

Prospectively frozen external benchmarking reveals benchmark-context dependence in somatic variant calling

Md Mahmudul Hasan

mhasan222@gmail.com

Ambiospire LLC <https://orcid.org/0000-0001-9327-0437>

Research Article

Keywords: somatic variant calling, whole-genome sequencing, cancer genomics, benchmarking, HG008, HCC1395, Mutect2, Strelka2, DeepSomatic, reproducibility

Posted Date: October 5th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11210923/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare potential competing interests as follows: Md Mahmudul Hasan is the founder of AmbioSpire LLC, the organization under which the P003 research program was developed.

Abstract

Somatic variant benchmarks can be difficult to interpret when analytical choices change after external truth is examined. We froze caller configurations, normalized exact allele matching, benchmark regions, consensus rules and spatial-bootstrap parameters after HCC1395/SEQC2 development and before protected HG008 evaluation. In the primary HG008 region, DeepSomatic and majority consensus achieved F1 scores of 0.8973 and 0.9136; their paired difference was concentrated in INDELs (delta F1, - 0.0631) rather than SNVs (- 0.0002). Analyses prespecified before HG008 access but executed after primary evaluation linked ensemble behavior to caller-support patterns and showed variation by INDEL length and benchmark context. The study does not establish clinical validity or general transportability. It shows how prospective freezing, support decomposition and spatial uncertainty can make benchmark-specific performance more interpretable.

Introduction

Somatic small-variant callers differ in statistical model, local assembly, learned representation and filtering strategy. Their measured performance can therefore change with variant class, tumor-normal characteristics, sequencing conditions, genome representation and truth-set construction¹⁻³. Most benchmarking describes performance within a system. Less is known about how a prospectively fixed analytical procedure behaves when applied to a different benchmark context without retuning.

The distinction matters because many reasonable analytical choices remain available after calls are generated. Normalization, callable-region selection, caller filtering, ensemble thresholds and comparison rules can all change reported performance. Such iteration can support method development, but it weakens the evidentiary value of an external resource as a test of a previously specified procedure. Representation is also consequential: exact allele matching is deterministic, yet complex variants with equivalent haplotypic interpretations can be expressed differently^{4,5}. The matching rule and benchmark region are therefore part of the evaluated method.

Uncertainty estimates are likewise conditional on design. Variant-level resampling treats events as exchangeable even though alignment difficulty, sequence context and benchmark accessibility are spatially structured. Genomic block resampling preserves some local dependence, but its intervals still describe the evaluated genomic domain rather than a population of tumors. A benchmark comparison is most interpretable when the estimand, comparison region, matching rule and resampling unit are specified together and their scope is kept separate from patient- or population-level claims.

Deterministic ensembles add another layer. Union, majority and strict-intersection rules are fixed by support thresholds, but their performance depends on the joint error structure of the component callers. Prior work has compared individual somatic callers and voting ensembles across benchmark datasets⁶. An aggregate score alone, however, does not reveal which support patterns enter or leave an ensemble as its threshold changes.

These concerns motivate two complementary requirements. First, the analytical procedure should be fixed before an external truth resource becomes available for inspection. Second, the resulting performance should be decomposed in ways that remain determined by the frozen procedure rather than selected after outcomes are known. In this setting, variant-class contrasts and complete caller-support patterns can explain aggregate behavior without learning a new model or tuning a new threshold on the external benchmark.

We used a protected external-test design to examine these issues. A three-caller procedure comprising Mutect2, Strelka2 and DeepSomatic was developed with HCC1395/HCC1395BL and SEQC2 resources^{1,2}. Before protected HG008 truth access, we fixed the scientific definitions, caller configurations, normalized exact CHROM/POS/REF/ALT comparator, benchmark-region roles, primary caller, consensus rules, endpoint and spatial-bootstrap parameters. We then evaluated the unchanged procedure against the NIST HG008 somatic small-variant benchmark⁷⁻⁹. The objective was to determine how the frozen procedures behaved in the external benchmark and which variant-class and caller-support structures accounted for their aggregate results.

Results

Prospectively frozen design separates development from external evaluation

HCC1395/HCC1395BL and SEQC2 v1.2.1 resources were used for development, whereas HG008 was reserved for external evaluation. The external analysis used the NIST HG008-T draft clonal/truncal somatic small-variant benchmark v0.3. Scientific definitions were fixed before protected HG008 truth access, and protected benchmark bytes were acquired only after external evaluation was opened. The retained evidence showed no HG008-driven change to callers, filters, normalization, region roles, ensemble thresholds, the primary method or bootstrap parameters (Fig. 1).

DeepSomatic was the frozen primary analytical caller; Mutect2 and Strelka2 were individual comparators. P003 performed no DeepSomatic training or tuning. The deterministic ensembles used fixed Mutect2/Strelka2/DeepSomatic order: union required at least one caller, majority at least two and strict intersection all three. The primary region was HG008 v0.3 nogermelineinterference; HG008 v0.3 all was the prespecified sensitivity region; and nogermelinewithin50bp was prespecified exploratory.

The primary endpoint was F1 under normalized exact CHROM/POS/REF/ALT identity. Spatial uncertainty used 5-Mb genomic blocks, 10,000 bootstrap replicates, seed 3003 and linear percentile intervals. These scientific definitions were frozen prospectively. The exact historical external scoring and bootstrap source bytes were not retained in the original checksum-bound pretruth archive. Retrospectively recovered command evidence was consistent with the frozen definitions, and later internal independent verification reproduced the reported numerical results, but that later evidence does not establish prospective custody of the original source bytes.

Caller and deterministic-consensus performance differ in the HG008 external benchmark

The primary region contained 14,014 truth variants, comprising 9,078 SNVs and 4,936 INDELs. The six procedures occupied distinct precision–recall positions (Fig. 2 and Extended Data Table 1). Mutect2 produced 12,027 true positives, 752 false positives and 1,987 false negatives (precision 0.9412, recall 0.8582, F1 0.8978). Strelka2 produced 11,757 true positives, 4,414 false positives and 2,257 false negatives (precision 0.7270, recall 0.8389, F1 0.7790). DeepSomatic produced 11,504 true positives, 124 false positives and 2,510 false negatives (precision 0.9893, recall 0.8209, F1 0.8973; 95% block-bootstrap CI, 0.8929–0.9015).

