Application Note

Whole Exome Sequencing Benchmark

Francesco Lescai, Ph.D., Leif Schauser, Ph.D. and Jonathan Jacobs, Ph.D.

QIAGEN Bioinformatics – Aarhus, Denmark

Summary

We have benchmarked QIAGEN® Bioinformatics' CLC Genomics Server (GxS) with GATK 4.0 (Broad Institute, MA) (1), with and without the use of Apache Spark™ in different computing environments, and compared runtime and performance when analyzing 100X coverage of human whole-exome data. We also compared the sensitivity, precision and overall accuracy of variant calling performed by a workflow developed with CLC to both GATK and Strelka2 (2).

When comparing runtime, CLC GxS performs competitively to GATK with Apache Spark on a 40-core server and Intel®-Optane™ SSD technology (60 versus 58 minutes, respectively). The CLC Genomics Server performs better overall than GATK without Apache Spark.

In terms of sensitivity, the CLC GxS outperforms the other platforms in both SNP and INDEL calling (99.6% and 91.9%, respectively, on PASS variants). GATK, however, scores highest for SNP calling precision (99.9% versus 99.1% by CLC) and Strelka2 performs higher for INDELs precision (95.2% versus 92.3% by CLC). In terms of accuracy, CLC GxS performs best with SNPs (99.4% versus 97.4% and 97.3% by GATK and Strelka2, respectively), while Strelka2 performs best for INDELs (93.3% versus 92.3% by CLC).

Introduction

Researchers frequently question the speed and accuracy of a bioinformatics analysis pipeline they depend on. The data and results presented in this benchmark in fact originate from a customer request to compare the speed and accuracy of the CLC Genomics Server (CLC GxS) and CLC Genomics Workbench (GxWB) to that of GATK, a popular pipeline for variant calling in human biomedical genomics data.

In this Application Note, we present a comparative analysis of both runtime and performance of the CLC Genomics Server and GATK v.4, both with and without the use of an Apache Spark framework. Apache Spark is considered an environment for in-memory distributed computing designed to accelerate "big data" analytics and processing. Since the most recent release of GATK replaced multi-threading processing with the use of Apache Spark, it seemed necessary to conduct this benchmark on a GATK + Apache Spark combination, as well. Also, we tested the software packages in three different computing environments, in order to describe the most common situations a user could face.

We also performed a comparison of the sensitivity, precision and resulting accuracy (F1 score) of the variants called by each pipeline. For this specific benchmark, we have also included Strelka2, a recently released high-performance variant caller.

To to execute this test, we used data from the Genome in a Bottle (GiAB) Consortium (3) sample (NA12878), used widely for accuracy, testing and pipeline comparisons (4, 5). For the study presented here, the GiAB sample was sequenced in-house at QIAGEN on an Illumina® sequencer, utilizing capture technology from Twist Bioscience™ (San Francisco, CA, USA). This capture technology was chosen based on the need for uniformity of coverage in ensuring higher quality results (6).

In general, it is essential to remember that speed is not the only important factor when designing an analytical pipeline, and that speed is not constituted by the runtime alone. Kawalia and colleagues offer a series of general considerations when describing their workflow (7).

Computing environments

For this benchmark we used three different computing environments, a standard reference sample throughout and repeated comparable workflows in all combinations three times each to account for variances in runtime.

Table 1 is a summary of the three environments. The Intel-Optane system is a high-performance server with 40 cores (80 threads) and a dedicated high-performance Solid State Disk (SSD) technology. This is a solution likely to accelerate algorithms capable of multi-threaded analysis which will also scale with threads, while minimizing any bottleneck due to disk Input/Output (I/O).

In contrast, and to illustrate the benefit of high-performance I/O while maintaining the same computing environment, we also used the Intel-Optane server but we moved the data to a network file system – thereby intentionally creating an I/O bottleneck.

In addition, to test the performance in a portable environment we conducted the benchmark on a MacBook Pro laptop.

