Evaluation Reports

```
SV_STAT command:
$ ./sv_stat -m 50000 -T asvclr;cuteSV;pbsv;Sniffles2;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_test/hg002_hg19_variants_asvclr_1.4.0_20240220.vcf
./data_test/output_cuteSV.vcf ./data_test/output_pbsv.vcf
./data_test/output_Sniffles.vcf ./data_test/output_svim_sorted.vcf
./data_test/HG002_SVs_Tier1_v0.6.vcf ./data_test/hs37d5.fa -o
HG19_all_exp
```

1. Benchmarking results

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

The evaluation 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 consistency (Seqcons) 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 assessment results for the variation detection methods.

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

Table 1 Structural Variation Detection Method Performance Evaluation

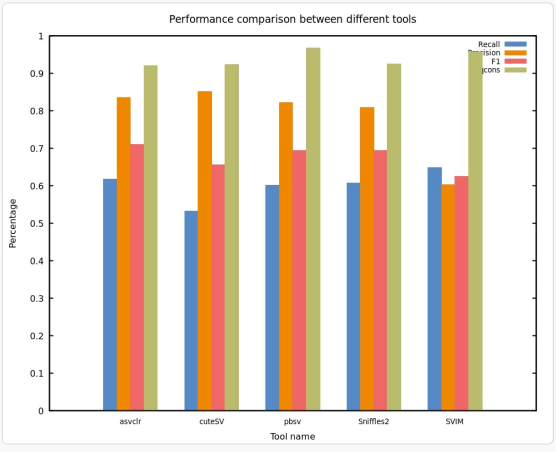
Tool	#SVs_bench	#SVs_user	#SVs_filtered_user	#TP_bench	#TP_user	#FP	#FN	Recall	Precision	F1 score	Seqcons
asvclr	74012	52857	52857	45694	44180	8677	28318	0.617386	0.835840	0.710194	0.920133
cuteSV	74012	44937	44928	39442	36955	6413	34570	0.532914	0.852126	0.655735	0.923026
pbsv	74012	52807	52741	44494	42927	9253	29518	0.601173	0.822672	0.694694	0.967706
Sniffles2	74012	54545	54458	44983	43114	10160	29029	0.607780	0.809288	0.694207	0.924712
SVIM	74012	116615	116427	48028	47230	30995	25984	0.648922	0.603771	0.625533	0.958157

The table 1 shows the evaluation 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. Seqcons represents the sequence consistency, which refers to calculating the sequence consistency score for matched SV pairs that include sequences.

(2) The evaluation 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 Seqcons; (b) the other category consists of #TP_bench, #TP_user, #FP, and #FN. The result statistics are as follows:

(a)



(b)

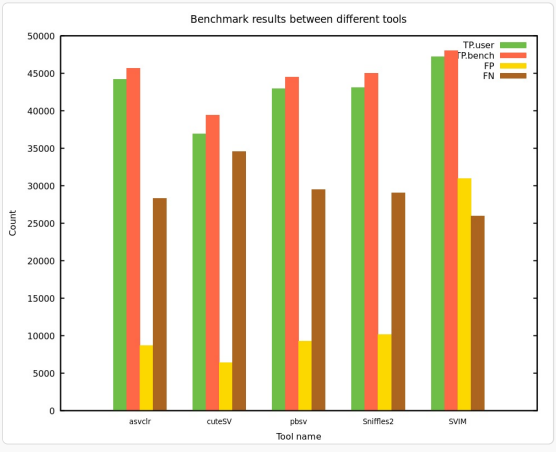


Figure 1 Evaluation results of the user-call set

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 center distance and the region size ratio of the overlapping variations.

