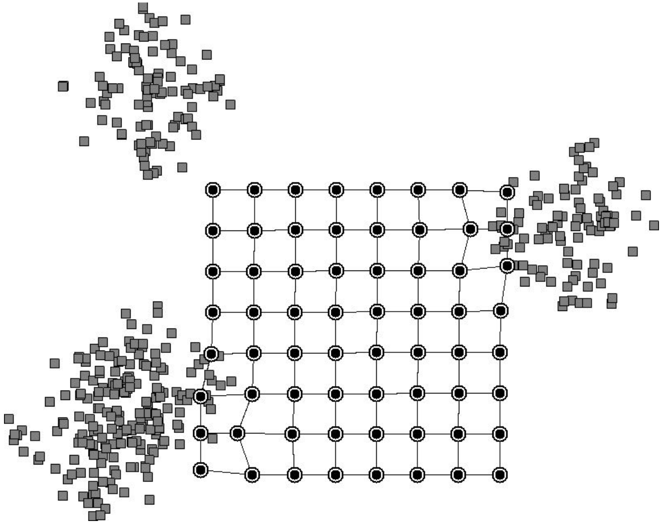
Self-Organizing Maps and Tandem Repeats

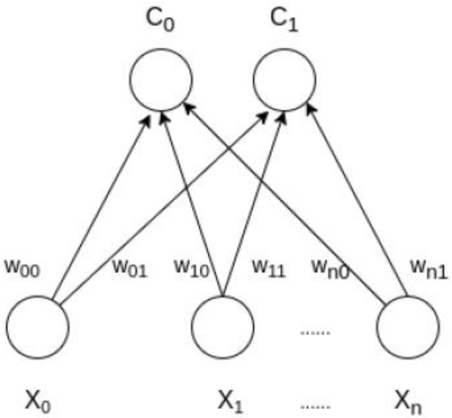
And also a benchmark counting alternative

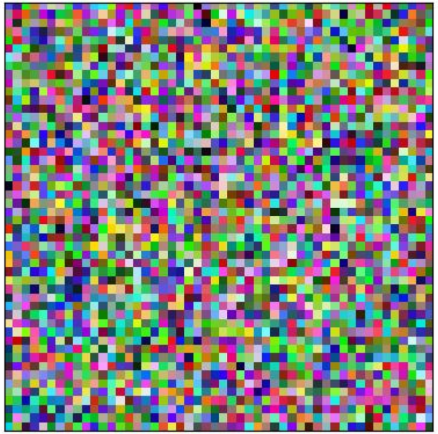
Adam English
Baylor College of Medicine
1/24/2023

Self-Organizing Map

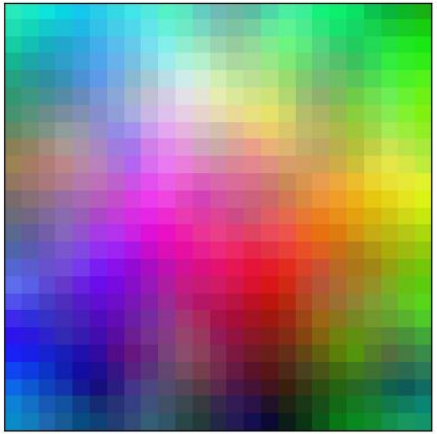


SOM is an unsupervised machine learning technique used to produce a low-dimensional (typically two-dimensional) representation of a higher dimensional data set while preserving the topological structure of the data. - Wikipedia





