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"
-o HG002_Tier1_eval hg002_asvclr.vcf hg002_SVDSS.vcf
hg002_Debreak.vcf hg002_Sniffles2.vcf hg002_pbsv.vcf
hg002_cuteSV.vcf hg002_SVIM.vcf HG002_SVs_Tier1_v0.6.vcf
hs37d5.fa
```

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, FP, FN, LP. 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

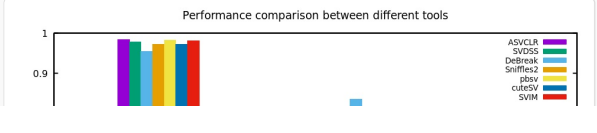
Tool	#SVs_bench	#SVs_user	#SVs_filtered_user	#TP	#FP	#FN	#LP	Identity	Recall	Precision	F1 score
ASVCLR	74012	60029	60029	48504	11806	25508	1842	0.984556	0.655353	0.804245	0.722205
SVDSS	74012	55876	55876	38154	14142	35858	2459	0.977983	0.515511	0.729578	0.604142
DeBreak	74012	51914	51878	44908	8886	29104	99	0.954804	0.606766	0.834814	0.702753
Sniffles2	74012	57555	57532	47362	12035	26650	1228	0.972551	0.639923	0.797380	0.710027
pbsv	74012	55378	55361	45507	12275	28505	718	0.983201	0.614860	0.787564	0.690578
cuteSV	74012	42286	42257	30261	11277	43751	1174	0.971649	0.408866	0.728514	0.523773
SVIM	74012	124008	123950	51510	42845	22502	3230	0.981050	0.695968	0.545917	0.611878

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. #LP represents variant calls that, referring to adjacent or overlapping variants, collectively resemble true positives. Identity represents the measure of sequence identity, which is calculated for matched SV pairs that include sequences.

