How one ONT read becomes primary, supplementary, clipped, and secondary CRAM records

A source-backed guide to raw-query length, per-record aligned segments, soft clipping, split alignments, and the nonduplicated aligned-query union.

The essential correction

A supplementary alignment is another alignment record for the same original read. It often covers sequence soft-clipped by the primary record. It does not create another read, and the CRAM does not merge the records into one row.

ONE MOLECULE

One ONT read
one QNAME, one raw-query coordinate system

SEVERAL RECORDS

1 primary + 0-many supplementary + 0-many secondary
different reference loci or strands are possible

STILL ONE READ

The union is a derived summary
not a newly combined CRAM record

Seven real reads selected from the existing deterministic 2% sample of the full-coverage Take101 HG002 ONT CRAM · no realignment · no synthetic examples

REPRESENTATION MODEL

A read is not an alignment record

Raw read / query
The original ONT molecule, identified by one QNAME. Its sequence positions form the canonical query coordinate system.

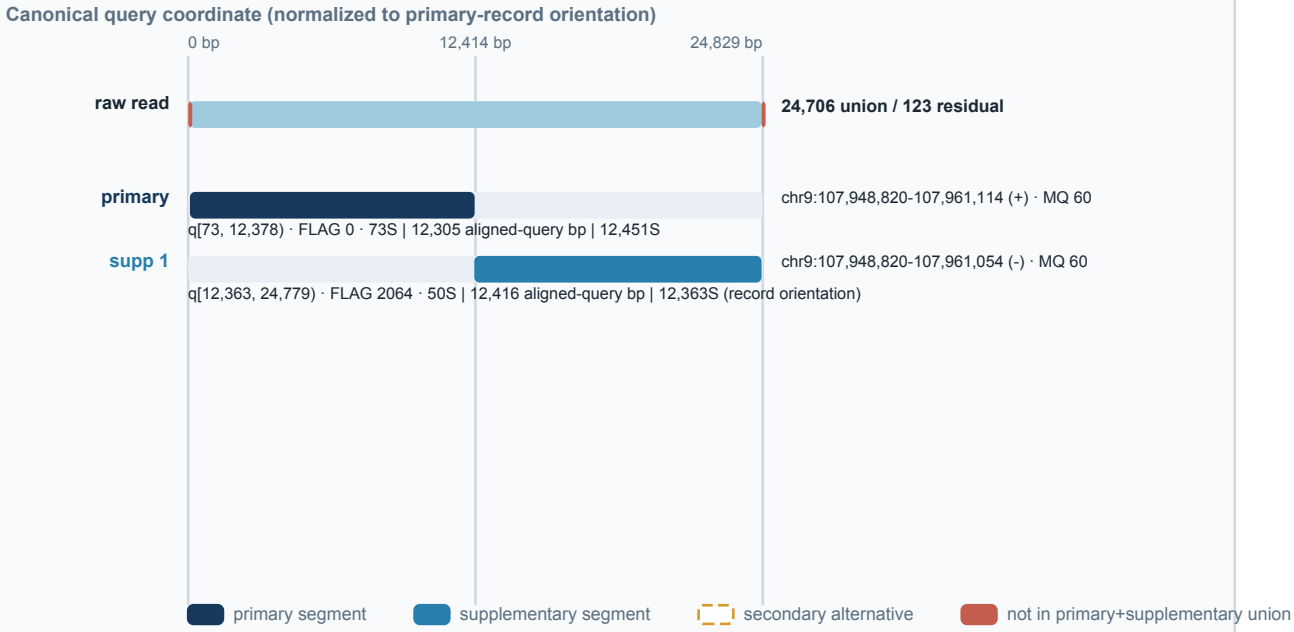
Primary record
minimap2's representative placement. It may describe only one part of the molecule.

Supplementary record
Another aligned segment of that same QNAME. The SAM term is supplementary.

