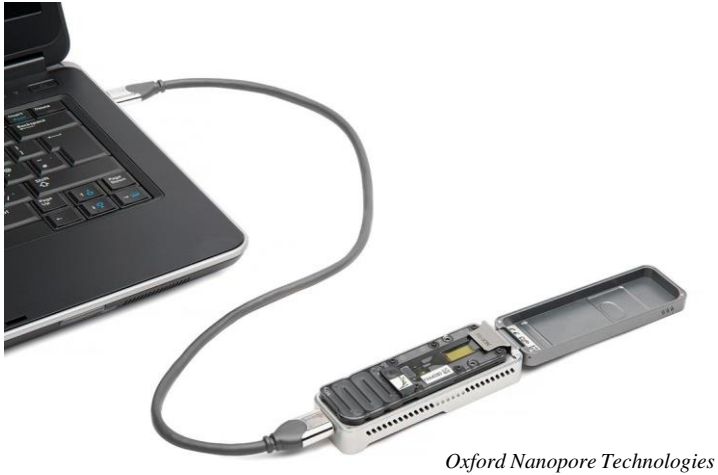
Diagnosis of Fasioscapulohumeral Dystrophy Through Nanopore Sequencing

Pavel Avdeyev
Computational Biology Institute
The George Washington University

Background on Nanopore

- Nanopore is a compact 3rd generation sequencing technology for both DNA and RNA
- Capable of sequencing ultra long reads (from 10kb - 100kb+)
- Provides compact, fast, and relatively low cost sequencing solution



