技术前沿 DOI:10.11699/cyxb20131202

全外显子组测序分析中预处理方法和变异识别方法的比较

闫 瑾,潘 琦,任 红
(重庆医科大学附属第二医院感染科、重庆医科大学病毒性肝炎研究所,重庆 400010)

【摘 要】**目的**:比较全外显子组数据分析中不同的预处理方法和变异过滤方法对变异识别的影响。**方法**:利用 2 例全外显子组测序数据,从使用不同的预处理方法(FASTX-Toolkit、Trimmomatic 及未做预处理)、修饰后不成对读长(single-end reads,SE)取舍策略以及变异过滤方法[Hard 过滤和变异质量得分重新校正(variant quality score recalibration,VQSR)]3 个方面,通过数据覆盖深度(depth of coverage,DP)、识别变异的数目、转换/颠换比值和基因型一致性等特征,比较他们对全外显子组变异识别结果的影响。**结果**:Trimmomatic 预处理后的读长测序 DP 与未预处理的原始数据接近,但明显高于 FASTX-Toolkit 预处理方法。当 DP≥10×且基因型质量分数(genotype quality score,GQ)≥20 时,经 Trimmomatic 预处理后识别到的单核苷酸变异(single nucleotide variant,SNV)数量比 FASTX-Toolkit 多,与未预处理组接近。当包含 SE 时,FASTX-Toolkit 组多识别出的 SNV 数量高于(28%)Trimmomatic 组(5%)。当样本量较少时,在所有实验组中 Hard 过滤方法滤掉的 SNV 要少于 VQSR。**结论**:Trimmomatic 修饰(过滤)原始序列更温和,而 FASTX-Toolkit 可能过度过滤了原始数据。保留 SE 有利于下游变异识别。Hard 过滤相较于 VQSR 表现出了更高的容忍度。

【关键词】全外显子组测序;预处理;变异识别
【中国图书分类法分类号】R857.3;Q344+.12 【文献标志码】A 【收稿日期】2013-03-26

Comparison of methods of pre-processing and variant filtering in analyzing whole exome sequencing data

YAN Jin, PAN Qi, REN Hong
(*Department of Infectious Disease, the Second Affiliated Hospital, Institute for Viral Hepatitis, Chongqing Medical University*)

【Abstract】***Objective***: To investigate effects of methods of pre-processing and variant filtering on variant recognition in analyzing whole exome sequencing data. ***Methods***: Through the calculation of depth of coverage (DP), number of variants, transition/transversion and

作者介绍:闫 瑾,Email:essq37-3-9@163.com,
研究方向:病毒性肝炎相关性肝细胞肝癌的全外显子组分析。
通信作者:任 红,Email:renhong0531@vip.sina.com。
基金项目:国家自然科学基金资助项目(编号:30930082)。

[7] Hirt B.Selective extraction of polyoma DNA from infected mouse cell culture[J].J Mol Biol,1967,26(2):365-369.
[8] 张秉强,吴 莹,唐 霓,等.向下转膜法在乙肝病毒 Southern 杂交中的应用[J].生物技术通报,2005(2):26-28.
Zhang B Q,Wu Y,Tang N,et al.Application of downward capillary transfer in the southern blotting of HBV[J].Biotechnology Bulletin,2005(2):26-28.
[9] 崔 静,胡接力,张文露,等.化学发光法检测体外 HBV 复制[J].生物技术,2009,19(2):26-28.
Cui J,Hu J L,Zhang W L,et al.Analysis of hepatitis B virus replication in vitro by chemiluminescent detection method[J].Biotechnology,2009,19(2):26-28.
[10] Lodish H,Berk A,Zipursky S L,et al.Molecular cell biology[M].4th ed.New York:W.H.Freeman,2000.
[11] Newbold J E,Xin H,Tencza M,et al.The covalently closed duplex form of the hepadnavirus genome exists in situ as a heterogeneous population of viral minichromosomes[J].J Virol,1995,69(6):3350-3357.
[12] Kock J,Theilmann L,Galle P,et al.Hepatitis B virus nucleic acids associated with human peripheral blood mononuclear cells do not originate from replicating virus[J].Hepatology,1996,23(3):405-413.
[13] Chong C L,Chen M L,Wu Y C,et al.Dynamics of HBV cccDNA expression and transcription in different cell growth phase[J].J Biomed Sci,2011,18(1):96.
[14] Jang J H,Rives C B,Shea L D.Plasmid delivery in vivo from porous tissue-engineering scaffolds:transgene expression and cellular transfection[J].Mol Ther,2005,12(3):475-483.
[15] He M L,Wu J,Chen Y,et al.A new and sensitive method for the quantification of HBV cccDNA by real-time PCR[J].Biochem Biophys Res Commun,2002,295(5):1102-1107.

