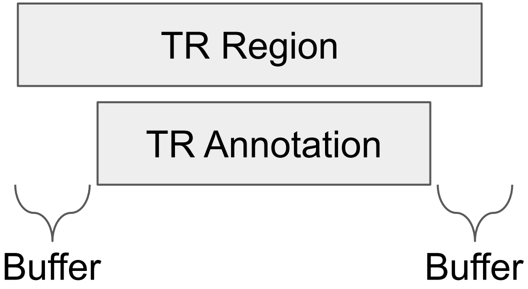
GIAB TR

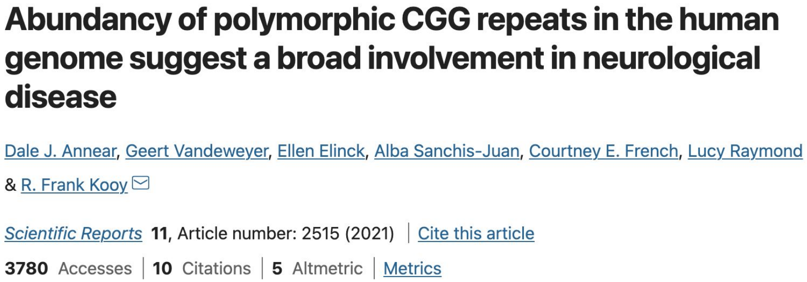
TR Regions Goals

1. Find regions of the genome with tandem repeats
2. Give descriptions of the regions about their 'benchmarkability'
   a. Is there "contamination" of interspersed repeats?
   b. Do we have a decent buffer of non-TR sequence? (aiming for ±25bp)
   c. Are the region’s repeat(s) complex?
3. Other annotations that may assist stratification or interesting summaries
   a. codis/patho/genes



‘Seeds’ to TR Regions

1. Nine sources
2. Self-merge 25bp slop & filter to span 10bp to 50kbp
3. Inter-merge
   a. ~~Remove within 5kbp of reference gaps~~
4. Run TRF/RepMask on regions’ sequence
5. **Improve annotations**
6. Keep regions with at least one TRF anno



Abundancy of polymorphic CGG repeats in the human genome suggest a broad involvement in neurological disease

Dale J. Annear, Geert Vandeweyer, Ellen Elinck, Alba Sanchis-Juan, Courtney E. French, Lucy Raymond & R. Frank Kooy

Scientific Reports **11**, Article number: 2515 (2021) | Cite this article

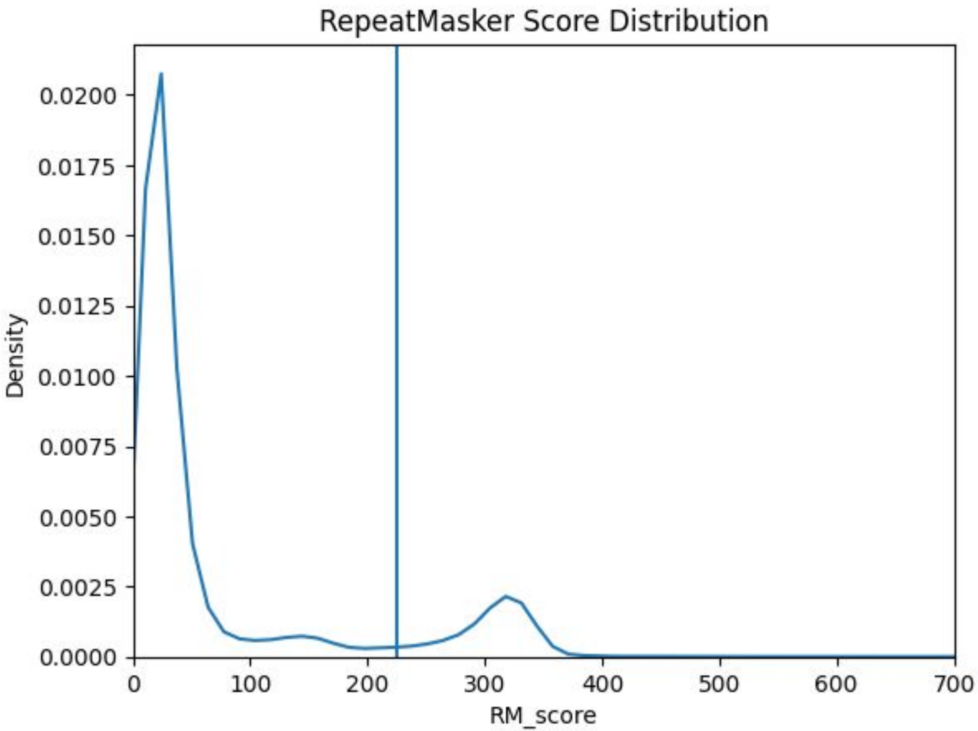
3780 Accesses | **10** Citations | **5** Altmetric | Metrics

Name	**Source**
GIAB	FTP
Baylor	UCSC Genome Browser
UCSC1 (Ensembl?)	???
UCSC2 (adVNTR?)	Github?
TRGT	Github
pbsv	Github
Illumina (from Egor)	Github
USC (Vamos)	Pre-print
abundant_pCGG	Paper

