CompStor Novos™

Accelerated, Highly-Accurate Variant Calling using a Memory Optimized, Multi-Node Appliance

Various hardware accelerators have been used to speed next generation sequencing (NGS) variant calling pipelines. OmniTier has developed a more flexible, memory-centric architecture to address compute bottlenecks in human whole genome sequencing (WGS). CompStor Novos™ significantly reduces analysis time while boosting accuracy in alignment and assembly based variant calling. CompStor Novos™ *alignment* completes in under 2 hours and *assembly* completes in 1-3 hours, using the same architecture and automated application interface. Alignment based variant calling shows significant accuracy gains compared to GATK *Best Practices* in benchmark testing.

Introduction

Whole genome sequencing informatics pipelines require both acceleration and accuracy to enable precision medicine efforts. A memory-centric approach, efficiently using tiers of solid state drives (SSDs) and DRAM, drives down run times and costs-per-genome. The compute advantage also allows for a dual pipeline — *de novo* assembly and reference alignment techniques followed by integrated variant calling. Here the benefits of both methodologies are available: highly accurate short variants via alignment and larger structural variants via assembly. CompStor Novos™ is a fully automated compute appliance, enabling complete variant discovery in germline, short-read sequencing. Results demonstrate:

1. Run time acceleration with WGS being completed in 1.5 - 3 hours;
2. Superior variant calling accuracy to the GATK Best Practices pipeline;
3. Structural variant discovery not found with traditional pipelines.

Genomics has been called a “four-headed beast” by University of Illinois practitioners for its data requirements around: acquisition, storage, distribution, and analysis[1]. The field has already eclipsed its largest big data competitors in astronomy and hyper-scaled social media (e.g. YouTube) in terms of sequencing capacity and overall data size. By 2025 an estimated 10 billion total genomes (human, plant and animal) could be completed. And the field’s 2025 ***annual*** sequencing capacity could reach 1 zetta-base pairs (10^{21}).[1]

CompStor Novos™ addresses these data challenges with a memory-centric solution that effectively treats expansive SSD storage as an *in-memory* compute resource, driving accelerated analysis. The solution is multi-noded, with 2-8 commercial servers, and scales for varying WGS throughput requirements. A web-application job scheduler allows users to flexibly select several analysis options including *de novo* assembly and reference-based alignment, as shown in Figure 1. This

ready-to-go solution may be used by bioinformatic non-specialists and avoids the complexities of maintaining multiple disparate open-source tools. High-speed ingress and egress of data is also optimized.

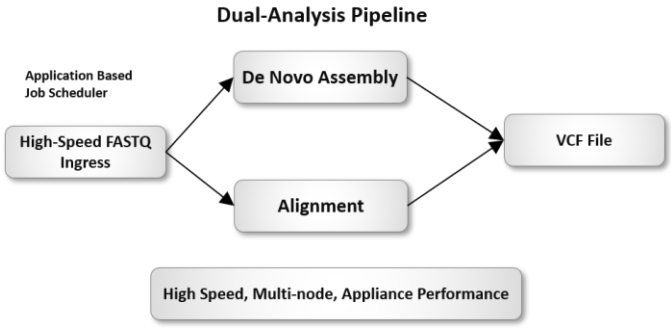

Figure 1: ***CompStor Novos™ Dual Pipeline*** *– A schematic showing high speed ingress of FASTQ files into an appliance cluster running either de novo assembly or reference alignment based analysis and VCF output.*

CompStor Novos™ Alignment – highly accurate variant calling

CompStor Novos™ alignment pipeline is optimized to perform faster and more accurately than the GATK Best Practices pipeline. Figure 2 shows a Receiver Operator Characteristic (ROC) curve measuring sensitivity (true positivity) and specificity (false positivity) in variant calling. High accuracy is achieved using a WGS-optimized deep learning technique. This performance extends to all Genome in a Bottle (GIAB) datasets where a truth set is available for comparison, as shown in Table 1.

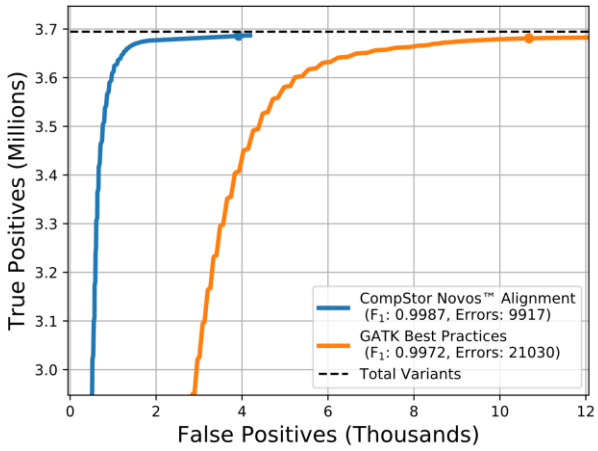

Figure 2: ***Receiver Operator Characteristic (ROC) Curve*** *– CompStor Novos™ alignment mode is compared to GATK Best Practices pipeline for false positives versus true positives on HG001 at 35x coverage. CompStor Novos™ shows 11,113 fewer errors and an increased F-Score compared to GATK.*

