TECHNICAL METHODS NOTE - LONG-READ WGS - REVIEWED 1 SEPTEMBER 2026

ONT alignment statistics need explicit units before they guide variant calling

A long-read CRAM contains alignment *records*, while the sequencer produced molecular *reads*. Suitable QC and variant-calling decisions depend on keeping record, segment, reference-depth, and molecule-level units separate.

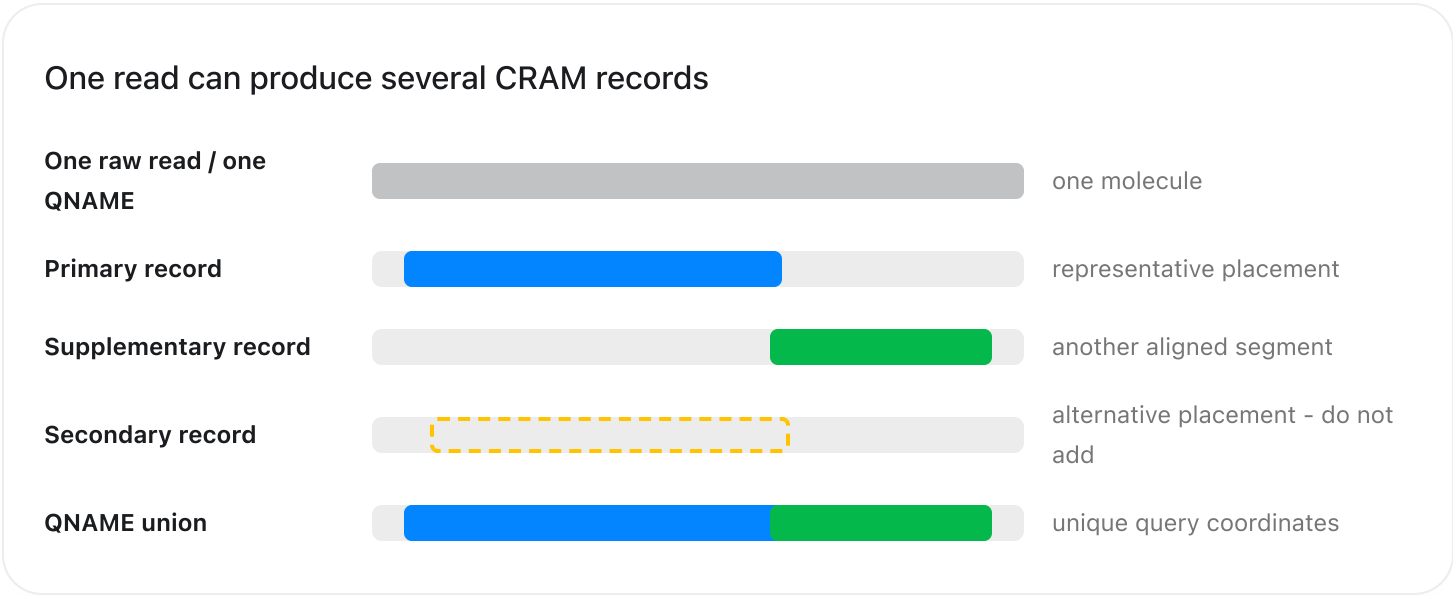
RECOMMENDATION **Keep the Samtools baseline, retain AlignStats only with corrected labels, add explicit mosdepth profiles for reference coverage, and use one-QNAME query-union metrics for raw-versus-aligned completeness.** Do not add a generic variant-caller flag to count secondary or supplementary records as extra reads.

Samtools baseline	AlignStats role	Depth views	Molecule union
Retain	Segment QC	Two profiles	97.96%

Answer first

The current tools are useful, but they do not all measure the same thing. samtools stats is deliberately primary-centric for major headline fields. **samtools idxstats** counts mapped read-segments. AlignStats 0.11 includes supplementary records as separate aligned-length observations and includes them in coverage. None of those outputs alone measures how much of each original ONT molecule is represented by the union of its primary and supplementary alignments.