Table 1. Computing environments

	Intel-Optane	Intel-network	Laptop
CPUs	2 x Intel Xeon Gold 6148 2.40GHz	2 x Intel Xeon Gold 6148 2.40GHz	Intel Core i7
Cores	40 cores/ 80 threads	40 cores/ 80 threads	4 cores
RAM	192 GB RAM	192 GB RAM	16 GB RAM
Disk	Intel-Optane technology SSD	Network Storage	500 GB Mac SSD

The data

We used exome sequencing of a NA12878 reference sample from the Genome in a Bottle Consortium. The sequencing data were generated in-house with Illumina HiSeq®, and the library captured with Twist Bioscience technology. The resulting dataset consisted of 77.5 M reads, 76 M of which were mapped at an average coverage of 107.6x with CLC.

Software

The following software versions were used for the comparison: The Genome Analysis Toolkit (GATK) v4.0.1.1; Apache Spark v.2.2.3; Strelka v.2.9; CLC Genomics Server v11.0 and CLC Command Line Tools v6.0.

To evaluate the accuracy of each pipeline, Hap.py v0.3.10 (8) was used. The analyses and plotting of these benchmarking results were primarily carried out using RStudio with R version 3.5.1 (2018-07-02).

Runtime benchmark

The analysis was repeated three times consecutively with each software to account for the variance in the execution. It is important to highlight that the CLC workflow includes steps not present in the GATK workflow (Figure 1), namely a structural variant analysis step and an optimized local realignment step. A local realignment is performed by Haplotype Caller, as part of haplotype-based genotype

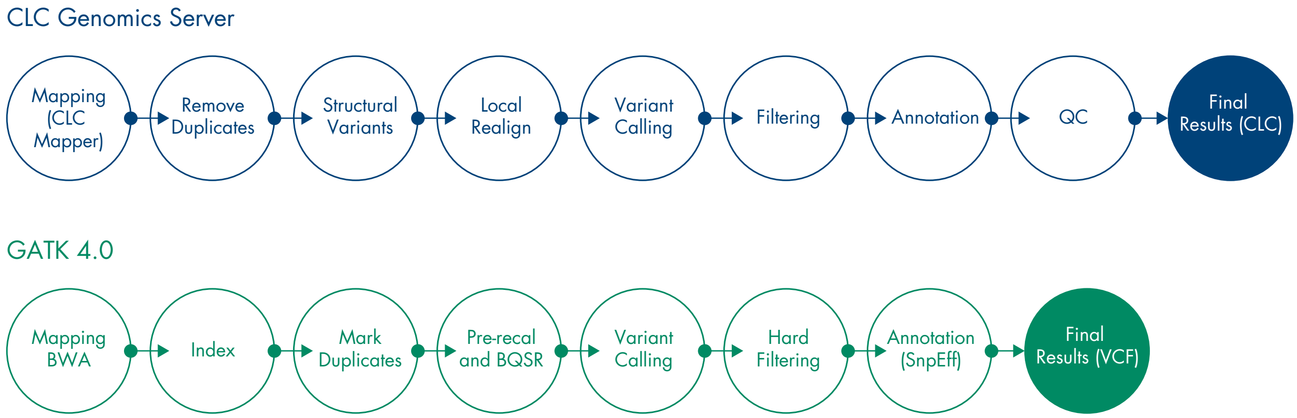


Figure 1. Comparison of the workflows.

