doi:10.3969/j.issn.1672-5565.03.10

一站式全基因组和外显子组测序数据自动分析软件(SeqMule)

李 鑫[1],李 凯[2],李一佳[1*],马 磊[2]
(1.云南舜喜再生医学工程有限公司,昆明 650000;
2.昆明理工大学信息工程与自动化学院,昆明 650500)

摘 要:SeqMule 可根据调用的人类基因组和外显子组数据自动调节变量,对所有测序数据的单核苷酸多态性(Single nucleotide polymorphism,SNP)进行分析和注释。目的:通过对两名痛风患者的实验数据进行分析,详细地为生物信息学研究人员介绍了 SeqMule 软件,以期为全基因组和外显子组测序数据提供一站式的分析途径。方法:基于 SeqMule 内置的 BWA(Burrows-Wheeler Aligner)、GATK(The Genome Analysis Toolkit)、SAMtools、Freebayes 比对和分析工具,以两名痛风患者的 DNA 测序数据分析为例,本文详细地论述了 SeqMule 的特点及操作,并对两名患者的外显子测序数据进行了自动化比对与 SNP 分析。发现 SeqMule 优化了很多分析软件存在的一些问题,可以对外显子组和全基因组测序数据实现全面、灵活、高效地自动化分析,能更好地分析高通量测序数据,最终提升数据分析的一致性和准确性。

关键词:基因;测序;SeqMule;外显子;SNP
中图分类号:Q343.1 **文献标志码**:A **文章编号**:1672-5565(2016)03-188-07

A one-stop analytic software for sequencing data of whole genome and exome: SeqMule

LI Xin [1], LI Kai [2], LI Yijia [1*], MA Lei [2]
(1.*Stem Cell And Regenerative Medicine Research Center*, *Yunnan Suns Regenerative Medicine Engineering Co. Kunming* 650000, *China*;
2.*School of Information Engineering and Automation*, *Kunming University of Science and Technology*, *Kunming* 650500, *China*)

Abstract: SeqMule can adjust variables automatically according to the data of the invoked human genomes and the exomes, and also can analyze and annotate SNPs (Single Nucleotide Polymorphism). Objectives: This paper introduces SeqMule software to researchers on bioinformatics in detail by analyzing the experimental data of two patients with gout, with the hope of providing a one-stop analytical approach for the whole genomes and exomes. Methods: This paper discusses the features and operations of the SeqMule taking the analysis of DNA data of two patients with gout using the BLAST and analysis softwares such as BWA, GATK, SAMtools, Freebayes embedded in SeqMule, and also we have carried out BLASTs for the their exomes automatically and analyzed SNPs for them. Conclusions: SeqMule has resolved some questions present in many softwares. It also can analyze the data from the whole genomes and the exomes automatically in a comprehensive, flexible and efficient way, better analyze the data from high throughput sequencing, and finally improve the consistency and accuracy of the data analysis.

Keywords: Gene; Sequencing; SeqMule; Exome; SNP

随着人类基因组计划的胜利完成和后基因组时代的来临[1],DNA 测序技术已成为人类探索生命秘密的重要手段之一,对生物、生命科学、医学等领域的技术发展起到了巨大的推动作用[2]。经过三十多年的努力,DNA 测序技术已经取得巨大的进展,在第一代和第二代测序技术的基础上,以单分子测

收稿日期:2016-04-05;**修回日期**:2016-06-12.
作者简介:李鑫,男,本科生,研究方向:二代测序技术;E-mail: 281528209@qq.com;
李凯,男,硕士研究生,研究方向:生物信息学;E-mail: 553234748@qq.com.
*通信作者:李一佳,男,博士,研究方向:干细胞和基因临床转化;E-mail: yijia.tsinghua@gmail.com.

