Are ONT long read human whole genomes ready for production scale?

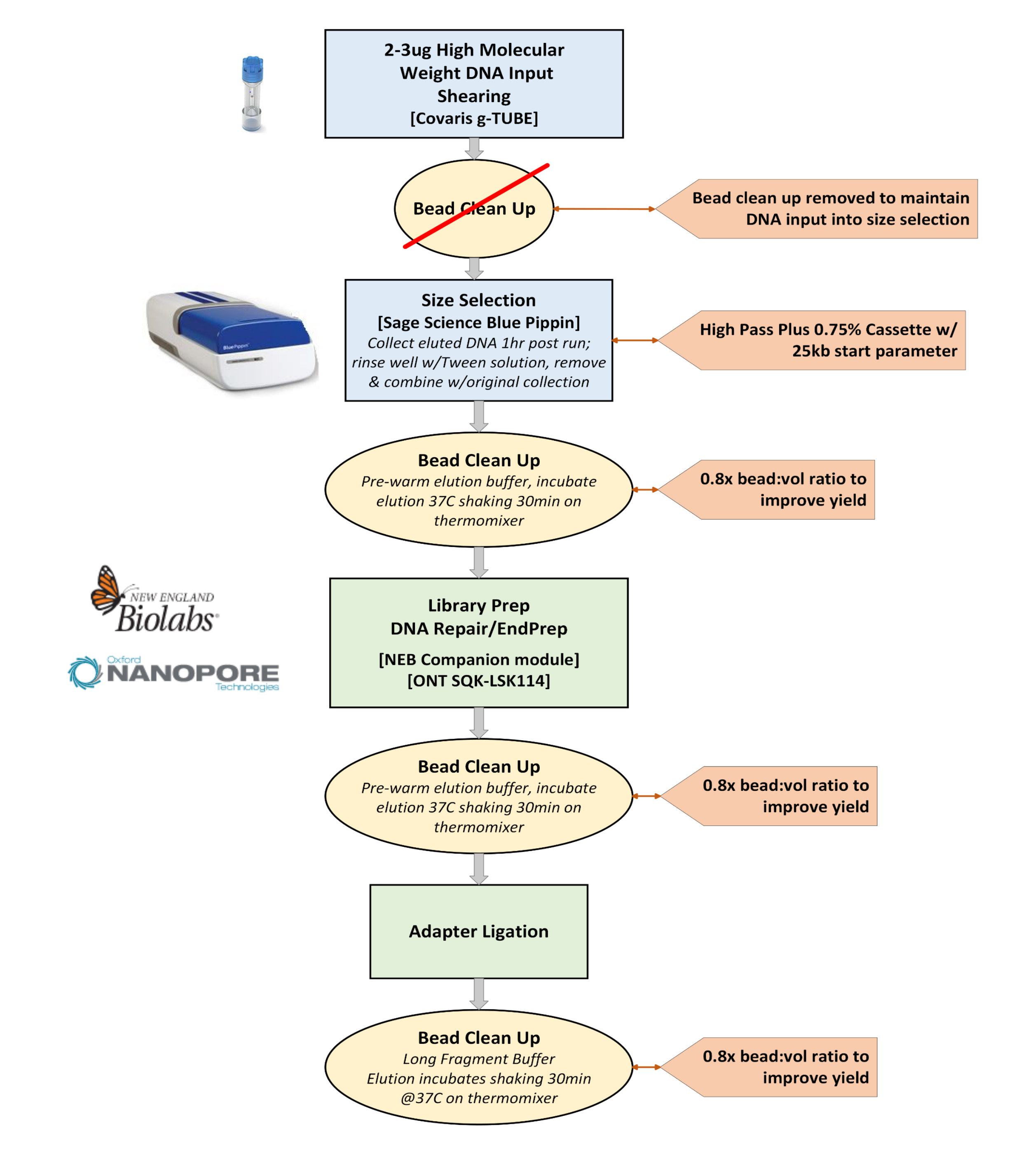
B Marosy[1], M Kokosinski[1], J Gearhart[1], J Paschall[1], B Craig[1], M Mawhinney[1], D Mohr[1], M Sheridan[1], P Witmer[1], D Muzny[2], A Scott[1], J Hosea[3], C Montano[3], L Morina[3], Q Li[3], A Klein[4], M Schatz[3], W Timp[3], K Doheny[1] and the *All of Us Research Program*

