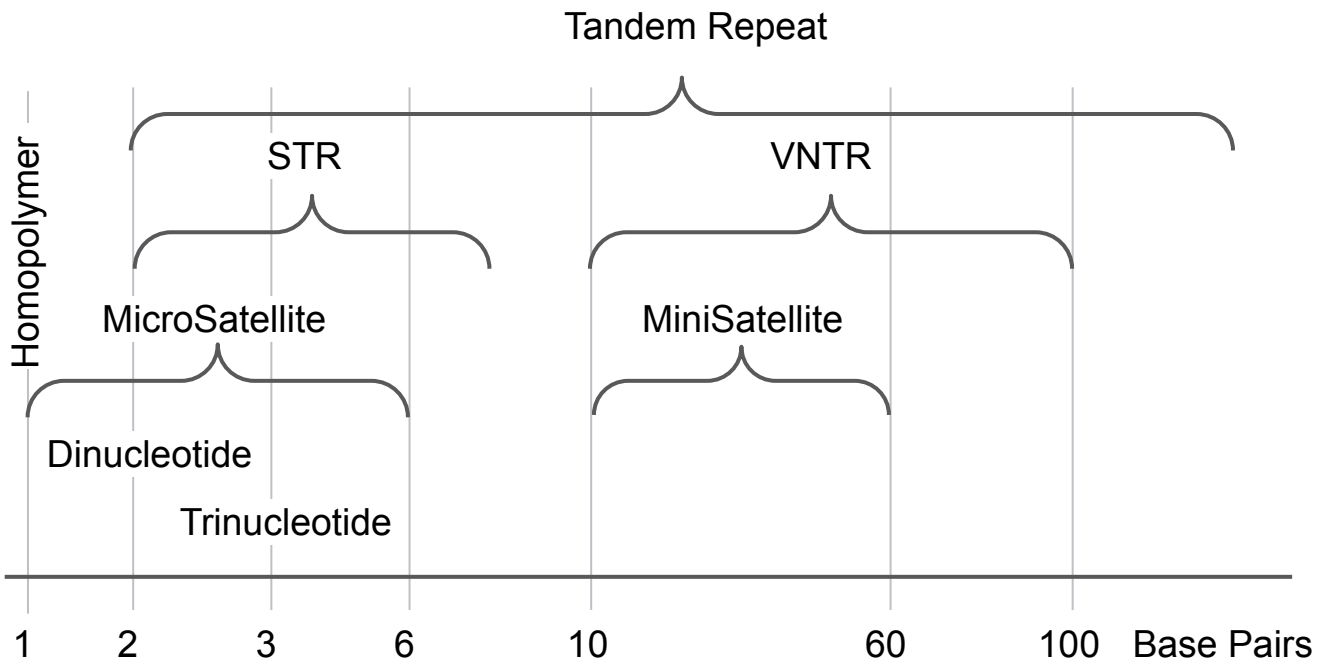
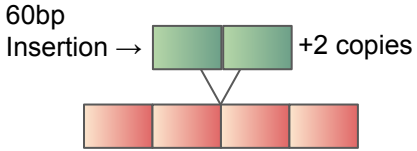
GIAB Tandem Repeat Benchmark

Adam English
Human Genome Sequencing Center
Baylor College of Medicine

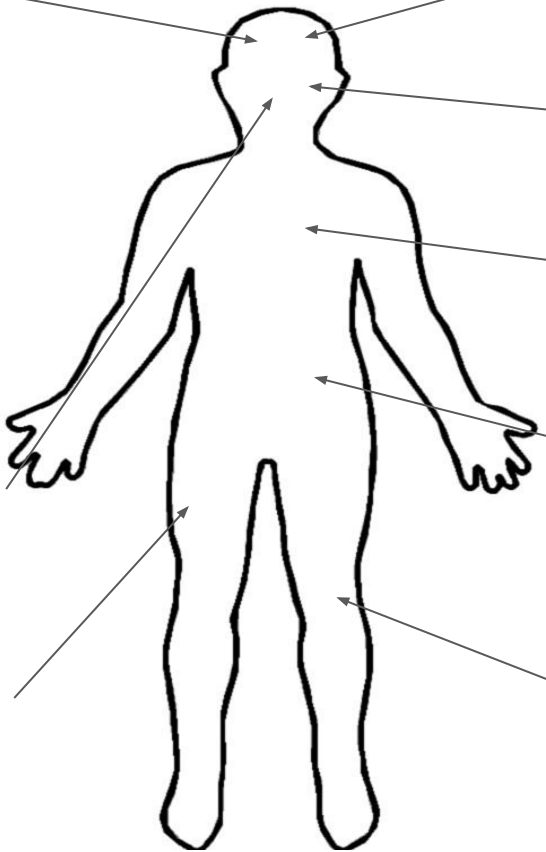




Repeat Motif >= 2bp
Variant Sequence (INDEL) >= 5bp

An **insertion/deletion** represents repeat **expansion/contraction**

Known Pathogenic Tandem Repeats



Neurological Disorders

Amyotrophic lateral sclerosis and/or frontotemporal dementia

Dentatorubral-pallidoluysian atrophy

Episodic ataxia

Friedreich ataxia

Hereditary essential tremor type 6

Huntington's disease

Myotonic dystrophy 1

Neuronal intranuclear inclusion disease

Spinocerebellar ataxia

Spinal and bulbar muscular atrophy

Panhypopituitarism and Growth Hormone Deficiency

X-linked mental retardation with isolated growth hormone

X-linked panhypopituitarism

Muscular Dystrophies

Duchenne muscular dystrophy

Oculopharyngeal muscular dystrophy

Epilepsy and Seizure Disorders

Developmental and epileptic encephalopathy

Familial adult myoclonic epilepsy

Mental retardation, FRA12A type

Ophthalmological Disorders

Blepharophimosis, ptosis, and epicanthus inversus syndrome

Fuchs endothelial corneal dystrophy-3

Cardiovascular Disorders

Tetralogy of Fallot

Myopathy

CANVAS syndrome

Cerebellar ataxia, neuropathy, and vestibular areflexia syndrome

Oculopharyngodistal myopathy 1

Connective Tissue Disorders

Cleidocranial dysplasia

Desbuquois dysplasia-2

Multiple epiphyseal dysplasia

Pseudoachondroplasia

Synpolydactyly

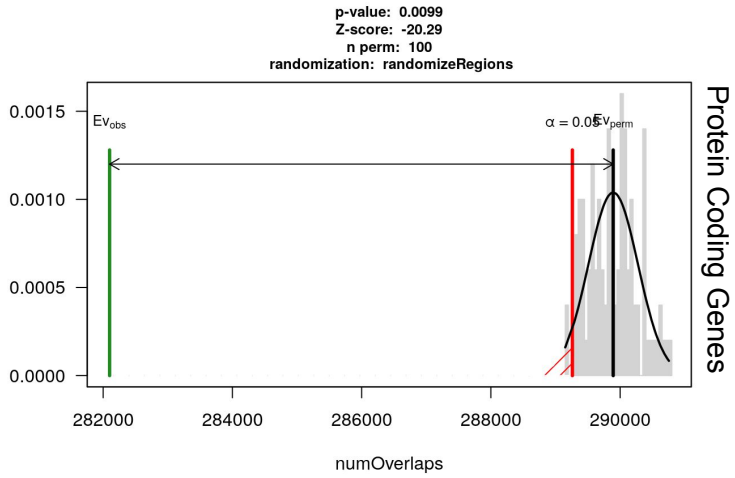
*** Deduped/Grouped by ChatGPT**

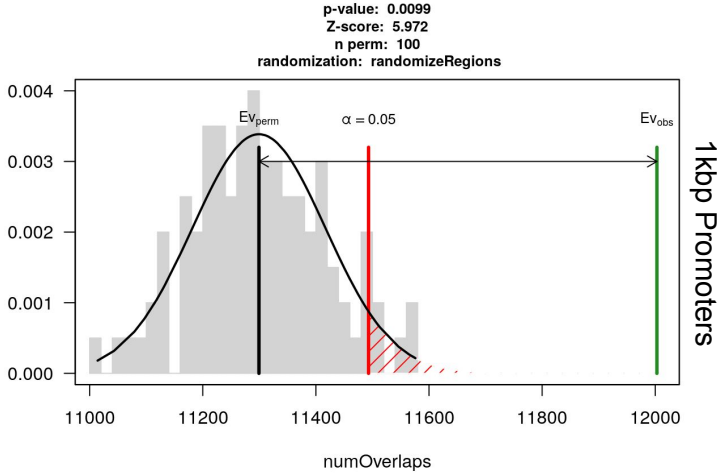
Interesting TR properties

- <10% of GRCh38 is tandem repeats
- Fewer TRs in Protein Coding Regions
- More TRs in Protein Coding Promoters
- TRs are highly polymorphic

Variant counts from 86 long-read assemblies

Size	Total	Within TR	Percent
SNP	106,410,565	19,448,255	18.28%
[1,5)	14,421,648	3,750,268	26.00%
[5,50)	3,138,977	2,192,451	69.85%
SV	763,953	567,553	74.29%





— UCSC Simple Repeat Track <50kb
– GENCODEV43 (1kbp for promoters) – RegionR

TR Benchmark Goals

Build a benchmark for Tandem Repeats in HG002

Three main components:

- Regions - Where are tandem repeats in the reference
- Variants - Which regions have tandem repeats expansions/contractions
- Tools - How to compare variants

TR Regions

The GIAB TR group provided ‘seeds’ for tandem repeat regions.

