

Sentieon DNA pipeline for variant detection - Software-only solution, over 20× faster than GATK 3.3 with identical results

Sentieon DNaseq Software is a suite of tools for running DNA sequencing secondary analyses. The Sentieon DNaseq Software produces identical results to the Genome Analysis Toolkit (GATK) Best Practice Workflow using HaplotypeCaller, with more than a 20-fold increase in processing speed on the same hardware.

This paper presents a benchmark analysis of the GATK and Sentieon DNaseq software packages using publically available datasets from the 1000 genomes database, and includes speed comparisons and output concordance analyses.

Sentieon DNA pipeline for variant detection

Software-only solution, over 20× faster than GATK 3.3 with identical results

Jessica A. Weber¹, Rafael Aldana⁵, Brendan D. Gallagher⁵, Jeremy S. Edwards^{2,3,4}

¹Department of Biology, University of New Mexico, Albuquerque, NM. ²Department of Chemistry & Chemical Biology, University of New Mexico, Albuquerque, NM. ³Department of Molecular Genetics and Microbiology, University of New Mexico, Albuquerque, NM. ⁴Department of Chemical & Nuclear Engineering, University of New Mexico, Albuquerque, NM. ⁵Sentieon Inc., Mountain View, CA

Corresponding author:

Rafael Aldana

465 Fairchild Dr, Suite 135 Mountain View, CA, USA

email: rafael.aldana@sentieon.com

Abstract

Sentieon DNaseq Software is a suite of tools for running DNA sequencing secondary analyses. The Sentieon DNaseq Software produces identical results to the Genome Analysis Toolkit (GATK) Best Practice Workflow using HaplotypeCaller, with more than a 20-fold increase in processing speed on the same hardware.

This paper presents a benchmark analysis of the GATK and Sentieon DNaseq software packages using publically available datasets from the 1000 genomes database, and includes speed comparisons and output concordance analyses.

1 Introduction

2 Recent advances in next generation sequencing (NGS)
3 technologies have dramatically increased the rate of data output
4 while significantly reducing costs. However, highly accurate
5 analysis of NGS data is computationally intensive and creates a
6 bottleneck in the overall sequencing workflow.

7 The current gold standard in variant calling is the Genome
8 Analysis Toolkit (GATK)¹ Best Practice Workflow pipeline using
9 HaplotypeCaller, which is regarded to have the highest accuracy
10 for both single nucleotide polymorphisms (SNPs) and small
11 insertions and deletions (indels).^{2,3} However, its slow computation
12 speed often makes adoption challenging.

13 To address these challenges, the Sentieon DNaseq Software
14 Package was developed to significantly decrease the analysis time
15 and the computational resource requirements for variant detection
16 without compromising accuracy; to that end, the software
17 implemented from the ground up the mathematics developed by
18 the Broad Institute¹. The result is a 20-to-50-fold increase in
19 processing speed on the same hardware with results that are
20 identical to the GATK pipeline, with differences within the
21 numerical noise.

22 Sentieon DNaseq Software package

23 The Sentieon DNaseq Software is a package of tools used to
24 perform ultrafast variant detection in genomic data obtained from
25 NGS. It is designed to run on generic CPUs, without the need for
26 specialized hardware (such as GPU, FPGAs, ASICs, etc.).

27 Sentieon DNA produces identical results to the GATK 3.3
28 pipeline with more than 20× speed improvement and includes all
29 individual stages of the pipeline, namely: sample quality metrics
30 calculation, duplicate read removal, indel realignment, base
31 quality recalibration, and variant calling. The usage of Sentieon

32 DNA is consistent with GATK and utilizes similar inputs, outputs,
33 and parameters.

34 Sentieon DNA benchmarking methodology

35 A benchmarking comparison of Sentieon DNA and GATK 3.3
36 was performed using publically available genomic data from the
37 1000 Genomes Project (Appendix 1). The data was first mapped to
38 the human reference genome hg19 using BWA⁴ 0.7.12 and
39 SAMtools⁵ 1.2. The sorted.bam files were then used in two
40 software scripts, which were created following the GATK Best
41 Practice Workflows^{3,6} (see Appendix 2 for scripts). Each stage in
42 GATK corresponds to a stage in Sentieon DNA, allowing for
43 detailed, step-by-step evaluations of the two packages.

