GIABTR

Comparing to the Benchmark

September 4, 2023

Goals:

- Description of the new tooling (Truvari, Laytr)
- Use the benchmark to demonstrate its utility
- Answer 4 main questions:
  - How much of a difference does refine make on the results?
    - Highlight how that some of these tools based on a catalog. Some of the catalogs are subsets of our full catalog and of the benchmark regions.
  - How informative are the reports?
    - Can we use variant / region / laytr to get a better understanding of what any particular tool is doing?
    - Want to bring attention to the subsets stratifications.
  - Are there sites inaccessible from one technology / technique vs another?
    - Sites where non-nist WGS all miss but every TR caller captures
    - Sites long reads capture but none of the short reads capture
  - How many locations are universally FP or FN?
    - These may be indicative of low quality sites to remove from the benchmark.

New Tooling

Truvari refine

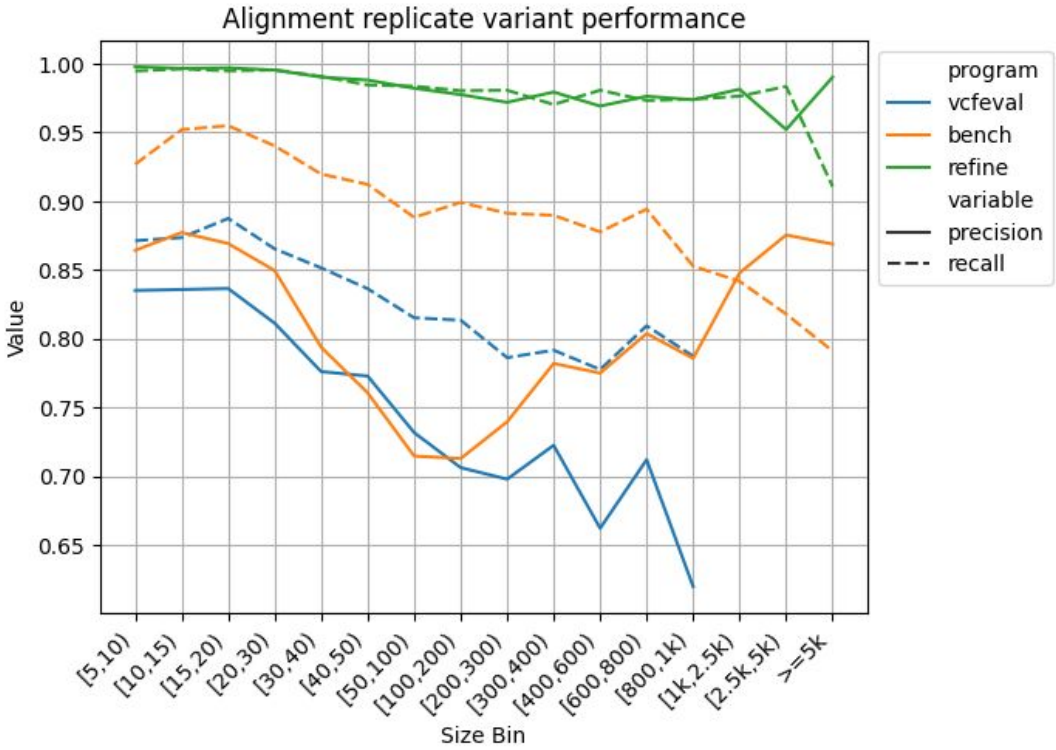
- For regions with FN/FP, harmonize variants with MAFFT and recompare
- Improves comparability of disparate variant representations
- Region counting for performance metrics

Laytr

- Report stratifications on region annotations

Refine main example

- Using the alignment replicate, report the precision/recall from bench, refine, and rtg vcfeval
  - Theoretically capable of ‘perfect’ performance



Introduction to Laytr

- Intersects refine.regions.txt, with benchmark and full catalog annotations
  - example report
- Will use the technical replicates as the primary example
  - Not much to interpret by way of performance.
  - Since they’re not produced by tools (e.g. hipstr/medakka) we can limit risk of results being interpreted as a bakeoff