Union increased recall to 0.9148 but reduced precision to 0.7125 (F1 0.8011). Strict intersection reached precision 0.9995 but recall 0.7555 (F1 0.8606). Majority yielded 11,880 true positives, 112 false positives and 2,134 false negatives (precision 0.9907, recall 0.8477, F1 0.9136; 95% CI, 0.9096–0.9175). Majority therefore had the highest F1 among the six procedures in this primary benchmark region, but the result is not a universal method ranking.

The components of F1 clarify the trade-offs. Union recovered more truth variants while admitting many calls supported by one method. Strict intersection removed nearly all false positives but missed more truth variants. Majority retained high precision and more truth variants than strict intersection. P003 did not optimize a clinical utility or downstream loss function, so these operating points should not be converted into a deployment recommendation.

The result also separates caller identity from aggregation strategy. DeepSomatic and Mutect2 had similar overall F1 values despite different balances of precision and recall, whereas Strelka2 contributed both true variants and a large number of false-positive calls to the union. Majority did not create a new probabilistic model; it excluded single-caller patterns by rule. Reporting counts and operating points together therefore avoids treating a higher benchmark F1 as evidence that one software architecture is generally preferable.

The DeepSomatic–majority difference is concentrated in INDELs

The frozen paired comparison used identical resampled genomic blocks for DeepSomatic and majority. The overall DeepSomatic-minus-majority delta F1 was -0.0164 (95% CI, -0.0185 to -0.0142 ; Fig. 3). For SNVs, DeepSomatic F1 was 0.9844, majority F1 was 0.9846 and the paired delta was -0.0002 (95% CI, -0.0019 to 0.0015). For INDELs, DeepSomatic F1 was 0.6895 and majority F1 was 0.7526, yielding a paired delta of -0.0631 (95% CI, -0.0699 to -0.0562).

The overall contrast thus combined near-equal SNV performance with a larger INDEL difference. Because the primary region contained more SNVs than INDELs, the aggregate estimate alone obscured this class-specific pattern. The intervals describe spatial variation across genomic blocks within the defined benchmark; they are not confidence intervals for a population of tumors, patients, laboratories or sequencing protocols.

This localization is important for interpretation of the primary endpoint. The overall paired difference was not evidence of a uniform shift across small variants. Instead, it arose from a variant class in which representation, alignment and local assembly can be especially consequential. The exact-match design cannot determine how much of the INDEL contrast reflects biological detection versus representation discordance, so the class-specific result motivates caution rather than a causal explanation.

Caller-support structure explains deterministic ensemble behavior

These analyses were prespecified before HG008 access and executed after completion of the primary external evaluation. All eight support patterns were retained in fixed Mutect2/Strelka2/DeepSomatic bit order, in which 1 denotes support by a caller and 0 denotes no support; thus 000 denotes a truth variant missed by all three callers (Fig. 4 and Extended Data Table 2).

The 000 pattern contained 1,194 false negatives, including 1,081 INDELs. The fully concordant 111 pattern contained 10,588 true positives and 5 false positives. The Strelka2-only 010 pattern contained 135 true positives and 4,311 false positives. The other single-caller patterns were smaller: 001 contained 57 true positives and 85 false positives, and 100 contained 748 true positives and 665 false positives. In this benchmark, single-caller support—especially 010—accounted for much of the false-positive burden, whereas 111 identified a highly concordant subset.

The same table provides a direct accounting of the consensus rules. Union includes every pattern except 000; majority includes 011, 101, 110 and 111; strict intersection includes only 111. These thresholds reconciled exactly to the frozen aggregate consensus counts. Support is not a calibrated probability or biological truth label: a one-caller pattern can contain true events, and a three-caller pattern can contain errors. Its value here is to make deterministic thresholding inspectable.

INDEL length and germline context refine the external benchmark pattern

These analyses were also prespecified before HG008 access and executed after completion of the primary external evaluation. The frozen nonmissing INDEL-length bins contained 4,456 one-base-pair variants, 322 variants of 2–5 bp, 112 variants of 6–15 bp and 46 variants longer than 15 bp (Fig. 5a and Extended Data Table 3). Majority F1 exceeded DeepSomatic F1 in each bin: 0.7380 versus 0.6707, 0.8200 versus 0.7971, 0.9772 versus 0.9401 and 0.9438 versus 0.8298, respectively. The > 15-bp bin was small and is descriptive; the strata do not replace the primary endpoint.

The predefined germline-context partition used the locked all region as its universe. “Absent” comprised records retained by the locked nogermelineinterference region, “present” comprised records in all but not in nogermelineinterference, and unclassifiable records remained missing. It did not substitute the separate nogermelinewithin50bp resource or derive a distance threshold from outcomes. All six procedures had lower F1 in the present stratum. DeepSomatic F1 was 0.0528 in the present stratum and 0.8973 in the

absent stratum; majority F1 was 0.0860 and 0.9136, respectively (Fig. 5b and Extended Data Table 4). These resource-defined strata differ in multiple coupled properties, so the contrast does not identify nearby germline variation as the cause.

Sensitivity analyses show benchmark-context dependence without establishing population generalization

The `nogermlinewithin50bp` analysis was prespecified exploratory. Its overall F1 values ranged from 0.8990 for strict intersection to 0.9487 for majority; full procedure-specific results are reported in Supplementary Table 1 and are not promoted to primary or confirmatory evidence.

In the prespecified all-region sensitivity analysis, overall F1 was 0.7917 for Mutect2, 0.6976 for Strelka2, 0.7756 for DeepSomatic, 0.7297 for union, 0.7940 for majority and 0.7387 for strict intersection. The prespecified all-region uncertainty analysis, executed after the primary external evaluation, produced 95% block-bootstrap intervals of 0.7854–0.7980, 0.6691–0.7205, 0.7689–0.7822, 0.7018–0.7521, 0.7875–0.8003 and 0.7316–0.7456, respectively (Supplementary Table 2). These results show sensitivity to benchmark-region definition without changing the primary region or endpoint.

For descriptive context only, completed HCC1395 and HG008 F1 values were compared post hoc. The procedures did not show a common direction or magnitude of change across the two systems (Supplementary Fig. 1). Because the systems differ in samples, data generation, truth construction and evaluated regions, this comparison is not an inferential estimate of a single transportability effect and does not establish population-level generalization.

