UKB WGS pilot data on BGISEQ

BGI has performed whole genome sequencing for 50 samples from UK Biobank on BGISEQ-500. The pilot project has been done in Shenzhen, China. The whole genome sequencing data generated from BGISEQ-500 of 50 DNA sample(s) is with averagely 142,649.94 Mb raw bases. After removing low-quality reads we obtained averagely 1,328,520,531 clean reads (132,852.05 Mb). The clean reads of each sample had high Q20 (~98.36%) and Q30 (~92.59%), which showed high sequencing quality. The average GC content was 40.30%.

Table 1 Summary of whole genome sequencing data

PARAMETER	BGISEQ-500 (Average)
Clean reads	1328520531
Clean bases (Mb)	132852.05
Clean data rate (%)	93.12
Clean read Q20 (%)	98.36
Clean read Q30 (%)	92.59
GC content (%)	40.3
Mapping rate (%)	99.98
Unique rate (%)	90.70
Duplication rate (%)	2.89
Average sequencing depth	42.16
Coverage (%)	99.10
Coverage at least 4X (%)	98.44
Coverage at least 15X (%)	95.99

Total clean reads per sample were aligned to the human reference genome (GRCh38/HG38) using Burrows-Wheeler Aligner (BWA). On average, 99.98% mapped successfully and 90.70% mapped uniquely. The duplicate reads were removed from total mapped reads, resulting in about 2.89% duplicate rate and 42.16-fold mean sequencing depth on the whole genome excluding gap regions. On average per sequencing individual, 99.10% of the whole genome excluding gap regions were covered by at least 1X coverage and 95.99% had at least 15X coverage. In addition, the distributions of per-base sequencing depth and cumulative sequencing depth were shown as Figure 1 and Figure 2, respectively.

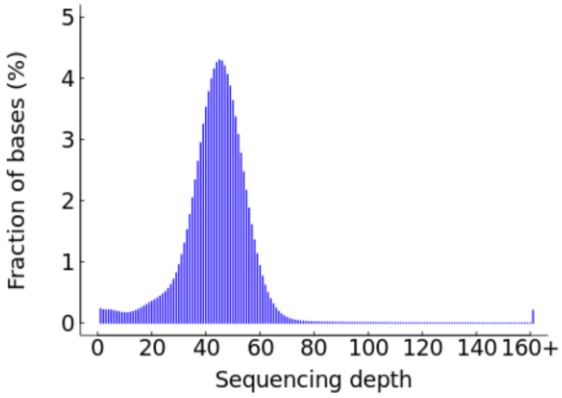

Figure 1 The distribution of per-base sequencing depth on the whole genome

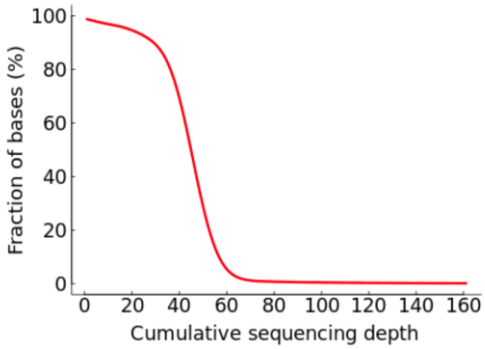

Figure 2 Cumulative depth distribution on the whole genome

BGISEQ-500, a BGI Sequencer, is an industry leading high-throughput sequencing system, powered by combinatorial Probe-Anchor Synthesis (cPAS) and improved DNA Nanoballs (DNB) technology. The cPAS chemistry works by linking a fluorescent probe to a DNA anchor on the DNB, followed by high-resolution digital imaging. This combination of linear amplification and DNB technology reduces the error rate while enhancing the signal. In addition, the size of the DNB is controlled in such a way that only one DNB is bound per active site in the flow cell. This patterned array technology not only provides sequencing accuracy, but it also increases the chip utilization and sample density.

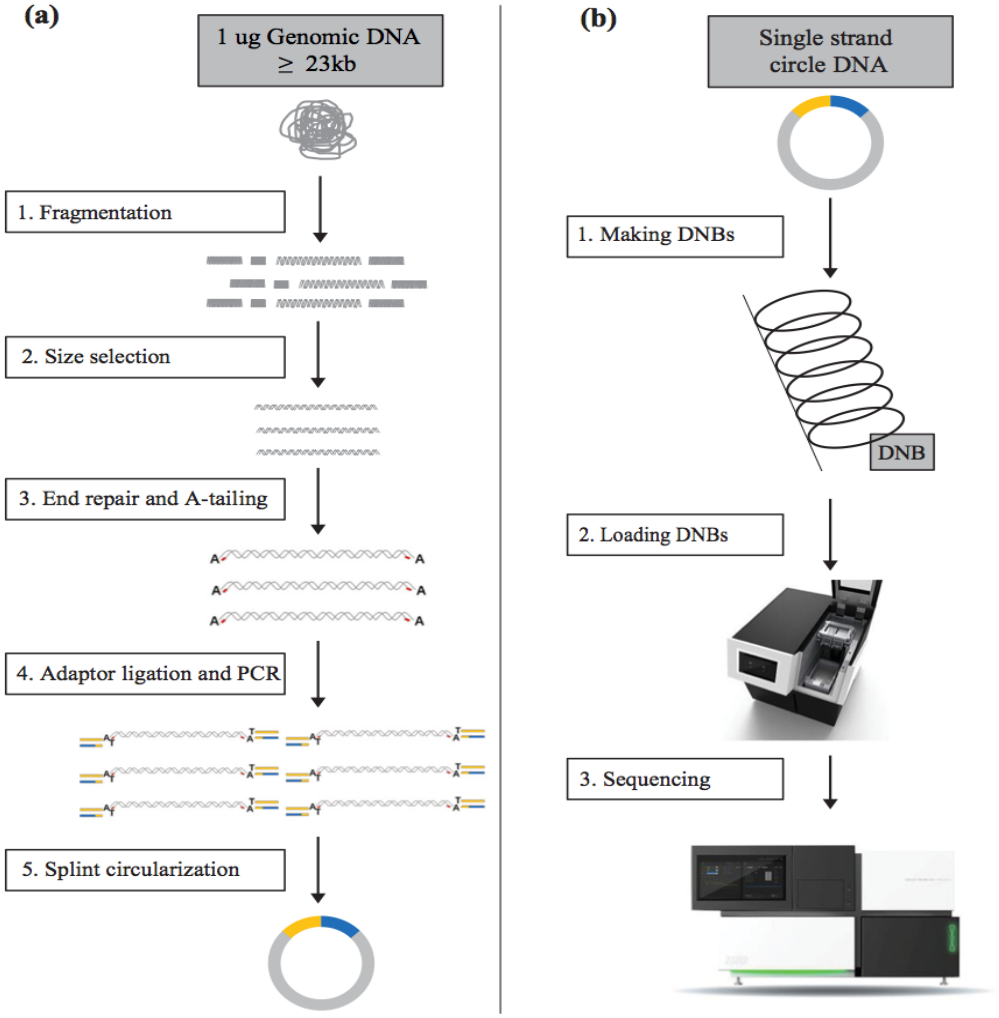

Figure 3: Flowchart of library construction and sequencing.