DOI:10.12113/202106001

群体基因组结构变异检测工作流

曹舒淇,刘诗琦,姜 涛*
(哈尔滨工业大学 计算学部,哈尔滨 150001)

摘 要:结构变异作为人类基因组上的一种大规模的变异类型,对分子与细胞进程、调节功能、基因表达调控、个体表型具有重要的影响,检测群体中基因组结构变异有助于绘制群体基因组变异图谱,刻画群体遗传进化特征,为疾病诊治、精准医疗的发展提供支撑。本研究提出一种面向高通量测序的群体基因组结构变异检测工作流,该工作流通过使用多种高性能基因组结构变异检测算法实现全面、精准的结构变异挖掘,使用多层融合与过滤获得高精度群体结构变异候选集合,利用基因型重新校正、变异修剪、类型校对,最终完整绘制群体基因组结构变异图谱。基于该工作流对由267个样本组成的人群进行群体结构变异检测,检测出了96 202个结构变异,其变异种类和频率分布与其他国际基因组计划相符,这些结果证明了本工作流具有良好的群体结构变异检测能力。同时,工作流通过并行的方式在内存可控的基础上显著降低了分析时间,为大规模人群基因组结构变异的高效检测提供了重要支撑。

关键词:群体基因组;结构变异;变异检测;变异融合
中图分类号:TP399 **文献标志码**:A **文章编号**:1672-5565(2021)04-232-08

Workflow of structural variation detection from population genomes

CAO Shuqi, LIU Shiqi, JIANG Tao*
(*Faculty of Computing*, *Harbin Institute of Technology*, *Harbin* 150001, *China*)

Abstract: Structural variation is an important type of genome variation, which affects molecular and cellular processes, regulatory functions, and brings great influence of the regulation of gene expression and individual phenotype. The accurate detection of population-scale structural variation helps to draw the full spectrum of population genome variation, which reveals the characteristics of population genetics and evolution, and gives support for disease analysis and precision medicine. This paper provides a workflow of structural variation detection from population genomes based on high-throughput sequencing data. The workflow achieves comprehensive and accurate structural variation detection through multiple high-performance structural variation detection algorithms. The multilayer integration and filter were applied to achieve set of candidate structural variation with high precision. By performing genotype correction, variation trimming, and type revising, the spectrum of structural variation of population genomes was obtained. In this study, structural variation detection was performed through the workflow on a population group containing 267 individuals, and 96 202 structural variations were reported. The types of variation and distributions of variation frequencies corresponded to those in other international genome projects, which indicates that the workflow has outstanding ability for structural variation detection from population genomes. Meanwhile, the parallel workflow significantly decreases the analysis time while maintaining the memory cost, which gives strong support for large-scale population structural variation detection.

Keywords: Population genomes; Structural variation; Variation detection; Variant integration

基因组结构变异(Structural Variation,SV)是基因组上大尺度的核苷酸序列重排性变化,它包括长度大于50 bp的插入(INS)、缺失(DEL)、倒位(INV)、重复(DUP)、易位(BND)[1]。相关研究表

收稿日期:2021-06-02;**修回日期**:2021-06-23.
基金项目:国家重点研发项目(No. 2017YFC0907503);国家自然科学基金项目(No. 32000467).
作者简介:曹舒淇,女,本科生,研究方向:生物信息学. E-mail:1170300326@stu.hit.edu.cn.
*通信作者:姜涛,男,助理教授,研究方向:生物信息学. E-mail:tjiang@hit.edu.cn.

明,平均每个人类个体上存在大约两万个结构变异[2],结构变异尽管相较于单核苷酸变异(SNV)、短插入缺失变异(INDEL)数量较少,但因其变异长度较大,因此对基因组上核苷酸序列的影响是最广泛的[3]。结构变异会改变基因序列信息,进而影响转录过程,改变蛋白质空间结构,从而引发性状与表型的改变[4]。此外,结构变异对基因表达调控[5]、种群多样性[6]等方面有着重要影响,同时与以自闭症[7]、阿尔兹海默症[8]等为代表的许多疾病的引发有密切的关系。

结构变异会对人类遗传、进化产生影响,形成个体之间的差异,影响种群的发展与演进。对于同一个群体,相当数量的结构变异对于群体中大部分个体是共享的,这些共享的结构变异可以有效对群体的特征与结构进行刻画[9]。此外,在群体中仍存在个体特有的结构变异,这些个体特有的结构变异反映了个体独有的特性,通过对特有结构变异以及个体表型的分析,能够发掘结构变异与表型、疾病之间的重要关系[10]。

