利用人类全基因组二代测序数据比较 BWA-MEM 和 NovoAlign *

王文雅　庞尔丽†
(生物多样性与生态工程教育部重点实验室,北京师范大学生命科学学院,100875,北京)

摘要　随着测序技术的发展,二代测序数据越来越多,将测序数据准确地比对到参考基因组是后续研究的基础. BWA-MEM 和 NovoAlign 作为 2 个最常用的 DNA 序列比对软件,还没有评估其在基因组中不同结构区域的表现. 本研究基于真实和模拟数据,对 2 个软件在人类基因组的低复杂度、片段性重复和其他区域进行了评估. 结果显示: BWA-MEM 将尽可能多的测序数据比对到基因组,且在低复杂度和片段性重复区域存在过度比对的现象,特别是在片段性重复区域的比对品质较低; 而 NovoAlign 比对到基因组的序列数量低于 BWA-MEM,但在片段性重复区域的比对品质较优,因此对于存在较多片段性重复区域的基因组来说,使用 NovoAlign 比对是一种更好的策略.

关键词　二代测序; 比对; BWA-MEM; NovoAlign; 全基因组
中图分类号　Q987　　**DOI**: 10.12202/j.0476-0301.2020338

0　引言

随着二代测序技术的快速发展,测序成本不断降低,一些中小型实验室不仅可以承担起全外显子组测序的费用,甚至可以对一定数量个体组成的多个群体进行全基因组测序. 相比于全外显子组测序,全基因组范围内的测序数据不仅能够发现位于蛋白质编码区域的兴趣位点,还可以分析占据基因组大部分的"垃圾 DNA"(junk DNA). 在肿瘤免疫学中,早期的研究认为,新生抗原(neoantigen)主要由位于外显子区域上的体细胞突变产生[1]; 随后研究发现,基因组的非编码区,特别是反转座子可以通过表冠遗传失调[2],转录并表达蛋白质,这些蛋白质同样可以作为新生抗原,与主要组织相容性复合体(MHC)结合,继而被相关免疫细胞识别. 因此得益于全基因组测序技术,研究范围可以从外显子组扩展到全基因组. 以人类全基因组为例, Yates 等[3] 在 Ensembl 公开的 GRCh37 版本人类参考基因组文件大小达 3GB, 2008 年由美国国立人类基因组研究所等科研机构发起的国际千人基因组计划,需要对来自 10 多个国家和地区的 2 504 个个体全基因组测序数据进行对比分析[4]. 完整的全基因组数据分析流程十分复杂,涉及多个分析步骤,其中最为核心的步骤是比对过程.

比对过程不仅在基因组数据分析中举足轻重,在转录组数据分析中也十分重要. 已有文献比较了 RNA-seq 数据常用的比对工具[5],包括 TopHat2、STAR、HISAT2 和 RASER 等. 而对于分析 DNA-seq 的重测序数据,比对的目的则是将测序数据追溯到它们在参考基因组中的原始位置. 测序数据的长度和错误率、参考基因组的重复序列及测序数据本身和参考基因组的序列差异等因素,会为比对过程带来一定的难度. 因此,为了准确地完成比对过程, DNA-seq 的比对软件也应运而生. 这些软件根据建立索引的方式可分为 2 类,即基于哈希表(Hash Tables)建立索引和基于前后缀树(Prefix/Suffix Tries)建立索引[6].

其中 NovoAlign(http://www.novocraft.com/products/novoalign/)为最常用的基于哈希表的比对软件,而 BWA-MEM[7] 为最常用的基于前后缀树的比对软件. 2 个比对工具的基本特征如表 1 所示. 2 个软件除在算法上不同外,比对质量分数(mapping quality, MQ)的取值范围也不同,默认条件下 BWA-MEM 的 MQ 分布在 [0, 60],而 NovoAlign 在 [0, 70] 分布. 其中: BWA-MEM 的 MQ 是短读序列与参考基因组错误配对的概率经对数函数($-10\lg P$)转换得到的; NovoAlign 的 MQ 是参考基因组上给定位置观测到短读序列的概率经对数函数($10\lg P$)转换得到的. 因而 MQ 越高,代表该短读序列比对至正确位置的可能性越大,比如

* 国家自然科学基金资助项目(31571361)
† 通信作者: 庞尔丽(1973—),女,博士,副教授. 研究方向: 生物信息学. E-mail: pangerli@bnu.edu.cn
收稿日期: 2020-09-16

MQ为60,意味着比对错误的概率仅为10^{-6},比对结果十分可信.尽管2种比对软件采用了不同算法建立比对打分系统,比对过程中参数的设置也会影响比对得分,但是最终都在输出结果中以概率的方式表示比对结果的优劣.另一方面,2个软件的比对相关参数具有不同默认值.在BWA-MEM中,错配罚分默认为4,空位打开罚分(gap opening penalty)为6,空位扩展罚分(gap extend penalty)为1.还可以通过设定带宽(band width)控制空位的长度.但需要注意的是,最大空位长度还与打分矩阵和匹配序列(hit)的长度有关.使用NovoAlign比对时,默认条件下,错配罚分为30,空位打开罚分为40,空位扩展罚分为6.比对分数阈值的计算公式为$(L-a)\times b$,其中L为短读序列的长度,a与参考基因组大小正相关,b一般小于空位扩展罚分.

