Penn
UNIVERSITY of PENNSYLVANIA

Exploring Quantification Methods for Simple and Complex Tandem Repeats on Nanopore Sequencing Data

Kyle Liao, PURM, College of Arts and Sciences, Class of 2025

Faculty Mentor: Kai Wang | Perelman School of Medicine at the University of Pennsylvania, Department of Pathology and Laboratory Medicine | Children's Hospital of Philadelphia

Children's Hospital of Philadelphia
RESEARCH INSTITUTE

INTRODUCTION

A tandem repeat is a sequence of DNA bases that is repeated many times within a chromosome. Tandem repeats can be grouped into variable number tandem repeats (VNTRs) and short tandem repeats (STRs). The key difference between these two groups is that VNTRs consist of comparatively longer repeating units. Because STRs and VNTRs each make up approximately 3% of the human genome and are highly variable from person to person, they are often analyzed as markers for identification. Outside of genetic identification, tandem repeats are particularly important because they are often associated with diseases like Huntington's disease, fragile X syndrome, and bipolar disorder.

Historically, genotyping of tandem repeats has included labor-intensive and costly methods like PCR followed by gel electrophoresis. The advent of next-generation sequencing, whereby genetic information is rapidly sequenced in short segments, allowed for more efficient analysis. However, because pathogenic repeats are often hundreds to thousands of base pairs long, long read sequencing can provide a better estimations of expansion length, especially for patient samples with highly expanded repeats.

This summer, my research involved analyzing large datasets by generating summary statistics, calculating sample enrichments, and estimating repeat counts. The focus of my project was to explore and evaluate different methods of quantifying tandem repeats (both VNTR and STR) based on long-read sequencing data, specifically from Oxford Nanopore Sequencing.

OBJECTIVES &SIGNIFICANCE

My main project involved working with Nanopore data generated by Dr. Egli's lab at Columbia University on INS and GFP VNTRs. The objective of my research was conduct analysis on two samples datasets to quantify how many repeats existed in specific regions of interest. Afterwards, I looked to evaluate repeat estimate results from multiple different methods to gauge their efficacy and precision.

Repeat estimation is important for many reasons. Diseases associated with STRs often display a phenomenon known as genetic anticipation, whereby the risk of disease increases as generations pass (and STRs proliferate). Oftentimes the probability of disease also increases alongside repeat count. Being able to accurately quantify repeats allows for a better understanding of STR expansion disorders.