随着国际千人基因组计划的实施与推动[11-13],各国也纷纷启动了本国的大规模人群基因组计划[14-17],希望通过分析和构建本国、本民族的基因组变异图谱,更加深入地解读本国人群在遗传、进化上的机理,为接下来开展的疾病诊治、精准健康发展提供支撑。结构变异作为对基因序列影响最为广泛的基因组变异类型,如何高效、精准的检测群体结构变异已成为当前群体基因组研究中的核心。因此我们基于多层过滤的质量控制,多种算法的联合检测、多维度变异融合和校对,开发了一个高性能的群体结构变异检测工作流,实现了群体基因组结构变异的全面、精准检测。该工作流总体分为四个环节:基因组测序片段比对,单样本基因组结构变异检测,单样本基因组结构变异融合以及群体基因组结构变异检测(见图 1)。

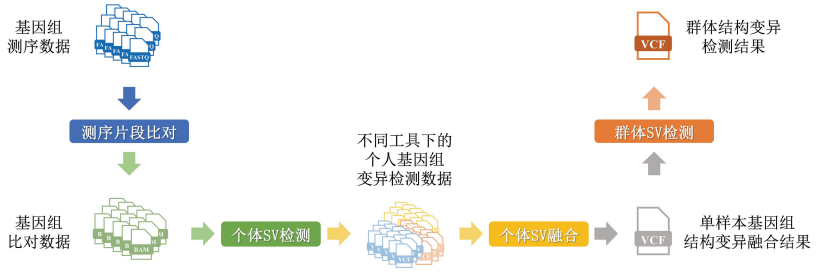


图 1　群体基因组结构变异检测工作流

Fig. 1　Workflow of structural variation detection from population genomes

1　基因组测序片段比对

高通量基因组测序片段比对是基因组数据分析的首要环节,测序片段的比对的精度将对变异检测、基因组拼接等下游分析产生重要的影响。因此,对基因组片段测序数据、片段比对数据等有效的质量控制,是保障以测序片段比对为基础的基因组数据分析的关键。为此,本研究设计了多重质量控制与过滤的基因组测序片段比对工作流,该工作流主要包含以下步骤(见图 2)。

(1)使用测序片段质量评价算法 FastQC(https://github.com/s-andrews/FastQC)(v0.11.9,默认参数)对各样本基因组测序数据进行质量控制,通过对测序片段中 GC 含量、重复性、碱基质量、片段长度分布等指标进行统计和阈值判定,若任意满足:测序片段 GC 含量与理论分布偏差 30%;测序片段重复度超过理论重复总量的 50%;测序片段任意位置的碱基质量下四分位数低于 5 或中位数低于 20;任意测序片段长度不足或长于 150 bp,则将其认定为低质量测序样本数据,并进行过滤处理。

(2)使用高通量测序片段比对算法 BWA(https://github.com/lh3/bwa)(v0.7.17,默认参数),完成各测序样本向参考基因组序列的比对。使用比对格式转换算法 Sambamba[18](v0.8.0,默认参数)对各样本比对结果进行格式转换和排序;使用测序重复片段标记算法 Samblaster[19](v0.1.2,默认参数)对各样本转换后的比对文件进行重复标记;使用 GATK(https://github.com/broadinstitute/gatk)(v4.2.0.0,默认参数)对测序片段比对中碱基质量校正形成最终的片段比对数据。

(3)使用测序片段比对质量评价算法 Qualimap[20](v2.2.1,默认参数)对各样本测序片段比对结果进行质量控制,通过对片段比对中测序覆盖度(不低于 30 × 测序深度)、片段重复性(不高于 5%)、片段比对率(不低于 95%)等指标进行统计和

阈值判定,进一步过滤低质量测序样本数据。

(4)使用 DNA 污染估计算法 Verifybamid[21](v2.0.1,默认参数)计算各样本中 DNA 污染程度,过滤高污染率(高于 3%)测序样本数据,形成最终用于下游分析的群体样本集合。

多重质量控制与过滤的基因组测序片段比对工作流在完成各样本测序片段比对任务的同时,将有效监控测序数据质量、比对数据质量、样本污染情况等多重指标,为高质量基因组结构变异检测奠定基础。

