Subsetting TRr_v1.1 to HG002 Benchmark

Adam English
Baylor College of Medicine

v1.1 !?

- `hom_span` is now `hom_pct`

- Added/consolidated with gnomAD/STRipy pathogenic repeats
  - 54 pathogenic regions unchanged, 2 changed, and 6 added.
  - Changed
    - NOTCH2NLC -> NOTCH2NLA
    - NOTCH2NL -> NOTCH2NLC
  - Added
    - EIF4A3, PRNP, TBX1, PRDM12, DMD, ZIC3
- Total of 66 known pathogenic repeats hitting 62 regions

chrom	start	end	Locus	Motifs 1	Motifs 2	Repeat type	Region	Path. repeats	Inheritance mode	Disease
TRregions 37bp upstream and 188bp downstream										
chr1	1435798	1435818	VWA1		GGCGC GGAGC		Coding	>=3	Autosomal recessive	Hereditary motor neuropathy
Single Region ARX → chrX:25013536-25013899										
chrX	25013649	25013698	ARX_1	GCN	GCG	Imperfect GCN	Coding	>=23, >=18	X-linked recessive	Developmental and epileptic encephalopathy-1 (DEE1)X-linked mental retardation with or without seizures (MRXARX)
chrX	25013529	25013565	ARX_2	GCN		Imperfect GCN	Coding	>=20, >=20, >=23	X-linked recessive	Developmental and epileptic encephalopathy-1 (DEE1)Partington syndrome (PRTS)X-linked mental retardation with or without seizures (MRXARX)
Single Region HOXA13 → chr7:27199614-27200230										
chr7	27199924	27199966	HOXA13_1	GCN	NGC	Imperfect GCN	Coding	>=22	Autosomal dominant	Hand-foot-genital syndrome (HFG)
chr7	27199825	27199861	HOXA13_2	GCN		Imperfect GCN	Coding	>=18	Autosomal dominant	Hand-foot-genital syndrome (HFG)
chr7	27199678	27199732	HOXA13_3	GCN		Imperfect GCN	Coding	>=24	Autosomal dominant	Hand-foot-genital syndrome (HFG)

Intersecting TRr_v1.1 with Assembly Coverage

- When selecting the regions for the HG002 benchmark, we need to have confident coverage (1x per-hap) from the HPRC haplotype-resolved assembly.
- We'll analyze two alignments of the assembly (dipcall, adotto) as well as their intersection.
  - How much of the genome is covered confidently?
  - How many TRregions are covered confidently?

Genome

	Span Count	Span Total BP	Genome %
dipcall	48,624	2,778,450,120	86.8%
adotto	328	2,668,392,630	83.4%
Both	45,870	2,615,712,814	81.7%

TRregions

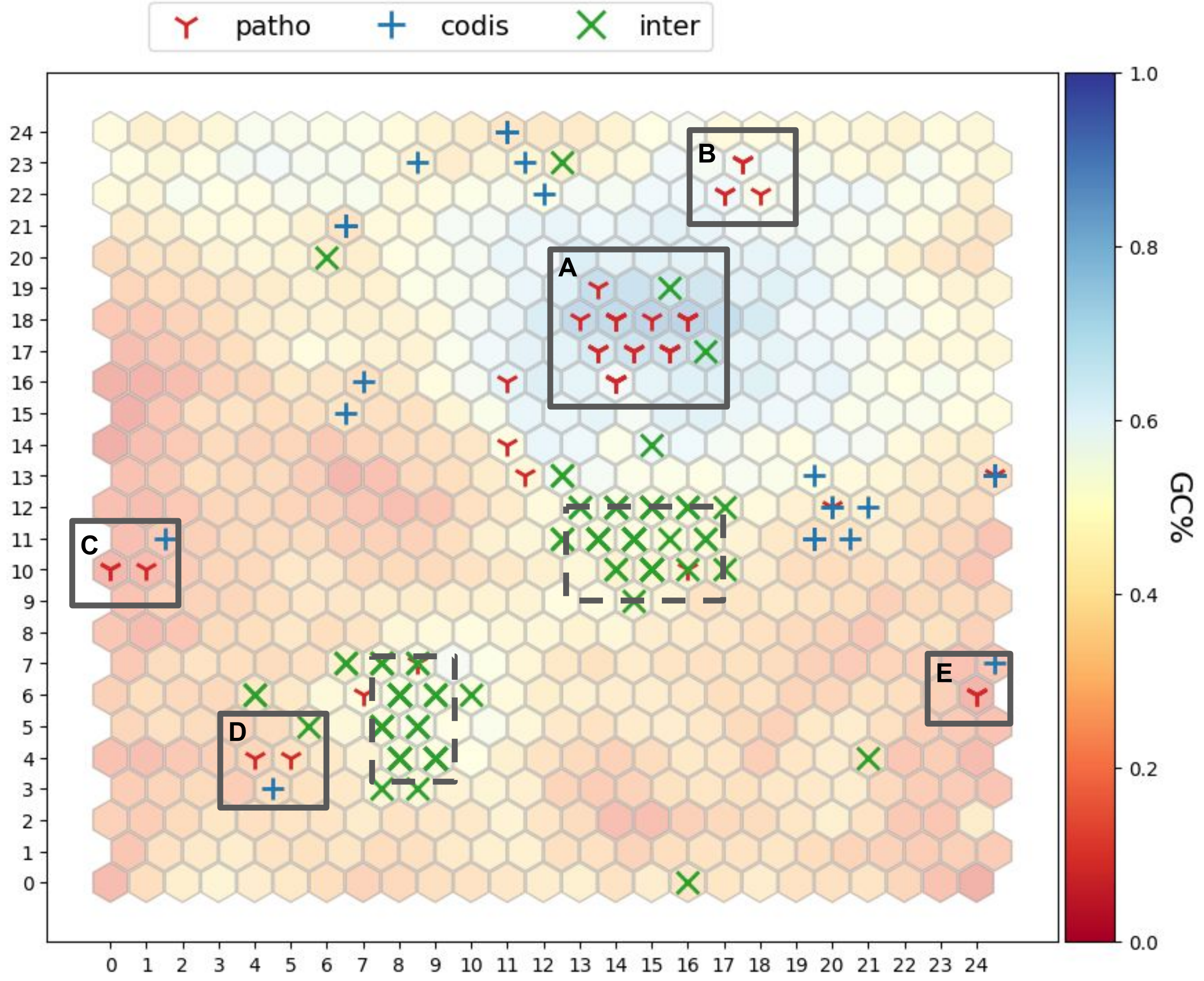
	Count	Span	Genome %	TRr Count %	TRr Span %
Total TRr	1,784,804	237,865,075	7.4%		
dipcall	1,707,318	212,853,127	6.7%	95.66%	89.48%
adotto	1,701,194	217,607,408	6.8%	95.32%	91.48%
Both	1,645,456	203,578,939	6.4%	92.19%	85.59%