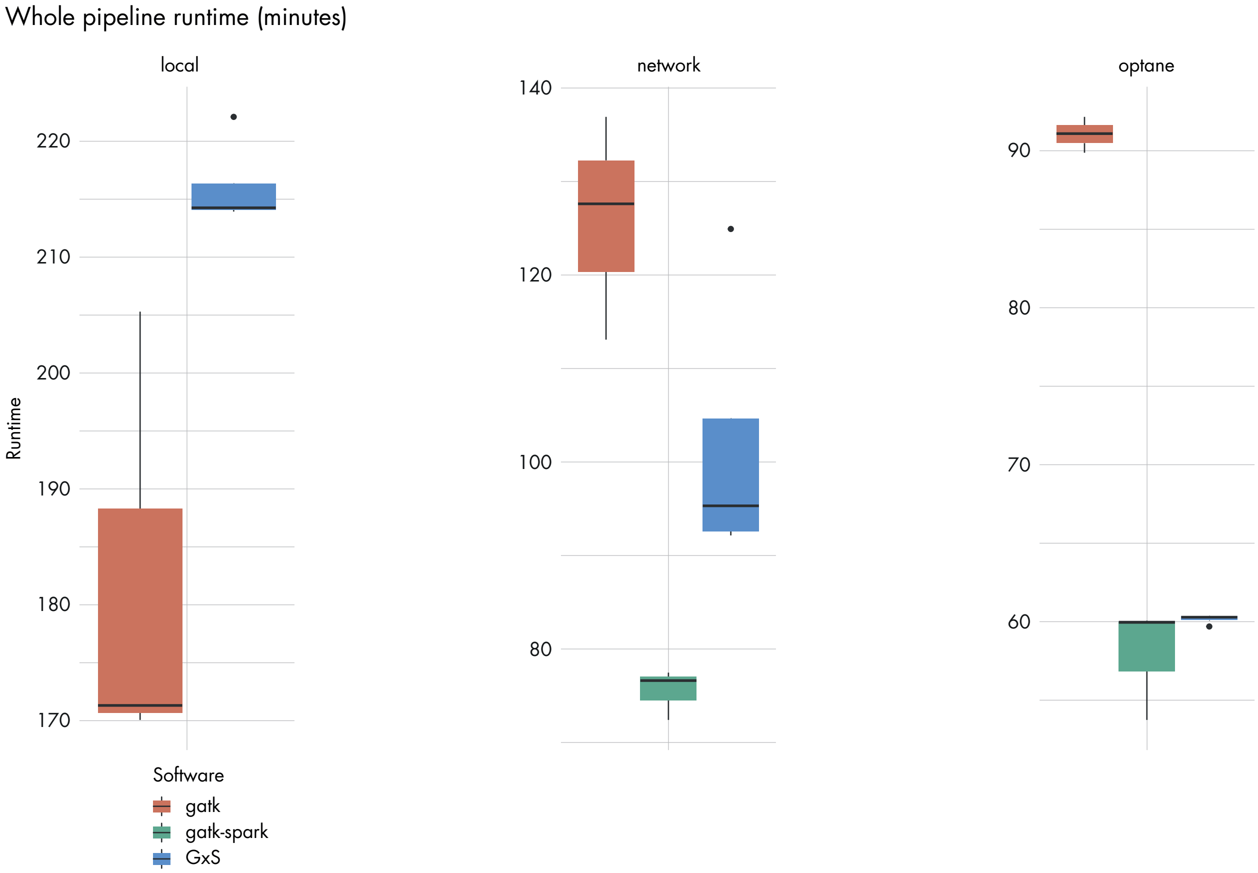


Figure 2. Comparison of pipeline runtimes.

calling in GATK, but an entirely local realigned mapping is not provided as a result, and no structural variant calling is included in the workflow. These elements are essential when considering the differences in runtime.

Despite the additional steps in the CLC workflow, GxS can run in a comparable time to the most advanced GATK + Apache Spark installation, when executed on a high-end server. CLC GxS completes the end-to-end analysis, including the structural variant and local realignment steps, in about 60 minutes. GATK using Apache Spark completes the analysis in 57 minutes on a high-performance server coupled with SSD storage (Figure 2). It is also worth highlighting that the CLC pipeline can be efficiently executed from a graphical interface, using a workflow editor, and is accessible by non-bioinformaticians, which may be an important feature for many users. The use of an advanced distributed computing solution like the Apache Spark environment adds an extra effort to the installation of GATK, and requires expertise in Linux system administration to be appropriately configured on a server or computing cluster.

