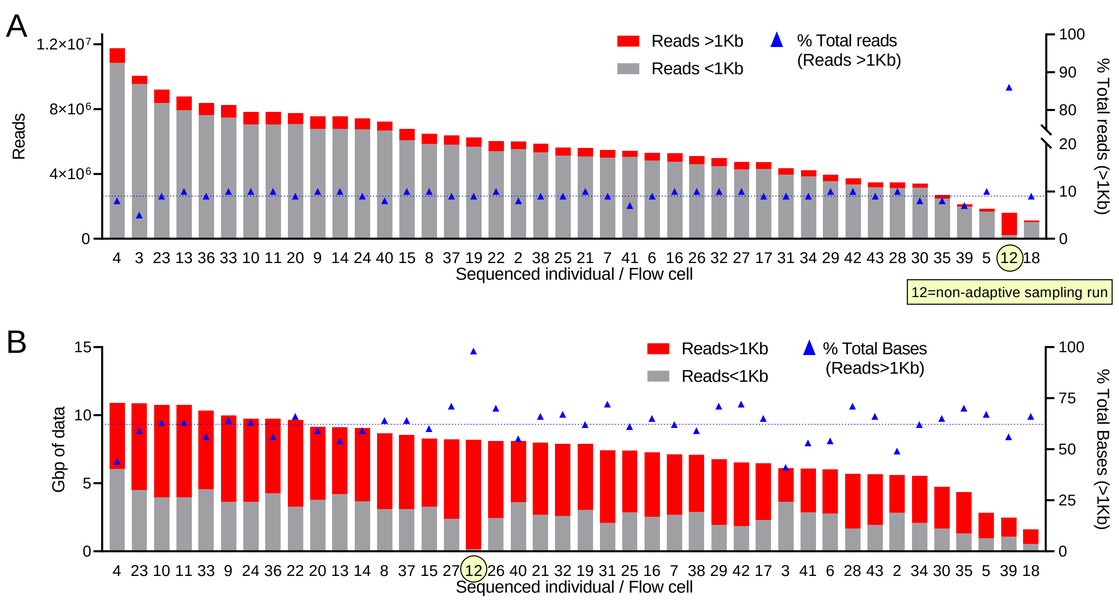
Adaptive sampling with the Oxford Nanopore MinION sequencer to identify new genes involved in male infertility

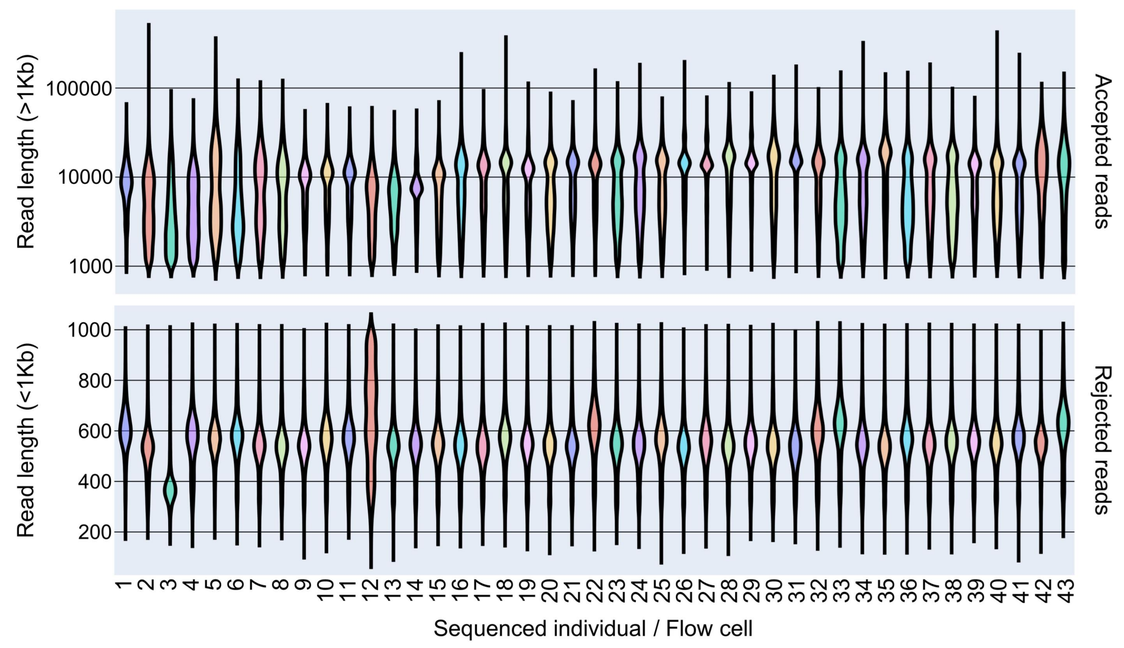
Medhat Mahmoud[1], Kimia Bakhshandeh[2], Larry I. Lipshultz[3], Fritz J. Sedlazeck[1,4], and Thomas X. Garcia[2,3]

