BIOINFORMATICS APPLICATION NOTE #1

Benchmark Results

OmniTier CompStor Novos® *Alignment* and *Assembly* pipelines shows greater variant calling accuracy than GATK across the seven GIAB datasets with accelerated runtimes

OmniTier Inc. | 1591 McCarthy Blvd, Milpitas, CA 95035 | 2720 Superior Dr NW, Rochester, MN 55901

SUMMARY

OmniTier's CompStor Novos® bioinformatics appliance calls short variants (SNVs & Indels <50bp) using either a de novo assembly or reference alignment methodology. Both pipelines show greater variant calling accuracy across the seven Genome in a Bottle (GIAB) datasets than GATK pipeline as measured by F_1 scores, minimum error counts, and ROC curves.

The CompStor Novos® appliance demonstrates accelerated run-times, completing variant calling in 1.8 hours on a 2-node configuration versus 9.4 hours for the GATK Best Practices pipeline. Further runtime reduction below 1 hour is feasible with more nodes.

1 INTRODUCTION

Accurate and fast variant calling in whole genome sequencing (WGS) drives precision medicine adoption. CompStor Novos® is a scalable compute appliance utilizing a tiered-memory architecture of DRAM and SSDs to accelerate and improve variant calling in germline WGS pipelines. It is a *dual pipeline* utilizing *de novo* assembly and alignment techniques. This strategy increases the opportunity to find unique variants while driving down runtimes. The appliance is controlled by an intuitive web application interface and enables batch processing automation. For variant calling, OmniTier uses a domain-optimized deep learning methodology to produce fewer false positives and more true positives in germline WGS. This improved performance is demonstrated across the 7 GIAB datasets made available by the National Institute of Standards and Technology (NIST)[1].

2 COMPARING DATA

GIAB datasets (HG001-HG007) were downloaded from NIST GIAB website[2]. The GATK 4.0.1.2 pipeline software was downloaded from the Broad Institute website[3]. F_1 scores are calculated from recall (fraction of true variants detected) and precision (fraction of variants called that are true):

$$F_1 = \frac{2}{(\text{recall}^{-1} + \text{precision}^{-1})}$$

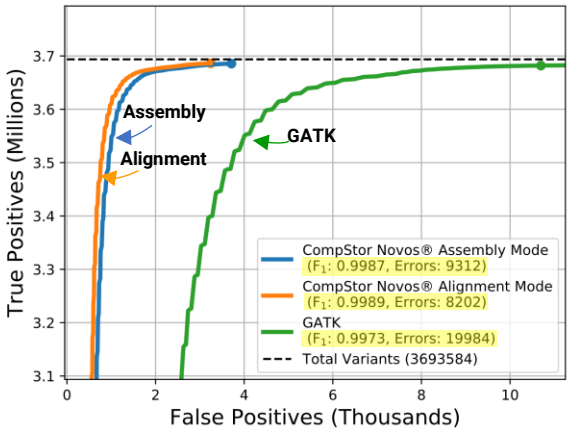
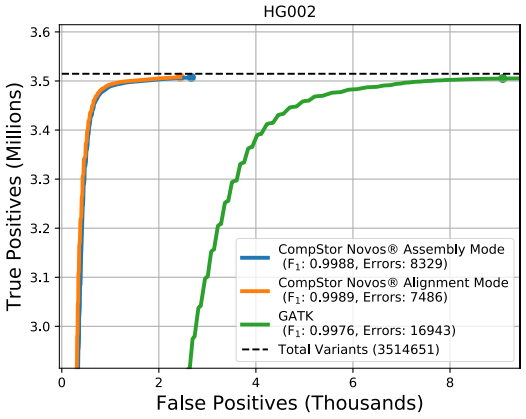
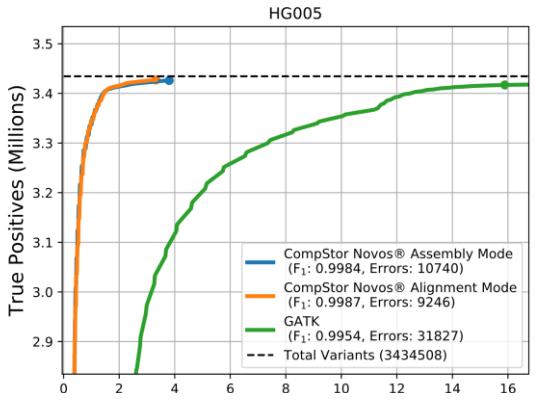


