Scalable Long-Read Low-Coverage Human Whole Genome Sequencing (lcWGS) for Genotype Imputation

Michelle Rahardja, Sabina Gude, Maura Costello, Christianto Putra, Yanyan Liu, Ashley Silvia, Zac Zwirko, Gavin Rush, and Joe Mellor

seqWell, Inc. Beverly, MA USA

Introduction

Low-coverage whole-genome sequencing (lcWGS) combined with long-read sequencing technologies represents a powerful approach for genomic analysis, facilitating the exploration of genetic variation with unprecedented depth and accuracy. lcWGS allows researchers to capture a broad overview of the genome at a lower cost, making it feasible to study large populations and rare variants. When integrated with long-read sequencing systems such as PacBio's Revio™, lcWGS enhances the resolution of single nucleotide polymorphisms (SNPs) and other structural variations. One key to unlocking the efficient use of lcWGS in long read sequencing is the ability to prepare long read sequencing libraries at scale.

Here we demonstrate the use of LongPlex Long Fragment Multiplexing kit in long read lcWGS application using the reference cohorts of human genomic DNA samples. The eight human genomic DNA samples were fragmented and barcoded with LongPlex, underwent additional size-selection step using PacBio's Short Read Eliminator (SRE) to increase HiFi read lengths prior to the SMRTbell prep kit 3.0. The sequencing results were downsampled to < 0.2X coverage and accuracy of SNP calls were determined for chromosome 22. The combined use of lcWGS and long read sequencing presents a cost-effective while producing reliable results for genotype imputation using the open-source GLIMPSE pipeline.

Methods

- Eight individual human genomic DNA from the reference cohorts of CEPH/Utah, Ashkenazi, and Han Chinese cohorts (Table 1) were fragmented and barcoded with LongPlex method.
- Size selection was done by Short Read Eliminator (PacBio) on pooled samples post LongPlex to eliminate fragments below 10kb, enhancing the sequencing efficiency and increase HiFi read lengths.
- The size-selected LongPlex library was directly processed using the PacBio SMRTbell Prep Kit 3.0. The resulting SMRTbell library was loaded into a SMRT cell-25M and sequenced in HiFi Whole Genome Sequencing mode with a 30-hour movie length.
- Sequencing data were individually down sampled to 40,000 random reads (0.16X coverage) before variant calling and imputation.
- The open-source GLIMPSE pipeline (v1.1.1) performed imputation using default settings. Leveraging reference data from the 1000 Genomes Project, the pipeline includes steps to genotype, impute, and phase variants.

Table 1. Summary of hgDNA samples assessed in the study

Coriell ID	NIST ID	Reference Cohort	DQN_{30kb}	DNA Input Amount (ng)
NA12878	HG001	Utah/Mormon	4.6	497
NA12891	N/A	Utah/Mormon	6.2	492
NA12892	N/A	Utah/Mormon	6.7	495
NA24385	HG002	Ashkenazi	6.8	480
NA24149	HG003	Ashkenazi	6.9	489
NA24631	HG005	Chinese	6.1	503
NA24694	HG006	Chinese	6.5	524
NA24695	HG007	Chinese	5.5	497

LongPlex Long Fragment Multiplexing Kit Workflow

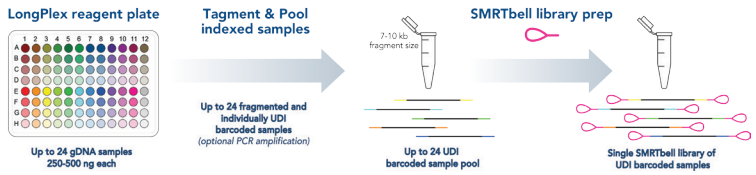

Figure 1. LongPlex uses Tn5 transposase to simultaneously fragment and index DNA samples in a rapid enzymatic workflow. Indexed samples are pooled together for further processing through PacBio SMRTbell Prep Kit 3.0. Post Revio sequencing, samples are demultiplexed using customized LIMA scripts followed by alignment and genotype imputation using the open-source GLIMPSE pipeline.