METHODS & DATA VISUALIZATION

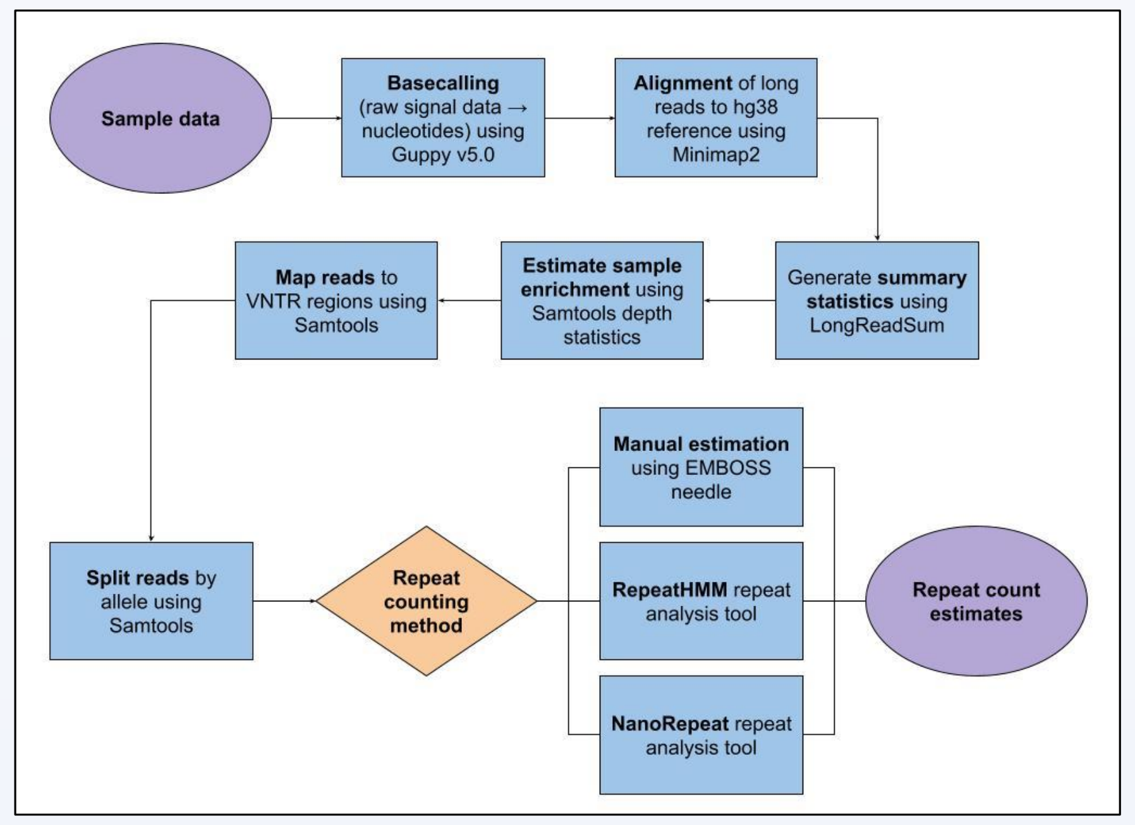
Regions of interest for Columbia VNTR data:

- chromosome: starting base-ending base – repeating motif
- chr9: 27573484-27573547 – GCCCCG
- chr11: 2161569-2161976 – CTGTCCCACACCC

Data was split into 2 samples:

- VNTRA48 data from 09/24/2021
- VNTRA48 data from 03/11/2021

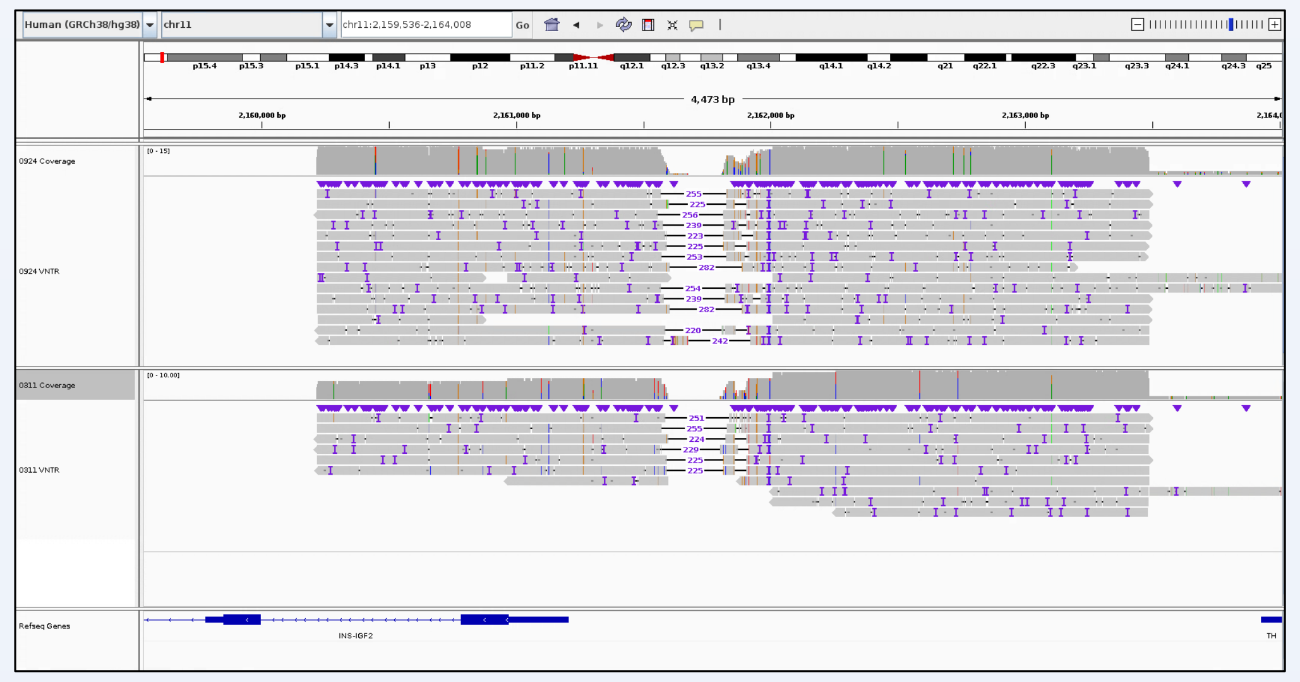
Fig. 1: Data analysis workflow for VNTR data



