不再赘述。

步骤 206, 确定各个群体基因组变异数据所对应的染色体区域。

步骤 207, 将处于同一个染色体区域的群体基因组变异数据进行合并处理, 以得到多个样本的目标群体基因组变异数据。

其中, 目标群体基因组变异数据可以但不限于 25 个染色区域各自对应的群体基因合并结果。其中, 每个染色区域各自对应的群体基因合并结果是通过对应染色区域所对应的群体基因组变异数据进行合并处理而得到的。

在本示例中, 在多个样本各自对应的基因组变异数据进行处理的过程中, 根据多个样本的群体信息, 确定基因组变异数据的划分方式, 并根据划分方式, 分别对各个基因组变异数据进行划分, 以得到各个基因组变异数据各自对应的多个变异数据片段以及将各个基因组变异数据中具有相同排序的变异数据片段进行合并, 以得到合并变异数据, 并分别对各个合并变异数据进行群体变异检测, 以得到各个合并变异数据各自的群体基因组变异数

据，按照染色体区域对各个群体基因组变异数据进行合并，以得到多个样本的目标群体基因组变异数据。由此，通过按照染色体区域对局部群体基因组变异数据进行合并，可方便简单地得到多个样本的每个染色区域所对应的全部群体基因组变异数据。

基于上述任意一个实施例的基础上，为了提高所得到的群体基因组差异数据的准确性，在一些示例中，可结合多个样本的群体信息来确定对各个基因组变异数据进行划分的方式，并基于所确定出的划分方式分别对各个基因组变异数据进行划分，并基于划分后的数据进行后续出，为了可以清楚理解该过程，下面结合图 3 对该过程进行示例性描述。

图 3 是根据本公开另一个实施例的对高性能测序数据 Joint-Calling 加速的方法的流程示意图。

如图 3 所示，该对高性能测序数据 Joint-Calling 加速的方法可以包括：

步骤 301，获取多个样本各自对应的基因组变异数据。

其中，关于步骤 301 的具体实现方式，可参见本公开实施例的相关描述，此处不再赘述。

步骤 302，比较多个样本的群体信息是否相同，以得到比较结果。

步骤 303，在比较结果为多个样本的群体信息相同的情况下，确定基因组变异数据的划分方式为第一划分方式。

其中，第一划分方式用于指示基于变异位点密度进行划分的方式。

步骤 304，在划分方式为第一划分方式的情况下，针对每个基因组变异数据，将基因组变异数据按照变异位点密度进行划分，以得到基因组变异数据对应的多个变异数据片段。

其中，关于步骤 304 的具体实现方式，可参见本公开实施例中的相关描述，此处不再赘述。

步骤 305，在比较结果为多个样本的群体信息不相同的情况下，确定基因组变异数据的划分方式为第二划分方式。

其中，第二划分方式用于指示基于等位基因密度进行划分的方式。

步骤 306，在划分方式为第二划分方式的情况下，针对每个基因组变异数据，将基因组变异数据按照等位基因密度进行划分，以得到基因组变异数据对应的多个变异数据片段。

其中，变异数据片段的排序是基于变异数据片段在基因组变异数据中的位置确定的。

其中，需要说明的是，关于步骤 306 的具体实现方式，可参考本公开实施例的相关描述，此处不再赘述。

步骤 307，将各个基因组变异数据中具有相同排序的变异数据片段进行合并，以得到合并变异数据。

其中，关于步骤 307 的具体实现方式，可参见本公开实施例的相关描述，此处不再赘述。

步骤 308，分别对各个合并变异数据进行群体变异检测，以得到各个合并变异数据各自

的群体基因组变异数据。

步骤 309, 对各个群体基因组变异数据进行合并, 以得到多个样本的目标群体基因组变异数据。

其中, 需要说明的是, 关于步骤 307 至步骤 309 的具体实现方式, 可参见本公开实施
5 例的相关描述, 此处不再赘述。

在本示例中, 基于多个样本的群体信息来判断多个样本是否来自于同一个群体, 并基于判断结果来确定对基因组变异数据的划分方式, 并基于所确定出划分方式对各个基因组变异数据进行划分, 得到各个基因组变异数据各自对应的多个变异数据片段以及将各个基因组变异数据中具有相同排序的变异数据片段进行合并, 以得到合并变异数据, 并分别对
10 各个合并变异数据进行群体变异检测, 以得到各个合并变异数据各自的群体基因组变异数据, 按照染色体区域对各个群体基因组变异数据进行合并, 以得到多个样本的目标群体基因组变异数据。由此, 可以提高所得到的群体基因组变异数据的准确性。

在本公开的一个实施例中, 在一些场景中, 基因组上可能有一段很长距离上没有变异位点的区域, 为了方便后续基于染色体区域对各个局部群体基因组变异数据进行合并处理,
15 在采用基于多个样本的群体信息所确定出的划分方式对基因组数据进行划分之前, 还可以基于预设划分长度对基因组数据进行划分, 为了可以清楚理解本公开, 下面结合图 4 对该实施例的方法进行示例性描述。

图 4 是根据本公开另一个实施例的对高性能测序数据 Joint-Calling 加速的方法的流程示意图。

20 如图 4 所示, 该对高性能测序数据 Joint-Calling 加速的方法可以包括:

步骤 401, 获取多个样本各自对应的基因组变异数据。

步骤 402, 根据多个样本的群体信息, 确定基因组变异数据的划分方式。

其中, 需要说明的是, 关于步骤 401 和步骤 402 的具体实现方式, 可参见本公开实施
例的相关描述, 此处不再赘述。

25 步骤 403, 针对每个基因组变异数据, 按照预设长度对基因组变异数据进行划分, 以得到基因组变异数据对应的多个变异子数据。

其中, 预设长度是在群体变异检测装置中预先设置的长度, 在实际应用中, 可根据实际应用需求来确定该预设长度的取值, 该实施例对此不作具体限定。

其中, 需要说明的是, 步骤 402 和步骤 403 的执行顺序不分先后。

30 步骤 404, 根据划分方式, 对多个变异子数据进行划分, 以得到基因组变异所对应的多个变异数据片段。

其中, 变异数据片段的排序是基于变异数据片段在基因组变异数据中的位置确定的。

步骤 405, 将各个基因组变异数据中具有相同排序的变异数据片段进行合并, 以得到合并变异数据。

步骤 406, 分别对各个合并变异数据进行群体变异检测, 以得到各个合并变异数据各自的群体基因组变异数据。

步骤 407, 确定各个群体基因组变异数据所对应的染色体区域。

5 步骤 408, 将处于同一个染色体区域的群体基因组变异数据进行合并处理, 以得到多个样本的目标群体基因组变异数据。

其中, 需要说明的是, 关于步骤 405 至步骤 408 的具体实施方式, 可参见本公开实施例的相关描述, 此处不再赘述。

在本示例中, 基于预设长度对基因组变异数据进行划分, 以得到基因组变异数据的多个变异子数据, 并基于多个样本的群体信息所确定出的划分方式继续对各个变异子数据进行划分, 以得到变异数据片段, 将各个基因组变异数据中具有相同排序的变异数据片段进行合并, 以得到合并变异数据, 分别对各个合并变异数据进行群体变异检测, 以得到各个合并变异数据各自的群体基因组变异数据, 确定各个群体基因组变异数据所对应的染色体区域, 以及将处于同一个染色体区域的群体基因组变异数据进行合并处理, 以得到多个样本的目标群体基因组变异数据。由此, 通过两种划分方式对基因组数据进行划分, 并基于
10 划分后的数据进行后续群体变异检测, 可提高后续所得到的目标群体基因组变异数据的准确性。