(1) Deviation of the center distance

As the center 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 center distance deviation

Tool	-200--151	-150--101	-100--51	-50--1	0-50	51-100	101-150	151-200
asvclr	248	402	725	5682	29455	4281	2144	1468
cuteSV	203	345	696	5255	25892	3471	1689	1178
pbsv	358	501	795	5123	31030	2546	1756	1382
Sniffles2	355	469	781	3732	31630	4556	2043	1462
SVIM	602	809	1283	6325	33782	4056	2535	2007

(2) Deviation of the region size ratio

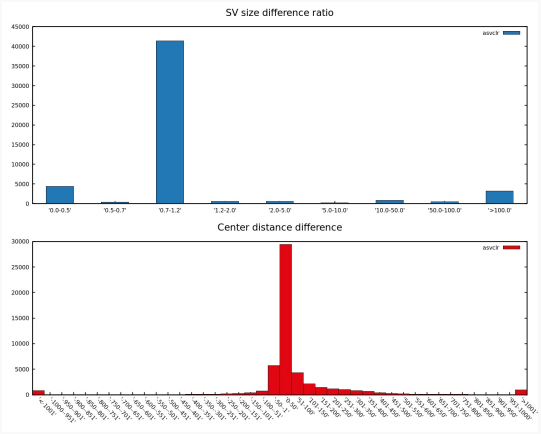
Calculating the region size ratio for two overlapping variations based on the length of SVs, the closer the ratio is to 1, the smaller the deviation, indicating a more precise and accurate identification result. Statistics results for nine size regions are presented in Table 3:

Table 3 Statistical results of deviation of the region size ratio

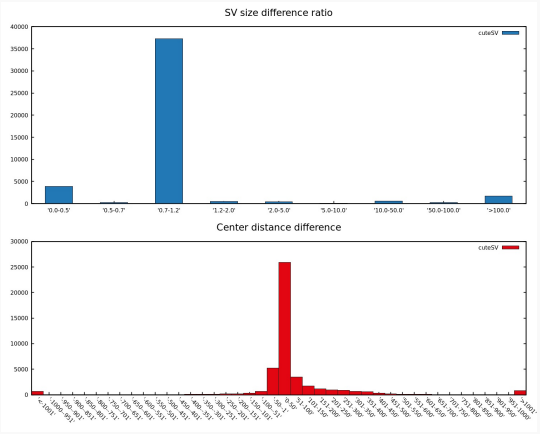
Tool	0.0-0.5	0.5-0.7	0.7-1.2	1.2-2.0	2.0-5.0	5.0-10.0	10.0-50.0	50.0-100.0	>100.0
asvclr	4373	398	41354	612	584	168	857	460	3178
cuteSV	3883	290	37294	500	417	81	553	255	1738
pbsv	4499	458	36190	721	715	175	4452	1329	3498
Sniffles2	5184	442	43846	604	509	133	633	254	1701
SVIM	8441	574	46891	691	750	265	1261	704	4842

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

(a) asvclr:

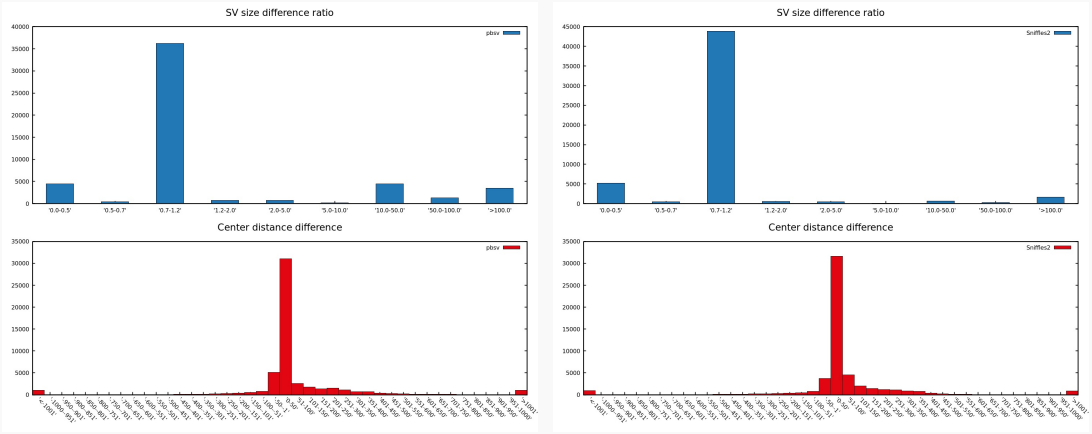


(b) cuteSV:



(c) pbsv:

(d) Sniffles2:



(e) SVIM:

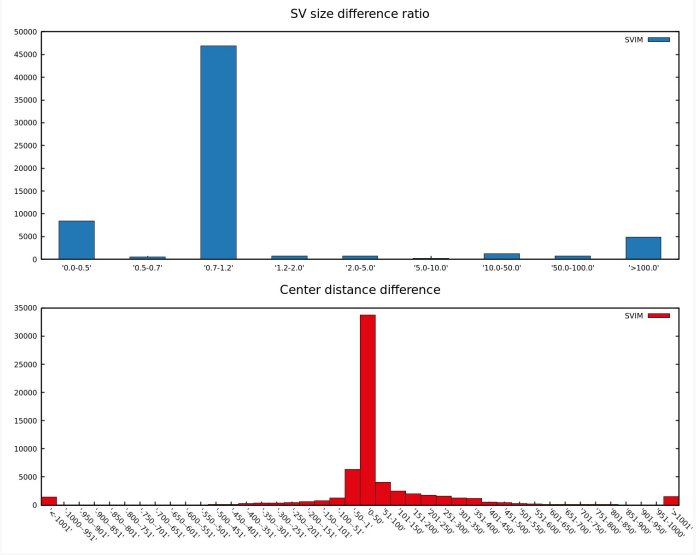


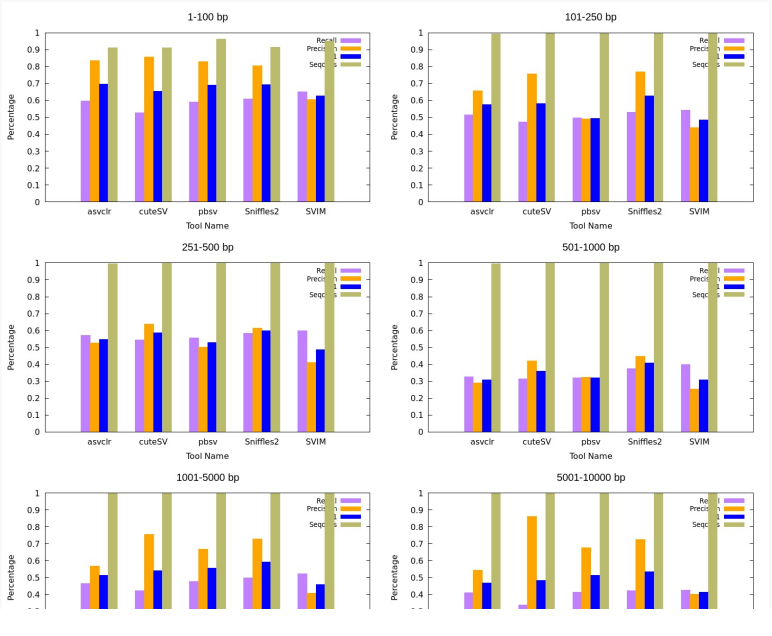
Figure 2 Deviation statistics with overlapping variations

3. Evaluation 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 evaluation results are presented in the table as followsï¼š