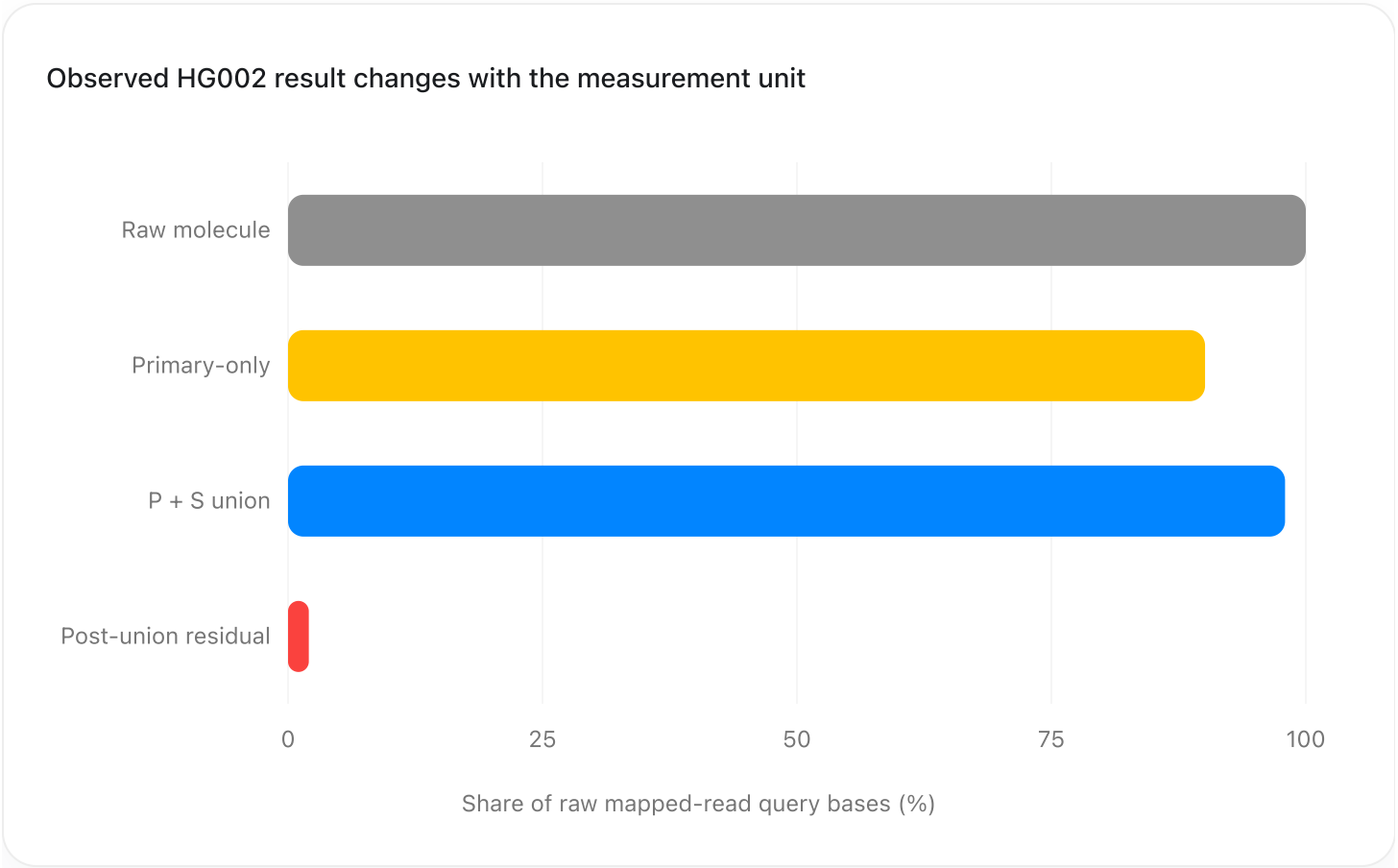
For variant calling, **there is no universal flag that should be enabled to use "everything beyond primary."** Structural-variant callers need split/SA evidence and generally handle it deliberately. Secondary alignments remain competing placements, not additional molecules. Small-variant callers should not be forced to count supplementary or secondary records independently because that can double-count one molecule.



