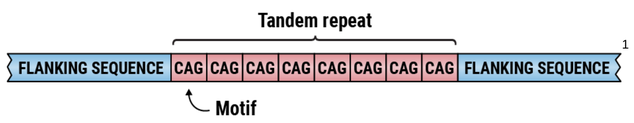
HMMSTR: A Modified Profile-HMM for Tandem Repeat Copy Number Determination in Nanopore Sequence Reads

Kinsey Van Deynze[1], Camille Mumm[2], Jessica Switzenberg[1], Alexandra Weber[1], Connor Maltby[3], Ryan Mills[1,2],Peter Todd[3], Alan P. Boyle[1,2]

[1]Department of Computational Medicine and Bioinformatics. [2]Department of Human Genetics. [3]Department of Neurology. University of Michigan, Ann Arbor, MI

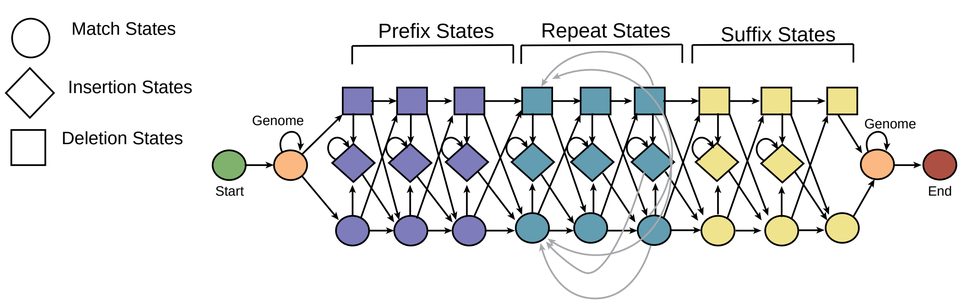
Introduction



- Known tandem repeats (TRs) makeup ~3% of the human genome[2]
- Cause over 50 human diseases including ALS, Ataxias, and Huntington's Disease
- Short read sequencing methods fail to accurately characterize TRs
- Oxford Nanopore Technologies (ONT) produce reads up to 2Mb
- ONT is compatible with multiple targeted enrichment techniques
- Accurate TR copy number determination is critical for understanding TR expansion pathogenesis and diagnostics

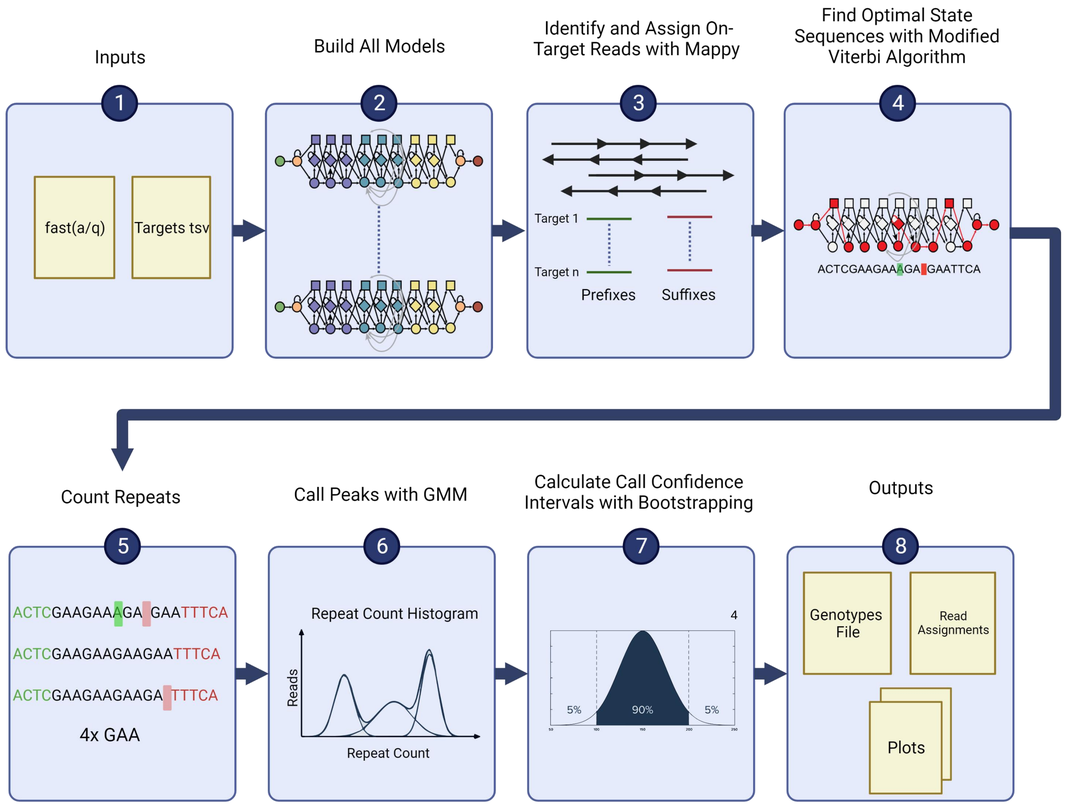
We propose a targeted technique for TR copy number determination that combines Cas9 targeted Nanopore sequencing with a modified profile-HMM based counting algorithm

