High Coverage, Ultra-long Read Sequencing of Human Genomes

Jennifer Idol, Liang Gong, Chee Hong Wong, Dave Harrison, Lauren Bellfy, Adam Mil-Homens, Mallory Ryan, Chew Yee Ngan, Charles Lee, Chia-Lin Wei

Genome Technologies, The Jackson Laboratory for Genomic Medicine, Farmington, Connecticut 06032 USA

ABSTRACT

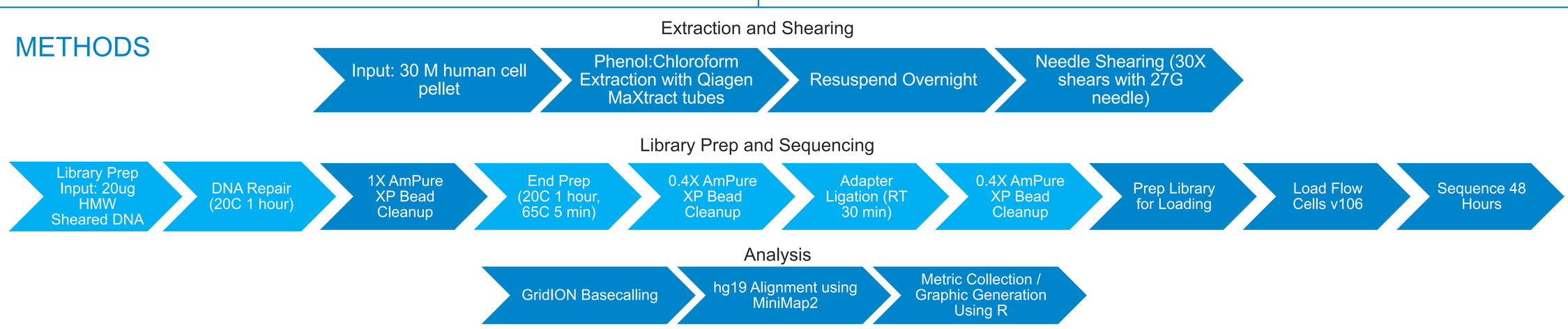
The long reads and improved data throughput enabled by the Oxford Nanopore platform have enabled many genomic analyses, including detection of SVs and de novo genome assembly. Here, we describe the use of GridIONs to generate high coverage human genome sequences of three individuals from the 1000 Genomes Project selected by the Human Genome Structural Variation Consortium for comprehensive haplotype-aware structural variation detection. To exploit Nanopore's potential of long read length and sequencing yield, we optimized protocols for preparing high-molecular weight genomic DNA library. Notably, we achieved multi-Gigabases (Gb) of reads of hundreds of kb of aligned sequences per run. To take advantage of the GridION highly parallel nature, we further built an efficient sample-to-data generation workflow and streamlined the data storage and QC pipeline. A single human genome of 30X coverage could be completed in two and a half weeks. With the high coverage Nanopore data from the well-characterized human genomes, we report the characterization and standardization of Oxford Nanopore performance matrices based on the latest technology updates.

INTRODUCTION

The purpose of the Human Genome Structural Variation Consortium (HGSV) is to sequence three trios from the 1000 Genomes Project for high-quality structural variation. The project displayed in this poster includes one human from each trio (GM19240, HG00514 and HG00733). GM19240 is a from Yoruba, HG00514 is Han Chinease and HG00733 is Puerto Rican. These three human samples therefore give a good distribution worldwide. The HGSV also aims to use new technologies to achieve their goal including taking advantage of the growing field of long reads.

Oxford Nanopore has been a leader in the field of long read sequencing and the GridION system allows for long read data in a high throughput sequencing capacity. The ability to sequence five flow cells simultaneously for 48 hours allows ten libraries to be run per week. This throughput enables up to 70 Gb of data to be sequenced on a single sample in one week. Adding time for high molecular weight extractions, a 30X human genome can be sequenced in as little as 2 weeks.

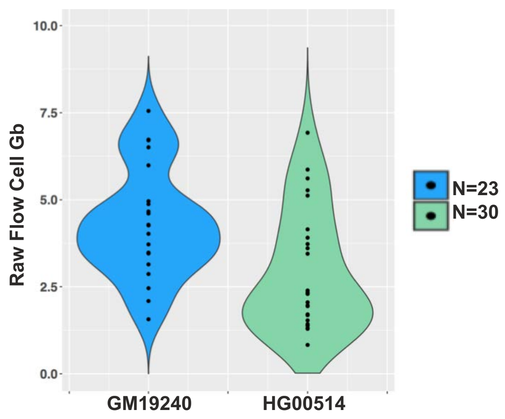
METHODS



Project Summary

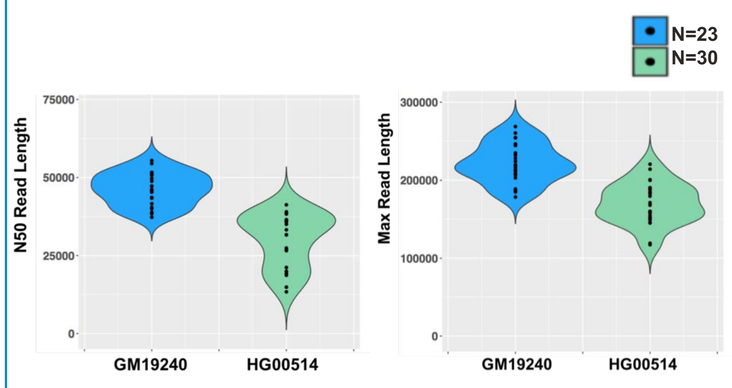
	GM19240	HG00514
# Extractions Needed	3	4
# Flow Cells for 30X Coverage	23	30
Total Yield (Gb)	99.1	97.4
Average Flow Cell Yield (Gb)	4.3 ± 1.6	2.9 ± 1.7
N50 Read Length (bp)	45,470 ± 5,117	30,111 ± 8,565
Max Read Length (bp)	268,840	220,670

Raw Yield Of Flow Cells



Raw yields for flow cells of each sample. HG00514 has more flow cells with less yield because of an older extraction.