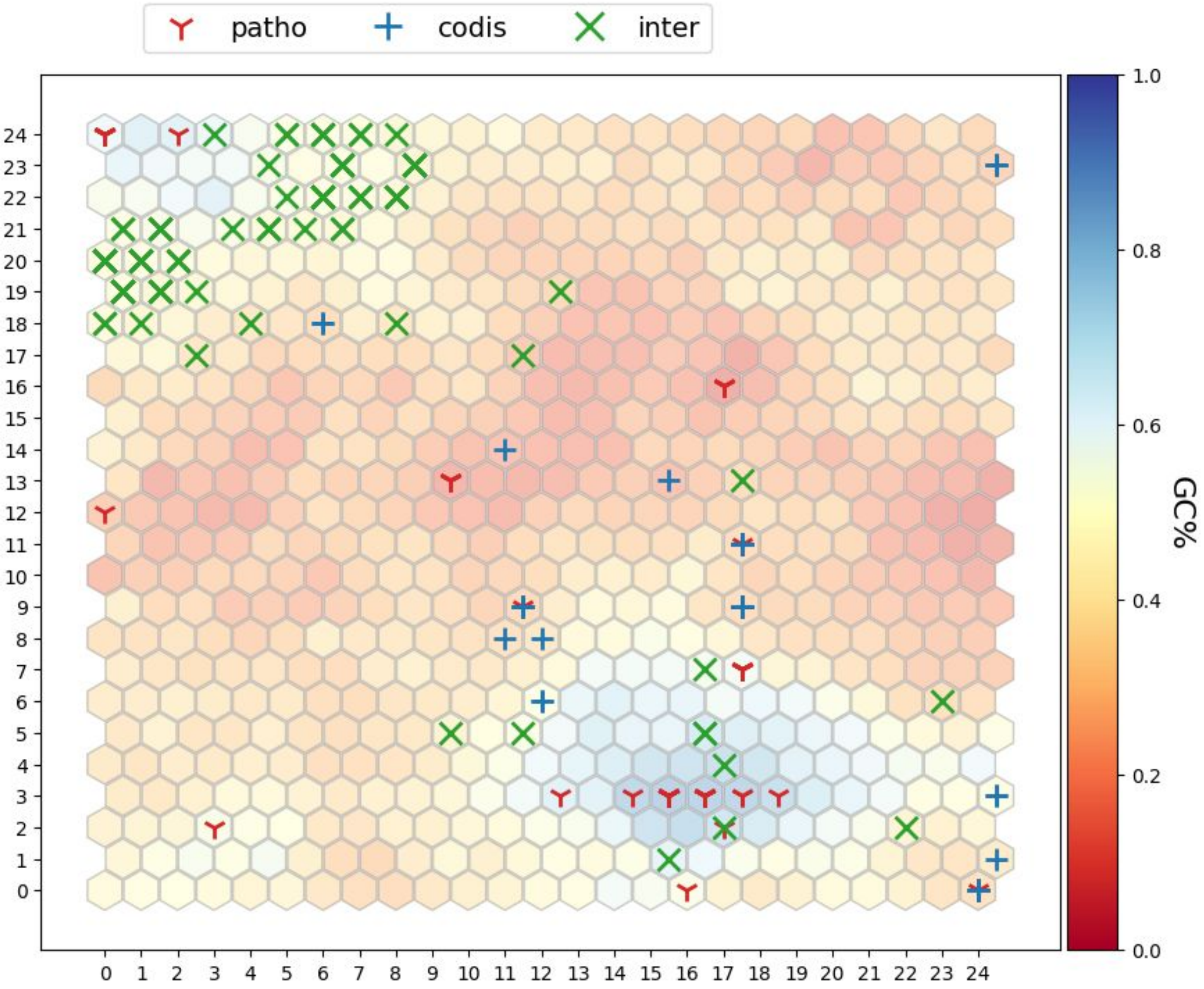
Train on RGB colors
(3 features)



Resulting SOM

SOM with Tandem Repeats

- Built a SOM with Kmer Featurization of sequence spanned by TRregions
  - k=3
  - n_features = 64
  - Frequency normalized
- Hyperparameters:
  - 25x25
  - sigma=1.5
  - learning_rate=1,
  - topology='hexagonal',
  - neighborhood_function='gaussian',
  - activation_distance='euclidean'
- Map Patho, Codis, and 100 randomly sampled ‘interspersed’ TRregions


patho
codis
inter
24
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
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
1.0
0.8
0.6
0.4
0.2
0.0
GC%