Secondary record
A competing alternative placement. It is not added to aligned-query coverage.

Soft clip (S)
Sequence retained in SEQ/QUAL but outside that particular record's aligned segment. With minimap2 `-Y`, supplementary records use S rather than H.

Actual Case C: two records, one molecule



What CRAM preserves

The two records retain distinct reference position, strand, MAPQ, FLAG, and CIGAR. A downstream analysis can join them by QNAME and normalize their query coordinates; that derived union does not replace either record.

INTERPRETATION

Three different things are often called a “gap”

1 · Reference-only gap

A CIGAR D or N consumes reference coordinates but no query bases. It is not missing sequence inside the raw read.



2 · Primary-record clip

The difference between raw length and the primary aligned span is often terminal S. Some or nearly all of that query sequence may align in a supplementary record.



3 · True query residual

A query interval covered by neither the primary nor any supplementary record remains genuinely outside the alignment union. Case E contains a 9,667-bp internal residual.



Preferred wording

Replace “gaps in primary aligned reads are filled by supplementary alignments” with: “Sequence soft-clipped by the primary alignment is often aligned by one or more supplementary records for the same QNAME.”

Case C illustrates meaning 2: primary alone covers 49.56%; the primary+supplementary union covers 99.50%.

Case E illustrates meaning 3: 9,796 query bases remain outside the nonsecondary union, including a 9,667-bp internal interval.

OBSERVED EXAMPLES

Simple reads: when clipping remains, and when a supplement recovers it

ACTUAL HG002 READ

Case A · One primary record, short terminal clip

QNAME 6d101124-298d-4fd8-8e5d-3c043efda745 raw 7,500 bp · union 99.04% · residual 72 bp



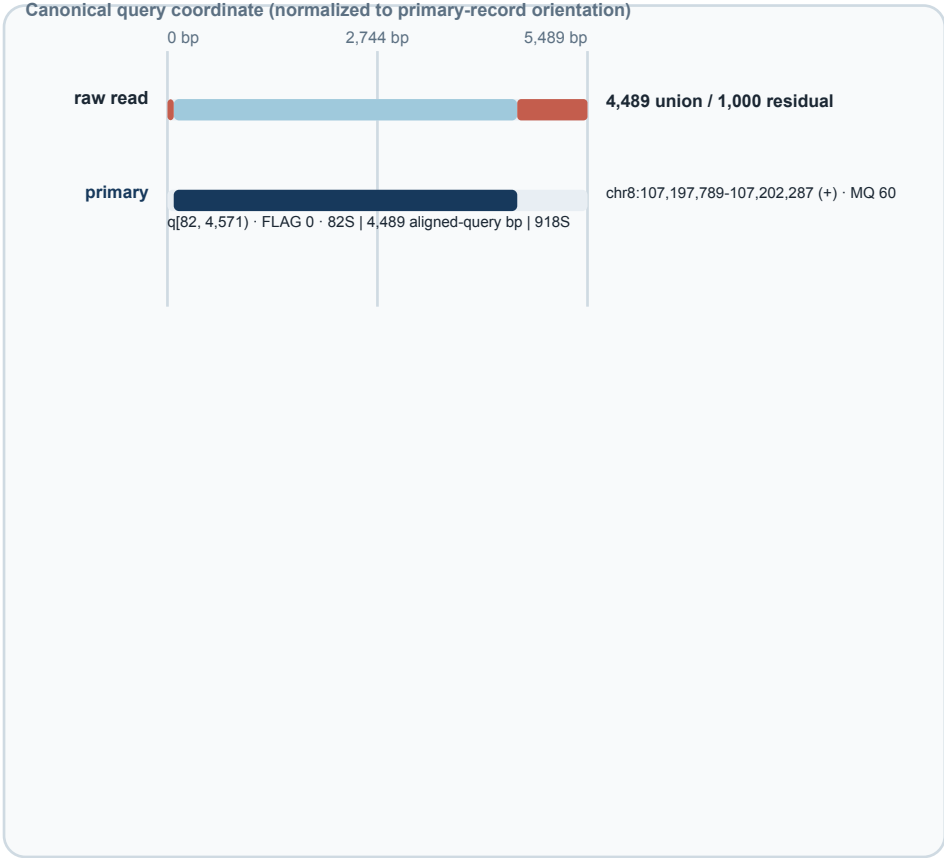
Interpretation

One primary record covers 99.04% of the query. The 72-bp terminal soft clip is the complete post-union residual.