序为特点的第三代测序技术已经诞生。第三代测序技术虽然解决了第二代测序技术读长短、速度慢等缺点,但由于其成本和错误率偏高、通量低,目前最常用的依然是以 Illumina 公司的 Solexa 技术[3]为标志的第二代测序技术。

第二代测序技术拥有相当高的测序通量,覆盖度高。得到的 reads 不仅长度短,数量又极为巨大,这给序列拼接带来了巨大的挑战,而基因组测序中的一个关键的步骤就是序列拼接[4]。拼接后,还需要对所有的 SNP 进行分析和注释。

针对 SNP 的分析,目前有一些基于云端的高通量测序数据分析平台,比如 Galaxy[5]。Galaxy 等现行的生物信息学平台,使大量的生物信息学工具易于操作,用户上传数据后可立即开始分析。但是,当用户拥有超大数据量时,存储限制了数据的传输速度,较长的工作排队时间使其变得不切实际。除了平台解决方案,还有其他独立途径可进行 SNP 的多样分析。例如 SeqMule[6]、HugeSeq[7]、Ngs _ backbone[8]和 Bcbio-nextgen[9]四款集成分析软件,可以运用自带的工具对 SNP 进行比对、注释、分析等。但是,由于部分软件集成某些专用工具,比如 Bcbio-nextgen 软件专有的比对工具 NovoAlign[10],不是对所有研究人员免费开放。四款集成分析软件相比,SeqMule 软件结合了 5 种 SNP 比对工具和 5 种 SNP 分析工具,其余三款分析软件只有一种或两种 SNP 比对工具和 SNP 分析工具。除此之外,只有 SeqMule 软件拥有可选且开源的 SNP 比对工具,具有更高的灵活性和可用性。

SeqMule 软件是以人类遗传病研究为背景,专门针对外显子组或全基因组序列分析设计的。它采用高度灵活的各种调用格式对 SNP 进行完全自动化的分析和注释,支持 Sun Grid Engine 并行处理,可以进行测序质量的检测、孟德尔错误率检测、一致性评估,生成最终的 HTML 报告。相比之下,SeqMule 是上述解决方案中较好的一款软件,推荐生物信息学人员使用。

1　SeqMule 软件

1.1　基本介绍

对测序数据进行分析的时候,除了测序平台的差异[11],仍要考虑算法间的差异。例如,5 种生物信息学算法(SOAP、BWA-GATK、BWA-SNVer、GNUMAP、BWA-SAMtools)分析 SNV(Single Nucleotide Variants)的一致性只有 57.4%,而每种计算途径间的变异数为 0.5%~5.1%[12]。在不同的测序错误率和 indel 标记下,校准也存在差异[13]。目前,公开发表的计算方法几乎没有提供两种或更多的比对和 SNP 分析方法。

分析软件的安装和配置是首要问题,而且这个问题的重要性已经被许多试图去使用它的人所证实,像 Bioconductor、Bioperl 和 Web-based 三款软件[14-16]。理论上,来自一个程序的输出结果很难被输入另一个和它类似的程序中。例如,GATK 不能接受来自 SOAP2 的输出。此外,软件的不同步更新,可能导致软件的不兼容。虚拟机和虚拟化技术为用户解决了该问题[17-19],然而,虚拟机系统不可避免地限制了客户系统可用的计算资源,减少了软件工具的灵活性。因此,对于没有计算机背景的普通用户来说,部署软件成为了一个很大的难题。针对普通用户,迫切需要一种易于执行和整合多种工具的分析途径。

在不影响易用性、高效性和重复性的前提下,由南加州大学的王凯实验室开发了一个全能的解决方案——SeqMule,能够执行一系列自动化的命令来分析高通量测序数据。它结合了 5 种比对工具:BWA(包括 BWA-backtrack 和 BWA-MEM)、Bowtie、Bowtie2、SOAP2、SNAP[20-24],5 种不同 SNP 分析工具:GATK(包括 GATKLite 和 version 3)、SAMtools、VarScan 2、Freebayes、SOAPsnp[25-28]和一些配件程序:FastQC、Picard、tabix、VCFtools 30,而且可以通过修饰配置文件来获得多种组合。通过不同工具结合而设置变量形成交叉,从而获得更高的准确性、敏感性和特异性。SeqMule 能提供建立在不同调用者之上的并行功能,还能够更好地分析高通量测序数据,提升分析的一致性和准确性。针对目前主流服务器(CPU:2 Intel Xeon X5650, 内存 48GB),只需 24 小时,SeqMule 可从设置好的全基因组数据生成带注释的 VCF 文件。

SeqMule 的工作流程如图 1 所示,分析过程中有很多可利用的工具。其中,先使用 FastQC 进行质量控制,再采用 BWA-backtrack、BWA-MEM、Bowtie 等工具进行初始校准,校准后可使用 Picard Tools 对质量控制进行评估,再使用 GATK、SAMtools、SOAPsnp、VarScan 工具进行突变调用和过滤,最后采用 GATK CombineVariants 交叉或合并。

1.2　SeqMule 安装方法

SeqMule 可在如下网址下载:http://seqmule.openbioinformatics.org.

