Re-genotyping structural variants through an accurate force-calling method

Shuqi Cao[1,2], Tao Jiang[1,2,*], Yadong Liu[1], Shiqi Liu[1] and Yadong Wang[1,*]

[1]Faculty of Computing, Harbin Institute of Technology, Harbin 150001, China.

[2]Joint first authors: Shuqi Cao and Tao Jiang.

*To whom correspondence should be addressed: tjiang@hit.edu.cn, ydwang@hit.edu.cn

Abstract

Long-read sequencing technologies have great potential for the comprehensive discovery of structural variation (SV). However, the accurate genotype assignment for SV is still a challenge since the unavoidable factors like specific sequencing errors or limited coverages. Herein, we propose cuteSV2, a fast and accurate long-read-based re-genotyping approach that is able to force-calling genotypes for the given records. cuteSV2 is an upgraded version of cuteSV and applies an improved strategy of the refinement and purification of the heuristic extracted signatures through spatial and allele similarity estimation. The benchmarking results on several baseline evaluations demonstrate that cuteSV2 outperforms the-state-of-art methods and has stable and robust capability in practice. cuteSV2 is available at https://github.com/tjiangHIT/cuteSV.

Keywords

Structural variant, re-genotyping, force-calling, long read sequencing

Background

Structural variant (SV) is a fundamental type of genomic mutations with a strong association with evolution, population structure, and clinical diseases [1-3]. Typically, the size of SV is over 50 bp and mainly contains insertion, deletion, inversion, duplication, translocation, and more complex variant [4]. Previous population studies like genome-wide association studies (GWAS) focus on the exhaustive characterization of single-nucleotide variant (SNV) underlying human traits and diseases and ignore the important heritability of SV[5, 6]. However, because of SV's variety and wide-range alteration, it usually affects much more nucleotides on the genome and has significant value in genome researches. Recently, long-read sequencing technologies enable the discovery of full-spectrum SVs on a single individual[7-10], which provides an opportunity to reveal the population-based SVs through large-scale sequencing[11-14].

Currently, the joint variant calling strategy is often applied for the construction of population-based SV genetic maps. It can consider all samples simultaneously and generates high-quality population-scale variant callsets, which can give an overall reveal of the variant distribution of population. Since the much more difficulty of obtaining reliable genotypes for a given SV across a population group, the critical step is to perform a re-genotyping for each sample at all SV positions, which benefits refining the original genotypes or filling them up due to insufficient coverage[15, 16].

Here, we introduce an accurate and fast re-genotyping method by force calling, named cuteSV2, to determine the zygosities of all SVs on each individual at the population scale. Using a bisection search strategy on the diverse alignment signatures, the zygosities are obtained by calculating the distribution likelihood of each candidate circumstance. The experiments are implemented on an Ashkenazim trio-family and a population group with 100 Chinese individuals respectively. The results indicate our re-genotyping method shows an outstanding and reliable performance compared to the homogeneous state-of-the-art methods, which represents it will assist in obtaining accurate allele frequencies of homology SVs for further population genetic measurement and estimation.

Results and Discussion

We first evaluated the re-genotyping performance for cuteSV2 and other state-of-the-art methods (i.e., Sniffles1[17], Sniffles2[18], and SVJedi[19]), on a well-known human sample HG002 from the Ashkenazim trio-family group based on the Genome in a Bottle (GiaB) National Institute of Standards and Technology (NIST) ground truthsets (SV v0.6). The target SV candidates were integrated from the Ashkenazim trio-family group (i.e., HG002, HG003, and HG004) detected by PBSV (https://github.com/PacificBiosciences/pbsv). The benchmarking results (Fig. 1A and Supplementary Table S1~3) indicate that cuteSV2 achieved the highest precision under each sequencing technology (96.73% for HiFi, 97.28% for CLR, 97.39% for ONT), while the other three methods reported about twice more false positive SVs. When it comes to recall rate, cuteSV2, Sniffles1, and Sniffles2 reached quite competitive performance against others, besides SVJedi. Overall, cuteSV2 obtained the best F1 scores under the diverse sequencing technologies (94.68% for HiFi, 94.97% for CLR, 95.76% for ONT) according to its pretty high precision and stable recall. Furthermore, we assessed the genotype concordance between the re-genotyped results and the ground truth. It is obvious that cuteSV2 achieved the most optimal concordance on HiFi and CLR datasets (98.14% and 97.75% respectively), followed by Sniffles2 (97.12% and 97.74%), while the other two methods perform worse on genotype concordance (Sniffles1: 46.20% for HiFi and 46.13% for CLR, SVJedi: 71,58% for HiFi and 91.36% for CLR). As for ONT datasets, Sniffles2 acquired the highest concordance at 96.53%, and cuteSV2 followed at 96.46% but is still much higher than the other two methods.

We then evaluated the performance of re-genotyping on GIAB Challenging Medically Relevant Gene Benchmark (CMRG) v1.00 (Fig. 1A and Supplementary Table S4~6). For precision, cuteSV2 still kept ahead among these methods regardless of the sequencing technologies. Besides, it is worth noting that cuteSV2 achieved 100% precision for CLR data, indicating all the SVs reported are accurate. When it comes to recall rate, cuteSV2 performed best on HiFi and ONT data, and was about 2% behind Sniffles1 and Sniffles2 for CLR data.

Considering the sensitivity and accuracy at the same time, cuteSV2 reached the highest F1 score on HiFi and ONT data, and the second best on CLR data (1% lower than that of Sniffles2). For the genotype concordance, cuteSV2 and Sniffles2 still performed better over 10% higher than the other two methods, while no more than 1% between these two methods.