Variant Callers

Program	Caller Type	Sequencing	Locations	Ref
gangstr	TR	short-reads	catalog	link
hipstr	TR	short-reads	catalog	link
medakaTR	TR	long-reads	catalog	n.a.
trgt	TR	long-reads	catalog	link
deepvariant	snp-indel	short-reads	WGS	link
biograph	SV (& small)	short-reads	WGS	link
sniffles	SV	long-reads	WGS	link
GIABv4.2.1	snp-indel	long-reads	WGS	link
GIABverkko*	all	assembly	WGS	n.a.
hipstr_sub	STR	short-reads	catalog	link
GIABv4.2.1_sub	snp-indel	long-reads	WGS	link
GIABverkko_sub*	all	assembly	WGS	n.a.

TR callers

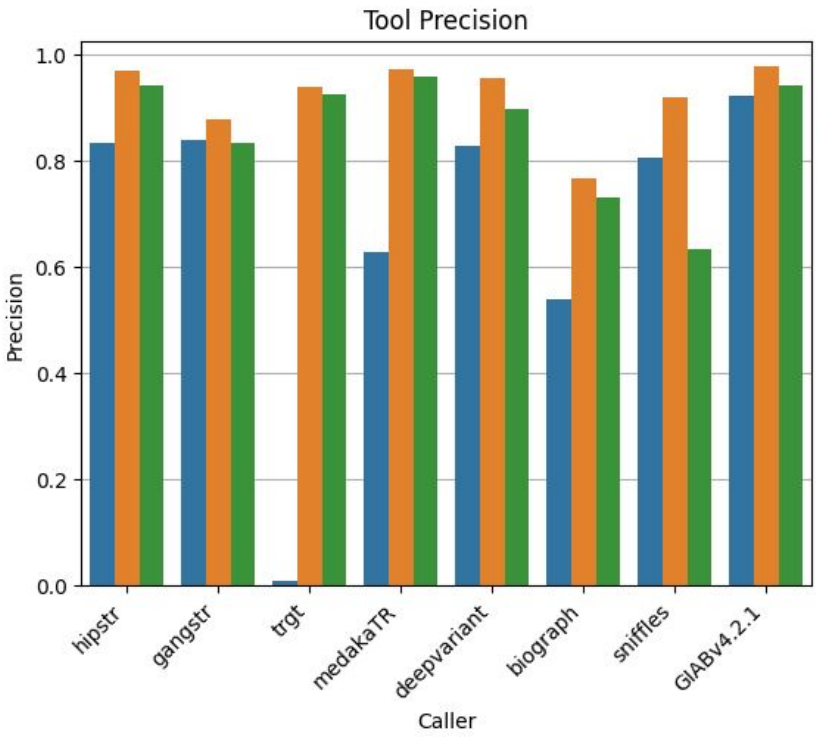
Happen to get TRs callers

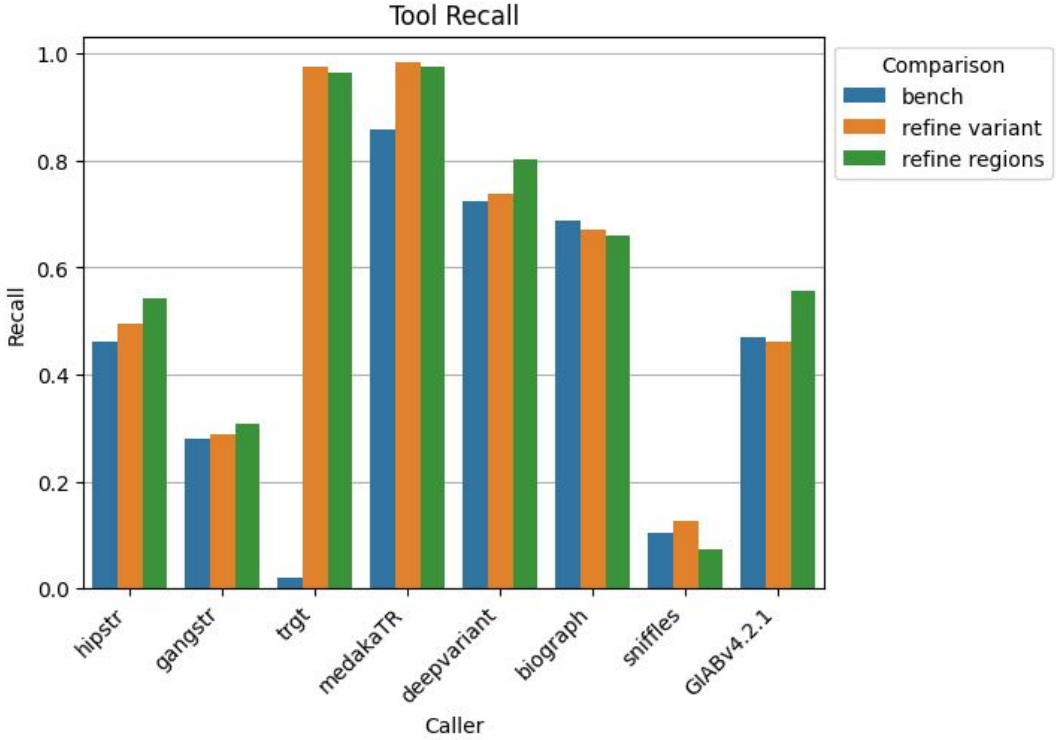
Truth sets

Subsets

*GIABverkko won't be in manuscript
Will use '_sub' to highlight `refine --regions`

Q1: How much of a difference does refine make on the results?





Refine Difference

- Refine improves bench performance by ~13p.p.
- BioGraph and GIABv4.2.1 have lower recall after refine.

tool	precision	recall
biograph	0.227	-0.017
GIABv4.2.1	0.054	-0.008
GIABv4.2.1_sub	0.064	0.525

- Lack of phasing is causing `bcftools consensus` (phab) to drop variants.

Refine Variant - Bench Difference		
	precision	**recall**
mean	0.247	0.141
std	0.293	0.332
min	0.039	-0.017
25%	0.099	0.003
50%	0.132	0.018
75%	0.257	0.055
max	0.932	0.956

Importance of Region Summary

Variant count before/after refine:

	Bench		Refine		Refine - Bench	
tool	**base**	**comp**	**base**	**comp**	**base**	**comp**
hipstr	139,372	78,458	143,090	74,038	+3,718	-4,420
