GIABTR Benchmark

The Final Stretch
(if you exclude writing the manuscript)

Creating Tiers

We’ve been working on the strawman. For the most part, the regions seem good. But there have been a few regions with lower confidence (FN/FP). These low-conf regions broadly fall into:

- Collapsed Hets / Missed Base-Calls
- TR region boundary issues
- Comparison issues

We have two replicates of the HG002 assembly from Eicher/Li. Additionally, we have two alignments of the HG002 HPRC assembly. How often do these agree?

Leveraging assembly/alignment replicates

- Compare the three replicates to the chr20 strawman.
- Intersect `refine.regions.txt` state.

	Eichler	Li	HPRC
TP	2,476	2,472	2,463
TN	35,888	35,883	35,909
FP	34	40	17
FN	35	34	39

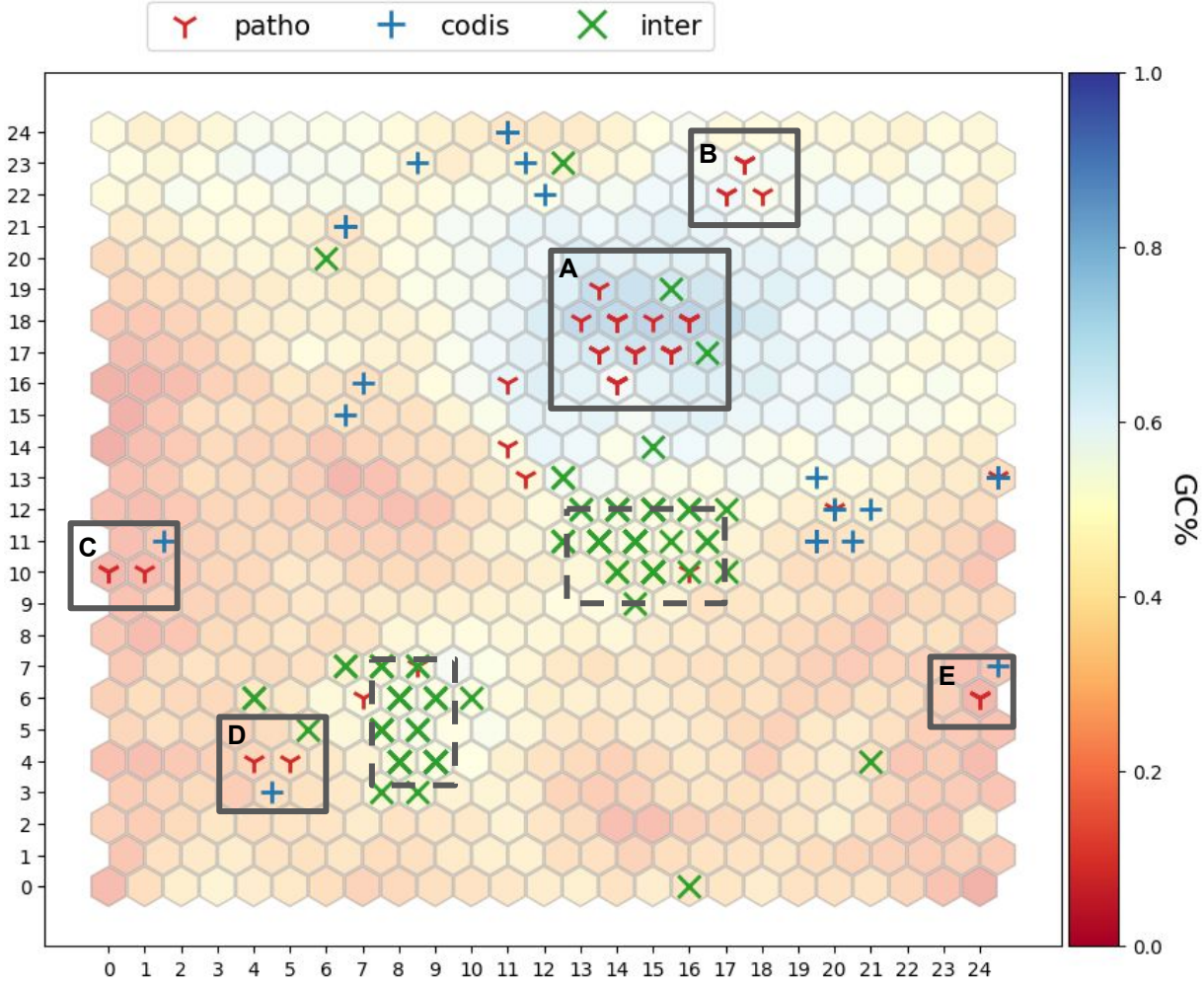
- Pasting the regions together, groupby state LINK. Notes LINK