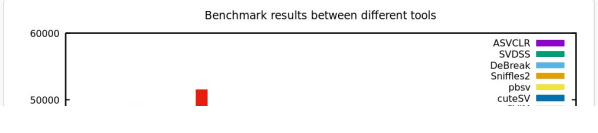
(2) The benchmarking results of two categorizes of metrics are shown in the figure:

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

(a)



(b)



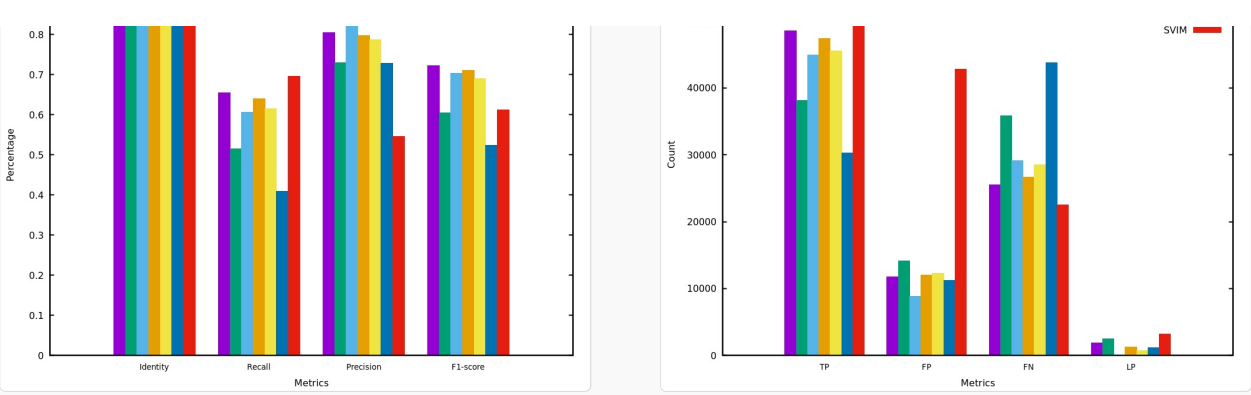


Figure 1 Benchmarking results of the user-call set

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

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

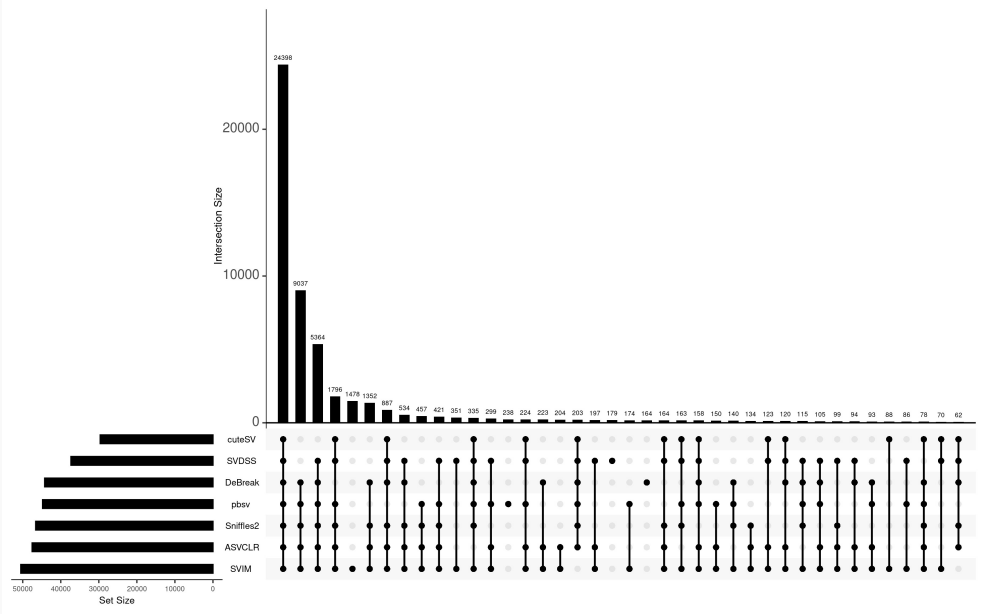


Figure 2 UpSet plot of TP across multiple callsets

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	252	340	614	3529	38593	2353	1693	1358
SVDSS	324	383	757	3961	31069	2779	1956	1586
DeBreak	205	309	543	3115	35754	2228	1372	980
Sniffles2	255	359	617	2558	37458	2211	1496	1205
pbsv	283	386	709	4977	32678	2244	1543	1174
cuteSV	266	351	741	4198	20348	2125	1510	1229
SVIM	504	592	987	4869	38848	3315	2520	2107

(2) Deviation of the variant size ratio

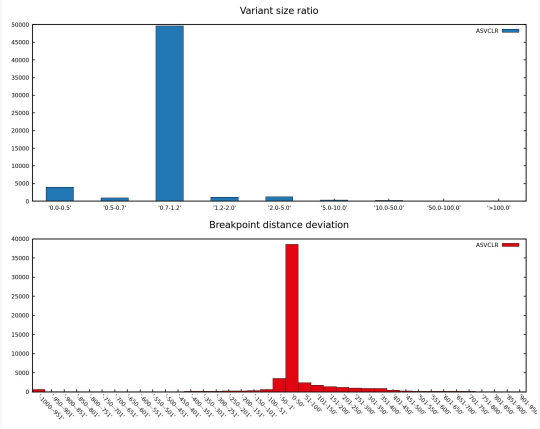
Calculating the variant size ratio for two overlapping variations based on the length of SVs, the closer the ratio is to 1, the smaller the deviation, indicating a more precise and accurate identification result. Statistics results for nine size regions are presented in Table 3:

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

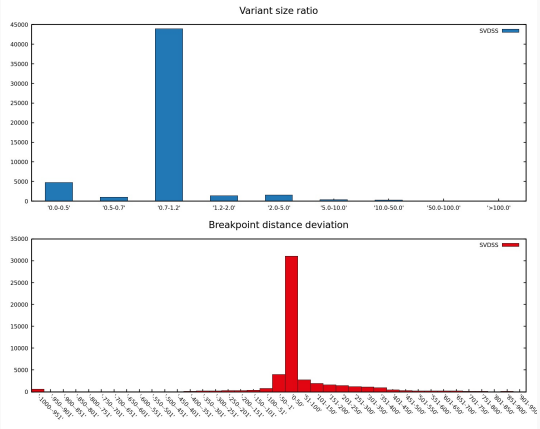
Tool	0.0-0.5	0.5-0.7	0.7-1.2	1.2-2.0	2.0-5.0	5.0-10.0	10.0-50.0	50.0-100.0	>100.0
