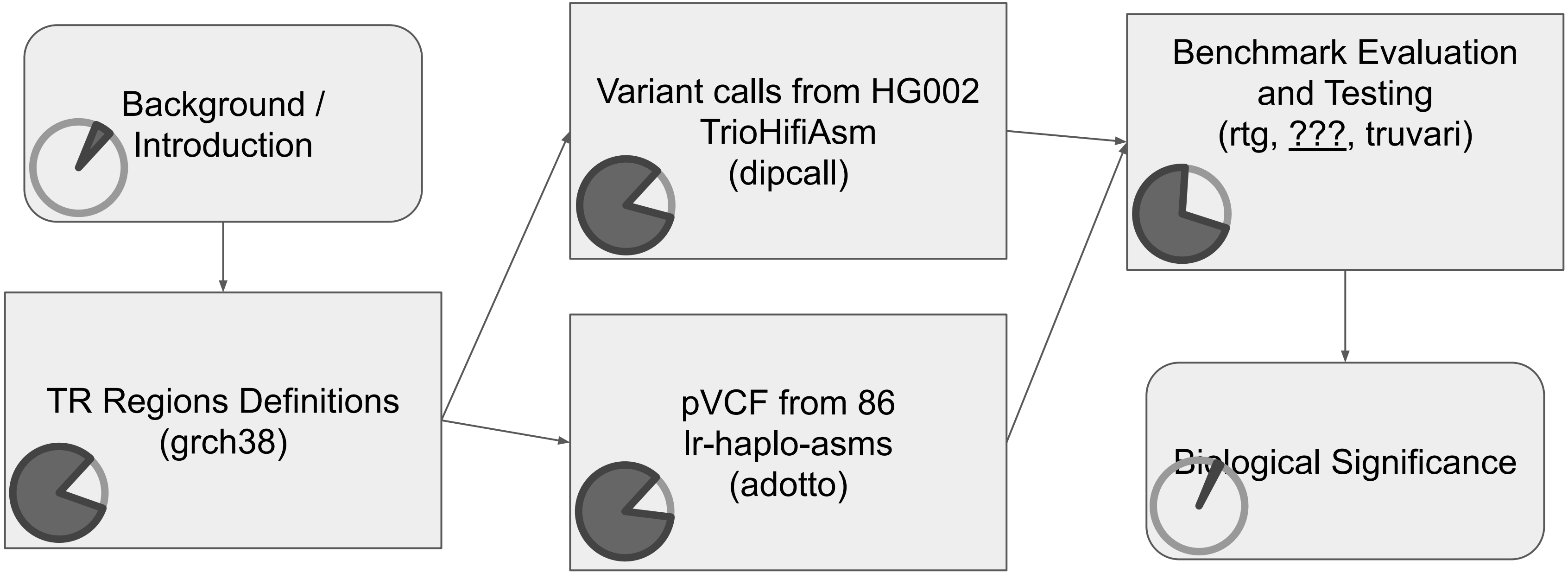
GIABTR

Adam English
HGSC@BCM

September 20, 2022

Revisiting the Roadmap



Background / Introduction

TR Regions Definitions (grch38)

Variant calls from HG002 TrioHifiAsm (dipcall)

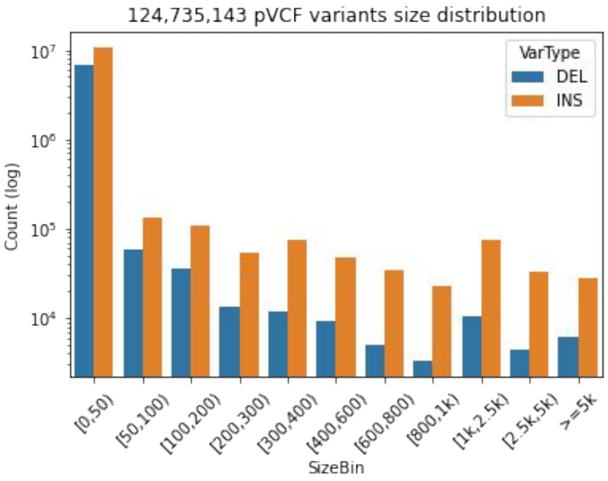
pVCF from 86 lr-haplo-asms (adotto)