(责任编辑:冉明会)

non-reference concordance, we compared the effects of different pre-processing methods (FASTX-Toolkit, Trimmomatic and non treatment) and strategies of single-end (SE) inclusion and 'Hard' filter and variants quality score recalibration (VQSR) on variants calling in variants filter using whole exome sequencing data from two test samples. ***Results***: Trimmomatic pre-processed reads showed similar DP to reads without pre-processing, but significantly higher than that of FASTX-Toolkit pre-processed reads. With DP ≥10× and genotype quality (GQ) ≥20, number of called single nucleotide variants (SNV) identified by Trimmomatic was greater than that identified by FASTX-Toolkit, but similar to that without pre-processing. With the inclusion of SE, number of variants increased significantly for FASTX-Toolkit pre-processing (28%) than Trimmomatic pre-processing (5%). In the all settings, 'Hard' filtering filtered less SNVs than VQSR filtering in small sample size. ***Conclusions***: Sequence reads are trimmed and/or filtered moderately by Trimmomatic, whereas they seemed to be over-filtered by FASTX-Toolkit. Keeping the SE is good for variants recognition in the downstream analysis. The 'Hard' filtering showed a more favorable tolerability profile than 'VQSR' filtering.

【Key words】 whole exome sequencing; pre-processing; variant filtering

自全外显子组技术出现以来，研究者们利用该技术揭示了众多孟德尔疾病发病的原因[1]。随着近年来测序技术飞速发展，第 2 代测序技术应用日趋成熟，费用成本逐渐下降，全外显子组测序被越来越多的实验室和临床检测所应用。虽然第 2 代测序技术通量大幅提升，测序深度不断提高，但在带来更高的碱基识别率的同时，又对生物信息学提出巨大的挑战。由于技术的高速发展以及学界的争议，仍然没有一套公认的、标准的第 2 代测序数据质量控制方法。如是否需要对测序得到的原始序列进行质量控制以及哪些因素对变异识别造成影响都尚无定论。本研究试图比较 FASTX-Toolkit（http://hannonlab.cshl.edu/fastx_toolkit/）、Trimmomatic [2]2 种不同的预处理方法以及预处理后产生的不成对读长（single-end reads，SE）的取舍策略和变异过滤方法，对全外显子组测序数据分析中的测序数据覆盖深度（depth of coverage，DP）、识别变异的数目、转换/颠换比值（transition/transversion ratio，Ti/Tv）和基因型一致性的影响。

1 材料和方法

1.1 样本

利用 2 组 1000 基因组计划中测序得到的全外显子组数据（NA12878 和 NA18967，http://depts.washington.edu/swansonw/Swanson_Lab/Data.html）作为样本进行比较。2 个样本均使用 NimbleGen® SeqCap EZ Exome probes（v1.0）进行外显子捕获，并在 Illumina® Genome Analyzer Ⅱx 测序仪上采用 76 bp 双末端测序方法进行测序[2-3]，分别生成 13.4 GB 和 17.0 GB fastq 格式正反向序列数据。

1.2 全外显子组测序分析流程

本文所有数据分析程序均在 Mac OS X（v10.8.2）命令行终端环境下运行。

参考 GATK 建议的第 2 代测序变异识别和基因分型步骤框架[4-5]，本研究设计了实验中使用的全外显子组分析流程（图 1），共分 3 个阶段。

第 1 阶段为外显子组原始数据处理。首先，对 2 例外显子组数据分别用 FASTX-Toolkit（v0.0.13）和 Trimmomatic（v0.20）进行预处理，并设无预处理对照组（w/o pre-processing）。预处理步骤包括测序接头/引物的剪切（所用引物序列见表 1）、依据碱基质量修饰和过滤、过滤低质量原始序列（定义质量值小于 20 为低质量）和滤过人工产物。由于预处理会将原长度一致的正反向配对序列变得长度不一，生成 SE。由于比对程序无法识别长短不一的读长，故需要把它们与成对读长（paired-end reads，PE）分开处理。然后，将预处理后生成的数据输入到 BWA（v0.5.9）[6]中，与参考序列（GRCH7）进行比对。因为 BWA 无法同时处理 PE 和 SE，故需要用"sampe"和"samse"2 种命令分别运算，最终生成二进制序列比对/图谱文件（binary SAM file，BAM）。鉴于目前还不清楚经非对称修饰后产生的 SE 对下游的数据分析能造成多大的影响，本研究将经过预处理过后的数据分为 PE&SE 组（pse）（使用 Picard[7] "MergeSamFiles"整合 SAM 文件）和 PE 组（pe）分别分析。

