Benchmarking Reports

```
ASVBM command:
$ ./asvbm -m 50000 -T
ASVCLR;SVDSS;DeBreak;Sniffles2;pbsv;cuteSV;SVIM -C
1;2;3;4;5;6;7;8;9;10;11;12;13;14;15;16;17;18;19;20;21;22;X;Y
./Data/output_ASVCLR.vcf ./Data/output_SVDSS.vcf
./Data/output_DeBreak.vcf ./Data/output_Sniffles2.vcf
./Data/output_pbsv.vcf ./Data/output_cuteSV.vcf
./Data/output_SVIM.vcf ./Data/HG002_SVs_Tier1_v0.6.vcf
./Data/hs37d5.fa -o Tier1_eval
```

1. Benchmarking results

Variant type match mode: **loose (allow type match between DUPLICTION and INSERTION)**

The benchmarking metrics has two categories after filtering long SV regions: one category is used to highlight performance by metrics including Recall, Precision, F1 score, and sequence identity (Identity) and the other category presents benchmark results, which consists of TP_bench, TP_user, FP, FN. Visualizing these metrics through bar charts provides a more intuitive representation of the benchmarking results for the variation detection methods.

(1) The benchmarking results of the user-called set are as follows:

Table 1 Structural Variation Detection Method Performance Benchmarking

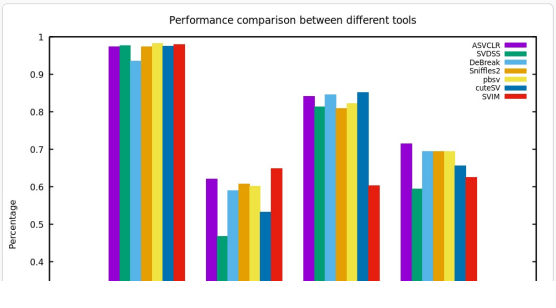
Tool	#SVs_bench	#SVs_user	#SVs_filtered_user	#TP_bench	#TP_user	#FP	#FN	Identity	Recall	Precision	F1 score
ASVCLR	74012	54423	54423	45986	45807	8616	28026	0.973949	0.621332	0.841685	0.714914
SVDSS	74012	45787	45787	34684	37221	8566	39328	0.977112	0.468627	0.812916	0.594524
DeBreak	74012	49868	49709	43644	41248	7565	30368	0.936123	0.589688	0.845021	0.694634
Sniffles2	74012	54545	54458	44973	43106	10168	29039	0.973332	0.607645	0.809138	0.694063
pbsv	74012	52807	52741	44492	42927	9253	29520	0.983104	0.601146	0.822672	0.694676
cuteSV	74012	44937	44928	39438	36952	6416	34574	0.975146	0.532860	0.852057	0.655674
SVIM	74012	116615	116427	48022	47230	30995	25990	0.979928	0.648841	0.603771	0.625495

The table 1 shows the benchmarking results of the variation identification result. Where #SVs_bench represents the number of identified structural variations (SVs) in the benchmark set, #SV_user represents the number of SVs in the called set, and #SV_filtered_user represents the number of SVs after filtering out large SVs. #TP stands for the number of True Positives, indicating correctly identified targets or events. #FP stands for the number of False Positives, representing falsely identified targets or events. #FN represents the number of False Negatives, referring to the targets or events that were missed or not identified correctly. Identity represents the measure of sequence identity, which is calculated for matched SV pairs that include sequences.

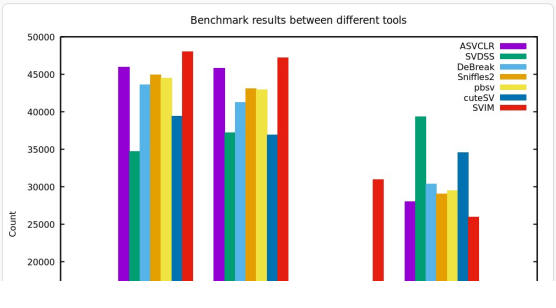
(2) The benchmarking results of two categorizes of metrics are shown in the figure:

Two categories of metrics are independently calculated: (a) one category includes Recall, Precision, F1 Score, and Identity; (b) the other category consists of #TP_bench, #TP_user, #FP, and #FN. The result statistics are as follows:

(a)



(b)



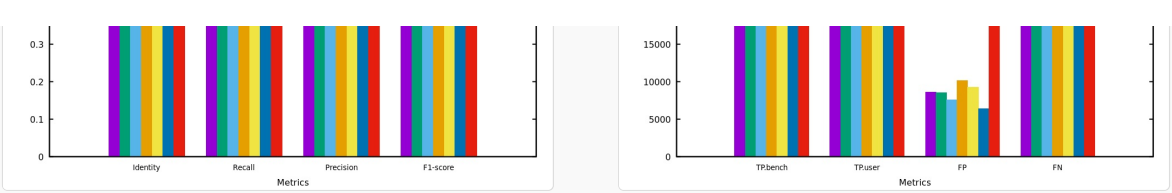


Figure 1 Benchmarking results of the user-call set

(3) The UpSet plot generated by TP_bench for multiple callsets benchmarking is shown as follows:

The UpSet plot illustrates the benchmarking of TP_bench variants generated by TP_bench across multiple user callsets. The plot displays the distribution and intersection of high-confidence variants within the benchmark set.