1)笔者使用的是 CentOS 7 系统,安装 SeqMule 之前要先安装必要的软件和环境,相关命令如下:

sudo yum install-y gcc gcc-c++ make cmake

ncurses-devel ncurses R unzip automake autoconf git-core gzip tar

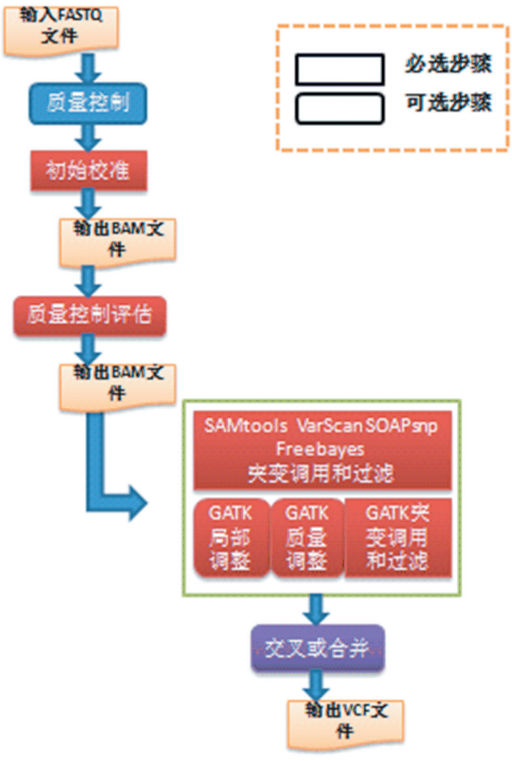

图 1 SeqMule 的工作流程图

Fig. 1 Scheme of SeqMule workflow

2)下载 SeqMule 程序 git clone https://github.com/WGLab/SeqMule.git,如果 Https 不支持也可以用 git 模式 git clone git://github.com/WGLab/SeqMule.git。

3)进入 SeqMule 文件夹,利用./Build freshinstall 进行初始安装。

4)安装一次后,利用./Build installexes 安装 missing 的部分。

5)由于 GATK 要单独安装,利用./Build gatk 看安装命令,核心就是把 GATK 的 jar 文件拷贝到制定文件夹。

6)把环境变量写到用户的 bashrc 里面,然后用 source 命令更新以下环境变量,这样 Seqmule 就可以不制定他的绝对位置来使用了,否则会出现 command 找不到的情况。

echo 'export PATH= $ PATH:absolute_path_to _seqmule/bin' >> ~/.bashrc source ~/.bashrc

7)下载 Seqmule 要使用到的 Database,seqmule download -down hg19all。

此部分利用 terminal 下载很缓慢,建议使用专用下载工具下载,大概有 40 G 左右。将下载的文件放到 SeqMule 的 database 文件里,再利用 SeqMule download-down hg19all 命令进行解压缩,然后把文件再按名称放到指定文件夹。另外,笔者已将下载好的 database 文件夹都放置到百度云,读者可以通过 shunxirm@163.com 获取下载秘钥。

1.3 SeqMule 软件的运行

SeqMule 软件运行在 Linux 系统平台下,命令简单且易于掌握。根据测序方法不同,SeqMule 的分析方式也不同。SeqMule 主要包括三种分析方式,分别是:典型的外显子组分析、快速转换的全基因组分析和基于家系的三人外显子组分析。此外,SeqMule 软件可以一次性分析多个样本,大大简化了生物信息学研究人员的操作。