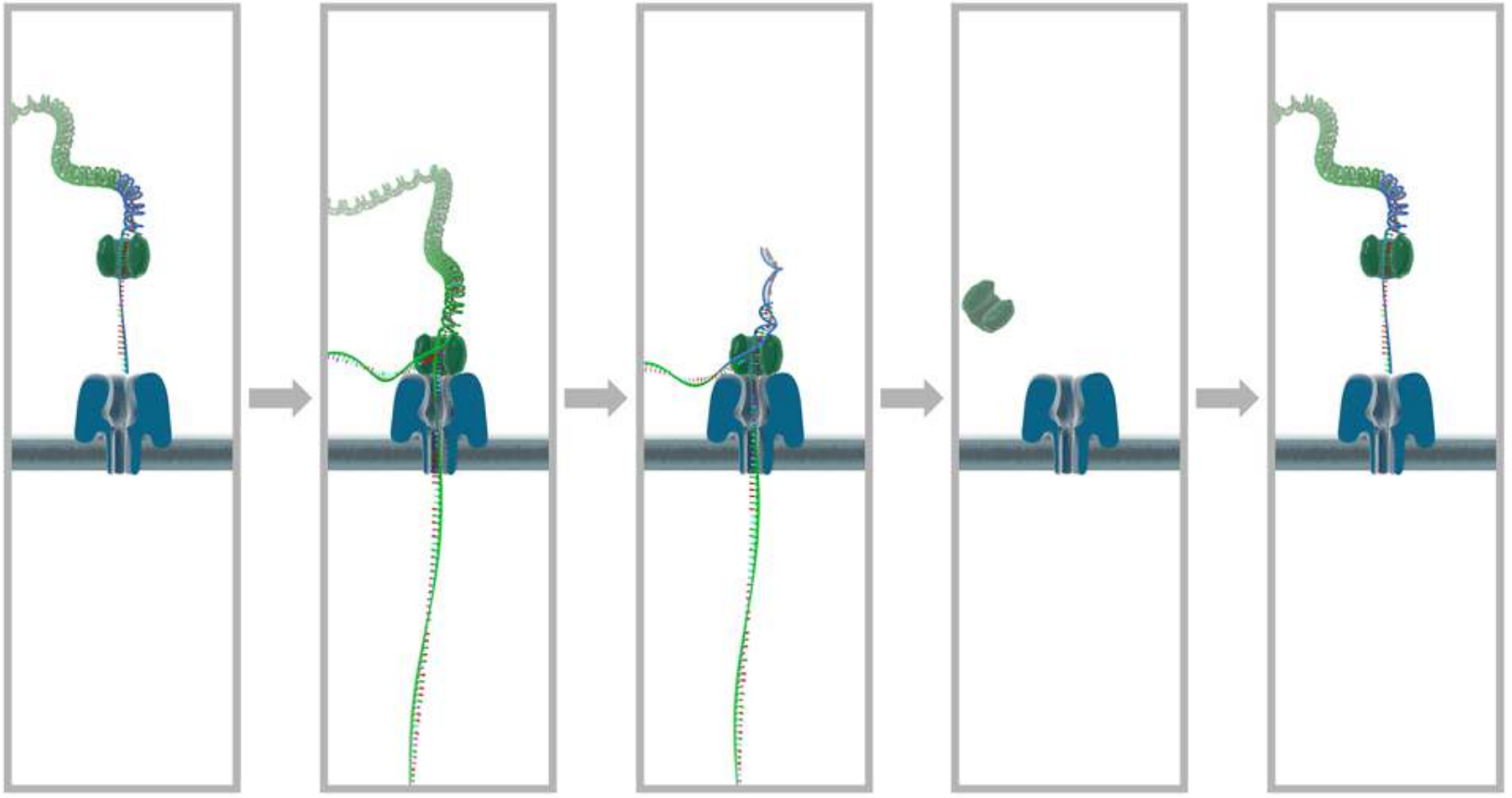
Oxford Nanopore Technologies

Products

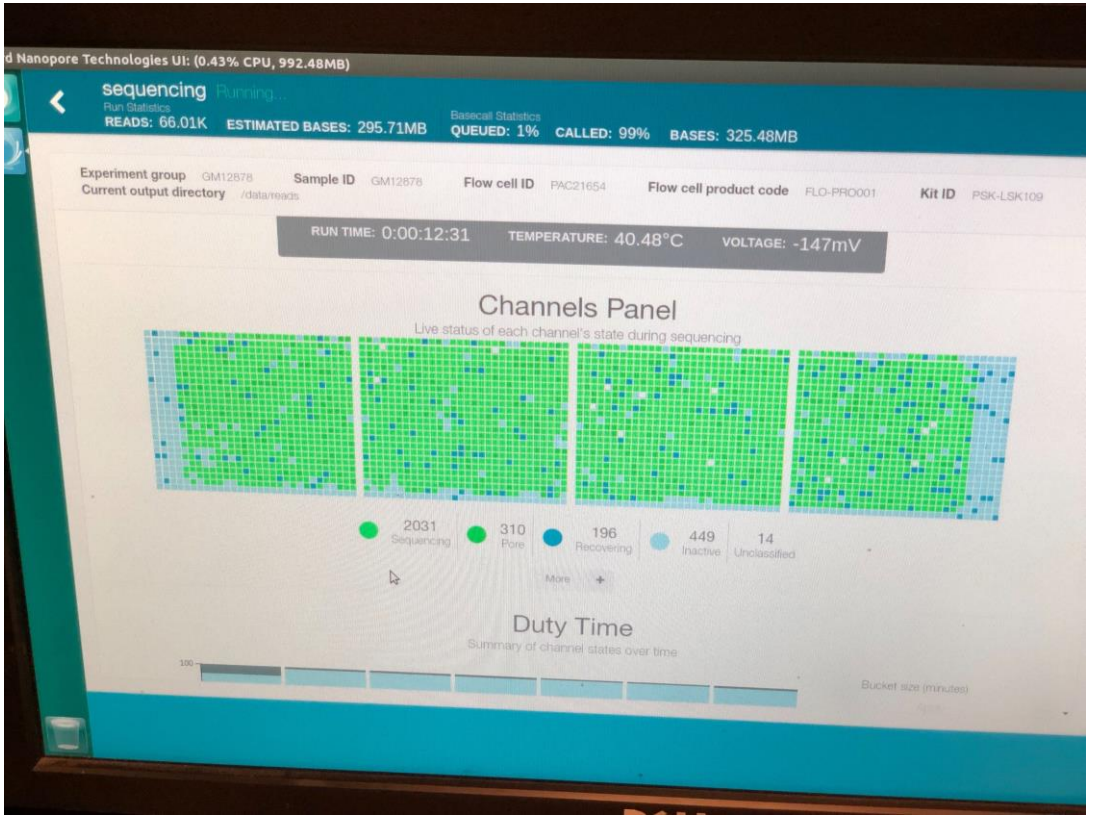
Products

- Automatic library preparation;
- ``Hands-free''.

How it Works

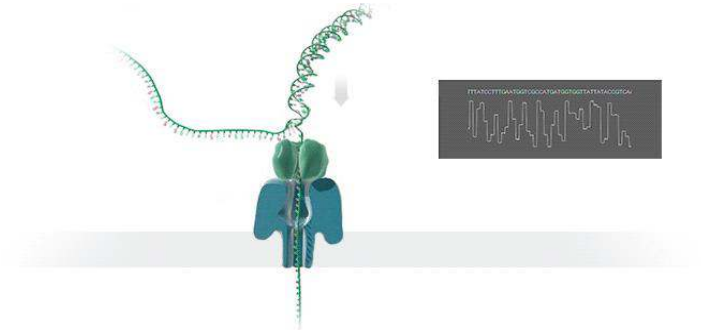


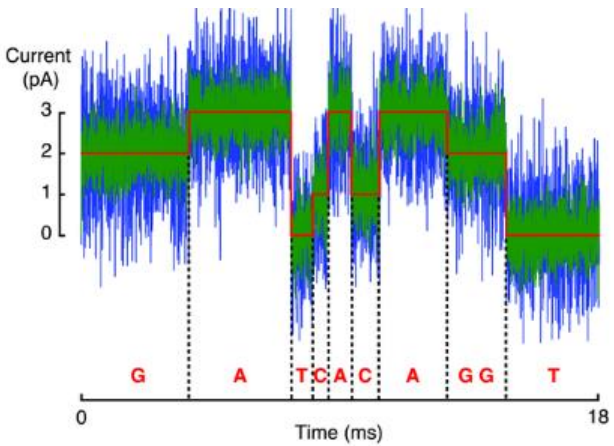
How it Works



1. Thar she blows! Ultra long read method for nanopore sequencing http://lab.loman.net/2017/03/09/ultrareads-for-nanopore/
2. Nanopore Sequencing Book: DNA extraction and purification methods http://lab.loman.net/2018/05/25/dna-extraction-book-chapter/

