Genomic Medicine Short Course Kansas City, MO September 2019



# Answering questions with nanopore sequencing: From bacteria to sequoias

Winston Timp Department of Biomedical Engineering Johns Hopkins University

#### **Revolutions in Science: Genomics**







- Draft of the human genome was completed in 2001
- ~3 billion bases in size
- Think about this like the first transistor (1947) the watershed after which genomic and epigenomic engineering has exploded



#### **Revolutions in Genomics: Single Molecule Sequencing**



- First patented back in 1995, commercialized in 2014
- No theoretical upper limit to sequencing read length, practical limit only in delivering DNA to the pore intact
- Palm sized sequencer
- Sequencing output 5-20Gb





Disclosure: Timp has two patents (US Patent 8,748,091; US Patent 8,394,584) licensed to ONT

# **Nanopore Library Prep**



- Library prep is very similar to methods for short-read sequencing
- For DNA shearing we used Covaris gTubes or Diagenode Megaruptor .
- After end-repair and A-tailing, leader adapter with motor protein is ligated .
- MinION arrays 512 channels (with 4 pores possible per channel) (shown bottom left from running software); dark green pores are sequencing, light green available, other colors inactive.

#### **Sequencing Operation**





Oxford Nanopore Technologies

- Protein nanopores on a synthetic polymer
- Multiple base-pairs at a time ("k-mers")
- Characteristic current signature is converted to nucleotide sequences

#### Nanopore Sequencing Workflow





• Nanopolish : uses alignment and current signal to improve base-calls

Alignment

#### **Problems with Nanopore basecalling**



- Multiple bases influence the current passing through the pore.
- Through simulation with Brownian Dynamics, we calculated the contribution from triplets of DNA in a solid-state nanopore - 64 current levels.
- Not all of these different currents are distinguishable



Comer and Aksimentiev J. of Phys Chem C 116(5) 3376-3393 (2012)

#### **Hidden Markov Model**



- Markov chains represent a series of states occurring in sequence, with defined transition probabilities. In a hidden Markov model the state is only observed indirectly.
- As an example, consider inferring the weather (without going outside) by observing whether people coming in from outside are wearing gloves or not

8

#### **Prior Information for Decoding**





- With no prior information, a given current value may not be called correctly (333pA would be called as GGG)
- If we know the previous triplet, the next triplet is well defined, leaving only four possibilities, resulting in the correct call of TCG



# **Nanopore HMM basecalling**

- By using a sequence of observables and maximizing the total joint probability given below, we find the sequence of states.
- This is done using the Viterbi algorithm which grows, finding the most likely path for each step, saving the probabilities, to avoid recalculation.
- 1<sup>st</sup> generation basecallers from Oxford used a HMM for basecalling similar to the one detailed in our Biophysical paper
- Transition probability matrix for oxford seems to allow for a 0, 1 (most common), 2, or 5 (reset) move.
- We think that Oxford trained its basecalling model on unmethylated lambda



10



# **Basecalling shifting to RNN**

- Recently (over the past year) there has been a shift to neural network based basecalling
- A *recurrent* neural network is still one with memory, that has a dependence on past computations
- Specifically two layers of Bidirectional Long Short Term Memory (BLSTM)
- These still require the same "training" data to learn what current distributions correspond to which k-mers – and the results are still k-mer based, as multiple bases still influence the current.



