Supplemental material to: FocalSV: 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], Xin Maizie Zhou[1,2,*]

[1] *Department of Biomedical Engineering, Vanderbilt University, Nashville, TN, USA*
[2] *Department of Computer Science, Vanderbilt University, Nashville, TN, USA*
[†] *These authors contributed equally to this work.*
[*] *Corresponding author: maizie.zhou@vanderbilt.edu.*

Contents

1 Supplemental Results

S1 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. Phasing accuracy was first assessed using phased SNPs, with the GIAB v4.2.1 phased SNP set as the ground truth (Wagner et al. 2022), and evaluated using a publicly available phasing evaluation toolkit from (Choi 2020), based on metrics such as switch error rate (SER), the number of long switch errors, the number of point switch errors, the number of total switch errors, and the fraction of correctly phased heterozygous variants. Detailed definitions of these metrics and the evaluation procedure are described in the Supplemental Methods section. All three libraries (HiFi_L1, CLR_L1, and ONT_L1) demonstrated low SERs (median: 0%; mean: 2.2%-2.4%) and high fractions of correctly phased heterozygous variants (median: 97.8%-100%; mean: 93.3%-95.6%) across all regions (Supplemental Fig. S38). Notably, approximately 68% of regions achieved 0% SER. The median total switch error count was 0, while the mean was close to 1 (Supplemental Fig. S38). Overall, these phasing accuracy metrics indicate high phasing quality, supporting reliable downstream SV detection.

We also analyzed the distribution of phasing quality scores (PhaseQ), derived from Strand-seq data, across all phase blocks on the Hifi_L1 dataset generated by FocalSV(auto). The distribution was highly skewed, indicating strong enrichment for high-confidence phasing in the resulting local diploid assemblies. The detailed procedure for computing PhaseQ is described in the Supplemental Methods section and Supplemental Fig. S37. As shown in Supplemental Fig. S39, 52.8% of the phase blocks achieved a PhaseQ value greater than 0.95, reflecting near-perfect separation between the two haplotypes. Additionally, approximately 83.2% of blocks had PhaseQ values exceeding 0.8, demonstrating that the majority of regions exhibited robust phasing. In contrast, low-quality phase blocks (PhaseQ < 0.4) were rare, accounting for fewer than 3% of all blocks. These findings further underscore the robust phasing quality achieved by local assemblies across the majority of target regions.

2 Supplemental Methods

S1 Somatic SV discovery and evaluation in cancer genome

To assess the effectiveness of various SV calling methods in identifying somatic mutations, we utilized PacBio and ONT Tumor-Normal paired libraries of sample HCC1395. We also incorporated the high-confidence HCC1395 somatic SV call set as a gold standard, following the methodology outlined by Talsania et al (Talsania et al. 2022). Initially, each benchmarked tool was independently applied to detect SVs in tumor and matched normal samples using the GRCh38 reference genome. Subsequently, the SV filtering and merging tool SURVIVOR (Jeffares et al. 2017) was employed to extract somatic mutations. The resulting somatic SV call sets were then compared to the high-confidence somatic SV call set to assess performance.

The pipeline for employing SURVIVOR to identify somatic mutations is outlined as follows. To reduce false positives, we first extracted SV records annotated with "PASS" in the filter column to ensure that high confidence SV set was used for somatic calling. We then segmented the filtered VCF files according to SV type (translocation and non-translocation) and size windows (for non-translocations). The size windows recommended by Talsania et al. are as follows: 50-100bp, 101-500bp, 501-1000bp, 1001-30000bp, >30000bp. We utilized the following SURVIVOR command to split each VCF file:

```
SURVIVOR filter <input_vcf> \
NA  <min_sv_size> <max_sv_size> 0 -1 \
<out_vcf>
```

Next, somatic mutations were extracted by merging the tumor and normal SV call sets. Each VCF file of a specific size window employed the lower bound of its size window as the maximum threshold of breakpoint distance when merging SVs. For instance, in a 50-100bp VCF file, the threshold for breakpoint distance was set to 50bp. This strategy ensures consistency and accuracy in the merging process across different size windows.

```
SURVIVOR merge <VCFlist> \
<dist_thresh> 1 1 0 0 <min_sv_size> \
<merged_vcf>
```

The <VCFlist> denotes the VCF file path for the normal and tumor SV call sets. Following the merging procedure, SVs supported exclusively by the tumor call set, without corresponding support in the normal call set, were filtered out as somatic SVs. This filtering procedure was executed on non-TRA VCF files across various size windows. Subsequently, the five resulting filtered VCF files for non-TRA plus the TRA VCF file were concatenated into a single VCF file representing the final somatic mutation call set.

Before evaluation, post-processing steps were applied to the somatic INV and TRA callsets to standardize event representations and reduce redundant or ambiguous calls. During SV merging, SURVIVOR by default converts any breakend (BND) event with both breakpoints on the same chromosome into an INV, without validating the strand orientation of the breakpoints. However, true INVs require that the two breakpoints map to opposite strands. To address this, we applied an additional filter to retain only INV calls where breakpoints are located on opposite strands, improving the specificity of the INV callset. For TRAs, some SV callers report both breakend representations for the same interchromosomal event, depending on the ordering of breakpoints in the VCF. For example, a TRA linking `Chr1:1000` and `Chr2:2000` may be represented as `Chr1 1000 N[Chr2:2000[` or `Chr2 2000 ]Chr1:1000]N`. Since SURVIVOR does not automatically collapse such redundant BND pairs into a single TRA call, we implemented a custom post-processing procedure to identify and eliminate these duplicate entries prior to evaluation.

In the final step, the evaluation of somatic mutations against the high-confidence gold standard call set was conducted using distinct criteria. Specifically, for TRAs, assessment was carried out solely at the breakend level. The breakend shift (r) between TRAs from the call set and the gold standard set was calculated. If the value of r is less than or equal to 1kb, the TRA is classified as true positive (TP); otherwise, it is labeled as false positive (FP). For non-TRA SVs, both the breakend shift (r) and size similarity (P) are calculated.

$$\begin{cases} r = max(|Start_1 - Start_2|, |End_1 - End_2|) \\ P = \dfrac{min(SVlen_1, SVlen_2)}{max(SVlen_1, SVlen_2)} \end{cases} \tag{1}$$

For non-TRA SVs, if the value of r is less than or equal to 500bp and P is greater than or equal to 0.5, the non-TRA SV is classified as true positive (TP); otherwise, it is categorized as false positive (FP). This stringent assessment approach ensures the robustness and accuracy of the evaluation process, effectively distinguishing true somatic mutations from potential artifacts or false positives.

S2 Phasing evaluation

To assess phasing accuracy, we utilized the *measure_phasing_performance.pl* script from the publicly available phasing evaluation toolkit (Choi 2020). The tool compares predicted phased variants against high-confidence reference haplotypes and computes multiple phasing performance metrics, as detailed below. The evaluation was performed using our phased VCF files for sample HG002, with the Genome in a Bottle (GIAB) v4.2.1 phased SNP set serving as the ground truth reference. For each chromosome, chromosome-specific VCFs were evaluated by executing the following command:

```
measure_phasing_performance.pl \
-i <predicted.vcf.gz> \
-r <truth.vcf.gz> \
-o <output_prefix>
```

The evaluation reports multiple fine-grained phasing metrics, from which we selected five representative statistics to summarize performance: (1) switch error rate (SER, %), reflecting the proportion of incorrectly phased heterozygous variant transitions; (2) the number of long switch errors, representing major haplotype block misassignments; (3) the number of point switch errors, capturing isolated phasing inconsistencies at individual heterozygous sites; (4) the total switch error count, computed as the sum of long switch errors, twice the number of point switch errors, and any undetermined errors, following the convention described by Choi et al. (Choi 2020); and (5) the fraction of correctly phased heterozygous variants, indicating the overall accuracy of phasing across the genome. Collectively, these metrics provide a comprehensive evaluation of phasing accuracy at both local and long-range genomic scales.

To further evaluate the phasing accuracy of diploid local assemblies across all target regions by FocalSV, we developed a quantitative assessment pipeline that integrates Strand-seq data with a high-quality telemere-to-telemere (T2T) diploid assembly of HG002 (Supplemental Fig. S37). The complete, gapless T2T HG002 assembly provides an unambiguous haplotype backbone, while Strand-seq offers orthogonal, chromosome-length phase information that enables robust validation of long-range phasing accuracy. The T2T HG002 diploid assembly was first aligned to the hg19 reference genome using minimap2. Based on the resulting alignment records, we extracted the T2T contig segments corresponding to genomic regions where FocalSV had performed local diploid assemblies to detect SVs for HG002 (Supplemental Fig. S37A).

Next, Strand-seq reads from individual cells were independently aligned to the T2T diploid assembly (Supplemental Fig. S37B). For each contig, we calculated a strand-state metric, the CW ratio, defined as:

$$\mathrm{CW} = \frac{C - W}{C + W}$$

where W and C denote the number of read 1 alignments mapped to the T2T contig in the Watson (–) and Crick (+) orientations, respectively. Since the T2T diploid assembly includes both maternal and paternal contigs for each chromosome, we assessed CW ratios separately for each haplotype to identify informative Strand-seq cells. A cell was considered informative for a given chromosome if:

$$\mathrm{CW}_{\mathrm{mat}} \cdot \mathrm{CW}_{\mathrm{pat}} < 0 \quad \& \quad \min\left(|\mathrm{CW}_{\mathrm{mat}}|,\ |\mathrm{CW}_{\mathrm{pat}}|\right) > 0.75$$

Where $\mathrm{CW}_{\mathrm{pat}}$ and $\mathrm{CW}_{\mathrm{mat}}$ represent the CW ratio on the paternal and maternal contigs of a chromosome, respectively. Informative cells were then classified by haplotype polarity: cells with $\mathrm{CW}_{\mathrm{pat}} > 0$ were labeled as P cells (indicating the paternal strand was in the forward orientation and the maternal strand was in the reverse orientation), while cells with $\mathrm{CW}_{\mathrm{pat}} < 0$ were labeled as M cells (indicating the maternal strand was in the forward orientation, and the paternal strand was in the reverse orientation). From these informative cells, we extracted reads overlapping the previously identified T2T assembly regions and realigned them to the corresponding FocalSV phase blocks (i.e., haplotype 1 (HP1) and haplotype 2 (HP2) contigs) of each chromosome (Supplemental Fig. S37C).

To quantify haplotypic strand bias, we defined read classification based on strand orientation and cell polarity. Specifically, a read 1 from a P cell was labeled a P read if it aligned to the contig in the forward orientation, and as an M read if it aligned in the reverse orientation. For reads from M cells, this rule was reversed: forward alignments were considered M reads, and reverse alignments were considered P reads. Using these definitions, the PM ratio for each phase block contig was calculated as:

$$\mathrm{PM} = \frac{P - M}{P + M}$$

where P and M are the total numbers of P and M reads collected from all informative cells, respectively.

Finally, we calculated the phasing quality score (PhaseQ) for each phase block of FocalSV as:

$$\text{PhaseQ} = \frac{|\text{PM}_{\text{HP1}} - \text{PM}_{\text{HP2}}|}{2}$$

where PM_{HP1} and PM_{HP2} denote the PM ratios computed for the haplotype 1 and haplotype 2 contigs, respectively. PhaseQ values range from 0 to 1, with higher values indicating stronger haplotypic separation (Supplemental Fig. S37C). Ideally, in a perfectly phased block, the two haplotypes (HP1 and HP2) should exhibit distinct and high haplotype polarity. That is, the majority of Strand-seq reads aligned to HP1 should originate from one parental haplotype (either P or M), while those mapped to HP2 should predominantly derive from the opposite haplotype. Under such conditions, the resulting PhaseQ value is expected to approach 1. Conversely, a PhaseQ value near 0 may indicate either low haplotype polarity - where Strand-seq reads from both parental haplotypes are mixed within HP1 and HP2 - or that HP1 and HP2 share the same haplotype polarity, suggesting they represent redundant or incorrectly phased representations of the same haplotype sequence. Only phase blocks with at least one Strand-seq read coverage on both haplotype contigs were included in the final evaluation.

3 Supplemental Tables

Supplemental Table S1: Large deletions (DELs) and insertions (INSs) (≥50bp) calling performance across three Hifi datasets. The table presents True Positive (TP), False Positive (FP), False Negative (FN), Recall, Precision, F1 score (in %), Genotype TP, Genotype FP, 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. 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	sawfish	cuteSV	SVIM	PBSV	Sniffles2	SKSV
DEL	Hifi_L1	**TP**	3925	3901	3837	3865	3812	3924	3641	3654	3846	3484	3509
		FP	289	267	321	350	408	291	487	530	288	676	605
		FN	191	215	279	251	304	192	475	462	270	632	607
		Precision	93.14%	**93.59**%	92.28%	91.7%	90.33%	93.1%	88.2%	87.33%	93.03%	83.75%	85.29%
		Recall	**95.36**%	94.78%	93.22%	93.9%	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%	94.2%	88.33%	88.05%	93.24%	84.2%	85.27%
		Genotype TP	3782	3864	3726	3782	3688	3860	3607	3607	3758	3364	3438
		Genotype FP	143	37	111	83	124	64	34	47	88	120	71
		GT Concord	99.1%	99.05%	97.11%	97.85%	96.75%	98.37%	99.07%	98.71%	97.71%	96.56%	97.98%
	Hifi_L2	**TP**	3910	3911	3874	3892	3818	3923	3586	3926	3860	3904	3248
		FP	310	258	335	338	419	280	236	809	326	309	537
		FN	206	205	242	224	298	193	530	190	256	212	868
		Precision	92.65%	93.81%	92.04%	92.01%	90.11%	93.34%	**93.83**%	82.9%	92.21%	92.67%	85.81%
		Recall	95.0%	95.02%	94.12%	94.56%	92.76%	95.31%	87.12%	95.4%	93.78%	94.85%	78.91%
		F1	93.81%	**94.41**%	93.07%	93.27%	91.42%	94.31%	90.35%	88.7%	92.99%	93.74%	82.22%
		Genotype TP	3871	3872	3768	3838	3698	3853	3495	3844	3790	3838	3166
		Genotype FP	39	39	106	54	120	70	91	82	70	66	82
		GT Concord	99.0%	99.0%	97.26%	98.61%	96.86%	98.22%	97.46%	97.9%	98.19%	98.31%	97.48%
	Hifi_L3	**TP**	3919	3905	3848	3887	3818	3917	3722	3924	3866	3911	3352
		FP	310	262	348	352	427	288	228	1287	314	303	551
		FN	197	211	268	229	298	199	394	192	250	205	764
		Precision	92.67%	93.71%	91.71%	91.7%	89.94%	93.15%	**94.23**%	75.3%	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%	93.2%	93.9%	83.6%
		Genotype TP	3879	3866	3718	3832	3697	3856	3649	3859	3812	3851	3287
		Genotype FP	40	39	130	55	121	61	73	65	54	60	65
		GT Concord	98.98%	99.0%	96.62%	98.59%	96.83%	98.44%	98.04%	98.34%	98.6%	98.47%	98.06%
INS	Hifi_L1	**TP**	4961	4958	4917	4916	4824	4977	4348	3677	4088	4053	4572
		FP	501	644	778	766	1801	956	620	645	455	1075	736
		FN	320	323	364	365	457	304	933	1604	1193	1228	709
		Precision	90.8%	88.5%	86.3%	86.5%	72.8%	83.89%	87.5%	85.1%	90.0%	79.0%	86.1%
		Recall	93.9%	93.88%	93.1%	93.1%	91.4%	**94.24**%	82.3%	69.6%	77.4%	76.8%	86.6%
		F1	92.4%	91.11%	89.6%	89.7%	81.0%	88.76%	84.9%	76.6%	83.2%	77.9%	86.4%
		Genotype TP	4866	4878	4699	4809	3755	4598	4307	3309	3586	3621	4495
		Genotype FP	95	80	218	107	1069	379	41	368	502	432	77
		GT Concord	98.1%	98.39%	95.6%	97.8%	77.8%	92.38%	99.1%	90.0%	87.7%	89.3%	98.3%
	Hifi_L2	**TP**	4952	4952	4929	4907	4837	4967	4351	4838	4050	4849	3257
		FP	563	655	779	774	1776	894	391	2976	457	642	522
		FN	329	329	352	374	444	314	930	443	1231	432	2024
		Precision	89.8%	88.32%	86.4%	86.4%	73.1%	84.75%	91.8%	61.9%	89.9%	88.3%	86.2%
		Recall	93.8%	93.77%	93.3%	92.9%	91.6%	**94.05**%	82.4%	91.6%	76.7%	91.8%	61.7%
		F1	91.7%	90.96%	89.7%	89.5%	81.3%	89.16%	86.8%	73.9%	82.8%	90.0%	71.9%
		Genotype TP	4868	4863	4737	4793	3798	4617	4248	4546	3864	4703	3182
		Genotype FP	84	89	192	114	1039	350	103	292	186	146	75
		GT Concord	98.3%	98.2%	96.1%	97.7%	78.5%	92.95%	97.6%	94.0%	95.4%	97.0%	97.7%
	Hifi_L3	**TP**	4953	4958	4912	4919	4846	4966	4539	4866	4048	4901	3751
		FP	574	656	781	777	1795	923	415	5940	499	628	556
		FN	328	323	369	362	435	315	742	415	1233	380	1530
		Precision	89.6%	88.31%	86.3%	86.4%	73.0%	84.33%	91.6%	45.0%	89.0%	88.6%	87.1%
		Recall	93.8%	93.88%	93.0%	93.2%	91.8%	**94.04**%	86.0%	92.1%	76.7%	92.8%	71.0%
		F1	91.7%	91.01%	89.5%	89.6%	81.3%	88.92%	88.7%	60.5%	82.4%	90.7%	78.2%
		Genotype TP	4878	4879	4661	4818	3794	4591	4466	4586	3894	4757	3699
		Genotype FP	75	79	251	101	1052	375	73	280	154	144	52
		GT Concord	98.5%	98.41%	94.9%	98.0%	78.3%	92.45%	98.4%	94.3%	96.2%	97.1%	98.6%

Supplemental Table S2: Large deletions (DELs) and insertions (INSs) (≥50bp) calling performance across three CLR datasets. The table presents True Positive (TP), False Positive (FP), False Negative (FN), Recall, Precision, F1 score (in %), Genotype TP, Genotype FP, 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. 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	**TP**	3876	3770	3576	3544	866	3679	3774	3782	3779
		FP	348	270	811	460	185	318	339	242	307
		FN	240	346	540	572	3250	437	342	334	337
		Precision	91.76%	93.32%	81.51%	88.51%	82.4%	92.04%	91.8%	**93.99**%	92.49%
		Recall	**94.17**%	91.59%	86.88%	86.1%	21.04%	89.38%	91.7%	91.89%	91.81%
		F1	**92.95**%	92.45%	84.11%	87.29%	33.52%	90.69%	91.7%	92.92%	92.15%
		Genotype TP	3828	3722	3433	3375	740	3629	3694	3655	3729
		Genotype FP	48	48	143	169	126	50	80	127	50
		GT Concord	**98.76**%	98.73%	96.0%	95.23%	85.45%	98.64%	97.9%	96.64%	98.68%
	CLR_L2	**TP**	3903	3868	3664	3775	3579	3786	3863	3826	3796
		FP	338	272	353	328	451	355	273	275	289
		FN	213	248	452	341	537	330	253	290	320
		Precision	92.03%	**93.43**%	91.21%	92.01%	88.81%	91.43%	93.4%	93.29%	92.93%
		Recall	**94.83**%	93.97%	89.02%	91.72%	86.95%	91.98%	93.9%	92.95%	92.23%
		F1	93.41%	93.7%	90.1%	91.86%	87.87%	91.7%	93.6%	93.12%	92.57%
		Genotype TP	3853	3829	3605	3717	3377	3759	3808	3742	3696
		Genotype FP	50	39	59	58	202	27	55	84	100
		GT Concord	98.72%	98.99%	98.39%	98.46%	94.36%	**99.29**%	98.6%	97.8%	97.37%
	CLR_L3	**TP**	3809	3793	3472	3469	1907	3297	3265	3758	3224
		FP	311	282	459	360	389	278	245	302	203
		FN	307	323	644	647	2209	819	851	358	892
		Precision	92.45%	93.08%	88.32%	90.6%	83.06%	92.22%	93.0%	92.56%	**94.08**%
		Recall	**92.54**%	92.15%	84.35%	84.28%	46.33%	80.1%	79.3%	91.3%	78.33%
		F1	92.5%	**92.61**%	86.29%	87.33%	59.48%	85.74%	85.6%	91.93%	85.48%
		Genotype TP	3704	3705	3321	3336	1607	3223	3226	3637	3095
		Genotype FP	105	88	151	133	300	74	39	121	129
		GT Concord	97.24%	97.68%	95.65%	96.17%	84.27%	97.76%	98.8%	96.78%	96.0%
INS	CLR_L1	**TP**	4903	4796	4815	4750	1097	4548	3760	4213	4576
		FP	597	660	1868	1172	479	693	295	310	623
		FN	378	485	466	531	4184	733	1521	1068	705
		Precision	89.15%	87.9%	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%	71.2%	79.78%	86.65%
		F1	**90.96**%	89.34%	80.49%	84.8%	32.0%	86.45%	80.55%	85.94%	87.33%
		Genotype TP	4753	4684	4418	4370	746	4459	2320	3689	4426
		Genotype FP	150	112	397	380	351	89	1440	524	150
		GT Concord	96.94%	97.66%	91.75%	92.0%	68.0%	**98.04**%	61.7%	87.56%	96.72%
	CLR_L2	**TP**	4939	4935	4846	4928	4654	4852	1460	4114	4272
		FP	675	889	913	896	2317	983	863	664	993
		FN	342	346	435	353	627	429	3821	1167	1009
		Precision	**87.98**%	84.74%	84.15%	84.62%	66.76%	83.15%	62.85%	86.1%	81.14%
		Recall	**93.52**%	93.45%	91.76%	93.32%	88.13%	91.88%	27.65%	77.9%	80.89%
		F1	**90.67**%	88.88%	87.79%	88.75%	75.97%	87.3%	38.4%	81.8%	81.02%
		Genotype TP	4792	4877	4646	4769	3146	4802	644	3683	4057
		Genotype FP	147	58	200	159	1508	50	816	431	215
		GT Concord	97.02%	98.82%	95.87%	96.77%	67.6%	**98.97**%	44.11%	89.52%	94.97%
	CLR_L3	**TP**	4829	4831	4843	4774	2497	4216	3986	4221	3719
		FP	561	711	1294	1083	1637	566	362	599	436
		FN	452	450	438	507	2784	1065	1295	1060	1562
		Precision	89.59%	87.17%	78.91%	81.51%	60.4%	88.16%	**91.67**%	87.57%	89.51%
		Recall	91.44%	91.48%	**91.71**%	90.4%	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%
		Genotype TP	4591	4648	4322	4290	1355	4057	3495	3742	3445
		Genotype FP	238	183	521	484	1142	159	491	479	274
		GT Concord	95.07%	96.21%	89.24%	89.86%	54.27%	**96.23**%	87.68%	88.65%	92.63%

