Evaluating Gencove's Low-Pass Whole Genome Sequencing and analysis platform performance with MGIEasy Fast PCR-FREE FS Library Prep Set

Introduction

The goal of whole genome sequencing (WGS) is to identify genetic variants in the population that influence phenotypes, such as disease susceptibility or traits in the breeding. Historically, this has posed a challenge. The only cost-effective technology for profiling large numbers of individuals has been genotyping arrays, which only measure a small fraction of the genome. Sequencing-based technologies, which enable more comprehensive profiling of genetic variation, remain too expensive for routine use.

In response to this, Gencove developed a Low-Pass WGS (LPWGS) and analysis software platform to make high-throughput and cost-effective whole genome sequencing more accessible and interpretable. Gencove's Low-Pass Sequencing and analysis platform returns more data affords more statistical power than genotyping arrays and is <99% concordant to WGS.

Gencove's platform works as follows:
1. A large numbers of DNA samples are multiplexed in a single lane or run of a sequencer.
2. Samples are sequenced at a low coverage (typically, <1X)
3. FASTQ files are uploaded into the Gencove platform where imputation to a haplotype reference panel generates a VCF file with highly accurate variant calls across the whole genome enabling further downstream analysis.
This application note demonstrates the performance of Gencove LPWGS pipeline when adapted to MGI low-pass whole genomic sequencing workflow using the MGIEasy Fast PCR-FREE FS Library Prep Set and Genetic Sequencer DNBSEQ-T7*. This study was performed by Gencove and MGI.

Methods

Library preparation and sequencing

Libraries were prepared from 200 ng of NA12878 genomic DNA using the MGIEasy Fast PCR-FREE FS Library Prep Set (Catalog No. 940-000021-00, MGI) and the corresponding standard automated workflow on the MGISP-960 High-throughput Automated Sample Preparation System. The resulting libraries were pooled using equal volumes of each library. The pool was then quantified and diluted for circularization and making DNB.

The circularization and DNB making was using the DNBSEQ onestep DNB Make Reagent Kit (OS-DB) (Catalog No. 1000026466, MGI). Sequencing was performed on the Genetic Sequencer DNBSEQ-T7* with 2×100 bp paired-end reads using the DNBSEQ-T7RS High Throughput Sequencing Reagent Kit (FCL PE100)* (Catalog No. 1000028455, MGI).

Data analysis

The data analysis was performed by the Gencove low-pass WGS analysis pipeline. The input file was the FASTQ for a standard sample NA12878 with 5.5Gb of sequence. As part of the standard analysis pipeline, an imputed VCF and aligned BAM files were generated, along with several QC statistics.

Evaluation of the analysis

To deeply evaluate the performance of the Gencove platform, NA12878 high confidence data was divided into repetitive/non-repetitive regions, regions of different GC content, and exonic/intergenic/intronic regions. The sensitivity and precision of LPWGS was evaluated based on the databases.

Database	Linkage
Genome version	GRCh38
NA12878 high confidence data (VCF and BED files)	https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878_HG001/NISTv4.2.1/GRCh38/
NA12878 Low Complex Region	https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/genome-stratifications/v3.0/GRCh38/LowComplexity/
NA12878 Different GC Content	https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/genome-stratifications/v3.0/GRCh38/GCcontent/
Genome Components analysis	http://ftp.ensembl.org/pub/release-105/gtf/homo_sapiens/Homo_sapiens.GRCh38.105.chr.gtf.gz

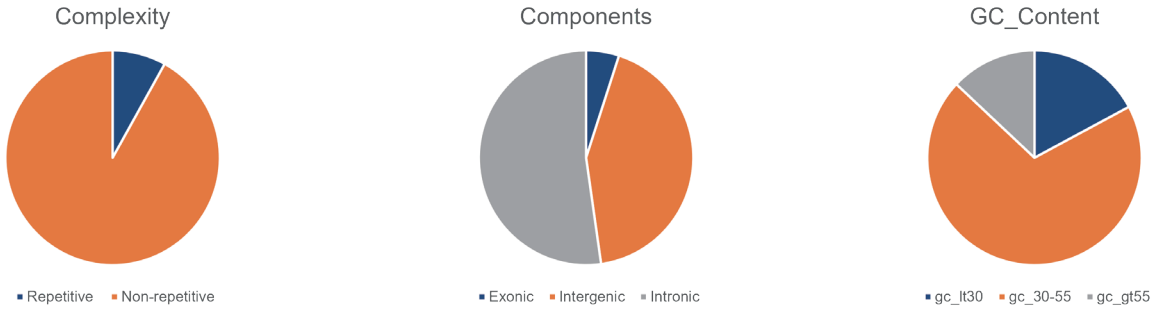

Figure 1. Databases being used for the evaluation.