表1 BWA-MEM和NovoAlign的比对特征

比对工具	版本[1)]	引用次数[2)]	发布年份	错配[3)]	空位	双末端测序	多线程运行	比对质量分数	算法
BWA-MEM	v0.7.16	25 369	2013	Y	Y	Y	Y	0~60	Burrows-Wheeler Transform + prefix/suffix tree
NovoAlign	v3.09.00	−	2010	Y	Y	Y	N	0~70	Needleman-Wunsch + Hash Table

[1)] BWA-MEM和NovoAlign的最新版本和引用次数均于2020年6月20日获取. [2)] BWA-MEM引用次数查询于Web of Science(https://webofknowledge.com),NovoAlign没有公开发表的文献,可在官方网站http://www.novocraft.com/products/novoalign/上获取. [3)] 错配、空位、双末端测序和多线程运行展示了比对工具是否具备该特性,Y代表存在该特性,N表示不允许该特性.其中,二者皆允许比对过程中存在错配和空位,皆可比对双末端测序数据.BWA-MEM可以多线程运行,NovoAlign的非商业版本不能分配运行线程数.同时,二者的比对质量分数和算法也存在差异.

这2个软件也具有相同性,都以FASTQ文件作为输入,都以SAM文件作为输出.比对参数选择上:1)能够比对双末端测序数据;2)可以清除低质量碱基.

目前已有许多研究从不同角度比较了这些软件:Li等[6]在算法层面上从建立索引方式、空位比对和长测序数据比对等角度回顾总结了几种比对算法的发展;Ruffalo等[8]使用模拟数据,设定变异类型的频率与长度和比对质量分数阈值,比较了这些软件在运行时间和准确性上的差异,发现BWA在牺牲准确性的基础上具有极高的比对率,而NovoAlign则能稳健地比对长插入缺失;其他研究[9−10]针对如何判定短读序列是否比对正确,分别提出了不同的评估方法;Hatem等[11]的研究从默认参数对比对结果的影响进行了评估,他们使用模拟和真实数据以默认参数或相同的比对参数评估9种软件,得到没有任何一种软件能够全方面优于其他软件的结论;2016年的一个研究[12]则评估了比对工具在人类基因组高度多态性区域上的比对率和准确性的表现,发现NovoAlign在高度多态区域有最高的比对率.除上述比较研究外,还有许多研究从数据品质、短读序列长度、基因组大小和串联重复序列比例角度评估了比对工具,发现NovoAlign和BWA比对性能存在差异,如:相同算法的比对结果更相近,NovoAlign对数据质量更为敏感[13];BWA均能以较快的速度处理不同长度的短读序列,NovoAlign分别对36、50、72及>100 bp测序数据敏感性更高,同时发现串联重复与错误比对率无关或有很低的相关性[14].

结合以上研究的结论,NovoAlign能以相对慢的运行速度得到整体较为准确的比对结果,BWA比对长测序数据时在速度上占有优势.

尽管已有许多研究比较了比对软件,但由于这些软件不断在更新,因此需要重新评估新版本的比对工具.其次,在基因组中的不同特征区域,如低复杂度区域和片段性重复区域,这些软件的表现还尚待评估.由于复制和重组的作用,低复杂度区域具有不稳定性[15],这类区域不受控制的扩展会导致一些人类疾病.比如,多聚谷氨酰胺疾病(亨廷顿病、脊髓和延髓肌萎缩症等)的致病原因是低复杂度区域扩张引起的蛋白质功能改变[16].片段性重复区域在灵长类动物进化、创造新基因和塑造人类遗传变异方面也发挥了重要作用[17],因为片段性重复区域之间的同源性可以启动非等位基因之间的同源重组,产生重复、缺失、倒位和移位等变异类型.并且已有许多研究证实了片段性重复区域与疾病相关,Velocardioface/DiGeorge综合症与21号染色体q11上3个高度(99%)相似的片段性重复区域有关[18].也有数据表明片段性重复区域及其相关的结构变异具有有益的作用,因重复而增加CCL3L1的拷贝数与HIV易感性的显著降低有关[19].有研究发现,在多种比对工具与变异识别工具组成的分析流程中,BWA-MEM和NovoAlign作为比对工具的结果准确性相对更高[20].基于此,本研究利用不同测序深度的NA12878人类个体真实测序数据和模拟数据,探讨BWA-MEM和NovoAlign在人类基因组中不同特征区域的表现,并从比对率、比对质量分数及基因组不同区域上的比对质量分数分布,对NovoAlign和BWA-MEM进行比较.希望评估结果可为研究人员根据特定基因组选择准确的比对工具提供参考.