Sample data → **Basecalling** (raw signal data → nucleotides) using Guppy v5.0 → **Alignment** of long reads to hg38 reference using Minimap2 → Generate **summary statistics** using LongReadSum → **Estimate sample enrichment** using Samtools depth statistics → **Map reads** to VNTR regions using Samtools → **Split reads** by allele using Samtools → **Repeat counting method** → **Manual estimation** using EMBOSS needle / **RepeatHMM** repeat analysis tool / **NanoRepeat** repeat analysis tool → **Repeat count estimates**

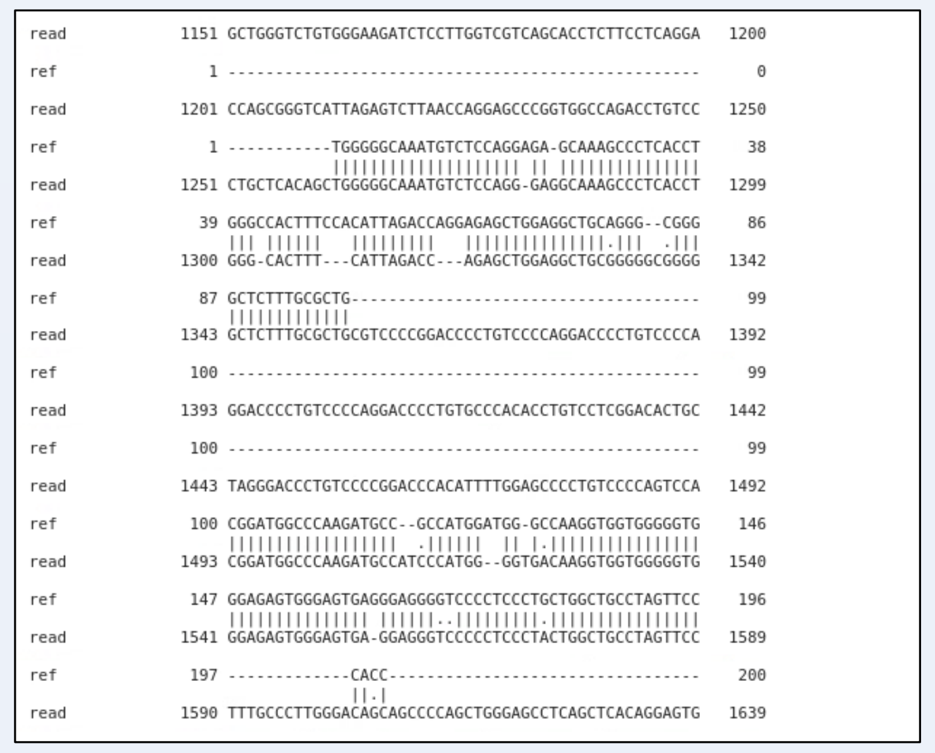
Integrated genomics viewer (IGV) was used to visualize sequence data. The breaks in the sequence are the investigated repeat expansions.