ACTUAL HG002 READ

Case B · One primary record, larger terminal residual

QNAME c4afd752-13d2-4ed5-b0af-ae646094db1c raw 5,489 bp · union 81.78% · residual 1,000 bp



Interpretation

No supplementary record exists. Here the 82-bp and 918-bp terminal soft clips remain genuinely outside the primary-plus-supplementary union.

OBSERVED EXAMPLE

Case C: the primary record alone looks like half a read

ACTUAL HG002 SPLIT ALIGNMENT

Case C · One supplementary segment recovers the apparent half-read loss

QNAME f06edc8e-3d1d-4c05-9272-91b38e525f04

raw 24,829 bp · union 99.50% · residual 123 bp



Record-level view

Primary: 12,305 aligned-query bp (49.56%); its CIGAR shell exposes a 12,451-bp terminal soft clip.

Molecule-level view

The supplementary segment aligns most of that apparent loss. The two segments overlap by 15 query bases; their nonduplicated union is 24,706 bp (99.50%), leaving only 123 bp residual.

COMPLEX OBSERVED EXAMPLE

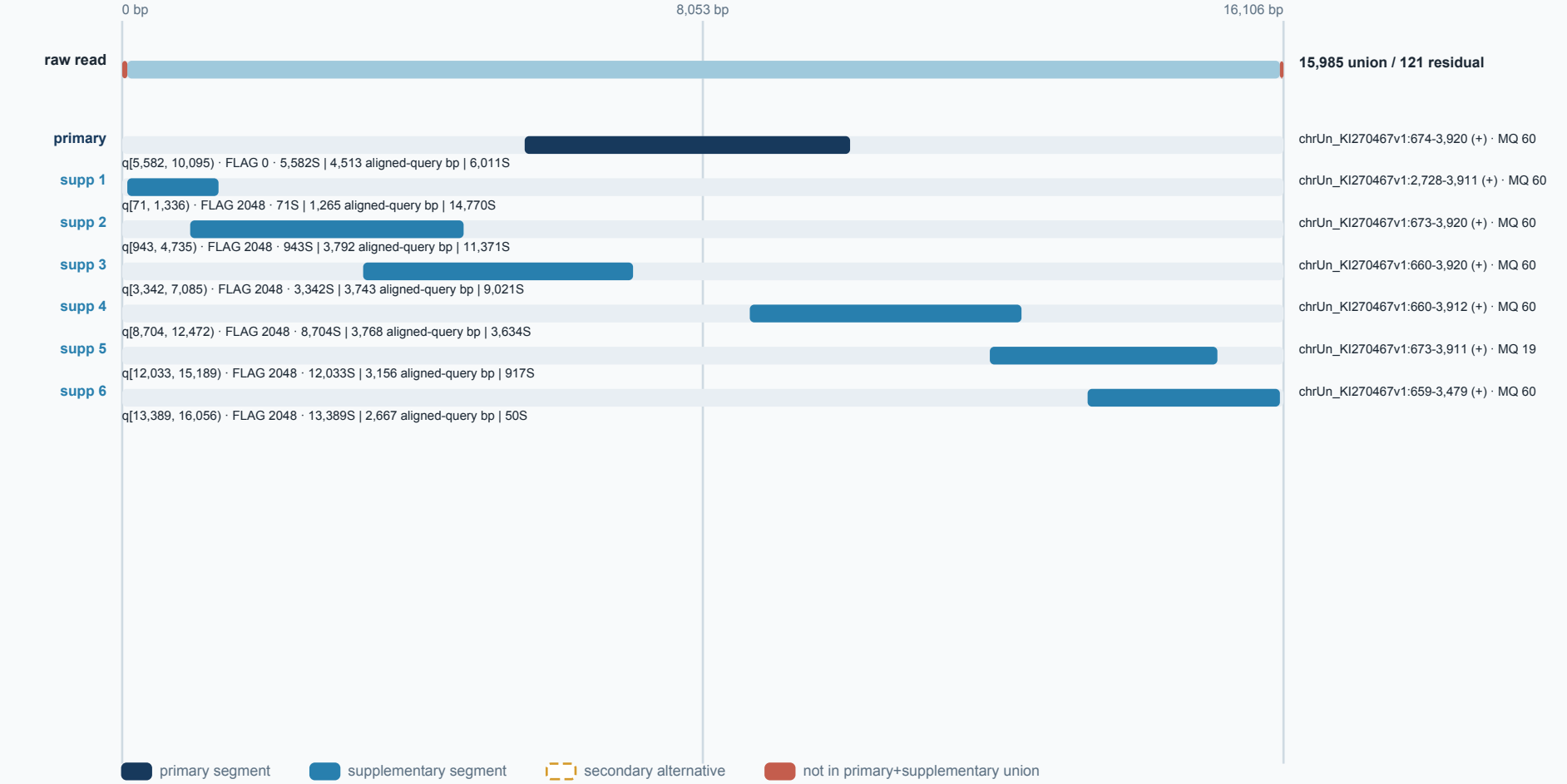
Case D: overlapping records make naive span sums exceed raw length