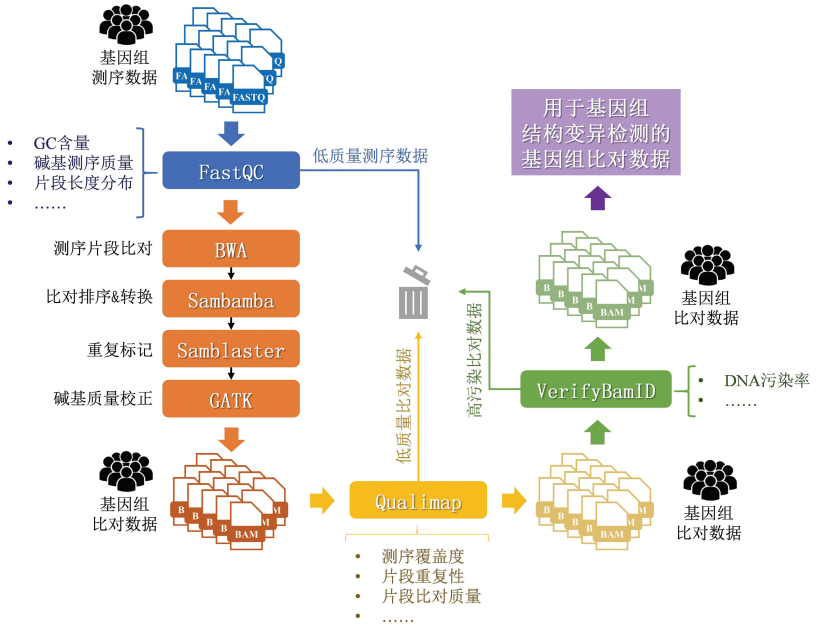


图 2 基因组测序片段比对流程

Fig. 2 Workflow of sequencing read alignment

2 单样本基因组结构变异检测

受限于高通量测序数据读长与系统性测序误差的限制,采用单一结构变异检测工具识别各样本基因组中的结构变异往往存在敏感性与准确性较低的问题,这将制约结构变异的检测能力和向下游研究转化的水平。针对这一问题,本研究采用三款当前性能最好的个体基因组结构变异检测算法,全面挖掘各样本基因组结构变异,主要步骤如下(见图 3)。

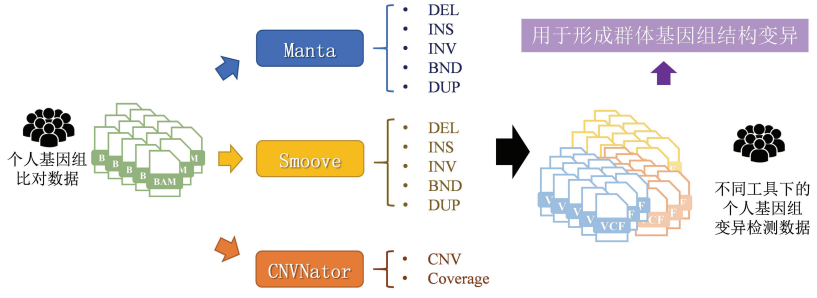


图 3 单样本基因组结构变异检测流程

Fig.3 Workflow of structural variation detection from individual sample

(1)使用快速检测基因组结构变异检测算法 Manta[22](v1.6.0,默认参数)识别各基因组中 DEL 变异、INS 变异、INV 变异、BND 变异、DUP 变异。

(2)使用简化集成基因组结构变异检测与基因分型算法 Smoove(https://github.com/brentp/smoove)(v0.2.7,默认参数)识别各基因组中 DEL 变异、INS 变异、INV 变异、BND 变异、DUP 变异。

(3)使用拷贝数变异检测算法 CNVNator[23]

(v0.4.1,默认参数)识别各基因组中的拷贝数变异(CNV),并计算各基因组区域上测序覆盖度信息。

分别使用 Manta、Smoove、CNVNator 三种结构变异检测算法,将有效挖掘每个样本基因组中多种类型结构变异,为融合形成群体基因组结构变异提供支撑。

3　单样本基因组结构变异融合

群体基因组结构变异主要由各单样本基因组结构变异融合产生。如何对来自不同样本、不同检测算法形成的结构变异进行融合,是当前产生高精度群体基因组结构变异的核心。为此,本研究分别对群体样本中由同种检测算法与不同检测算法预测的结构变异分层次整合,从而产生最终群体结构变异候选位点,主要步骤如下(见图 4)。

(1)分别对由 Manta、Smoove 检测算法产生的结构变异按照基因组坐标进行排序,完成在相同检测算法上不同样本结构变异的融合。若相邻两个结构变异存在交叠,则将两个变异合并为一个变异,直至所有变异均不存在交叠性。对于合并后的结构变异,分别记录来源样本标号,同时以累加方式累积各来源样本在此变异上的变异质量数、测序片段支持度等信息。