Fig. 2: IGV view of chr11 VNTR



EMBOSS Needle was used for pairwise alignment; repeat count was then inferred by counting repeats between anchor sequences (where lines connect ref and read).

Fig. 3: EMBOSS Needle alignment

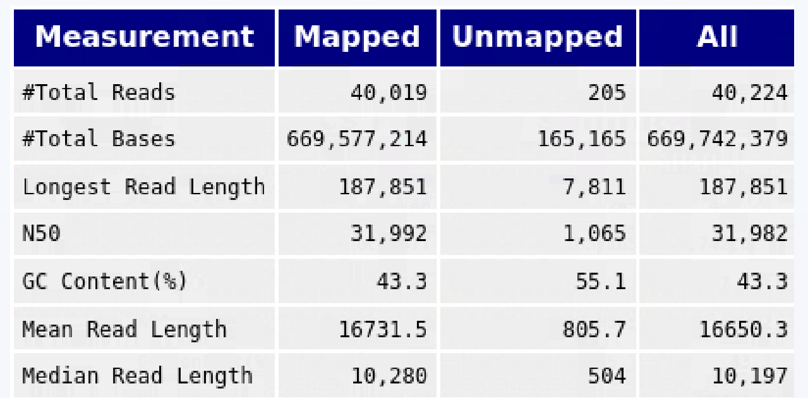


```
read            1151 GCTGGGTCTGTGGGAAGATCTCCTTGGTCGTCAGCACCTCTTCCTCAGGA   1200

ref                1 --------------------------------------------------      0

read            1201 CCAGCGGGTCATTAGAGTCTTAACCAGGAGCCCGGTGGCCAGACCTGTCC   1250

ref                1 -----------TGGGGGCAAATGTCTCCAGGAGA-GCAAAGCCCTCACCT     38
                                |||||||||||||||||||| || |||||||||||||||
read            1251 CTGCTCACAGCTGGGGGCAAATGTCTCCAGG-GAGGCAAAGCCCTCACCT   1299

ref               39 GGGCCACTTTCCACATTAGACCAGGAGAGCTGGAGGCTGCAGGG--CGGG     86
                     ||| ||||||   |||||||||   |||||||||||||||.|||  .|||
read            1300 GGG-CACTTT---CATTAGACC---AGAGCTGGAGGCTGCGGGGGCGGGG   1342

ref               87 GCTCTTTGCGCTG-------------------------------------     99
                     |||||||||||||
read            1343 GCTCTTTGCGCTGCGTCCCCGGACCCCTGTCCCCAGGACCCCTGTCCCCA   1392

ref              100 --------------------------------------------------     99

read            1393 GGACCCCTGTCCCCAGGACCCCTGTGCCCACACCTGTCCTCGGACACTGC   1442

ref              100 --------------------------------------------------     99

read            1443 TAGGGACCCTGTCCCCGGACCCACATTTTGGAGCCCCTGTCCCCAGTCCA   1492

ref              100 CGGATGGCCCAAGATGCC--GCCATGGATGG-GCCAAGGTGGTGGGGGTG    146
                     ||||||||||||||||||  .||||||  || |.||||||||||||||||
read            1493 CGGATGGCCCAAGATGCCATCCCATGG--GGTGACAAGGTGGTGGGGGTG   1540

ref              147 GGAGAGTGGGAGTGAGGGAGGGGTCCCCTCCCTGCTGGCTGCCTAGTTCC    196
                     ||||||||||||||| ||||||..|||||||||.||||||||||||||||
read            1541 GGAGAGTGGGAGTGA-GGAGGGTCCCCCTCCCTACTGGCTGCCTAGTTCC   1589

ref              197 ------------CACC----------------------------------    200
                                 ||.|
read            1590 TTTGCCCTTGGGACAGCAGCCCCAGCTGGGAGCCTCAGCTCACAGGAGTG   1639
```