When the data are being processed on a network drive, the CLC Genomics Server reduces analysis time by 20% compared to GATK alone (~102 minutes versus ~126 minutes), while GATK with Apache Spark runs 25% faster (~76 minutes). Most likely, Apache Spark is able to compensate for part of the overhead generated into the I/O when reading from network drives.

CLC analysis on a laptop runs in about three and a half hours, compared to GATK running in three hours. This is a very acceptable time, considering the additional data produced by CLC and the QC step included in the workflow. Additionally, it is worth noting that we used a dataset with higher coverage (100X) than typical exome applications. The analysis performed on a laptop does not include a comparison with Apache Spark, because the environment would create more overhead on a 4-cores machine by partitioning the data, outweighing the advantages in the computation.

Table 2. Runtime benchmarking results (in minutes)

Environment	Analysis name	GATK	GATK-Apache Spark	GxS
Intel-Optane SSD	ALIGN	11.23	11.23	16.34
Intel-network drive	ALIGN	17.27	11.65	21.97
Laptop SSD	ALIGN	81.23	NA	90.26
Intel-Optane SSD	DEDUP	18.29	16.06	8.69
Intel-network drive	DEDUP	38.68	22.77	18.40
Laptop SSD	DEDUP	14.67	NA	10.73
Intel-Optane SSD	SV	NA	NA	6.78
Intel-network drive	SV	NA	NA	13.26
Laptop SSD	SV	NA	NA	24.14
Intel-Optane SSD	LOCALREALIGN	NA	NA	7.71
Intel-network drive	LOCALREALIGN	NA	NA	10.66
Laptop SSD	LOCALREALIGN	NA	NA	25.06
Intel-Optane SSD	RECAL	29.71	24.12	NA
Intel-network drive	RECAL	33.77	33.36	NA
Laptop SSD	RECAL	36.14	NA	NA
Intel-Optane SSD	QC	NA	NA	5.40
Intel-network drive	QC	NA	NA	10.78
Laptop SSD	QC	NA	NA	8.95
Intel-Optane SSD	CALLING	30.06	5.01	12.28
Intel-network drive	CALLING	34.13	5.97	17.21
Laptop SSD	CALLING	47.61	NA	52.71
Intel-Optane SSD	FILTER	0.24	0.21	0.35
Intel-network drive	FILTER	0.31	0.26	0.78
Laptop SSD	FILTER	0.36	NA	0.65
Intel-Optane SSD	ANNO	1.48	1.29	2.42
Intel-network drive	ANNO	1.66	1.52	15.39
Laptop SSD	ANNO	2.19	NA	3.44
Intel-Optane SSD	PIPELINE	91.02	57.91	60.14
Intel-network drive	PIPELINE	125.81	75.53	101.91
Laptop SSD	PIPELINE	182.21	NA	216.10

Figure 3 illustrates the execution on a high-end server, with a breakdown of each pipeline by analysis step. Figure 4 shows in a similar way the performance on a laptop.