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

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



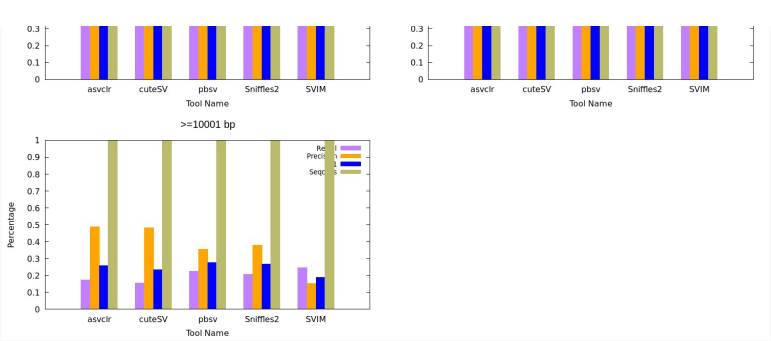


Figure 3 Statistics of metrics of different SV size region

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

Table 4 The metric evaluation results of asvclr in different SV regions

Region	#TP_bench	#TP_user	#FP	#FN	Recall	Precision	F1 score	Seqcons
1-100bp	38240	36972	7288	25904	0.596159	0.835337	0.695766	0.911476
101-250bp	2032	1870	972	1926	0.513391	0.657987	0.576764	0.992708
251-500bp	1701	1662	1504	1271	0.572342	0.524953	0.547624	0.995176
501-1000bp	364	358	870	748	0.327338	0.291531	0.308399	0.996147
1001-5000bp	577	578	439	661	0.466074	0.568338	0.512151	0.999387
5001-10000bp	138	138	116	198	0.410714	0.543307	0.467797	1.000000
>10000bp	44	44	46	208	0.174603	0.488889	0.257310	1.000000

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

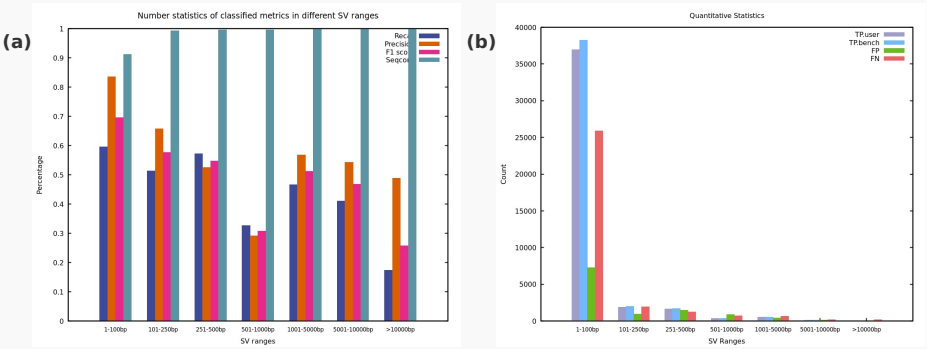


Figure 4 Result statistics of different SV size region

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

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

Table 5 The metric evaluation results of cuteSV in different SV regions

Region	#TP_bench	#TP_user	#FP	#FN	Recall	Precision	F1 score	Seqcons
1-100bp	33850	31715	5248	30294	0.527719	0.858020	0.653505	0.912349
101-250bp	1866	1686	546	2092	0.471450	0.755376	0.580559	1.000000
251-500bp	1618	1577	891	1354	0.544415	0.638979	0.587918	1.000000
501-1000bp	349	341	472	763	0.313849	0.419434	0.359040	1.000000

1001-5000bp	522	512	166	716	0.421648	0.755162	0.541145	1.000000
5001-10000bp	113	113	18	223	0.336310	0.862595	0.483940	1.000000
>10000bp	39	40	43	213	0.154762	0.481928	0.234287	1.000000