Furthermore, we randomly down-sampled the datasets above to 30×, 20×, 10× and 5× and re-genotyped the HG002 on them (Supplementary Table S1~6). While decreasing the sequencing coverage, the performance of all the methods descended to some extent (Fig. 1B), however, cuteSV2 still remained almost the best performance in most conditions. Especially focusing on precision, cuteSV2 only dropped less than 1%, even a little raising when the coverage varies from the origin coverage to 5× (from 96.73% to 96.96% for HiFi, from 97.28% to 97.33% for CLR, from 97.39% to 97.37% for ONT), which is all higher than the other methods. The recall rate of each method fell more because of the absence of pivotal SV signatures in low-coverage data. cuteSV2 and Sniffles2 dropped about 6% when the coverage descends to 10×, while Sniffles1 and SVJedi dropped much more. Comprehensively considering precision and recall, cuteSV2 still reached the best F1 scores on various coverage data under all sequencing technologies. When it comes to genotype concordance, Sniffles2 performed best on low coverage datasets (dropped by 8.24% for HiFi, 15.48% for CLR, 12.98% for ONT), with cuteSV2 following close (dropped by 12.31% for HiFi, 16.82% for CLR, 14.10% for ONT).

To evaluate the re-genotyping performance on a large-scale cohort, we applied the methods (SVJedi was discarded due to its relatively low capability) on a real population group consisting of 100 Chinese individuals that were sequenced by ONT on about 10~15×. After re-genotyping, Bcftools[20] was adopted to calculate the allele frequency (AF), the test score of excess heterozygosity (ExcHet), and the test score of Hardy-Weinberg Equilibrium (HWE) of each variant. After removing the nonexistent SVs whose AF equals 0, cuteSV2 remained 102,095 SVs which is much more than that of Sniffles2 (95,184) and Sniffles1 (73,962), which indicates the ability of cuteSV2 to detect abundant SVs (Fig. 2A and Supplementary Table S7). CuteSV2 had the smallest number of SVs with ExcHet test scores lower than 0.05, and the largest number of SVs with ExcHet test scores higher than 0.95, indicating its outstanding genotype assignments for heterozygous (Fig. 2B and Supplementary Table S8). In the HWE test, Sniffles2 (14.26%) and cuteSV2 (16.56%) reported a less proportion of SVs with HWE test scores lower than 0.05, while Sniffles1 have much more bad cases (30.63%) compared with the other two (Fig. 2B and Supplementary Table S8~10).

We randomly selected three samples from this large-scale group to further assessed the consistency of SV calling and corresponding genotyping through orthogonal PacBio HiFi sequencing. From the distribution of consistency rate (CR) and SV amount under different AF intervals shown in Fig. 2C, it is obvious that cuteSV2 and Sniffles2 showed the outstanding ability to detect SVs with reporting over 20,000 SVs for each sample (20,429 for cuteSV2 and 23,068 for Sniffles2, in average), while Sniffles1 only reported about 15,398 SVs in average (Fig. S2 and Supplementary Table S11~13). When it comes to consistency rate, Sniffles1 performed the highest (about 80%) and cuteSV2 followed close (about 78%), while Sniffles2 performed much lower (about 70%). Taking these two aspects into account, cuteSV2 detected the largest number of consistent SVs with both high total amount and consistency rate. Furthermore, we benched the CR of different SV types and Fig. 2C shows that for insertions and deletions, the conclusion is similar to that above, that is, cuteSV2 detected the most consistent SVs (Supplementary Table S14). cuteSV2 discovered much more inversions than other methods (about 70 for cuteSV2, 30 for Sniffles1, and 40 for Sniffles2), meanwhile, there are about twice as consistent inversions. When it comes to duplications, cuteSV2 reported about ten times more, and there are about 30 consistent duplications while Sniffles1 and Sniffles2 only discovered less than 10 duplications in total. Hence, cuteSV2 not only performed better in insertions and deletions, but also in longer and more complex SVs. Furthermore, we took the various genomic regions where SVs were involved into account. It is obvious that cuteSV2 achieved almost the highest consistency rate for those SVs located in gene-related regions (Fig. 2C and Supplementary Table S15). This consequence indicates the highly accurate SV detection ability, especially in high genetic conservation regions, which would benefit the studies that highly rely on functional SVs.

Finally, we examined the computational performance of these tools in different datasets. From the assessments of elapsed time on the HG002 sample (Fig. 2D and Supplementary Table S16), Sniffles2 had the fastest speed regardless of the dataset (i.e., HiFi: 871s, CLR: 8,263s, ONT: 4,655s), with cuteSV2 (i.e., HiFi: 1,256s, CLR: 19,134s, ONT: 11,400s) followed at approximately one time slower and Sniffles1 (i.e., HiFi: 5,705s, CLR: 24,978s, ONT: 16,380s) and SVJedi (i.e., HiFi: 96,765s, CLR: 93,062s, ONT: 103,471s, all under 16 threads) were much slower. Notably, cuteSV2 and Sniffles2 both achieved a quasilinear speedup and time reduction with the increasing threads. Hence, there would be no significant time discrepancy between cuteSV2 and Sniffles2 when applying more CPU threads. In terms of memory footprint, Sniffles1 (from 2.6 to 38.7 GB) and Sniffles2 (from 0.6 to 4.3 GB) were variable and highly related to the size of the data. SVJedi cost a larger memory that was about 7 GB. By contrast, cuteSV2 used a relatively small and stable memory (about 3 GB) regardless of the size of the dataset, which is mainly due to the about 3 GB size of the human reference genome involved in the reference allele generation. For the assessments of the 100 Chinese individuals (Fig. 2E and Supplementary Table S17), cuteSV2 and Sniffles2 achieved equivalent and fast speed (i.e., 320s and 281s respectively), whereas the average memory footprint for cuteSV2 (3.53 GB) was smaller than that for Sniffles2 (4.14 GB). Sniffles1 cost much more time (10,162s) and memory (5.45 GB) on each sample significantly.

Through re-genotyping SVs via long reads, we supply an accurate and fast re-genotyping solution, which is promoting the studies like population genetic measurement and estimation effectively. In order to better apply cuteSV2 in practice, there are still several advantages and disadvantages that need to explain in detail.