GIAB Dataset	CompStor Novos™	GATK Best Practices
HG001	**.9987**	.9972
HG002	**.9987**	.9976
HG003	**.9984**	.9971
HG004	**.9984**	.9970
HG005	**.9983**	.9954
HG006	**.9991**	.9980
HG007	**.9990**	.9978
F_1 Average	**.9987**	.9972
Total Errors Average	**9,726**	20,165

Table 1: *CompStor Novos™ F_1 Score comparison with GATK Best Practices across all GIAB datasets including Total Errors*

The CompStor Novos™ alignment pipeline is both faster and more accurate than GATK Best Practices, especially in higher coverage genomes. CompStor Novos alignment shows superior variant calling to both open source options (E.g. GATK Best Practices) and commercial options. F-scores in Table 1 demonstrate superior accuracy to GATK in all GIAB datasets.

CompStor Novos™ Assembly – unique variant discovery

De novo DNA assembly reconstructs genomes from DNA sequence reads without the use of a reference template. It is, therefore, unbiased by inherent differences between the personal genome and the reference template. *De novo* methods are widely considered too computationally intensive for routine, high throughput variant calling applications, requiring 12-72 hours depending on coverage depth and computing resources. Therefore, *de novo* assembly techniques are often restricted to lower throughput applications: model organism assembly; long-read sequencing; linked-reads; and Hi-C/3C data. In addition, variant calling from *de novo* assembly is difficult to accurately implement. CompStor Novos™ changes this dynamic with accurate 1-3 hour variant calling, offering new variant discovery opportunities.

Assembly- variant calling performance

CompStor Novos™ assembly accurately calls short variants (SNVs and Indels) and produces contigs where complex structural variants (SV) may be identified. Our results show:

1. Comparable variant calling accuracy to the most reliable alignment based approaches for SNVs and indels with less than 50 base pairs;

2. Unique subset of truth-set-verified short variants is present, approximately 2,000 per genome. These variants are disproportionately Indels;

3. Structural variant (SV) detection with base pair level resolution not possible from short-read alignment based methods.

The data analyzed came from two well-characterized subjects, NA12878 and NA24385, made available through GIAB Consortium. Sequencing coverages of 35X and 100X were also analyzed for both subjects. The full results are in academic press review with collaborators from the Mayo Clinic[3].

Novos: Accurate Variant Calling SNVs and Indels

Novos nearly matches the well-developed GATK Haplotype Caller[2] Best Practices [2] across the whole genome in SNV and Indels of less than 50 base pairs, as shown in Figure 3. De novo assemblers have not previously been able to match alignment based variant calling given technical difficulties inherent in relating to a reference genome.

CompStor Novos™ assembly also showed little to no variant allele frequency (VAF) bias in its variant calling. VCF bias is the tendency for some alignment-based protocols to misrepresent heterozygous indels of increasing length[3].

Within GIAB datasets the assembly pipeline discovers approximately 2,000 unique short variants per genome not found with common alignment pipelines. These variants are disproportionality indels (40%) and indicate the assembly pipeline's ability to discover more complex short variants.

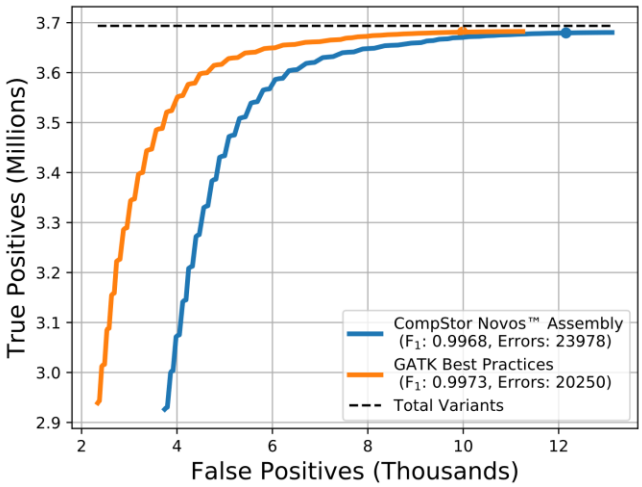

***Figure 3: ROC curve** comparison of true positives versus false positives for CompStor Novos' de novo assembly compared to a GATK Haplotype caller pipeline using NA12878 (HG001) with 35x coverage.*

Novos Assembly: Revealing Structural Variants

Structural and other complex variation in the personal genome are preserved in the CompStor Novos™ assembled contigs. For HG002, 386 SVs were detected with sizes between 50 and 21,046 bps. Of these 386 SVs, 346 were confirmed in the GIAB truth set. Table 2 shows four of these Novos-discovered variants which were not discoverable in the GATK alignment based genome.