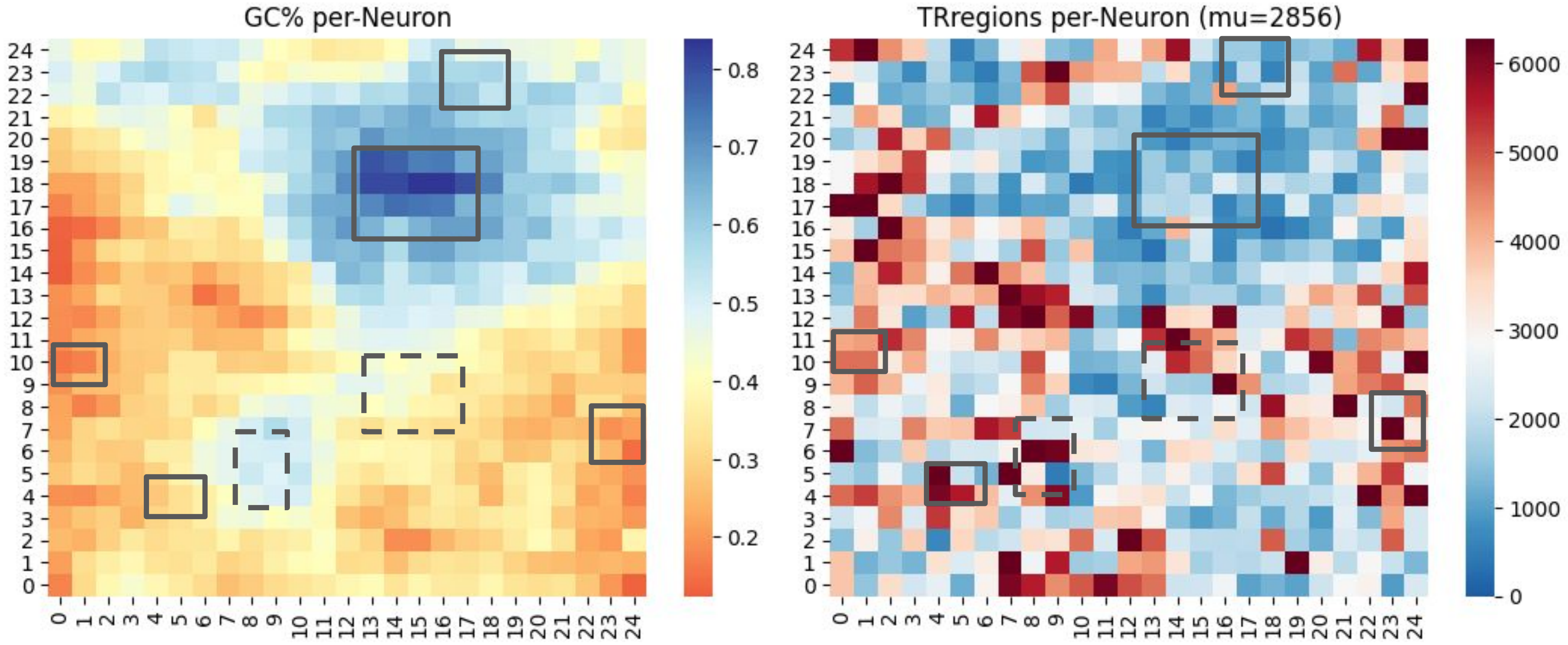
Patho/Codis

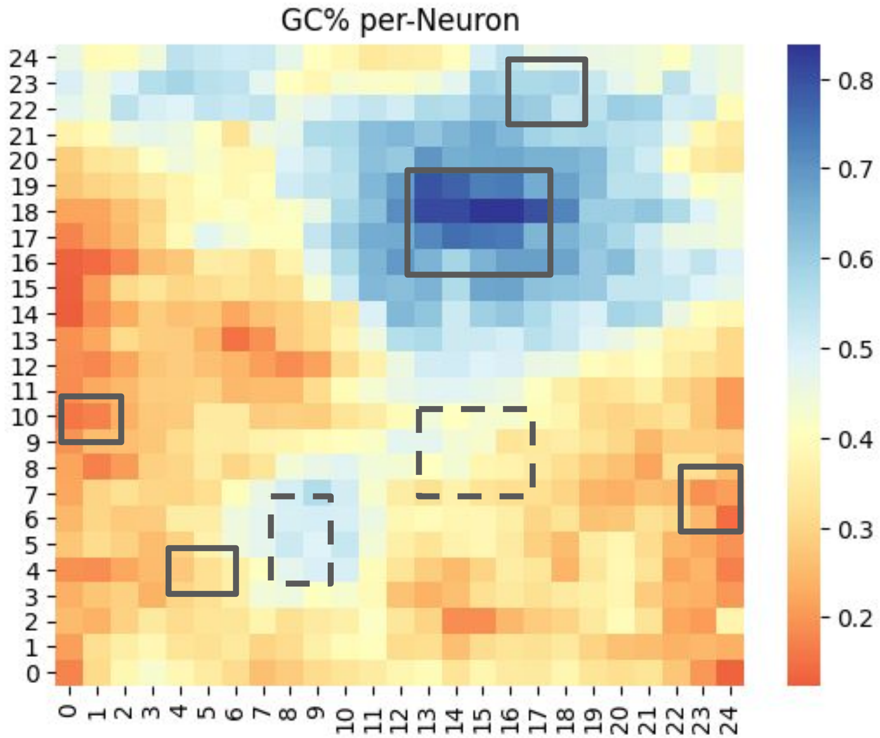
	Patho	Patho %	Codis	Codis %
Total TRr	62		51	
dipcall	50	80.65%	44	86.27%
adotto	52	83.87%	24	47.06%
Both	42	67.74%	23	45.10%