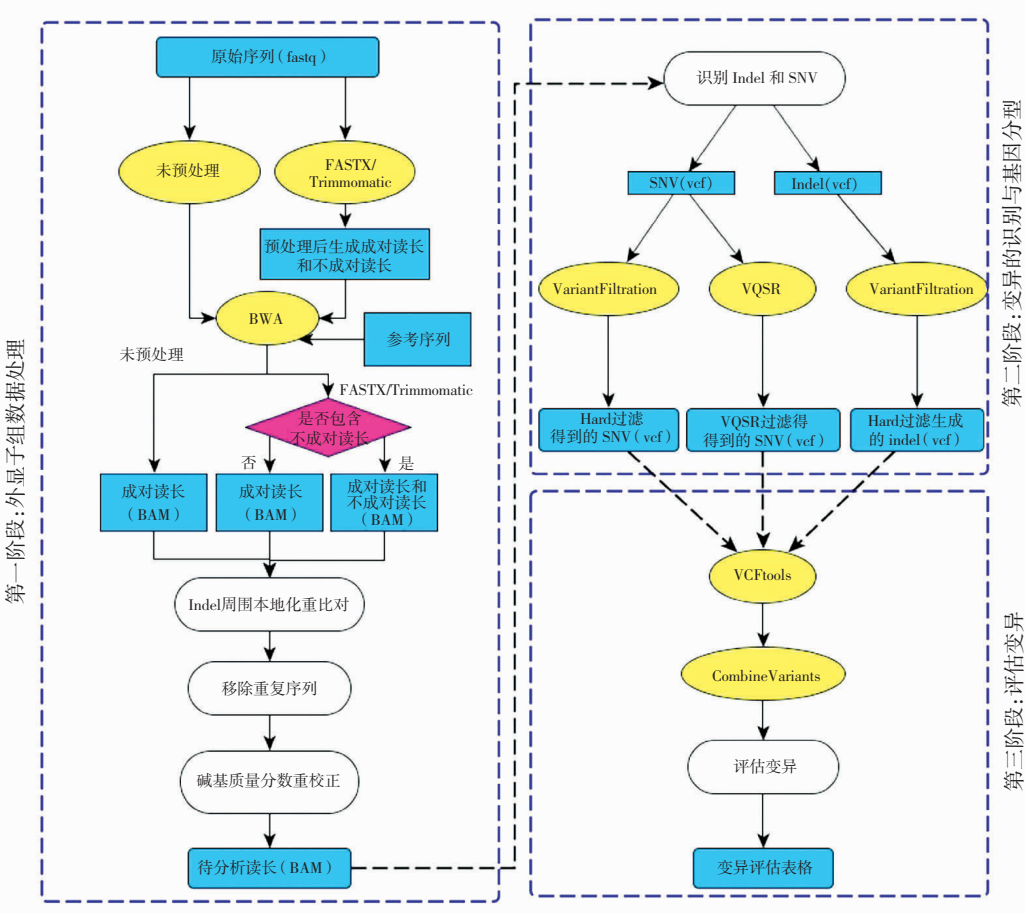
BWA 在进行比对时难以避免会产生错误，尤其在插入和缺失变异（insertion and deletion，Indel）的片段周围。若不进行校正，这些错误很容易被误认为是单核苷酸变异（single nucleotide variant，SNV）。使用 GATK[5]（v1.6）"RealignerTarget Creator"对这些区段进行本地化重比对，Picard 中的"Mark Duplicates"工具移除重复序列，使用"CountCovariates"和"TableRecalibration"基于测序循环次数和周围序列情况对碱基质量分数重校正。最后使用 SAMtools[7]对 BAM 文件进行索引编辑。

第 2 阶段为变异识别和基因分型。以人类基因组（GRCh37）和 dbSNP 数据库（Release 132，Build 37）作为参考序列，使用GATK"UnifiedGneotyper"工具处理第 1 阶段生成的结果，识别 2 个样本 SNV 和 Indel。"UnifiedGenotyper"的参数"stand_emit_conf "和"stand_call_ conf"均使用 30.0，以忽略

所有低质量(phred-based<30)的变异,若有基因型未能识别均标记为“N”。

使用 GATK“selectVariants”对生成的所有变异结果进行挑选,把 SNV 结果挑选出来。选择 Hard 过滤和变异质量分数重校正(variant quality score recalibration,VQSR)2 种方法来过滤原始 SNV 以比较两者过滤效果。将 Hard 过滤工具“VariantFiltration”6 个过滤参数(表 2)设为固定值:“QD<2.0”、“MQ<40.0”、“FS>60.0”、“HaplotypeScore>13.0”、“MQRankSum<-12.5”以及“ReadPosRankSum<-8.0”。使用固定参数:“maxGaussinas 4”和“percentBadVariants 0.05”运行 VQSR“VariantRecalibrator”和“ApplyRecalibration”。

第 3 阶段为变异结果统计和质量评价。依据 SNV 和 Indel 上的标记特征和参数,使用 VCFtools(v0.1.10)[8]进行过滤。针对 Hard 过滤后得到的结果,使用“QDFilter”、“MQFilter”、“FSFilter”、“HaplotypeScoreFilter”、“MQRankSumFilter”和“ReadPosRankSumFilter”对 SNV 进行过滤。为降低假阳性率,设错误发现率(false discovery rate,FDR)为 1%,故把VQSR 标记的变异中凡是满足真实敏感度大于 99%的 SNV(含有“TruthSensitivityTranche99.90to100.00”和“TruthSensitivityTranche99.00to99.90”标记)滤掉。最后,运行“VariantEval”工具中的“GenotypeConcordance”模块将实验中识别出的变异分别与 NA12878 和 NA18967 在 HapMap3 和 Omni 的变异数据[9]进行比较。生成的评价指标有 SNV 数量、Ti/Tv、dbSNP 率(即能在 dbSNP 数据中找到的变异数目的比例)、非参考变异



蓝色矩形:输入或输出文件,括号内为文件格式;黄色椭圆:软件名称;紫色菱形:条件判断;白色圆框:分析步骤;实线箭头:各阶段内部程序运行方向;虚线箭头:各阶段间程序运行方向

图 1　实验设计及全外显子组数据分析流程图

Fig.1　Design of the present study and the flow chart of whole exome sequencing data analysis

表 1　在 FASTX-Toolkit 和 Trimmomatic 中使用的引物(接头)序列

Tab.1　Adapters/primers used in FASTX-Toolkit and Trimmomatic

引物(接头)	序　列
Paried_End_Sequencing_Primer_SP1	ACACTCTTTCCCTACACGACGCTCTTCCGATCT
Paired_End_Sequencing_Primer_SP2	CGGTCTCGGCATTCCTACTGAACCGCTCTTCCGATCT
Paired_End_PCR_Primer_f	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
Paired_End_PCR_Primer_r	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT
Paired_End_Adapter_A2	GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG

敏感度(non-reference sensitivity,NRS)、非参考变异不一致率(non-reference discrepancy rate,NRD)。

1.3 数据 DP 分析

DP 是测序技术的重要指标之一。当测序长度不变时,提高测序深度可以获得更好的比对率和变异识别率。NimbleGen SeqCap EZ Exome (v1.0) 探针中包含了 25.2 MB 目标区域(targeted-region)和 34.1 MB 平铺区域(tiled-region),囊括了 CCDS 数据库(v20080430)中所公布的[10]所有无重复的蛋白质编码区域(包括外显子上下游 200 bp 以内),以及约 550 个miRNAs 序列。应用基因组坐标转换工具“liftover”[11],将 NimbleGen SeqCap EZ Exome (v1.0)(hg18) tiled 区域的坐标系转换成hg19 标准,共生成了 176 159 个连续的捕获基因组区域(总计 34 108 798 bp)。将 GRCh37/hg19 人类基因组序列 refGene(从 UCSC 基因组浏览器中下载,2012-6-3)作为参考基因序列,对 SeqCap EZ Exome (v1.0)基因组捕获的区间(从 NimbleGen 公司获得)进行定位。最后使用 GATK “Depthof-Coverage”工具对预处理后生成的 BAM 文件进行 DP分析。

2 结 果

2.1 预处理后的读长

使用 FASTX-Toolkit、Trimmomatic 预处理 2 个样本测序数据后读长剩余数量见表 3。FASTX-Toolkit 滤过后保留下的读长数明显少于经 Trimmomatic 处理后的。在保留下的读长中,经 FASTX-Toolkit 处理后 SE 所占比例(在 NA12878 为 30.4%,在 NA18967 中为 35.5%)要显著高于 Trimmomatic 处理后的(在 NA12878 为 3.3%,在 NA18967 中为 8.0%)。若将删除重复序列后的原始序列定义为唯一的读长,由比对后的结果可知,经 FASTX-Toolkit 处理后比对上参考基因组的读长要显著低于 Trimmomatic 处理后的以及未预处理的。

2.2 数据 DP

每个样本有 5 个 BAM 文件被用来检测 DP(图 2 和表4),包括未预处理的原始读长、FASTX-Toolkit 预处理生成的读长(pse 组和 pe 组)以及 Trimmomatic 预处理后生成的读长(pse 组和 pe 组)。从图 2A 中可看出未处理的原始序列的 DP 最高,NA12878 和 NA18967 均有 84%的碱基>20×,平均 DP

表 2 变异过滤中使用的过滤参数

Tab.2 Filter parameters used in variants filtering

过滤参数名称	描 述
QD(quality by depth)	变异质量值/测序 DP,分值越高表明该变异准确性越高
MQ(mapping quality)	比对质量
FS(Fisher strand)	使用 Fisher 确切检验检测读长标准差得到的 phred 尺度 *P* 值
HaplotypeScore	单倍体分值,用于表现拥有 2 个且仅有 2 个分离单倍体型位点的一致性。分值越高表明该区域序列比对效果越差
MQRankSum(mapping quality rank sum test)	对发生变异的读长的比对质量进行 Mann-Whitney *U* 检验所得 *Z* 值
ReadPosRankSum(read position rank sum test)	对读长中变异发生位置到读长末端的距离进行 Mann-Whitney *U* 检验所得 *Z* 值
maxGaussinas	在进行变分贝叶斯算法时使用的高斯最大数量
percentBadVariants	在构建差质量变异的高斯混合模型时,使用最差分值的变异的比例

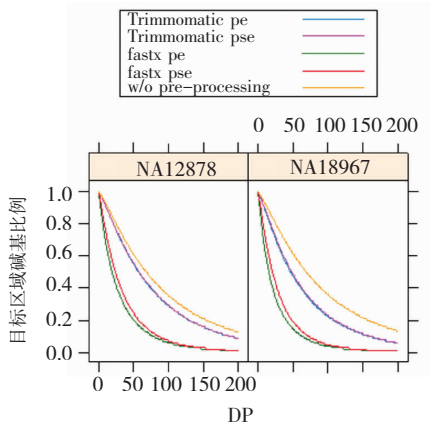
表 3 预处理后、重复序列及比对后读长数量统计

Tab.3 Number of reads after pre-processing,mapped to the human genome and duplicative reads

