GIAB TR

Testing Benchmarking Tools

- On chr20 - look at performance of discovery tools on TR regions (v0.3)
  - 32,088 Regions cover 5,685,093bp (8.8%) of chr20
  - No filtering
    - poorly covered by base VCF or ‘complex’ regions remain
    - Not all regions considered by the discovery tools, so we expect low recall
- Use RTG, Truvari bench, Truvari rebench and evaluate if reported matches are correct.

Comp Tool	Params
RTG	–squash-ploidy, –no-roc, –all-records
Truvari bench	–no-ref c, –sizemin 5
Truvari rebench	–use-originals

- Input comparison VCFs were pre-processed to split multi-allelics and normalize “REPL” representations with (`bcftools norm -m-any -f`)

Truvari bench summary

	HipSTR	GangSTR	TRGT
base cnt	4,522	4,522	4,522
call cnt	3,557	1,175	1,306
TP-base	2,063	1,004	1,221
TP-call	2,063	1,004	1,221
FP	1,494	171	85
FN	2,459	3,518	3,301
precision	0.580	0.854	0.935
recall	0.456	0.222	0.270
f1	0.511	0.352	0.419
gt_concordance	0.940	0.940	0.975

RTG Summary

Manually filtered to calls >=5bp

Truvari rebench f1's:

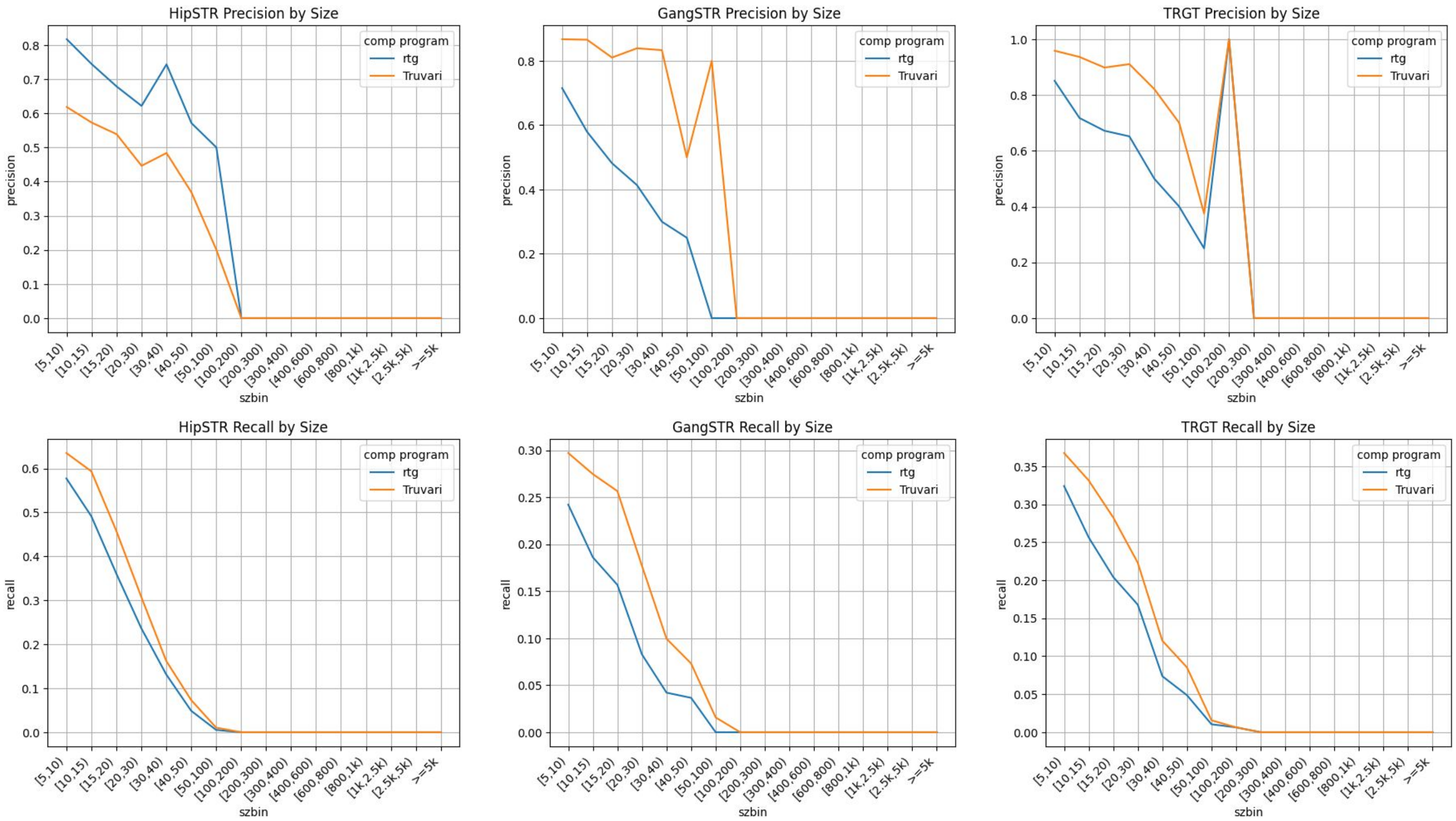
- HipSTR: 0.536
- GangSTR: 0.357
- TRGT: 0.422