Basecallers

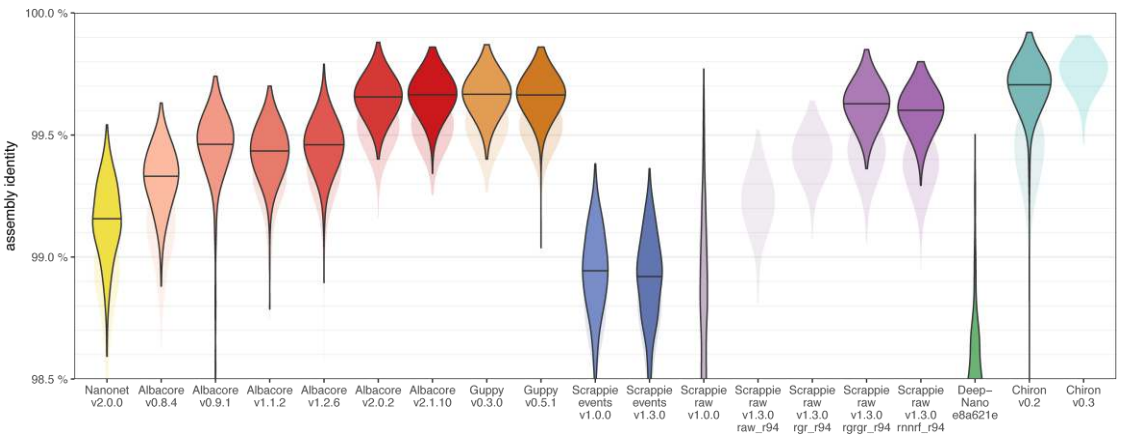




- Albacore
- Scrappie

Comparison of Oxford Nanopore basecalling tools

https://github.com/rrwick/Basecalling-comparison/blob/master/README.md



DNA Sequencing

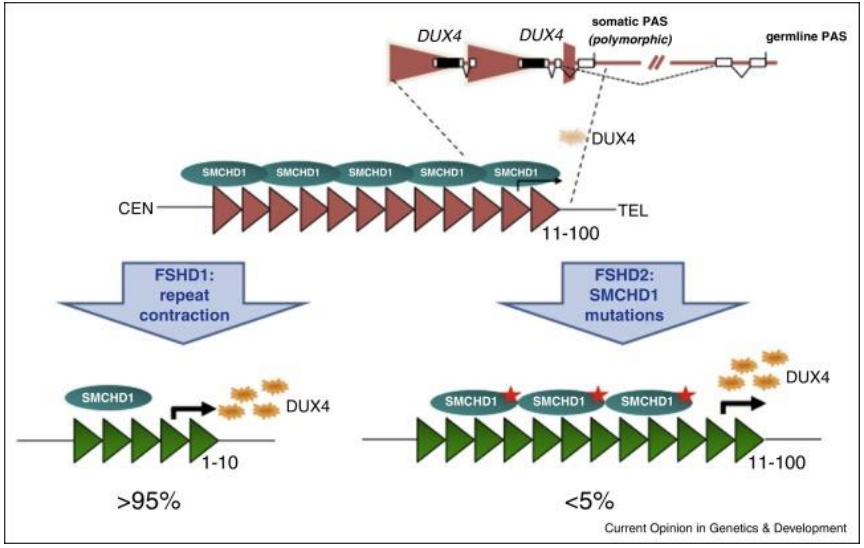
- DNA Sequencing (capable of generating reads from 40kb to 100kb+)
  - Error rates are roughly equivalent to PacBio at 10-15%[1], but are random leading to areas with high error rates
  - Still problems resolving highly methylated DNA
- Can be used for de novo whole genome sequencing
- Used for Structural Variation research in genomes

RNA/Future sequencing

- Capable of sequencing full transcripts of RNA
  - Direct from RNA
  - cDNA (with or without PCR)
- Can be used to characterize RNA viruses
- Also used to study gene expression and gene isoforms
- Working on improving resolution for methylated DNA through bioinformatic analysis
- Research being conducted on using nanopores to sequence proteins
- Real time analysis while sequencing

FSHD

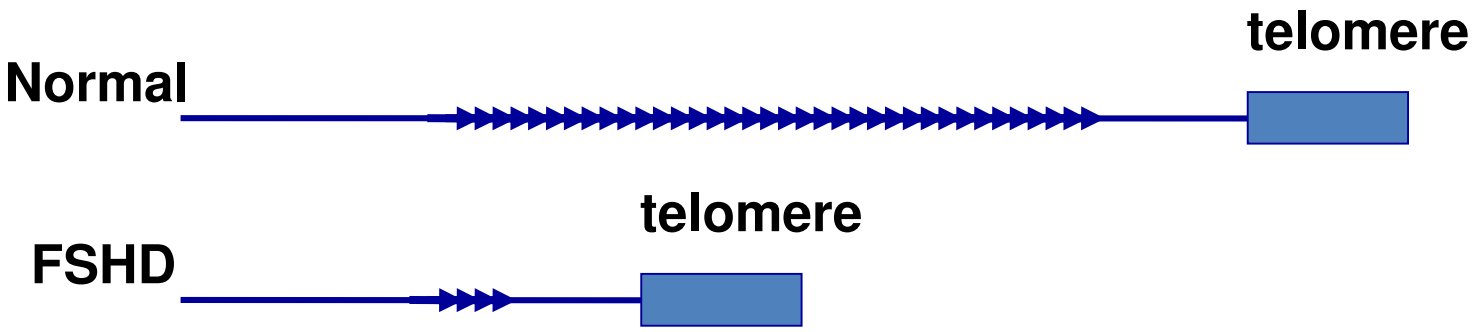
- Classifying Structural Variation to diagnose Facioscalpulohumeral Dystrophy (FSHD)
  - Disease is caused by a contraction of the D4Z4 repeated region w/ a 3.3kb repeat unit (from 10-50+ repeats to less than 10 units)
- Diagnosis currently requires a Pulse-Field Gel Electrophoresis in combination with Southern Blotting
  - Shown this can be done
- Nanopore sequencing may provide a method that is faster and cheaper than the current diagnosis



