GIAB TR

Revisiting Tiering

- Separated the regions into Tier1/Tier2
  - Two sequencing replicates (eichler/li assemblies)
  - One alignment replicate (Adotto / dipcall+NIST curation)
  - Tiering rules based on benchmarking of the three replicates
- Verkko HG002 assembly comparison by NIST showed a number of mis-classified sites which may not satisfy Reliable IDentification of Errors (RIDE)

Original Tiering Rules

Using replicate states (e.g. TP_FN_TP), place regions in Tier2 if:

- The alignment-replicate can't match to itself at loci where Truvari refine doesn’t assign TN|TP
- There is no agreement between any of the repl states
- Seq-replicates agreed to something different from the alignment replicate
- A FP in a seq-replicate but TP agreement between other two replicates suggested the possibility of a collapsed het

These rules assumed seq-replicates had adequate coverage to be informative.

Additional Tiering Rules

- **TN_TN_TN** : Need at least one seq-replicate to have adequate coverage
- **FN_FN_TP** : Assumed it was FN because of a lack of coverage. If at least one seq-replicate is well covered, this suggests aln-rep could have a FP
- **FP_TN_TN** / **TN_FP_TN** : Assumed the seq-replicate had coverage that confirmed the th negative. If it doesn't, 50/50 on who is right between the FP seq-replicate and TN align-replicate
- **FN_FP_TP** / **FP_FN_TP** : No agreement. Rule was missed originally

These rules made independent of Verkko assembly results.

Run `truvari refine` on `--regions` well covered by sequencing replicates.

Original Tiering

Region counts
Tier1: 1,700,175 (95% of catalog)
Tier2: 6,678

(3xTNerr + 3xTPerr) / TotalError

566 / 1,209 = **46.8%**

Less than half the verkko errors are in ‘repl’ states with the highest confidence.

state repl	FN	FN,FP	FP	TN	TP	total	'error'
FN_UNK_TP	0	0	0	0	1	1	0
TN_TP_TP	0	0	0	5	43	48	0
TP_UNK_TP	0	0	0	0	3	3	0
FP_FN_TP	5	0	1	0	27	33	6
FN_FP_TP	4	1	2	0	28	35	7
FN,FP_TP_TP	9	23	17	0	134	183	49
TP_FN,FP_TP	9	27	21	1	197	255	57
FN_TP_TP	47	6	9	0	966	1028	62
TP_FN_TP	71	7	5	1	818	902	83
FN_FN_TP	86	11	1	0	294	392	98
TN_FP_TN	0	0	135	1040	0	1175	135
FP_TN_TN	0	0	146	914	1	1061	146
TN_TN_TN	2	1	168	1585496	6	1585673	171
TP_TP_TP	159	74	162	14	89183	89592	395

New Tiering

Region counts
Tier1: 1,638,508 (91% of catalog)
Tier2: 68,345

650 / 1,187 = **54.8%**

More than half the verkko errors are in ‘repl’ states with the highest confidence.

state repl	FN	FN,FP	FP	TN	TP	UNK	total	'error'
TP_UNK_TP	0	0	0	0	1	0	1	0
TN__TN	0	0	1	3159	1	0	3161	1
TP__TP	0	0	1	0	188	0	189	1
_TP_TP	2	1	1	0	616	0	620	4
_TN_TN	0	1	5	13247	0	0	13253	6
FN,FP_TP_TP	9	22	18	0	116	0	165	49
TP_FN,FP_TP	9	28	19	0	176	0	232	56
FN_TP_TP	49	6	9	0	575	0	639	64
TP_FN_TP	71	7	4	2	655	0	739	82
TN_FP_TN	0	0	131	756	0	0	887	131
FP_TN_TN	0	0	143	588	1	0	732	143
TN_TN_TN	4	0	176	1509653	49	0	1509882	180
TP_TP_TP	207	85	178	34	88752	1	89257	470

Differences

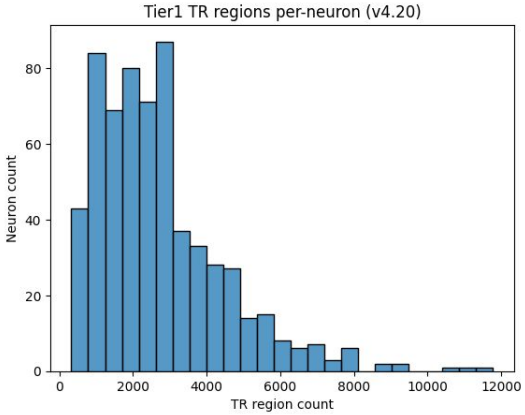
tier	tier_new	count
Tier1	Tier1	1638199
	Tier2	61976
Tier2	Tier2	6369
	Tier1	309

