Improvements to the GenapSys Sequencing Platform to Enable Low Cost, Highly Accurate DNA Sequencing

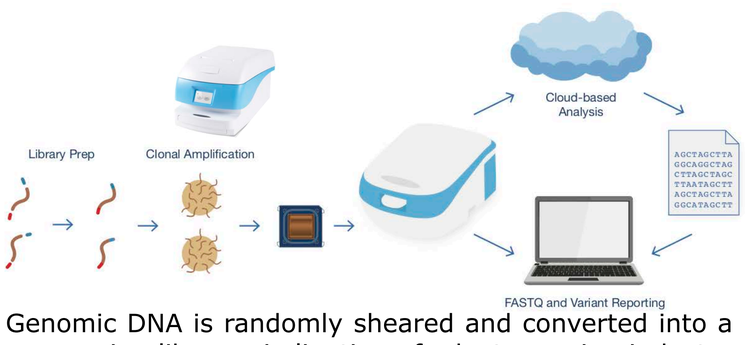
Tyson A. Clark, Kosar B. Parizi, Hamid Rategh, Caterina Schweidenback, Linda Hsie, Meysam Rezaei Barmi, Maryam Jouzi, Seth Stern, Nicolas Monier, Saurabh Paliwal, Mohammad Fallahi, and Hesaam Esfandyarpour

Abstract

Advances in sequencing technologies have revolutionized genomic medicine and decreased the cost of generating data in large quantities. However, achieving the lowest cost per base typically requires a large initial capital investment. Furthermore, high run costs on these large instruments necessitates batching large numbers of samples to keep per sample costs low. In contrast, the GenapSys Sequencing Platform employs a novel, electrical-based detection method that produces highly accurate sequence data on a small, flexible, and affordable instrument. The detection method, which utilizes CMOS chips, enables the system to be compact, accessible, and affordable. Using its 16M sensor chip, the platform is capable of generating up to 2 Gb of high-quality nucleic acid sequence in a single run, and we routinely generate sequence data that >80% of bases exceeds Q30 (i.e., raw accuracy 99.9%) with average read lengths of 150 bp.

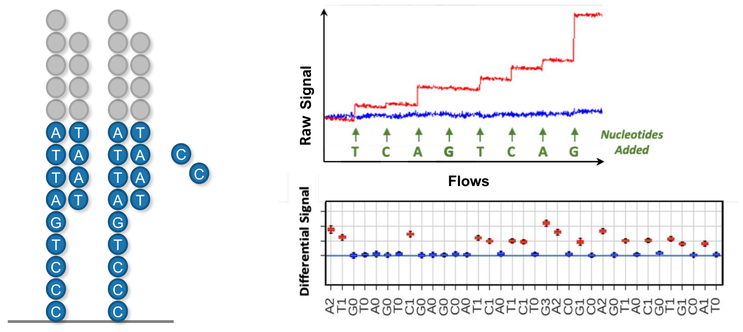
Here, we describe further innovations of the sequencing system that improve data quality, throughput, speed, and ease of use. Next generation sequencing chemistries and an updated base calling algorithm have improved raw read accuracy and homopolymer calling while decreasing total run time. We demonstrate the performance and functionality of the GenapSys sequencing technology using microbial and human genomes as well as using multiplexed samples. We also highlight the performance of this platform for germline and somatic variant detection in standard samples.

GenapSys Sequencing Workflow



Genomic DNA is randomly sheared and converted into a sequencing library via ligation of adapters using industry standard methods. Individual library molecules are clonally amplified onto beads and loaded into the sequencing chip. Automated sequencing is carried out on the GenapSys Sequencer which utilizes cloud-based analysis methods that ultimately deliver FASTQ and variant reporting files.