1　材料和方法

1.1　参考基因组及注释信息　人类 GRCh37 版本的 FASTA 文件作为参考基因组, 下载于 ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_reference_assembly_sequence/hs37d5.fa.gz. 由于大多数错误比对的短读序列都与 X 或 Y 染色体相关, 特别是有许多来源于 X 染色体的短读序列比对至 Y 染色体上[21]. 所以本研究只评估比对工具比对常染色体的性能, 并用 NovoAlign 和 BWA-MEM 的 index 功能对参考基因组建立索引文件, 用于后续的比对过程.

人类 GRCh37 版本的注释文件下载于 NCBI, 版本索引号为 GCF_000001405.25. 低复杂度区域文件下载于http://figshare.com/articles/Low_complexity_regions_in_hs37d5/969685[22], 片段性重复区域文件下载于 http://humanparalogy.gs.washington.edu/build37/build37.htm.

1.2　真实数据　本研究同时采用真实和模拟测序数据. 真实测序数据集下载于 National Institute of Standards and Technology(NIST)的 GIAB FTP site, 选择 NA12878 个体 Illumina 平台产生的测序文件. 在 NA12878 个体 HiSeq 目录下, 每个亚目录存有约 20×～30×的测序数据, 这些数据可以拼接在一起用于后续的研究. 本研究选取前 5 个亚目录中的测序文件, 合并后用 seqtk (https://github.com/lh3/seqtk)随机抽取短读序列, 组成测序深度分别为 10×、15×、20×、30×和 40×的测序数据, 作为真实数据进行评估.

此外, GIAB 还提供了 NA12878 个体的标准(benchmark)变异文件, 文件囊括单核苷酸多态性(SNPs)和小插入缺失(InDels), 可为评估变异识别流程的结果提供黄金参考标准[23].

1.3　模拟数据　本研究使用 VarSim(0.8.4)[24] 和 ART (2.5.8)[25] 模拟不同测序深度的短读测序数据. 其中, VarSim 可在参考基因组中插入不同类型的变异, 包括 SNPs、InDels 和大的结构性变异(structural variations). 本文模拟 2 套突变基因组: 第 1 套仅引入 SNPs 与 InDels2 种类型的变异; 第 2 套除上述变异外, 还加入结构性变异, 如>50 bp 的插入、缺失和重复等. 首先, 利用 VarSim 生成包含上述变异类型的模拟突变基因组. 以人类 GRCh37/hg19 版本参考基因组(hs37d5)为基础, 以 dbSNP[26] 中真实的突变信息、GIAB 公布的 NA12878 个体高置信度标准 vcf 变异文件和下载于 Database of Genomic Variants的结构性变异文件作为突变数据库, 生成得到模拟突变基因组. VarSim 软件的主要运行参数为

```
python varsim.py\
—vc_in_vcf dbSNP.vcf \
—vc_in_vcf NA12878.vcf \
—reference hs37d5.fa \
—sv_dgv GRCh37_hg19_supportingvariants.txt
—id simu
```

接着, 将模拟基因组的 FASTA 文件输入到 ART 软件中, 参考 HiSeq2500 测序平台碱基质量分数的概率分布, 模拟测序深度分别为 10×、15×、20×、30×和 40×, 长度为 150 bp 的双末端短读序列. 将双末端短读序列片段化的平均长度设置为 300 bp, 长度标准差为 50 bp. ART 软件的运行参数为

```
art_illumina -ss HS25 -l 150 -f 10 -p -m 300 -s 50 -i simu.fa
```

1.4　品质检查与控制　首先使用 fastQC 检查测序数据的基本特征: 碱基质量分数及分布、4 种碱基数量比例、GC 含量、Ns 比例和短读序列长度等, 然后使用 trimmomatic 删除低品质的碱基及序列.

fastQC 和 trimmomatic 的运行命令及参数分别为

```
fastqc -o./ -t 8 fq.gz
java -jar trimmomatic.jar PE -threads 4 -phred33 -trimlog./log \
read1.fq.gz read2.fq.gz trimmed_read1.fq.gz \
unpair1fq.gz trimmed_read2.fq.gz unpair2fq.gz \
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 \
MINLEN:36 AVGQUAL:5
```

1.5　比对参数　NovoAlign(3.09.00)与 BWA-MEM (0.7.16)均以默认的参数运行, 不调整与错配及空位相关的分数.

具体的运行命令如下:

```
NovoAlign:
novoindex $indexfile $ref
novoalign -d $indexfile -f $read1 $read2 -o SAM
BWA:
bwa index $ref
bwa mem -t 4 $ref $read1 $read2
```

其中 BWA-MEM -t 分配运行的线程数, NovoAlign 非商业版本无法调整线程数.

1.6　评估方法　从 2 个角度评估 2 种比对工具: 比对率和比对质量分数. 比对率为总比对率, 即短读序列比对至参考基因组的百分比.

统计 NovoAlign 和 BWA-MEM 的比对质量分数分布. 由于 GATK 官方的最佳流程中, 推荐使用 MQ≥40 的标准过滤变异结果, 而 SAMtools 自带的变异过滤脚本中, MQ 阈值默认为 10, 因此, 本文将比对质量分

数区间划分为 [0, 10)、[10, 20)、[20, 30)、[30, 40)和[40, max], max 在 BWA-MEM 中为 60, 在 NovoAlign 中为 70.