ASVCLR	3934	890	49680	1159	1273	291	204	21	18
SVDSS	4802	1001	43913	1428	1581	366	268	28	16
DeBreak	2725	631	41472	2909	1991	526	380	38	31
Sniffles2	4090	839	47058	1170	1270	288	215	21	25
pbsv	4771	914	43436	1590	1483	364	265	36	35
cuteSV	4260	631	31921	1250	1289	349	294	47	123
SVIM	10225	1334	54310	1674	1785	448	342	52	80

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

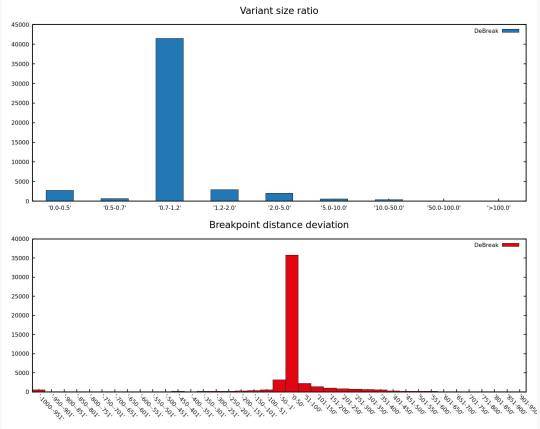
(a) ASVCLR:



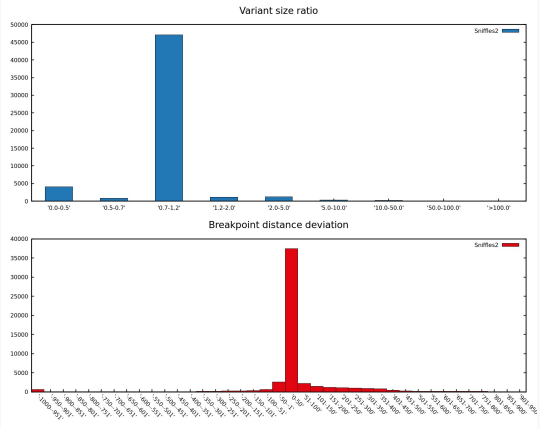
(b) SVDSS:



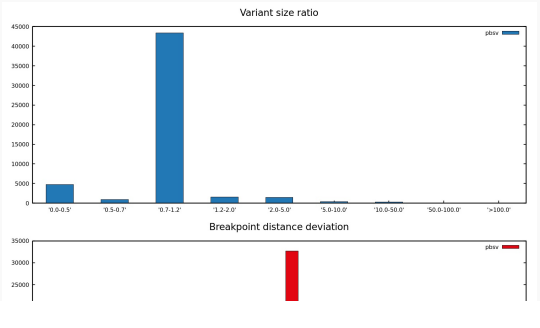
(c) DeBreak:



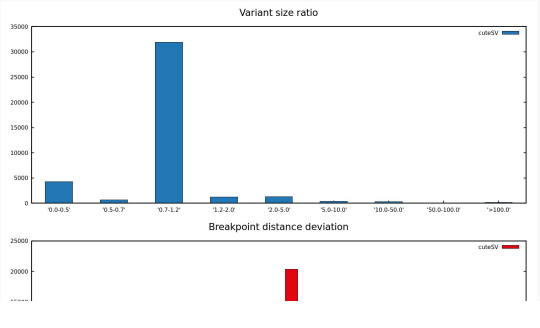
(d) Sniffles2:

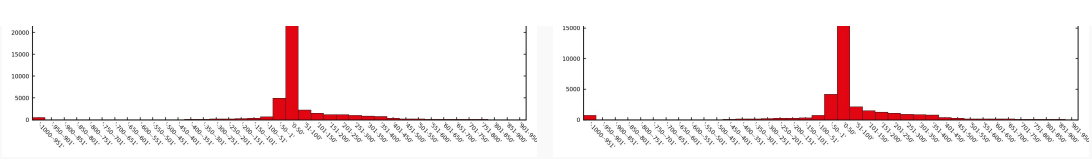


(e) pbsv:



(f) cuteSV:





(g) SVIM:

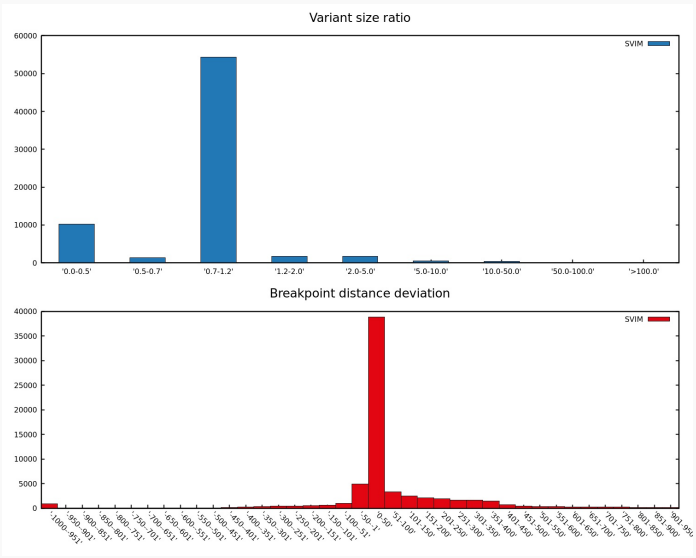


Figure 3 Deviation statistics with overlapping variations

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:

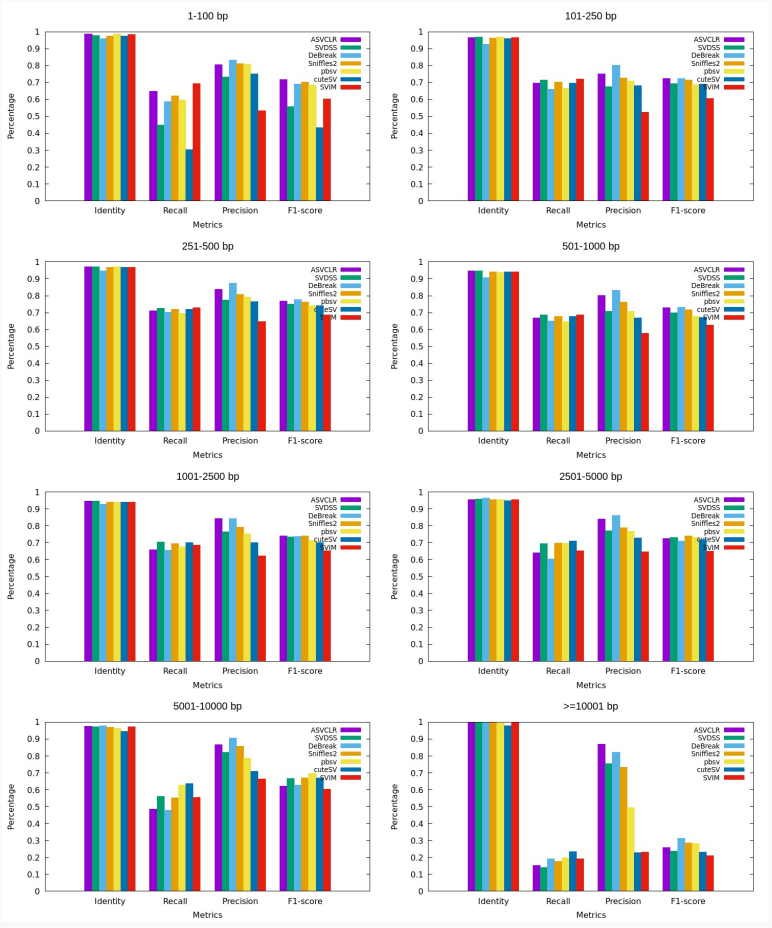


Figure 4 Statistics of metrics of different SV size region

(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	35069	34818	8414	19152	0.986515	0.646779	0.806499	0.717862
101-250bp	5198	4843	1716	2275	0.967097	0.695571	0.751808	0.722597
251-500bp	4413	4266	858	1784	0.971029	0.712119	0.837223	0.769620
501-1000bp	1682	1592	420	832	0.947247	0.669053	0.800190	0.728769
1001-2500bp	1346	1180	250	698	0.945650	0.658513	0.843358	0.739560
2501-5000bp	525	475	101	296	0.956150	0.639464	0.838658	0.725639