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

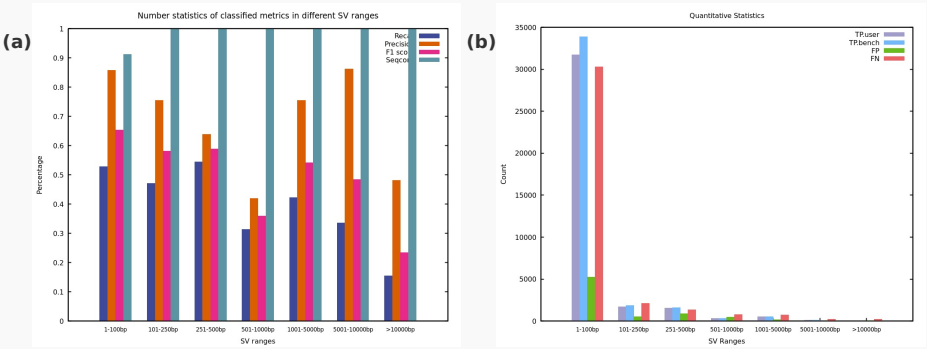


Figure 5 Result statistics of different SV size region

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

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

Table 6 The metric evaluation results of pbsv in different SV regions

Region	#TP_bench	#TP_user	#FP	#FN	Recall	Precision	F1 score	Seqcons
1-100bp	37819	35709	7265	26325	0.589595	0.830944	0.689767	0.962946
101-250bp	1971	1794	1853	1987	0.497979	0.491911	0.494926	1.000000
251-500bp	1656	1620	1607	1316	0.557201	0.502014	0.528170	1.000000
501-1000bp	355	351	738	757	0.319245	0.322314	0.320772	1.000000
1001-5000bp	591	582	288	647	0.477383	0.668966	0.557165	1.000000
5001-10000bp	139	140	67	197	0.413690	0.676328	0.513368	1.000000
>10000bp	57	59	107	195	0.226190	0.355422	0.276449	1.000000

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

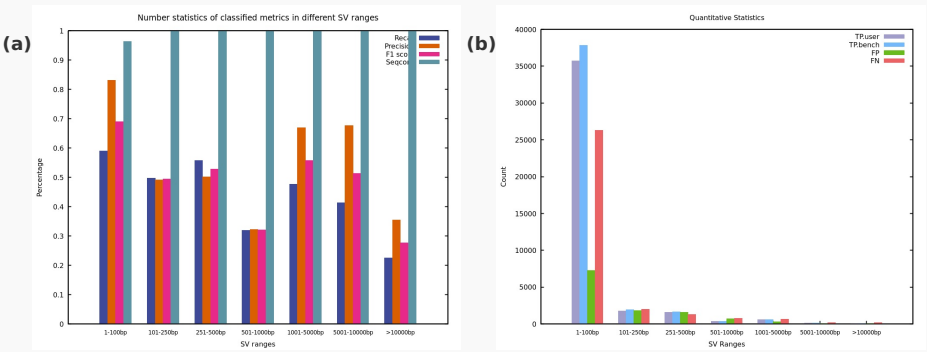


Figure 6 Result statistics of different SV size region

Figure (a) shows the statistical results of Recall, Precision, F1 score and Seqcons;
(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 evaluation results of Sniffles2 in different SV regions

Region	#TP_bench	#TP_user	#FP	#FN	Recall	Precision	F1 score	Seqcons
1-100bp	38996	36911	8905	25148	0.607945	0.805636	0.692966	0.914317
101-250bp	2098	1936	582	1860	0.530066	0.768864	0.627514	1.000000
251-500bp	1737	1705	1066	1235	0.584455	0.615301	0.599482	1.000000
501-1000bp	416	419	518	696	0.374101	0.447172	0.407386	1.000000
1001-5000bp	615	636	238	623	0.496769	0.727689	0.590454	1.000000
5001-10000bp	142	152	58	194	0.422619	0.723810	0.533650	1.000000
>10000bp	52	56	92	200	0.206349	0.378378	0.267058	1.000000

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

(a) (b)

