Briefings in Bioinformatics, 00(00), 2020, 1–11

doi: 10.1093/bib/bbaa148
Method Review

Benchmarking variant callers in next-generation and third-generation sequencing analysis

Surui Pei, Tao Liu, Xue Ren, Weizhong Li, Chongjian Chen and Zhi Xie

Corresponding authors: Chongjian Chen, Annoroad Gene Technology (Beijing) Co., Ltd, Beijing 100176, China. Tel.: +86 18612465077. E-mail: cchen@annoroad.com; Zhi Xie, State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou 510060, China. Tel.: +86 2087335131. E-mail: xiezhi@gmail.com

Abstract

DNA variants represent an important source of genetic variations among individuals. Next- generation sequencing (NGS) is the most popular technology for genome-wide variant calling. Third-generation sequencing (TGS) has also recently been used in genetic studies. Although many variant callers are available, no single caller can call both types of variants on NGS or TGS data with high sensitivity and specificity. In this study, we systematically evaluated 11 variant callers on 12 NGS and TGS datasets. For germline variant calling, we tested DNAseq and DNAscope modes from Sentieon, HaplotypeCaller mode from GATK and WGS mode from DeepVariant. All the four callers had comparable performance on NGS data and 30× coverage of WGS data was recommended. For germline variant calling on TGS data, we tested DNAseq mode from Sentieon, HaplotypeCaller mode from GATK and PACBIO mode from DeepVariant. All the three callers had similar performance in SNP calling, while DeepVariant outperformed the others in InDel calling. TGS detected more variants than NGS, particularly in complex and repetitive regions. For somatic variant calling on NGS, we tested TNscope and TNseq modes from Sentieon, MuTect2 mode from GATK, NeuSomatic, VarScan2, and Strelka2. TNscope and Mutect2 outperformed the other callers. A higher proportion of tumor sample purity (from 10 to 20%) significantly increased the recall value of calling. Finally, computational costs of the callers were compared and Sentieon required the least computational cost. These results suggest that careful selection of a tool and parameters is needed for accurate SNP or InDel calling under different scenarios.

Key words: variant callers; germline variant; somatic variant

Introduction

DNA variants include single nucleotide variants (SNVs), small insertions and deletions (InDels), and structural variations (SVs), representing an important source of genetic variations among individuals [1]. Based on the cell type where variants occur, there are two types of variants, germline and somatic variants, which occur in germ cells and somatic cells, respectively [2].

Surui Pei (PhD) is a joint post-doctoral fellow in Zhongshan Ophthalmic Center at Sun Yat-sen University and Annoroad Gene Technology (Beijing) Co., Ltd. Her research area is high-throughput omics data analysis.
Tao Liu (PhD) is a bioinformatician in Annoroad Gene Technology (Beijing) Co., Ltd. His research area is high-throughput omics data analysis.
Xue Ren is a bioinformatician in Annoroad Gene Technology (Beijing) Co., Ltd. Her research area is high-throughput omics data analysis.
Weizhong Li (PhD) is a professor of Bioinformatics in Zhongshan School of Medicine at Sun Yat-sen University. He is interested in understanding and interpreting the relationships between genomic factors and disease phenotypes through computational approaches.
Chongjian Chen (PhD) is chief technology officer in Annoroad Gene Technology (Beijing) Co., Ltd. His research is focused on high-throughput omics data analysis and clinical laboratory techniques.
Zhi Xie (MD, PhD) is a professor of Bioinformatics in Zhongshan Ophthalmic Center at Sun Yat-sen University. He is interested in applying big data analytics in biology and medicine.
Submitted: 11 March 2020; **Received (in revised form):** 11 June 2020

Next-generation sequencing (NGS) is by far the most popular technology for genome-wide variant calling. Whole genome or whole exome sequencing based on NGS is routinely used to identify patient-specific germline variants in Mendelian diseases and somatic variants in cancer [3–5]. To date, the NGS technology has dominated by the Illumina-like NGS platforms, that are accurate and high-throughput, with relatively low cost, but the length of reads is less than 200 base-pair (bp). Many germline and somatic variant callers have been developed for Illumina-like NGS data due to their popularity (The following NGS refers to Illumina-like NGS). For germline variant calling, HaplotypeCaller in Genome Analysis Tool kit (GATK) is one of the most commonly used callers, which aligns NGS reads against a reference genome and calls SNVs and InDels simultaneously via local *de novo* assembly of haplotypes [6, 7]. DeepVariant is a recent and the first variant caller based on the deep convolutional neural network that was originally designed to call SNVs and InDels from NGS data [8]. Sentieon is a commercial variant caller that is designed as an accelerated software for GATK [9]. Sentieon has DNAseq mode that exactly matches the haplotype mode of GATK without down-sampling and DNAscope mode that has improved accuracy by machine learning models. For somatic variant calling, many callers have also been developed, including MuTect2 in GATK, TNseq mode and TNscope mode in Sentieon, NeuSomatic and Strelka2. MuTect2 is a somatic SNP and InDel caller that combines MuTect with the assembly based machinery of HaplotypeCaller [10]. TNscope combines haplotype-based variant calling for variant candidate calling and machine learning for variant filtration to improve the accuracy of variant calling. TNseq provides matching results to MuTect2, but without down-sampling for improved accuracy and consistency [11]. NeuSomatic is the first convolutional neural network approach for somatic variant calling, which summarizes sequence alignments into small matrices and incorporates many features to capture variant signals [12]. Strelka2 is a variant caller for the analysis of germline variants in small cohorts and somatic variant in tumor/normal sample pairs, which uses a Bayesian approach to represent continuous allele frequencies for both tumor and normal samples, while leveraging the expected genotype structure of the normal samples [13].

In addition to NGS, third-generation sequencing (TGS) technology has also been used in genetic studies in the last few years [14]. Two major TGS platforms, including single-molecule real-time (SMRT) technology from PacBio and Oxford Nanopore Technology, are characterized by long read length (10–100 kb) but with a high sequencing error rate (~15%). TGS has significant advantage in identifying SVs due to long read length compared to NGS. The recently developed circular consensus sequencing (CCS) mode from SMRT technology considerably improved the accuracy of sequences. CCS generates high fidelity (HiFi) reads to provide base-level resolution with less than 1% error rate for the calling of variant types from SNVs and InDels to SVs [15]. Recent studies have shown that SNV calling from TGS reads improved variant calling and provided independent validation to NGS reads, particularly generated high-confidence variant calls in repetitive regions of the genome, which is inaccessible to NGS [15]. Compared with NGS, there are a few software designed for SNP and InDel calling in TGS data. DeepVariant recently introduced a machine learning model particularly designed to call SNP and InDel for the CCS mode of SMRT long sequencing reads [16].

With many variant callers available, several benchmarking studies have been conducted. Hwang *et al.* [17] compared GATK, Samtools and Freebayes on datasets generated from different platforms using the same sample (HG001) and observed different biases toward specific types of SNP genotyping errors by different variant callers. Chen *et al.* [18] tested three variant callers, GATK, Strelka2 and Samtools on datasets generated from different platforms also using HG001. Bian *et al.* [19] tested five open-source somatic variant callers on four synthetic datasets and concluded that MuTect2 performed the best among the five callers. Despite the existence of these benchmarking studies, the new deep learning-based caller, DeepVariant and the most popular commercial variant caller, Sentieon, have not been included in most of the benchmarking studies. In addition, previous benchmarking studies of SNP and InDel callers focused on NGS data, but TGS data have been ignored.