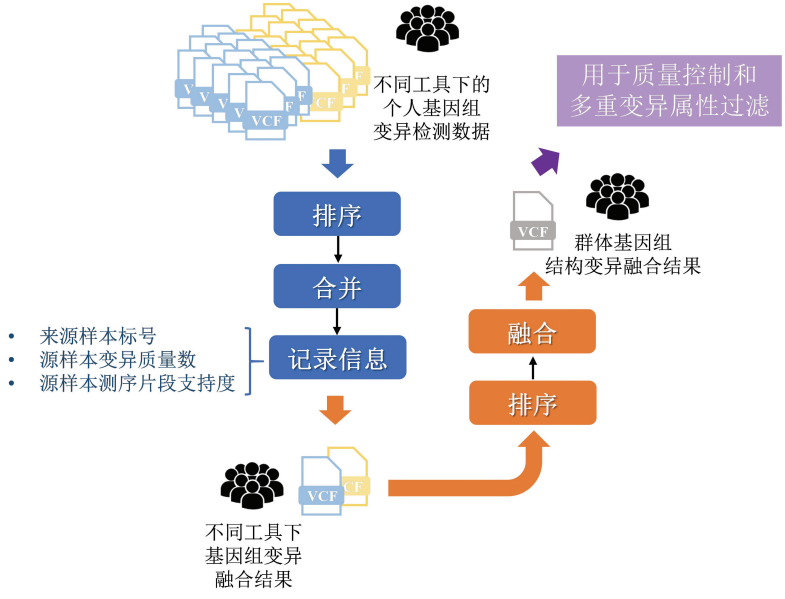


图 4　单样本基因组结构变异融合流程

Fig. 4　Workflow of integration of individual structural variation

(2)再次对由 Manta、Smoove 检测算法产生的基因组结构变异融合结果进行排序,完成对不同检测算法产生的融合结构变异数据的二次融合。

通过使用不同工具对各样本基因组中的结构变异双重融合,在充分保留各样本基因组中潜在的结构变异的同时,在融合过程中记录每个结构变异的样本支持情况、变异质量情况等信息,为过滤低质量群体结构变异提供了保障。

4　群体基因组结构变异检测

在完成单样本基因组结构变异融合后,进行质量控制和多重变异属性过滤,是完成高质量群体基因组结构变异检测,绘制高精度人群基因组结构变异图谱的核心。本研究实现这一目标主要采用如下 3 个步骤(见图 5)。

(1)将融合后的基因组结构变异按照变异类型分别拆分为 DEL、INS、INV、DUP、BND 五种类型。使用基因组结构变异断点基因型计算算法 SVTyper[24](v0.7.1,默认参数),分别从单样本层面对以上融合后的五种类型结构变异重新计算基因型。使用 CNVNator 计算的单样本层面的测序覆盖度信息对重新校准基因型信息的各结构变异进行注释。

(2)对重新校正基因型和测序覆盖度信息的各类型结构变异合并,计算各变异在人群中的变异频率。将合并后群体结构变异检测结果转换为 bedpe 文件格式并排序,对存在变异区域交叠的结构变异进行聚类,保留聚类中具有最大变异频率的结构变异,将聚类中其余结构变异修剪删除。

