FocalSV enables target region-based structural variant assembly and refinement using single-molecule long-read sequencing data

Can Luo, Zimeng Jamie Zhou, Yichen Henry Liu, et al.

Genome Res. published online August 13, 2025
Access the most recent version at doi:10.1101/gr.280282.124

METHODOLOGY ARTICLE

FocalSV enables target region-based structural variant assembly and refinement using single-molecule long-read sequencing data

Can Luo[1†], Zimeng Jamie Zhou[2†], Yichen Henry Liu[2] and Xin Maizie Zhou[1,2*]

Abstract

Structural variants (SVs) play a critical role in shaping the diversity of the human genome and their detection holds significant potential for advancing precision medicine. Despite notable progress in single-molecule long-read sequencing technologies, accurately identifying SV breakpoints and resolving their sequence remains a major challenge. Current alignment-based tools often struggle with precise breakpoint detection and sequence characterization, while whole genome assembly-based methods are computationally demanding and less practical for targeted analyses. Neither approach is ideally suited for scenarios where regions of interest are predefined and require precise SV characterization. To address this gap, we introduce FocalSV, a targeted SV detection framework that integrates both assembly- and alignment-based signals. By combining the precision of local assemblies with the efficiency of region-specific analysis, FocalSV enables more accurate SV detection. FocalSV supports user-defined target regions and can automatically identify and expand regions with potential structural variants to enable more comprehensive detection. FocalSV was evaluated on ten germline datasets and two paired normal-tumor cancer datasets, demonstrating superior performance in both precision and efficiency.

Keywords: structural variants; phasing; region-based; diploid assembly; long reads

*Correspondence:
maizie.zhou@vanderbilt.edu
[1]Department of Biomedical Engineering, Vanderbilt University, Nashville, TN 37235, USA
Full list of author information is available at the end of the article
†Equal contributor

Introduction

Single nucleotide variants (SNVs) constitute the most abundant form of genetic variation in humans and can be efficiently detected using short-read sequencing technologies. Therefore, genome-wide association studies (GWAS) have primarily focused on SNVs to investigate the genetic basis of phenotypic traits. In contrast, structural variants (SVs) - larger genomic alterations of 50 base pairs (bp) or more, including insertions (INS), deletions (DEL), duplications (DUP), inversions (INV), and translocations (TRA) - represent a major source of genetic diversity but are more challenging to detect accurately with short-read sequencing (Alkan et al. 2011). Despite their considerable biomedical relevance (Weischenfeldt et al. 2013), short-read-based approaches fail to identify over half of the SVs within an individual genome (Chaisson et al. 2019), constraining our ability to fully delineate the genetic landscape associated with complex disease phenotypes.

In recent years, long-read sequencing technologies have significantly advanced SV detection by providing extended read lengths, making it possible to resolve complex genomic regions that were previously challenging to analyze (Jain et al. 2016; Rhoads and Au 2015). Technologies such as PacBio's HiFi and Continuous Long Reads (CLR) and Oxford Nanopore Technology (ONT) offer distinct advantages. PacBio's HiFi reads, for instance, achieve accuracy of up to 99.9% with read lengths up to 20kb, making them particularly well-suited for capturing SVs in complex genomic contexts (Pacbio 2023). ONT, on the other hand, can generate even longer reads - exceeding 1 megabase on some cases - though with slightly lower accuracy. These long-read technologies allow for more accurate genomic alignment and improved assembly, significantly enhancing confidence in SV detection and enabling a more comprehensive view of the genomic landscape (Logsdon et al. 2020). However, despite these advantages, long-read technologies come with notable limitations. High-quality long reads remain costly and require substantial computational resources for processing, particularly when performing whole-genome assembly. These challenges present significant barriers in large-scale studies, where analyzing SVs across entire genomes for hundreds to thousands of samples is computationally intensive and time-consuming. Consequently, there is increasing demand for more efficient approaches that leverage long-read capabilities without the high computational cost of whole-genome assembly.

Alternative approaches, based on alignment-based methods, detect SVs by mapping long reads directly to a reference genome and identifying breakpoints based on alignment discrepancies. These methods are generally less computationally intensive and are well-suited for large-scale datasets. However, they face challenges in accurately identifying SV breakpoints and sequences compared to assembly-based methods (Liu et al. 2024). Limitations of existing whole genome assembly-based and alignment-based methods make them suboptimal for scenarios requiring precise SV characterization in predefined genomic regions.

To address this gap, we introduce FocalSV, a target region assembly-based SV detection tool that combines the precision of assembly-based methods with the efficiency of a region-specific approach. The region-specific

design of FocalSV is particularly valuable for clinical and genomic research, enabling users to focus on medically relevant SVs in specific loci or regions with SVs of interest. FocalSV enables both user-defined target analysis and automatic identification of SV-enriched regions, supporting more comprehensive and accurate structural variant detection. We benchmarked FocalSV against several state-of-the-art alignment-based and assembly-based tools, demonstrating overall superior accuracy and robustness across a wide variety of long-read datasets. FocalSV provides a practical and scalable alternative to whole-genome assembly, making it suitable for analyzing regions of interest in both individual samples and large-scale population studies.

Results

FocalSV is an efficient, region-aware tool for structural variant detection, offering two complementary modes - auto and target - to address diverse analytical goals (Figs. 1A-B and 2A-C). In target mode, users can specify regions of interest for focused SV detection. In auto mode, FocalSV autonomously detects and refines SV-rich regions by integrating population-level SV patterns with read-level signals from individual long-read data. To demonstrate the superior performance of FocalSV, we benchmarked both modes against four state-of-the-art assembly-based and five alignment-based SV detection tools across multiple long-read sequencing datasets. The assembly-based tools evaluated include FocalSV (v1.0.0), PAV (freeze2) (Ebert et al. 2021), SVIM-asm (v1.0.2) (Heller and Vingron 2020), Dipcall (v0.3) (Li et al. 2018), and sawfish (v1.0.1)(Saunders et al. 2025), while the alignment-based tools comprise cuteSV (v1.0.11) (Jiang et al. 2020), SVIM (v1.4.2) (Heller and Vingron 2019), pbsv (v2.6.2) (Biosciences 2021), Sniffles2 (v2.0.6) (Smolka et al. 2024), and SKSV (v1.0.2) (Liu et al. 2021).

The datasets include nine libraries derived from the well-characterized HG002 sample, sequenced with PacBio HiFi, CLR, and ONT technologies; one HiFi library from the Han Chinese male sample HG005; and two paired tumor-normal datasets (CLR and ONT) from the HCC1395 breast cancer cell line. Specifically, the HG002 datasets encompass three HiFi libraries (Hifi_L1, Hifi_L2, and Hifi_L3) with coverages ranging from 30× to 56×, three CLR libraries (CLR_L1, CLR_L2, and CLR_L3) with coverages between 29× and 89×, and three ONT libraries (ONT_L1, ONT_L2, and ONT_L3) with coverages between 47× and 57×. Further details on the tools and datasets are provided in Table 1.

We assessed the performance of all tools based on key metrics, including breakpoint identification and SV sequence accuracy. As illustrated in Figs. 1A-B and 2A-C, FocalSV's pipeline leverages region-based strategy to collect diverse SV signatures, enabling precise detection. A rigorous filtering process ensures that the final SV call sets are both comprehensive and accurate, highlighting FocalSV's ability to capture a wide spectrum of structural variants. The complete methodology and implementation details are described in the Methods section.

Evaluation of SV calls in FocalSV and comparisons with existing SV callers

To assess the performance of insertion and deletion SV detection, we evaluated five assembly-based tools (FocalSV, PAV, SVIM-asm, Dipcall, and sawfish) and five alignment-based tools (cuteSV, SVIM, pbsv, Sniffles2, and SKSV) using nine long-read sequencing libraries of the HG002 sample. For FocalSV, we evaluated two operational modes: FocalSV(target), which uses predefined regions, and FocalSV(auto), which automatically identifies and expands regions with potential SVs. SV calls were benchmarked against the Genome in a Bottle (GIAB) SV gold standard (Zook et al. 2014) using Truvari (v4.0.0) (English et al. 2022), a widely adopted structural variant evaluation tool. Truvari compares SV calls from any tool's call set against a gold standard call set, both provided in Variant Call Format (VCF) files, by analyzing key metrics: reference distance, reciprocal overlap, size similarity, and sequence similarity.

In this study, we applied a moderate-tolerance parameter set in Truvari for SV comparisons, with the following settings: p=0.5, P=0.5, r=500, and O=0.01. Specifically, the parameter p (pctstim), which ranges from 0 to 1.0, controls the minimum required sequence similarity for two SVs to be considered identical. P (pctsize), also ranging from 0 to 1.0, defines the minimum allowable allele size similarity. O (pctovl), ranging from 0 to 1.0, establishes the minimum reciprocal overlap ratio, a key measure for comparing deletions and evaluating their breakpoint alignment. Lastly, r (refdist), which can range from 0 to 1000bp, sets the maximum allowable difference between the reference positions of two SVs, helping assess breakpoint shifts in insertion events.

We first evaluated the average performance across different PacBio Hifi, CLR, and ONT datasets (Tables 2-4). In the HiFi datasets (Table 2), FocalSV(target)demonstrated the third-highest average F1 score for deletions (93.99%) and the highest for insertions (91.93%). FocalSV(auto) attained the highest average F1 score for deletions (94.29%) and the second-highest for insertions (91.03%). In the CLR datasets (Table 3), FocalSV(target) outperformed the other tools with the highest average F1 scores for deletions (92.95%) and insertions (90.71%). FocalSV(auto) ranked second, with comparable performance for deletions (92.92%) and slightly lower performance for insertions (89.16%). In the ONT datasets (Table 4), FocalSV(target) achieved the highest average F1 score for deletions (92.94%), and ranked third for insertions (89.69%). FocalSV(auto) ranked second in average F1 score for deletions (92.50%), and achieved the highest F1 score for insertions (90.01%). These results highlight the robustness of both FocalSV modes across different sequencing platforms, with either FocalSV(target) or FocalSV(auto) consistently ranking first or second in average F1 scores for deletions and insertions across nearly all datasets. In terms of overall genotyping accuracy, cuteSV and FocalSV consistently performed best, with accuracies frequently in the 98-99% range.

When examining each dataset individually (Tables 2-4 and Supplemental Tables S1-S3), FocalSV outperformed all other tools, achieving the highest F1 scores for both deletions and insertions across all Hifi, CLR,

and ONT libraries. In the context of three Hifi datasets (Table 2 and Supplemental Table S1), FocalSV emerged as the top overall performer. For deletions, FocalSV(auto) outperformed all other tools on Hifi_L2 and Hifi_L3, exceeding the F1 score of the second-best tool, sawfish, by an average of 0.12%. On Hifi_L1, FocalSV(target) achieved the highest performance, closely followed by sawfish and FocalSV(auto). For insertions, both FocalSV(target) and FocalSV(auto) consistently outperformed all other tools across all three libraries, with FocalSV(target) achieving F1 score that surpassed the next-best tools by an average of 1.80%.

In the three CLR datasets (Table 3 and Supplemental Table S2), FocalSV emerged as the top-performing tool in terms of F1 score and recall across nearly all libraries, demonstrating clear advantages. For deletions, FocalSV(auto) outperformed the second-ranked tools in F1 score on CLR_L2 and CLR_L3 by an average of 0.39%. On CLR_L1, FocalSV(target) exhibited the highest F1 score, closely followed by PBSV and FocalSV(auto). For deletion recall, FocalSV(target) ranked first, outperforming the second-ranked tools across all three libraries by an average of 1.48%. For insertions, both FocalSV(target) and FocalSV(auto) excelled across all libraries, with FocalSV(target) achieving an average F1 score 3.45% higher than the second-ranked tools. For insertion recall, FocalSV(target) exceeded the second-ranked tools on CLR_L1 and CLR_L2 by an average of 0.93%. On CLR_L3, FocalSV(auto) exhibited the second-highest recall, trailing the top by 0.23%.

In the three ONT datasets (Table 4 and Supplemental Table S3), both FocalSV(target) and FocalSV(auto) maintained substantial leads. FocalSV(target) outperformed the second-ranked tool across ONT_L1, ONT_L2, and ONT_L3 by an average of 0.58% in deletion F1 score. For deletion recall, FocalSV(target) outperformed the second-ranked tools on ONT_L1 and ONT_L2 by an average of 0.58%. On ONT_L3, FocalSV(target) achieved the second-highest recall, trailing the top by 0.10%. For insertion F1, FocalSV(auto) achieved the highest F1 scores on ONT_L1, ONT_L2, and ONT_L3, outperforming the second-ranked tool, cuteSV, by an average of 0.30%. In terms of insertion recall, FocalSV(target) outperformed the second-ranked tool on ONT_L2 and ONT_L3 by an average of 0.51%. On ONT_L1, FocalSV(target) achieved the second-highest recall, trailing the top recall by 0.49%.

For these benchmarks, FocalSV(target) used regions of interest defined by SVs identified by GIAB. To avoid biased evaluation, we also measured false positive (FP) counts within regions lacking known SVs. Recognizing that some tools may report fewer SVs - leading to lower recall but artificially high precision, we further normalized FP counts by penalizing them according to each tool's average F1 score (Powers 2020; Sasaki et al. 2007), using the formula: $normalized\ FP\ count = raw\ FP\ count \times (1 - F1)$. To evaluate performance in SV-negative regions, we randomly selected 1,000 regions without known SVs using the GIAB gold standard. All tools produced very few false positives. FocalSV(target) achieved the lowest FP counts across both raw (CLR_L1: 5; ONT_L1: 3) and normalized values (CLR_L1: 0.4; ONT_L1: 0.26). For HiFi data, it showed the second-lowest raw (Hifi_L1: 6) and lowest normalized FP value (Hifi_L1: 0.4)

(Supplemental Fig. S1). These results demonstrate FocalSV(target)'s consistently high precision across both SV-rich and SV-negative regions.

In summary, FocalSV consistently outperformed other tools across nine HG002 datasets (Hifi, CLR, and ONT) based on average F1 scores and library-specific performance, demonstrating strong detection capabilities for both deletions and insertions. Although it occasionally ranked second or third in recall or precision, these variations were minor and did not diminish its overall effectiveness. In terms of genotyping accuracy, FocalSV consistently ranked first or second, with accuracy rates frequently reaching 98-99%. These results highlight FocalSV as a reliable and high-performing tool for structural variant detection.

FocalSV demonstrates excellent resilience to different SV size ranges

So far, we evaluated SVs by averaging across all size ranges. However, we found that the tools we assessed showed varying levels of accuracy in detecting SVs of different sizes. To demonstrate the effect of SV size and compare the resilience of each tool, we plotted the F1 score against the range of SV sizes (Fig. 3 and Supplemental Fig. S2). We utilized the same modest tolerance parameters in Truvari for benchmarking. FocalSV(auto) and FocalSV(target), along with all other tools, are visualized in Fig. 3 and Supplemental Fig. S2, respectively. Both modes of FocalSV exhibited comparable performance across all variant size ranges and libraries.