Supplemental Table S3: Large deletions (DELs) and insertions (INSs) (≥50bp) calling performance across ONT datasets. The table presents True Positive (TP), False Positive (FP), False Negative (FN), Recall, Precision, F1 score (in %), Genotype TP, Genotype FP, 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. 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	**TP**	3865	3846	3819	3865	3734	3819	3828	3805
		FP	341	323	373	389	584	363	384	349
		FN	251	270	297	251	382	297	288	311
		Precision	91.89%	**92.25**%	91.1%	90.86%	86.48%	91.32%	90.88%	91.6%
		Recall	93.9%	93.44%	92.78%	93.9%	90.72%	92.78%	93.0%	92.44%
		F1	**92.89**%	92.84%	91.94%	92.35%	88.55%	92.05%	91.93%	92.02%
		Genotype TP	3824	3798	3771	3818	3459	3784	3763	3774
		Genotype FP	41	48	48	47	275	35	65	31
		GT Concord	98.94%	98.75%	98.74%	98.78%	92.64%	99.08%	98.3%	**99.19**%
	ONT_L2	**TP**	3868	3842	3665	3857	3647	3803	3809	3801
		FP	335	322	368	379	661	642	602	970
		FN	248	274	451	259	469	313	307	315
		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%	92.4%	92.54%	92.35%
		F1	**92.99**%	92.8%	89.95%	92.36%	86.59%	88.84%	89.34%	85.54%
		Genotype TP	3834	3788	3618	3808	3289	3760	3734	3764
		Genotype FP	34	54	47	49	358	43	75	37
		GT Concord	**99.12**%	98.59%	98.72%	98.73%	90.18%	98.87%	98.03%	99.03%
	ONT_L3	**TP**	3853	3792	3741	3857	3669	3804	3820	3804
		FP	323	348	372	380	628	478	480	516
		FN	263	324	375	259	447	312	296	312
		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%
		Genotype TP	3819	3743	3694	3811	3341	3770	3761	3778
		Genotype FP	34	49	47	46	328	34	59	26
		GT Concord	99.12%	98.71%	98.74%	98.81%	91.06%	99.11%	98.46%	**99.32**%
INS	ONT_L1	**TP**	4903	4859	4896	4929	4724	4835	4671	4813
		FP	773	529	795	851	2492	518	984	595
		FN	378	422	385	352	557	446	610	468
		Precision	86.38%	90.18%	86.03%	85.28%	65.47%	**90.32**%	82.6%	89.0%
		Recall	92.84%	92.01%	92.71%	**93.33**%	89.45%	91.55%	88.45%	91.14%
		F1	89.5%	**91.09**%	89.25%	89.12%	75.6%	90.93%	85.42%	90.06%
		Genotype TP	4812	4804	4760	4840	3011	4781	4003	4618
		Genotype FP	91	55	136	89	1713	54	668	195
		GT Concord	98.14%	98.87%	97.22%	98.19%	63.74%	**98.88**%	85.7%	95.95%
	ONT_L2	**TP**	4943	4861	4713	4906	4584	4783	4587	4763
		FP	789	551	850	862	2598	540	1031	625
		FN	338	420	568	375	697	498	694	518
		Precision	86.24%	89.82%	84.72%	85.06%	63.83%	**89.86**%	81.65%	88.4%
		Recall	93.6%	92.05%	89.24%	92.9%	86.8%	90.57%	86.86%	90.19%
		F1	89.77%	**90.92**%	86.92%	88.8%	73.56%	90.21%	84.17%	89.29%
		Genotype TP	4859	4807	4580	4804	2716	4711	3785	4538
		Genotype FP	84	54	133	102	1868	72	802	225
		GT Concord	98.3%	**98.89**%	97.18%	97.92%	59.25%	98.49%	82.52%	95.28%
	ONT_L3	**TP**	4911	4832	4791	4894	4594	4809	4693	4790
		FP	746	541	860	899	2537	520	1062	613
		FN	370	449	490	387	687	472	588	491
		Precision	86.81%	89.93%	84.78%	84.48%	64.42%	**90.24**%	81.55%	88.65%
		Recall	**92.99**%	91.5%	90.72%	92.67%	86.99%	91.06%	88.87%	90.7%
		F1	89.8%	**90.71**%	87.65%	88.39%	74.03%	90.65%	85.05%	89.67%
		Genotype TP	4831	4749	4652	4801	2783	4759	4081	4581
		Genotype FP	80	83	139	93	1811	50	612	209
		GT Concord	98.37%	98.28%	97.1%	98.1%	60.58%	**98.96**%	86.96%	95.64%

Supplemental Table S4: **Deletion calling performance evaluation on HG005 by tool support.** The table presents number of SVs in the bench set, Recall, Precision, F1 score (in %) for all benchmarked SV callers. The highest scores for Recall, Precision, F1 are highlighted in bold.

number of supporting tools	metric	FocalSV	PAV	SVIM-asm	Dipcall	sawfish	cuteSV	SVIM	PBSV	Sniffles2	SKSV
2	**bench**	4710	4710	4710	4710	4710	4710	4710	4710	4710	4710
	Recall	93.93%	94.44%	**94.54**%	90.74%	94.37%	92.34%	93.42%	93.25%	94.20%	90.25%
	Precision	98.77%	97.76%	97.35%	94.56%	97.46%	**98.91**%	98.30%	98.74%	98.23%	95.83%
	F1	**96.29**%	96.07%	95.93%	92.61%	95.89%	95.51%	95.80%	95.92%	96.17%	92.96%
3	**bench**	4625	4625	4625	4625	4625	4625	4625	4625	4625	4625
	Recall	95.57%	**95.91**%	95.46%	91.70%	95.55%	94.01%	95.05%	94.77%	95.76%	91.52%
	Precision	98.68%	97.49%	96.52%	93.83%	96.89%	**98.89**%	98.21%	98.54%	98.05%	95.42%
	F1	**97.10**%	96.70%	95.99%	92.75%	96.21%	96.39%	96.60%	96.62%	96.89%	93.43%
4	**bench**	4552	4552	4552	4552	4552	4552	4552	4552	4552	4552
	Recall	96.62%	**96.99**%	96.05%	92.38%	96.75%	95.25%	96.22%	96.05%	96.88%	92.82%
	Precision	98.19%	97.03%	95.58%	93.03%	96.56%	**98.61**%	97.86%	98.29%	97.63%	95.24%
	F1	**97.40**%	97.01%	95.81%	92.70%	96.65%	96.90%	97.03%	97.16%	97.25%	94.01%
5	**bench**	4505	4505	4505	4505	4505	4505	4505	4505	4505	4505
	Recall	97.29%	97.49%	96.60%	92.87%	97.38%	96.03%	96.89%	96.65%	**97.58**%	93.45%
	Precision	97.86%	96.53%	95.15%	92.57%	96.19%	**98.39**%	97.52%	97.89%	97.32%	94.91%
	F1	**97.57**%	97.01%	95.87%	92.72%	96.78%	97.19%	97.21%	97.26%	97.45%	94.17%
6	**bench**	4472	4472	4472	4472	4472	4472	4472	4472	4472	4472
	Recall	97.74%	**97.99**%	96.96%	93.25%	97.65%	96.47%	97.27%	96.94%	**97.99**%	93.85%
	Precision	97.59%	96.31%	94.80%	92.26%	95.75%	**98.11**%	97.18%	97.46%	97.01%	94.61%
	F1	**97.67**%	97.14%	95.87%	92.75%	96.69%	97.28%	97.23%	97.20%	97.50%	94.23%
7	**bench**	4440	4440	4440	4440	4440	4440	4440	4440	4440	4440
	Recall	98.09%	98.22%	97.18%	93.56%	97.95%	96.87%	97.57%	97.18%	**98.31**%	94.21%
	Precision	97.23%	95.85%	94.34%	91.90%	95.35%	**97.82**%	96.78%	97.01%	96.63%	94.30%
	F1	**97.66**%	97.02%	95.74%	92.72%	96.63%	97.34%	97.17%	97.10%	97.47%	94.25%
8	**bench**	4375	4375	4375	4375	4375	4375	4375	4375	4375	4375
	Recall	98.54%	**98.58**%	97.67%	94.10%	98.06%	97.76%	98.06%	97.51%	98.47%	95.06%
	Precision	96.25%	94.79%	93.42%	91.08%	94.06%	**97.27**%	95.84%	95.91%	95.37%	93.76%
	F1	97.38%	96.65%	95.50%	92.57%	96.02%	**97.51**%	96.94%	96.70%	96.90%	94.40%
9	**bench**	4279	4279	4279	4279	4279	4279	4279	4279	4279	4279
	Recall	**98.81**%	98.76%	98.43%	94.88%	98.20%	98.60%	98.39%	97.76%	98.67%	95.98%
	Precision	94.40%	92.88%	92.09%	89.82%	92.13%	**95.95**%	94.06%	94.04%	93.47%	92.58%
	F1	96.55%	95.73%	95.15%	92.28%	95.07%	**97.26**%	96.17%	95.86%	96.00%	94.25%
10	**bench**	3985	3985	3985	3985	3985	3985	3985	3985	3985	3985
	Recall	**99.57**%	99.10%	98.72%	95.93%	98.59%	99.40%	98.92%	99.32%	99.15%	98.52%
	Precision	88.59%	86.79%	86.01%	84.58%	86.14%	**90.08**%	88.07%	88.98%	87.47%	88.50%
	F1	93.76%	92.54%	91.93%	89.90%	91.95%	**94.51**%	93.18%	93.87%	92.94%	93.24%

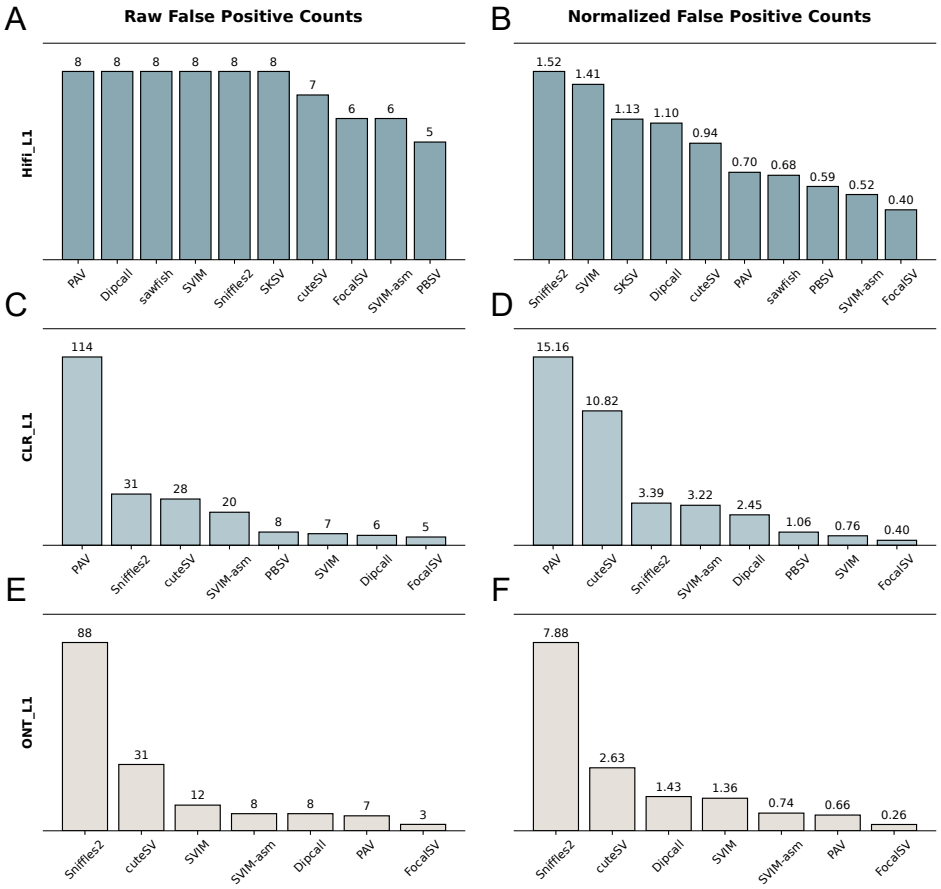
Supplemental Table S5: **Insertion calling performance evaluation on HG005 by tool support.** The table presents number of SVs in the bench set, Recall, Precision, F1 score (in %) for all benchmarked SV callers. The highest scores for Recall, Precision, F1 are highlighted in bold.

number of supporting tools	metric	FocalSV	PAV	SVIM-asm	Dipcall	sawfish	cuteSV	SVIM	PBSV	Sniffles2	SKSV
2	**bench**	6238	6238	6238	6238	6238	6238	6238	6238	6238	6238
	Recall	**93.76**%	93.27%	92.87%	74.91%	88.81%	89.39%	90.14%	74.00%	91.82%	86.26%
	Precision	96.61%	**97.11**%	96.36%	78.16%	91.63%	96.02%	95.79%	96.49%	95.87%	94.02%
	F1	**95.17**%	95.15%	94.58%	76.50%	90.20%	92.59%	92.88%	83.76%	93.80%	89.98%
3	**bench**	6141	6141	6141	6141	6141	6141	6141	6141	6141	6141
	Recall	**95.02**%	94.45%	94.02%	75.85%	89.82%	90.64%	91.26%	74.82%	93.05%	87.33%
	Precision	96.38%	**96.81**%	96.04%	77.91%	91.23%	95.85%	95.47%	96.05%	95.63%	93.71%
	F1	**95.69**%	95.61%	95.02%	76.86%	90.52%	93.17%	93.31%	84.12%	94.32%	90.41%
4	**bench**	6042	6042	6042	6042	6042	6042	6042	6042	6042	6042
	Recall	**95.90**%	95.56%	94.60%	76.17%	91.06%	91.94%	92.47%	75.90%	94.31%	88.60%
	Precision	95.71%	**96.38**%	95.08%	76.97%	91.00%	95.66%	95.18%	95.86%	95.36%	93.53%
	F1	95.80%	**95.97**%	94.84%	76.57%	91.03%	93.76%	93.80%	84.72%	94.83%	91.00%
5	**bench**	5966	5966	5966	5966	5966	5966	5966	5966	5966	5966
	Recall	**96.60**%	96.40%	95.14%	76.55%	91.84%	92.86%	93.19%	76.52%	95.17%	89.46%
	Precision	95.19%	**95.99**%	94.41%	76.38%	90.62%	95.40%	94.72%	95.42%	95.03%	93.26%
	F1	95.89%	**96.19**%	94.77%	76.47%	91.23%	94.11%	93.95%	84.93%	95.10%	91.32%
6	**bench**	5918	5918	5918	5918	5918	5918	5918	5918	5918	5918
	Recall	**96.87**%	96.72%	95.40%	76.73%	92.13%	93.39%	93.71%	76.93%	95.71%	89.95%
	Precision	94.70%	**95.54**%	93.91%	75.95%	90.18%	95.18%	94.48%	95.17%	94.79%	93.01%
	F1	95.77%	**96.13**%	94.65%	76.34%	91.14%	94.28%	94.10%	85.09%	95.25%	91.45%
7	**bench**	5851	5851	5851	5851	5851	5851	5851	5851	5851	5851
	Recall	**97.28**%	97.16%	95.80%	76.93%	92.33%	94.10%	94.21%	77.44%	96.02%	90.69%
	Precision	94.02%	**94.89**%	93.23%	75.28%	89.35%	94.82%	93.90%	94.71%	94.03%	92.71%
	F1	95.62%	**96.01**%	94.50%	76.09%	90.81%	94.46%	94.05%	85.21%	95.01%	91.69%
8	**bench**	5720	5720	5720	5720	5720	5720	5720	5720	5720	5720
	Recall	**97.57**%	97.55%	96.24%	77.20%	92.43%	95.47%	95.21%	77.94%	96.35%	92.27%
	Precision	92.19%	93.14%	91.57%	73.86%	87.45%	**94.04**%	92.78%	93.19%	92.23%	92.22%
	F1	94.80%	**95.30**%	93.85%	75.49%	89.87%	94.75%	93.98%	84.88%	94.25%	92.25%
9	**bench**	5469	5469	5469	5469	5469	5469	5469	5469	5469	5469
	Recall	97.97%	**98.08**%	97.31%	77.75%	92.65%	96.58%	95.81%	79.50%	96.60%	94.46%
	Precision	88.50%	89.53%	88.52%	71.12%	83.81%	**90.96**%	89.27%	90.89%	88.42%	90.27%
	F1	93.00%	93.61%	92.71%	74.28%	88.01%	**93.69**%	92.42%	84.81%	92.33%	92.32%
10	**bench**	4231	4231	4231	4231	4231	4231	4231	4231	4231	4231
	Recall	**98.35**%	98.23%	97.49%	76.01%	92.46%	96.88%	97.35%	97.99%	97.09%	96.45%
	Precision	68.73%	69.37%	68.61%	53.79%	64.70%	70.59%	70.17%	**86.66**%	68.75%	71.31%
	F1	80.91%	81.31%	80.54%	63.00%	76.13%	81.67%	81.56%	**91.98**%	80.50%	82.00%

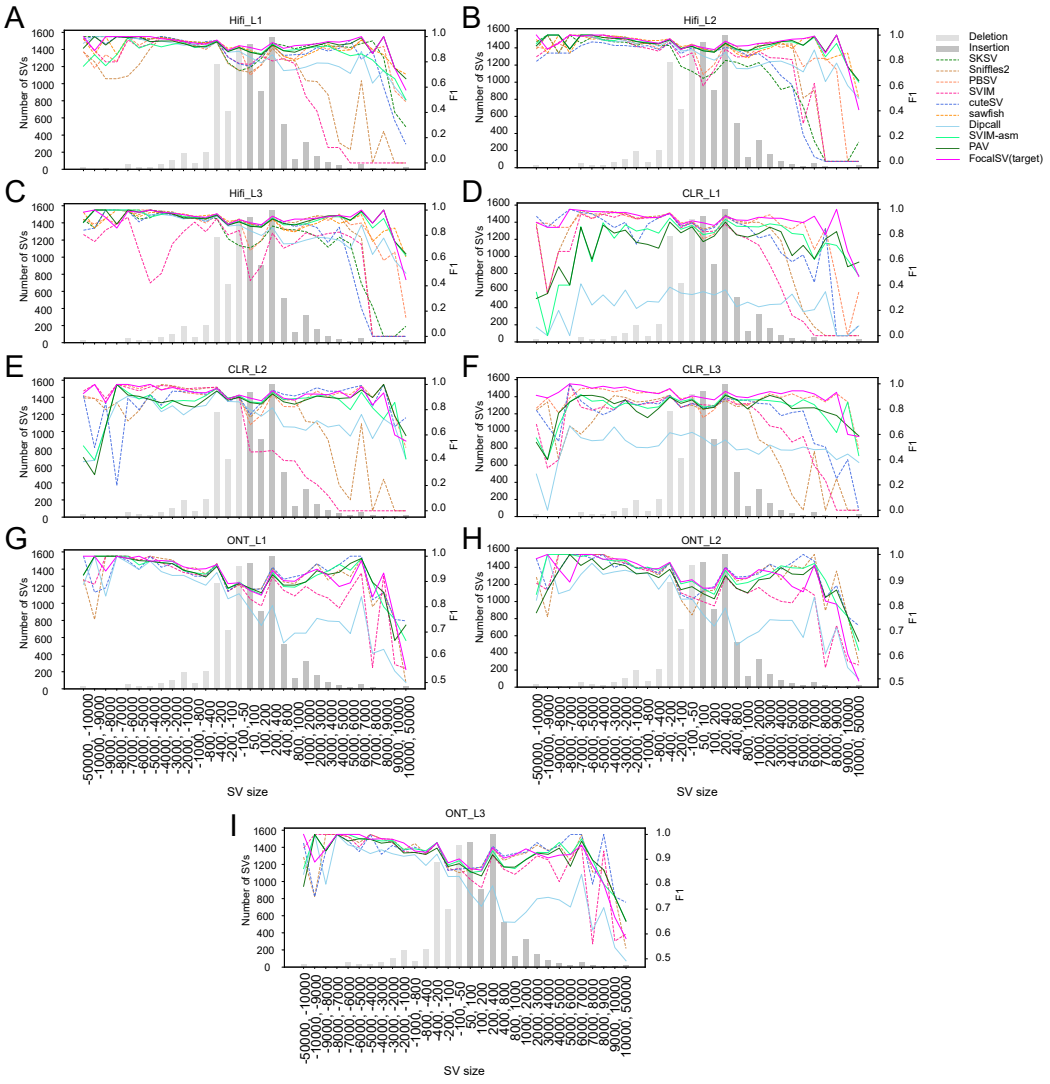
Supplemental Table S6: **Large deletions (DELs) and insertions (INSs) (≥50bp) calling performance in high and low phasing groups across different datasets.** The table presents True Positive (TP), False Positive (FP), False Negative (FN), Recall, Precision, F1 score (in %), Genotype TP, Genotype FP, and Genotype accuracy (measured by Genotype Concordance) for all benchmarked SV callers. Benchmarking was conducted using Truvari with the following parameter settings: p = 0.5, P = 0.5, r = 500, and O = 0.01.