5001-10000bp	231	214	36	246	0.975358	0.484277	0.865169	0.620968
>10000bp	40	41	6	225	0.996031	0.150943	0.869565	0.257235

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

(a) (b)

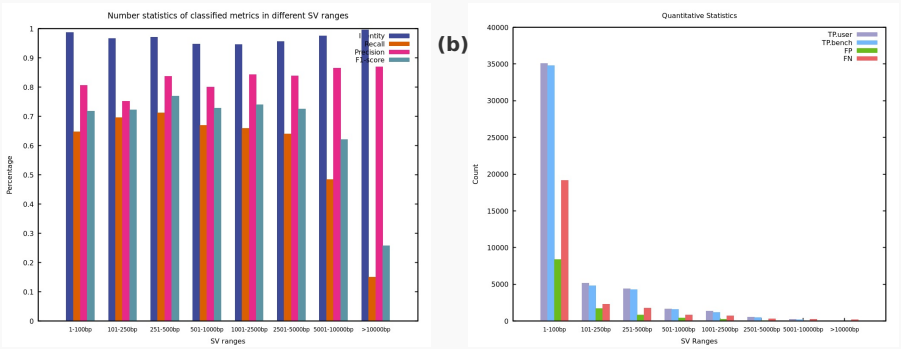


Figure 5 Result statistics of different SV size ranges

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	24269	26318	8860	29952	0.978548	0.447594	0.732561	0.555673
101-250bp	5350	5522	2588	2123	0.967806	0.715911	0.673973	0.694309
251-500bp	4490	4724	1300	1707	0.970847	0.724544	0.775475	0.749145
501-1000bp	1729	1718	712	785	0.946364	0.687749	0.708316	0.697881
1001-2500bp	1441	1330	442	603	0.945414	0.704990	0.765268	0.733894

2501-5000bp	571	536	170	250	0.958087	0.695493	0.770580	0.731114
5001-10000bp	267	257	58	210	0.971830	0.559748	0.821538	0.665835
>10000bp	37	34	12	228	0.996059	0.139623	0.755102	0.235669

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

(a) (b)

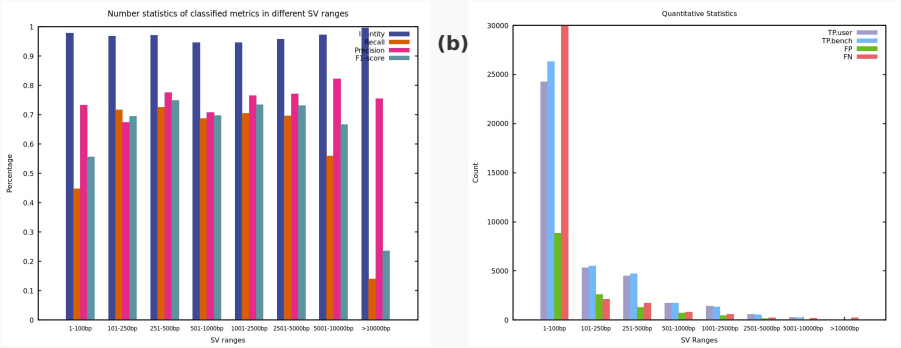


Figure 6 Result statistics of different SV size ranges

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	31879	30318	6355	22342	0.959475	0.587946	0.833787	0.689611
101-250bp	4935	4366	1212	2538	0.925625	0.660377	0.802831	0.724670
251-500bp	4343	4196	624	1854	0.946072	0.700823	0.874371	0.778037
501-1000bp	1638	1536	329	876	0.906491	0.651551	0.832740	0.731087
1001-2500bp	1338	1246	251	706	0.927120	0.654599	0.842039	0.736581
