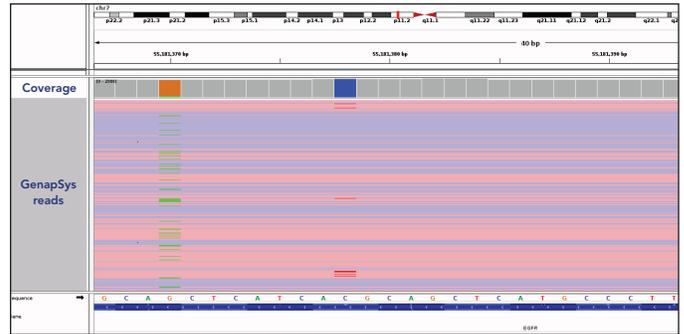# GenapSys™ Sequencing Platform:

Accurate Variant Calling at Lower Coverages

- Higher accuracy sequencing enables superior variant calling at lower coverage
- Lower mismatch error rates compared to established sequencing technologies
- Low price per run and low price per sample
- Up to 2 Gb of highly accurate DNA sequence data per run

Fig. 1 IGV visualization of two EGFR variants called following sequencing and analysis of a Pan Cancer panel library for the Horizon Oncospan Reference Standard. Variants shown are EGFR: p.Q787Q (left) with expected allele frequency of 15% and EGFR: p.T790M (right) with expected allele frequency of 1%.

Introduction

Genetic medicine relies upon the ability to confidently identify differences between a patient's genome and the reference. Next-Generation Sequencing (NGS) has become the tool of choice for discovering and identifying novel single nucleotide polymorphisms (SNPs) and call known SNPs in genomic samples. Accurately calling these SNPs requires high-quality sequencing data, high coverage, and a thorough bioinformatics approach to identify the SNPs in a statistically relevant Manner. Variant calling can be used for carrier screening, identifying rare diseases, and provides the foundation for personalized medicine.

GenapSys™ has developed a novel, scalable, low cost NGS platform that is capable of generating high quality data necessary for variant calling applications. Using whole exome sequencing and targeted cancer panel sequencing, we demonstrate higher accuracy compared to another sequencing technology, especially at lower coverages, highlighting superior variant calling and lower mismatch error rates of the GenapSys Platform.
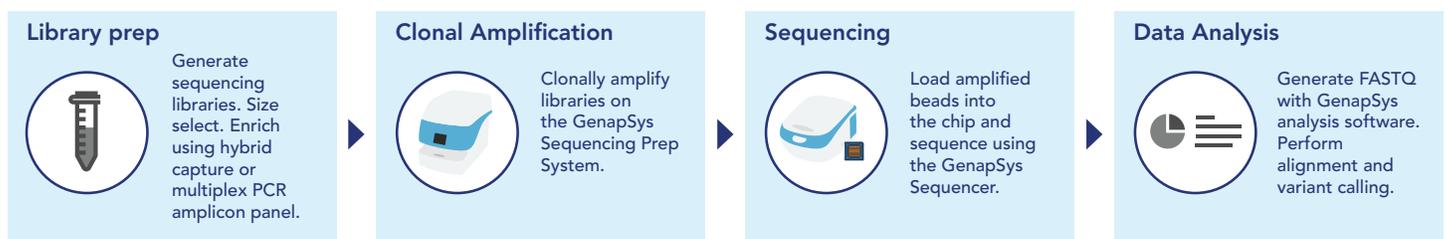
Technology

The Genapsys™ Sequencer employs a novel electrical detection method that is capable of generating highly accurate DNA sequence information. With a CMOS based detector, simple fluidics, and low computational requirements, the GenapSys instrument is small, affordable, and accessible even to novice genomic scientists. Inside the sequencing chip are millions of individual sensors, each loaded with a single bead coated in thousands of clonal copies of a particular DNA sequence. Individual nucleotides are flowed across the chip in succession and successful incorporation is detected by changes in impedance as the complementary DNA strand grows.

Experimental Methods

Germline variant calling was performed on data from Whole Exome Sequencing (WES) of the Genome in a Bottle (GIAB)

Streamlined targeted sequencing workflow on the GenapSys Sequencing Platform

Library prep
Generate sequencing libraries. Size select. Enrich using hybrid capture or multiplex PCR amplicon panel.

Clonal Amplification
Clonally amplify libraries on the GenapSys Sequencing Prep System.

Sequencing
Load amplified beads into the chip and sequence using the GenapSys Sequencer.