- cuteSV2 implements a two-round purification on SV signatures to select the reads that support the target SV. The first round follows the refinement step of cuteSV to distinguish signatures with various allele lengths, while the second round performs an enhanced purification strategy by inspecting the allele similarity comprehensively. An example (Fig. S3) shows that a target deletion (1,620bp) on chromosome 13 was obtained correctly genotype only by cuteSV2. This is mainly due to the strength of the two-round purification strategy for figuring out the signatures on the complex genomic region.

- cuteSV2 applies a merging strategy for those fragile signatures affected by the erroneous read-alignment and generates agglomerated signatures. This will improve the sensitivity of re-genotyping through the maintenance of the potential signatures as many as possible. An example (Fig. S4) shows that a target insertion (299bp) on chromosome X was determined as homozygous only by cuteSV2 and is concordant with the corresponding ground truth. For the absence or limitation of the merging strategy, Sniffles1 and Sniffles2 reported wrong genotypes.

- The heuristic SV signature extraction module enables cuteSV2 to discover more amount and more accurate signatures, especially for longer and more complex SVs such as inversions and duplications, which helps to carry out high-quality genotype assignments. From Fig. S5, there is an inversion of 1,200bp on chromosome 1 and only cuteSV2 reported the real genotype successfully. Another example of duplication of 7,759bp on chromosome 1 is shown in Fig. S6, all the other methods reported the genotype as 0/0 except cuteSV2. This duplication is located on the *HRNR* gene that is highly associated with gene compression, and only cuteSV2 achieved the genotype assignment correctly for this, which indicates that cuteSV2 has the potential for functional SV discovery in those accurate clinical studies.
- Although cuteSV2 contains a great ability of re-genotyping, however, there are still left several limitations both for it and the current re-genotyping technologies. On the one hand, cuteSV2 can widely handle most of the SV types including insertion, deletion, inversion, duplication, and translocation, but it is difficult to accomplish the re-genotyping of Copy Number Variation (CNV) till now. On the other hand, it is the main mission for cuteSV2 and other methods to complete the re-genotyping on the diploid genome but not on the polyploid genome like plants, and this is also a bottom-neck in the variant calling field. Hence, we will further focus on these topics and try to solve them in the future.

Conclusions

In this article, we introduce an accurate and fast re-genotyping method cuteSV2 to assign genotypes for population-based SVs via long-read sequencing. It is owing to the two algorithms of heuristic signature purification and specific-designed scanning line, that cuteSV2 achieves accurate signature marking and effective read distribution statistics, benefitting to the reliable likelihood estimation of SV genotypes. The benchmarking results indicate that cuteSV2 retains outstanding re-genotyping performance compared with the-state-of-art methods in various aspects. Especially, it has high accuracy with all sequencing technologies regardless of the sequencing depth. These strengths demonstrate that cuteSV2 has a reliable genotype recognition ability contributing to high-quality SV detection for large-scale population studies and precise clinical practice. We believe that cuteSV2 will have wide application in cutting-edge genomic research.

Methods

The force-calling re-genotyping method as an expansion module is integrated into our previous SV detector cuteSV[21, 22]. It can parallelly assign new genotypes referring to the given population-scale SVs, for each SV type on each chromosome among every individual. The approach has four major steps as follows (see Fig. S1).

Step 1. Signatures extraction of various SV types

We follow the signature extraction module of cuteSV to comprehensively collect various types of SV signatures including insertions, deletions, inversions, duplications, and translocations.

Step 2. Candidate signatures marking for SVs

For each re-genotyped SV call, cuteSV2 marks all candidate signatures extracted in Step 1 through specific-designed spatial similarity measurement and allele-similarity estimation. First, a binary search strategy is implemented in the ordered signature list to find the signature Sig_{flag} which has the nearest coordinate as the given SV. Then cuteSV2 measures the spatial similarity of signatures upstream and downstream and collects them recursively. For a collected signature Sig_{NN}, its nearest neighborhood signature Sig_{NN}' will be collected when it satisfies:

$$\begin{cases} \left|BP_{NN} - BP_{flag}\right| \le 2000 \\ \left|BP_{NN} - BP_{NN}'\right| \le max\_cluster\_bias \end{cases} \quad (1)$$

where BP_{NN}, BP_{NN}' and BP_{flag} are the breakpoints of Sig_{NN}, Sig_{NN}' and Sig_{flag}, respectively; $max\_cluster\_bias$ indicates a threshold which is mutative due to various SV types and sequencing platforms (by default, 200, 100, 500, 500 and 50 for DEL, INS, DUP, INV and TRA, respectively). It is worth noting that, for translocations, an additional condition that constrains the same transferred chromosome ID of signatures is designed to ensure the spatial similarity of collected signatures.

The above-collected signatures only consider the spatial similarity but ignore the similarity in alleles. Therefore, cuteSV2 then purifies the collected signatures through the allele-similarity estimation and only retains the signatures with high allele-similarity as candidate signatures. For insertions and deletions, to restore the signatures as many as possible, the fragile signatures from the same read are merged together to generate novel potential signatures. Then cuteSV2 implements the refinement step of cuteSV on the signatures to discard them into various clusters and measures the allele-similarity of each cluster to re-genotyped SV as below:

$$Similarity_{cluster} = \frac{\min(\frac{1}{n} \times \sum_{i=1}^{n} LEN_i\,, LEN_{given})}{\max(\frac{1}{n} \times \sum_{i=1}^{n} LEN_i\,, LEN_{given})} \quad (2)$$

where LEN_{given} and LEN_i indicate the length of the given re-genotyped SV and the ith signature in the cluster, respectively. cuteSV2 selects the cluster with the highest allele-similarity as once purified collected signatures. For inversions, duplications, and translocations, cuteSV2 uses the collected signatures directly. In the second purification, cuteSV2 measures the allele-similarity of each purified collected signature through (3) and remains the signatures whose allele similarity exceeds 0.7 as the final candidate signatures.

$$Similarity_{allele} = \frac{\min(LEN_{allele}, LEN_{given})}{\max(LEN_{allele}, LEN_{given})} \quad (3)$$

Step 3. Overlapped reads marking for SVs

cuteSV2 computes the distribution of reads around each re-genotyped SV breakpoint by an overlap scanning line. Specifically, for each SV, cuteSV2 records all alignment reads that cover the SV on the chromosome.