In this study, we systematically evaluated four germline variant and six somatic variant callers on NGS datasets generated from the Illumina HiSeq sequencer as well as three germline variant callers on TGS datasets generated from the CCS mode of the PacBio Sequel II sequencer. We compared their performance using a number of evaluation metrics on 12 NGS and TGS datasets. In addition, the computing usages were also compared. Our systematic evaluation of variant callers will provide useful recommendations for their use under different scenarios.

Material and methods

Data source

All the datasets used in this study are summarized in Table 1. Germline variants were evaluated on NGS data (Illumina) and TGS data (PacBio). The sample HG001 was a B lymphocyte cell line from a woman of Utah's CEPH lineage. HG002, HG003 and HG004 were samples from a Jewish family, and HG005, HG006 and HG007 were samples from a China family [20]. For the NGS datasets, HG001 used for germline variant analysis was downloaded from European Nucleotide Archive (ENA; http://www.ebi.ac.uk/ena), and it was sequenced to a depth of approximately 50-fold coverage by the Illumina HiSeq 2000 sequencing platform. HG001 used for mixture data was downloaded from Genome in a Bottle (GIAB; ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/) and it was sequenced to a depth of 100-fold coverage by the Illumina HiSeq 2500 sequencing platform. The FASTQ data of HG002-HG007 were downloaded from GIAB, which were all sequenced by the Illumina HiSeq 2500 sequencing. The simulated sets of NGS data were generated from HG002 and HG005 using an in-house script at coverages of 2×, 5×, 10×, 15×, 30× and 50× for germline variant calling. The TGS data of HG001, HG002 and HG005 were generated by the CCS mode of PacBio Sequel II platform, obtained from the GIAB (ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/), which was aligned with hs37d5 BAM files for subsequent evaluation [21]. The true set (variant call format [VCF]) file of each sample was downloaded from GIAB. The link to HG001 is ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878_HG001/latest/GRCh37/. The link to HG002 is ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/AshkenazimTrio/HG002_NA24385_son/latest/GRCh37/. The link to the other samples is ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/.

As the true positive and negative variants are not possible to be obtained in tumor samples, a recent study suggested that *in-silico* mixtures worked nearly as well as *in-vitro* mixtures for benchmarking [11]. We therefore mixed germline datasets from the GIAB reference samples *in-silico* to create synthetic tumor samples to simulate true-positive somatic variants comprised of variants that were unique in one sample and absent from

Table 1. Summary of datasets used in this study

Data set name	Platform	Sample source	Note	Coverage
NGS001	Illumina HiSeq2000	HG001[a]	B lymphocytes cell line	50×
NGS002	Illumina HiSeq2500	HG002	Ashkenazim Trio son	30×
NGS003	Illumina HiSeq2500	HG003	Ashkenazim Trio father	30×
NGS004	Illumina HiSeq2500	HG004	Ashkenazim Trio mother	30×
NGS005	Illumina HiSeq2500	HG005	Chinese Trio son	30×
NGS006	Illumina HiSeq2500	HG006	Chinese Trio father	30×
NGS007	Illumina HiSeq2500	HG007	Chinese Trio mother	30×
TGS001	PacBio Sequel II (CCS)	HG001	B lymphocytes cell line	30×
TGS002	PacBio Sequel II (CCS)	HG002	Ashkenazim Trio son	32×
TGS005	PacBio Sequel II (CCS)	HG005	Chinese Trio son	30×
MIX010	Illumina HiSeq2500	10% HG001[b] with 90% HG002	Mixed data	100×
MIX020	Illumina HiSeq2500	20% HG001[b] with 80% HG0002	Mixed data	100×
MIX040	Illumina HiSeq2500	40% HG001[b] with 60% HG002	Mixed data	100×
MIX060	Illumina HiSeq2500	60% HG001[b] with 40% HG0002	Mixed data	100×

[a]HG001 (used for germline variant analysis) was downloaded from ENA (http://www.ebi.ac.uk/ena), and it was sequenced to a depth of approximately 50-fold coverage by the Illumina HiSeq 2000 sequencing platform.
[b]HG001 (used for mixture data) was downloaded from GIAB, and it was sequenced to a depth of 100-fold coverage by the Illumina HiSeq 2500 sequencing platform.

the other. The original data of two samples (HG001 and HG002) were generated by GIAB, which were sequenced by Illumina HiSeq 2500. The true set of *in-silico* tumor sample was the loci, which existed in high-confidence variant call sets of HG001 but not existed in all variant call sets of HG002. The depth of coverage in each *in-silico* tumor sample was 100× coverage. Four tumor samples with 10, 20, 40 and 60% tumor purity were mixed separately into HG002 to obtain *in-silico* mixture data of different target depths for the subsequent analysis. For example, 10× depth data from sample HG001 and 90× depth data from sample HG002 were mixed together to produce the 10% tumor sample. The same procedure was applied in other three tumor samples.

Data process

The FASTQ file was aligned with the human reference genome (hs37d5) using the BWA-MEM algorithm [22]. The BAM file was processed followed these steps: mark duplication, InDel realignment and base quality score recalibration (BQSR). All steps were performed according to standard instructions (see Supplementary Information).

Evaluation of variant callers

A total number of 11 callers used in the study (Table 2). For germline variant calling on the NGS datasets, we used DNAseq and DNAscope modes from Sentieon, HaplotypeCaller mode from GATK and WGS mode from DeepVariant. For germline variant calling on the TGS datasets, we used DNAseq mode from Sentieon, HaplotypeCaller mode from GATK and PACBIO mode from DeepVariant. To evaluate the influence of different sequencing depths on variation calling, DNAseq was used at different sequencing depths.

To detect somatic variants on the NGS datasets, we used TNscope and TNseq modes from Sentieon, MuTect2 mode from GATK, NeuSomatic, VarScan2 and Strelka2. The details of the mode and software are listed in Table 2.

Evaluation of variant calling accuracy was performed using RTG Tools (Real Time Genomics Tools) [23]. RTG Tools contain utilities to manipulate and compare multiple VCF files, as well as utilities for processing common NGS data formats. The SNP and InDel were evaluated separately. In the evaluation process, QUAL (quality) value was used to draw the precision-recall (PR) curve for germline variant, while AF (allele frequency) was selected to draw the PR curve for somatic variant.

We defined true positive (TP), true negative (TN), false positive (FP), false negative (FN) variants and F1 score as follows:

TP: variants called by a variant caller as the same genotype as the positive variants.

FP: variants called by a variant caller but not in the positive variants.

FN: positive variants that were not called by a variant caller.

Precision: TP/(TP + FP).

Recall: TP/(TP + FN).

F1 = 2∗(precision ∗ recall)/(precision + recall).

The vcfeval function in RTG Tools was used for evaluation. The input file was the VCF file of truth set (base set) and call set. The output files are used for ROC curve analysis.

Evaluation of SNPs and InDels in germline variant calling

We used the vcfeval function of RTG Tools to evaluate and compare the performance of different callers [23]. Hard filter is a method recommended to filter the raw calls emitted by the GATK callers and it has been shown to have an impact on the GATK performance [7]. We conducted hard filter on HG001 and checked the effect before and after hard filter. We also evaluated the effect of two parameters emit_conf and call_conf of Sentieon DNAseq mode, which determined the threshold of variant quality to emit or call a variant, respectively. The filter parameters were shown in the Supplementary Information. To evaluate the difference in accuracy and calling ability between NGS and TGS, the genome was divided into the GIAB high_conf region, the GIAB filtered region, and the outside GIAB region. GIAB high_conf region was a region that includes a subset of variant calls that are easier to detect defined by GIAB; the GIAB filtered region was the variable region that was included in GIAB, but was not included in the high-confidence region; and the outside GIAB region was other genomic regions except for the above two regions.