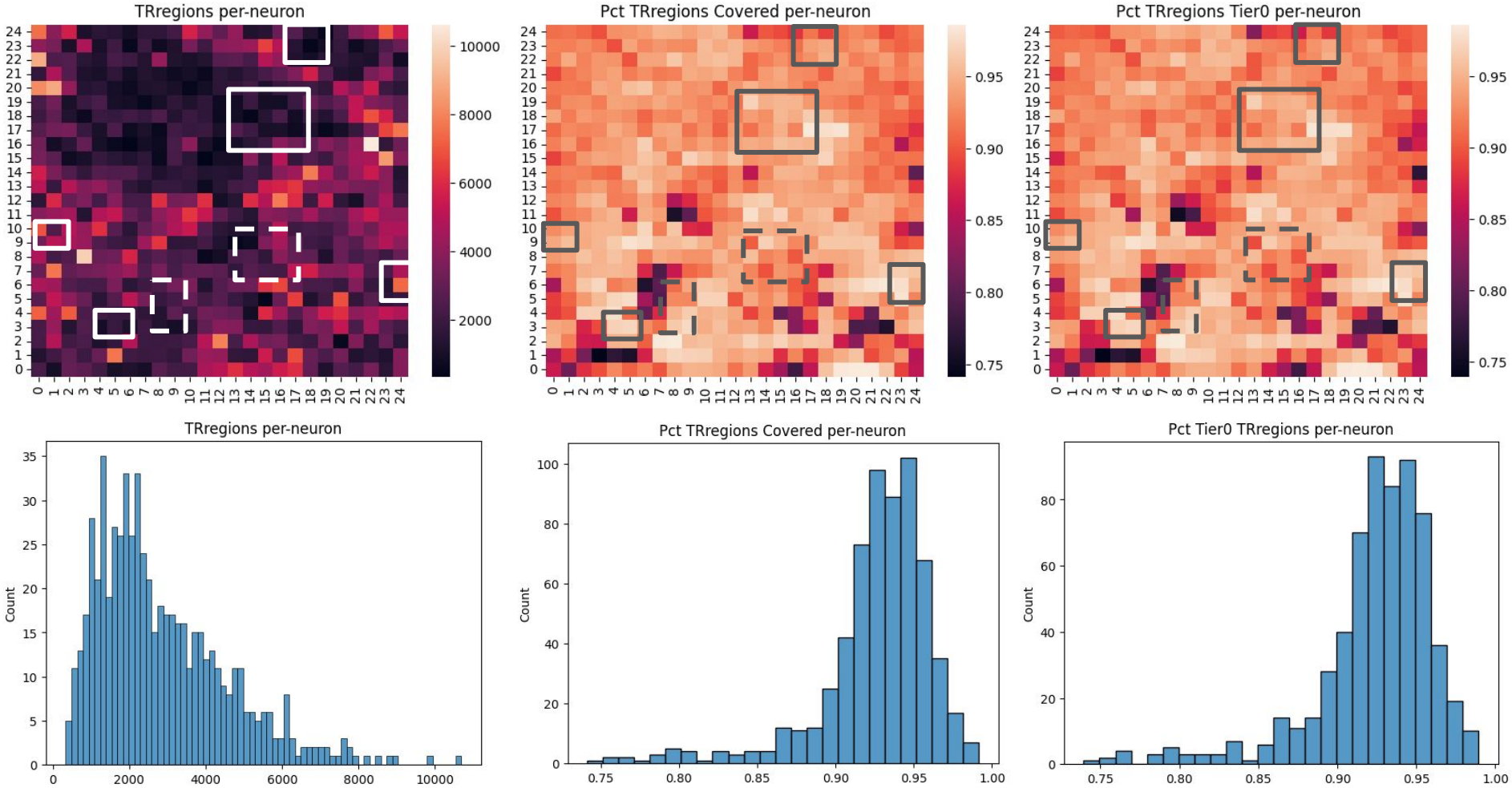
GIAB HG002 TR Benchmark…

		Count	Pct	Mean Entropy
	Tier2	4,879	-	0.870
	Tier0	1,640,577	100%	0.888
	Positives	100,998	6.2%	0.865