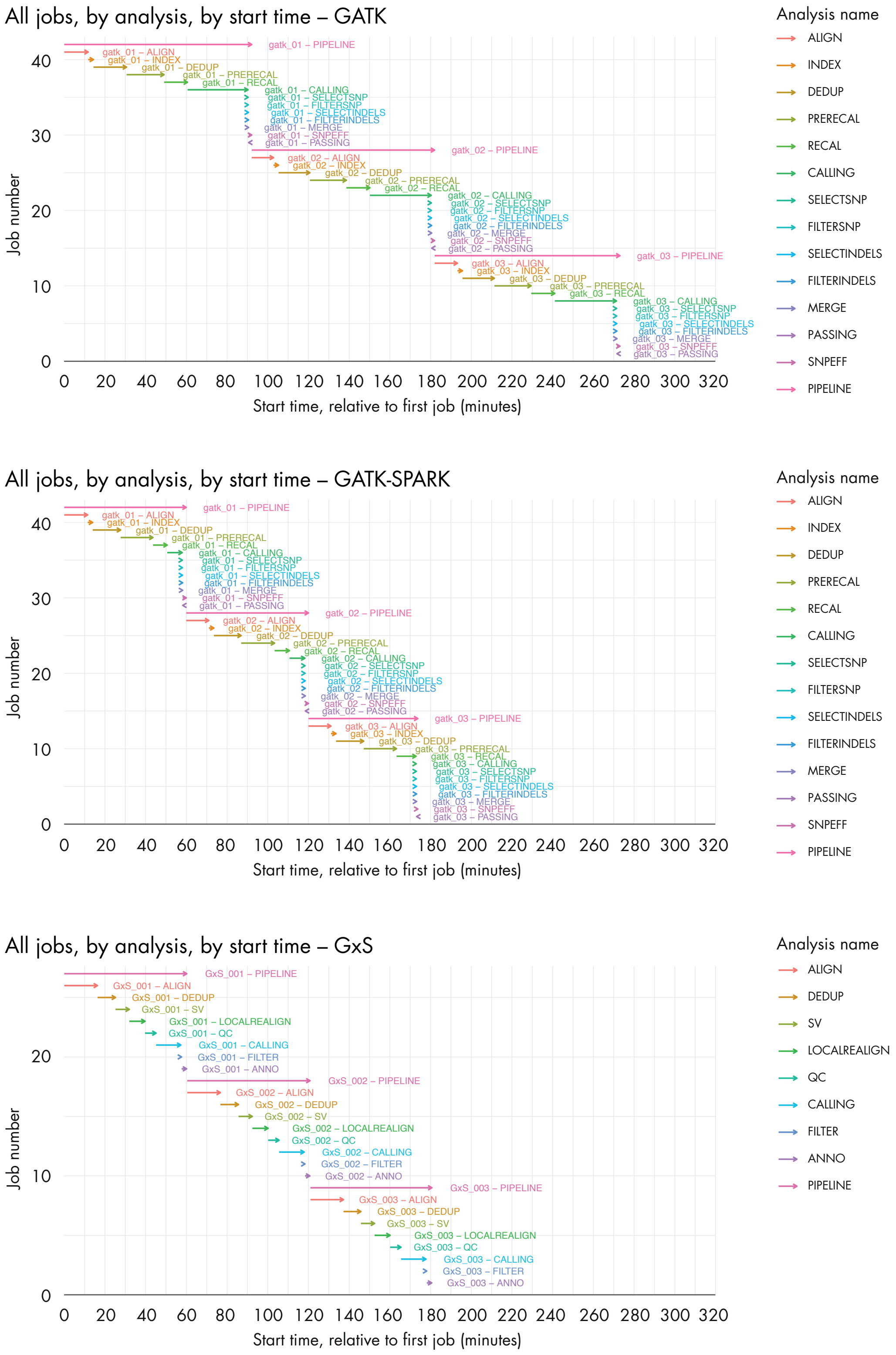
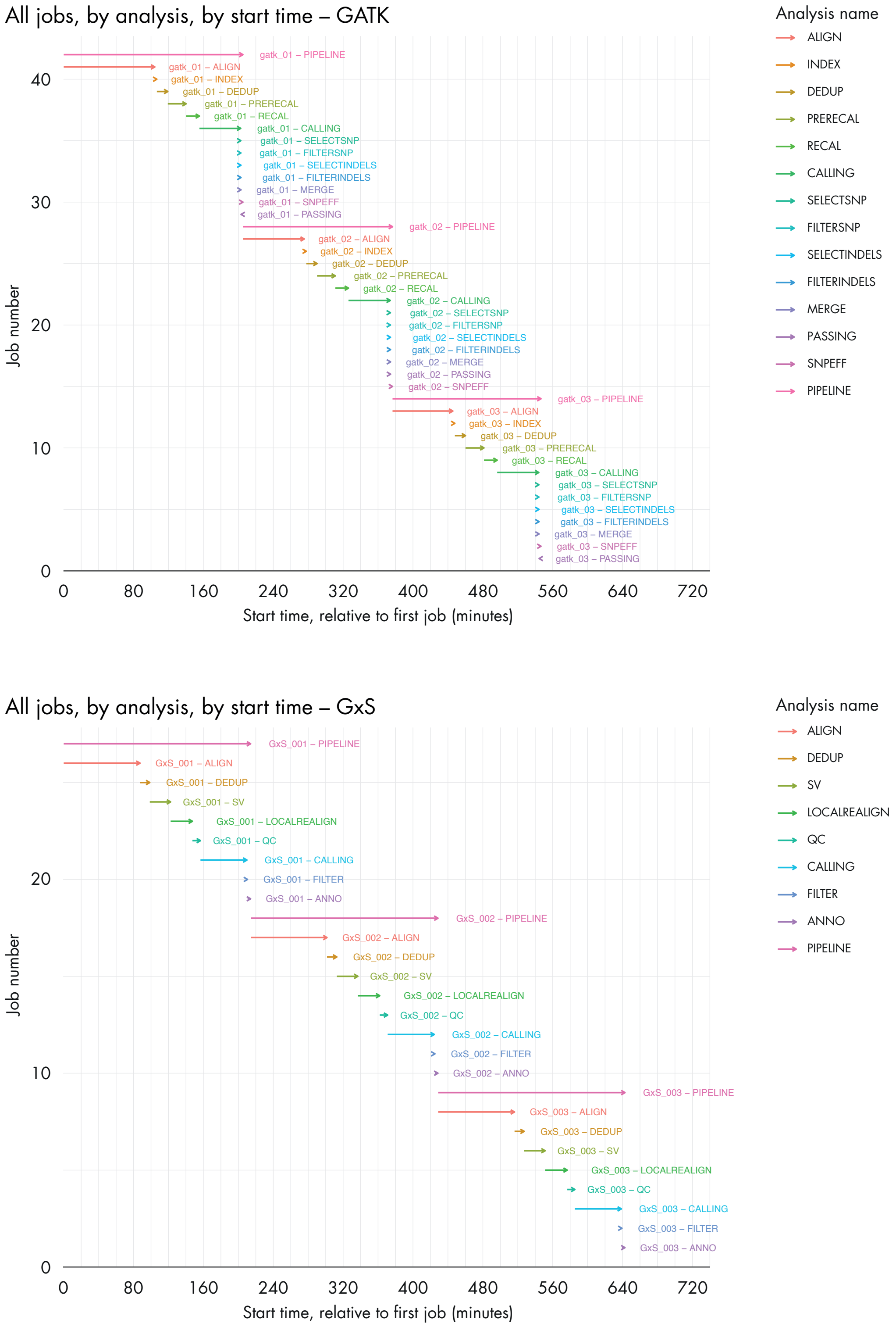