Why the unit changes the scientific answer

The existing full-coverage HG002 audit demonstrates the consequence. Looking only at each primary record reports 90.10% of mapped-read query bases aligned. Unioning primary and supplementary query coordinates once per QNAME raises recovery to 97.96%, leaving 2.04% residual. The supplementary records are not extra reads; they explain much of the apparent raw-to-primary length gap.

This is why a field named “aligned read length” is unsafe when its observations are actually separate primary and supplementary alignment segments.



Evaluation of suitable statistics tools

A layered stack is preferable to a single replacement. Samtools supplies robust format/flag reconciliation. Mosdepth supplies explicitly filtered reference coverage. Raw-read tools describe the sequencer output without alignment-selection bias. ONT error/identity tools can add per-alignment diagnostics. A QNAME-aware union is still required for the specific raw-versus-aligned morphology question.

What each tool can and cannot establish

Tool	Best use	Unit	Supplementary behavior	Disposition
samtools flagstat	Flag and record reconciliation	Alignment records, with primary/secondary/supplementary categories	Reported separately	Retain; prefer JSON or TSV output
samtools stats	General CRAM integrity, primary-centric length/MAPQ/error summaries	Mixed sections; headline total/read-length fields exclude secondary and supplementary	Separate count; excluded from major headline fields	Retain as baseline QC
samtools idxstats	Reference/contig distribution and index sanity	Mapped read-segments per reference	A molecule may be counted more than once	Retain; never call this unique-read accounting
AlignStats 0.11	Historical WGS coverage and alignment-segment summaries	Alignment record/segment	Included in aligned-length and coverage calculations; secondary/QC-failed/unmapped excluded internally	Retain for continuity, but relabel and do not use as molecule completeness
mosdepth	Authoritative reference-base depth and coverage thresholds	Reference positions covered by records surviving an explicit flag mask	Included by default mask 1796; excluded by mask 3844	Preferred depth engine with two named profiles
pomoxis stats_from_bam	ONT-oriented per-alignment identity/error summaries	Aligned record	Does not replace QNAME-level union accounting	Optional diagnostic companion
NanoPlot/NanoStat or pycoQC	Raw run/read length and base-quality QC from FASTQ or basecaller summary	Original read/molecule	Not applicable	Preferred raw-read QC; separate from alignment QC
QNAME query-union audit	Raw-versus-aligned completeness, clipping recovery, gaps and overlaps	One original ONT molecule	Primary and supplementary intervals unioned once; secondary excluded	Required for the scientific question raised here

KEEP

samtools flagstat/stats/idxstats for baseline integrity, flags and contig distribution.

RELABEL

AlignStats aligned length as segment length; coverage as nonsecondary segment depth including supplementary.

ADD

mosdepth plus QNAME union for declared reference depth and molecule completeness.

Make reference coverage explicit

A single unqualified “ONT coverage” number is ambiguous. Report at least two named values: primary-only reference depth and nonsecondary alignment-segment depth. Mosdepth makes the filter visible and reproducible.

```
# Summary depth from primary mapped QC-passing nonduplicate records only
mosdepth -n -f reference.fa -F 3844 sample.primary_only sample.cram

# Summary depth from primary + supplementary nonsecondary segments
mosdepth -n -f reference.fa -F 1796 sample.nonsecondary_segments sample.cram
```

The second profile is mosdepth's current default mask. It can legitimately be useful for “how much aligned sequence covers the reference,” but it is not a unique-molecule measurement and can count overlapping primary/supplementary segments at the same locus.

Recommended named depth profiles

Profile	-F mask	Hex	Excluded records	Supplementary	Interpretation
Primary-only reference depth	3844	0xF04	UNMAP, SECONDARY, QCFAIL, DUP, SUPPLEMENTARY	Excluded	Depth contributed by primary mapped QC-passing nonduplicate records
Nonsecondary segment depth	1796	0x704	UNMAP, SECONDARY, QCFAIL, DUP	Included	Depth from primary and supplementary segments; overlapping segments may both contribute
Current AlignStats 0.11 coverage	772	0x304	UNMAP, SECONDARY, QCFAIL	Included	Alignment-segment coverage; duplicates are retained unless separately removed

Variant calling: should flags be added to use more than primary?

No generic inclusion flag should be added. The correct record policy is caller- and variant-class-specific.