在本公开的一个实施例中, 图 5 是根据本公开另一个实施例的对高性能测序数据 Joint-Calling 加速的方法的流程示意图。

如图 5 所示, 该方法还可以包括:

20 步骤 501, 针对每个基因组变异数据, 按照预设长度, 分别对基因组变异数据的多个变异数据片段进行再次划分, 以得到基因组变异数据对应的多个变异数据子片段。

其中, 变异数据子片段的排序是基于变异数据子片段在其所属于的变异数据片段中的位置确定的。

25 其中, 预设长度是在群体变异检测装置中预先设置的长度, 在实际应用中, 可根据实际应用需求来确定该预设长度的取值, 该实施例对此不作具体限定。

步骤 502, 将各个基因组变异数据中具有相同排序的变异数据子片段进行合并, 以得到合并变异数据片段。

步骤 503, 分别对各个合并变异数据片段进行群体变异检测, 以得到各个合并变异数据片段各自的群体基因组变异数据。

30 步骤 504, 对各个合并变异数据片段各自的群体基因组变异数据进行合并处理, 以得到多个样本对应的合并基因组变异数据。

在本示例中, 基于多个样本的群体信息所确定出的划分方式对基因组数据进行划分, 得到基因组变异数据的多个变异数据片段后, 继续基于预设长度分别对基因组变异数据的多个变异数据片段进行划分, 以得到变异数据子片段, 将各个基因组变异数据中具有相同

排序的变异数据子片段进行合并，以得到合并变异数据片段，分别对各个合并变异数据片段进行群体变异检测，以得到各个合并变异数据片段各自的群体基因组变异数据，对各个合并变异数据片段各自的群体基因组变异数据进行合并处理，以得到多个样本对应的合并基因组变异数。由此，通过两种划分方式对基因组数据进行划分，并基于划分后的数据进行后续群体变异检测，可提高后续所得到的群体基因组变异数据的准确性。

为了可以清楚了解本公开，下面结合图 6，对该实施例的对高性能测序数据 Joint-Calling 加速的方法进行示例性描述。

图 6 是根据本公开另一个实施例的对高性能测序数据 Joint-Calling 加速的方法的流程示意图。

如图 6 所示，该对高性能测序数据 Joint-Calling 加速的方法可以包括：

步骤 601，分别多个样本各自对应的基因组变异数据进行划分，以得到每个基因组变异数据各自对应的多个变异数据片段。

其中，需要说明的是，关于对基因组变异数据进行划分的划分方式，可基于多个样本的群体信息来确定。

其中，关于基于多个样本的群体信息确定对基因组变异数据进行划分的划分方式的具体描述，可参见本公开实施例的相关描述，此处不再赘述。

步骤 602，将各个基因组变异数据中具有相同排序的变异数据片段进行合并，以得到合并变异数据。

步骤 603，分别对各个合并变异数据进行群体变异检测，以得到各个合并变异数据各自的群体基因组变异数据。

步骤 604，根据各个群体基因组变异数据所对应的染色体区域，将处于同一个染色体区域的群体基因组变异数据进行合并处理，以得到多个样本的对应染色体区域的目标群体基因组变异数据。

其中，需要说明的是，关于步骤 602 至步骤 604 的具体实现方式，可参见本公开实施例的相关描述，此处不再赘述。

在本示例中，在群体变异检测的过程中，通过分别对多个局部合并数据进行群体异常检测，以得到多个局部群体基因组异常数据，并对各个局部群体基因组异常数据进行合并处理，以得到完整群体基因组数据，提高了获取完整的群体基因组数据的效率。

在一些示例性的实施方式中，在得到多个样本对应的完整的群体基因组变异数据后，如果又收集到新样本，为了节省大量计算资源和时间消耗，在本示例中，还可以采用本公开的方式对新样本进行处理，以得到新样本对应的群体基因组变异数据，然后，将原先的多个样本对应的群体基因组变异数据和新样本对应的群体基因变异数据进行合并处理，以得到最终的群体的群体基因组变异数据。

与上述几种实施例提供的群体变异检测方法相对应，本公开的一种实施例还提供一种

对高性能测序数据 Joint-Calling 加速的装置，由于本公开实施例提供的对高性能测序数据 Joint-Calling 加速的装置与上述几种实施例提供的对高性能测序数据 Joint-Calling 加速的方法相对应，因此，在对高性能测序数据 Joint-Calling 加速的方法的实施方式也适用于本实施例的对高性能测序数据 Joint-Calling 加速的装置，在本实施例中不再详细描述。

5 图7是根据本公开一个实施例的对高性能测序数据 Joint-Calling 加速的装置的结构示意图。

如图7所示，该对高性能测序数据 Joint-Calling 加速的装置700包括：获取模块701、第一确定模块702、第一划分模块703、第一合并处理模块704、第一群体变异检测模块705和第二合并处理模块706，其中：

10 获取模块701，用于获取多个样本各自对应的基因组变异数据；

第一确定模块702，用于根据多个样本的群体信息，确定基因组变异数据的划分方式；

第一划分模块703，用于根据划分方式，分别对各个基因组变异数据进行划分，以得到各个基因组变异数据各自对应的多个变异数据片段，其中，变异数据片段的排序是基于变异数据片段在基因组变异数据中的位置确定的；

15 第一合并处理模块704，用于将各个基因组变异数据中具有相同排序的变异数据片段进行合并，以得到合并变异数据；

第一群体变异检测模块705，用于分别对各个合并变异数据进行群体变异检测，以得到各个合并变异数据各自的群体基因组变异数据；

20 第二合并处理模块706，用于对各个群体基因组变异数据进行合并，以得到多个样本的目标群体基因组变异数据。

在本公开的一个实施例中，第二合并处理模块706，具体用于：确定各个群体基因组变异数据所对应的染色体区域：将处于同一个染色体区域的群体基因组变异数据进行合并处理，以得到多个样本的目标群体基因组变异数据。

25 在本公开的一个实施例中，在图7所示的装置实施例的基础上，如图8所示，第一确定模块702可以包括：

比较单元7021，用于比较多个样本的群体信息是否相同，以得到比较结果；

确定单元7022，用于根据比较结果，确定基因组变异数据的划分方式。

30 在本公开的一个实施例中，确定单元7022，具体用于：在比较结果为多个样本的群体信息相同的情况下，确定基因组变异数据的划分方式为第一划分方式；在比较结果为多个样本的群体信息不相同的情况下，确定基因组变异数据的划分方式为第二划分方式，其中，第一划分方式用于指示基于变异位点密度进行划分的方式，第二划分方式用于指示基于等位基因密度进行划分的方式。