**Basecalling - RNN** 

Distributions learned from squiggle training data

> Bidirectional information flow (BLSTM layer)

Processing layer

Bidirectional information flow (BLSTM layer)

Multi-base prediction

Decode to sequence



#### **Building Genomes: Genomic Technology Development**





- We wanted to try this out to see how well we could build genomes now
- As part of a team with Steven Salzberg (JHU) and David Neale (UCDavis) we are trying to sequence the giant redwoods
- This is a hard problem! The genomes are big (3X or 10X human) and have lots of stretches that are hard to distinguish





#### Sequoia sempervirens Coast redwood California – central/north coast



**Sequoiadendron** giganteum Giant sequoia California – Sierra Nevada



Two species in our redwood genome project

California endemics

Economic, cultural, and conservation value

"Advanced management strategies"

Huge genomes: 30Gb Coast, 9Gb Giant



#### **Building Genomes: Genomic Technology Development**



Building a genome is like putting together a puzzle:

- Larger (blurry) pieces = easier puzzle
- Small pieces = hard to put together, can't figure out "blue sky"
- Small + large pieces = cleary and easier to put together





#### **Problem**

- Want: High molecular weight (HMW) DNA
  - 100kb + average
- Need: High yield
  - At least 10ug purified gDNA from 1 g of leaf tissue
- Need: High quality
  - Based on nanodrop spectra and gel migration
    - 260/280 ~1.8
    - 260/230 ~2.0+
    - No aberrant migration
- Reproducible/Robust protocol
  - Want no wizards
- High and consistent sequencing yield (- at least 3-5Gb per run)







# **Sample realities before optimization**





#### **Trials: DNA extraction**

- Detergent (SDS, SLS, CTAB, Triton)
- Extract nuclei first (yes or no)
- Phenol chloroform (yes or no)
- Salt used for alcohol precipitation (NaCl, sodium acetate, ammonium acetate, none)
- Modifications for improving fragment lengths (agarose embedding, nanobind)
- Synchronous coefficient of drag alteration (SCODA) for purification





#### THE REASON I AM SO INEFFICIENT





**Top 3 extraction protocols: Zhang** 

- Tissue ground in LN2
- Nuclei: Cell wall lysis -> filtration -> differential centrifugation
- DNA: Overnight SLS lysis, phenol chloroform, alcohol precipitation + sodium acetate





### **Top 3 extraction protocols: Healey**



- Preferred plant protocol: Skips the differential nuclei extraction
- CTAB (cetyl trimethylammonium bromide) cationic detergent
- CTAB extraction, chloroform, alcohol precipitation + NaCl, elute





# **Top 3 extraction protocols: Nanobind**

#### ocirculomics



#### **Unique Tentacle Binding Mechanism**

• Enhances binding capacity and protects DNA from shear forces



Low Input (10  $\mu$ g)





Medium Input (50 µg)

High Input (200 µg)

- Three material properties needed: low shear, non-porous, high surface area
- DNA tentacles form and extend from substrate to get high binding capacity
- Low shear unlike beads and columns

#### ocirculomics

# Nanobind: How does it work





23

#### **Extraction Comparison**



	Zhang	Nanobind	Healey
Reads	201k	500k	195k
Yield	0.51Gb	4.72Gb	1.08Gb
N50	7.1kb	12.3kb	8.6kb
Median	929	9.8kb	5.1kb

- Zhang seemed (in our hands) to fragment badly)
- Nanobind and Healey seemed to give reasonable read lengths, but doesn't match with PFGE profile size

24





# Improved extraction and sequencing methods affect assembly contiguity

Assembly with MaSuRCA 3.2.4+ scaffolding with HiRise(Dovetail genomics)

	Sequence in	N50	N50 scaffold	Number of	Number of
	contigs	contig	(Kbp)	contigs	scaffolds
	(Gbp)	(Kbp)			
Illumina only	7.9	12	65	2,507,175	1,007,217
Illumina+Nanopore	8.1	360	489	49,676	39,821
Illumina+Nanopore+	8.1	318	689,571	52,835	8,215
Chicago+Hi-C					

- Improvements in contig and scaffold size over other conifer assemblies afforded by long reads
- MaSuRCA assembler (Zimin et al, Bioinf 2013; <u>http://www.genome.umd.edu/masurca.html</u>)