RESULTS & CONCLUSIONS

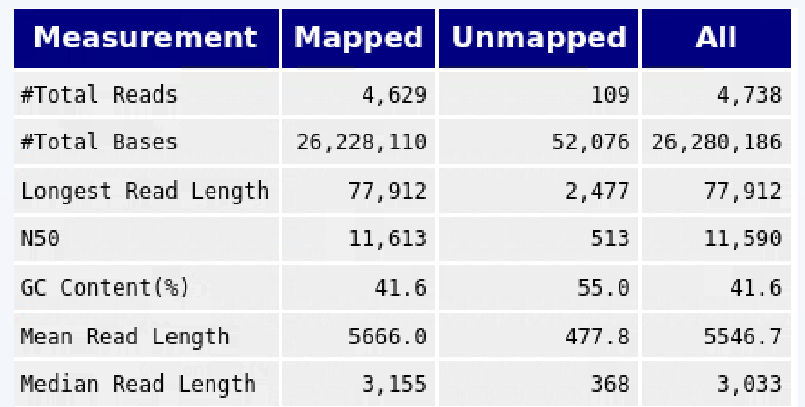
LongReadSum was used to generate summary statistics of the Columbia VNTR datasets, including information like total reads, total bases, and mean read length.

Fig. 4: Summary statistics for 09/24 data



Measurement	Mapped	Unmapped	All
#Total Reads	40,019	205	40,224
#Total Bases	669,577,214	165,165	669,742,379
Longest Read Length	187,851	7,811	187,851
N50	31,992	1,065	31,982
GC Content(%)	43.3	55.1	43.3
Mean Read Length	16731.5	805.7	16650.3
Median Read Length	10,280	504	10,197

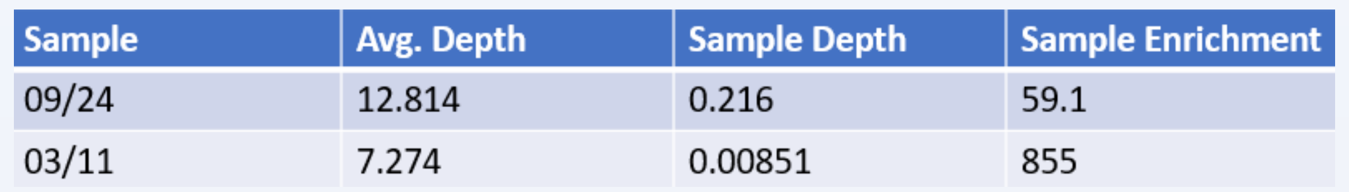
Fig. 5: Summary statistics for 03/11 data



Measurement	Mapped	Unmapped	All
#Total Reads	4,629	109	4,738
#Total Bases	26,228,110	52,076	26,280,186
Longest Read Length	77,912	2,477	77,912
N50	11,613	513	11,590
GC Content(%)	41.6	55.0	41.6
Mean Read Length	5666.0	477.8	5546.7
Median Read Length	3,155	368	3,033

Enrichment estimates were also calculated for each sample. This was done by calculating the ratio of the average depth of each nucleotide in the region of interest to the depth of the sample relative to the whole genome.