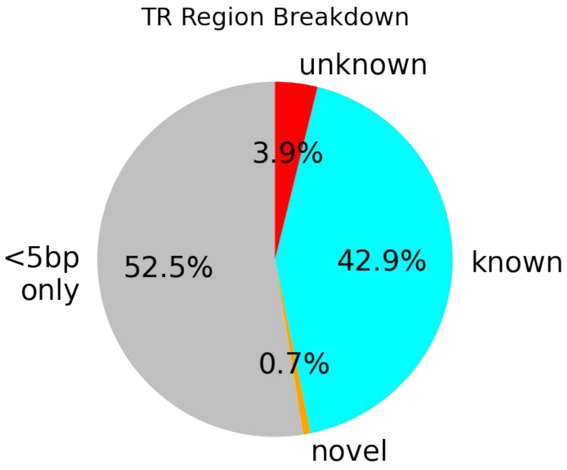
Benchmark Evaluation and Testing (rtg, ???, truvari)

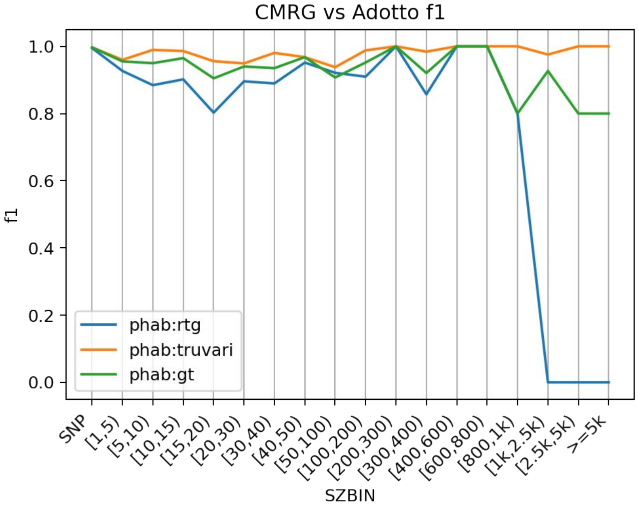
Biological Significance

Benchmarking

- Adotto v0.2 TR Regions
- HG002 Variants
  - GIAB TrioHifiAsm for HG002
  - Adotto HG002
- RTG / Truvari bench / Truvari phab
- Variant calls
  - TRGT
  - GangSTR
  - HipSTR





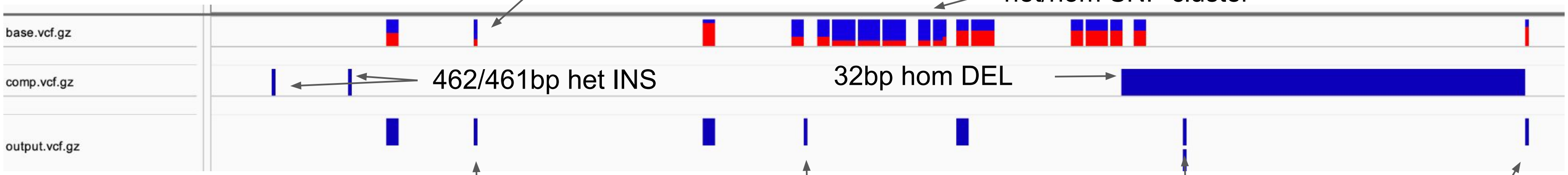


Base VCF

Both Adotto and GIAB TrioHifiAsm (THFA) use the HPRC's HG002 assembly.

With manual inspection and phab, we've found that they have largely the same alleles, but with different variant representations.