Library QC and Revio Sequencing Metrics

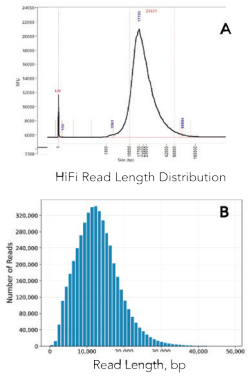

Table 2. Summary of library quantification QC and Revio sequencing run quality metrics

Yield Post LongPlex prep	2.2 µg
Yield Post Size Selection via SRE	1.1 µg
Yield Post SMRTbell prep kit 3.0	293 ng
Post SMRTbell prep Femto size (mode)	17.7 kb
Total HiFi Reads	4,737,266
Mean HiFi Read Length (bp)	12,850
Total HiFi Yield (Gb)	60.89
Median Read Quality	Q37

Figure 2. The electropherogram generated by an Agilent Femto Pulse gDNA 165kb Analysis shows trace of 8-plex LongPlex library pool using PCR-free workflow, size selected via SRE, post SMRTbell Prep Kit 3.0 (A). SMRT Link output (B) shows a mean read length at 12.9kb and a median HiFi read quality of Q37 (Table 2).

Genotyping from 0.2X Coverage of Well-characterized Human Genomes

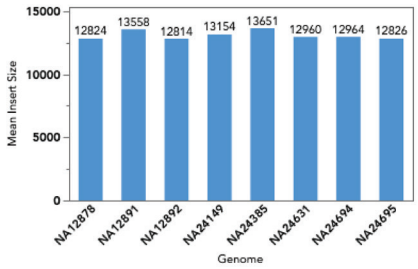

Figure 3. Insert size uniformity of an 8-plex LongPlex multiplexed library. A low CV (2.6%) of mean read length across multiple samples indicates consistency of fragmentation using Tn5 transposase.

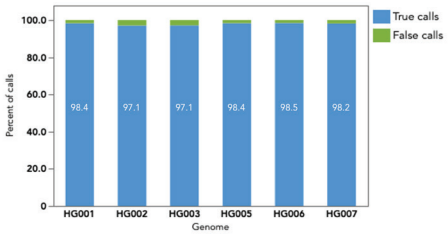

Figure 4. Accuracy of low-coverage WGS genotyping for eight individual human genomic samples with LongPlex on a Revio System. Data correspond to the confirmed heterozygous/homozygous chromosome 22 SNPs in the GIAB Consortium HG001, HG002, HG003, HG005, HG006, and HG007 truth data set.

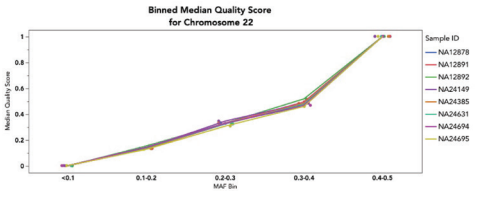

Figure 5. Median quality score for all variants in a certain minor allele frequency (MAF) range for chromosome 22 of eight individual hgDNA samples. The imputation quality score is an estimate of imputation quality on a scale of 0 to 1, where 1 indicates that a genotype has been imputed with high certainty. Variants with very low MAF may lead to less reliable imputation results, as there may not be enough data to accurately predict the missing genotype.

Summary and Conclusions

- The LongPlex Long Fragment Multiplexing Kit enables multiplexed, highly scalable construction of long read library pools for low-coverage WGS by simultaneously fragmenting and indexing up to 96 samples prior to SMRTbell Prep Kit 3.0 (Figure 1), increasing throughput and reducing time, labor and total sequencing costs.
- A high proportion of imputed variant calls (> 97%) were identical to those in the truth data set (Figure 4), confirming that a minimal amount of WGS data — only 40,000 read pairs or < 0.2x genome coverage — can achieve highly accurate genotyping.

More Information on LongPlex