We also analyzed the influence of GC content and duplication rate on the effect of variant calling using DNAseq from DeepVariant. The difference in GC content between NGS and

Table 2. Software and mode for evaluation

Short name	Mode	Software	Version	Mutation type	Sequencing type
DNAseq_S	DNAseq	Sentieon	201808.05	Germline	NGS and TGS
DNAscope_S	DNAscope	Sentieon	201808.05	Germline	NGS
HC_GATK	HaplotypeCaller	GATK	4.0.7	Germline	NGS and TGS
WGS_DV	WGS	DeepVariant	0.8	Germline	NGS
PACBIO_DV	PACBIO	DeepVariant	0.8	Germline	TGS
TNscope_S	TNscope	Sentieon	201808.05	Somatic	NGS
TNseq_S	TNseq	Sentieon	201808.05	Somatic	NGS
Mutect2_GATK	Mutect2	GATK	4.0.7	Somatic	NGS
NeuSomatic	–	NeuSomatic	0.2.0	Somatic	NGS
VarScan2	–	VarScan2	2.3.9	Somatic	NGS
Strelka2	–	Strelka2	2.8.4	Somatic	NGS

TGS data in FP loci was calculated. Different GC content region files were downloaded from https://doi.org/10.1038/s41587-019-0054-x [24], and the number of variations in these regions was calculated using Bedtools. Firstly, overlapping regions in each BED file were merged, and then SNP and InDel loci information was transformed into BED format (one line for each locus). At last, intersect of Bedtools (v2.26.0) was used to calculate the number of overlapping regions between these two BED files. Meanwhile, the ratio of the number of variations in TGS data to that in NGS data was calculated. The dup95, dup99 files were downloaded from UCSC, Dup95 means repetitive regions in which the duplication rate was more than 95%; Dup99 means repetitive regions in which the duplication rate was more than 99%; dup_all and dup_gt10k files were downloaded from https://doi.org/10.1038/s41587-019-0054-x [24], of which Dup_all means all repetitive regions; Dup_gt10k means repetitive regions greater than 10 kb. The number of variations in these regions was calculated using Bedtools in NGS and TGS data.

Results

Overview of workflow

To evaluate the accuracy of variant callers for germline variant and somatic variant on NGS and TGS datasets, we designed and implemented different test schemes (Figure 1). We first detected variants and different procedures were used for the NGS and TGS datasets. Next, the sensitivity and specificity of each software were evaluated. In addition, the effects of different parameters on the calling of variation results were evaluated. Finally, the analysis results of the NGS and TGS data were compared.

Evaluation of germline variants from NGS data

DNAseq and DNAscope from Sentieon, HaplotypeCaller from GATK and WGS from DeepVariant were compared for germline variant calling on NGS data. All four callers showed high accuracy in both SNPs and InDels calling on seven datasets (Figure 2, see Supplementary Figure S1 and Supplementary Table S1 available online at https://academic.oup.com/bib). F1 scores of SNP calling were all above 0.99, and that of InDel calling were all above 0.98, which agreed with findings from a previous study [12]. While NGS001 had 50× coverage and the other six datasets had 30× coverage, the performance was indistinguishable among the datasets. We further evaluated the effects of different sequencing depths on variant calling. We down-sampled datasets with coverages of 2×, 5×, 10×, 15× and 30× from the NGS002 and NGS005 datasets (see Supplementary Figure S2 and Supplementary Table S2 available online at https://academic.oup.com/bib). At low sequencing depth of less than 15×, the precision did not change much, but the recall value was low. When the sequencing depth exceeded 15×, the improvement in F1 score was small. As the sequencing depth increased from 30× to 50×, F1 scores were similar. Therefore, the sequencing depth of 30× is recommended for germline variant calling.

Previously, hard filter have been suggested to filter variants on variant calling, which may have an influence on the accuracy of calling [7]. We also evaluated how the filtration process impacted on variant calling results. After filtering the SNPs using hard filter, the precision value increased from 0.9979 to 0.9991, the recall value decreased from 0.9985 to 0.9866, and the F1 score was slightly decreased. Similar results were obtained in InDel filtering (see Supplementary Figure S3A and B available online at https://academic.oup.com/bib). In addition, adjustment parameters of emit_conf and call_conf from 30 to 10, the precision value was also decreased while the recall value was increased, with unchanged F1 scores of SNPs and InDels (see Supplementary Figure S4A and B available online at https://academic.oup.com/bib). We also assessed the effect of BQSR on germline variant calling, and the results showed that BQSR had little effect on the F1 score (see Supplementary Figure S5A and B available online at https://academic.oup.com/bib).

Evaluation of germline variants from TGS

DNAseq mode of Sentieon, HaplotyperCaller mode of GATK, and PACBIO mode of DeepVariant were used to detect SNP and InDel from TGS data (Figure 3, see Supplementary Table S3 available online at https://academic.oup.com/bib). SNP calling results showed that F1 scores of the three software were all above 0.99, and the difference between the three software was marginal, which was consistent with a previous study [25]. However, InDel calling results were significantly different among the three software. The F1 scores of DeepVariant were the highest on all three datasets (0.9902, 0.9927, 0.9924 for TGS001, TGS002 and TGS005), followed by DNAseq mode of Sentieon (0.9433, 0.9390, 0.9393), whereas that of HaplotypeCaller from GATK (V4.0.7) were only 0.8437, 0.8223 and 0.8078, illustrating accurate and consistent performance of PACBIO DeepVariant in detecting germline variants on PacBio data. After filtering the SNPs using hard filter, the F1 score slightly increased (0.9979–0.9982); however, the F1 score of InDels significantly increased from 0.8437 to 0.9512 after filtering with hard filter (see Supplementary Figure S3C and D

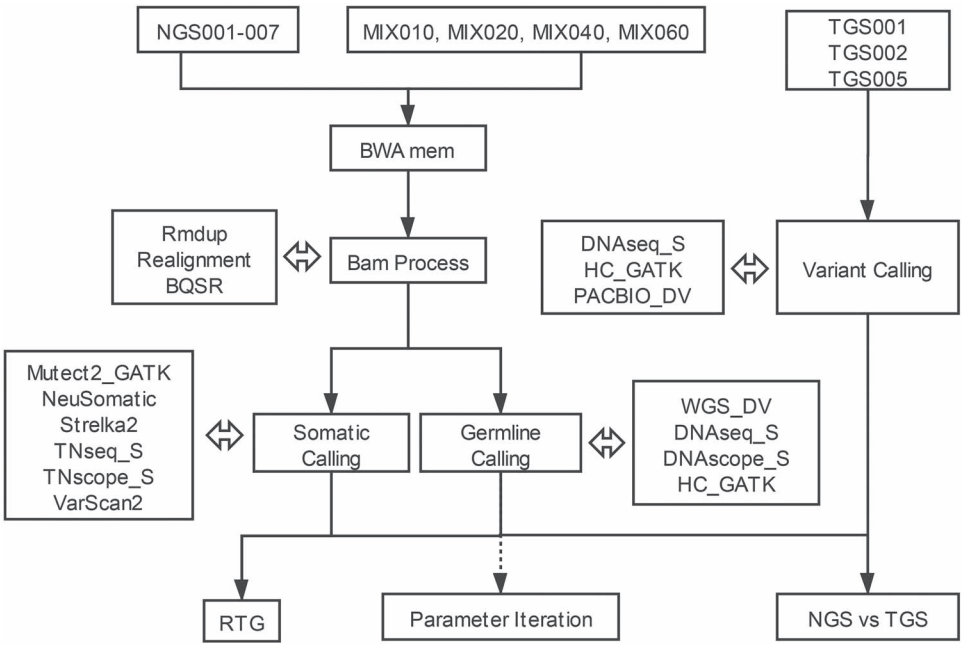

Figure 1. A schematic diagram of variant calling. The FASTQ data were aligned with the human reference genome (hs37d5) using the BWA-MEM algorithm. The BAM file from NGS data was processed followed these steps: mark duplication, InDel realignment and BQSR. And then germline and somatic variants were detected by different variant callers. The BAM file from TGS data was processed to detect germline variant by different variant callers. The sensitivity and specificity of each caller were evaluated using RTG tools. Finally, the detected variants the NGS and TGS data were compared.

available online at https://academic.oup.com/bib), suggesting its importance in InDel identification from TGS data.

