Accelerated somatic mutation calling for whole-genome and whole-exome sequencing data from heterogenous tumor samples

Shuangxi Ji, Wenyi Wang, Tong Zhu, et al.

Genome Res. published online April 8, 2024
Access the most recent version at doi:10.1101/gr.278456.123

Accelerated somatic mutation calling for whole-genome and whole-exome sequencing data from heterogenous tumor samples

Shuangxi Ji[1], Tong Zhu[2], Ankit Sethia[2], Wenyi Wang[1*]

[1]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA.

[2]NVIDIA Corporation, Santa Clara, CA, USA.

*Correspondence to: Wenyi Wang, wwang7@mdanderson.org

Running title: MuSE with multi-step parallelization

Abstract

Accurate detection of somatic mutations in DNA sequencing data is a fundamental prerequisite for cancer research. Previous analytical challenge was overcome by consensus mutation calling from four to five popular callers. This, however, increases the already nontrivial computing time from individual callers. Here, we launch MuSE 2, powered by multi-step parallelization and efficient memory allocation, to resolve the computing time bottleneck. MuSE 2 speeds up 50 times than MuSE 1 and 8-80 times than other popular callers. Our benchmark study suggests combining MuSE 2 and the recently accelerated Strelka2 achieves high efficiency and accuracy in analyzing large cancer genomic datasets.

Introduction

Cancer arises and evolves by accumulating various types of genetic alterations, such as single nucleotide variation (SNV), copy number alteration (CNA) and structural variation (SV). The high-throughput sequencing (HTS) technology has revolutionized the way we look at many human diseases, particularly cancer. With its constantly improved capacity and reduced cost, HTS is enabling investigations of genetic alterations within large human patient cohorts, hence advancing both basic and translational cancer research. Many computational tools have been developed for calling somatic variants (Xu 2018), which typically require, as input, whole-genome sequencing (WGS) or whole-exome sequencing (WES) data from the tumor tissue, as well as from the blood of the patient to serve as the germline control. WGS provides the most comprehensive coverage to sequence both protein-coding and non-coding regions across the entire genome; whereas WES provides an efficient alternative to WGS by targeting only protein-coding regions that accounts for 1-2% of the genome (Alfares et al. 2018), hence achieving both higher read depth (Sun et al. 2021; Barbitoff et al. 2020) and lower sequencing cost.

We previously launched MuSE 1 (Fan et al. 2016), a statistical approach for somatic mutation calling, where we introduced a combination of nucleotide base-specific Markov substitution model for molecular evolution and a tumor sample-specific error model to estimate tier-based cutoffs for selecting SNVs. Due to its high sensitivity and specificity, especially for calling subclonal SNVs, MuSE 1 was adopted in multiple pipelines, including as a major contributing caller to reach final consensus calls by the Cancer Genome Atlas (TCGA) PanCanAtlas project (Ellrott et al. 2018) across ~13,000 tumor samples, and the International Cancer Genome Consortium Pan-Cancer Analysis of Whole Genomes (ICGC-PCAWG) initiative (The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium 2020) across ~2,700 tumor samples.

One major limitation of MuSE 1, like many other mutation callers (Larson et al. 2012; Cibulskis et al. 2013; Koboldt et al. 2012), is the computational speed. It takes 2-3 days to finish running

the WGS data of a tumor-normal pair on a typical Linux server with an Intel Xeon processor and more than 100 gigabytes (GB) random access memory (RAM), which explains the commonly seen long wait-times for completing mutation calling before any downstream analysis in large patient cohort studies. Here, we present MuSE 2, which maintains the same input, output and mathematical model as MuSE 1, but accelerates significantly for both WES and WGS data by adopting a new algorithmic programming backbone. MuSE 2 employs a multithreaded producer-consumer model and the OpenMP library for parallel computing, including parsing and uncompressing reads from binary sequence alignment/map formatted (BAM) files, detecting and filtering variants, and writing output. It is also optimized by adopting a more efficient memory allocator. In this paper, we have benchmarked the accuracy of MuSE 2 against three somatic mutation callers, i.e., MuTect2 (Cibulskis et al. 2013), SomaticSniper (Larson et al. 2012) and VarScan2 (Koboldt et al. 2012), which are the other highlighted somatic mutation callers in the National Cancer Institute Genomic Data Commons (GDC) DNA-seq analysis pipeline: https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/DNA_Seq_Variant_Calling_Pipeline/, as well as a recently accelerated mutation caller Strelka2 (Kim et al. 2018). We use the consensus mutation calls generated by previous consortial studies with 3-5 un-accelerated callers (Ellrott et al. 2018; The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium 2020; Craig et al. 2016). Here we demonstrate the improved utility of our new caller using WES data generated from 5 tumor-normal pairs and WGS data generated from 7 tumor-normal pairs.

Results

Overview of approach. MuSE 2 takes as input the indexed reference genome FASTA file, the BAM format sequencing data from a pair of tumor-normal tissues (**Supplemental Fig**. **S1**) and the dbSNP (Sherry et al. 2001) variant call format (VCF) file, which is bgzip compressed, tabix (Li 2011) indexed using the same reference genome. Unlike MuSE 1, which can only utilize one

core, MuSE 2 takes advantage of the multi-core resources in a modern computer or a computing node for somatic SNV calling from WES/WGS data (**Fig. 1A-B**).

Since our benchmarking study requires a large amount of computational resources to cover multiple callers and scenarios, we only include results for WES data from 5 tumor-normal pairs and WGS data from 5 tumor-normal pairs, which are randomly selected and downloaded from the GDC (Genomic Data Commons) data portal and the ICGC data portal, respectively. The sequencing depths from these samples reflect the wide ranges presented in both datasets (**Supplemental Table S1**, **Supplemental Fig. S2**). We further include WGS data from 2 tumor-normal pairs to evaluate mutation calling performances on newer sequencing platforms.

We compare the SNV entries in the output VCF files generated by MuSE 2 with those by MuSE 1 for each patient sample with the same or different number of CPU cores. Since each SNV entry is denoted by one line of string in a VCF file, we compare the strings from both methods line by line. The result shows that all the entries from the two methods are identical with the same number or different number of cores (**Supplemental Fig. S3**).