Patho Clustering

Cluster A

Num TRs	*Motif*
6	CTG
1	CNG
1	CTA,CTG
1	ACCTCGCTGTGCCGCTGCCG

Cluster B

Num TRs	*Motif*
11	CGG
9	CNG
9	CCG
1	CCCCGG
1	AGCGGCGCGG
1	CGCGGGGCGGGG

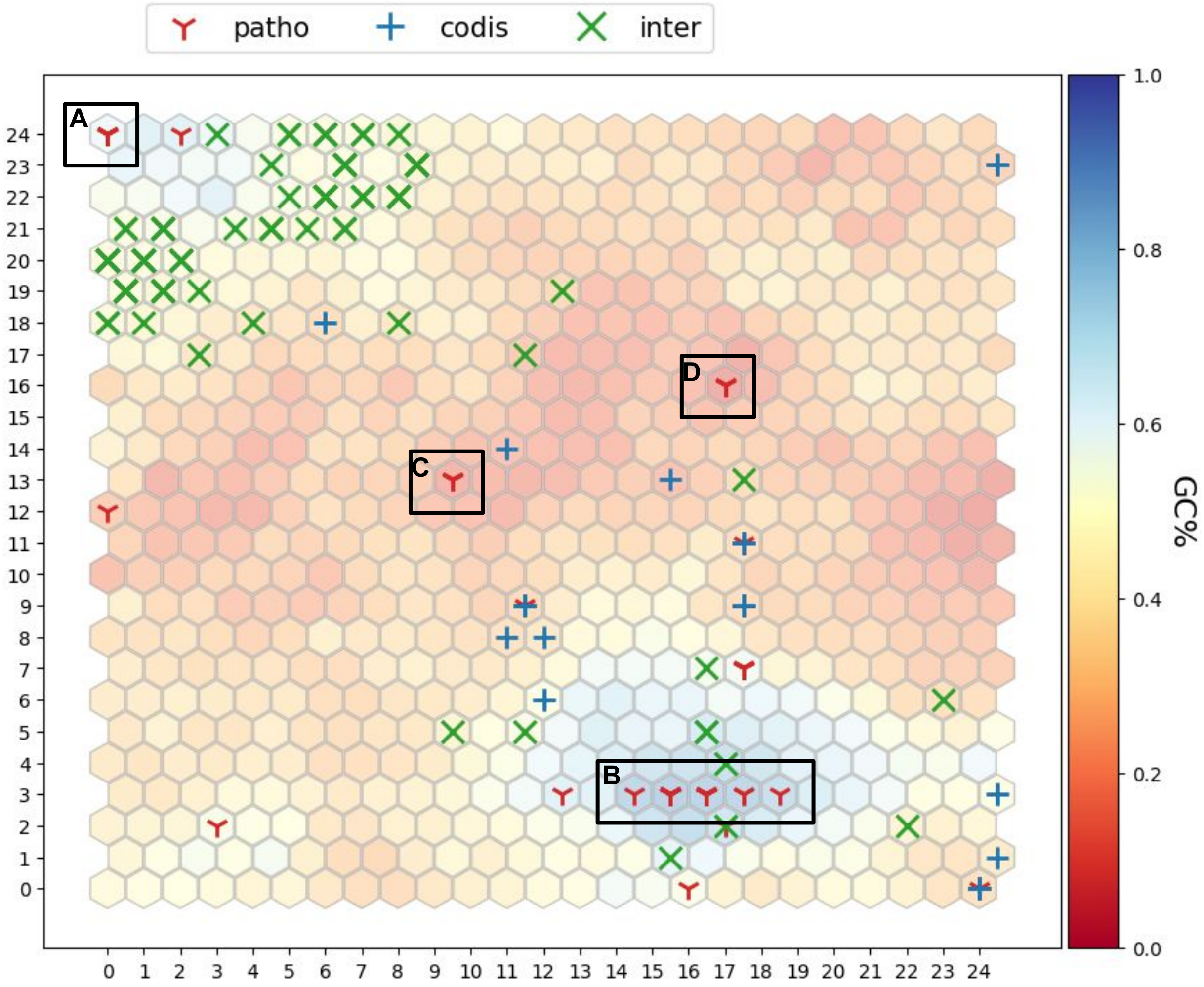
Cluster C

Num TRs	*Motif*
4	TGAAA