gangstr	139,372	46,685	142,321	46,701	+2,949	+16
trgt	139,372	316,718	154,183	160,909	+14,811	-155,809
medakaTR	139,372	225,881	152,035	175,396	+12,663	-50,485
deepvariant	139,372	122,650	155,810	120,885	+16,438	-1,765
biograph	139,372	173,819	156,093	133,438	+16,721	-40,381
sniffles	139,372	15,825	148,243	18,538	+8,871	+2,713
GIABv4.2.1	139,372	71,177	152,290	72,087	+12,918	+910

1. The benchmark VCF's representations tend to split after harmonization (increase in base cnt)
2. Comparison VCFs' representations may split or combine (pos/neg comp cnt diff)
3. trgt/deepvariant have reference homozygous calls counted in the set, but not in metric counts
4. medakaTR reports homozygous variants as two hets

Region Summary

	base P	base N	comp P	comp N	PPV	TPR	TNR	NPV	ACC	BA	F1
hipstr	106,461	1,600,392	61,214	1,645,639	0.943	0.542	0.999	0.972	0.971	0.771	0.689
gangstr	105,843	1,601,010	38,960	1,667,893	0.835	0.307	0.998	0.958	0.955	0.653	0.450
trgt	107,253	1,599,600	111,634	1,595,219	0.927	0.965	0.997	0.999	0.995	0.981	0.945
medakaTR	104,858	1,601,995	106,646	1,600,207	0.959	0.975	0.998	1.000	0.997	0.987	0.967
deepvariant	107,376	1,599,477	95,936	1,610,917	0.898	0.802	0.999	0.992	0.987	0.901	0.848
biograph	107,159	1,599,694	96,990	1,609,863	0.730	0.661	0.995	0.988	0.974	0.828	0.694
sniffles	105,081	1,601,772	11,949	1,694,904	0.635	0.072	1.000	0.945	0.943	0.536	0.130
GIABv4.2.1	104,340	1,602,513	61,654	1,645,199	0.944	0.558	1.000	0.974	0.973	0.779	0.701