Accuracy benchmarking for real tumor samples. We evaluate the performance of MuSE 2 and compare to other callers using the consensus SNV calls from TCGA (for WES data) and PCAWG (for WGS data) as truth sets. The truth sets include 168-2,553 somatic SNVs (mean=1,394, median=932) for WES and 3,813-19,081 somatic SNVs (mean=10,146, median=8,073) for WGS data. We first compared among callers whose predecessors contributed to the consensus call: MuSE 2, MuTect2, Somatic Sniper, VarScan2 for TCGA WES, MuSE 2 and MuTect2 for PCAWG WGS. We divided the mutation positions into multiple bins defined by variant allele frequency (VAF), or sequencing read depth, by classes of variant effects or clonality (**Methods**), and calculated precision, recall and F1 score, i.e., the harmonic mean of precision and recall for each bin. For both WES and WGS data, MuSE 2 achieves a higher precision at a similar or higher recall, hence a higher F1 score (**Supplemental Fig. S4A** for TCGA**, Supplemental Fig. S4B** for PCAWG) across all bins of VAF and read depth, and

across variant classes. It also achieves a higher recall for subclonal consensus SNV calls from PCAWG WGS (Dentro et al. 2021, **Methods**). We then compared the performance between MuSE 2 and Strelka2 (Kim et al. 2018), which used machine learning and curated data to train a position-specific error score, and was developed after the release of PCAWG consensus calls. Note that MuSE 2 utilizes the same Bayesian Markov model as MuSE 1 to explicitly define an evolutionary process and estimate model-based parameters based on data. Overall, Strelka2 performs well in both WES and WGS data. Compared to MuSE 2, its performance is lower in WES (**Fig. 2**) across all bins of VAF, read depth and different variant class. However, in the case of WGS data, it either matches or surpasses MuSE 2 in precision, at a similar recall rate (**Fig. 3A**). On the other hand, Strelka2 still shows a slightly lower recall than MuSE 2 at positions with low VAF (<0.2) or low read depth (<20×).

Accuracy benchmarking for tumor cell line data. As the TCGA and PCAWG data used above were generated by Illumina HiSeq 2000 more than a decade ago, we further obtained a set of tumor cell line sequencing data that were generated using newer sequencing platforms to compare the caller performances. A cell line tumor-normal pair COLO829/COLO829BL was used to generate WGS data on HiSeq X Ten, called COLO829 Illumina, and NovaSeq, called COLO829 10X, and the consensus calls from three mutation callers were generated through a multi-institutional effort (Craig et al. 2016, **Methods**). Although the sequencing data is newer, the consensus call effort was relatively old and the three callers did not entirely match what was established by the TCGA and PCAWG projects. They include MuTect2's predecessor MuTect (Cibulskis et al. 2013), Strelka2's predecessor Strelka (Saunders et al. 2012) and Seurat (Christoforides et al. 2013). We therefore implemented two strategies to benchmark the accuracy of MuSE 2 and Strelka2. First, we took the consensus calls as the truth set, which includes 35,543 SNVs from two pairs of samples, COLO829 Illumina and COLO829 10X. Second, we put aside the consensus calls and instead used the mutation calls made in the tumor cell line (100%) as the truth set. This truth set then includes 45,853 SNVs from MuSE 2

and 44,257 SNVs from Strelka2, respectively called from COLO829 Illumina. We then evaluated how many of these initial calls were recovered in a *in silico* diluted dataset where the tumor cell proportion decreases to 75%, 50%, 25%, 20% and 10% (**Methods**). MuSE 2 and Strelka2 did equally well with both truth sets. MuSE 2 presented a slightly higher recall than Strelka2 at positions with low VAF (<0.2) or in samples with low tumor cell proportion (≤10%) (**Fig. 3B, C**). We note that this slight advantage in MuSE 2 is consistently observed across all WGS data, whereas an advantage of higher precision in MuSE 2 is consistently observed in WES data.

Speed benchmarking. We compare the speed of running MuSE 2 against MuSE 1, MuTect2, SomaticSniper, VarScan2 and Strelka2 on a computing cluster. Each method is tested with the number of CPU cores at 1, 5, 10, 20, 28, 40 and 80. All methods are assigned with the same RAM of 50GB for WES data and 150GB for WGS data. The time cost of each method for each pair of data is shown in **Fig. 4A**. Except for COLO829 10X, both MuSE 2 and Strelka2 continue to gain computational advantages with increasing number of CPU cores, while the other four methods do not. We examine the overall speed performances of these methods with MuSE 2 at 80 cores, Strelka2 at 80 scores, and the average time cost across multiple runs over the different numbers of cores except for core=1 (where the computing resource is too limited) for the other methods (**Fig. 4B**). Both MuSE 2 and Strelka2 achieve much faster SNV calling compared to all other methods. For WES data, MuSE 2 accelerates 28-58 times (mean=44) than MuSE 1, 68-83 times (mean=77) than MuTect2, 5-8 times (mean=8) than SomaticSniper, 33-39 times (mean=36) than VarScan2. Similarly, for WGS data, it accelerates 48-59 times (mean=57) than MuSE 1, 33-44 times (mean=41) than MuTect2, 7-8 times (mean=8) than SomaticSniper, 33-43 times (mean=37) than VarScan2 for WGS data. On the other hand, Strelka2 is faster than MuSE 2 for all the WES data and the WGS data except for COLO829 10X: it is about two-fold speedup on average compared to MuSE 2. For COLO829

10X, however, Strelka2 stopped gaining speed after 5 CPU cores, while MuSE 2 continued accelerating; its computing time is 20 times of MuSE 2 at the number of cores of 80.

Overall, MuSE 2 and Strelka2 are the two top accelerated methods compared to others in the above benchmarking. We further examine the difference between the SNV calls reported by them for the same patient sample (**Fig. 4C**). For WES data, 46-78% (mean=66%) of the calls are identified by both; 2-16% (mean=7%) of the calls are unique to MuSE 2, and 13-51% (mean=27%) of the calls are unique to Strelka2. For WGS data, 41-77% (mean=62%) of the calls are identified by both; 6-18% (mean=11%) of the calls are unique to MuSE 2, and 13-45% (mean=26%) of the calls are unique to Strelka2.

We further investigate the feasibility of using the intersect calls from these two methods to reproduce the consensus calls for these data generated by previous studies (Ellrott et al. 2018; The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium 2020; Craig et al. 2016). For WES data, these calls notably improve the F1 scores of Strelka2 calls, while maintaining a comparable F1 score with the MuSE 2 calls (**Fig. 4D** and **Supplemental Table S2**). This suggests that running MuSE 2 alone might be sufficient for WES data when the computing resource is limited. For WGS data, on the contrary, Strelka2 calls (0.76-0.92) have higher F1 scores than MuSE 2 calls (0.63-0.88) in all the patient samples; while the intersection calls outperform the two individual callers and reach the highest F1 scores (0.91-0.96). Also, the intersect calls achieve the highest precision values (0.92-0.96 for WES, 0.87-0.95 for WGS) for all the data benchmarked, despite the differences of read depths and sequencing platform they were generated (**Supplemental Table S1**). The intersect calls maintain good recall values at 0.74-0.89 (median =0.86) for WES data, and at 0.96-0.98 (median=0.96) for WGS data. The recall values of either the intersect call sets, or the individual call sets from the two methods are consistently higher for WGS data than WES (**Fig. 4D**, all results from the WGS data fall in the top rectangle). Also, MuSE 2 call sets achieve higher precisions and F1 scores, lower recalls for WES, but higher recalls, lower precisions and F1 scores for WGS data, when compared to the

calls from Strelka2. Since the F1 scores of intersect calls are lower in WES than in WGS, we further investigated whether the inclusion of a third caller could improve the accuracy for WES. This led to an interesting observation that adding VarScan2 or SomaticSniper as a third caller did not improve precision, recall or F1, while adding MuTect2 improved recall but decreased precision and F1 (**Supplemental Table S3**). In the project of Multi-Center Mutation Calling in Multiple Cancers (MC3) organized by TCGA, the latter strategy was further expanded to include five callers, maximizing recall, and followed by a post-filtering pipeline to further improve precision, independent of any callers (Ellrott et al. 2018). This post-filtering pipeline removes potentially false positive variants that are caused by germline contamination, sequence artifacts, low read depth in the normal sample, as well as non-exonic variants. In summary, combining mutation calls from the two accelerated callers MuSE 2 and Strelka2, e.g., by simply taking an intersection of the calls, is promising to achieve optimized mutation calling in a significantly shorter wait-time. This strategy is particularly useful for WGS data and for analysis of large patient cohorts. With WES data, running MuSE 2 alone can be a cost-effective strategy to obtain mutation calls with high precision and a reasonable F1 score.