Library	Metric	High phasing group (DEL)	Low phasing group (DEL)	High phasing group (INS)	Low phasing group (INS)
Hifi_L1	benchmark	2074	2010	2570	2664
	comp	2106	2032	2740	2818
	TP	1988	1887	2421	2501
	FP	118	145	319	317
	FN	86	123	149	163
	Precision	94.40%	92.86%	88.36%	88.75%
	Recall	95.85%	93.88%	94.20%	93.88%
	F1	95.12%	93.37%	91.19%	91.24%
	Genotype TP	1967	1872	2380	2463
	Genotype FP	21	15	41	38
	GT concord	98.94%	99.21%	98.31%	98.48%
CLR_L1	benchmark	2025	2057	2544	2689
	comp	1990	2018	2606	2780
	TP	1872	1870	2320	2436
	FP	118	148	286	344
	FN	153	187	224	253
	Precision	94.07%	92.67%	89.03%	87.63%
	Recall	92.44%	90.91%	91.19%	90.59%
	F1	93.25%	91.78%	90.10%	89.08%
	Genotype TP	1849	1845	2272	2372
	Genotype FP	23	25	48	64
	GT concord	98.77%	98.66%	97.93%	97.37%
ONT_L1	benchmark	1928	2168	2423	2834
	comp	1940	2211	2458	2908
	TP	1827	2003	2252	2589
	FP	113	208	206	319
	FN	101	165	171	245
	Precision	94.18%	90.59%	91.62%	89.03%
	Recall	94.76%	92.39%	92.94%	91.35%
	F1	94.47%	91.48%	92.28%	90.18%
	Genotype TP	1811	1971	2232	2554
	Genotype FP	16	32	20	35
	GT concord	99.12%	98.40%	99.11%	98.65%

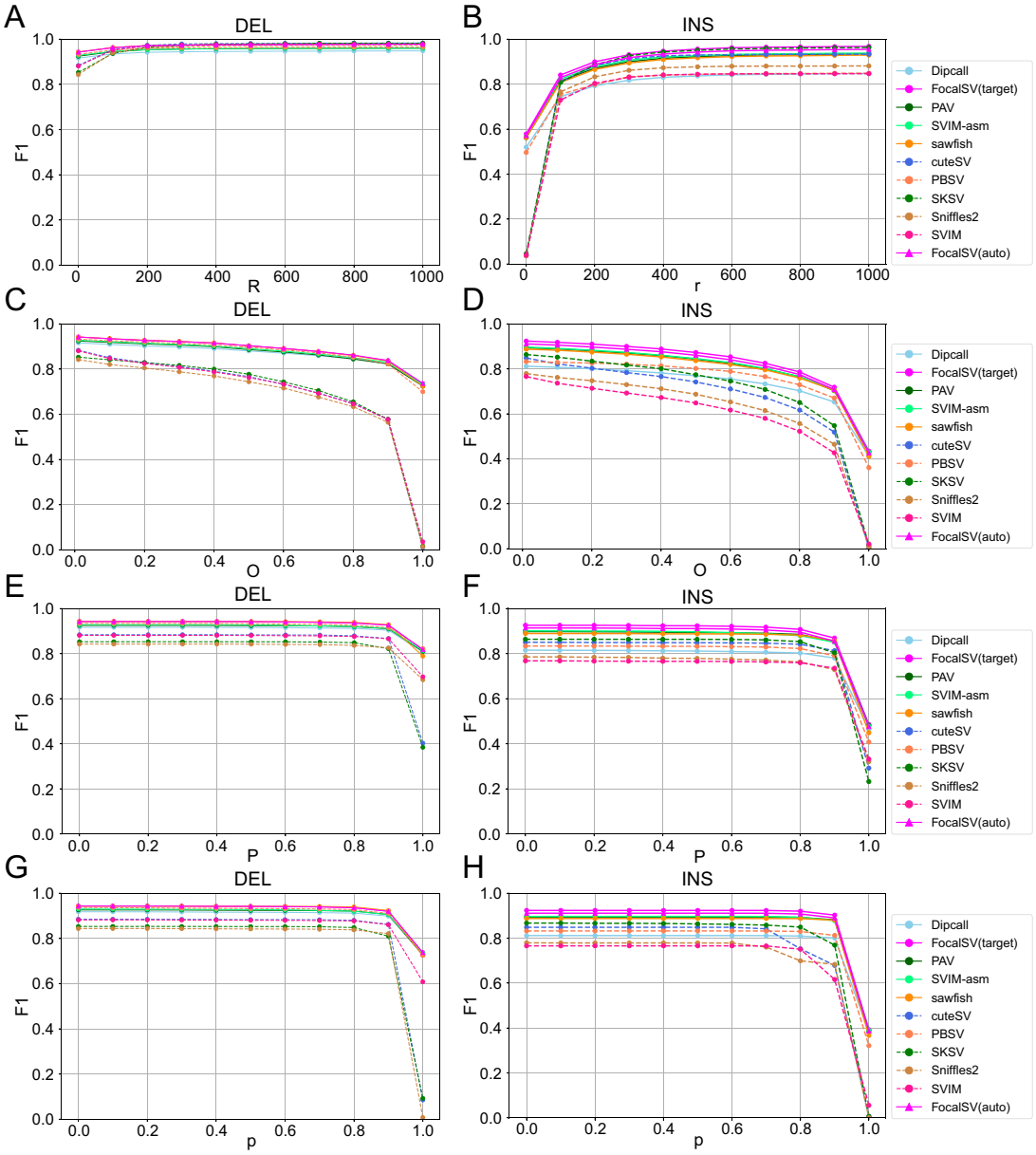
4 Supplemental Figures



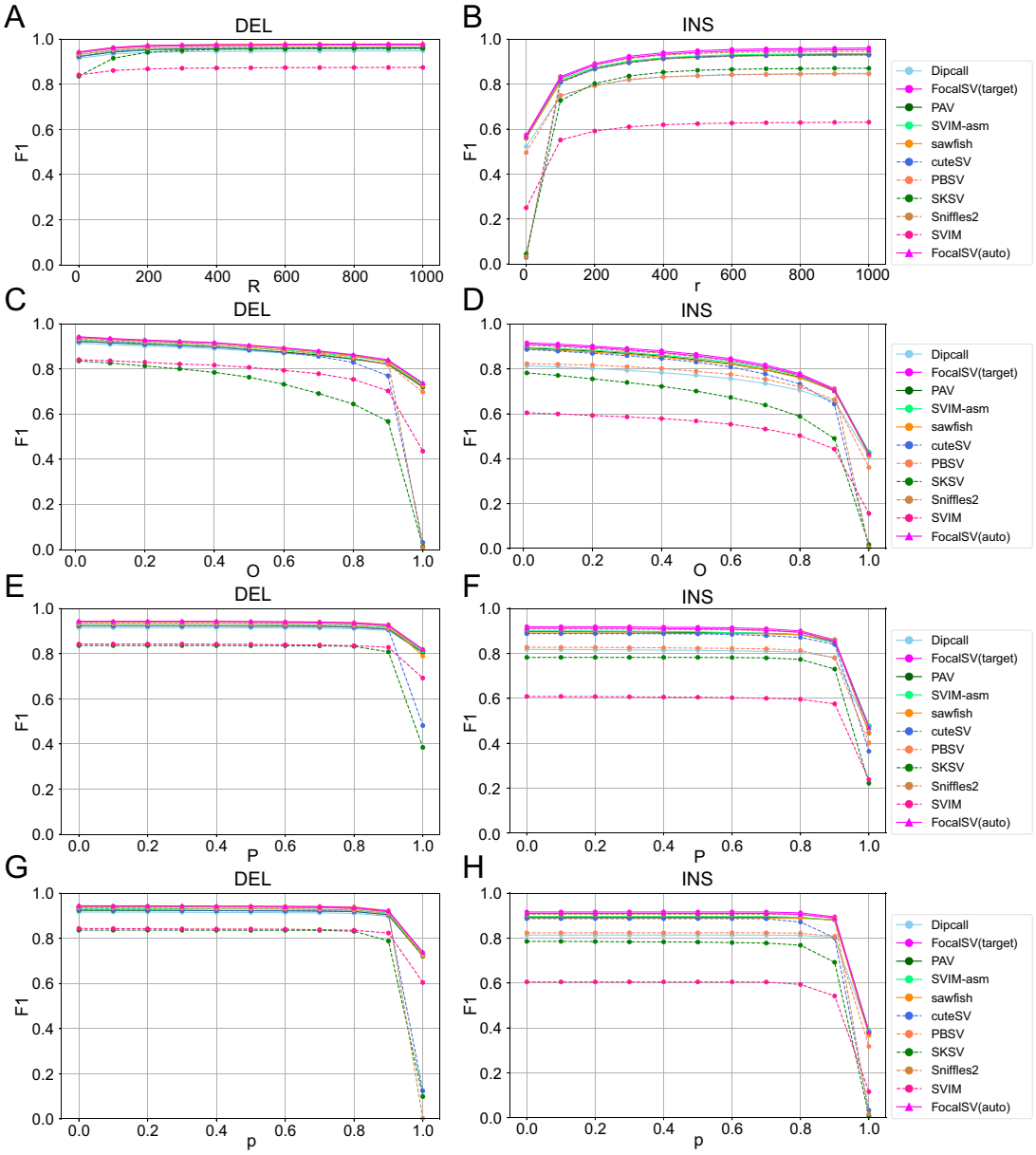
Supplemental Figure S1: **Comparison of false positive SV calls across tools and datasets in SV-negative regions.** **(A-B)** Raw and normalized false positive SV call counts on Hifi_L1 across tools. **(C-D)** Raw and normalized false positive SV call counts on CLR_L1 across tools. **(E-F)** Raw and normalized false positive SV call counts on ONT_L1 across tools.



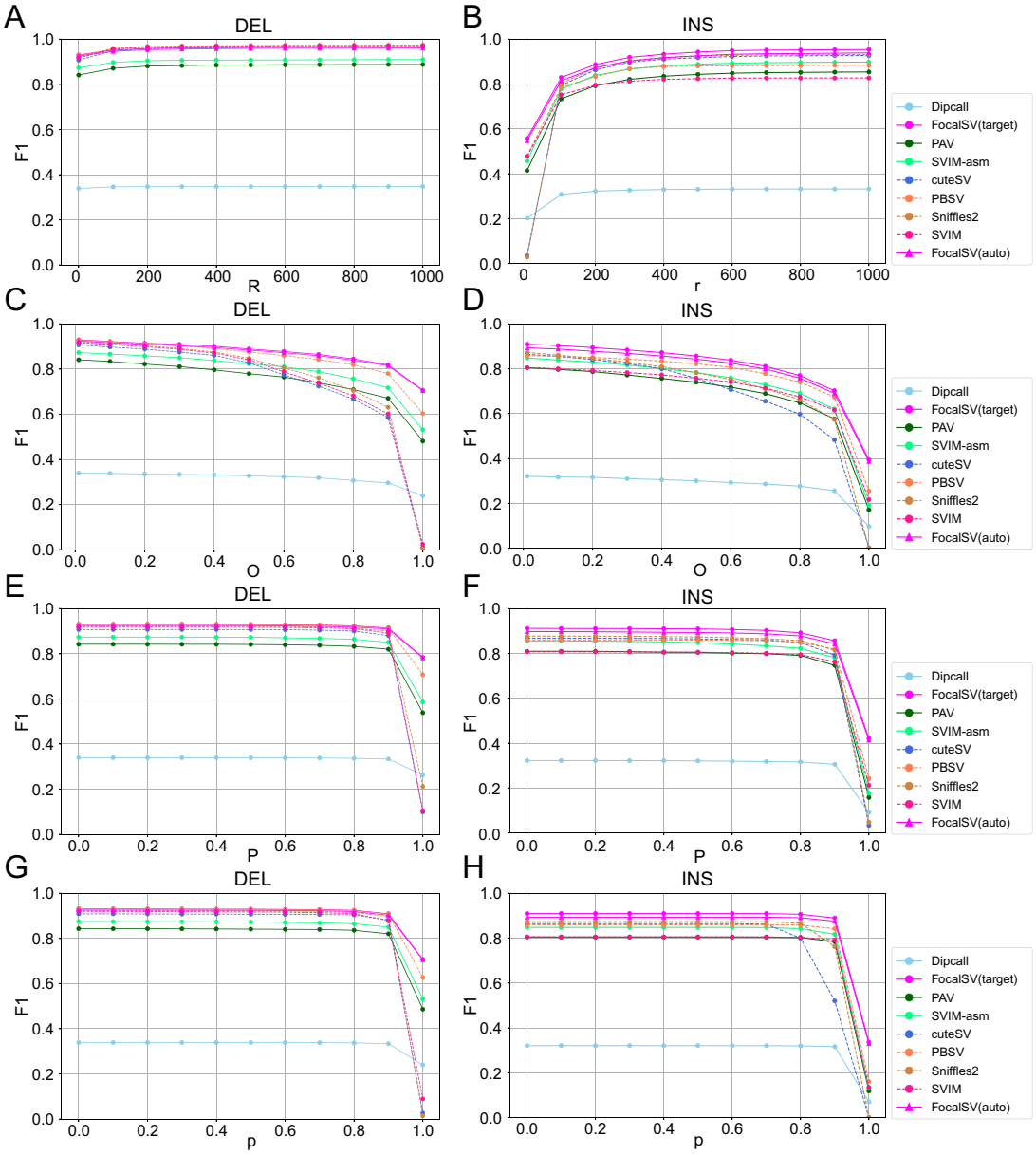
Supplemental Figure S2: **F1 accuracy of SV detection by FocalSV(target) and all other tools across various size ranges on nine long-read datasets. (A-C)** F1 accuracy plot for three Hifi datasets. Negative ranges denote deletions and the positive ranges denote insertions. The bar plot illustrates the 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.



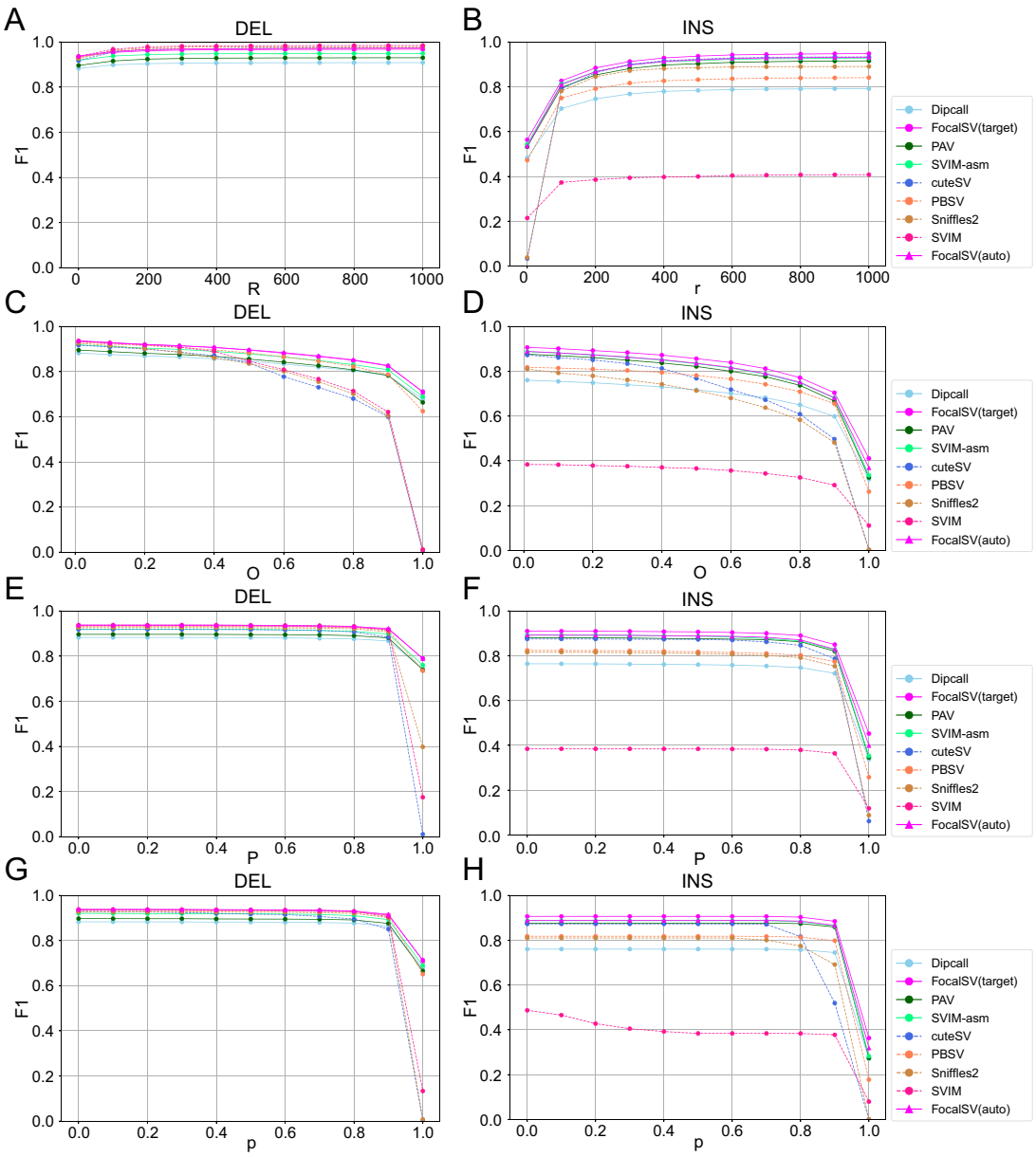
Supplemental Figure S3: **F1 accuracy by changing four different evaluation parameters on Hifi_L1. (A-B)** F1 score curves for deletions (DEL) and insertions (INS) across all tools as r is varied. r is the maximum reference location distance between SV call and gold standard SV. r varies from 0–1000 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.



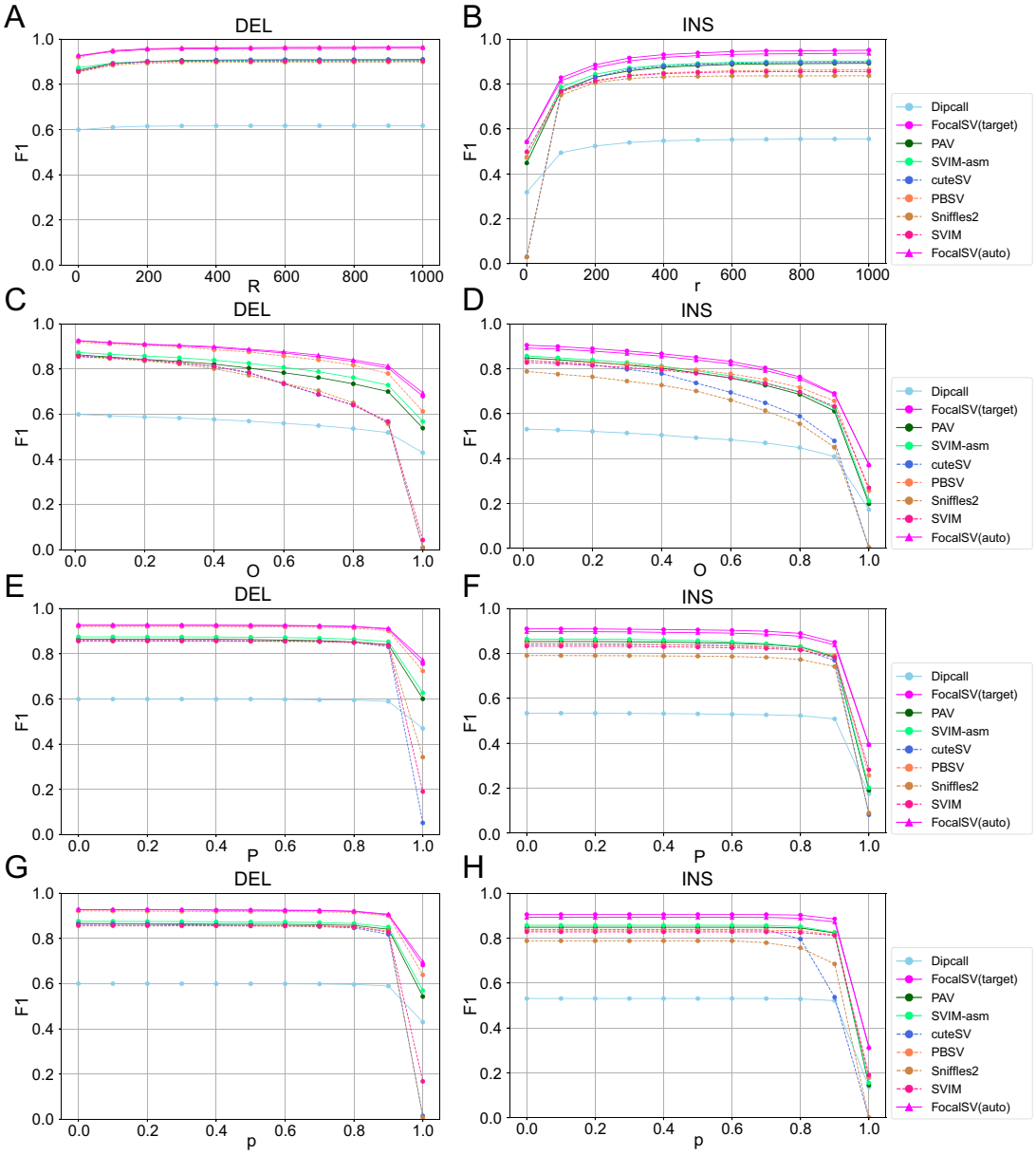
Supplemental Figure S4: **F1 accuracy by changing four different evaluation parameters on Hifi_L3. (A-B)** F1 score curves for deletions (DEL) and insertions (INS) across all tools as r is varied. r is the maximum reference location distance between SV call and gold standard SV. r varies from 0–1000 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.



