Picard_MarkDup-2.7.0

The DE Quick Start tutorial provides an introduction to basic DE functionality and navigation.
Please work through the tutorial and add your comments on the bottom of this page. Or send comments per email to xwang@cshl.edu.
Thank you.

Rationale and background:

Picard: http://broadinstitute.github.io/picard

Picard is a set of command line tools for manipulating high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM and VCF.

The MarkDuplicates tool is to locate and tag duplicate reads in a BAM or SAM file, where duplicate reads are defined as originating from a single fragment of DNA. Duplicates can arise during sample preparation e.g. library construction using PCR. See also EstimateLibraryComplexity for additional notes on PCR duplication artifacts. Duplicate reads can also result from a single amplification cluster, incorrectly detected as multiple clusters by the optical sensor of the sequencing instrument.

The program can take either coordinate-sorted or query-sorted inputs, however the behavior is slightly different. When the input is coordinate-sorted, unmapped mates of mapped records and supplementary/secondary alignments are not marked as duplicates. However, when the input is query-sorted (actually query-grouped), then unmapped mates and secondary/supplementary reads are not excluded from the duplication test and can be marked as duplicate reads.

MarkDuplicates also produces a metrics file indicating the numbers of duplicates for both single- and paired-end reads.

If desired, duplicates can be removed using the REMOVE_DUPLICATE and REMOVE_SEQUENCING_DUPLICATES options.

Pre-Requisites

1. A CyVerse account. (Register for an CyVerse account here - user.cyverse.org)
2. Mandatory arguments
    a. Input directory: Directory of SAM/BAM files to analyze. The alignment files must be coordinate sorted.
3. Optional arguments:
    a. VALIDATION_STRINGENCY: Validation stringency for all SAM files read by this program. Setting stringency to SILENT can improve performance when processing a BAM file in which variable-length data (read, qualities, tags) do not otherwise need to be decoded. Default value: LENIENT. This option can be set to 'null' to clear the default value. Possible values: {STRICT, LENIENT, SILENT}
    b. REMOVE_DUPLICATES: If true do not write duplicates to the output file instead of writing them with appropriate flags set. Default value: true.

Sample data

The following test data are provided for testing BWA-index-mem here /iplant/home/xiaofei_iplant/Sorghum_chr8/chr8_test:

1. BWA-index-mem_0.7.10_Apr10_Test/G3_P_H3_chr8_BWA_bam and BWA-index-mem_0.7.10_Apr10_Test//G3_P_K4me3_chr8_BWA_bam
   Note: These are the outputs of BWA-index-mem_0.7.10.

Results

Successful execution of the Picard_MarkDup_2.7.0 will create 2 directories named out for BAM files and metrics.

Outputs

1. BAM files
   G3_P_H3_chr8_BWA_bam_rmDup_BAM:
    a. G3_P_H3_rep1_chr8_rmDup.sorted.bam
    b. G3_P_H3_rep1_chr8_rmDup.sorted.bam.bai
    c. G3_P_H3_rep2_chr8_rmDup.sorted.bam
    d. G3_P_H3_rep2_chr8_rmDup.sorted.bam.bai
   G3_P_K4me3_chr8_BWA_bam_rmDup_BAM:
    a. G3_P_K4me3_rep1_chr8_rmDup.sorted.bam
    b. G3_P_K4me3_rep1_chr8_rmDup.sorted.bam.bai
    c. G3_P_K4me3_rep2_chr8_rmDup.sorted.bam

    d. G3_P_K4me3_rep2_chr8_rmDup.sorted.bam.bai
2. Metrics files
   G3_P_H3_chr8_BWA_bam_rmDup_metrics:
    a. G3_P_H3_rep1_chr8_rmDup_metrics.txt
    b. G3_P_H3_rep2_chr8_rmDup_metrics.txt
   G3_P_K4me3_chr8_BWA_bam_rmDup_metrics:
    a. G3_P_K4me3_rep1_chr8_rmDup_metrics.txt
    b. G3_P_K4me3_rep2_chr8_rmDup_metrics.txt