PCR Duplicate Removal With Consensus Functionality

Release 202503.02

Sentieon, Inc

Jan 27, 2026

Contents

1 Introduction

This document describes how to remove PCR duplicates using Sentieon® Genomics tool. This step uses two individual commands to collect read information and perform the deduplication. The option `--consensus` of `LocusCollector` controls whether to output the consensus of PCR duplicates. If the unique molecular identifiers (UMI) tag is applicable, use the option `--umi_tag` for `LocusCollector` to turn on the barcode-aware deduplication.

2 Non-consensus-based deduplication

With Non-consensus-based deduplication, a representative read from a group of duplicate reads is selected as the primary read.

2.1 Non-consensus-based deduplication without UMI

This workflow matches the default outcome of Picard `MarkDuplicates`.

```
sentieon driver -t NUMBER_THREADS -i SORTED_BAM \
  --algo LocusCollector --fun score_info SCORE.gz
sentieon driver -t NUMBER_THREADS -i SORTED_BAM \
  --algo Dedup [--rmdup] --score_info SCORE.gz  \
  --metrics DEDUP_METRIC_TXT DEDUPED_BAM
```

There is a special 3-pass deduplication flow to mark both primary and non-primary reads. This workflow, however, is only available for non-consensus-based deduplication without UMI.

```
sentieon driver -t NUMBER_THREADS -i SORTED_BAM \
  --algo LocusCollector --fun score_info SCORE.gz
sentieon driver -t NUMBER_THREADS -i SORTED_BAM \
  --algo Dedup --score_info SCORE.gz --output_dup_read_name \
  --metrics DEDUP_METRIC_TXT TMP_DUP_QNAME.gz
sentieon driver -t NUMBER_THREADS -i SORTED_BAM \
  --algo Dedup --dup_read_name TMP_DUP_QNAME.gz \
  DEDUPED_BAM
```

2.2 Non-consensus-based deduplication with UMI

This workflow utilizes the UMI information in addition to the 5' positions of both reads and read-pairs to determine PCR duplicates. Use the option `--umi_tag` in `LocusCollector` to specify the logic UMI tag.

```
sentieon driver -t NUMBER_THREADS -i SORTED_BAM \
  --algo LocusCollector --umi_tag XR --fun score_info SCORE.gz
sentieon driver -t NUMBER_THREADS -i SORTED_BAM \
  --algo Dedup [--rmdup] --score_info SCORE.gz  \
  --metrics DEDUP_METRIC_TXT DEDUPED_BAM
```

3 Consensus-based deduplication

With consensus-based deduplication, one single consensus read is generated from a group of duplicate reads. This process will correct errors introduced by PCR and sequencing. It will also estimate base quality scores at each position to reflect the new probability that a base in the consensus read is called incorrectly. Additional base quality adjustments should not be performed after consensus-based deduplication.

Set the option `--consensus` in `LocusCollector` to turn on the consensus-based deduplication function. Moreover, the reference FASTA file is now required for `Dedup`.

Note

The mate information of consensus reads may be incorrect after consensus-based deduplication. Mate information is not used by Sentieon® variant callers, but may be used by other bioinformatics software.

If correct mate information is required, `samtools fixmate` can be used to update the mate information of reads after consensus-based deduplication.

```
samtools collate -@ NUMBER_THREADS -Ou DEDUPED_BAM \
  | samtools fixmate --reference REFERENCE -@ NUMBER_THREADS - - \
  | sentieon util sort -r REFERENCE -o FIXMATE_BAM -t NUMBER_THREADS -i -
```

3.1 Consensus-based deduplication without UMI

Without UMI, this workflow uses the alignment coordinates alone to cluster sequencing reads.

```
sentieon driver -t NUMBER_THREADS -i SORTED_BAM \
  --algo LocusCollector --consensus --fun score_info SCORE.gz
sentieon driver -t NUMBER_THREADS -r REFERENCE -i SORTED_BAM \
  --algo Dedup [--rmdup] --score_info SCORE.gz  \
  --metrics DEDUP_METRIC_TXT DEDUPED_BAM
```

3.2 Consensus-based deduplication with UMI

With UMI, this workflow uses both the alignment coordinates and their UMIs to cluster sequencing reads.

```
sentieon driver -t NUMBER_THREADS -i SORTED_BAM \
  --algo LocusCollector --consensus --umi_tag XR --fun score_info SCORE.gz
sentieon driver -t NUMBER_THREADS -r REFERENCE -i SORTED_BAM \
  --algo Dedup [--rmdup] --score_info SCORE.gz  \
  --metrics DEDUP_METRIC_TXT DEDUPED_BAM
```

3.2.1 UMI barcode error correction

Errors in the UMI barcodes are corrected automatically based on the edit distance with other barcodes. To disable this function, use the option `--umi_ecc_dist 0` in `LocusCollector`.

