The Long Road to Long Reads: Challenges to Implementing a Long Read Sequencing Service

M Kokosinski[1], B Marosy[1], J Gearhart[1], J Paschall[1], B Craig[1], M Mawhinney[1], D Mohr[1], M Sheridan[1], P Witmer[1], D Muzny[2], A Scott[1], J Hosea[3], C Montano[3], L Morina[3], Q Li[3], A Klein[4], M Schatz[3], W Timp[3], the *All of Us Research Program,* and K Doheny[1]
[1]Johns Hopkins University, Department of Genetic Medicine, Baltimore, MD, USA; [2]Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX, USA;
[3]Johns Hopkins University, Department of Biomedical Engineering, Baltimore, MD, USA; [4]Johns Hopkins University, Department of Oncology, Baltimore, MD, USA

Introduction

Long read sequencing has been a steadily developing technology with several advantages over short read sequencing including resolving large and complex structural variants, tandem repeats, phasing of variants and native detection of base modifications. Implementation of Oxford Nanopore Technologies long read sequencing in a high throughput core facility has required standardization and optimization of protocols to achieve 30x coverage whole genomes with an N50 >25kb in a reproducible manner. Typically, long read sequencing protocols require new extractions to facilitate high molecular weight DNA and increased DNA inputs to ensure higher N50s. Studies which use previously extracted DNA require optimization of protocols in order to achieve similar N50s. Here we provide insight from implementing this technology utilizing biobanked DNA from the NIH All of Us Research Program (AoU). The AoU protocol was optimized for longer N50s, utilizing 2-6 ug of input DNA, g-Tube shearing and size selection on the Blue Pippin prior to LSK114 library prep and sequencing on a single R10.4 flowcell with 3 loads over 72 hours. To date, we have processed over 691 samples achieving an average of 36.3x coverage and N50 of 25.2kb.

Challenges to production-scale work have included the time required for super-high-accuracy live basecalling as well as inconsistencies in total yield of R10.4 flowcells. A second high priority study is examining familial pancreatic cancer samples with an expected total of at least 400 long read WGS. A pilot using 66 cell line derived DNA samples (fresh extractions optimized for long reads) achieved an average of 64.8x coverage and N50s of 25.8kb. In order to achieve this higher depth, we adapted our methods by increasing DNA input to LSK114 library prep to at least 6 ug, creating 2 libraries per sample, and sequencing 2 R10.4 flowcells per sample. Currently, we are actively exploring adaptive sampling, to address clinical application for repeat expansion testing, methylation profiling, and elucidation of "unidentified” 2nd variants in individuals with cystic fibrosis, to inform treatment options.

Methods Improvement to Library Prep – AoU

The AoU protocol, developed at Baylor HGSC and optimized for longer N50s at CIDR, utilizes 2-3ug of input DNA, g-Tube shearing and size selection on the Blue Pippin prior to LSK114 library prep and sequencing on a single R10.4 flow cell with 3 loads over 72hours.

Library prep improvements were first designed to maximize yield and longer N50s while using a limited amount of DNA (2-3ug input). Initial experiments conducted using pooled samples from the AoU biobank saw a loss of 27% of DNA post shear clean up and additional loss of 59% of DNA post Blue Pippin size selection and clean up. Eliminating the shear clean up allowed for less loss of DNA throughout the process. Once library yield was stabilized, size selection parameters were adjusted in order to achieve the largest fragments while still maintaining library yield. Size selection start parameters were tested at 20, 22, and 25kb. A start size of 25kb allowed for the highest N50s while still maintaining adequate library yield. Processing of 300 samples was completed using the above methods. Library prep redo rate using this method was 12.3% where redo is defined by producing <16 fmol yield.

