HG002 full-coverage ONT alignment morphology audit

Existing-data review of HG002 ONT [0,25) reads, focused on raw versus aligned read length, soft clipping, split-alignment morphology, base quality, GC and perfect-repeat composition, MAPQ 20 routing, and released Ganon2 outcomes.

Bottom line: The broad claim that minimap2 soft clipping is prematurely truncating a large fraction of ONT reads is **not supported** in this dataset. Primary alignments cover 90.10% of mapped-read bases; supplementary segments recover most of the apparent long clips and raise unique query coverage to **97.96%**. Only **0.90%** of mapped reads retain at least 1 kb unaligned after the union.

ONT COVERAGE	PRIMARY READS	UNION ALIGNED	RESIDUAL
12.19x	**4,968,928**	**97.96%**	**2.04%**
mosdepth mean; median 12x	exact whole-CRAM count	base-weighted, mapped reads	after supplementary recovery

Scope and dataset

This is the **full-coverage Bjuice-prevalidation HG002 result**, AU `HG002-ck2ky6p4wk6exh`, not the approximately 2.9x one-chip validation CRAM. The released Take101 QC labels the long-read input **ONT [0,25)** and reports 12.19x mean genome coverage, 12x median coverage, 95.02% mapped alignment records, and 4,968,928 primary reads.

The CRAM header records minimap2 2.26-r1175 with Sentieon's HybridIlluminaONT2.0 model and `-Y`. Exact routing totals use the entire existing CRAM. Detailed CIGAR, query-coordinate, and QUAL morphology uses a deterministic 2% QNAME sample of 99,052 primary reads, including 93,004 human-mapped primaries.

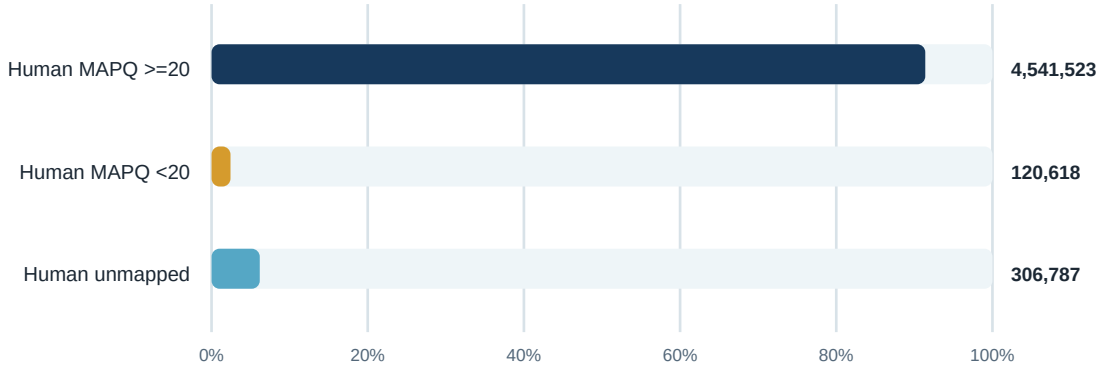
Evidence tier	Unit	What it supports
Whole CRAM	6,165,969 SAM records	Exact primary, secondary, supplementary, mapping, and MAPQ totals
Deterministic 2% sample	99,052 primary QNAMEs	Query unions, residual lengths, split morphology, quality, GC, and perfect-repeat composition
Released QC/Ganon2	Existing reports and logs	Coverage and aggregate taxonomic outcomes

No new alignment or taxonomic classification contributed to this report. The analysis is descriptive and read-only with respect to the released data.

1. EXACT READ ACCOUNTING

Human routing and Ganon2 input reconcile exactly

Primary reads are the unique-read denominator. The three branches below are mutually exclusive and exhaustive; secondary and supplementary SAM records are evidence attached to those reads, not extra reads.



Primary-read branch	Reads	Share	Disposition
Human MAPQ >=20	4,541,523	91.3984%	Terminal human branch
Human mapped MAPQ <20	120,618	2.4274%	Human-aligned, also screened by Ganon2
Human unmapped	306,787	6.1741%	Screened by Ganon2
Total primary	4,968,928	100.0000%	Exact reconciliation

Ganon2 input: 120,618 low-MAPQ mapped + 306,787 human-unmapped = **427,405**. There are no MAPQ 255 primary alignments. This proves the production routing count without double counting.