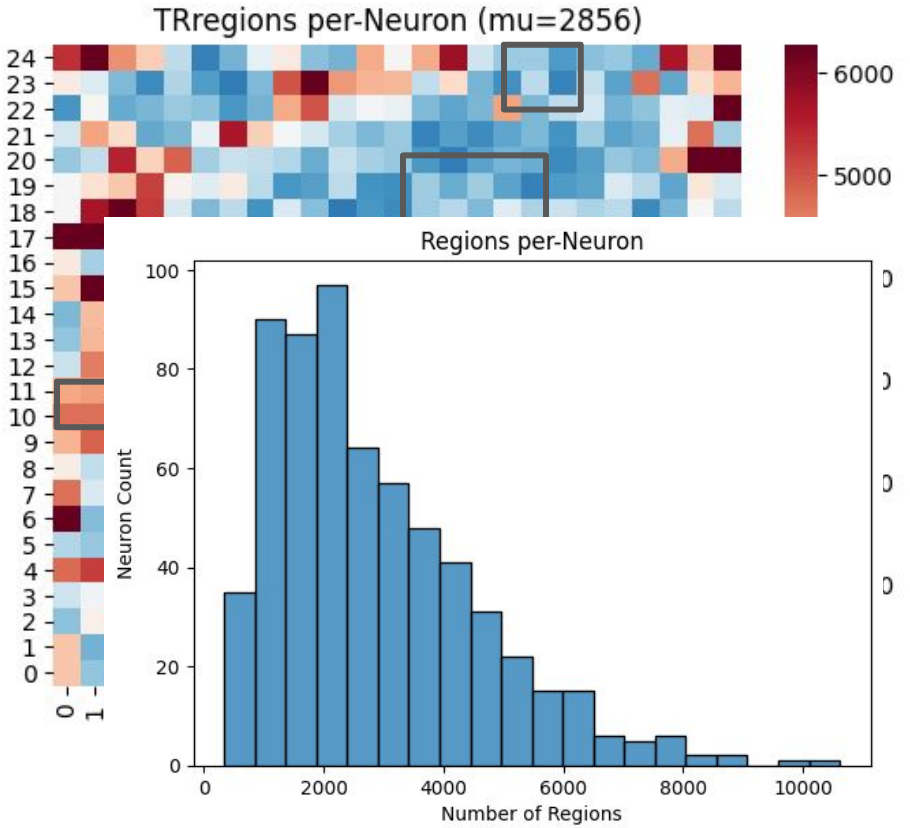
Cluster	Motif	Count
A	CGG	10
	CCG	10
	CNG	7
	CTG	7
	GCN	2
	ACCTCGCTGTG CCGCTGCCG	1
	GGCCTG	1
	CGCGGGGCGG GG	1
	CCCCGG	1
B	AGC	6
C	AAAAT	3
D	AAAAG	1
	AAG	1
E	TTTTA	3

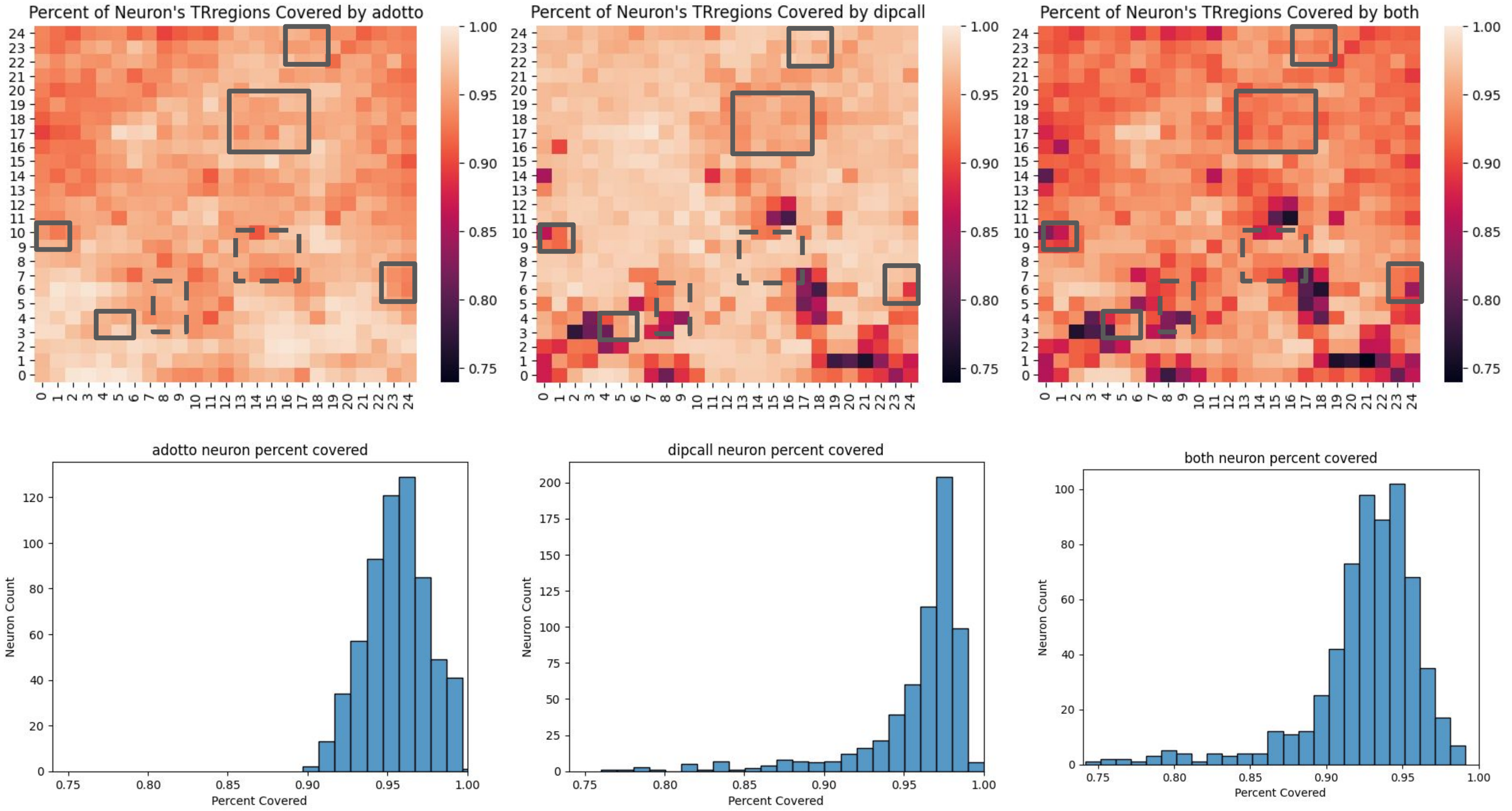
- 54 of 62 Patho TRr in 5 clusters
- Interspersed TRr concentrated in two clusters