Discussion

P003 supports a general methods lesson: a benchmark conclusion depends not only on the caller, but also on benchmark context, variant class and aggregation rule. Prospectively fixing the analysis before protected external truth access provides a stronger test of this dependence than retrospectively retuning against the external resource. In HG008, majority had a higher benchmark-conditional F1 than the frozen primary DeepSomatic procedure, but the paired difference was near zero for SNVs and concentrated in INDELs. Support-pattern accounting then showed which subsets entered or left each deterministic ensemble.

Prospective freezing narrows the analytical choices that remain after truth is examined. It does not make the frozen choices uniquely correct, nor does it remove benchmark bias. Instead, it preserves the external resource's role as a test of an existing procedure. This principle is relevant to comparative computational studies in which filtering, normalization, region selection, matching and uncertainty definitions can all change the reported result.

The design also changes how a negative or mixed external result should be used. If the external benchmark exposes lower performance, a rank change or a class-specific failure, the immediate response

should be to report and diagnose that result rather than revise the frozen procedure and continue calling the same resource an external test. Improvements motivated by the external result can be valuable, but they belong to a new development lineage with another protected evaluation. This separation makes methodological learning compatible with honest external validation.

The class-specific result argues for reporting beyond an aggregate F1 score. Small-variant benchmarks can be numerically dominated by SNVs even when INDEL behavior differs materially. The late-executed bounded analysis showed the DeepSomatic–majority pattern across the predefined nonmissing INDEL-length bins, although the longest bin was small. Such decomposition is secondary evidence, not a new primary endpoint, but it makes the aggregate contrast more interpretable.

Caller-support analysis similarly clarifies what a deterministic ensemble does. Majority voting itself is not novel, and prior studies have compared voting ensembles across datasets⁶. Here, the contribution is the combination of a protected, prospectively frozen external test with complete support-pattern accounting. The large false-positive burden in 010 and the high concordance of 111 explain the observed threshold trade-offs for these callers in this benchmark. They do not show that majority will outperform learned, weighted or representation-aware ensembles elsewhere.

Benchmark context also mattered. Performance differed between the predefined germline-resource strata, the broader all-region sensitivity analysis and the within50bp exploratory region. These resources encode correlated biological, technical and truth-set differences. No single factor can be assigned a causal role. The post-hoc HCC1395-to-HG008 comparison likewise describes two nonexchangeable systems rather than estimating performance across a patient population.

These findings do not imply that benchmark systems are arbitrary. Carefully constructed truth sets remain essential for measuring errors under defined conditions. The point is that an observed score inherits the resource’s variant composition, regions and comparison conventions. Methods papers should therefore report enough structure to distinguish a stable procedure from a stable rank, and a benchmark-specific interval from an uncertainty statement about future samples.

The primary comparator was normalized exact CHROM/POS/REF/ALT identity. This rule is transparent and auditable, but it can classify representation-equivalent complex alleles as discordant^{4,5}. Aardvark, hap.py, scmp-somatic and other representation-aware comparisons were not performed. Any such analysis would be complementary post-hoc sensitivity work in a separate scientific lineage.

The study also separates scientific reproducibility from archival completeness. Scientific definitions were frozen before HG008 access, but the exact historical external scoring and bootstrap source was not retained in the original checksum-bound pretruth archive. Retrospective command evidence was consistent with the frozen definitions, and a separate internal implementation reproduced the primary counts, consensus arithmetic, selected bootstrap results and bounded secondary outputs. That review strengthens numerical traceability but is not external laboratory replication and does not retroactively establish custody of the original source bytes.

Independent arithmetic reproduction is useful because it tests whether retained inputs and machine-readable outputs support the reported numbers without relying on the original calculator. It cannot recreate unavailable historical execution details or establish biological replication. Treating design reproducibility, input identity, numerical reproducibility and executable custody as separate dimensions provides a more precise account than a single undifferentiated reproducibility claim.

The evidence remains bounded. It covers one development and one external tumor–normal benchmark system, not patient-level or population-level validation. The systems differ along multiple dimensions, preventing causal attribution of their differences. Representation-aware sensitivity analysis was not performed. DeepSomatic was externally developed and unchanged within P003, but independence of its upstream training from HCC1395/SEQC2 is not established. Prespecified VAF, tumor-depth, normal-depth, mapping-quality and homopolymer strata were deferred because adequate retained resource or implementation provenance was unavailable; tandem-repeat analysis remained unresolved. Runtime observations are descriptive operational telemetry, not a controlled computational-performance comparison.

Future work should preserve exact executable custody prospectively, use multiple protected external tumor–normal systems and predeclare representation-aware sensitivity analyses. Within the present boundaries, P003 shows why external benchmarks are most informative when they remain tests rather than tuning opportunities and when aggregate scores are decomposed into the variant classes and support structures that produced them.

Online Methods

Study design and prospective freeze

P003/T001 was a computational benchmark with separate development and external-evaluation roles. HCC1395/HCC1395BL and SEQC2 v1.2.1 were used for development^{1,2}; HG008 was reserved for protected external evaluation^{7–9}. Before HG008 truth access, the scientific specification fixed caller configurations, PASS filtering, normalization, normalized exact allele comparison, benchmark-region roles, deterministic ensembles, primary endpoint and bootstrap parameters. No HG008 result was used to change these choices.

Development and external resources

Development used HCC1395 tumor and HCC1395BL matched-normal data with SEQC2 v1.2.1 truth and high-confidence resources^{1,2}. HCC1395 read accessions were SRR7890825 and SRR7890828. Development-to-external comparisons are post-hoc descriptive because HCC1395 and HG008 are nonexchangeable benchmark systems.

External evaluation used the NIST HG008 matched tumor–normal resource^{7,8} and the HG008-T Draft Clonal/Truncal Somatic Small Variant Benchmark v0.3, release identity NIST_HG008-T_somatic-smvar_DraftBenchmark_V0.3-202604259. The truth VCF SHA-256 was

6ff5b215e0dc634c85da6164cffbd5790233258fcc48c88a598331430b061cad. The primary nogermlineinterference, sensitivity all and exploratory nogermlinewithin50bp BED SHA-256 values were cad04330da8a2966410824f7d3101172fa30d3c4dbcd1abbf3c30730539f5ea1, c6100933430322eb62c64e826a349c1861e26a029853502bd247f86d27e22726 and 08fc5974466c63db61cb45628272a5a73a7aa8ca6cd07fe87b0f6085e126d2a9. Protected bytes are not redistributed.