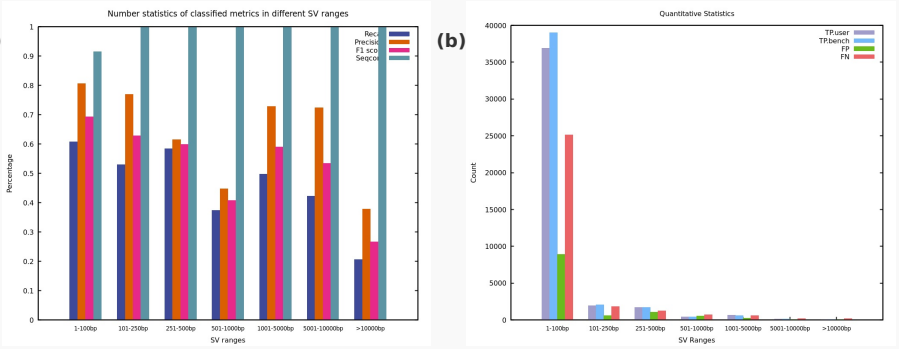


Figure 7 Result statistics of different SV size region

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

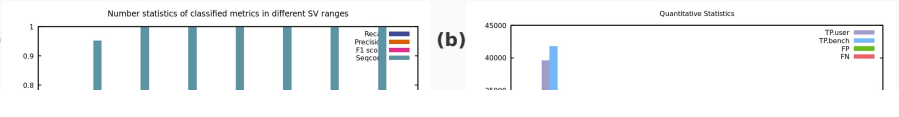
(6) The user-called set (SVIM) of basic metrics results statistics

Table 8 The metric evaluation results of SVIM in different SV regions

Region	#TP_bench	#TP_user	#FP	#FN	Recall	Precision	F1 score	Seqcons
1-100bp	41774	39594	25735	22370	0.651253	0.606071	0.627850	0.952254
101-250bp	2141	1960	2500	1817	0.540930	0.439462	0.484945	1.000000
251-500bp	1776	1747	2502	1196	0.597577	0.411156	0.487140	1.000000
501-1000bp	442	442	1308	670	0.397482	0.252571	0.308875	1.000000
1001-5000bp	645	656	953	593	0.521002	0.407707	0.457444	1.000000
5001-10000bp	143	146	218	193	0.425595	0.401099	0.412984	1.000000
>10000bp	62	71	393	190	0.246032	0.153017	0.188684	1.000000

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

(a) (b)



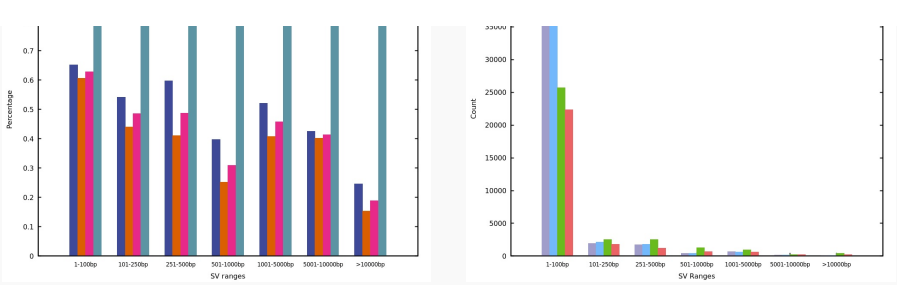


Figure 8 Result statistics of different SV size region

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

4. Quantitative statistics

(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

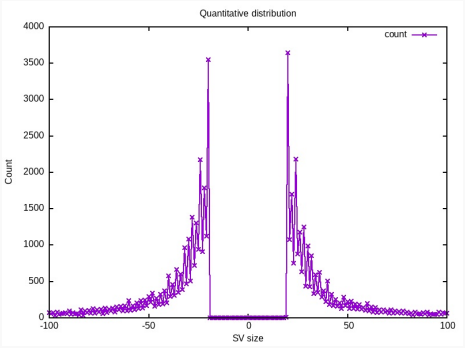


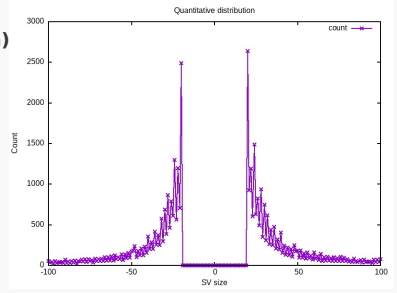
Figure 9 The quantity distribution of the benchmark set

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 numberï¼š52857
The SV reference region size statistics after filtering for user-called set (asvclr):Total SVs numberï¼š52857

(a)



(b)