2501-5000bp	495	492	80	326	0.965038	0.602923	0.860870	0.709169
5001-10000bp	229	224	24	248	0.979235	0.480084	0.905138	0.627397
>10000bp	51	51	11	214	1.000000	0.192453	0.822581	0.311927

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

(a) (b)

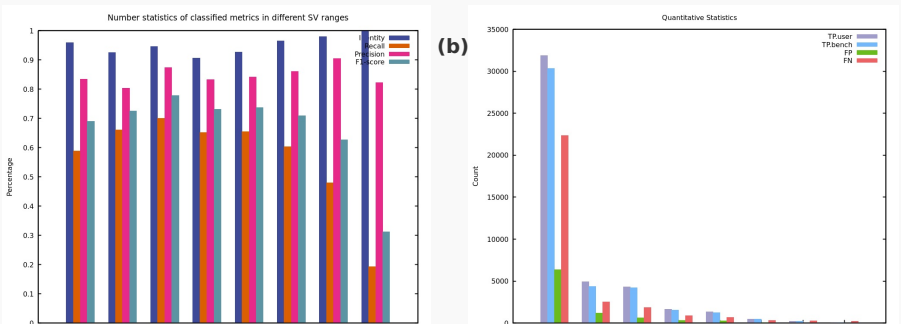


Figure 7 Result statistics of different SV size ranges

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	33643	31910	7884	20578	0.973524	0.620479	0.810148	0.702740
101-250bp	5248	4740	1969	2225	0.962269	0.702261	0.727172	0.714500
251-500bp	4464	4233	1059	1733	0.967578	0.720349	0.808256	0.761775
501-1000bp	1701	1528	535	813	0.941136	0.676611	0.760733	0.716211
1001-2500bp	1422	1234	372	622	0.940630	0.695695	0.792642	0.741011
2501-5000bp	574	514	155	247	0.954654	0.699147	0.787380	0.740645
5001-10000bp	263	240	44	214	0.970526	0.551363	0.856678	0.670918
>10000bp	47	44	17	218	1.000000	0.177358	0.734375	0.285714

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

(a) (b)

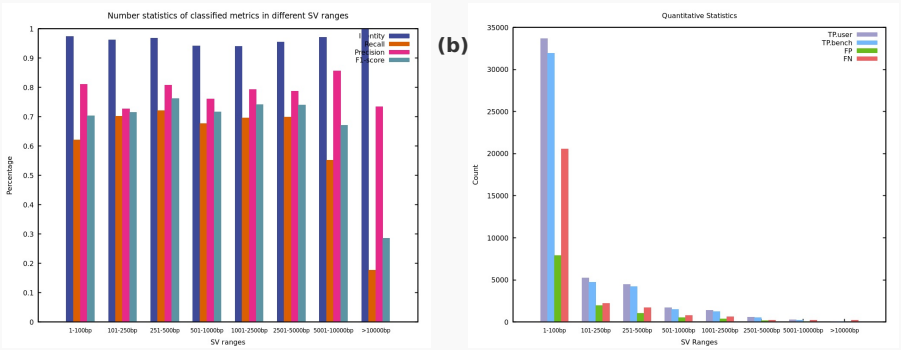


Figure 8 Result statistics of different SV size ranges

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	32294	30427	7671	21927	0.987942	0.595599	0.808057	0.685749
101-250bp	4977	4439	2034	2496	0.967948	0.665998	0.709884	0.687241
251-500bp	4302	4109	1130	1895	0.972192	0.694207	0.791973	0.739874
501-								

1000bp	1628	1485	674	886	0.939327	0.647574	0.707211	0.676080