1p.p - 10p.p. difference

RTG is finding fewer matches between base/comp.

	HipSTR	GangSTR	TRGT
TP-base	1,784	724	1,004
TP-call	1,887	724	1,005
FN	2,666	3,726	3,446
FP	574	451	301
base count	4,450	4,450	4,450
call count	2,461	1,175	1,306
precision	0.767	0.616	0.770
recall	0.401	0.163	0.226
f1	0.527	0.257	0.349

RTG x Truvari Comparison


HipSTR Precision by Size

GangSTR Precision by Size

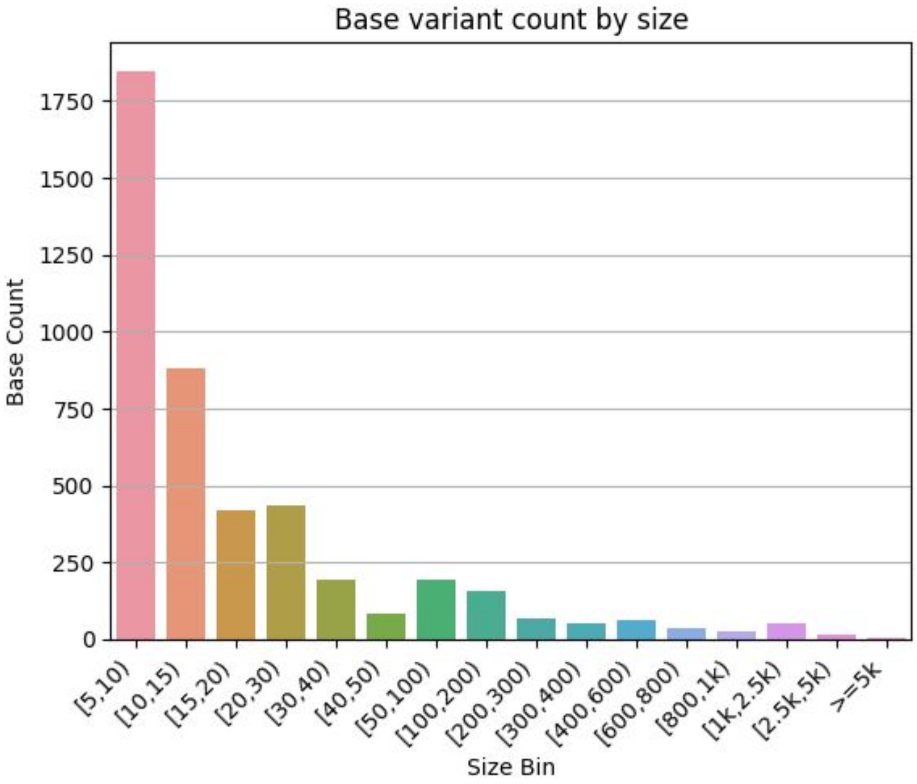
TRGT Precision by Size

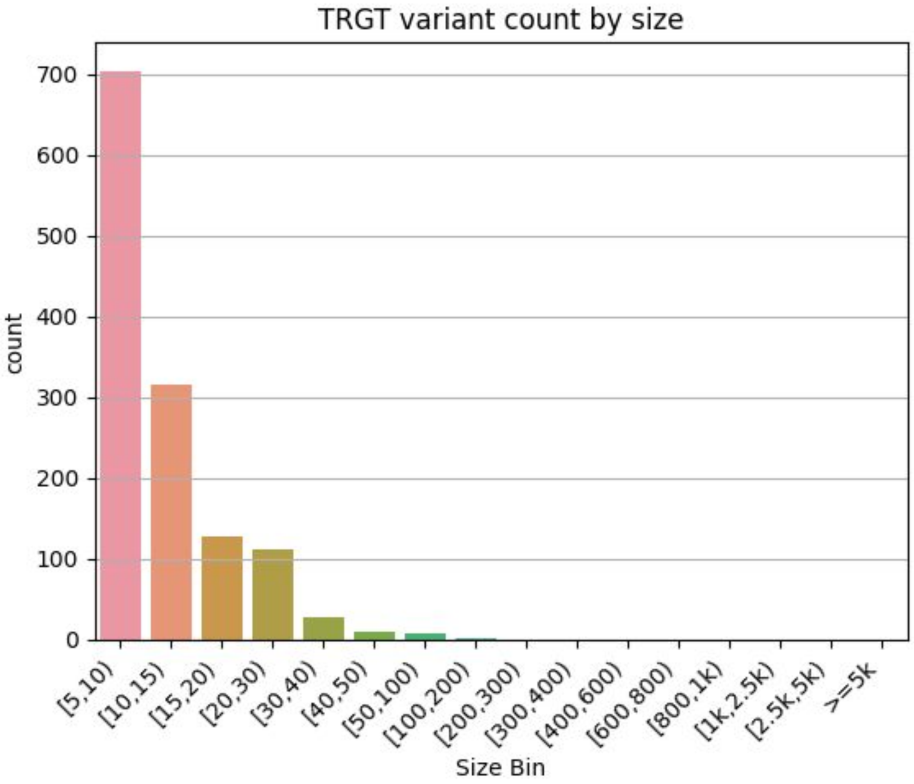
HipSTR Recall by Size

GangSTR Recall by Size

TRGT Recall by Size

No Matches Larger than 300bp (?)





HipSTR Precision Anomaly

	Truvari	RTG (>=5bp)
TP-base	2,063	1,784
TP-call	2,063	1,887
FN	2,459	2,666
FP	1,494	574
Base Count	4,522	4,450
Call Count	3,557	2,461

RTG drops variants due to ‘complex evaluation’.

For HipSTR, 72 base and 1,096 comp are dropped

This amounts to ~30% of comparison calls and seems to be biased against FPs.

Fewer FPs gives an inflated precision.

Intersection of RTG and Truvari for TRGT

Assume RTG is 100% precise in matching variants.

Handful of variants RTG matches which Truvari does not.

6 of the 7 RTG unique TP-calls are recovered by `rebench`.

	Truvari	Shared	RTG (>=5bp)
TP-base	223	998	6
TP-call	223	998	7
FN	81	3,220	226
FP	7	78	223

Why did Truvari bench miss the RTG matches?