Fig. 6: Enrichment estimates

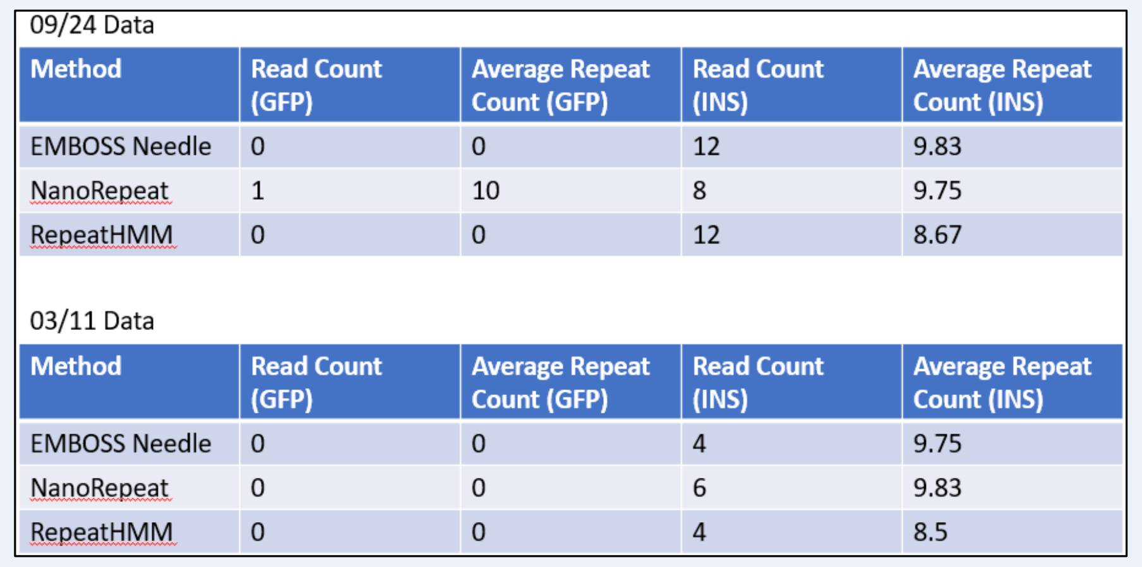


Sample	Avg. Depth	Sample Depth	Sample Enrichment
09/24	12.814	0.216	59.1
03/11	7.274	0.00851	855

For the 2 samples, the region of interest at chr11 has approximately 60- and 850-times enrichment for 09/24 and 03/11 respectively.

After using EMBOSS Needle pairwise alignment followed by manual repeat counting between anchors (fig. 3), as well as feeding reads split by allele into NanoRepeat and RepeatHMM, results were stored in the following table.

Fig. 7: Comparison of VNTR Repeat Estimate Results



09/24 Data

Method	Read Count (GFP)	Average Repeat Count (GFP)	Read Count (INS)	Average Repeat Count (INS)
EMBOSS Needle	0	0	12	9.83
NanoRepeat	1	10	8	9.75
RepeatHMM	0	0	12	8.67

03/11 Data

Method	Read Count (GFP)	Average Repeat Count (GFP)	Read Count (INS)	Average Repeat Count (INS)
EMBOSS Needle	0	0	4	9.75
NanoRepeat	0	0	6	9.83
RepeatHMM	0	0	4	8.5

Overall, samples of VNTRA48 had very few supporting reads for the GFP allele, making comparison between tools hard if only GFP allele read count was available. The repeat count for INS allele reads in both samples averaged between 8 and 10 across all methods of repeat detection. NanoRepeat estimates are closer to manual estimates in terms of average repeat count.

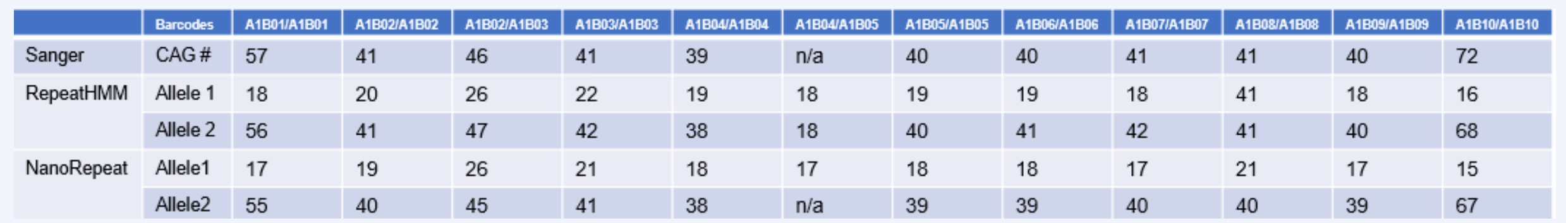
DISCUSSION & EXTENSION

This analysis of VNTR datasets yielded evidence that both RepeatHMM and NanoRepeat are accurate for estimating VNTR repeat count and calculated similar results to manual repeat estimation.