[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 – Project: All of Us Research Program (AoU)

Long read sequencing has been a steadily developing technology with several advantages over short read sequencing including detection of structural variants and expansion repeats, phasing of variants, and direct detection of base modifications in both DNA and RNA. The Center for Inherited Disease Research (CIDR) quickly implemented high-throughput human WGS sequencing on the ONT platform last year to meet the needs of the All of Us Research Program as part of the AoU Baylor-Hopkins Center for Clinical Genetics.

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, developed at Baylor HGSC and optimized for longer N50s at CIDR, utilizes 2-3 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 (**Figure 1**).


2-3ug High Molecular Weight DNA Input Shearing [Covaris g-TUBE]

~~Bead Clean Up~~ ← Bead clean up removed to maintain DNA input into size selection

Size Selection [Sage Science Blue Pippin]
Collect eluted DNA 1hr post run; rinse well w/Tween solution, remove & combine w/original collection ← High Pass Plus 0.75% Cassette w/ 25kb start parameter

Bead Clean Up
Pre-warm elution buffer, incubate elution 37C shaking 30min on thermomixer ← 0.8x bead:vol ratio to improve yield

Library Prep
DNA Repair/EndPrep
[NEB Companion module]
[ONT SQK-LSK114]

Bead Clean Up
Pre-warm elution buffer, incubate elution 37C shaking 30min on thermomixer ← 0.8x bead:vol ratio to improve yield