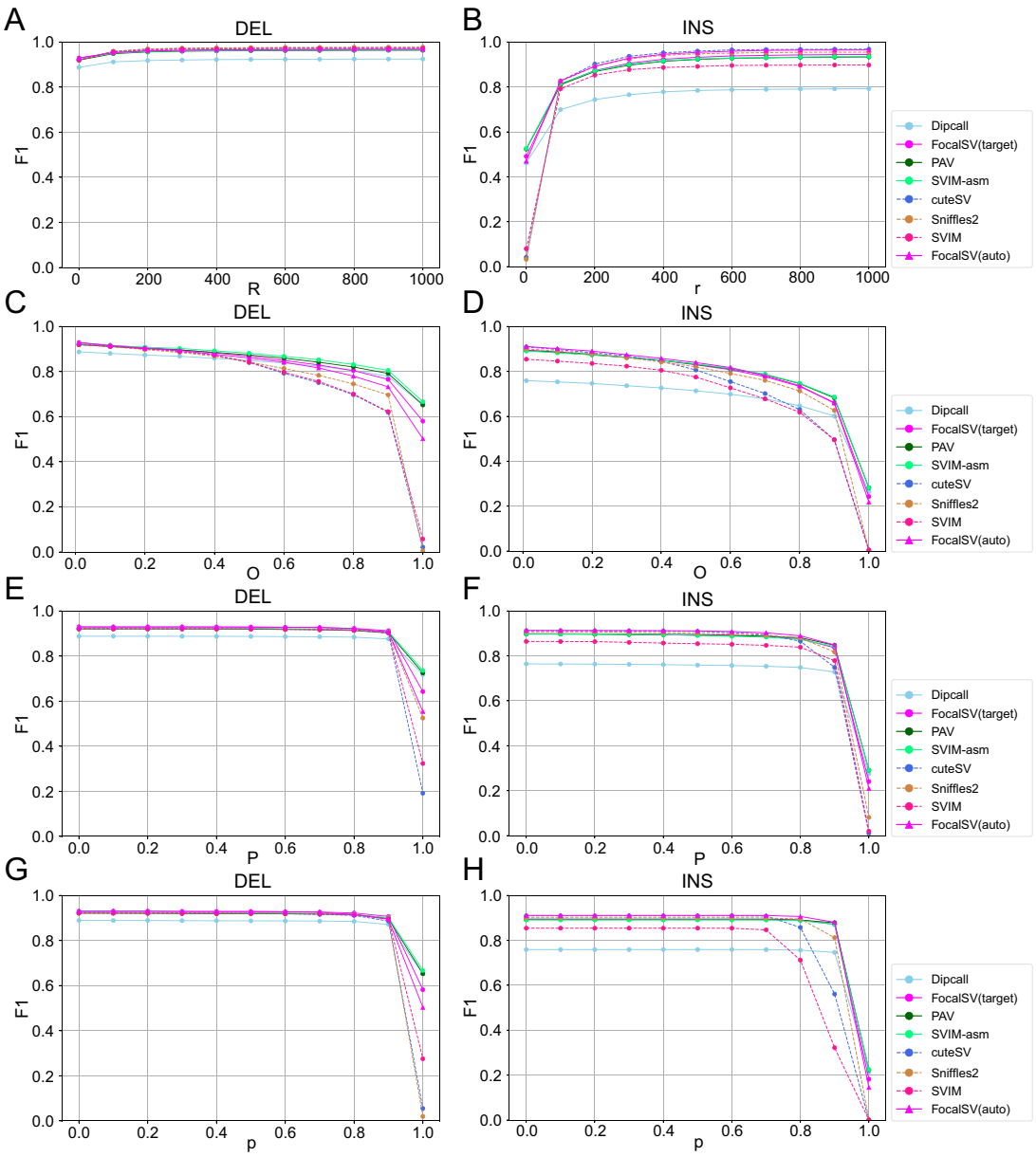
Supplemental Figure S5: **F1 accuracy by changing four different evaluation parameters on CLR_L1. (A-B)** F1 score curves for deletions (DEL) and insertions (INS) across all tools as r is varied. r is the maximum reference location distance between SV call and gold standard SV. r varies from 0–1000 bp 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.



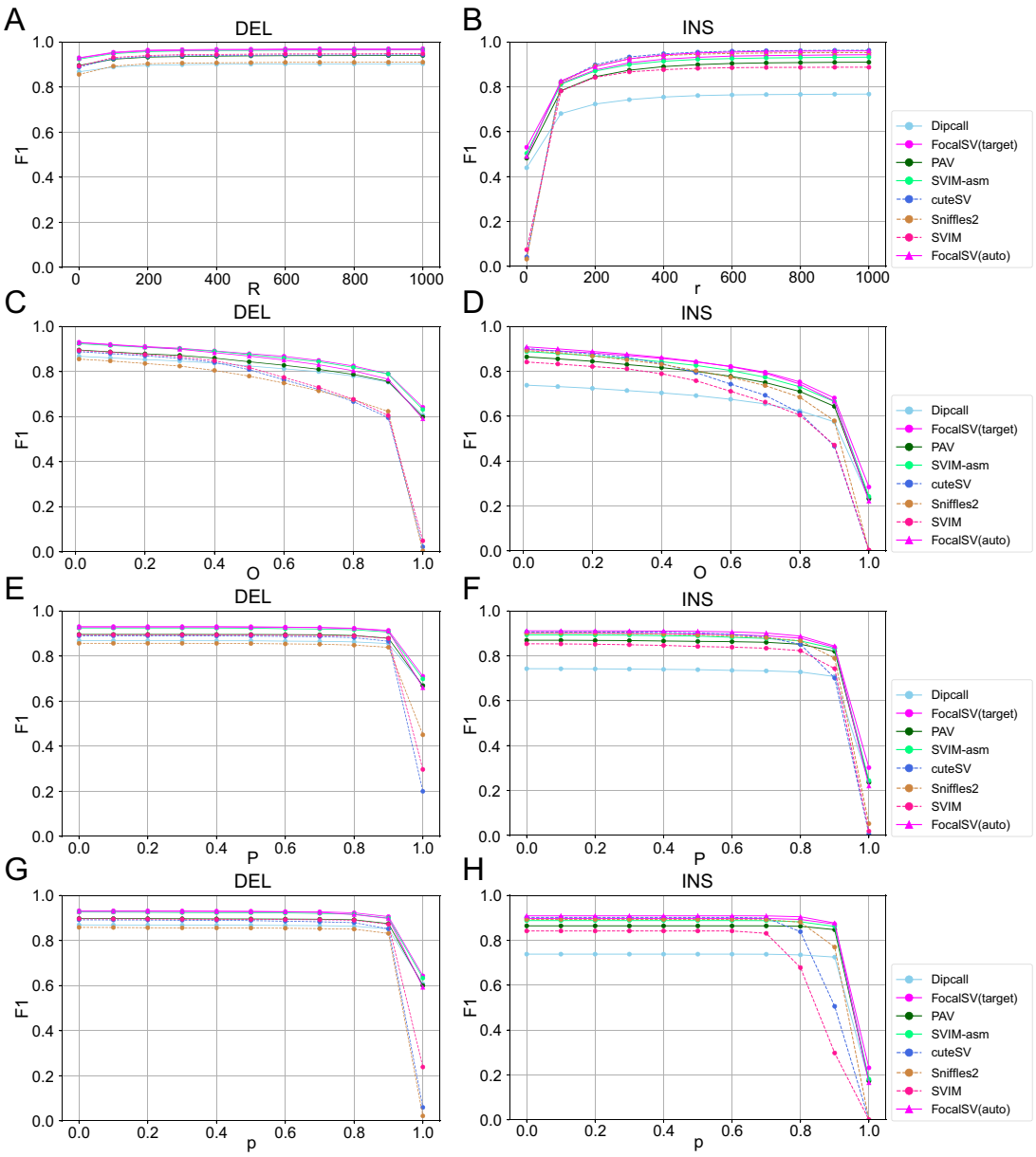
Supplemental Figure S6: **F1 accuracy by changing four different evaluation parameters on CLR_L2. (A-B)** F1 score curves for deletions (DEL) and insertions (INS) across all tools as r is varied. r is the maximum reference location distance between SV call and gold standard SV. r varies from 0–1000 bp 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.



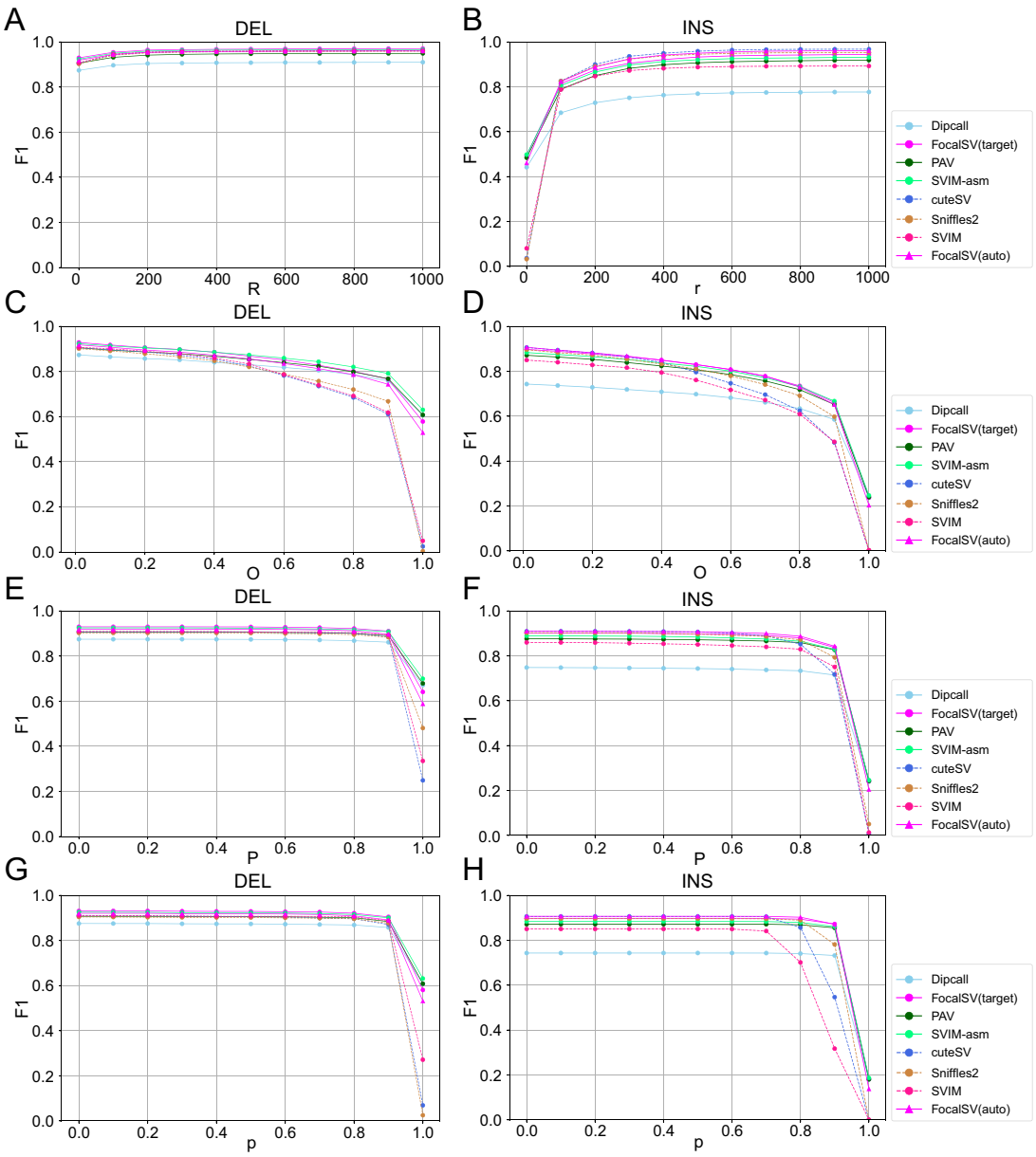
Supplemental Figure S7: **F1 accuracy by changing four different evaluation parameters on CLR_L3. (A-B)** F1 score curves for deletions (DEL) and insertions (INS) across all tools as r is varied. r is the maximum reference location distance between SV call and gold standard SV. r varies from 0–1000 bp 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.



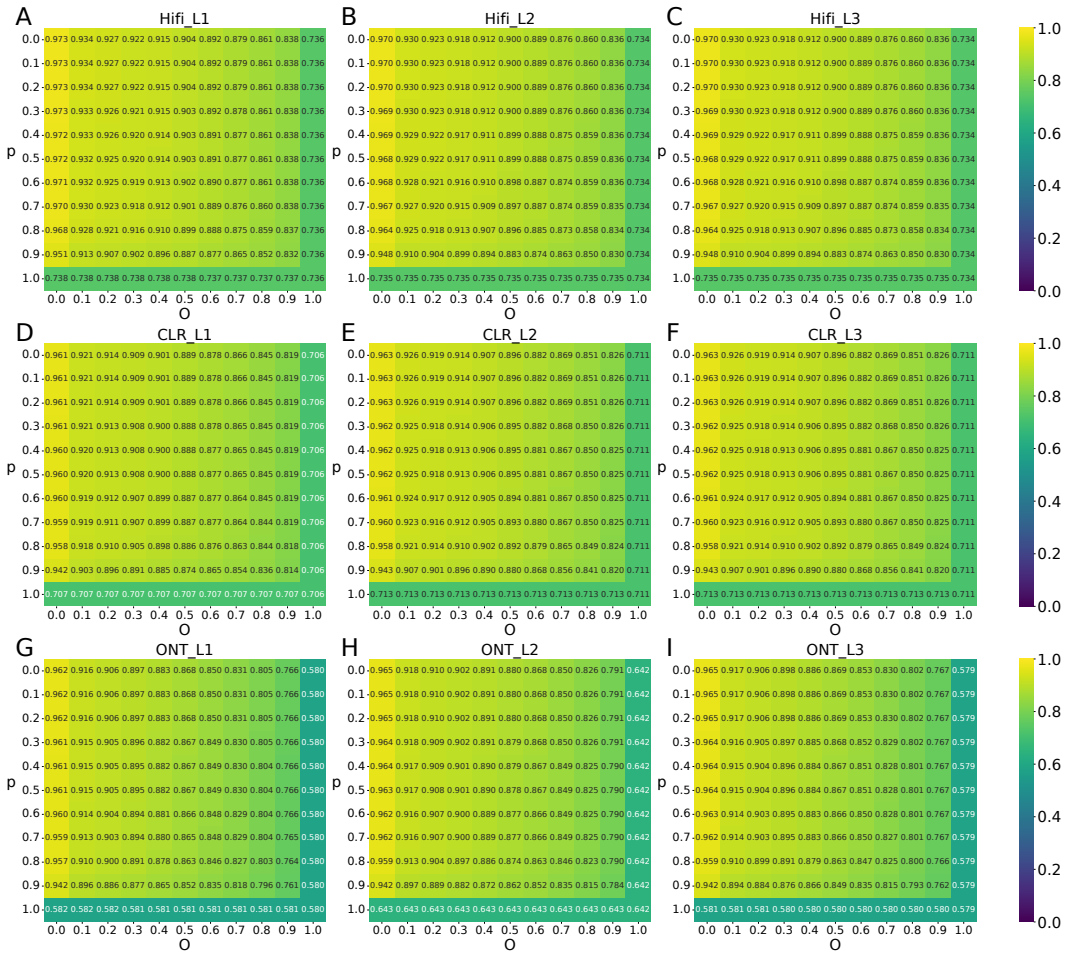
Supplemental Figure S8: **F1 accuracy by changing four different evaluation parameters on ONT_L1. (A-B)** F1 score curves for deletions (DEL) and insertions (INS) across all tools as r is varied. r is the maximum reference location distance between SV call and gold standard SV. r varies from 0–1000 bp 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.



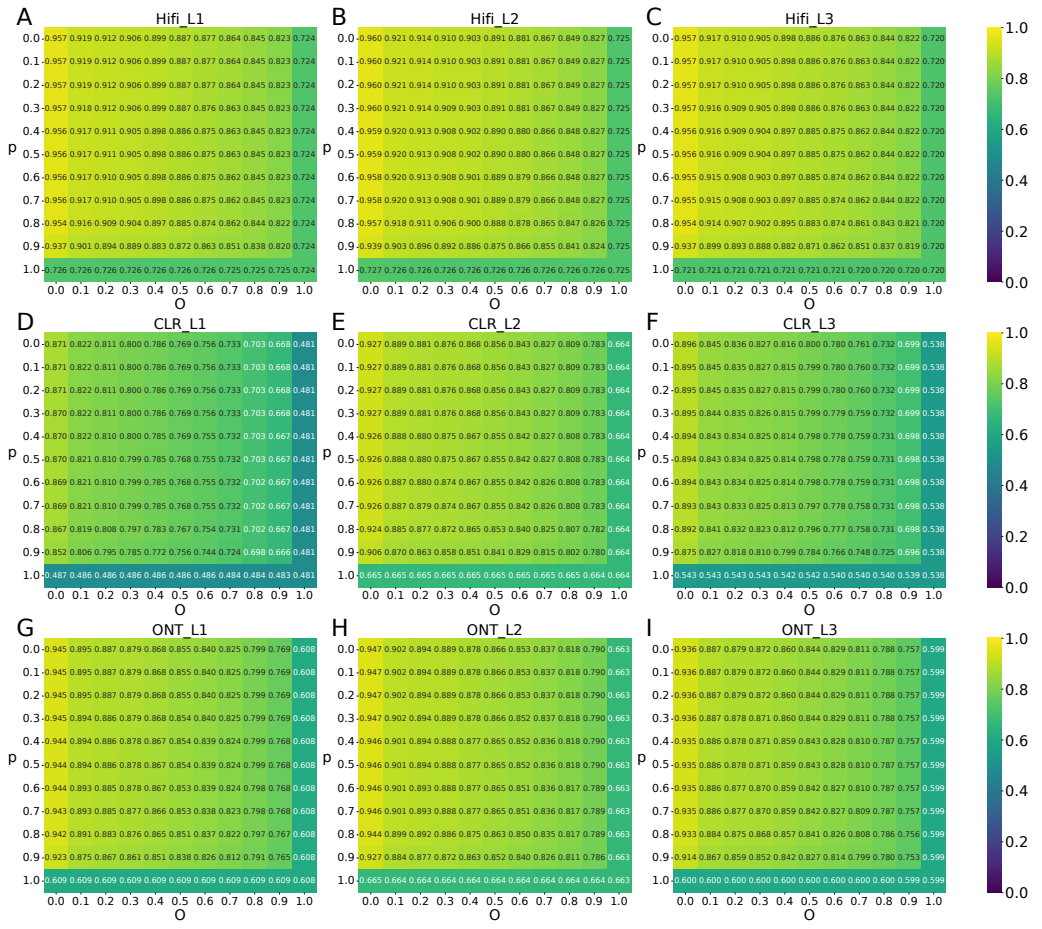
Supplemental Figure S9: **F1 accuracy by changing four different evaluation parameters on ONT_L2. (A-B)** F1 score curves for deletions (DEL) and insertions (INS) across all tools as r is varied. r is the maximum reference location distance between SV call and gold standard SV. r varies from 0–1000 bp 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.



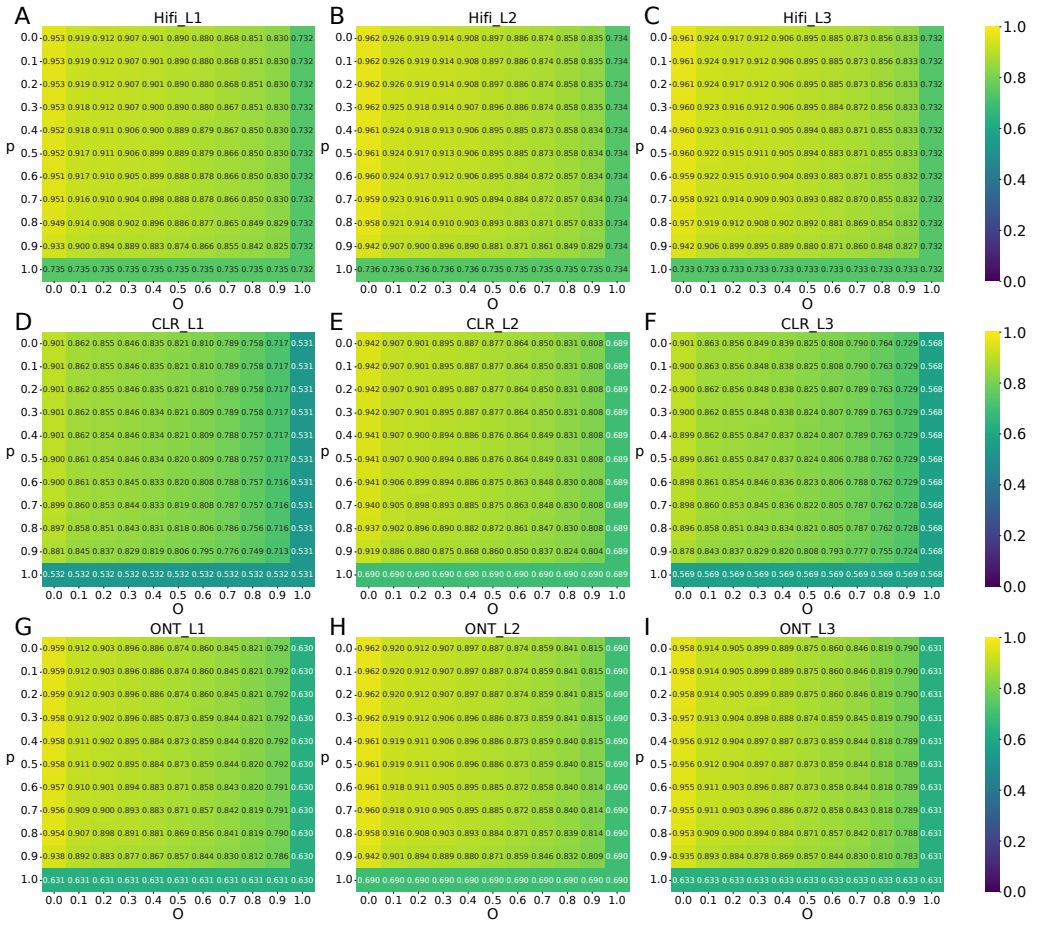
Supplemental Figure S10: **F1 accuracy by changing four different evaluation parameters on ONT_L3. (A-B)** F1 score curves for deletions (DEL) and insertions (INS) across all tools as r is varied. r is the maximum reference location distance between SV call and gold standard SV. r varies from 0–1000 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.