ACTUAL HG002 READ

Case D · Six overlapping supplementary records form one contiguous query union

QNAME d354a16a-43a8-4d09-bdae-f233ca87dd2c

raw 16,106 bp · union 99.25% · residual 121 bp



Why the union matters

The seven record spans sum to 22,904 bp on a 16,106-bp molecule because 6,919 query bases are represented in overlaps. The nonduplicated primary+supplementary union is 15,985 bp, leaving only 121 bp residual. Never add record spans as if they were independent reads.

COMPLEX OBSERVED EXAMPLE

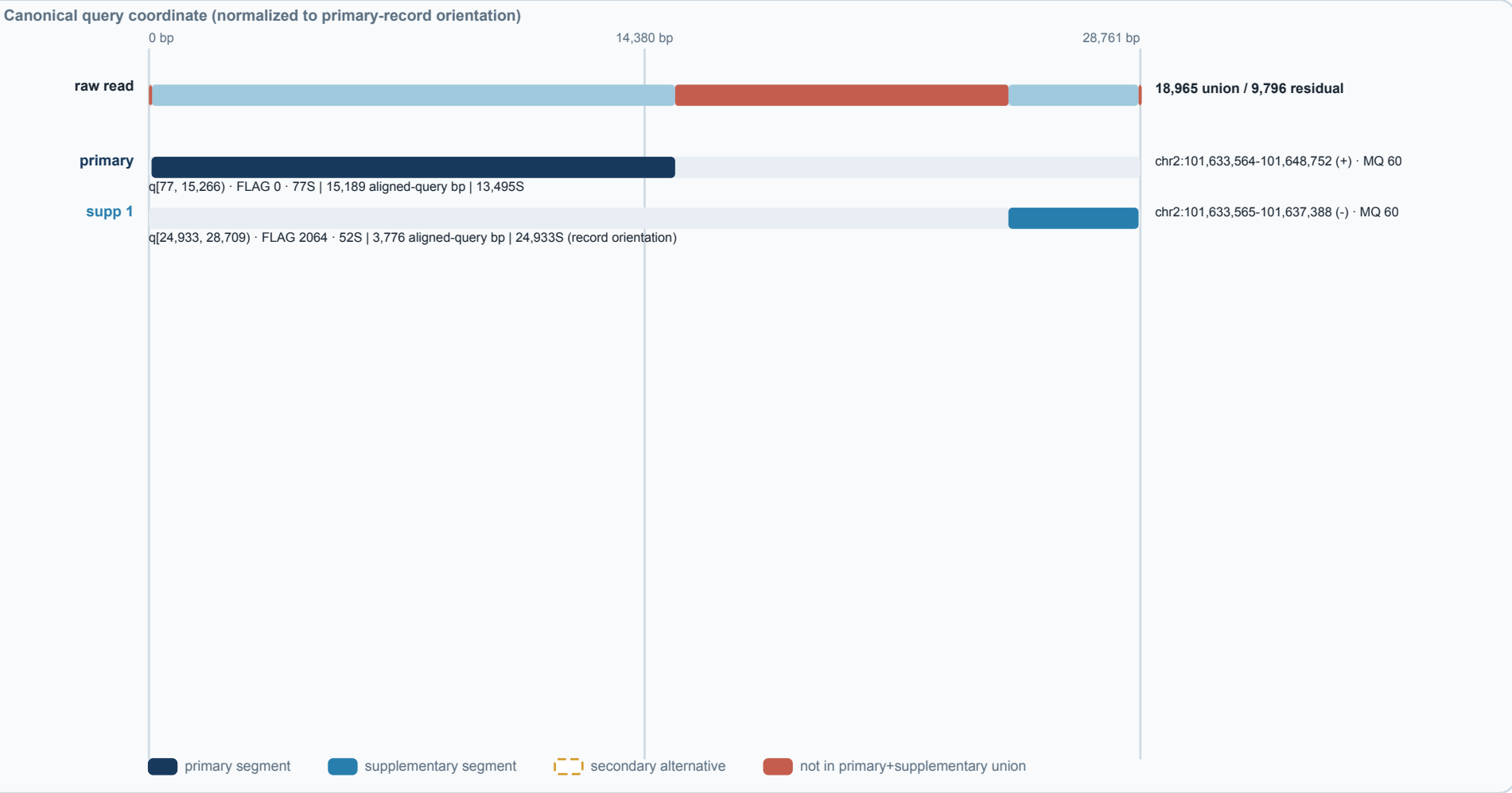
Case E: a genuine 9,667-bp internal query residual

ACTUAL HG002 READ

Case E · A genuinely disjoint alignment leaves a 9,667-bp internal residual

QNAME 73825a55-e15f-4691-92fb-6388067b48c8

raw 28,761 bp · union 65.94% · residual 9,796 bp



This is the literal internal-gap case

The primary covers q[77, 15,266); the supplementary covers q[24,933, 28,709). Sequence q[15,266, 24,933)—9,667 bp—is covered by neither nonsecondary record. This differs fundamentally from Case C, where the apparent primary-record loss is recovered by a supplementary segment.