Interspersed Repeats

- Want to detect interspersed repeats ‘contaminating’ our TR regions
- Process:
  - Run RepeatMasker on the sequence spanned by a TR region
  - Filter to only hits with score >= 225 and an Interspersed repeat class
    - Interspersed: SINE, LINE, LTR, DNA, Retroposon, [srp sc sn t r]RNA
    - Tandem Repeat: Simple_repeat, Low_complexity, Satellite, Unknown
- Add column to TR regions of the interspersed repeat class with the highest score over the sequence
  - Will be keeping them in catalog

Interspersed Repeats



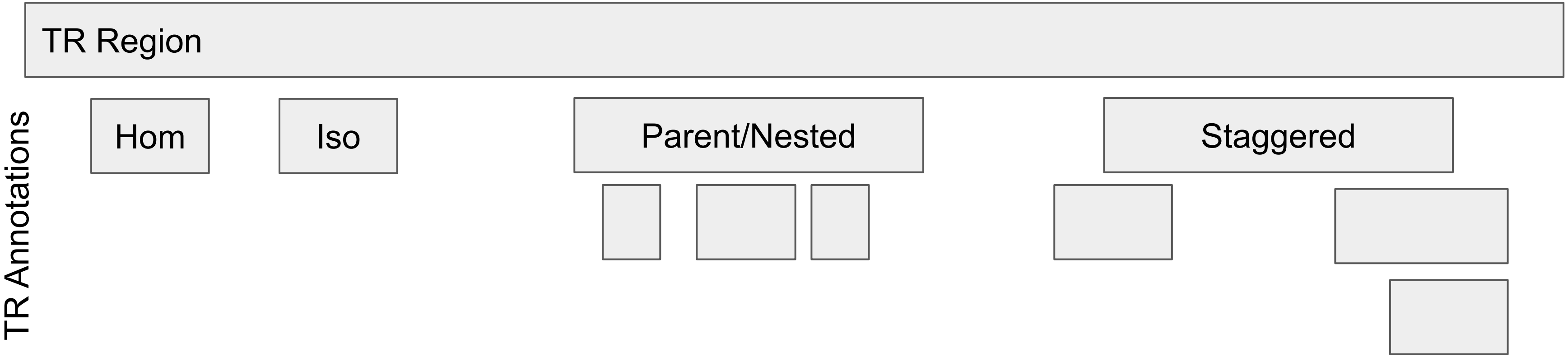
Repeat Class	Count	Percent
.	1,684,443	94.38%
SINE	89,671	5.02%
LTR	6,576	0.37%
Retroposon	2,076	0.12%
LINE	2,003	0.11%
DNA	35	0.00%

Simplifying TR Annotations

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

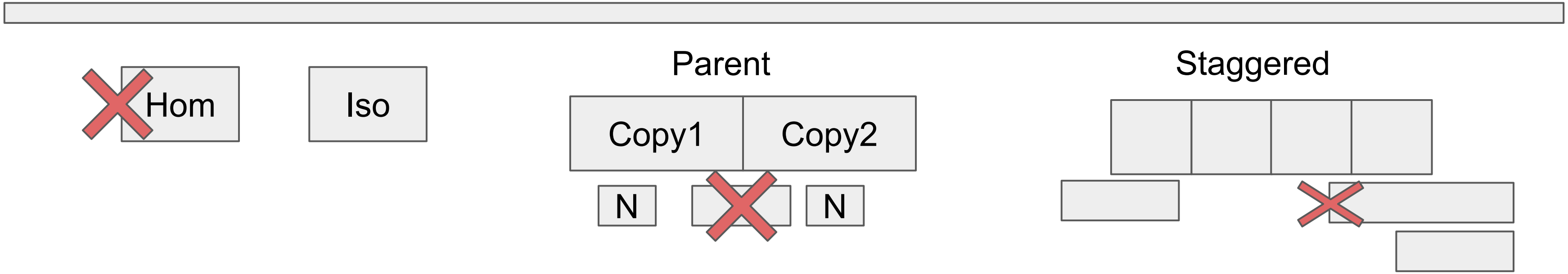
- Homopolymer: mononucleotide repeat
- Isolated: repeat is by itself
- Parent/Nested: repeats within a repeat
- Staggered: repeat overlapping repeats

If no annotation remains in a region after filtering, remove it.



Simplifying TR Annotations

- Homopolymer: removed
- Isolated: kept (these are the best)
- Parent/Nested:
  - annotations contained within longer spanning annotations AND overlapping repeat copies are removed.
  - The longest spanning **‘parent’** and annotations within its copies are kept
- Staggered:
  - annotations overlapping boundaries that aren’t ‘copy’ adjacent are removed



Simplification Results

- Remove regions with only homopolymer annotations
  - 386,985 regions removed from 2,171,789 (17.8%)
- Remaining 1,784,804 regions cover 237,865,075 bp (~7.43% of genome)
  - Started with 3,626,555 annotations
  - Removed 1,781,487 annotations (49.1%)
  - End with 1,845,068 annotations
- Without interspersed: 1,684,443 covering 173,160,739bp (~5.41%)

Per-region summary post-simplification

	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

- During the simplification process, we populate a bitflag describing an annotation’s overlap with other annotations
- We record for the region the ovl_flag to summarize the overlapping complexity of its annotations