A Snakemake pipeline for somatic SNV calling. Finally, we introduce a fully automated mutation calling pipeline for general users who do not have the time or expertise to learn about the nuances in optimizing mutation calling accuracy, using the Snakemake workflow management system (Köster and Rahmann 2012) (**Supplemental Fig. S5**). This user-friendly pipeline allows for running all preprocessing, postprocessing, as well as MuSE 2 and Strelka2 together, in the background, without manual curation. It is compatible with typical Linux systems and computing clusters, and optimizes memory and CPU utilization by parallelizing independent tasks.

Discussion

Precision medicine and personalized cancer treatments have advanced remarkably in the last decade, which greatly benefited from the accurate identification of genetic variations in the tumor tissue using HTS data. An efficient and accurate somatic mutation caller is crucial to the scientific studies of all cancers and their clinical management. Previously the accuracy and utility of MuSE 1, either alone (Fan et al. 2016) or as a member of a multi-caller consensus calling strategy has been validated by multiple consortial projects (Ellrott et al. 2018; The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium 2020). This study further develops MuSE 2, in order to fully utilize resources on a high-performance computing machine, including both the CPU cores and memory allocation. The producer-consumer model behind the parallelization implemented in the step of 'MuSE call' gives MuSE 2 the ability to manage multiple processes (workers) at the same time: they run independently at their own rates without being affected by the computing load of other processes. Since the calculation in the step of 'MuSE sump' is more straightforward – the computing speed bottlenecks only reside in several for-loop iterations, we therefore use the OpenMP library, with which the parallelization is relatively trivial. The speed-up by MuSE 2 becomes evident when it is run on at least 4-5 cores to take advantage of the multi-step parallelization. In summary, MuSE 2 improves the mutation calling utility of MuSE 1 by accelerating its computing speed by up to 50-60 times for both WES and WGS data. MuSE 2 reduces the computational time cost of a somatic mutation calling project from ~40 hours to < 1 hour for WGS data, and from 2-4 hours to ~5 minutes for WES data, from each pair of tumor-normal samples.

MuSE 2 is much faster than the other three benchmarked callers, i.e., MuTect2, SomaticSniper and VarScan2. It is slightly slower than Strelka2 for the sequencing data generated by HiSeq 2000 and HiSeq X Ten, but much faster than the latter with NovaSeq. Since we only include one pair of tumor-normal WGS data from HiSeq X Ten and NovaSeq, respectively, more data are needed to validate this result in the future study. The intersection of MuSE 2 and Strelka2 calls can substantially improve precisions without much loss in recalls,

hence improving the overall F1 scores for both the WES and WGS data benchmarked. We therefore demonstrate the utility of the intersection calls from these two fast callers, as compared to using each caller individually or using un-accelerated callers.

In contrast to the consensus calls of TCGA and PCAWG and the cell line study, running MuSE 2 and Strelka2 to generate intersect calls may greatly improve the efficiency of genomic data analysis for large patient cohorts, especially for those with WGS data. We also found running MuSE 2 alone is a cost-effective solution for mutation calling in WES data, as it would otherwise require 4-5 callers plus post-filters to achieve much higher recall and precision. Finally, we provide a Snakemake workflow pipeline that automatically runs preprocessing, intersect mutation calling using the two accelerated callers, and postprocessing without human intervention, in order to improve accessibility by general users. We note that the hg19 genome assembly was used throughout the study due to the fact that all consensus calls were based on hg19. Given the underlying models of MuSE 2 and Strelka2, we expect the performance of variant calling of both methods to be insensitive to genome assemblies. As the switch of assembly build from hg19 to hg38 can impact preprocessing and read mapping to generate the input data, some difference in variant calls could be observed, which should not substantially affect the conclusions (Gao et al., 2019). Future development of MuSE 2 includes indel calling, SNV calling from formalin-fixed paraffin-embedded (FFPE) samples and from tumor samples only, all of which require an advanced error model construction as well as further benchmarking. In summary, we expect the proposed MuSE 2 to significantly accelerate the variant calling process and benefit the cancer research and clinical communities.

Methods

Sample selection

The consensus mutation calls of the TCGA portion of the PCAWG project were downloaded from the ICGC data portal (https://dcc.icgc.org/releases/PCAWG/consensus_snv_indel). The consensus mutation calls of the TCGA MC3 project were downloaded from GDC (https://gdc.cancer.gov/about-data/publications/mc3-2017). We randomly selected 5 patient samples from each of the two repositories, and downloaded the BAM files from https://dcc.icgc.org/repositories and https://portal.gdc.cancer.gov/repository, respectively. The sequencing data of these 10 patient samples were generated by Illumina HiSeq 2000. We further downloaded BAM files of WGS data from the metastatic melanoma cell line COLO829 and the matched normal lymphoblastoid line COLO829BL from European Nucleotide Archive (ENA) (accession code PRJEB27698, https://www.ebi.ac.uk/ena/browser/view/PRJEB27698). We downloaded the latest WGS dataset (Espejo Valle-Inclan et al. 2022). The sequencing libraries were either prepared with Illumina TruSeq Nano reagent kit and sequenced on the HiSeq X Ten platform (COLO829 Illumina), or prepared on the 10X Chromium platform and sequenced on the NovaSeq platform (COLO829 10X). For COLO829 Illumina, we also downloaded the BAM files of the cell line with mixed tumor cell proportions of 75%, 50%, 25%, 20% and 10%. These data were simulated by *in silico* mixing of reads from COLO829 (100% tumor) and COLO829BL (normal sample) with different ratios.

BAM pre-processing

MuSE 2 adopts the same preprocessing steps for the unaligned sequencing reads of the tumor-normal pair as MuSE 1, which include trimming poor-quality bases, removing adapters, marking duplicate reads, performing local indel realignment for the paired tumor-normal BAM files jointly, and recalibrating base quality scores (**Supplemental Fig. S1**). In this study, the sequencing reads are aligned against the hg19 reference genome build using BWA-MEM (Li 2013).