1.3.1 SeqMule 软件的使用命令

在存放需要分析的 FASTQ 格式文件的文件夹下,打开系统终端,输入以下命令运行 SeqMule:

seqmule pipeline-a normal_R1.fastq.gz-b normal_R2.fastq.gz-prefix example-N 2-capture default-threads 4-e

其中,normal_R1.fastq.gz 和 normal_R2.fastq.gz 是 DNA 经过测序仪测序后产生的 FASTQ 格式的压缩文件,分别是一条 DNA 上两条链的基因数据。参数"-prefix example"是告诉 SeqMule 软件你的样本名称是 example;"-capture default"是让 SeqMule 软件使用默认的区域定义文件——hg19 外显子区,对应文件可从安捷伦 SureSelect 工具包中下载;"-threads 4"是令 SeqMule 软件在运行时使用该计算机的四个线程;"-e"的意思是这个数据集是外显子组数据或者捕获的测序数据,而不是全基因组数据。

1.3.2 典型的外显子组分析命令

通过测序仪对外显子组测序后,得到四个 FASTQ 文件的压缩文档,在终端下运行以下命令进行典型外显子组分析:

seqmule pipeline-a sample _ lane1 _ R1. fq. gz, sample_lane2 _ R1. fq. gz-b sample _ lane1 _ R2. fq. gz, sample _ lane2 _ R2. fq. gz-capture seqmule/database/hg19-

nimblegen/nexterarapidcapture _ exome _ targetedregions _ v1. 2. bed-m-e-advanced seqmule/misc/predefined_con-

fig/bwa _ gatk _ HaplotypeCaller. config-quick-t 4-prefix mySample

命令中"-advanced seqmule/misc/predefined _ config

/bwa _ gatk _ HaplotypeCaller. config"表示使用 BWA 和 GATK 这两个可选工具包进行 SNP 的比对和分析。参数中-quick 是使软件使用更多的计算

机内存来进行快速分析;“-t 4”表示分析时使用计算机的四个 CPU;“-m”是合并两个数据集。

1.3.3　快速转换的全基因组分析命令

通过测序仪对全基因组测序后,得到两个 FASTQ 文件的压缩文档,在终端下运行以下命令进行快速转换的全基因组分析:

seqmule pipeline-a sample_R1.fq.gz-b sample_R2.fq.gz-advanced seqmule/misc/predefined_c-onfig/snap_freebayes.config-quick-t 12-g-prefix mySample

命令中“-advanced seqmule/misc/predefined_config/snap_freebayes.config”表示使用 SNAP 和 FreeBayes 这两个可选工具包进行全基因组 SNP 的比对和分析。参数“-g”表示全基因组分析;“-t 12”是令 SeqMule 使用计算机的 12 个 CPU 进行比对分析,因为 SNAP 工具使用时非常消耗内存,因此采用多个 CPU 来提高软件运行速度。

1.3.4　三人外显子组分析命令

对同一个家庭的三个人进行外显子组测序后,使用 SeqMule 软件对三人的测序数据进行分析来发现致病基因,命令如下:

seqmule pipeline-a fa_R1.fq.gz,mo_R1.fq.gz,son_R1.fq.gz-b fa_R2.fq.gz,mo_R2.fq.gz,son_R2.fq.gz-ms-e-q-t 4-prefix father, mother, son-capture default-sge "qsub-V-cwd-pe smp XCPUX"

命令中“-sge "qsub-V-cwd-pe smp XCPUX"”表示使用 SG 工具包来进行分析。参数中“-ms”表示针对多样本的基因突变识别,可以更加准确地分析来自同一家庭的三个人的外显子组数据。

2　使用 SeqMule 软件进行 SNP 分析

2.1　数据准备

患者数据来自舜喜再生医学工程有限公司。昆明医科大学第一附属医院的两名痛风患者在云南舜喜再生医学工程有限公司抽血并提取 DNA 后,使用 Illumina 公司的 Hiseq3000 测序仪进行外显子组测序。测序后得到 FASTQ 格式文件的压缩文件,作为实验前准备数据。使用 SeqMule 软件进行基因的比对、拼接并进行 SNP 分析。