	Negatives	1,539,579	93.8%	0.889

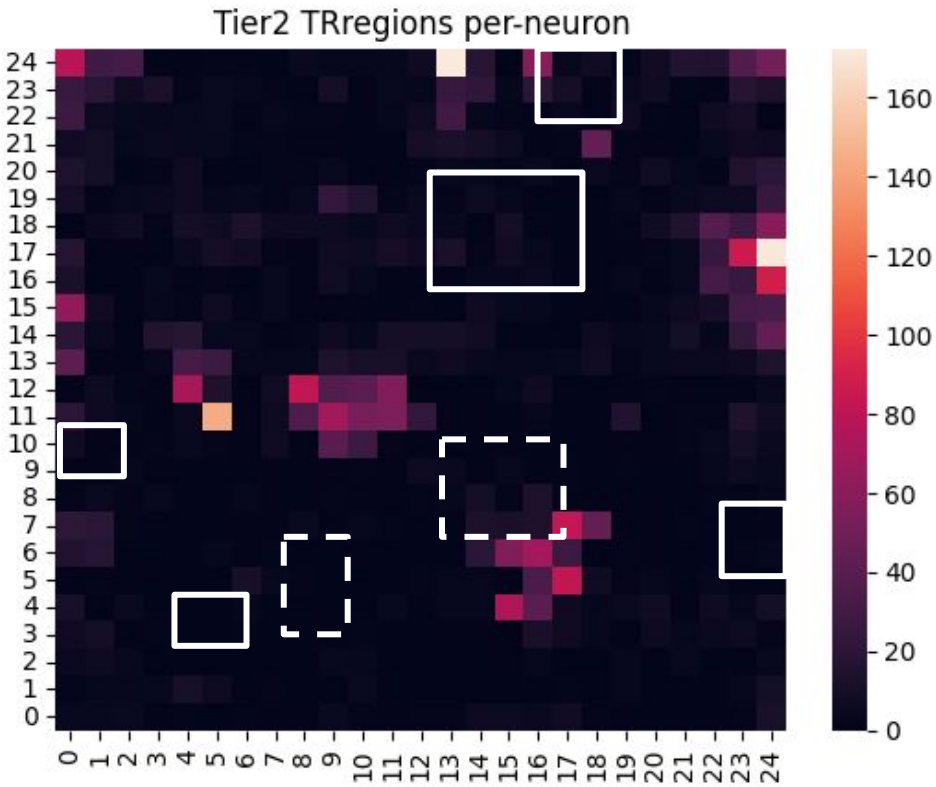
- 54 Patho TRr in 5 neighborhoods
- Interspersed TRr concentrated in two neighborhoods