All the re-genotyped SVs and alignment reads are collected together as 2-tuple (s_{SV}, e_{SV}), (s_{read}, e_{read}), where s_{read} and s_{SV} represent the start coordinate of read/SV, e_{read} and e_{SV} represent the end coordinate of read/SV, respectively. The tuples are grouped by chromosome, and in each group, all the breakpoints in tuples are sorted in order according to their coordinates. Then cuteSV2 designs a line scanning the ordered breakpoints from head to tail. While moving the scanning line, cuteSV2 uses a read set R to record the reads that stride over the line and $R = \{read|s_{read} < bp_{line} < e_{read}\}$, where bp_{line} represents the breakpoint that the scanning line reaches. In detail, the read is added into read set when the scanning line reaches s_{read} and removed from read set when the line reaches e_{read}. And when the line reaches s_{SV} or e_{SV}, cuteSV2 records reads in the read set R as the reads cover the SV breakpoint. The intersection of reads covering both the start and end breakpoint of an SV is marked as the reads that cover the SV.

Step 4. Assign genotypes based on likelihood estimation

In step2 and step3, candidate signatures and cover reads are marked for the re-genotyped SVs. Due to the fact that candidate signatures represent the reads that support the SV, the difference set between the cover reads and the reads where candidate signatures come from represents the reads that support the reference. Using these two read sets, cuteSV2 uses the genotyping module from cuteSV to assign genotype through the bi-allelic assumption on the likelihood of various zygosity.

Implementation of benchmark

For the Ashkenazim Trio, we obtain the PacBio HiFi sequencing alignment data of HG002, HG003, and HG004 from GiaB and apply PBSV to detect the individual SVs. The cohort-level structural variations are generated by merging the three individuals via SURVIVOR[23]. Then we obtain the CLR and ONT sequencing alignment data of HG002 from GiaB and re-genotype the cohort-level structural variations by four methods (cuteSV2, Sniffles1, Sniffles2, SVJedi). The re-genotyping results are benchmarked via Truvari.

For the 100 Chinese population group, we implement ONT sequencing on 100 samples and PacBio HiFi sequencing on three of these samples (i.e., Sample #1, #2, and #3). Four different callers (cuteSV, SVIM[24], Sniffles1, and Nanovar[25]) are applied for individual SV calling and SURVIVOR is applied for multiple samples merging. With the merged cohort-level SVs, we implement re-genotyping methods on 100 samples and apply SURVIVOR the second time to achieve revised cohort-level SVs. Then we implement PBSV on three HiFi sequencing sample to achieve the ground truth. The HWE and ExcHet are achieved via bcftools and the consistent rate is obtained from Truvari.

The benchmark is implemented with a server having 2 Intel(R) Xeon(R) Gold 6240 CPU @ 2.60GHz (32 cores in total), 15.8 Terabytes RAM and 6221 Terabytes hard disk space (no SSD is used), running on CentOS Linux release 7.5.1804 Operating System. The runtime and memory footprint were assessed by using the "/usr/bin/time -v" command of the Linux Operating System.

The detailed commands can be referred in Supplementary Note.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The force calling module in cuteSV2 has been implemented in Python and can be easily installed via bio-conda and PyPI. Its source code is available at github.com/tjiangHIT/cuteSV. The alignment data in the Ashkenazim Trio experiment is available at https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/. The original data of 100 Chinese individuals is restricted access, please contact ydwang@hit.edu.cn for permission acquirement.

Competing interests

The authors declare that they have no competing interests

Funding

This work has been supported by the National Natural Science Foundation of China (Grant number 32000467), China Postdoctoral Science Foundation (Grant Number 2020M681086), and Heilongjiang Postdoctoral Foundation (Grant Number LBH-Z20014).

Authors' contributions

SC, TJ, and YL designed the method. SC and TJ implemented the method. SC and SL performed the experiments and data analysis. SC and TJ wrote the manuscript. The authors read and approved the final manuscript.

References

[1] De Vries B B, Pfundt R, Leisink M, et al. Diagnostic genome profiling in mental retardation [J]. Am J Hum Genet, 2005, 77(4): 606-616.

[2] Sebat J, Lakshmi B, Malhotra D, et al. Strong association of de novo copy number mutations with autism [J]. Science, 2007, 316(5823): 445-449.

[3] Lupski J R. Structural variation mutagenesis of the human genome: Impact on disease and evolution [J]. Environmental and Molecular Mutagenesis, 2015, 56(5):

[4] Sudmant P H, Rausch T, Gardner E J, et al. An integrated map of structural variation in 2,504 human genomes [J]. Nature, 2015, 526(7571): 75-81.

[5] Patron J, Serra-Cayuela A, Han B, et al. Assessing the performance of genome-wide association studies for predicting disease risk [J]. PLoS ONE, 2019, 14(12): e0220215.

[6] Halvorsen M, Huh R, Oskolkov N, et al. Increased burden of ultra-rare structural variants localizing to boundaries of topologically associated domains in schizophrenia [J]. Nature Communications, 2020, 11(1): 1842.

[7] Huddleston J, Chaisson M J, Steinberg K M, et al. Discovery and genotyping of structural variation from long-read haploid genome sequence data [J]. Genome Research, 2016, 27(5):

[8] Liu Y, Jiang T, Su J, et al. SKSV: ultrafast structural variation detection from circular consensus sequencing reads [J]. Bioinformatics, 2021, 37(20): 3647-3649.