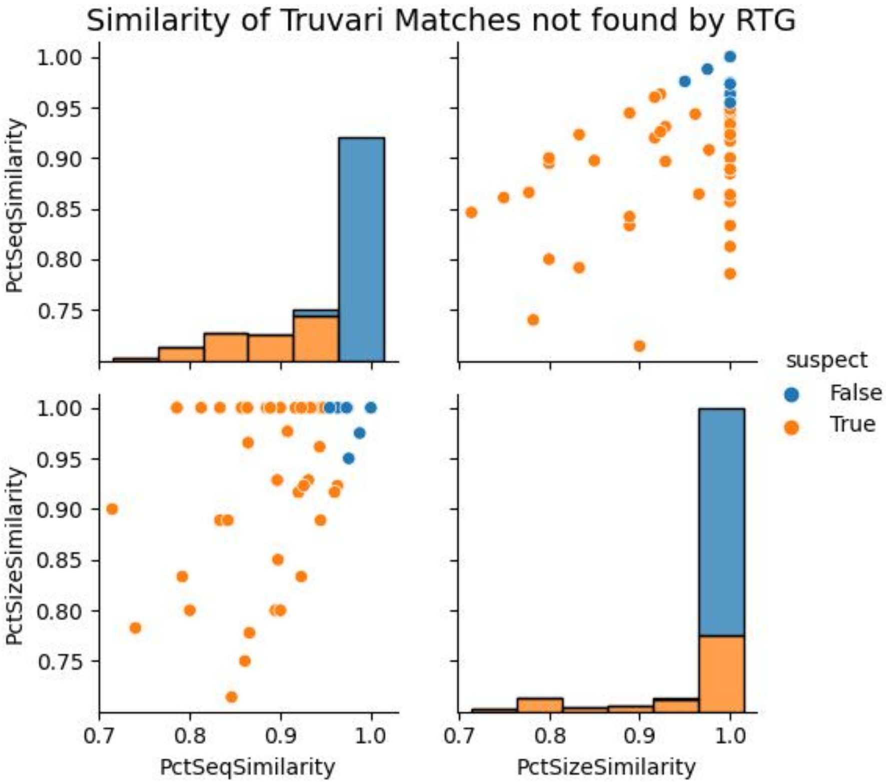
CHROM	POS	REF	ALT	PctSeqSimilarity	PctSizeSimilarity	StartDistance	EndDistance	SizeDiff	PctRecOverlap
chr20	11790786	GTTTA	ATTTATTTGTTTG	0.6818	1	-24	-28	0	0
chr20	1221412	C	CTTTCT	.	.	.	.	.	.
chr20	43083322	ATATATATATATATATATATAT	GTATATATATATATATATACATATATATATATATATATATAC	0.6667	1	-1	-22	0	0.4762
chr20	48610841	TTCCTCTTCCTCTTCCTCTTCCTCTTCCTCTTCCTCTTCT	CTCTTCCTCTTCCTCTTCCTCTTCCTCTTCC	0.4	1	-6	-36	0	0.1
chr20	4986653	CACACACACACACACACACACACACACACACACACACAC	TACACACACACACACACACACACACACACACACACACACACACACACAGAG	0.4545	1	36	0	0	0.2941
chr20	801857	TAAATAAATAAATAAATAAATAAATAAATAAATAAAT	ATAAATAAATAAATAAATAAATAAATAAAC	0.3556	1	-3	-32	0	0.1351
chr20	9078348	CTTTTCTTC	GTTTTCTTCTCTTT	0.6	1	8	0	0	0.3846

What did Truvari find that RTG didn't?

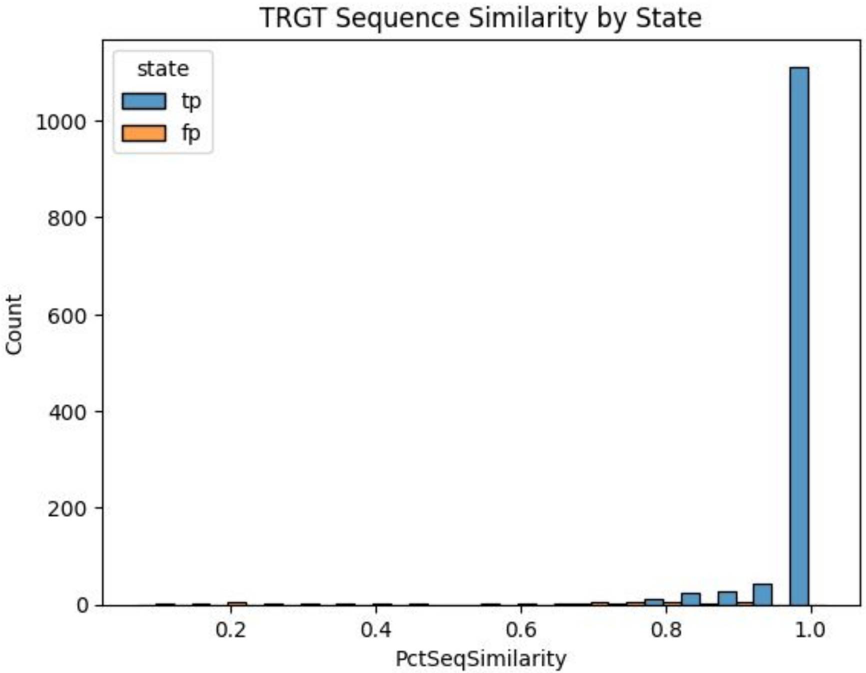
Looking at TRGT calls:

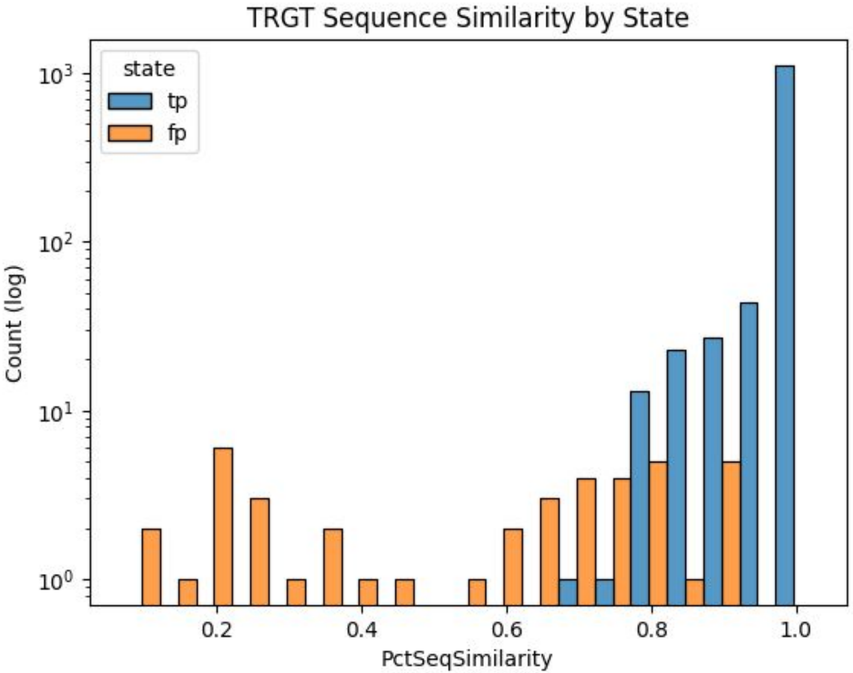
- 223 Truvari unique TP-base
- 148 (66.4%) have sequence and size similarity >= 0.95

Higher similarity thresholds would lower performance of discovery tools

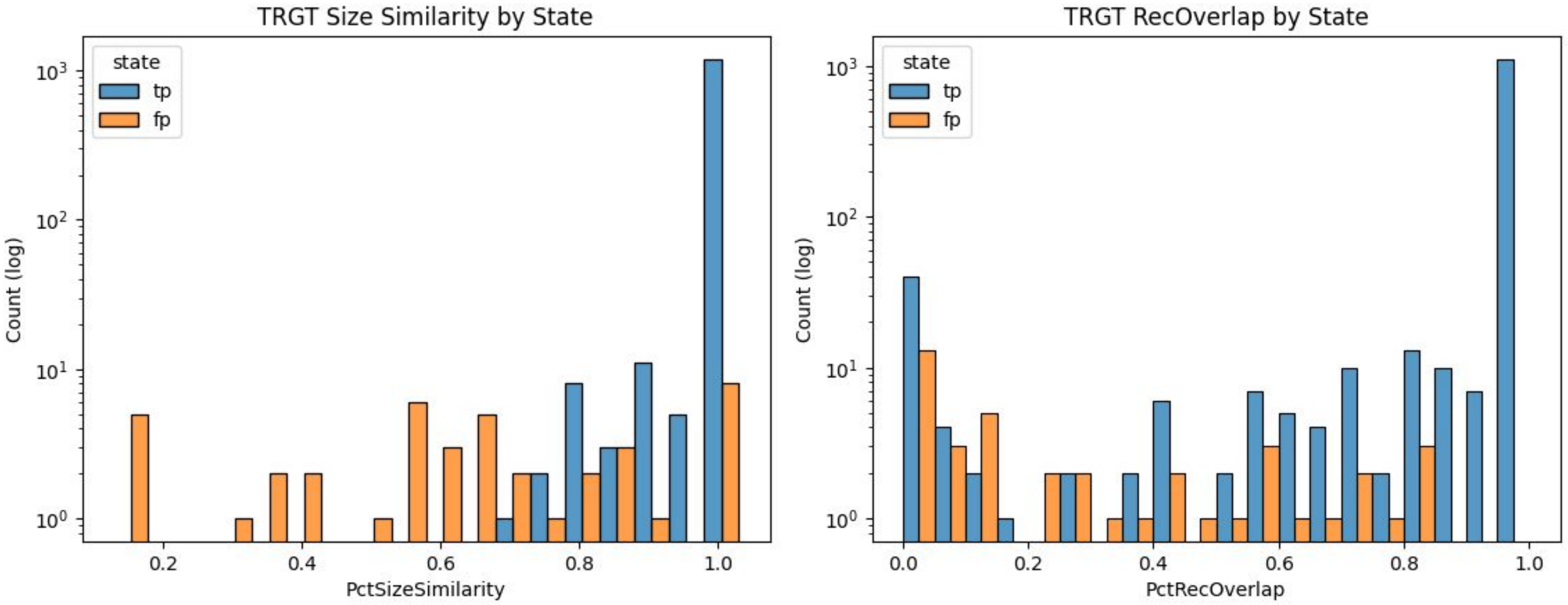


How similar is similar enough?