在本公开的一个实施例中，在划分方式为第一划分方式的情况下，第一划分模块703，具体用于：针对每个基因组变异数据，将基因组变异数据按照变异位点密度进行划分，以

得到基因组变异数据对应的多个变异数据片段。

在本公开的一个实施例中，在划分方式为第二划分方式的情况下，第一划分模块 703，具体用于：针对每个基因组变异数据，将基因组变异数据按照等位基因密度进行划分，以得到基因组变异数据对应的多个变异数据片段。

5 在本公开的一个实施例中，如图 8 所示，该装置还可以包括：

第二划分模块 707，用于针对每个基因组变异数据，按照预设长度对基因组变异数据进行划分，以得到基因组变异数据对应的多个变异子数据；

第一划分模块 703，具体用于：根据划分方式，对多个变异子数据进行划分，以得到基因组变异所对应的多个变异数据片段。

10 在本公开的一个实施例中，如图 8 所示，该装置还可以包括：

第三划分模块 708，用于针对每个基因组变异数据，按照预设长度，分别对基因组变异数据的多个变异数据片段进行再次划分，以得到基因组变异数据对应的多个变异数据子片段，其中，变异数据子片段的排序是基于变异数据子片段在其所属的变异数据片段中的位置确定的；

15 第三合并处理模块 709，用于将各个基因组变异数据中具有相同排序的变异数据子片段进行合并，以得到合并变异数据片段；

第二群体变异检测模块 710，用于分别对各个合并变异数据片段进行群体变异检测，以得到各个合并变异数据片段各自的群体基因组变异数据。

20 在本公开的一个实施例中，在通过多个计算节点对多个合并变异数据进行群体变异检测的情况下，第一群体变异检测模块 705，具体用于：针对每个计算节点：从多个合并变异数据中获取一个未处理的目标合并变异数据；调用计算节点对目标合并变异数据进行群体变异检测，以得到目标合并变异数据所对应的群体基因组变异数据；从多个合并变异数据获取下一个未处理的目标合并变异数据，直至多个合并变异数据均完成群体变异检测。

25 本公开实施例的对高性能测序数据 Joint-Calling 加速的装置，在多个样本各自对应的基因组变异数据进行处理的过程中，根据多个样本的群体信息，确定基因组变异数据的划分方式，并根据划分方式，分别对各个基因组变异数据进行划分，以得到各个基因组变异数据各自对应的多个变异数据片段以及将各个基因组变异数据中具有相同排序的变异数据片段进行合并，以得到合并变异数据；分别对各个合并变异数据进行群体变异检测，以得到各个合并变异数据各自的群体基因组变异数据，以及对各个群体基因组变异数据进行合并，
30 以得到多个样本的目标群体基因组变异数据。由此，在群体变异检测的过程中，通过分别对多个局部合并数据进行群体异常检测，并对检测结果进行合并，可提高获取多个样本的完整的群体基因组变异数据的效率。

根据本公开的实施例，本公开还提供了一种电子设备和一种可读存储介质。

图 9 是根据本公开一个实施例的电子设备的结构框图。

如图 9 所示, 该电子设备 900 包括: 存储器 910、处理器 920 及存储在存储器 910 上并可在处理器 920 上运行的计算机指令。

处理器 920 执行指令时实现上述实施例中提供的对高性能测序数据 Joint-Calling 加速的方法。

5 进一步地, 电子设备 900 还包括:

通信接口 930, 用于存储器 910 和处理器 920 之间的通信。

存储器 910, 用于存放可在处理器 920 上运行的计算机指令。

存储器 910 可能包含高速 RAM 存储器, 也可能还包括非易失性存储器 (non-volatile memory), 例如至少一个磁盘存储器。

10 处理器 920, 用于执行程序时实现上述实施例的对高性能测序数据 Joint-Calling 加速的方法。

如果存储器 910、处理器 920 和通信接口 930 独立实现, 则通信接口 930、存储器 910 和处理器 920 可以通过总线相互连接并完成相互间的通信。总线可以是工业标准体系结构 (Industry Standard Architecture, 简称为 ISA) 总线、外部设备互连 (Peripheral Component, 简称为 PCI) 总线或扩展工业标准体系结构 (Extended Industry Standard Architecture, 简称为 EISA) 总线等。总线可以分为地址总线、数据总线、控制总线等。为便于表示, 图 9 中
15 仅用一条粗线表示, 但并不表示仅有一根总线或一种类型的总线。

可选的, 在具体实现上, 如果存储器 910、处理器 920 及通信接口 930, 集成在一块芯片上实现, 则存储器 910、处理器 920 及通信接口 930 可以通过内部接口完成相互间的通
20 信。

处理器 920 可能是一个中央处理器 (Central Processing Unit, 简称为 CPU), 或者是特定集成电路 (Application Specific Integrated Circuit, 简称为 ASIC), 或者是被配置成实施本公开实施例的一个或多个集成电路。

本公开另一方面实施例提出了一种计算机可读存储介质, 其上存储有计算机程序, 该
25 程序被处理器执行时实现本公开实施例任一的对高性能测序数据 Joint-Calling 加速的装置。

在本说明书的描述中, 参考术语“一个实施例”、“一些实施例”、“示例”、“具体示例”、或“一些示例”等的描述意指结合该实施例或示例描述的具体特征、结构、材料或者特点包含于本发明的至少一个实施例或示例中。在本说明书中, 对上述术语的示意性表述不必须针对的是相同的实施例或示例。而且, 描述的具体特征、结构、材料或者特点可以在任一
30 个或多个实施例或示例中以合适的方式结合。此外, 在不相互矛盾的情况下, 本领域的技术人员可以将本说明书中描述的不同实施例或示例以及不同实施例或示例的特征进行结合和组合。

此外, 术语“第一”、“第二”仅用于描述目的, 而不能理解为指示或暗示相对重要性或者隐含指明所指示的技术特征的数量。由此, 限定有“第一”、“第二”的特征可以明示或者隐

含地包括至少一个该特征。在本发明的描述中，“多个”的含义是至少两个，例如两个，三个等，除非另有明确具体的限定。

尽管上面已经示出和描述了本发明的实施例，可以理解的是，上述实施例是示例性的，不能理解为对本发明的限制，本领域的普通技术人员在本发明的范围内可以对上述实施例

5 进行变化、修改、替换和变型。

权利要求书

1.一种对高性能测序数据 Joint-Calling 加速的方法,其特征在于,所述方法包括:

获取多个样本各自对应的基因组变异数据;

5 根据所述多个样本的群体信息,确定所述基因组变异数据的划分方式;

根据所述划分方式,分别对各个所述基因组变异数据进行划分,以得到各个所述基因组变异数据各自对应的多个变异数据片段,其中,所述变异数据片段的排序是基于所述变异数据片段在所述基因组变异数据中的位置确定的;

10 将各个所述基因组变异数据中具有相同排序的变异数据片段进行合并,以得到合并变异数据;

分别对各个所述合并变异数据进行群体变异检测,以得到各个所述合并变异数据各自的群体基因组变异数据;

