LightSpeed – enabling affordable genome sequencing analysis at scale

Whole genome sequencing (WGS) and whole exome sequencing (WES) strategies have entered mainstream clinical- and population-genomics. Yet, the required data analysis tasks remain a bottleneck in terms of bioinformatics expertise, execution speed and operation costs.

QIAGEN® CLC Genomics Workbench Premium is a user-friendly platform for next-generation sequencing (NGS) secondary analysis with a broad menu of applications and visualization options. Here, we introduce the platform's new QIAGEN CLC LightSpeed Module, which empowers genomics laboratories to perform hereditary secondary analysis with high accuracy at unprecedented runtimes. What's more, it enables this without requirements for expensive and energy-demanding specialized hardware, such as Field Programmable Gate Arrays (FPGAs) or Graphical Processing Units (GPUs).

LightSpeed can process FASTQ files to produce variant call format (VCF) files containing single nucleotide variant (SNV), insertion–deletion mutation (InDel) and structural variant (SV) calls. LightSpeed performs quality and adapter trimming, read mapping, deduplication, local realignment, quality control and germline variant calling. Down-sampling strategies are not used by LightSpeed.

Depending on local IT policies, hardware configurations and Internet access, CLC LightSpeed is deployable using local computers or Amazon Web Services (AWS®) cloud. Highly-scalable server and high-performance-computing–grid (HPC) deployments are also supported for on-premise parallelization of LightSpeed.

In this application note, we present benchmark results on the performance of LightSpeed with regards to speed, costs using AWS and accuracy on gold standard WGS and WES datasets. We used three deployment scenarios for the study: Laptop, workstation and cloud.

Benchmark study

Datasets

Benchmarking for WGS was performed with publicly available HG001 and HG002 datasets and the associated truth sets provided by the Genome in a Bottle (GIAB) Consortium (Table 1). Where appropriate, FASTQ files were down-sampled for apples-to-apples comparison to published benchmarks of other accelerated FASTQ-to-VCF pipelines. WES was benchmarked on representative samples produced by QIAGEN and GIAB (Table 1).

Hardware specifications

AWS: The AWS-EC2 c6id.32xlarge instance with 244 GB RAM (100 GB reserved for Java) was used at spot price ($1.2771/hour, Ohio location, December 14, 2022). The time required for transferring data to the EC2 instance was approximately 6 minutes for WGS samples and approximately 20 seconds for WES samples.

Workstation: 2x Intel Xeon 6238R (28 core, 2.2GHz base, 4.0GHz boost) processors with 192GB RAM.

Laptop: MacBook Pro 2021 with an M1 Pro processor and 32GB RAM.

Execution time, cloud computing costs, energy consumption and carbon footprint

LightSpeed can be deployed as part of QIAGEN CLC Genomics Workbench Premium either on-premise or in the cloud using AWS. We benchmarked LightSpeed execution times using the AWS cloud and monitored the associated AWS EC2 instance costs (Figure 1, Figure 3) and compared them to data from vendors of hardware-accelerated pipelines.

Next, we benchmarked LightSpeed execution times using a workstation and a laptop (Figure 2). The hardware-accelerated pipelines from Vendor 1 and Vendor 2 are incompatible with generic workstation and laptop

Table 1. Characteristics of datasets used in this benchmark.

Data type	Sample	Details	Coverage	Reads (M)	Size (Gb)	Source
WGS	HG001	-	48x	1086	161	PrecisionFDA Truth Challenge 2016
WGS	HG002	-	37x	830	125	PrecisionFDA Truth Challenge V2 2020
WES	HG001	QIAseq Exome	50x	36	5	QIAGEN data
WES	HG001	Garvan; Nextera Exome	49x	82	8	GIAB

All data are FASTQ files produced by Illumina sequencing machines. Coverage was determined using the QIAGEN CLC Genomics platform.

