High molecular weight DNA extraction and long-read next-generation sequencing of human genomic reference standards

Matthew W. Mitchell[1]*, Primo Baybayan[2]*, Aaron Wenger[2], William Rowell[2], Lei Zhu[2], Gretchen Smith[1], Deborah V. Requesens[1]

[1]*Coriell Institute for Medical Research; 403 Haddon Avenue, Camden, NJ 08103; mmitchell@coriell.org*

[2]*Pacific Biosciences, Menlo Park, CA 94025; pbaybayan@pacificbiosciences.com*

ASHG VIRTUAL MEETING 2020 OCTOBER 27-30

INTRODUCTION

Recent advances in long-read next-generation sequencing have made it possible to produce sequence reads greater than 100 kilobases (kb). This has in turn sparked a growing interest in producing high (100-300kb) and ultra-high (>300kb) molecular weight DNA to sequence with this technology. Long reads allow researchers to characterize structural variation that are more challenging or impossible to detect with other approaches, such as inversions, translocations, duplications, and insertions of mobile repetitive elements (Chaisson et al. 2014). Longer read lengths also improve the accuracy of haplotype phasing analyses and can be used to better characterize highly polymorphic regions with complex linkage disequilibrium structure (e.g. human leukocyte antigen genes in the major histocompatibility complex) (Hosomichi et al. 2015). Here, we have extracted high molecular weight (HMW) DNA from a highly characterized reference standard cell line (GM24385) in the NIGMS Human Genetic Cell Repository using Circulomics Nanobind technology (Zhang et al. 2016) and compared it to a matched control (NA24385) extracted using 'standard' methods (Miller et al. 1988). DNA was sheared to 20-25kb and constructed to a SMRTbell library using Express Prep 2.0. The SMRTbell library was sequenced on the Sequel II System generating >25 Gb of highly accurate reads with mean HiFi read length of 20-25 kb and >99% accuracy. HiFi reads were mapped to the human reference genome with pbmm2. More than 3 million small variants were called with DeepVariant, and more than 20,000 structural variants were called with pbsv, which is 3-5 times more than identified in typical short-read callsets. Variants were phased with WhatsHap into phase blocks with N50 length >200 kb. The assembled sequences from these highly characterized samples will serve as important and vital benchmarks for future research utilizing HMW DNA and long-read next-generation sequencing.

HIGH MOLECULAR WEIGHT DNA EXTRACTION

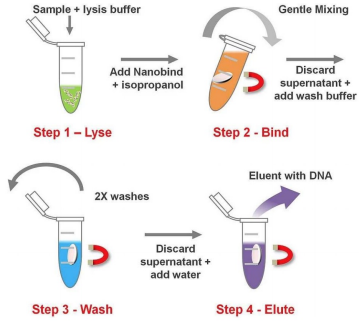

Figure 1. We performed HMW DNA extraction using the Circulomics Nanobind CBB Big DNA Kit. The Nanobind extraction process produces DNA molecules ranging from 50-300+ kb in length with ultra-low damage and high purity in under an hour. Extractions were performed using an optimized protocol provided by Circulomics on the KingFisher Flex. DNA quality was measured on the Agilent 2200 TapeStation System. Double stranded DNA was measured using the Qubit dsDNA BR Assay Kit.

JUSTIFICATION

Long-read sequencing technology has enabled researchers to significantly advance the field of genome biology (Amarsinghe et al. 2020; Burgess 2018). One of the most impactful uses of this technology is that it has vastly improved reference genome assembly and has been utilized to generate and improve human genome benchmarks using reference standard samples (Bowden et al. 2019; Wang et al. 2019; Zook et al. 2016). The Coriell Institute of Medical Research houses a number of repositories (including the NIGMS Human Genetic Cell Repository) that contain highly characterized samples that serve as reference materials for human genomics researchers. By making HMW DNA available for these samples that are suitable for use on long-read sequencing platforms, including the PacBio Sequel II System, we can establish a standardized, centralized, and reproducible resource that will help advance studies of human genomics.