Overlap	Bit
Isolated	1
Nested	2
Parent	4
Staggered_dn	8
Staggered_up	16

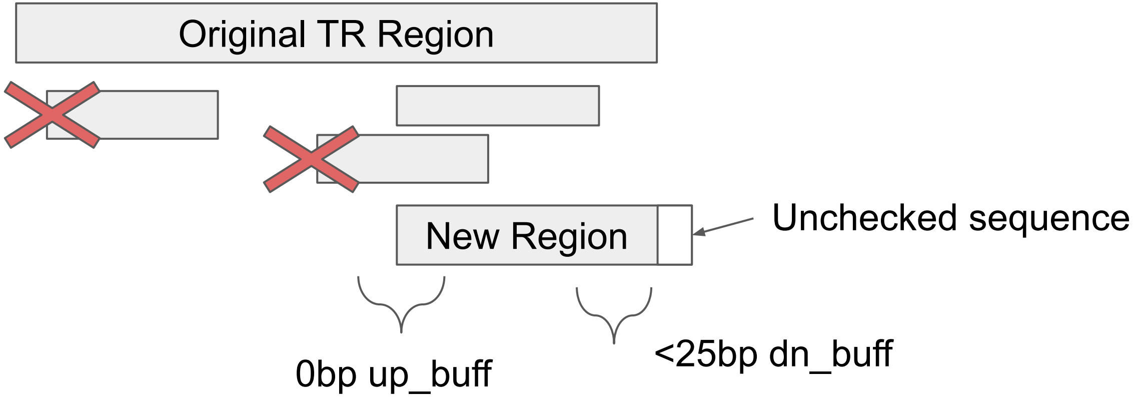
Region Overlap Flag Summary

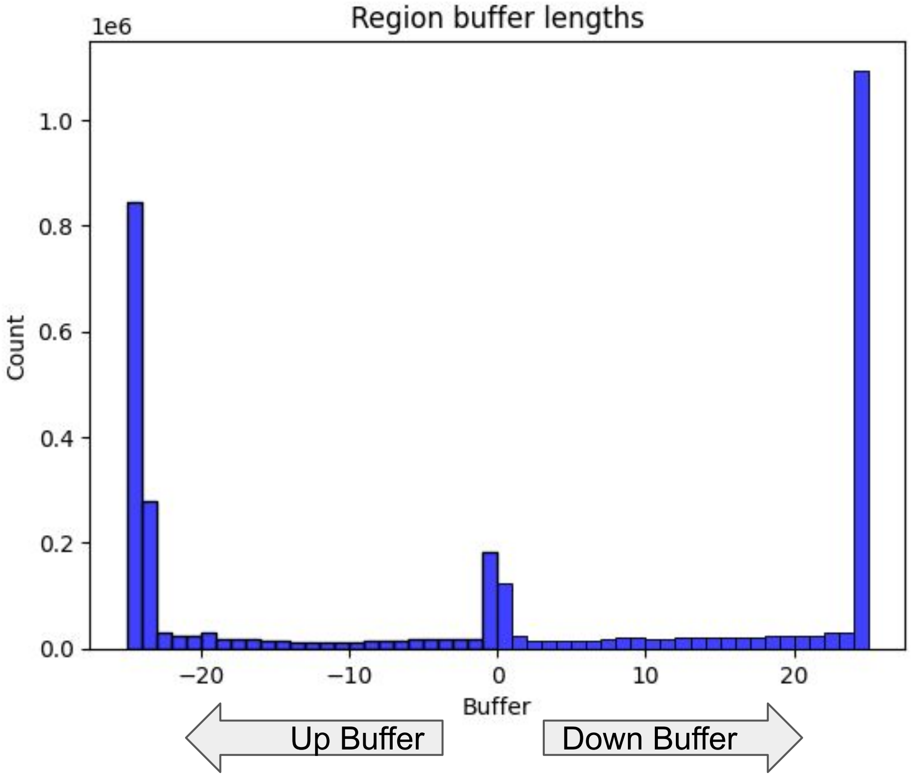
Bit					Region Count	Percent of Regions
isolated	**nested**	**parent**	**staggered_dn**	**staggered_up**		
isolated	**.**	**.**	**.**	**.**	**1,160,810**	**68.91%**
			staggered_dn	**.**	5,202	0.31%
				staggered_up	22,926	1.36%
		parent	**.**	**.**	103,958	**6.17%**
			staggered_dn	**.**	6,429	0.38%
				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%