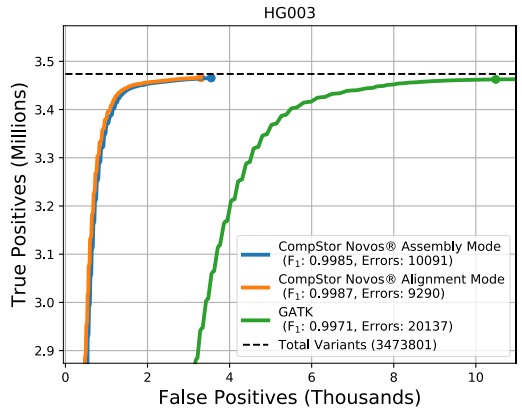
Figure 1. Receiver Operator Characteristics (ROC) Curve comparing HG001 true positives and false positives between CompStor Novos® Alignment and GATK Best Practices

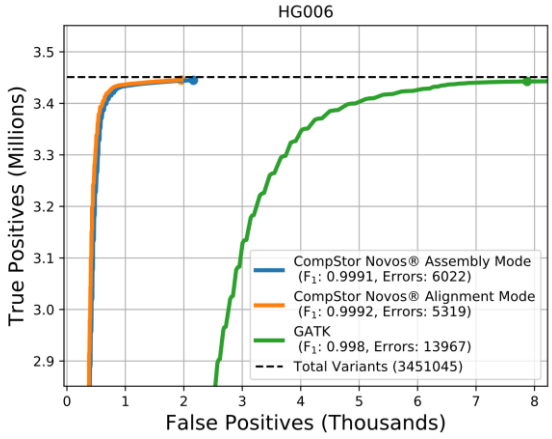
GIAB Dataset	CompStor Novos®	GATK	Percentage Improvement
HG001	**8,202**	19,984	59.0%
HG002	**7,486**	16,943	55.8%
HG003	**9,290**	20,137	53.9%
HG004	**9,685**	21,169	54.2%
HG005	**9,246**	31,827	70.9%
HG006	**5,319**	13,967	61.9%
HG007	**5,351**	14,918	64.1%
Total Errors Avg	**7,797**	19,849	60.7%
F_1 Average	**.9989**	.9972	

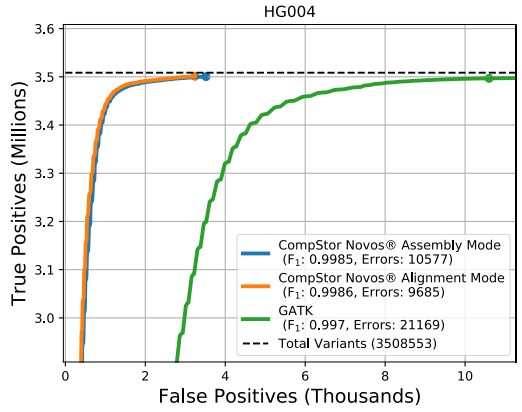
Table 1. Total Errors and F_1 Scores comparing CompStor Novos® Alignment with GATK Best Practices for all NIST GIAB datasets











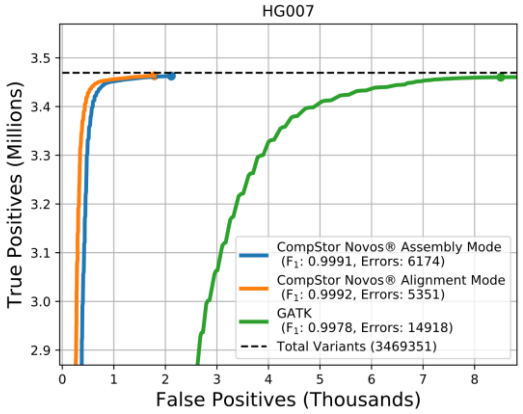


Figure 2. ROC Curves for HG002-HG007 comparing CompStor Novos® Alignment with GATK Pipeline

REFERENCES

[1] J. M. Zook, B. Chapman, J. Wang et al., "Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls," Nature Biotechnology, vol. 32, no. 3, pp. 246–251, 2014.

[2] National Institute of Standards and Technology, US Department of Commerce. https://www.nist.gov/programs-projects/genome-bottle.

[3] Broad Institute, https://software.broadinstitute.org/gatk/.