DNA QUALITY CONTROL

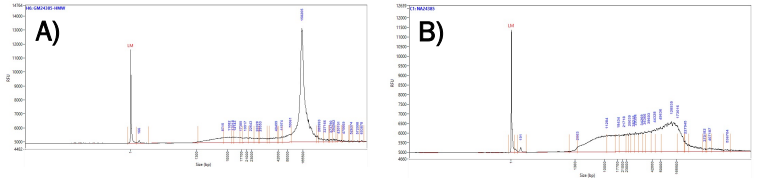

Figure 2. DNA fragment size was measured using the Femto Pulse System (Agilent Technologies) for two DNA samples extracted from GM24385 cells. **(A)** HMW DNA extracted using Circulomics Nanobind technology is highly enriched for molecules of ~150kb, well within the range expected when using the Nanobind CBB Big DNA Kit. **(B)** DNA extracted using a Miller salting-out procedure (Miller et al. 1988) shows a wide range of DNA fragment sizes, ranging from 2-200kb. These results demonstrate the benefit of the Circulomics HMW DNA extraction, which enriches the resulting DNA for ultra-long molecules, well above the 20-25kb fragment size targeted for HiFi SMRTbell library prep and sequencing on the Sequel II System.

ACKNOWLEDGEMENTS

We would like to thank José Santana (Coriell), Luke Hickey (PacBio), Jeffrey Burke (Circulomics), and Kelvin Liu (Circulomics) for their contributions to this work. This work was funded by the National Institutes of Health and the National Institute of General Medical Sciences (Project# 2U42GM115336-06).

SIZE SELECTION & LIBRARY PREP

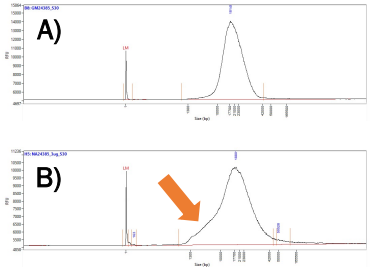

Figure 3. Genomic DNA was sheared on the Megaruptor 3 (Diagenode Inc.) for HMW DNA (A) and standard DNA (B). The shoulder shown by the orange arrow indicates that standard DNA is more fragmented prior to shearing.

Table 1. Post size selection, HMW DNA demonstrated better yield compared to the matched standard control.

Library Prep Yield	HMW DNA (GM24385)	Standard DNA (NA24385)
Input (ng)	3,520	2,500
SMRTbell Library Yield (ng)	858	723
Final Size-Selected HiFi Library (ng)	490	112

HIFI READ LENGTH DISTRIBUTION

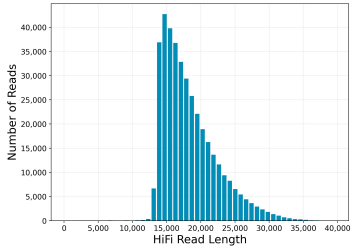

Figure 4. HiFi read length distribution of the HMW DNA library sequenced on the Sequel II System. The data shows a mean HiFi read length of ~18kb, a median HiFi Read Quality of Q31, and a mean HiFi number of passes of 9. This read length distribution illustrates that the HMW DNA extracted from GM24385 is a suitable standard for use in long-read sequencing applications.

REFERENCES

- Amarasinghe, S.L. et al. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*, 21(1), 30.
- Bowden, R. et al. (2019). Sequencing of human genomes with nanopore technology. *Nature Communications*, 10(1), 1869.
- Burgess, D. J. (2018). Next regeneration sequencing for reference genomes. *Nature Reviews Genetics*, 19(3), 125-125.
- Chaisson, M.J.P. et al. (2014). Resolving the complexity of the human genome using single-molecule sequencing. *Nature*, 517, 608.
- Hosomichi, K., Shiina, T., Tajima, A., & Inoue, I. (2015). The impact of next-generation sequencing technologies on HLA research. *Journal Of Human Genetics*, 60, 665.
- Miller, S. A., Dykes, D. D., & Polesky, H. F. (1988). A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Research*, 16(3), 1215.
- Wang, Y.-C. et al. (2019). High-coverage, long-read sequencing of Han Chinese trio reference samples. *Scientific Data*, 6(1), 91.
- Zhang, Y. et al. (2016). A Simple Thermoplastic Substrate Containing Hierarchical Silica Lamellae for High-Molecular-Weight DNA Extraction. *Advanced Materials*, 28(48), 10630-6.
- Zook, J. M. et al. (2016). Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific Data*, 3(1), 160025.