Overall, assembly-based tools such as FocalSV(auto), FocalSV(target), PAV, SVIM-asm, Dipcall, and sawfish showed better resilience than alignment tools such as SKSV, Sniffles2, PBSV, SVIM, and cuteSV when processing HiFi data (Fig. 3A-C and Supplemental Fig. S2A-C). Most alignment-based tools had a significant drop in performance with large INSs in the range of 2kb-50kb, while most assembly-based tools typically encountered performance challenges only with large INSs in the range of 9kb-50kb. Among all assembly-based tools, FocalSV, in both modes, continued to distinguish itself with exceptional performance across nearly all size ranges, except for 8kb-10kb deletions and 10kb-50kb insertions.

On CLR data (Fig. 3D-F and Supplemental Fig. S2D-F), assembly-based tools except Dipcall still outperformed alignment-based tools in terms of resilience, especially for INSs. Across all size ranges, FocalSV, in both modes, maintained a top-tier performance. Notably, FocalSV's performance is much better than other tools on CLR_L3, where it had the highest F1 score across all size ranges except for INSs in the range of 9kb-10kb and showed the least fluctuation.

On ONT data (Fig. 3G-I and Supplemental Fig. S2G-I), the difference between alignment-based and assembly-based tools was less pronounced. Two alignment-based tools, cuteSV and Sniffles2, demonstrated resilience comparable to assembly-based tools. FocalSV, in both modes, emerged as a top-performing tool

for small to large-sized deletions (50bp-8kb) and insertions (50bp-2kb). For large deletions (8kb-50kb) and insertions (2kb-50kb), FocalSV, in both modes, remained within the range of the best tools.

FocalSV achieves robust SV detection across evaluation parameters

While assessing different SV callers against a gold standard, we recognized the potential impact of breakpoint shifts and sequence similarity issues on the evaluation. This acknowledgment arises from the fact that SVs often span substantial genomic regions. In previous evaluations, we selected a set of moderate-tolerance but fixed parameters to demonstrate the overall performance for each tool. However, the choice of parameters such as breakpoint shift tolerance and sequence similarity for identifying a call as a true positive, varies subjectively. Therefore, to comprehensively evaluate the effectiveness and stability of SV callers, we adjusted key evaluation parameters (p, P, r, and O) using Truvari for benchmarking (English et al. 2022). Specifically, we systematically evaluated the impact of four key parameters by conducting two complementary experiments: (1) a univariate sensitivity analysis, where each parameter was varied individually to assess its effect on F1 scores while holding others constant; and (2) a grid-based benchmarking experiment, in which all pairwise parameter combinations were explored to evaluate tool performance across the full parameter space. Parameters p, P, and O were adjusted in increments of 0.1 from 0 to 1, while r was adjusted in 100bp increments from 0 to 1000bp. Our goal was to examine how different levels of stringency or leniency in these parameters might impact SV detection across the nine datasets. We expected FocalSV to demonstrate relatively consistent and stable performance, even under more stringent conditions.

We first systematically varied each parameter and plotted curves to illustrate the changes in F1 scores across all libraries. When varying one parameter, the remaining parameters were held at their default moderate values. For deletion and insertion detection on Hifi_L1 (Fig. 4A-B), increasing the stringency of the matching threshold by reducing r from 1000bp to 0bp led to a consistent decline in F1 scores for all tools, with insertions showing greater sensitivity to r than deletions. When varying O, P, or p from 0 to 1.0 - corresponding to increasing stringency (Fig. 4C-H) - F1 scores of all tools decreased. Notably, for P and p, F1 scores for both deletions and insertions dropped sharply when the thresholds exceeded 0.8. Changes in O had an even stronger impact: both deletions and insertions exhibited a substantial decline in F1 performance when O increased, with insertions again being more sensitive than deletions. Across all parameter settings, FocalSV, in both modes, consistently achieved the highest F1 scores. Similar trends were observed across other libraries, with FocalSV maintaining top performance in nearly all scenarios (Supplemental Figs. S3-S10).

We next varied each parameter pair and plotted F1 heatmaps to illustrate performance changes across all libraries. For deletion calls on Hifi_L1, The F1 heatmaps for FocalSV(auto) and FocalSV(target) are shown in Fig. 5A-I and Supplemental Fig. S11, respectively. Both modes demonstrated comparable performance

across all libraries under varying evaluation parameters. Based on the comprehensive benchmarking study by Liu et al. (Liu et al. 2024), which thoroughly examined the impact of various parameter combinations, such as p-O, P-r, O-r, p-P, p-r, and P-O, on deletion evaluation, we selected p and O as the representative parameter pair, as they had the most substantial influence on deletion performance. As shown in Fig. 5A-I and Supplemental Fig. S11, increasing the values of p and O imposed stricter correspondence requirements between the SV call and the gold standard, leading to a decline in F1 scores. Liu et al.'s heatmap and gradient analysis demonstrated distinct performance patterns among tools as more stringent thresholds were applied. They showed that all read alignment-based SV callers, except for pbsv, exhibited significant performance drops when p or O exceeded 0.7, with F1 scores falling below 5% when either parameter reached 1.0. In contrast, FocalSV, in both modes, and other assembly-based SV callers from Liu et al.'s study, namely Dipcall, SVIM-asm, and PAV, along with pbsv, maintained stable performance across the parameter grid. Even under the strictest conditions (p = 1.0, O = 1.0), these tools achieved F1 scores greater than 69%. Notably, under this strict exact-match criterion in terms of sequence similarity and reciprocal overlap ratio, FocalSV, in both modes, excelled with an exceptional F1 score above 74% (Fig. 5A and Supplemental Fig. S11A), outperforming other stable tools and demonstrating its robustness. Across nearly all parameter settings, FocalSV consistently achieved higher F1 scores compared to Dipcall, SVIM-asm, PAV, pbsv, and sawfish (Supplemental Figs. S12-S15A), demonstrating superior performance across the grid. The overall performance trends for these tools were consistent with those previously reported by Liu et al (Supplemental Figs. S12-S20A).

For insertion evaluations on Hifi_L1 (Fig. 6A and Supplemental Figs. S21-S30A), the pair p and r were similarly selected as the most representative parameters based on Liu et al.'s findings (Liu et al. 2024). Their study revealed that insertion detection was generally more sensitive to parameter variations compared to deletions. Combining their results and our experiment, we found that most tools, including FocalSV, experienced a decline in F1 scores when p exceeded 0.8 (i.e. when the required allele sequence similarity between the called SV and the benchmark exceeded 80%) (Fig. 6A and Supplemental Figs. S21-S30A). Additionally, F1 scores decreased when r was reduced. Once r dropped to 200bp or less, indicating a more stringent reference distance threshold, performance deteriorated significantly across all tools. Similar to the trends observed for deletions, FocalSV, other assembly-based tools, and pbsv displayed greater robustness to stringent parameters compared to read alignment-based tools. Among all robust performing tools, FocalSV, in both modes, consistently achieved higher F1 scores than Dipcall, SVIM-asm, PAV, and pbsv, showing outstanding performance over the grid (Fig. 6A and Supplemental Figs. S21-S25A). We have also observed similar patterns on grid searches for FocalSV on other libraries (Fig. 5B-I, Fig. 6B-I, and Supplemental Figs. S12-S30).

The robustness of assembly-based tools to evaluation parameters may indicate their capacity to accurately detect SV breakpoints and alternative allele sequences, as highlighted by Liu et al. To address this, we analyzed the distribution of SV breakpoint shift and alternate allele sequence similarity for FocalSV. Breakpoint shifts were defined as the maximum reference location difference between two compared SV calls, calculated as the greatest start/end location difference between true positive SVs and their corresponding benchmark SVs. Sequence similarity was quantified using the edit distance between compared SV calls, directly extracted from Truvari. Our results revealed that FocalSV achieved a near-zero breakpoint shift and near 100% SV sequence similarity with the benchmark callset (Supplemental Figs. S31-S34). In line with previous findings, our results confirm that FocalSV, like other assembly-based tools, demonstrates strong resilience across various parameter settings, underscoring its capacity to accurately capture both SV breakpoints and alternative alleles. Notably, FocalSV consistently outperformed other assembly-based tools, including Dipcall, SVIM-asm, and PAV, across all parameter configurations, further emphasizing its superior accuracy and reliability.

FocalSV exhibits reliable SV detection on HG005, validated by multi-tool consensus benchmarking

To further assess the robustness and generalizability of FocalSV, we extended our evaluation to the Chinese male sample HG005. As the GIAB consortium has not provided a gold-standard SV callset for this sample, we developed a consensus-based evaluation framework leveraging agreement across multiple established SV callers. We first applied all benchmarked tools to the HG005 HiFi dataset (Table 1) and collected their respective callsets based on the GRCh38 reference genome. From each, we extracted indel SVs ($\geq$50 bp) marked as "PASS" in the VCF filter field and located within the high-confidence regions defined by GIAB. Although originally curated for HG002, this BED file encompasses regions with consistent support across sequencing platforms and excludes problematic areas such as segmental duplications and low-mappability areas. Given its stringent design, we applied it to HG005 to enable fair and broadly comparable evaluations. Following filtering, we merged and sorted the resulting VCF files, then applied distance-based clustering to group SVs of the same type within a 500bp window. These clusters served as proxy gold-standard sets for HG005, with each cluster annotated by the number of supporting tools. By adjusting the minimum support threshold, we generated benchmark sets at varying confidence levels. SV callers were evaluated by comparing their predictions to these clusters: a call was considered a true positive if it overlapped a benchmark cluster and matched the majority genotype. SVs without matching clusters were treated as false positives. We computed recall, precision, and F1 scores across multiple confidence thresholds to quantify performance.

Overall, FocalSV(auto) demonstrated leading performance in deletion detection across a wide range of confidence thresholds. As shown in Supplemental Fig. S35, FocalSV consistently achieved the highest F1 scores

when the minimum number of supporting tools ranged from 2 to 7. For instance, at the stringent threshold of 7 supporting tools, FocalSV reached an F1 score of 97.66% (Supplemental Table S4), outperforming the second-best tool, Sniffles2, by 0.19%. At thresholds above 7, cuteSV ranked highest, with FocalSV following closely in second place. The number of benchmark deletions ranged from 3,985 to 4,710 (Supplemental Fig. S35D and Supplemental Table S4), closely matching the 4,116 tier 1 deletions reported for HG002 in the high-confidence BED file. This consistency supports the reliability of our filtering and evaluation framework.

For insertion calls, FocalSV maintained strong performance, ranking first or second in F1 score across most confidence thresholds. As shown in Supplemental Fig. S36 and Supplemental Table S5, FocalSV achieved the highest F1 scores when the number of supporting tools was 2 or 3. At thresholds between 4 and 8, FocalSV consistently ranked second, closely following PAV, with an average F1 difference of 0.3%. When the threshold increased to 9 supporting tools, FocalSV ranked third, trailing the top performer by 0.7%. At the 10-tool threshold, F1 scores dropped substantially for most tools, except PBSV, which reported significantly fewer insertions than others. This suggests the benchmark set at this threshold may be incomplete. As a result, recall becomes a more informative metric under such stringent criteria, and FocalSV achieved the highest recall among all tools. The number of benchmark insertions ranged from 4,231 to 6,238 as the support threshold increased from 2 to 10, closely aligning with the 5,281 tier 1 insertions reported in the HG002 high-confidence BED file, further validating the soundness of our evaluation framework.

In summary, FocalSV achieved the highest overall performance in deletion detection and consistently ranked among the top performers for insertion detection on the HG005 dataset, highlighting its robustness, reliability, and broad applicability across diverse genomic contexts.

FocalSV maintains strong SV detection performance regardless of variations in phasing quality

FocalSV utilizes Longshot for phasing and read partitioning into two haplotypes prior to indel SV assembly; therefore, phasing quality can directly influence downstream SV detection accuracy. We assessed phasing accuracy using SNP-based switch error metrics (Choi 2020) and PhaseQ scores derived from Strand-seq data, as detailed in the Supplemental Methods and Supplemental Fig. S37. Across HiFi, CLR, and ONT datasets, we observed low switch error rates (median: 0%; mean: 2.2–2.4%) and high proportions of correctly phased heterozygous variants (median: 97.8–100%; mean: 93.3–95.6%), with 68% of regions achieving 0% SER (Supplemental Fig. S38). Strand-seq analysis further confirmed robust phasing, with over 80% of phase blocks showing PhaseQ values >0.8 (Supplemental Fig. S39). Additional results are provided in the Supplemental Results.

To evaluate phasing quality, we further examined the distribution of phased read percentages across all target regions. Both HiFi and ONT datasets exhibited high phasing rates (HiFi_L1: mean 83-87%, median

97-98% across both modes; ONT_L1: mean 82-88%, median 84-90%; Supplemental Fig. S40), whereas the CLR dataset displayed lower phasing rates (mean 49-52%, median 57-58%; Supplemental Fig. S40). To assess the effect of phasing on SV detection, we ranked all target regions by phased read percentage and divided them into high and low phasing groups. SV detection performance was then compared between these groups. Regions with higher phasing percentages showed moderately improved F1 scores across HiFi, CLR, and ONT datasets, particularly for deletions. For instance, deletion F1 scores for high vs. low phasing regions were 95.1% vs. 93.4% (HiFi_L1), 93.3% vs. 91.8% (CLR_L1), and 94.5% vs. 91.5% (ONT_L1) (Supplemental Table S6). In contrast, insertion detection was less affected, with F1 scores of 91.2% vs. 91.2% (HiFi_L1), 90.1% vs. 89.1% (CLR_L1), and 92.3% vs. 90.2% (ONT_L1) (Supplemental Table S6). Overall, FocalSV demonstrated high phasing accuracy for both SNP phasing and local assemblies, while maintaining robust structural variant detection performance even in regions with lower proportions of phasing-informative reads.

FocalSV achieves superior performance in complex somatic SV detection in cancer data

To extend the evaluation to include SV detection beyond deletions and insertions, we analyzed additional SV types involving complex DNA rearrangements, such as translocations (TRA), inversions (INV), and duplications (DUP). FocalSV and other relevant tools were applied to detect somatic SVs in two publicly available cancer libraries (Fang et al. 2021). Dipcall was excluded from this analysis, as it was not designed to detect these three types of SVs. Although PAV is designed to detect inversions, it failed to detect any inversions in the cancer data and was therefore also excluded from this analysis.

We conducted a comparative analysis by applying the selected tools to two publicly available tumor-normal paired libraries (PacBio CLR and ONT), as provided by Talsania et al (Talsania et al. 2022). These libraries, along with the high-confidence HCC1395 somatic SV call set serving as the benchmark gold standard, formed the basis for evaluating the detection of three classes of somatic SVs. Each benchmarked tool was first applied independently to each library, generating VCF files. These VCFs were then processed using SURVIVOR (Jeffares et al. 2017) to identify somatic variants by comparing paired normal-tumor VCFs. Detailed method to detect somatic SVs is provided in the Supplemental Methods section. The identified somatic SVs were subsequently compared to the gold standard, which contained 137 translocations, 133 inversions, and 230 duplications.