2 结果与讨论

2.1 模拟测序数据 本研究模拟了 2 套突变基因组,其中只包括 SNPs 和 InDels 的基因组(以下称为简单基因组),同时还包括复杂结构性变异的基因组(以下称为复杂基因组). 其中: 简单基因组总计引入 2 998 727 个 SNPs, 239 351 个小插入和 248 400 个小缺失; 复杂基因组包括 2 979 716 个 SNPs, 不同长度的插入与缺失各 237 694 和 246 673 个, 长度>100 bp 的串联重复 8 233 个, 复杂变异 12 324 个.

2.2 测序深度对比对率的影响 在下游变异识别过程中,特别是对于识别体细胞变异,测序数据的深度会显著影响结果的敏感性和准确性,因此本节首先探究不同测序深度与比对率的关系. 同时把基因组按照其特征分成低复杂度、片段性重复和全基因组区域 3 类. 低复杂度区域一般由简单重复的氨基酸序列或碱基序列构成. 研究发现低复杂度区域会受到进化力量的选择[17]. 片段性重复区域则是长度>1 000 bp, 相似性>90% 的序列, 分布在基因组的多个位置. 其中低复杂度区域和片段性重复区域在基因组中的比例分别为 8.75% 和 4.47%.

从表 2 中不难发现, 对于每种比对工具, 无论是使用真实数据还是模拟数据, 随着测序的深度增加, 3 类区域的比对率并没有随之变化, 说明测序深度不会影响比对工具的比对率.

表 2 BWA-MEM 和 NovoAlign 在基因组不同区域上的比对率 %

测序深度	基因组区域	简单基因组		复杂基因组		真实测序数据	
		BWA-MEM	NovoAlign	BWA-MEM	NovoAlign	BWA-MEM	NovoAlign
10×	全基因组	100.00	98.82	99.91	97.60	96.91	91.59
	片段性重复区[1)]	4.84	3.89	4.99	3.59	6.81	5.34
	低复杂度区[2)]	9.04	9.06	9.23	9.13	9.25	8.82
15×	全基因组	100.00	98.82	99.91	97.59	96.91	91.59
	片段性重复区	4.84	3.89	4.99	3.59	6.81	5.34
	低复杂度区	9.04	9.06	9.22	9.12	9.25	8.82
20×	全基因组	100.00	98.82	99.91	97.59	96.91	91.59
	片段性重复区	4.84	3.89	4.99	3.59	6.81	5.34
	低复杂度区	9.04	9.06	9.23	9.12	9.25	8.82
30×	全基因组	100.00	98.82	99.91	97.59	96.91	91.59
	片段性重复区	4.84	3.89	4.99	3.59	6.81	5.34
	低复杂度区	9.04	9.06	9.22	9.13	9.25	8.82
40×	全基因组	100.00	98.82	99.91	97.59	96.91	91.59
	片段性重复区	4.84	3.89	4.99	3.60	6.81	5.34
	低复杂度区	9.04	9.06	9.22	9.12	9.25	8.82

[1)]基因组中片段性重复区域的比例为4.47; [2)]基因组中低复杂度区域的比例为8.75.

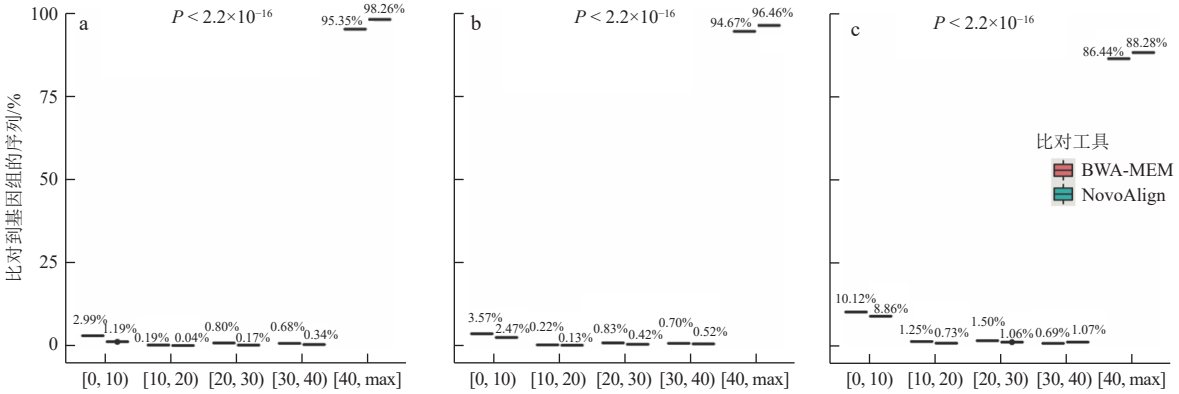
进一步比较 2 种比对工具在基因组不同特征区域的比对率, BWA-MEM 的比对率都高于 NovoAlign, 只有在简单基因组的低复杂度区域上, 2 种比对工具的比对率相差较小. 反映出在默认参数的情况下, BWA-MEM 会尽可能地将测序数据回溯至参考基因组.