Program	Comp	True pos baseline	True pos-call	False-pos	False-neg	Precision	Sensitivity	F-measure
RTG	**hprc**	4,476,465	4,658,725	119,799	186,176	0.9749	0.9601	0.9674
Truvari	**hprc**	22,384	22,384	2,768	7,540	0.889	0.748	0.812



426bp hom INS

het/hom SNP cluster

462/461bp het INS

32bp hom DEL

145bp hom INS

226bp hom INS

23bp hom INS

33/32bp het INS

Ranking TR regions

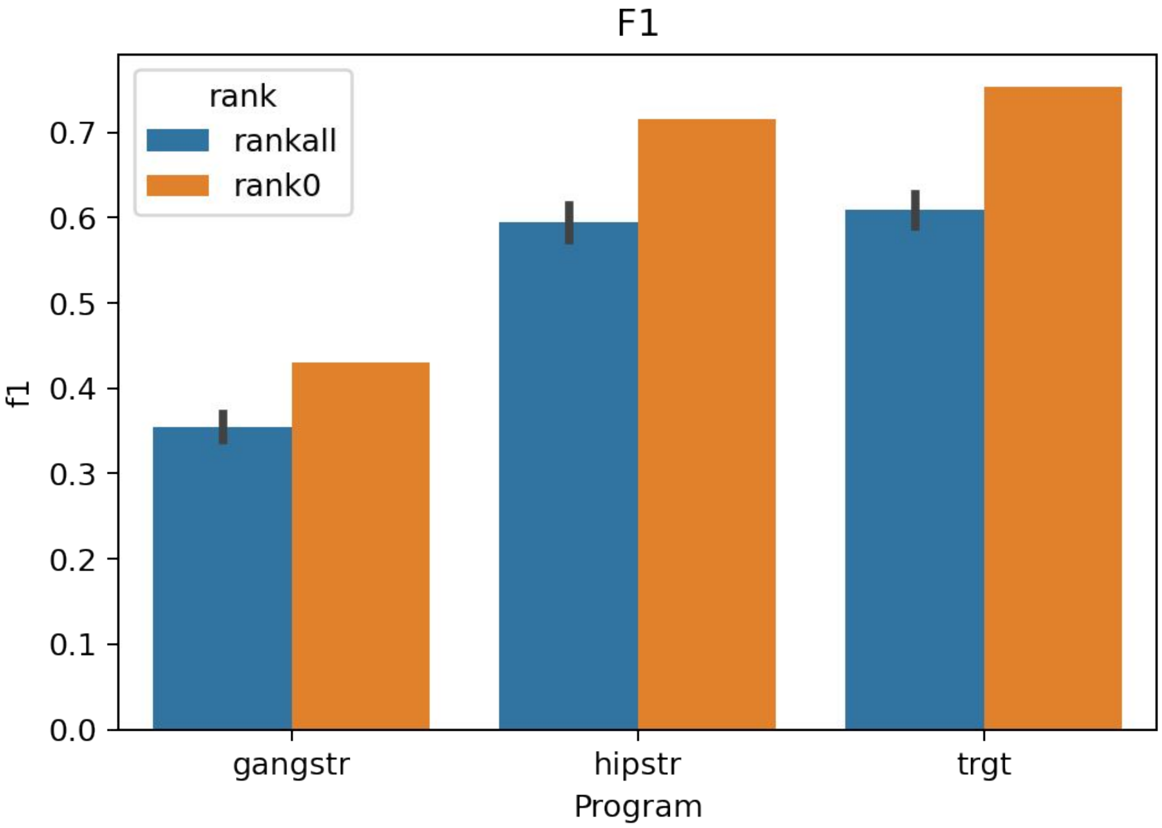
- We have a total of 938,088 regions in our TR catalog
- We are going to classify the regions into different ranks
  - Higher the rank, the harder it is to compare.
  - These ranks should be highly correlated with traditional GIAB Tiers
  - Ranks will be disjointed sets
- Ranks
  - 5: < 5bp variants
  - 4: >=5bp variant from Adotto or THFA
  - 3: >=5bp from Adotto **AND** THFA
  - 2: **SAME NUMBER** of >=5bp in Adotto **AND** THFA
  - 1: Only `truvari bench` **MATCHING** variants between Adotto and THFA
  - 0: Rank1 AND >= 1 variant matching TR catalog using `truvari anno trf`