FocalSV(auto) outperformed other tools (SVIM-asm, cuteSV, SVIM, pbsv, and Sniffles2) in terms of F1 score across nearly all sequencing libraries and SV types, except for inversions in ONT data (Table 5). Specifically, on PacBio CLR data, FocalSV(auto) outperformed the second-best tools in F1 score by 4.7%, 1.2%, and 7.6% for translocations, inversions, and duplications, respectively. On ONT data, FocalSV(auto) exceeded the second-best tools by 29.6% and 6.3% for translocations and duplications, respectively. For

inversions, Sniffle2 achieved the highest F1 score (30.4%), followed by FocalSV(auto) (20.2%). We also applied FocalSV(target) to directly detect the high-confidence somatic SV call set based on predefined target regions. As expected, it achieved the highest F1 scores across all scenarios, followed by FocalSV(auto).

Computation cost of FocalSV

Finally, we evaluated FocalSV's memory requirement and runtime performance. For a relatively large SV - an insertion of 12.6kb on Chromosome 21 - FocalSV finished execution in 32 seconds of elapsed time and 5 minutes and 20 seconds of CPU time, using a maximum 5.6MB of memory when run with 8 threads. In comparison, detecting the smallest SV - a deletion of 50bp on Chromosome 21 - took a similar elapsed time of 33 seconds and required 5 minutes and 30 seconds for CPU time, with a higher memory usage of 0.76GB. On average, the CPU time was approximately 5.5 minutes for one target region, reflecting consistent computational effort regardless of SV size.

Based on this estimate, detecting 20,000 SVs would require approximately 1,800 total CPU hours - comparable to the runtime for FocalSV(auto) to assemble a single library. As such, assembling all SVs across an entire genome using FocalSV can still be more computationally demanding than alignment-based methods. Importantly, we did not compare FocalSV's runtime to existing SV callers, as its time efficiency is specifically optimized for targeted analyses involving a limited number of regions rather than whole-genome discovery. While assembly-based SV detection is generally resource-intensive, FocalSV's region-based design enables efficient and rapid analysis when users focus on selected loci of interest.

FocalSV is designed to support multi-level parallelization, allowing tasks to be efficiently distributed across multiple servers, with each server assigned to distinct genomic regions. Within each region, FocalSV utilizes multithreading to manage assembly and variant detection tasks, maximizing CPU core utilization.

Discussion

FocalSV is a high-performance, region-aware tool for structural variant detection. It supports two complementary operational modes - auto and target - designed to accommodate distinct analytical scenarios. The auto mode autonomously identifies and refines SV-enriched regions by integrating population-scale patterns with individual read-level signals. This makes it particularly suitable for discovery-driven analyses where no prior SV information is available, offering a scalable alternative to whole-genome assembly. In contrast, the target mode is optimized for hypothesis-driven studies, allowing users to input specific regions of interest, such as those derived from existing SV annotations or clinically relevant loci. By combining the accuracy of assembly-based methods with the efficiency of region-specific processing, FocalSV delivers robust and

scalable performance across diverse sequencing technologies and SV types. Together, these modes establish FocalSV as a flexible and effective solution for both exploratory and targeted structural variant analyses.

Across nine HG002 long-read datasets, FocalSV demonstrated consistently strong and reliable performance in detecting large indel SVs. Both auto and target modes ranked among the top two tools in F1 score across PacBio HiFi, CLR, and ONT platforms, while maintaining high genotyping accuracy (>98%) across nearly all datasets. FocalSV also showed robust performance across a broad range of SV sizes and evaluation criteria, highlighting its versatility for structural variant analysis in long-read sequencing. Furthermore, its strong performance on the HG005 sample underscores the tool's robustness and generalizability across different individuals.

FocalSV also showed strong capabilities in detecting translocations, inversions, and duplications in tumor-normal paired cancer datasets. Across multiple sequencing platforms, both modes outperformed existing tools. The target mode, guided by known SV regions, achieved the highest accuracy across all variant types, while the auto mode performed robustly without prior region knowledge. It is worth noting that FocalSV is designed to rely primarily on read alignment-based signals rather than contig-based assemblies for detecting translocations and inversions. This decision is motivated by the practical challenge of assembling these SV types, even when sufficient supporting reads are present. We recover most duplications indirectly, through assembled insertions. Our recent benchmark study (Liu et al. 2024) further supports this strategy, showing that most assembly-based tools are not optimized for translocations, inversions, and duplications, whereas alignment-based approaches often yield better sensitivity and precision. These findings support our choice to prioritize alignment-based evidence for these SV classes.

Methods

We present FocalSV, a targeted SV detection framework that integrates assembly- and alignment-based signals through region-specific analysis, enabling accurate detection and characterization across all SV types: insertions (INSs), deletions (DELs), duplications (DUPs), translocations (TRAs), and inversions (INVs). FocalSV operates in two modes: target and auto. In target mode (referred to as FocalSV(target)), it uses user-defined regions of interest, while in auto mode (FocalSV(auto)), it automatically identifies and expands regions likely to contain SVs.

Large INS and DEL assembly

FocalSV takes the whole genome aligned reads BAM file along with either user-specified target regions or automatically detected candidate regions to perform assembly and refinement of large indel SV (≥50bp) (Fig. 1). FocalSV(auto) employs a customized algorithm to detect potential SV regions, with details provided in

the following section. These candidate regions are saved in a BED file. Once target regions are defined - either automatically or manually - they are used for a target region-based workflow.

This workflow for identifying a single target SV consists of the following steps: (1) Extraction of region-specific BAM file, (2) partitioning of reads, (3) haplotype-aware local assembly, (4) detection of candidate SV, and (5) filtering of indel SVs and refinement of genotypes.

Automatic identification of regions with potential SVs

To identify genomic regions likely to harbor SVs, we integrated two complementary sources of evidence: (1) read-level SV breakpoint signatures extracted from individual long-read alignments, and (2) population-scale SV breakpoints obtained from pangenome-based multi-sample SV catalogs. Each source was independently processed to capture both sample-specific and recurrent population variation, and their results were subsequently merged through distance-based clustering to generate a comprehensive, high-confidence set of candidate SV intervals as input for FocalSV(auto).

Read-based SV signature clustering. For each sequencing library, we first extracted candidate SV signatures from the read-to-reference alignment (BAM) file. To comprehensively detect insertions and deletions from long-read alignments, we integrated both intra-alignment and inter-alignment signatures. Intra-alignment signatures were derived directly from the primary alignment's CIGAR string, where insertions were identified from 'I' operations and deletions from 'D' operations, each exceeding a user-defined minimum length threshold (default: 30bp). Insertions were represented as triplets (P, L, S), where P is the aligned reference position, L is the inserted sequence length, and S is the inserted sequence retrieved from the read. Deletions were recorded as pairs (P, L), where P indicates the reference start position and L the deletion length. To mitigate redundancy caused by alignment fragmentation, adjacent intra-alignment signals within the same read were merged if their positions were within a user-defined proximity threshold. In contrast, inter-alignment signatures were extracted from split-read mappings by utilizing supplementary alignment (SA) tags to identify reads with multiple alignment segments. For each pair of consecutive segments mapped to the same chromosome and strand, we calculated the differences between the read and reference coordinates. A large gap in the read combined with a small gap in the reference indicated an deletion, whereas a large gap in the reference with minimal read gap signified an insertion. These inter-alignment signatures enabled detection of SVs that span multiple alignment blocks, large insertions or deletions, or complex breakpoints often fragmented by long-read alignment algorithms. By integrating both intra- and inter-alignment signals, the method improves sensitivity for detecting large and alignment-fragmented SVs.

Subsequently, for each large indel SV type, SV signatures were initially clustered based on genomic proximity using a fine-scale distance threshold (dt_{fine}, default: 500bp) to group nearby events. Clusters supported

by fewer than a minimum number of individual signatures (min_sig, default: 0.2 × average coverage) were discarded. Each retained cluster was represented as a genomic interval spanning the first to last breakpoint in the group. To address potential redundancy and positional uncertainty, a second, coarser round of clustering was performed using a broader threshold (dt_{coarse}, default: 15kb), yielding a refined set of candidate SV regions derived from read-level evidence.

Pangenome-based SV clustering. Inspired by recent advances in pangenome graph-based SV genotyping for short-read data (Ebler et al. 2022), we leveraged population-scale variation data by incorporating breakpoint positions from a multi-sample VCF file to supplement individual-level SV signals. This same VCF also provides the variant catalog used for pangenome graph construction in Pangenie (Ebler et al. 2022), enabling improved sensitivity and consistency in genotyping across diverse individuals. All non-header records were parsed to extract SV breakpoints, which were grouped by chromosome and sorted by genomic position. A clustering procedure, similar to read-based breakpoint clustering, was applied using the same coarse distance threshold (dt_{coarse}, default: 15kb). Specifically, SV breakpoints located within the specified threshold were merged into clusters, producing a population-informed set of candidate SV regions.

To avoid bias in our evaluation, samples HG002 and HG005 were excluded from the multi-sample VCF input. The VCF was constructed using haplotype-resolved assemblies from nine individuals (HG01109, HG01243, HG02055, HG02080, HG02109, HG02145, HG02723, HG03098, and HG03492). SVs were identified from the assemblies and merged using the Pangenie vcf-merging pipeline (Ebler et al. 2022). This multi-sample VCF serves as the default population reference for FocalSV(auto), though users may substitute any population-scale, pangenome graph-based SV catalog suitable for their specific study population. As additional high-quality HiFi assemblies become available, FocalSV(auto) can leverage increasingly comprehensive population variation, improving sensitivity for recurrent, low-frequency, or difficult-to-resolve SVs that may be missed in single-sample analyses.

Integration of candidate SV regions from read-level and pangenome-based evidence. The final set of SV candidate regions was generated by merging the refined intervals from both sources: read-level signatures extracted from individual BAM files and population-based SV calls. For each chromosome, intervals from both sources were combined and re-clustered using the same coarse distance threshold (dt_{coarse}) to consolidate overlapping or proximal regions. To accommodate breakpoint uncertainty inherent to long-read alignments and SV merging, each resulting interval was symmetrically extended by $\pm fl$ (default: 7kb). This integration strategy ensures that both individual-specific and population-informed variation contribute to a comprehensive set of candidate SV regions. The merged intervals were output in BED format and used as input for the subsequent target region-based workflow.

Extraction of region-specific BAM file

FocalSV(target) is designed for scenarios where regions suspected to contain SVs have been identified, but the precise details of the SVs within that region have yet to be determined. FocalSV(target) assists in more accurately inferring SV breakpoint and sequence based on prior knowledge. To use FocalSV(target), users must provide regions suspected SV-containing regions, which serve as input. FocalSV(target) then extracts region-specific BAM files for subsequent reads partitioning, SV assembly, and refinement. In our experiments, to validate the accuracy of SV calling for FocalSV(target), we adopted the benchmark VCF file of the sample HG002 from Genome in a Bottle (GIAB) as prior knowledge.

Specifically, for INSs, the start and end positions of the target regions we utilized are defined as follows:

$$\begin{cases} Start = (Chrom_{INS}, Breakpoint - Flanking) \\ End = (Chrom_{INS}, Breakpoint + Flanking) \end{cases} \tag{1}$$

where the breakpoint refers to the insertion breakpoint on the reference genome, and the flanking size we used is 50kb by default.

In terms of DELs, the target regions we utilized are defined as follows:

$$\begin{cases} Start = (Chrom_{DEL}, Start_{DEL} - Flanking) \\ End = (Chrom_{DEL}, End_{DEL} + Flanking) \end{cases} \tag{2}$$

where the $start_{DEL}$ and end_{DEL} refer to the starting and ending positions of a target DEL on the reference genome and the flanking size we used is 50kb. When the breakpoints of INSs and DELs are unknown, users can estimate breakpoints to define the target regions and adjust the flanking regions to ensure regions are sufficiently large to encompass the potential SV.

In contrast, FocalSV(auto) automatically scans the aligned BAM file to detect regions suspected of harboring SVs using a custom detection method introduced earlier. It then extracts region-specific BAM files for subsequent reads partitioning, SV assembly, and refinement.

Partitioning of reads

The FocalSV pipeline integrates Longshot (Edge and Bansal 2019), a haplotype estimation tool, to perform phasing. Longshot builds on the read-based haplotype phasing algorithm, HapCUT2 (Edge et al. 2017) and employs a pair-Hidden Markov Model (pair-HMM) to mitigate uncertainties in local alignment. This approach enables the estimation of precise base quality values, which are crucial for genotype likelihood calculations. However, since Longshot is specifically designed to detect and phase single nucleotide variants (SNVs) and only approximately 75% of the reads can be partitioned into two distinct haplotypes, we assume

the remaining 25% of reads correspond to highly homogeneous regions. As a result, FocalSV assigns the remaining reads directly to both haplotypes of the nearest phase block.

Haplotype-aware local assembly

After reads partitioning, every read within the region BAM file is assigned to a certain phase block and haplotype. FocalSV then performs local assembly based on these phase blocks and haplotypes. For PacBio Hifi reads, FocalSV adopts Hifiasm (v0.14) (Cheng et al. 2021) for local assembly, while for PacBio CLR and Nanopore (ONT) reads, Flye (v2.9.1) (Kolmogorov et al. 2019) is utilized. Haplotype-resolved contigs are generated at the end of this procedure.

In rare cases where the target region is either too small or highly homogeneous, Longshot may fail to produce a phased BAM file. To address this limitation, we adopted a dual-assembly strategy. For HiFi data, we used Hifiasm (v0.16) in dual-haplotype mode to generate haplotype-resolved assemblies. For CLR and ONT data, we employed a combination of Flye and Hapdup to construct dual assemblies.

Detection of Candidate indel SVs

The haplotype-resolved contigs are then aligned to the human reference genome using minimap2 (Li 2018) and SAMtools (Danecek et al. 2021) in FocalSV. For HG002, we used hg19 reference genome to evaluate SVs because the GIAB gold standard SV callset is available only for hg19. The example command is as follows:

```
minimap2 -a -x asm5 --cs -r2k -t 30 \
     <ref_genome> \
     <contigs_fasta> \
      | samtools sort > contigs.bam
samtools index contigs.bam
```

Notably, the contigs-to-reference BAM file shares similar features with the reads-to-reference BAM file, as both files contain intra- and inter-alignments inferred by CIGAR operations, which indicate potential SVs. However, two key differences exist in terms of coverage and length: the reads-to-reference BAM file typically exhibits much higher coverage, while the contigs are generally much longer than individual reads. To reliably collect SV-related signatures from contig-to-reference BAM file, we adapt the contig-based signature collection methodology from VolcanoSV (Luo et al. 2024), which refines and optimizes the conventional reads-based SV signature collection methods.