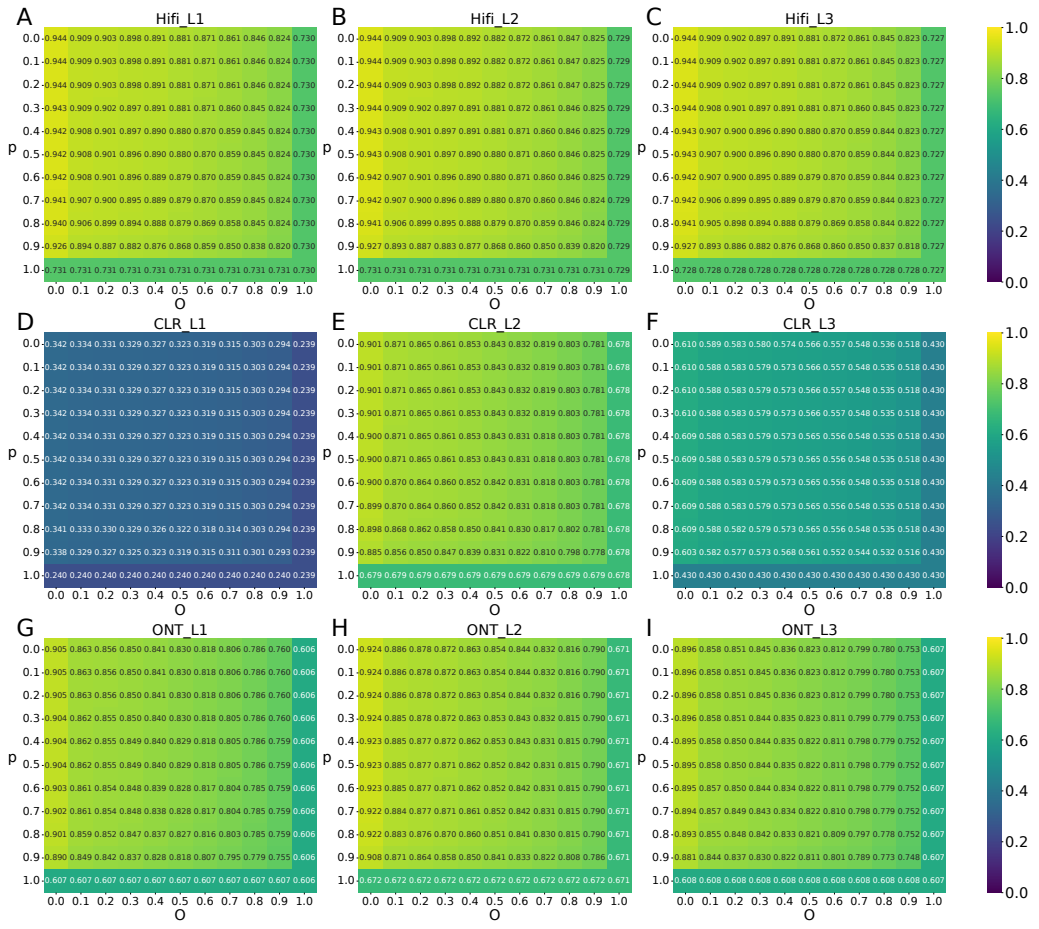
Supplemental Figure S11: **Deletion F1 accuracy of FocalSV(target) by tuning different evaluation parameters (p and O).** O is the minimum reciprocal overlap between SV call and gold standard SV. p is the minimum percentage of allele sequence similarity between SV call and gold standard SV. O and p vary from 0-1 with a 0.1 interval. **(A-C)** The F1 heatmap for deletions by FocalSV(target) 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(target) on three CLR datasets. **(G-I)** The F1 heatmap for deletions by FocalSV(target) on three ONT datasets.



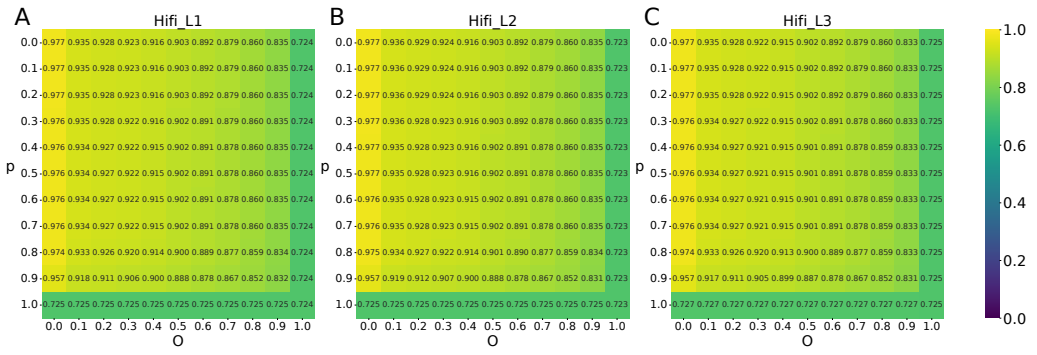
Supplemental Figure S12: **Deletion F1 accuracy of PAV by tuning different evaluation parameters (p and O).** O is the minimum reciprocal overlap between SV call and gold standard SV. p is the minimum percentage of allele sequence similarity between SV call and gold standard SV. O and p vary from 0-1 with a 0.1 interval. **(A-C)** The F1 heatmap for deletions by PAV 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 PAV on three CLR datasets. **(G-I)** The F1 heatmap for deletions by PAV on three ONT datasets.



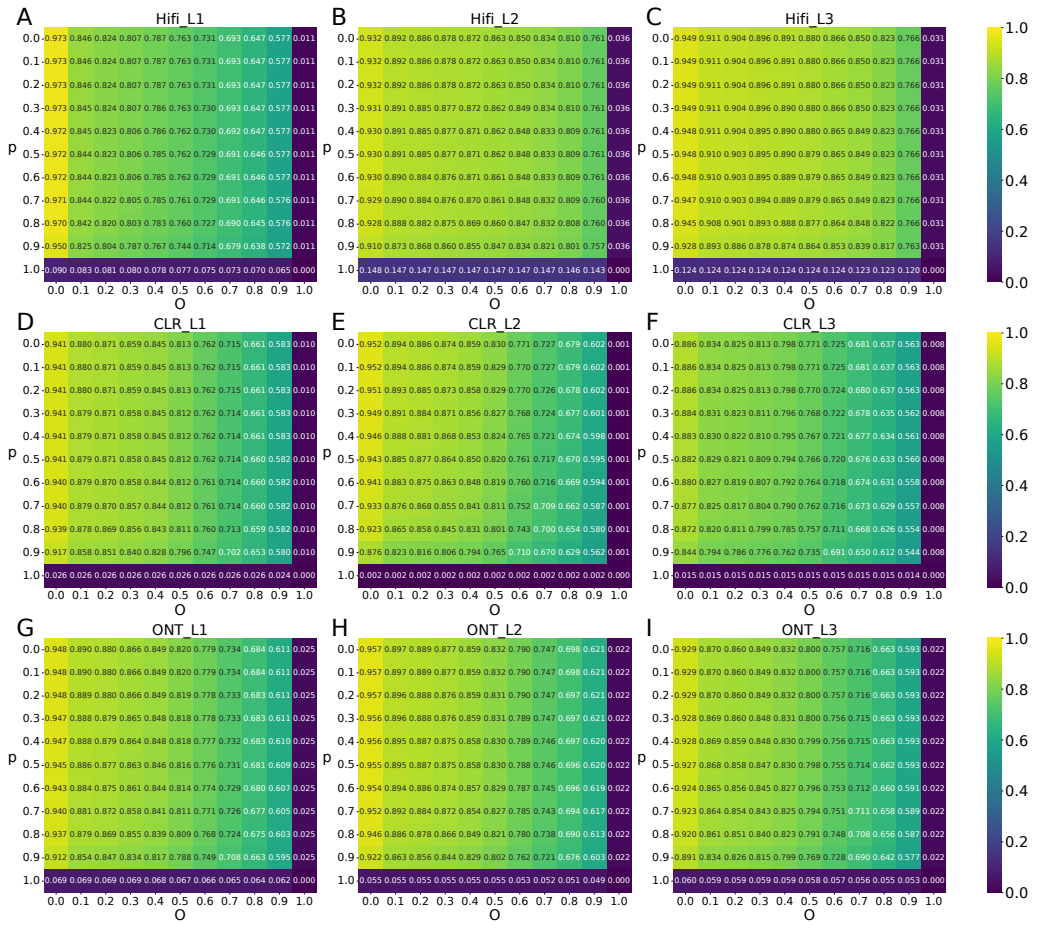
Supplemental Figure S13: **Deletion F1 accuracy of SVIM-asm by tuning different evaluation parameters (p and O).** O is the minimum reciprocal overlap between SV call and gold standard SV. p is the minimum percentage of allele sequence similarity between SV call and gold standard SV. O and p vary from 0-1 with a 0.1 interval. **(A-C)** The F1 heatmap for deletions by SVIM-asm 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 SVIM-asm on three CLR datasets. **(G-I)** The F1 heatmap for deletions by SVIM-asm on three ONT datasets.



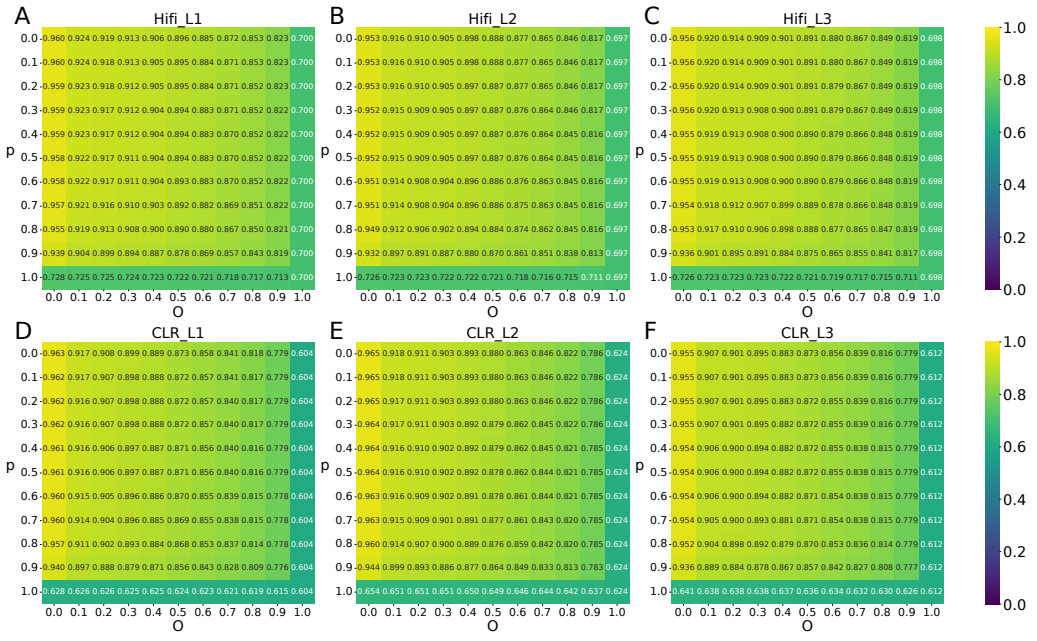
Supplemental Figure S14: **Deletion F1 accuracy of Dipcall by tuning different evaluation parameters (p and O).** O is the minimum reciprocal overlap between SV call and gold standard SV. p is the minimum percentage of allele sequence similarity between SV call and gold standard SV. O and p vary from 0-1 with a 0.1 interval. **(A-C)** The F1 heatmap for deletions by Dipcall 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 Dipcall on three CLR datasets. **(G-I)** The F1 heatmap for deletions by Dipcall on three ONT datasets.



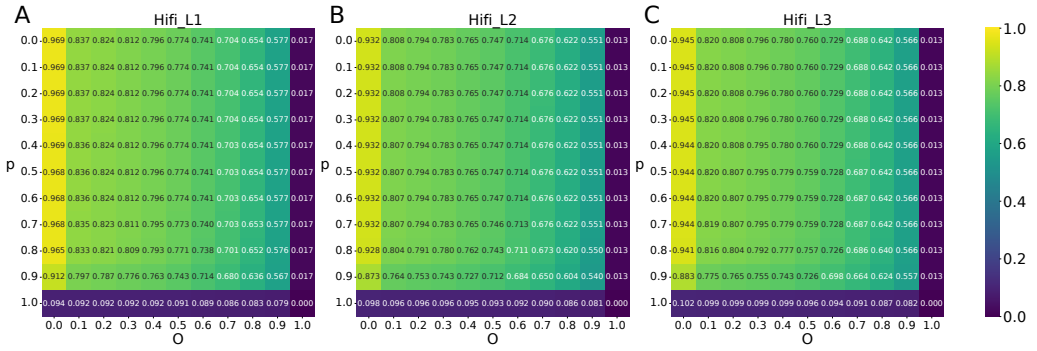
Supplemental Figure S15: **Deletion F1 accuracy of sawfish by tuning different evaluation parameters (p and O).** O is the minimum reciprocal overlap between SV call and gold standard SV. p is the minimum percentage of allele sequence similarity between SV call and gold standard SV. O and p vary from 0-1 with a 0.1 interval. **(A-C)** The F1 heatmap for deletions by sawfish on three Hifi datasets. Every cell in the heatmap represents the F1 score under a specific pair of p and O evaluation.



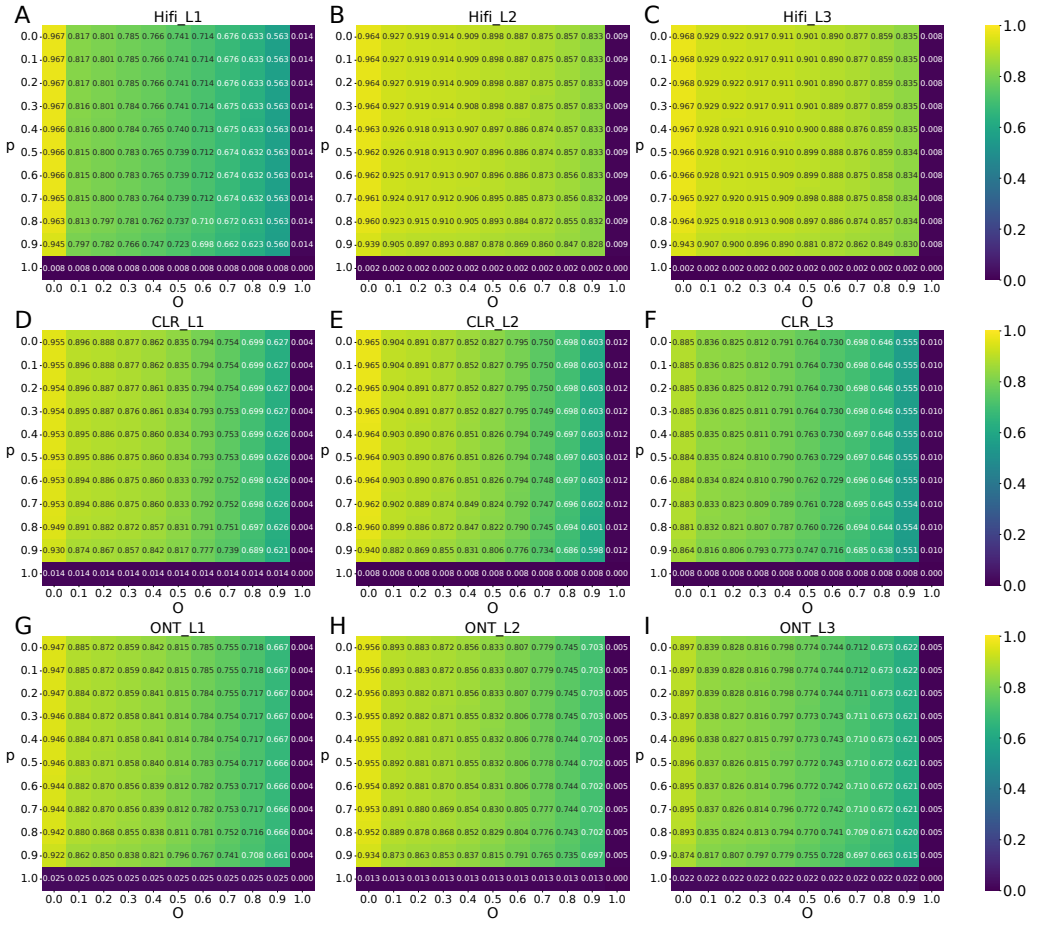
Supplemental Figure S16: **Deletion F1 accuracy of cuteSV by tuning different evaluation parameters (p and O).** O is the minimum reciprocal overlap between SV call and gold standard SV. p is the minimum percentage of allele sequence similarity between SV call and gold standard SV. O and p vary from 0-1 with a 0.1 interval. **(A-C)** The F1 heatmap for deletions by cuteSV 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 cuteSV on three CLR datasets. **(G-I)** The F1 heatmap for deletions by cuteSV on three ONT datasets.



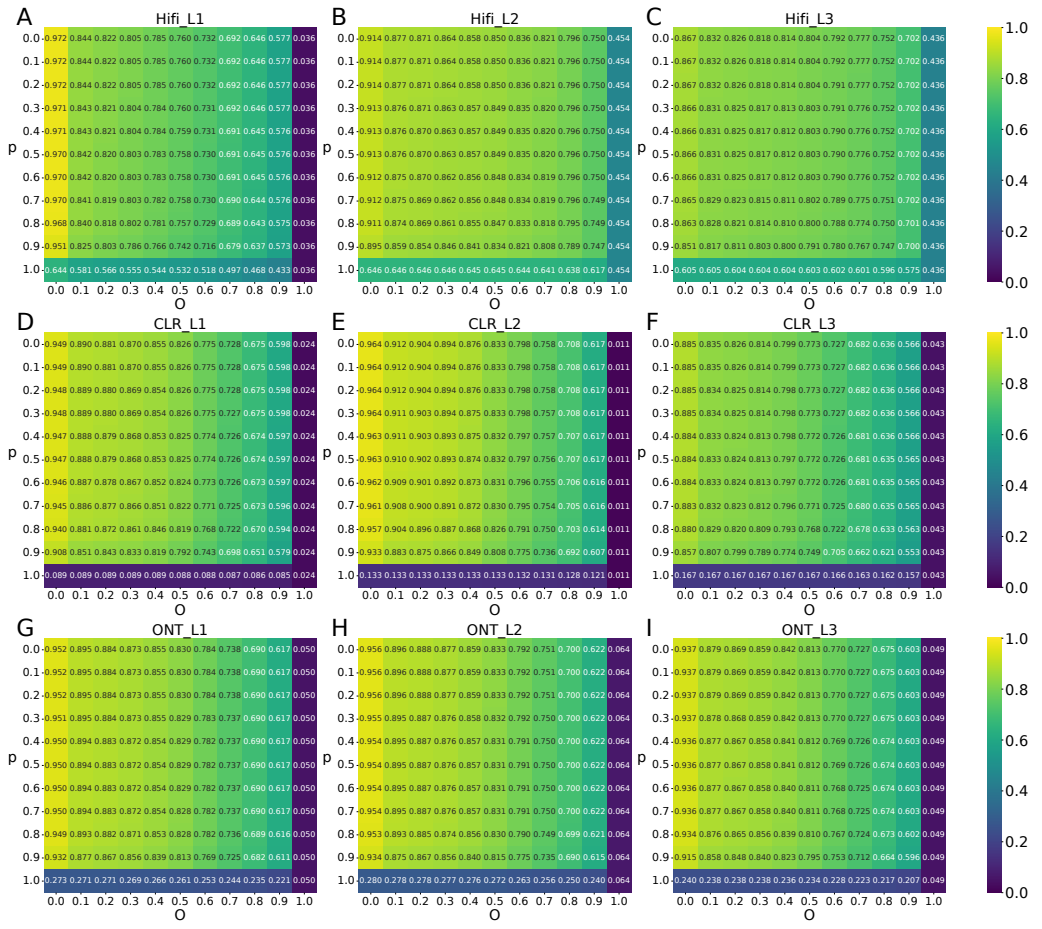
Supplemental Figure S17: **Deletion F1 accuracy of PBSV by tuning different evaluation parameters (p and O).** O is the minimum reciprocal overlap between SV call and gold standard SV. p is the minimum percentage of allele sequence similarity between SV call and gold standard SV. O and p vary from 0-1 with a 0.1 interval. **(A-C)** The F1 heatmap for deletions by PBSV 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 PBSV on three CLR datasets.