- **SNVs and small indels:** the primary placement is the main evidence. A supplementary segment may be relevant at its own locus, but blindly counting every record as an independent read would inflate support. Secondary records should not be promoted to evidence.
- **Structural variants:** primary plus supplementary/SA geometry is essential for many breakpoints, inversions and translocations. The caller should associate segments by QNAME/SA/CIGAR rather than count them as unrelated reads.
- **Secondary alignments:** these are alternative placements. Retain them for diagnostic morphology, but do not enable them as extra variant-supporting molecules.

The current DayOA commands already pass the whole CRAM to each caller; there is no upstream primary-only **samtools view** filter.

Caller-specific evidence and flag recommendation

Caller	Class	DayOA input	Supplementary / split evidence	Secondary	Recommended flag action
Sentieon DNAscope LongRead / hybrid DNAscope	SNV and small indel	Entire ONT CRAM (and SR CRAM for hybrid)	Exact internal treatment not stated in public documentation	Exact internal treatment not stated publicly; must not be interpreted as independent molecules	Do not add an undocumented inclusion flag. Retain the full CRAM and obtain vendor confirmation if an auditable record policy is required.
Sniffles2	Structural variant	Entire ONT CRAM	Used directly and/or reconstructed from the primary SA tag for split evidence	Explicitly skipped in current source	No enable flag is needed. Avoid an exclusion mask containing 0x800 unless validated. --output-rnames is optional for auditability, not inclusion.
Sentieon LongReadSV	Structural variant	Entire ONT CRAM	Split evidence is scientifically relevant; exact internal policy is not public	No public record-level policy located	Do not invent a flag. Keep the full CRAM and request the supported policy from Sentieon if release acceptance depends on it.

Sniffles2

No flag is required to “turn on” supplementary evidence. Current Sniffles2 source skips secondary records, accepts supplementary records that pass MAPQ/length filters, and reconstructs split evidence from SA relationships on nonsupplementary records to avoid duplicate split processing. Under normal QC its current defaults are MAPQ 20 and minimum alignment length 1,000 bp. These are evidence thresholds, not supplementary-enable switches.

Do **not** add **--exclude-flags 2308 / 0x904** merely as cleanup: that mask contains **0x800** and therefore removes supplementary records. Sniffles may still reconstruct many splits from the primary SA tag, but the behavior is no longer equivalent. If traceability is desired, **--output-rnames** exposes supporting read names in the VCF; it changes auditability, not which record types are eligible.

Sentieon DNAscope and LongReadSV

DayOA already supplies the complete CRAM. The public Sentieon material reviewed does not document a record-level flag that enables supplementary alignments or states the exact internal secondary/supplementary policy. The defensible action is **not to invent a flag.** If clinical/release acceptance requires an auditable guarantee, obtain a version-specific statement from Sentieon and record it with the caller version/model.

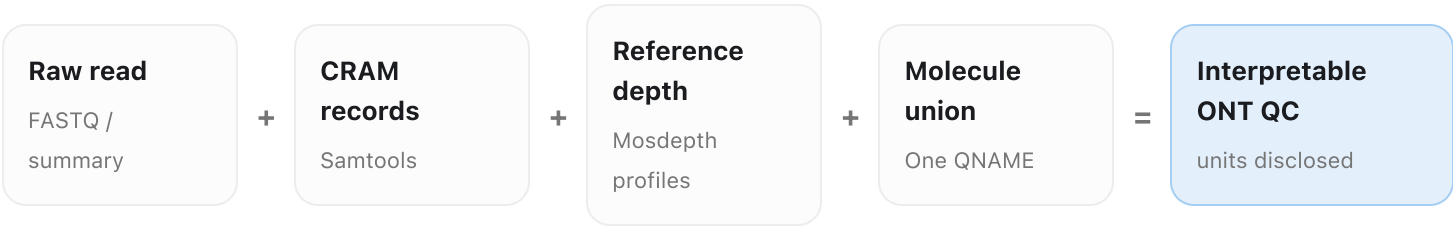
Minimap2 -Y

This mapper option controls supplementary CIGAR representation: bases outside a supplementary segment are soft-clipped rather than hard-clipped. It does not enable a variant caller to treat secondary records as evidence and does not itself change alignment endpoints. Keep its representational effect separate from caller record-selection policy.