Methods



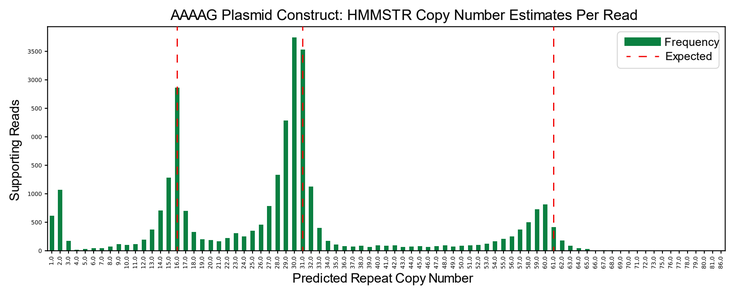
- Proposed modified Profile-HMM. It is built on 3 distinct regions: upstream flanking "Prefix", repeat motif, and downstream flanking "Suffix"
- Model varies with inputted prefix, repeat motif, and suffix
- Additional edges allow for paths through variable repeat copy number
- Emission and transition probabilities based on up-to-date ONT sequencing error rates or custom input

Workflow

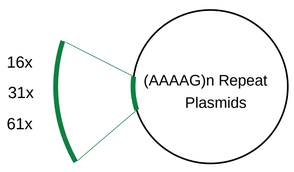


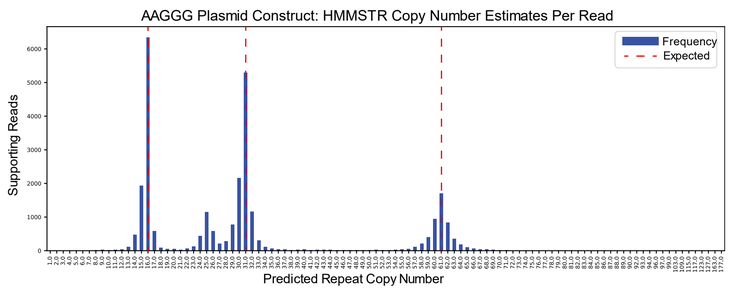
Results: Plasmid Constructs

- Plasmids constructed with variable motifs and copy number
- Dashed lines indicate intended insertion size