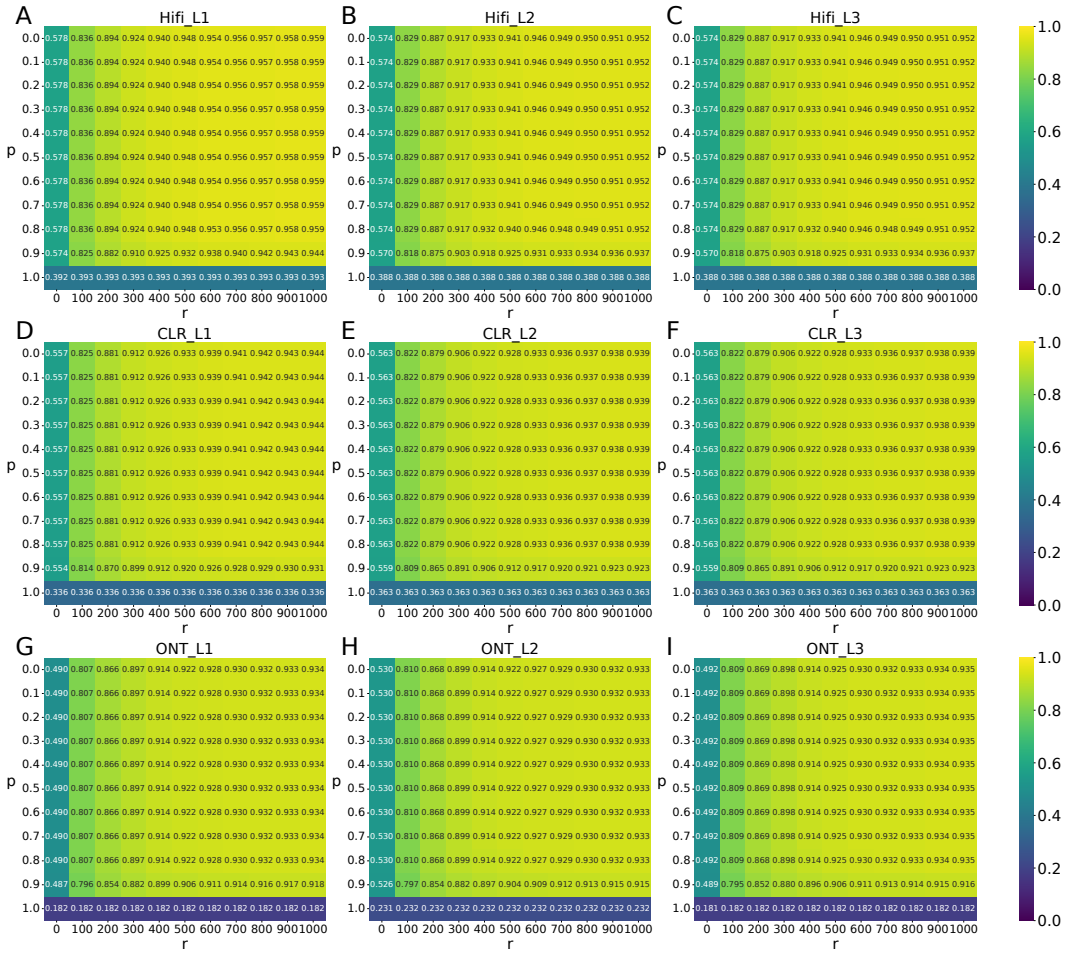
Supplemental Figure S18: **Deletion F1 accuracy of SKSV by tuning different evaluation parameters (p and O).** O is the minimum reciprocal overlap between SV call and gold standard SV. p is the minimum percentage of allele sequence similarity between SV call and gold standard SV. O and p vary from 0-1 with a 0.1 interval. **(A-C)** The F1 heatmap for deletions by SKSV on three Hifi datasets. Every cell in the heatmap represents the F1 score under a specific pair of p and O evaluation.



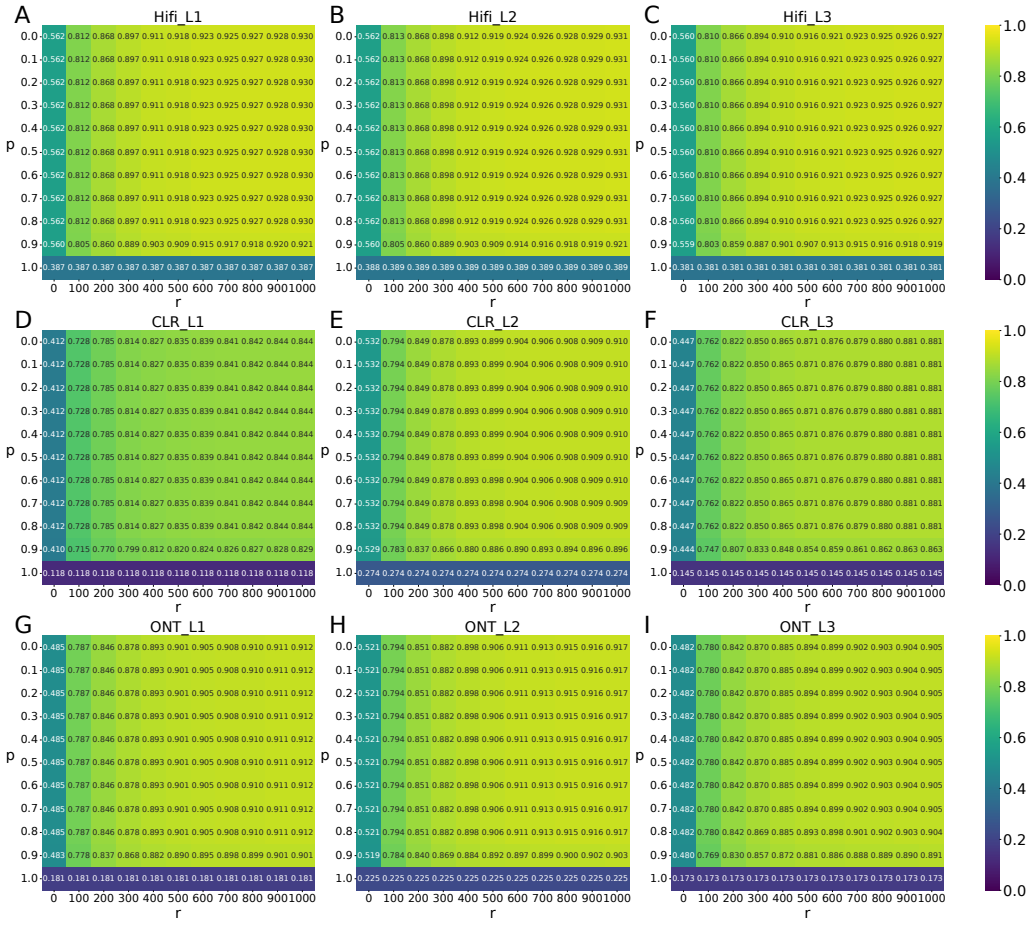
Supplemental Figure S19: **Deletion F1 accuracy of Sniffles2 by tuning different evaluation parameters (p and O).** O is the minimum reciprocal overlap between SV call and gold standard SV. p is the minimum percentage of allele sequence similarity between SV call and gold standard SV. O and p vary from 0-1 with a 0.1 interval. **(A-C)** The F1 heatmap for deletions by Sniffles2 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 Sniffles2 on three CLR datasets. **(G-I)** The F1 heatmap for deletions by Sniffles2 on three ONT datasets.



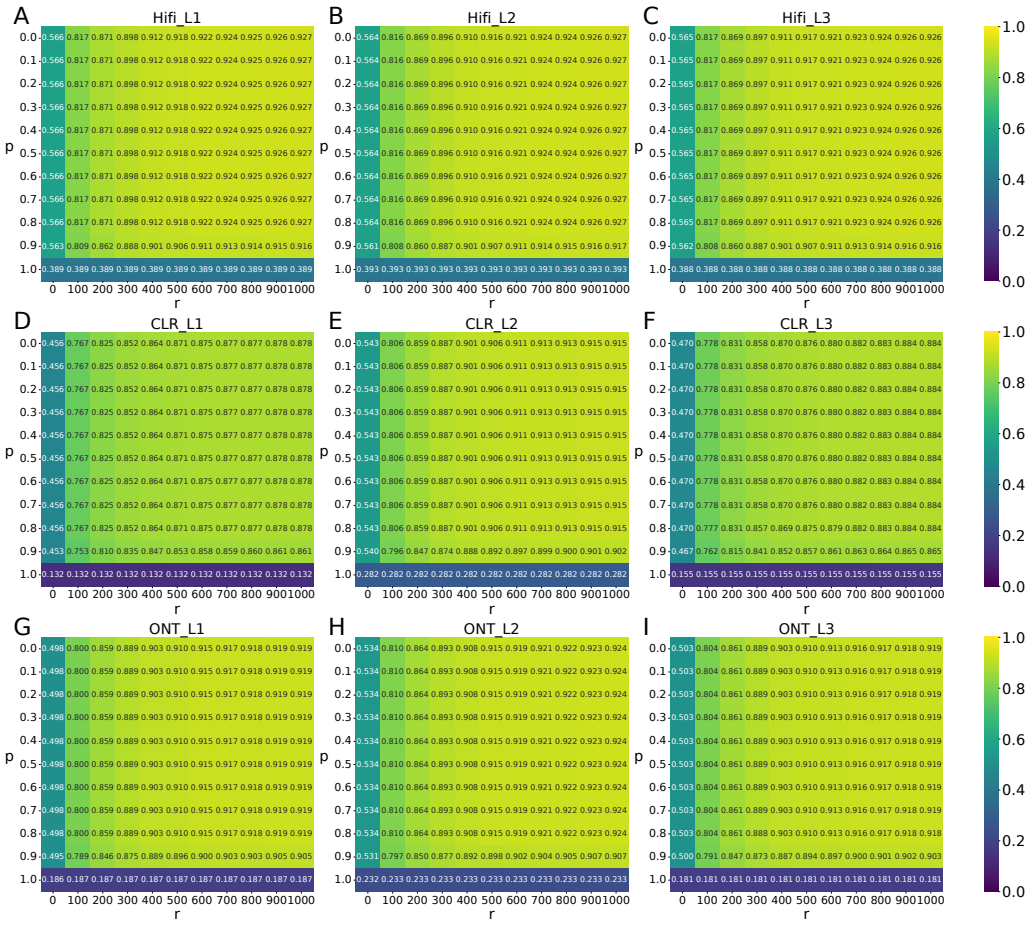
Supplemental Figure S20: **Deletion F1 accuracy of SVIM by tuning different evaluation parameters (p and O).** O is the minimum reciprocal overlap between SV call and gold standard SV. p is the minimum percentage of allele sequence similarity between SV call and gold standard SV. O and p vary from 0-1 with a 0.1 interval. **(A-C)** The F1 heatmap for deletions by SVIM 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 SVIM on three CLR datasets. **(G-I)** The F1 heatmap for deletions by SVIM on three ONT datasets.



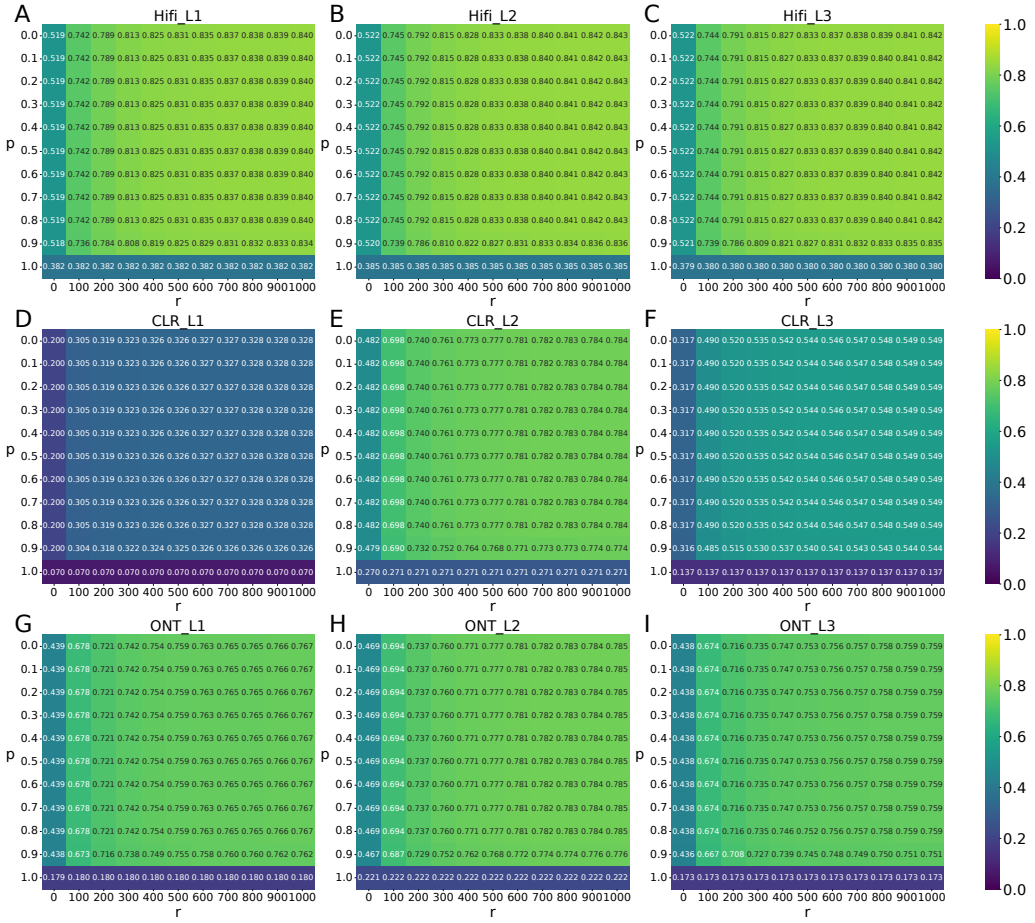
Supplemental Figure S21: **Insertion F1 accuracy of FocalSV(target) by tuning different evaluation parameters (p and r).** r is the maximum reference location distance between SV call and gold standard SV. p is the minimum percentage of allele sequence similarity between SV call and gold standard SV. 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(target) 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(target) on three CLR datasets. **(G-I)** The F1 heatmap for insertions by FocalSV(target) on three ONT datasets.



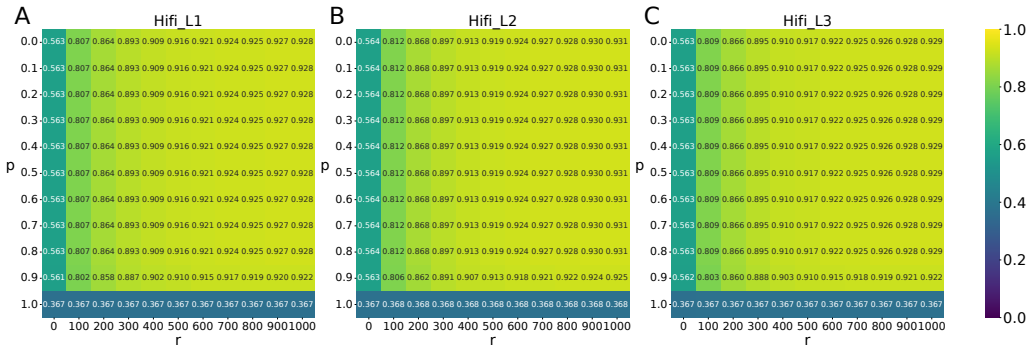
Supplemental Figure S22: **Insertion F1 accuracy of PAV by tuning different evaluation parameters (p and r).** r is the maximum reference location distance between SV call and gold standard SV. p is the minimum percentage of allele sequence similarity between SV call and gold standard SV. p vary from 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 PAV 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 PAV on three CLR datasets. **(G-I)** The F1 heatmap for insertions by PAV on three ONT datasets.



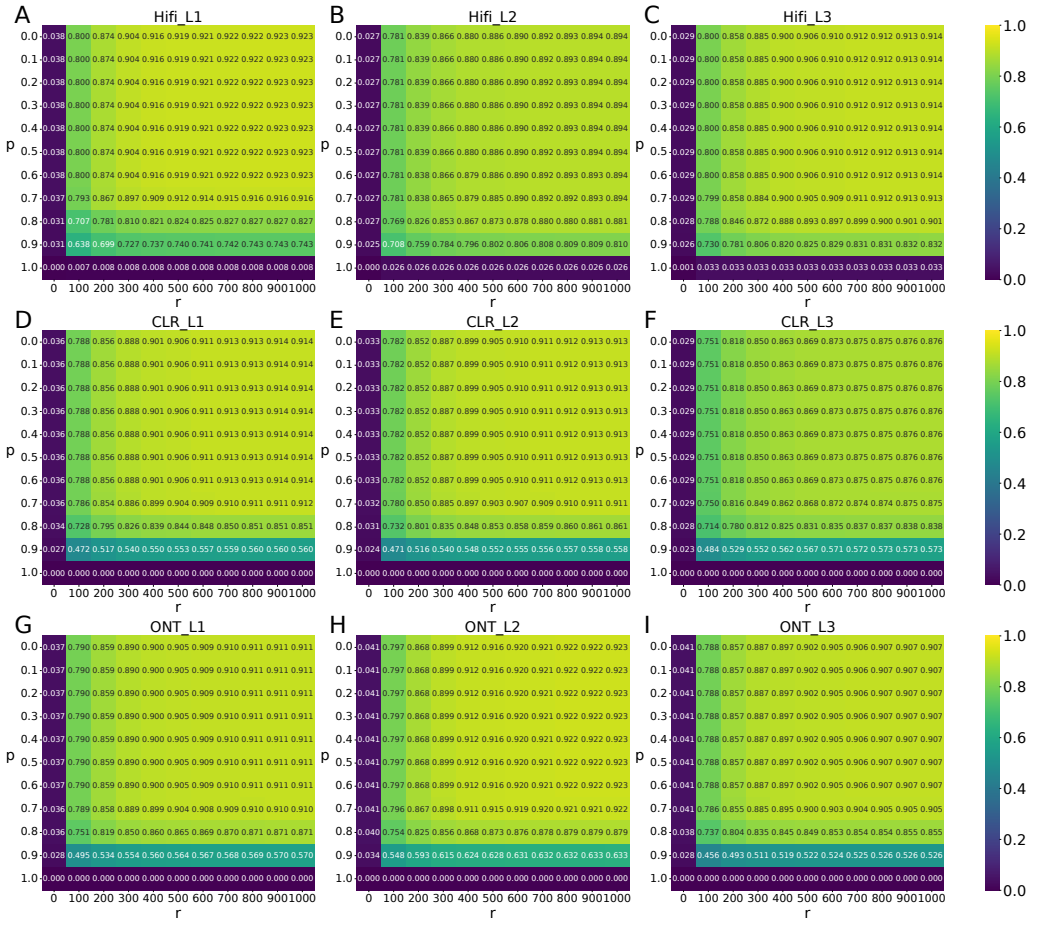
Supplemental Figure S23: **Insertion F1 accuracy of SVIM-asm by tuning different evaluation parameters (p and r).** r is the maximum reference location distance between SV call and gold standard SV. p is the minimum percentage of allele sequence similarity between SV call and gold standard SV. p vary 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 SVIM-asm 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 SVIM-asm on three CLR datasets. **(G-I)** The F1 heatmap for insertions by SVIM-asm on three ONT datasets.



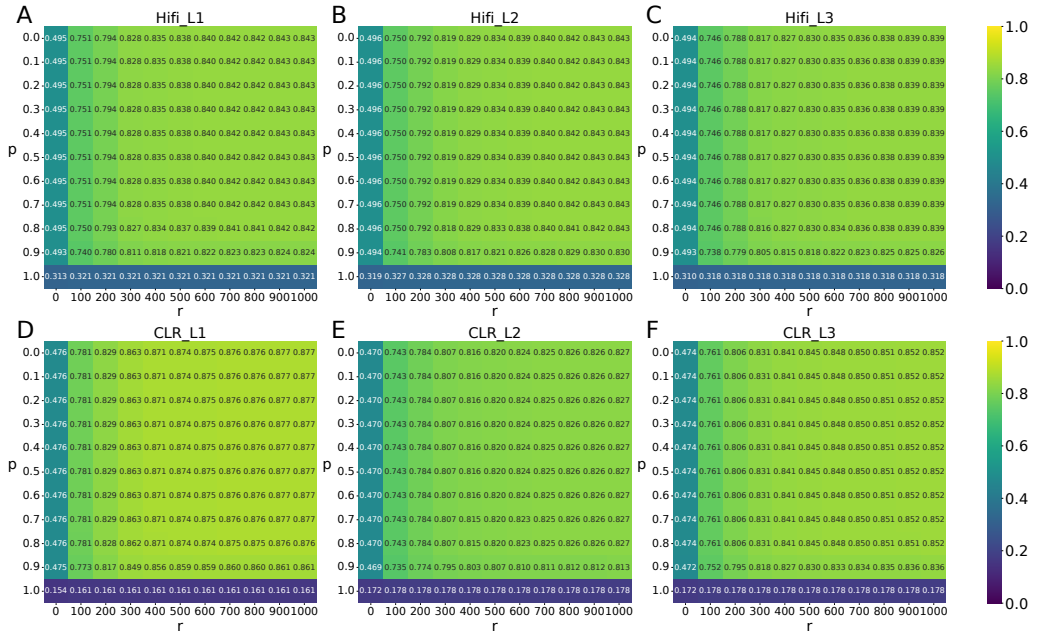
Supplemental Figure S24: **Insertion F1 accuracy of Dipcall by tuning different evaluation parameters (p and r).** r is the maximum reference location distance between SV call and gold standard SV. p is the minimum percentage of allele sequence similarity between SV call and gold standard SV. p vary 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 Dipcall 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 Dipcall on three CLR datasets. **(G-I)** The F1 heatmap for insertions by Dipcall on three ONT datasets.



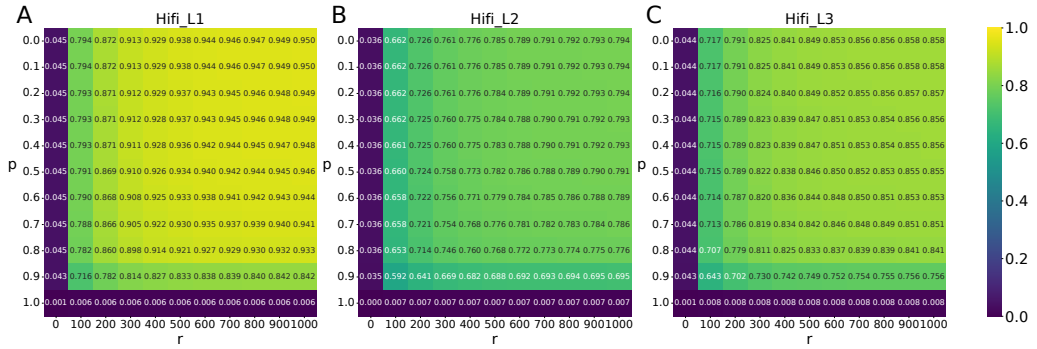
Supplemental Figure S25: **Insertion F1 accuracy of sawfish by tuning different evaluation parameters (p and r).** r is the maximum reference location distance between SV call and gold standard SV. p is the minimum percentage of allele sequence similarity between SV call and gold standard SV. p vary 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 sawfish on three Hifi datasets. Every cell in the heatmap represents the F1 score under a specific pair of p and r evaluation.