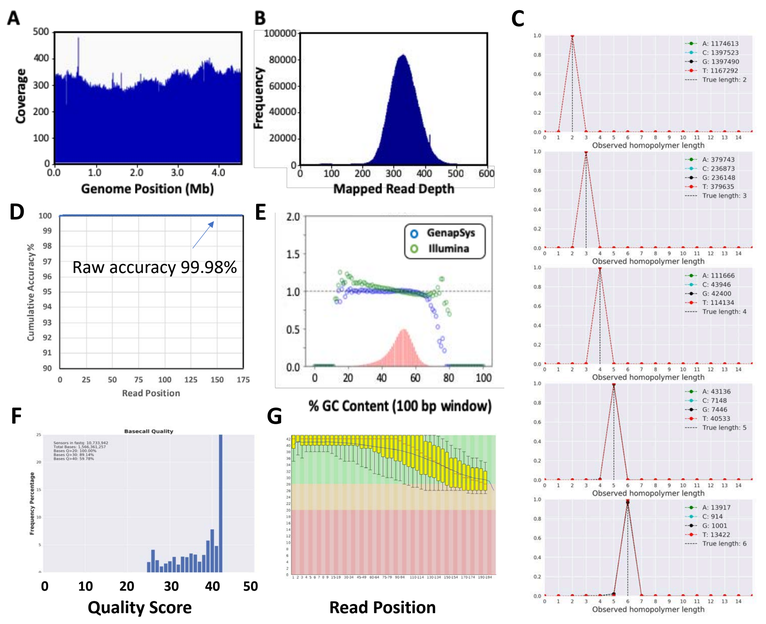
The CMOS sequencing chip contains millions of individual sensors, which are each loaded with a clonally amplified bead. The electrodes in each sensor are capable of measuring minute changes in impedance when nucleotides are incorporated opposite the bead-bound templates.



Nucleotides are injected one base at a time. When a nucleotide is incorporated, the measured impedance value of that sensor will jump, creating a graph that resembles a staircase. The magnitude of the differential signal correlates with the number of incorporated nucleotides.

Exceptional Sequencing Performance

Microbial Genome Sequencing Statistics

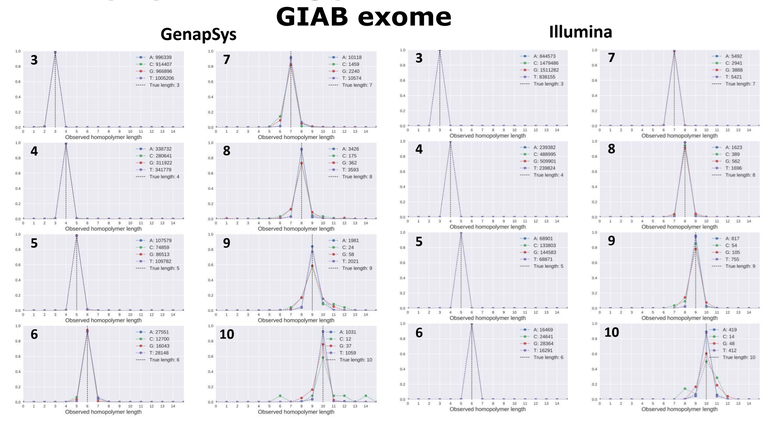


(A) Coverage plot across the *E. coli* genome. (B) Histogram of mapped read depth (C) Homopolymer calling performance (D) Average cumulative sequencing accuracy by read position. (E) Normalized coverage: Coverage is binned by percent GC content in a 100 bp window. The red histogram represents the abundance of each GC content bin in the genome. (F) Histogram of the basecall Qscores. (G) Basecall quality as a function of read position. (H) Summary alignment statistics.

H

Number of Reads	10.7M
Number of Bases	1.6G
Number of Mapped Reads	10.1M
Average Mapped Read Length	~150 bp
Raw Accuracy at Position 150	99.9%
Percentage of Bases with Q>20	99.6%
Percentage of Bases with Q>30	86.0%
Genome Size	4.69 Mb
Average Coverage	315x

Homopolymer calling performance for NA12878 GIAB exome



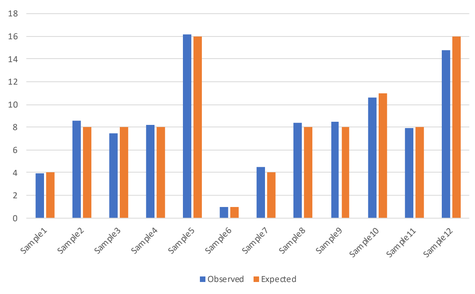
Using data from a human exome sequencing run, base calls from homopolymer regions (3-10 nt) were compared to the hg38 reference for each of the four different bases. Performance is shown relative to the same library sequenced with Illumina.

Sample Multiplexing

Sample demultiplexing accuracy >99.99%

Sample multiplexing in NGS is important for optimal throughput and cost per sample. The GenapSys platform is capable of multiplexing using standard single and dual index kits, including 96 sample kits. To characterize multiplexing accuracy, we designed 12 unique templates (spanning low, medium and high GC inserts) with different combinations of 8bp dual indexes (NEBNext Multiplex Oligo Set 1). Dual index based sample predictions were compared with the insert sequence to calculate demux accuracy.