Figure 3. Gantt plot on 40-cores server and SSD storage.


All jobs, by analysis, by start time – GATK

All jobs, by analysis, by start time – GxS

Figure 4. Gantt plot on laptop (MacBook Pro 15").

Benchmark of accuracy

To benchmark the accuracy on this dataset, we also included Strelka2. This is a new caller that is gaining interest because of its flexibility in calling variants under different circumstances. We ran Strelka2 on the BAM files generated by GATK after recalibration, and before the use of HaplotypeCaller. For each pipeline we present the results before (ALL) and after filtering (PASS), to appreciate both the original sensitivity of the caller, as well as the effect on precision and sensitivity that each of the filters applied has on the final results. Table 2 presents the results in detail.

If we look at the filtered variants in terms of sensitivity, the CLC GxS performs better than any supplier on both SNPs and INDELs (99.62% and 91.86%, respectively). GATK scores highest for precision of SNPs (99.85% followed closely by CLC with 99.11%) and Strelka2 for precision of INDELS (95.16% compared to 92.33% achieved by CLC).

In terms of accuracy, CLC GxS scores highest for SNPs (99.36% versus 97.37% and 97.33% achieved by GATK and Strelka2, respectively), while Strelka2 scores highest for INDELs (93.31% versus 92.33% by CLC).