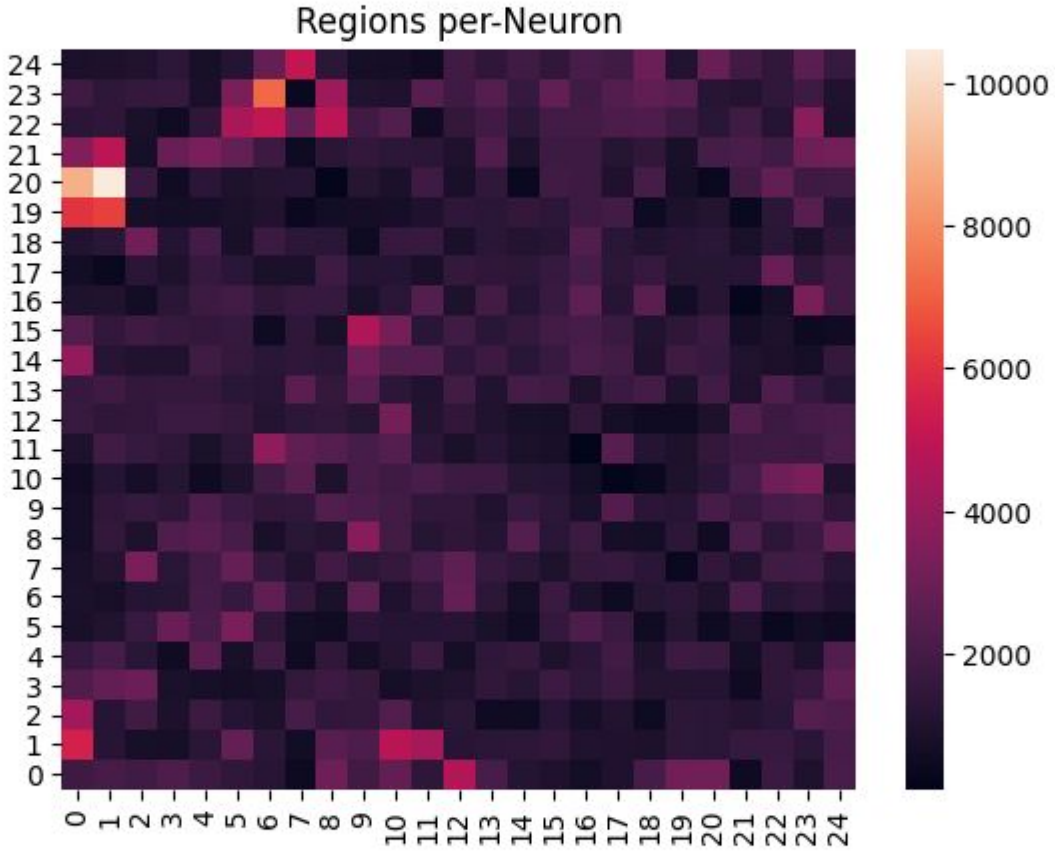
Cluster D

Num TRs	*Motif*
4	TTTTA

- 54 of 65 Pathogenic TRs clustered
- Interspersed TRs are concentrated

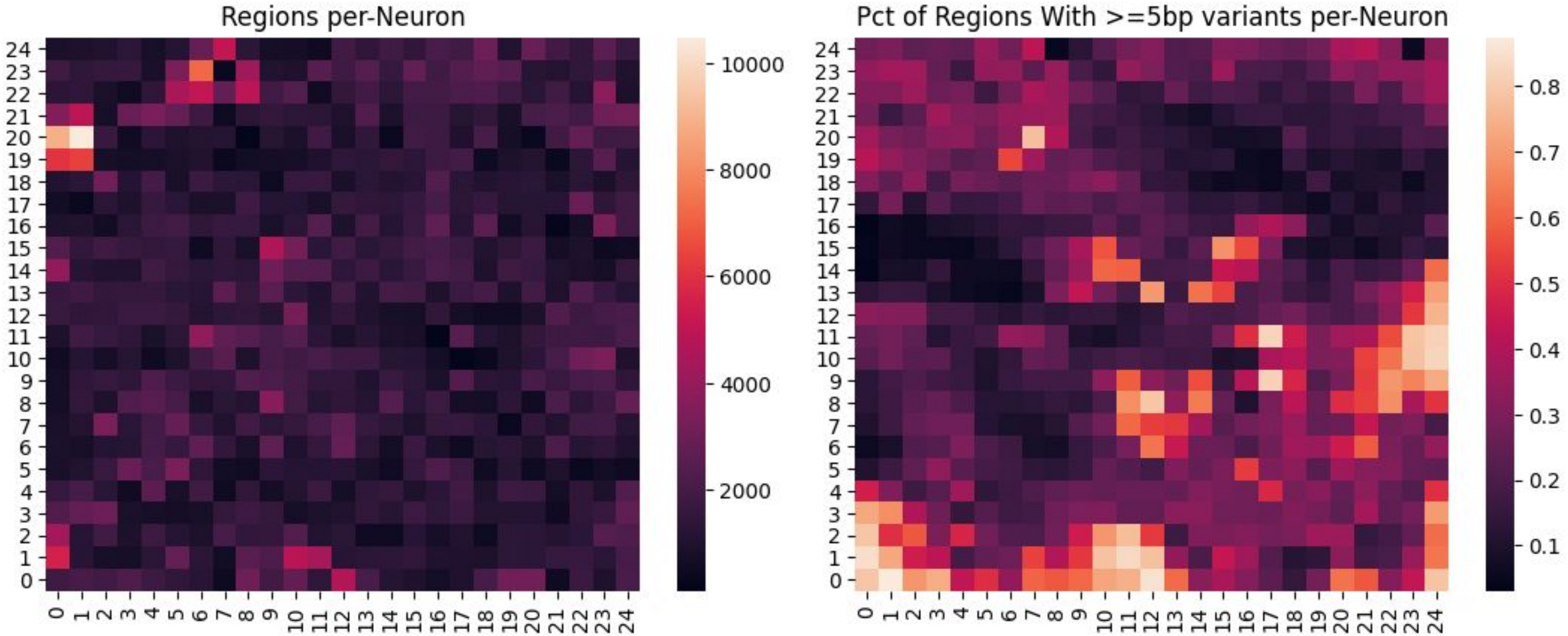


Mapping the 1,784,803* v1.0 TRregions to their neurons.