44 Six exome samples ranging from 3-347× coverage, and two full
45 genome samples, with 6× and 14× coverage, were selected for the
46 comparison of the two pipelines. The eight samples were analyzed
47 individually using SAMtools 1.2/GATK 3.3 and Sentieon DNA
48 201505.02 on a 24 core, 2.4 GHz AMD Opteron 6234, 96GB
49 memory server running Ubuntu 14.04.2 at the University of New
50 Mexico.

51 Sentieon DNA is >20× faster than GATK 3.3

52 The runtime for the two pipelines using HaplotypeCaller variant
53 calling was measured in core minutes. Exome runtime ranged
54 from 108-2126 minutes for GATK 3.3 and 3-47 minutes for
55 Sentieon DNA, while genome runtime was 2188 and 3978
56 minutes for GATK 3.3 and 66 and 198 minutes for Sentieon DNA
57 (Appendix 3). Overall, Sentieon DNA provided a speed
58 improvement over GATK 3.3 of 34-51× on the six exome samples
59 and of 20-33× on the genome samples (Figure 1). For a
60 comparison of UnifiedGenotyper variant calling, see Table 1 and
61 Appendix 3.

Figure 1: Runtime Comparison

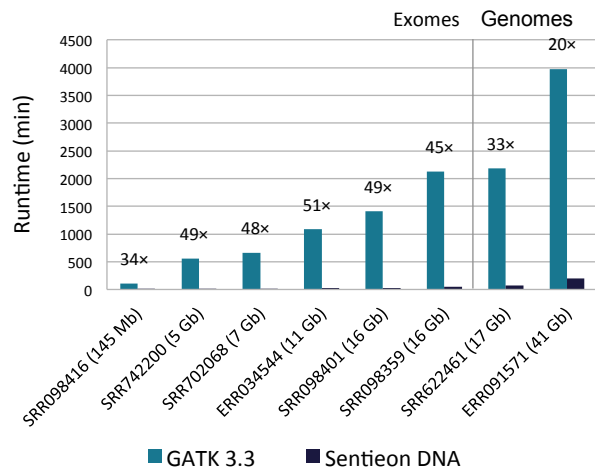


Figure 1. Runtime comparison of the pipelines using HaplotypeCaller variant calling in both software packages. Speed improvement of Sentieon DNA over GATK 3.3 is provided above each sample. Sentieon DNA runtime improvement ranges from 20–51× faster than GATK.

Figure 2: Accuracy Comparison

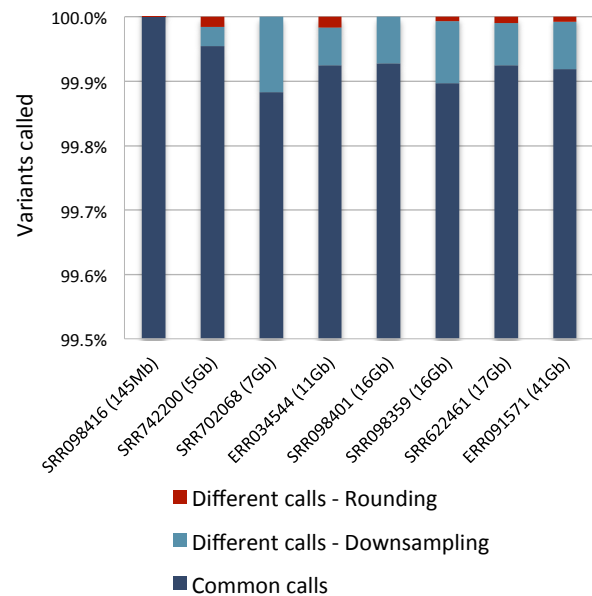


Figure 2. Concordance analysis for the variants called using the HaplotypeCaller in both software packages. Over 99.8% of the variant calls produced by GATK 3.3 and Sentieon DNA were identical. After removing the differences from GATK downsampling, the variant calls were over 99.99% concordant

Sentieon DNA produces identical results to GATK 3.3

The variant calling results of the two pipelines were analyzed for concordance using the program VarSeq™ from Golden Helix. Variants with quality-by-depth smaller than 2 and depth smaller than 5 were removed from the comparisons, as were variants called outside the exome capture area in the six exome samples.