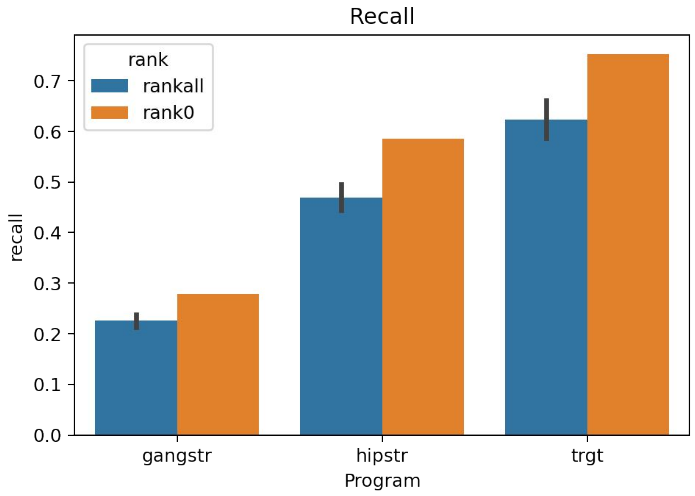
TR regions HG002 ranks

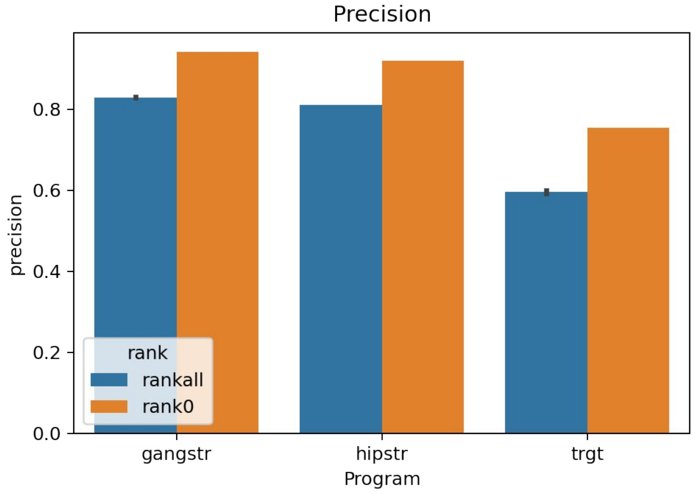
- Rank5: **72.9%** of the TR regions with HG002 variants contain only < 5bp events
- Rank0: **21.3%** HG002 with likely resolvable TR variants
- Rank 1-4: **5.7%** are essentially filters

set	count	percent
Region Total	938,088	
HG002 Total	546,777	58.3%
rank5	398,825	42.5%
rank4	11,129	1.2%
rank3	11,955	1.3%
rank2	4,713	0.5%
rank1	3,660	0.4%
rank0	116,495	12.4%

Results


F1


Recall


Precision

Results

- Rank0 improves performance
- TR Variants not in rank0
  - Short reads - ~21%
  - Long reads - 31%
- Only interested in precision for now

disc	rank	base	TP-call	FP	Precision
gangstr	rank0	adotto	38,482	2,379	0.942
		thfa	38,474	2,387	0.942
	AllTR	adotto	43,320	8,885	0.830
		thfa	43,216	8,989	0.828
hipstr	rank0	adotto	80,922	7,073	0.920
		thfa	80,881	7,114	0.919
	AllTR	adotto	89,892	20,950	0.811
		thfa	89,914	20,928	0.811
trgt	rank0	adotto	103,913	33,944	0.754
		thfa	103,984	33,873	0.754
	AllTR	adotto	120,056	80,321	0.599
		thfa	118,759	81,618	0.593