Results

Alignment to the reference genome

Alignment of raw reads to a reference genome is one of the key steps in NGS data analysis. A good alignment algorithm can map reads onto a reference genome rapidly and accurately. Table 1 shows the QC and mapping statistics of NA12878 LPWGS data against human reference genome hg38 aligned by Gencove platform. The result shows the mapping rate at 99.95%.

Table 1. Alignment results.

Metrics	Value
[Total] Raw Reads (All reads)	55,353,582
[Total] QC Fail reads	0
[Total] Raw Data (Mb)	5535.36
[Total] Paired Reads	55,353,582
[Total] Mapped Reads	55,324,529
[Total] Fraction of Mapped Reads	99.95%
[Total] Fraction of PCR duplicate reads	4.62%

Coverage and Depth

For genome sequencing data, assessing the depth and breadth of DNA sequence coverage is a necessary precursor to variant discovery as depth of coverage drives the power to detect genetic variation, especially in the case of heterozygous sites. Coverage information is critical when detecting copy number variation (CNV) and structural variation (SV). Figure 2 shows the result of coverage and average coverage across different chromosomes. The average depth across chromosomes is about 1X and the average coverage was 77%. Among chromosomes, chromosome 3 had the highest coverage rate (83.17%) with chromosome 22 having the lowest coverage rate (56.54%).

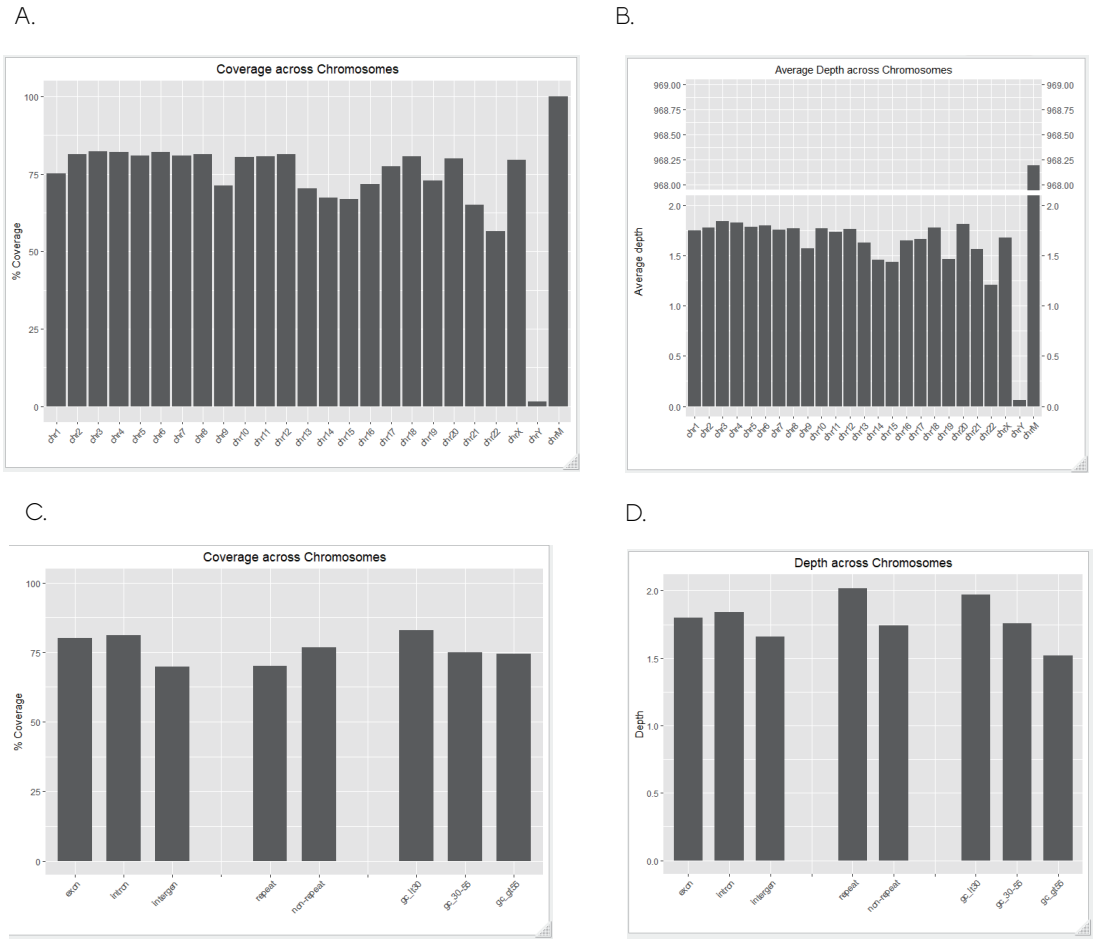

Figure 2. Depth and coverage calculation. A) Depth across the chromosomes. B) Coverage across the chromosomes. C) Coverage for different part of the genome. D) Depth for different part of the genome.