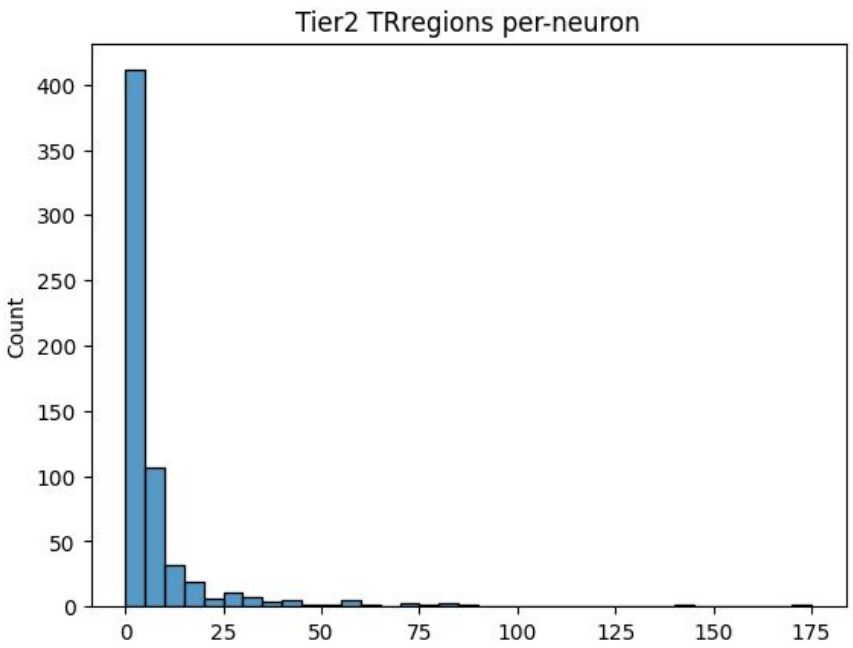
Neighborhood	Motif	Count
A	CGG	10
	CCG	10
	CNG	7
	CTG	7
	GCN	2
	ACCTCGCTGTG CCGCTGCCG	1
	GGCCTG	1
	CGCGGGGCGG GG	1
	CCCCGG	1
B	AGC	6
C	AAAAT	3
D	AAAAG	1
	AAG	1
E	TTTTA	3




TRregions per-neuron
Pct TRregions Covered per-neuron
Pct TRregions Tier0 per-neuron
TRregions per-neuron
Pct TRregions Covered per-neuron
Pct Tier0 TRregions per-neuron
Count
Count
Count

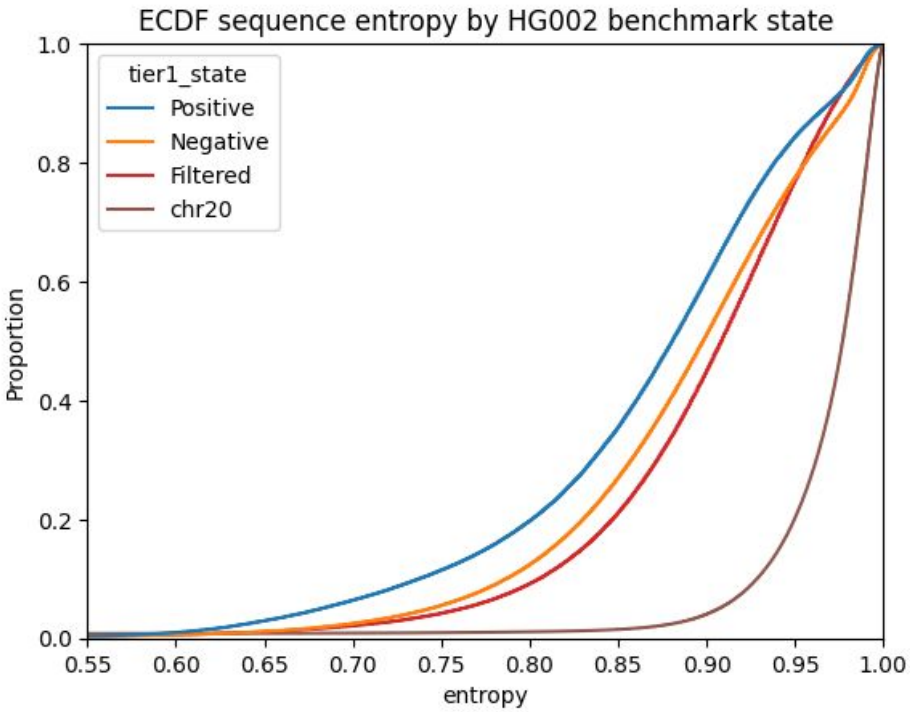
Tier2 Regions (N=4,879)