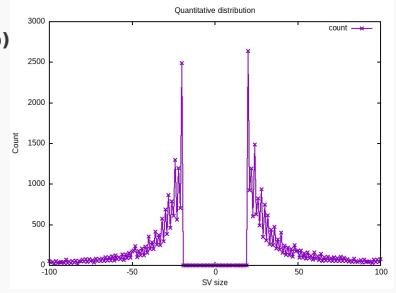


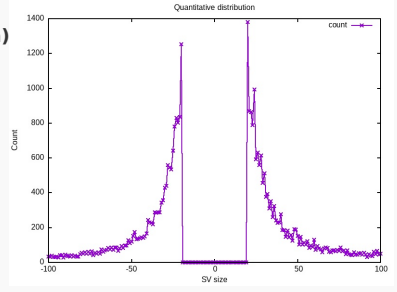
Figure 10 The quantity distribution of the user-called set

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 (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

(a)



(b)

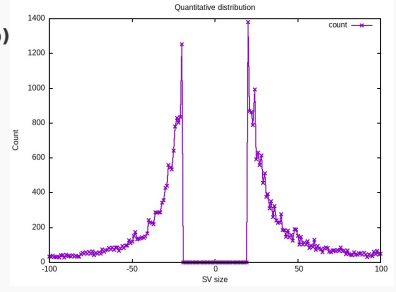


Figure 11 The quantity distribution of the user-called set

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 (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

(a) **(b)**

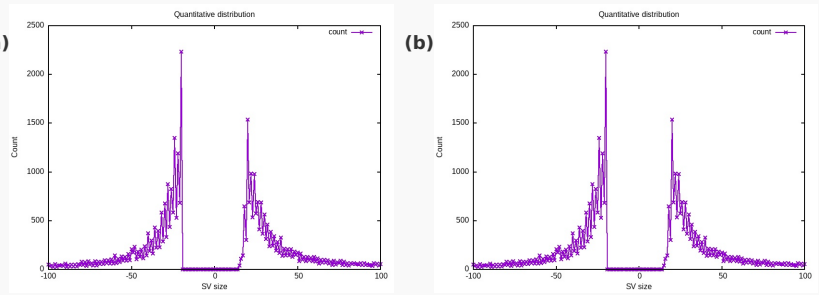


Figure 12 The quantity distribution of the user-called set

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

(a) **(b)**

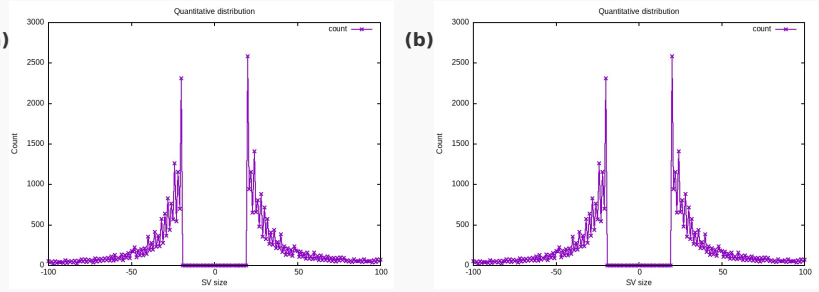


Figure 13 The quantity distribution of the user-called set

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 (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

(a) **(b)**

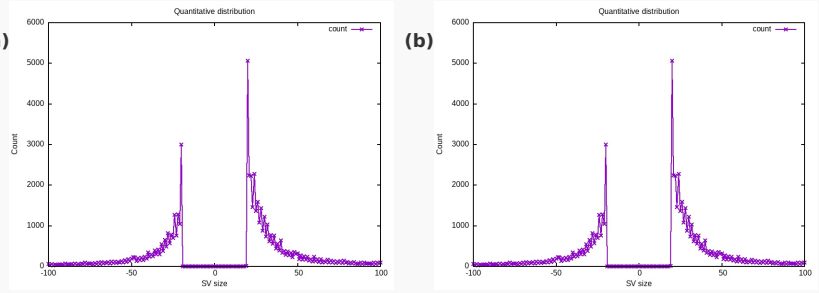


Figure 14 The quantity distribution of the user-called set

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.

More information

- For more detailed evaluation results, please refer to the generated result information in the respective folders.
- For more detailed experiment information, please refer to the github repositories: sv_stat and sv_stat-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 Evaluation Reports. ----------