[9] Jiang T, Liu B, Li J, et al. rMETL: sensitive mobile element insertion detection with long read realignment [J]. Bioinformatics, 2019, 18): 18.

[10] Jiang T, Fu Y, Liu B, et al. Long-Read based Novel Sequence Insertion Detection with rCANID [J]. IEEE Transactions on NanoBioscience, 2019, 1-1.

[11] Beyter D, Ingimundardottir H, Oddsson A, et al. Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits [J]. Nature Genetics, 2021, 53(6): 779-786.

[12] Audano P A, Sulovari A, Graves-Lindsay T A, et al. Characterizing the Major Structural Variant Alleles of the Human Genome [J]. Cell, 2019, 176(3): 663-675.e619.

[13] Chaisson M J P, Sanders A D, Zhao X, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes [J]. Nature Communications, 2019, 10(1): 1784.

[14] Ebert P, Audano P A, Zhu Q, et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation [J]. Science, 2021, 372(6537): eabf7117.

[15] Koboldt D C. Best practices for variant calling in clinical sequencing [J]. Genome Medicine, 2020, 12(1):

[16] Jiang T, Liu S, Cao S, et al. Long-read sequencing settings for efficient structural variation detection based on comprehensive evaluation [J]. BMC Bioinformatics, 2021, 22(1): 1-17.

[17] Sedlazeck F J, Rescheneder P, Smolka M, et al. Accurate detection of complex structural variations using single-molecule sequencing [J]. Nature Methods, 2018,

[18] Smolka M, Paulin L F, Grochowski C M, et al. Comprehensive Structural Variant Detection: From Mosaic to Population-Level [J]. bioRxiv, 2022,

[19] Lecompte L, Peterlongo P, Lavenier D, et al. SVJedi: Genotyping structural variations with long reads [J]. Cold Spring Harbor Laboratory, 2019, 17):

[20] Danecek P, Bonfield J K, Liddle J, et al. Twelve years of SAMtools and BCFtools [J]. GigaScience, 2021, 10(2):

[21] Jiang T, Liu Y, Jiang Y, et al. Long-read-based human genomic structural variation detection with cuteSV [J]. Genome Biology, 2020, 21(1): 189.

[22] Jiang T, Liu S, Cao S, et al. Structural Variant Detection from Long-Read Sequencing Data with cuteSV [M]//Ng C, Piscuoglio S. Variant Calling: Methods and Protocols. New York, NY; Springer US. 2022: 137-151.

[23] Jeffares D C, Jolly C, Hoti M, et al. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast [J]. Nature Communications, 2017, 8(1): 14061.

[24] David H, Martin V. SVIM: structural variant identification using mapped long reads [J]. Bioinformatics, 17): 2907-2915.

[25] Tham C Y, Tirado-Magallanes R, Goh Y, et al. NanoVar: accurate characterization of patients' genomic structural variants using low-depth nanopore sequencing [J]. Genome Biol, 2020, 21(1): 56.