In order to decrease runtime, GATK employs downsampling in areas of high coverage, which results in run-to-run variation in the variants called (Appendix 4). Sentieon DNA, however, does not downsample and produces identical results between runs.

To identify the number of differing variant calls between the two pipelines that can be attributed to this downsampling, the GATK 3.3 pipeline was run an additional seven times for each sample. If all eight GATK 3.3 runs did not consistently call a variant, Sentieon DNA differences in these calls were attributed to downsampling by GATK.

The VarSeq™ analyses revealed over 99.8% concordance between the GATK 3.3 and Sentieon DNA variant calls (Figure 2). After removing the variation caused by GATK downsampling, the concordance between the two software packages increased to more than 99.99% (Table 1). In total, there were less than 1 in 10,000 true differences between the GATK 3.3 and Sentieon DNA analyses, which were caused by rounding differences between the two different software paths. This level of variance is 10× less than the numerical noise caused by run-to-run variation within GATK (Appendix 4).

Conclusions

The Sentieon DNaseq Software package for variant detection produces identical SNP and indel variant identification to GATK 3.3 at >20× the speed. Transitioning pipelines from GATK to Sentieon DNA is easy due to consistent pipeline stages and similar user interface. Thus, Sentieon DNA enables drastically higher productivity, faster turn around time, and an order of magnitude increase in effective computing power of existing systems.

Figure 3: Sentieon Laptop vs. GATK Server Runtimes

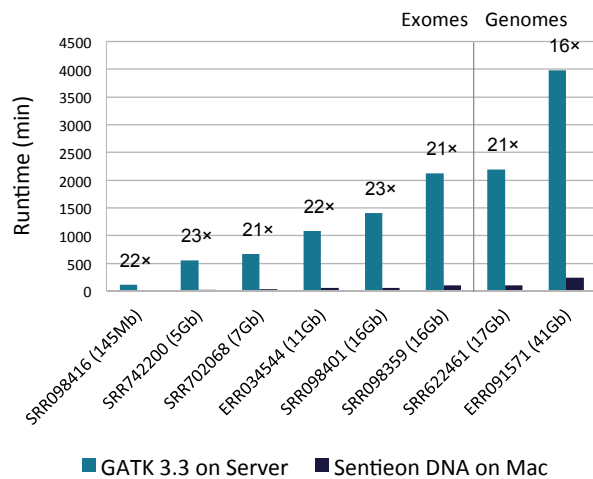


Figure 3. Runtime comparison of the pipelines using HaplotypeCaller variant calling for GATK run on the server compared to Sentieon DNA run on the MacBook. Speed improvement of Sentieon DNA over GATK 3.3 is provided above each sample. Sentieon DNA run on an 8-core laptop is >16x faster than GATK run on a 24 core server.

1

References

- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernysky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*. 2010;20:1297-303.
- Yi M, Zhao Y, Jia L, He M, Kebebew E, Stephens RM. Performance comparison of SNP detection tools with illumina exome sequencing data – an assessment using both family pedigree information and sample-matched SNP array data. *Nucleic Acids Research*. 2014;42(12):e101.
- Van der Auwera GA, Carneiro M, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella K, Altshuler D, Gabriel S, DePristo M. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Current Protocols in Bioinformatics*. 2013;43:11.10.1-11.10.33.
- Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2010;26(5):589-95.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-9.
- DePristo M, Banks E, Poplin R, Garimella K, Maguire J, Hartl C, Philippakis A, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell T, Kernysky A, Sivachenko A, Cibulskis K, Gabriel S, Altshuler D, Daly M. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*. 2011;43:491-498.