Reference and alignment

HG008 reads were aligned to GRCh38_GIABv3_no_alt_analysis_set_maskedGRC_decoys_MAP2K3_KMT2C_KCNJ18 (SHA-256 0713026dfdd61fb1cd68be4b9689a9e5c55c8708a3843248b0b574506fd0f0e9) with BWA-MEM2 release v2.3 (ref. 10). An approved reporting exception records that the embedded version reporter returned 2.2.1; the source release, upstream commit 7aa5ff6c3330490e5629ab9b7327683d2dce02d6 and source-archive SHA-256 support v2.3. The final runtime registry digest and a complete public alignment command/flag record were not retained in the locked release.

Variant callers and retained software identities

Mutect2 used GATK 4.6.2.0 (ref. 11) in broadinstitute/gatk:4.6.2.0 with digest sha256:71b17ee42d149e8ec112603f5305c873ab60d93949ef8bb62a4fff85427f56fb4. Strelka2 used version 2.9.10 (ref. 12) in quay.io/biocontainers/strelka:2.9.10-h9ee0642_1 with digest sha256:e6133fa935d444aaf3aa04b57256a580d9b3498298fc17cb5f4a9818116105ec5. DeepSomatic used version 1.10.0 (ref. 13) in google/deepsomatic:1.10.0 with digest sha256:6cd98c41781cdc1a2916d8029bedef70a57b87d41f933bd4cf259c80f291e27b6. Its bundled whole-genome model manifest SHA-256 was 3d909228d0224370959222cfc7f67d92ab76436d4455be92b9fb90cdb9ff7a66.

P003 performed no DeepSomatic training or tuning. HG008 was external to P003 development, but P003 does not claim that upstream DeepSomatic training was independent of HCC1395/SEQC2. The locked public release does not contain complete executable caller command lines, runtime flags, Strelka configuration steps or DeepSomatic sharding details. Version, container and model identities are retained; unavailable command detail is not reconstructed.

PASS filtering, normalization and exact comparison

Only PASS calls were evaluated. The retained workflow order was PASS restriction, sites-only conversion, splitting of multiallelic records, left normalization against the benchmark-matched reference and exact duplicate removal with bcftools 1.20 (ref. 14). The historical exact bcftools binary or container digest was not retained. Original caller VCF identities and normalized derived-VCF identities were checksum-bound. The frozen comparator required equality of CHROM, POS, REF and ALT after normalization. MNPs remained atomic after minimal representation. No representation-aware comparator was used.

Consensus definitions and metrics

Support used fixed Mutect2/Strelka2/DeepSomatic order. Union required support count ≥ 1 , majority ≥ 2 and strict intersection = 3. No weighting, adaptive threshold or post-external redefinition was used. TP denoted an allele present in call and truth sets, FP a called allele absent from truth and FN a truth allele absent from the call set within the evaluated region. Precision was $TP/(TP + FP)$, recall $TP/(TP + FN)$ and F1 their harmonic mean. Results were reported overall and for SNVs and INDELs.

Benchmark regions

The primary analysis used HG008 v0.3 nogermelineinterference. HG008 v0.3 all was the prespecified sensitivity region. nogermelinewithin50bp was prespecified exploratory and was not promoted to primary or confirmatory status.

Genomic block bootstrap and paired delta F1

The frozen primary bootstrap used 5,000,000-bp genomic blocks intersecting the benchmark, 10,000 replicates, seed 3003, 95% linear percentile intervals and a minimum defined-replicate fraction of 0.95. The retained artifact reports 601 eligible primary-region blocks. For DeepSomatic-minus-majority delta F1, each replicate used identical sampled blocks for both methods. The locked release does not retain block-level sufficient statistics or metadata sufficient to resolve the exact historical handling of contig boundaries, partial terminal blocks and empty intersections. Those details are an archival reproducibility limitation and are not inferred. The intervals quantify spatial variation within the evaluated benchmark, not patient-population uncertainty.

Prespecified analyses executed after primary evaluation

Five analyses were prespecified before HG008 access and executed after completion of the primary external evaluation: 000–111 caller-support patterns, INDEL-length stratification, nearby germline-interference stratification, the nogermelinewithin50bp exploratory analysis and the all-region genomic block bootstrap. Their execution source was frozen before that bounded execution, inputs and outputs were checksum-bound and later review confirmed conformance to the bounded scope. They provide secondary mechanistic, exploratory or sensitivity evidence and do not redefine the primary result.

Caller support retained all patterns in Mutect2/Strelka2/DeepSomatic order. 000 represented truth variants missed by all callers; no pattern performance score was created. INDEL length used frozen bins of 1 bp, 2–5 bp, 6–15 bp, > 15 bp and missing. Missing values remained explicit, and display cells with fewer than five observations followed the prespecified small-cell rule.

The nearby-germline analysis used the locked all-region resource as its universe. Absent meant membership in the locked nogermelineinterference BED; present meant membership in the locked all BED but not the primary BED; unresolved records remained missing. This partition was distinct from nogermelinewithin50bp, and no new distance threshold was derived from performance. The exploratory within50bp analysis used only its locked BED. The all-region bootstrap used the same block size, replicate count, seed, interval method and minimum defined fraction as the primary bootstrap and introduced no new pairwise contrast.

Internal independent numerical reproduction

An internal G10 reviewer used a separate minimal implementation that did not import the original scoring or bounded-analysis calculators. The review reproduced primary overall, SNV and INDEL integer counts; deterministic consensus counts; requested primary bootstrap intervals; and all five bounded result groups. No numerical mismatch or scientific retuning was found. This was internal procedural reproducibility review, not external biological or laboratory replication.

Executable-source custody limitation

Scientific definitions were prospectively frozen before HG008 access. The exact historical external scoring and bootstrap source was not retained within the original checksum-bound pretruth workflow archive. Retrospectively recovered command evidence was consistent with the frozen analytical definitions, and the later independent implementation reproduced the reported numbers. Neither fact establishes prospective custody of the exact historical source bytes.