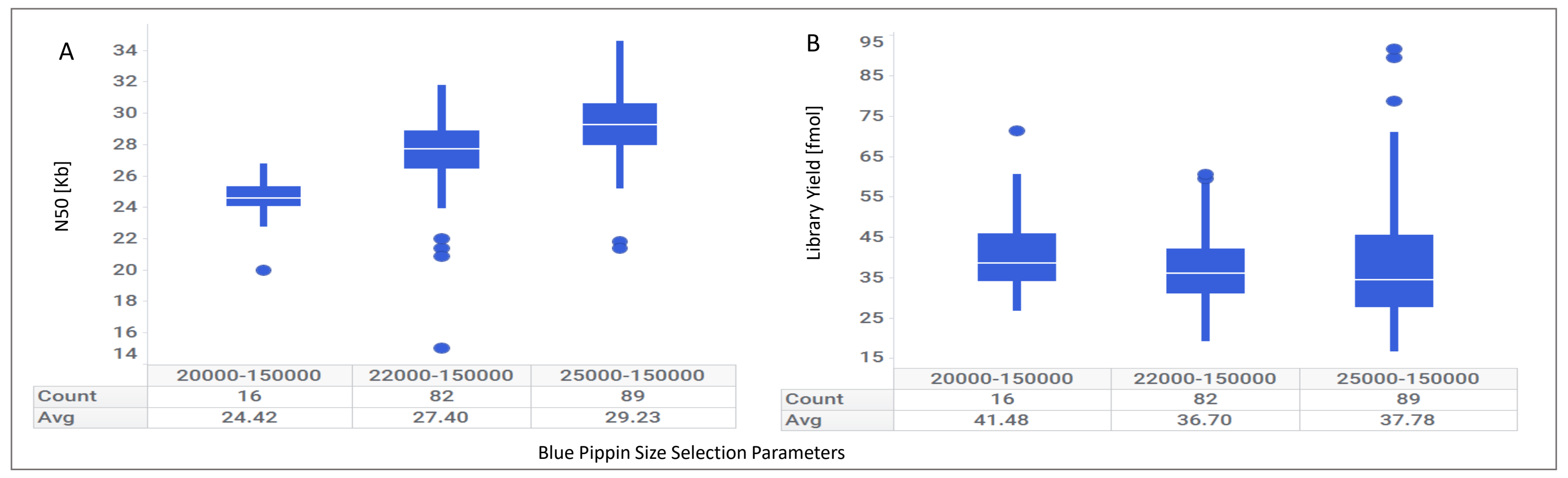

Figure 1A/B. (A) N50 [Kb] and (B) Library Yield fmol, when using 3ug of DNA input and omitting the bead clean up post shear, across variable size selection (20, 22 & 25kb) parameters.

A second batch of samples provided from the AoU biobank arrived with greater quantity of DNA available for library prep. With ample DNA provided, we began processing samples with 6ug input DNA in order to lower the rate of library redos and increase sequencing loading concentration to 25fmol/load. Due to input volume constraints on the Blue Pippin, shear clean up was reinstated to reduce volume from shear to size selection. Although we saw similar loss at this step, the increase in overall input was enough to overcome this loss. Batch one samples saw an average library prep yield of 37.4 fmol and batch 2 had an average of 115.1fmol. Using the same redo criteria of <16 fmol yield, the redo rate dropped to 0.2%.