31

Table 1: Comparison of Sentieon DNA and GATK 3.3 Variant Calls

A. HaplotypeCaller										
Sample	Type	Sequencing Coverage	Sequenced Bases	Identical Variants	Differences Caused by Downsampling	Concordance: Sentieon vs. GATK without Downsampling	Differences: Missed INDEL	Differences: Added INDEL	Differences: Missed SNP	Differences: Added SNP
SRR098416	Exome	3x	145M	1420	0	100.000%	0	0	0	0
SRR742200	Exome	102x	5G	26454	8	99.985%	2	0	1	1
SRR702068	Exome	140x	7G	27296	32	100.000%	0	0	0	0
ERR034544	Exome	251x	11G	23995	14	99.983%	0	0	3	1
SRR098401	Exome	341x	16G	25067	18	100.000%	0	0	0	0
SRR098359	Exome	347x	16G	29104	28	99.993%	0	0	0	2
SRR622461	Whole Genome	6x	17G	1776194	1162	99.991%	13	7	58	90
ERR091571	Whole Genome	14x	41G	4317568	3159	99.992%	27	23	110	188

B. UnifiedGenotyper										
Sample	Type	Sequencing Coverage	Sequenced Bases	Identical Variants	Differences Caused by Downsampling	Concordance: Sentieon vs. GATK without Downsampling	Differences: Missed INDEL	Differences: Added INDEL	Differences: Missed SNP	Differences: Added SNP
SRR098416	Exome	3x	145M	591	0	100.000%	0	0	0	0
SRR742200	Exome	102x	5G	33475	4	100.000%	0	0	0	0
SRR702068	Exome	140x	7G	34729	12	100.000%	0	0	0	0
ERR034544	Exome	251x	11G	29803	9	99.993%	0	1	1	0
SRR098401	Exome	341x	16G	31632	3	100.000%	0	0	0	0
SRR098359	Exome	347x	16G	36938	9	99.997%	0	1	0	0
SRR622461	Whole Genome	6x	17G	2352529	565	99.979%	112	92	141	146
ERR091571	Whole Genome	14x	41G	5341272	1180	99.981%	247	193	330	243

Appendix 1: 1000 Genomes samples used in benchmarking

Individual	Sample	Type	Technology	Sequenced Bases	Sequencing Coverage	Link
NA11930	SRR098416	Exome	Illumina HiSeq	145M	3x	ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data/NA11930/sequence_read/SRR098416*
NA12046	SRR742200	Exome	Illumina HiSeq	5G	102x	ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data/NA12046/sequence_read/SRR742200*
NA12155	SRR702068	Exome	Illumina HiSeq	7G	140x	ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data/NA12155/sequence_read/SRR702068*
NA11932	ERR034544	Exome	Illumina HiSeq	11G	251x	ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data/NA11932/sequence_read/ERR034544*
NA12878	SRR098401	Exome	Illumina HiSeq	16G	341x	ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data/NA12878/sequence_read/SRR098401*
NA12891	SRR098359	Exome	Illumina HiSeq	16G	347x	ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data/NA12891/sequence_read/SRR098359*
NA12878	SRR622461	Whole Genome	Illumina HiSeq	17G	6x	ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data/NA12878/sequence_read/SRR622461*
NA12878	ERR091571	Whole Genome	Illumina HiSeq	41G	14x	ftp://ftp-trace.ncbi.nih.gov/giab/ftp/technical/NA12878_data_other_projects/sequence_read/ERP001229/ILLUMINA/sequence_read/ERR091571*

Appendix 2: Pipeline scripts

Read Alignment (used by both pipelines)

```
bwa mem -M -R "@RG\tID:$group\tSM:$sample\tPL:$pl" -t 24 $fasta $fastq_1 $fastq_2 | samtools view -Sb - >align.bam
samtools sort -@ 24 align.bam sorted
```