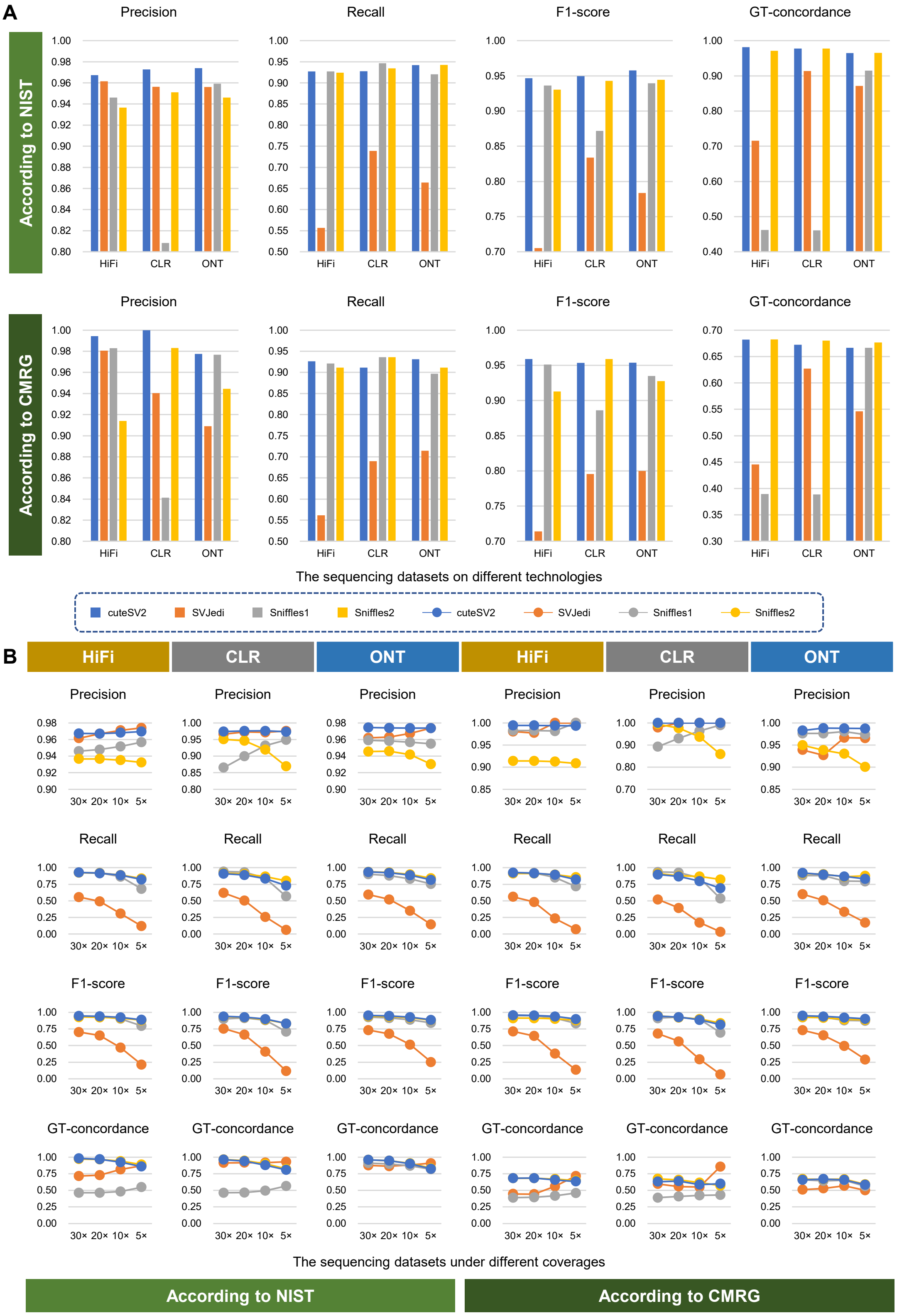


Fig 1. The benchmark results of the re-genotyping performance on the HG002 sample. A. The precision, recall, F1 score, and genotype concordance on different sequencing technologies including HiFi (PacBio HiFi sequencing), CLR (PacBio CLR sequencing), and ONT (Oxford Nanopore Technologies sequencing). The ground truth is Genome in a Bottle (GiaB) National Institute of Standards and Technology (NIST) ground truthsets (SV v0.6) and

the GIAB Challenging Medically Relevant Gene Benchmark (CMRG) v1.00, respectively. **B.** The precision, recall, F1 score, and genotype concordance under different sequencing coverages (i.e., 30×, 20×, 10×, and 5×) on different sequencing technologies (i.e., HiFi, CLR, and ONT) based on NIST and CMRG ground truth.