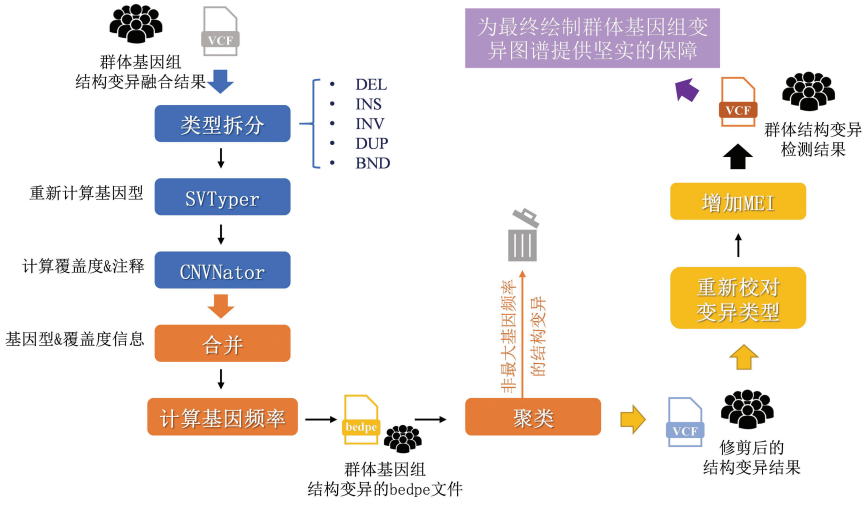


图 5 群体基因组结构变异检测流程

Fig. 5 Workflow of structural variation detection from population genomes

(3)将经过修剪的结构变异集合重新转换为 vcf 格式,依据测序覆盖度信息和基因型一致性信息对各结构变异重新校对变异类型,新增移动元件变异(MEI)类型,形成最终的群体结构变异检测结果。

经过对不同类型结构变异基因型的重新校正、过滤和变异类型校对,有效消减检测形成的假阳性结构变异预测结果,最大限度反映群体中真实的结构变异位点、类型和变异频率,为最终绘制群体基因组变异图谱提供了坚实的保障。

5 结果与分析

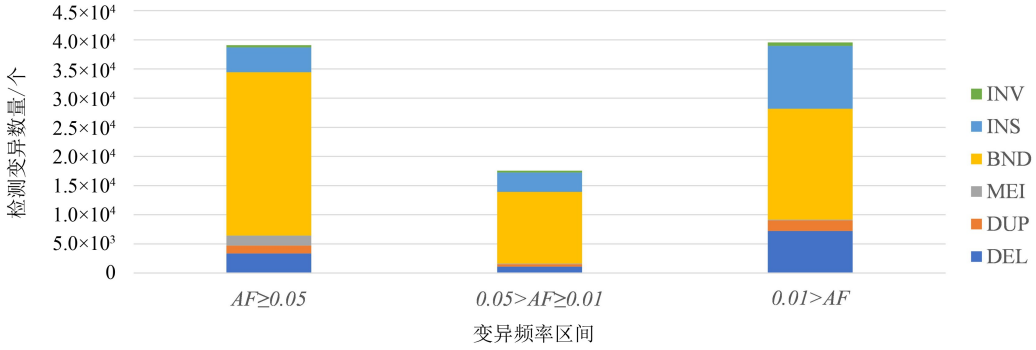
为了验证群体基因组结构变异检测工作流的真实效果,本研究构建了由 267 个样本组成的人群,使用 Illumina 高通量测序平台对该人群样本进行了 30×高深度全基因组测序,并使用本研究提出的群体基因组结构变异检测工作流对此 267 个样本进行群体结构变异检测(见表 1),合计检测出了 96 202 个结构变异,其中包括:11 697 个DEL 变异、18 385 个INS 变异、3 563 个DUP 变异、1 278 个 INV 变异、2 007 个MEI 变异、59 272 个BND 变异。

在该 267 个样本构成的人群中(见图 6),常见变异($AF \geqslant 0.05$)占总体检出变异的 41%(39 086/96 202),低频变异($0.05 > AF \geqslant 0.01$)占总体检出变异的 18%(17 554/96 202),罕见变异($0.01 > AF$)占总体检出变异的 41%(39 562/96 202)。值得关注的是,在 DEL、DUP、INS、INV 四种类型结构变异中,罕见变异的占比基本是均超过总体检验出变异的 50%,相比之下,在 MEI、BND 两种类型结构变异中,检测出的常见变异数量是总体可检测变异数量的主要占比。这些结果与过去开展的基因组计划发现的结果相一致[25-26],说明本研究建立的群体基因组结构变异检测工作流具有良好的检测能力。