Sequencing depth

The sequencing depth of each BAM file after pre-processing is estimated by SAMtools with the 'depth' command. For WGS data, the overall depth was calculated as the average of the read

depths of all genomic locations. For WES data, the overall depth was calculated as the average of the read depths of the genomic locations in the exon regions defined by the exome capture kit downloaded from GDC (https://gdc.cancer.gov/about-data/publications/mc3-2017).

Parallel computing implementation for MuSE

MuSE call. We implement a multithreaded producer-consumer model which deploys threads for parsing and uncompressing reads from BAM files, variant filtering and detection, writing outputs and monitoring the whole process. The model connects all the threads concurrently by thread-safe queues and atomic variables. We also adopt a faster and more efficient memory allocator (i.e., TCMalloc: https://github.com/google/tcmalloc) rather than use the default malloc in C and new in C++ in this step. The parallelization model starts with creating of 6 threads, 3 for the BAM of tumor sample and the other 3 for the BAM of normal sample: 1 of the 3 threads parses the compressed binary data and sends its reference to two queues, namely ChunkReadQueue and ChunkUnzipQueue; the other two threads take the data from the ChunkUnzipQueue, decompress it and label it as 'processed'. This change is also effective for the data in ChunkReadQueue, since these two queues in fact store the same data. Another thread (i.e., read) is then created, which takes uncompressed data from ChunkReadQueue and recover it to read format for both the BAM tumor sample and the BAM of normal sample, and pushes them to the same queue, ReadQueue. A new thread named processReads is created; it parses the reads from ReadQueue and sends them to the queue, processQ. *n* threads named workers are created to take the reads from processQ and process them following the same pre-filtering and the evolutionary model as MuSE 1. The last thread is named as 'monitor', which prints the sizes of the queues every second. Here, users can specify *n* according to the number of cores available in the input of MuSE 2 (**Supplemental Fig. S6**).

MuSE sump. We use the OpenMP library to parallelize the three most time-consuming parts in MuSE sump. The first is the loading of candidate variants, the corresponding estimates of equilibrium frequencies for all four alleles (A, C, T, G) for each variant from MuSE call, and

filtering out the variants whose ratio between the variant allele frequency from the normal sample and the variant allele frequency from the tumor sample above a predefined cutoff (Fan et al. 2016) (0.05). The second is scanning for the remaining variants in the dbSNP and marked as 'true' or 'false' if they appear in the database or not. For WGS data, MuSE 1 fits a two-component Gaussian mixture model to the allele frequencies of the post-filtered variants to separate true mutations from background noise. The parameters (e.g., mean, standard deviation and proportion) of the two components are estimated using the expectation-maximization algorithm which are repeated 50 times with random initializations. For the three parts, we parallel the for-loop iterations using the 'omp parallel for' clause from OpenMP in MuSE 2 to deploy the computation on multiple cores.

Speed benchmarking settings

For all the benchmarked methods, if the number of cores requested lies in {1, 5, 10, 20, 28}, the processor is Intel(R) Xeon(R) Gold 6132 CPU @ 2.60GHz; if the number of cores requested lies in {40, 80}, the processor is Intel(R) Xeon(R) Platinum 8380 CPU @ 2.30GHz. We run each method by submitting the LSF (Load Sharing Facility) job script using the bsub command, with which we can easily control the random-access memory (RAM) and the number of cores specified for each method. The options for the 6 callers can be found in **Supplemental Table S4**.

Precision and recall

For the samples from TCGA and PCAWG, we used the consensus SNV calls published previously (Ellrott et al. 2018; The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium 2020) as a truth set; for the cell line COLO829/COLO829BL, we used the SNV calls downloaded from (Craig et al. 2016) as a truth set. The version information of the benchmarked callers is listed in **Supplemental Table S5**. Indel calls were removed from the call set before any comparison. For WGS data, we took calls from all tiers in MuSE 2 (Fan et al. 2016), and

only calls from the PASS category from the other callers for each patient sample. We filtered out low quality SNVs from the consensus calls from PCAWG WGS data that are labeled as 'LOWSUPPORT', 'OXOGFAIL', 'bSeq', 'bPcr', 'GERM1000G', 'GERMOVLP', 'NORMALPANEL' or 'REMAPFAIL'(https://dcc.icgc.org/releases/PCAWG/consensus_snv_indel). Consistently, we used consensus calls for the cell line COLO829/COLO829BL which had already gone through a similar post-filtering process (Craig et al. 2016). For WES data, we selected calls from all the categories except for 'Tier5' from MuSE 2, and only calls from the PASS category from the other callers for each patient sample. We also used the consensus calls of TCGA WES that had already gone through post-filtering (Ellrott et al. 2018). The intersection between any two sets from the same patient sample were identified by matching the SNV ids, which combined the columns of CHROM, POS, REF and ALT from the two VCF files. For WES data, we removed the SNVs from the intersection calls outside the regions defined by the exome capture kit of TCGA.

We considered any calls reported by the consensus, but not by the intersection calls as false negatives, any calls reported by the intersection calls, but not by the consensus as false positives. We calculated precision, recall and F1 score to evaluate the accuracy of a call set against a truth set.

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Definitions of variant allele frequency bins, sequencing depth bins, clonality and variant effect annotation for SNVs

To resolve the issue that the read depth (including the number of reads supporting the reference allele and the alternate allele) can be different for the same SNVs from MuSE 2, Strelka2 and the consensus calls, we used alleleCount (https://github.com/cancerit/alleleCount) to recalculate the read depth and VAF for all the unique SNVs from MuSE 2, Strelka2 and the consensus calls. We finally generated the bins of VAFs (i.e., 0-0.2, 0.2-0.3, 0.3-0.4, 0.4-0.5 and >0.5) and read

depths (<40×, 40-80×, 80-120×, 120-160× and >160× for WES data, <20×, 20-40×, 40-60×, 60-80~~X~~× and >80× for WGS data) for the calls of each method for detailed comparisons.

We downloaded the consensus subclonal reconstruction results (Dentro et al. 2021) for the consensus calls of PCAWG WGS data from ICGC data portal: https://dcc.icgc.org/releases/PCAWG/subclonal_reconstruction/. Each SNV in the consensus calls is defined as either clonal or subclonal when such information is available. To compare the performance between MuSE 2 and other callers, we restricted the SNVs from each caller overlapping with the ones from the consensus calls such that they can be annotated as clone or subclone. Therefore, only recall is evaluated.

We used Ensembl VEP (McLaren et al. 2016) (v101) to predict the effect of a SNV. For simplicity, we merged nonsense and missense variants into nonsynonymous variants; variants in 3'UTR (untranslated region), 5' UTR, 3' Flank and 5' Flank into untranslated region variants; variants in splice region, translation start site and RNA variants into others category. We also renamed silent variants to synonymous variants. We have four classes for the SNVs from TCGA WES data: nonsynonymous, synonymous, untranslated region and others, and six classes for the SNVs from PCAWG or cell line COLO829/COLO829BL WGS data: nonsynonymous, synonymous, intergenic region, intron, untranslated region and others.

