NGS 分析流程性能及准确度评估 (v0.3)

Edited from 《edico 性能及准确度评估 (v0.2)》

大数据计算组　　马强华

2017 年 9 月 7 日

目录

1 目的

对比评估 NGS 分析流程的性能及其变异检测精度（snp & indel 突变类型）。

2 说明

2.1 数据

1. Zebra500 NA12878 38X
   - /hwfssz1/BIGDATA_COMPUTING/data/NA12878/Zebra500_NA12878_WGS/NA12878_read_1.fq.gz
   - /hwfssz1/BIGDATA_COMPUTING/data/NA12878/Zebra500_NA12878_WGS/NA12878_read_2.fq.gz

2.2 reference 版本

hg19

2.3 分析流程及所用计算资源

1. GATK3
   - 说明: 即 GATK Best Practices（GATK3.7）
   - 计算节点: cngb-hadoop-a16-4(10.53.20.164)
     CPU: Intel(R) Xeon(R) CPU E5-2650 v4 @ 2.20GHz, CPU 核数: 12, 逻辑 CPU: 48
     内存: 256G
2. GATK3_2
   - 说明: 使用 biobambam2 bamsormadup 作为排序去重工具, 其他步骤同 GATK Best Practices
   - 计算节点: cngb-compute-e03-9 (10.53.0.39)
     CPU: Intel(R) Xeon(R) CPU E7-4830 v4 @ 2.00GHz, CPU 核数: 14, 逻辑 CPU: 112
     内存: 512G
3. GATK4
   - 说明: beta 版本
   - 计算节点: 同 GATK3
4. GaeaHC ： Gaea + hadoop streaming GATK HC
   - 计算节点: 10.53.20.[12-32] 共 20 个计算节点
     CPU: Intel(R) Xeon(R) CPU E5-2650 v4 @ 2.20GHz, CPU 核数: 12, 逻辑 CPU: 48
     内存: 256G
5. edico

- 计算节点：cngb-edico-a23-1 (10.53.4.148)
  CPU: Intel(R) Xeon(R) CPU E5-2690 v4 @ 2.60GHz, CPU 核数: 14, 逻辑 CPU: 56
  内存: 256G

6. Sentieon

- 计算节点：同 GATK3

注 1： 各流程所使用变异检测算法均为 HaplotyperCaller，均不对 vcf 结果做额外处理（矫正过滤等）。

注 2： 存储均使用：华为 os9000 （/hwfssz1/BIGDATA_COMPUTING）

3 性能评估

3.1 各流程耗时

表 1: 各流程耗时

流程 / 步骤	GATK4	GATK3[1]	GATK3_2[2]	GaeaHC[3]	Sentieon	edico
比对	04:01:02	04:01:02	04:41:10	00:19:49	04:08:36	-
排序和去重	14:16:59	14:16:59	0	00:06:10	00:30:11	-
重比对	0	0	0	0	00:21:15	-
质量值矫正	05:57:02	02:38:17	02:41:46	00:12:56	01:51:44	-
PrintRead	05:52:13	15:26:21	26:20:33	0	0	-
变异检测	08:27:55	14:59:22	08:41:08	00:26:00	00:41:33	-
总计	38:59:33	41:22:02	42:24:37	01:05:00	07:33:19	00:36:00

[1] GATK3 的 BaseRecalibrator 处理的数据来源于 GATK4 的 MarkDuplicates 处理后得到的 bam 文件数据
[2] 该流程过滤耗时包含在比对步骤中
[3] 不包含解压步骤

4 变异检测精度评估

4.1 精度评估软件及指标说明

1. 评估软件

- RTG vcfeval

2. 参考集

- GIAB NIST3.3

3. 评估指标说明

- 真阳性位点（True positives）：在标准集中存在，test.vcf 中也存在的变异数。
- 假阴性位点（False negatives）：在标准集中存在，test.vcf 中不存在的变异数。
- 假阳性位点（False positives）：在标准集中不存在，test.cf 中存在的变异数。