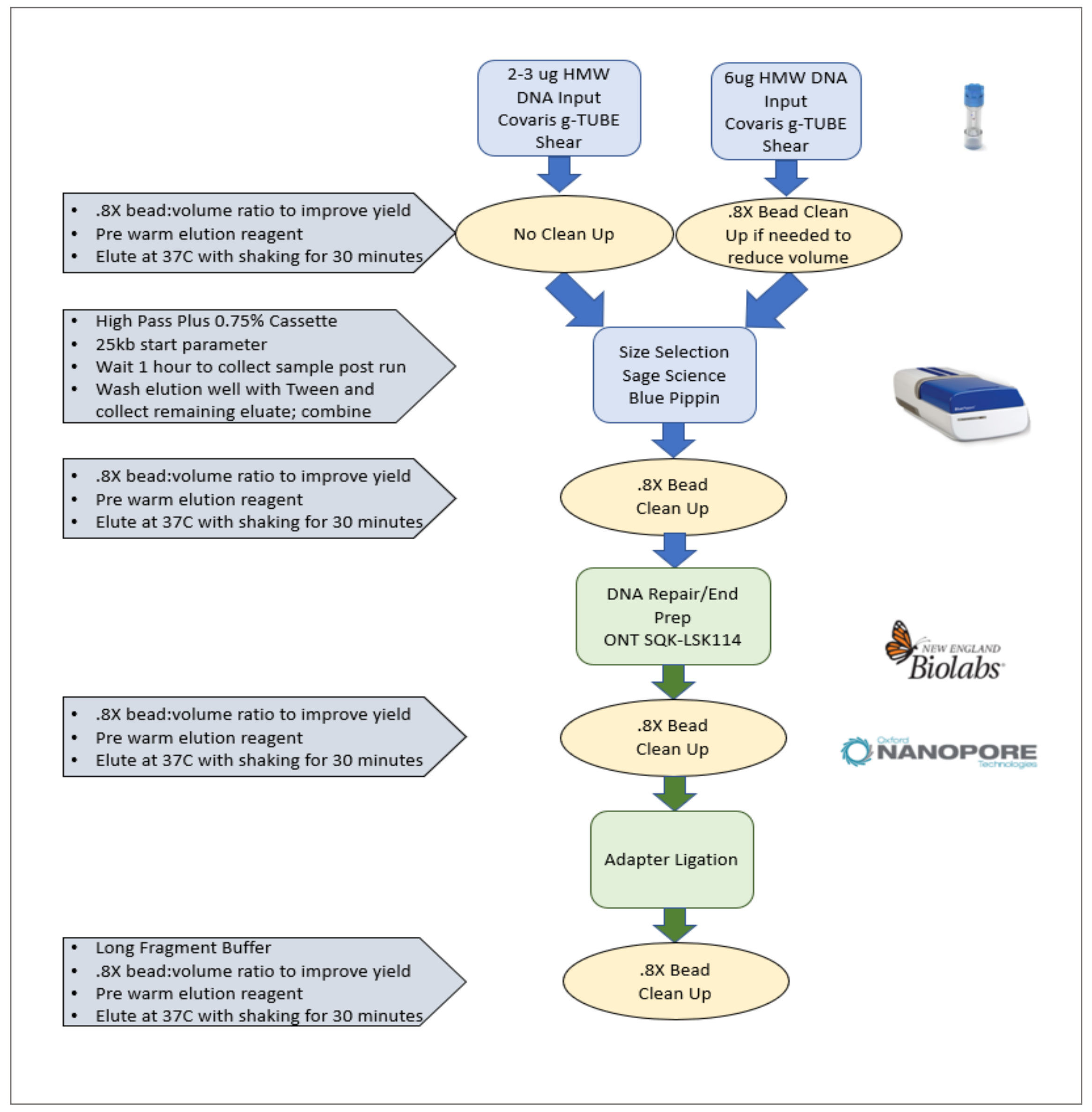
2-3 ug HMW DNA Input Covaris g-TUBE Shear
6ug HMW DNA Input Covaris g-TUBE Shear
No Clean Up
.8X Bead Clean Up if needed to reduce volume
Size Selection Sage Science Blue Pippin
.8X Bead Clean Up
DNA Repair/End Prep ONT SQK-LSK114
.8X Bead Clean Up
Adapter Ligation
.8X Bead Clean Up
.8X bead:volume ratio to improve yield
Pre warm elution reagent
Elute at 37C with shaking for 30 minutes
High Pass Plus 0.75% Cassette
25kb start parameter
Wait 1 hour to collect sample post run
Wash elution well with Tween and collect remaining eluate; combine
.8X bead:volume ratio to improve yield
Pre warm elution reagent
Elute at 37C with shaking for 30 minutes
.8X bead:volume ratio to improve yield
Pre warm elution reagent
Elute at 37C with shaking for 30 minutes
Long Fragment Buffer
.8X bead:volume ratio to improve yield
Pre warm elution reagent
Elute at 37C with shaking for 30 minutes

Figure 2. AoU protocol workflow depicting each major step along with enhancements to increase yield of high molecular weight fragments. Arrows in grey depict additional optimization at CIDR to increase yield and produce longer N50s.

Methods Improvements to Sequencing – Project AoU

Sequencing for all samples occurred on PromethION R10.4 flowcells. Flowcells were run for 72 hours with a wash/reload at 24 and 48 hours. When enough library was available, flowcells were loaded with 25 fmol for all three loads. As expected, flowcells with enough library yield to load 3 fresh loads outperformed those requiring 1 sample recovery and significantly outperformed those requiring 2 sample recoveries. The higher input in batch 2 allowed for both higher loading concentration and fewer samples requiring recovery. In batch one, 29.3% of samples required additional sequencing on a second flow cell to achieve >30x coverage while batch 2 dropped to 11.5% of samples requiring additional sequencing.