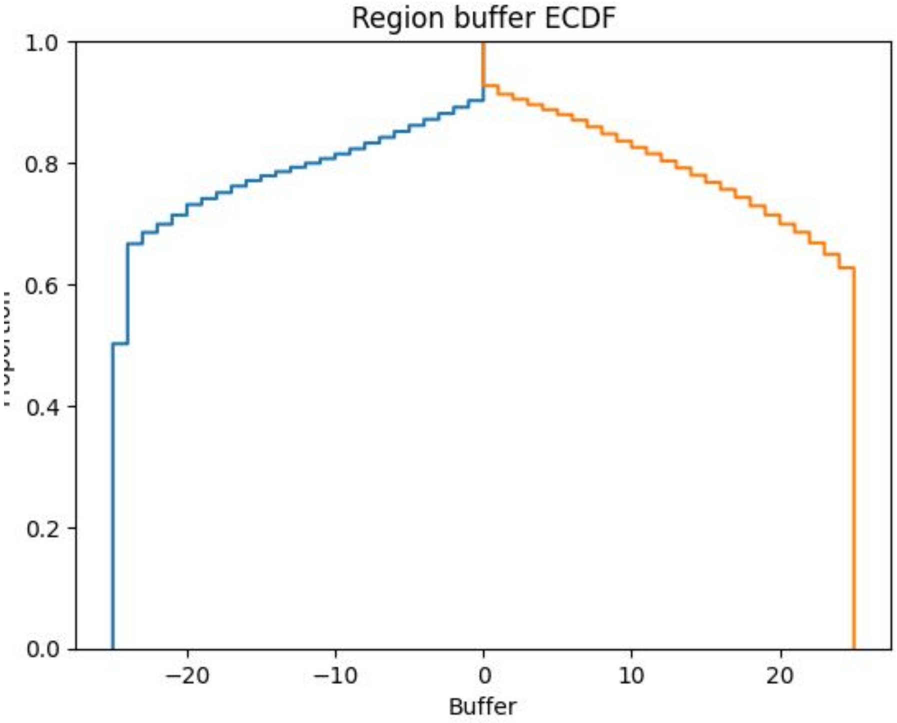
Regions with “simple” annotations:
1,548,933 (91.95%)

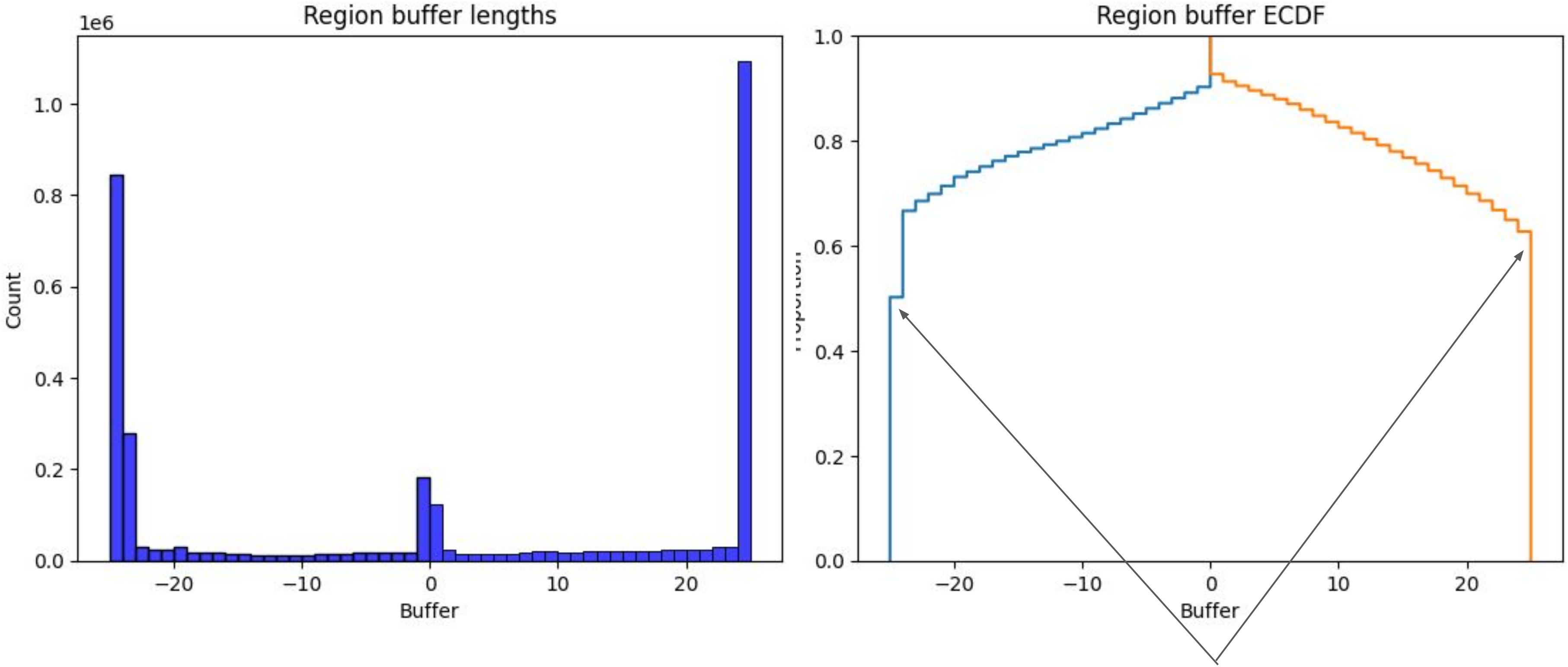
Refining boundaries

After altering the annotations within a region, we try to update the region's boundaries to ±25bp from the `min(annotation_starts)` and `max(annotation_ends)`. However, this may place boundaries inside of previously removed annotations. Therefore, we add a column to the region with information about our start/end buffer length.