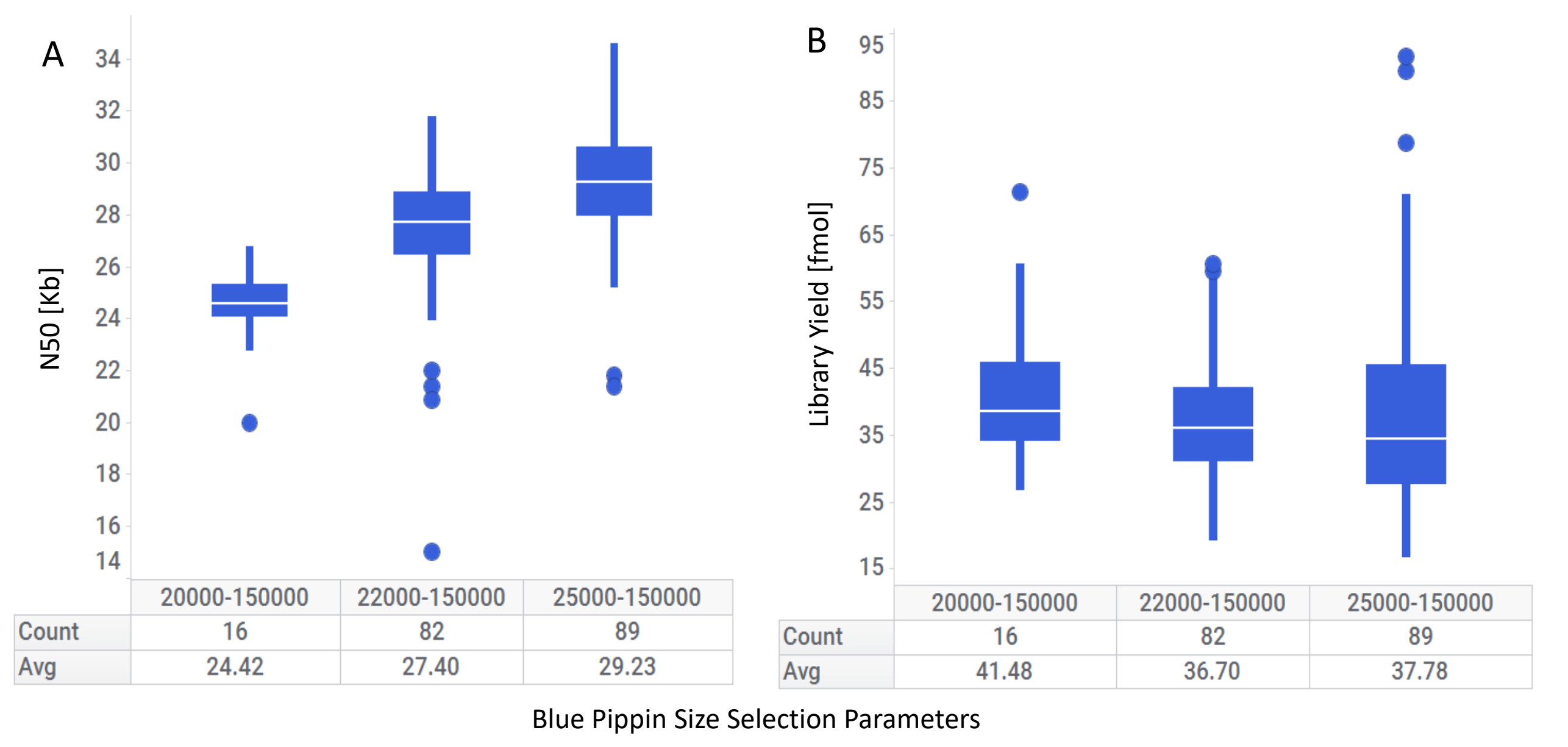
Adapter Ligation

Bead Clean Up
Long Fragment Buffer Elution incubates shaking 30min @37C on thermomixer ← 0.8x bead:vol ratio to improve yield

Figure 1. AoU protocol workflow depicting each major step along with enhancements to increase yield of high molecular weight fragments. Arrows in orange depict additional optimization at CIDR to increase yield and longer n50s.

Methods Improvement to Library Prep – Project AoU

In order to optimize for low inputs (2-3ug) and attempt higher N50s, we first addressed DNA yield going from shear and size selection into library prep. Our initial experiments with test DNA (pooled samples representative of sample source from the AoU biobank) indicated up to a 27% loss of DNA post shear clean up, and an additional 59% loss of DNA after Pippin size selection and clean up. ONT recommends starting with 1ug into library prep, after shearing and size selection are performed. By eliminating the post shear clean up, DNA input was maintained going into size selection with less overall loss prior to library prep. We also observed improved library yields when starting with more than 1ug of DNA into library prep, preferably 1.5ugs. In addition we transferred the process from 1.5ml Eppendorf tubes to 96 well plate format where possible and tested the use of lo bind plates throughout the process. We observed stable library prep yields when using lo bind plates for End Repair and Ligation reactions, and using MIDI style deep well plates for clean ups (data not shown). Utilizing this optimized protocol our library redo rate is 11% (redo if library yield is <16 fmol yield). Once library yield was stabilized we adjusted size selection parameters to confirm highest setting that would still maintain yield. **Figure 2a/b** depicts the N50 size achieved for 3ug DNA input when altering the size selection start parameters to 20, 22 and 25kb. Using the optimized protocol we could achieve similar yields when increasing the size selection parameters.