Evaluation of the Precision and Sensitivity

The RTG vcfeval was used for the precision and sensitivity analysis of SNV and Indel. The result shows F-score as 98.79% and 93.56% for SNVs and Indels, respectively. To evaluate the performance of the Gencove platform for different genomic regions and hard-to-detect regions, the F-measure for exon/intron/intergenic region, repetitive/non-repetitive and regions with different GC content were calculated.

The result shows that intronic, intergenic, repetitive, and higher GC-content region have slightly lower F-score value for both SNV and Indel, compare to easier detected regions, as expected. In addition, SNV for all types has F-score exceeding 96% and the maximum percentage over 99%, the same values for Indel are 91% and 96% (criteria for SNV and Indel are 95% and 85%, respectively). These results demonstrate that the Gencove platform can detect variants with significantly high precision, for every region of the genome including hard-to-detect regions.

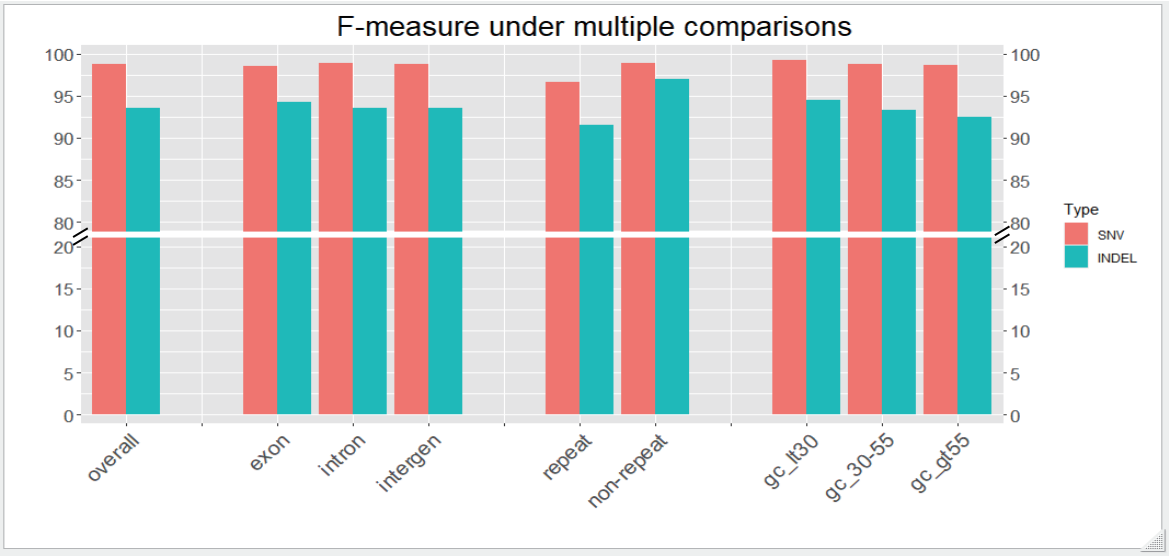

Figure 3. F-measure score calculation.

Conclusion

In conclusion, the MGIEasy Fast PCR-FREE FS Library Prep Set provides a fast, flexible, and automatable library preparation workflow that enables the completion of library preparation for 96 samples in 90 minutes and can support flexible throughput for multiplexing 8-384 samples. Together with DNBSEQ sequencing platform and the Gencove platform, it presents a cost-effective turnkey solution for LPWGS.

The Gencove platform is a user-friendly cloud-based software-as-a-service, which requires little bioinformatic support. The platform is available by subscription via web app at gencove.com, or can be interacted with in an automated manner via the API and command line interface. The coverage and F-measure analysis performed above show that the Gencove platform delivers highly accurate and precise variant detection, which enables many downstream applications, including the analysis of agriculture and breeding.

In sum, the Gencove platform can perform accurate alignment and variant calling analysis and is compatible with MGI's WGS workflow for the genotyping of various species.

Ordering Information

Procedure	Vendor	Kit	Catalog Number
Library Preparation	MGI	MGIEasy Fast PCR-FREE FS Library Prep Set	940-000021-00
Circularization and DNB making	MGI	DNBSEQ onestep DNB Make Reagent Kit (OS-DB)*	1000026466
Sequencing	MGI	DNBSEQ-T7RS* High Throughput Sequencing Reagent Kit (FCL PE100)	1000028455

Reference

1. Gencove data analysis configurations https://docs.gencove.com/main/data-analysis-configurations/

MGI Tech Co.,Ltd | Building 11, Beishan Industrial Zone, Yantian District, Shenzhen, CHINA, 518083
en.mgi-tech.com | MGI-service@mgi-tech.com

*Unless otherwise informed, StandardMPS and CoolMPS sequencing reagents, and sequencers for use with such reagents are not available in Germany, USA, Spain, UK, Hong Kong, Sweden, Belgium, Italy, Finland, Czech Republic, Switzerland, Portugal, Austria and Romania.

FOR RESEARCH USE ONLY. NOT FOR USE IN DIAGNOSTIC PROCEDURES.