GC% per-Neuron
TRregions per-Neuron (mu=2856)
0.8
0.7
0.6
0.5
0.4
0.3
0.2
6000
5000
4000
3000
2000
1000
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24

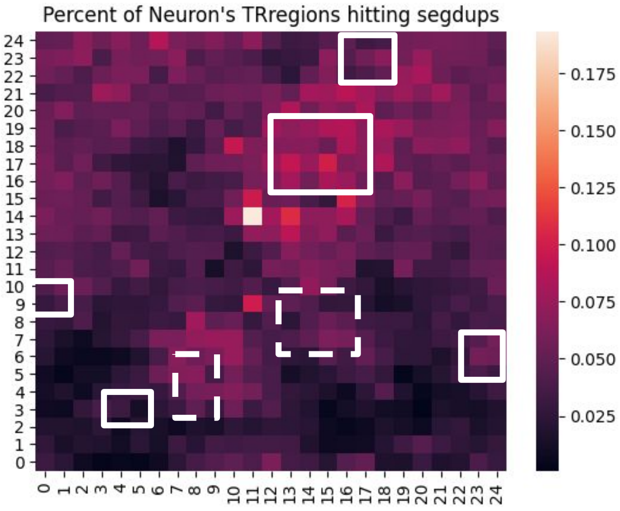

GC% per-Neuron
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
0.8
0.7
0.6
0.5
0.4
0.3
0.2


TRregions per-Neuron (mu=2856)
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
6000
5000
Regions per-Neuron
Neuron Count
0
20
40
60
80
100
Number of Regions
0
2000
4000
6000
8000
10000

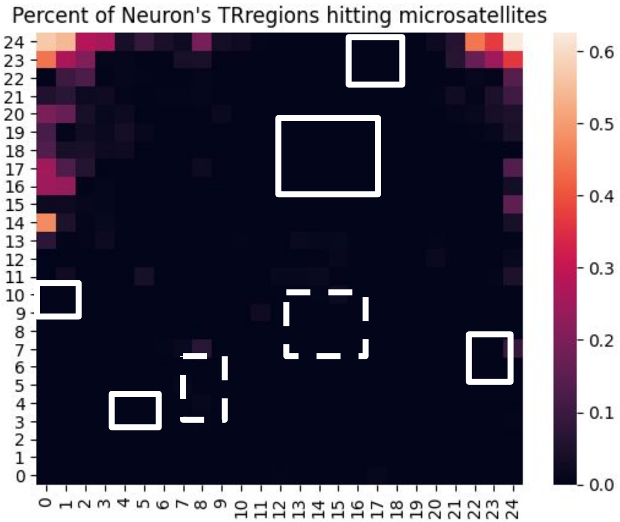

Percent of Neuron's TRregions Covered by adotto
Percent of Neuron's TRregions Covered by dipcall
Percent of Neuron's TRregions Covered by both
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
1.00
0.95
0.90
0.85
0.80
0.75
adotto neuron percent covered
dipcall neuron percent covered
both neuron percent covered
Neuron Count
Percent Covered
0
20
40
60
80
100
120
25
50
75
125
150
175
200
0.75
0.80
0.85
0.90
0.95
1.00

What types of TRr are in these dipcall 'dryspots'?

