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

Matthew W. Mitchell*, Primo Baybayan*, Aaron Wenger2, William Rowell2, Lei Zhu2, Gretchien Smith1, Deborah V. Requeens1
1Coriell Institute for Medical Research; 403 Haddon Avenue, Camden, NJ 08103; mmitchell@coriell.org
2PacBio Biosciences, Menlo Park, CA 94025; pbaybayan@pacificbiosciences.com

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 (300-3000x) range of highly pure, high 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 shorter reads including insertions, deletions, 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. 2016). 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. 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 research. 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.

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 Agilent 7500 Solution Series DNA Fragmentation System. The shoulders shown by the orange arrow demonstrate better yield compared to the standard DNA extraction, which enrichesthe 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.

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 research. 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
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# 2U42GM153366-06).

ACKNOWLEDGEMENTS
REFERENCES