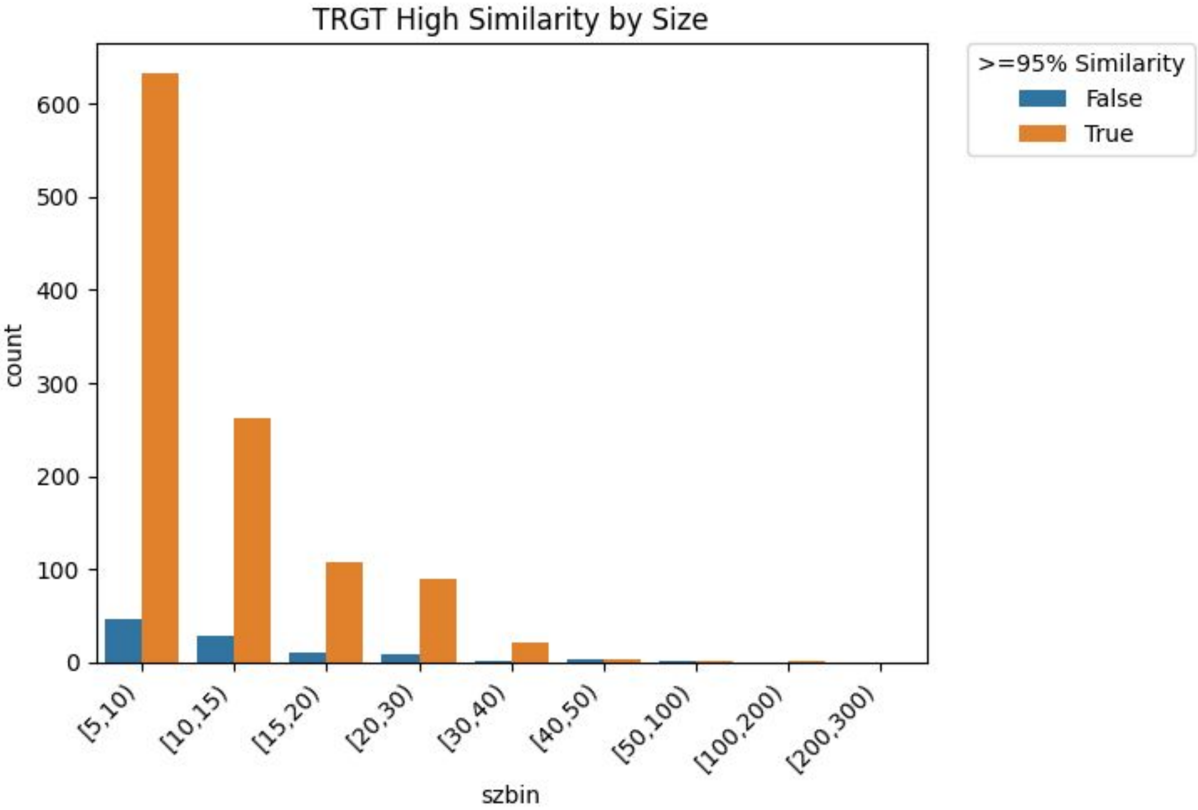


How similar is similar enough?



Similarity by Size

- Doesn’t seem to be a strong pattern with size and similarity.
- Will need more test data with larger variants to make sure.
- Looks like truvari is okay to compare [5-50)bp (with caution)



Truvari rebench

How many regions needed rebench’ing?