因模拟数据保证了基因组不同区域测序深度均匀, 因此可以排除测序不均匀对比对率的影响. BWA-MEM 在低复杂度和片段性重复区域上的比对率均超过这 2 个区域在基因组中的占比(低复杂度区域和片段性重复区域在基因组中的比例分别为 8.75% 和 4.47%), 具体来说, 在真实数据中, BWA-MEM 甚至将 6.81% 的短读序列比对至片段性重复区域, 而 NovoAlign 在片段性重复区域比对率低于该片段在基因组中的占比(比对率在该区域<4%). 因此, NovoAlign 在片段性重复区域上的比对效果稍显不足, BWA-MEM 则能将测序数据过渡比对至这 2 个区域上.

表 2 也展示出比对率与基因组复杂程度的关系. 因无法明确 NA12878 个体真正的结构性变异数量, GIAB 公布的标准结构性变异文件只保留了极高可信度的变异结果, 因此不讨论真实测序数据的基因组复

杂度对比对率的影响, 仅比较简单基因组和复杂基因组的比对率(简单基因组只包括简单的变异类型, 而复杂基因组还引入了结构性变异). BWA-MEM 几乎可以将模拟简单基因组的全部测序数据比对至参考基因组, 而在模拟复杂基因组测序数据中, 比对率下降至 99.91%; NovoAlign 在模拟简单基因组上的序列总比对率为 98.82%, 在模拟复杂基因组上序列的总比对率为 97.60%. 可能由于复杂的结构性变异打乱基因组原始序列内容, 因此会使得一部分测序数据不能比对到参考基因组. 但有趣的是, 当基因组复杂性增加后, 低复杂度区域和片段性重复区域的比对率略有上升; 只有 NovoAlign 在片段性重复区域上的比对率是下降的, 从 3.89% 下降至 3.59%.

2.3　测序深度对比对质量分数分布的影响　当设置比对质量分数的阈值对短读序列质控时, 每个区间上分布的短读序列数量十分重要, 如果过滤区间设置过高, 则会丢失大量的短读序列, 反之则会增加比对结果的假阳性. 图 1 展示的是不同测序深度的真实和模拟数据经 2 种比对工具处理后, 比对质量分数分布的情况, 左侧图例为 BWA-MEM, 右侧为 NovoAlign. 从图 1 中观察到, 2 种比对工具针对不同测序深度的数据, 比对质量分数区间的分布没有显著差异.



a 简单基因组的比对质量分数分布; b 复杂基因组的比对质量分数分布; c 真实测序数据的比对质量分数分布. 每一区间上, 左侧为 BWA-MEM, 右侧为 NovoAlign.

图 1　BWA-MEM 和 NovoAlign 不同比对质量分数在 5 个测序深度分布

与真实测序数据(图 1-c)相比, 模拟数据(图 1-a 和 b)有更多的测序数据(94.67% ~ 98.26%)落在高比对质量分数区间([40, max]). 基因组中结构性变异也会影响测序数据在比对质量分数区间上的分布, 图 1-a 和 b 分别展示的是简单基因组和复杂基因组比对质量分数的分布, 可以看到当加入复杂的结构性变异后, BWA-MEM 和 NovoAlign 在低比对质量分数区间上的比例均有不同程度的增加, 使得 [40, max] 区间上的比例有所降低. 说明因结构性变异的存在, 使得短序列比对错误率提高.

观察 2 种比对工具在各区间上的分布, NovoAlign 有更多的测序数据分布在高比对质量分数区间, 其中对于比对质量分数 [40, max], 使用模拟数据比对时, BWA-MEM 有约95% 的测序数据位于该区间, NovoAlign 则有超过 96% 的测序数据处于该区间; 对于真实测序数据, BWA-MEM 和 NovoAlign 有显著差异(χ^2-test, $P< 2.2\times10^{-16}$).

