

## **Picard.MarkDuplicates Documentation**

**Description:** Flags duplicate reads in a SAM or BAM file.

Author: Picard team, gp-help@broadinstitute.org

### Summary

Examines aligned records in the supplied SAM or BAM file to locate duplicate reads. All records are then written to the output file with the duplicate records flagged. For more details on the SAM/BAM format, see the specification here: <a href="http://samtools.sourceforge.net/">http://samtools.sourceforge.net/</a>.

This module wraps the Picard MarkDuplicates function.

#### **Parameters**

Name	Description
input.file (required)	Input file (BAM or SAM).  NOTE: The input file must be coordinate sorted.
remove. duplicates (required)	If yes, do not write duplicates to the output file. If no, write duplicates to the output file with appropriate flags set. Default: no
max.file. handles (required)	Maximum number of file handles to keep open during execution of the module. Set this number a little lower than the perprocess maximum number of files that may be open. This number can be found via the ulimit -n command on a Unix system. Default: 8000
read.name. regex	Regular expression that can be used to parse read names in the input file. Read names are parsed to extract three variables: tile/region, x-coordinate, and y-coordinate. These values are used to estimate the rate of optical duplication in order to give a more accurate estimated library size. The regular expression should contain three capture groups for the three variables, in order. An example might be: $   [a-zA-Z0-9]+:[0-9]:([0-9]+):([0-9]+):([0-9]+).*  $

# GenePattern

optical. duplicate. distance (required)	The maximum offset between two duplicate clusters in order to consider them optical duplicates. This should usually be set to a fairly small number (e.g., 5-10 pixels) unless you are using later versions of the Illumina pipeline that multiply pixel values by 10, in which case 50-100 is more appropriate.
sorting.collectio n.size.ratio	This number, plus the maximum RAM available to the JVM, determine the memory footprint used by some of the sorting collections. If you are running out of memory, try reducing this number. Default value: 0.25.
output.prefix (required)	The prefix of the output SAM or BAM file.

### **Output Files**

1. SAM/BAM file

A SAM or BAM file (depending on the input format) with the duplicate reads either removed or flagged. For more details on the SAM/BAM format, see the specification here: http://samtools.sourceforge.net/.

<input.file\_basename>.metrics.txtText file containing duplication metrics.

## **Platform Dependencies**

Module type: Preprocess & Utilities

CPU type: any OS: any

Language: Java (minimum version 1.6)