对各个所述群体基因组变异数据进行合并,以得到所述多个样本的目标群体基因组变异数据。

15 2.如权利要求1所述的方法,其特征在于,所述对各个所述群体基因组变异数据进行合并,以得到所述多个样本的目标群体基因组变异数据,包括:

确定各个所述群体基因组变异数据所对应的染色体区域;

将处于同一个染色体区域的所述群体基因组变异数据进行合并处理,以得到所述多个样本的目标群体基因组变异数据。

20 3.如权利要求1所述的方法,其特征在于,所述根据所述多个样本的群体信息,确定所述基因组变异数据的划分方式,包括:

比较所述多个样本的群体信息是否相同,以得到比较结果;

根据所述比较结果,确定所述基因组变异数据的划分方式。

25 4.如权利要求3所述的方法,其特征在于,所述根据所述比较结果,确定所述基因组变异数据的划分方式,包括:

在所述比较结果为所述多个样本的群体信息相同的情况下,确定所述基因组变异数据的划分方式为第一划分方式;

30 在所述比较结果为所述多个样本的群体信息不相同的情况下,确定所述基因组变异数据的划分方式为第二划分方式,其中,所述第一划分方式用于指示基于变异位点密度进行划分的方式,所述第二划分方式用于指示基于等位基因密度进行划分的方式。

5.如权利要求4所述的方法,其特征在于,在所述划分方式为所述第一划分方式的情况下,所述根据所述划分方式,分别对各个所述基因组变异数据进行划分,以得到各个所述基因组变异数据各自对应的多个变异数据片段,包括:

针对每个所述基因组变异数据,将所述基因组变异数据按照变异位点密度进行划分,

以得到所述基因组变异数据对应的多个变异数据片段。

6.如权利要求5所述的方法，其特征在于，针对每个所述基因组变异数据，将所述基因组变异数据按照变异位点密度进行划分，以得到所述基因组变异数据对应的多个变异数据片段，包括：

5 针对每个所述基因组变异数据，对所述基因组变异数据中的变异位点的分布情况进行统计，以得到该基因组变异数据的变异位点统计结果；

根据所述变异位点统计结果，基于变异位点密度对所述基因组变异数据进行划分，以得到所述基因组变异数据对应的多个变异数据片段，其中，所述多个变异数据片段所对应的长度是不同的，并且所述多个变异数据片段上的变异位点数量之间的数量差在第一预设数量之内。

7.如权利要求4所述的方法，其特征在于，在所述划分方式为第二划分方式的情况下，所述根据所述划分方式，分别对各个所述基因组变异数据进行划分，以得到各个所述基因组变异数据各自对应的多个变异数据片段，包括：

针对每个所述基因组变异数据，将所述基因组变异数据按照等位基因密度进行划分，以得到所述基因组变异数据对应的多个变异数据片段。

8.如权利要求7所述的方法，其特征在于，针对每个所述基因组变异数据，将所述基因组变异数据按照等位基因密度进行划分，以得到所述基因组变异数据对应的多个变异数据片段，包括：

针对所述基因组变异数据，确定出所述基因组变异数据中每个变异位点上所对应的等位基因信息；

对所述基因组变异数据中的等位基因信息进行统计，以得到所述基因组变异数据的等位基因信息的统计结果；

根据所述统计结果，基于基因等位基因密度对所述基因组变异数据进行划分，以得到所述基因组变异数据对应的多个变异数据片段，其中，所述多个变异数据片段所对应的长度是不同的，并且，所述多个变异数据片段中等位基因的数量之间的差值在第二预设数量内。

9.如权利要求2所述的方法，其特征在于，根据所述划分方式，分别对各个所述基因组变异数据进行划分，以得到各个所述基因组变异数据各自对应的多个变异数据片段之前，所述方法还包括：

30 针对每个所述基因组变异数据，按照预设长度对所述基因组变异数据进行划分，以得到所述基因组变异数据对应的多个变异子数据；

所述根据所述划分方式，分别对各个所述基因组变异数据进行划分，以得到各个所述基因组变异数据各自对应的多个变异数据片段，包括：

根据所述划分方式，对所述多个变异子数据进行划分，以得到所述基因组变异所对应

的多个变异数据片段。

10.如权利要求1所述的方法，其特征在于，所述方法还包括：

针对每个基因组变异数据，按照预设长度，分别对所述基因组变异数据的多个变异数据片段进行再次划分，以得到所述基因组变异数据对应的多个变异数据子片段，其中，所述变异数据子片段的排序是基于所述变异数据子片段在其所属的变异数据片段中的位置确定的；

将各个所述基因组变异数据中具有相同排序的变异数据子片段进行合并，以得到合并变异数据片段；

分别对各个所述合并变异数据片段进行群体变异检测，以得到各个所述合并变异数据片段各自的群体基因组变异数据。

11.如权利要求1-10中任一项所述的方法，其特征在于，在通过多个计算节点对多个所述合并变异数据进行群体变异检测的情况下，所述分别对各个所述合并变异数据进行群体变异检测，以得到各个所述合并变异数据各自的群体基因组变异数据，包括：

针对每个计算节点：

从多个所述合并变异数据中获取一个未处理的目标合并变异数据；

调用所述计算节点对所述目标合并变异数据进行群体变异检测，以得到所述目标合并变异数据所对应的群体基因组变异数据；

从多个所述合并变异数据获取下一个未处理的目标合并变异数据，直至多个所述合并变异数据均完成群体变异检测。

12.一种对高性能测序数据 Joint-Calling 加速的装置，其特征在于，所述装置包括：

获取模块，用于获取多个样本各自对应的基因组变异数据；

第一确定模块，用于根据所述多个样本的群体信息，确定所述基因组变异数据的划分方式；

第一划分模块，用于根据所述划分方式，分别对各个所述基因组变异数据进行划分，以得到各个所述基因组变异数据各自对应的多个变异数据片段，其中，所述变异数据片段的排序是基于所述变异数据片段在所述基因组变异数据中的位置确定的；

第一合并处理模块，用于将各个所述基因组变异数据中具有相同排序的变异数据片段进行合并，以得到合并变异数据；

第一群体变异检测模块，用于分别对各个所述合并变异数据进行群体变异检测，以得到各个所述合并变异数据各自的群体基因组变异数据；

第二合并处理模块，用于对各个所述群体基因组变异数据进行合并，以得到所述多个样本的目标群体基因组变异数据。

13.如权利要求12所述的装置，其特征在于，所述第二合并处理模块，具体用于：

确定各个所述群体基因组变异数据所对应的染色体区域；

将处于同一个染色体区域的所述群体基因组变异数据进行合并处理，以得到所述多个样本的目标群体基因组变异数据。

14.如权利要求 12 所述的装置，其特征在于，所述第一确定模块，包括：

比较单元，用于比较所述多个样本的群体信息是否相同，以得到比较结果；

5 确定单元，用于根据所述比较结果，确定所述基因组变异数据的划分方式。

15.如权利要求 14 所述的装置，其特征在于，所述确定单元，具体用于：

在所述比较结果为所述多个样本的群体信息相同的情况下，确定所述基因组变异数据的划分方式为第一划分方式；

10 在所述比较结果为所述多个样本的群体信息不相同的情况下，确定所述基因组变异数据的划分方式为第二划分方式，其中，所述第一划分方式用于指示基于变异位点密度进行划分的方式，所述第二划分方式用于指示基于等位基因密度进行划分的方式。