Record inventory

SAM record class	Records	Interpretation
Primary	4,968,928	One unique read record per QNAME
Secondary	743,401	Alternative placement evidence; never added to coverage
Supplementary	453,640	Additional segment evidence from the same raw query
All records	6,165,969	Record total, not read total

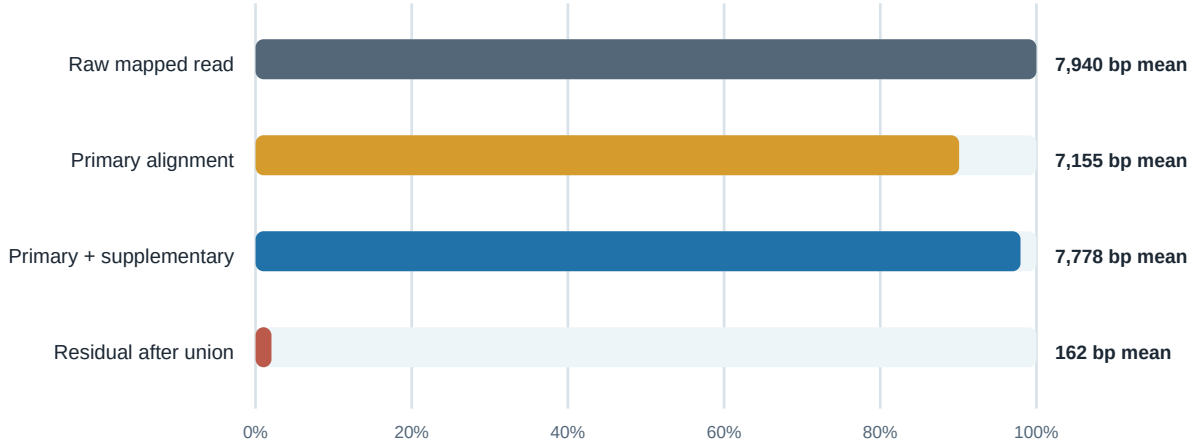
MAPQ 20 interpretation

MAPQ is confidence in the selected genomic placement, not the probability that a read is human. A repeat-derived human read can map end to end with low MAPQ. The current threshold is therefore defensible as a **permissive screening route**, but not as a biological human/nonhuman cutoff.

2. RAW LENGTH VERSUS ALIGNED LENGTH

Supplementary segments explain most of the apparent gap

The correct unit is the union of query coordinates from primary and supplementary alignments, counted once per QNAME. Summing alignment-record soft clips - especially in a -Y CRAM - repeatedly counts the same raw sequence and overstates loss.



Mapped-read measure	Mean	Median	5th percentile	95th percentile
Raw length	7,940 bp	6,290 bp	1,133 bp	18,802 bp
Primary aligned query	7,155 bp	5,871 bp	993 bp	16,440 bp
Union-aligned query	7,778 bp	6,143 bp	1,018 bp	18,588 bp
Residual query	162 bp	118 bp	70 bp	138 bp
Union-aligned fraction	96.83% read-mean	98.30%	89.26%	99.49%

What -Y does - and does not do

The minimap2 -Y option represents sequence outside a supplementary segment as **soft clips** instead of hard clips, preserving the query sequence and QUAL in the supplementary SAM record. It does **not** tell minimap2 to stop alignment early or change the alignment endpoints. In this sample no primary or supplementary record contains hard clipping.