2.2　分析报告

SeqMule 分析完成后,生成 HTML 格式的详细分析报告(SeqMule Report)。分析报告网页上有分析总结、样本分析报告、分析途径、分析参数和帮助文件按钮,点开后即可查看详细信息。

样本分析结果展示了统计资料、SNV 与 NON-SNV 韦恩图和覆盖度图。统计资料里包含基因校准数据表、基因覆盖率统计数据表和基因突变数据表。其中,表 1 是 SeqMule 软件对该患者的基因数据进行初始校准得到的校准统计表,包含通过的质量控制读长数、失败的读长数、比对的读长数及和数据库匹配上的读长数等数据。表 2 是该患者的基因覆盖率度统计数据表,包括总的长度、目标区域的平均覆盖度等数据。表 3 是该患者的基因突变数据表,包含该患者的所有突变位点数、SNV 突变位点数和插入/缺失位点数等数据。

表 1　患者 1 的校准统计表

Table 1　Alignment stats of the NO.1 patient

校准项	校准统计值/条
QC-passed reads	138 070 884
QC-failed-reads	0
Duplicates	0
Mapped reads	115 341 601(83.54%)
Paired reads	138 070 884
Read1	69 040 101
Read2	69 030 783
Properly paired	110 030 592(79.69%)
Reads with itself and mate mapped	115 304 036
Singletons	37 565(0.03%)
Reads with mate mapped to a different chromosome	15 315 418
Reads with mate mapped to a different chr (mapQ at least 5)	15 315 418

表 2 患者 1 的覆盖率统计数据表

Table 2 Coverage stats of the NO.1 patient

覆盖项	覆盖统计(%)
total length (defined by nexterarapidcapture_expandedexome_targetedregions.chrmod.nooverlap.bed)	62.09 Mb
Fraction of reads mapped to target region	33.43
Average coverage in target region	89.19
Percentage above 30	86.89
Percentage above 20	91.25
Percentage above 10	94.46
Percentage above 5	96.05

表 3 患者 1 的突变数据表

Table 3 Variant stats of the NO.1 patient

样本	外显子数据/条
Number of variants	28 757
Number of SNVs	23 236
Number of indels	5 521
Transitions	16 404
Transversions	6 843
Ti/Tv Ratio	2.40
Total heterozygotes	15 126
Ref/Alt heterozygotes	14 880
Alt/Alt heterozygotes	246
Homozygotes	13 631

SNV 与 NON-SNV 韦恩图是 SeqMule 结合 3 种不同的 SNP 分析工具得出的 SNV 和 NON-SNV 突变重叠图。图 2 是两名患者的 SNV 与 NON-SNV 韦恩图。从图中可以得出,基于 3 种分析工具单独分析出的基因突变结果、两两之间分析出的相同突变基因的结果以及三种分析工具分析出的相同突变基因的个数。患者 1 的数据中,GATK、SAMtools 和 freebayes 三种分析工具的分析结果中都出现 SNV 突变的位点有22 011个,NON-SNV 突变的位点有2 291个。患者 2 的数据中,三种分析工具的分析结果中都出现 SNV 突变的位点有 29 111 个,NON-SNV 突变的位点有2 269个。