16.如权利要求 15 所述的装置，其特征在于，在所述划分方式为所述第一划分方式的情况下，所述第一划分模块，具体用于：

15 针对每个所述基因组变异数据，将所述基因组变异数据按照变异位点密度进行划分，以得到所述基因组变异数据对应的多个变异数据片段。

17.如权利要求 16 所述的装置，其特征在于，所述第一划分模块，具体用于：

针对每个所述基因组变异数据，对所述基因组变异数据中的变异位点的分布情况进行统计，以得到该基因组变异数据的变异位点统计结果；

20 根据所述变异位点统计结果，基于变异位点密度对所述基因组变异数据进行划分，以得到所述基因组变异数据对应的多个变异数据片段，其中，所述多个变异数据片段所对应的长度是不同的，并且所述多个变异数据片段上的变异位点数量之间的数量差在第一预设数量之内。

18.如权利要求 15 所述的装置，其特征在于，在所述划分方式为第二划分方式的情况下，所述第一划分模块，具体用于：

25 针对每个所述基因组变异数据，将所述基因组变异数据按照等位基因密度进行划分，以得到所述基因组变异数据对应的多个变异数据片段。

19.如权利要求 18 所述的装置，其特征在于，所述第一划分模块，具体用于：

针对所述基因组变异数据，确定出所述基因组变异数据中每个变异位点上所对应的等位基因信息；

30 对所述基因组变异数据中的等位基因信息进行统计，以得到所述基因组变异数据的等位基因信息的统计结果；

根据所述统计结果，基于基因等位基因密度对所述基因组变异数据进行划分，以得到所述基因组变异数据对应的多个变异数据片段，其中，所述多个变异数据片段所对应的长度是不同的，并且，所述多个变异数据片段中等位基因的数量之间的差值在第二预设数量

内。

20.如权利要求 14 所述的装置，其特征在于，所述装置还包括：

第二划分模块，用于针对每个所述基因组变异数据，按照预设长度对所述基因组变异数据进行划分，以得到所述基因组变异数据对应的多个变异子数据；

5 所述第一划分模块，具体用于：

根据所述划分方式，对所述多个变异子数据进行划分，以得到所述基因组变异所对应的多个变异数据片段。

21.如权利要求 15 所述的装置，其特征在于，所述装置还包括：

10 第三划分模块，用于针对每个基因组变异数据，按照预设长度，分别对所述基因组变异数据的多个变异数据片段进行再次划分，以得到所述基因组变异数据对应的多个变异数据子片段，其中，所述变异数据子片段的排序是基于所述变异数据子片段在其所属于的变异数据片段中的位置确定的；

第三合并处理模块，用于将各个所述基因组变异数据中具有相同排序的变异数据子片段进行合并，以得到合并变异数据片段；

15 第二群体变异检测模块，用于分别对各个所述合并变异数据片段进行群体变异检测，以得到各个所述合并变异数据片段各自的群体基因组变异数据。

22.如权利要求 12-21 中任一项所述的装置，其特征在于，在通过多个计算节点对多个所述合并变异数据进行群体变异检测的情况下，所述第一群体变异检测模块，具体用于：

针对每个计算节点：

20 从多个所述合并变异数据中获取一个未处理的目标合并变异数据；

调用所述计算节点对所述目标合并变异数据进行群体变异检测，以得到所述目标合并变异数据所对应的群体基因组变异数据；

从多个所述合并变异数据获取下一个未处理的目标合并变异数据，直至多个所述合并变异数据均完成群体变异检测。

25 23.一种电子设备，其特征在于，包括：

存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，其特征在于，所述处理器执行所述程序时实现如权利要求 1-11 中任一所述的方法。