- Precision (PPV)： (true positives) / (true positives + false positives)
- Recall (sensitivity)： (true positives) / (true positives + false negatives)
- F-measure： 2 * precision * recall / (precision + recall)

根据 F-measure（Precision 和 Recall 的调和平均数）的值来判断结果优劣，值越高越好。Threshold 是变异的 QUAL 值，作为结果最优情况下的过滤阈值。

4.2 snp 评估

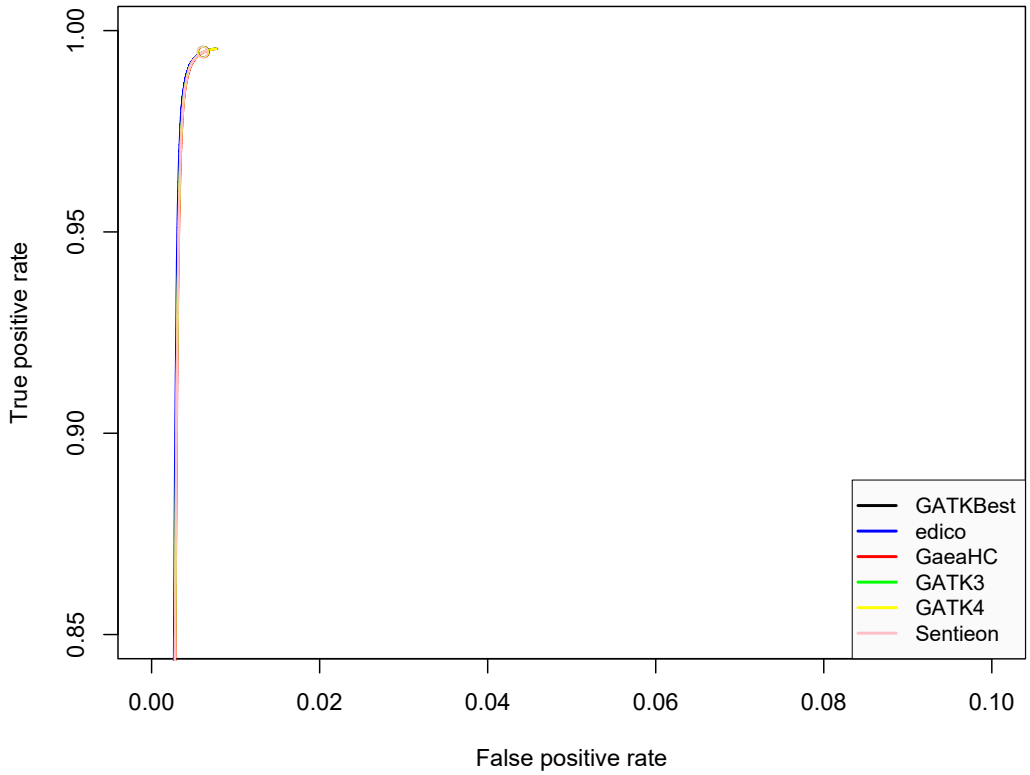
4.2.1 评估指标数值

表 2: 各流程评估指标

评估指标 / 流程	Threshold[1]	True-pos	False-pos	False-neg	Precision	Sensitivity	F-measure
GATK3	47.280	3176500	19675	16449	0.9938	0.9948	0.9943
GATK3_2	47.280	3176487	19672	16462	0.9938	0.9948	0.9943
GATK4	49.790	3176298	19676	16651	0.9938	0.9948	0.9943
GaeaHC	50.740	3176310	19902	16639	0.9938	0.9948	0.9943
edico	37.860	3176461	19765	16488	0.9938	0.9948	0.9943
Sentieon	50.740	3176222	19434	16727	0.9939	0.9948	0.9943