GATK 3.3

Sentieon DNA

Sample metrics calculation	<pre>java -jar picard.jar CollectAlignmentSummaryMetrics INPUT=sorted.bam OUTPUT=aln_metrics.txt REFERENCE_SEQUENCE=\$fasta ADAPTER_SEQUENCE=null VALIDATION_STRINGENCY=SILENT java -jar picard.jar CollectGcBiasMetrics INPUT=sorted.bam OUTPUT=gc_metrics.txt SUMMARY_OUTPUT=gc_summary.txt CHART_OUTPUT=gc_bias.pdf REFERENCE_SEQUENCE=\$fasta ASSUME_SORTED=true VALIDATION_STRINGENCY=SILENT java -jar picard.jar MeanQualityByCycle INPUT=sorted.bam OUTPUT=mq_metrics.txt CHART_OUTPUT=meanq_cycle.pdf REFERENCE_SEQUENCE=\$fasta VALIDATION_STRINGENCY=SILENT PF_READS_ONLY=true java -jar picard.jar QualityScoreDistribution INPUT=sorted.bam OUTPUT=qd_metrics.txt CHART_OUTPUT=qscore_dist.pdf REFERENCE_SEQUENCE=\$fasta VALIDATION_STRINGENCY=SILENT PF_READS_ONLY=true java -jar picard.jar CollectInsertSizeMetrics INPUT=sorted.bam OUTPUT=is_metrics.txt REFERENCE_SEQUENCE=\$fasta</pre>	<pre>\$driver -r \$fasta -t 24 -i sorted.bam --algo MeanQualityByCycle mq_metrics.txt --algo QualDistribution qd_metrics.txt --algo GCBias --summary gc_summary.txt gc_metrics.txt --algo AlignmentStat aln_metrics.txt --algo InsertSizeMetricAlgo is_metrics.txt python \$dir/bin/plot.py metrics -o metrics-report.pdf gc=gc_metrics.txt qd=qd_metrics.txt mq=mq_metrics.txt isize=is_metrics.txt</pre>
Duplicate read removal	<pre>java -jar picard.jar MarkDuplicates M=dup_reads I=sorted.bam O=dedup.bam samtools index dedup.bam</pre>	<pre>\$driver -t 24 -i sorted.bam --algo LocusCollector --fun score_info score.txt \$driver -t 24 -i sorted.bam --algo Dedup --rmdup --score_info score.txt deduped.bam</pre>
Indel realignment	<pre>java -jar GenomeAnalysisTK.jar -T RealignerTargetCreator -R \$fasta -I dedup.bam -known \$dbsnp_Mill -o realigner.intervals java -jar GenomeAnalysisTK.jar -T IndelRealigner -R \$fasta -I dedup.bam -known \$dbsnp_Mill -targetIntervals realigner.intervals -o realigned.bam</pre>	<pre>\$driver -r \$fasta -t 24 -i deduped.bam --algo Realigner -k \$dbsnp_Mill realigned.bam</pre>
Base Quality Score Recalibration	<pre>java -jar GenomeAnalysisTK.jar -T BaseRecalibrator -nct 24 -R \$fasta -I realigned.bam -knownSites \$dbsnp -knownSites \$dbsnp_Mill -o recal.table java -jar GenomeAnalysisTK.jar -T PrintReads -nct 24 -R \$fasta -I realigned.bam -BQSR recal.table -o recal.bam java -jar GenomeAnalysisTK.jar -T BaseRecalibrator -nct 24 -R \$fasta -I realigned.bam -knownSites \$dbsnp -knownSites \$dbsnp_Mill -BQSR recal.table -o after_recal.table java -jar GenomeAnalysisTK.jar -T AnalyzeCovariates -R \$fasta -before recal.table -after after_recal.table -plots recal_plots.pdf</pre>	<pre>\$driver -r \$fasta -t 24 -i realigned.bam --algo QualCal -k \$dbsnp -k \$dbsnp_Mill recal_data.table \$driver -r \$fasta -t 24 -i realigned.bam -q recal_data.table --algo QualCal -k \$dbsnp -k \$dbsnp_Mill --pre recal_data.table --csv recal.csv recal_data.table.post python \$dir/bin/plot.py b24qsr -o recal_plots.pdf recal.csv</pre>
Variant calling – HaplotypeCaller	<pre>java -jar GenomeAnalysisTK.jar -T HaplotypeCaller -nct 24 -R \$fasta -I recal.bam -o HC.vcf</pre>	<pre>\$driver -r \$fasta -t 24 -i realigned.bam -q recal_data.table --algo Haplotype output-hc.vcf --algo ReadWriter recal.bam</pre>
Variant calling – UnifiedGenotyper	<pre>java -jar GenomeAnalysisTK.jar -T UnifiedGenotyper -nt 24 -R \$fasta -I recal.bam -o UG.vcf -glm BOTH</pre>	<pre>\$driver -r \$fasta -t 24 -i realigned.bam -q recal_data.table --algo Genotyper output-ug.vcf</pre>