Max and N50 Read Length

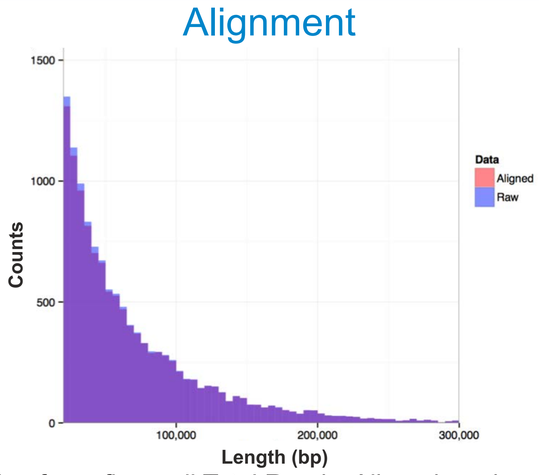


Max and N50 read length for flow cells of each sample. This can vary according to extraction, library and flow cell quality.

Read Quality Scores

Qscores of a single flow cells for each sample with each dot representing a single read. Qscores average 10.2 for GM19240 and 10.0 for HG00514.

Alignment

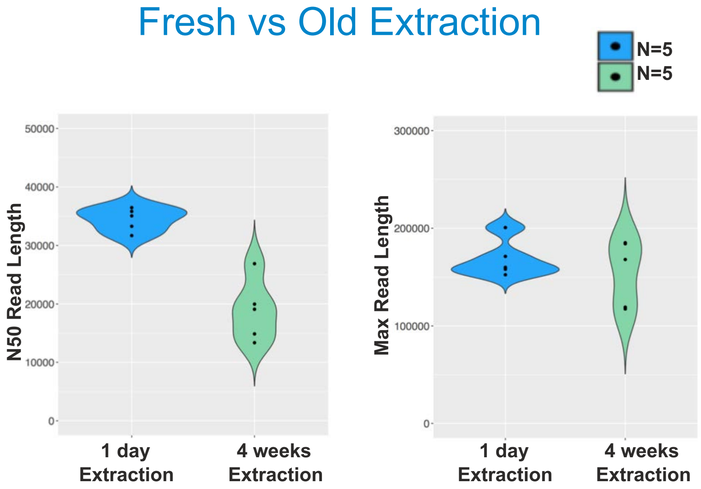


Example of one flow cell Total Reads. Aligned reads are shown in the overlap of Aligned (red) and Raw (blue) which forms purple. >90% of all reads align.

CONCLUSIONS

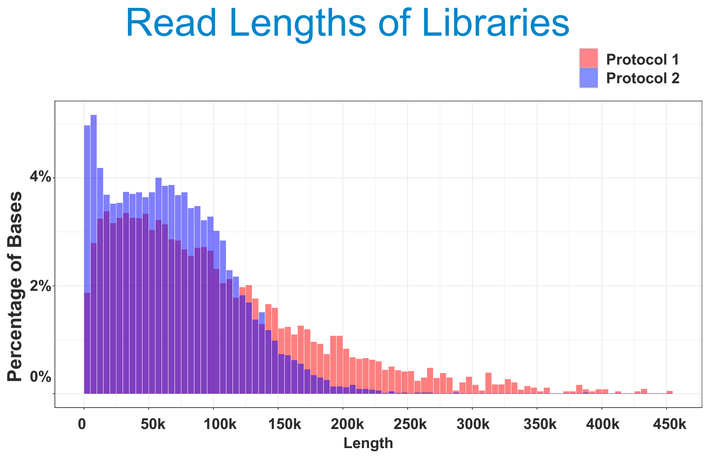
- 30X Coverage on a Human Genome can be completed in 2 weeks with 40X coverage taking 2.5 weeks using 1 GridION.
- High molecular weight extractions produce data with a longer length and more yield when used fresh for library prep.
- Phenol:Chlorofom extractions combined with needle shearing produce high quality sequences.
- Stringent bead cleanups (0.4X) help to remove smaller reads from sequencing data.
- There is a tradeoff on length vs yield on GridION. However, even with increased yield, max length is between 100kb – 450kb.
- Length of libraries is sample dependent. For our work we saw one sample had slightly shorter N50 and max lengths. Yields were consistent even with these differences across samples.

Fresh vs Old Extraction



5 libraries made from both old and fresh extractions. Fresh extractions produce better N50 and Max lengths.

Read Lengths of Libraries



Example of two flow cells (one blue and one red) complete read length graph. Some flow cells get up to 250kb while others have shown up to 450kb.

LITERATURE

- For more information on the 1000 Genomes Project please visit: http://www.internationalgenome.org/
- For more information on the Human Genome Structural Variation Consortium (HGSV) please visit: http://www.internationalgenome.org/human-genome-structural-variation-consortium/
- For more information on the Oxford Nanopore GridION please visit: https://nanoporetech.com/products/gridion

ACKNOWLEDGEMENTS AND FUNDING

The authors would like to thank James Brayer at Oxford for his continuing support throughout this project.
Genome Technologies at The Jackson Laboratory for Genomic Medicine is partially supported by the National Cancer Institute of the National Institutes of Health under Award Number P30CA034196.

The Jackson Laboratory
Leading the search for tomorrow's cures