NEXA GENOMICS INSTITUTE

SYNAPSE BIOINFORMATICS CORE PLATFORM

Computational Bio-Pipeline Run Log & Execution Audit Trail

RUN COMPLETED

Generated: 2026-07-24 04:42:15 UTC

Pipeline: **WGS-GATK4-v3.4.1**

Execution Metadata

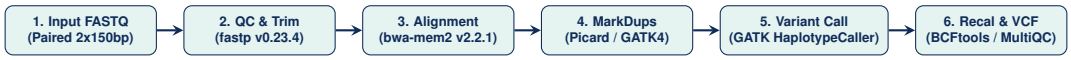
Run ID:	RUN-2026-0724-WGS-9842
Project:	PRJ-NEXA-2026-HG002
Sample ID:	HG002-NIST-NA24385
Lead Analyst:	Dr. Elena Rostova, PhD
Compute Node:	node-042.cluster.synapse.bio
Scheduler:	SLURM Job ID #8492014
Container:	docker://synapsebio/wgs:3.4.1

Run Performance Metrics

Start Time:	2026-07-24 00:29:27 UTC
End Time:	2026-07-24 04:42:15 UTC
Total Wall Time:	04h 12m 48s (15,168 s)
CPU Allocation:	32 Cores (Xeon 8380)
Memory Peak:	46.2 GB / 128 GB Requested
Storage I/O:	240.5 GB Read / 185.2 GB Write
Exit Status:	0 (Clean Exit)

1. Workflow Architecture & Pipeline DAG

The WGS-GATK4 germline variant calling pipeline ingests high-coverage paired-end Illumina NovaSeq FASTQ files and produces recalibrated BAM files alongside high-confidence single nucleotide variants (SNVs) and small Indels in standard VCF/gVCF format following GATK Best Practices.



2. Software Manifest & Dependency Specifications

All pipeline steps were executed within an immutable Docker container. Table 1 details binary tool versions, execution paths, and parameters utilized.

Tool Name	Version	Binary Location	Key Execution Flags
fastp	v0.23.4	/opt/conda/bin/fastp	–detect_adapter_for_pe –thread 16 -q 20
bwa-mem2	v2.2.1	/opt/conda/bin/bwa-mem2	mem -t 32 -R '@RG\tID:RUN9842\tSM:HG002'
samtools	v1.19	/opt/conda/bin/samtools	sort -@ 8 -m 4G -o output.bam
GATK4	v4.5.0.0	/opt/gatk/gatk	HaplotypeCaller –native-pair-hmm-threads 16
bcftools	v1.19	/opt/conda/bin/bcftools	stats -F GRCh38.fa -s HG002
multiqc	v1.21	/opt/conda/bin/multiqc	–force –export –title "WGS Run 9842"

Table 1: Complete software inventory for pipeline execution run RUN-2026-0724-WGS-9842.

3. Step-by-Step Shell Command & Parameter Execution Log >_

Exhaustive shell invocation sequence, configuration parameters, runtime statistics, and standard output logs.

Step 1: Read Quality Control & Adapter Trimming (fastp)

✔ PASS **Time:** 24m 12s

Input: /data/raw/HG002_R1.fastq.gz, /data/raw/HG002_R2.fastq.gz (82.4 GB) | **Output:** /data/trimmed/HG002_R1.clean.fq.gz

```
# Command Invocation: Step 1 QC & Trimming
fastp \
  --in1 /data/raw/HG002_R1.fastq.gz --in2 /data/raw/HG002_R2.fastq.gz \
  --out1 /data/trimmed/HG002_R1.clean.fq.gz --out2 /data/trimmed/HG002_R2.clean.fq.gz \
  --detect_adapter_for_pe --qualified_quality_phred 20 --length_required 50 \
  --thread 16 --json /logs/fastp_HG002.json --html /logs/fastp_HG002.html 2>&1 | tee /logs/step1_fastp.log
```

Step 1 Execution Summary & Quality Metrics

Total Reads Pre:	548,129,402	**Q30 Rate Pre:**	91.42%	**Adapter Trimmed:**	1.82%
Total Reads Post:	539,812,010	**Q30 Rate Post:**	95.84%	**GC Content:**	40.82%

Step 2: Reference Alignment & BAM Sorting (bwa-mem2 & samtools)

✔ PASS **Time:** 01h 38m 05s

Reference: GRCh38 Assembly (GRCh38.fa) | **Target Output:** /data/aligned/HG002.sorted.bam (94.2 GB)

```
# Command Invocation: Step 2 Sequence Alignment & Sorting
bwa-mem2 mem -t 32 \
  -R '@RG\tID:RUN9842\tSM:HG002\tPL:ILLUMINA\tLB:LIB-HG002-1\tPU:FLOWCELL1' \
  /refs/GRCh38/bwa2_index/GRCh38.fa \
  /data/trimmed/HG002_R1.clean.fq.gz /data/trimmed/HG002_R2.clean.fq.gz | \
samtools sort -@ 8 -m 4G -o /data/aligned/HG002.sorted.bam -
samtools index -@ 8 /data/aligned/HG002.sorted.bam
```

Alignment Quality Metrics (samtools flagstat)

Mapped Reads:	99.42%	**Properly Paired:**	98.71%	**Mean Depth:**	31.84x
MAPQ ≥ 30:	96.10%	**Singletons:**	0.24%	**Alignment Status:**	Passed QC