The contig-based signature collection method can be intuitively explained. When an INS is present in the individual's genome, the contig can be conceptualized as a structure of "reference sequence A + INS sequence

+ reference sequence B". If the INS sequence is short, it is typically supported by an intra-alignment, which can be directly extracted from the CIGAR operation. However, longer INS sequences are supported by an inter-alignment (split alignments), where reference sequences A and B align independently to the human reference, while the INS sequence remains unaligned. Similarly, for a DEL in the individual's genome, the human reference genome can be viewed as having a structure of "sequence A + sequence B + sequence C", while the corresponding contig can be considered approximately as "sequence A + sequence C", with sequence B deleted. This deletion can be directly identified from the CIGAR operation. DELs can also be inferred from split alignments, where sequences A and C (adjacent on the contig) are aligned independently to two distant regions on the reference genome.

An INS signature, inferred from a pair of split-alignment events, is collected as follows:

$$\begin{cases} Chrom = Chrom_{INS} \\ Breakpoint = \dfrac{RefEnd_{seg1} + RefStart_{seg2}}{2} \\ SVlen = D_{cont} - D_{ref} \end{cases} \tag{3}$$

when

$$\begin{cases} D_{cont} - D_{ref} \geq 30 \text{ bp} \\ OLP_{ref} \leq 3000 \text{ bp} \end{cases} \tag{4}$$

A DEL signature is collected as follows:

$$\begin{cases} Chrom = Chrom_{DEL} \\ Start = RefEnd_{seg1} \\ End = RefEnd_{seg1} + |D_{cont} - D_{ref}| \\ SVlen = D_{cont} - D_{ref} \end{cases} \tag{5}$$

when

$$\begin{cases} D_{ref} - D_{cont} \geq 30bp \\ \quad OLP_{ref} \leq 3000bp \end{cases} \tag{6}$$

where $RefStart_{segi}$ and $RefEnd_{segi}$ denote the start and end coordinates of i-th aligned sequence segment relative to the human reference ($i \in [1, 2]$). The segment with the smaller start coordinate on the reference genome is designated as the first aligned segment. D_{cont} and D_{ref} represent the distance between two segments relative to the contig and the human reference, respectively. OLP_{ref} refers to the overlap size between two segments on the human reference.

We then incorporated the clustering pipeline from VolcanoSV (Luo et al. 2024) to organize the collected signatures by SV type and generate a candidate SV pool. The pipeline involves three steps. Firstly, a chaining clustering algorithm groups signatures from the same haplotype, using SV size similarity and breakpoint distance as clustering criteria. Next, a pairing algorithm matches signatures from different haplotypes for genotyping. Finally, a stringent one-to-K clustering algorithm is applied to reduce the redundancy in the call set.

Filtering of indel SVs and refinement of genotypes

To further refine SV calls, FocalSV employs VolcanoSV's reads-based signature support algorithm (Luo et al. 2024) to filter out false positive SVs and enhance genotyping precision.

For false positive control, the process begins by gathering read-based SV signatures near the potential SV breakpoints. A similarity score is then calculated between the read-based signature and the candidate SV. Signatures exceeding the similarity threshold are considered supporting signatures. A candidate INS is retained if it has at least one supporting signature, while a candidate DEL must meet the supporting coverage threshold (calculated as the ratio of supporting signatures to local read coverage) to be considered valid.

For genotype refinement, a genotype decision tree model (Luo et al. 2024) is employed to correct the genotype of indel SVs. This model considers five key parameters: contig-based genotype (heterozygous or homozygous), SV size (small or large indel), SV type (insertion or deletion), sequencing technology (Hifi, CLR, or ONT), and a supporting read-based signature ratio (number of supporting read-based signatures/local read depth). With all combinations of the first four parameters, the decision tree consists of 24 leaf nodes. Each leaf node is associated with an empirically determined threshold for the supporting signature ratio. This ratio, ideally reflecting the genotype, is then utilized to predict the final genotype of the SV.

Multiple-region mode

FocalSV is also designed to detect SVs across multiple regions. When provided with multiple target regions either automatically or manually, FocalSV first retrieves region-specific BAM files, and applies the same steps for each region: reads partitioning, local assembly, and candidate SV detection. This process is parallelized in FocalSV for greater efficiency. Afterward, FocalSV merges all VCF files from each region into a single VCF file. A clustering algorithm is then applied to the merged VCF file to eliminate redundant variants. Finally, the SV filtering and genotype refinement are performed to produce the final VCF file.

Detection, recovery, and breakend refinement of duplications, translocations, and inversions in the target mode

Duplication detection and recovery

For duplication detection in the target mode, FocalSV integrates information from both contigs and read-based BAM files. The user provides the approximate start and end positions of the target duplication as prior knowledge. Using this input, contigs are generated through a region-based assembly pipeline, the same approach used for large indel detection. These contigs are subsequently aligned to a human reference genome to produce a contig-based BAM file. Simultaneously, a read-based BAM file is generated by extracting reads from the user-defined target duplication region along with an extended flanking region (default: 25kb). In the alignment-based information derived from the contig-based BAM file, duplication (DUP) can be regarded as a special instance of insertion (INS), particularly when the inserted sequence closely resembles a reference segment near the insertion breakpoint. As a result, if two adjacent segments on the contig align to the same region on the reference genome, a duplication can be directly inferred. However, in practice, due to the limitation of existing aligners, only one of the duplicated segments may be correctly aligned to the reference, while the other segment is "skipped", leading to it being incorrectly labeled as an INS in the BAM file. To address this, we designed a dedicated duplication recovery pipeline from insertion calls. Specifically, FocalSV extracts the alternate alleles from all insertion calls and realigns them to the reference genome. If an alternate allele aligns close to its corresponding insertion breakpoint, it indicates the presence of a duplication (Fig. 2A). The duplication's start and end coordinates are defined by the alignment coordinates of the alternate allele.

This recovery procedure can recover a considerable amount of DUPs missed by the aligner. However, in cases where the contigs are collapsed (i.e. do not contain the duplicated segments) due to misassembly, we designed an additional reads alignment-based approach to recover those DUPs. In this approach, FocalSV identifies events in the target region where a read aligns more than once on the same strand, generating an alignment records pool for each muti-aligned read (MAR). For each MAR, the positional relationship between any two alignment records is evaluated to determine whether they suggest a duplication. A DUP signature is inferred when the pair of read segment alignment records meet the below conditions:

$$\begin{cases} ReadEnd_{seg1} - ReadStart_{seg2} \leq Th_{olp} \\ RefEnd_{seg1} > RefStart_{seg2} \end{cases} \tag{7}$$

where $ReadStart_{segi}$, $ReadEnd_{segi}$, $RefStart_{segi}$ and $RefEnd_{segi}$ represent the start and end coordinates of the i-th segment relative to read and reference, respectively (Fig. 2A). Th_{olp} denotes the maximum allowed

overlap between two segments (500 bp by default). The segments are ordered such that $ReadStart_{seg1} < ReadStart_{seg1}$.

Next, FocalSV applies a distance-based clustering algorithm to group DUP signatures. Two DUP signatures are clustered together if their breakpoint shift is less than a specified distance threshold (1000bp by default). The average start and end positions within each cluster are selected as the final breakpoints of the DUP. It is important to note that this reads-based approach is only used for DUPs estimated to be smaller than 5Mb, as split-alignments with gaps larger than 5Mb on the reference are likely misalignments and unreliable. Finally, FocalSV merges the result from both the conitg-based and read-based BAM files to generate the final set of DUP calls.

Translocation detection

The format for a typical translocation (TRA) is:

$$(ChromA, PosA, ChromB, PosB) \tag{8}$$

To accurately detect breakends (BND) of TRAs in the target mode, FocalSV requires prior knowledge of the approximate regions for both BNDs, $PosA$ and $PosB$:

$$\begin{cases} (ChromA, PosA - Flanking, PosA + Flanking) \\ (ChromB, PosB - Flanking, PosB + Flanking) \end{cases} \tag{9}$$

To achieve the detection and refinement, FocalSV identifies TRA signatures from read-based BAM files and calculates the optimal BND. Specifically, TRAs are inferred when two adjacent segments on the read align to different chromosomes. FocalSV first collects reads aligned to both BND regions using equation 9. It then calculates the segment distance between two alignment records of each read. A TRA event is inferred if the two alignment records satisfy the below conditions:

$$\begin{cases} |ReadEnd_{seg1} - ReadStart_{seg2}| \leq 1000bp \\ \text{or} \\ |ReadEnd_{seg2} - ReadStart_{seg1}| \leq 1000bp \end{cases} \tag{10}$$

where $ReadStart_{seg1}$, $ReadEnd_{seg1}$, $ReadStart_{seg2}$, and $ReadEnd_{seg2}$ represent the start and end positions of each aligned segment on the read, respectively. A TRA signature is represented as follows:

$$\begin{cases} (Chrom_{seg1}, RefEnd_{seg1}, Chrom_{seg2}, RefStart_{seg2}) \\ \qquad \text{for } |ReadEnd_{seg1} - ReadStart_{seg2}| \leq 1000bp \\ \qquad\qquad\qquad\qquad\qquad or \\ (Chrom_{seg1}, RefStart_{seg1}, Chrom_{seg2}, RefEnd_{seg2}) \\ \qquad \text{for } |ReadEnd_{seg2} - ReadStart_{seg1}| \leq 1000bp \end{cases} \tag{11}$$

where $RefStart_{seg1}$, $RefEnd_{seg1}$, $RefStart_{seg2}$, and $RefEnd_{seg2}$ denote the start and end positions of each aligned segment on the reference genome, respectively (Fig. 2B). A clustering algorithm is then employed to merge TRA signatures. Any two TRA signatures are grouped into a single cluster if they meet the following criteria:

$$\begin{cases} ChromA_{TRA1} = ChromA_{TRA2} \\ ChromB_{TRA1} = ChromB_{TRA2} \\ |PosA_{TRA1} - PosA_{TRA2}| \leq 100bp \\ |PosB_{TRA1} - PosB_{TRA2}| \leq 100bp \end{cases} \tag{12}$$

The average coordinate for each BND is selected as the cluster center, serving as the final TRA call.

Inversion detection

The format for a typical inversion (INV) is:

$$(Chrom_{INV}, Start_{INV}, End_{INV}) \tag{13}$$

Similar as TRA, the target region in the target mode for INV is defined as follows:

$$(Chrom_{INV}, Start_{INV} - Flanking, End_{INV} + Flanking) \tag{14}$$

FocalSV then identifies INV signatures from the read-based BAM file. Specifically, INV events are identified when two adjacent segments of a read align to two distant locations on the reference in reverse orientations.

The criteria for inferring an INV and its corresponding signature are outlined below:

$$\begin{cases} (Chrom, RefStart_{rev}, RefStart_{fwd}) \\ \text{for } |ReadEnd_{rev} - ReadStart_{fwd}| \leq 100bp \\ or \\ (Chrom, RefEnd_{fwd}, RefEnd_{rev}) \\ \text{for } |ReadEnd_{fwd} - ReadStart_{rev}| \leq 100bp \end{cases} \tag{15}$$

where $ReadStart_{fwd}$, $ReadEnd_{fwd}$, $ReadStart_{rev}$, and $ReadEnd_{rev}$ denote the start and end positions of the forward- and reverse-aligned segment on the read, while $RefStart_{fwd}$, $RefEnd_{fwd}$, $RefStart_{rev}$, and $RefEnd_{rev}$ represent the start and end positions on the reference (Fig. 2C). A clustering algorithm is then applied to merge INV signatures. Two INV signatures are grouped into a single cluster if they meet the following criteria:

$$\begin{cases} Chrom_{INV1} = Chrom_{INV2} \\ |Start_{INV1} - Start_{INV2}| \leq 100bp \\ |End_{INV1} - End_{INV2}| \leq 100bp \end{cases} \tag{16}$$

The average position for each BND is selected as the cluster center, serving as the final INV call.

Automatic detection and refinement of duplications, translocations, and inversions

When target regions are not provided or available, the auto mode of FocalSV can be used to detect duplications, translocations, and inversions. Its SV detection workflow consists of three modules: (1) signature extraction, (2) signature clustering, and (3) post-processing.

Signature extraction

In auto mode, to localize potential signals of duplications, translocations, and inversions across the genome, FocalSV first divides each chromosome into contiguous, nonoverlapping 1Mb windows. Within each window, all long reads aligned to that region are examined for evidence of split alignments. Intuitively, a split alignment indicates that a single read maps to two distinct genomic loci, potentially reflecting an underlying SVs such as inversion, duplication, or translocation.

Within each 1Mb window, FocalSV identifies reads carrying a supplementary alignment tag (SA tag) and meeting a minimum mapping quality threshold (MAPQ $\geq$ 10). SA tags indicate additional alignments of the same read, commonly arising from SVs that cause discontinuous or ambiguous mappings. Compared

to the primary alignment, the supplementary alignment may reside on the same chromosome - indicating a potential inversion or duplication, or on a different chromosome - suggesting a possible translocation. By scanning all windows in parallel, FocalSV accumulates three sets of candidate reads per window:

- **Inversion candidates:** Reads with supplementary alignment on the same chromosome but in the opposite orientation relative to the primary alignment.
- **Duplication candidates:** Reads with supplementary alignment on the same strand and in the same orientation as the primary alignment.
- **Translocation candidates:** Reads with supplementary alignment mapping to a different chromosome.

Each candidate read is annotated with mapping start and end position on both the reference genome and the read, along with strand orientation and mapping quality for primary and supplementary alignments. After processing all windows on a given chromosome, candidate reads are aggregated into chromosome-wide collections by SV type, serving as input for downstream signature inference. To detect SV signatures corresponding to duplications, translocations, and inversions, FocalSV applies type-specific pairing rules based on reference and read-level coordinates, as defined in Equations 7, 11, and 15. Specifically, a duplication (DUP) signature is inferred when two same-strand segments align to the same chromosome in order, with minimal separation on the read and substantial overlap on the reference (Equation 7). A translocation (TRA) signature is inferred when a single read maps to two distinct chromosomes, with the segments located within 1000bp on the read (Equation 11). An inversion (INV) signature is inferred when two segments align to opposite strands of the same chromosome, with no more than 100bp separation on the read and close proximity on the reference (Equation 15).

Signature clustering

To consolidate redundant signatures and minimize noise, FocalSVS applies a distance-based clustering algorithm to the inferred duplications, translocations, and inversions signatures. For each SV type, the signatures are grouped into clusters if their genomic coordinates lay within a fixed distance threshold (default: 100bp), ensuring that nearby signatures likely originating from the same event are merged (Equations 12, 16). Each resulting cluster is treated as a candidate SV call. Within each cluster, summary statistics are computed, including the mean start and end coordinates (or breakend positions), the number of supporting reads, average mapping quality, and the standard deviation of breakpoint positions across supporting alignments. These features are then used for downstream filtering and prioritization of high-confidence SVs.