Intersecting Tier0 with pVCF

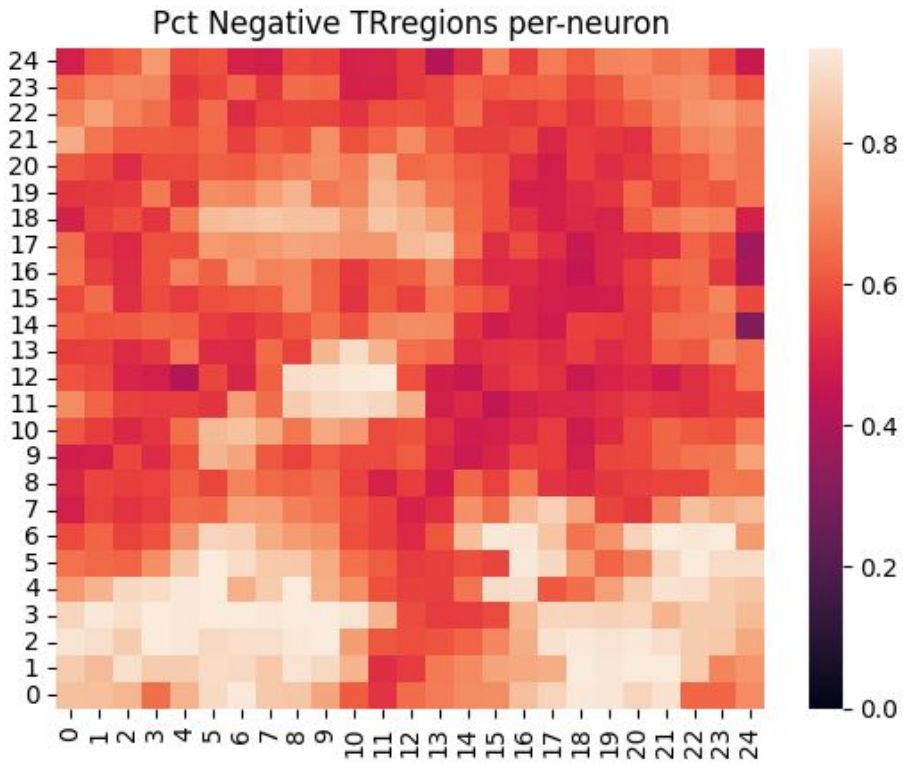
		Count	Pct	Mean Entropy
	Tier2	4,879	-	0.870
	Tier0	1,640,577	100%	0.888
P	**HG002 >= 5**	100,998	6.1%	0.865
N	**Other >= 5**	89,949	5.5%	0.880
	HG002 [0,5)	337,949	20.6%	0.886
	Other [0,5)	653,093	39.5%	0.888
F	**No Var**	458,588	28.0%	**0.895**