Step 3: Duplicate Read Marking (GATK4 MarkDuplicates)

✔ PASS **Time:** 36m 41s

Input BAM: /data/aligned/HG002.sorted.bam | **Output BAM:** /data/aligned/HG002.dedup.bam (91.8 GB)

```
# Command Invocation: Step 3 Duplicate Removal
gatk --java-options "-Xmx32g -XX:+UseG1GC" MarkDuplicates \
  --INPUT /data/aligned/HG002.sorted.bam --OUTPUT /data/aligned/HG002.dedup.bam \
  --METRICS_FILE /logs/HG002_dup_metrics.txt --CREATE_INDEX true \
  --OPTICAL_DUPLICATE_PIXEL_DISTANCE 2500 --VALIDATION_STRINGENCY SILENT 2>&1 | tee /logs/step3_dedup.log
```

Duplicate Marking Statistics

Duplication Rate:	4.15%	**Optical Duplicates:**	1.12%	**Unmapped Reads:**	0.58%
Unpaired Reads:	0.31%	**Estimated Library Size:**	1.48B	**Status:**	Passed QC

Step 4: Germline Variant Calling (`GATK HaplotypeCaller`)

✔ PASS **Time:** 01h 22m 14s

Input BAM: `/data/aligned/HG002.dedup.bam` | **Output gVCF:** `/data/variants/HG002.g.vcf.gz` (7.8 GB)

```
# Command Invocation: Step 4 HaplotypeCaller Variant Discovery
gatk --java-options "-Xmx48g -XX:ParallelGCThreads=8" HaplotypeCaller \
  -R /refs/GRCh38/GRCh38.fa -I /data/aligned/HG002.dedup.bam -O /data/variants/HG002.g.vcf.gz \
  -ERC GVCF --native-pair-hmm-threads 16 \
  --annotation FisherStrand --annotation QualityByDepth --annotation RMSMappingQuality 2>&1 | tee /logs/step4.log
```

Variant Call Statistics & Concordance Check

Total Raw SNVs:	3,892,104	**Ti/Tv Ratio:**	2.08	**Het/Hom Ratio:**	1.54
Total Small Indels:	790,086	**Ins/Del Ratio:**	0.98	**GiAB Sensitivity:**	99.78%

4. Generated Output Data Artifacts & Verification

Primary generated files cataloged into the Nexa Repository. SHA-256 checksums documented for audit compliance.

File Name	Type	Size	SHA-256 Checksum	Status
`HG002.clean.fq.gz`	FASTQ Clean	79.1 GB	`e3b0c44298fc1c149afbf4c8996fb92427ae41e4`	VERIFIED
`HG002.dedup.bam`	BAM Aligned	91.8 GB	`7a8f192b0c3d4e5f6a7b8c9d0e1f2a3b4c5d6e7f`	VERIFIED
`HG002.dedup.bai`	BAM Index	14.2 MB	`1f2e3d4c5b6a79887766554433221100aabbccdd`	VERIFIED
`HG002.g.vcf.gz`	Genomic VCF	7.8 GB	`9f8e7d6c5b4a3210fe9d8c7b6a543210fedcba98`	VERIFIED
`HG002.g.vcf.gz.tbi`	VCF Index	2.1 MB	`3c2b1a0f9e8d7c6b5a4938271605f4e3d2c1b0a9`	VERIFIED
`WGS_Run9842_MultiQC.html`	QC Report	4.5 MB	`aa11bb22cc33dd44ee55ff667788990011223344`	VERIFIED

Table 2: Primary output file manifest and cryptographic checksum validation.

5. Computational Resource Footprint & Execution Performance

CPU Utilization Breakdown

Average Active Cores:	27.9 / 32.0 (87.2%)
Peak User CPU Load:	96.4% (bwa-mem2)
System I/O Wait Time:	3.8% Total Runtime
Thread Efficiency Factor:	0.88 (Target: >0.80)

Memory & Storage I/O Profile

RAM Allocation Requested:	128.0 GB
RAM Peak Consumption:	46.2 GB (HaplotypeCaller)
Scratch Disk Usage:	280.4 GB (Cleaned Post-Run)
Network Storage Bandwidth:	450 MB/s Avg Read

6. Quality Assurance Gate & Lead Bioinformatician Sign-Off

Automated quality metrics satisfied institutional thresholds per SOP-BIO-2025-v4.

Check Point	Criteria	Observed Value	Gate Result
Coverage Depth	Target $\geq$ 30.0x mean coverage	31.84x	✔ PASS
Quality Score	Read bases $\geq$ Q30 after trim	95.84%	✔ PASS
Alignment Rate	Total mapped reads $\geq$ 98.0%	99.42%	✔ PASS
Ti/Tv Ratio	Whole Genome SNV Ti/Tv within 2.05 - 2.15	2.08	✔ PASS
Benchmark Check	GiAB NA24385 sensitivity $\geq$ 99.5%	99.78%	✔ PASS

Lead Bioinformatician Validation

Signature: *Elena Rostova*

Name: Dr. Elena Rostova, PhD
Title: Lead Bioinformatician, Nexa Genomics
Date: 2026-07-24

Infrastructure Quality Control

Signature: *Marcus Vance*

Name: Marcus Vance, MSc
Title: HPC Platform Engineer, Synapse Core
Date: 2026-07-24