Software availability

MuSE 2 is implemented in C++ and is available at GitHub https://github.com/wwylab/MuSE with the GPL-2.0 license. A Dockerfile is included in the repository for building MuSE 2 into a Docker container running on Linux machines. A Snakemake pipeline for somatic SNV calling, MuSE.Snakemake 1.0, is also available on the GitHub repository. The source code of the repository can also be found in the **Supplemental Material**.

Competing interest statement

Tong Zhu and Ankit Sethia are employees of NVIDIA Corp. and own NVIDIA stock as part of the standard compensation package.

Acknowledgements

S.J. is supported by the MD Anderson Colorectal Cancer Moon Shot Program and National Institutes of Health (NIH) R01CA268380. W.W. is supported by DoD PC210079, NIH R01CA268380, P30CA016672. S.J. and W.W are also supported by The MD Anderson Cancer Center SPORE in Gastrointestinal Cancer Grant P50 CA221707. The authors thank the Biomedical Visualization team at Anderson Cancer Center for their assistance in creating Figure 1. Copyright used with the permission of The Board of Regents of the University of Texas System through The University of Texas MD Anderson Cancer Center.

Author contributions

S.J. implemented the parallelization for the ‘MuSE sump’ step, performed the speed benchmarking, analyzed the results, and wrote the manuscript in collaboration with other authors. T.Z. and A.S. implemented the parallelization for the ‘MuSE call’ step. W.W. conceived the project, planned, and supervised the work, wrote the manuscript, in collaboration with all other authors. All authors commented on and approved the final manuscript.

References

Alfares A, Aloraini T, subaie L Al, Alissa A, Qudsi A Al, Alahmad A, Mutairi F Al, Alswaid A, Alothaim A, Eyaid W, et al. 2018. Whole-genome sequencing offers additional but limited clinical utility compared with reanalysis of whole-exome sequencing. *Genetics in Medicine* **20**: 1328–1333.

Barbitoff YA, Polev DE, Glotov AS, Serebryakova EA, Shcherbakova I V, Kiselev AM, Kostareva AA, Glotov OS, Predeus A V. 2020. Systematic dissection of biases in whole-exome and whole-genome sequencing reveals major determinants of coding sequence coverage. *Sci Rep* **10**: 1–13.

Christoforides A, Carpten JD, Weiss GJ, Demeure MJ, Von Hoff DD, Craig DW. 2013. Identification of somatic mutations in cancer through Bayesian-based analysis of sequenced genome pairs. *BMC Genomics* **4**: 302.

Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, Getz G. 2013. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* **31**: 213–219.

Craig DW, Nasser S, Corbett R, Chan SK, Murray L, Legendre C, Tembe W, Adkins J, Kim N, Wong S, et al. 2016. A somatic reference standard for cancer genome sequencing. *Sci Rep* **6**: 24607.

Dentro SC, Leshchiner I, Haase K, Tarabichi M, Wintersinger J, Deshwar AG, Yu K, Rubanova Y, Macintyre G, Demeulemeester J, et al. 2021. Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. *Cell* **184**: 2239-2254.e39.

Ellrott K, Bailey MH, Saksena G, Covington KR, Kandoth C, Stewart C, Hess J, Ma S, Chiotti KE, McLellan MD, et al. 2018. Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell Syst* **6**: 271-281.e7.

Espejo Valle-Inclan J, Besselink NJM, de Bruijn E, Cameron DL, Ebler J, Kutzera J, van Lieshout S, Marschall T, Nelen M, Priestley P, et al. 2022. A multi-platform reference for somatic structural variation detection. *Cell Genomics* **2**: 100139.

Fan Y, Xi L, Hughes DST, Zhang J, Zhang J, Futreal PA, Wheeler DA, Wang W. 2016. MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biol* **17**: 178.

Gao GF, Parker JS, Reynolds SM, Silva TC, Wang LB, Zhou W, Akbani R, Bailey M, Balu S, Berman BP, et al. 2019. Before and After: Comparison of Legacy and Harmonized TCGA Genomic Data Commons' Data. *Cell Syst* **9**: 24-34.e10.

Kim S, Scheffler K, Halpern AL, Bekritsky MA, Noh E, Källberg M, Chen X, Kim Y, Beyter D, Krusche P, et al. 2018. Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods* **15**: 591–594.

Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK. 2012. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* **22**: 568–576.

Köster J, Rahmann S. 2012. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **28**: 2520–2522.

Larson DE, Harris CC, Chen K, Koboldt DC, Abbott TE, Dooling DJ, Ley TJ, Mardis ER, Wilson RK, Ding L. 2012a. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* **28**: 311–317.

Larson DE, Harris CC, Chen K, Koboldt DC, Abbott TE, Dooling DJ, Ley TJ, Mardis ER, Wilson RK, Ding L. 2012b. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* **28**: 311–317.

Li, H. 2011. Tabix: Fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics* **27**: 718–719.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:13033997*.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, Cunningham F. 2016. The Ensembl Variant Effect Predictor. *Genome Biol* **17**: 1–14.

Saunders CT, Wong WSW, Swamy S, Becq J, Murray LJ, Cheetham RK. 2012. Strelka: Accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* **28**: 1811–1817.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: The NCBI database of genetic variation. *Nucleic Acids Res* **29**: 308–311.

Sun Y, Liu F, Fan C, Wang Y, Song L, Fang Z, Han R, Wang Z, Wang X, Yang Z, et al. 2021. Characterizing sensitivity and coverage of clinical WGS as a diagnostic test for genetic disorders. *BMC Med Genomics* **14**: 1–13.

The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. 2020. Pan-cancer analysis of whole genomes. *Nature* **578**: 82–93.

Xu C. 2018. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Comput Struct Biotechnol J* **16**: 15–24.

Figure legends

Figure 1. Assembly line illustration of the multi-step parallelization implemented in MuSE 2. *(A)* 'MuSE call'. Workers (threads) keep fetching chunks from the input BAM files from the tumor and normal samples and unzipping them to the text format of reads. Downstream workers combine the reads from tumor and normal samples and send to a queue; from there, other workers detect candidate variants. *(B)* 'MuSE sump'. Multiple workers are used to take the candidate variants and their corresponding estimated summary statistic $\pi's$ and scan them against dbSNP database, labeling those appearing in the database. For candidate variants from the WGS data, we fit two-component Gaussian Mixture Models (GMMs) with multiple initializations, distributed to multiple workers, in order to separate true variants from background noise; for candidate variants from the WES data, no parallelization is implemented due to computational simplicity as we simply fit a beta distribution to π's.