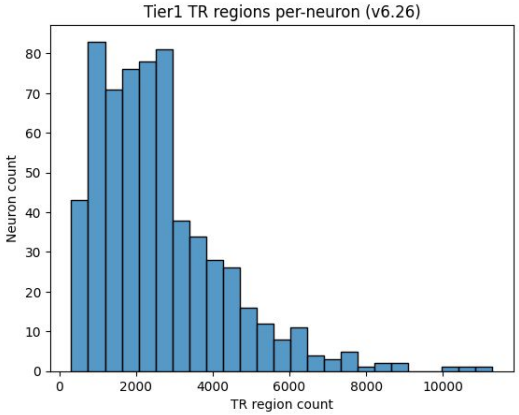
Truvari v4.1 difference

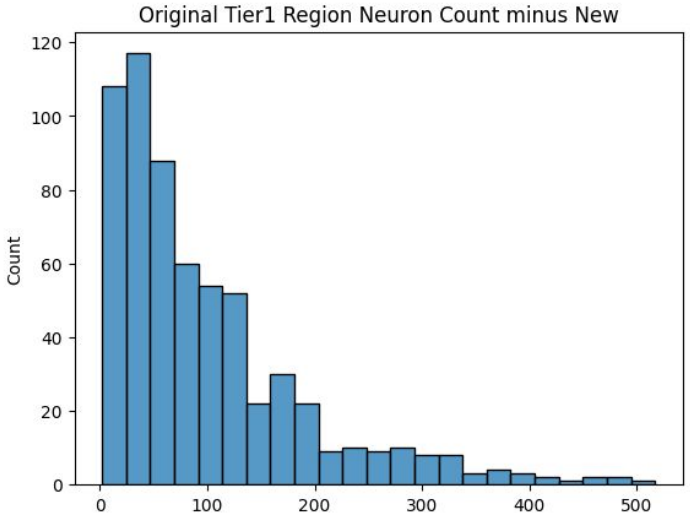
	repl_new
__TN	60633
__TP	451
FN_FN_TP	233
_FP_TN	217
FP__TN	177

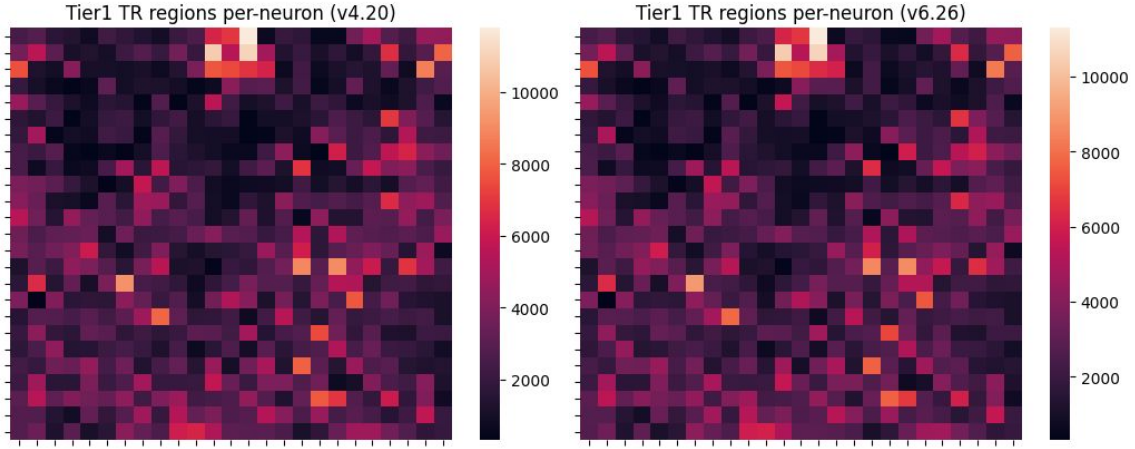
repl	repl_new	
TP_TP_FN	TP_TP_TP	65
TN_TN_TP	TN_TN_TN	45
TP_TP_FP	TP_TP_TP	29
TP_TP_FN,FP	TP_TP_TP	28
TP_TN_TN	_TN_TN	22
TP_TP_TN	TP_TP_TP	20

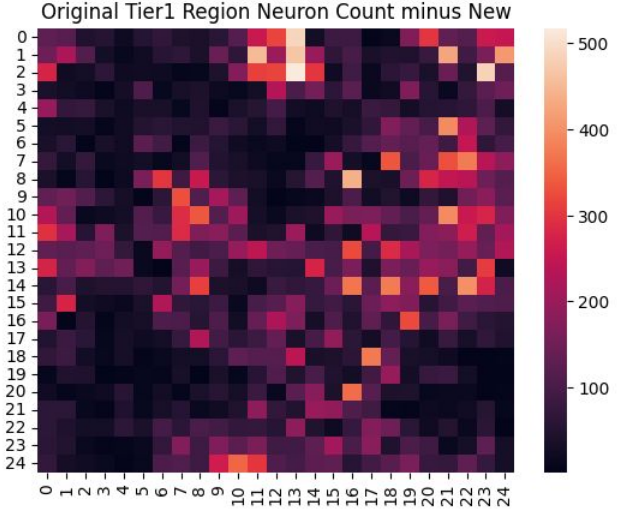
Doesn’t appear to be a sequence-context bias added with the new tiering.











Laytr

`laytr giabTR` supplementary files (som|map) are packaged with the benchmark.

Also, fixed a header issue.

Will send out benchmark v6.23 this week