Comparison of germline variant calling results between NGS and TGS data

Having demonstrated the consistent performance of DeepVariant on both NGS and TGS datasets, we next sought to compare the SNP and InDel loci of NGS and TGS data by DeepVariant. We found that the TGS data could obtain more variations than that of the NGS data for all three samples (Table 3). While the number of SNPs was 3 815 558 in NGS data, the number of SNPs was 3 898 442 on average in TGS data, yielding around 2.13% more than NGS. The number of InDels was 901 477 on average in TGS data, yielding around 3.89% more than that of NGS (866 423).

We further evaluated the differences of SNPs and InDels from the NGS and TGS datasets. We divided the genome into the GIAB high_conf region, the GIAB filtered region and the outside GIAB region (Method). It showed that more SNP and InDel loci could be detected in TGS data in all three regions (Table 3). In the high_conf region, the average number of SNPs was 3 101 630 in the TGS data, while the average number of SNPs was 3 098 137 in the NGS data, 0.11% more SNP loci could be detected in TGS data than that of in NGS data; however, the average number of SNPs in TP loci in TGS and NGS data was 3 097 639 and 3 096 786, respectively, an increase of only 0.03%. In InDels calling, 0.05% more InDel loci could be detected in TGS data (443 559) than that of in NGS data (443 344); however, the average number of InDels in TP loci in TGS data (440 313) decreased by 0.05% compared with the NGS data (440 509). In the GIAB_filter region, 12.66% more SNP loci could be detected in TGS data (679 042) than that of in NGS data (593 102), and 8.23% more InDel loci could be detected in TGS data (419 806) than that of in NGS data (385 276). In the outside region, 19.29% more SNP loci could be detected in TGS data (122 584) than that of in NGS data (98 938), and 28.74% more InDel loci could be detected in TGS data (36 715) than that of in NGS data (26 162) (Table 3).

We also divided the genome into different regions based on different GC contents. It also showed that the average number of SNP and InDel loci in high-GC-content and low-GC-content regions were higher in TGS data (Figure 4**A** and **B**). Although the number of variations detected in each sample was different, resulting in a larger SD, the overall trend was consistent across all samples. In the low-GC-content region (less than 20%), the average number of SNPs in TGS data (37 295) increased by 11.70% compared with the NGS data (33 388), and the average number of InDels in TGS data (28 649) increased by 8.82% compared with the NGS data (26 326). In the high-GC-content region (more than 60%), the average number of SNPs in TGS data (293 017) increased by 7.95% compared with the NGS data (271 432), and the average number of InDels in TGS data (73 219) increased by 18.45% compared with the NGS data (61 812).

In the highly repetitive region, the number of SNP and InDel was higher in TGS data, too (Figure 4**C** and **D**). In the dup_all region, the average number of SNPs in TGS data (300 425) increased by 21.79% compared with the NGS data (246 673), and the average number of InDels in TGS data (54 491) increased by 30.80% compared with the NGS data (41 659). With the increase in duplication rate, the number of SNP and InDel loci in TGS data increased more than that of in NGS data. In the dup99 region, 43.04% more SNP loci and 80.95% more InDel loci could be detected in TGS data than that of in NGS data. In the dup_gt10k region, 25.86% more SNP loci and 35.81% more InDel loci could be detected in TGS data than that of in NGS data.

The results found that specific SNP and InDel loci were detected only in TGS data. In order to check the reliability of SNP and InDel loci from TGS, we visualized 60 randomly selected loci using integrative genomics viewer (IGV). The visualization results showed that all the loci detected from TGS data were supported by more than five reads by IGV inspection

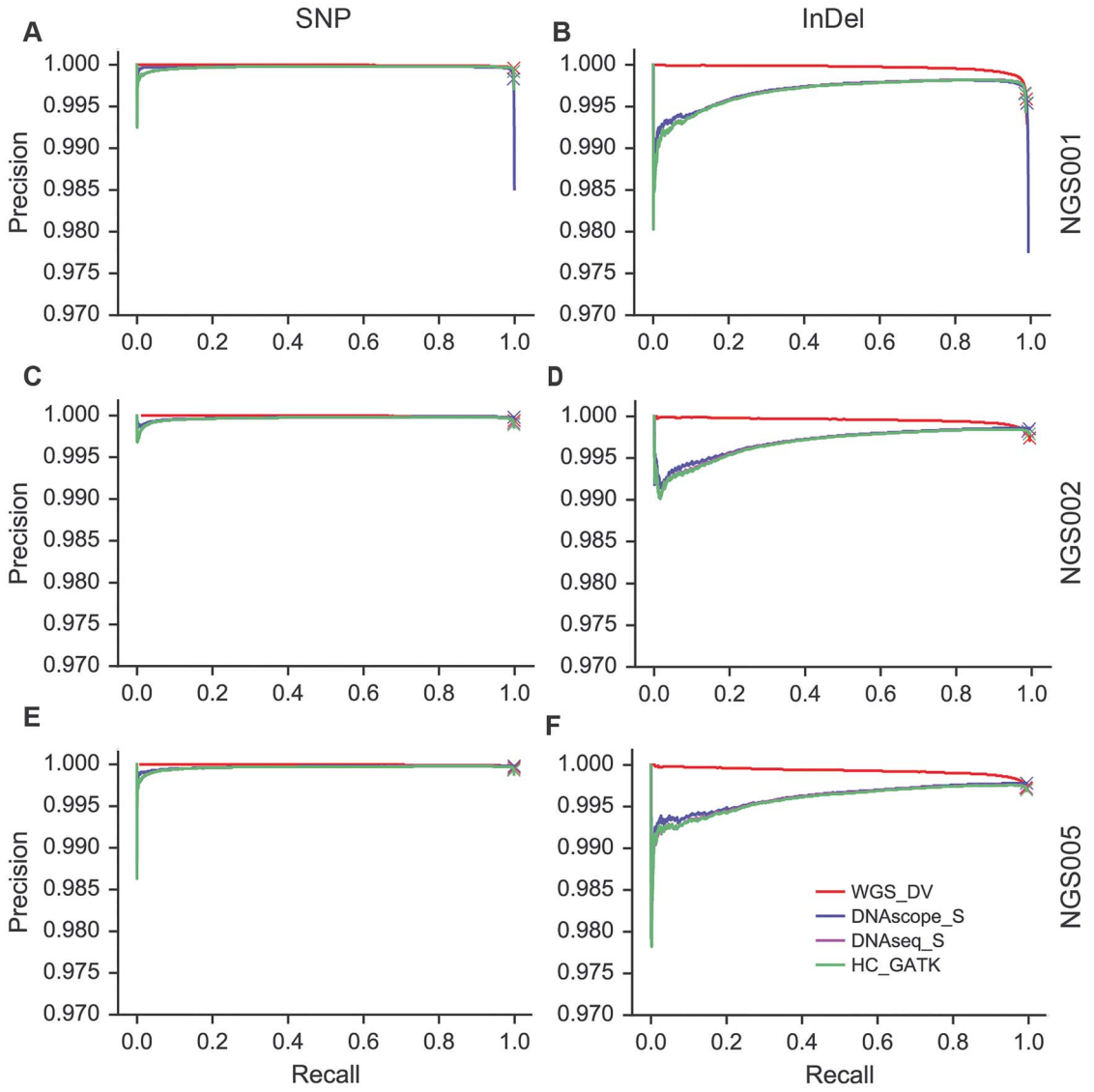

Figure 2. Precision-recall curves for germline variant calling on NGS datasets. 'WGS_DV' is the DeepVariant v0.8 variant caller with WGS mode; 'DNAscope_S' is Sentieon DNAscope variant caller; 'DNAseq_S' is Sentieon DNAseq variant caller; 'HC_GATK' is GATK HaplotypeCaller variant caller. 'X' marks the maximum F1-score for each caller. (**A–F**) SNPs and InDels in datasets NGS001, NGS002 and NGS005.