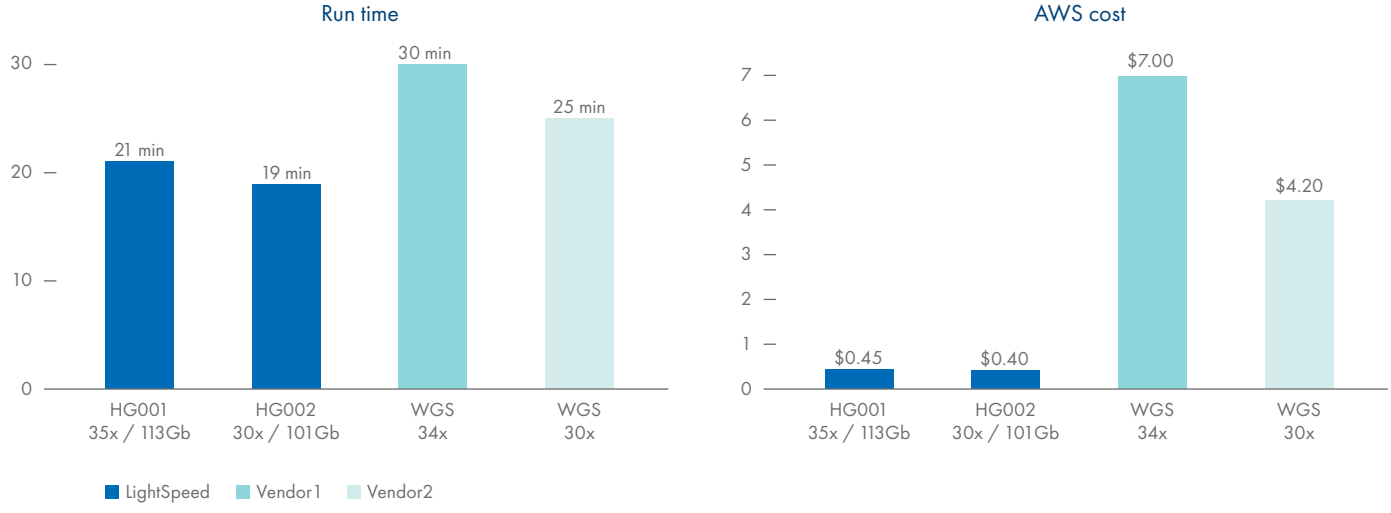

Figure 1. Execution times and associated costs for executing WGS analysis using LightSpeed on AWS compared to other vendor's accelerated FASTQ-to-VCF pipelines. HG001 was down-sampled to 34x coverage and HG002 was down-sampled to 30x coverage for fair comparison with published data from Vendor 1 and 2. The time shown is CPU time and does not include FASTQ file transfer from S3 bucket to the AWS EC2 instance. The data for Vendor 1 and Vendor 2 was retrieved from the vendor's specifications at their corresponding homepages (accessed December 14, 2022).

deployment, and thus it was not possible to include these in the comparison. LightSpeed is the only software able to perform 34x coverage WGS hereditary analysis on laptop computers within a reasonable timeframe.

We also monitored the total power consumption of the workstation when executing LightSpeed on 34x WGS data. The energy consumption was 0.190 kWh. This corresponds to a CO_2 equivalent of only 27g CO_2 (emission data from the Danish Energy Agency). Hence, LightSpeed has the smallest carbon footprint currently available for secondary NGS hereditary analysis.