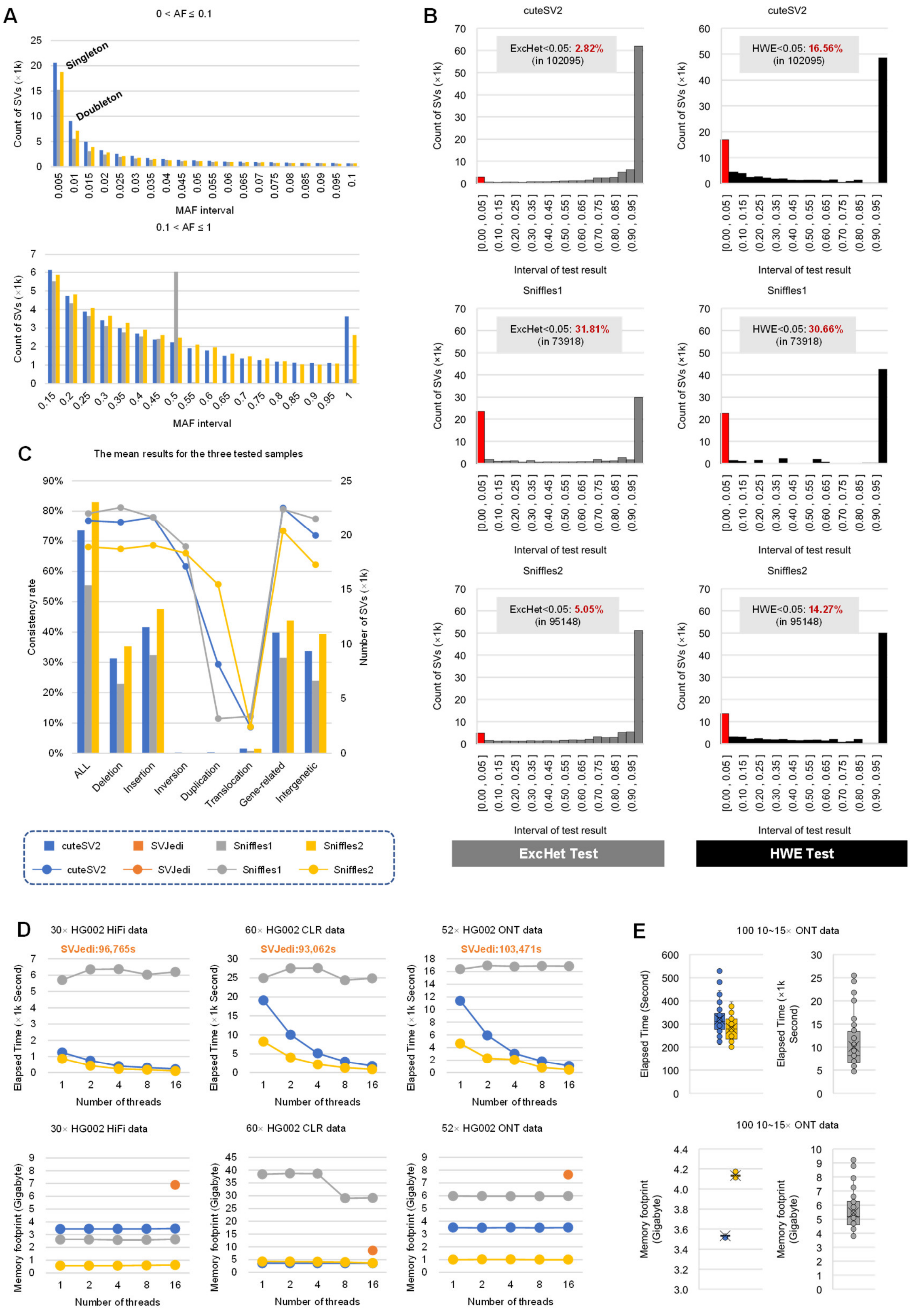
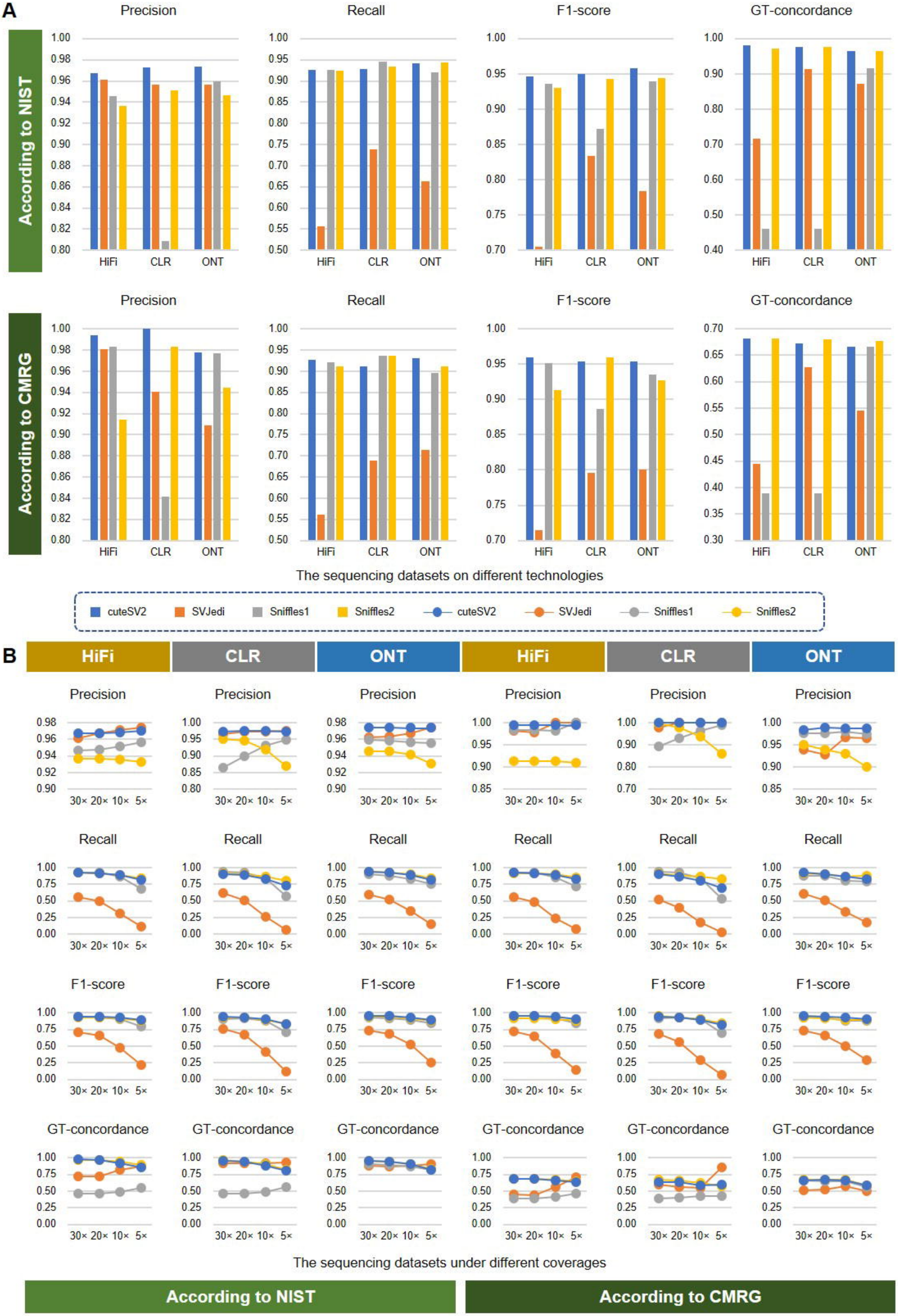
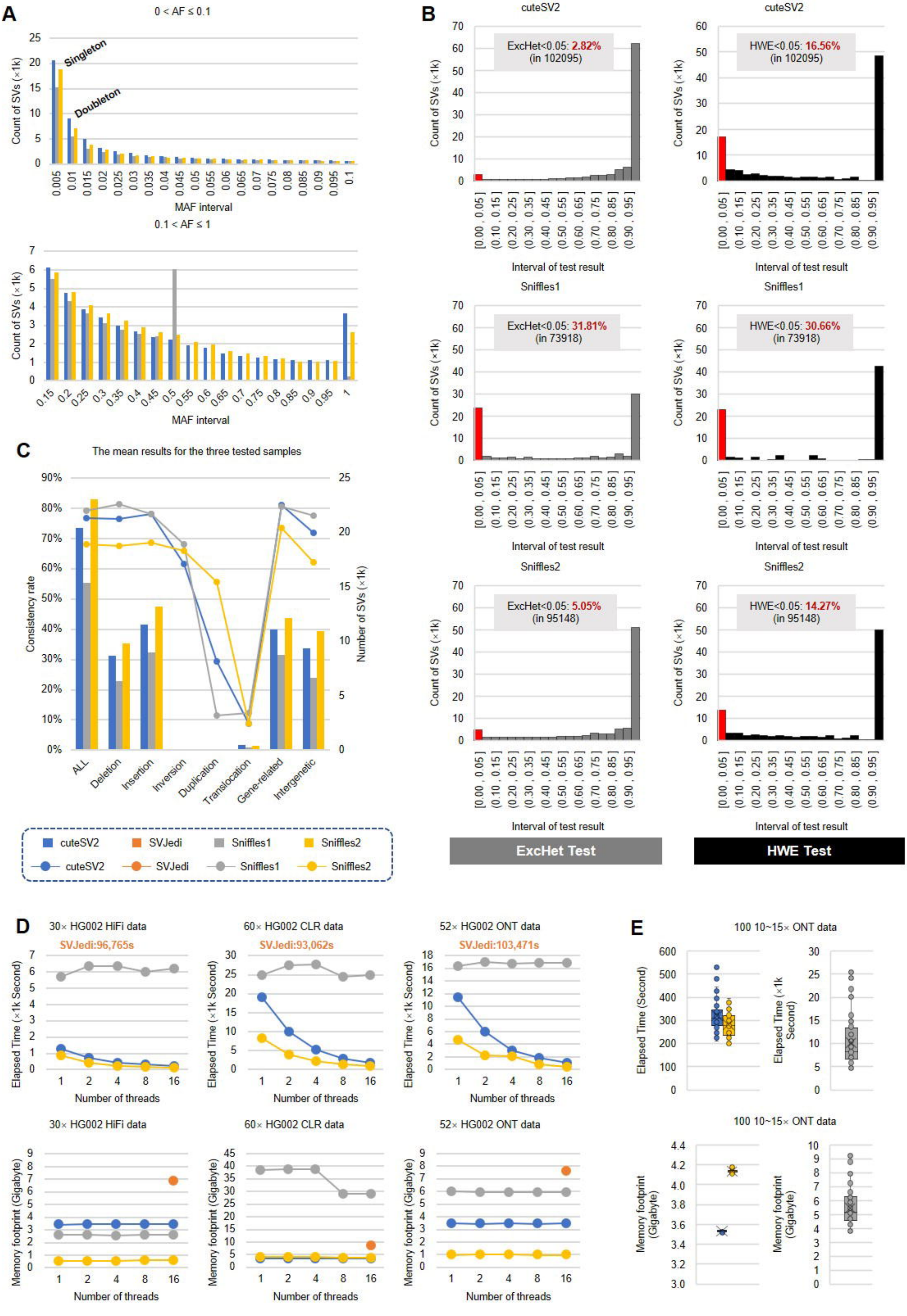