Region buffer lengths
1e6
1.0
0.8
0.6
0.4
0.2
0.0
Count
−20
−10
0
10
20
Buffer
Up Buffer
Down Buffer


Region buffer ECDF
1.0
0.8
0.6
0.4
0.2
0.0
−20
−10
0
10
20
Buffer

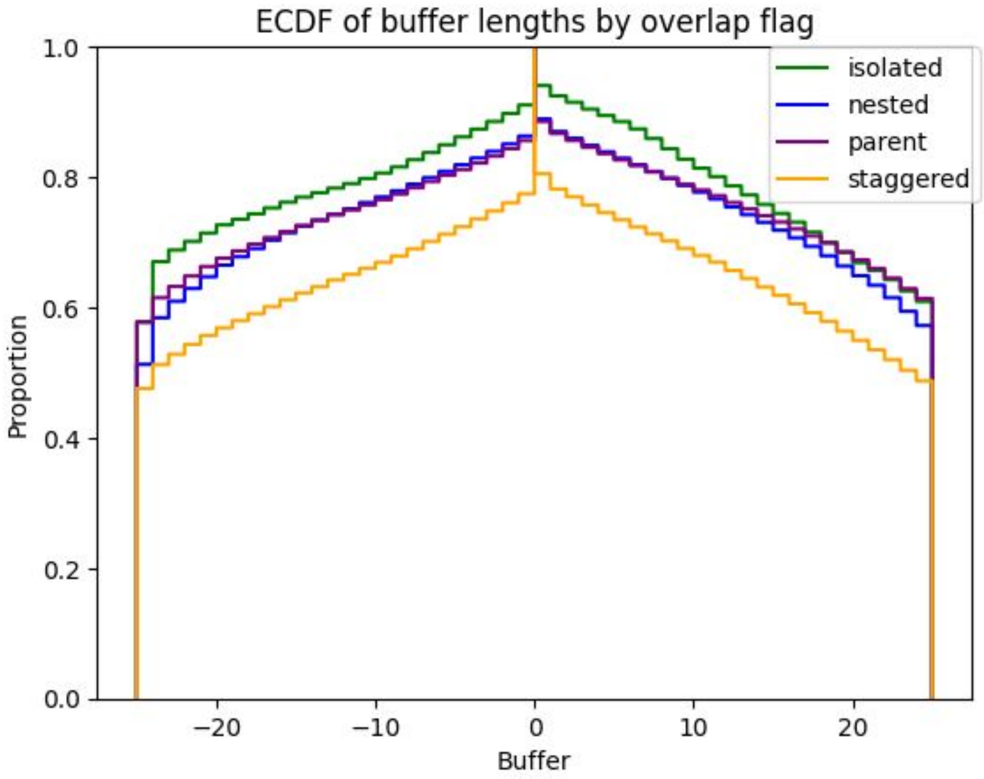


~50%/60% of regions have 25bp up_buff **OR** 25bp dn_buff

Boundary summary

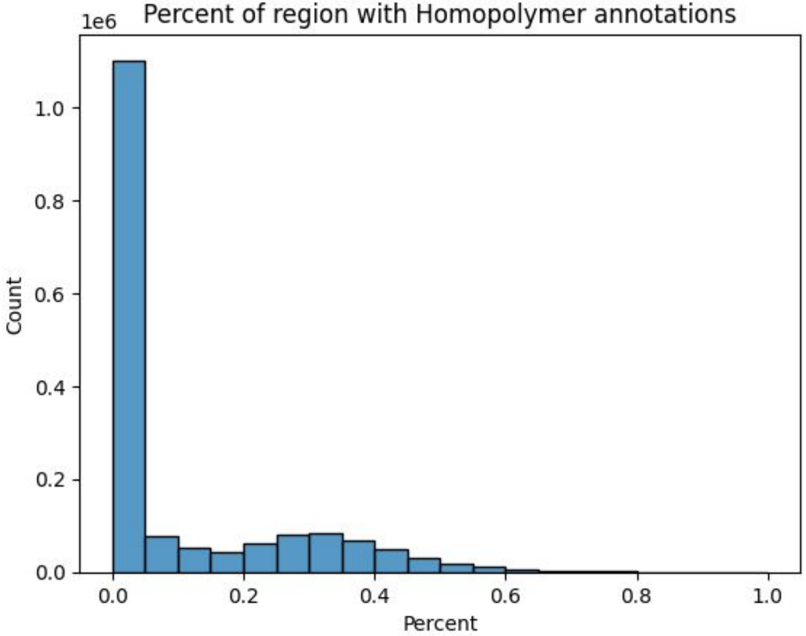
Min Buffer (Both Ends)	Count	Percent
5	1,297,178	77.01%
10	1,156,925	68.68%
20	892,996	53.01%
25	543,786	32.28%