Data Availability

The publication-supporting dataset and reproducibility package is available at Zenodo: <https://doi.org/10.5281/zenodo.22998005>. The archive contains release-safe derived aggregate source data, result tables, figure-source data, methods metadata, provenance summaries and checksums from the locked P003/T001 scientific publication release v1.0.0; it does not constitute a complete end-to-end executable reconstruction of the historical workflow. Protected HG008/NIST truth VCF and BED bytes, the benchmark-matched reference, FASTQ, BAM/CRAM, large caller VCFs and private custody records are not redistributed by AmbioSpire. Their checksums retain identity but neither confer access nor alter redistribution rights. HG008 resources remain available through NIST/GIAB and original providers subject to original terms. HCC1395 development reads are available under SRR7890825 and SRR7890828.

Code Availability

The publication-supporting Zenodo archive available at <https://doi.org/10.5281/zenodo.22998005> contains release-safe methods and workflow documentation, software identities, machine-readable aggregate results, source-data tables, checksums and provenance records. It contains no executable caller workflow, original external primary-scoring or bootstrap source, bounded-analysis source, or internal-verification source. The exact historical primary scoring and bootstrap source bytes were not retained in the original pretruth archive. Third-party tools remain available from their original providers. The MIT boundary applies only to original publication-specific P003 software if later intentionally included in a reviewed archive; it does not establish that executable code is present in this archive.

Declarations

Competing Interests

Md Mahmudul Hasan is the founder of AmbioSpire LLC, the organization under which the P003 research program was developed.

Funding

This research received no external funding specifically for P003.

Author Contributions

Md Mahmudul Hasan: Conceptualization; Methodology; Software; Validation; Formal analysis; Investigation; Data curation; Writing – original draft; Writing – review & editing; Visualization; Project administration; Resources.

Acknowledgements

The author has no additional acknowledgements.

References

1. Fang, L. T. et al. Establishing community reference samples, data and call sets for benchmarking cancer mutation detection using whole-genome sequencing. *Nat. Biotechnol.* 39, 1151–1160 (2021). <https://doi.org/10.1038/s41587-021-00993-6>
2. Xiao, W. et al. Toward best practice in cancer mutation detection with whole-genome and whole-exome sequencing. *Nat. Biotechnol.* 39, 1141–1150 (2021). <https://doi.org/10.1038/s41587-021-00994-5>
3. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* 31, 213–219 (2013). <https://doi.org/10.1038/nbt.2514>
4. Tan, A., Abecasis, G. R. & Kang, H. M. Unified representation of genetic variants. *Bioinformatics* 31, 2202–2204 (2015). <https://doi.org/10.1093/bioinformatics/btv112>
5. Krusche, P. et al. Best practices for benchmarking germline small-variant calls in human genomes. *Nat. Biotechnol.* 37, 555–560 (2019). <https://doi.org/10.1038/s41587-019-0054-x>
6. Guille, A. et al. A benchmarking study of individual somatic variant callers and voting-based ensembles for whole-exome sequencing. *Brief. Bioinform.* 26, bbae697 (2025). <https://doi.org/10.1093/bib/bbae697>
7. McDaniel, J. et al. Development and extensive sequencing of a broadly-consented Genome in a Bottle matched tumor-normal pair. *Sci. Data* 12, 1195 (2025). <https://doi.org/10.1038/s41597-025-05438-2>
8. Wagner, J. et al. A complete human pancreatic cancer genome. *bioRxiv* (2026). <https://doi.org/10.64898/2026.05.01.722316>

9. 9. National Institute of Standards and Technology. Cancer Genome in a Bottle: HG008-T Draft Clonal/Truncal Somatic Small Variant Benchmark v0.3, release NIST_HG008-T_somatic-smvar_DraftBenchmark_V0.3-20260425. <https://www.nist.gov/programs-projects/cancer-genome-bottle>; benchmark directory: https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data_somatic/HG008/Liss_lab/analysis/NIST_HG008-T_somatic-smvar_DraftBenchmark_V0.3-20260425/ (accessed 27 September 2026).
10. 10. Vasimuddin, M. et al. Efficient architecture-aware acceleration of BWA-MEM for multicore systems. In 2019 IEEE International Parallel and Distributed Processing Symposium (2019). <https://doi.org/10.1109/IPDPS.2019.00041>
11. 11. GATK Team. Mutect2. Genome Analysis Toolkit documentation for GATK 4.6.2.0, Broad Institute. <https://gatk.broadinstitute.org/hc/en-us/articles/35967636854939-Mutect2> (accessed 28 September 2026).
12. 12. Kim, S. et al. Strelka2: fast and accurate calling of germline and somatic variants. Nat. Methods 15, 591–594 (2018). <https://doi.org/10.1038/s41592-018-0051-x>
13. 13. Park, K. et al. Accurate somatic small variant discovery for multiple sequencing technologies with DeepSomatic. Nat. Biotechnol. 44, 1569–1578 (2026). <https://doi.org/10.1038/s41587-025-02839-x>
14. 14. Danecek, P. et al. Twelve years of SAMtools and BCFtools. GigaScience 10, giab008 (2021). <https://doi.org/10.1093/gigascience/giab008>
15. Extended Data