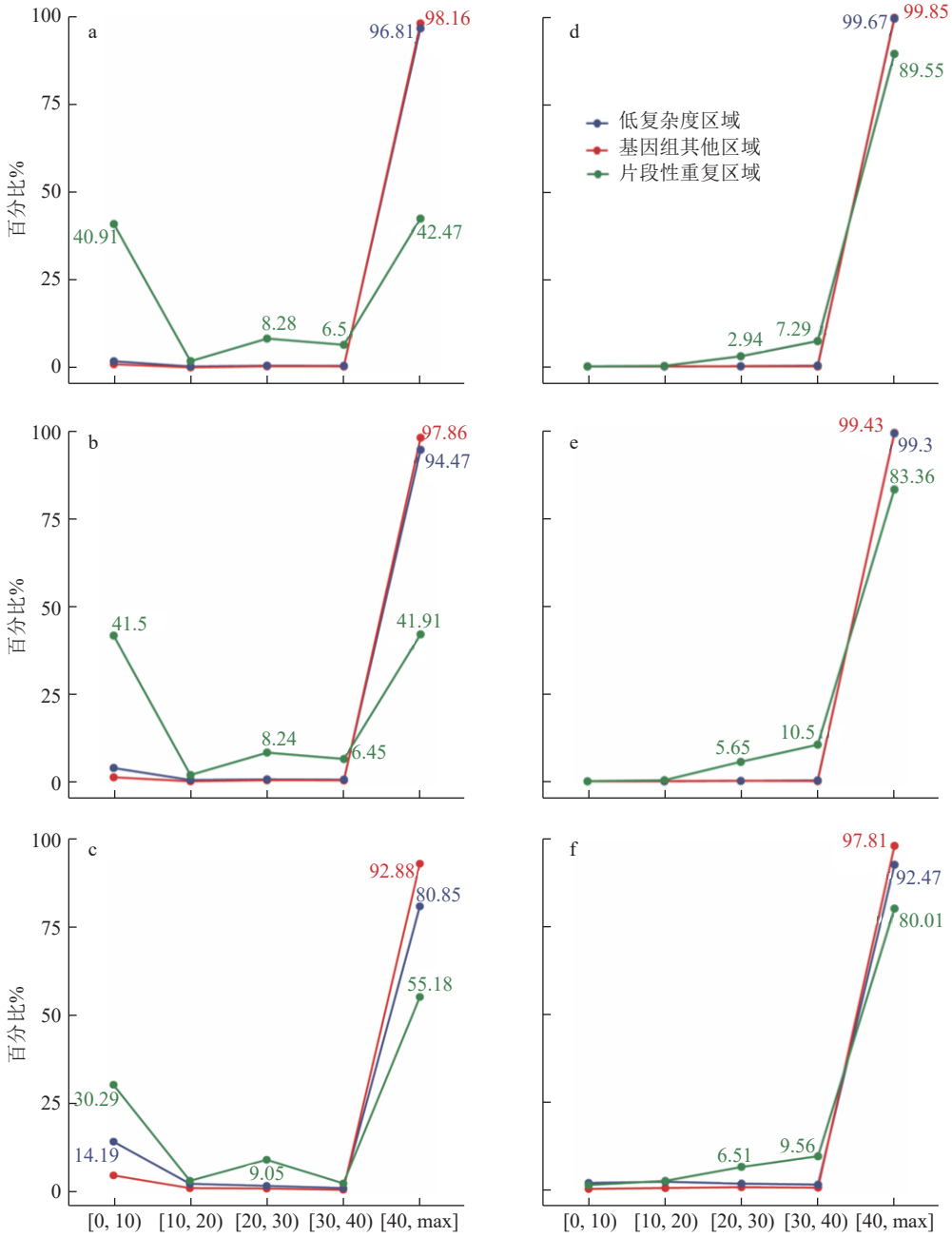
结合 2.2 节的结果, NovoAlign 虽然整体比对率不如 BWA-MEM, 但是有更多的测序数据获得较高的比对质量分数. 因 GATK 流程中不推荐保留 MQ<40 的变异结果, SAMtools 则是保留 MQ≥10 的变异结果. 而本研究中, 比对质量分数在 [10, 40)区间上的测序数据最多只占据 3.44%, 因此考虑敏感性和准确性的平衡关系, 对于下游识别变异时, 是否保留这部分短读序列有待进一步调查.

2.4　不同基因组区域的比对质量分数　由于比对质量分数的分布不随测序深度发生变化, 因此下面只分析 30×测序数据的比对结果. 去掉没有比对至参考基因组的测序数据, 统计比对成功的短读序列在低复杂度区域、片段性重复区域和基因组其余区域的比对质量分数分布情况.

对比 BWA-MEM(图 2-a~c)与 NovoAlign(图 2-d~f)的结果, 可以看到: 在片段性重复区域, 对于 BWA-MEM 来说, 在该区域上有大量比对质量分数<10 的短读序列, 意味着这些短读序列比对至参考基因组的多个位置或比对质量较差. 具体来说, BWA-MEM 在模拟数据中, 片段性重复区域仅有约 40% 的短读序列比对质量分数≥40, 这个比例甚至近似等于比对质量分数<10 的短读序列; 真实数据在该区域上有 55.18% 的短读序列比对质量分数≥40. 但在其他区域中, 大部分比对上的序列具有较高的比对质量分数, 如 BWA-MEM 在低复杂度区域和基因组其他区

域,皆有>80% 的测序数据比对质量分数≥40.

对于 NovoAlign 来说,在片段性重复区域上有较多的测序数据比对质量分数≥40,但是在 [20, 40)区间上还是存在一定比例的短读序列. 因此,NovoAlign 和 BWA-MEM 在片段性重复区域上的比对质量分数不如基因组其他区域,但是 NovoAlign 能够将更多的测序数据以≥40 的比对质量分数比对至片段性重复区域.



a. BWA-MEM 简单基因组的比对质量分数分布; b. BWA-MEM 复杂基因组的比对质量分数分布; c. BWA-MEM 真实测序数据的比对质量分数分布; d. NovoAlign 简单基因组的比对质量分数分布; e. NovoAlign 复杂基因组的比对质量分数分布; f. NovoAlign 真实测序数据的比对质量分数分布.