Figure 3

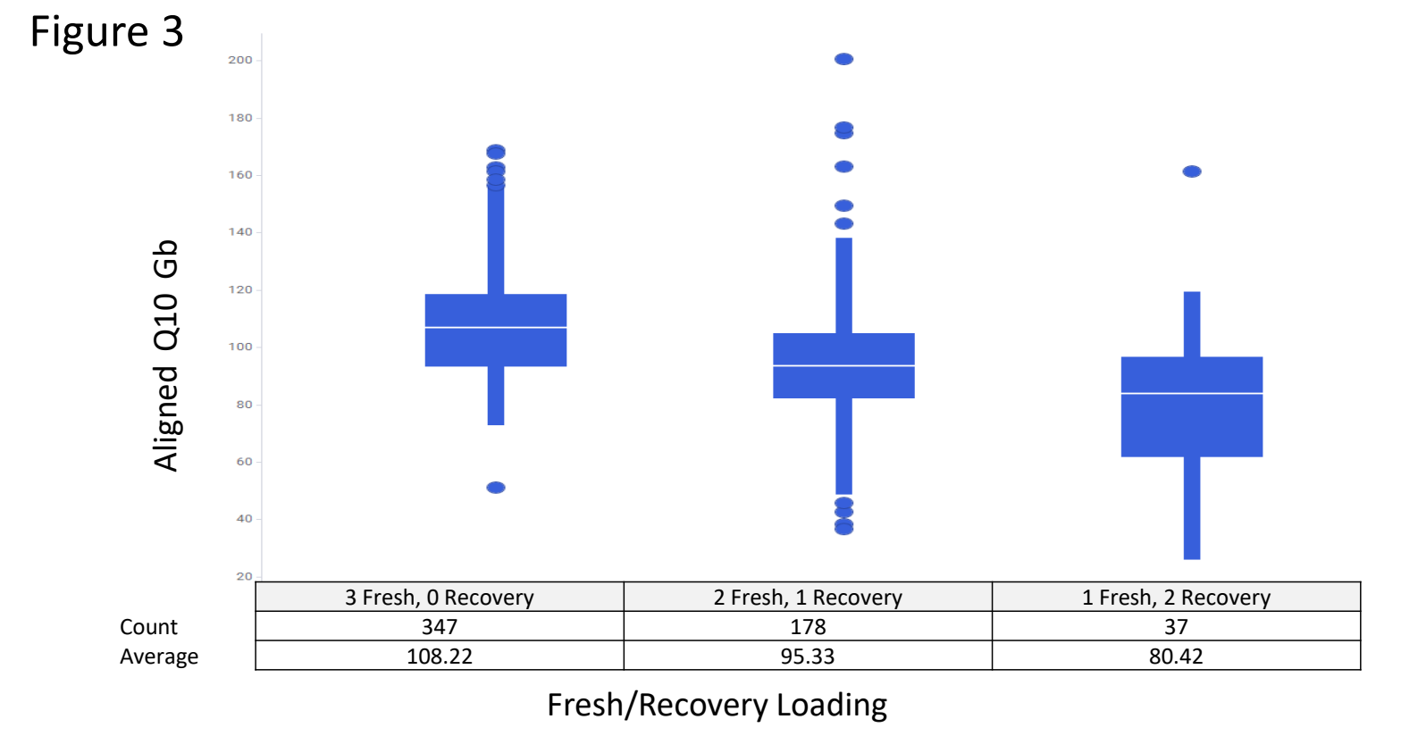

Table 1

	Batch 1	Batch 2
Number of Samples	300	391
Average loading conc (fmol)	15.4	24.3
Load 2 samples requiring recovery	24	5
Load 3 samples requiring recovery	149	22
Total aligned Q10 GB average	94.9	112.7
% Samples requiring additional sequencing	29.3	11.5

Figure 3. Sequencing yield (Aligned Q10 Gb) from flowcells with Fresh/Recovery loading conditions. Table 1. Comparison of Batch 1 and Batch 2 AoU.

During the processing of batch 1, ONT introduced light shields to improve sequencing yield. A comparison of flowcells run with and without the light shields shows improvement to the overall yield. Table 2 compares samples with similar library yield (total fmol) and loading concentration. 24% of batch 1 samples processed using light shields required additional sequencing, an improvement from the overall rate of 29.3% additional sequencing in batch 1.