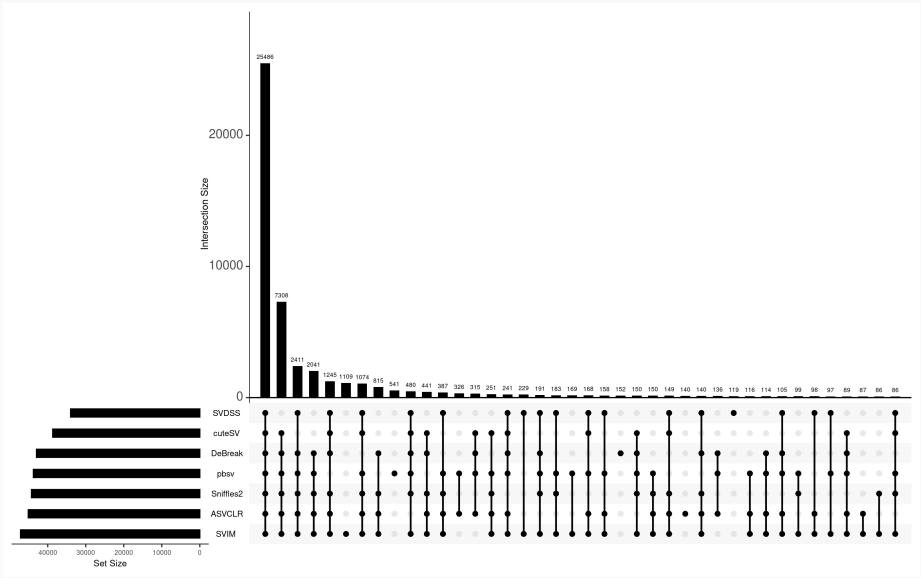


Figure 2 UpSet plot of TP_bench across multiple callsets

2. Statistical results of deviations for overlapping variants

For variations that overlap between the user-called set and the benchmark set, the deviations between them are quantified by calculating the breakpoint distance and the variant size ratio of the overlapping variations.

(1) Deviation of the breakpoint distance

As the breakpoint distance approaches 0, the deviation decreases, indicating a more precise identification result. Statistics results for eight size regions are presented in Table 2:

Table 2 Statistical results of breakpoint distance deviation

Tool	-200--151	-150--101	-100--51	-50--1	0-50	51-100	101-150	151-200
ASVCLR	241	358	756	6408	31364	4160	1924	1313
SVDSS	281	373	902	5463	27417	2188	1520	1122
DeBreak	278	451	849	3435	31649	3741	1546	963
Sniffles2	276	355	646	3154	32890	4559	2006	1394
pbsv	288	438	812	5027	32552	2314	1476	1071
cuteSV	182	276	604	4750	27134	3410	1541	1107
SVIM	553	686	1159	5988	35797	3855	2291	1737

(2) Deviation of the variant size ratio

Calculating the variant size ratio for two overlapping variations based on the length of SVs, the closer the ratio is to 1, the smaller the deviation, indicating a more precise and accurate identification result. Statistics results for nine size regions are presented in Table 3:

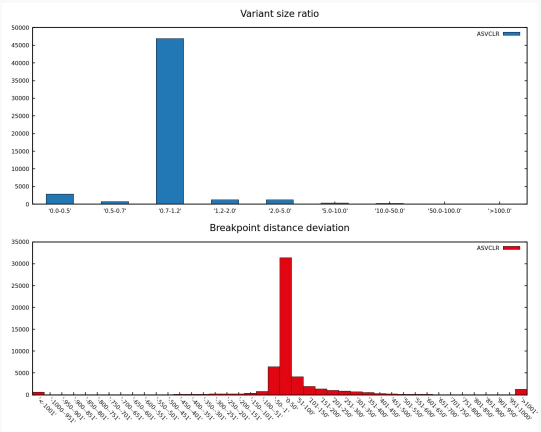
Table 3 Statistical results of deviation of the variant region size ratio

Tool	0.0-0.5	0.5-0.7	0.7-1.2	1.2-2.0	2.0-5.0	5.0-10.0	10.0-50.0	50.0-100.0	>100.0

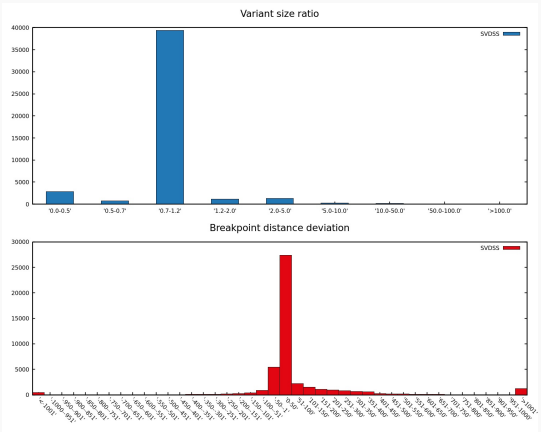
ASVCLR	2885	773	46869	1287	1222	306	248	31	23
SVDSS	2874	737	39357	1186	1330	283	154	6	6
DeBreak	2559	657	40170	2723	1841	456	292	21	40
Sniffles2	3257	752	46170	1244	1236	342	231	31	43
pbsv	3500	840	43660	1724	1489	401	321	40	62
cuteSV	2916	615	39079	1039	934	213	175	16	24
SVIM	7184	1303	51061	1730	1752	607	524	85	173

The statistical results of user-called sets are as follows:

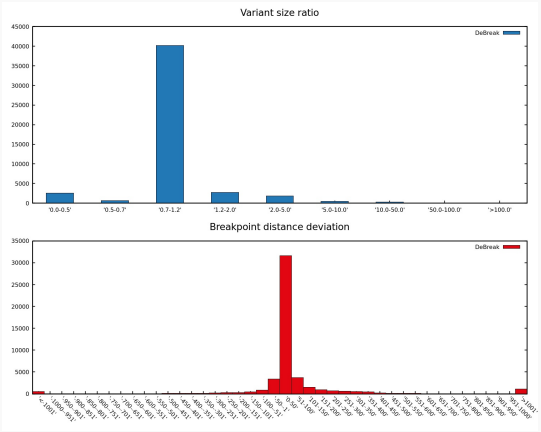
(a) ASVCLR:



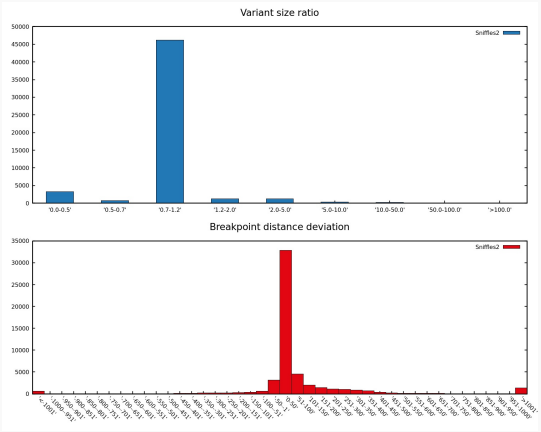
(b) SVDSS:



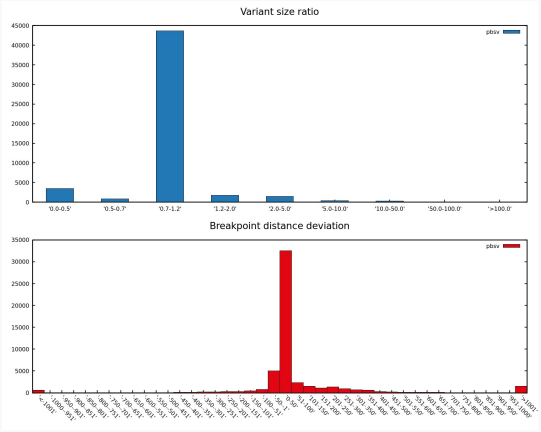
(c) DeBreak:



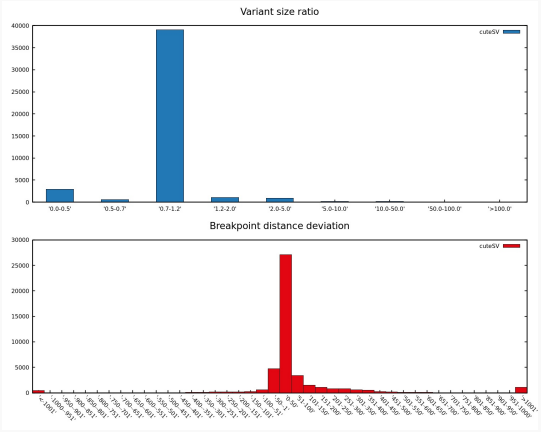
(d) Sniffles2:



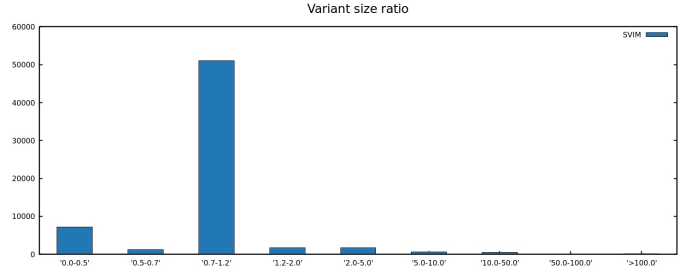
(e) pbsv:



(f) cuteSV:



(g) SVIM:



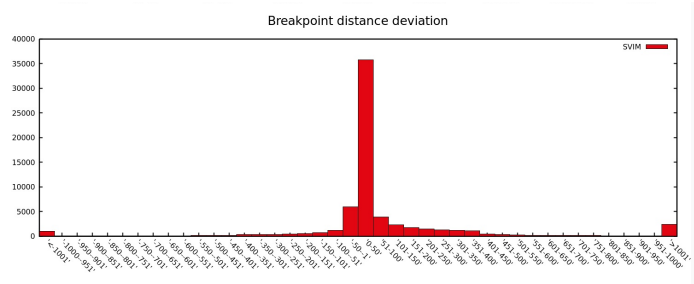


Figure 3 Deviation statistics with overlapping variations

3. Benchmarking results for metrics of different SV size regions

The SV identification results typically contain variations of various sizes, and categorize these variations into different size ranges could be used to explore the identification results more detailed in a fine-grained manner, and could provide new insights into the sensitivity of SV callers to variations of different sizes. Detailed benchmarking results are presented in the table as followsï¼š

(1) Benchmarking results for metrics of different SV size regions with different methods

Variations are categorized into eight size regions and metrics are computed for comprehensive benchmarking for different detection methods within each region. The benchmarking results are as follows:

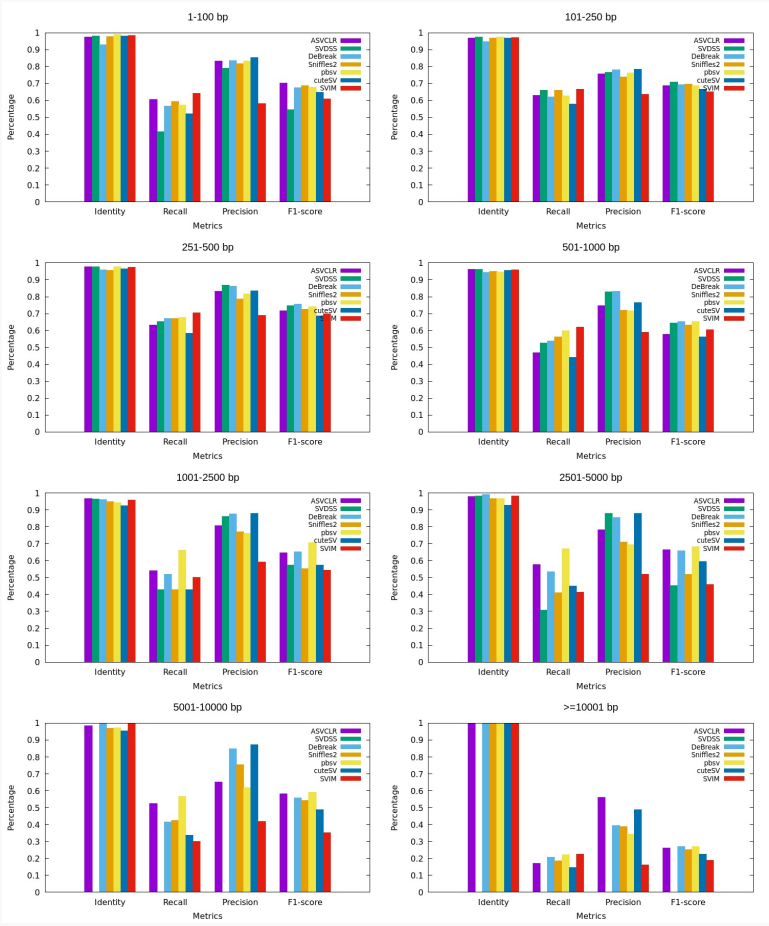


Figure 4 Statistics of metrics of different SV size region

(2) The user-called set (ASVCLR) of basic metrics results statistics

Table 4 The metric benchmarking results of ASVCLR in different SV regions

Region	#TP_bench	#TP_user	#FP	#FN	Identity	Recall	Precision	F1 score
1-100bp	32744	31699	6409	21190	0.975563	0.607112	0.831820	0.701921
101-250bp	4771	4553	1453	2804	0.970400	0.629835	0.758075	0.688031
251-500bp	3905	4724	960	2282	0.977244	0.631162	0.831105	0.717464

501-1000bp	1228	1380	466	1386	0.963267	0.469778	0.747562	0.576976
1001-2500bp	1128	1292	307	958	0.967407	0.540748	0.808005	0.647898
2501-5000bp	478	554	155	349	0.979408	0.577993	0.781382	0.664472
5001-10000bp	263	249	133	237	0.983324	0.526000	0.651832	0.582195
>10000bp	49	50	39	240	0.999628	0.169550	0.561798	0.260486

Benchmarking results for metrics of different SV size regions show as following figures:

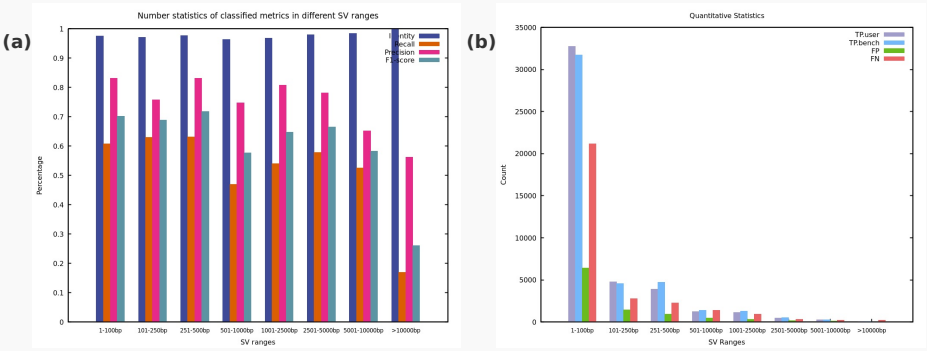


Figure 5 Result statistics of different SV size ranges

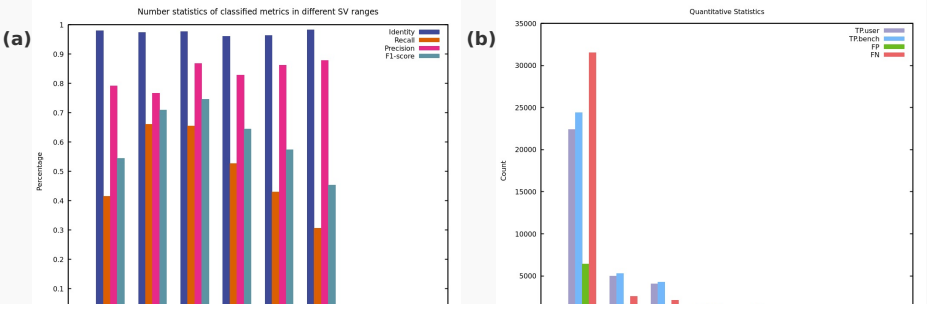
Figure (a) shows the statistical results of Recall, Precision, F1 score and Identity; (b) shows the statistical results of #TP_benchmark, #TP_user,#FP and #FN.

(3) The user-called set (SVDSS) of basic metrics results statistics

Table 5 The metric benchmarking results of SVDSS in different SV regions

Region	#TP_bench	#TP_user	#FP	#FN	Identity	Recall	Precision	F1 score
1-100bp	22422	24398	6436	31512	0.980052	0.415730	0.791269	0.545078
101-250bp	4998	5296	1620	2577	0.973636	0.659802	0.765761	0.708843
251-500bp	4050	4290	657	2137	0.976850	0.654598	0.867192	0.746046
501-1000bp	1376	1438	298	1238	0.960658	0.526396	0.828341	0.643720
1001-2500bp	897	927	148	1189	0.963567	0.430010	0.862326	0.573858
2501-5000bp	253	245	34	574	0.981830	0.305925	0.878136	0.453767
5001-10000bp	0	0	0	500	0.000000	0.000000	0.000000	0.000000
>10000bp	0	0	0	289	0.000000	0.000000	0.000000	0.000000

Benchmarking results for metrics of different SV size regions show as following figures:



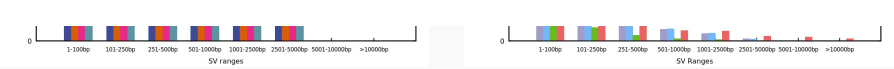


Figure 6 Result statistics of different SV size ranges

Figure (a) shows the statistical results of Recall, Precision, F1 score and Identity;
(b) shows the statistical results of #TP_benchmark, #TP_user,#FP and #FN.