Fig 2. The benchmark results of the re-genotyping performance on the Chinese population group and the computational performance tests. A. The distribution of SV amount under different minor allele frequency (MAF) intervals. The above one shows the distribution of SVs whose MAF is smaller than 0.1 and the below one shows more common SVs with MAF larger than 0.1. It is also worth mentioning that the SVs in the regions of

MAF=0.005 and MAF=0.01 indicate singletons and doubletons respectively. **B.** The distribution of test scores of excess heterozygosity (ExcHet) and Hardy-Weinberg Equilibrium (HWE) for these three methods in the population group. The left three figures with gray columns represent the ExcHet scores, and the right with black columns represent the HWE scores. The red column indicates the SVs with test scores lower than 0.05 and demonstrates failures in the ExcHet or HWE test. **C.** The mean consistency rate for the three tested samples on different SV types (deletion, insertion, inversion, duplication, translocation) and different genomic regions (gene-related and inter-genetic regions). Each column and line represent the mean number of SV and mean consistency rate of the three samples, respectively. **D.** The elapsed time and memory footprint on the different HG002 datasets (PacBio HiFi, PacBio CLR, and ONT) with various threads (1, 2, 4, 8, 16). **E.** The box plot of the elapsed time and memory footprint on the 100 Chinese individuals.


A
According to NIST
Precision
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86
0.84
0.82
0.80
HiFi
CLR
ONT
Recall
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
HiFi
CLR
ONT
F1-score
1.00
0.95
0.90
0.85
0.80
0.75
0.70
HiFi
CLR
ONT
GT-concordance
1.00
0.90
0.80
0.70
0.60
0.50
0.40
HiFi
CLR
ONT
According to CMRG
Precision
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86
0.84
0.82
0.80
HiFi
CLR
ONT
Recall
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
HiFi
CLR
ONT
F1-score
1.00
0.95
0.90
0.85
0.80
0.75
0.70
HiFi
CLR
ONT
GT-concordance
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
HiFi
CLR
ONT
The sequencing datasets on different technologies
cuteSV2
SVJedi
Sniffles1
Sniffles2
cuteSV2
SVJedi
Sniffles1
Sniffles2
B
HiFi
CLR
ONT
HiFi
CLR
ONT
Precision
0.98
0.96
0.94
0.92
0.90
30× 20× 10× 5×
Precision
1.00
0.95
0.90
0.85
0.80
30× 20× 10× 5×
Precision
0.98
0.96
0.94
0.92
0.90
30× 20× 10× 5×
Precision
1.00
0.95
0.90
0.85
30× 20× 10× 5×
Precision
1.00
0.90
0.80
0.70
30× 20× 10× 5×
Precision
1.00
0.95
0.90
0.85
30× 20× 10× 5×
Recall
1.00
0.75
0.50
0.25
0.00
30× 20× 10× 5×
F1-score
1.00
0.75
0.50
0.25
0.00
30× 20× 10× 5×
GT-concordance
1.00
0.75
0.50
0.25
0.00
30× 20× 10× 5×
The sequencing datasets under different coverages
According to NIST
According to CMRG


A
0 < AF ≤ 0.1
Singleton
Doubleton
Count of SVs (×1k)
MAF interval
0.1 < AF ≤ 1
Count of SVs (×1k)
MAF interval
B
cuteSV2
ExcHet<0.05: 2.82% (in 102095)
cuteSV2
HWE<0.05: 16.56% (in 102095)
Sniffles1
ExcHet<0.05: 31.81% (in 73918)
Sniffles1
HWE<0.05: 30.66% (in 73918)
Sniffles2
ExcHet<0.05: 5.05% (in 95148)
Sniffles2
HWE<0.05: 14.27% (in 95148)
Count of SVs (×1k)
Interval of test result
ExcHet Test
HWE Test
C
The mean results for the three tested samples
Consistency rate
Number of SVs (×1k)
ALL
Deletion
Insertion
Inversion
Duplication
Translocation
Gene-related
Intergenetic
cuteSV2
SVJedi
Sniffles1
Sniffles2
D
30× HG002 HiFi data
SVJedi:96,765s
60× HG002 CLR data
SVJedi:93,062s
52× HG002 ONT data
SVJedi:103,471s
Elapsed Time (×1k Second)
Number of threads
Memory footprint (Gigabyte)
E
100 10~15× ONT data
Elapsed Time (Second)
Elapsed Time (×1k Second)
Memory footprint (Gigabyte)