Recommended ONT QC contract

1. **Raw molecule QC:** FASTQ/basecaller-summary read count, yield, N50, length and Q-score using NanoPlot/NanoStat or pycoQC.
2. **Record reconciliation: samtools flagstat -O json**, with exact primary, secondary, supplementary, mapped and QC-failed counts.
3. **Primary-centric alignment QC:** retain **samtools stats**; expose its definitions rather than treating every section as the same denominator.
4. **Reference distribution:** retain **samtools idxstats**, labeled “mapped read-segments by reference.”
5. **Reference depth:** mosdepth primary-only and nonsecondary-segment profiles with the flag mask, reference identity and denominator recorded.
6. **ONT alignment identity/error:** optionally add pomoxis **stats_from_bam** or an equivalent caller-validated tool; label it per alignment.
7. **Molecule morphology:** one row per QNAME containing raw length, primary aligned-query length, primary-plus-supplementary query-coordinate union, residual, segment count, gaps/overlaps, and secondary-alternative count.

This contract lets MultiQC show comparable operational metrics without collapsing incompatible units into one “aligned read length” or “coverage” field.



Current DayOA evidence

The read-only checkout review confirms that DayOA generates the Samtools trio, invokes a vendored AlignStats binary with **-C -U**, and passes complete LR CRAMs to DNAscope, Sniffles2 and LongReadSV. No caller is currently fed a prefiltered primary-only CRAM.

Observed workflow command boundaries

Component	Location	Observed command shape	What it proves
Samtools metrics	workflow/rules/samtools_metrics.smk:33	samtools stats; samtools flagstat; samtools idxstats	Baseline metrics are generated together.
AlignStats	workflow/rules/alignstats.smk:86	resources/alignstats/alignstats -C -U ...	No -F exclusion mask is supplied; the v0.11 internal filters determine supplementary treatment.
ONT DNAscope	workflow/rules/sent_snv_ont.smk:149	dnascope-longread -i <ONT CRAM> --tech ONT	Whole CRAM passed; no upstream flag filter.
Hybrid DNAscope	workflow/rules/sent_hybrid_ilmn_ont_modular2.smk:1145	DNAscope -i <LR CRAM> -i <SR CRAM>	Whole LR and SR CRAMs passed.
Sniffles2	workflow/rules/sent_hybrid_ilmn_ont_modular2.smk:1504	sniffles --input <ONT CRAM>	No supplementary-exclusion option supplied.
Sentieon LongReadSV	workflow/rules/sent_hybrid_ilmn_ont_modular2.smk:1701	sentieon driver --input <LR CRAM> --algo LongReadSV	Whole LR CRAM passed.

Final recommendation

No immediate variant-caller inclusion flag change is supported by the evidence. Leave the full-CRAM caller inputs intact. Sniffles2 already has the split-evidence behavior required for SVs; secondary records are deliberately excluded. For Sentieon callers, document the present uncertainty and seek vendor confirmation rather than guessing.

For QC, treat AlignStats as a continuity output, not the authoritative ONT molecule metric. Add mosdepth with two explicit flag profiles and promote QNAME-level primary-plus-supplementary union accounting to the authoritative answer for raw-versus-aligned completeness. Preserve the Samtools outputs because they remain the best compact reconciliation of file integrity and SAM flags.

Sources, scope and limitations

Primary references: Samtools stats, flagstat, idxstats, SAM specification, AlignStats 0.11 aligned-length source, AlignStats 0.11 coverage source, mosdepth documentation, ONT pomoxis, Sniffles2 lead extraction, Sniffles2 options/defaults, and Sentieon DNAscope LongRead.

The HG002 percentages are reviewed existing-data results: exact whole-CRAM totals were kept separate from deterministic sampled morphology. No new alignment, variant calling, Ganon2 run, workflow launch, configuration edit, release, or AWS mutation was performed. Public Sentieon documentation did not establish exact secondary/supplementary record handling; this report deliberately labels that policy as unknown.

Prepared 1 September 2026 - technical methods note