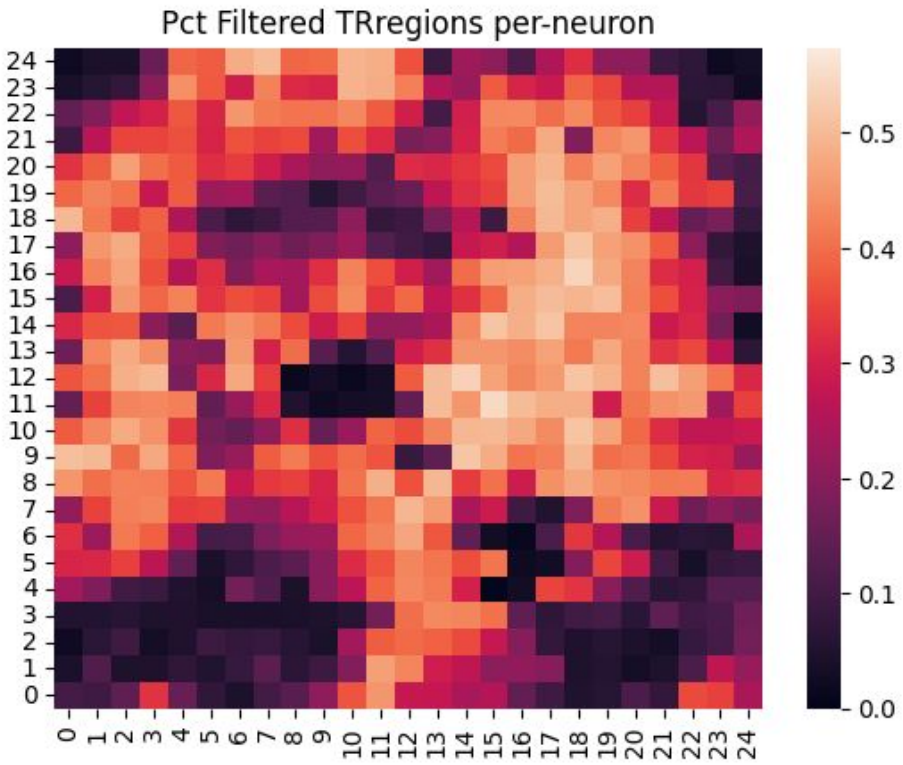


chr20: 200bp sliding window with 50bp step

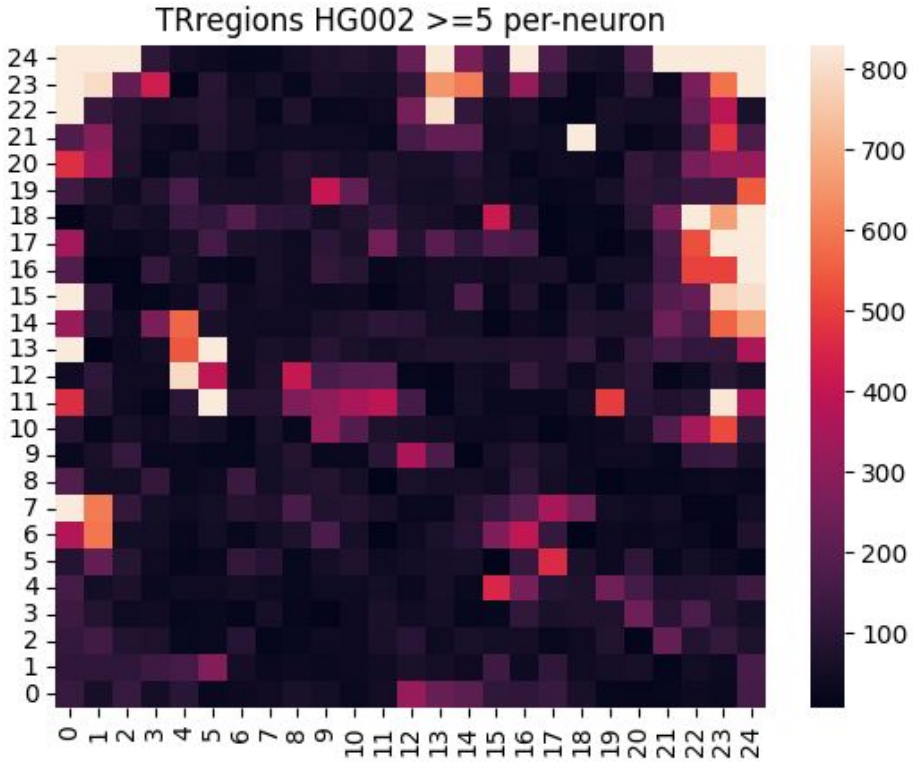
Intersecting Tier0 with dbSNP153 (chr1)