[1] 根据 QUAL 阈值，取 F-measure 最大时的值

4.2.2 snp 精度 ROC 曲线图



[1] 该 ROC 曲线根据 QUAL 值绘制
[2] 曲线上圆圈位置为 F-measure 最大时的 QUAL 值

4.2.3 snp 评估结论

各流程在 snp 上的结果 F-measure 均相同，其精度没有明显差异。

4.3 indel 评估

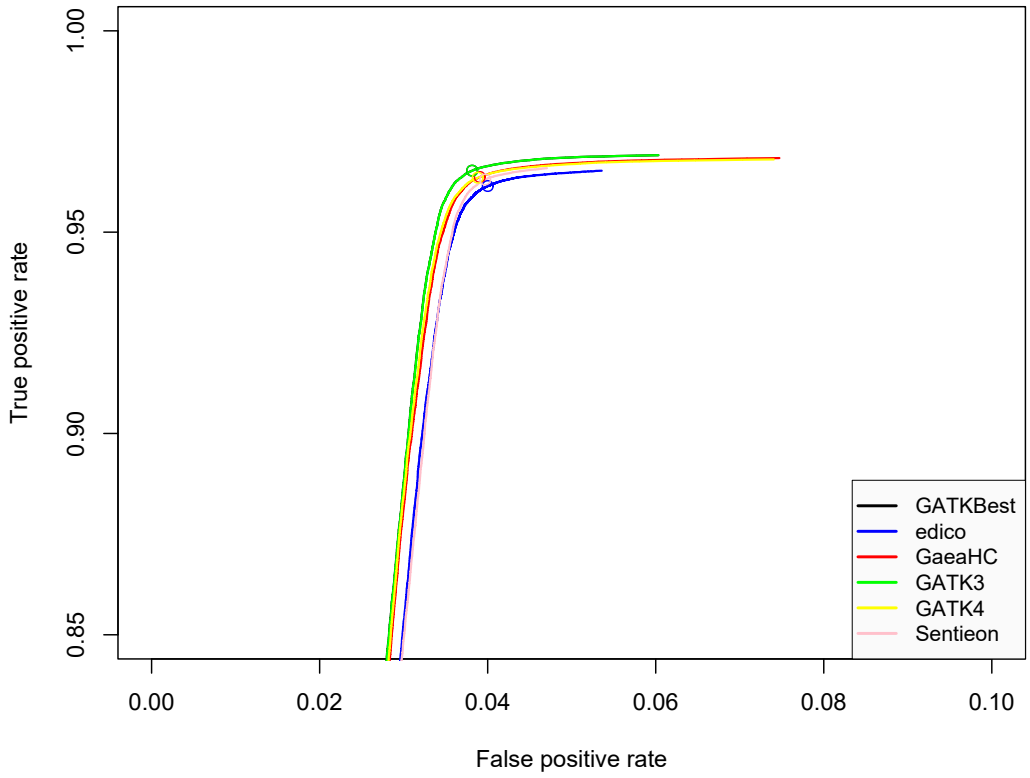
4.3.1 评估指标数值

表 3: 各流程评估指标

评估指标 / 流程	Threshold[1]	True-pos	False-pos	False-neg	Precision	Sensitivity	F-measure
GATK3	59.770	356685	14103	12794	0.9620	0.9654	0.9637
GATK3_2	59.770	356685	14100	12794	0.9620	0.9654	0.9637
GATK4	73.730	356042	14345	13437	0.9613	0.9636	0.9624
GaeaHC	73.730	356105	14437	13374	0.9610	0.9638	0.9624
edico	73.200	355273	14789	14206	0.9600	0.9616	0.9608
Sentieon	59.770	355874	14674	13605	0.9604	0.9632	0.9618

[1] 根据 QUAL 阈值，取 F-measure 最大时的值

4.3.2 indel 精度 ROC 曲线图



[1] 该 ROC 曲线根据 QUAL 值绘制
[2] 曲线上圆圈位置为 F-measure 最大时的 QUAL 值

4.3.3 indel 评估结论

GATK3 流程在 indel 检测上结果最优，edico 最差。

5 总结

edico 性能最好，用时只有 36 分钟。各流程在 snp 的检测上差异很微小，从 indel 结果看 GATK3 流程精度最优，edico 最差。总体看，结果相差并不很大，在可接受范围内。