(4) The user-called set (DeBreak) of basic metrics results statistics

Table 6 The metric benchmarking results of DeBreak in different SV regions

Region	#TP_bench	#TP_user	#FP	#FN	Identity	Recall	Precision	F1 score
1-100bp	30552	29153	5740	23382	0.930498	0.566470	0.835497	0.675171
101-250bp	4706	4327	1202	2869	0.946854	0.621254	0.782601	0.692656
251-500bp	4160	4088	651	2027	0.959120	0.672378	0.862629	0.755713
501-1000bp	1409	1361	277	1205	0.943603	0.539021	0.830891	0.653863
1001-2500bp	1085	1004	140	1001	0.962569	0.520134	0.877622	0.653163
2501-5000bp	443	410	69	384	0.991700	0.535671	0.855950	0.658955
5001-10000bp	208	200	36	292	1.000000	0.416000	0.847458	0.558060
>10000bp	60	61	94	229	1.000000	0.207612	0.393548	0.271826

Benchmarking results for metrics of different SV size regions show as following figures:

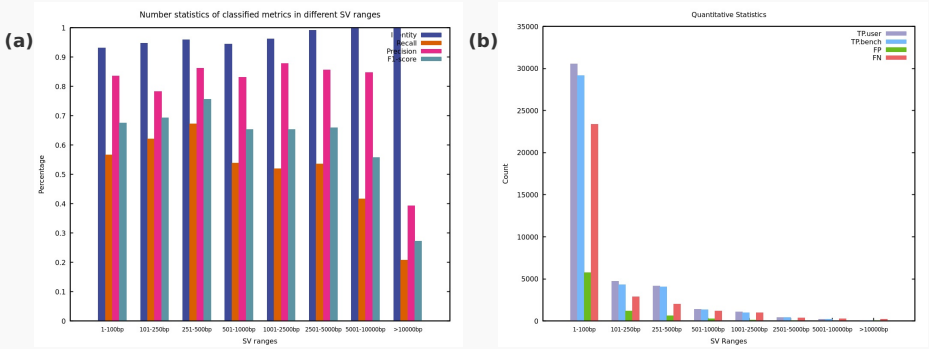


Figure 7 Result statistics of different SV size ranges

Figure (a) shows the statistical results of Recall, Precision, F1 score and Identity;
(b) shows the statistical results of #TP_benchmark, #TP_user,#FP and #FN.

(5) The user-called set (Sniffles2) of basic metrics results statistics

Table 7 The metric benchmarking results of Sniffles2 in different SV regions

Region	#TP_bench	#TP_user	#FP	#FN	Identity	Recall	Precision	F1 score
1-100bp	32046	30472	6809	21888	0.978751	0.594171	0.817360	0.688120
101-250bp	4999	4744	1666	2576	0.968367	0.659934	0.740094	0.697719
251-500bp	4157	4361	1173	2030	0.956056	0.671893	0.788038	0.725345
501-1000bp	1468	1455	568	1146	0.948392	0.561591	0.719229	0.630709
1001-2500bp	896	871	258	1190	0.950466	0.429530	0.771479	0.551825

2501-5000bp	339	331	135	488	0.967038	0.409915	0.710300	0.519834
5001-10000bp	212	213	69	288	0.968087	0.424000	0.755319	0.543119
>10000bp	54	58	91	235	1.000000	0.186851	0.389262	0.252499

Benchmarking results for metrics of different SV size regions show as following figures:

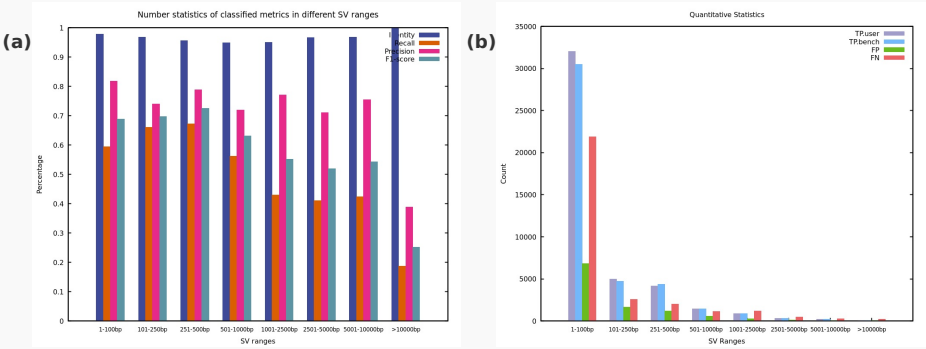


Figure 8 Result statistics of different SV size ranges

Figure (a) shows the statistical results of Recall, Precision, F1 score and Identity;
(b) shows the statistical results of #TP_benchmark, #TP_user,#FP and #FN.

(6) The user-called set (pbsv) of basic metrics results statistics

Table 8 The metric benchmarking results of pbsv in different SV regions

Region	#TP_bench	#TP_user	#FP	#FN	Identity	Recall	Precision	F1 score
1-100bp	30906	29258	5872	23028	0.990925	0.573034	0.832849	0.678934
101-250bp	4748	4499	1399	2827	0.974117	0.626799	0.762801	0.688144
251-500bp	4188	4562	1024	1999	0.977928	0.676903	0.816685	0.740253
501-1000bp	1567	1753	692	1047	0.946939	0.599464	0.716973	0.652974
1001-2500bp	1378	1341	418	708	0.941614	0.660594	0.762365	0.707840
2501-5000bp	555	512	226	272	0.966663	0.671100	0.693767	0.682245
5001-10000bp	283	267	164	217	0.974003	0.566000	0.619490	0.591538
>10000bp	64	66	127	225	0.997893	0.221453	0.341969	0.268822

Benchmarking results for metrics of different SV size regions show as following figures:

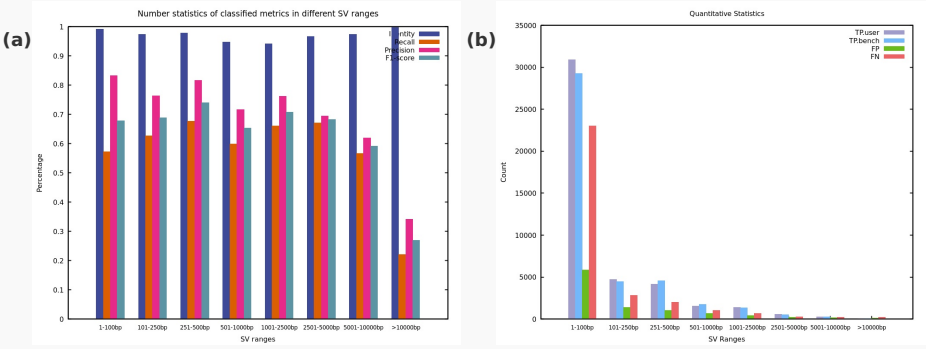


Figure 9 Result statistics of different SV size ranges

Figure (a) shows the statistical results of Recall, Precision, F1 score and Identity;
(b) shows the statistical results of #TP_benchmark, #TP_user,#FP and #FN.

(7) The user-called set (cuteSV) of basic metrics results statistics

Table 9 The metric benchmarking results of cuteSV in different SV regions

Region	#TP_bench	#TP_user	#FP	#FN	Identity	Recall	Precision	F1 score
1-100bp	28083	26487	4512	25851	0.980815	0.520692	0.854447	0.647067
101-250bp	4386	4003	1099	3189	0.968367	0.579010	0.784594	0.666305
251-500bp	3605	3565	698	2582	0.963644	0.582673	0.836266	0.686809
501-1000bp	1158	1081	333	1456	0.956337	0.442999	0.764498	0.560949
1001-2500bp	891	813	113	1195	0.925314	0.427133	0.877970	0.574683
2501-5000bp	372	341	47	455	0.928014	0.449819	0.878866	0.595070
5001-10000bp	169	164	24	331	0.955794	0.338000	0.872340	0.487220
>10000bp	42	43	45	247	1.000000	0.145329	0.488636	0.224028

Benchmarking results for metrics of different SV size regions show as following figures:

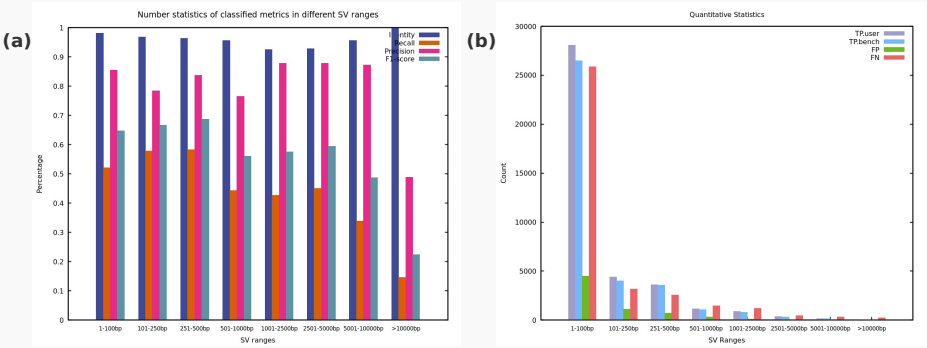


Figure 10 Result statistics of different SV size ranges

Figure (a) shows the statistical results of Recall, Precision, F1 score and Identity;
(b) shows the statistical results of #TP_benchmark, #TP_user,#FP and #FN.

(8) The user-called set (SVIM) of basic metrics results statistics

Table 10 The metric benchmarking results of SVIM in different SV regions

Region	#TP_bench	#TP_user	#FP	#FN	Identity	Recall	Precision	F1 score
1-100bp	34558	32953	23695	19376	0.983789	0.640746	0.581715	0.609805
101-250bp	5054	5349	3078	2521	0.971451	0.667195	0.634745	0.650566
251-500bp	4359	4858	2175	1828	0.975566	0.704542	0.690744	0.697575
501-1000bp	1619	1720	1199	995	0.959684	0.619357	0.589243	0.603925
1001-2500bp	1043	1032	708	1043	0.958859	0.500000	0.593103	0.542587
2501-5000bp	341	326	300	486	0.982840	0.412334	0.520767	0.460250
5001-10000bp	151	154	214	349	0.999300	0.302000	0.418478	0.350824

>10000bp	65	75	389	224	1.000000	0.224913	0.161638	0.188097

Benchmarking results for metrics of different SV size regions show as following figures:

(a) (b)

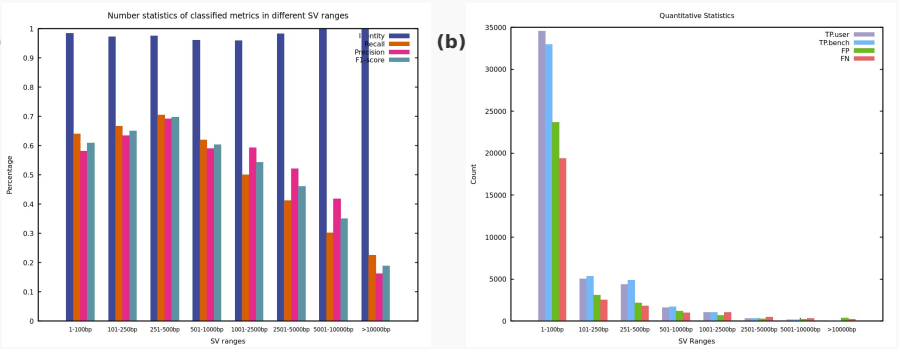


Figure 11 Result statistics of different SV size ranges

Figure (a) shows the statistical results of Recall, Precision, F1 score and Identity; (b) shows the statistical results of #TP_benchmark, #TP_user,#FP and #FN.