- About half of our regions have at least ±20bp buffer



Homopolymers

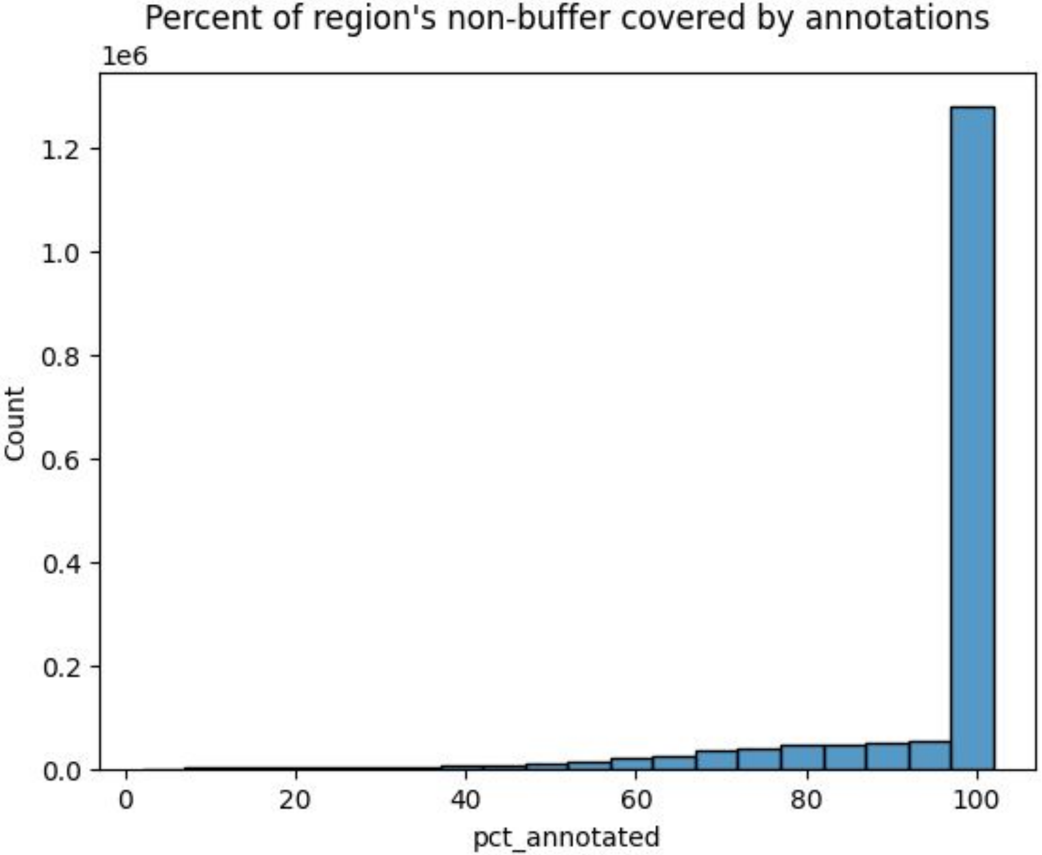
- We ignored homopolymers when calculating boundaries, which may affect our non-TR buffer sequence
- hom_span - total bases of region that *had* homopolymer annotations
- 39,761 (2.3%) of regions have >=50% of span with homopolymer annotations



Percent of non-buffer sequence annotated

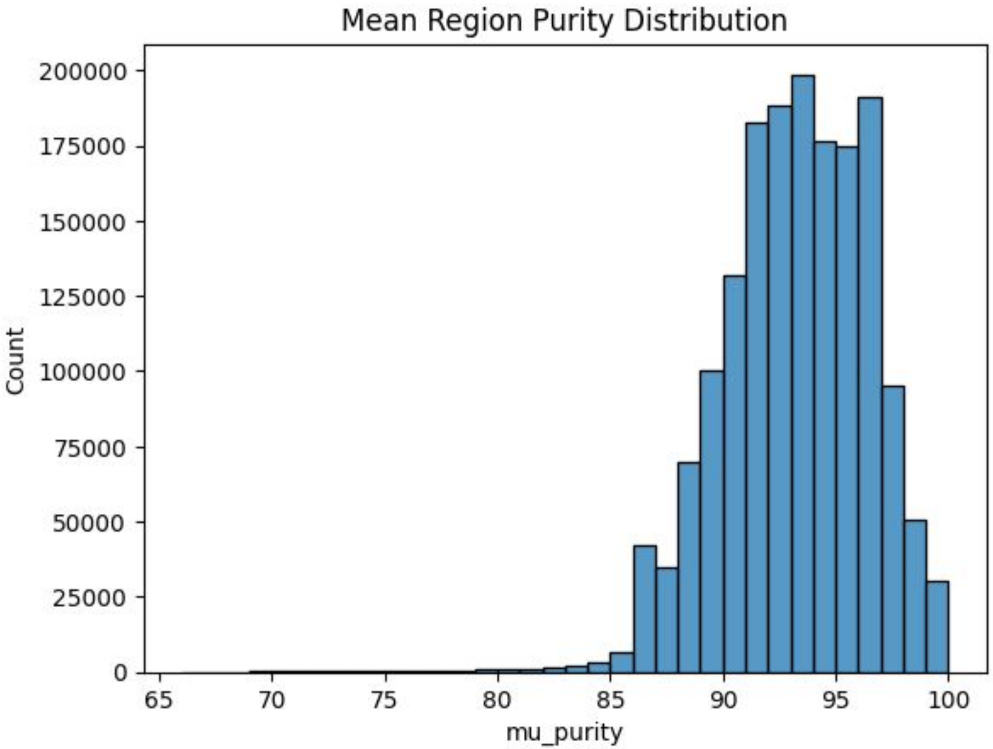
- Estimate of ‘density’
- Indicative of ‘gaps’ between the annotations of a region