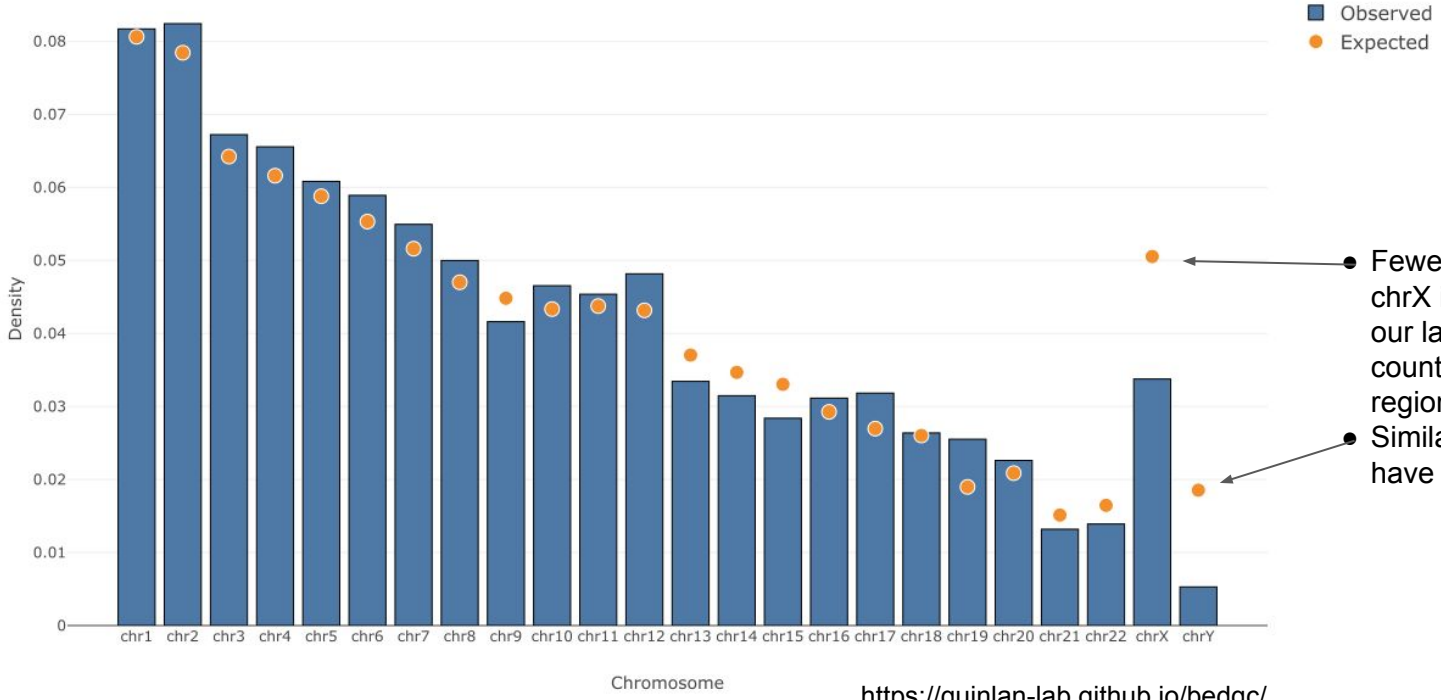
Processing:

- Span between 10bp, 50kbp
- Add 25bp slop, intra-merge bed
- Inter-merge bed files
- Run TRF/RepeatMasker
- Simplify annotations

Name	Source	Count	~Genome Cov
GIAB	FTP	1,269,585	4.42%
Baylor	UCSC Genome Browser	652,137	2.54%
UCSC1	Ensembl	1,738,251	1.20%
UCSC2	Github	10,259	0.02%
TRGT	Github	163,400	0.15%
pbsv	Github	316,944	4.55%
Illumina	Github	163,355	0.15%
USC	Pre-print	467,104	1.62%
pCGG	Paper	5,765	0.01%

TR Regions Summary

Total of 1,784,803 regions spanning 237.9 Mbp. ~8% of GRCh38.



- Fewer observed regions in chrX is partially attributable to our largest ‘seed’ source by count did not have chrX regions.
- Similarly, 5 sources did not have chrY regions.

https://quinlan-lab.github.io/bedqc/

Tandem Repeat Catalog

- Simplify annotations per-region
- Give descriptions of the regions
  - Are the region’s repeat(s) complex?
  - Do we have a decent buffer of non-TR sequence? (aiming for ±25bp)
  - Is there "contamination" of interspersed repeats?
- Other information that may assist stratification or interesting summaries
  - CODIS - 53 sites
  - Known Pathogenic Repeats - 66 sites
  - Genes - Ensembl v105

Overlapping Repeat Motifs

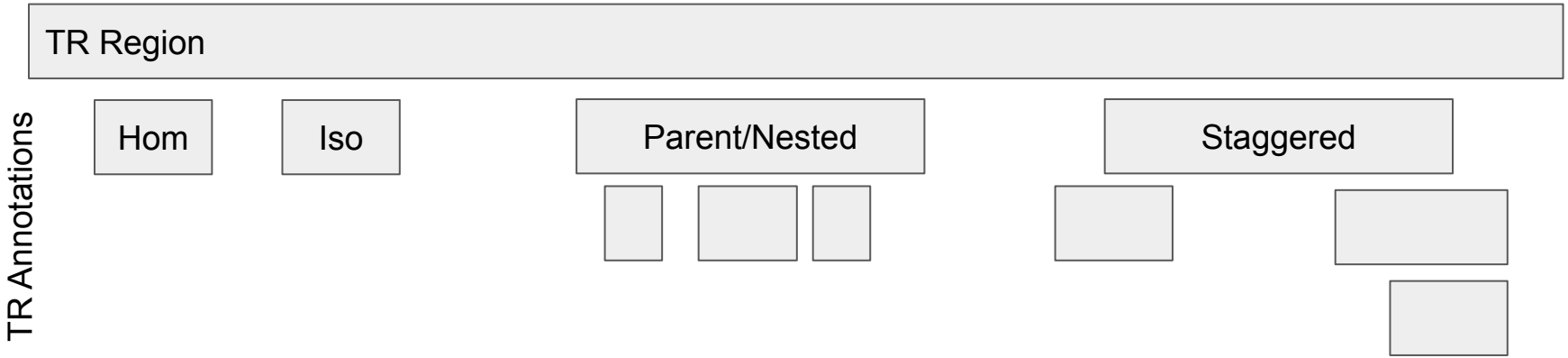
chr1	72120	72163	16	2.7	59	1.47	ATATATACATACACAC
chr1	72124	72164	12	3.3	62	1.49	ATATATACATAC
chr1	72128	72163	4	8.8	52	1.48	ATAC

```
>chr1:72120-72164
ATATATATATACACACATATATACATACATACATACATACATAT
ATATATAcATACACACATATATACATACAcACATAtATACATA
    ATAcATACAtACATATATACATACATAtATACATACATAT
        ATACAtACATAcATACATACATACATACATACATA
```

Simplifying TR Annotations

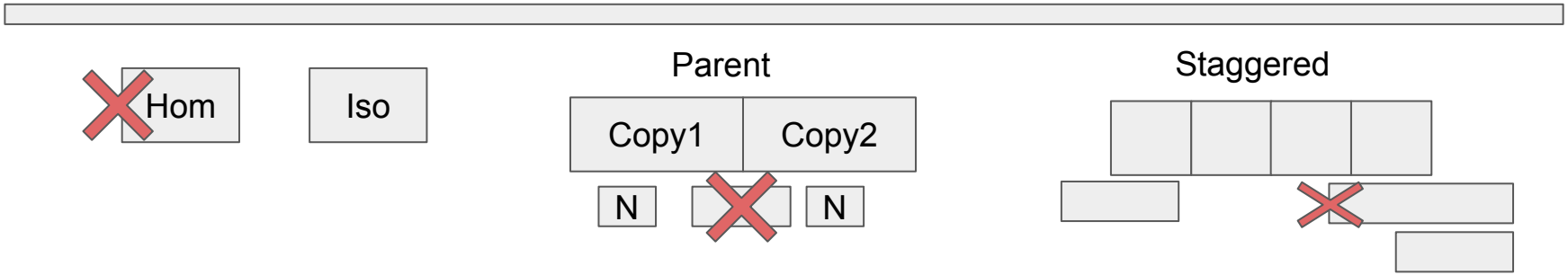
Classify region annotations for filtering into four classes based on motif and overlap with other annotations.