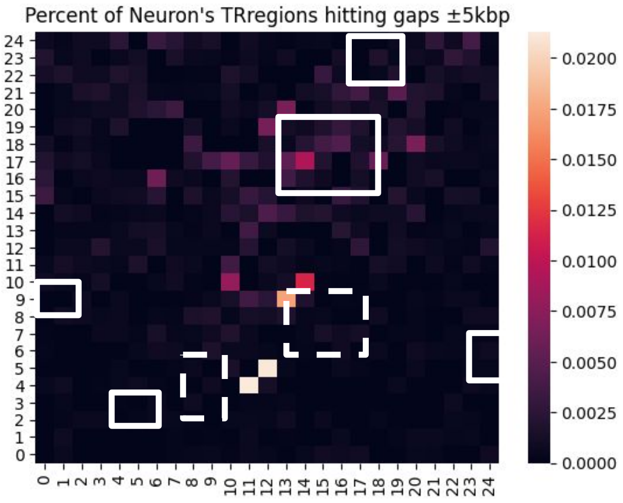
SegDups
73,736 regions



Microsatellites
40,225 regions



Gaps
1,333 regions

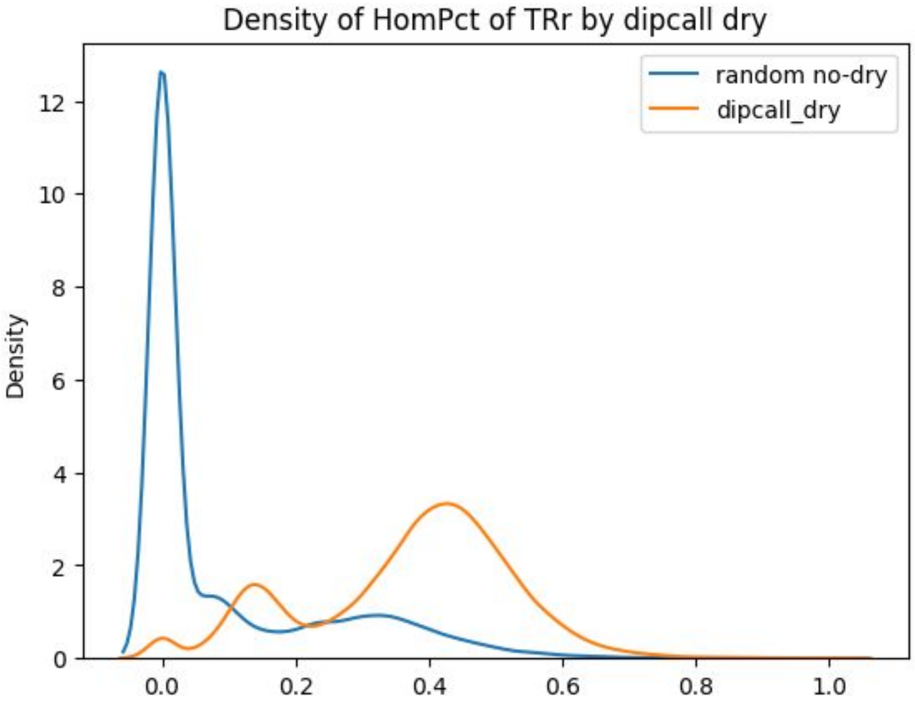


Tracks from UCSC Table Browser

Homopolymers

“[The regions have] imperfect homopolymers at the edge of SINEs. We exclude these from our benchmark regions even though they are covered by the assembly because we exclude perfect or imperfect homopolymers longer than 30bp due to higher HiFi error rates” - Justin Zook

We excluded homopolymer annotations from our TR regions, but if TRF also found other non-homopolymer annotations in the region, it stayed in the catalog.



22,787 regions <80% captured vs 30k random.

adotto_TRregions_v1.1_HPRC_HG002_Covered.bed

Run Truvari on callers against all HG002 variants in pVCF

```
bench --no-ref a --sizemin 5
      --pick ac --includebed $BED
refine --use-original --reference $REF
```

Too many TN regions *may* make it difficult to interpret

Variant Summary			
	GangSTR	TRGT	HipSTR
TP-base	39,792	139,477	68,735
TP-comp	39,798	140,455	69,642
FP	5,620	16,727	6,225
FN	100,599	9,186	71,075
precision	0.876	0.894	0.918
recall	0.283	0.938	0.492
f1	0.428	0.915	0.640
base cnt	140,391	148,663	139,810
comp cnt	45,418	157,182	75,867

Region Summary			
	GangSTR	TRGT	HipSTR
TP	31,314	94,836	55,340
TN	1,537,514	1,529,488	1,535,837
FP	5,159	15,054	6,037
FN	72,993	6,790	48,611
base P	104,701	104,249	104,607
base N	1,540,755	1,541,207	1,540,849
comp P	37,757	111,327	62,711
comp N	1,607,699	1,534,129	1,582,745
PPV	0.829	0.852	0.882
TPR	0.299	0.910	0.529
TNR	0.998	0.992	0.997
NPV	0.956	0.997	0.970
ACC	0.953	0.987	0.967
BA	0.648	0.951	0.763
F1	0.440	0.880	0.661

Finding TRs.

We’ve been analyzing variants based on length >=5bp. However, not all INDELs >=5bp are tandem repeat expansions/contractions. Therefore, we need a way to find TRs. `truvari anno trf` is designed to annotate if INDELs are TR exp/con.

Process:

- For each variant within TRregions over `--sizemin` (we’re using 5):
- Compare the variant to the TRregion’s annotations
- If no match to TRregion annotations, alter the reference sequence spanned by the TRregion with the variant, rerun TRF, and try to match
- If still no match, but a TRF annotation overlapping the variant, add that annotation to entry

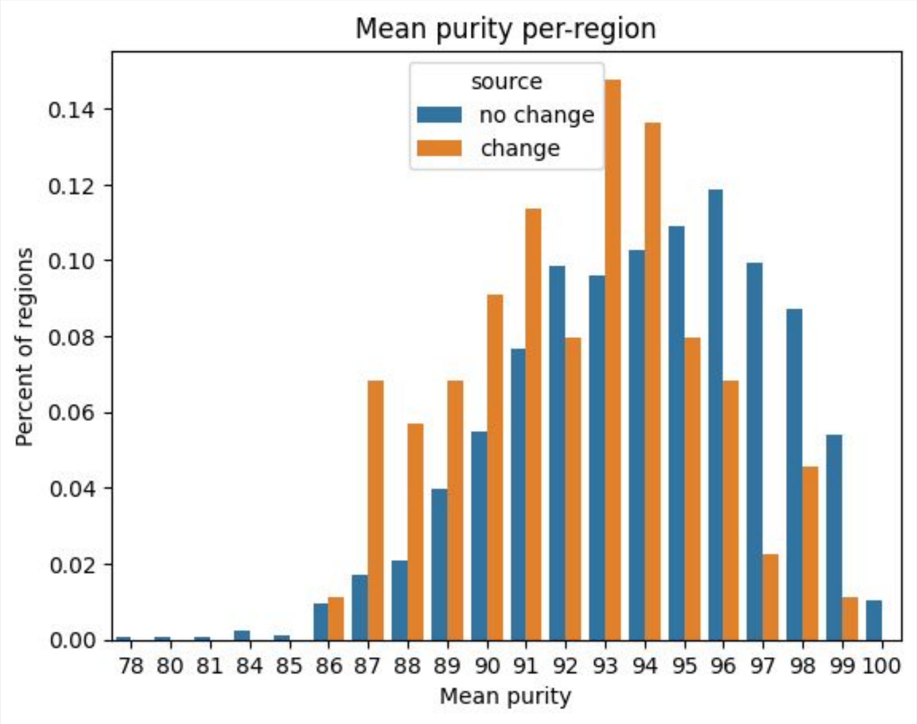
INFO	**Definition**
TRF	Entry hits a tandem repeat region
TRFdiff	ALT TR copy difference from reference
TRFrepeat	Repeat motif
TRFovl	Percent of ALT covered by TRF annotation
TRFstart	Start position of discovered repeat
TRFend	End position of discovered repeat
TRFperiod	Period size of the repea
TRFcopies	Number of copies aligned with the consensus pattern
TRFscore	Alignment score
TRFentropy	Entropy measure
TRFsim	Similarity of ALT sequence to generated motif faux sequence