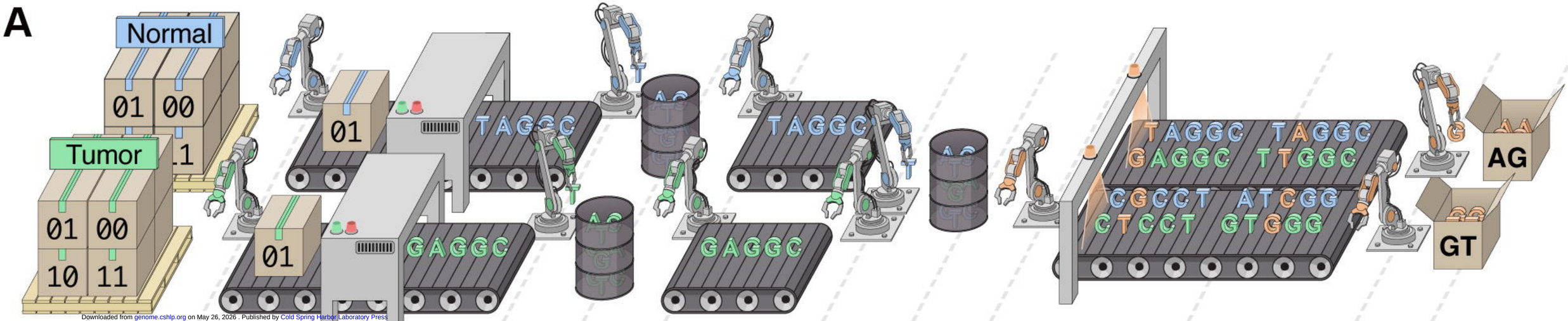
Figure 2. Comparisons of F1 score, precision and recall between MuSE 2 and Strelka2 within each bin of variant allele frequency (VAF, top) or sequencing read depth (middle), variant class (bottom) for TCGA WES data. The calls of each method, and the consensus calls, which are used as the truth set, are pooled from the WES data of 5 patient samples in TCGA.

Figure 3. Accuracy benchmarking of MuSE 2 and Strelka2 within each bin of VAF (first row) or sequencing read depth (second row), class of variant effects (third row) for PCAWG WGS (A) and cell line WGS data (B). Comparison of recall between the two methods within different clonality for PCAWG WGS data is shown in the last row of (A). The calls of each method, and the truth set, are pooled from the WGS data of the selected 5 patient samples from PCAWG (A), or the WGS data of the cell line COLO829/COLO829BL generated from two platforms (B). The number of consensus

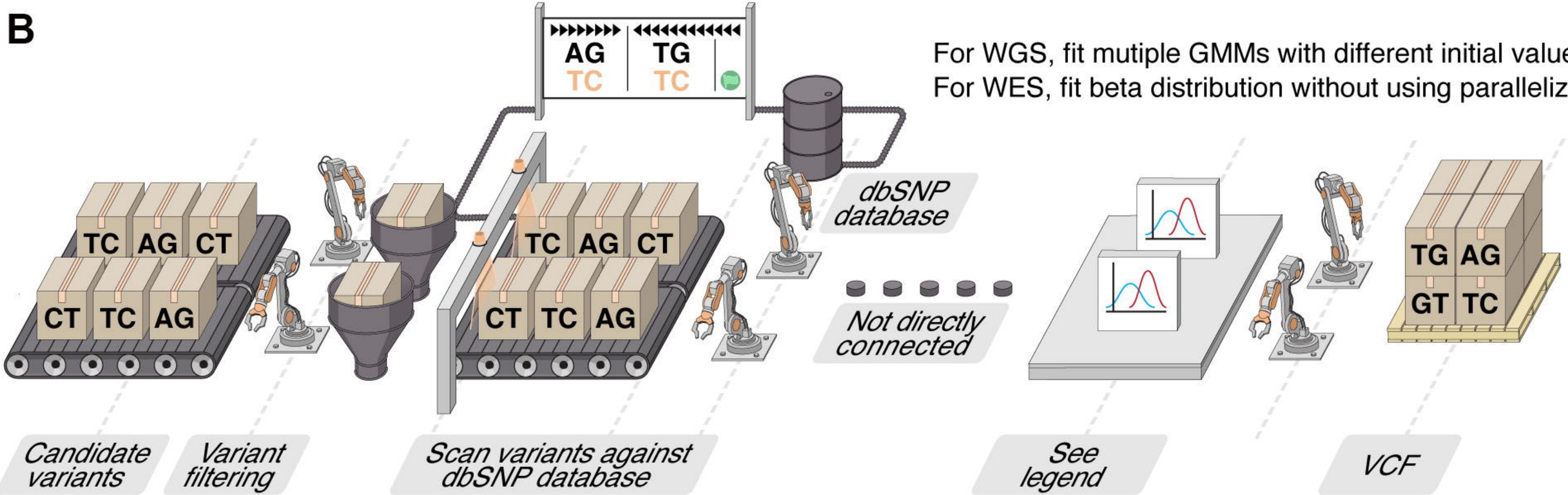
calls for a variant class is included in the x-axis labels. (C) F1 scores of MuSE 2 and Strelka2 with varied tumor cell proportions. Calls from the tumor purity of 100% were used as the truth set.