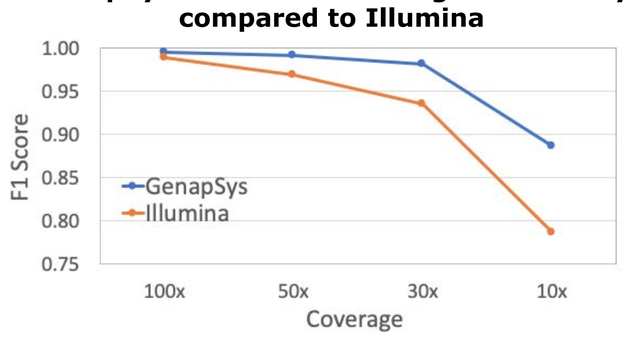
12 dual indexed templates were combined with variable ratios (16-fold variation) in a library. Sample reads were identified based on indexes with **>99.99%** accuracy. Observed abundances closely match expected sample abundances.



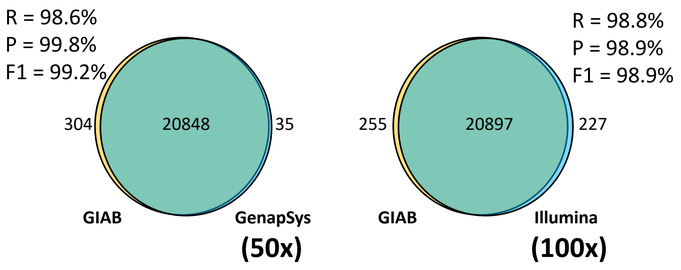
Gold Standard Variant Calling

GenapSys SNP calls show higher accuracy compared to Illumina

A

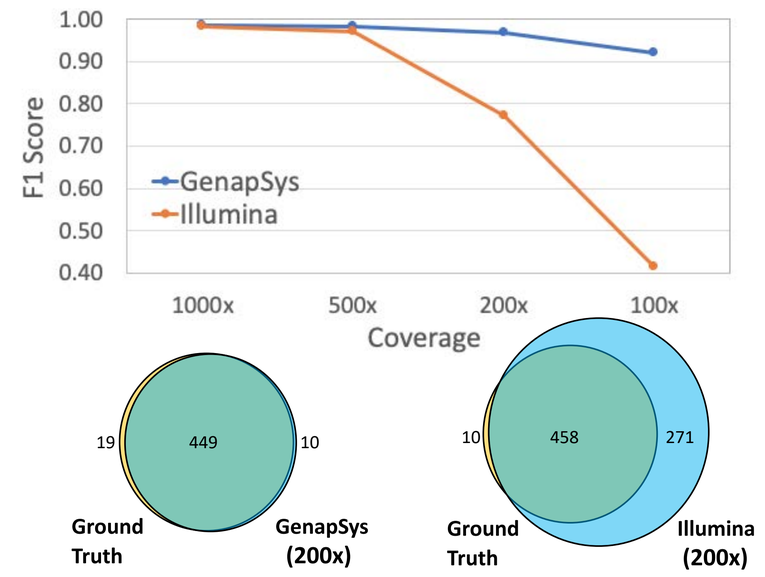


B



A The NA12878 WES library was sequenced on both the GenapSys Sequencer and Illumina NextSeq. Single Nucleotide Polymorphisms (SNPs) relative to the hg38 reference were called using DeepVariant trained on respective technologies' data. **GenapSys SNV calls show higher F1-score accuracy as compared to Illumina**, especially at lower coverages, highlighting lower mismatch error rates in GenapSys sequencing data. **B** GenapSys results show better concordance with the GIAB high-confidence variants (IDT xGen Exome panel) at half the coverage compared to Illumina.

GenapSys SNV calls show higher accuracy compared to Illumina on Hybrid Capture Cancer Panel



The Horizon HD827 Oncospan reference DNA standard was used to generate a library using the IDT xGen Pan Cancer panel v1.5 and sequenced on both the GenapSys Sequencer and Illumina. Single Nucleotide Variants (SNVs) down to 2% were called using VarDict. Ground truth vcf file was generated based on the common variants detected in high coverage Illumina and GenapSys sequencing data. At 200x, the Illumina data shows 271 false positive calls compared to GenapSys's 10. Higher accuracy in GenapSys data, especially at lower coverages, highlights **lower mismatch error**.

Summary

The GenapSys Sequencing Platform offers:

- Low price per run and low price per sample
- Highly accurate sequencing data with 80% >Q30
- Highly accurate demultiplexing capability
- High confidence SNP and SNV variant calling with a fraction of coverage compared to established sequencing technologies

www.genapsys.com