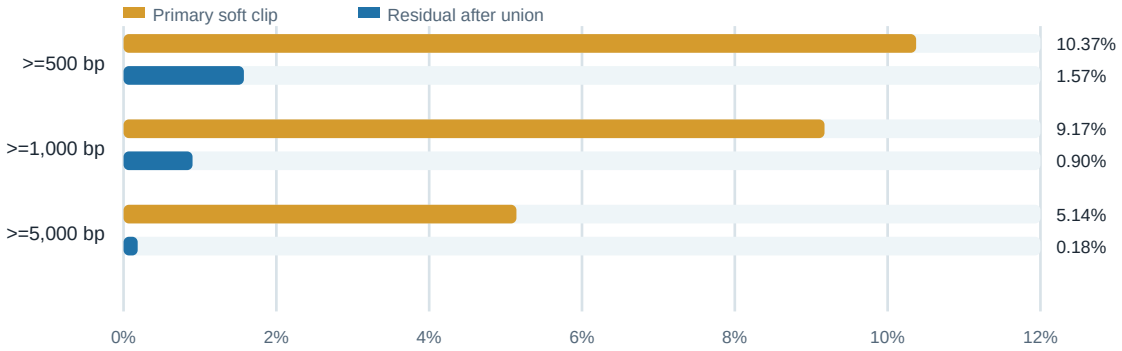
Every sampled mapped primary record has at least one soft-clipped base. That prevalence is real at the CIGAR-record level, but it is not equivalent to large unique-read loss. The mapped-read median residual after the union is only 118 bp: a 74 bp median leading residual plus a 48 bp median trailing residual. The median internal gap is 0 bp.

> Among the 8,584 reads with supplementary alignments, the primary soft clip averages **6,969 bp**; supplementary segments recover **6,759 bp**, leaving only **210 bp** residual on average. This is the dominant explanation for the large raw-versus-primary gap.

3. RESIDUAL TAIL AND ALIGNMENT MORPHOLOGY

Large primary clips usually contain already-aligned supplementary sequence

The paired bars compare the apparent primary soft clip with genuinely residual query sequence after primary-plus-supplementary union. The post-union tail is small but nonzero.



Residual threshold	Mapped reads	Share of mapped reads	Interpretation
>=500 bp	1,465	1.575%	Bounded review tail
>=1,000 bp	840	0.903%	Less than 1 in 100 mapped reads
>=5,000 bp	172	0.185%	Rare long-residual tail

One contiguous alignment is the norm after union

Query-union morphology	Mapped reads	Share
One contiguous union	91,720	98.6194%
Two unions	1,238	1.3311%
Three unions	40	0.0430%
Four or more unions	6	0.0065%

Supplementary and secondary evidence

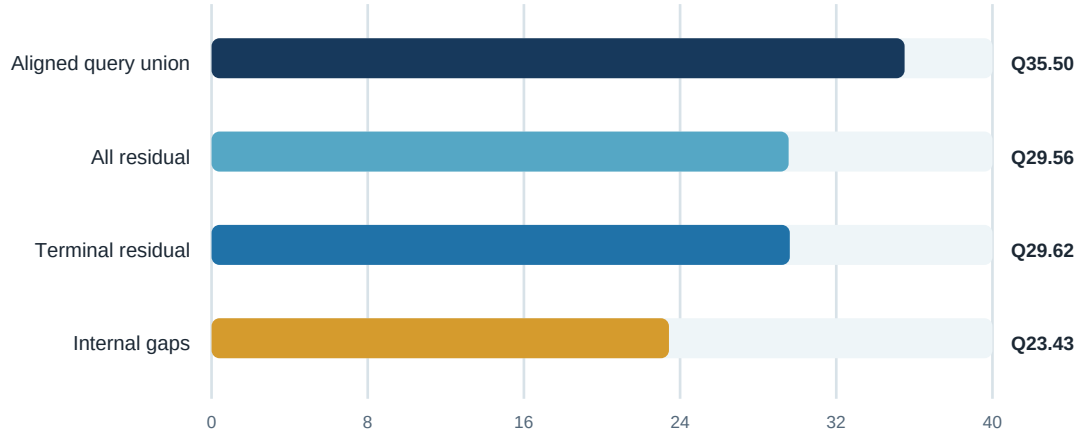
Supplementary phenotype	Records	Share of supplementary
Same-locus split (within 100 kb)	7,168	79.04%
Distant same-contig split	168	1.85%
Cross-contig split	1,733	19.11%

8,584 split reads carry 9,069 supplementary records. Separately, 5,061 mapped reads carry 15,000 secondary alternatives. Secondary records are never added to query coverage.

4. RESIDUAL PHENOTYPES AND BASE QUALITY

Residual sequence is mainly short terminal sequence

Of 41.22 million residual bases in the deterministic sample, 40.85 million (99.11%) are terminal and 0.37 million (0.89%) are internal query gaps. Internal gaps occur in 1,284 mapped reads, or 1.38% of mapped reads.