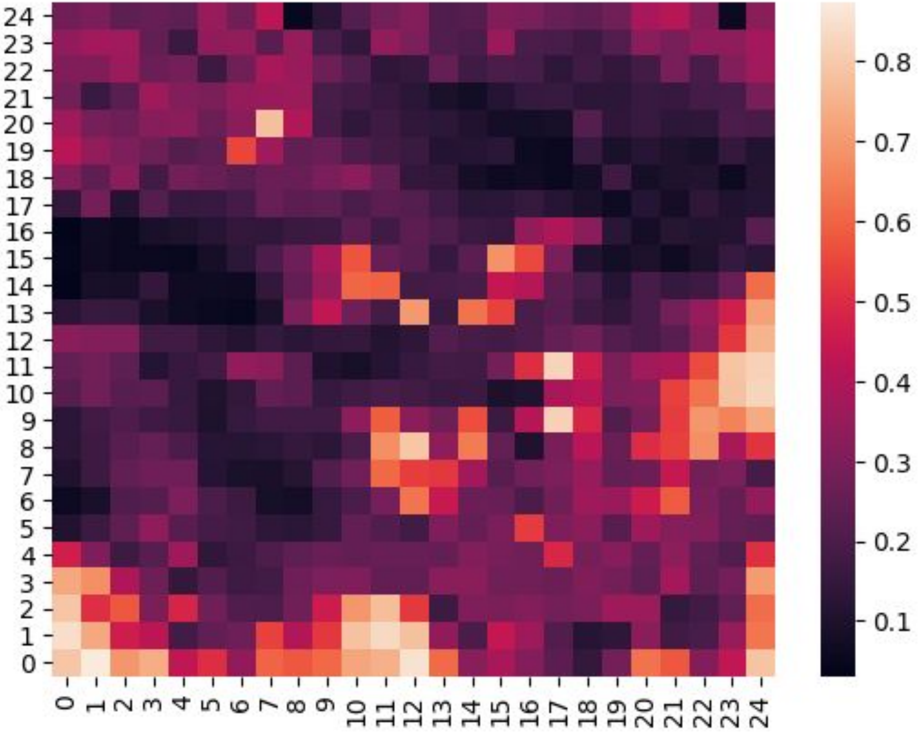
Making a ‘straw man’ benchmark

- Subset to regions with filtered variants from pVCF:
  - Set genotypes of sites without single coverage per-haplotype to missing
  - Remove variants with MISSINGNESS >= 10%
  - Variant must be contained entirely within TRregion boundaries.
  - Variant must be at least 5bp long
- 1,784,804 Regions reduce to 301,262 (16.9%) with filtered variants
- Additional region filters
  - >= 90% annotated → 225,825 (12.7%)
  - Up and Dn buff >= 10bp → 172,771 (9.7%)
  - Subregions <= 2 → 166,588 (9.3%)
  - No interspersed repeat ‘contamination’ → **161,794 (9.1%)**

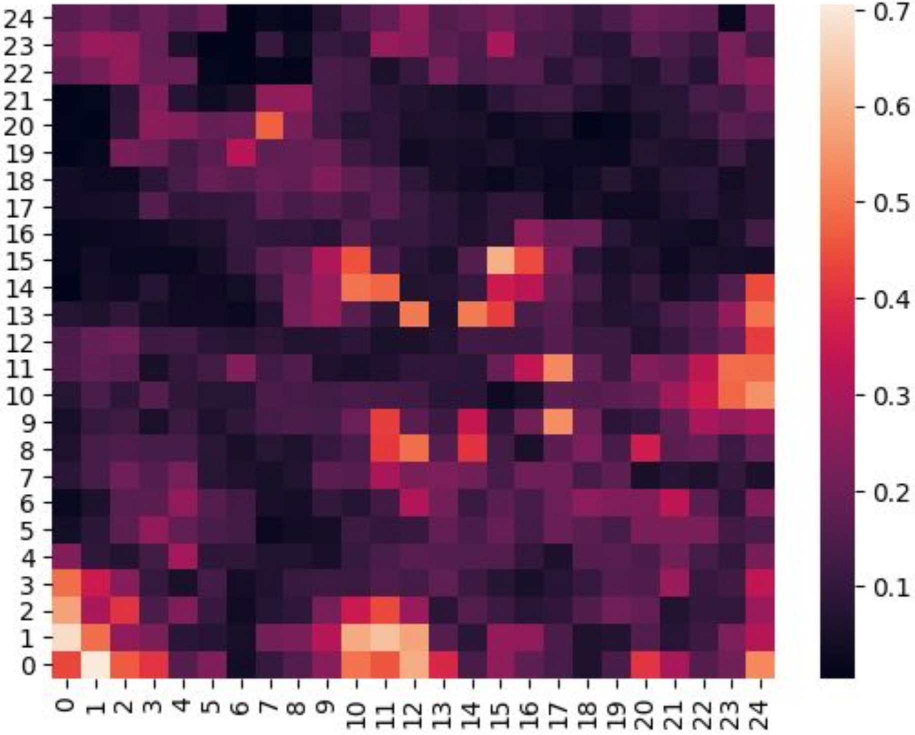


Regions per-Neuron

Pct of Regions With >=5bp variants per-Neuron



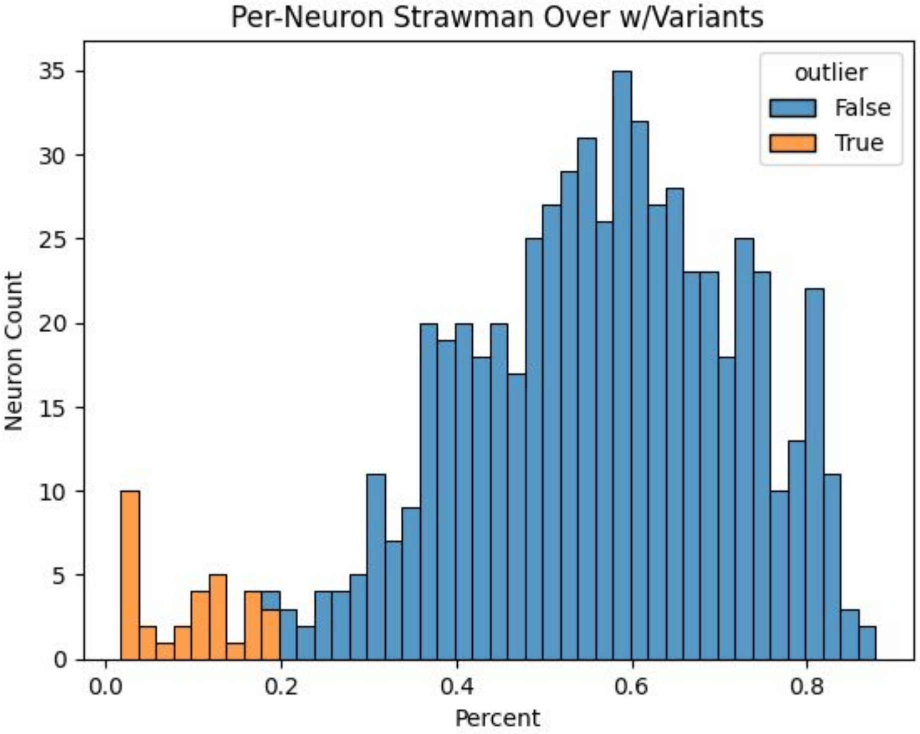
Pct of Regions With >=5bp variants per-Neuron