Post processing

To refine duplication candidates, FocalSV employs a structured post-processing pipeline that integrates read-level evidence with both local and global coverage statistics. First, the genome-wide average read depth ($globalRD$) is estimated from the corresponding BAM file to enable subsequent feature adjustment. For each duplication event, a minimum number of supporting reads is required - by default, at least $0.1 \times globalRD$ - carrying valid SA tags and sufficient mapping quality (default ≥ 50). Coverage is then extracted from the candidate duplication region (COV_{DUP}) as well as from fixed 1kb flanking windows on the left (COV_{left}) and right (COV_{right}) sides. Three coverage-based features are computed: (1) Coverage shift: defined as $2 \times COV_{DUP}/(COV_{left} + COV_{right})$, this metric captures the expected local increase in coverage due to duplication. (2) Relative duplication coverage: calculated as $COV_{DUP}/globalRD$, providing a normalized estimate of local copy gain; and (3) Breakpoint variability: measured as the standard deviation of the left and right breakpoints within the original DUP cluster, used to detect irregular mapping patterns indicative of potential artifacts. Duplication candidates that do not meet empirically optimized thresholds for any of these features are filtered out. This multi-feature framework ensures that only high-confidence duplication calls - with strong read support and consistent coverage profiles - are retained.

Filtering of translocation (TRA) and inversion (INV) candidates is based on two criteria: a minimum number of supporting reads and a minimum mapping quality threshold. Specifically, each candidate is required to have at least $r \times globalRD$ supporting split reads, each with mapping quality greater than or equal to a platform-specific threshold. The values of r and the minimum mapping quality vary depending on the sequencing platform and are empirically optimized to account for their distinct error profiles. For example, for INV calls, r and the minimum mapping quality are set to 0.25 and 60 on Hifi data, 0.3 and 58 on CLR data, and 0.35 and 58 on ONT data.

Sequencing data

PacBio CLR, Hifi, and ONT sequencing reads for HG002 are available at GIAB and NCBI. PacBio Hifi sequencing reads for HG005 are available at NCBI. The high-confidence HCC1395 somatic SV callset and the PacBio and ONT Tumor-Normal paired libraries of HCC1395 are publicly accessible at NCBI. Table 1 lists hyperlinks for all 14 previously mentioned real datasets. The Tier1 benchmark SV callset and high-confidence HG002 region were obtained from https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/analysis/NIST_SVs_Integration_v0.6/. The haplotype-resolved assemblies for the nine samples used in constructing the multi-sample VCF are publicly available at https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=working/HPRC_PLUS/. These assemblies can be found within the `assemblies/year1_freeze_assembly_v2` subdirectory corresponding to each sample.

T2T HG002 assembly was obtained from https://s3-us-west-2.amazonaws.com/human-pangenomics/T2T/HG002/assemblies/hg002v1.1.fasta.gz. Strand-seq data for HG002 involved in the phasing quality evaluation are publicly available at https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=NHGRI_UCSC_panel/HG002/hpp_HG002_NA24385_son_v1/Strand_seq/. All VCF files that support the findings of this study are available from https://doi.org/10.5281/zenodo.15740572.

Software availability

FocalSV is available freely at GitHub (https://github.com/maiziezhoulab/FocalSV). All scripts necessary to reproduce this study are available as Supplemental Code and are also available at GitHub https://github.com/maiziezhoulab/FocalSV/tree/main/evaluation.

Competing interests

The authors declare no competing interests.

Acknowledgements

This work was supported by the NIH NIGMS Maximizing Investigators' Research Award (MIRA) R35 GM146960.

Author contributions

X.M.Z. conceived and led this work. C.L., Z.J.Z., and X.M.Z. designed the framework. C.L. and Z.J.Z. implemented the framework. C.L., Z.J.Z., and Y.H.L performed all analyses. C.L., Z.J.Z., and X.M.Z. wrote the manuscript. All authors reviewed the manuscript.

Author details

[1]Department of Biomedical Engineering, Vanderbilt University, Nashville, TN 37235, USA. [2]Department of Computer Science, Vanderbilt University, Nashville, TN 37235, USA.

References

Alkan C, Coe BP, and Eichler EE. 2011. Genome structural variation discovery and genotyping. *Nature reviews genetics* **12**: 363–376.

Biosciences P. 2021. pbsv-pacbio structural variant (sv) calling and analysis tools.

Chaisson MJ, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, Gardner EJ, Rodriguez OL, Guo L, Collins RL, et al.. 2019. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nature communications* **10**: 1784.

Cheng H, Concepcion GT, Feng X, Zhang H, and Li H. 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature methods* **18**: 170–175.

Choi YW. 2020. Phasing performance evaluation scripts. https://github.com/ywchoi/phasing. Accessed: 2025-05-27.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al.. 2021. Twelve years of SAMtools and BCFtools. *GigaScience* **10**. Giab008.

Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, Sulovari A, Ebler J, Zhou W, Serra Mari R, et al.. 2021. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**: eabf7117.

Ebler J, Ebert P, Clarke WE, Rausch T, Audano PA, Houwaart T, Mao Y, Korbel JO, Eichler EE, Zody MC, et al.. 2022. Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nature genetics* **54**: 518–525.

730 Edge P, Bafna V, and Bansal V. 2017. Hapcut2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome research* **27**:
731 801–812.
732 Edge P and Bansal V. 2019. Longshot enables accurate variant calling in diploid genomes from single-molecule long read sequencing. *Nature*
733 *communications* **10**: 4660.
734 English AC, Menon VK, Gibbs RA, Metcalf GA, and Sedlazeck FJ. 2022. Truvari: refined structural variant comparison preserves allelic diversity.
735 *Genome Biology* **23**: 271.
736 Fang LT, Zhu B, Zhao Y, Chen W, Yang Z, Kerrigan L, Langenbach K, de Mars M, Lu C, Idler K, et al.. 2021. Establishing community reference
737 samples, data and call sets for benchmarking cancer mutation detection using whole-genome sequencing. *Nature biotechnology* **39**: 1151–1160.
738 Heller D and Vingron M. 2019. Svim: structural variant identification using mapped long reads. *Bioinformatics* **35**: 2907–2915.
739 Heller D and Vingron M. 2020. Svim-asm: structural variant detection from haploid and diploid genome assemblies. *Bioinformatics* **36**: 5519–5521.
740 Jain M, Olsen HE, Paten B, and Akeson M. 2016. The oxford nanopore minion: delivery of nanopore sequencing to the genomics community. *Genome*
741 *biology* **17**: 1–11.
742 Jeffares DC, Jolly C, Hoti M, Speed D, Shaw L, Rallis C, Balloux F, Dessimoz C, Bähler J, and Sedlazeck FJ. 2017. Transient structural variations
743 have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nature communications* **8**: 14061.
744 Jiang T, Liu Y, Jiang Y, Li J, Gao Y, Cui Z, Liu Y, Liu B, and Wang Y. 2020. Long-read-based human genomic structural variation detection with
745 cutesv. *Genome biology* **21**: 1–24.
746 Kolmogorov M, Yuan J, Lin Y, and Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. *Nature biotechnology* **37**: 540–546.
747 Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094–3100.
748 Li H, Bloom JM, Farjoun Y, Fleharty M, Gauthier L, Neale B, and MacArthur D. 2018. A synthetic-diploid benchmark for accurate variant-calling
749 evaluation. *Nature methods* **15**: 595–597.
750 Liu Y, Jiang T, Su J, Liu B, Zang T, and Wang Y. 2021. Sksv: ultrafast structural variation detection from circular consensus sequencing reads.
751 *Bioinformatics* **37**: 3647–3649.
752 Liu YH, Luo C, Golding SG, Ioffe JB, and Zhou XM. 2024. Tradeoffs in alignment and assembly-based methods for structural variant detection with
753 long-read sequencing data. *Nature Communications* **15**: 2447.
754 Logsdon GA, Vollger MR, and Eichler EE. 2020. Long-read human genome sequencing and its applications. *Nature Reviews Genetics* **21**: 597–614.
755 Luo C, Liu YH, and Zhou XM. 2024. Volcanosv enables accurate and robust structural variant calling in diploid genomes from single-molecule long
756 read sequencing. *Nature Communications* **15**: 6956.
757 Pacbio. 2023. Revio system: Reveal more with accurate long-read sequencing at scale.
758 Powers DM. 2020. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061* .
759 Rhoads A and Au KF. 2015. Pacbio sequencing and its applications. *Genomics, Proteomics & Bioinformatics* **13**: 278–289.
760 Sasaki Y et al.. 2007. The truth of the f-measure. *Teach tutor mater* .
761 Saunders CT, Holt JM, Baker DN, Lake JA, Belyeu JR, Kronenberg Z, Rowell WJ, and Eberle MA. 2025. Sawfish: Improving long-read structural
762 variant discovery and genotyping with local haplotype modeling. *Bioinformatics* **41**: btaf136.
763 Smolka M, Paulin LF, Grochowski CM, Horner DW, Mahmoud M, Behera S, Kalef-Ezra E, Gandhi M, Hong K, Pehlivan D, et al.. 2024. Detection of
764 mosaic and population-level structural variants with sniffles2. *Nature biotechnology* pp. 1–10.
765 Talsania K, Shen Tw, Chen X, Jaeger E, Li Z, Chen Z, Chen W, Tran B, Kusko R, Wang L, et al.. 2022. Structural variant analysis of a cancer
766 reference cell line sample using multiple sequencing technologies. *Genome biology* **23**: 1–33.
767 Weischenfeldt J, Symmons O, Spitz F, et al.. 2013. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nature*
768 *Reviews Genetics* **14**: 125–138.
769 Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, and Salit M. 2014. Integrating human sequence data sets provides a resource of
770 benchmark snp and indel genotype calls. *Nature biotechnology* **32**: 246–251.

Tools	Version	Resource	Variant Types	Link
Dipcall	0.3	Li et al. 2018	DEL, INS, small indel, SNP	https://github.com/lh3/dipcall
SVIM-asm	1.0.2	Heller et al. 2020	DEL, INS, DUP, INV, TRA	https://github.com/eldariont/svim-asm
PAV	2.0.0	Ebert et al. 2021	DEL, INS, INV, small indel, SNP	https://github.com/EichlerLab/pav
sawfish	1.0.1	Saunders et al. 2025	INS, DEL, INV, DUP, BND	https://github.com/PacificBiosciences/sawfish
FocalSV	1.0.0	This paper	DEL, INS, INV, TRA, DUP	https://github.com/maiziezhoulab/FocalSV
SVIM	1.4.2	Heller et al. 2019	DEL, INS, DUP, INV, TRA	https://github.com/eldariont/svim
cuteSV	1.0.11	Jiang et al. 2020	DEL, INS, DUP, INV	https://github.com/tjiangHIT/cuteSV
pbsv	2.6.2		DEL, INS, DUP, INV, TRA, CNV	https://github.com/PacificBiosciences/pbsv
SKSV	1.0.2	Liu et al. 2021	DEL, INS, DUP, INV, TRA	https://github.com/ydLiu-HIT/SKSV
Sniffles2	2.0.6	Smolka et al. 2022	DEL, INS, DUP, INV, TRA	https://github.com/fritzsedlazeck/Sniffles

Dataset	Abbreviation in the paper	Coverage	Source
HG002 PacBio CCS 15kb+20kb	Hifi_L1	56x	https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG002_NA24385_son/PacBio_CCS_15kb_20kb_chemistry2/reads/
HG002 PacBio CCS 10kb	Hifi_L2	30x	https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG002_NA24385_son/PacBio_CCS_10kb/
HG002 PacBio CCS 11kb	Hifi_L3	34x	https://www.ncbi.nlm.nih.gov/sra/SRR8833180
HG002 MtSinai	CLR_L1	65x	https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG002_NA24385_son/PacBio_MtSinai_NIST/
HG002 PacBio CLR	CLR_L2	89x	https://www.ncbi.nlm.nih.gov/sra/SRX7668835
HG002 PacBio CLR	CLR_L3	29x	https://www.ncbi.nlm.nih.gov/sra/SRX6719924
HG002 Nanopore UCSC_lt100kb	ONT_L1	51x	https://s3-us-west-2.amazonaws.com/human-pangenomics/NHGRI_UCSC_panel/HG002/hpp_HG002_NA24385_son_v1/nanopore/downsampled/greater_than_100kb/HG002_ucsc_ONT_lt100kb.fastq.gz
HG002 Nanopore PRJNA678534	ONT_L2	48x	https://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRP292617&o=acc_s%3Aa
HG002 Nanopore Standard_Unsheared_UCSC	ONT_L3	47x	https://s3-us-west-2.amazonaws.com/human-pangenomics/NHGRI_UCSC_panel/HG002/hpp_HG002_NA24385_son_v1/nanopore/downsampled/standard_unsheared/HG002_ucsc_Jan_2019_Guppy_3.4.4.fastq.gz
HG005 HiFi	HG005_Hifi	50x	ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/ChineseTrio/HG005_NA24631_son/PacBio_CCS_15kb_20kb_chemistry2/uBAMs/
HCC1395 Tumor PacBio	HCC1395_PB	39x	https://www.ncbi.nlm.nih.gov/sra/?term=SRR8955953
HCC1395 Normal PacBio	HCC1395BL_PB	44x	https://www.ncbi.nlm.nih.gov/sra/?term=SRR8955954
HCC1395 Tumor ONT	HCC1395_ONT	12x	https://www.ncbi.nlm.nih.gov/sra/?term=SRR16005301
HCC1395 Normal ONT	HCC1395BL_ONT	19x	https://www.ncbi.nlm.nih.gov/sra/?term=SRR17096031

Table 1 Resource for different tools and long-read datasets. Top panel: The SV calling tools used in this paper. Tool version number, cited article, variant types called by each tool, and tool links are shown in the table. Bottom panel: The long-read datasets used in this paper. The abbreviation, coverage, and source link for each dataset are shown in the table. The abbreviation is used to refer to each dataset in the main text.