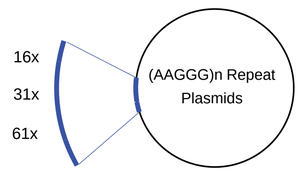


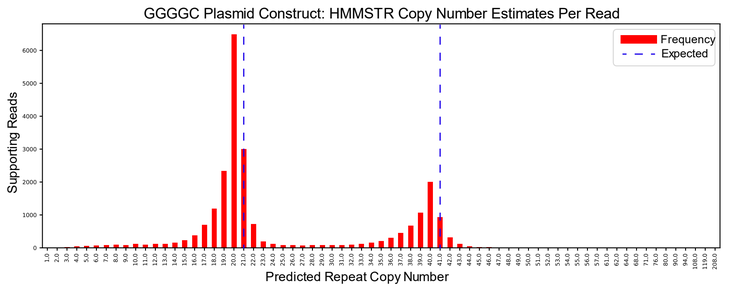
HMMSTR calls: 16, 30, 60



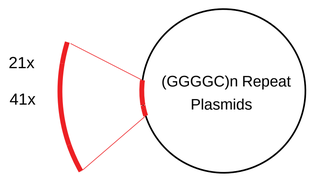


HMMSTR calls: 16, 25, 31, 61





HMMSTR calls: 20, 40

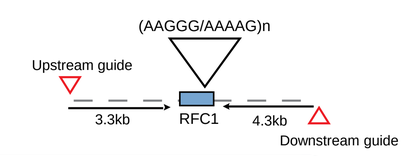


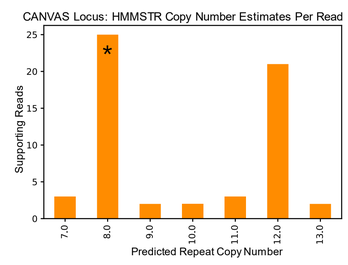
Results: Disease Associate Loci in GM12878

- HMMSTR was run on a series of Nanopore Cas9-Targeted Sequencing runs targeting 1 or 48 disease associated loci in the GM12878 lymphoblastoid cell line

CANVAS Locus

- Cerebellar ataxia, neuropathy, vestibular areflexia syndrome (CANVAS)[3]
- Homozygous expansion in an *RFC1* intron
- Normal motif and range: AAAAG 11 - 200
- Pathogenic motif and range: AAGGG 400 - >2000



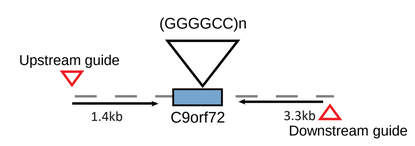


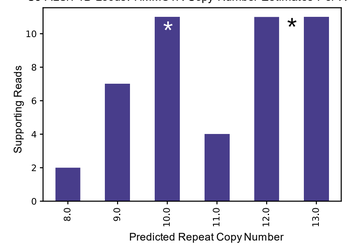
GM12878 repeat counts
- 60 reads spanning the repeat
- HMMSTR call: 8 and 12

* indicates TRF count from PacBio CCS assembly of GM12878

C9 ALS/FTD Locus

- Amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD)[4]
- Heterozygous expansion in an *C9orf72* intron
- Normal motif and range: GGGGCC < 12
- Pathogenic motif and range: GGGGCC > 30



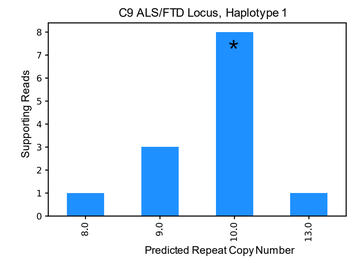


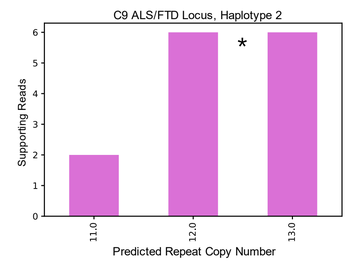
GM12878 repeat counts
- 49 reads spanning the repeat
- HMMSTR call: 10 and 12.5

Phased reads validate heterozygous repeat count

- Haplotypes assigned to reads using LRphase
- HMMSTR run on each haplotype separately