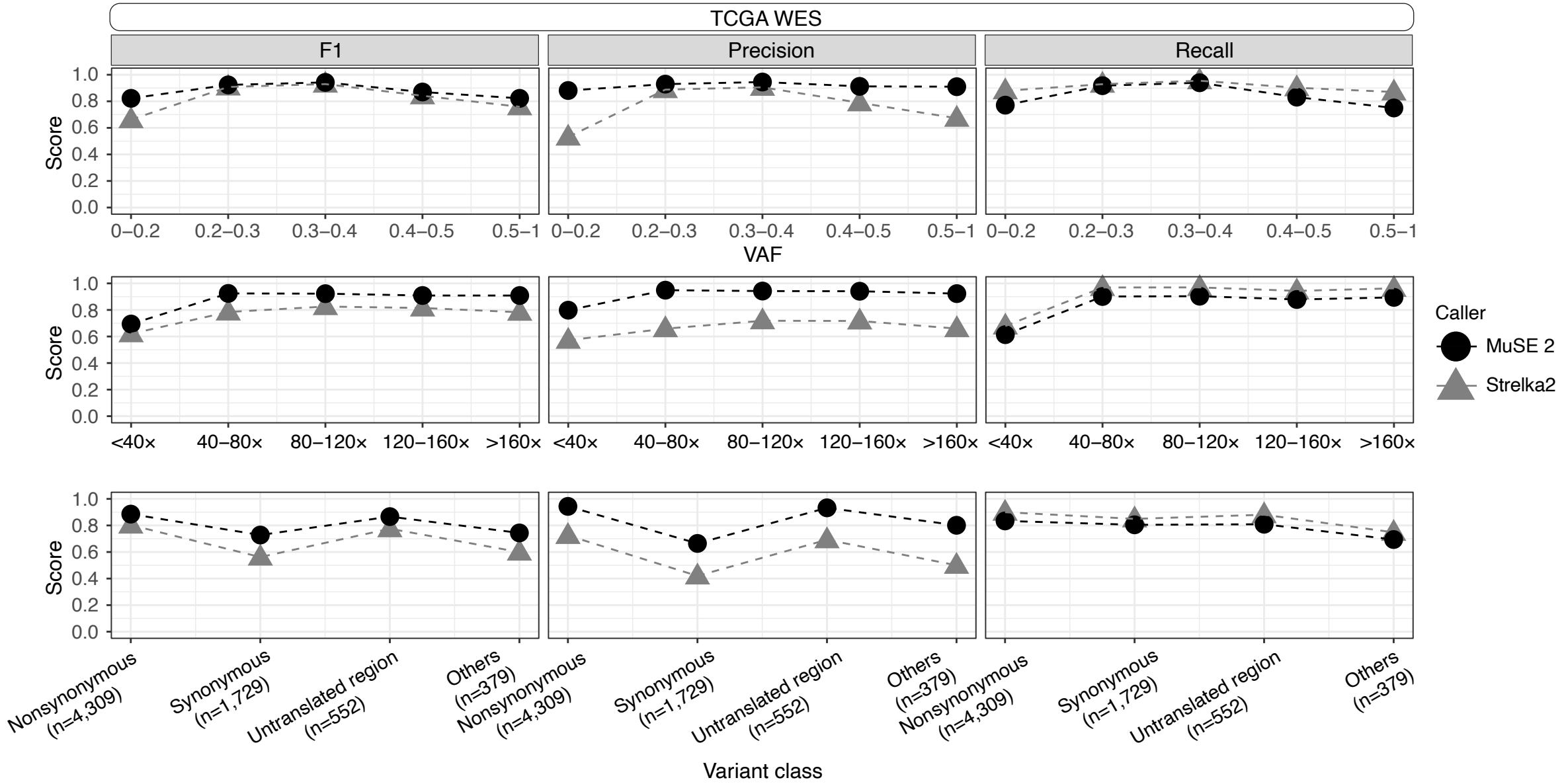
Figure 4. Benchmarking the speed and usability of MuSE 2. *(A)* The runtime of MuSE 2 against MuSE 1 and the other four methods for both WES and WGS data across different numbers of cores. The numbers in the plot represent the fold speedup of MuSE 2 (with 80 cores) relative to the other methods whose time cost is averaged across different numbers of cores (excluding No. of core=1). For Strelka2, only the time cost with 80 cores is considered. *(B)* A simplified version of *(A)* in which the time cost of each method is averaged across all samples (excluding COLO829 10X as an outlier, see (A)) and different numbers of cores. *(C)* Venn diagrams showing the unique and shared SNV calls of MuSE 2 and Strelka2. *(D)* Scatter plot of the precision, recall for the intersection calls from MuSE 2 (in red) and Strelka2 (in blue), the calls from each of the two methods (in purple) against the previously reported consensus calls which are considered as the benchmark. For both WES (circle) and WGS (triangle) data, the median F1 scores of the intersection calls, calls from each individual method are shown. Two shaded rectangles highlight the difference of the performance metrics between WES and WGS data. Results from the WGS data are located in the top rectangle.