图 2 30×测序数据 BWA-MEM 和 NovoAlign 在基因组不同区域上的比对质量分数分布

在低复杂度与基因组其余区域上,比对质量分数的分布很类似,均有 90% 以上的短读序列以≥40 的比对质量分数比对至该区域上(BWA-MEM 比对到低复杂度区域的真实数据除外,只有 80.85%). 在真实测序数据中,基因组余下区域比对质量分数≥40 的短读序列比例略高于低复杂度区域;在模拟测序数据中,2 类基因组区域在各比对质量分数上的比例几乎相等.

总之,在默认条件下,相比于 BWA-MEM, NovoAlign 能在基因组的不同区域上获得较高的比对质量分数,

尽管 NovoAlign 的比对率不如 BWA-MEM, 但仍有理由相信, NovoAlign 可以保留大部分的变异信息, 特别是在片段性重复区域上. 不考虑运行时间, 当基因组中存在大量的片段性重复区域时, 推荐优先使用 NovoAlign 作为比对工具处理测序数据.

3　结论

经上述分析, 可以得到以下结论:

1)BWA-MEM 和 NovoAlign 可以稳定地处理不同深度的测序文件. 一般研究考虑实际经济因素, 会选择 30×的测序深度, 仅从比对率和比对质量分数分布的角度观察, 选择该深度是可行的.

2)BWA-MEM 相比于 NovoAlign 能够最大程度地利用测序数据, 将其比对至参考基因组上. 至少存在 2 个因素使得一部分短读序列不能比对至参考基因组: 一个是受到基因组本身复杂性的影响, 部分源自片段性重复区域的短读序列, 无法比对或唯一比对至正确位置; 另一个是比对工具参数的设置也会拒绝一些短读序列比对至参考基因组上.

3)BWA-MEM 和 NovoAlign 的比对质量分数分布存在差异, NovoAlign 有更多的短读序列分布在高分数区间. 在后续的变异识别过程中, GATK 在最佳流程中推荐保留比对质量分数≥40 的变异结果, 而 SAMtools 变异过滤程序默认将比对质量分数的阈值设置为 10, 在本研究中, 区间 [10,40) 的短读序列<4%, 因此是否有必要过滤该区间的测序数据有待考量. 片段性重复区域的比对质量分数低于基因组的其他区域, 而 NovoAlign 相较于 BWA-MEM, 能将更多的短读序列以高比对质量分数比对至该区域.

此外, 在实际运行过程中, 受到 2 种比对工具本身算法的影响, 处理同样的测序数据, BWA-MEM 的速度优于 NovoAlign. 这是因为在对测序数据和参考基因组建立索引时, BWA-MEM 采用较为复杂的树结构, 将部分重复的序列合并存储, 快速地比对和查询. 同时当参考基因组或测序数据较大时, 由于 BWA-MEM 可以指定线程数, 能够快速地得到比对结果, 优先推荐使用 BWA-MEM. 但是, 结合本文结论和之前的文献研究结果, NovoAlign 处理较长空位和复杂基因组时, 如片段性重复区域, 表现更加优越. 在本研究中, 短读序列的长度为 150(或 148) bp, 根据 NovoAlign 的比对打分系统, 默认参数下, 一对短读序列的比对结果中, 最多允许出现 43 个错配, 或一个 206 bp 长的缺失, 或与参考基因组的最低相似性(identity)为 86%. 当比对时允许存在较低的相似性, 比对的准确性会随着测序错误率增加; 但是当基因组存在较为复杂的结构性变异时, 宽松的阈值设定能够提高精确度(保留比对过程中出现的长空位而不是连续的单碱基错配)并且节省运行时间. 特别是通过对测序数据的拆分来实现并行处理, NovoAlign 的速度劣势也可以得到解决. 对二代甚至是三代测序数据, 数据的分析处理远不止比对这一步骤, 还涉及后续的变异识别与注释, 所以还需要进一步地计算和评估哪种比对工具与变异识别工具的搭配具有更好的准确性和敏感性.

4　参考文献

[1] SCHUMACHER T N, SCHREIBER R D. Neoantigens in cancer immunotherapy[J]. Science, 2015, 348(6230): 69

[2] SMITH C C, SELITSKY S R, CHAI S J, et al. Alternative tumour-specific antigens[J]. Nature Reviews Cancer, 2019, 19(8): 465

[3] YATES A D, ACHUTHAN P, AKANNI W, et al. Ensembl 2020[J]. Nucleic Acids Research, 2020, 48(D1): D682

[4] SUDMANT P H, RAUSCH T, GARDNER E J, et al. An integrated map of structural variation in 2,504 human genomes[J]. Nature, 2015, 526(7571): 75

[5] SAHRAEIAN S M E, MOHIYUDDIN M, SEBRA R, et al. Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis[J]. Nature Communications, 2017, 8(1): 59

[6] LI H, HOMER N. A survey of sequence alignment algorithms for next-generation sequencing[J]. Briefings in Bioinformatics, 2010, 11(5): 473