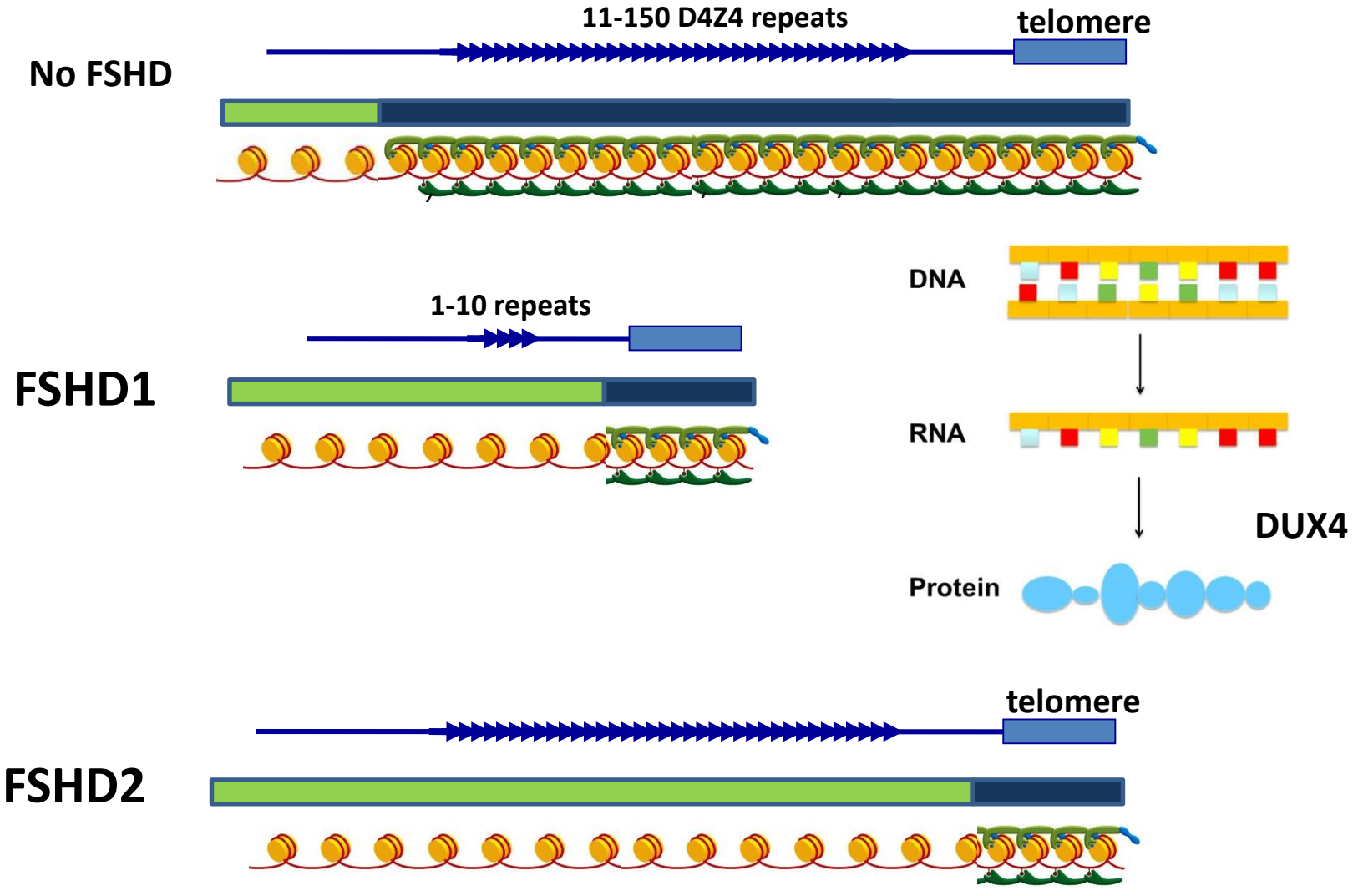
Genetic and Epigenetic contributors to FSHD, Science Direct

FSHD

- Incidence of 1:8,000 to 1:20,000 individuals
- Two types (we are concerned with the first)
  - FSHD1 caused by contraction of D4Z4 array
  - FSDH2 caused by mutations in gene that helps maintain structure of DNA in this region (SMCHD1)
- Onset varies, but severity of disease is proportional to size of contraction.