- Homopolymer: polyN repeat
- Isolated: repeat is by itself
- Parent/Nested: repeats within a repeat
- Staggered: repeat overlapping repeats



Simplifying TR Annotations

- Homopolymer: removed
- Isolated: kept (these are the best)
- Parent/Nested:
  - The longest spanning parent and annotations within its copies are kept
- Staggered:
  - Abbutting annotations that are ‘copy adjacent’ are kept



After Simplifying:

- n_annos = 7
- n_subregions = 3

Simplification Results

- Remove regions with only homopolymer annotations
  - 386,985 regions removed from 2,171,789 (17.8%)
- Remaining 1,784,803 regions cover 237,865,075 bp (~8.13% of GRCh38)
  - Started with 3,626,555 annotations
  - Collapsed 1,781,487 annotations (49.1%)
  - End with 1,845,068 annotations

Per-region annotation summary

	mean	std	min	25%	50%	75%	max
n_annos	1.64	1.96	1	1	1	2	296
n_subregions	1.46	1.05	1	1	1	2	191

Overlap Flag Summary

Bit					Region Count	Percent of Regions
isolated	nested	parent	staggered_dn	staggered_up		
isolated	**.**	**.**	**.**	**.**	**1,160,810**	**68.91%**
			staggered_dn	**.**	5,202	0.31%
				staggered_up	22,926	1.36%
		parent	**.**	**.**	103,958	**6.17%**
			staggered_dn	**.**	6,429	0.38%
				staggered_up	19,407	1.15%
	nested	**parent**	**.**	**.**	13,233	0.79%
			staggered_dn	**.**	6,130	0.36%
				staggered_up	13,394	0.80%
.	**.**	**parent**	**.**	**.**	284,165	**16.87%**
			staggered_dn	**.**	1,513	0.09%
				staggered_up	163	0.01%
	nested	**parent**	**.**	**.**	36,228	2.15%
			staggered_dn	**.**	10,501	0.62%
				staggered_up	384	0.02%

Regions with “simple” annotations:
1,548,933 (91.95%)

v1.1 Available -> https://github.com/ACEnglish/adotto/

chr/start/end	region coordinates (3 columns)
ovl_flag	overlap categories of annotations inside the region
up_buff	number of bases upstream of the first annotation's start that are non-TR sequence
dn_buff	number of bases downstream of the last annotation's end that are non-TR sequence
hom_pct	percent of region’s range annotatable as homopolymer
n_filtered	number of annotations removed from the region
n_annos	number of annotations remaining in the region
n_subregions	number of subregions in the region
mu_purity	average purity of annotations in region
pct_annotated	percent of the region's range (minus buffer) annotated
interspersed	name of interspersed repeat class found within region by RepeatMasker
patho	name of gene affected by a pathogenic tandem repeat in region
codis	name of codis site contained in region
gene_flag	gene features intersecting region (Ensembl v105)
biotype	comma separated gene biotypes intersecting region (Ensembl v105)
annos	JSON of TRF annotations in the region (list of dicts with keys: motif, entropy, ovl_flag, etc)

TR Benchmark Goals

Build a benchmark for Tandem Repeats in HG002

Three main components:

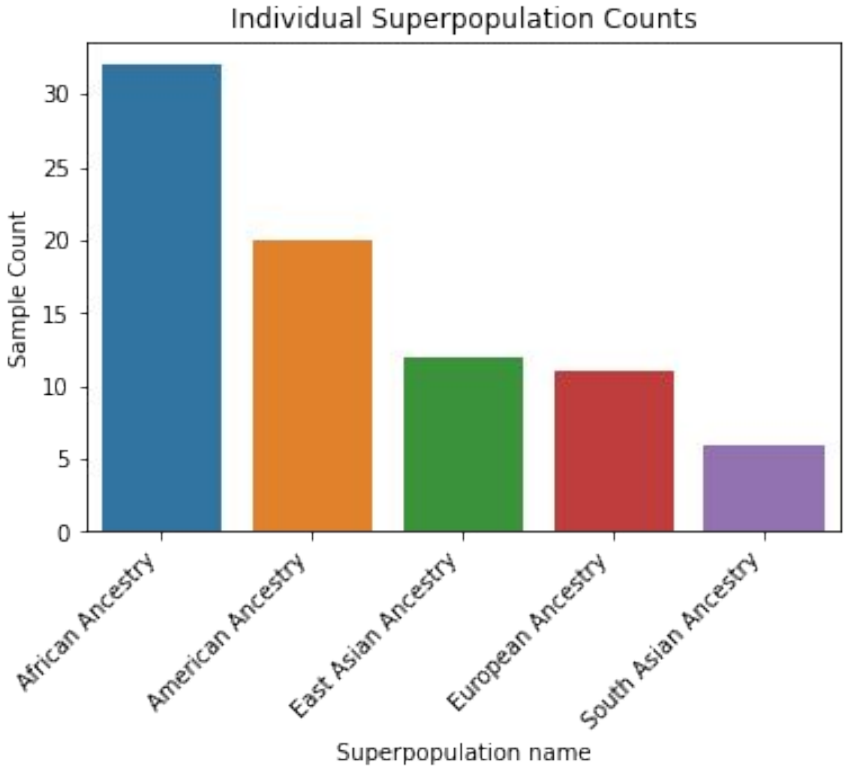
- ✅ Regions - Where are tandem repeats in the reference
- Variants - Which regions have tandem repeats expansions/contractions
- Tools - How to compare variants

Variants - Sample Data

Collected haplotype resolved long-read assemblies from multiple projects covering multiple individuals.

- 3 Projects
  - HPRC (47)
  - Eichler (34)
  - Li (4)
- 172 haplotypes
- 86 samples
- 78 individuals

Replicates	
HG00733	3
NA19240	2
NA24385	3
HG03486	2
HG02818	2
NA12878	2



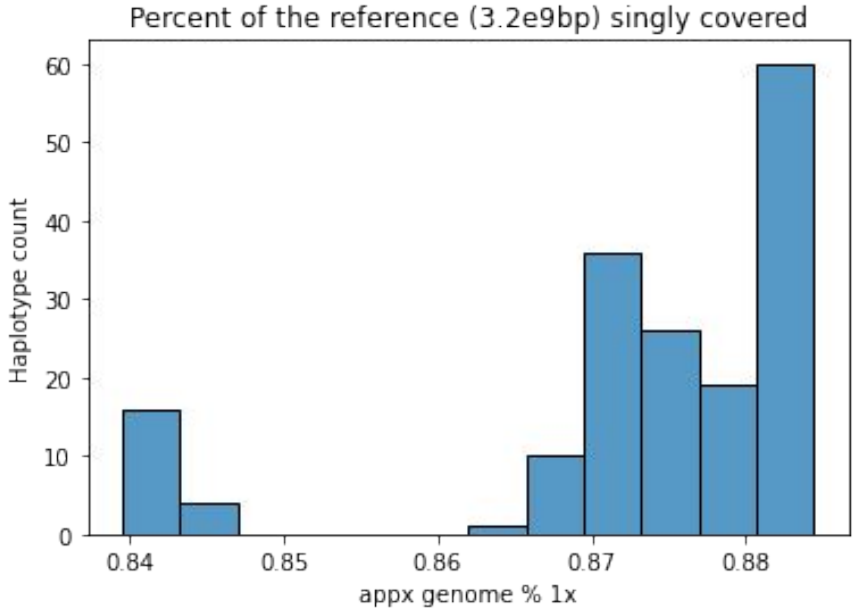
Variant Calling Pipeline

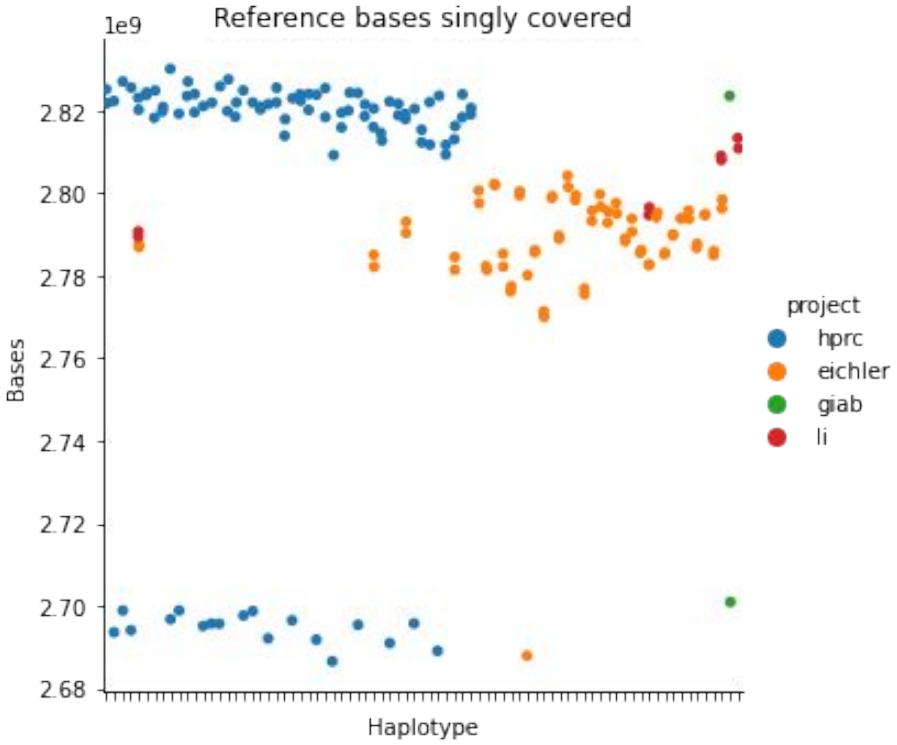
- Built a pVCF from haplotype-resolved long-read assemblies.
- Choosing alignment parameters for minimap2

Parameter Performance GIAB HG002 SV v0.6 (hg19)

Project	Params	True-pos baseline	True-pos call	False-pos	False-neg	Precision	Recall	F-measure
li	**giab**	9,273	10,516	1,093	368	0.906	0.962	0.933
li	**tru**	9,251	10,477	945	390	0.917	0.960	0.938
li	**cust**	9,338	10,595	890	303	0.923	0.969	0.945
li	**pan**	9,335	10,647	712	306	**0.937**	**0.968**	**0.953**
eich	**giab**	9,241	10,448	1,053	400	0.908	0.959	0.933
eich	**tru**	9,217	10,403	935	424	0.918	0.956	0.936
eich	**pan**	9,316	10,590	700	325	**0.938**	**0.966**	**0.952**