3.2.2 RNA sequence data

Set the option `--rna` in `LocusCollector` when using RNA sequence data aligned with STAR.

```
sentieon driver -t NUMBER_THREADS -i SORTED_BAM \
  --algo LocusCollector --rna [--consensus] [--umi_tag XR] --fun score_info SCORE.gz
sentieon driver -t NUMBER_THREADS -r REFERENCE -i SORTED_BAM \
  --algo Dedup [--rmdup] --score_info SCORE.gz DEDUPED_BAM
```

4 Appendix

4.1 Determine the read structure and extract UMI barcode sequences

As a first step, you need to extract the barcode sequences from the input reads. This is performed with the Sentieon® `umi extract` command, which extracts the barcode sequence information from the reads and adds it to the read description. The adapter sequence should be removed from the input reads before running umi tag extraction. This can be performed by other third-party tools.

The output of `umi extract` is in FASTQ format with interleaved R1 and R2 reads. By default, the output from the extract command will be sent to standard output unless otherwise defined with the option `-o`.

The syntax of the `umi extract` command is as follows:

```
sentieon umi extract [options] read_structure fastq1 [fastq2] [fastq3]

Options:
  -o      Output file (default: stdout)
  -d      Turn on duplex mode
  --umi_tag     Logic umi tag (default 'XR')
  --output_format       Output format FASTQ or SAM (default 'FASTQ')
```

The first argument of the `umi extract` command defines the read structure. For paired-end reads, two read structures separated by comma ‘,’ should be specified.

A read structure is defined by `<number><operator>` pairs. The number can be any digit or ‘+’ to indicate the end of the read. Possible operators include:

- **T** The template sequence.
- **B** The cell barcode sequence.
- **M** The molecular barcode sequence.
- **S** A sequence of bases that should be ignored.

The `-d` option is used to extract duplex UMI and label the strand from which it originates. Duplex UMI extraction requires an identical read structure for both strands.

As an example, the following command demonstrates a single-ended UMI extraction on pair-ended reads. In this case, the first read in the pair contains an 8bp molecular barcode followed by a 12bp spacer and then the template sequence. The second read contains only the template sequence. The paired reads will be interleaved in the output file. Please note that in this example the output is piped to `gzip` to generate a compressed FASTQ file. In general, we recommend piping the output directly to the next step (Sentieon® `bwa mem`).

```
sentieon umi extract "8M12S+T,+T” \
  sample_R1.fastq.gz \
  sample_R2.fastq.gz | \
  gzip -c \
  > sample_extracted_pair.fastq.gz
```

The command below demonstrates duplex UMI extraction where both reads contain 4bp molecular barcode followed by template sequence.

```
sentieon umi extract \
  -d \
  "4M+T,4M+T” \
  sample_R1.fastq.gz \
  sample_R2.fastq.gz | \
  gzip -c \
  > sample_extracted_pair.fastq.gz
```

Below is a code example including both UMI extraction and alignment to the reference genome.

```
sentieon umi extract \
  8M12S+T,+T \
  sample_R1.fastq.gz \
  sample_R2.fastq.gz | \
sentieon bwa mem \
  -R "@RG\tID:$GROUP\tSM:$SAMPLE\tLB:$LIBRARY\tPL:$PLATFORM” \
  -t $NT \
  -K $BWA_K_SIZE \
  -p \
  -C \
  $REF \
  - | \
sentieon util sort \
  -i - \
```

(continues on next page)

(continued from previous page)

```
  -o sorted.bam \
  --sam2bam
```

Below is a use case when the UMI sequence is already in a separate FASTQ file `sample_I1.fastq.gz`. When running in this mode, only one additional UMI index read is allowed. The UMI index read should contain no template sequences. This mode does not allow duplex UMI extraction.

```
sentieon umi extract \
  "+M,+T,+T" \
  sample_I1.fastq.gz \
  sample_R1.fastq.gz \
  sample_R2.fastq.gz | \
  gzip -c \
  > sample_extracted_pair.fastq.gz
```

The command below produces a SAM format output. This format is useful when using the RNA-seq aligner `STAR`. In this example, the first read in the pair contains a 16bp cell barcode and a 10bp molecular barcode. The second read contains only the template sequence. The output is single-end reads in the SAM format.

```
sentieon umi extract \
  --output_format SAM \
  "16B10M+S,+T" \
  sample_R1.fastq.gz \
  sample_R2.fastq.gz \
  > sample_extracted.sam
```

The output of `umi extract` contains additional tags. By default, the output contains a CR tag for the cell or sample barcode and an XR tag for UMI sequence to be used by `umi consensus`.

Table 4.1: Additional tags generated by `umi extract`

Tags	Meaning
RX	Extracted UMI sequence bases.
XR	UMI tag for grouping in `umi consensus`.
CR	Cell barcode.

Sentieon

160 E Tasman Dr STE 208, San Jose, CA 95134-1619
www.sentieon.com