4. SV size distribution statistics

(I) Distribution of SV of multiple user callsets

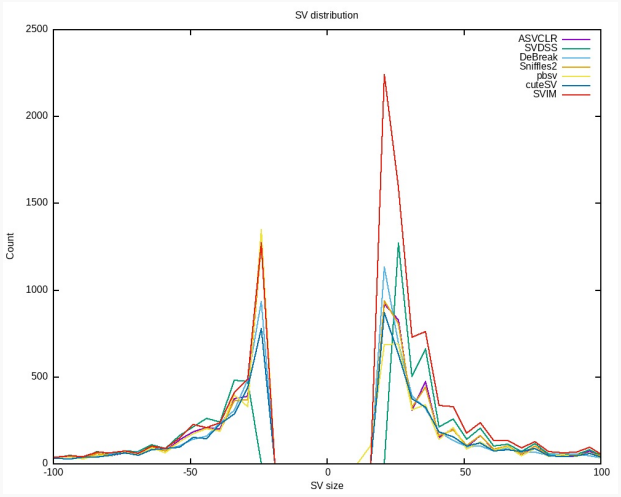


Figure 12 SV distribution for mutiple user callsets

(II) Distribution of SV of benchmark set and user callsets

(1) Statistics of the count of different SV lengths in the benchmark set:

The SV reference region size statistics for benchmark set: Total SVs numberï¼š74012

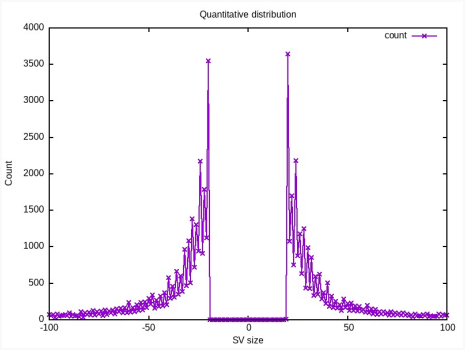


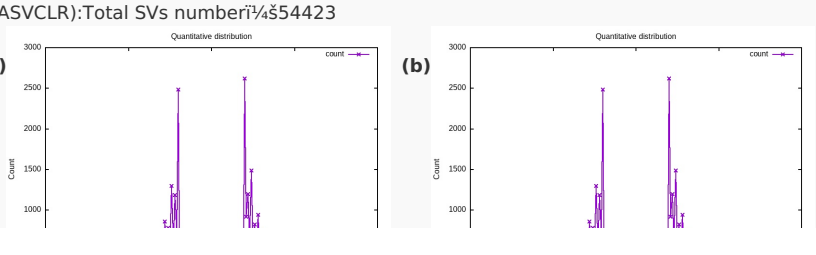
Figure 13 The quantity distribution of the benchmark set

The figure shows the distribution of SV counts of the benchmark set.

(2) Statistics of the count of different SV lengths in the user-called set (ASVCLR):

The SV reference region size statistics before filtering for user-called set (ASVCLR):Total SVs numberï¼š54423
The SV reference region size statistics after filtering for user-called set (ASVCLR):Total SVs numberï¼š54423

(a) (b)



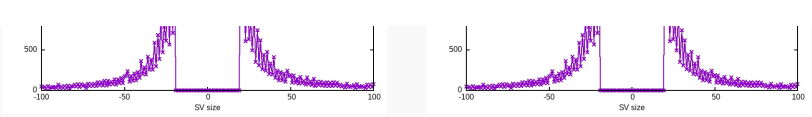


Figure 14 The quantity distribution of the user-called set

The figures show the distribution of SV counts, where (a) represents the result statistics before filtering large SVs, and (b) shows the result statistics after filtering large SVs.

(3) Statistics of the count of different SV lengths in the user-called set (SVDSS):

The SV reference region size statistics before filtering for user-called set (SVDSS):Total SVs numberï¼š45787
The SV reference region size statistics after filtering for user-called set (SVDSS):Total SVs numberï¼š45787

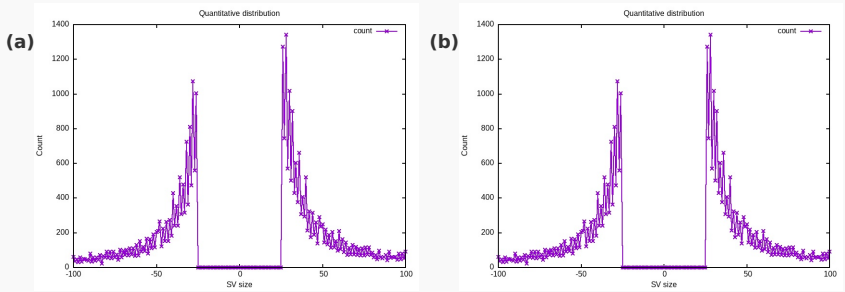


Figure 15 The quantity distribution of the user-called set

The figures show the distribution of SV counts, where (a) represents the result statistics before filtering large SVs, and (b) shows the result statistics after filtering large SVs.

(4) Statistics of the count of different SV lengths in the user-called set (DeBreak):

The SV reference region size statistics before filtering for user-called set (DeBreak):Total SVs numberï¼š49868
The SV reference region size statistics after filtering for user-called set (DeBreak):Total SVs numberï¼š49709

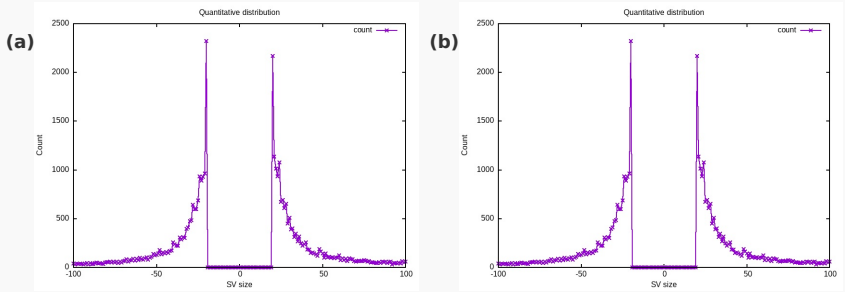


Figure 16 The quantity distribution of the user-called set

The figures show the distribution of SV counts, where (a) represents the result statistics before filtering large SVs, and (b) shows the result statistics after filtering large SVs.