Haplotype Coverage





Intersecting TR Regions 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.
- Analyze minimap2 alignment coverage
  - adotto
    - Maximize variant representation consistency with v0.6 SV
  - dipcall + NIST curation
    - Maximize alignment continuity + masking of ‘problematic’ regions
  - 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	95.03%
adotto	328	2,668,392,630	91.27%
Both	45,870	2,615,712,814	89.47%

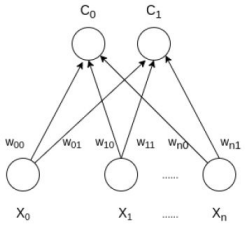
TRregions

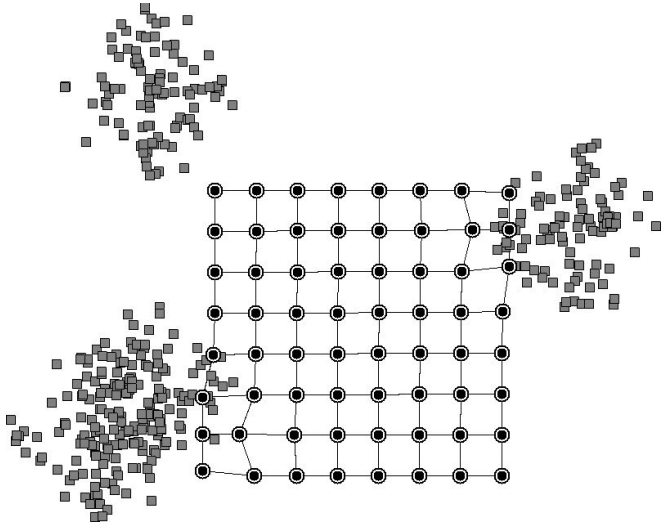
	Count	Span	Genome %	TRr Count %	TRr Span %
Total TRr	1,784,804	237,865,075	8.14%		
dipcall	1,707,318	212,853,127	7.28%	95.66%	89.48%
adotto	1,701,194	217,607,408	7.44%	95.32%	91.48%
Both	1,645,456	203,578,939	6.96%	92.19%	85.59%

Patho/Codis

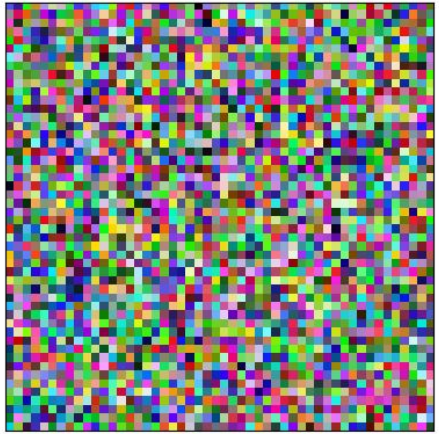
	Patho	Patho %	Codis	Codis %
Total TRr	62		51	
dipcall	50	80.65%	44	86.27%
adotto	52	83.87%	24	47.06%
Both	42	67.74%	23	45.10%

Self-Organizing Map

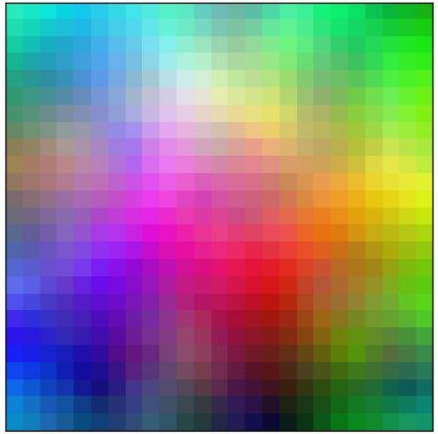




- Built a SOM with Kmer Featurization of sequence spanned by TRregions
  - k=3; n_features = 64; frequency normalized
- Hyperparameters:
  - Shape 25x25; sigma=1.5; learning_rate=1; topology='hexagonal'; neighborhood_function='gaussian'; activation_distance='euclidean'
- Map Patho, Codis, and 100 randomly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



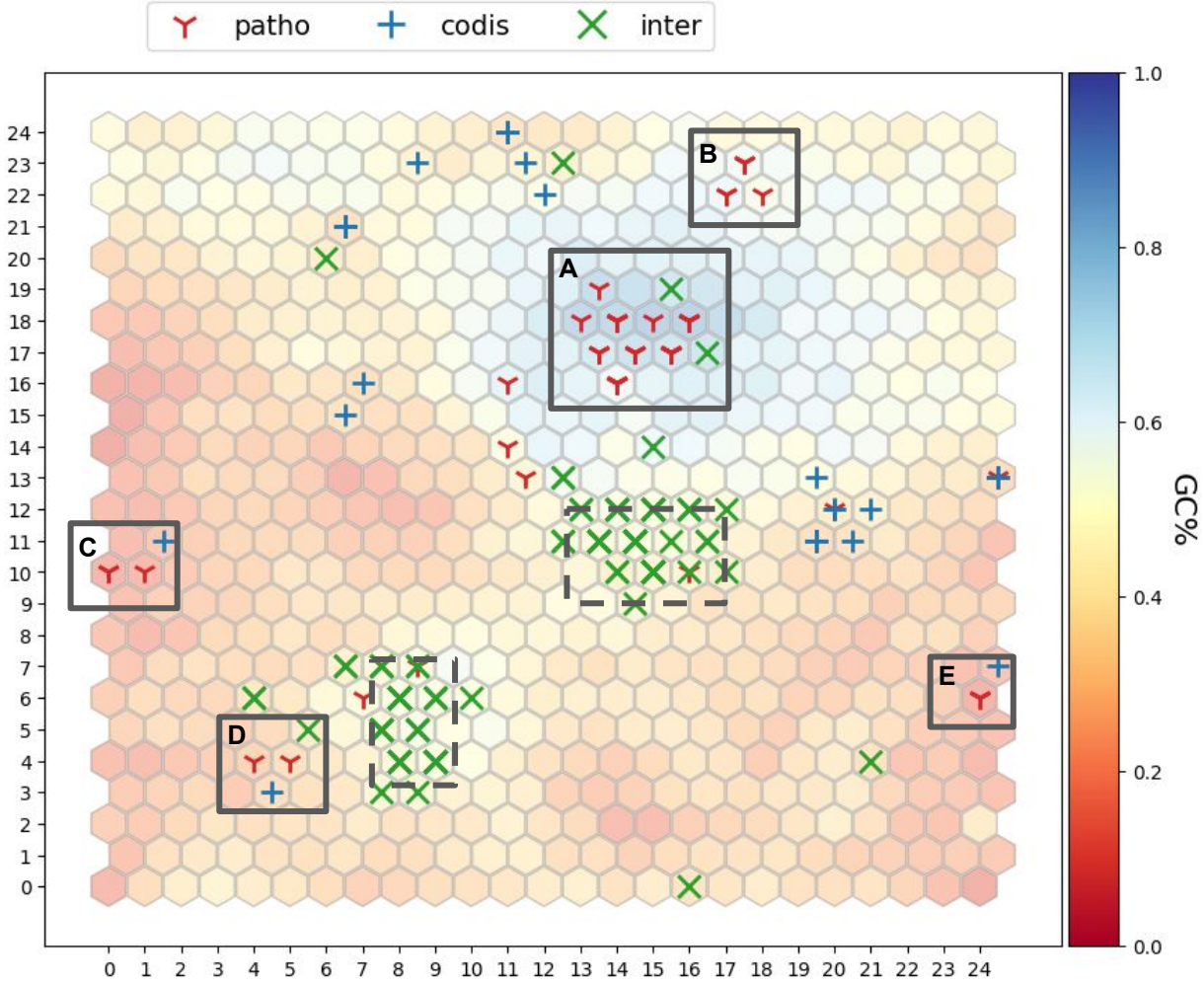
Train on RGB colors
(3 features)