Data Analysis
Generate FASTQ with GenapSys analysis software. Perform alignment and variant calling.

sample NA12878. Somatic variant calling was performed on sequencing data from a Pan Cancer panel library generated using the Oncospan Reference Standard. The Oncospan Reference Standard (HD827) was ordered from Horizon Discovery which is composed of DNA from multiple cell lines, and contains 386 variants across 152 key cancer genes. The allele frequencies range from 1% to 100% and frequencies of 25 of these variants are confirmed with ddPCR. Genomic libraries were generated following mechanical fragmentation, adapter ligation, size selection with a median insert size of ~200 bp and PCR. Hybrid capture-based enrichment was done using the IDT Exome Research panel (39 Mb region, 19,396 genes) for the Whole Exome Sequencing (WES) library, or using the IDT Pan Cancer panel V1.5 (800 kb region, 127 genes) for the Cancer Panel library. Post-capture PCR was performed to generate the final libraries. Individual library molecules were clonally amplified onto beads using the
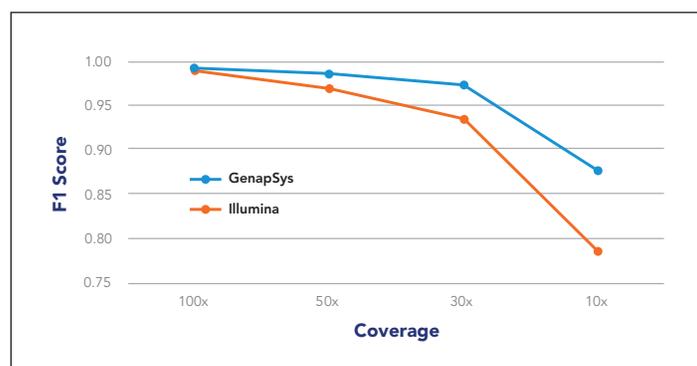
Fig. 2 Variant Calling accuracy following whole exome sequencing of the GIAB NA12878 sample. GenapSys has a lower mismatch error rate compared to Illumina, resulting in improved F1-scores at lower mean coverages.

GenapSys Sequencing Prep System, and enriched amplified beads were loaded onto the G3 chip for sequencing on the GenapSys Sequencer. Illumina cancer panel libraries were generated and sequenced on an Illumina NextSeq instrument, to compare sequencing performance.

Sequencing reads were aligned to the hg38 reference genome using BWA MEM (v 0.7.17). DeepVariant (v 0.9.0) was used for germline variant calling based on pre-trained model (Illumina) and custom-trained model (GenapSys) respectively with a cutoff allele frequency of 12%. The GIAB high confidence variants within the IDT xGen Exome panel were used as the ground truth. Vardict (v1.7.0) was used for Somatic variant calling (both Illumina and GenapSys) with a minimum allele frequency of 2%. The ground truth for high confidence somatic variants was generated based on the common variants detected in high coverage Illumina and GenapSys

data within the IDT xGen Pan Cancer Panel coding exons. High confidence variants down to 2% allele frequencies were considered for the analysis.

Results

A Whole Exome library was generated from the GIAB (Genome In A Bottle) NA12878 sample and sequencing was done on the GenapSys Sequencer and Illumina NextSeq platform. Sequencing data was downsampled to different mean coverage levels: 100x, 50x, 30x, and 10x. Variants were called using the same bioinformatics workflow for both data sets, as described in the Methods section. F1 scores were calculated by comparing the variant calls to the high confidence variants from the GIAB. As shown in Fig. 2, the F1 scores for SNP calls are higher for GenapSys as compared to Illumina, with the improved variant calling (higher F1 scores)
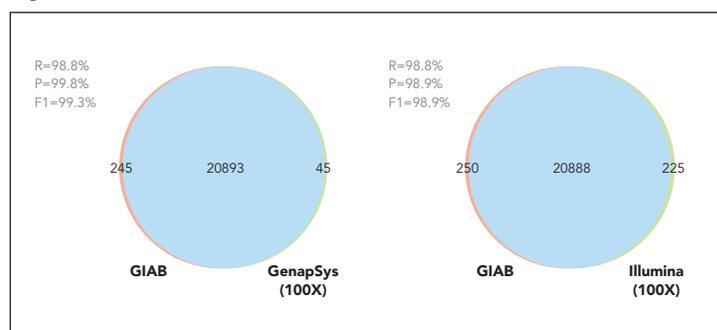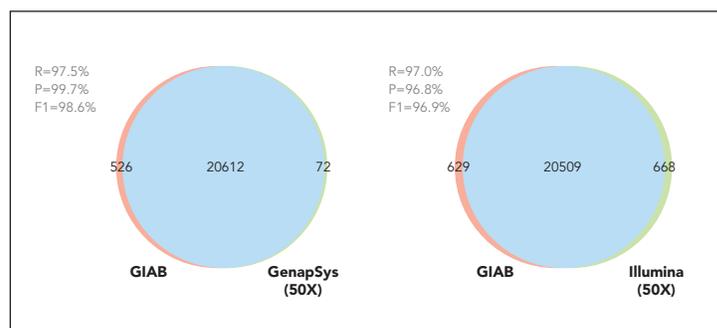
Fig. 3A