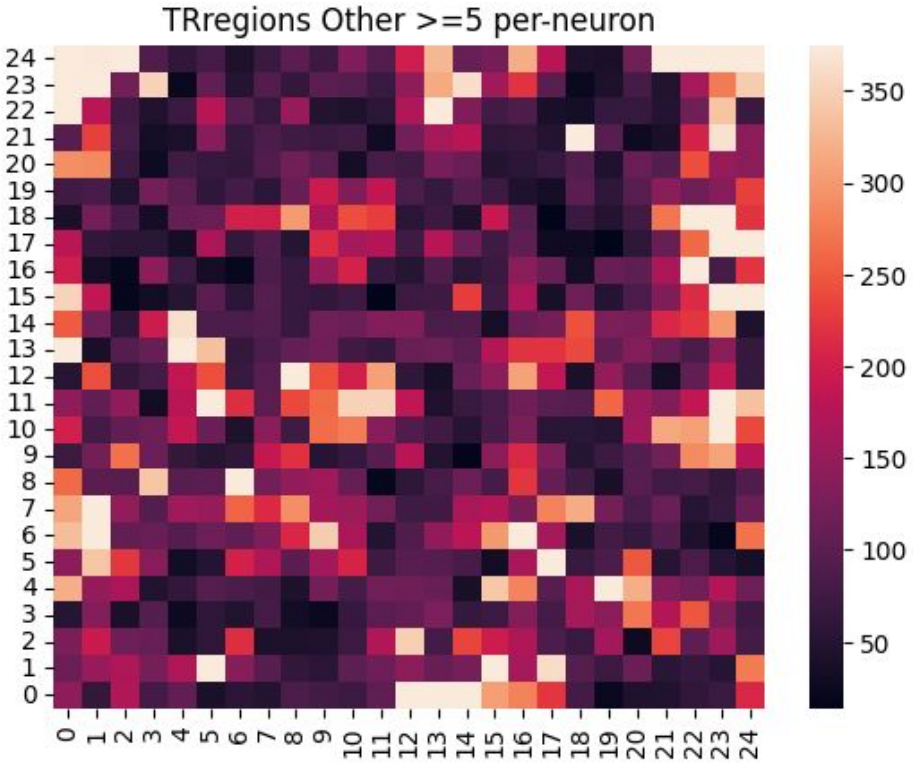
		All		NoVar		AnyVar	
		Count	Pct	Count	Pct	Count	Pct
All	Regions	138,130	-%	39,738	28.77%	98,392	71.23%
	Rare	138,109	99.98%	39,729	99.98%	98,380	99.99%
	Common	74,784	54.14%	3,145	7.91%	71,639	72.81%
	dbSNPs	5,052,879	9.66%	629,545	1.20%	4,423,334	8.46%
	Rare	4,891,034	96.80%	628,619	99.85%	4,262,415	96.36%
	Common	231,670	4.58%	3,430	0.54%	228,240	5.16%
non-SNV							
	Regions	126,772	91.78%	33,275	26.25%	93,497	73.75%
	Rare	125,988	99.38%	33,213	99.81%	92,775	99.23%
	Common	41,282	32.56%	1,153	3.47%	40,129	42.92%
	dbSNPs	1,116,397	24.00%	83,000	1.78%	1,033,397	22.22%
	Rare	1,039,965	93.15%	82,304	99.16%	957,661	92.67%
	Common	97,940	8.77%	1,257	1.51%	96,683	9.36%