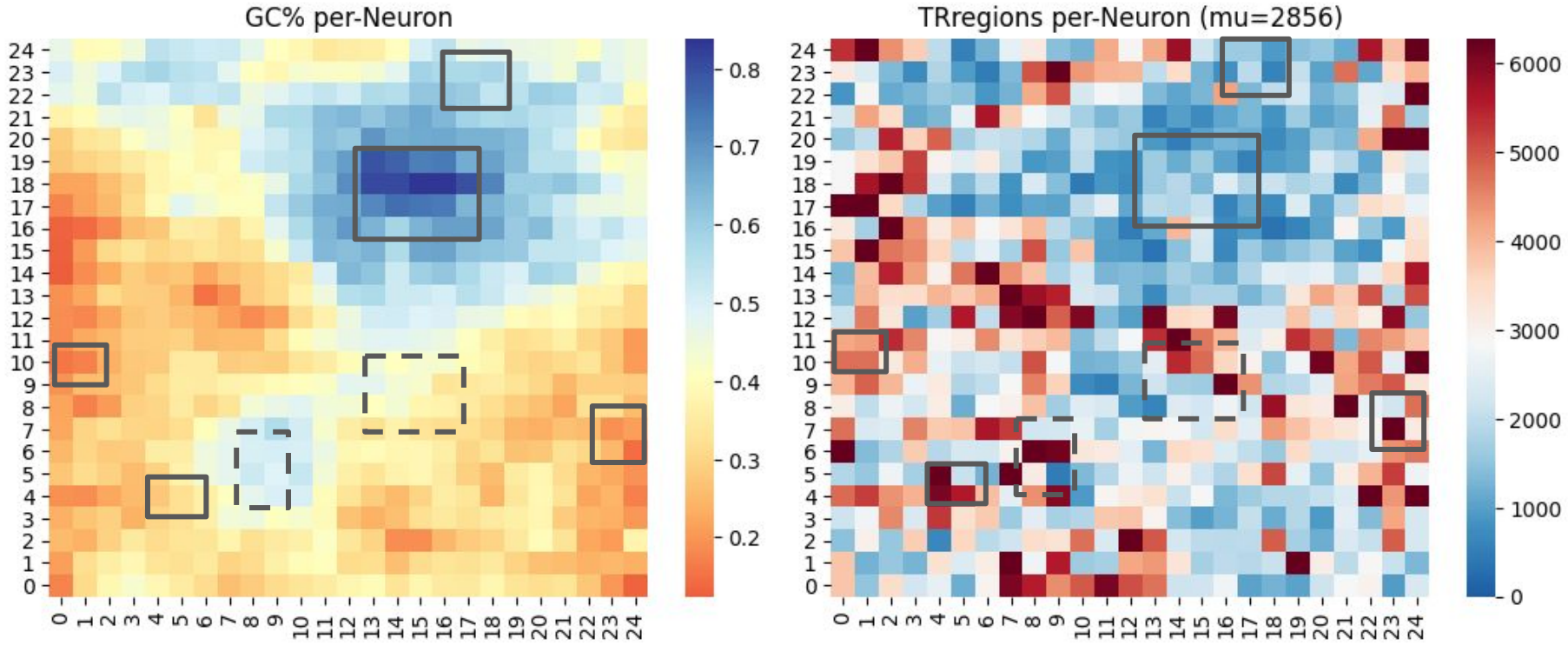


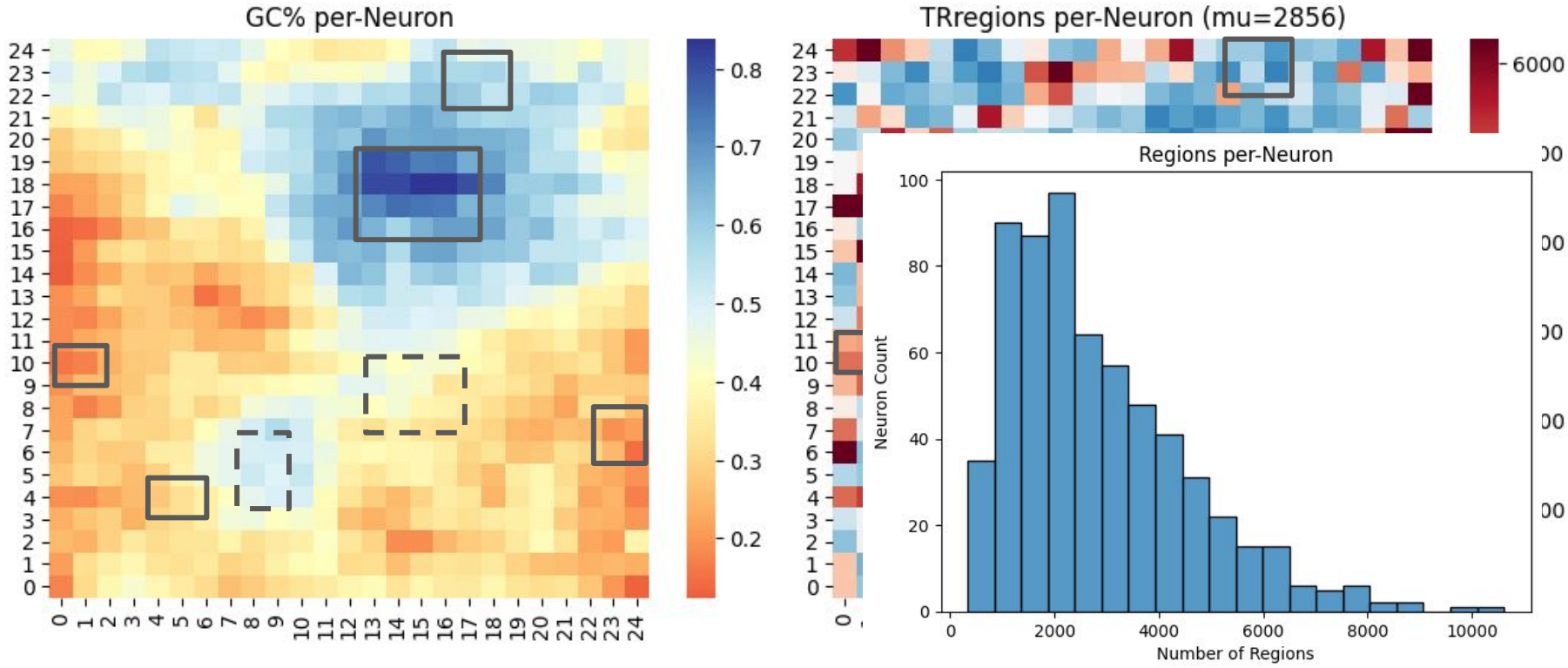
Resulting SOM

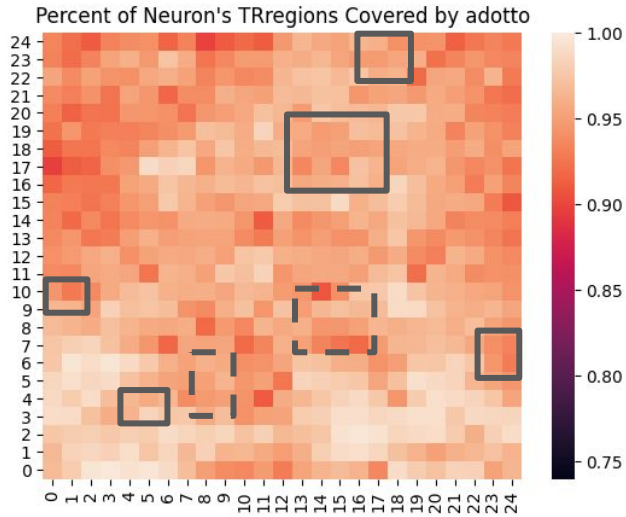
- 54 Patho TRr in 5 neighborhoods
- Interspersed TRr concentrated in two neighborhoods

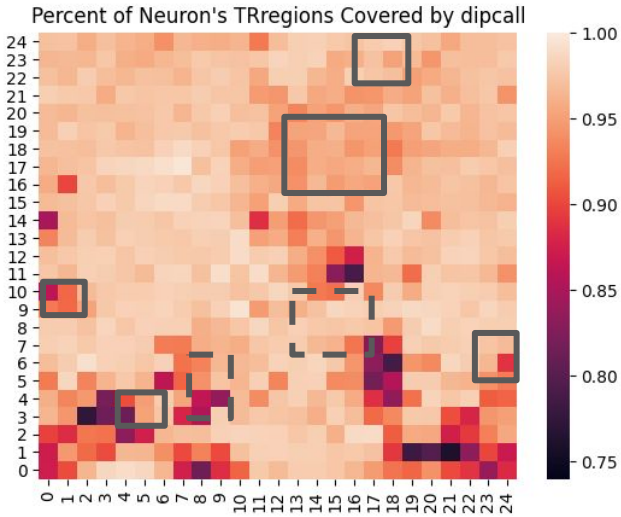
Neighborhood	Motif	Count
A	CGG	10
	CCG	10
	CNG	7
	CTG	7
	GCN	2
	ACCTCGCTGTG CCGCTGCCG	1
	GGCCTG	1
	CGCGGGGCGG GG	1
	CCCCGG	1
B	AGC	6
C	AAAAT	3
D	AAAAG	1
	AAG	1
E	TTTTA	3

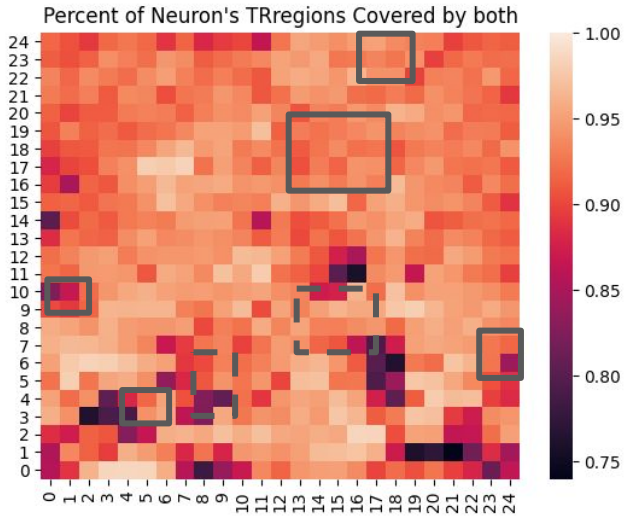


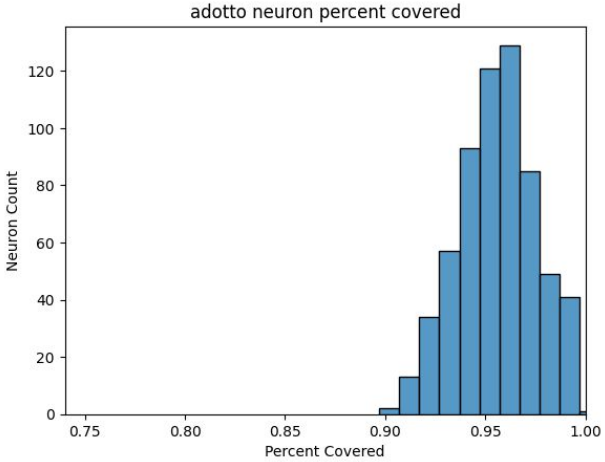

GC% per-Neuron
0.8
0.7
0.6
0.5
0.4
0.3
0.2
TRregions per-Neuron (mu=2856)
6000
5000
4000
3000
2000
1000
0