COMPLEX OBSERVED EXAMPLE

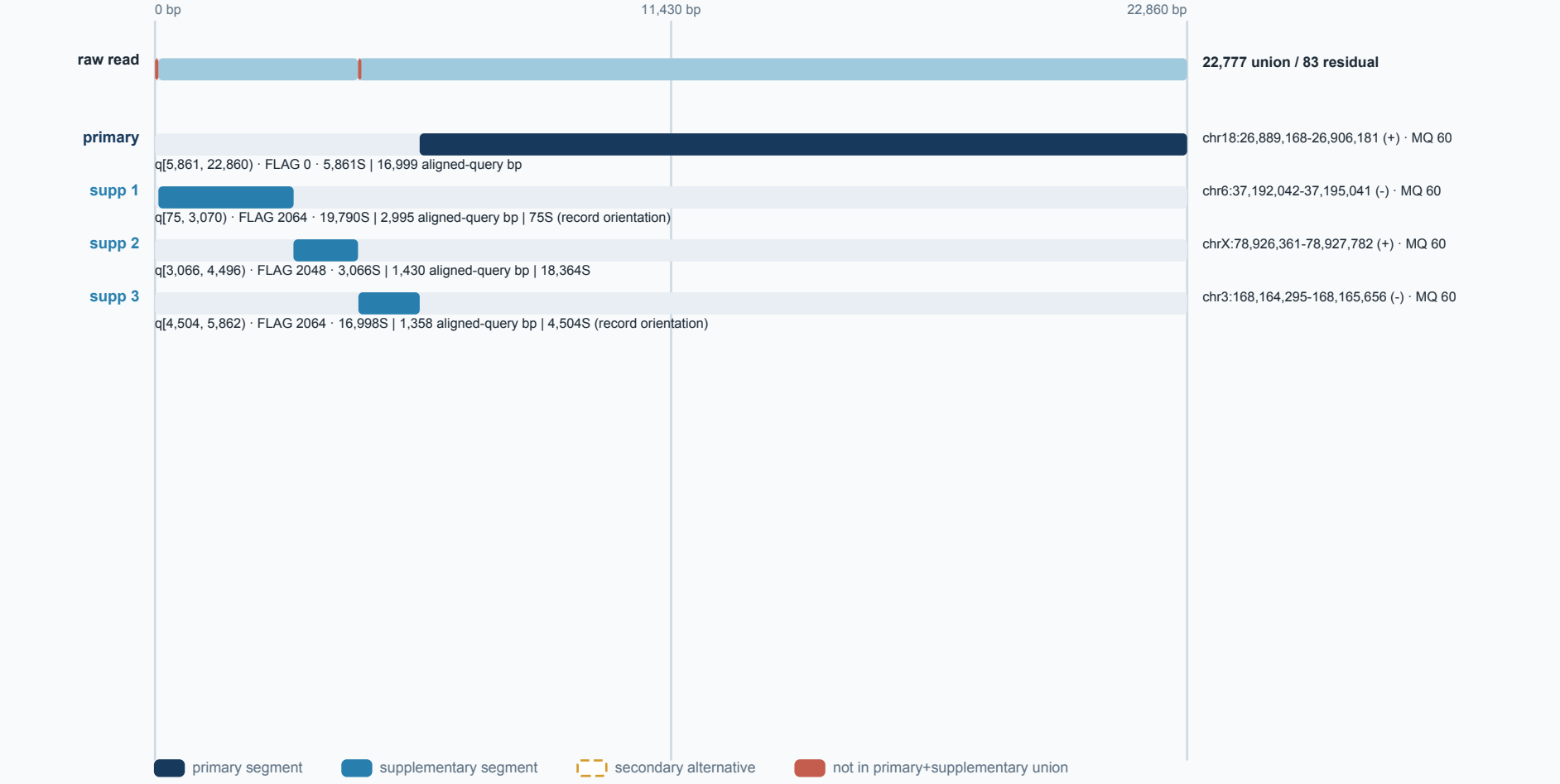
Case F: one molecule, four chromosomes, nearly complete query coverage

ACTUAL HG002 READ

Case F · Four chromosomes cover almost the entire query

QNAME 3422730d-c11f-4a61-a82d-e12b0c9af465

raw 22,860 bp · union 99.64% · residual 83 bp



Still one read

One primary record and three supplementary records preserve four reference placements for the same QNAME. Their query-coordinate union covers 99.64% of the molecule; only 83 bp remain residual. The alignment alone does not determine whether the cross-contig morphology reflects biology, a chimera, repeat ambiguity, or alignment error.