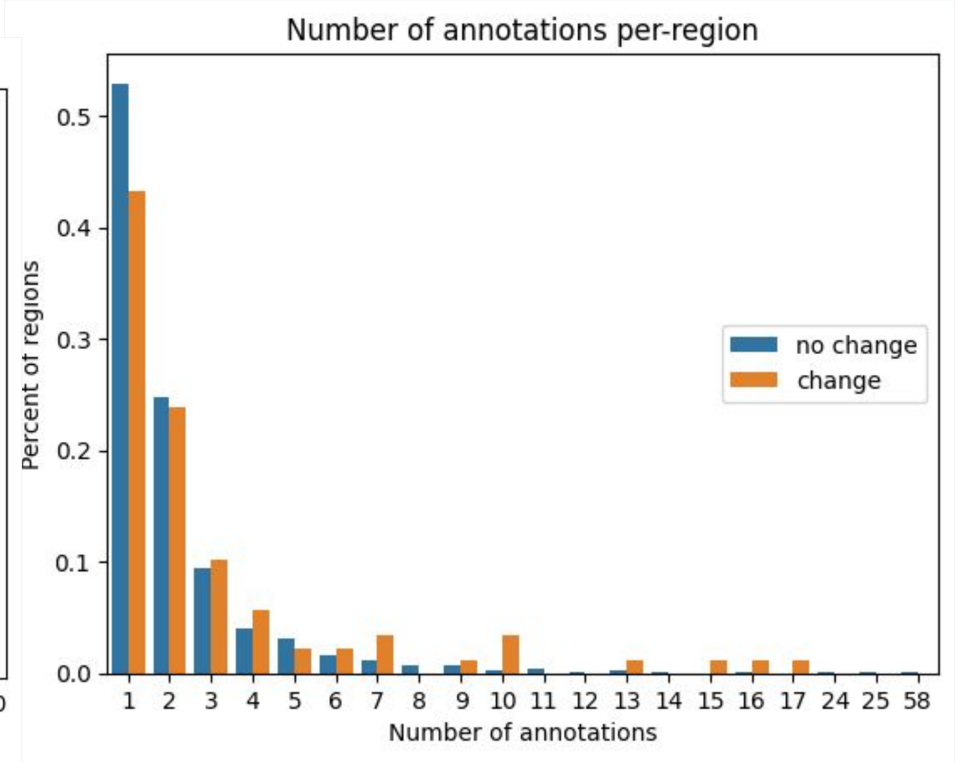
Testing `truvari anno trf` - Finding ‘simple’ cases

Use phab to harmonize pVCF. Identify regions where HG002 has one INDEL >=5bp in original and phab pVCF.

- 37,393 covered regions on chr20
- 2,381 with >=5bp variant (6.4% of all)
- 1,731 with one >=5bp variant (72.7% of var-regions)
- 1,643 with one >=5bp variant post-phab (69%)

Comparing 1,643 simple regions with 88 regions where phab changed variant count. The changed TRregions appear to be slightly more ‘complex’ on average.





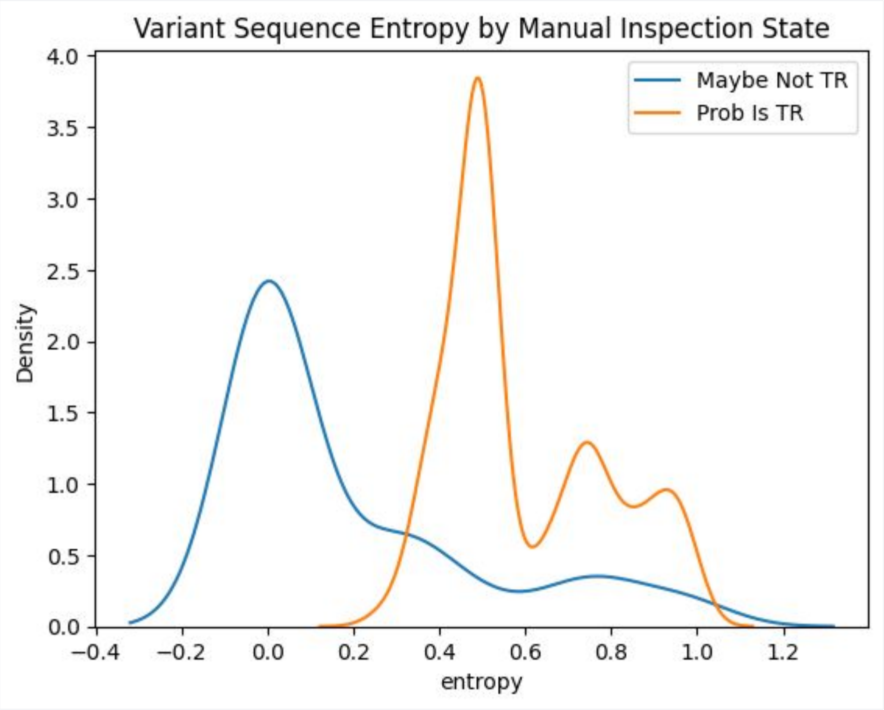
Manual inspection of Variants

- 47 not_annotated (2.9%) - may not be tandem repeats
    - 24 homopolymers
    - 3 missed (likely TR)
    - 4 missed (unlikely TR)
    - 13 not TRs
    - 3 huge, ignored (583bp, 1424bp, 2962bp insertions)
- 228 not_in_annos (13.9%) - TRdiff == 0, couldn’t match to TRrep-annos
    - 60 homopolymers
    - 14 DEL with long period 167-412, but little SVLEN (<50)
        - 11 are homopolymers
    - 153 INS - missed - not in catalog.
- 1,368 annotated (83.3%) - Possible real TRs.
    - No homopolymers
    - 781 DEL
    - 587 INS