Figures

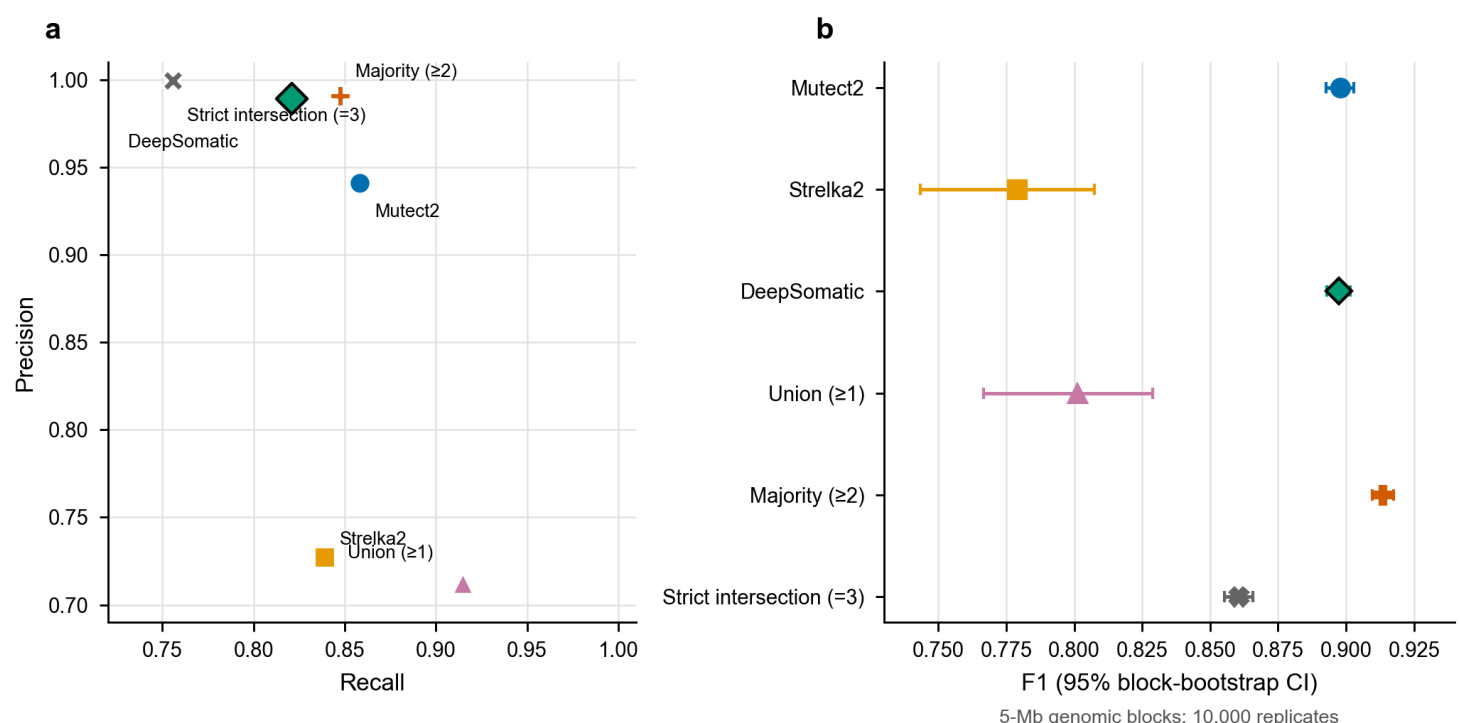


Figure 1

Figure 1 | Prospectively frozen external-test design. a, HCC1395/HCC1395BL and SEQC2 resources were used for development. Scientific definitions were fixed before protected HG008 v0.3 truth access; primary evaluation, prespecified secondary analyses executed after that evaluation and independent internal numerical verification followed. This chronology does not imply prospective custody of every executed source byte. b, Frozen components comprised three callers, deterministic union (support count ≥ 1), majority (≥ 2) and strict-intersection ($=3$) ensembles, normalized exact CHROM/POS/REF/ALT matching, and distinct primary, sensitivity and exploratory benchmark-region roles. G10 was an internal reproducibility review, not external laboratory replication

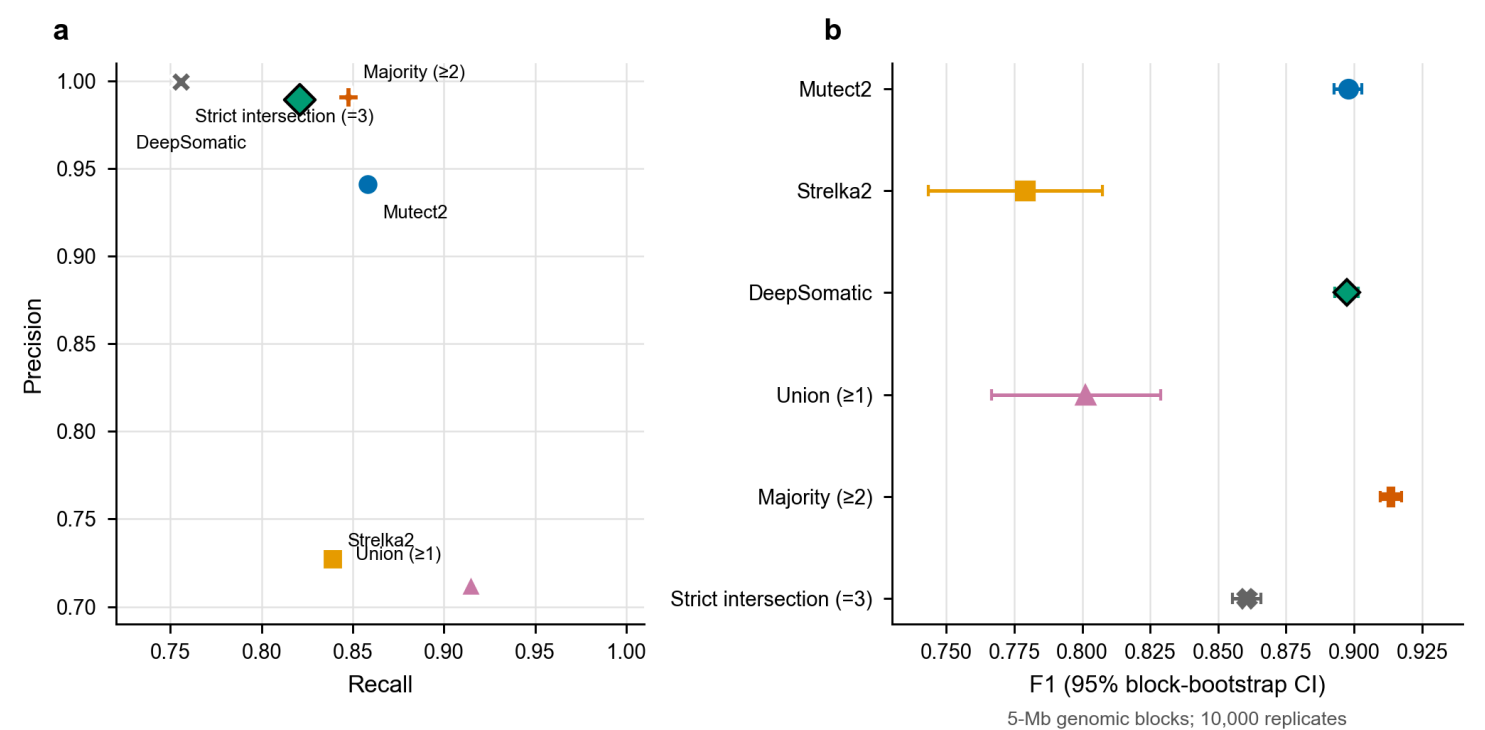


Figure 2

Figure 1 | Prospectively frozen external-test design. a, HCC1395/HCC1395BL and SEQC2 resources were used for development. Scientific definitions were fixed before protected HG008 v0.3 truth access; primary evaluation, prespecified secondary analyses executed after that evaluation and independent internal numerical verification followed. This chronology does not imply prospective custody of every executed source byte. b, Frozen components comprised three callers, deterministic union (support count ≥ 1), majority (≥ 2) and strict-intersection ($=3$) ensembles, normalized exact CHROM/POS/REF/ALT matching, and distinct primary, sensitivity and exploratory benchmark-region roles. G10 was an internal reproducibility review, not external laboratory replication