10 nm

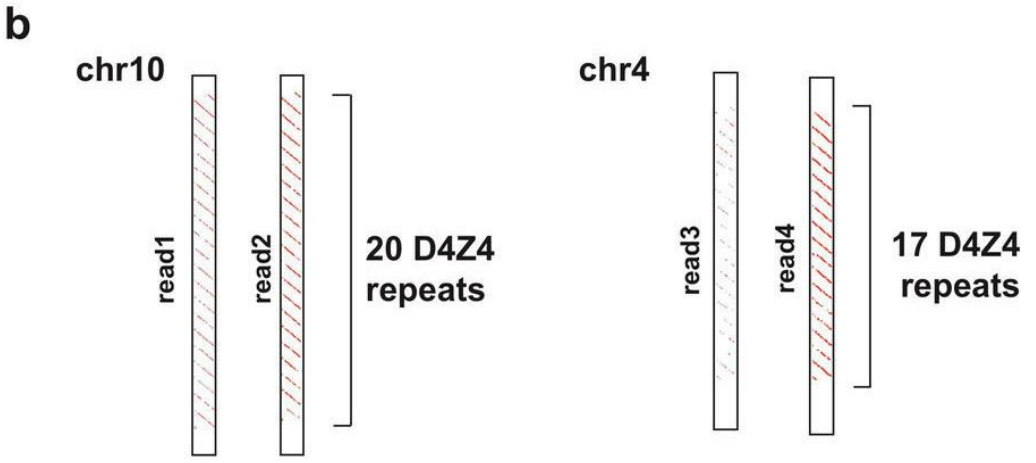

No FSHD
11-150 D4Z4 repeats
telomere
FSHD1
1-10 repeats
DNA
RNA
DUX4
Protein
FSHD2
telomere
Mutations in SMCHD1

Challenges

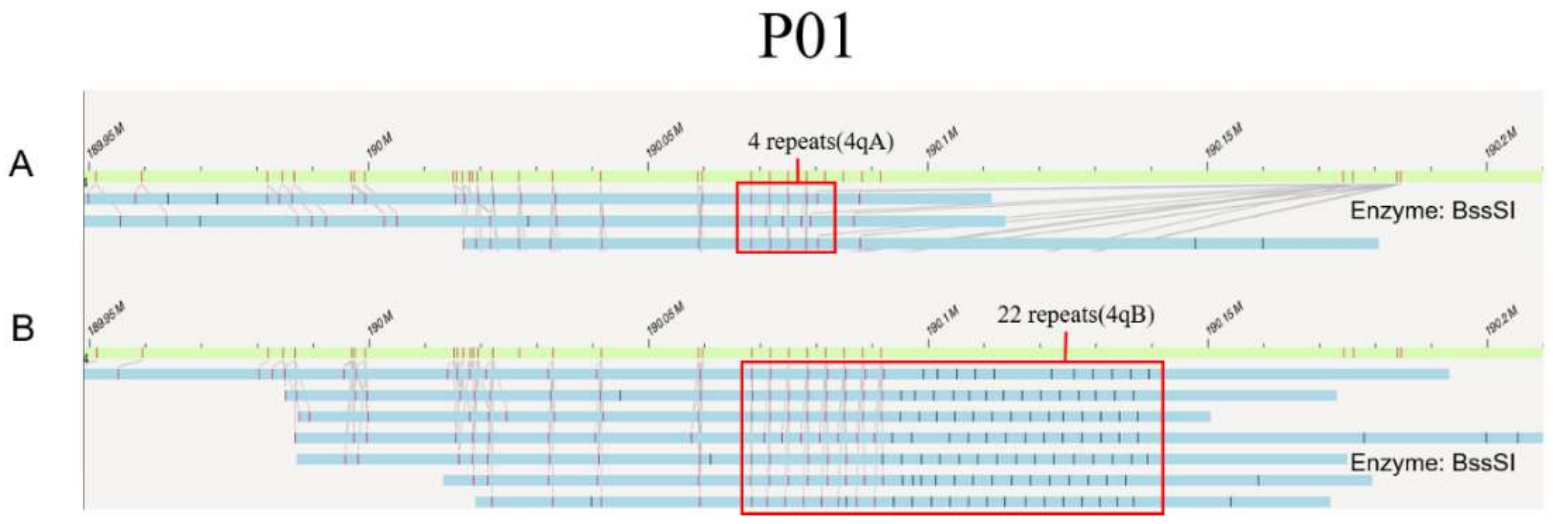
- D4Z4 repeat unit is 3.3kb, so reads of length 33kb+ are needed to detect an array that is 10 units or less
  - To add to the difficulty, we need a read that spans the whole region to definitively say we have a reduced region
- D4Z4 is not unique to chromosome 4, D4Z4-like sequences are found in many places in the genome
  - This means our problem is to not only find a read spanning the whole region, but also to say it is certainly from chromosome 4
- This repeat is on the telomeric end of chromosome 4, so DNA is heavily methylated
- Furthermore, we must uniquely identify between two haplotypes at the end of the array, pLAM A and B, as only A leads to the disease