Pct Negative TRregions per-neuron
0.8
0.6
0.4
0.2
0.0

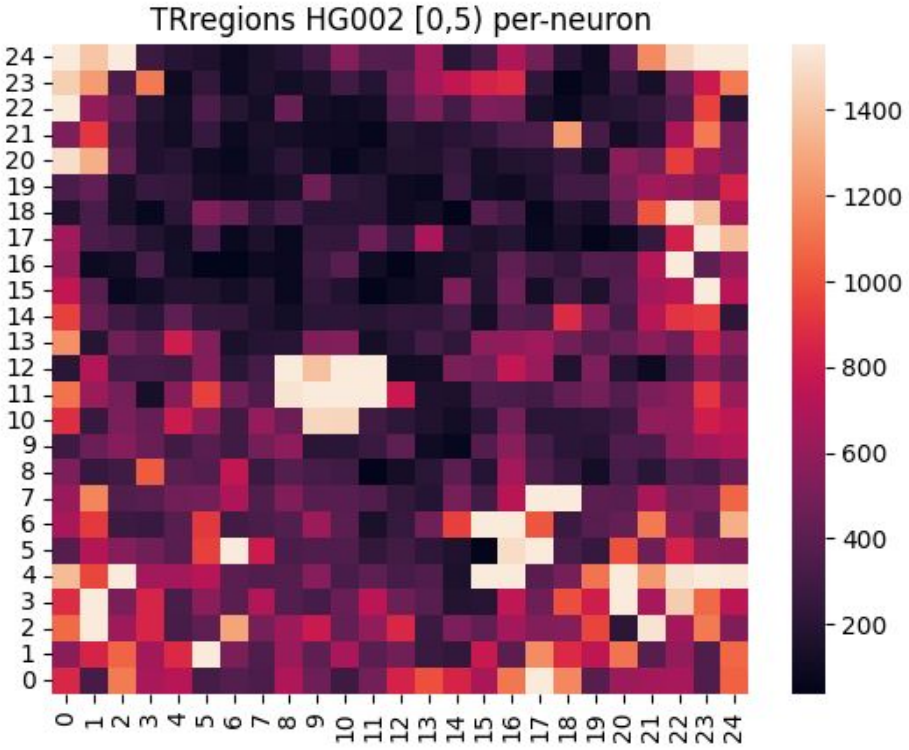

Pct Filtered TRregions per-neuron
0.5
0.4
0.3
0.2
0.1
0.0

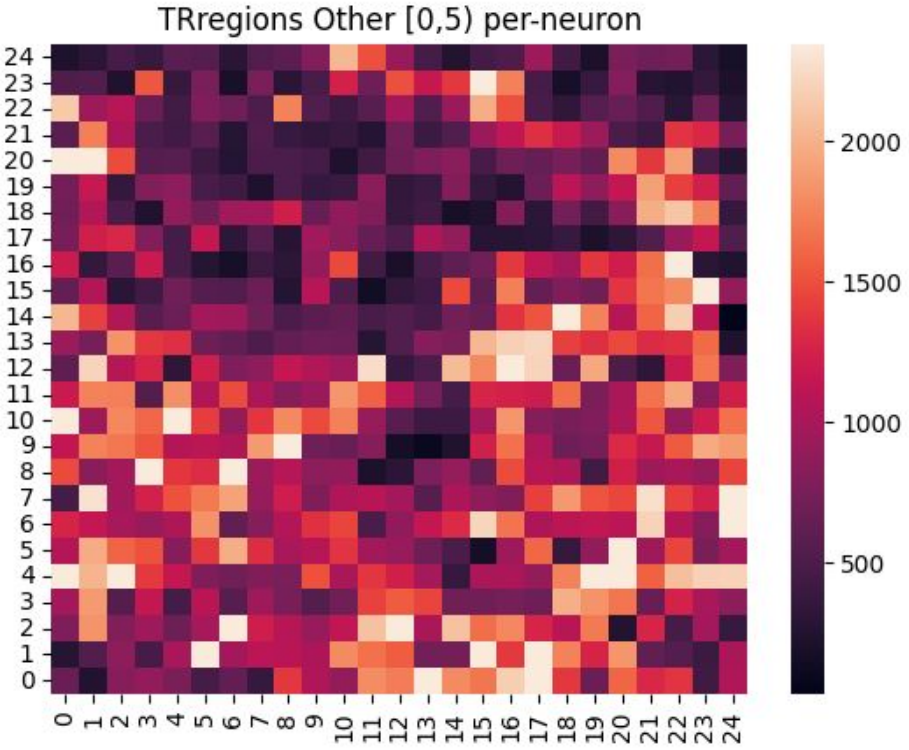
Are Positives representational of TR diversity?





To better visualize if, “Our HG002 has a representative subset of TR variation”, need to redo >=5 SOM with HG002 exclusive, Other exclusive, Shared.


TRregions HG002 [0,5) per-neuron
1400
1200
1000
800
600
400
200


TRregions Other [0,5) per-neuron
2000
1500
1000
500

Summary/Next Steps

- Proposed idea for excluding some Negative regions
  - Just use Tier0 or OK with excluding No Var negatives
  - Whole genome benchmark ready either way
- Have Tiering
  - Evaluate chr20 Tier2 or move to whole genome?
- Working on a stratification tool, which will be ready Laytr
- Manuscript