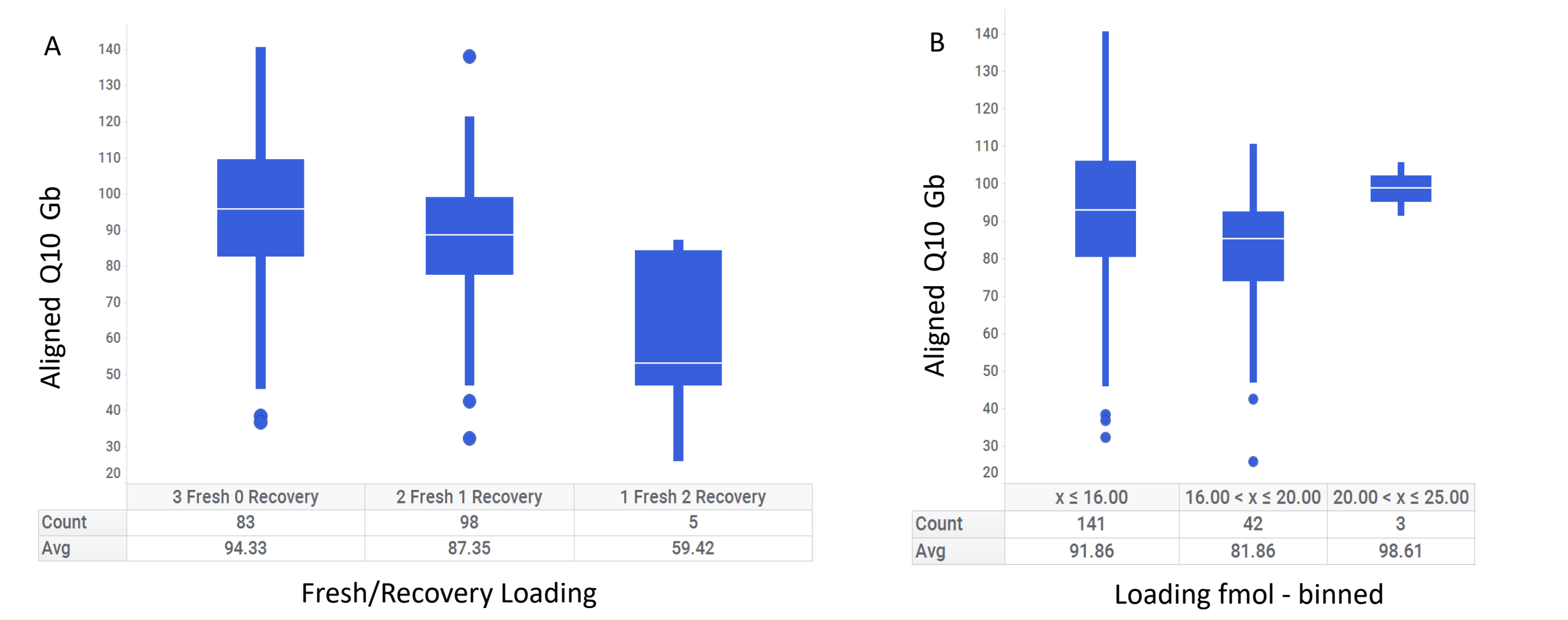
	20000-150000	22000-150000	25000-150000
Count	16	82	89
Avg	24.42	27.40	29.23

	20000-150000	22000-150000	25000-150000
Count	16	82	89
Avg	41.48	36.70	37.78

Figure 2A/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.

Methods Improvements to Sequencing – Project AoU

Following Library Prep samples are prepared for loading on the PromethION R10.4 flowcells. Flowcells are run for 72hrs with a wash/reload performed every 24 hours for a total of 3 loads. Typical loading uses 16fmol, but can use up to 25fmol when library yield allows for it. For libraries with lower yield, recovery of the loaded DNA can be done on the second or third load and used to reload. We compared coverage and sequencing yield (**Figure 3A**) when using variable loading inputs and either fresh or recovered library for each load. We observed that 3 fresh loads resulted in higher sequencing yield compared to 1 fresh load with 2 recovery loads. When restricting to samples that had either 3 or 2 fresh loads, we did not observe any benefit when loading higher amounts of DNA (**Figure 3B**). However, even with these optimizations, we observed a 41% redo/resequencing rate of flowcells (sequencing yield <90Gb). These ‘failed’ flowcells had an average of 59Gb produced per flowcell.