[7] LI H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM [EB/OL]. [2020-09-10]. https://figshare.com/articles/poster/Aligning_sequence_reads_clone_sequences_and_assembly_con_gs_with_BWA_MEM/963153/1. doi: 10.6084/m9.figshare.963153.v1

[8] RUFFALO M, LAFRAMBOISE T, KOYUTÜRK M. Comparative analysis of algorithms for next-generation sequencing read alignment[J]. Bioinformatics, 2011, 27(20): 2790

[9] FONSECA N A, RUNG J, BRAZMA A, et al. Tools for mapping high-throughput sequencing data[J]. Bioinformatics, 2012, 28(24): 3169

[10] HOLTGREWE M, EMDE A K, WEESE D, et al. A novel and well-defined benchmarking method for second generation read mapping[J]. BMC Bioinformatics, 2011, 12(1): 210

[11] HATEM A, BOZDAĞ D, TOLAND A E, et al. Benchmarking short sequence mapping tools[J]. BMC Bioinformatics, 2013, 14(1): 184

[12] TIAN S L, YAN H H, NEUHAUSER C, et al. An analytical workflow for accurate variant discovery in highly

divergent regions[J]. BMC Genomics, 2016, 17(1): 703
[13] YU X Q, GUDA K, WILLIS J, et al. How do alignment programs perform on sequencing data with varying qualities and from repetitive regions?[J]. BioData Mining, 2012, 5(1): 6
[14] THANKASWAMY-KOSALAI S, SEN P, NOOKAEW I. Evaluation and assessment of read-mapping by multiple next-generation sequencing aligners based on genome-wide characteristics[J]. Genomics, 2017, 109(3/4): 186
[15] ELLEGREN H. Microsatellites: simple sequences with complex evolution[J]. Nature Reviews Genetics, 2004, 5(6): 435
[16] GATCHEL J R, ZOGHBI H Y. Diseases of unstable repeat expansion: mechanisms and common principles[J]. Nature Reviews Genetics, 2005, 6(10): 743
[17] BAILEY J A, EICHLER E E. Primate segmental duplications: crucibles of evolution, diversity and disease[J]. Nature Reviews Genetics, 2006, 7(7): 552
[18] PAVLICEK A, HOUSE R, GENTLES A J, et al. Traffic of genetic information between segmental duplications flanking the typical 22q11.2 deletion in velo-cardio-facial syndrome/DiGeorge syndrome[J]. Genome Research, 2005, 15(11): 1487
[19] GONZALEZ E, KULKARNI H, BOLIVAR H, et al. The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility[J]. Science, 2005, 307(5714): 1434
[20] BUSH S J, FOSTER D, EYRE D W, et al. Genomic diversity affects the accuracy of bacterial single-nucleotide polymorphism-calling pipelines[J]. GigaScience, 2020, 9(2): giaa007
[21] LEE H, LEE K W, LEE T, et al. Performance evaluation method for read mapping tool in clinical panel sequencing[J]. Genes & Genomics, 2018, 40(2): 189
[22] CHAUDHRY S R, LWIN N, PHELAN D, et al. Comparative analysis of low complexity regions in Plasmodia[J]. Scientific Reports, 2018, 8(1): 335
[23] ZOOK J M, MCDANIEL J, OLSON N D, et al. An open resource for accurately benchmarking small variant and reference calls[J]. Nature Biotechnology, 2019, 37(5): 561
[24] MU J C, MOHIYUDDIN M, LI J, et al. VarSim: a high-fidelity simulation and validation framework for high-throughput genome sequencing with cancer applications[J]. Bioinformatics, 2015, 31(9): 1469
[25] HUANG W C, LI L P, MYERS J R, et al. ART: a next-generation sequencing read simulator[J]. Bioinformatics, 2012, 28(4): 593
[26] SHERRY S T, WARD M H, KHOLODOV M, et al. dbSNP: the NCBI database of genetic variation[J]. Nucleic Acids Research, 2001, 29(1): 308

Comparison of BWA-MEM and NovoAlign using human whole-genome next-generation sequencing reads

WANG Wenya PANG Erli†

(Key Laboratory for Biodiversity Science and Ecological Engineering of Ministry of Education, College of Life Sciences, Beijing Normal University, 100875, Beijing, China)

Abstract The development of sequencing technology has led to an increasing amount of next-generation sequencing data. To align numerous reads to reference genomes accurately is the basis for downstream analysis. BWA-MEM and NovoAlign, the most widely used DNA-seq alignment tools, have not been evaluated for applications to different structural regions in a genome. In the present work, we estimate the two alignment tools in low complexity region, segmental duplication region and the remaining region of the human genome using real and simulated data. BWA-MEM could align reads to reference genome, and even excessively align reads to low complexity regions and segmental duplication regions with low mapping quality. Compared with BWA-MEM, NovoAlign could align relatively fewer reads to reference genome, but most aligned reads have higher mapping quality. It is suggested that NovoAlign should be used when genomes are interspersed high segmental duplications.

Keywords next-generation sequencing; alignment; BWA-MEM; NovoAlign; whole-genome

【责任编辑: 武 佳】