Figure 4

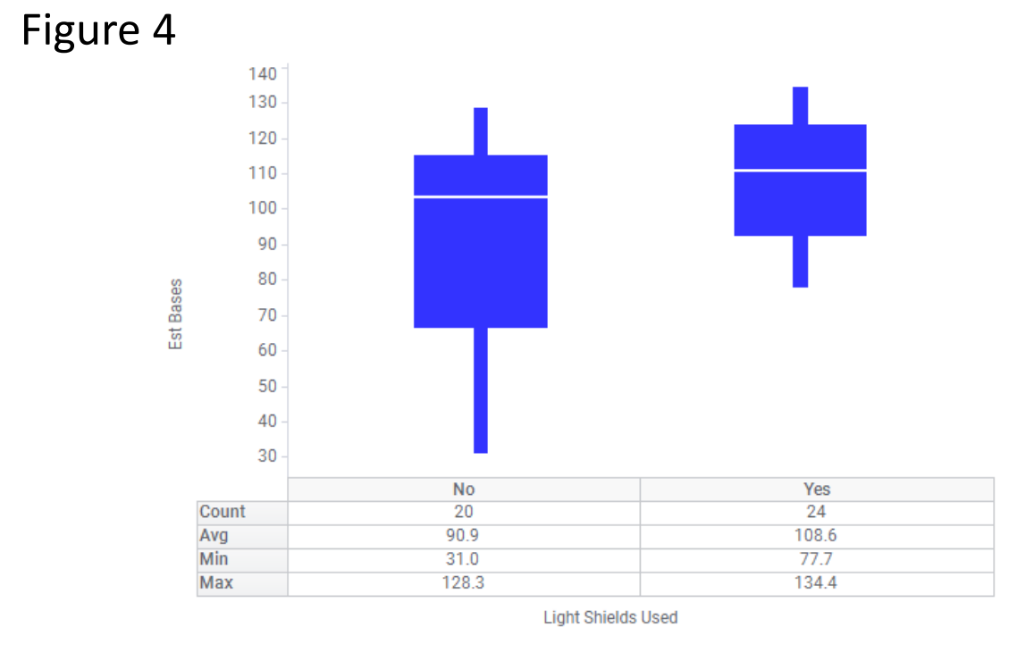

Table 2

	No Light Shields	Light Shields
Number of Samples	20	24
Average Library Yield (fmol)	37.1	41.1
Average fmol/load	13.6	15.2
Average Flow cell pore count	7884	7113
Est Bases (Gb)	90.9	108.6

Figure 4/Table 2. Comparison of flowcell yield with and without light shields

Pancreatic Cancer Study

A second study aims to examine 400 samples using WGS long reads to study familial pancreatic cancer. To date, we have completed 66 samples for that study. This study aims for WGS coverage of 60x. In order to achieve the higher depth, we used 6ug input and created 2 libraries via our adapted LSK114 protocol. The libraries are then sequenced on 2 R10.4 flowcells.

As in many HMW samples, DNA was particularly viscous creating a challenge for shearing the DNA. Many of the samples clogged G-tubes and needed to be diluted significantly to perform shearing. Samples that had shear issues often fell below 16 fmol library yield, causing our overall library prep redo rate for this project to be 25.8%. The use of 2 flow cells meant that some underperforming flow cells had a library pair with a higher performing flowcell. Being able to combine the data from 2 flow cells meant a reduction of samples requiring additional sequencing. This led to 9% of samples requiring additional sequencing. Overall coverage for this project is 64.8x with an average N50 of 25.8kb.

Throughput Challenges

Inconsistent flow cell yield and computational intensity of using super-accurate basecalling (SUP) have caused the largest challenges to production workflow. At the time we began processing the AoU project, a P24 PromethION with an A100 tower could only run 8 flow cells using SUP and methylation calling via MinKnow/Guppy. Additionally, at the time, basecalling for those 8 flow cells typically took 4 days after the 72 hour run. ONT released Dorado software for basecalling which saw improvements of basecalling speed of ~25%. Using Dorado and MinKnow version 23.11.7 we are now able to sequence up to 15 flow cells using SUP + methylation.

One of the P24s had an additional A100 tower installed, splitting the deck between the two processors (12 deck positions for each tower). This configuration allows all deck positions to be used as each A100 tower receives data from 12 positions. Currently, we process samples for 36 flow cells per batch. Since it is still fairly common to have poor performing flow cells, this allows for movement onto a new flow cell if needed without overloading the sequencers.