TR expansion/contraction identification heuristics

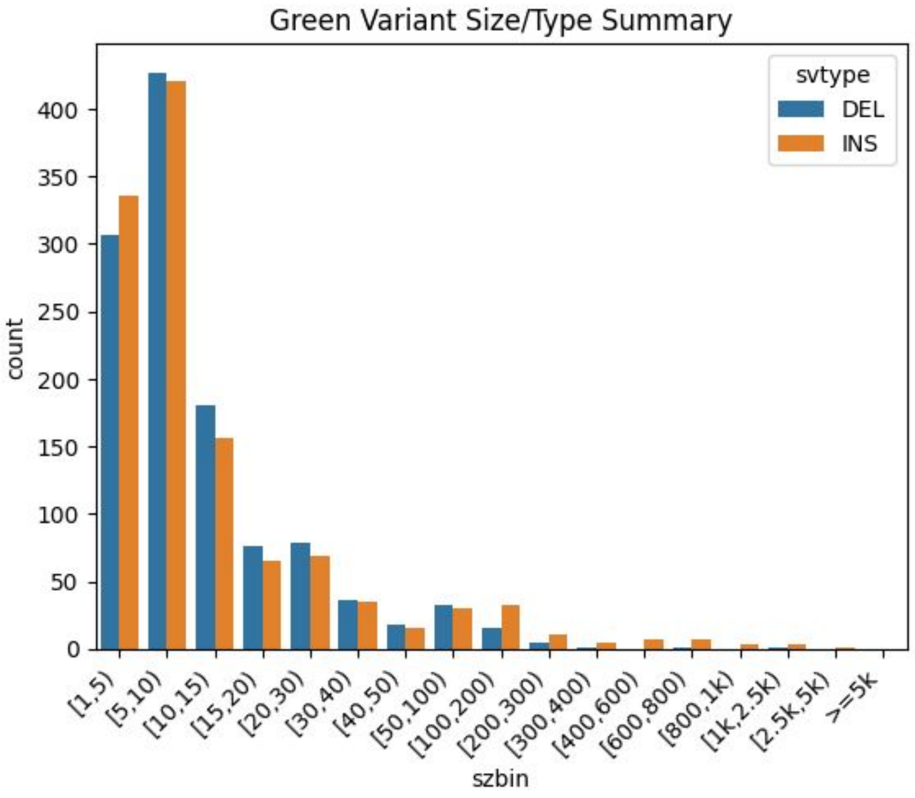
- Found three heuristics that best separate ‘Prob’ TR from ‘Maybe Not’ TR.
    - Is annotated by `truvari anno trf`
    - TRFperiod length > 1
    - Variant Sequence Entropy >= 0.25
- **Assuming** accurate manual inspection, estimate TR-identification performance

TP	1,561
TN	150
FP	19 (13?)
PPV	**0.988**
FPR	**0.112** (**0.08**?)