[1]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA
[2]Department of Pathology & Immunology, Baylor College of Medicine, Houston, TX, USA
[3]Scott Department of Urology, Baylor College of Medicine, Houston, TX, USA
[4]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA

ABSTRACT: Male infertility is a widespread problem that is often hard to diagnose as multiple genes can impact this diverse phenotype. The advent of long-read sequencing using the MinION or PromethION platforms provides us a methodology to cost efficiently assess mutations across genes having an impact in male infertility at speed. In the present work we evaluated and implemented both platforms across control and patient samples. For this we utilized adaptive sampling on the MinION platform to target 21.1 megabase pairs (Mbp) of search space including 101 known male infertility genes, large regions of the Y chromosome known to contain microduplications and microdeletions (e.g. AZFb/AZFc), and 13 protein-coding genes that we hypothesize underly the cause of infertility. Over a four-month period we were able to sequence and analyze 48 men with male infertility and fertile control samples including reference DNA samples (HG001,HG002,HG003,HG004) together with replicates of certain samples amounting to 64 flow cells so far. All samples were reported with SNV and phased SV based on Princess. Thus, providing most comprehensive information across these genes. All our samples show consistently high data quality with 97% of mapped reads and high F-scores (precision and sensitivity) across the control samples for SNV and SV. This is possible due to the effectiveness of the selected sequencing averaging a 5-6 fold increase of coverage across the 113 genes of interest over non-regions-of-interest. Currently we are in the process of using phase data and deleteriousness scores (CADD, DANN) to filter and rank variants in each individual for compound heterozygosity and loss-of-function. To further optimize the speed of this approach we have now implemented the P2-solo in our pipeline showing a significant increase in data generation. Overall, in our presentation we will show how we our implementation of the MinION system yielded meaningful data that can be used to identify new mutations involved in male infertility.

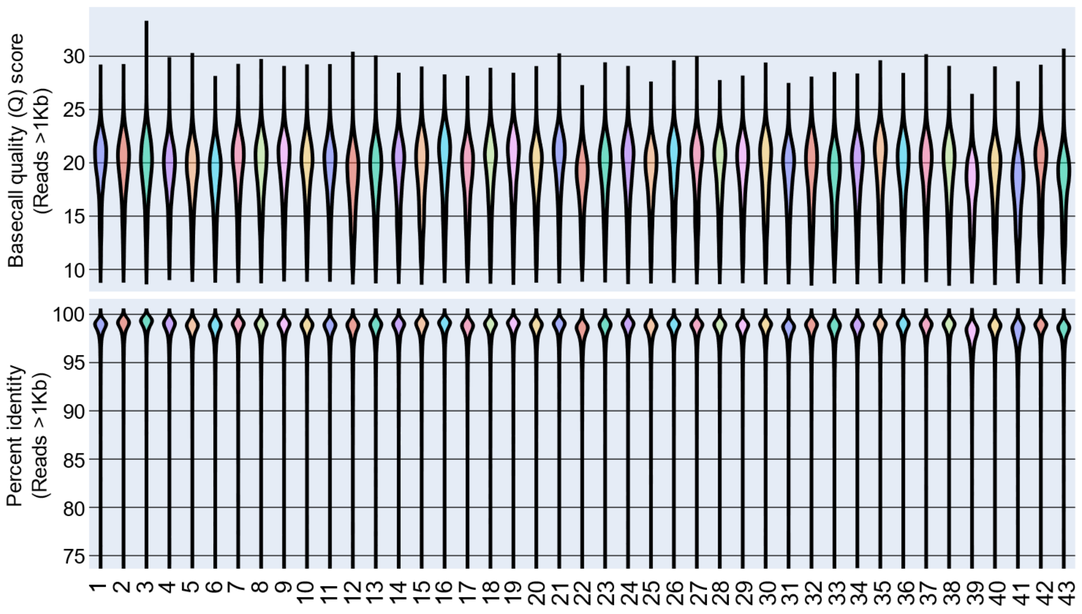
Read Length Size Distribution Across Samples