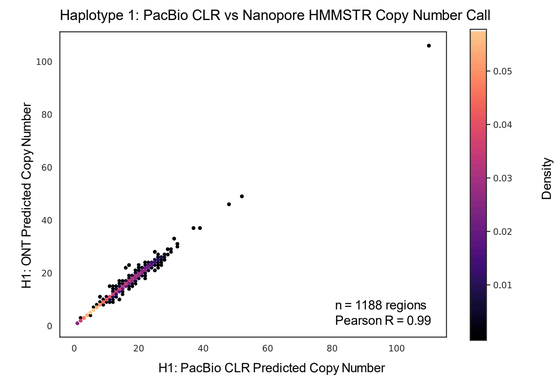
HMMSTR Calls: 10 and 12

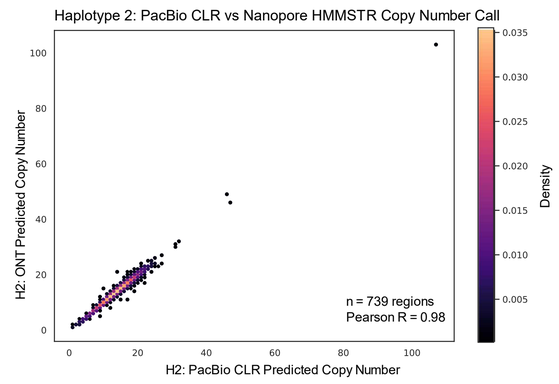


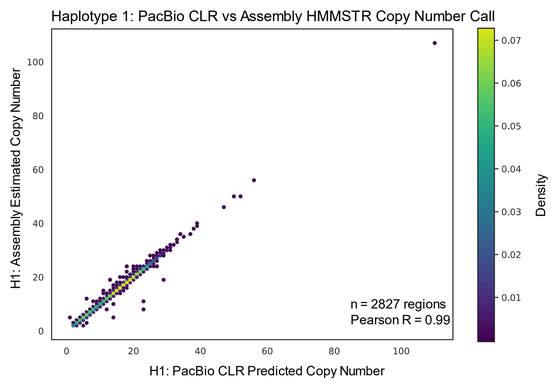


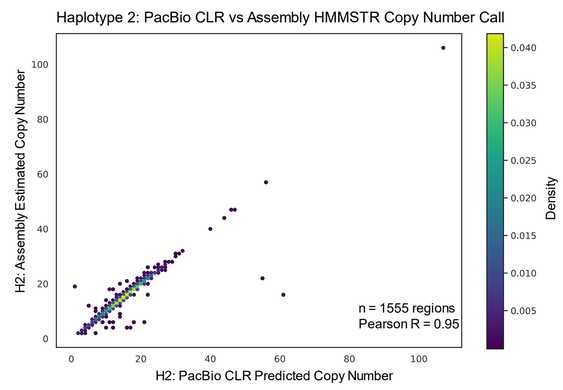
Results: WGS ONT, PacBio, and PacBio CCS Assembly Concordance

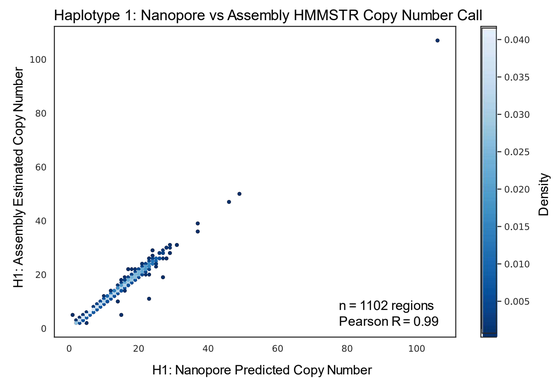
- HMMSTR was run on 11 thousand regions overlapping repeats identified in Sanger sequencing dataset by TRF[5] on:
  (1) GM12878 Whole Genome Sequencing Nanopore data from the Nanopore WGS Consortium [6]
  (2) GM12878 WGS PacBio (CLR) data
- Flanking sequences for each repeat target were aligned to a GM12878 PacBio CCS Assembly from the Human Genome Structural Variation Consortium (HGSVC) to extract repeat regions. TRF was then ran on each to get all assembly copy number estimations [7]
- Regions were filtered by the following per comparison:
  (1) Repeats detected in region for both datasets
  (2) > 5 supporting reads in experimental dataset(s)
  (3) Input prefix and suffix < 70% masked by TRF

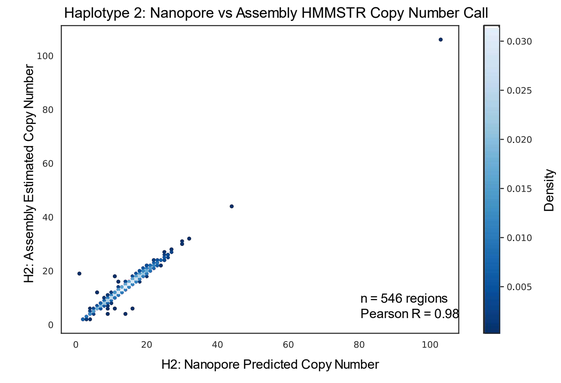












Acknowledgements and References

- Boyle lab for sequencing, Todd lab for plasmid constructs, Mills lab for PacBio CLR Data
- Genome Science Training Program (GSTP, NHGRI T32)
- Taubman Institute

1. (Dolzhenko, E. et al. 2021)
2. (Balendra et al 2018)
3. (Chintalaphani et al., 2021)
4. https://sixsigmadsi.com/glossary/confidence-interval/
5. (Benson 1999)
6. (Jain, M. et al. 2018)
7. (Porubsky, D. 2020)

the BOYLE LAB