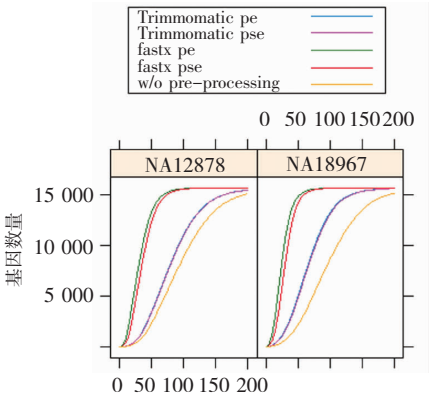
处理方式	过滤后读长数量			重复序列	比对后读长[b]
	正向序列	反向序列	总数[a]		
NA12878					
未预处理	37 520 750	37 520 750	75 041 500	6 270 719	61 879 633(90.0%)
fastx pse	15 901 894+7 524 420	15 901 894+6 385 715	45 713 923(60.9%)	10 458 902	35 061 064(51.0%)
fastx pe	15 901 894	15 901 894	31 803 788(42.4%)	2 389 424	29 295 863(42.6%)
Trimmomatic pse	30 224 628+1 173 192	30 224 628+919 252	62 541 700(83.3%)	5 970 087	56 178 578(81.7%)
Trimmomatic pe	30 224 628	30 224 628	60 449 256(80.6%)	4 682 178	55 407 250(80.6%)
NA18967					
未预处理	47 628 042	47 628 042	95 256 084	5 413 177	65 097 743(72.5%)
fastx pse	13 150 803+7 340 826	13 150 803+7 117 666	40 760 098(42.8%)	8 768 770	31 841 000(35.4%)
fastx pe	13 150 803	13 150 803	26 301 606(27.6%)	1 249 364	24 968 823(27.8%)
Trimmomatic pse	25 761 311+2 211 181	25 761 311+2 250 735	55 984 538(58.8%)	4 871 681	50 297 732(56.0%)
Trimmomatic pe	25 761 311	25 761 311	51 522 622(54.1%)	2 414 402	48 339 432(53.8%)

注:a,总数:等于正向序列与反向序列数目之和;括号中百分比表示过滤后剩余读长数目占原始读长的比例;b,比对后读长:为经过预处理或未经过预处理并去除重复序列后能与参考基因组比对上的读长;括号中的百分比表示比对上的读长占原始读长的比例

分别为 99.00× 和 101.88×。经 Trimmomatic 预处理后的读长与原始读长 DP 接近，也远高于 FASTX-Toolkit。SE 的加入对 FASTX-Toolkit 数据的 DP 提升有较显著的影响(14%~24%)。以上结果说明使用 FASTX-Toolkit 做预处理会极大的降低原始数据的碱基位点 DP 和平均 DP。



A. 目标区域中碱基比例与测序 DP 的关系曲线



B. 目标区域基因数量与平均 DP 的关系曲线

图 2　目标区域碱基和基因与测序 DP 的关系

Fig.2　Relationship between the targeted sequence and DP

使用 refGene(hg19，含有 23 652 个唯一的基因)作为 DP 分析的参考序列，总共有 15 737 个(占 refGene 的 66.5%)参考基因被捕获，少于 CCDS 作为参考序列捕获的蛋白质编码基因(16 188 个)[12]。NimbleGen SeqCap EZ Exome (v1.0)捕获到的基因经过预处理后 DP 分布见图 2B。当测序 DP<20× 时，FASTX-Toolkit 处理后的基因数目最多。例在 NA18967 中，FASTX-Toolkit 过滤后有 23.3%(pse)和 38.1%(pe)的捕获基因 DP<20×；而使用 Trimmomatic，仅有 4.2%(pse) and 4.7%(pe)的捕获基因 DP<20×。以上结果证明不同的预处理方法会对文库捕获的基因测序 DP 造成显著的差别，因此也会影响了识别重要(医学相关)基因的效能。

2.3　变异评估

对测序数据基因型质量影响最大的 2 个因素是 DP 和基因型质量分数(genotype quality score，GQ)[13]。GATK 认为外显子组测序数据中比对错误和 DP 之间的关系并不明确，故不建议在外显子组数据处理中直接应用 DP 过滤器[5]。然而，序列中某一位点的 DP 和该样本识别出的变异数目之间的关系尚不清楚，故分析位点测序深度的阈值与识别变异数目的关系并作图(图 3A)。图中可见在不同的方案组中，SNV 的数量均随着 DP 的增加而迅速减少，特别是在当 DP≤5×。当DP≥10×时，识别变异的数量趋于相似。图 3B 表现了 SNV 数目和 GQ 关系。

以 DP≥10×、GQ≥20 为标准，统计识别到的 SNV 数目、Ti/Tv 以及与 HapMap 和 Omni 中已识别的基因型一致性。从表 5 中可以看到不同的预处理组间变异数目均有明显差别，未经过预处理的和用 Trimmomatic 预处理能识别出更多的变异。未预处理组中的变异数目约是使用 FASTX-Toolkit 预处理得到的 2 倍。当分析中包含 SE 时，在 FASTX-Toolkit 组中 SNV 识别率升高 15.5%~28.3%，而是在 Trimmomatic 中仅多识别了 0.7%~2.8%。而在比较 2 种基因型过滤方法中，Hard 过滤则能识别出更多的变异(表 6)，尤其是在未预处理组和 Trimmomatic 组。dbSNP 率虽不能直接用于衡量变异的质量优

表 4　使用不同的预处理方法后 tiled 区域的碱基 DP 分析

Tab.4　Analysis of base coverage in the tiled-region under different pre-processing methods