	Library	Metric	FocalSV (target)	FocalSV (auto)	PAV	SVIM-asm	Dipcall	sawfish	cuteSV	SVIM	PBSV	Sniffles2	SKSV
DEL	Hifi_L1	Precision	93.14%	**93.59%**	92.28%	91.70%	90.33%	93.10%	88.20%	87.33%	93.03%	83.75%	85.29%
		Recall	**95.36%**	94.78%	93.22%	93.90%	92.61%	95.34%	88.46%	88.78%	93.44%	84.65%	85.25%
		F1	**94.24%**	94.18%	92.75%	92.79%	91.46%	94.20%	88.33%	88.05%	93.24%	84.20%	85.27%
		GT Concordance	**99.10%**	99.05%	97.11%	97.85%	96.75%	98.37%	99.07%	98.71%	97.71%	96.56%	97.98%
	Hifi_L2	Precision	92.65%	93.81%	92.04%	92.01%	90.11%	93.34%	**93.83%**	82.90%	92.21%	92.67%	85.81%
		Recall	95.00%	95.02%	94.12%	94.56%	92.76%	95.31%	87.12%	**95.40%**	93.78%	94.85%	78.91%
		F1	93.81%	**94.41%**	93.07%	93.27%	91.42%	94.31%	90.35%	88.70%	92.99%	93.74%	82.22%
		GT Concordance	**99.00%**	**99.00%**	97.26%	98.61%	96.86%	98.22%	97.46%	97.90%	98.19%	98.31%	97.48%
	Hifi_L3	Precision	92.67%	93.71%	91.71%	91.70%	89.94%	93.15%	**94.23%**	75.30%	92.49%	92.81%	85.88%
		Recall	95.21%	94.87%	93.49%	94.44%	92.76%	95.17%	90.43%	**95.34%**	93.93%	95.02%	81.44%
		F1	93.92%	**94.29%**	92.59%	93.05%	91.33%	94.15%	92.29%	84.14%	93.20%	93.90%	83.60%
		GT Concordance	98.98%	**99.00%**	96.62%	98.59%	96.83%	98.44%	98.04%	98.34%	98.60%	98.47%	98.06%
	Overall	Precision	92.82%	**93.70%**	92.01%	91.80%	90.13%	93.20%	92.09%	81.84%	92.58%	89.74%	85.66%
		Recall	95.19%	94.89%	93.61%	94.30%	92.71%	**95.27%**	88.67%	93.17%	93.72%	91.51%	81.87%
		F1	93.99%	**94.29%**	92.80%	93.04%	91.40%	94.22%	90.32%	86.96%	93.14%	90.61%	83.70%
		GT Concordance	**99.03%**	99.02%	97.00%	98.35%	96.81%	98.34%	98.19%	98.32%	98.17%	97.78%	97.84%
INS	Hifi_L1	Precision	**90.80%**	88.50%	86.30%	86.50%	72.80%	83.89%	87.50%	85.10%	90.00%	79.00%	86.10%
		Recall	93.90%	93.88%	93.10%	93.10%	91.40%	**94.24%**	82.30%	69.60%	77.40%	76.80%	86.60%
		F1	**92.40%**	91.11%	89.60%	89.70%	81.00%	88.76%	84.90%	76.60%	83.20%	77.90%	86.40%
		GT Concordance	98.10%	98.39%	95.60%	97.80%	77.80%	92.38%	**99.10%**	90.00%	87.70%	89.30%	98.30%
	Hifi_L2	Precision	89.80%	88.32%	86.40%	86.40%	73.10%	84.75%	**91.80%**	61.90%	89.90%	88.30%	86.20%
		Recall	93.80%	93.77%	93.30%	92.90%	91.60%	**94.05%**	82.40%	91.60%	76.70%	91.80%	61.70%
		F1	**91.70%**	90.96%	89.70%	89.50%	81.30%	89.16%	86.80%	73.90%	82.80%	90.00%	71.90%
		GT Concordance	**98.30%**	98.20%	96.10%	97.70%	78.50%	92.95%	97.60%	94.00%	95.40%	97.00%	97.70%
	Hifi_L3	Precision	89.60%	88.31%	86.30%	86.40%	73.00%	84.33%	**91.60%**	45.00%	89.00%	88.60%	87.10%
		Recall	93.80%	93.88%	93.00%	93.20%	91.80%	**94.04%**	86.00%	92.10%	76.70%	92.80%	71.00%
		F1	**91.70%**	91.01%	89.50%	89.60%	81.30%	88.92%	88.70%	60.50%	82.40%	90.70%	78.20%
		GT Concordance	98.50%	98.41%	94.90%	98.00%	78.30%	92.45%	98.40%	94.30%	96.20%	97.10%	**98.60%**
	Overall	Precision	90.07%	88.38%	86.33%	86.43%	72.97%	84.32%	**90.30%**	64.00%	89.63%	85.30%	86.47%
		Recall	93.83%	93.84%	93.13%	93.07%	91.60%	**94.11%**	83.57%	84.43%	76.93%	87.13%	73.10%
		F1	**91.93%**	91.03%	89.60%	89.60%	81.20%	88.95%	86.80%	70.33%	82.80%	86.20%	78.83%
		GT Concordance	98.30%	98.33%	95.53%	97.83%	78.20%	92.59%	**98.37%**	92.77%	93.10%	94.47%	98.20%

Table 2 Large deletions (DELs) and insertions (INSs) ($\geq$50bp) calling performance across three Hifi datasets. The table presents Recall, Precision, F1 score (in %), and Genotype accuracy (measured by Genotype Concordance) for all benchmarked SV callers. The highest scores for Recall, Precision, F1, and Genotype accuracy are highlighted in bold. Additionally, the overall average performance for each SV type and library is summarized at the bottom of the table. Benchmarking was conducted using Truvari with the following parameter settings: p = 0.5, P = 0.5, r = 500, and O = 0.01.

	Library	Metric	FocalSV (target)	FocalSV (auto)	PAV	SVIM-asm	Dipcall	cuteSV	SVIM	PBSV	Sniffles2
DEL	CLR_L1	Precision	91.76%	93.32%	81.51%	88.51%	82.40%	92.04%	91.80%	**93.99%**	92.49%
		Recall	**94.17%**	91.59%	86.88%	86.10%	21.04%	89.38%	91.70%	91.89%	91.81%
		F1	**92.95%**	92.45%	84.11%	87.29%	33.52%	90.69%	91.70%	92.92%	92.15%
		GT Concordance	**98.76%**	98.73%	96.00%	95.23%	85.45%	98.64%	97.90%	96.64%	98.68%
	CLR_L2	Precision	92.03%	**93.43%**	91.21%	92.01%	88.81%	91.43%	93.40%	93.29%	92.93%
		Recall	**94.83%**	93.97%	89.02%	91.72%	86.95%	91.98%	93.90%	92.95%	92.23%
		F1	93.41%	**93.70%**	90.10%	91.86%	87.87%	91.70%	93.60%	93.12%	92.57%
		GT Concordance	98.72%	98.99%	98.39%	98.46%	94.36%	**99.29%**	98.60%	97.80%	97.37%
	CLR_L3	Precision	92.45%	93.08%	88.32%	90.60%	83.06%	92.22%	93.00%	92.56%	**94.08%**
		Recall	**92.54%**	92.15%	84.35%	84.28%	46.33%	80.10%	79.30%	91.30%	78.33%
		F1	92.50%	**92.61%**	86.29%	87.33%	59.48%	85.74%	85.60%	91.93%	85.48%
		GT Concordance	97.24%	97.68%	95.65%	96.17%	84.27%	97.76%	**98.80%**	96.78%	96.00%
	Overall	Precision	92.08%	**93.28%**	87.01%	90.37%	84.76%	91.90%	92.73%	**93.28%**	93.17%
		Recall	**93.85%**	92.57%	86.75%	87.37%	51.44%	87.15%	88.30%	92.05%	87.46%
		F1	**92.95%**	92.92%	86.83%	88.83%	60.29%	89.38%	90.30%	92.66%	90.07%
		GT Concordance	98.24%	98.47%	96.68%	96.62%	88.03%	**98.56%**	98.43%	97.07%	97.35%
INS	CLR_L1	Precision	89.15%	87.90%	72.05%	80.21%	69.61%	86.78%	92.73%	**93.15%**	88.02%
		Recall	**92.84%**	90.82%	91.18%	89.95%	20.77%	86.12%	71.20%	79.78%	86.65%
		F1	**90.96%**	89.34%	80.49%	84.80%	32.00%	86.45%	80.55%	85.94%	87.33%
		GT Concordance	96.94%	97.66%	91.75%	92.00%	68.00%	**98.04%**	61.70%	87.56%	96.72%
	CLR_L2	Precision	**87.98%**	84.74%	84.15%	84.62%	66.76%	83.15%	62.85%	86.10%	81.14%
		Recall	**93.52%**	93.45%	91.76%	93.32%	88.13%	91.88%	27.65%	77.90%	80.89%
		F1	**90.67%**	88.88%	87.79%	88.75%	75.97%	87.30%	38.40%	81.80%	81.02%
		GT Concordance	97.02%	98.82%	95.87%	96.77%	67.60%	**98.97%**	44.11%	89.52%	94.97%
	CLR_L3	Precision	89.59%	87.17%	78.91%	81.51%	60.40%	88.16%	**91.67%**	87.57%	89.51%
		Recall	91.44%	91.48%	**91.71%**	90.40%	47.28%	79.83%	75.48%	79.93%	70.42%
		F1	**90.51%**	89.27%	84.83%	85.72%	53.04%	83.79%	82.79%	83.58%	78.83%
		GT Concordance	95.07%	96.21%	89.24%	89.86%	54.27%	**96.23%**	87.68%	88.65%	92.63%
	Overall	Precision	88.91%	86.60%	78.37%	82.11%	65.59%	86.03%	82.42%	**88.94%**	86.22%
		Recall	**92.60%**	91.92%	91.55%	91.22%	52.06%	85.94%	58.11%	79.20%	79.32%
		F1	**90.71%**	89.16%	84.37%	86.42%	53.67%	85.85%	67.25%	83.77%	82.39%
		GT Concordance	96.34%	97.56%	92.29%	92.88%	63.29%	**97.75%**	64.50%	88.58%	94.77%

Table 3 Large deletions (DELs) and insertions (INSs) ($\geq$50bp) calling performance across three CLR datasets. The table presents Recall, Precision, F1 score (in %), and genotype accuracy (measured by Genotype Concordance) for all benchmarked SV callers. The highest scores for Recall, Precision, F1, and Genotype accuracy are highlighted in bold. Additionally, the overall average performance for each SV type and library is summarized at the bottom of the table. Benchmarking was conducted using Truvari with the following parameter settings: p = 0.5, P = 0.5, r = 500, and O = 0.01.

	Library	Metric	FocalSV (target)	FocalSV (auto)	PAV	SVIM-asm	Dipcall	cuteSV	SVIM	Sniffles2
DEL	ONT_L1	Precision	91.89%	**92.25%**	91.10%	90.86%	86.48%	91.32%	90.88%	91.60%
		Recall	**93.90%**	93.44%	92.78%	**93.90%**	90.72%	92.78%	93.00%	92.44%
		F1	**92.89%**	92.84%	91.94%	92.35%	88.55%	92.05%	91.93%	92.02%
		GT Concordance	98.94%	98.75%	98.74%	98.78%	92.64%	99.08%	98.30%	**99.19%**
	ONT_L2	Precision	92.03%	**92.27%**	90.88%	91.05%	84.66%	85.56%	86.35%	79.67%
		Recall	**93.97%**	93.34%	89.04%	93.71%	88.61%	92.40%	92.54%	92.35%
		F1	**92.99%**	92.80%	89.95%	92.36%	86.59%	88.84%	89.34%	85.54%
		GT Concordance	**99.12%**	98.59%	98.72%	98.73%	90.18%	98.87%	98.03%	99.03%
	ONT_L3	Precision	**92.27%**	91.59%	90.96%	91.03%	85.39%	88.84%	88.84%	88.06%
		Recall	93.61%	92.13%	90.89%	**93.71%**	89.14%	92.42%	92.81%	92.42%
		F1	**92.93%**	91.86%	90.92%	92.35%	87.22%	90.59%	90.78%	90.18%
		GT Concordance	99.12%	98.71%	98.74%	98.81%	91.06%	99.11%	98.46%	**99.32%**
	Overall	Precision	**92.06%**	92.04%	90.98%	90.98%	85.51%	88.57%	88.69%	86.44%
		Recall	**93.83%**	92.97%	90.90%	93.77%	89.49%	92.53%	92.78%	92.40%
		F1	**92.94%**	92.50%	90.94%	92.35%	87.45%	90.49%	90.68%	89.25%
		GT Concordance	99.06%	98.68%	98.73%	98.77%	91.29%	99.02%	98.26%	**99.18%**
INS	ONT_L1	Precision	86.38%	90.18%	86.03%	85.28%	65.47%	**90.32%**	82.60%	89.00%
		Recall	92.84%	92.01%	92.71%	**93.33%**	89.45%	91.55%	88.45%	91.14%
		F1	89.50%	**91.09%**	89.25%	89.12%	75.60%	90.93%	85.42%	90.06%
		GT Concordance	98.14%	98.87%	97.22%	98.19%	63.74%	**98.88%**	85.70%	95.95%
	ONT_L2	Precision	86.24%	89.82%	84.72%	85.06%	63.83%	**89.86%**	81.65%	88.40%
		Recall	**93.60%**	92.05%	89.24%	92.90%	86.80%	90.57%	86.86%	90.19%
		F1	89.77%	**90.92%**	86.92%	88.80%	73.56%	90.21%	84.17%	89.29%
		GT Concordance	98.30%	**98.89%**	97.18%	97.92%	59.25%	98.49%	82.52%	95.28%
	ONT_L3	Precision	86.81%	89.93%	84.78%	84.48%	64.42%	**90.24%**	81.55%	88.65%
		Recall	**92.99%**	91.50%	90.72%	92.67%	86.99%	91.06%	88.87%	90.70%
		F1	89.80%	**90.71%**	87.65%	88.39%	74.03%	90.65%	85.05%	89.67%
		GT Concordance	98.37%	98.28%	97.10%	98.10%	60.58%	**98.96%**	86.96%	95.64%
	Overall	Precision	86.48%	89.98%	85.18%	84.94%	64.57%	**90.14%**	81.93%	88.68%
		Recall	**93.14%**	91.85%	90.89%	92.97%	87.75%	91.06%	88.06%	90.68%
		F1	89.69%	**90.91%**	87.94%	88.77%	74.40%	90.60%	84.88%	89.67%
		GT Concordance	98.27%	98.68%	97.17%	98.07%	61.19%	**98.78%**	85.06%	95.62%

Table 4 Large deletions (DELs) and insertions (INSs) ($\geq$50bp) calling performance across three ONT datasets. The table presents Recall, Precision, F1 score (in %), and Genotype accuracy (measured by Genotype Concordance) for all benchmarked SV callers. The highest scores for Recall, Precision, F1, and genotype accuracy are highlighted in bold. Additionally, the overall average performance for each SV type and library is summarized at the bottom of the table. Benchmarking was conducted using Truvari with the following parameter settings: p = 0.5, P = 0.5, r = 500, and O = 0.01.