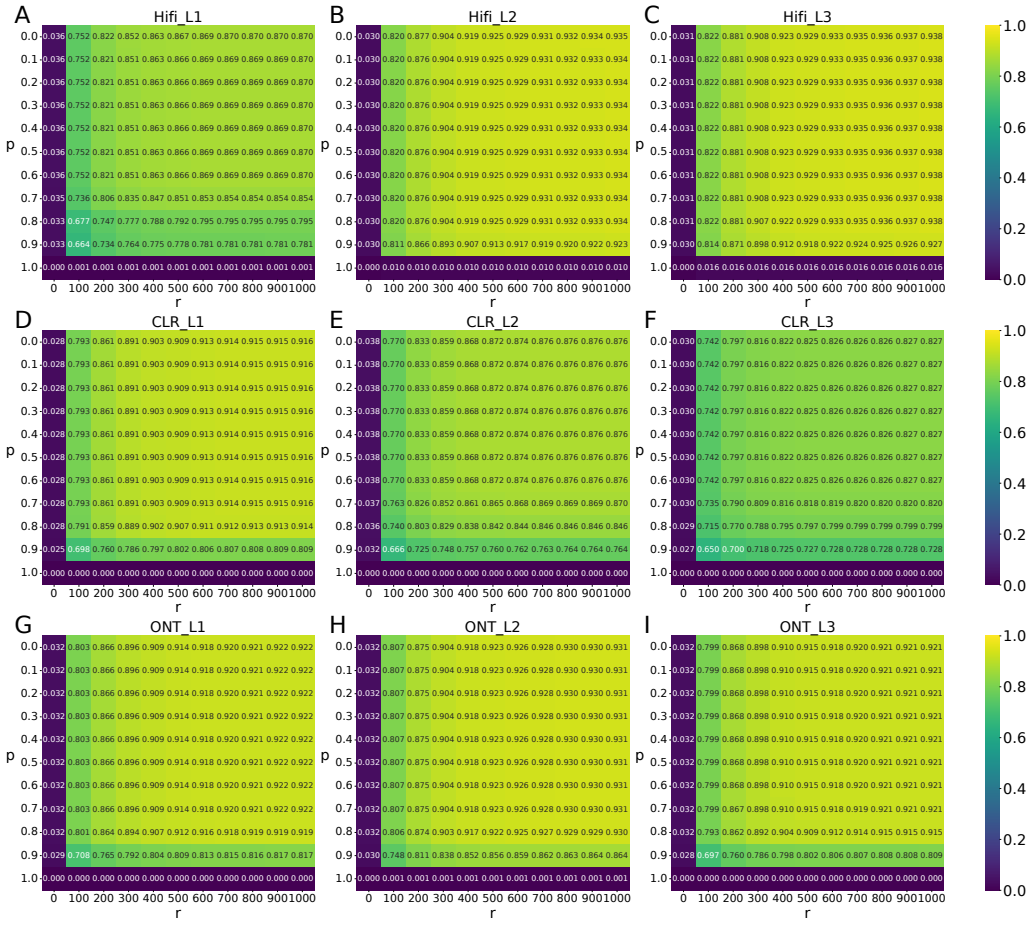
Supplemental Figure S26: **Insertion F1 accuracy of cuteSV by tuning different evaluation parameters (p and r).** r is the maximum reference location distance between SV call and gold standard SV. p is the minimum percentage of allele sequence similarity between SV call and gold standard SV. p vary from 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 cuteSV 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 cuteSV on three CLR datasets. **(G-I)** The F1 heatmap for insertions by cuteSV on three ONT datasets.



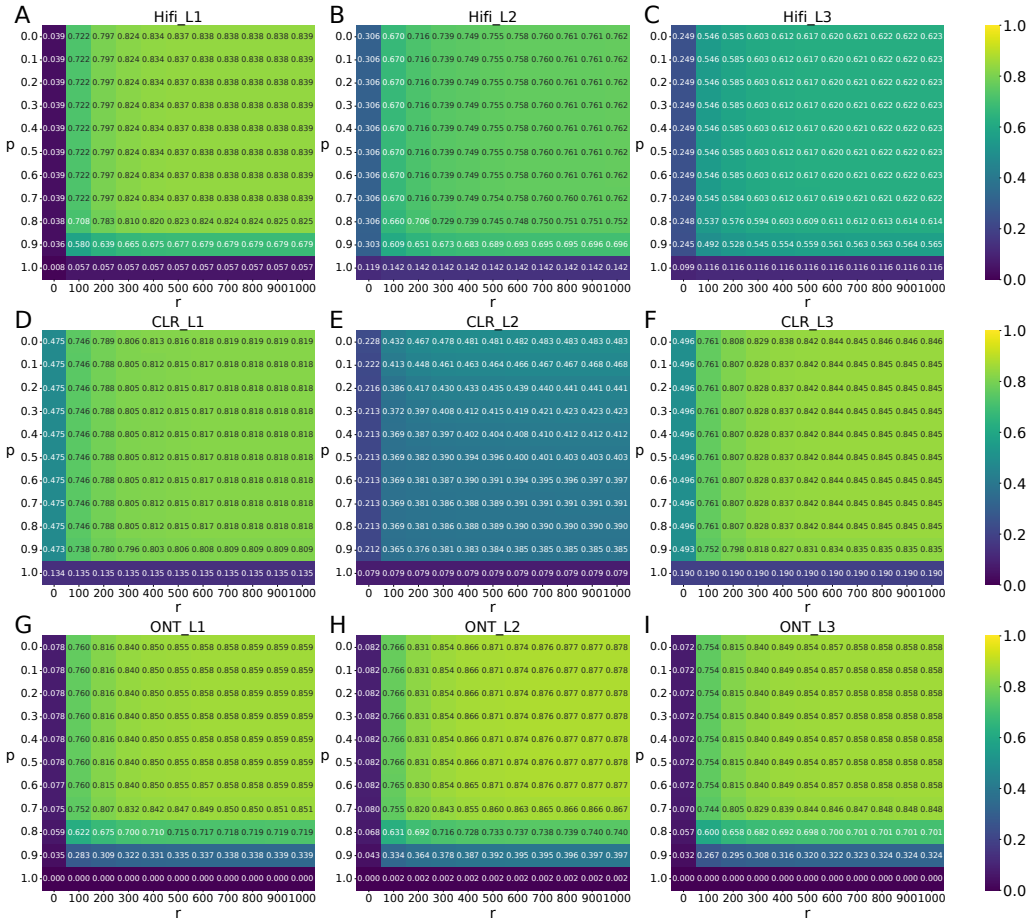
Supplemental Figure S27: **Insertion F1 accuracy of PBSV by tuning different evaluation parameters (p and r).** r is the maximum reference location distance between SV call and gold standard SV. p is the minimum percentage of allele sequence similarity between SV call and gold standard SV. p vary from 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 PBSV 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 PBSV on three CLR datasets.



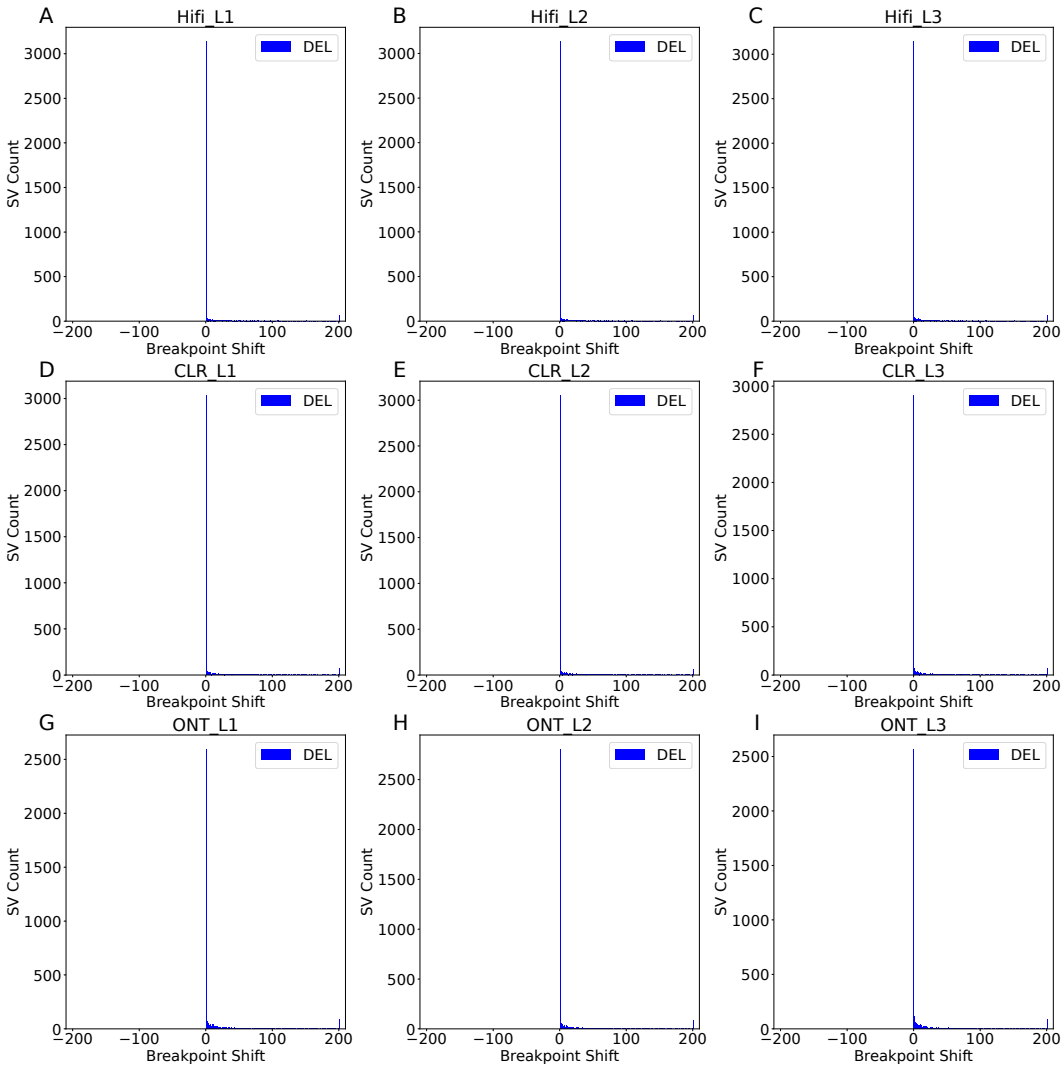
Supplemental Figure S28: **Insertion F1 accuracy of SKSV by tuning different evaluation parameters (p and r).** r is the maximum reference location distance between SV call and gold standard SV. p is the minimum percentage of allele sequence similarity between SV call and gold standard SV. p vary from 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 SKSV on three Hifi datasets. Every cell in the heatmap represents the F1 score under a specific pair of p and r evaluation.



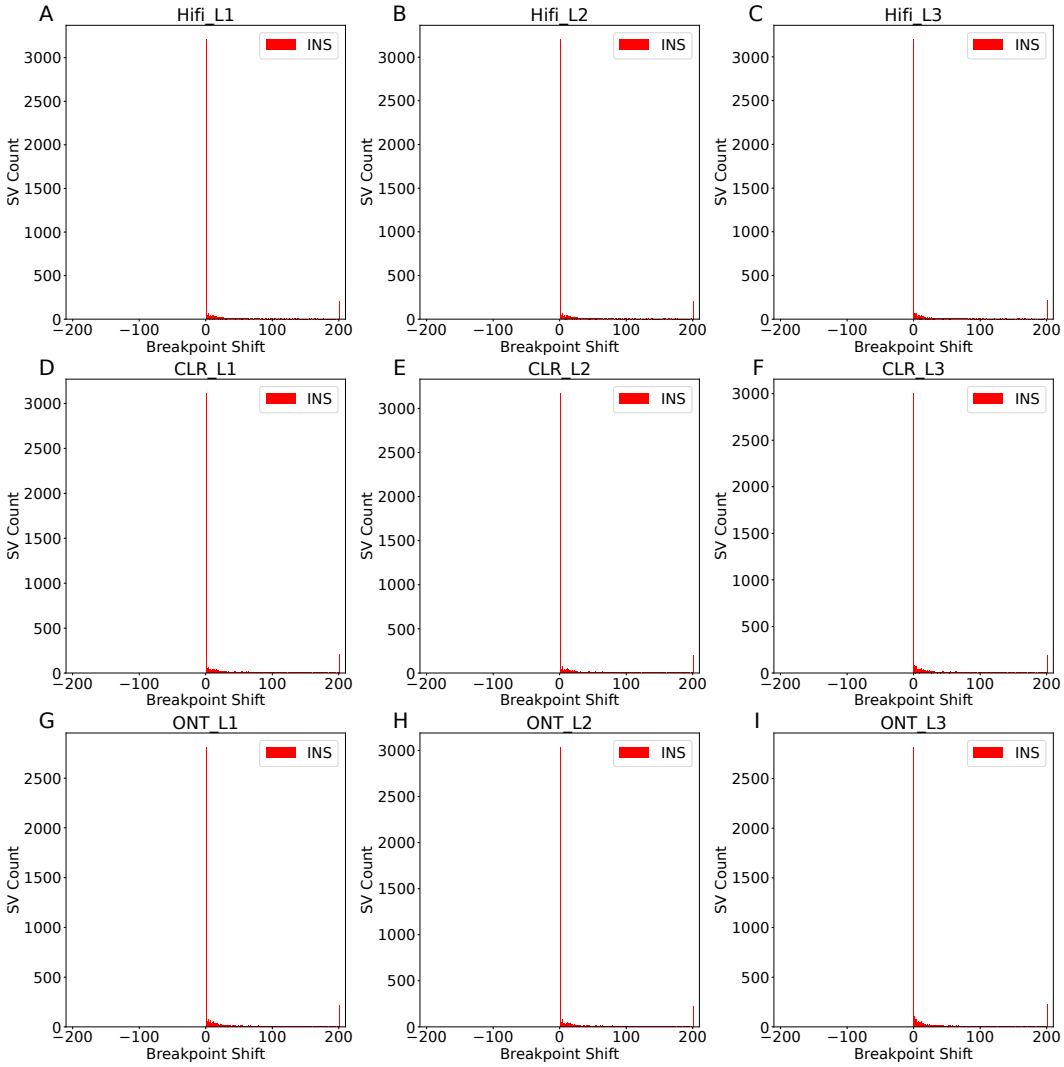
Supplemental Figure S29: **Insertion F1 accuracy of Sniffles2 by tuning different evaluation parameters (p and r).** r is the maximum reference location distance between SV call and gold standard SV. p is the minimum percentage of allele sequence similarity between SV call and gold standard SV. p vary 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 Sniffles2 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 Sniffles2 on three CLR datasets. **(G-I)** The F1 heatmap for insertions by Sniffles2 on three ONT datasets.



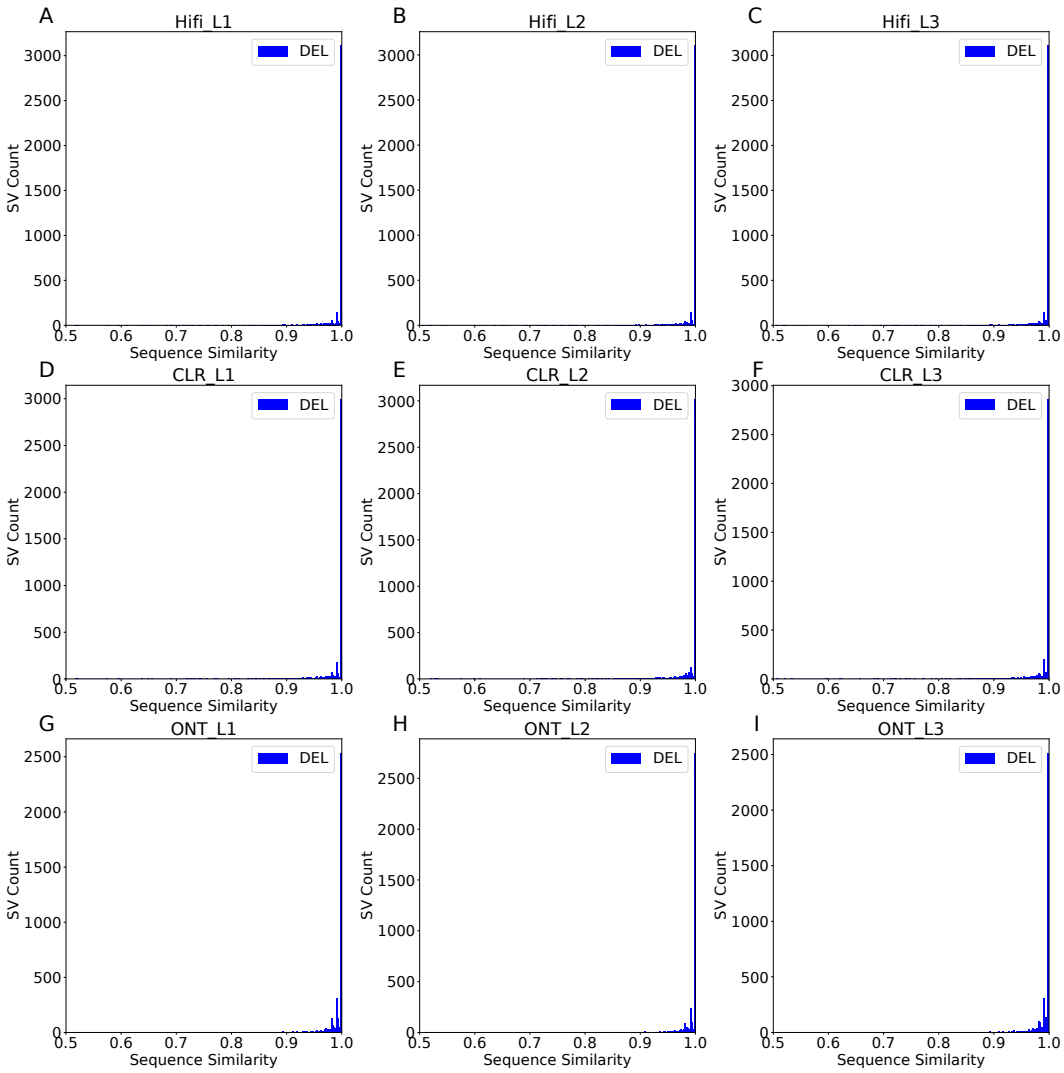
Supplemental Figure S30: **Insertion F1 accuracy of SVIM by tuning different evaluation parameters (p and r).** r is the maximum reference location distance between SV call and gold standard SV. p is the minimum percentage of allele sequence similarity between SV call and gold standard SV. p vary from 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 SVIM 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 SVIM on three CLR datasets. **(G-I)** The F1 heatmap for insertions by SVIM on three ONT datasets.



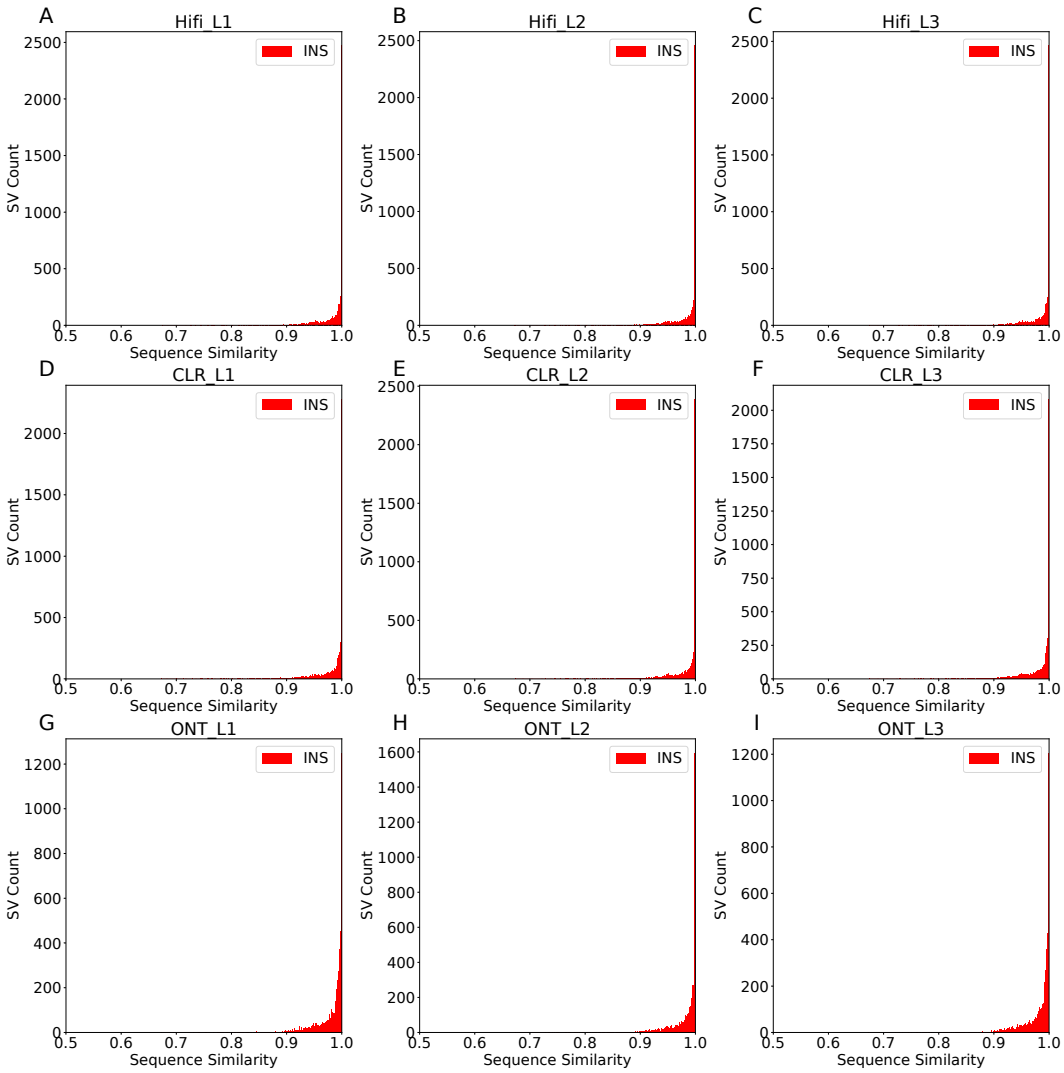
Supplemental Figure S31: **Distribution of breakpoint shift for deletion SVs across all nine libraries. (A-C)** Distribution of breakpoint shift for three Hifi datasets. **(D-F)** Distribution of breakpoint shift for three CLR datasets. **(G-I)** Distribution of breakpoint shift for three ONT datasets. The evaluation was done by moderate Truvari parameters (p=0.5, P=0.5, O=0.01, r=500).