#### Size Selection – Going even longer



27

#### **Read Length**

- Working with Circulomics we have been trying to get the read length up
- Using a size selection with their Nanobind material, read N50 can be substantially improved
- There is still room for improvement often still difficult to get both high yield and high read length





#### Methodology extensible to other plants





Maize results courtesy of B. Vaillancourt and Krystle Wiegert-Rininger of the C. Robin Buell Lab at Michigan State University

#### **Building Genomes: Genomic Technology Development**



hummingbirds to giant sequoia





- Modern Definition of epigenetics involves heritable changes other than genetic sequence, e.g., positive feedback, high order structure, chromatin organization, histone modifications, DNA methylation.
- An analogy to a computer system:
  - DNA Sequence = Hardware
  - User input = Environment
  - Systems Biology = Running programs
  - Epigenetics = RAM

#### **Nanopore Sequencing of Modifications**





#### Nanopore: nanopolish methyltrain



• Where  $S_m$  is the probability methylated for a given observable D and  $S_r$  the probability unmethylated)

• We then take the log of this likelihood ratio, and threshold for >2.5 as methylated; <2.5 as unmethylated

#### **Nanopolish Methylation**



N = 658621 r = 0.895

34

#### **Cas9 enrichment Method**





#### Using a panel of guideRNAs

- Yield from
- 3ug GM12878 gDNA
- MinION Flow cell





#### Using a panel of guideRNAs

- Yield from
- 3ug GM12878 gDNA
- Flongle Flow cell



# **Enrichment of hTERT region**

- We observe an "erosion" of the unmethylated (blue) CpG island in the promoter of hTERT in progressive cancer samples
- In the late metastasis a mutation in the ETS binding site of the promoter occurs in one of the alleles
- The mutant allele appears to have a more unmethylated island than the WT allele





#### **Structural Variants in Cancer**

- Structural variants (SV), large insertions, deletions or translocations in the genome, are hard to detect with short-read sequencing
- Nanopore sequencing can map them well, and with targeted sequencing we can observe these



Breast Cancer: 8kb deletion, chr7

#### **Structural Variation Detection**



 Bias likely due to length of reads input into enrichment



#### **Single Nucleotide Variants**

176 known SNVs exist in in span of 140kb in GM12878



MinION		ТР	Sensitivity	FP	PPV
default	SAMTOOLS	170	0.97	12	0.93
variant calls	NANOPOLISH	169	0.96	17	0.91
dual-strand	SAMTOOLS	142	0.81	3	0.98
filter	NANOPOLISH	156	0.89	1	0.99







#### NanoNOMe: Chromatin Accessibility with Nanopore

• NOMe-seq : Nucleosome Ocupancy and Methylome sequencing (Kelly et. al. Genome Res. 2012) Simultaneously measures DNA methylation (CpG) and nucleosome occupancy (GpC)



#### **Nanonome Signal**



#### NanoNOMe – DNAse Hypersensitive



nanoNOMe signal near DNAse-seq peaks validates the methodology



#### NanoNOMe: Aggregate CTCF binding sites

GpC Methylation

#### Chromatin Protection (1-GpC)



#### Endogenous Methylation (CpG)





#### **Methylation in Repetitive Regions**





Regions unmappable by NGS are mappable with long reads

#### **Repeats: BRCA1**

7947 bp





Reference genome doesn't have many of these repeats properly – for BRCA1 region we aligned our reads against a custom GM12878 genome assembly (Jain et al)

#### **Allele Specific Chromatin and Methylation**

CpG Methylation

p21.3 p15.3 p14.2 p12.3 p11.1

94,656,000 bp

PEG10

,655,000 bp

SGCE

q11.23

4.827

94,657,000 bp

GpC Accessibility



- Using long reads, we are likely to encounter a SNP
- This allows for phased methylation and chromatin data
- Near PEG10 (imprinted gene):
  - Maternal copy is methylated and inaccessible
  - Paternal copy is unmethylated and accessible

# **Coordinated Enhancers and Promoters** 10 kb

Using long reads, we can examine methylation and chromatin at some promoters and enhancers at the same time





#### **Antimicrobial Resistance**



Mutations of otherwise benign genes can confer resistance.