For future extension, I am currently working with data generated by Dr. Li Fang, Dr. Mas Monteys, and Dr. Davidson at CHOP. There are 11 Huntington's Disease (HTT) cell lines where repeat count is inferred/validated through Sanger sequencing. HTT is known to be caused by a STR expansion of trinucleotide CAG repeats. Comparing results from different tools would allow for a more standardized comparison between tools when evaluating their repeat detection ability.

Preliminary results of running RepeatHMM and NanoRepeat on each cell line are depicted below. Each cell line is uniquely identified by two barcodes (format: barcode/barcode). A1B04/A1B05 is a normal human cell line. The HTT region of interest is chr4: 3074876-3074939. Both tools provide similar estimates to validated results.

Fig. 8: Comparison of repeat Estimates for HD lines



	Barcodes	A1B01/A1B01	A1B02/A1B02	A1B02/A1B03	A1B03/A1B03	A1B04/A1B04	A1B04/A1B05	A1B05/A1B05	A1B06/A1B06	A1B07/A1B07	A1B08/A1B08	A1B09/A1B09	A1B10/A1B10
Sanger	CAG #	57	41	46	41	39	n/a	40	40	41	41	40	72
RepeatHMM	Allele 1	18	20	26	22	19	18	19	19	18	41	18	16
	Allele 2	56	41	47	42	38	18	40	41	42	41	40	68
NanoRepeat	Allele1	17	19	26	21	18	17	18	18	17	21	17	15
	Allele2	55	40	45	41	38	n/a	39	39	40	40	39	67

REFERENCES

1. Guppy https://nanoporetech.com/
2. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34:3094-3100. doi:10.1093/bioinformatics/bty191
3. Li, H. et al. The Sequence Alignment/Map format and SAMtools, *Bioinformatics*, Volume 25, Issue 16, 15 August 2009, Pages 2078–2079, https://doi.org/10.1093/bioinformatics/btp352
4. LongReadSum https://github.com/WGLab/LongReadSum
5. Robinson, J., Thorvaldsdóttir, H., Winckler, W. *et al.* Integrative genomics viewer. *Nat Biotechnol* **29,** 24–26 (2011). https://doi.org/10.1038/nbt.1754
6. Madeira F, Pearce M, Tivey ARN, et al. Search and sequence analysis tools services from EMBL-EBI in 2022. Nucleic Acids Research. 2022 Apr:gkac240. DOI: 10.1093/nar/gkac240.
7. NanoRepeat https://github.com/WGLab/NanoRepeat
8. Liu, Q., Zhang, P., Wang, D. *et al.* Interrogating the "unsequenceable" genomic trinucleotide repeat disorders by long-read sequencing. *Genome Med* **9,** 65 (2017). https://doi.org/10.1186/s13073-017-0456-7
9. Fang, L., Liu, Q., Monteys, A.M. *et al.* DeepRepeat: direct quantification of short tandem repeats on signal data from nanopore sequencing. *Genome Biol* **23,** 108 (2022). https://doi.org/10.1186/s13059-022-02670-6

ACKNOWLEDGMENTS

I would like to thank my mentor, Dr. Wang, as well as Umair Ahsan, Dr. Li Fang, Dr. Qian Liu, Dr. Egli, Dr. Mas Monteys, and Dr. Davidson for their patience, guidance, and help ☺.

CONTACT

Liao, Kyle
University of Pennsylvania
Email: kyleliao@sas.upenn.edu