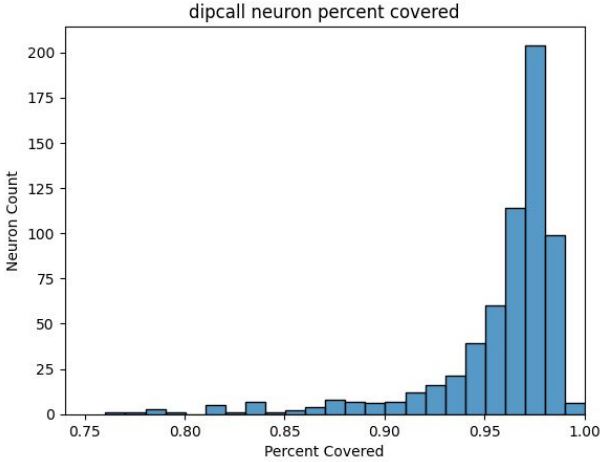

GC% per-Neuron
0.8
0.7
0.6
0.5
0.4
0.3
0.2
TRregions per-Neuron (mu=2856)
6000
Regions per-Neuron
Neuron Count
Number of Regions
0
2000
4000
6000
8000
10000
0
20
40
60
80
100

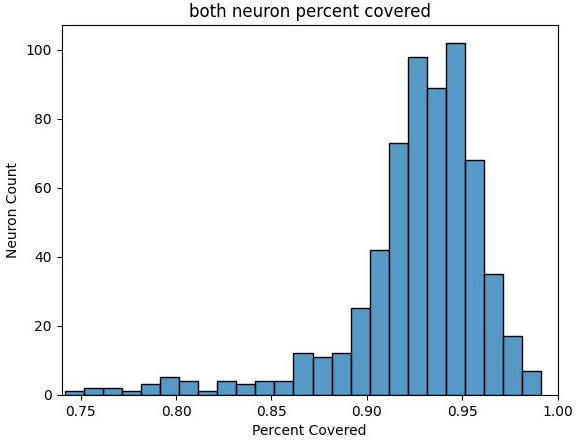

Percent of Neuron's TRregions Covered by adotto


Percent of Neuron's TRregions Covered by dipcall


Percent of Neuron's TRregions Covered by both


adotto neuron percent covered
Neuron Count
Percent Covered


dipcall neuron percent covered
Neuron Count
Percent Covered

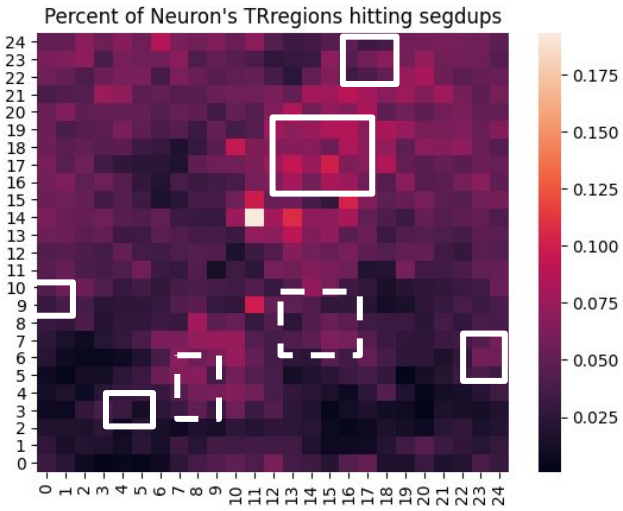

both neuron percent covered
Neuron Count
Percent Covered

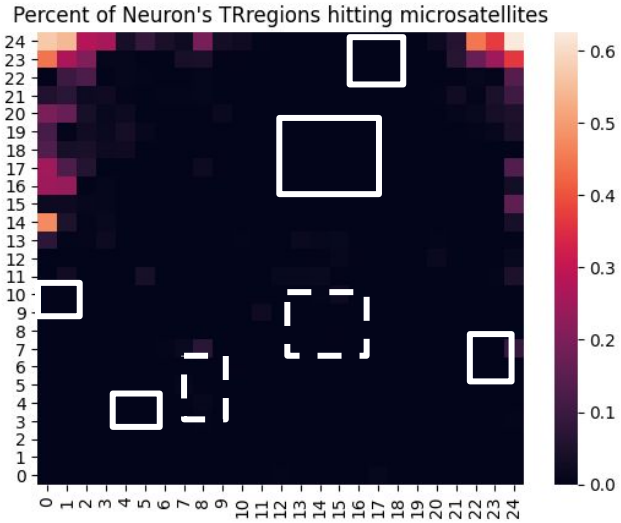
What types of TRr are in these dipcall ‘dryspots’?

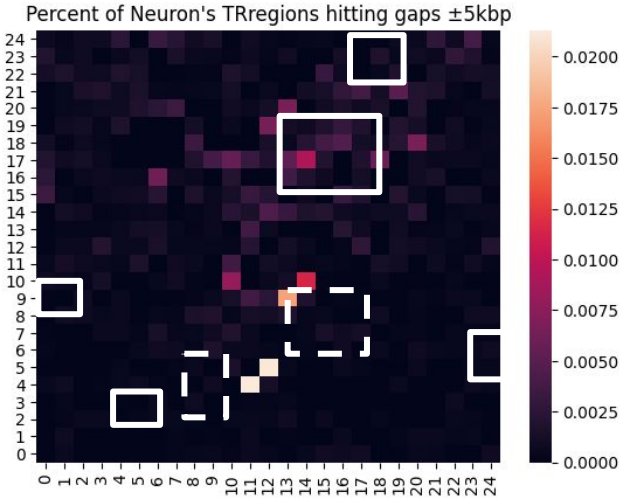
SegDups
73,736 regions

Microsatellites
40,225 regions

Gaps
1,333 regions



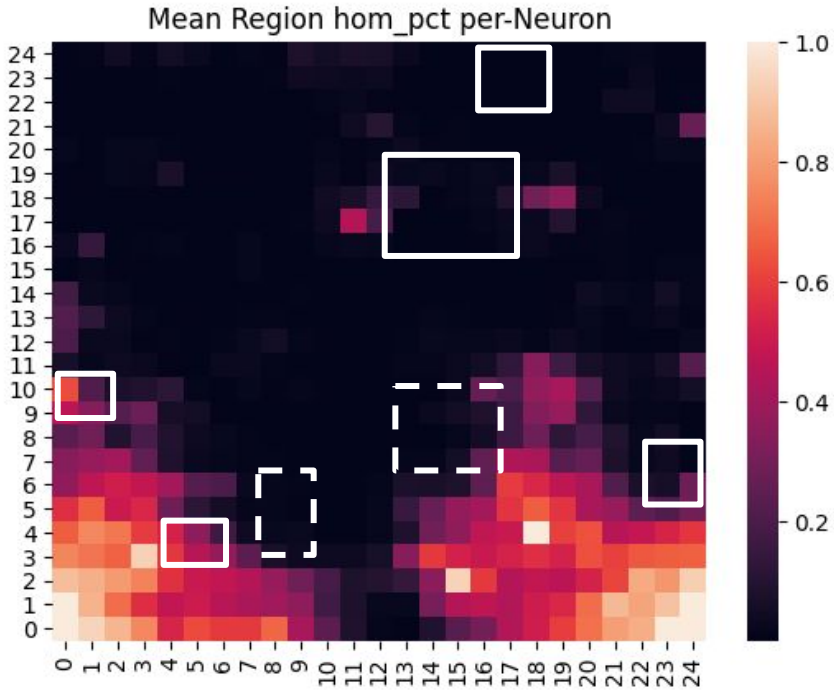


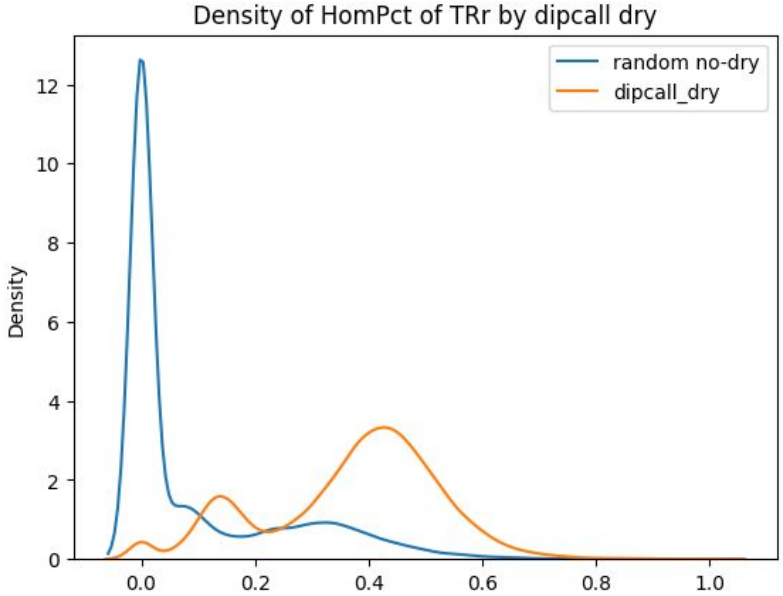


Tracks from UCSC Table Browser

Homopolymers

The ‘dry’ spots often have regions with 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 error rates