COMPLEX OBSERVED EXAMPLE

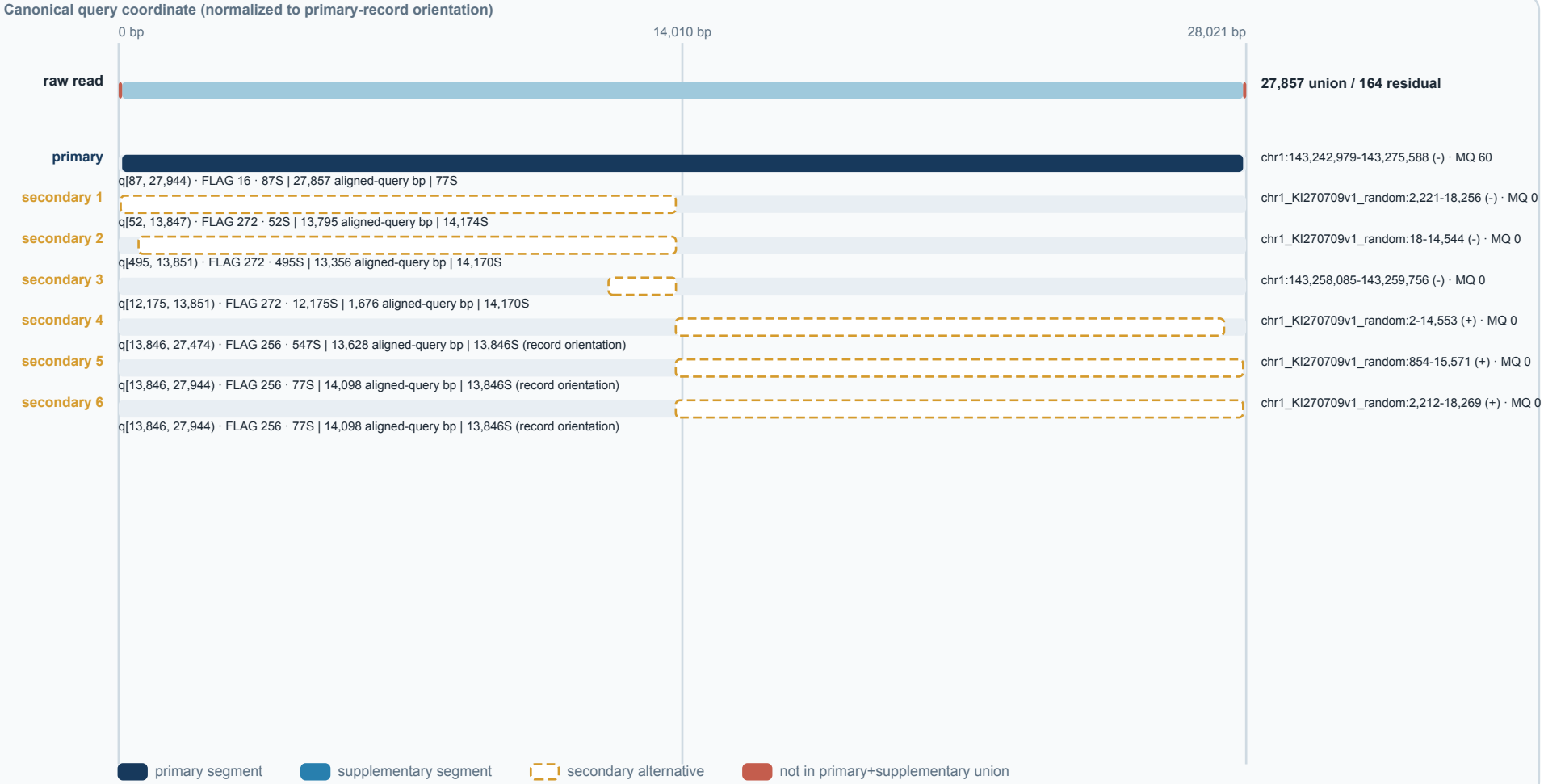
Case G: six secondary records are alternatives, not extra aligned pieces

ACTUAL HG002 READ

Case G · Six secondary alternatives are not six extra pieces of the read

QNAME b772b6d9-8174-4f5b-996c-cd5198fbb33b

raw 28,021 bp · union 99.41% · residual 164 bp



Secondary is not supplementary

The primary already covers 99.41% of the query. Six MAPQ-0 secondary records are competing placements for overlapping portions of the same molecule. They are excluded from the primary+supplementary query union and never counted as six additional pieces of the read.