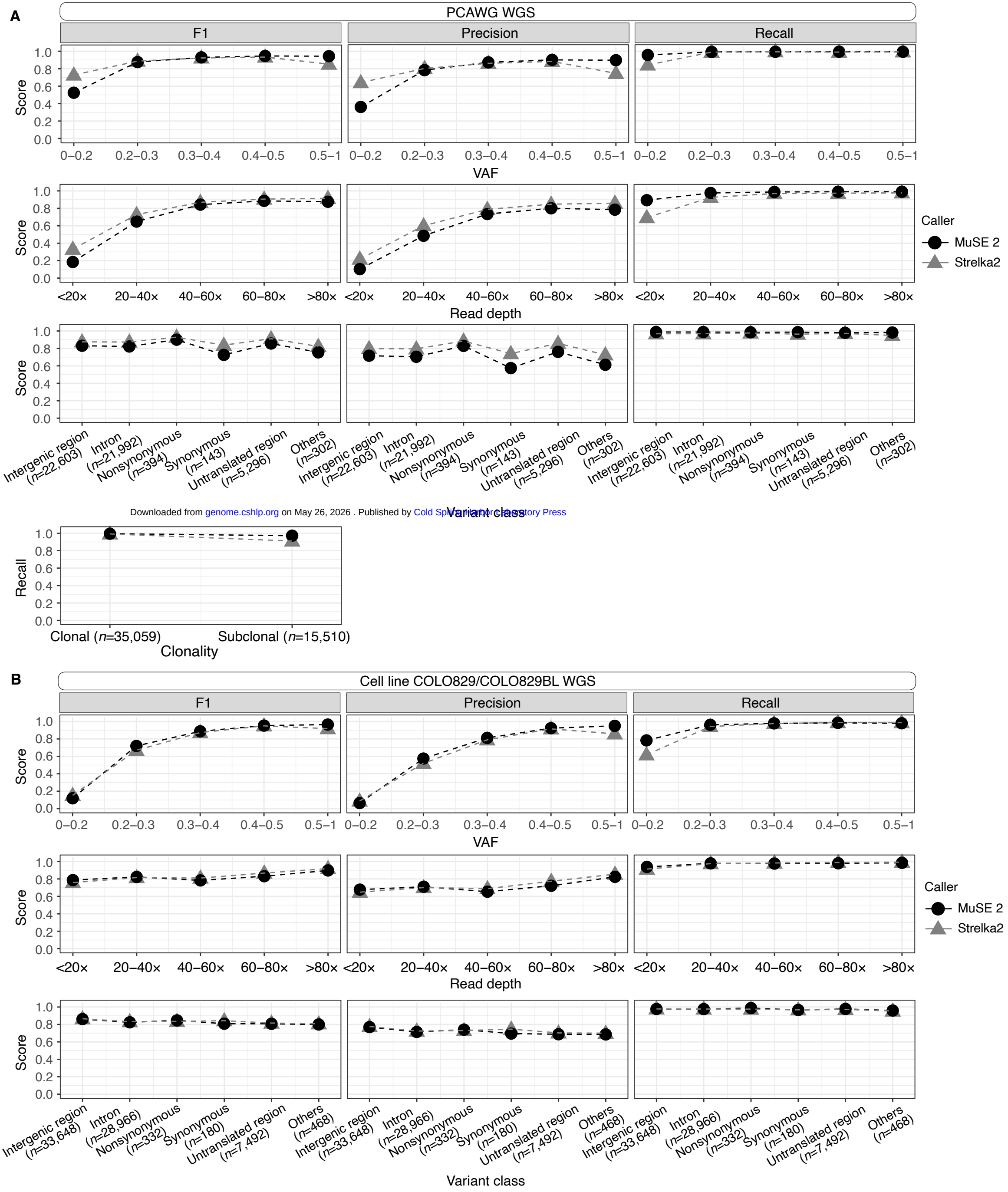
A



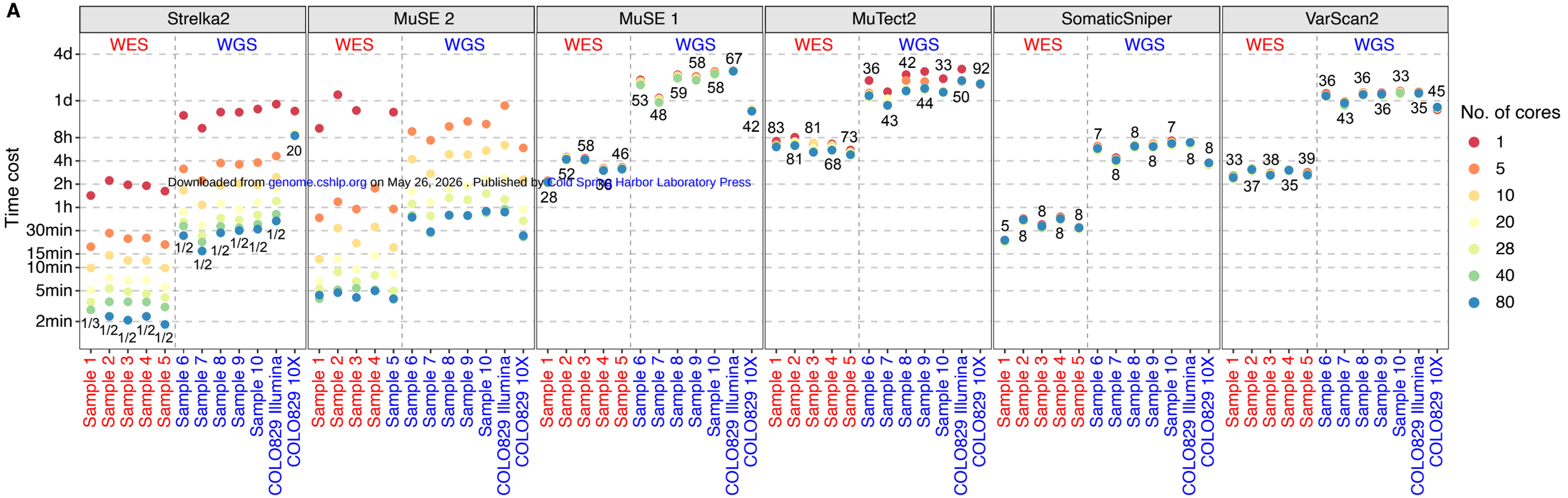
B



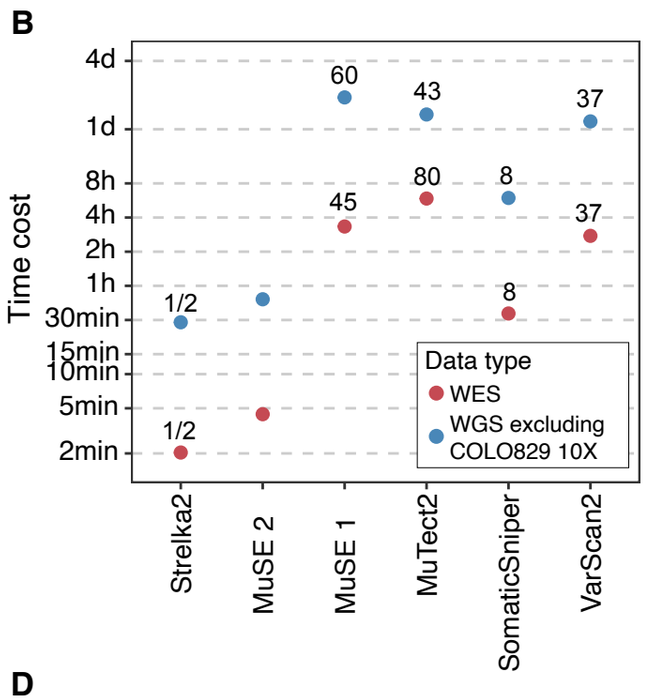

TCGA WES
F1
Precision
Recall
Score
1.0
0.8
0.6
0.4
0.2
0.0
0–0.2
0.2–0.3
0.3–0.4
0.4–0.5
0.5–1
VAF
<40×
40–80×
80–120×
120–160×
>160×
Caller
MuSE 2
Strelka2
Nonsynonymous (n=4,309)
Synonymous (n=1,729)
Untranslated region (n=552)
Others (n=379)
Variant class



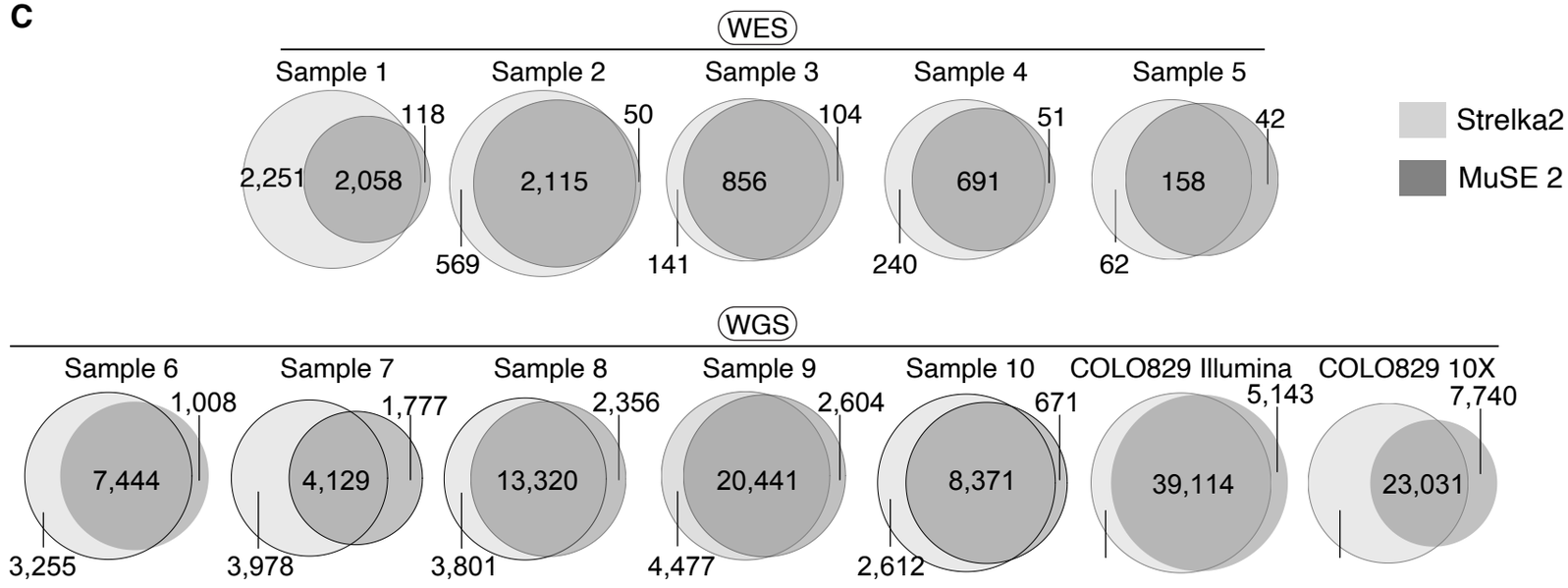
A



B



C



D