State of Field

- A group in Japan managed to sequence a D4Z4 unit by using specific restriction sites for the region on chr4, BAC cloning the DNA they received, using Nanopore to sequence this cloned DNA
  - Using the sequence they produced, they were able to find 8 reads in Rel4 (ultra long WGS Nanopore data set with 1.4 million reads) that had alignment longer than 2kb, of which 2 reads aligned to chromosome 4 on GRCh38
  - They were able to use a D4Z4 sequence to measure the repeat region to be 17 units in length using these 2 chr4 reads
  - We consider this a proof of concept, though we would like to apply it to actual clinical data
- Another group managed to count the repeat region using Bionano, which uses fluorescent staining, not sequencing by staining methylation of array after restriction
  - Were able to measure the repeat region and identify between A and B haplotypes after one staining run

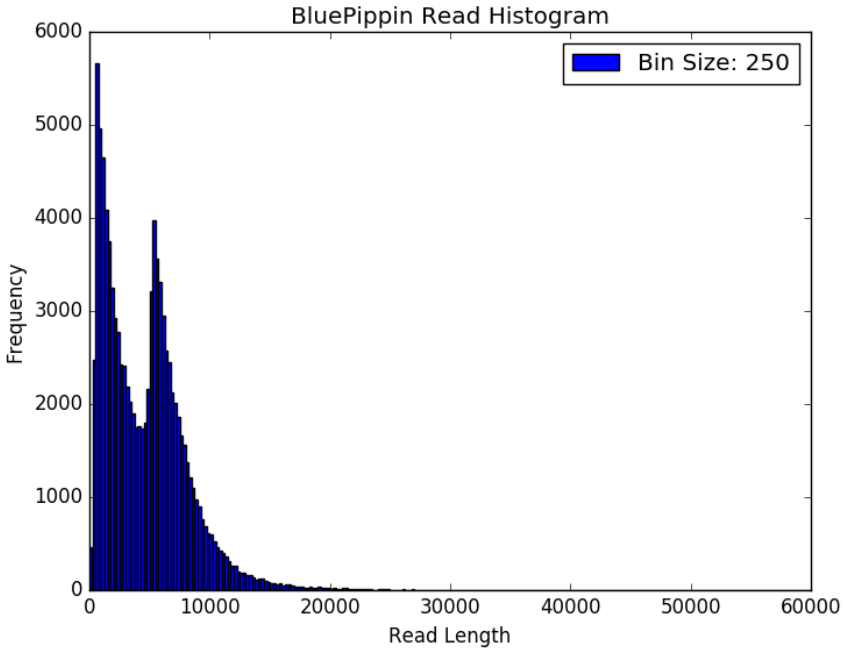


Mitsuhashi et al. 2017

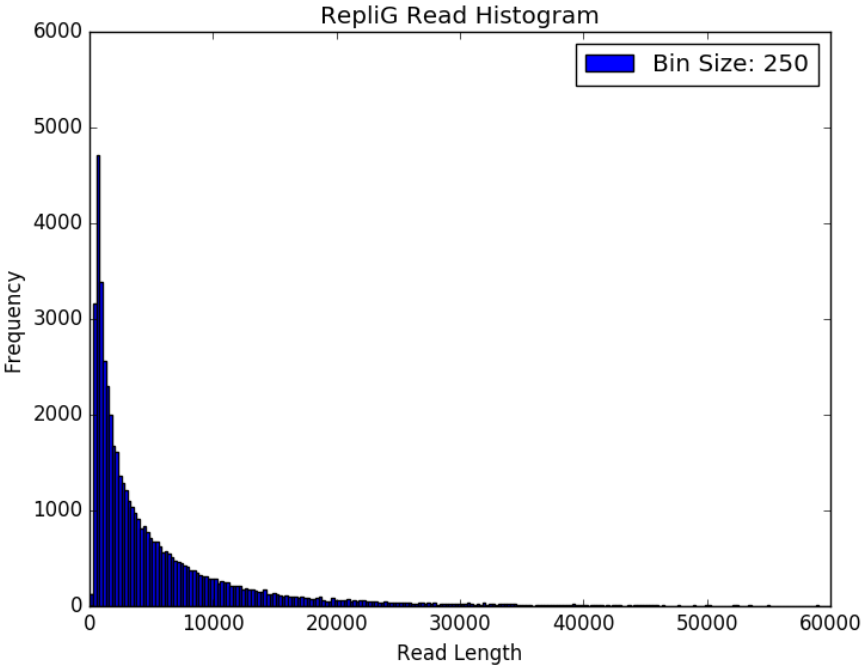


Dai et al. 2018

Data Sets

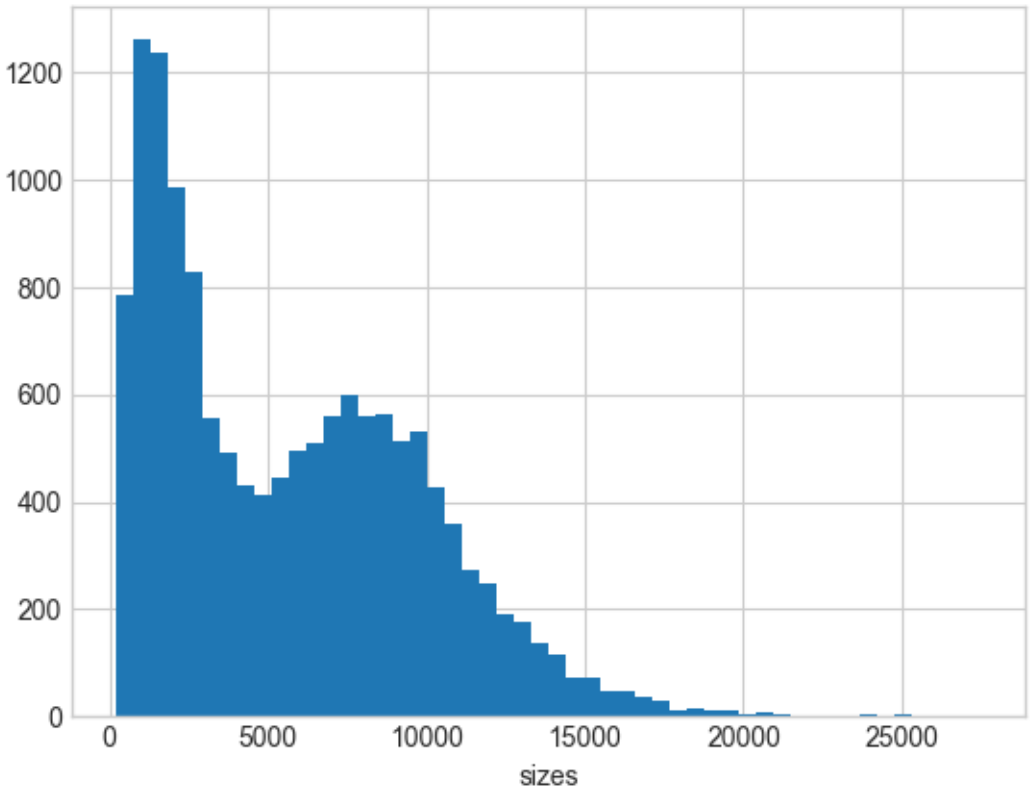


- 97k total reads
- Longest read is 604kb



- 28k total reads
- Longest read is 160kb

How It Should Look

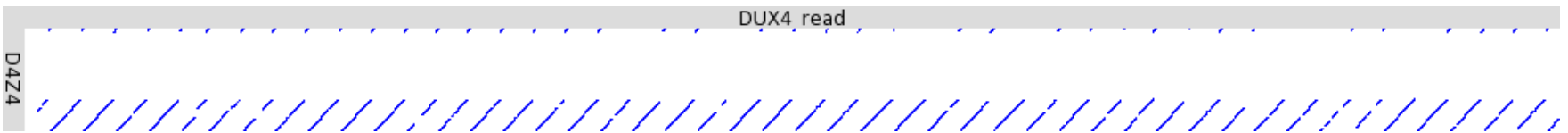


Alignment Tools

LAST	An extension of BLAST-like tools. Comes with tools to train alignment parameters, visualize alignments, and perform genotyping. Good at dealing with repeated regions.	http://last.cbrc.jp/