Table 3. Variation detection results using DeepVariant in different datasets and regions

Type	Data	PASS[d]	GIAB_highconf[e]	TP[f]	GIAB_filter[g]	Outside[h]
SNP	NGS[a]	3 815 558 (±19 386)	3 098 136 (±91.906)	3 096 786 (±93 320)	593 102 (±90 456)	98 937 (±93 318)
	TGS[b]	3 898 442 ± 17 873	3 101 630 (±95 163)	3 097 639 (±94 065)	679 042 (±66 886)	122 583 (±87 689)
	Ratio (%)[c]	2.13	0.11	0.03	12.66	19.29
InDel	NGS[a]	866 423 ± 21 372	443 344 (±48 325)	440 509 (±46 578)	385 276 (±45 259)	26 162 (±19 373)
	TGS[b]	901 476 ± 1468	443 559 (±48 148)	440 313 (±47 702)	419 806 (±27 181)	36 715 (±23 764)
	Ratio (%)[c]	3.89	0.05	−0.04	8.23	28.74

[a]NGS denotes the average number of NGS001, NGS002 and NGS005.
[b]TGS denotes the average number of TGS001, TGS002 and TGS005.
[c]Ratio (%) = (TGS − NGS)/NGS∗100%.
[d]PASS denotes the site number of PASS in the FILTER column in the VCF file.
[e]GIAB_highconf denotes PASS in the GIAB highconf region.
[f]TP denotes PASS in GIAB_highconf region consistent with the true set.
[g]GIAB_filter denotes PASS in the GIAB_filter region.
[h]Outside denotes PASS outside the GIAB_highconf and GIAB_filter region.

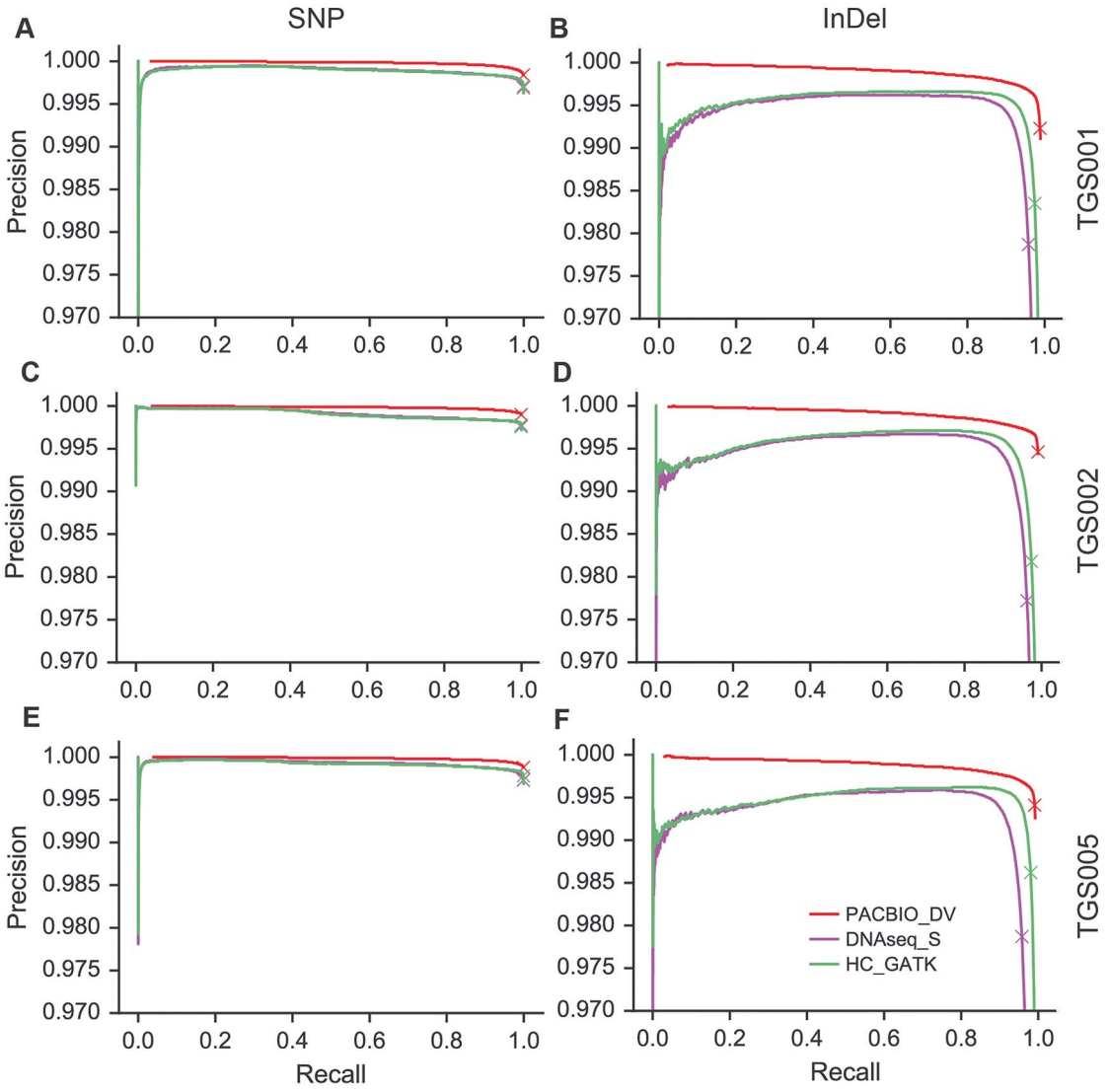

Figure 3. Precision-recall curves for germline variant calling on TGS datasets. 'PACBIO_DV' is the DeepVariant v0.8 variant caller with PACBIO mode; 'DNAseq_S' is Sentieon DNAseq variant caller; 'HG_GATK' is GATK HaplotypeCaller variant caller. 'X' marks the maximum F1-score for each caller. (**A**–**F**) SNPs and InDels in datasets TGS001, TGS002 and TGS005.

(see Supplementary Table S7 available online at https://academic.oup.com/bib), confirming reliability of variants detected from TGS datasets but missed by NGS (see Supplementary Figure S6 available online at https://academic.oup.com/bib).

Evaluation of somatic variants

To evaluate the effect of different tumor sample purities on somatic variant calling, we prepared four different *in-silico* mixture data, with 10, 20, 40 and 60% mixed ratios respectively (Material and methods). As expected, somatic variant calling from tumor sample with 20% tumor purities got a higher F1 score than 10% sample for all the callers. With the increase in tumor sample purity, the precision value did not change much, while the recall value increased considerably. F1 score of SNP calling significantly increased when the sample purity was increased from 10 to 20%, while the increase of F1 score was marginal when the sample purity was increased from 20 to 60%. For InDel calling, F1 score significantly increased when the sample purity was increased from 10 to 40%, while F1 scores lightly increased from 40 to 60% of tumor sample purities (Figure 5 and see Supplementary Table S4 available online at https://academic.oup.com/bib). All callers except TNseq showed high F1 score of SNPs when the tumor sample purity was 40 or 60%. When the sample purity was 20%, TNscope and Mutect2 had the highest F1 score (0.9799 and 0.9777, respectively), followed by NeuSomatic (0.9294). The other callers did not perform well. For InDel calling, when the sample purity was 40%, Mutect2 had the highest F1 score (0.9205), followed by TNscope (0.8545), and the other callers did not perform well (Figure 5 and see Supplementary Table S4 available online at https://academic.oup.com/bib).