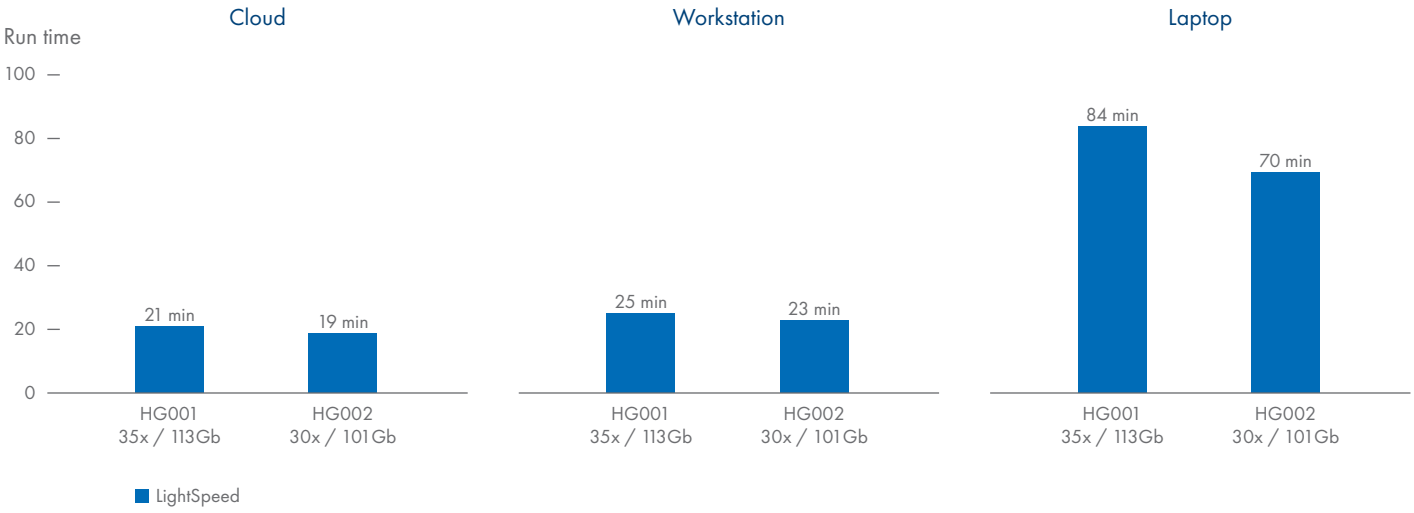

Figure 2. Execution (CPU) times of LightSpeed deployment in the cloud, a workstation and a laptop using WGS samples.

Figure 3. Execution (CPU) times and associated costs for executing LightSpeed using AWS for WES samples compared to another vendor's accelerated FASTQ-to-VCF pipeline.

Accuracy

We evaluated the accuracy of LightSpeed variant calling for HG001 and HG002 datasets against the corresponding truth sets provided by the GIAB Consortium and achieved 99% accuracy for more than 90% of the genome (Table 2).

Table 2. Accuracy of LightSpeed on WGS datasets.

Sample	Reference	SNVs			InDels			Combined
		F1 score	Sensitivity	Precision	F1 score	Sensitivity	Precision	F1 score
HG001	hg19	99.74%	99.63%	99.84%	98.33%	97.88%	98.78%	99.11%
HG002	hg19	99.01%	98.54%	99.48%	98.15%	97.75%	98.56%	98.89%
HG001	hg38	99.18%	98.92%	99.44%	98.04%	97.48%	98.62%	99.03%
HG002	hg38	98.87%	98.47%	99.26%	98.08%	97.69%	98.47%	98.76%

The genome assembly references used were hs37d5 (hg19) and hg38_no_alt_analysis_set (hg38). Accuracy was calculated on output from hap.py with vcfeval engine using the gold standard provided by GIAB (v4.2.1).

Conclusion

LightSpeed is the fastest germline calling FASTQ-to-VCF pipeline currently available and provides the most flexible deployment options for both on-premise and in the cloud use, as it is not dependent on specialized acceleration hardware such as FPGAs or GPUs. This speed enables instant scaling to match any sequencing throughput using only commodity hardware. Together, speed and deployment flexibility ensure the cheapest and most energy-efficient pipeline currently available for hereditary WGS and WES secondary data analysis. Importantly, improvements in computation time and costs do not compromise the accuracy of variant calling (Table 2).

For up-to-date licensing information and product-specific disclaimers, see the respective QIAGEN kit handbook or user manual. QIAGEN kit handbooks and user manuals are available at **www.qiagen.com** or can be requested from QIAGEN Technical Services or your local distributor.