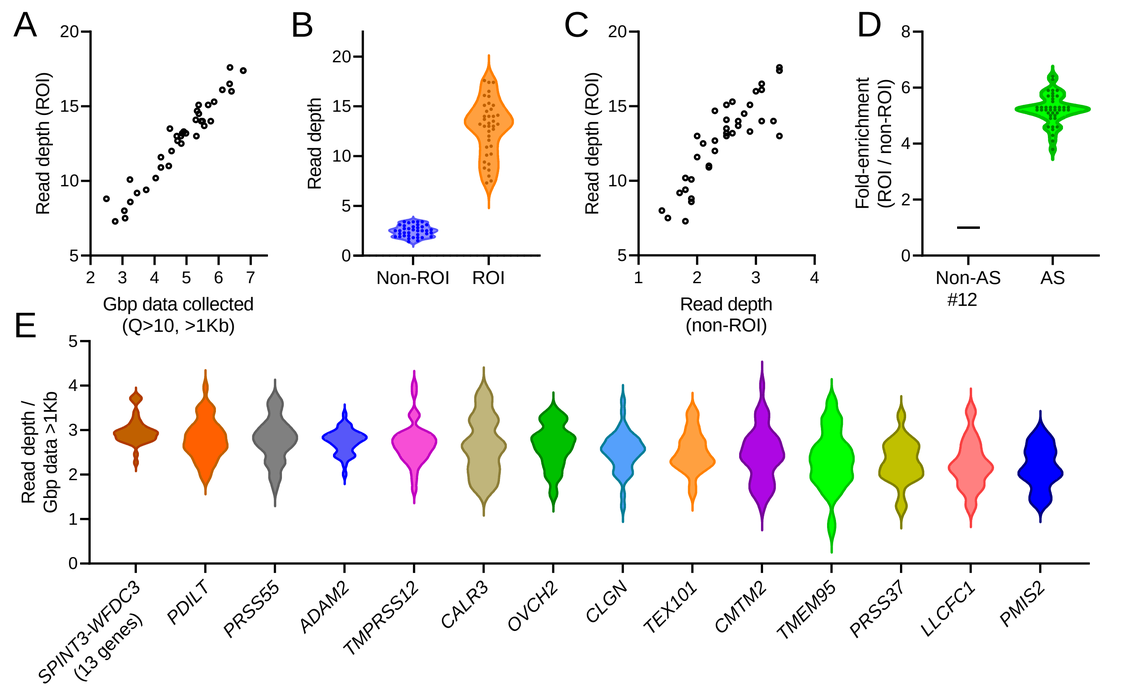


In three months, from August 2022 to December 2022, with a cumulative total of sixty-four R10.4.1 MinION flow cells, using 1 MinION sequencer and 1 GridION sequencer, both running in parallel, we were able to sequence 48 infertile men (4 of these men in duplicate) and 6 controls [2 fertile men, 1 in duplicate, and 4 reference DNA controls (HG001, n=3; HG002, n=4; HG003, n=1; and HG004, n=2)] and process this data through Guppy and Princess (our Python pipeline for alignment, SNV and SV calling, phasing, and QC). Our results show that adaptive sampling works as anticipated, with the majority of reads rejected. This is in contrast to one reference DNA run we made without AS enabled, which showed an overall average and peak read size of ~9 kilobases (Kb). Quantification of the amount of reads less than and greater than 1 Kb across all AS runs and the 1 non-AS run shows that AS causes ~77% of reads > 1 Kb, that would have normally been sequenced, to be rejected, and because of the huge rejections of reads, causes 37-38% of total throughput to come from reads < 1 Kb, instead of 2% as seen in our non-AS run. When considering reads only > 1 Kb in an AS run, then a clear peak in the abundance of reads at 9-12 Kb in size is apparent.

Read Quality Across Samples



ROI vs non-ROI Across Samples



With our SUP-bascalled fastq files filtered for a Q-score of 10 or greater, we observed mean for all samples at approximately 20. After mapping, the percent identity across all of our samples was ~97%, which demonstrated consistently good read alignment across all of our runs. Furthermore, visualizing data with a bivariate plots using Nanoplot we see that although percent identity of reads increases as a function of average base quality, read map quality is consistently high regardless of read length. Obtaining adequate read depth (RD) (number of overlapping reads) in our regions of interest (ROI) is critical for being able to confidently call variants. Among the most important questions we asked in analyzing our data was 1) what was our average weighted read depth across all and within each ROI, and 2) whether we obtained adequate read depth to confidently call variants. We found that runs can yield data amounts as low as 2.5 Gbp data (Q score >10; Read size > 1 Kb) to as high as 6.8 Gbp data with average weighted read depths from as low as 7.3 reads to as high as 17.6 reads. Average weighted read depth in our ROI is directly proportional to the amount of data collected in that run. When comparing non-ROI to ROI, we found that non-ROI had an average weighted read depth of 2.5 reads versus ROI that had an average weighted read depth of 13 reads, which is directly proportional across samples. When comparing ROI to non-ROI in the 1 sample we ran without adaptive sampling, we found that ROI and non-ROI had a near identical average weighted read depth of 2.5, giving us a fold-enrichment in ROI of 1. When comparing average weighted read depth in ROI and non-ROI across all our samples we found a consistent average 5.2-fold enrichment across all of the ROI relative to non-ROI. Next, we wondered whether any of the individual gene regions we enriched for showed significantly better or worse enrichment. No single gene or gene region was consistently better or worse in any given run. However, some genes such as PDILT and PRSS55, on average, demonstrated significantly higher overall average normalized read depth (2.8 ± 0.5 reads per Gbp data >1kb) in comparison to other genes like LLCFC1 and PMIS2, which are lower (2.2 ± 0.5 reads per Gbp data >1kb). Based on these results, we do not anticipate that this will adversely affect our ability to identify variants in any given gene or gene region.

The work presented herein was supported by the Male Contraceptive Initiative (David Sokal Innovation Award 2021-302) and the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD R01HD106056) to Thomas X. Garcia.
Oxford Nanopore Technologies products are not intended for use for health assessment or to diagnose, treat, mitigate, cure, or prevent any disease or condition. PO_1162(EN)_V3_14Nov2022