We further studied the effect of BQSR on somatic variant calling. SNP calling was carried out on the data before and after BQSR, and the results showed that BQSR had little effect on F1 score (see Supplementary Figure S5C and D and Supplementary Table S5 available online at https://academic.oup.com/bib).

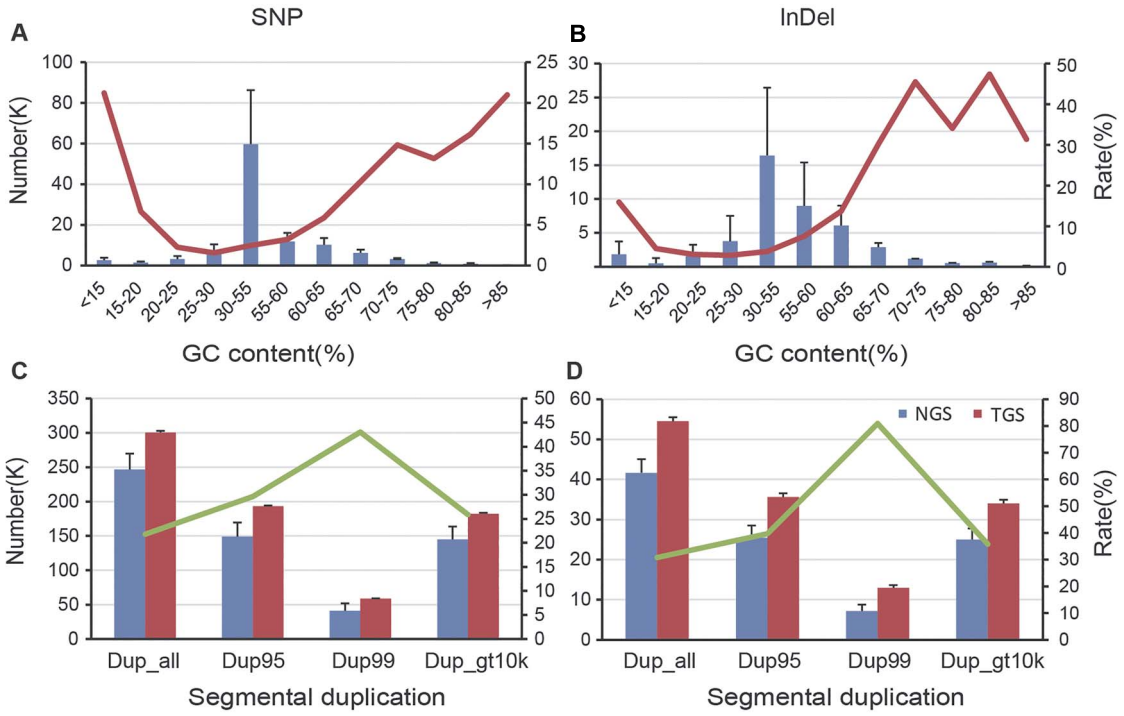

Figure 4. Comparison of the average number of variations in different GC content genomic regions and repetitive genomic regions in NGS and TGS data. The left axis (**A** and **B**) represents differences in the number of variations between TGS and NGS data; the left axis (**C** and **D**) represents the number of variations in TGS and NGS data, and the right axis (**A**–**D**) represents the proportional increase in the number of variations (TGS versus NGS data). NGS denotes the average number of NGS001, NGS002 and NGS005; TGS denotes the average number of TGS001, TGS002 and TGS005; Dup95 means repetitive regions in which duplication rate was more than 95%; Dup99 means repetitive regions in which duplication rate was more than 99%; Dup_all means all repetitive regions; Dup_gt10k means repetitive regions greater than 10 kb. (**A** and **B**) The difference values of SNPs and InDels between TGS and NGS in different GC contentgenomic regions. (**C** and **D**) The number of SNPs and InDels in different repetitive genomic regions for NGS and TGS data.

Evaluation of computational cost

Finally, the computational costs of different callers were compared. Four CPUs and 40 Gb of memory were allocated, and the computational costs of each caller were shown in Figure 6**A** and **B**. The computational resource consumption of Sentieon was minimal, and the running time was also the shortest among all callers. DeepVariant consumed more computational costs than GATK in TGS data, more efficient in NGS data.

In somatic variant calling, four CPUs were allocated to each caller, but the peak memory usage of each caller was varied, which was shown in Figure 6**C** and **D** and see Supplementary Table S6 available online at https://academic.oup.com/bib. CPU core hours of TNscope from Sentieon were the least among all callers, followed by Strelka2 and TNseq from Sentieon. In contrast, GATK and VarScan2 consumed the most CPU core hours (see Supplementary Table S6 available online at https://academic.oup.com/bib).

Discussion

In this study, we evaluated different variant callers for germline and somatic variant calling using NGS and TGS data. For germline variant calling on NGS data, F1 scores of Sentieon, GATK and DeepVariant were all above 0.99 with 30× coverage, indicating that researchers could obtain highly accurate and sensitive germline variant calling results using all three callers on NGS data. This result was consistent with a previous study [10]. F1 scores changed significantly with different sequencing depths, and the depth of 30× was recommended for a balance between cost and accuracy. For germline variant calling on TGS, DeepVariant had the highest F1 score (SNPs and InDels) of the three callers in TGS data. The possible reason is that DeepVariant uses a deep learning model, which does not assume any specific distribution, so it works also efficiently on TGS data. In contrast, GATK and Senteion used the model that was designed for NGS data, so they did not work well on TGS data.

We also compared germline variant calling on both NGS and TGS data. The number of SNP and InDel in TGS data was higher in high-GC-content regions and low-GC-content regions than in NGS data, which showed that single-molecule sequencing without PCR amplification could better solve the problem of GC bias. Similar results have been achieved in highly repetitive regions, indicating that the long reads of TGS have more advantages for variant calling in highly repetitive regions. At the same time, we also found that some unidentified variation sites in GIAB were detected in TGS data, indicating that some sites in the true set of GIAB need to be modified.

Somatic variant calling of the *in-silico* mixture data with 10% tumor cells showed that while the precision score could be above 0.99, the recall value was only about 0.5, with F1 scores were below 0.9. This indicated that a portion of somatic variant loci were not detected because of insufficient coverage data when the mix proportion of tumor cells was low (10%). Therefore, the recall value of the loci should be improved by increasing the sequencing depth in the case of low tumor purity. Indeed, F1 scores of SNP calling using Mutect2 and TNscope increased to 0.97 in 20% tumor purity sample. With the increase of tumor purity to 40 and 60%, all callers except TNseq showed high F1 score of SNPs. For InDel calling, Mutect2 and TNscope had the