Fig. 3B

Fig. 3 Improved germline variant calling by the GenapSys platform at both (A) 100x and (B) 50x mean coverage. This is demonstrated by lower false positive and false negative variant calls as compared to Illumina variant calls.

demonstrated at lower mean coverages, such as, 50x, 30x and 10x. This is the result of lower mismatch error rates in the GenapSys sequencing platform. A detailed analysis of the variant calls reveals that the GenapSys platform shows lower false positive and lower false negative variant calls at both 100x mean coverage and 50x mean coverage (Fig. 3).

To learn more about the GenapSys Sequencing Platform, visit GenapSys.com

Somatic variant calling performance, especially for low frequency variants, was demonstrated by sequencing a Pan Cancer panel library generated from the Horizon Discovery Oncospan Reference Standard. Variant calls were generated using Vardict based on data from both the GenapSys and the Illumina NextSeq platform. A 'Ground Truth' set of high confidence variants was generated as the common variants observed at >1000x coverage in each platform. To compare variant calling at lower coverages, the sequencing data was then downsampled to 1000x, 500x, 200x and 100x coverage. As shown in (Fig. 4), the F1-scores of both platforms are comparable at high coverages (1000x), but the GenapSys platform shows consistently high F1 scores at low coverages, even as low as 200x and 100x. Illumina data, on the other hand showed a significant drop in F1-score accuracy, as the

mean coverage dropped to 200x or 100x. This highlights the low mismatch error rate of GenapSys sequencing data, which leads to high confidence variant calling at lower coverages.

Summary

The performance of the GenapSys system was evaluated for germline and somatic variant calling using whole exome sequencing and cancer panel library sequencing. The superior SNV calling capability of the GenapSys platform was demonstrated by higher accuracy at a lower coverage compared to the Illumina technology. The GenapSys solution delivers high-quality NGS data in a compact, accessible and cost-effective platform.
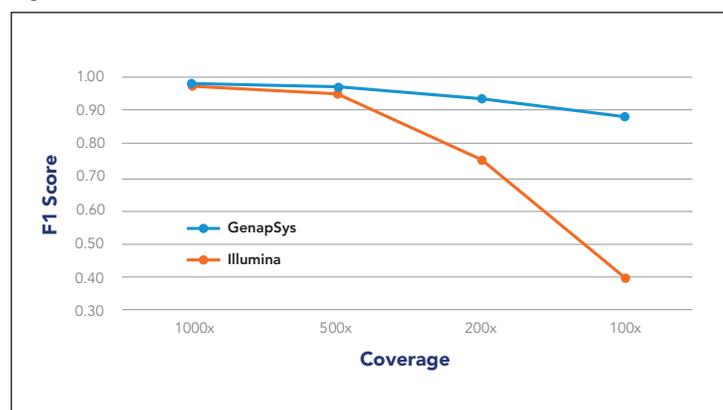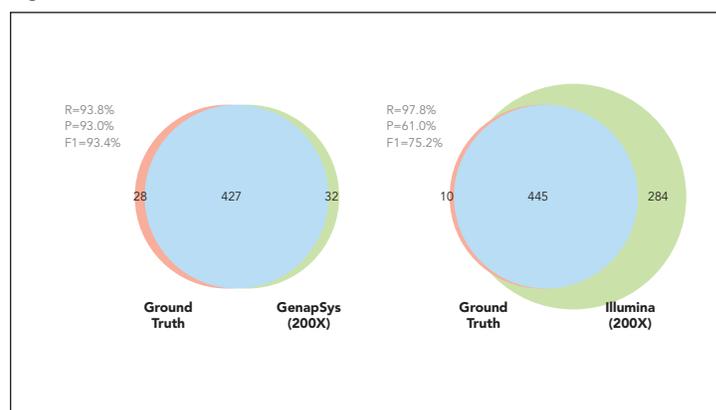
Fig. 4A

Fig. 4B

Fig. 4 Somatic variant calling using the GenapSys and Illumina platforms. (A) Lower mismatch error rates of the GenapSys platform as compared to Illumina, lead to higher F1-scores of variant calling at different coverages. (B) At 200x coverage, the GenapSys platform demonstrates a significantly higher F1-score as compared to Illumina, when compared with a 'Ground Truth' set of high coverage variants. Data was generated with a Pan Cancer panel library using the Horizon Oncospan Reference Standard.

To learn more about the GenapSys Sequencing Platform, visit GenapSys.com