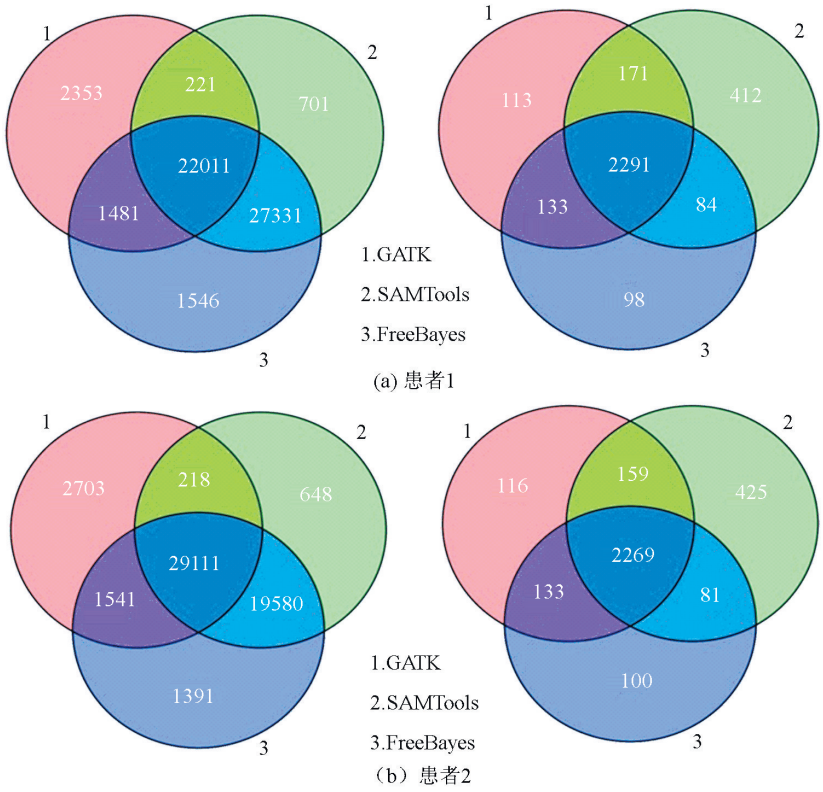

图 2 两名患者的 SNV 与 NON-SNV 韦恩图

Fig. 2 Venn Diagram (SNV and NON-SNV) of two patients

3 结　语

从实验结果可以看出,SeqMule 可对外显子组测序数据实现全面、简易、灵活、高效的一站式自动化分析。分析结果采用 HTML 报告的方式,展示出详细、美观的图表,简单易读。除了外显子组测序,SeqMule 还支持对全基因组测序进行一站式自动分析,更加多元化。SeqMule 解决了大部分分析软件存在的软件兼容性、配置复杂及不能访问高性能计算设施等问题,能更好地分析高通量测序数据,提升基因数据分析的一致性和准确性。该软件的 5 种比对方式、5 种 SNP 分析工具和多种多样的配件程序给用户提供了众多选择,内置的并行处理能力可加快分析的进程。除了上述特点外,SeqMule 使用单行命令完成复杂的任务,使其成为易于下载、安装、配置和运行的生物信息学的工具。

笔者已经用 SeqMule 来分析测序数据,并且获得了有意义的结果。随着新一代测序技术的快速发展和部署,我们期望 SeqMule 能够促进即将来临的大量测序数据分析,从而为人类遗传病研究奠定基础,并促进人类遗传病的诊断方法的完善。

参考文献(References)

[1]唐旭清, 朱平. 后基因组时代生物信息学的发展趋势[J]. 生物信息学, 2008, 6(3): 142-144.
TANG Xuqing, ZHU Ping. The development trends of bioinformatics in post-genomic era [J]. China Journal of Bioinformatics, 2008, 6(3): 142-144.

[2]陈文辉, 罗军, 赵超. 固态纳米孔:下一代 DNA 测序技术——原理、工艺与挑战[J]. 中国科学:生命科学, 2014(7): 649-662.
CHEN Wenjun, LUO Jun, ZHAO Chao. Solid nano pore: Next generation DNA sequencing technology-principle, techonology and challenge[J]. Science in China: Life Sciences, 2014(7): 649-662.

[3]CAPORASO J G, LAUBER C L, WALTERS W A, et al. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms[J]. Isme Journal, 2012, 6(8): 1621-1624.

[4]逯雯雯, 卢志远, 王亚旭,等. 面向新一代基因组测序技术的序列拼接算法[J]. 生物信息学, 2010, 8(3): 248-253.
LU Wenwen, LU Zhiyuan, WANG Yaxu, et al. Facing the sequence stitching algorithm of new generation genome sequencing technology[J]. China Journal of Bioinformatics, 2010, 8(3): 248-253.

[5]AFGAN E, BAKER D, CORAOR N, et al. Galaxy CloudMan: delivering cloud compute clusters[J]. BMC Bioinformatics, 2010, 12(6): S4-S4.

[6]GUO Y, DING X, SHEN Y, et al. SeqMule: automated pipeline for analysis of human exome/genome sequencing data[J]. Scientific Reports, 2015, 5: 14283. DOI: 10.1038/srep14283.