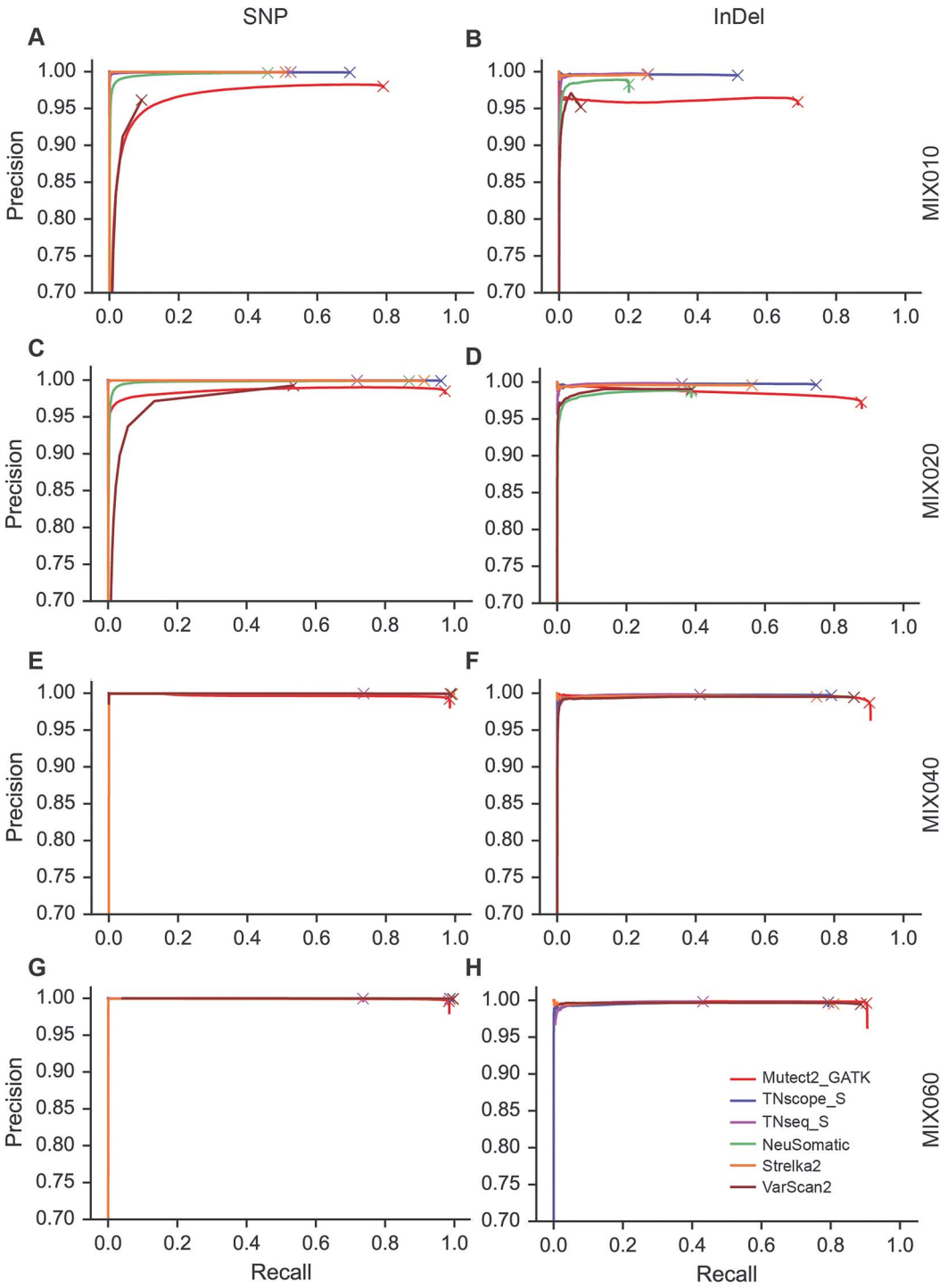

Figure 5. Precision-recall curves for somatic variant calling NGS datasets with different mixed tumor ratios. 'Mutect2_GATK' is the GATK Mutect2 variant caller; 'TNscope_S' is Sentieon TNscope variant caller; 'TNseq_S' is Sentieon TNseq variant caller; 'NeuSomatic' is NeuSomatic variant caller; 'Strelka2' is Strelka2 variant caller; 'VarScan2' is VarScan2 variant caller. 'X' marks the maximum F1-score for each caller. (**A** and **B**) SNPs and InDels in dataset MIX010. (**C** and **D**) SNPs and InDels in dataset MIX020. (**E** and **F**) SNPs and InDels in dataset MIX040. (**G** and **H**) SNPs and InDels in dataset MIX060.

highest F1 score when the sample purity was 20%, while Mutect2 and VarScan2 had the highest F1 score when the sample purity was 40 and 60%. Therefore, the higher tumor sample purity, the better the accuracy of SNP and InDel calling.

We also compared the computational resource consumption and time cost of all the callers. As an acceleration software, Sentieon reduced computing resource consumption and shortened the computation time without compromising the accuracy of

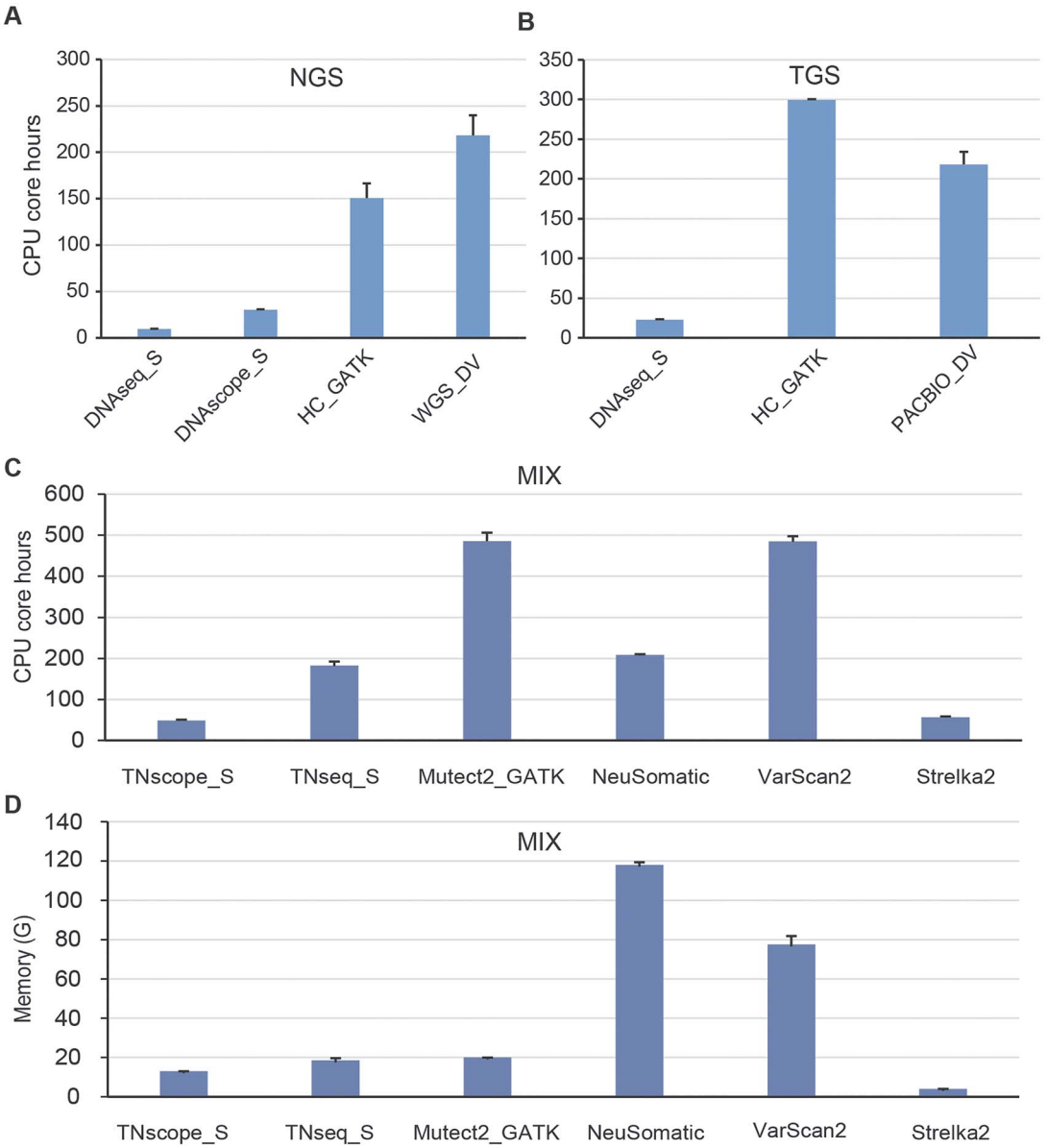

Figure 6. Computational costs of different callers. (**A** and **B**) CPU core hours of different callers for germline variant calling using NGS data and TGS data. Jobs were run on node with 4 CPUs and 40 Gb of memory. (**C**) CPU core hours of different callers for somatic variant calling using simulated NGS data. (**D**) Peak memory of different callers for somatic variant calling using simulated NGS data.