Minimap2	Modification of BWA-Mem for Nanopore. Very fast, but may have difficulties with repeated regions.	https://github.com/lh3/minimap2
NGMLR	Has interesting design choices (gap-penalty) aimed at dealing with high errors rates. Also has a tool called Sniffles for structural variant detection	https://github.com/philres/ngmlr
GraphMap	Offers transcriptome mapping to work with RNA sequencing, as well as DNA sequences.	https://github.com/isovic/graphmap

Alignment

- Used a host of aligners (NGMLR, Minimap2, LAST) to align sequence
- Can find some repeats of D4Z4, but we are having issues with finding reads that completely cover region
- Read pictured below is an example of one with a repeated region with 45 repeats, however read does not span region



Alignment Results

	Region / Tool	LAST	NGMLR	Minimap2
Rel4	5’ sequence (5.7 kb) (Japan)	12 reads	4 reads	13 reads
	D4Z4 repeat unit (3.3 kb) (NCBI)	385 reads	180 reads	356 reads
BluePippin	5’ sequence	0 reads	0 reads	0 reads
	D4Z4	36 reads	25 reads	31 reads
RepliG (enriched)	5’ sequence	0 reads	0 reads	0 reads
	D4Z4	19 reads	7 reads	19 reads

Alignment on Chromosomes

Minimap 2		GRCh38			GRCh37			
		Chr4	Chr10	Both	Chr4	Chr10	Both	GL000194.1
Rel4:	5prime	9	8	8	8	9	8	0
	D4Z4	100	107	51	72	64	27	17
BluePippin:	5prime	0	0	0	0	0	0	0
	D4Z4	5	9	4	0	2	0	0
RepliG:	5prime	0	0	0	0	0	0	0
	D4Z4	2	3	2	2	0	0	0