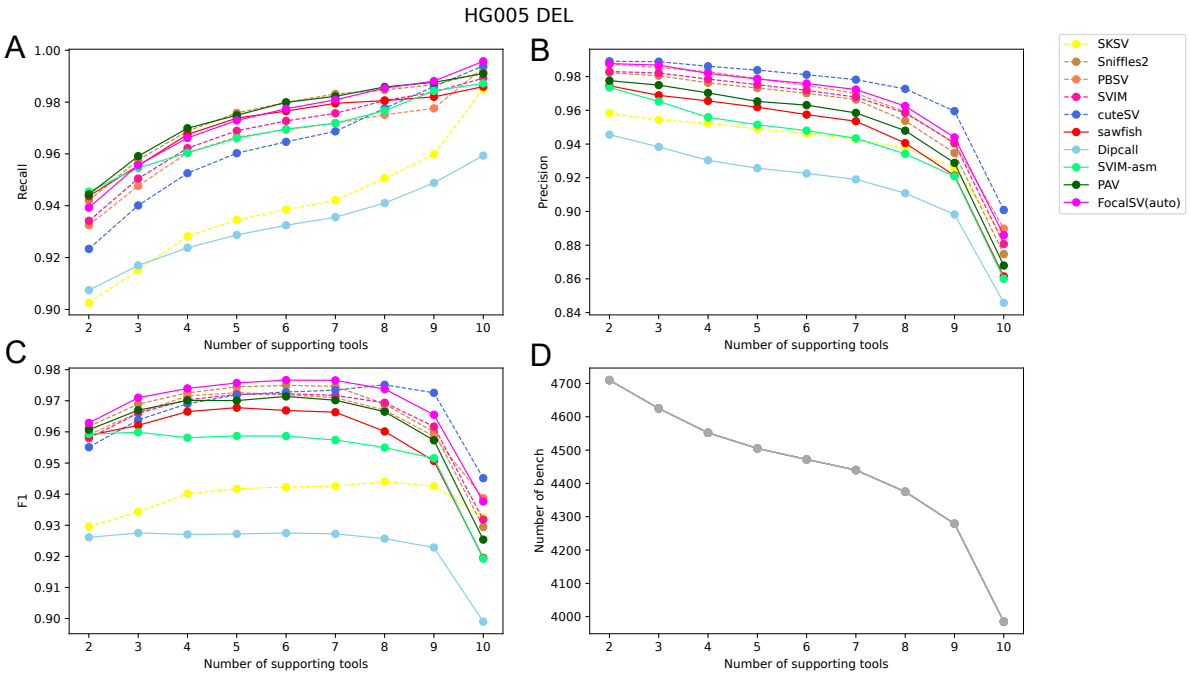
Supplemental Figure S32: **Distribution of breakpoint shift for insertion SVs across all nine libraries. (A-C)** Distribution of breakpoint shift for three Hifi datasets. **(D-F)** Distribution of breakpoint shift for three CLR datasets. **(G-I)** Distribution of breakpoint shift for three ONT datasets. The evaluation was done by moderate Truvari parameters (p=0.5, P=0.5, O=0.01, r=500).



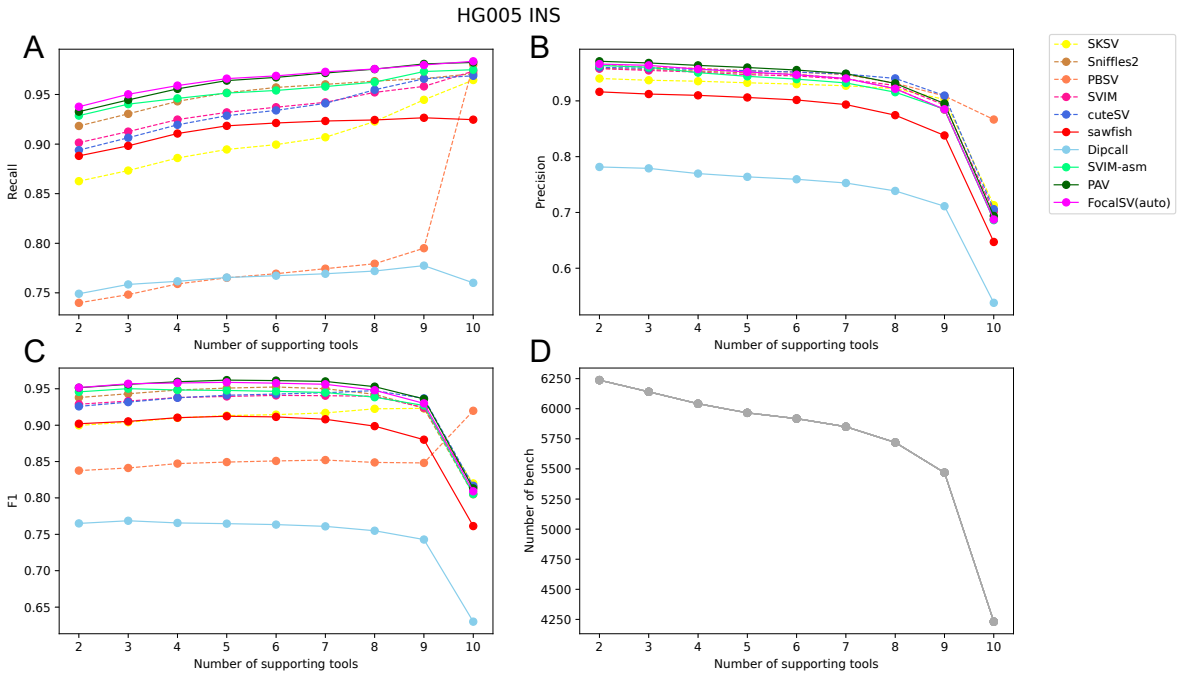
Supplemental Figure S33: **Distribution of sequence similarity for deletion SVs across all nine libraries.** **(A-C)** Distribution of sequence similarity for three Hifi datasets. **(D-F)** Distribution of sequence similarity for three CLR datasets. **(G-I)** Distribution of sequence similarity for three ONT datasets. The evaluation was done by moderate Truvari parameters (p=0.5, P=0.5, O=0.01, r=500).



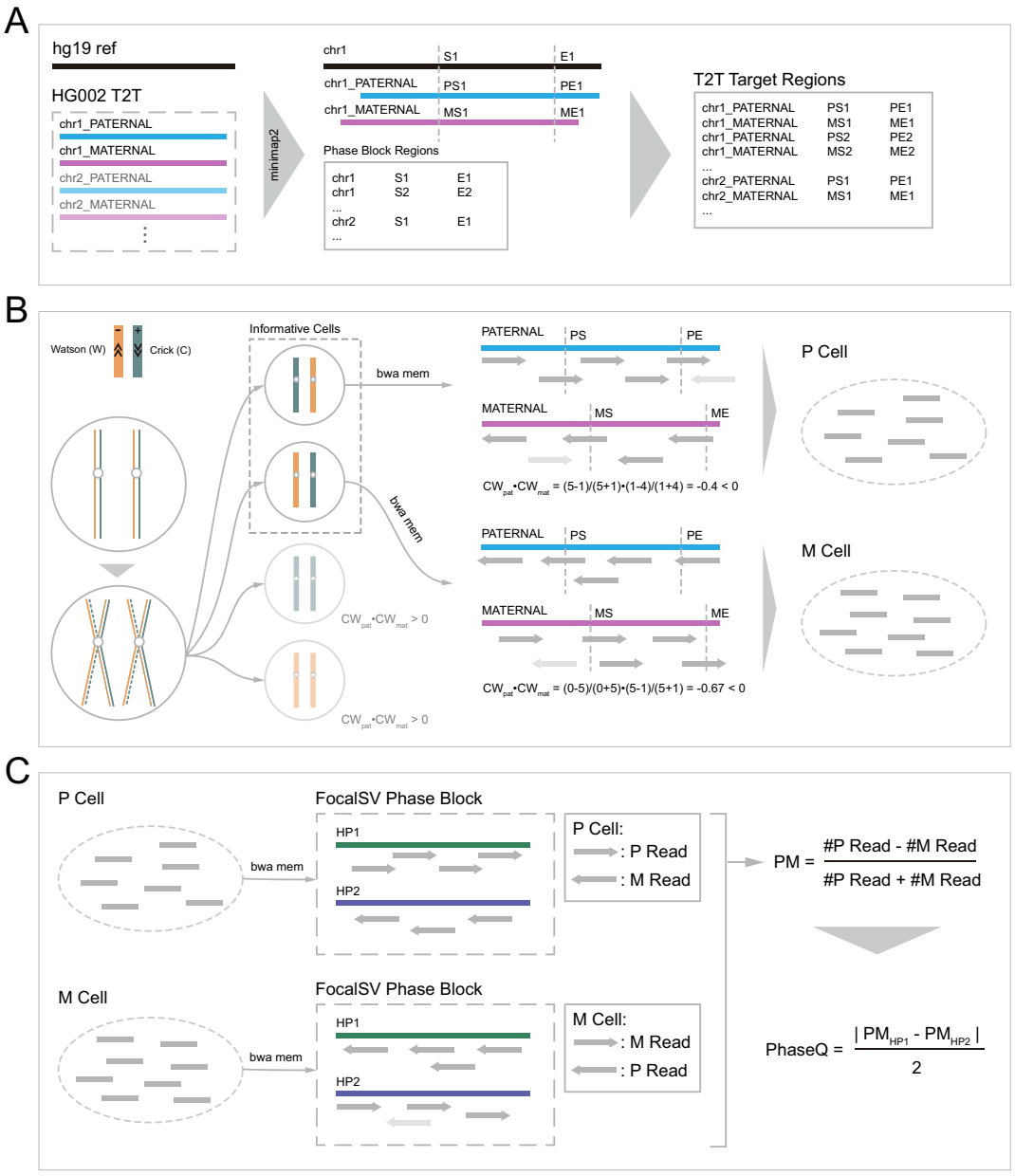
Supplemental Figure S34: **Distribution of sequence similarity for insertion SVs across all nine libraries.** **(A-C)** Distribution of sequence similarity for three Hifi datasets. **(D-F)** Distribution of sequence similarity for three CLR datasets. **(G-I)** Distribution of sequence similarity for three ONT datasets. The evaluation was done by moderate Truvari parameters (p=0.5, P=0.5, O=0.01, r=500).



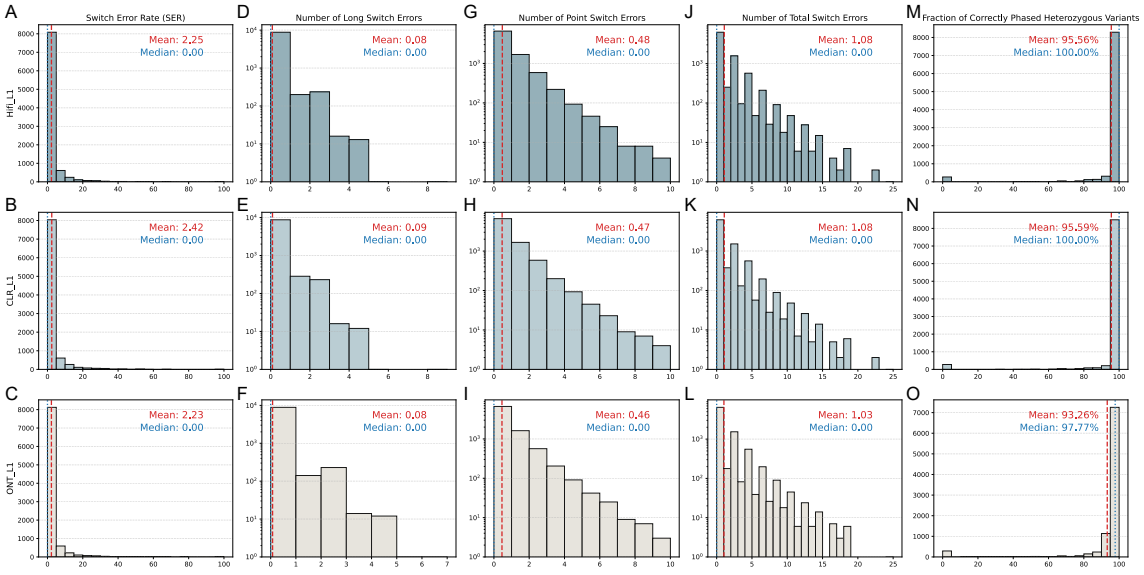
Supplemental Figure S35: **Evaluation of HG005 deletions (DEL) based on the number of supporting tools.** **(A)** Recall across varying numbers of supporting tools. **(B)** Precision across varying numbers of supporting tools. **(C)** F1 across varying numbers of supporting tools. **(D)** Number of deletions in the benchmark set across different levels of tool support.



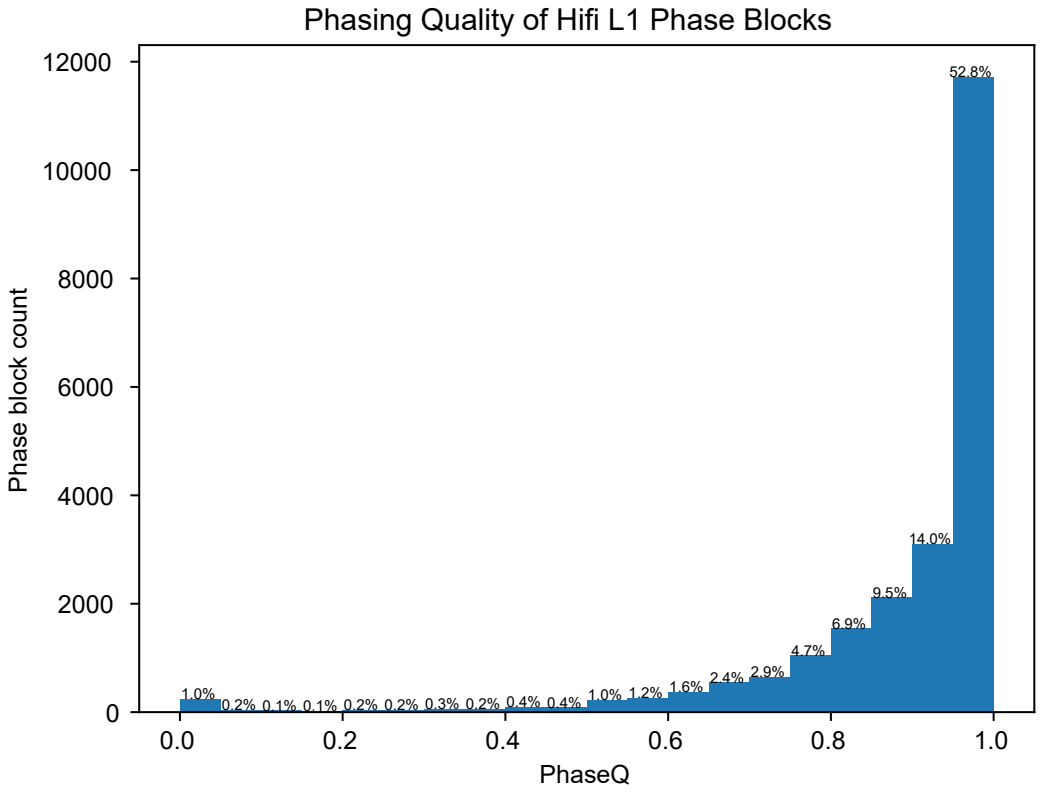
Supplemental Figure S36: **Evaluation of HG005 insertions (INS) based on the number of supporting tools.** **(A)** Recall across varying numbers of supporting tools. **(B)** Precision across varying numbers of supporting tools. **(C)** F1 across varying numbers of supporting tools. **(D)** Number of deletions in the benchmark set across different levels of tool support.



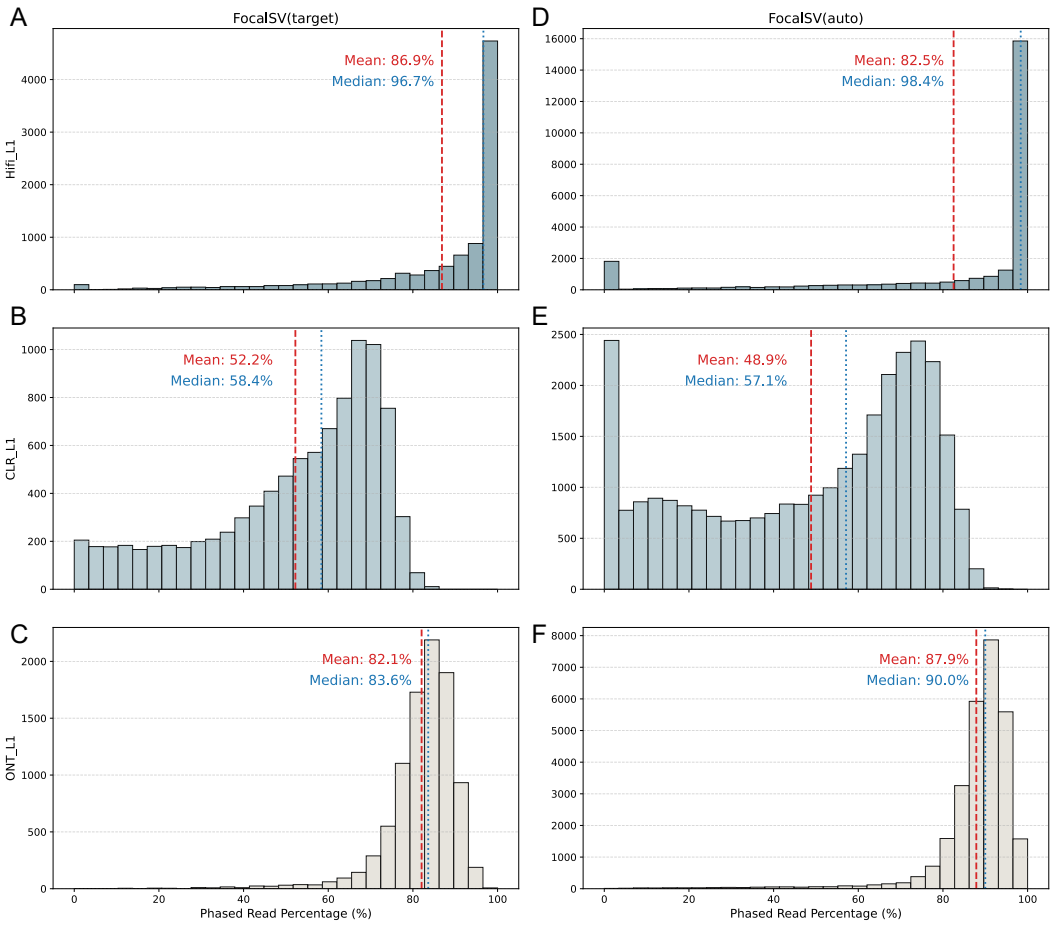
Supplemental Figure S37: **Overview of the Strand-seq-based PhaseQ Evaluation Pipeline.** **(A)** Align the T2T HG002 diploid assembly to the reference genome and extract the contig segments corresponding to genomic regions where local diploid assemblies were performed by FocalSV. **(B)** Strand-seq reads from individual cells were independently aligned to the T2T HG002 diploid contigs to identify informative cells. These cells were further classified into P and M types based on haplotype polarity. Reads mapping to the relevant contig segments corresponding to local assemblies were then extracted for downstream analysis. **(C)** Strand-seq reads from informative cells were then aligned to the HP1 and HP2 contigs within each phase block to compute the final PhaseQ score. A detailed method is described in the corresponding Method section.



Supplemental Figure S38: **Distribution of phasing evaluation metrics across different libraries. (A-C)** Distribution of switch error rate (SER) on Hifi_L1, CLR_L1 and ONT_L1. **(D-F)** Distribution of number of long switch errors on Hifi_L1, CLR_L1 and ONT_L1. **(G-I)** Distribution of number of point switch errors on Hifi_L1, CLR_L1 and ONT_L1. **(j-l)** Distribution of number of total switch errors on Hifi_L1, CLR_L1 and ONT_L1. **(m-o)** Distribution of fraction of correctly phased heterozygous variants on Hifi_L1, CLR_L1 and ONT_L1. The red dashed line indicates the mean, and the blue dashed line indicates the median for each metric.



Supplemental Figure S39: **Distribution of PhaseQ scores from Strand-seq-based phasing evaluation on Hifi_L1 by FocalSV(auto).** Distribution of PhaseQ scores across phase blocks generated by FocalSV(auto), with 83.2% of blocks achieving a PhaseQ score above 0.8.



Supplemental Figure S40: **Distribution of phased read percentages across different libraries. (A-C)** Distribution of phased read percentages for FocalSV(target) on Hifi_L1, CLR_L1, and ONT_L1 datasets. **(D-F)** Distribution of phased read percentages for FocalSV(auto) on Hifi_L1, CLR_L1, and ONT_L1 datasets.

5 Supplemental References

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

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 have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nature communications* **8**: 14061.

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 reference cell line sample using multiple sequencing technologies. *Genome biology* **23**: 1–33.

Wagner J, Olson ND, Harris L, Khan Z, Farek J, Mahmoud M, Stankovic A, Kovacevic V, Yoo B, Miller N, et al.. 2022. Benchmarking challenging small variants with linked and long reads. *Cell genomics* **2**.