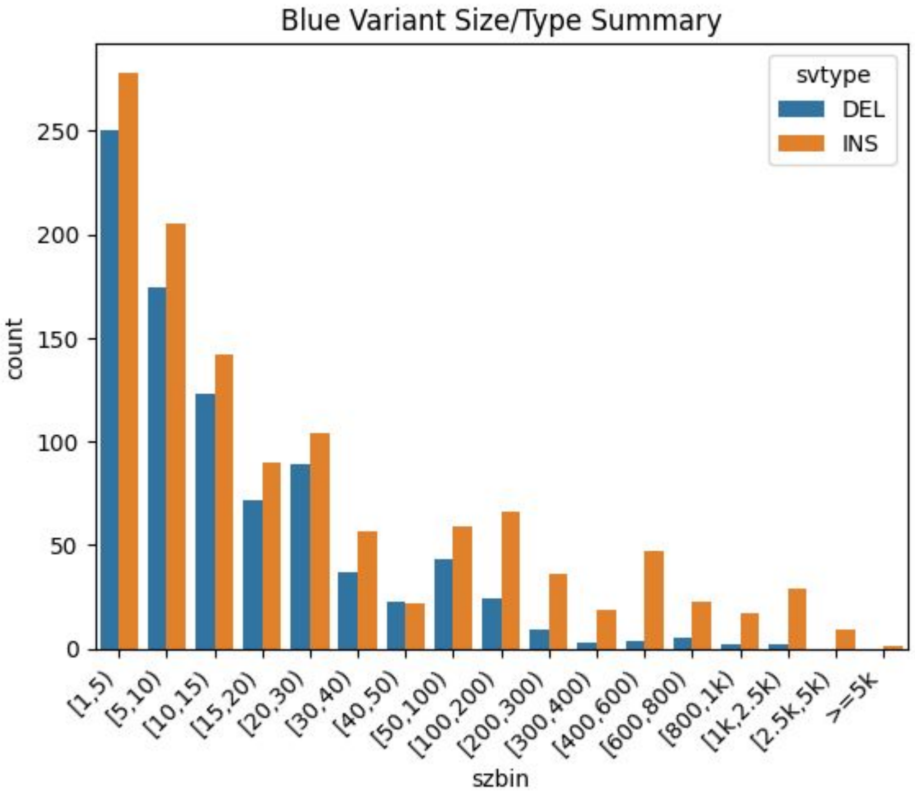


Tiers: Green and Blue

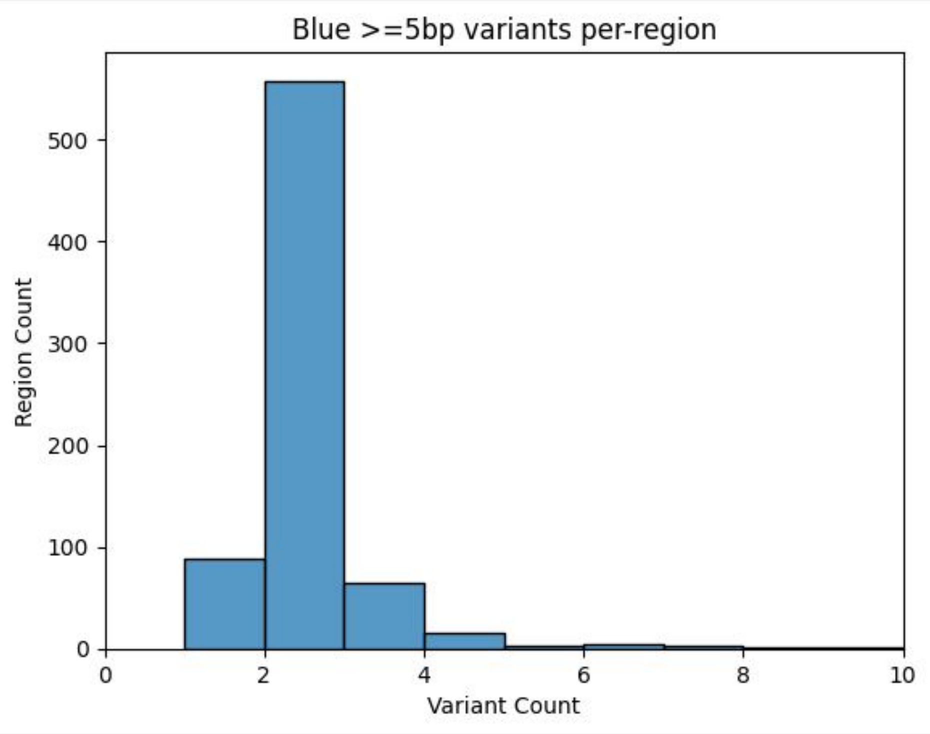
- Of the 37,393 well-covered TRregions on chr20 (92.7%)
  - Green : 1,731 ‘simpler’ regions (4.6%)
  - Blue : 738 ‘complex’ regions (1.9%)
  - Controls : 35,012 have no HG002 variants >= 5bp (95.6%)
    - 28,684 (81.9%) have no variants at all.

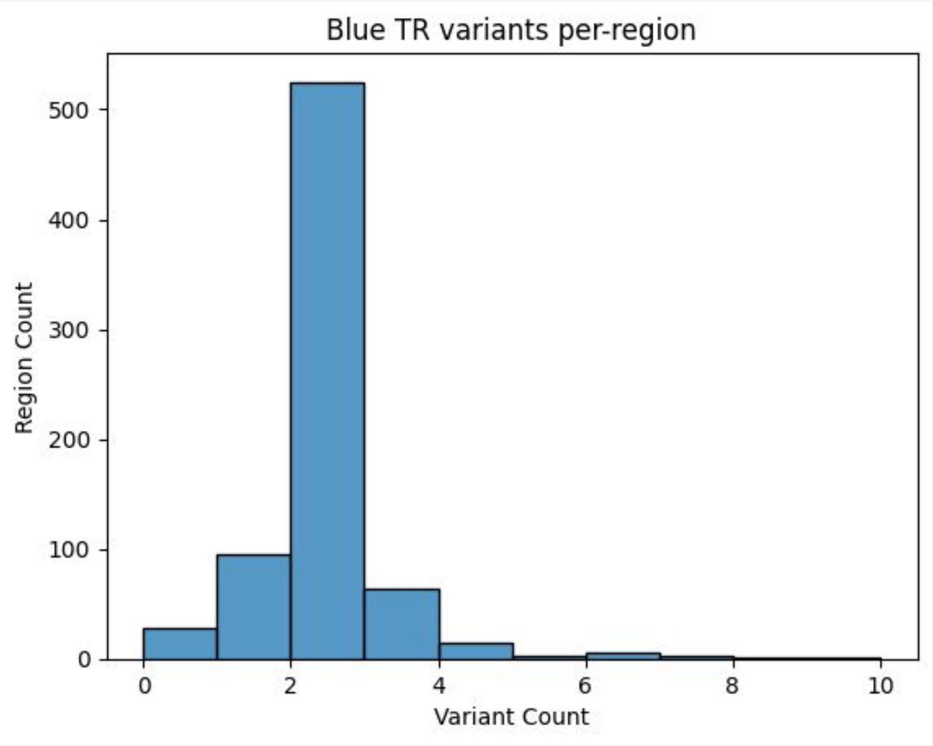


Green Variant Size/Type Summary



Blue Variant Size/Type Summary





- Green regions have one >=5bp variant per-region by definition.
- 764 (44%) have exactly one variant of any size.
- 150 regions’ variant is filtered by the heuristics.

TRGT Benchmarking

Variant Summary

	TP-base	TP-comp	FP	FN	precision	recall	f1	base cnt	comp cnt
Control	5	5	308	0	0.02	1.00	0.03	5	313
Green	1,733	1,787	29	36	0.98	0.98	0.98	1,769	1,816
Blue	1,687	1,657	41	51	0.98	0.97	0.97	1,738	1,698

Region Summary

	TP	TN	FP	FN	base P	base N	comp P	comp N	PPV	TPR	TNR	NPV	ACC	BA	F1
Control	5	35,860	177	0	5	36,037	182	35,860	0.03	1.00	1.00	1.00	1.00	1.00	0.05
Green	1,668	12	25	35	1,719	12	1,695	36	0.98	0.97	1.00	0.33	0.97	0.99	0.98
Blue	674	14	32	33	723	15	719	19	0.94	0.93	0.93	0.74	0.93	0.93	0.93

- Need to resolve variants inside control regions.
- Blue number of variants vs number of regions.

Next Steps

- Evaluate the chr20 strawman benchmark
  - Internal review
  - Focused effort on usability
    - Create documentation on files/tools and tutorials for users.
  - Curating control regions
    - Currently 15:1 baseN to baseP
    - Exclude regions with any HG002 variants?
    - Observed TR exp/con in other samples?
- Assembly realignment?
  - Given that there are a number of Codis sites that the adotto alignment parameters failed to get through, but dipcall's did, should we regenerate the pVCF with dipcall parameters?
- Phab on all TRregions?
  - Explore if phab over the pVCF helps `truvari anno trf` identify expansions / contractions.
- Region 'bleed'
  - Need to explore if closely neighboring regions are causing comparison anomalies.