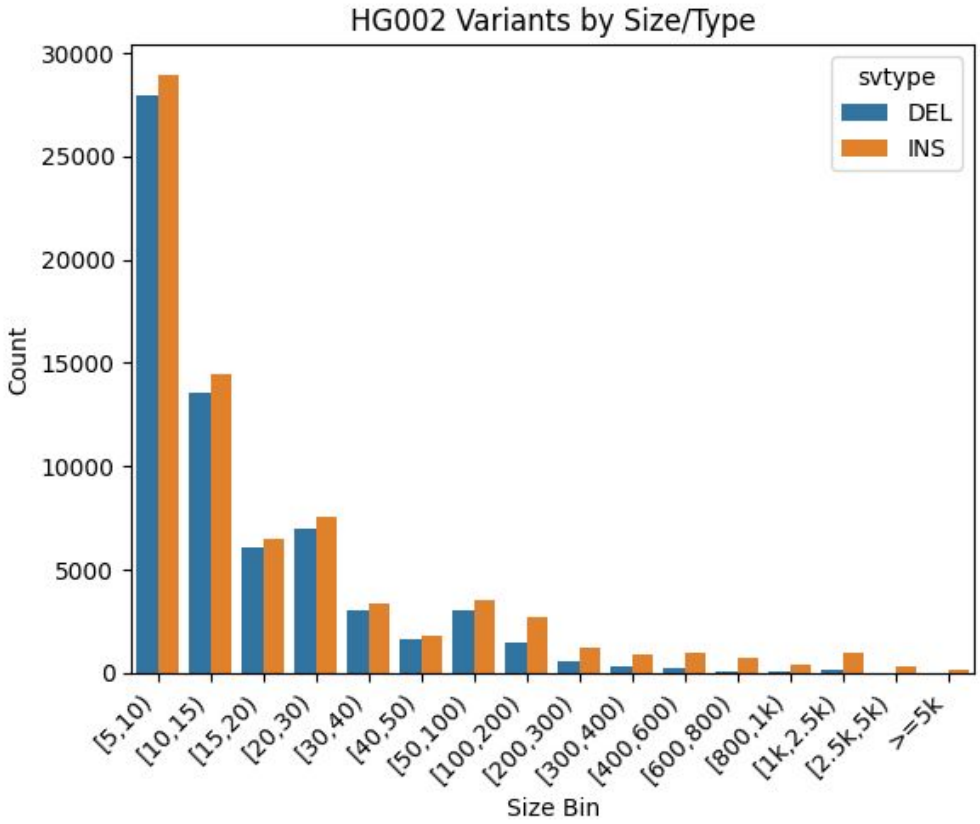


22,787 regions <80% captured vs 30k random.

HG002 HPRC Assembly Covered TR Regions' Variants

- 1,645,456 (92%) of TR regions covered
- ~90% are homozygous reference
  - Negative controls

Type	Count (>=5bp)
Insertion	74,256
Deletion	65,061
Total	139,317



TR Benchmark Goals

Build a benchmark for Tandem Repeats in HG002

Three main components:

- ✅ Regions - Where are tandem repeats in the reference
- ✅ Variants - Which regions have tandem repeats expansions/contractions
- Tools - How to compare variants

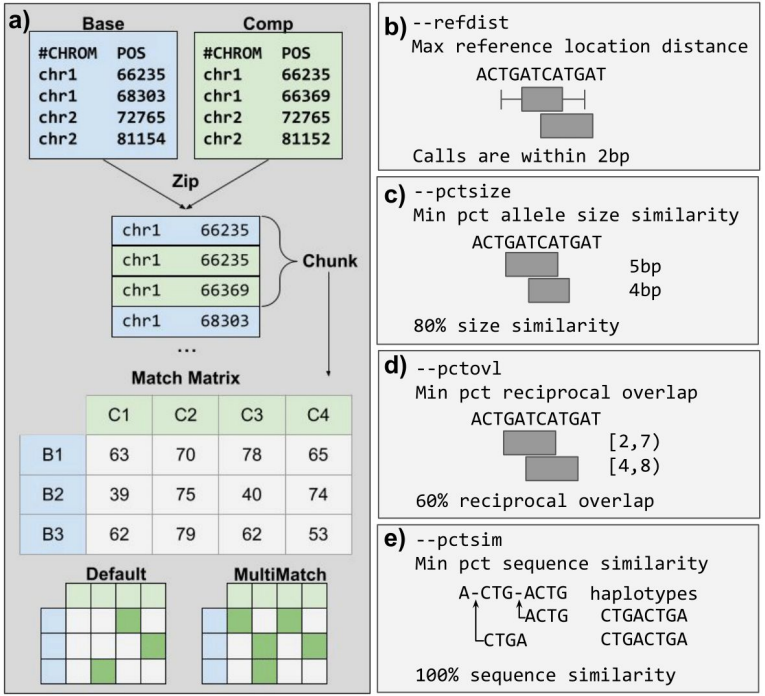
Truvari

Method | Open Access | Published: 27 December 2022

Truvari: refined structural variant comparison preserves allelic diversity

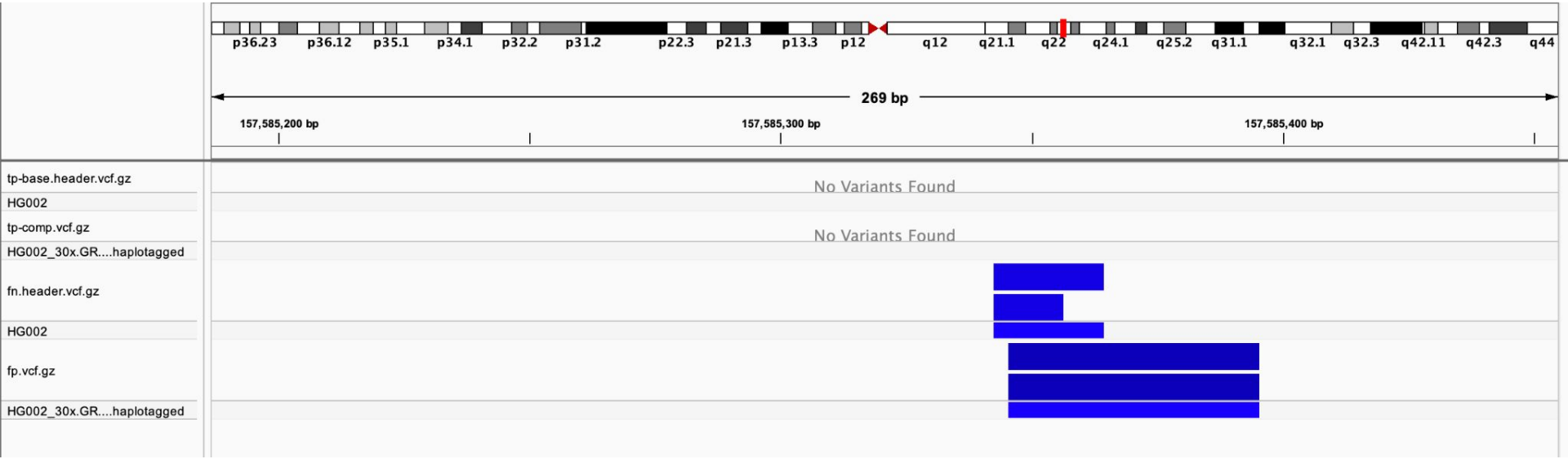
Adam C. English, Vipin K. Menon, Richard A. Gibbs, Ginger A. Metcalf & Fritz J. Sedlazeck

Genome Biology **23**, Article number: 271 (2022) | Cite this article



	CallerC	CallerB	CallerA
TP-base	88,239	39,220	64,489
TP-comp	88,532	39,227	65,379
FP	52,134	6,166	11,138
FN	51,133	100,152	74,883
precision	0.629	0.864	0.854
recall	0.633	0.281	0.463
f1	0.631	0.425	0.600
base cnt	139,372	139,372	139,372
comp cnt	140,666	45,393	76,517

Variant Representation - Example 1



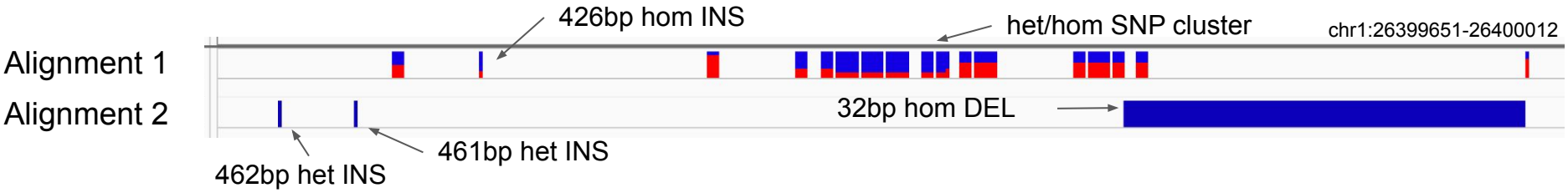
FN:
14bp DEL
chr1:157585342
REF: CTCTCTTTCTTTCTT
ALT: C
GT: 1/0

FP:
14bp DEL
chr1:157585346
REF: CTTTCTTTCTTTCTTTCTTTCTTTCTTTCTTTCTTTCTTTCTTTCTTTCC
ALT: TTCTTTCTTTCTTTCTTTCTTTCTTTCTTTCTTTCT
GT: 0/1