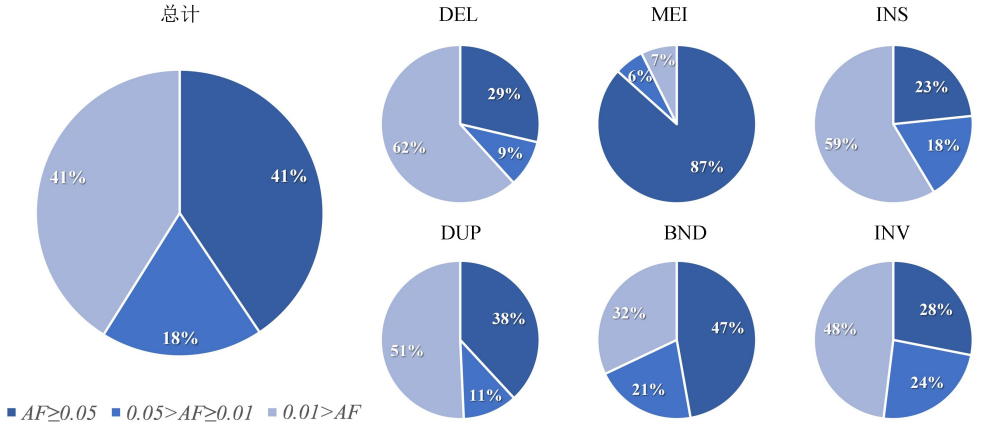
表 1 267 样本人群中结构变异检测结果统计

Table 1 Results of structural variation detection of 267 samples

结构变异类型	变异频率(AF)			合计
	$AF \geqslant 0.05$	$0.05 > AF \geqslant 0.01$	$0.01 > AF$	
DEL	3 355	1 121	7 221	11 697
DUP	1 359	397	1 807	3 563
MEI	1 739	121	147	2 007
BND	27 984	12 284	19 004	59 272
INS	4 290	3 326	10 769	18 385
INV	359	305	614	1 278
合计	39 086	17 554	39 562	96 202



(a)不同变异频率中各类型结构变异检测数量



(b) 不同变异类型中结构变异频率分布

图 6　不同变异频率中结构变异分布统计

Fig. 6　The distribution of structural variation among various allele frequencies

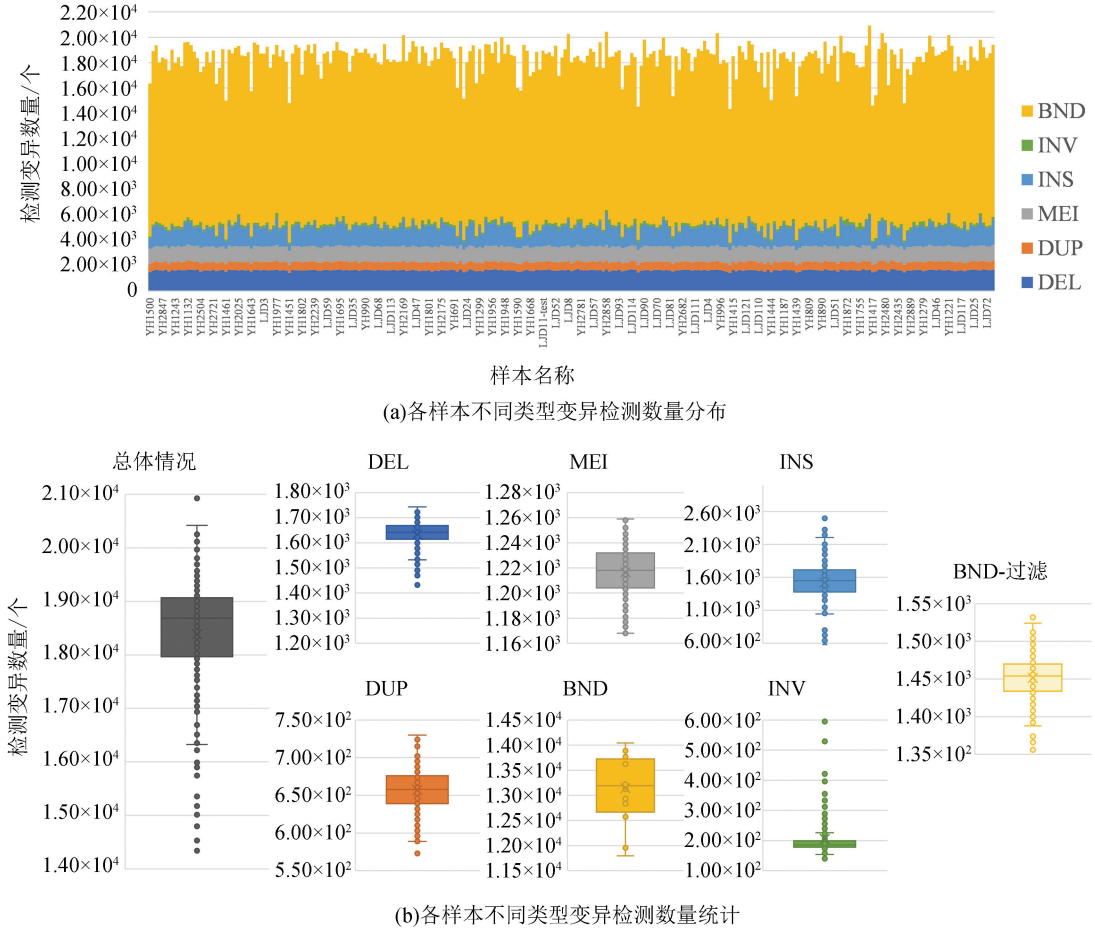
就每个样本可检测的结构变异而言,平均每个样本可以检测出 18 388 个结构变异,其中包含 1 634 个DEL 变异、657 个 DUP 变异、1 216 个MEI 变异、13 155 个BND 变异、1 521 个 INS 变异、206 个 INV 变异(见图 7)。受限于高通量测序技术中读长的限制,基因组重复片段区域中的结构变异难以检测和精确分型,其中仅可检测到断点连接关系的结构变异均归结为 BND 变异,因此导致了每个样本中包含了相当数量的 BND 变异。然而,仅获取变异断点连接关系,无法解析结构变异精准结构(如:是否为平衡变异,DNA 变化方向等)将严重影响 BND 变异的可信度和准确性。经过对 BND 变异按照置信度进行过滤(见图 7),总计移除 55 492 个BND 变异,仅保留3 780 个高置信度 BND 变异(移除率93.62%)。平均每个样本移除11 703 个 BND 变异,仅保留1 451 个高置信度 BND 变异(移除率88.97%)。