- Of 1,644,682 regions with non-interspersed and hom_pct < 0.50
- 30,627 (1.9%) regions have <50% annotated



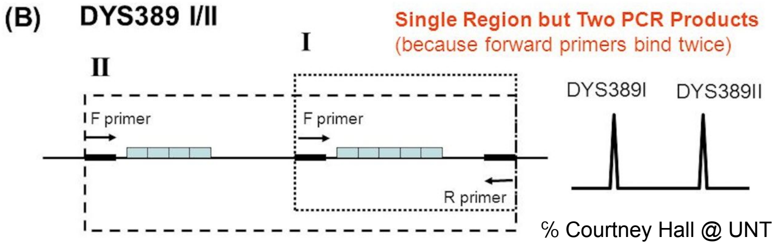
Repeat Purity

- Sequence similarity of `motif*copies` against annotation’s reference span.
- Region purity is average of its annotations’ purity



Patho/Codis

- 56 known pathogenic repeats
- 53 codis regions
- All are inside of the TR regions
- 4 codis are weird and within 2 TR regions


(B)
DYS389 I/II
Single Region but Two PCR Products
(because forward primers bind twice)
I
II
F primer
F primer
R primer
DYS389I
DYS389II

% Courtney Hall @ UNT

chrom	start	end	codis
chrY	12500448	12500495	DYS389I
chrY	12500448	12500611	DYS389II
chrY	18888804	18888851	DYS461
chrY	18888956	18888995	DYS460

Self overlapping

chrY:18888771-18889019
n_annos = 1
pct_annotated = 100

Known pathogenic repeats

- Four pathos pop out as suspicious
  - 4 intersect SINE
  - 2 have low up_buff
  - All are ‘sparse’ (multiple subregions, low pct_annotated)

						Hom	Num	Num	Num sub	Mu	Pct		
chrom	**start**	**end**	**ovl_flag**	**up_buff**	**dn_buff**	**span**	**filtered**	**annos**	**regions**	**purity**	**annotated**	**interspersed**	**patho**
chr16	17470858	17472830	29	13	25	35	9	10	9	91	18	SINE	XYLT1
chr2	96196773	96197458	5	1	25	21	5	5	5	95	18	SINE	STARD7
chr3	129171978	129172783	5	25	25	37	5	5	5	93	29	SINE	CNBP
chr3	183711960	183712250	1	25	25	12	3	2	2	97	28	SINE	YEATS2

Patho motif QC check

The reference set of pathogenic repeats have ‘known motifs’. Check how many match our TR region annotation motifs

- 41 of 56 motifs match
    - 18 needed ‘roll’ (e.g. GCC -> CCG)
- 7 patho motifs have ‘N’
    - All pass manual match (e.g. GCN -> GCC)
- 8 without a match
- **‘Correct’ Motifs: 48 / 56 = 87.5%**

Patho ID	Known Motifs	Adotto Motifs
NOTCH2NLC	GGC	CGC
ZIC2	GCN	**CAGCGGCGG**
NIPA1	GCG	CCCCCGTCCCG, GCGGCGGCA, GCC
FOXL2	NGC	GGCATG, **GCGGCTGCAGCCGCA**
AFF3	GCC	AG, GCCC, CGCCCGCCCCG, CGCGCGCCC, GCCCTGGGGCG
HOXD13	GCN	CCT, GGCCGGA, AGAGGGAG, TCG, GGCTAC, CGCTCCAG, CGGGGGCGC, CGCGC, GGGCCAGGGCC, **CGG**, **CGG**
CSTB	CGCGGGGCGGGG	GGGGCGC
PHOX2B	GCN	**CCG**, TGGGGT, **GCC**, CGCTGCCGCTGCG, GCCCCCGG
RUNX2	GCN	GGCGGCGGCGGCTGC, CCGCCCC, GCA
HOXA13	NGC	GCCCGCCGG, **GCG**, **GCC**, **GCC**, GGCGCCAAG, GGCCGG, GGCGGGATCGCGCCA
ZNF713	CGG	GCGGCGG*G*CGGCG, CCGCTGTCC
FXN	A,GAA	TACAAAAAAA
ARX	GCG	GCCC, GCT, GCCG, GCGGCC
SOX3	NGC	GTCTTG, GCCCAC, **GCGGCA**
TMEM185A	GCC	TCC, GCGCCA, **GCCGCC**

CAGCGGCGG--
|*||*||*||*
--GCN**gcn**GCN

Adotto motif is 3 copies of known motif

"Broken up" Annotations

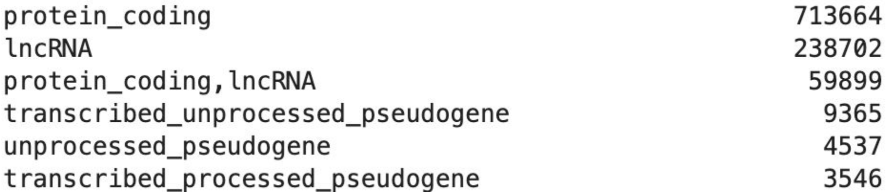
Truly non-parsimonious

Genes

Intersect with Ensembl-105 and record `gene_flag` and `gene_biotype`. **1,045,093** (62%) 'non-interspersed' TRs intersect genes.