Appendix 3: Runtime data per stage (in minutes)

Sample Name	Type	Sequenced Bases	Sequencing Coverage	Stage	Sentieon Runtime on server	GATK Runtime on server
SRR098416	Exome	145M	3×	Sample metrics calculation	0.2	2.2
				Duplicate read removal	0.3	1.8
				Indel realignment	0.2	28.8
				Base Quality Score Recalibration	2.1	21.0
				Variant calling – UnifiedGenotyper	0.3	10.5
				Variant calling – HaplotypeCaller	0.5	54.5
SRR742200	Exome	5G	102×	Sample metrics calculation	0.5	14.6
				Duplicate read removal	1.3	36.3
				Indel realignment	1.5	69.2
				Base Quality Score Recalibration	2.5	219.7
				Variant calling – UnifiedGenotyper	0.7	15.2
				Variant calling – HaplotypeCaller	5.4	213.1
SRR702068	Exome	7G	140×	Sample metrics calculation	0.7	20.7
				Duplicate read removal	1.9	55.5
				Indel realignment	1.9	86.7
				Base Quality Score Recalibration	3.9	297.6
				Variant calling – UnifiedGenotyper	0.8	14.7
				Variant calling – HaplotypeCaller	5.7	204.1
ERR034544	Exome	11G	251×	Sample metrics calculation	0.9	31.7
				Duplicate read removal	2.8	81.2
				Indel realignment	3.1	123.0
				Base Quality Score Recalibration	4.5	478.3
				Variant calling – UnifiedGenotyper	1.2	18.1
				Variant calling – HaplotypeCaller	9.9	370.5
SRR098401	Exome	16G	341×	Sample metrics calculation	1.0	31.4
				Duplicate read removal	2.8	80.9
				Indel realignment	4.6	140.4
				Base Quality Score Recalibration	4.6	406.9
				Variant calling – UnifiedGenotyper	1.2	23.1
				Variant calling – HaplotypeCaller	15.7	748.7
SRR098359	Exome	16G	347×	Sample metrics calculation	1.6	56.4
				Duplicate read removal	4.9	146.1
				Indel realignment	7.9	284.5
				Base Quality Score Recalibration	8.3	727.4
				Variant calling – UnifiedGenotyper	1.8	27.5
				Variant calling – HaplotypeCaller	24.6	912.0
SRR622461	Whole Genome	17G	6×	Sample metrics calculation	1.4	46.6
				Duplicate read removal	3.9	114.0
				Indel realignment	6.9	202.5
				Base Quality Score Recalibration	7.6	748.2
				Variant calling – UnifiedGenotyper	2.0	31.5
				Variant calling – HaplotypeCaller	46.6	1076.9
ERR091571	Whole Genome	41G	14×	Sample metrics calculation	3.1	111.4
				Duplicate read removal	11.0	296.9
				Indel realignment	18.0	448.6
				Base Quality Score Recalibration	15.8	1811.0
				Variant calling – UnifiedGenotyper	5.2	47.8
				Variant calling – HaplotypeCaller	149.7	1310.3

Appendix 4: Run to run differences in GATK due to downsampling

The run-to-run variation caused by re-running GATK 3.3 HaplotypeCaller 20 times is plotted below. Each run was compared to all other runs using the program VarSeq™ to determine the number of variants called in the run but not appearing in the other runs. The range of the variation in number of calls and the mean plus 1 standard deviation of the variation are shown. The difference between each run and the Sentieon DNA run is within the statistical variation due to the run-to-run differences.