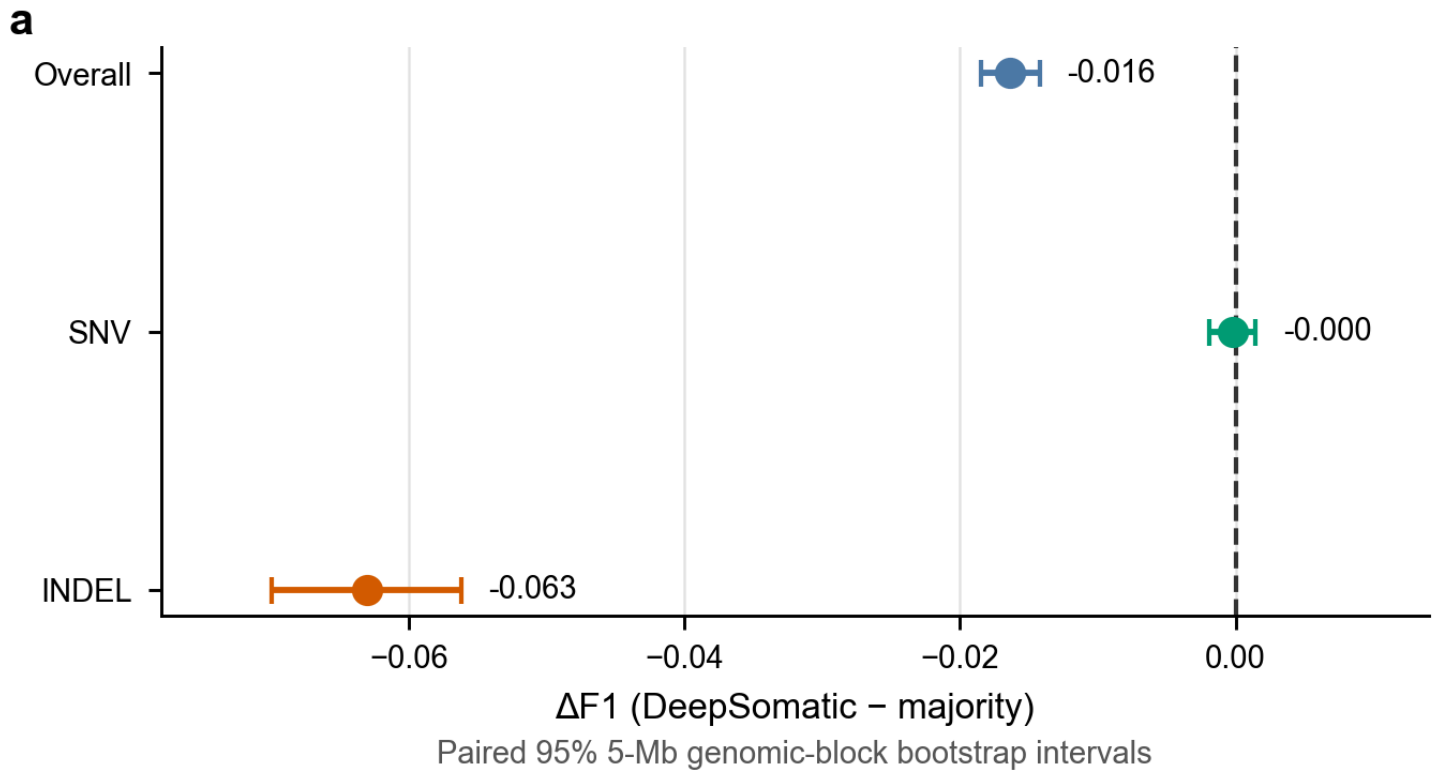


Figure 3

Figure 3 | Paired DeepSomatic-minus-majority differences by variant class. Overall, SNV and INDEL point differences are DeepSomatic F1 minus majority F1 in the HG008 v0.3 no germline interference region. Bars are frozen 95% paired block-bootstrap intervals using identical resampled 5-Mb blocks for both methods within each of 10,000 replicates (seed 3003). The vertical dashed line marks zero. No P values or additional method contrasts were calculated.