(5) Statistics of the count of different SV lengths in the user-called set (Sniffles2):

The SV reference region size statistics before filtering for user-called set (Sniffles2):Total SVs numberï¼š54545
The SV reference region size statistics after filtering for user-called set (Sniffles2):Total SVs numberï¼š54458

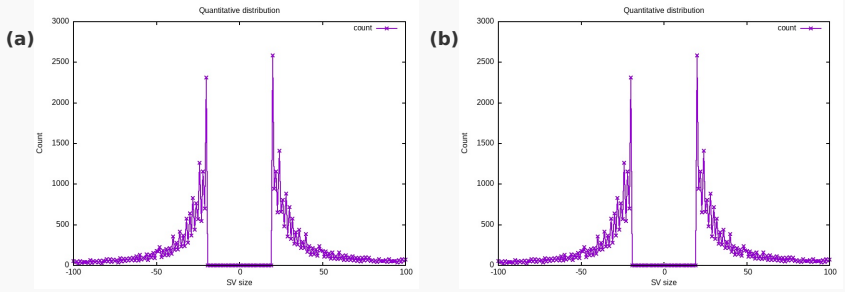


Figure 17 The quantity distribution of the user-called set

The figures show the distribution of SV counts, where (a) represents the result statistics before filtering large SVs, and (b) shows the result statistics after filtering large SVs.

(6) Statistics of the count of different SV lengths in the user-called set (pbsv):

The SV reference region size statistics before filtering for user-called set (pbsv):Total SVs numberï¼š52807
The SV reference region size statistics after filtering for user-called set (pbsv):Total SVs numberï¼š52741

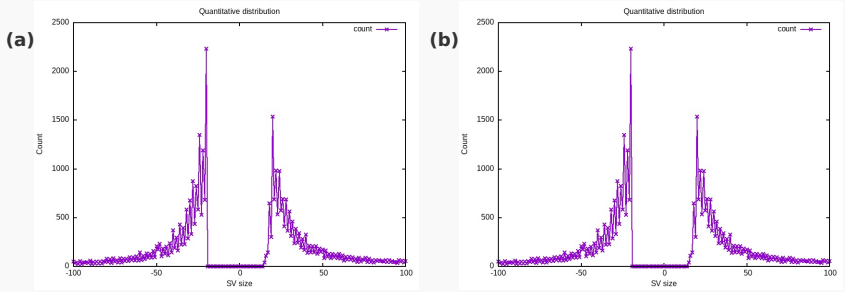


Figure 18 The quantity distribution of the user-called set

The figures show the distribution of SV counts, where (a) represents the result statistics before filtering large SVs, and (b) shows the result statistics after filtering large SVs.

(7) Statistics of the count of different SV lengths in the user-called set (cuteSV):

The SV reference region size statistics before filtering for user-called set (cuteSV):Total SVs numberï¼š44937
The SV reference region size statistics after filtering for user-called set (cuteSV):Total SVs numberï¼š44928

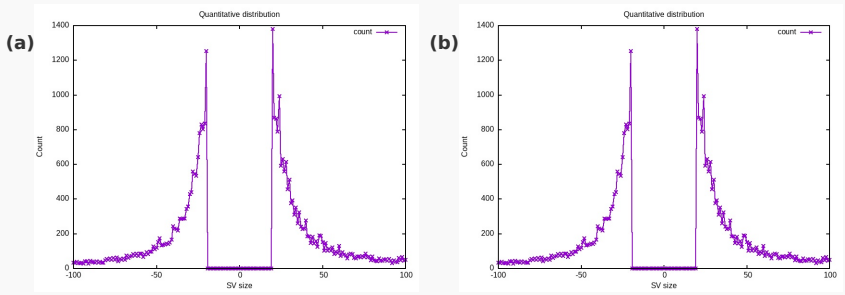


Figure 19 The quantity distribution of the user-called set

The figures show the distribution of SV counts, where (a) represents the result statistics before filtering large SVs, and (b) shows the result statistics after filtering large SVs.

(8) Statistics of the count of different SV lengths in the user-called set (SVIM):

The SV reference region size statistics before filtering for user-called set (SVIM):Total SVs numberï¼š116615
The SV reference region size statistics after filtering for user-called set (SVIM):Total SVs numberï¼š116427

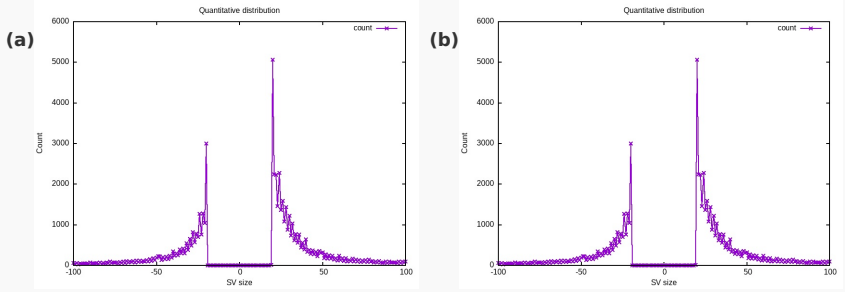


Figure 20 The quantity distribution of the user-called set

The figures show the distribution of SV counts, where (a) represents the result statistics before filtering large SVs, and (b) shows the result statistics after filtering large SVs.

More information

- For more detailed benchmarking results, please refer to the generated result information in the respective folders.
- For more detailed experiment information, please refer to the github repositories: asvbm and asvbm-experiments.
- If you have any problems, comments, or suggestions, please contact xzhu@ytu.edu.cn without hesitation. Thank you very much!

---------- This is the end of the Benchmarking Reports. ----------
-