Inconsistent performance of flow cells is a major driver of samples requiring additional sequencing. Figure 5/Table 3 shows the difference in performance among two flow cells loaded with the same library. The low performing flow cell starts with a relatively similar % of pores sequencing, but as sequencing continues, the low performing flow cell pore performance drops off significantly more.

Figure 5

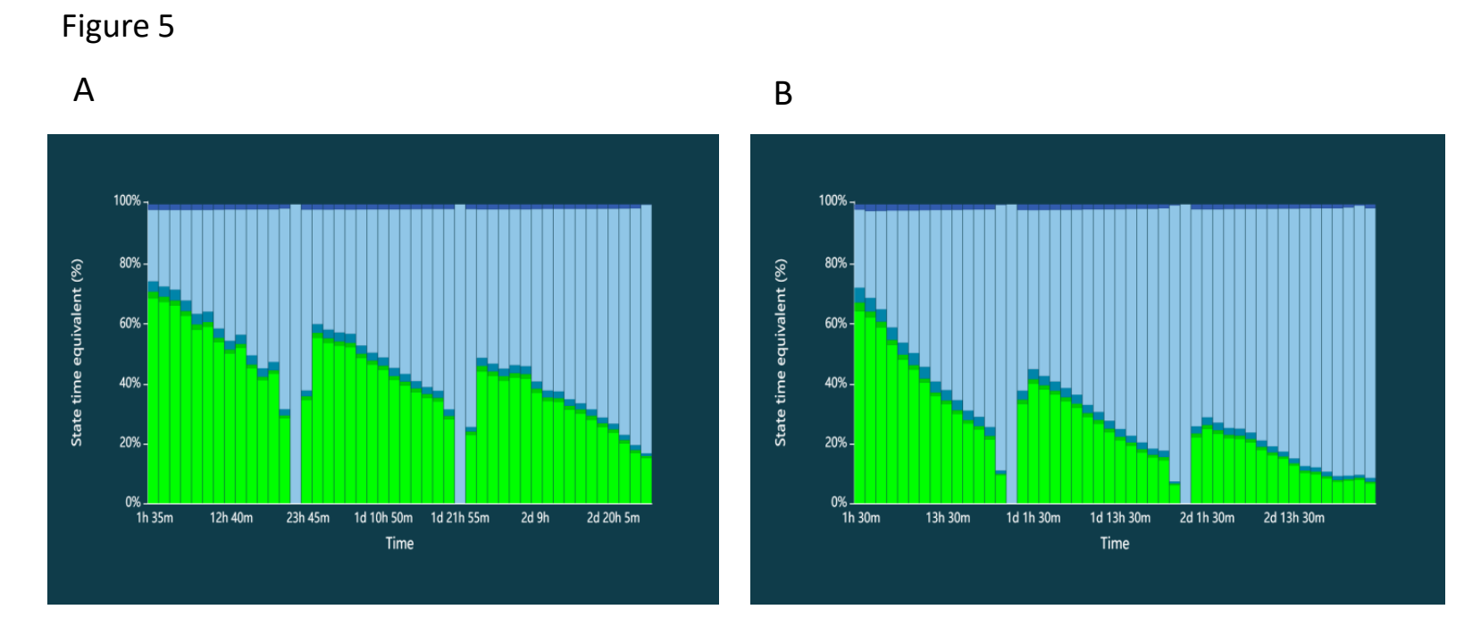

Table 3

	High Performance	Low Performance
Total yield (GB)	127.2	77.2
% Pores Sequencing - Load 1 start	68.6	64.3
% Pores Sequencing - Load 1 end	43.3	21.4
% Pores Sequencing - Load 2 start	55.3	39.9
% Pores Sequencing - Load 2 end	34	15.3
% Pores Sequencing - Load 3 start	44.1	24.9
% Pores Sequencing - Load 3 end	15.1	6.7

Figure 5. Pore activity plots for a high performing (A) and low performing (B) flow cell. Table 3. Metrics for flowcells shown in Figure 4.

Future applications – Adaptive Sampling

We are actively exploring adaptive sampling to address clinical applications for repeat expansion testing and elucidation of "unidentified” 2nd variants in individuals with cystic fibrosis. A specimen with known *CFTR* variants underwent library prep using the described protocol here modified without size selection. Previous sequencing with Illumina short reads could not fully delineate the 1.6kb insertion. Adaptive sampling of the *CFTR* region in conjunction with an all known repeat expansion panel based on the Miyatake et al paper (PMID: 36289212) resulted in 80x coverage of the targeted region and an overall genome coverage of 23x. Long read sequencing was able to fully characterize the insertion (Figure 6).