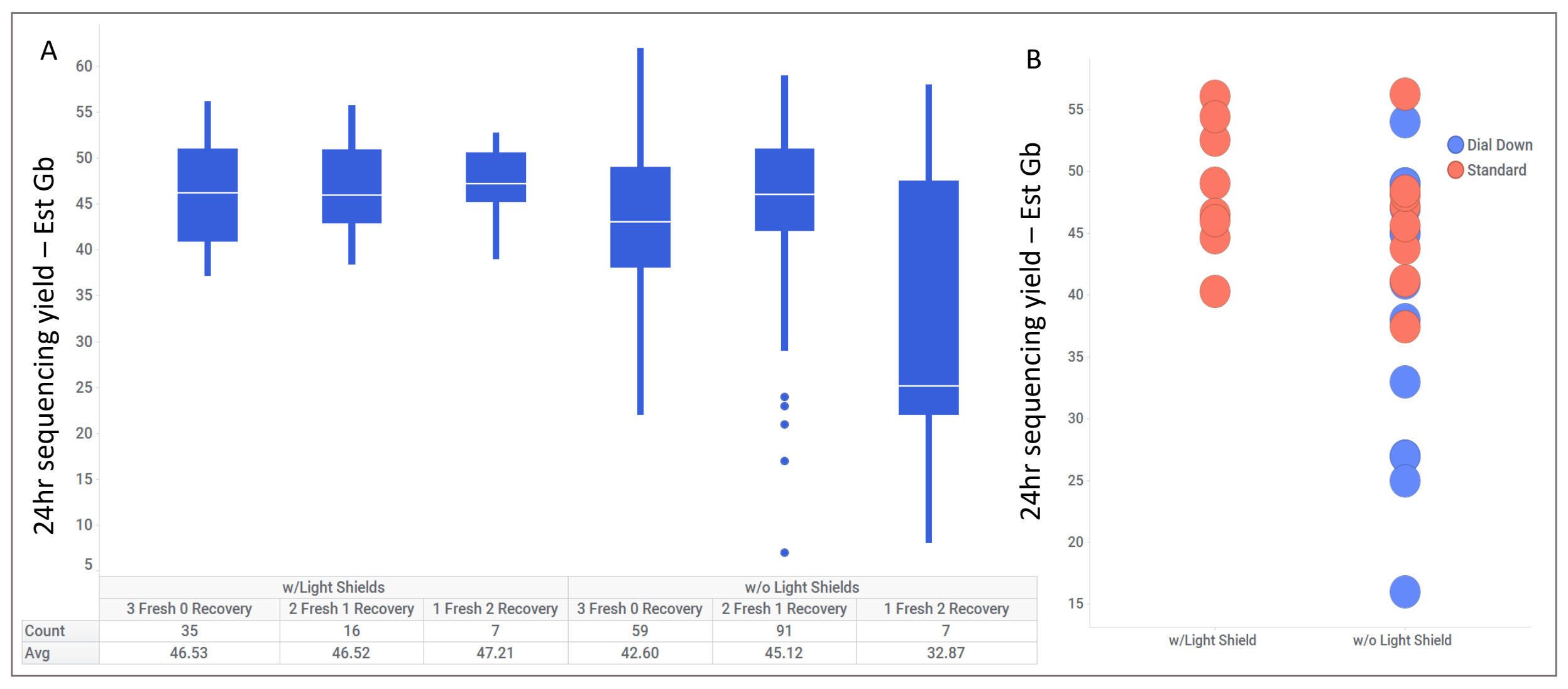


	3 Fresh 0 Recovery	2 Fresh 1 Recovery	1 Fresh 2 Recovery
Count	83	98	5
Avg	94.33	87.35	59.42

	x ≤ 16.00	16.00 < x ≤ 20.00	20.00 < x ≤ 25.00
Count	141	42	3
Avg	91.86	81.86	98.61

Figure 3A/B. Sequencing yield (Aligned Q10 Gb) from flowcells with (A) respective Fresh/Recovery loading conditions and (B) binned by fmol input for each loading.

Concurrently, ONT began providing light shields to be placed over the flowcells during processing to improve sequencing yield. When accounting for samples that used the light shield we could see improvement over yield regardless of 3, 2, or 1 fresh loads (**Figure 4A**). Incorporation of this practice into the workflow reduced the redo/resequencing rate to 24%, with an avg of 77Gb for those flowcells that were flagged/fell below 90Gb. This allowed for continued stabilization of the process to ensure performance. Additionally we investigated how light shield and dial down pipetting technique for loading flowcells affected sequencing yield. **Figure 4B** shows results of samples that used light shields w/wo the dial down method and the corresponding sequencing yield. We found speed of loading did not impact performance, however the use of light shields had the most impact on performance. To date we have released data on 187 samples achieving an average of 29.7x coverage and N50 of 28kb and Aligned Q10 Gb of 89.6 (**Figure 5**, **Table 1).**



	w/Light Shields			w/o Light Shields		
	3 Fresh 0 Recovery	2 Fresh 1 Recovery	1 Fresh 2 Recovery	3 Fresh 0 Recovery	2 Fresh 1 Recovery	1 Fresh 2 Recovery
Count	35	16	7	59	91	7
Avg	46.53	46.52	47.21	42.60	45.12	32.87

Figure 4A/B. Sequencing yield (estimated Gb from first 24hrs sequencing) from (A) flowcells with or without using light shields and the respective fresh/recovery loading condition, and (B) impact of pipetting speed vs light shield usage.