Variant Counts Benchmark Average: 150,508 **± 5,418**
Region Counts Benchmark Average: 106,046 **± 1,191**

The consistency of the region counts is tighter

Also highlight balanced accuracy as being a more useful metric than Accuracy due to the imbalance of negative and positive regions

Q3: Are there sites accessible to one technology but not another?

- Combine refine.regions.txt and make subsets
  - ShortRead TR agreement (gangstr/hipstr)
  - LongRead TR agreement (trgt/medaka)
- Compare subsets
  - With/without TN makes a difference.

Within Read Length TR caller agreement

Compare gangstr to hipstr and trgt to medakaTR. How often do the tools have the same benchmarking state?

		Short Read TR	Long Read TR
All	**Agree**	1,666,643	1,692,724
	Disagree	40,210	14,129
	Percent	97.64%	99.17%
Tier1	**Agree**	1,600,717	1,628,832
	Disagree	37,791	9,676
	Percent	97.69%	99.41%
Tier1 & HG002 ≥5bp	**Agree**	67,706	98,044
	Disagree	33,998	3,660
	Percent	66.57%	96.40%

Across Read Length TR caller agreement

Compare the short-read agreement sites with the long-read agreement sites. How often do the two read lengths agree? (Tier1 HG002 ≥5bp Regions)

Long Read	FN	FN,FP	FP	TP
Short Read				
TP	5	1	2	26,479
FN,FP	0	13	0	30
FP	0	0	41	2
TN	0	0	0	0
FN	188	58	107	38,279

~40% of regions are resolved regardless of read length used

~58% of regions are only resolved by long reads

(top) **8** regions are resolved only by short reads

The measurement of regions resolved by read length may be confounded by the catalogs used by gangstr/hipstr being a subset of the benchmark's regions whereas trgt/medaka analyze all regions.

Are there any patterns to what's read length resolvable?

What is it about the 58% long-read only resolved regions that's different from the 40% both resolvable? The length of the change and sequence entropy.

		Both	Long-read only
	count	26,907	38,421
max allele delta	**mean**	15.3	**105.9**
	std	187.9	398.2
	median	9	23
entropy	**mean**	0.887	**0.846**
	std	0.059	0.109
	median	0.892	0.865

Q3 part 2: Are there sites inaccessible from one technique vs another?

- Classify the Tier1 HG002 ≥5bp Regions by
  - ANY of the WGS callers (DeepVariant, BioGraph, Sniffles) match the benchmark
  - ANY of the TR callers match the benchmark
- How many are resolved by each set?
  - **105** (0.1%) are only resolved by the WGS callers
  - **10,304** (10.1%) are only resolved by the TR callers
  - **90,846** of 101,704 TR regions resolved by both.

Q4: Consistently FN/FP sites

Any place where all the callers are disagreeing with the benchmark are candidates for demotion to Tier2. This demotion would boost the reliability of the benchmark's Tier1 regions.

- 894 of the 1,706,853 regions (0.05%) are FP/FN/unanalyzed on all callers
  - 888 are also FP/FN in GIAB v4.2.1
  - 490 (54.8%) are already Tier2

Q2: How informative are the reports?

1) Need to describe each of the laytr stratifications
    a) Subsets
    b) Entropy
    c) Gene
    d) Interspersed
    e) Repeat Complexity
    f) Motif Length
    g) SOMs
    h) Expansion / Contraction (type)
    i) Max Sizebin (length)
2) Most of this can go in the methods, but I need to figure out one or two examples that I can put in the main text.