1001-2500bp	1383	1210	458	661	0.940671	0.676614	0.751222	0.711969
2501-5000bp	572	510	174	249	0.953735	0.696711	0.766756	0.730057
5001-10000bp	299	272	81	178	0.963078	0.626834	0.786842	0.697783
>10000bp	52	52	53	213	0.993388	0.196226	0.495238	0.281081

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

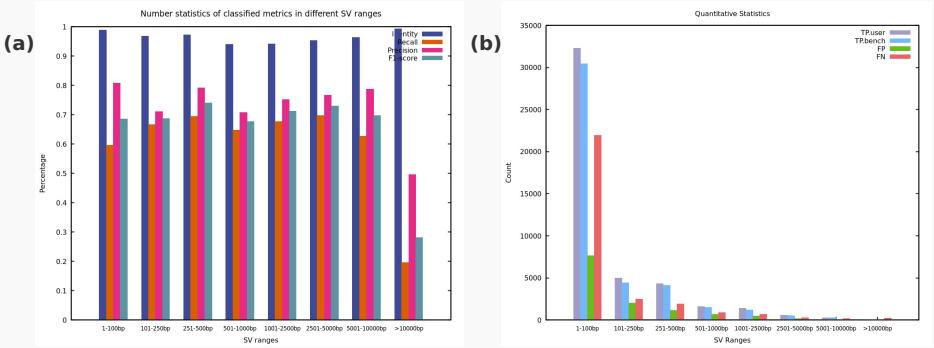


Figure 9 Result statistics of different SV size ranges

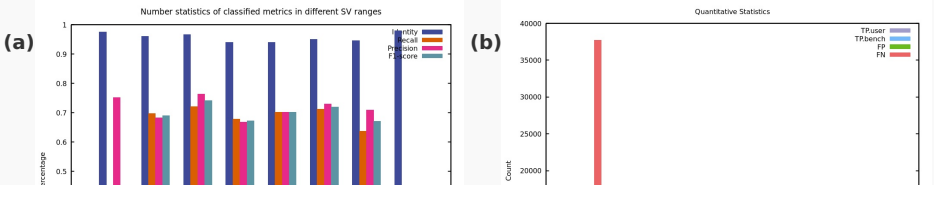
Figure (a) shows the statistical results of Recall, Precision, F1 score and Identity; (b) shows the statistical results of #TP_benchmark, #TP_user,#FP and #FN.

(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	16488	15335	5454	37733	0.975279	0.304089	0.751436	0.432966
101-250bp	5216	4694	2433	2257	0.960889	0.697979	0.681919	0.689856
251-500bp	4470	4246	1381	1727	0.967089	0.721317	0.763972	0.742032
501-1000bp	1704	1558	846	810	0.939824	0.677804	0.668235	0.672986
1001-2500bp	1433	1253	610	611	0.940033	0.701076	0.701419	0.701248
2501-5000bp	584	529	217	237	0.949925	0.711328	0.729089	0.720099
5001-10000bp	304	278	125	173	0.946387	0.637317	0.708625	0.671082
>10000bp	62	62	211	203	0.980085	0.233962	0.227106	0.230483

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



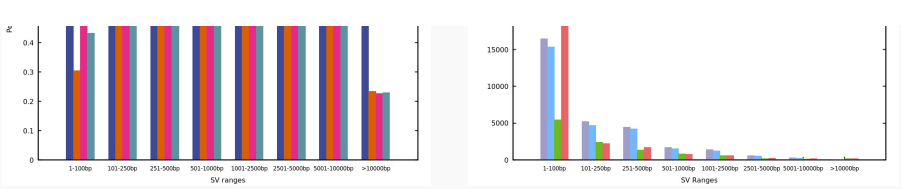


Figure 10 Result statistics of different SV size ranges

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.