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 11 cell line derived DNA samples (fresh extractions optimized for long reads) achieved an average of 58x coverage and N50s of 27.4kb. In order to achieve this higher depth, we adapted our methods by increasing DNA input to LSK114 library prep to 6 ug, creating 2 libraries per sample, and sequencing 2 R10.4 flowcells per sample (**Table 1, Figure 5**) which reduced the sequencing redo rate to 3%. This project also used the long-read sequencing technique to explore the importance of variations in pancreatic cancer. For more detailed information please see Poster #102.

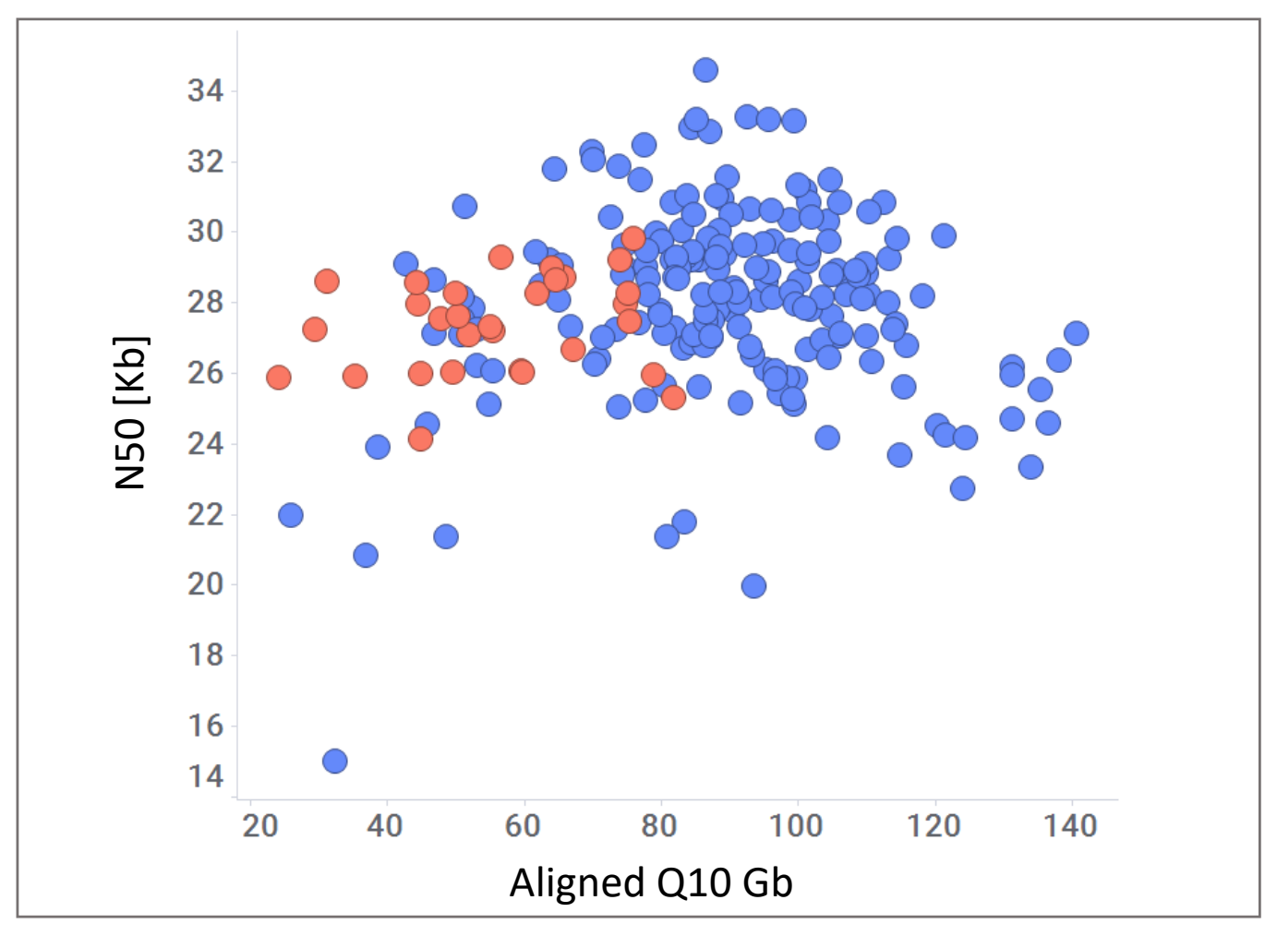


Figure 5. Sequencing yield and N50 [Kb] data produced for both AoU (blue) and Pancreatic Cancer Study (red) projects.

Project	AoU	Pancreatic Cancer Study
Samples	187	30
DNA Input [ug]	3	6
FCs/Sample	1	2
N50 [Kb]	28.0	27.4
Aligned_Q10 Gb	89.6	172.0
Mean Coverage	29.7	56.4

Table 1. Project QC summary stats

Methods Improvements to increase throughput

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. An earlier version of the MinKnow/Guppy software only allowed for 8 flowcells to be sequenced on 1 PromethION-24 with SUP + methylation basecalling which required an additional 4 days of processing time post 72hr sequencing run, resulting in a max throughput of 8 ~25X, N50~25kb WGS per week per P24. ONT then provided an improvement to increase the ‘chunks per runner’ parameter to achieve better use of the GPU. Once this configuration was in place we could process 11 samples/PromethION-24 using SUP plus Methylation calling in 5 days including the sequencing run time. To further increase our throughput, an additional A100 tower was connected to one of the PromethIONs, splitting the deck between the 2 towers (12 deck positions/tower), which allowed for simultaneous processing of 33 flowcells using live SUP basecalling across 2 PromethION-24 and 3 A100 towers. In reality it was infrequently possible to achieve this max throughput of 33 flowcells/week due to flowcell position issues or software issues.