PctSizeSimilarity=1
PctSeqSimilarity=0.4615

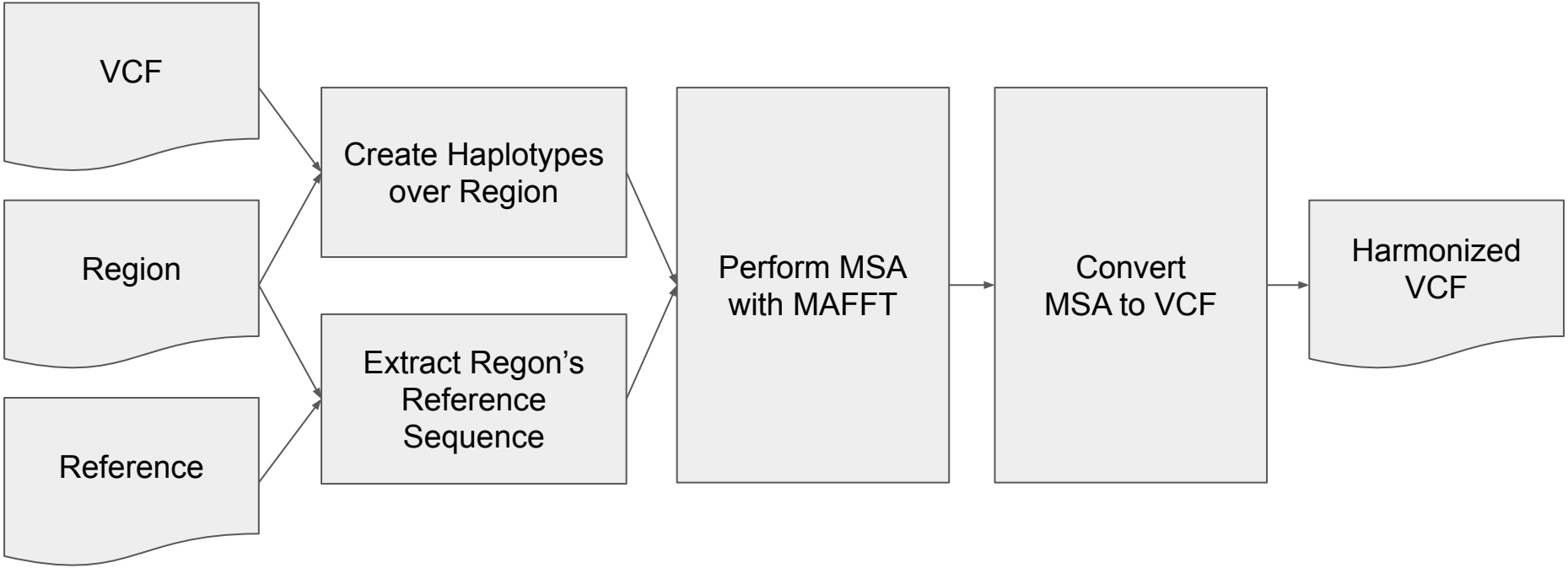
Variant Representation - Example 2

- Identical HPRC HG002 input assembly
- Different minimap2 alignment parameters

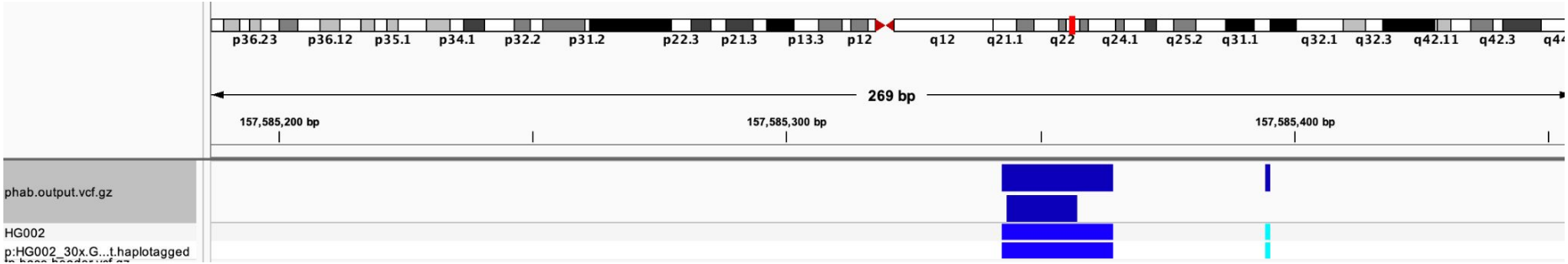


Truvari phab

Variant Harmonization with Multiple Sequence Alignment



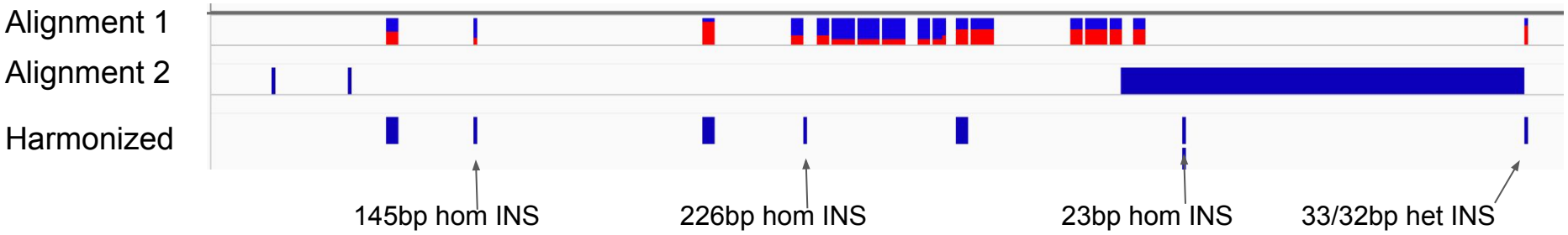
Phab Output - Example 1



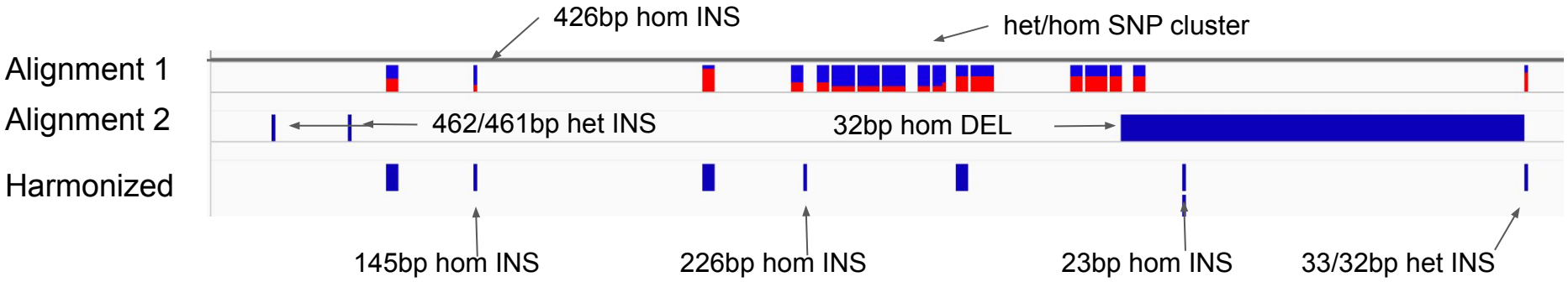
				Base	Comp
chr1	157585342	CTCTCTTTCTTTCTTTCTTTCTT	C	0/1	0/1
chr1	157585343	TCTCTTTCTTTCTTT	T	1/0	1/0
chr1	157585395	C	T	1/1	1/1

100% Sequence and Size Similarity

Phab Output - Example 2



Phab Output - Example 2



New Problems: Counting and Speed

Variant Count by State:

Alignments: 1 FN, 3 FP

Harmonized: 4 TP

Region Count by State:

Alignments: 1 FP/FN

Harmonized: 1 TP

Truvari refine

Automated benchmark result refinement with phab

- Find regions with at least one FP or one FN
- Runs regions through `phab`
- Re-runs `truvari bench` on harmonized variants
- Recalculate performance metrics

```
$ truvari bench ... -o result/
$ truvari refine --reference $REF result/
```

```
result/
   [ .. bench files .. ]
   refine.variant_summary.json
   refine.region_summary.json
   refine.regions.txt
   phab.output.vcf.gz
   phab_bench/
```

Truvari refine results on ‘strawman’ benchmark

Variant Summary			
	Caller1	Caller2	Caller3
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

Refine - Bench Summary Difference			
precision	0.012	0.265	0.064
recall	0.002	0.305	0.029
f1	0.003	0.284	0.040

Region Summary			
	Caller1	Caller2	Caller3
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
(specificity) **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

Summary

- ✅ Regions - Where are tandem repeats in the reference
- ✅ Variants - Which regions have tandem repeats expansions/contractions
- ✅ Tools - How to compare variants
- Remaining Work
  - Internal Review of a ‘Strawman’ benchmark
    - 5bp INDEL not necessarily a TR expansion/contraction
    - More curation of regions on HPRC HG002 specifically
  - Additional tooling for stratifications/reporting
  - Documentation
  - Do it all again against CHM13

Acknowledgements

Tandem Repeat Benchmark Group

- Fritz Sedlazeck
- Justin Zook
- Egor Dolzhenko
- Mark Chaisson
- Justin Wagner
- Helya neh Ziaei-jam
- Jonghun Park
- Nathanael Olson
- Nathan Dwarshuis
- Wouter De Coster
- Michael Eberle
- Don Freed