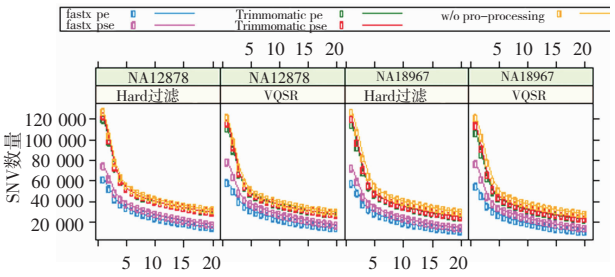
处理方法	DP			≥20×(%)	平均值(×)
	≥1×(bp)	≥1×(%)	≥20×(bp)		
NA12878					
未预处理	33 755 909	99	28 690 819	84	99.00
fastx pse	33 239 024	98	19 250 604	56	37.19
fastx pe	32 827 083	96	16 621 019	49	32.50
Trimmomatic pse	33 723 550	99	27 703 478	81	84.05
Trimmomatic pe	33 713 018	99	27 582 142	81	83.47
NA18967					
未预处理	33 699 948	99	28 702 984	84	101.88
fastx pse	33 067 279	97	17 421 302	51	31.52
fastx pe	32 478 468	95	14 069 586	41	26.27
Trimmomatic pse	33 598 036	99	26 335 580	77	71.10
Trimmomatic pe	33 561 322	98	25 970 491	76	69.72

劣，但在变异数目相近时 dbSNP 率更高则可说明假阳性率更低。经 Trimmomatic 预处理后虽然变异数量下降但 dbSNP 率上升（除 NA18967 pse VQSR 组持平），而 FASTX-Toolkit 变异数量和 dbSNP 率均下降。dbSNP 率在 VQSR 组中要比 Hard 过滤组中相对要高。

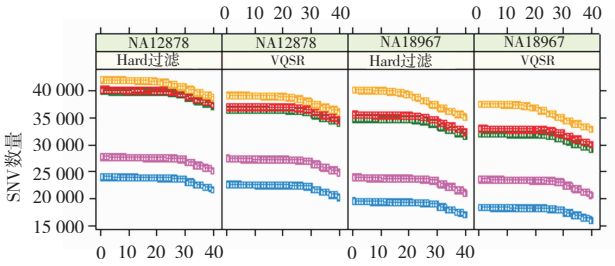
Ti/Tv 是对于评估新 SNV 变异特异性的关键指标，Ti/Tv 越高表明识别出的 SNV 越准确[4]。对于在外显子里的SNV，Ti/Tv 约为 3.0，其余的约为 2.0[14]。在本实验的新变异位点中，Trimmomatic 预处理后识别的新变异Ti/Tv 均比使用相同 SE 取舍策略的 FASTX-Toolkit 组高。已知变异位点 Ti/Tv 则均与之前基于 1000 基因组数据得到的估计值 2.8 相近[9]，但却与使用相同样本不同外显子捕获方法和不同测序方法[Illumina GA®Ⅱ（约 150×，93%碱基≥20×]的结果要低[4]。所有变异均能在 HapMap 和 Omni 数据库中找到（NRS=100%），差异性也与之前的研究一致，均小于 1%（表 5）[4]。

2.4 在性染色体上的变异

男性性染色体上发生的变异应是纯合子，而在女性性染色体上的变异显然不能出现在 Y 染色体上。所以本研究以此标准检查 2 个实验对象（NA12878 为女性，NA18967 为男性）在本流程中发生在性染色体上的错误（图 4）。在 NA18967 中，最多有 79 个和 25 个杂合性 SNV 分别标明在 X 染色体



A. 在不同的测序 DP 下变异数量的变化曲线



B. 在不同基因型质量值下变异数量的变化曲线

图 3 在不同测序 DP 和基因型质量值下 SNV 数量的变化曲线

Fig.3 Curve of number of SNVs under the different DP and GQ

表 5 经不同分析分案生成的测序 DP≥10×、基因型质量≥20 的 SNV 数量

Tab.5 Number of called SNVs with depth of coverage ≥10× and genotype quality ≥20 under different scenarios