In continued response, ONT released new basecalling software, Dorado, which provided substantial speed improvements improving run times by 25%. However, in order to accommodate the new software there was an initially unknown requirement for 220V power to the A100 towers. In addition, an alternate racked 4x NVIDIA A100 GPU system was installed to provide additional processing support when re-basecalling is required due to sequencer drops, new basecalling validation or additional methylation calling needs. This allows the sequencing towers to remain dedicated to maintaining the weekly sequencing through put.

Future applications – Adaptive Sampling

Currently, we are actively exploring adaptive sampling, to address clinical use for repeat expansion testing and elucidation of "unidentified” 2nd variants in individuals with cystic fibrosis, to inform treatment options. A sample with known repeat expansions for Huntington’s disease was prepared using Kit 14 UL with shearing and no size selection and sequenced on the PromethION-24 R10.4 flowcell. Adaptive sampling targeting all know repeat expansions based on the Miyatake et al paper (PMID: 36289212) was performed. The Oxford Nanopore wf-human-variation workflow was utilized for downstream analysis. We observed 3 fold enrichment of the HTT and JPH3 region (**Table 2**) and were able to correctly resolve the repeat expansions. We are currently transitioning this workflow to assess the CFTR region.

Adaptive Sampling	HD
Coverage of Genome	22x
Coverage of Targeted Regions	62x
Coverage of HTT Region	70x
Coverage of JPH3 Region	68x

Table 2. Coverage Metrics from Adaptive Sampling of Repeat Expansion panel.

Discussion/Conclusion

- Implementation of any new technology requires optimization and validation of overall performance and quality within a laboratory. Despite the challenges of utilization of this technology in a higher through-put workflow there continue to be improvements to the consistency of the technology.
- Removal of clean up steps when possible can improve library yields when DNA amounts and quality are limited.
- Sequencing with 20-25fmol fresh loads vs recovery is advantageous, however, utilization of light shields overall enables higher sequencing yield.
- Increasing the DNA input to 6ug and sequencing 2 flowcells/sample optimizes sequencing yield and reduces redo rate.
- Implementation of appropriate compute infrastructure is key to through-put and minimizes bottlenecks.