此外,本研究还对群体基因组结构变异检测工作流中变异融合与群体变异检测两个关键环节的计算开销和内存使用进行了统计(见表 2)。该工作流完成群体基因组结构变异检测和融合兼容串行分析和并行分析两种方式,其中串行计算方式需要约 173.4 h,最大内存开销30 GB,而采用并行计算方式仅需不足3 h,并维持最大30 GB的内存开销。这一结果表明,对于大规模人群基因组结构变异检测分析,在保持有限内存消耗的前提下,采用并行方式运行该工作流将显著提升计算速度,为高效、快速的群体基因组结构变异检测提供了保证。

表 2　267 样本人群中结构变异检测运行时间及内存统计

Table 2　Time and memory cost of structural variation detection of 267 samples

检测环节	运行时间/h		内存开销/ GB
	串行方式	并行方式	
单样本基因组结构变异融合	0.8	0.8	10
群体基因组结构变异检测	172.6	2.1	30



(a)各样本不同类型变异检测数量分布

(b)各样本不同类型变异检测数量统计

图 7 各样本不同类型结构变异检测数量分布和统计

Fig.7 The quantitative distribution of various structural variation types among samples

6 结 论

1)本研究构建了一套高效、精准的群体基因组结构变异检测工作流,该工作流通过多层过滤的质量控制,为高质量群体基因组结构变异检测提供支撑。

2)该工作流通过使用多种高性能结构变异检测算法,提高了结构变异检测的准确性与敏感性,并通过双重融合实现了群体结构变异候选位点的精准定位。

3)该工作流通过多维度重新校正结构变异候选位点的基因型与变异类别,进一步保障群体结构变异图谱的高质量构建。

4)利用该工作流对由 267 个样本组成的人群进行基因组结构变异检测,结果表明该工作流具有良好、快速、高效的检测能力。通过并行分析策略在控制内存消耗的基础上,提高了工作流的计算速度,为大规模群体基因组研究提供了可能。

参考文献(References)

[1]AUTON A, ABECASIS G R, ALTSHULER D M, et al. A global reference for human genetic variation[J]. Nature, 2015, 526(7571): 68-74. DOI:10.1038/nature15393.

[2]SEDLAZECK F J, RESCHENEDER P, SMOLKA M, et al. Accurate detection of complex structural variations using single-molecule sequencing [J]. Nature Methods, 2018. DOI:10.1038/s41592-018-0001-7.

[3]ALKAN C, COE B P, EICHLER E E. Genome structural variation discovery and genotyping [J]. Nature Reviews Genetics, 2011, 12(5): 363-376. DOI:10.1038/nrg2958.

[4]ZICHNER T, GARFIELD D, RAUSCH T, et al. Impact of genomic structural variation in Drosophila melanogaster based on population-scale sequencing [J]. Genome Research, 2013, 23(3): 568-79. DOI: 10.1101/gr.142646.112.

[5]WEISCHENFELDT J, SYMMONS O, SPITZ F O, et al. Phenotypic impact of genomic structural variation: insights from and for human disease[J]. Nature Reviews Genetics, 2013, 14(2): 125-38. DOI: 10.1038/nrg3373.

[6]MAHMOUD M, GOBET N, CRUZ-DÁVALOS D I, et al.