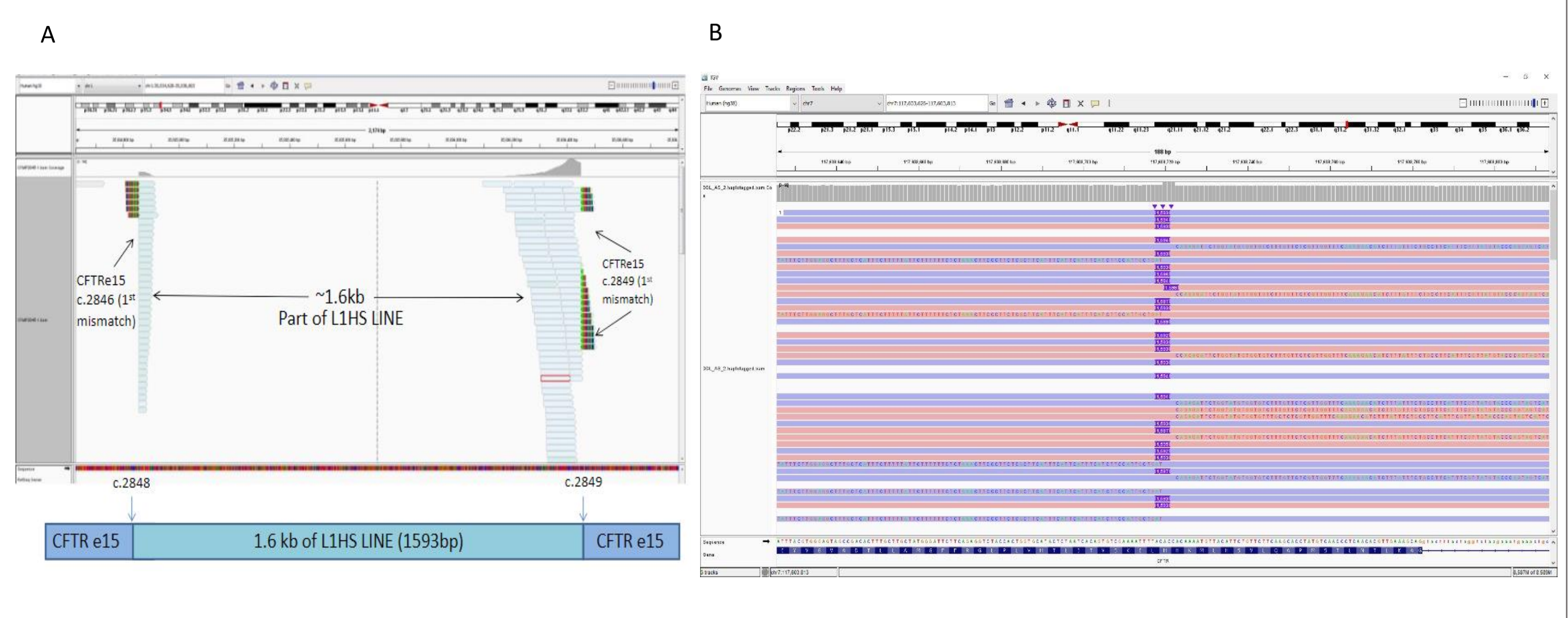
A
B
CFTRe15 c.2846 (1st mismatch)
~1.6kb
Part of L1HS LINE
CFTRe15 c.2849 (1st mismatch)
c.2848
c.2849
CFTR e15
1.6 kb of L1HS LINE (1593bp)
CFTR e15

Figure 6A/B. Insertion of 1,593 nucleotides from LINE (L1HS) into *CFTR*. (A) Results from short read sequencing vs (B) long read sequencing.

Discussion/Conclusion

- ❖Careful protocol optimization allows for lower DNA input, but still comes with significantly higher incidence of redos of both library prep and sequencing to achieve desired coverage.
- ❖Improvements to basecalling have been significant, but still leave considerable room for improvement as all deck positions on a P24 cannot be used while running SUP basecalling unless connected to two A100 towers.
- ❖Despite challenges, continuing improvements are making high-throughput long read sequencing services more achievable.