GL000194.1 is included as an alternate assembly of chromosome 4

Assembly Tools

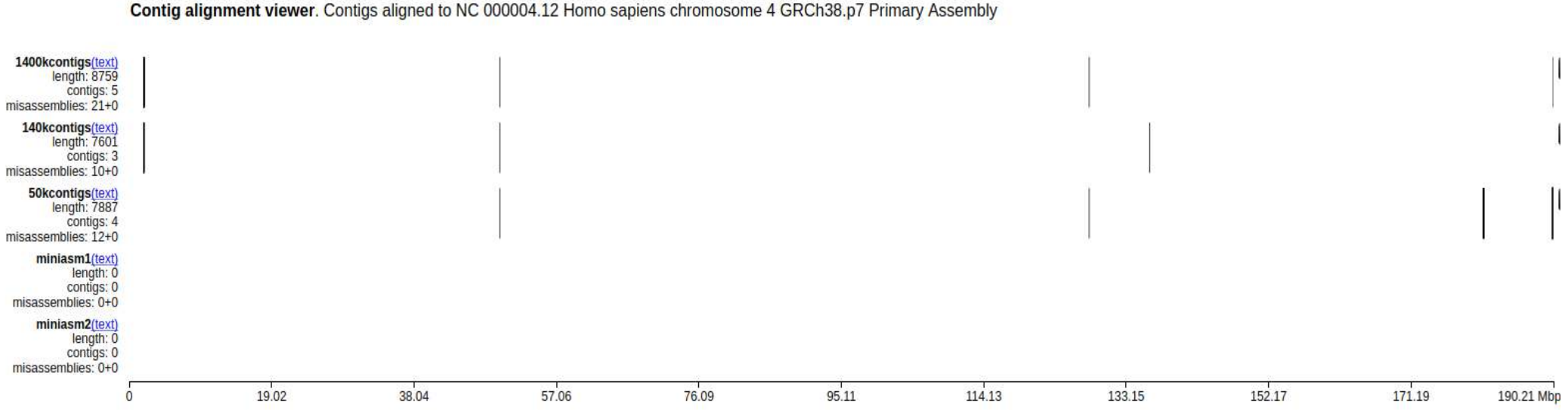
Canu	Assembler for 3rd generation reads. Can perform hybrid assembly. Performs trimming, error correction and consensus assembly. Can also be run distributed.	http://canu.readthedocs.io/en/latest/quick-start.html
MaSuRCA	Illumina assembler which can also perform hybrid assembly with long reads. Similar pipeline to Canu, but no distributed capabilities.	http://www.genome.umd.edu/masurca.html
Falcon	A relatively light assembler by Pacific Biosciences with a simple front end intended for smaller genomes.	https://github.com/PacificBiosciences/FALCON
Miniasm	Fast assembler, works in tandem with minimap assembler, no error correction	https://github.com/lh3/miniasm
FLYE	Assembler for Nanopore and PacBio reads, can use both corrected and uncorrected reads. It's predecessor ABruijn, could run distributed.	https://github.com/fenderglass/Flye

De Novo Assembly

- Of the assemblers mentioned previously, chose Canu which we ran on GWU's cluster
  - Both can use solely nanopore reads, as opposed to hybrid assembly
- Results indicated we did not have high enough coverage to assemble region

Assembly Results

- We found at most 26 contigs with Canu, with the longest being 22kb in length.
- The contig that aligns to this repeated region is too short, only 8kb in length >50% of it remaining unaligned to chr4

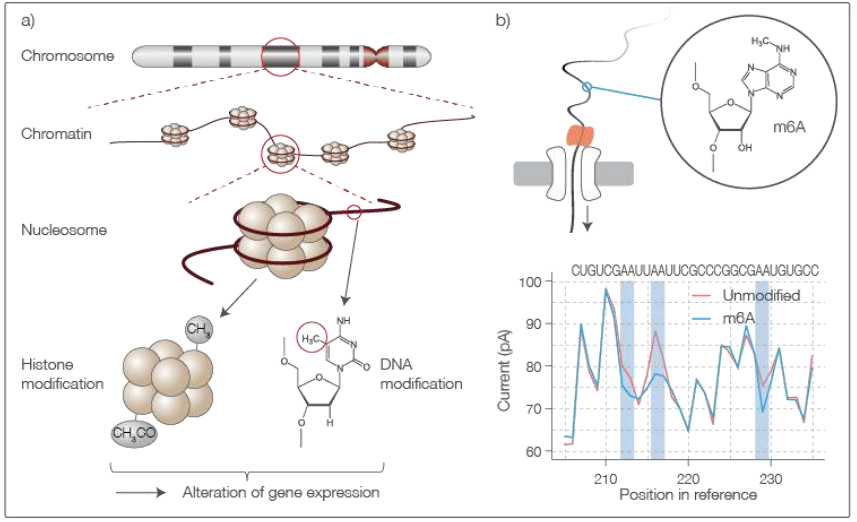


Future Plans

- First and foremost, we are looking to obtain more data sets with longer reads
- Explore methylation of our reads (as only D4Z4 from chr4 should be methylated)
- Return to assembly of region when we are sure we have reads that span repeats

Tools for Methylation

- Tombo – Released by Nanopore, used to analyze methylation on reads.[5]
- Nanopolish – used for polishing methylated assemblies by using raw signal data for reads.[6]



Oxford Nanopore Technology

References

1. https://blog.genohub.com/2017/06/16/pacbio-vs-oxford-nanopore-sequencing/
2. TechBlog: The nanopore toolbox https://github.com/rrwick/Basecalling-comparison/blob/master/README.md
3. Jain, M., Olsen, H. E., Paten, B. & Akeson, M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.* **17,** 239 (2016).
4. Jain, Miten, et al. "Nanopore sequencing and assembly of a human genome with ultra-long reads." *Nature biotechnology* 36.4 (2018): 338.
5. Jain, Miten, et al. "Linear assembly of a human centromere on the Y chromosome." *Nature biotechnology* 36.4 (2018): 321.

Thank you!