24.一种计算机可读存储介质，其上存储有计算机程序，其特征在于，该程序被处理器执行时实现如权利要求 1-11 中任一所述的方法。

30

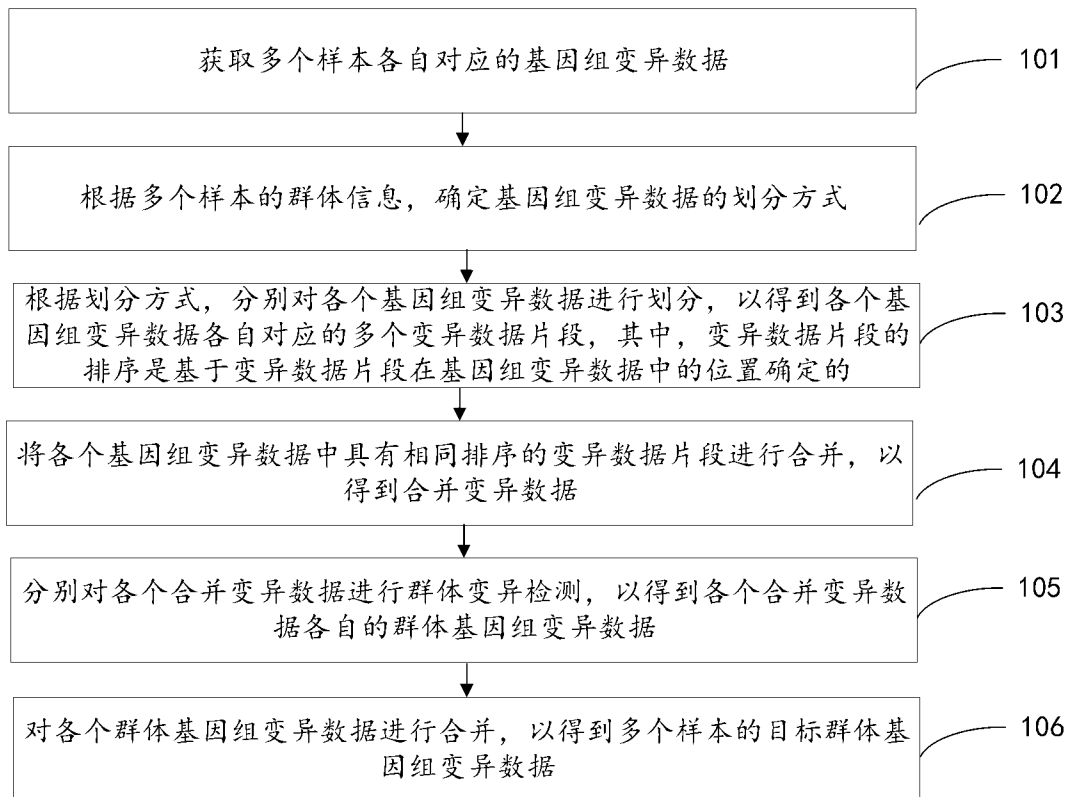


图 1

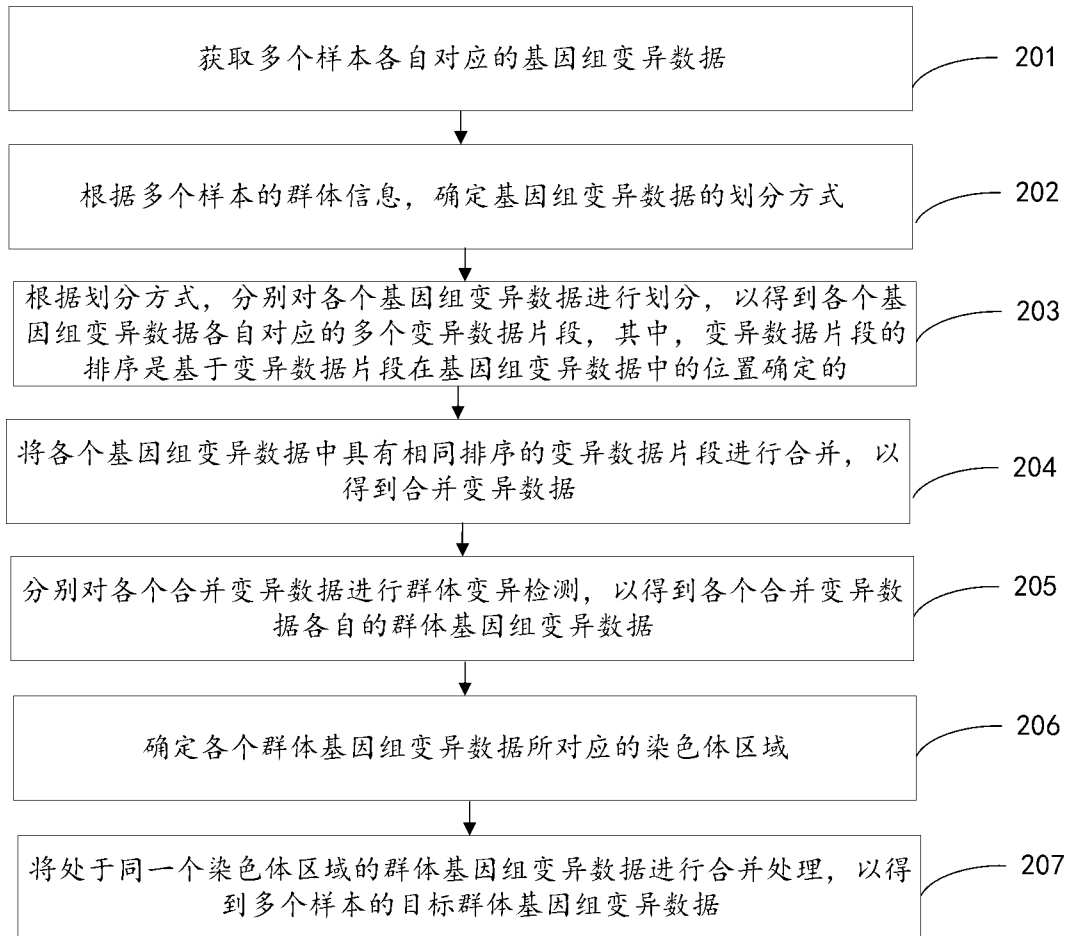


图 2

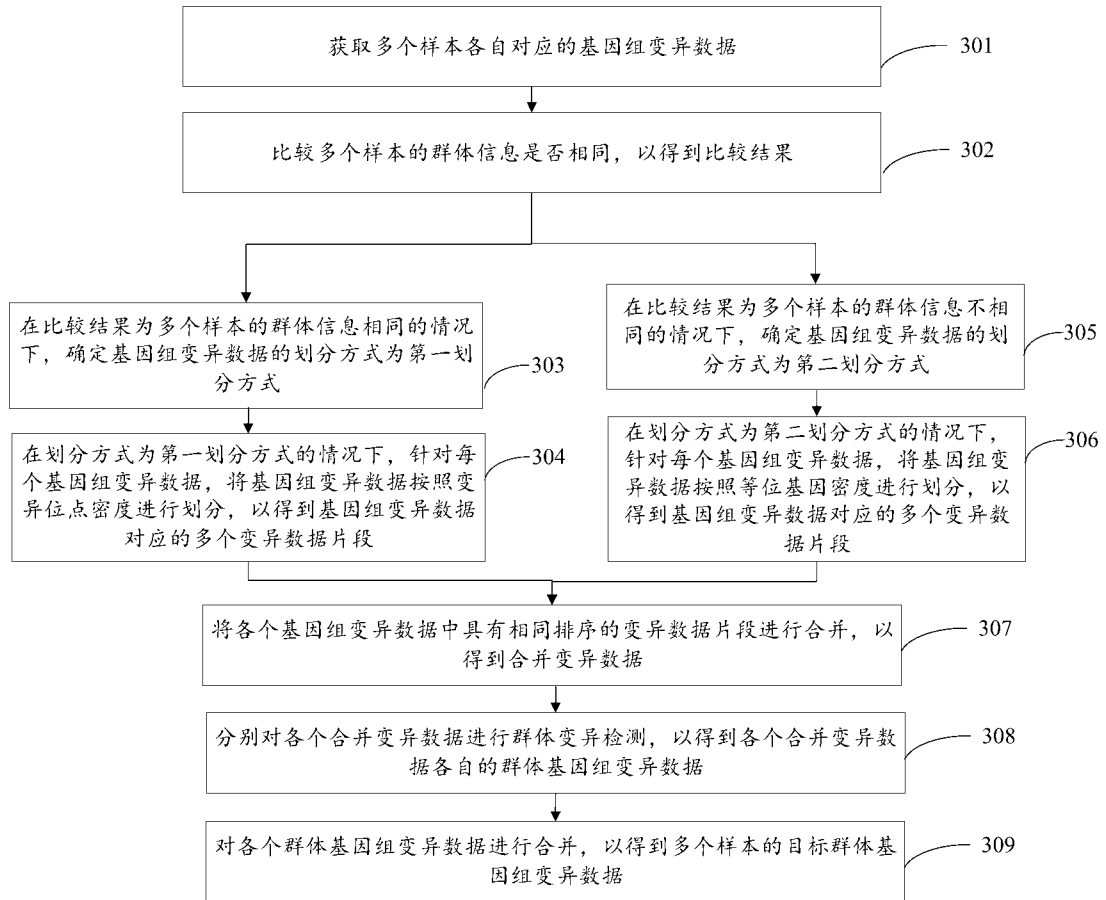


图 3



图 4

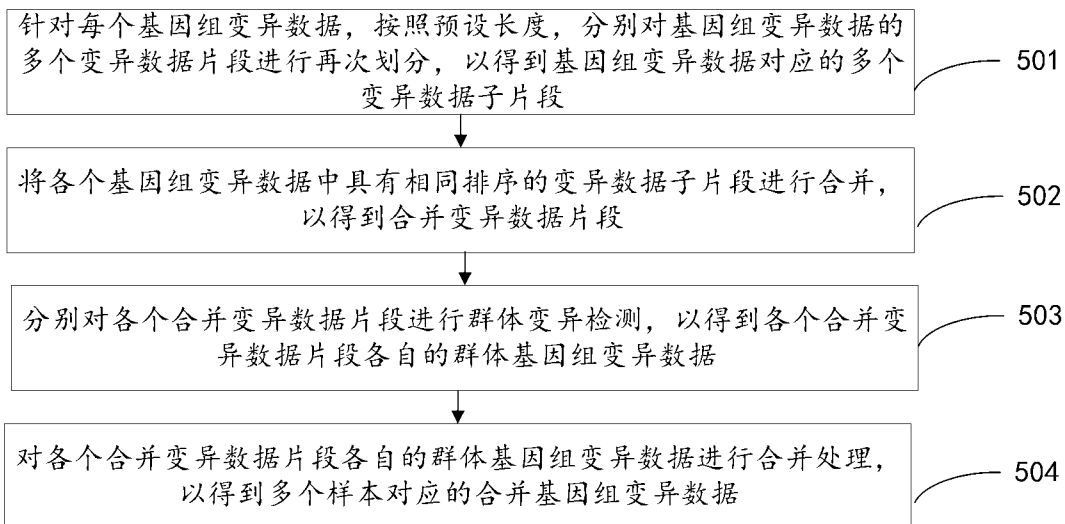


图 5

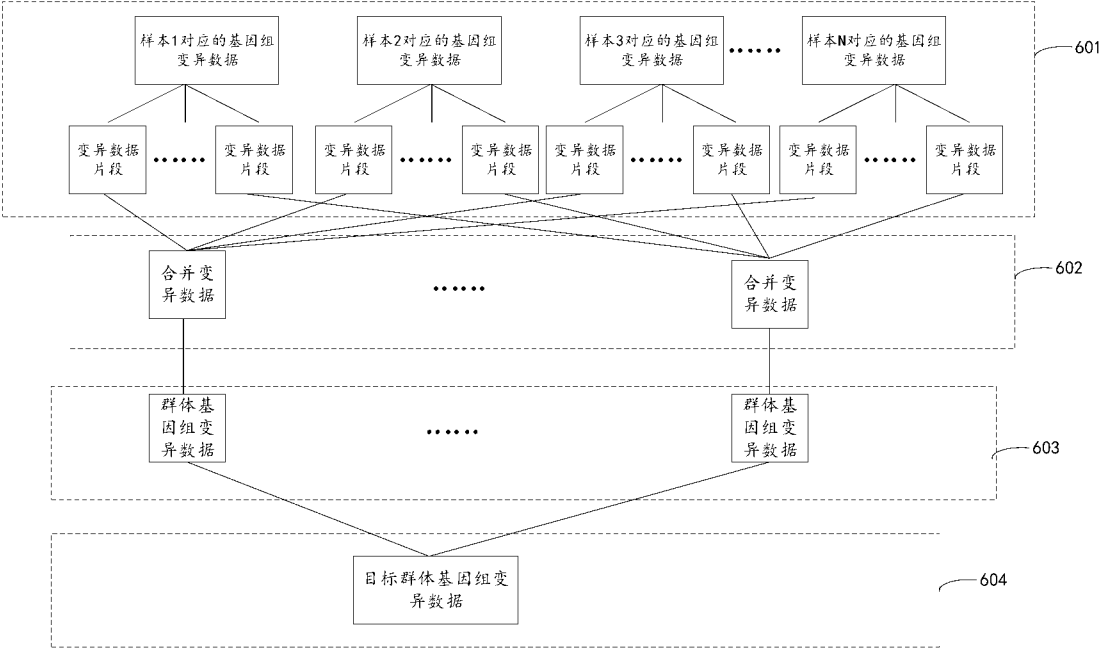


图 6

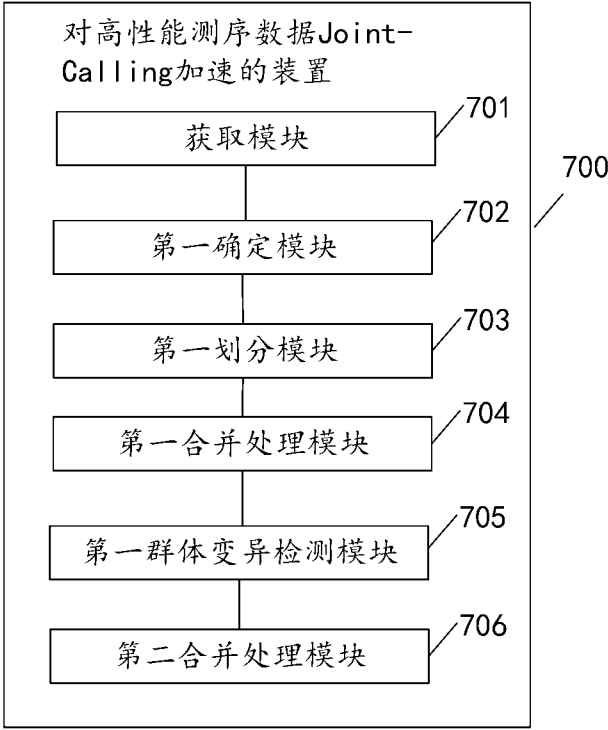


图 7

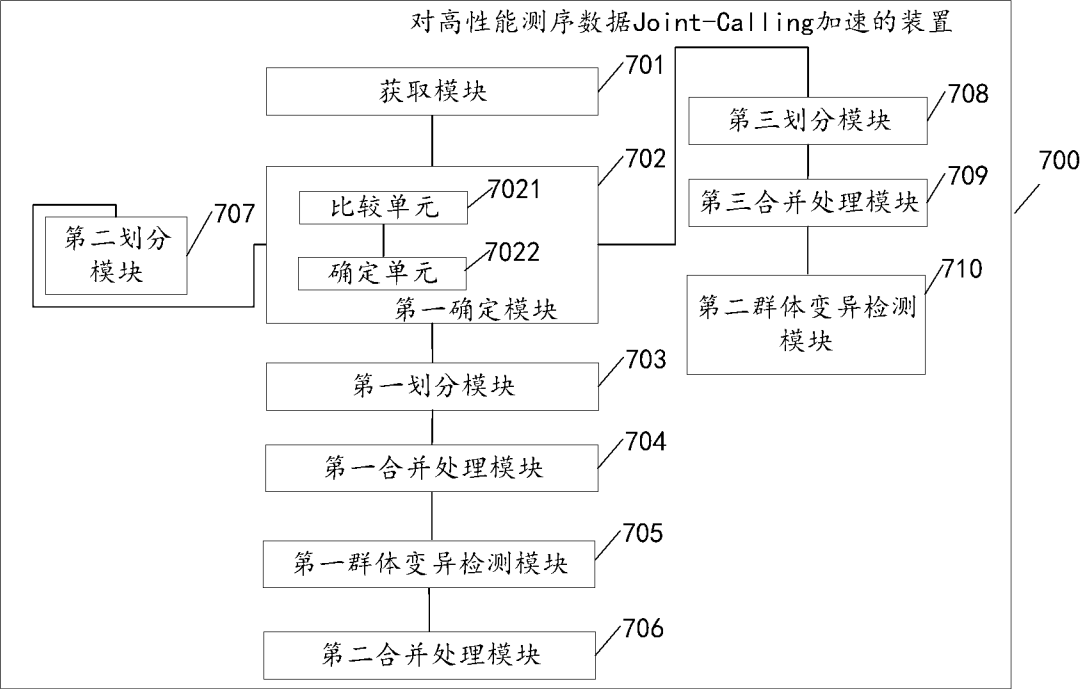


图 8

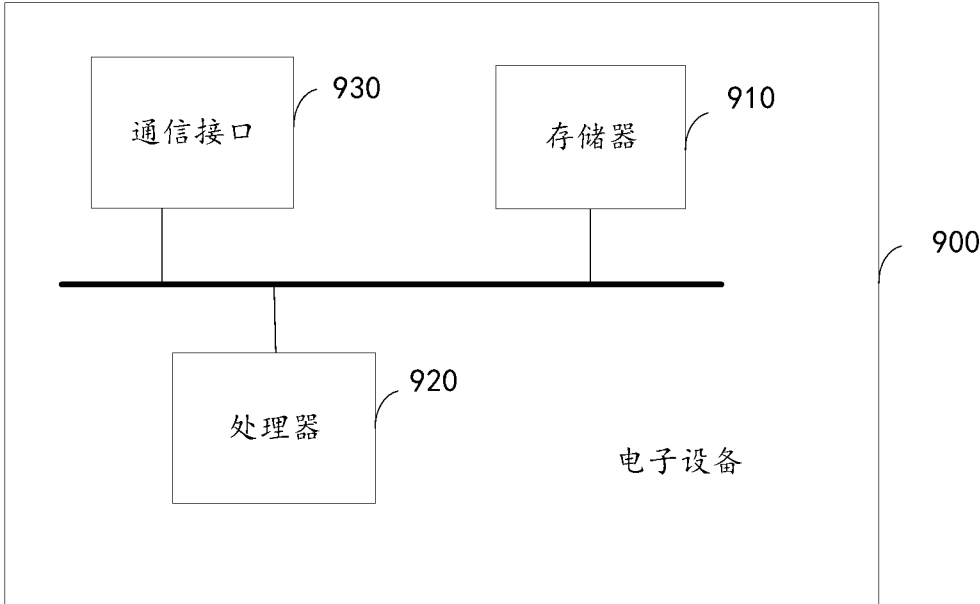


图 9

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2023/105780

A. CLASSIFICATION OF SUBJECT MATTER G16B20/20(2019.01)i According to International Patent Classification (IPC) or to both national classification and IPC		
B. FIELDS SEARCHED Minimum documentation searched (classification system followed by classification symbols) G16B G06F Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched Electronic data base consulted during the international search (name of data base and, where practicable, search terms used) VEN, CNABS, CNTXT, WOTXT, EPTXT, USTXT: 基因, 变异, 检测, 划分, 拆分, 位置, 合并, 群体, 基因组, 长度, gene, variation, detect+, split+, location, merg+, group, genome, length		
C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 114913918 A (CENTER FOR EXCELLENCE IN BRAIN SCIENCE AND INTELLIGENCE TECHNOLOGY, CHINESE ACADEMY OF SCIENCES) 16 August 2022 (2022-08-16) description, pages 6-12	1-24
A	CN 112735516 A (SHANGHAI PERSONAL BIOTECHNOLOGY CO., LTD.) 30 April 2021 (2021-04-30) entire document	1-24
A	CN 115331730 A (ZHENGZHOU JINYU CLINIC EXAMINATION CENTER CO., LTD.) 11 November 2022 (2022-11-11) entire document	1-24
A	CN 115910199 A (HARBIN INSTITUTE OF TECHNOLOGY) 04 April 2023 (2023-04-04) entire document	1-24