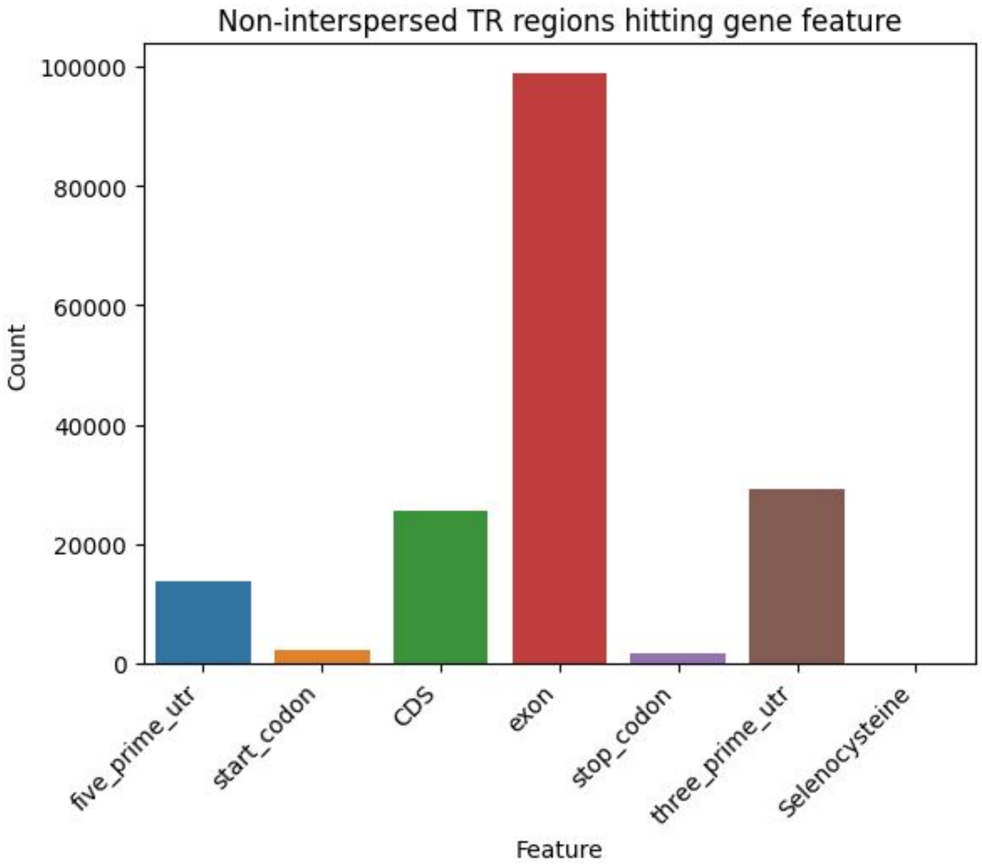
Feature	Bit	Feature	Bit
gene/transcript	1	exon	16
five_prime_utr	2	stop_codon	32
start_codon	4	three_prime_utr	64
CDS	8	Selenocysteine	128

biotype counts



```
protein_coding                                              713664
lncRNA                                                      238702
protein_coding,lncRNA                                        59899
transcribed_unprocessed_pseudogene                            9365
unprocessed_pseudogene                                        4537
transcribed_processed_pseudogene                              3546
```

…

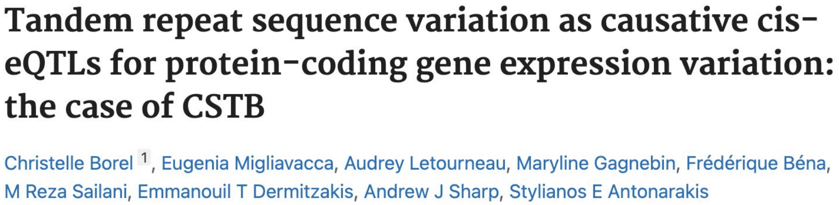


90.5% are 'intron_only'

One weird pathogenic TR

Our patho TR **CSTB** doesn’t intersect its gene ENSG00000160213

- Ensembl gene region:
  - chr21 43,772,511 43,776,330
- TR regions intersecting gene:
  - chr21 43,775,844 43,775,904
  - chr21 43,776,262 43,776,346
- CSTB TR region:
  - chr21 43,776,412 43,776,503 (downstream)



Tandem repeat sequence variation as causative cis-eQTLs for protein-coding gene expression variation: the case of CSTB

Christelle Borel [1], Eugenia Migliavacca, Audrey Letourneau, Maryline Gagnebin, Frédérique Béna, M Reza Sailani, Emmanouil T Dermitzakis, Andrew J Sharp, Stylianos E Antonarakis

“... the rare expansion of a repeat ... in **the promoter region** of the CSTB gene causes a silencing of the gene, resulting in progressive myoclonus epilepsy.”

v1.0 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_span	number of bases of the region found to be homopolymer repeats
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 (Enseml v105)
biotype	comma separated gene biotypes intersecting region (Enseml v105)
annos	JSON of TRF annotations in the region (list of dicts with keys: motif, entropy, ovl_flag, etc)