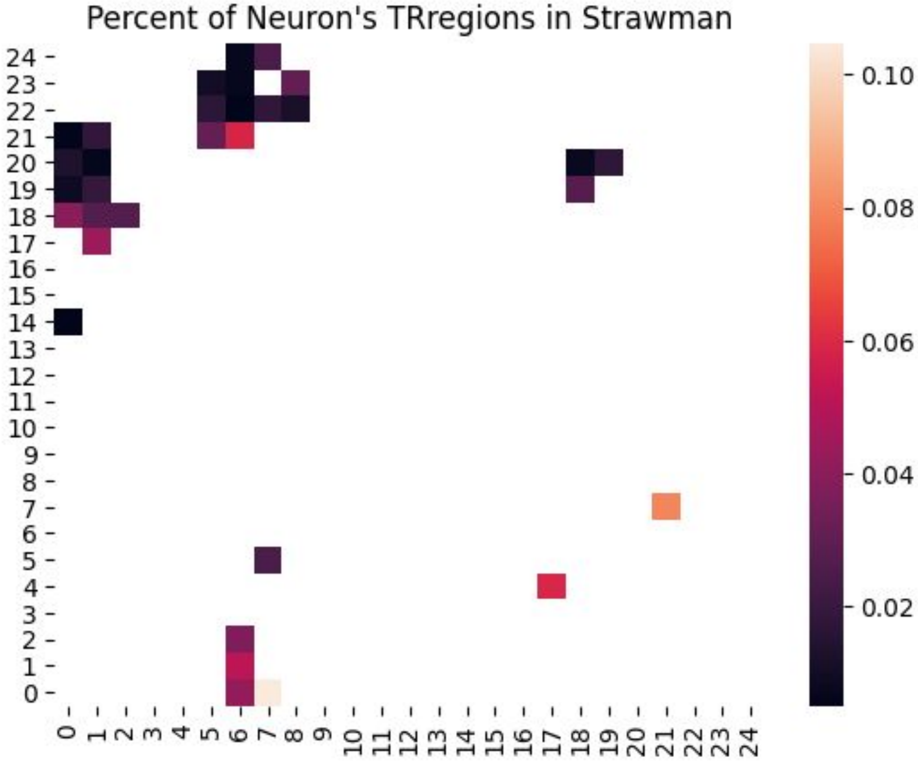


Percent of Neuron's TRregions in Strawman

Identifying Outliers

Of the regions with variants, are their neurons less represented in the straw man benchmark? (bottom 5%)





Bench Result

	TRGT	HipSTR	GangSTR
TP-base	55,454	41,230	27,845
TP-comp	55,595	41,732	27,834
FP	23,627	4,857	2,561
FN	24,990	39,214	52,599
precision	0.702	0.896	0.916
recall	0.689	0.513	0.346
f1	0.696	0.652	0.502
base cnt	80,444	80,444	80,444
comp cnt	79,222	46,589	30,395

Refine Result

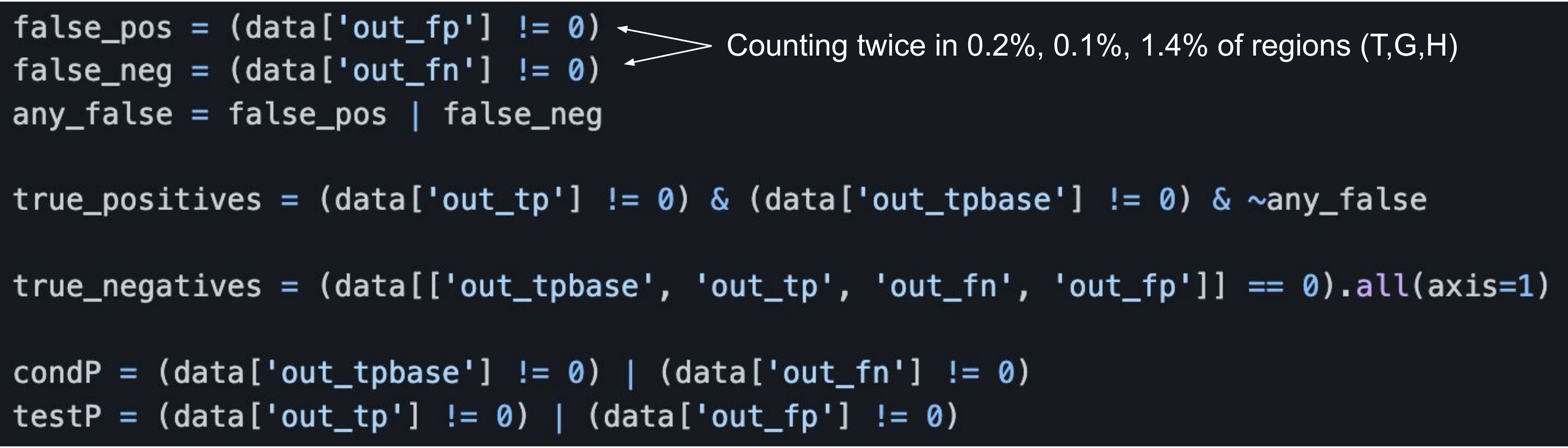
	TRGT	HipSTR	GangSTR
TP-base	81,325	43,542	28,146
TP-comp	81,785	44,057	28,137
FP	5,992	2,256	2,282
FN	3,600	18,434	7,313
precision	0.932	0.951	0.925
recall	0.958	0.703	0.794
f1	0.944	0.808	0.854
base cnt	84,925	61,976	35,459
comp cnt	87,777	46,313	30,419