the calling. Therefore, Sentieon is a choice if the speed is an important factor to be considered and the cost of the software is affordable. Otherwise, both DeepVairant and GATK are alternatives without decreasing accuracy compared to Sentieon. The result of TGS showed that Sentieon has the minimum resources consumption, while DeepVariant has the highest accuracy, and researchers can choose a caller based on actual needs in TGS data analysis.

To sum, by systematically evaluating performance of different callers, our study suggested that careful selection of callers, analysis parameters and sequencing platforms is required for reliable variant calling under different scenarios.

Key Points

- We systematically evaluate four germline variant and six somatic variant callers on NGS datasets and three germline variant callers on TGS datasets.
- Four germline variant callers have comparable performance on NGS datasets, where 30× coverage of WGS data is recommended for a balance of accuracy and cost.
- Three germline variant callers have similar performance on TGS datasets for SNP calling while DeepVariant outperformed the others for InDel calling.

More variants can be detected on TGS than NGS, particularly in complex and repetitive regions.
- TNscope in Sentieon and Mutect2 in GATK outperformed the other somatic variant callers. The higher tumor sample purity, the better the accuracy of SNP and InDel calling.
- Careful selection of a tool and parameters is required for accurate calling of SNPs and InDels under different scenarios.

Supplementary data

Supplementary data are available online at https://academic.oup.com/bib.

Acknowledgements

We thank useful suggestions from Dr Hongwei Wang and Dr Zhikun Wu. We also thank the support from Center for Precision Medicine, Sun Yat-sen University.

Funding

National Natural Science Foundation of China (31829002); National Key Research and Development Program of China (2019YFA0904400, 2016YFC0901604 to Z.X.).

Abbreviations

NGS, next-generation sequencing; TGS, third-generation sequencing; HiFi reads, highly accurate long reads; CCS, circular consensus sequencing; GATK, genome analysis tool kit; SNP, single-nucleotide polymorphisms; InDel, insertions and deletions; GIAB, genome in a bottle; BQSR, base quality score recalibration; NIST, National Institute of Standards and Technology; EMBL-EBI, European Bioinformatics Institute; RTG Tools, real time genomics tools; TP, true positive; FP, false positive; FN, false negative; PR curve, precision recall curves; RPRS, ReadPosRankSum; IGV, integrative genomics viewer.

References

1. Hofmann AL, Behr J, Singer J, *et al*. Detailed simulation of cancer exome sequencing data reveals differences and common limitations of variant callers. *BMC Bioinformatics* 2017;**18**:8.
2. Griffiths AJ, Miller JH, Suzuki DT, *et al*. Inheritance of organelle genes and mutations. In: *An Introduction to Genetic Analysis*, 7th edn. NewYork: WH Freeman, 2000.
3. Pereira PCB, Melo FM, De Marco LAC, *et al*. Whole-exome sequencing as a diagnostic tool for distal renal tubular acidosis. *J Pediatr (Versao em Portugues)* 2015;**91**:583–9.
4. Renkema KY, Stokman MF, Giles RH, *et al*. Next-generation sequencing for research and diagnostics in kidney disease. *Nat Rev Nephrol* 2014;**10**:433–44.
5. Kroigard AB, Thomassen M, Laenkholm AV, *et al*. Evaluation of nine somatic variant callers for detection of somatic mutations in exome and targeted deep sequencing data. *PLoS One* 2016;**11**:e0151664.
6. Warden CD, Adamson AW, Neuhausen SL, *et al*. Detailed comparison of two popular variant calling packages for exome and targeted exon studies. *PeerJ* 2014;**2**:e600.
7. Van der Auwera GA, Carneiro MO, Hartl C, *et al*. From FastQ data to high confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr Protoc Bioinformatics* 2013;**43**:11 10 11–33.
8. Poplin R, Chang P-C, Alexander D, *et al*. A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol* 2018;**36**:983–7.
9. Freed DN, Aldana R, Weber JA, *et al*. The Sentieon genomics tools—a fast and accurate solution to variant calling from next-generation sequence data. *BioRxiv* 2017;115717.
10. Cibulskis K, Lawrence MS, Carter SL, *et al*. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* 2013;**31**:213–9.
11. Freed D, Pan R, Aldana R. TNscope: accurate detection of somatic mutations with haplotype-based variant candidate detection and machine learning filtering. *bioRxiv* 2018; 250647.
12. Sahraeian SME, Liu R, Lau B, *et al*. Deep convolutional neural networks for accurate somatic mutation detection. *Nat Commun* 2019;**10**:1041.
13. Kim S, Scheffler K, Halpern AL, *et al*. Strelka22: fast and accurate calling of germline and somatic variants. *Nat Methods* 2018;**15**:591–4.
14. Mitsuhashi S, Matsumoto N. Long-read sequencing for rare human genetic diseases. *J Hum Genet* 2019;**65**:1.
15. Edge P, Bansal V. Longshot enables accurate variant calling in diploid genomes from single-molecule long read sequencing. *Nat Commun* 2019;**10**:1–10.
16. Wenger AM, Peluso P, Rowell WJ, *et al*. Highly-accurate long-read sequencing improves variant detection and assembly of a human genome. *BioRxiv* 2019;519025.
17. Hwang S, Kim E, Lee I, *et al*. Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Sci Rep* 2015;**5**:17875.
18. Chen J, Li X, Zhong H, *et al*. Systematic comparison of germline variant calling pipelines cross multiple next-generation sequencers. *Sci Rep* 2019;**9**:9345.
19. Bian X, Zhu B, Wang M, *et al*. Comparing the performance of selected variant callers using synthetic data and genome segmentation. *BMC bioinformatics* 2018;**19**:1–11.
20. Zook JM, Catoe D, McDaniel J, *et al*. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data* 2016;**3**:160025.
21. Auton A, Abecasis GR, Altshuler D, *et al*. A global reference for human genetic variation. *Nature* 2015;**526**:68–74.
22. Vasimuddin M, Misra S, Li H, *et al*. Efficient architecture-aware acceleration of BWA-MEM for multicore systems. In: *IEEE International Parallel and Distributed Processing Symposium (IPDPS): 2019 IEEE*, Rio de Janeiro, Brazil, 2019, 314–24.
23. Cleary JG, Braithwaite R, Gaastra K, *et al*. Comparing variant call files for performance benchmarking of next-generation sequencing variant calling pipelines. *BioRxiv* 2015; 023754.
24. Krusche P, Trigg L, Boutros PC, *et al*. Best practices for benchmarking germline small-variant calls in human genomes. *Nat Biotechnol* 2019;**37**:555–60.
25. Wenger AM, Peluso P, Rowell WJ, *et al*. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* 2019;**37**:1155–62.