DIRECT ANSWER

No: supplementary records do not become a newly combined read

What is combined—and where

The original molecule was already one read. CRAM stores its alignment segments as separate records so each can retain its own reference location, strand, FLAG, MAPQ, and CIGAR. An analyst can combine primary and supplementary query intervals into a nonduplicated union, but that is a derived measurement—not a new CRAM record.

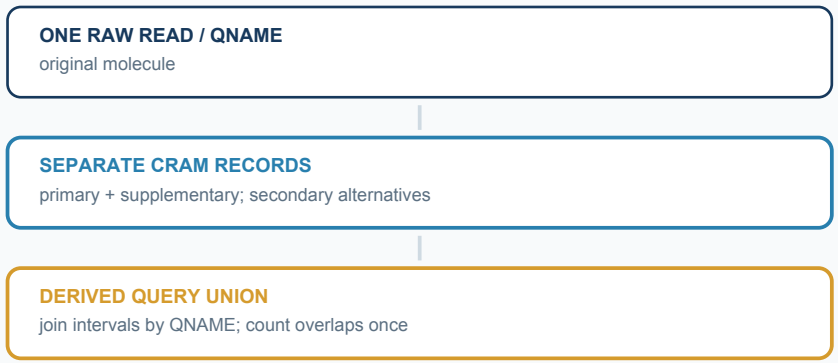
Why separate records are necessary

A split read can map to different positions, opposite strands, or different chromosomes. One SAM/CRAM row has only one reference start, strand flag, MAPQ, and CIGAR; therefore each segment needs its own record.

Why “aligned read length” is risky wording

It can mean the primary record's aligned-query span, the sum of record spans (which can double count overlaps), or the primary+supplementary union. The report uses the last quantity when asking how much of the original molecule is represented by at least one nonsecondary alignment.

The record-to-read relationship



The molecule never stopped being one read.

Bottom line

Use read for the raw molecule/QNAME, alignment record for each CRAM row, and aligned-query union for the nonduplicated span covered by primary plus supplementary records. Then “soft-clipped in the primary but aligned in a supplement” is precise and no longer contradictory.

How these examples were selected and represented

Source data

Seven actual QNAMEs selected from the existing deterministic 2% BAM sample of the full-coverage Take101 HG002 ONT CRAM. The sample contains 99,052 primary QNAMEs.

Fields shown

Record flags, MAPQ, reference coordinates, exterior soft clips, and aligned-query spans were read directly from the existing BAM records. Reverse-strand record intervals were normalized to the primary record's query orientation so all rows share one left-to-right molecule coordinate system.

Coverage arithmetic

Primary and supplementary intervals were unioned per QNAME and overlapping positions counted once. Secondary records were excluded from the union. Each selected read reconciles exactly: raw length = union-aligned bases + residual bases.

Intentionally not done

- no minimap2 rerun
- no residual-clip remapping
- no Ganon2 rerun or taxonomic classification
- no DayOA or DYEC change
- no synthetic alignment example

Interpretive limits

The diagrams explain record morphology, not biological cause. A cross-contig split can be consistent with structural variation, a chimera, repeat ambiguity, reference differences, or alignment error; causation requires additional evidence.

Companion report

This supplement explains the record structure behind the quantitative morphology results in HG002_Bjuice_Prevalidation_ONT_Alignment_Morphology_Audit.pdf.

Generated 2026-09-01 · descriptive analysis of existing data only