(gyrA)

Aldred et al Biochemistry (Feb

#### **Project Overview**

Correlate resistance phenotypes with genomic features  $\rightarrow$ Patients Nanopore (+ Illumina) Assemble genome, **DNA Sequencing** find resistance genes and mutations Rectal/ Perirectal CIM/mCIM Antibiotic Swab Culture  $\rightarrow$  Susceptibility MALDI-TOF to Profiles **ID** organisms **Disc diffusion** Etest Standard of Care Tests



#### The Pipeline so far





#### **Assemblies**



#### **KLPN: Phylogeny in the context of NCBI genomes**

- A comparison of our isolate assemblies to NCBI reference genomes for K. pneumo gave clear clustering with specific strains.
- Core whole-genome alignment with parsnp against complete *K. pneumo*



54



#### **Alignment Error Comparison**

- We used bowtie2 for Illumina and minimap2 for Pacbio and ONT alignment, then used samtools to compare to reference. For this comparison we used Illumina & Pacbio provided data compared to data from Nick Loman on *E. Coli*.
- Illumina reads which align do so nearly perfectly, with a per read median of 99.3% correct.
- PacBio reads which have an median of 89.2% of the read correct. The most frequent error type is insertions (7.45% median) with mismatches only 1.5% median % of read.
- ONT reads from v9.4 (Nick Loman's data) have a per read median of 92.4% correct, with deletions (9%) and mismatches (4.5%) both at a relatively high median per read.





#### Nanopolish







#### **Nanopore: Methylated Error**

•We sequenced samples from NEB ER2796 (E. Coli with KO of dam/dcm)

- Different methyltransferases are transformed in.
- •Notably, mismatch error rate and deletions seem higher on methylated samples than unmethylated.
- •The lower shift in 4mC and 6mA may be do to relative infrequency of those motifs.







#### **Methylation Associated Error**

We can address this problem by training models specifically for methylation motifs, using a similar HMM scheme to align electrical data to a reference.









models need to be trained on methylated k-mers in order to correct MTase motifs (dcm MTase= **CCWGG**)

Raw	Nanopolish	Methylation Aware
Assembly	Corrected	Nanopolish Corrected
98.89%	99.57%	99.76%



#### **Differences report** Mutation or sequencing artifact?





#### **Building Diagnoses: Real-time Detection**



As reads can be identified as they come off the sequencer, we can identify AMR rapidly Our retrospective analysis showed the resistance was identified for all of our isolates within 15 minutes



#### Summary

- Nanopore technology is full of potential for sequencing, but always choose the right tool for the right job.
  Often multiple approaches with complementary data yield the best results.
- Multiple bases affect the electrical signal from nanopores; rather than a problem, this can be an advantage, as each base is interrogated multiple times.
- Modifications to the primary DNA sequence (e.g. cytosine methylation) can be detected directly using nanopores
- Targeted sequencing with Cas9 allows for long reads in targeted regions, sidestepping issues of cost.
- Exogenous labeling allows simultaneous detection of chromatin and methylation state using nanopore sequencing



#### **Acknowledgments**





National Human Genome Research Institute 1R01HG010538 1R01HG009190

- Jared Simpson
- Paul Tang
- P.C. Zuzarte
- Michael Molnar
- **Trish Simner**
- **Chris Umbrich**
- Fritz Sedlazeck



- Jawara Allen
- **Brittany Avin** 
  - Sheridan Cavalier •
- Yunfan Fan

- Ariel Gershman
- Timothy Gilpatrick
- Isac Lee
  - Brittany Pielstick
- Roham Razaghi
- Norah Sadowski
- **Rachael Workman**

Baylor College of Medicine