Refine - Bench Difference

	TRGT	HipSTR	GangSTR
Precision	0.23	0.06	0.01
Recall	0.27	0.19	0.45
F1	0.25	0.16	0.35

Alternative Benchmark Counting

- `truvari phab` can (does) alter the base/comp variant counts.
- `truvari refine` (optionally) subsets `--includebed` to tool’s `--regions`
- This makes comparing performance between tools a little iffy.
- Instead, let’s measure performance on a per-region basis (refine.regions.txt)



```
false_pos = (data['out_fp'] != 0)
false_neg = (data['out_fn'] != 0)
any_false = false_pos | false_neg

true_positives = (data['out_tp'] != 0) & (data['out_tpbase'] != 0) & ~any_false

true_negatives = (data[['out_tpbase', 'out_tp', 'out_fn', 'out_fp']] == 0).all(axis=1)

condP = (data['out_tpbase'] != 0) | (data['out_fn'] != 0)
testP = (data['out_tp'] != 0) | (data['out_fp'] != 0)
```

Counting twice in 0.2%, 0.1%, 1.4% of regions (T,G,H)

Performance - Variants vs Regions

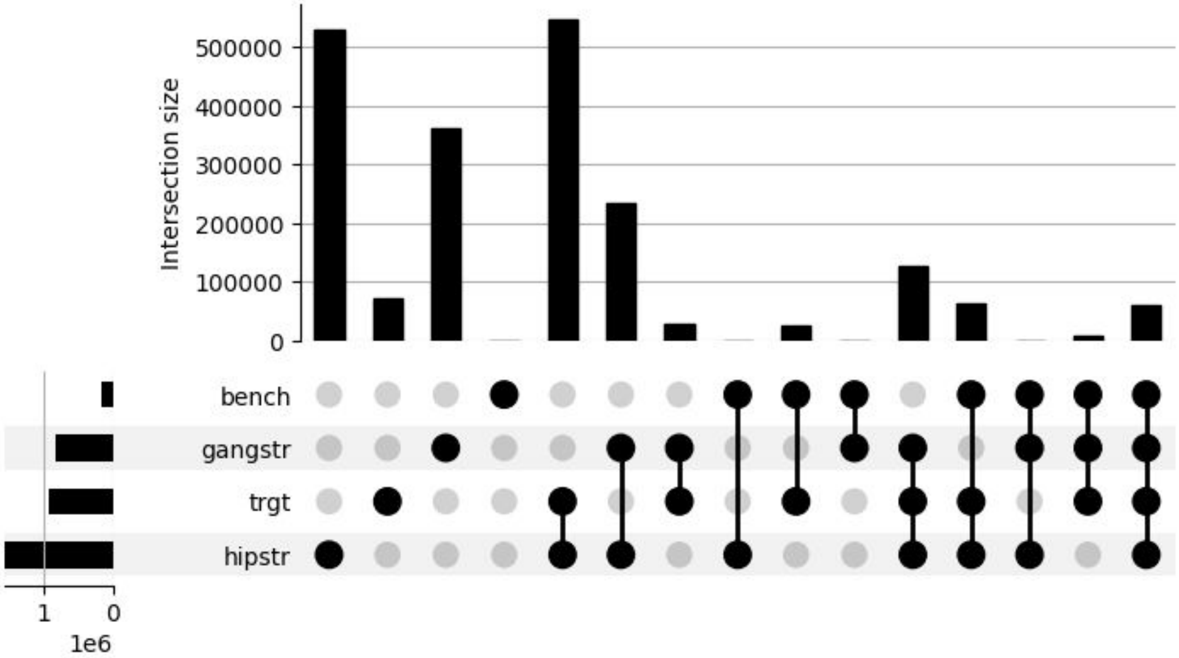
Variant Summary						
	precision	**recall**	.	.	.	**f1**
TRGT	0.932	0.958	.	.	.	0.944
HipSTR	0.951	0.703	.	.	.	0.808
GangSTR	0.925	0.794	.	.	.	0.854
Region Summary						
	PPV	**TPR**	**TNR**	**NPV**	**ACC**	**F1**
TRGT	0.905	0.934	0.965	0.984	0.953	0.919
HipSTR	0.931	0.746	0.980	0.872	0.890	0.828
GangSTR	0.891	0.811	0.980	0.925	0.913	0.849

Disjointed Regions

TR callers typically define their own sets of reference regions to evaluate.