[7]LAM H Y, PAN C, CLARK M J, et al. Detecting and annotating genetic variations using the HugeSeq pipeline[J]. Nature Biotechnology, 2012, 30(3): 226-229.

[8]BLANCA J M, PASCUAL L, ZIARSOLO P, et al. ngs_backbone: a pipeline for read cleaning, mapping and SNP calling using next generation sequence[J]. BMC Genomics, 2011, 12(1): 193-201.

[9]GUIMERA R V. Bcbio-nextgen: Automated, distributed next-gen sequencing pipeline[J]. Embnet Journal, 2012, 18(Supplement B):1-153.DOI: 10.14806/ej.17.B.286.

[10]RAMOS E, LEVINSON B T, CHASNOFF S, et al. Population-based rare variant detection via pooled exome or custom hybridization capture with or without individual indexing[J]. BMC Genomics, 2012, 13(1): 1-15.

[11]LAM H Y K, CLARK M J, CHEN R, et al. Performance comparison of whole-genome sequencing platforms[J]. Nature Biotechnology, 2012, 30(1): 78-82.

[12]O'RAWE J, JIANG T, SUN G, et al. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing[J]. Genome Medicine, 2013, 5(13): 1735-1742.

[13]RUFFALO M, LAFRAMBOISE T, KOYUTURK M. Comparative analysis of algorithms for next-generation sequencing read alignment[J]. Bioinformatics, 2011, 27(27): 2790-2796.

[14]GENTLEMAN R C , CAREY V J , BATES D M, et al. Bioconductor: open software development for computational biology and bioinformatics[M]. Genome Biology, 2004, 5: R80. DOI: 10.1186/gb-2004-5-10-r80.

[15]STAJICH J E, BLOCK D, BOULEZ K, et al. The Bioperl toolkit: Perl modules for the life sciences[J]. Genome Research, 2002, 12(10): 1611-1618.

[16]CHANG X, WANG K. wANNOVAR: annotating genetic variants for personal genomes via the web[J]. Journal of Medical Genetics, 2012, 49(7): 433-436.

[17]KRAMPIS K, BOOTH T, CHAPMAN B, et al. Cloud BioLinux: pre-configured and on-demand bioinformatics computing for the genomics community[J]. BMC Bioinformatics, 2012, 13(1): 1-8.

[18]NOCQ J, CELTON M, GENDRON P, et al. Harnessing virtual machines to simplify next-generation DNA sequencing analysis[J]. Bioinformatics, 2013, 29(17): 2075-2083.

[19]ANGIUOLI S V, MATALKA M, GUSSMAN A, et al.

CLOVR: a virtual machine for automated and portable sequence analysis from the desktop using cloud computing [J]. BMC Bioinformatics, 2011, 12(49): 356-356.

[20] LI H, DURBIN R. Fast and accurate short read alignment with Burrows-Wheeler transform [J]. Bioinformatics, 2009, 25(14): 1754-1760.

[21] LANGMEAD B, TRAPNELL C, POP M, et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome[J]. Genome Biology, 2009, 10(3): 1-10.

[22] LANGMEAD B, SALZBERG S L. Fast gapped-read alignment with Bowtie 2[J]. Nature Methods, 2012, 9(4): 357-359.

[23] LI R, YU C, LI Y, et al. SOAP2: an improved ultrafast tool for short read alignment[J]. Bioinformatics, 2009, 25(15): 1966-1967.

[24] ZAHARIA M, BOLOSKY W J, CURTIS K, et al. Faster and more accurate sequence alignment with SNAP [J]. ARXIV, 2011(1):1-10.

[25] MCKENNA A, HANNA M, BANRS E, et al. The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data[J]. Genome Research, 2014, 20(9): 1297-1303.

[26] LI H, HANDSAKER B, WYSOKER A, et al. The sequence alignment-map format and SAMtools[J]. Bioinformatics, 2009, 25(16): 2078-2079.

[27] KOBOLDT D C, ZHANG Q, LARSON D E, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing[J]. Genome Research, 2012, 22(3): 568-576.

[28] LI R, LI Y, FANG X, et al. SNP detection for massively parallel whole-genome resequencing [J]. Genome Research, 2009, 19(6): 545-552.