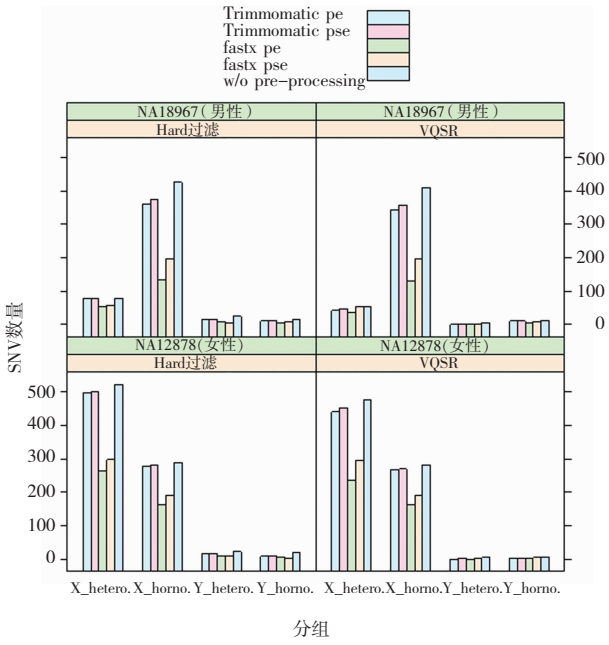
预处理方法	变异过滤方法	SNV				Ti/Tv			基因型一致性			
		总数	已知 SNV	新 SNV	dbSNP（%）	总数	已知 SNV	新 SNV	HapMap		Omni	
									NRS	NRD（%）	NRS	NRD（%）
NA12878												
未预处理	Hard	41 844	41 026	818	98.0	2.67	2.69	1.97	100	0.12	100	0.30
	VQSR	39 026	38 720	306	99.2	2.72	2.73	2.12	100	0.11	100	0.28
fastx pse	Hard	27 648	27 241	407	98.5	2.79	2.82	1.66	100	0.39	100	0.58
	VQSR	27 310	26 977	333	98.8	2.80	2.83	1.47	100	0.38	100	0.56
fastx pe	Hard	23 930	23 543	387	98.4	2.84	2.87	1.60	100	0.40	100	0.64
	VQSR	22 550	22 354	196	99.1	2.90	2.92	1.58	100	0.39	100	0.60
Trimmomatic pse	Hard	40 146	39 421	725	98.2	2.69	2.70	1.94	100	0.11	100	0.30
	VQSR	37 066	36 803	263	99.3	2.74	2.75	1.80	100	0.09	100	0.28
Trimmomatic pe	Hard	39 867	39 146	721	98.2	2.69	2.71	1.98	100	0.11	100	0.30
	VQSR	36 508	36 256	252	99.3	2.75	2.75	1.96	100	0.10	100	0.28
NA18967												
未预处理	Hard	39 445	37 973	1 472	96.3	2.74	2.75	2.31	100	0.29	100	0.37
	VQSR	36 969	36 021	948	97.4	2.79	2.80	2.44	100	0.26	100	0.35
fastx pse	Hard	23 787	22 994	793	96.7	2.86	2.89	2.17	100	0.59	100	0.71
	VQSR	23 469	22 747	722	96.9	2.87	2.89	2.17	100	0.58	100	0.72
fastx pe	Hard	19 386	18 726	660	96.6	2.91	2.94	2.25	100	0.56	100	0.68
	VQSR	18 299	17 816	483	97.4	2.95	2.97	2.40	100	0.56	100	0.65
Trimmomatic pse	Hard	35 506	34 227	1 279	96.4	2.77	2.78	2.58	100	0.24	100	0.34
	VQSR	32 919	32 066	853	97.4	2.82	2.82	2.76	100	0.21	100	0.30
Trimmomatic pe	Hard	34 764	33 513	1 251	96.4	2.77	2.78	2.50	100	0.26	100	0.36
	VQSR	32 026	31 219	807	97.5	2.83	2.83	2.77	100	0.23	100	0.33

上和 Y 染色体上；在 NA12878 中，最多有 40 个 SNV 发生在 Y 染色体上。

表 6 Hard 过滤和 VQSR 过滤后同组内 SNV 结果相同的和不同的数目

Tab.6 Number of shared and private SNVs by 'Hard' and 'VQSR' filtering

预处理方法	相同	不同	
		Hard过滤	VQSR
NA18967			
未预处理	36 858	2 587	111
fastx pse	22 855	932	614
fastx pe	17 986	1 400	313
Trimmomatic pse	32 869	2 637	50
Trimmomatic pe	31 991	2 773	35
NA12878			
未预处理	38 958	2 886	68
fastx pse	26 611	1 037	699
fastx pe	22 187	1 743	363
Trimmomatic pse	37 019	3 127	47
Trimmomatic pe	36 473	3 394	35



Hetero. 杂合子；Homo. 纯合子

图 4 两样本中性染色体上 SNV 的分布柱状图

Fig.4 Barchart of SNVs called on sex chromosomes of two samples

3 讨 论

之前，Liu 等[15]通过 5 例乳腺癌患者样本分析发现，利用 BWA 自带的质量值过滤功能，将低质量值的碱基（质量值小于 15）剪切掉后，并不能提高变异识别的准确性。而本次实验中，尝试利用 2 种流行的数据预处理工具，剪切掉引物（接头）序列并滤掉测序质量、过低的碱基，再进行比较变异识别的准确率。之所以选择 NA12878 和 NA18967 这 2 例样本作为测试对象，不仅是因为他们的数据公开，同时还因为他们已经被作为实验对象反复分析过，有既有的数据可以参考[4]。FASTX-Toolkit 与 Trimmomatic 都是应用广泛的质控预处理软件。通过本次实验，本研究发现原始数据在经过预处理后，DP 下降明显，会直接导致识别变异数目降低，特别是 FASTX-Toolkit。Trimmomatic 的优势在于能利用配对信息同时处理正反向序列，采用滑动窗口剪切，故比 FASTX-Toolkit 逐碱基剪切方法灵活。并且 Trimmomatic 经过质量控制后，识别到的变异假阳性率更低，表现为 dbSNP 率升高，NRD 比例降低，且新 SNV 的 Ti/Tv 也高于 FASTX-Toolkit。虽然 Trimmomatic 在本次实验中体现出了较大的优势，但是质量值低并不意味着某位点变异就一定是错误的，所以还不足以证明对原始的测序数据使用预处理软件后就一定会增加变异的正确识别率。序列比对程序在计算时需要耗费大量的计算机内存，当原始序列中引入了过多的引物序列和人工产物后，肯定会花更多的运算时间，而且还会降低序列比对的正确率。因此，若仅对原始序列剪切掉已知污染物序列而不对质量值进行修饰过滤，也不失为一种折中的办法。

