markdup

Table of contents

Mark duplicated reads in a BAM/CRAM file.

This tool locates and tags duplicate reads in a BAM or SAM file, where duplicate reads are defined as originating from a single fragment of DNA.

markdup supports the marking of duplicates in two ways, assuming the sort order to be coordinate (the default) or queryname (**--markdups-assume-sortorder-querynamer**).

The input BAM/CRAM must be sorted by queryname. If it is not, please run **pbrun bamsort** with **--sort-order queryname** to preprocess the input file.

Quick Start

```
# This command assumes all the inputs are in INPUT_DIR and all the outputs go to OUTPUT_DIR. docker run --rm --gpus all --volume INPUT_DIR:/workdir --volume OUTPUT_DIR:/outputdir \ --workdir /workdir \ nvcr.io/nvidia/clara/clara-parabricks:4.3.1-1 \ pbrun markdup \ --ref /workdir/${REFERENCE_FILE} \ --in-bam /workdir/${INPUT_BAM} \ --out-bam /outputdir/${OUTPUT_BAM}
```

Compatible Baseline Command

The command below is the GATK counterpart of the Parabricks command above. Note that the corresponding baseline command is different between marking by coordinate and by queryname. Choose the correct one based on your case. The first **gatk SortSam** command is listed here to guarantee the order of the input file to MarkDuplicates. Feel free to ignore it if your file order is correct.

Coordinate Sort Order

```
gatk SortSam \ -R <INPUT_DIR>/${REFERENCE_FILE} \ -I <INPUT_DIR>/${INPUT_BAM} \ -O <INPUT_DIR>/${SORTED_BAM} \ -SO coordinate gatk MarkDuplicates \ -I <INPUT_DIR>/${SORTED_BAM} \ -O <OUTPUT_DIR>/${MARKED_BAM} \ -M <OUTPUT_DIR>/${METRICS_FILE} \ -ASO coordinate
```

Queryname Sort Order

```
gatk SortSam \ -R <INPUT_DIR>/${REFERENCE_FILE} \ -I <INPUT_DIR>/${INPUT_BAM} \ -O <INPUT_DIR>/${SORTED_BAM} \ -SO queryname gatk MarkDuplicates \ -I <INPUT_DIR>/${SORTED_BAM} \ -O <OUTPUT_DIR>/${MARKED_BAM} \ -M <OUTPUT_DIR>/${METRICS_FILE} \ -ASO queryname gatk SortSam \ -R <INPUT_DIR>/${REFERENCE_FILE} \ -I <OUTPUT_DIR>/${MARKED_BAM} \ -O <OUTPUT_DIR>/${FINAL_BAM} \ -SO coordinate
```

markdup Reference

Mark duplicate reads in BAM file. The input file should be sorted by queryname.

Input/Output file options

--in-bam IN_BAM

Path of BAM/CRAM for marking duplicate. Need to be sorted by queryname already. This option is required. (default: None)

Option is required.

--out-bam OUT_BAM

Path of BAM/CRAM file after marking duplicate. (default: None)

Option is required.
--ref REF

Path to the reference file. (default: None)

Option is required.

--out-duplicate-metrics OUT_DUPLICATE_METRICS

Path of duplicate metrics file after Marking Duplicates. (default: None)

Tool Options:

--markdups-assume-sortorder-queryname

Assume the reads are sorted by queryname for Marking Duplicates. This will mark secondary, supplementary, and unmapped reads as duplicates as well. This flag will not impact variant calling while increasing processing times. (default: None)

--optical-duplicate-pixel-distance OPTICAL_DUPLICATE_PIXEL_DISTANCE

The maximum offset between two duplicate clusters in order to consider them optical duplicates. (default: None)

Performance Options:

--num-zip-threads NUM_ZIP_THREADS

Number of CPUs to use for zipping BAM/CRAM files in a run (default 10). (default: None)

--num-worker-threads NUM_WORKER_THREADS

Number of CPUs to use for markdup in a run (default 10). (default: None)

--mem-limit MEM_LIMIT

Memory limit in GBs during sorting and postsorting. By default, the limit is half of the total system memory. (default: 62)

Common options:

--logfile LOGFILE

Path to the log file. If not specified, messages will only be written to the standard error output. (default: None)

--tmp-dir TMP_DIR

Full path to the directory where temporary files will be stored.

--with-petagene-dir WITH_PETAGENE_DIR

Full path to the PetaGene installation directory. By default, this should have been installed at /opt/petagene. Use of this option also requires that the PetaLink library has been preloaded by setting the LD_PRELOAD environment variable. Optionally set the PETASUITE_REFPATH and PGCLOUD_CREDPATH environment variables that are used for data and credentials (default: None)

--keep-tmp

Do not delete the directory storing temporary files after completion.

--no-seccomp-override

Do not override seccomp options for docker (default: None).

--version

View compatible software versions.

GPU options:

--num-gpus NUM_GPUS

Number of GPUs to use for a run. GPUs 0..(NUM_GPUS-1) will be used.