Library	SV type	metric	FocalSV (target)	FocalSV (auto)	SVIM-asm	cuteSV	SVIM	pbsv	Sniffles2
HCC1395 PacBio	TRA	benchmark	137	137	137	137	137	137	137
		comp	119	74	25	30	58	452	173
		TP	91	58	11	14	49	73	41
		FP	28	16	14	16	9	379	132
		FN	46	79	126	123	88	64	96
		Recall	**66.42%**	42.34%	8.03%	10.22%	35.77%	53.28%	29.93%
		Precision	76.47%	78.38%	44.00%	46.67%	**84.48%**	16.15%	23.70%
		F1	**71.09%**	54.98%	13.58%	16.77%	50.26%	24.79%	26.45%
	INV	benchmark	133	133	133	133	133	133	133
		comp	104	49	19	101	27	62	88
		TP	55	30	8	9	23	31	29
		FP	49	19	11	92	4	31	59
		FN	78	103	125	124	110	102	104
		Recall	**41.35%**	22.56%	6.02%	6.77%	17.29%	23.31%	21.80%
		Precision	52.88%	61.22%	42.11%	8.91%	**85.19%**	50.00%	32.95%
		F1	**46.41%**	32.97%	10.53%	7.69%	28.75%	31.79%	26.24%
	DUP	benchmark	230	230	230	230	230	230	230
		comp	311	130	12	83	111	897	100
		TP	190	114	3	40	77	99	92
		FP	121	16	9	43	34	798	8
		FN	40	116	227	190	153	131	138
		Recall	**82.61%**	49.57%	1.30%	17.39%	33.48%	43.04%	40.00%
		Precision	61.09%	87.69%	25.00%	48.19%	69.37%	11.04%	**92.00%**
		F1	**70.24%**	63.33%	2.48%	25.56%	45.16%	17.57%	55.76%
HCC1395 ONT	TRA	benchmark	137	137	137	137	137	-	137
		comp	99	146	346	13	4	-	1862
		TP	82	57	10	8	3	-	43
		FP	17	89	336	5	1	-	1819
		FN	55	80	127	129	134	-	94
		Recall	**59.85%**	41.61%	7.30%	5.84%	2.19%	-	31.39%
		Precision	**82.83%**	39.04%	2.89%	61.54%	75.00%	-	2.31%
		F1	**69.49%**	40.28%	4.14%	10.67%	4.26%	-	4.30%
	INV	benchmark	133	133	133	133	133	-	133
		comp	71	65	21	11	5	-	38
		TP	49	20	3	2	2	-	26
		FP	22	45	18	9	3	-	12
		FN	84	113	130	131	131	-	107
		Recall	**36.84%**	15.04%	2.26%	1.50%	1.50%	-	19.55%
		Precision	**69.01%**	30.77%	14.29%	18.18%	40.00%	-	68.42%
		F1	**48.04%**	20.20%	3.90%	2.78%	2.90%	-	30.41%
	DUP	benchmark	230	230	230	230	230	-	230
		comp	257	110	9	27	13	-	84
		TP	168	93	2	11	9	-	76
		FP	89	17	7	16	4	-	8
		FN	62	137	228	219	221	-	154
		Recall	**73.04%**	40.43%	0.87%	4.78%	3.91%	-	33.04%
		Precision	65.37%	84.55%	22.22%	40.74%	69.23%	-	**90.48%**
		F1	**68.99%**	54.71%	1.67%	8.56%	7.41%	-	48.41%

Table 5 Somatic SV detection performance on paired normal-tumor HCC1395 breast cancer data using PacBio and ONT sequencing. The table summarizes the benchmark (ground truth set), comp (called set), TP (true positives), FP (false positives), FN (false negatives), recall, precision, and F1 scores for three SVs types - translocations (TRA), inversions (INV), and duplication (DUP) - evaluated across multiple benchmarked tools:

Figure 1 Schematic diagram of the FocalSV large indel detection pipeline. The workflow for large indel detection includes two modes: **(A)** single region mode and **(B)** multi-region mode. **(A) Single region mode:** the input data include a high-quality reference genome and a BAM file containing aligned long reads. The reads extraction module isolates long reads aligned to the region of interest. The haplotyping module partitions these reads into distinct parental haplotypes. The local assembly module uses the phased reads to perform independent de novo local assemblies. Finally, the variant calling module identifies indel structural variants (SVs) by comparing the assembled contigs to the reference genome, followed by filtering and genotype (GT) correction in postprocessing steps. **(B) Multi-region mode:** the input data includes a high-quality reference genome and a BAM file containing aligned long reads. FocalSV retrieves region-specific BAM files and processes each region independently through reads partitioning, local assembly, and SV detection. The VCF files from all regions are merged into a single file, and redundant variants are removed using a clustering algorithm. SV filtering and genotype refinement are then applied to produce the final VCF file.

Figure 2 Schematic diagram of duplication, translocation, and inversion breakpoint signatures detected by FocalSV. **(A) Duplication:** duplications can be identified using both contig-based and read-based BAM files. In a contig-based BAM file, a duplication is detected when an insertion call shows the alternate allele mapped to the surrounding sequence of the insertion breakpoint. In a read-based BAM file, a duplication signature is identified when two adjacent segments of a read align to overlapping regions on the reference genome in the same orientation. **(B) Translocation:** translocations are inferred when two adjacent segments of a read align to different chromosomes. **(C) Inversion:** inversions are detected when two adjacent segments of a read align in opposite orientations. Additional details on detecting duplications, translocations, and inversions can be found in the Methods section.

Figure 3 F1 accuracy of SV detection across various size ranges on nine long-read datasets. (A-C) F1 accuracy plot for three Hifi datasets. Negative ranges denote deletions and the positive ranges denote insertions. The bar plot illustrates the benchmark SV distribution across these size ranges. The line plot displays the F1 scores for four distinct detection methods. Dashed lines indicate alignment-based, while solid lines represent assembly-based methods. **(D-F)** F1 accuracy plot for three CLR datasets. **(G-I)** F1 accuracy plot for three ONT datasets.

Figure 4 F1 accuracy by changing four different evaluation parameters on Hifi_L2. (A-B) F1 score curves for deletions (DEL) and insertions (INS) across all tools as r is varied. r is the maximum reference location distance between SV call and gold standard SV. r varies from 0-1000 bp with a 100 bp interval. **(C-D)** F1 score curves for deletions (DEL) and insertions (INS) across all tools as O is varied. O is the minimum reciprocal overlap between SV call and gold standard SV. **(E-F)** F1 score curves for deletions (DEL) and insertions (INS) across all tools as P is varied. P is the minimum allowable allele size similarity between SV call and gold standard SV. **(G-H)** F1 score curves for deletions (DEL) and insertions (INS) across all tools as p is varied. p is the minimum percentage of allele sequence similarity between SV call and gold standard SV. O, P, and p vary from 0-1 with a 0.1 interval.

Figure 5 Deletion F1 accuracy of FocalSV(auto) by tuning different evaluation parameters (p and O). O is the minimum reciprocal overlap between SV call and gold standard SV. p is the minimum percentage of allele sequence similarity between SV call and gold standard SV. O and p vary from 0-1 with a 0.1 interval. **(A-C)** The F1 heatmap for deletions by FocalSV(auto) on three Hifi datasets. Every cell in the heatmap represents the F1 score under a specific pair of p and O evaluation. **(D-F)** The F1 heatmap for deletions by FocalSV(auto) on three CLR datasets. **(G-I)** The F1 heatmap for deletions by FocalSV(auto) on three ONT datasets.

Figure 6 Insertion F1 accuracy of FocalSV(auto) by tuning different evaluation parameters (p and r). r is the maximum reference location distance between SV call and gold standard SV. p is the minimum percentage of allele sequence similarity between SV call and gold standard SV. p vary from 0-1 with a 0.1 interval. r varies from 0-1000 bp with a 100 bp interval. **(A-C)** The F1 heatmap for insertions by FocalSV(auto) on three Hifi datasets. Every cell in the heatmap represents the F1 score under a specific pair of p and r evaluation. **(D-F)** The F1 heatmap for insertions by FocalSV(auto) on three CLR datasets. **(G-I)** The F1 heatmap for insertions by FocalSV(auto) on three ONT datasets.