TRGT FPs

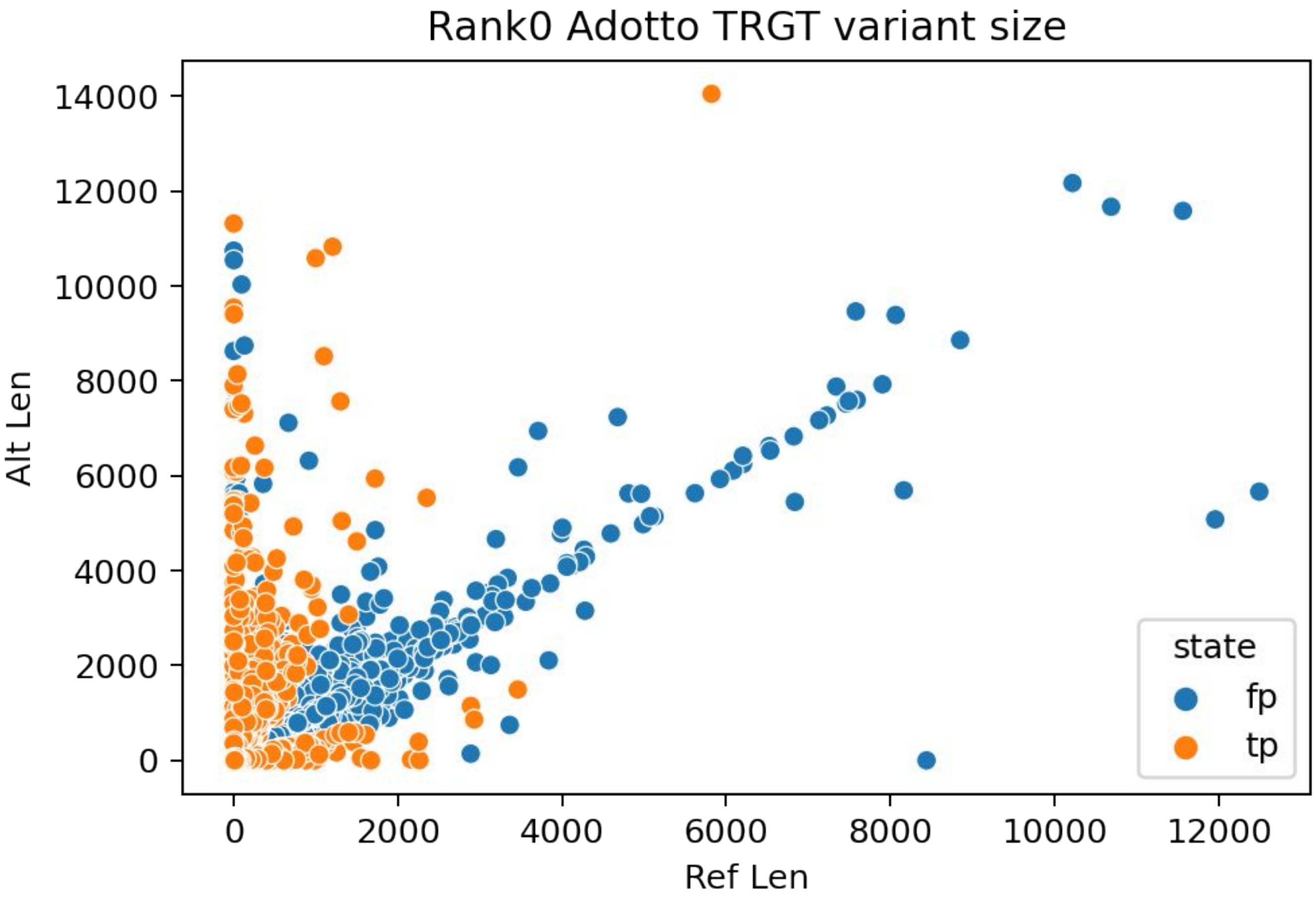
TRGT has lowest precision

86% of FP have REF & ALT >= 10bp

32% have <10bp diff between REF/ALT

‘**REPL**’ variant types

Excellent truvari phab candidates



Analyzing Adotto x TRGT Rank0 regions

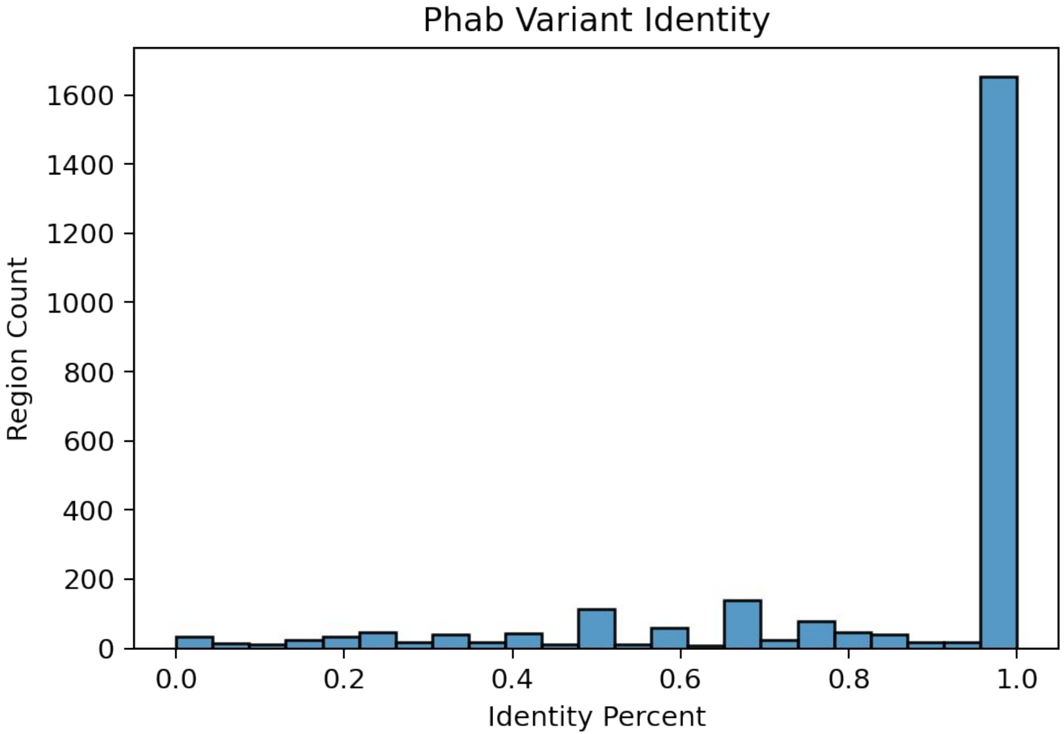
- Count the TP/FP/FN of TRGT results of region.
- Classify the regions by how many TP/FP/FN to find phab candidates

state	base count	comp count	region count
Mix of T/F	8,619	10,758	6,441
No TRGT calls	1,931	0	1,702
No FP/FN	94,719	94,721	83,118
Other	1	0	641
Unmatched	32,861	32,378	24,593

- ~24k regions with possibly unmatched calls (95% of TRGT FPs)
  - For testing, I'm just going to take 2,500 regions
- Also 641 'other' regions falling outside of classification (**??**)

TRGT FP phab

- 1,654 (66%) of regions have all variants matches after phab harmonization
- 296 (12%) have fewer than 50% of variants matching
- 8,233 of 12,444 (66%) input variants are in regions with >= 90% matches.



TRGT estimated performance

We can estimate tool’s performance by recategorizing 66% of FP as TP

	Original	After Phab Estimate
TP	103,913	126,316
FP	33,944	11,541
Precision	0.754	0.916

Summary

- Built first candidate benchmark dataset
- Software works well enough to get performance estimates
- True Negative regions?
    - Should we pepper in some of the TR regions where HPRC assembly confidently covers and only reference homozygous calls?
- Next steps:
    - Still need to document / upload benchmark
    - Evaluate these initial rough results
    - Work to formalize pipeline / metrics
    - Reach goal of user inputs VCF and pipeline outputs performance table