SAM/BAM Formats

SAM (Sequence Alignment/Map) Format

```
Header: @HD VN:1.6 SO:coordinate
@SQ SN:chr1 LN:248956422
Alignment: READ1 99 chr1 10001 60 76M = 10052 127 ACGT... IIII...
```

SAM (Text)

- Human-readable
- Tab-delimited
- Large file size
- 11 mandatory fields

BAM (Binary)

- Compressed SAM
- ~3-5x smaller
- Faster to process
- Requires indexing (.bai)

Key SAM Fields

- QNAME - Read name
- FLAG - Bitwise flag (paired, mapped, reverse, etc.)
- RNAME - Reference sequence name (chromosome)
- POS - Alignment position
- MAPQ - Mapping quality score
- CIGAR - Alignment string (M=match, I=insertion, D=deletion)

📊 Detailed Examples and Explanations by Field

📌 1. QNAME (Query Name) - Read Name

QNAME: HWUSI-EAS100R:6:73:941:1973
QNAME: SRR123456.1
QNAME: READ_00001

Description: Unique identifier for each read generated by the sequencing instrument.

- Typically contains sequencer information and coordinate data
- Paired reads from the same DNA fragment share the same QNAME
- Format: InstrumentID:RunNumber:FlowCell:Lane:Tile:X-coord:Y-coord

📌 2. FLAG - Bitwise Flag

FLAG Value Calculation Example

```
FLAG = 99 (0x63)
= 1 (paired) + 2 (properly paired) + 32 (mate reverse strand) + 64 (first in pair)
```

- `0x1 (1) = paired`
- `0x2 (2) = properly paired`
- `0x4 (4) = unmapped`
- `0x8 (8) = mate unmapped`
- `0x10 (16) = reverse strand`
- `0x20 (32) = mate reverse`
- `0x40 (64) = first in pair`
- `0x80 (128) = second in pair`

Description: Integer value representing various read attributes in binary format.

• Efficiently stores multiple attributes in a single number

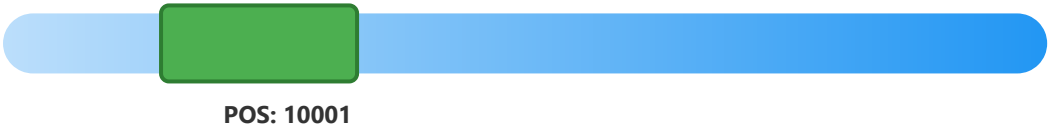
• Each bit has independent meaning (verified using AND operation)

• Example: FLAG=99 → Paired read, properly mapped, first read, mate is reverse strand

📌 3. RNAME & POS - Reference Sequence Name and Position

```
RNAME: chr1, chr2, chrX, chrY, chrM
POS: 10001 (1-based coordinate)
```

Chromosome Position Visualization



← chr1 start (1) ····· Read mapping position ····· chr1 end (248,956,422) →

Description:

• **RNAME:** Chromosome or contig name in reference genome (e.g., chr1, chr2, chrX)

• **POS:** 1-based coordinate where the read is mapped (chromosome starts at 1)

• '*' or '0' indicates an unmapped read

• BAM file indexing (.bai) enables rapid retrieval of specific regions

📌 4. MAPQ - Mapping Quality Score

MAPQ Value	Meaning	Error Probability	Confidence
60	Very high quality	0.0001% (1/1,000,000)	✓✓✓✓✓
40	High quality	0.01% (1/10,000)	✓✓✓✓
20	Medium quality	1% (1/100)	✓✓✓
10	Low quality	10% (1/10)	✓✓
0	No mapping confidence	-	✗

Description: Phred-scaled mapping quality score ($-10 \times \log_{10}(P)$)

- Represents the probability that the read is mapped to an incorrect position
- MAPQ ≥ 30: Generally considered reliable mapping
- MAPQ 0: Multi-mapping (maps equally well to multiple locations)
- Frequently used as filtering criteria in variant calling

📌 5. CIGAR - Alignment String

CIGAR: 50M2I48M1D25M

50M 2I 48M 1D 25M

Visual Representation:

💡 SAM/BAM formats are the standard formats for NGS data analysis, serving as the foundation for all sequencing analyses including variant analysis, RNA-seq, and ChIP-seq.

```
Ref:  ACGTACGT-ACGTACGTACGTACGT
Read: ACGTACGTTTACGTACGT-CGTACGT
CIGAR: 8M2I8M1D7M
```

Operator	Meaning	Description
M	Match/Mismatch	Aligned to reference (includes matches/mismatches)
I	Insertion	Base inserted in read (not in reference)
D	Deletion	Base deleted in read (present in reference)
S	Soft Clipping	Unaligned read ends (sequence retained)
H	Hard Clipping	Unaligned read ends (sequence removed)
N	Skipped Region	Skipped reference region (RNA-seq introns)

Description: Compact representation of alignment between read and reference sequence

- Each operation: number (length) + character (operation type)
- Critical information for structural variant and splice junction detection
- In RNA-seq, 'N' operation represents splicing junctions

📌 6. Complete SAM Line Example

```
READ1 99 chr1 10001 60 76M = 10052 127
ACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGT
```

```
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
NM:i:0 MD:Z:76 AS:i:76
```

Field	Value	Description
QNAME	READ1	Read identifier
FLAG	99	Paired, properly mapped, first read, mate reverse strand
RNAME	chr1	Chromosome 1
POS	10001	Start position
MAPQ	60	Very high mapping quality
CIGAR	76M	76 bases perfectly aligned
RNEXT	=	Mate also mapped to same chromosome
PNEXT	10052	Mate start position
TLEN	127	Template length (insert size)
SEQ	ACGT...	Read sequence
QUAL	IIII...	Per-base quality scores (Phred+33)

📌 7. BAM File Operations Example

```
# Convert SAM to BAM
samtools view -bS input.sam > output.bam

# Sort BAM
```

```
samtools sort output.bam -o sorted.bam

# Index BAM (creates .bai file)
samtools index sorted.bam

# Extract specific region
samtools view sorted.bam chr1:10000-20000

# View statistics
samtools flagstat sorted.bam
```

Description: BAM files are essential for efficient data processing.

- Only sorted BAM files can be indexed
- .bai index file enables rapid retrieval of specific regions
- Visualization tools like IGV require BAM + index
- Typical compression ratio is 1/3 to 1/5 compared to SAM