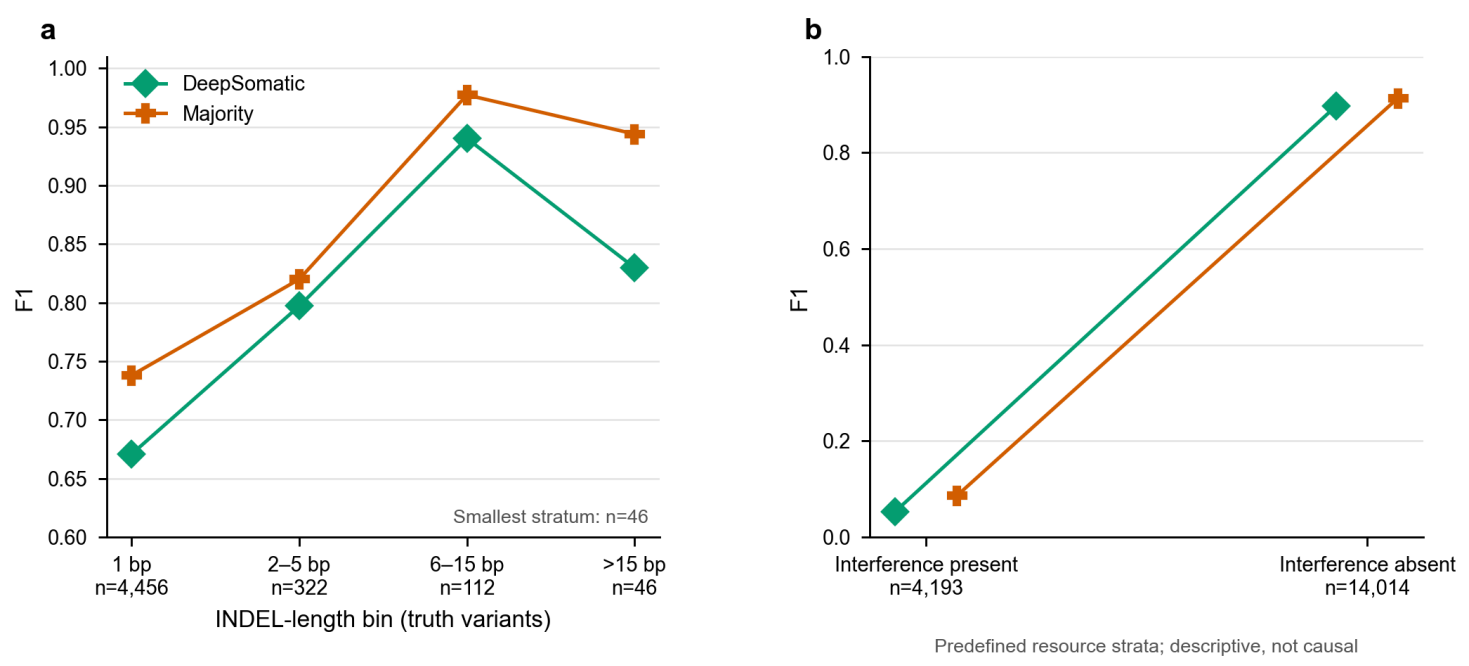


Figure 4

Figure 4 | Caller-support architecture in the primary HG008 region. These analyses were prespecified before HG008 access and executed after completion of the primary external evaluation. a, TP, FP and FN event counts for all eight support states in fixed Mutect2/Strelka2/DeepSomatic bit order. The 000 state retains 1,194 truth variants missed by all callers. The symlog axis displays zero and nonzero counts without changing their values. b, Caller membership and deterministic ensemble inclusion for each support state: union includes support count ≥ 1 , majority ≥ 2 and strict intersection $=3$. Dots indicate membership, not calibrated probability or a performance score.

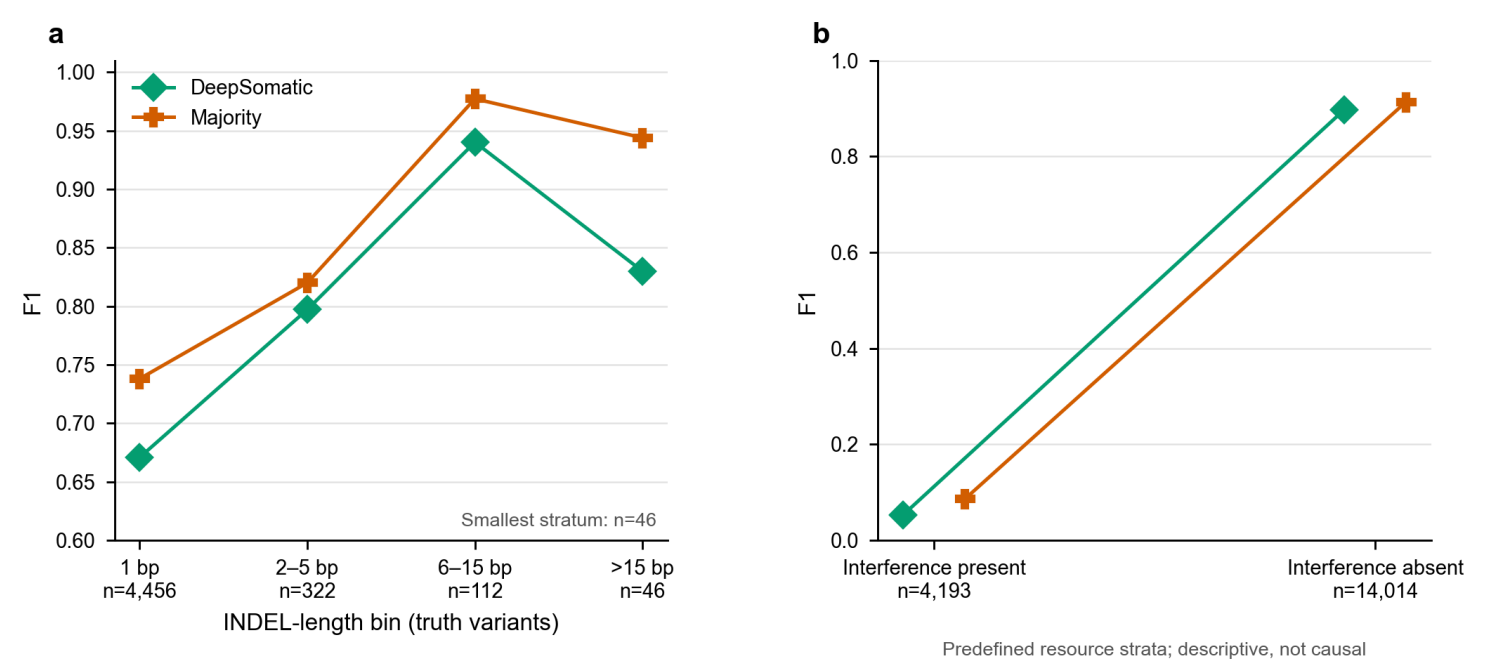


Figure 5

Figure 4 | Caller-support architecture in the primary HG008 region. These analyses were prespecified before HG008 access and executed after completion of the primary external evaluation. a, TP, FP and FN event counts for all eight support states in fixed Mutect2/Strelka2/DeepSomatic bit order. The 000 state retains 1,194 truth variants missed by all callers. The symlog axis displays zero and nonzero counts without changing their values. b, Caller membership and deterministic ensemble inclusion for each support state: union includes support count ≥ 1 , majority ≥ 2 and strict intersection $=3$. Dots indicate membership, not calibrated probability or a performance score.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [P003ResearchSquareSUPPLEMENTARYINFORMATION.pdf](#)
- [ExtendedData.docx](#)