Mean residual quality is lower than aligned sequence but remains substantial. Internal gaps are the lowest-quality residual class. Quality alone therefore does not explain all clipping, but it is consistent with difficult sequence being enriched outside the aligned union.

Descriptive residual phenotype	Sampled reads	Share of all sampled reads	Meaning
Low-quality terminal residual	1,897	1.92%	Conservative quality flag
Low-complexity / homopolymer proxy	345	0.35%	Entropy <1.2 bits/base or homopolymer >=50 bp
Internal query gap	1,284	1.30%	Separated aligned query intervals

What can and cannot be inferred

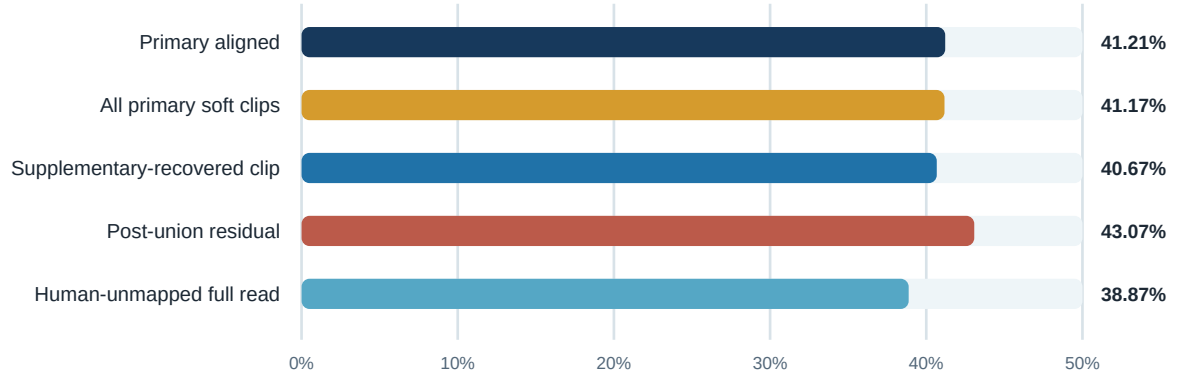
These phenotypes support three bounded conclusions: (1) most residual bases are terminal; (2) a minority is lower quality or simple sequence; and (3) long unexplained residuals are rare. They do **not** establish whether a particular residual is additional human sequence, adapter, contaminant, repeat, or molecule artifact. That organism- or locus-level attribution would require targeted remapping or classification, which was explicitly excluded from this one-off descriptive analysis.

> The evidence argues against a dataset-wide minimap2 truncation phenotype. It does support targeted review of the **0.90%** of mapped reads with at least 1 kb residual, particularly the **0.19%** with at least 5 kb residual.

5. GC AND PERFECT-REPEAT COMPOSITION

The final residual is compositionally distinct from ordinary soft clipping

Canonical primary sequences were partitioned with the existing CIGAR query intervals only. The comparison separates all primary soft clips from the subset recovered by supplementary alignments and from the final post-union residual.



GC content is nearly identical between all primary soft clips and primary-aligned sequence (-0.05 percentage points). The supplementary-recovered clip is also close. The small final residual is GC-richer (+1.86 points), whereas fully human-unmapped reads are GC-poorer (-2.34 points).

Sequence class	GC	Mono >=5 nt	Mono >=10 nt	Dinuc >=5 units	Dinuc >=10 units
Primary aligned	41.213%	4.451%	0.525%	0.334%	0.129%
All primary soft clips	41.167%	3.821%	0.601%	0.445%	0.261%
Supplementary-recovered clip	40.672%	4.311%	0.486%	0.336%	0.134%
Post-union residual	43.073%	1.931%	1.043%	0.867%	0.752%
Human-unmapped full read	38.873%	2.414%	0.111%	0.164%	0.052%

Strongest repeat signal: 0.752% of post-union residual bases lie in perfect dinucleotide repeats of at least 10 units, versus **0.129%** of aligned bases - a **5.82-fold enrichment**. Homopolymers of at least 10 nt are enriched 1.99-fold even though shorter >=5 nt homopolymers are less common, indicating fewer but longer tracts when present.