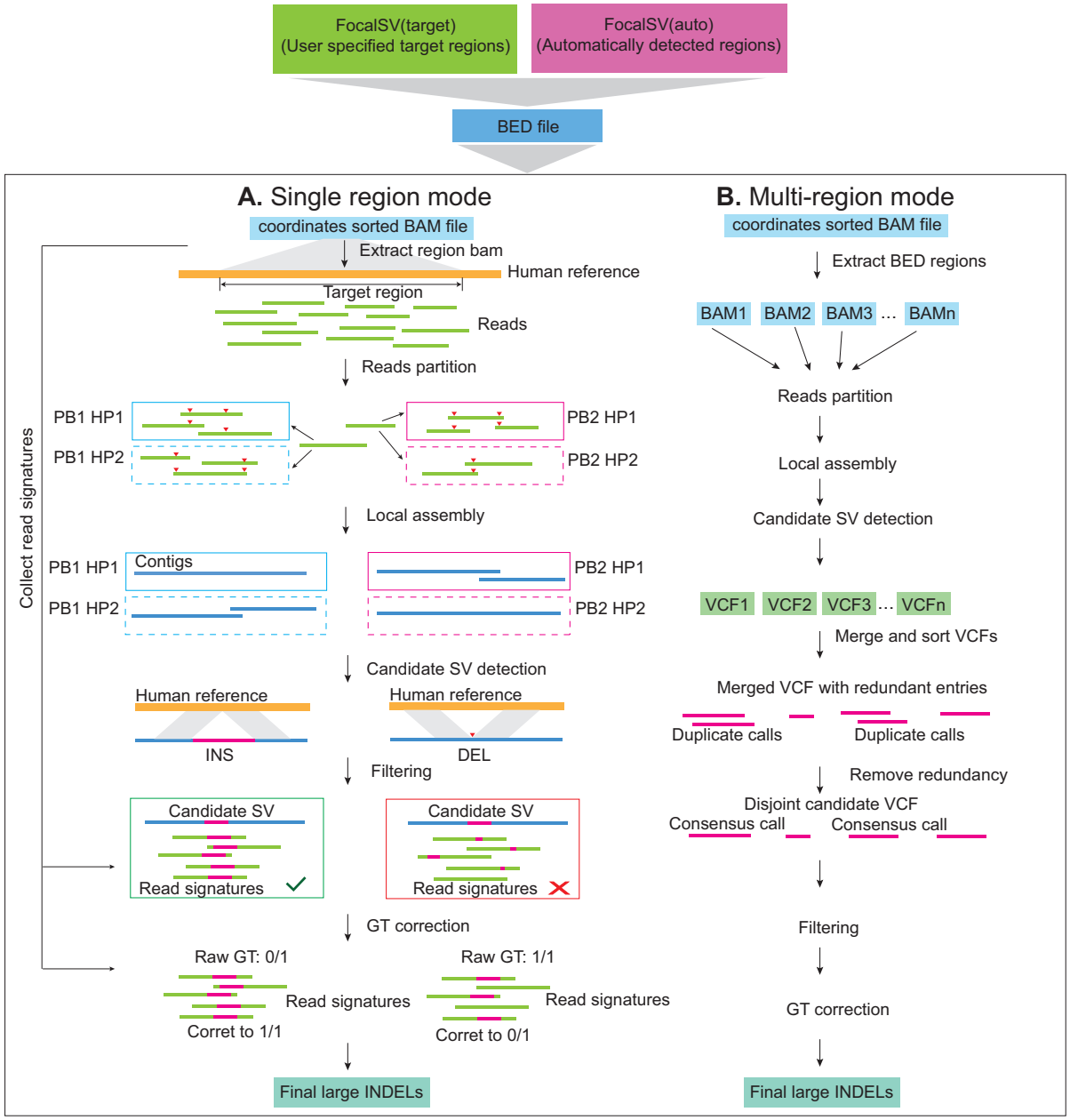

Figure 1

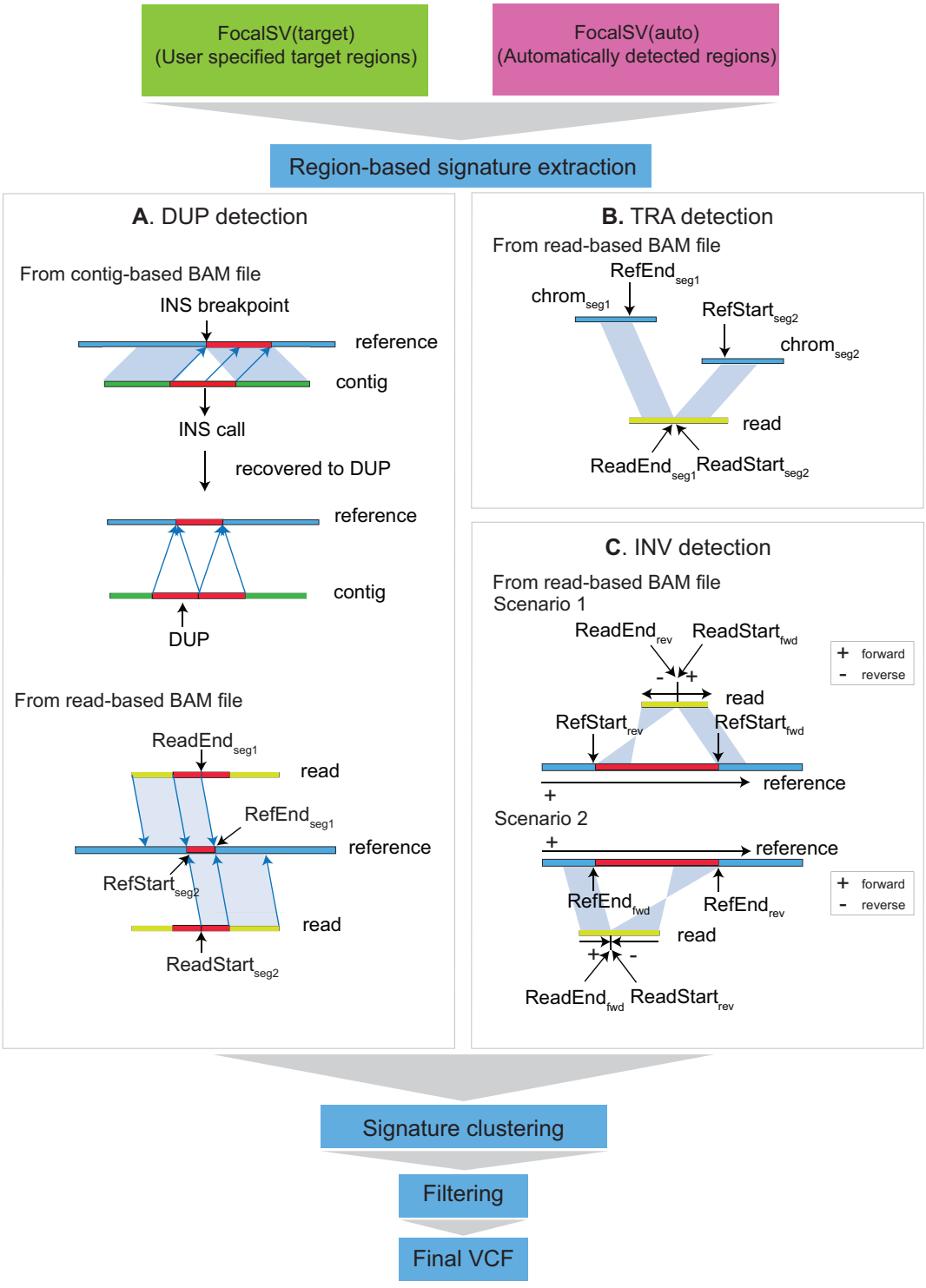

Figure 2

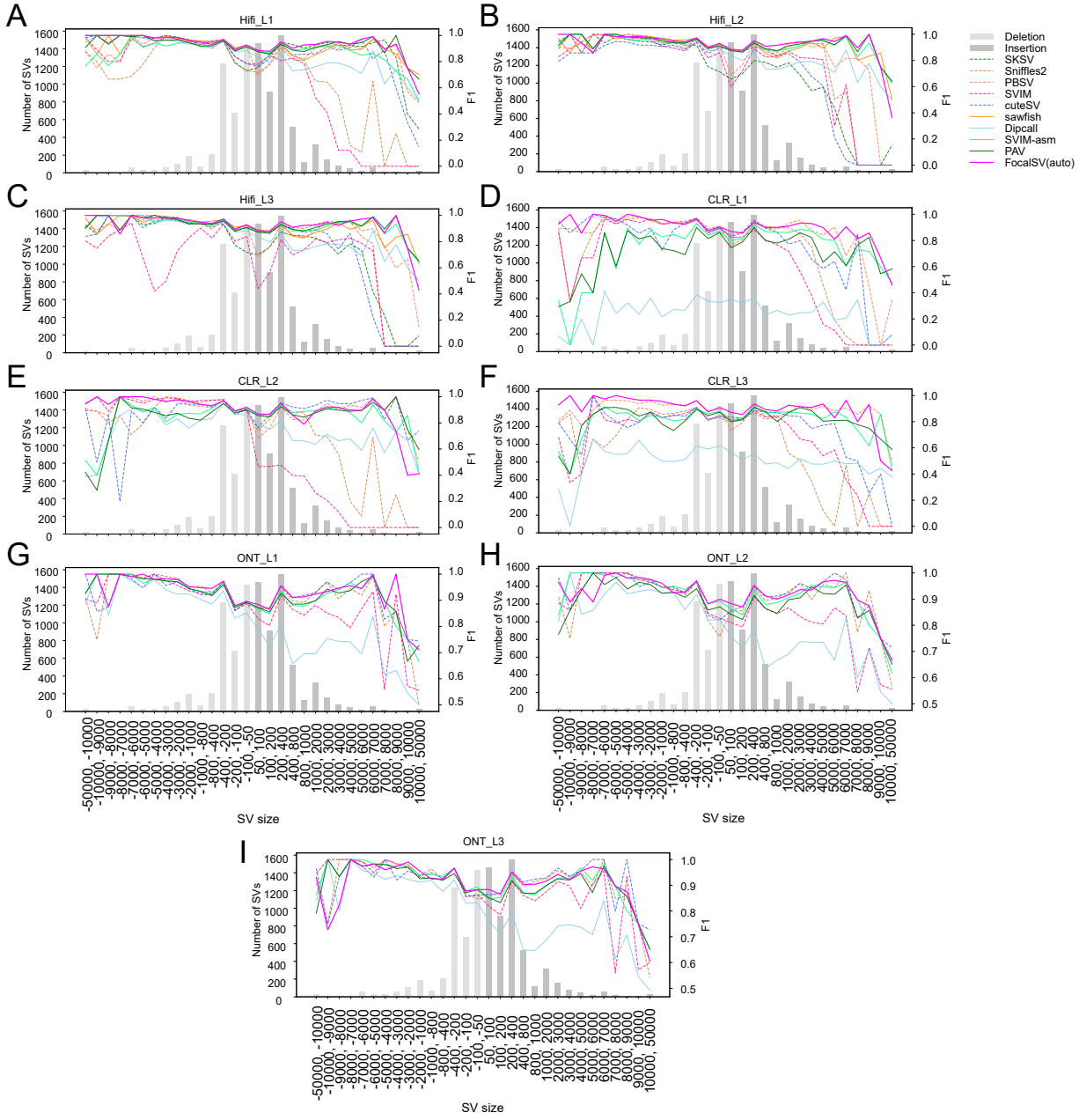

Figure 3

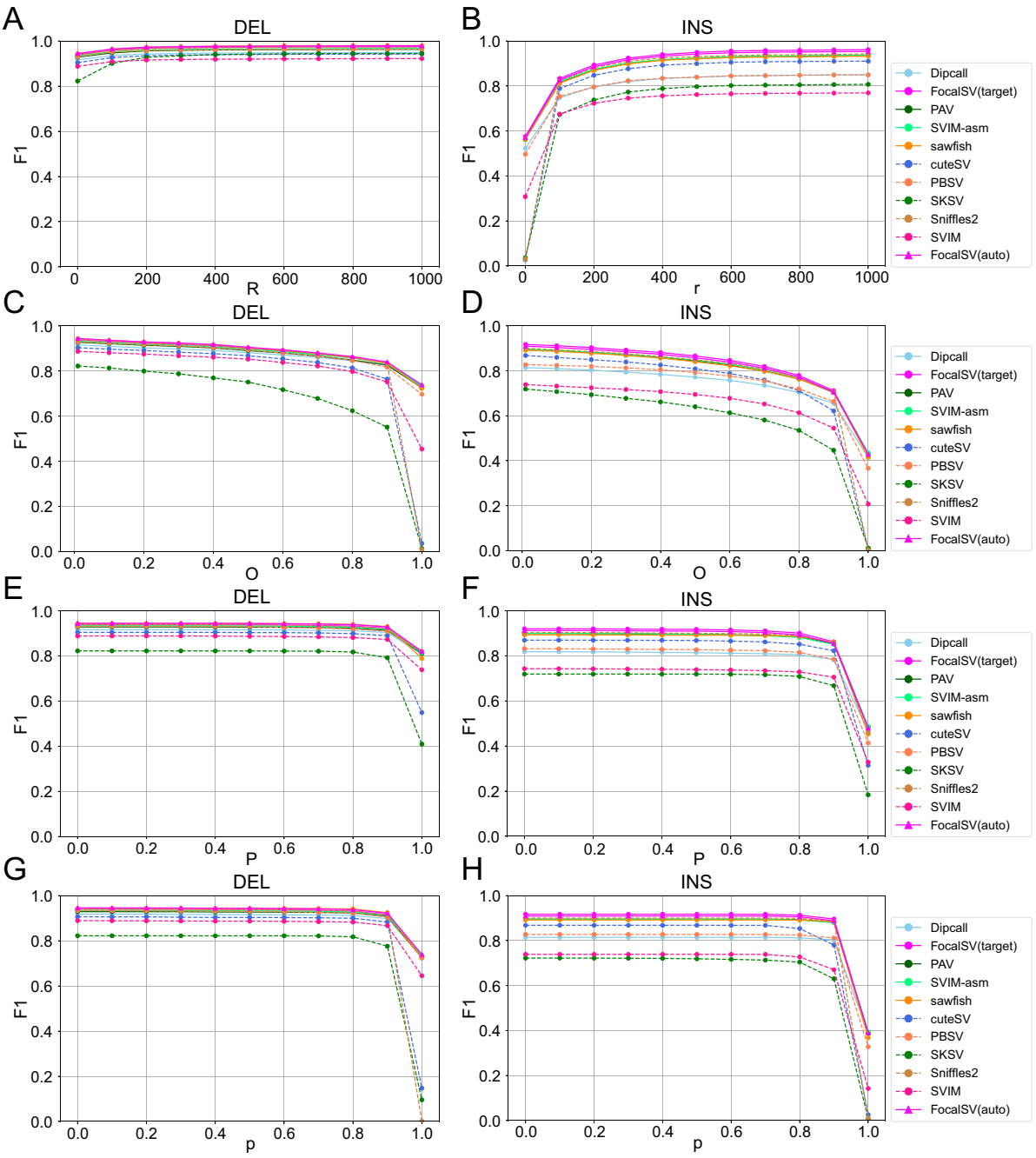

Figure 4

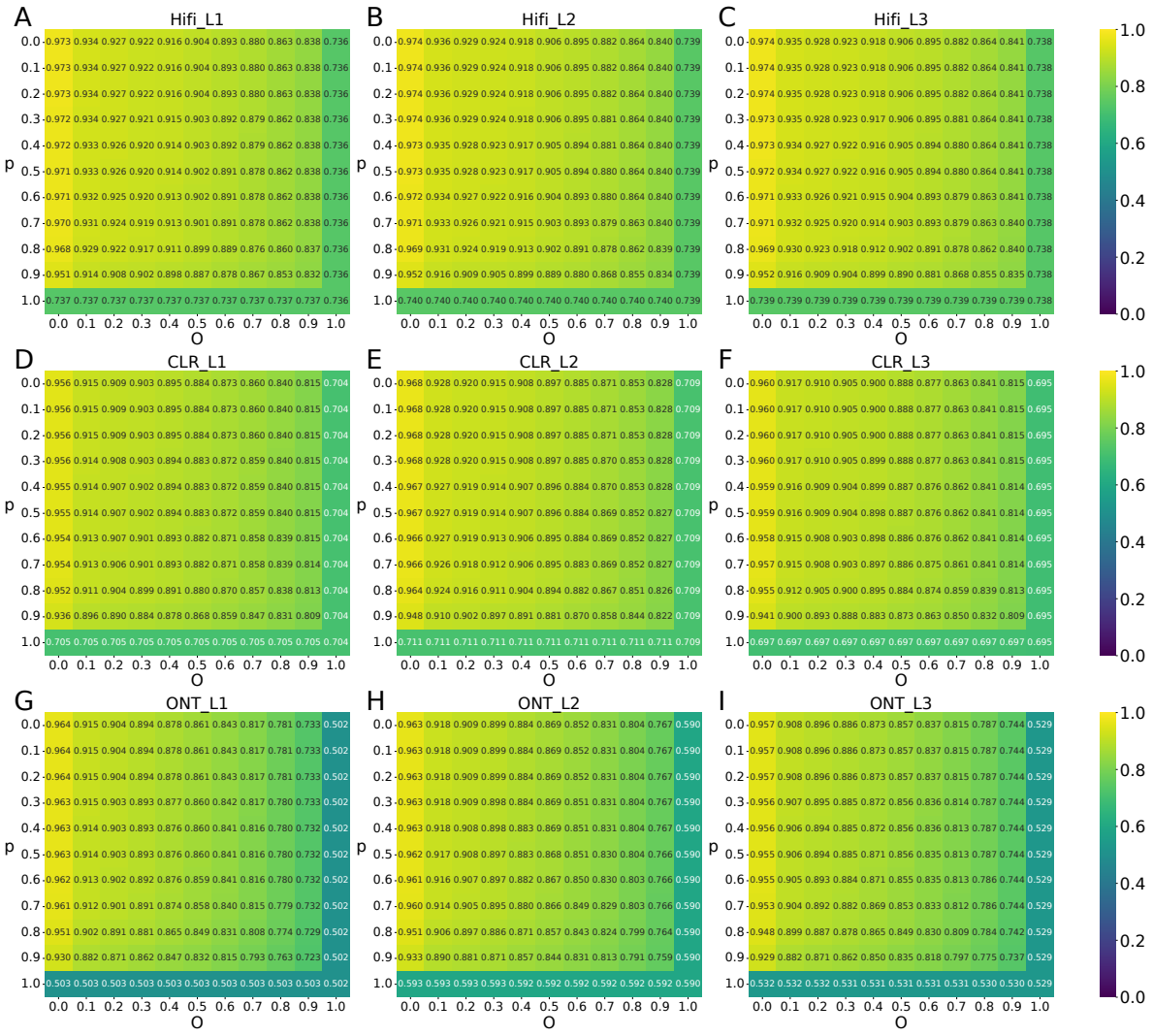

Figure 5

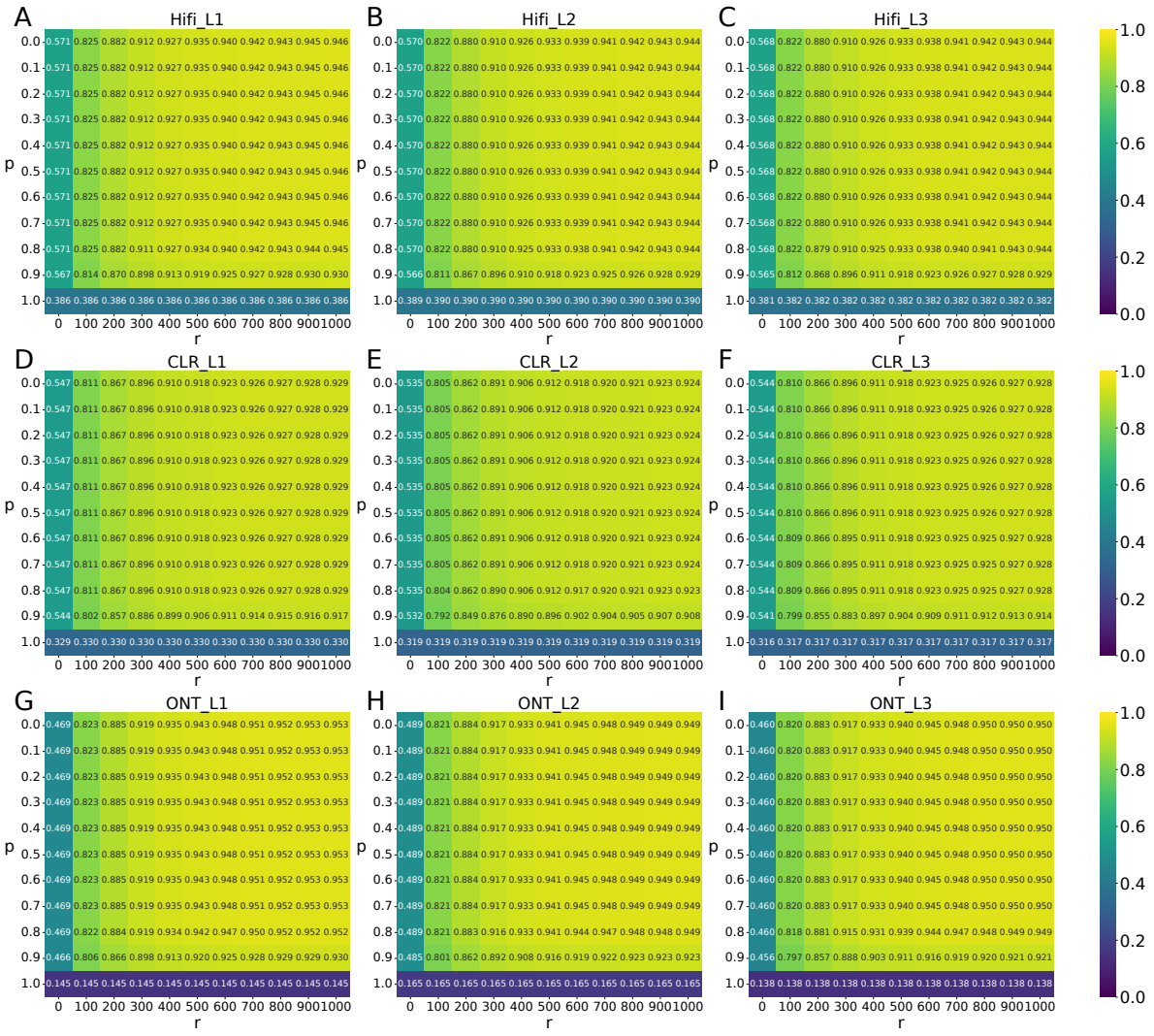

Figure 6