A	US 2015132295 A1 (POPULATION DIAGNOSTICS INC.) 14 May 2015 (2015-05-14) entire document	1-24
☐ Further documents are listed in the continuation of Box C. ☒ See patent family annex.		
* Special categories of cited documents: “A” document defining the general state of the art which is not considered to be of particular relevance “D” document cited by the applicant in the international application “E” earlier application or patent but published on or after the international filing date “L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified) “O” document referring to an oral disclosure, use, exhibition or other means “P” document published prior to the international filing date but later than the priority date claimed “T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention “X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone “Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art “&” document member of the same patent family		
Date of the actual completion of the international search 20 December 2023		Date of mailing of the international search report 28 December 2023
Name and mailing address of the ISA/CN China National Intellectual Property Administration (ISA/CN) China No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing 100088		Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2023/105780

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	114913918	A	16 August 2022	None			
CN	112735516	A	30 April 2021	None			
CN	115331730	A	11 November 2022	None			
CN	115910199	A	04 April 2023	None			
US	2015132295	A1	14 May 2015	US	2022380851	A1	01 December 2022
				DK	3511422	T3	06 February 2023
				EP	2891722	A1	08 July 2015
				US	2019249252	A1	15 August 2019
				FI	3511422	T3	31 January 2023
				EP	3511422	A2	17 July 2019

A. 主题的分类

G16B20/20(2019.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G16B G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用))

VEN, CNABS, CNTXT, WOTXT, EPTXT, USTXT: 基因, 变异, 检测, 划分, 拆分, 位置, 合并, 群体, 基因组, 长度, gene, variation, detect+, split+, location, merg+, group, genome, length

C. 相关文件

类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
A	CN 114913918 A (中国科学院脑科学与智能技术卓越创新中心) 2022年8月16日 (2022 - 08 - 16) 说明书第6-12页	1-24
A	CN 112735516 A (上海派森诺生物科技股份有限公司) 2021年4月30日 (2021 - 04 - 30) 全文	1-24
A	CN 115331730 A (郑州金域临床检验中心有限公司) 2022年11月11日 (2022 - 11 - 11) 全文	1-24
A	CN 115910199 A (哈尔滨工业大学) 2023年4月4日 (2023 - 04 - 04) 全文	1-24
A	US 2015132295 A1 (POPULATION DIAGNOSTICS INC.) 2015年5月14日 (2015 - 05 - 14) 全文	1-24

☐ 其余文件在C栏的续页中列出。

☒ 见同族专利附件。

★ 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“D” 申请人在国际申请中引证的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件(如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“p” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期

2023年12月20日

国际检索报告邮寄日期

2023年12月28日

ISA/CN的名称和邮寄地址

中国国家知识产权局
中国北京市海淀区蓟门桥西土城路6号 100088

授权官员

李楠

电话号码 (+86) 010-53961438

PCT/ISA/210 表(第2页) (2022年7月)

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2023/105780

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
CN	114913918	A	2022年8月16日	无			
CN	112735516	A	2021年4月30日	无			
CN	115331730	A	2022年11月11日	无			
CN	115910199	A	2023年4月4日	无			
US	2015132295	A1	2015年5月14日	US	2022380851	A1	2022年12月1日
				DK	3511422	T3	2023年2月6日
				EP	2891722	A1	2015年7月8日
				US	2019249252	A1	2019年8月15日
				FI	3511422	T3	2023年1月31日
				EP	3511422	A2	2019年7月17日