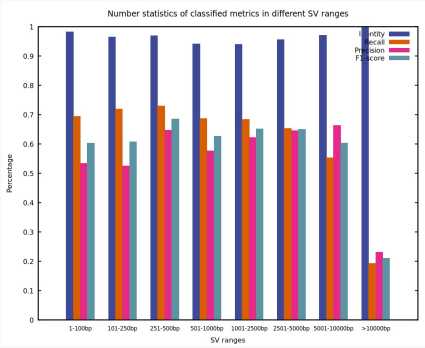
(8) The user-called set (SVIM) of basic metrics results statistics

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	37629	35946	32791	16592	0.982897	0.693993	0.534351	0.603798
101-250bp	5379	4843	4870	2094	0.964799	0.719791	0.524832	0.607042
251-500bp	4523	4275	2468	1674	0.969269	0.729869	0.646975	0.685927
501-1000bp	1729	1540	1269	785	0.941789	0.687749	0.576718	0.627358
1001-2500bp	1399	1201	851	645	0.940513	0.684442	0.621778	0.651607
2501-5000bp	536	472	293	285	0.955922	0.652862	0.646562	0.649697
5001-10000bp	264	243	134	213	0.970924	0.553459	0.663317	0.603429
>10000bp	51	49	169	214	1.000000	0.192453	0.231818	0.210309

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

(a)



(b)

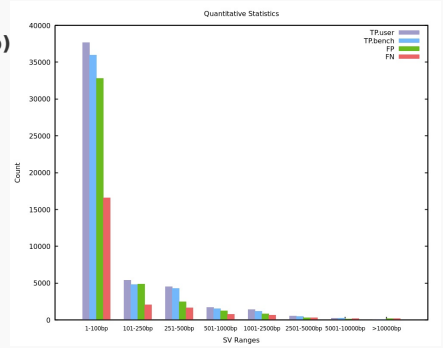


Figure 11 Result statistics of different SV size ranges

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

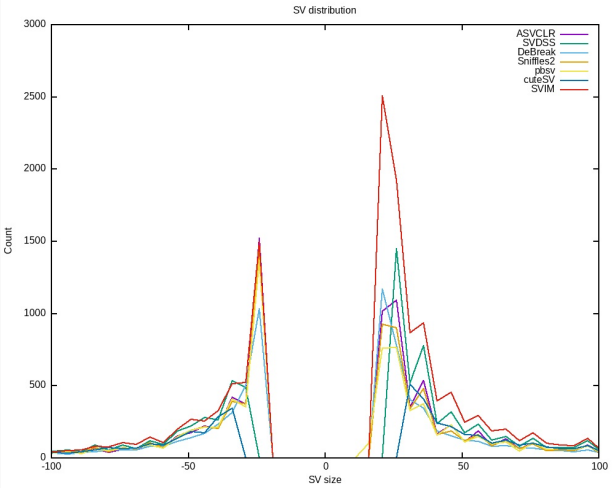


Figure 12 SV distribution for mutiple user callsets

(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

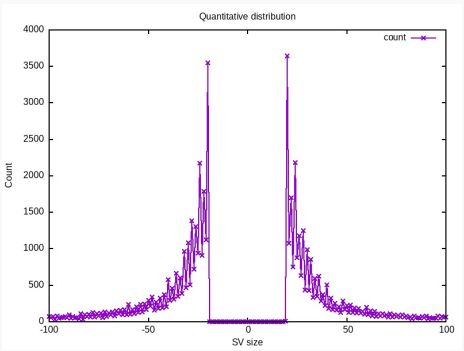


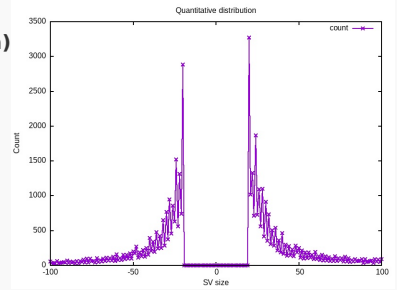
Figure 13 The quantity distribution of the benchmark set

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 numberï¼š60029
The SV reference region size statistics after filtering for user-called set (ASVCLR):Total SVs numberï¼š60029

(a)



(b)

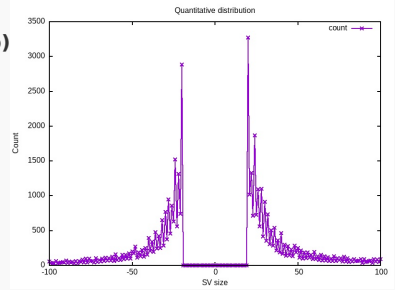


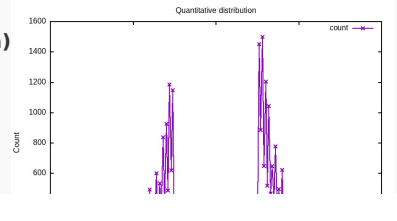
Figure 14 The quantity distribution of the user-called set

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.