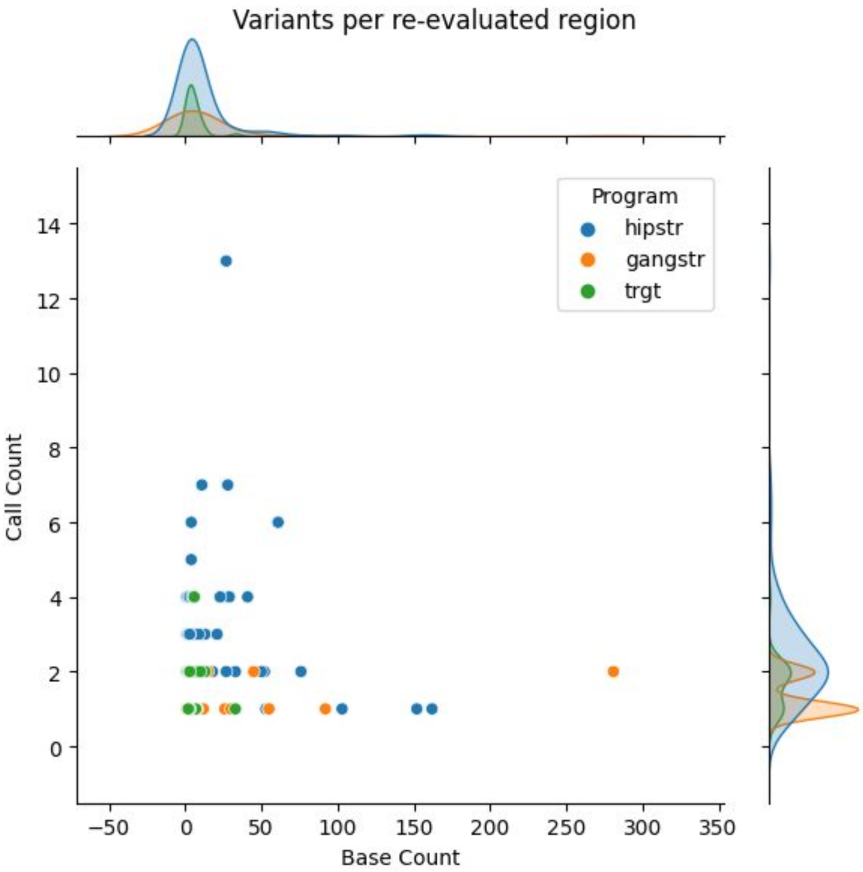
Had Change

	HipSTR	GangSTR	TRGT
Total Regions	32,088	32,088	32,088
Tool Regions	25,555	7,946	3,605
Reevaluated Regions	122	56	26
TPBase	83	15	9
TPCall	83	15	8
FN	84	15	11
FP	117	14	9
Any Change	117	21	13

Less than 1% of regions meet re-evaluation criteria of FN > 0 & FP > 0

Re-evaluated variants

- Parameter `--use-original` pulls variants from the input VCFs per-region
  - Allows variants filtered during bench (e.g. less than sizemin) to be analyzed
- On average, there are 11 base and 2 comparison calls per-regions.
  - Median 4 base and 2 comp
- Generally more ‘variant dense’ regions
  - Non-reevaled regions average 2 base 1.5 comp calls, and median of 1 each.
- Probably finding split representations.



	HipSTR		**GangSTR**		**TRGT**	
	Original	**Rebench**	**Original**	**Rebench**	**Original**	**Rebench**
TP-base	2,063	2,148	1,004	1,019	1,221	1,229
TP-call	2,063	2,149	1,004	1,020	1,221	1,229
FP	1,494	1,324	171	157	85	79
FN	2,459	1,451	3,518	594	3,301	83
precision	0.580	0.619	0.854	0.867	0.935	0.940
recall	0.456	0.597	0.222	0.632	0.270	0.937
f1	0.511	0.608	0.352	0.731	0.419	0.938
base cnt	4,522	3,599	4,522	1,613	4,522	1,312
call cnt	3,557	3,473	1,175	1,177	1,306	1,308

Conclusions

- I think we’re clear to use Truvari for variants >=5bp
  - Rough estimate 1% FN matches, rebench or lower thresholds would recover
  - Maybe 5% of matches are FP, higher thresholds would prevent
- Truvari thresholds at 70% similarity seems fine(?).
  - Giving recommended params of 95% will boost matching precision
- Truvari rebench helps recover missed matches
  - Not many `bench` missed matches
  - Helps reduce unanalyzed FNs
  - Useful in ‘variant dense’ regions
- Need to document Truvari better
- Can now move to creating the benchmark regions