These regions may not intersect with our benchmark regions.

Upset Plot of Beds' Intersection



Regions investigated by a tool but not in the benchmark are ignored by `--includebed`

Regions included in benchmark but not investigated by a tool create False Negatives (Lower Recall)

Performance of regions analyzed by all tools

Summary of 59,577 Shared Regions

	TRGT	HipSTR	GangSTR
TP	**22,636**	20,742	20,452
TN	35,067	**35,335**	35,135
FP	1,165	**699**	1,385
FN	**796**	2,878	3,134
base P	23,766	23,792	23,796
base N	35,811	35,785	35,781
comp P	24,002	21,607	22,231
comp N	35,575	37,970	37,346
PPV	0.943	**0.960**	0.920
TPR	**0.952**	0.872	0.859
TNR	0.979	**0.987**	0.982
NPV	**0.986**	0.931	0.941
ACC	**0.969**	0.941	0.933
BA	**0.966**	0.930	0.921
F1	**0.948**	0.914	0.889

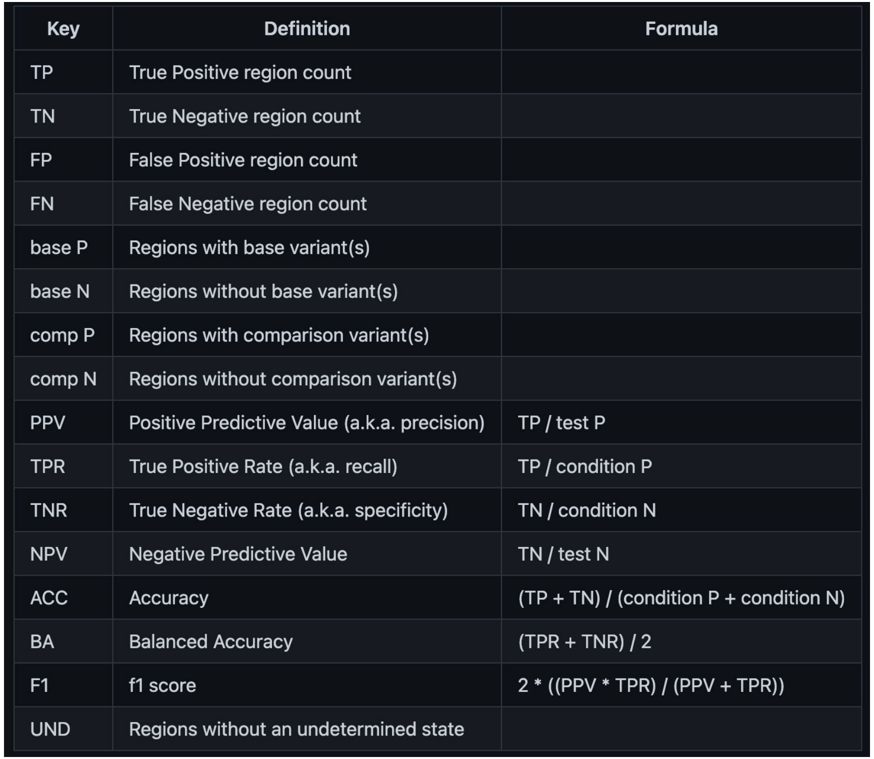
- 36.8% of Strawman analyzed by all tools

- Still have counting issues (e.g. P/N)
    - Phab is sometimes resizing variants to < sizemin
        - `refine --use-original`
    - Phab is moving some variants out of their regions?
    - A few msa2vcf are failing
- But these edge cases becoming less frequent
    - 30 regions with baseP difference between TRGT and HipSTR
- Note: These tools are trying to discover different variant types. By comparing their union, we are somewhat biasing to the most ‘focused’ tool’s area.

Next Steps

- Can we build a SOM out of TRregions features?
  - Could help creating/visualizing stratifications
- Strawman works, but we can do better
  - Checking that regions are covered by HPRC HG002 assembly
  - >=5bp != TR expansion/contraction. Work on `truvari anno trf`.
- Closer to cutting a Truvari v4.0 release

Region Summary Definitions



Key	Definition	Formula
TP	True Positive region count	
TN	True Negative region count	
FP	False Positive region count	
FN	False Negative region count	
base P	Regions with base variant(s)	
base N	Regions without base variant(s)	
comp P	Regions with comparison variant(s)	
comp N	Regions without comparison variant(s)	
PPV	Positive Predictive Value (a.k.a. precision)	TP / test P
TPR	True Positive Rate (a.k.a. recall)	TP / condition P
TNR	True Negative Rate (a.k.a. specificity)	TN / condition N
NPV	Negative Predictive Value	TN / test N
ACC	Accuracy	(TP + TN) / (condition P + condition N)
BA	Balanced Accuracy	(TPR + TNR) / 2
F1	f1 score	2 * ((PPV * TPR) / (PPV + TPR))
UND	Regions without an undetermined state	