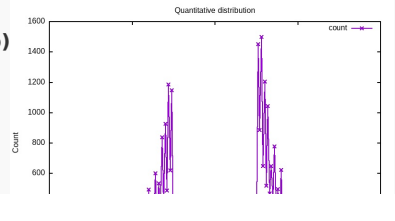
(3) Statistics of the count of different SV lengths in the user-called set (SVDSS):

The SV reference region size statistics before filtering for user-called set (SVDSS):Total SVs numberï¼š55876
The SV reference region size statistics after filtering for user-called set (SVDSS):Total SVs numberï¼š55876

(a)



(b)



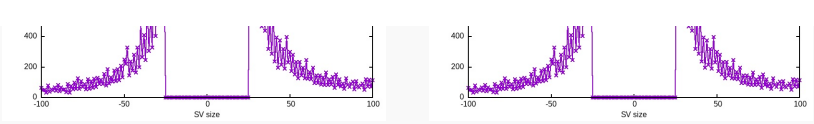


Figure 15 The quantity distribution of the user-called set

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 numberï¼š51914
The SV reference region size statistics after filtering for user-called set (DeBreak):Total SVs numberï¼š51878

(a) **(b)**

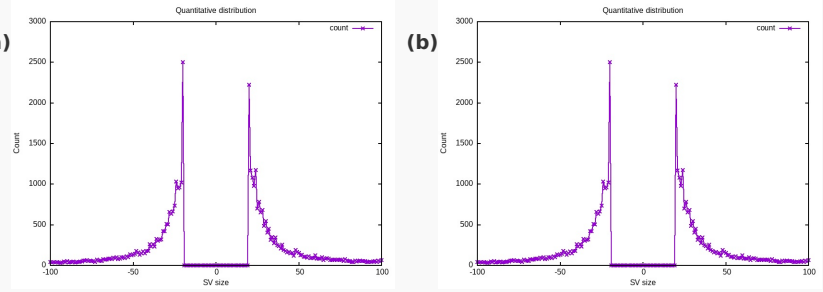


Figure 16 The quantity distribution of the user-called set

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 numberï¼š57555
The SV reference region size statistics after filtering for user-called set (Sniffles2):Total SVs numberï¼š57532

(a) **(b)**

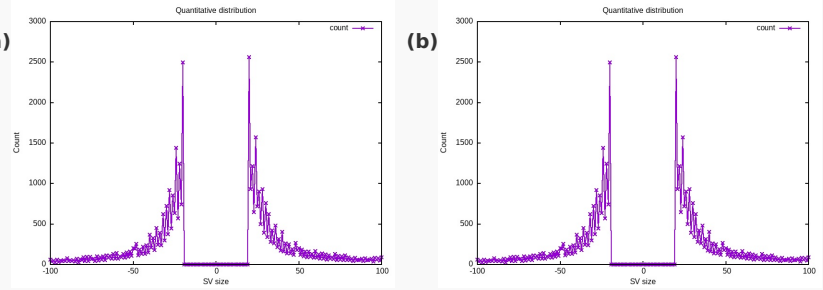


Figure 17 The quantity distribution of the user-called set

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 numberï¼š55378
The SV reference region size statistics after filtering for user-called set (pbsv):Total SVs numberï¼š55361

(a) **(b)**

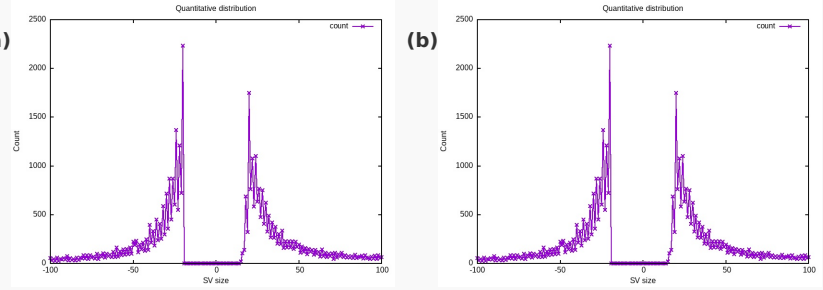


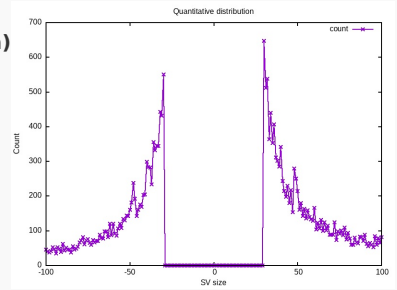
Figure 18 The quantity distribution of the user-called set

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 numberï¼š42286
The SV reference region size statistics after filtering for user-called set (cuteSV):Total SVs numberï¼š42257

(a)



(b)

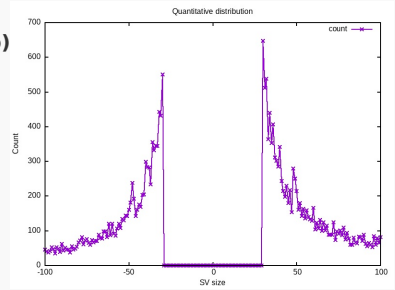


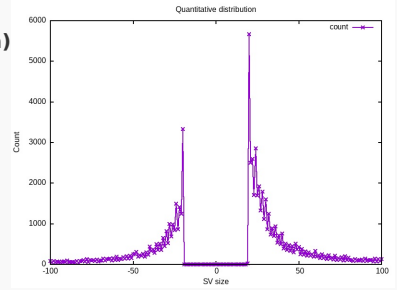
Figure 19 The quantity distribution of the user-called set

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 numberï¼š124008
The SV reference region size statistics after filtering for user-called set (SVIM):Total SVs numberï¼š123950

(a)



(b)

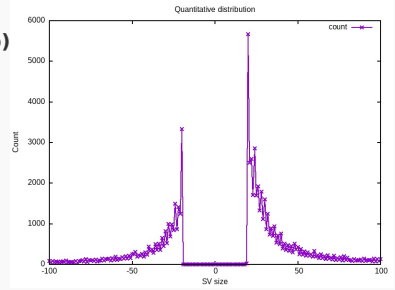


Figure 20 The quantity distribution of the user-called set

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. ----------