预处理必定会造成原本成对的双末端序列长短不一，产生 SE。FASTX-Toolkit 预处理后产生的 SE 数目明显高于 Trimmomatic，故其在包含 SE 后测序深度增加了接近 5×，比对上参考序列的读长升高了 19.7%，识别的 SNV 也升高了 15.5%~28.3%。但在 Trimmomatic 中此提升并不明显，所以保留 SE 对 FASTX-Toolkit 预处理后的数据更加重要。

若依据 GATK 的建议，使用 VQSR 要达到最佳结果至少需要 30 个样本。但是在现实操作中会因为诸多限制，如实验对象数量过少、经费有限、计算机硬件资源不足等实际原因，而很难达到此标准。若样本量与建议的相差不是太大，除了增加样本量以外，加深测序深度，提高测序覆盖率，效果应该也等同于增加样本数目。而在处理小样本测序数据时，则只能选择使用固定了参数的 VQSR（如 maximum Gaussians、percentBadVariants）或使用 Hard 过滤。虽然本实验中仅使用了 2 例样本，但预处理方法皆是在每一个独立样本内进行，且与 GATK 推荐

的在 VQSR 中需要大量样本以建立高斯模型不同，固定参数后识别效果并不因为样本量多少而发生改变，所以使用本实验方法识别到的每个样本变异数量与样本量大小并无太大关系。

本次研究存在以下局限性：(1)使用的样本量较少，提高样本量有助于更进一步验证预处理和变异过滤在全外显子组测序分析中的效果；(2)使用于外显子捕获的文库 NimbleGen SeqCap EZ Exome (v1.0) 发布于 2009 年，故其捕获效率及捕获基因数量可能会较新版本 NimbleGen SeqCap 和其他外显子组捕获试剂稍差[3,12,16]；(3)FASTX-Toolkit 方法将原始序列过度过滤的原因有待于进一步研究。最后，虽然本次实验中出现的性染色体错配的数量较小(仅约为所有 SNV 的 0.3%)，对结果影响不是太大，但还是有必要完善现有比对程序或开发一种新的软件，以识别并纠正错配到性染色体上的变异。当完成本文时，基因组测序分析小组(genome sequencing and analysis group，GSA)发布了GATK v4，并更新了变异识别的最佳流程。

综上所述，由于外显子组数据分析中使用不同的算法和策略，会导致识别到的变异有较大的差别。虽本次实验对这些差别进行了初步的分析，但还需要更深入的分析进行进一步的论证。而且不论是对全基因组分析还是全外显子组分析，建立一套标准化的分析流程仍然是将来努力的目标。

参 考 文 献

[1] Gilissen C，Hoischen A，Brunner H G，et al.Unlocking Mendelian disease using exome sequencing[J].Genome Biol，2011，12(9)：228.

[2] Lohse M，Bolger A M，Nagel A，et al.RobiNA：a user-friendly，integrated software solution for RNA-Seq-based transcriptomics[J].Nucleic Acids Res，2012，40(Web Server issue)：W622-W627.

[3] George R D，Mcvicker G，Diederich R，et al.Trans genomic capture and sequencing of primate exomes reveals new targets of positive selection[J].Genome Res，2011，21(10)：1686-1694.

[4] Depristo M A，Banks E，Poplin R，et al.A framework for variation discovery and genotyping using next-generation DNA sequencing data[J].Nat Genet，2011，43(5)：491-498.

[5] Mckenna A，Hanna M，Banks E，et al.The genome analysis toolkit：a mapreduce framework for analyzing next-generation DNA sequencing data[J].Genome Res，2010，20(9)：1297-1303.

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

[7] Li H，Handsaker B，Wysoker A，et al.The Sequence Alignment/Map format and SAMtools[J].Bioinformatics，2009，25(16)：2078-2079.

[8] Danecek P，Auton A，Abecasis G，et al.The variant call format and VCFtools[J].Bioinformatics，2011，27(15)：2156-2158.

[9] Abecasis G R，Altshuler D，Auton A，et al.A map of human genome variation from populationscale sequencing[J].Nature，2010，467(7319)：1061-1073.

[10] Pruitt K D，Harrow J，Harte R A，et al.The consensus coding sequence(CCDS) project：Identifying a common protein-coding gene set for the human and mouse genomes[J].Genome Res，2009，19(7)：1316-1323.

[11] Hinrichs A S，Karolchik D，Baertsch R，et al.The UCSC genome browser database：update 2006[J].Nucleic Acids Res，2006，34(Database issue)：D590-D598.

[12] Asan，Xu Y，Jiang H，et al.Comprehensive comparison of three commercial human whole-exome capture platforms[J].Genome Biol，2011，12(9)：R95.

[13] Guo Y，Long J，He J，et al.Exome sequencing generates high quality data in non-target regions[J].BMC Genomics，2012，13：194.

[14] Bainbridge M N，Wang M，Wu Y，et al.Targeted enrichment beyond the consensus coding DNA sequence exome reveals exons with higher variant densities[J].Genome Biol，2011，12(7)：R68.

[15] Liu Q，Guo Y，Li J，et al.Steps to ensure accuracy in genotype and SNP calling from Illumina sequencing data[J].BMC Genomics，2012，13 Suppl 8：S8.

[16] Sulonen A M，Ellonen P，Almusa H，et al.Comparison of solution-based exome capture methods for next generation sequencing[J].Genome Biol，2011，12(9)：R94.

（责任编辑：冉明会）