Contig length	Structural Variant	(Chr)	Position	Affected gene
3086	17210 bp **deletion**	7	109453901	---
2937	325 bp **deletion**	15	64633163	NG_051236.1
10974	542 bp **insertion**	8	27295979	NG_029510.1
4912	3630 bp **insertion**	4	97423310	---

Table 2: *CompStor Novos™ structural variant detection* — *Listing of four structural variants and their associated genomic information.*

Tiered-memory architecture

The per genome costs to assemble and call variants in precision medicine is a function of the:

1. number of bioinformatic pipelines required to obtain a complete set of genomic variants;
2. automation in the total informatics schema; and,
3. run time for the algorithms.

CompStor Novos™ addresses the first two with a multi-option variant discovery solution and an automated FASTQ→VCF user interface. The proprietary tiered-memory architecture shrinks # 3 *run time* to several hours depending upon the chosen cluster size and sequencing coverage, as shown in Table 3. The compute architecture, run times, and associated costs are discussed below.

As with most large-data problems, genome informatics benefits from extreme amounts of online memory. However, not all data in memory is accessed at the same rate, nor with the same criticality in overall performance, a situation akin to database caching. CompStor Novos' compute innovation is to optimize thread scheduling, data classification, data structure design, and algorithm choices in an Overlap Layout Consensus (OLC) based assembly algorithm. The nodes in each cluster perform as if they had 12.8 TB of DRAM, wherein up to 90% of that total memory is actually resident in high performance SSDs. This feature is the key factor for achieving supercomputer-like performance with commercial off-the-shelf (COTS) servers while maintaining low costs.

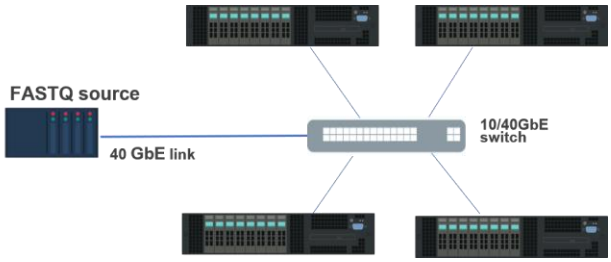

Figure 4: *A 4-node, CompStor Novos™ compute cluster*

Run times & per genome costs

CompStor Novos™ appliance solution is optimized for an automated and fast ingress of short read FASTQ data as shown in Figures 4. A web app schedules sequential jobs and lets the user automate and supervise the analysis process. Run times are shown for various node configurations in Table 3. The times compare well to supercomputing and hardware accelerator solutions.

ASSEMBLY		ALIGNMENT	
Configuration	**Run Time**	**Configuration**	**Run Time**
2 nodes	3 hours	2 nodes	2 hours
4 nodes	2 hours	4 nodes	1 hours
8 nodes	1 hour	8 nodes	<1 hours

Table 3: *CompStor Novos™ pipeline run times by node configuration for both alignment and assembly modes.*

Conclusion

Accurate and high speed variant calling is crucial for the routine implementation of precision medicine. CompStor Novos™ provides greater accuracy than the GATK Best Practices pipeline with significantly better F-scores over all GIAB datasets. Run times are also dramatically faster, completing in 1-3 hours depending on nodes and sequence coverages. The CompStor® architecture leverages significant SSD-based flash memory coordinated with DRAM to drive accelerated compute times for bioinformatic pipelines. A dual pipeline offers both de novo assembly and reference alignment based variant calling. Both pipelines deliver SNVs and short Indels with greater accuracy than open source and commercial solutions. In addition, SVs may be identified with fidelity in the assembled contigs of the *de novo* assembly.

Future software releases will include automated SV detection, and reference aided assembly methods. CompStor® is also a platform for massive data analytics jobs, performing singular value decomposition (SVD), principal component analysis (PCA), Least Squares (LSQR) solutions, and other AI-related analyses at TB scale.

About OmniTier

OmniTier Inc., founded in 2015, develops and supports integrated software solutions for memory-centric infrastructure applications, including high performance object caching, scientific analysis for machine learning, AI, and genomics. Its leadership team has a track record of delivering many industry firsts in data storage and access across different media types. The company has offices in Milpitas, California, and Rochester, Minnesota.

References

1 ***Big Data: Astronomical or Genomical? Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. (2015) PLoS Biol 13(7): e1002195.***

2 ***The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data*** *McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA, 2010 GENOME RESEARCH 20:1297-303*

3 ***CompStor Novos: a low cost yet fast assembly-based variant calling for personal genomes*** *– Academic Paper in press with Mayo Clinic collaborators, https://www.biorxiv.org/content/biorxiv/early/2018/12/04/486092.full.pdf*