PCR Duplicate Removal With Consensus Functionality

Release 202308.03

Sentieon, Inc

Oct 23, 2024

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.

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 step of 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*.

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
```

UMI barcode error correction

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

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
```