Interpretation and limits

The supplementary-recovered portion has aligned-like GC and repeat burden, supporting its interpretation as ordinary human-alignable sequence rather than mapper-truncated debris. The final residual is enriched for long perfect repeats, especially TG, TA, and AT dinucleotide motifs, but composition alone cannot determine whether a segment is human repeat sequence, adapter, contaminant, or another molecule artifact.

Operational definitions: GC is G+C divided by A+C+G+T; no ambiguous bases occurred. Mono values are fractions of callable bases in perfect homopolymer runs at least 5 or 10 nt long. Dinuc values are fractions in perfect non-homopolymer dinucleotide tandem repeats at least 5 units (10 nt) or 10 units (20 nt) long. These are transparent sequence metrics, not RepeatMasker or repeat-family annotations. Short terminal segment boundaries affect tract opportunity.

6. GANON2 OUTCOMES AND FINAL INTERPRETATION

Released Ganon2 outcomes: branch origin is unavailable

The existing Ganon2 run consumed exactly the 427,405 routed reads. Its aggregate results are exact, but the released per-read `.all` file was removed. Consequently, the classified reads cannot be assigned retrospectively to the low-MAPQ-mapped versus human-unmapped input branch without rerunning Ganon2.

Released Ganon2 outcome	Reads	Share of input	Accounting
Classified	289	0.0676%	Taxonomic evidence; branch origin unavailable
Unclassified	427,105	99.9298%	Engine-processed, no retained classification
Skipped over compiled length limit	11	0.0026%	Not processed by Ganon2 engine
Total input	427,405	100.0000%	Exact routing reconciliation

Unique-read accounting answer: 4,541,523 reads terminate in the MAPQ >=20 human branch. Another 120,618 are human-aligned but permissively screened, and 306,787 are human-unmapped and screened. Ganon2 classified 289 routed reads in aggregate. Existing files cannot say how many of those 289 came from each human branch.

Response to the clipping allegation

Comparing raw read length with only the primary alignment - or summing record-level soft clips from a `-Y` CRAM - materially overstates lost sequence. In full-coverage HG002, supplementary segments explain most of the gap: unique mapped-read query coverage rises from 90.10% primary-only to 97.96% after union. The remaining residual is typically 118 bp and terminal.

The data still contain a bounded tail worth studying: 840 of 93,004 sampled mapped reads retain at least 1 kb after union, and 172 retain at least 5 kb. Existing descriptive evidence cannot determine whether those tails are further alignable human sequence. That narrower experiment was not run.

Composition adds a bounded clue rather than a new classification: all primary clips are GC-matched to aligned sequence, the supplementary-recovered portion has aligned-like perfect-repeat burden, and only the small final residual is strongly enriched for long dinucleotide tracts.

Recommendation

Keep the current **MAPQ <20** Ganon2 screening rule provisionally. Report low-MAPQ human-mapped and human-unmapped inputs as separate branches, preserve future per-read classification output, report over-length omissions explicitly, and do not use record-level soft-clipped-base percentages as estimates of unique read loss.

Evidence identity and definitions

Source CRAM: Take101 HG002 Sentieon minimap2 ONT CRAM, 28,187,720,428 bytes; S3 ETag `b5db8bd255d86d5c03caa4f27e4458ce-673`. Reference: GRCh38 no-alt. Mapper: minimap2 2.26-r1175, HybridIlluminaONT2.0 model, `-Y`.

Coverage note: 12.19x is the reviewed Take101 mosdepth mean. Alignstats reports 11.6156x using its own genome/denominator convention; the two values are retained rather than silently merged.

Aligned query bases: CIGAR M, =, X, and I on query coordinates. **Union:** normalized primary and supplementary query intervals counted once per QNAME. **Residual:** complement of that union within the canonical primary sequence. Secondary records never contribute coverage.

Composition: GC is calculated over A/C/G/T. Mono-repeat burden is the base fraction in perfect homopolymers of at least 5 or 10 nt. Dinucleotide burden is the base fraction in perfect non-homopolymer tandem repeats of at least 5 or 10 units. These are sequence-composition metrics, not repeat-family annotations.

Scope: Existing-output descriptive analysis only. No new minimap2 run, Ganon2 run, DayOA/DYEC change, workflow launch, configuration change, release, or S3 mutation contributed to this report.