Structural variant calling: the long and the short of it[J]. Genome Biology, 2019, 20(1): 246. DOI: 10.1186/s13059-019-1828-7.

[7]HEDGES D J, HAMILTON-NELSON K L, SACHAROW S J, et al. Evidence of novel fine-scale structural variation at autism spectrum disorder candidate loci [J]. Molecular Autism,2012,3(1): 2. DOI:10.1186/2040-2392-3-2.

[8]ROVELET-LECRUX A, HANNEQUIN D, RAUX G, et al. APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy[J]. Nature Genetics,2006, 38(1): 24-26. DOI:10.1038/ng1718.

[9]AUDANO P A, SULOVARI A, GRAVES-LINDSAY T A, et al. Characterizing the major structural variant alleles of the human genome[J]. Cell, 2019, 176(3): 663-675. e19. DOI: 10.1016/j.cell.2018.12.019.

[10]HO S S, URBAN A E, MILLS R E. Structural variation in the sequencing era[J]. Nature Review Genetics,2020,21(3): 171-89.DOI:10.1038/s41576-019-0180-9.

[11]CONSORTIUM G P. A map of human genome variation from population-scale sequencing [J]. Nature,2010, 467(7319): 1061. DOI: 10.1038/nature09534.

[12]CONSORTIUM G P. An integrated map of genetic variation from 1,092 human genomes [J]. Nature, 2012, 491(7422): 56. DOI: 10.1038/nature11632.

[13]CONSORTIUM G P. A global reference for human genetic variation[J]. Nature, 2015, 526(7571): 68. DOI: 10.1038/nature15393.

[14]THE UK10K CONSORTIUM. The UK10K project identifies rare variants in health and disease[J]. Nature, 2015, 526(7571):82 - 90. DOI:10.1038/nature14962.

[15]LEK M, KARCZEWSKI K J, MINIKEL E V, et al. Analysis of protein-coding genetic variation in 60,706 humans[J]. Nature, 2016, 536(7616): 285-291. DOI: 10.1038/nature19057.

[16]Large-scale whole-genome sequencing of the Icelandic population[J]. Nature Genetics, 2015, 47(5): 435-444. DOI: 10.1038/ng.3247.

[17]MCCARTHY S, DAS S, KRETZSCHMAR W, et al. A reference panel of 64,976 haplotypes for genotype imputation [J]. Nature Genetics, 2016, 48(10): 1279-1283. DOI:10.1038/ng.3643.

[18]TARASOV A, VILELLA A J, CUPPEN E, et al. Sambamba: fast processing of NGS alignment formats[J]. Bioinformatics, 2015, 31(12): 2032-2034. DOI:10.1093/bioinformatics/btv098.

[19]FAUST G G, HALL I M. SAMBLASTER: Fast duplicate marking and structural variant read extraction[J]. Bioinformatics, 2014, 30(17): 2503-2505. DOI:10.1093/bioinformatics/btu314.

[20]GARCÍA-ALCALDE F, OKONECHNIKOV K, CARBONELL J, et al. Qualimap: evaluating next-generation sequencing alignment data[J]. Bioinformatics, 2012, 28(20): 2678-2679. DOI:10.1093/bioinformatics/bts503.

[21]ZHANG F, FLICKINGER M, TALIUN S A G, et al. Ancestry-agnostic estimation of DNA sample contamination from sequence reads [J]. Genome Research, 2020, 30(2):185-194. DOI: 10.1101/gr.246934.118.

[22]CHEN X, SCHULZ-TRIEGLAFF O, SHAW R, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications[J]. Bioinformatics, 2016, 32(8): 1220-1222. DOI: 10.1093/bioinformatics/btv710.

[23]ABYZOV A, URBAN A E, SNYDER M, et al. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing[J]. Genome Research, 2011, 21(6): 974-984. DOI: 10.1101/gr.114876.110.

[24]CHIANG C, LAYER R M, FAUST G G, et al. SpeedSeq: ultra-fast personal genome analysis and interpretation[J]. Nature Methods, 2015, 12(10): 966-968. DOI: 10.1038/nmeth.3505.

[25]ABEL H J, LARSON D E, REGIER A A, et al. Mapping and characterization of structural variation in 17,795 human genomes[J]. Nature, 2020, 583(7814): 83-89. DOI: 10.1038/s41586-020-2371-0.

[26]COLLINS R L, BRAND H, KARCZEWSKI K J, et al. A structural variation reference for medical and population genetics[J]. Nature, 2020, 581(7809): 444-451. DOI: 10.1038/s41586-020-2287-8.