Table 3. Results of Hap.py comparing the software in terms of recall, precision and accuracy (F1 score)

Software	Type	Filter	TRUTH.TOTAL	TRUTH.TP	TRUTH.FN	QUERY.TOTAL	QUERY.FP	Recall	Precision	F1_Score
GATK4	INDEL	ALL	295	266	29	317	44	0.901695	0.861199	0.880982
GATK4	INDEL	PASS	295	266	29	297	24	0.901695	0.919192	0.910359
GATK4	SNP	ALL	16570	15776	794	15863	88	0.952082	0.994452	0.972806
GATK4	SNP	PASS	16570	15741	829	15763	23	0.94997	0.998541	0.97365
STRELKA2	INDEL	ALL	295	272	23	304	27	0.922034	0.911184	0.916577
STRELKA2	INDEL	PASS	295	270	25	289	14	0.915254	0.951557	0.933053
STRELKA2	SNP	ALL	16570	15772	798	15901	129	0.951841	0.991887	0.971451
STRELKA2	SNP	PASS	16570	15730	840	15753	23	0.949306	0.99854	0.973301
GxS11	INDEL	ALL	295	282	13	300	18	0.955932	0.94	0.947899
GxS11	INDEL	PASS	295	271	24	292	21	0.918644	0.928082	0.923339
GxS11	SNP	ALL	16570	16521	49	16929	408	0.997043	0.975899	0.986358
GxS11	SNP	PASS	16570	16507	63	16655	148	0.996198	0.991114	0.993649

Conclusions

The benchmark presented shows both CLC Genomics Workbench and CLC Genomics Server offer competitive solutions compared to GATK, both in terms of runtime on high-end computing infrastructures or a laptop, as well as in terms of accuracy of called variants.

The impressive performance, even when compared to a sophisticated data distributed computing environment such as Apache Spark, shows the quality of algorithms implemented in CLC software. The high levels of accuracy demonstrated in the comparison of variant calls supports the reliability of the analysis pipeline for large scale applications like exome sequencing.

References

1. https://software.broadinstitute.org/gatk/gatk4
2. Kim, S. et al. (2018) Strelka2: Fast and Accurate Calling of Germline and Somatic Variants. Nature methods **15**, 591.
3. Zook, J. M. et al. (2014) Integrating Human Sequence Data Sets Provides a Resource of Benchmark SNP and Indel Genotype Calls. Nature Biotechnology **32**, 246.
4. Cornish, A. and Guda, C. (2015) A Comparison of Variant Calling Pipelines Using Genome in a Bottle as a Reference. Biomed Res Int, 456479.
5. Hwang, S. et al. (2015) Systematic Comparison of Variant Calling Pipelines using Gold Standard Personal Exome Variants. Scientific Reports **5**, 17875.
6. Kingsmore, S. F. et al (2008) The Importance of Coverage Uniformity Over On-Target Rate for Efficient Targeted NGS. Nature Reviews Drug Discovery **7**, 221.
7. Kawalia, A. et al. (2015) Leveraging the Power of High Performance Computing for Next Generation Sequencing Data Analysis: Tricks and Twists from a High Throughput Exome Workflow. PLoS ONE **10**, e0126321.
8. https://github.com/Illumina/hap.py

Discover more at **www.qiagenbioinformatics.com**.

To learn more, have a look at these informative tools:

Web resources

CLC Genomics Server:

www.qiagenbioinformatics.com/products/clc-genomics-server/

CLC Genomics Workbench

www.qiagenbioinformatics.com/products/clc-genomics-workbench/

CLC Microbial Genomics Module

www.qiagenbioinformatics.com/products/clc-microbial-genomics-module/

Tutorials

www.qiagenbioinformatics.com/support/tutorials/

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.

The CLC Genomics Workbench and CLC Genomics Server are intended for molecular biology applications. These products are not intended for the diagnosis, prevention or treatment of a disease.