AgilentMBCDedup

Read Me File

AgilentMBCDedup is a program to process the Molecular Barcode (MBC) information of a HaloPlex^{HS} Illumina(c) run. AgilentMBCDedup will tag read pairs in a bam/sam file with their MBC sequences read out of the index 2 FASTQ file(s), and mark or remove MBC duplicates from that sam/bam file. This is the very same module that is part of the SureCall package, offered as a Windows or MacOS X application.

Command line syntax

```
java -Xmx24G -jar AgilentMBCDedup.jar [-X temp_directory] [-t temp_directory]
[-D] [-U] [-IB] [-OS] -b amplicons.bed -o output_file_name
input_bam_file_name index2_fastq_file_1[.gz] [... index2_fastq_file_N[.gz]]
```

This jar was compiled using Java version 8, make sure your java command is at least at that version number by running "java –version". 12GB RAM is enough for most 4M read files, tailor that parameter to your data size and complexity (large designs use more memory than small panels).

Options

-X temp_directory: location of temporary intermediate bam files used to store overflow of matches. Intermediate files will be deleted at program exit.

-t temp_directory: location of temporary intermediate bam files used to store overflow of matches. Intermediate files are not deleted.

-D: Duplicates out, mark duplicates but output all input records annotated. Default is merge duplicates and output only the consensus read pair per barcode.

-U: unsorted bam/sam output - Faster and requires less RAM.

-q5: MBC base quality threshold expressed in 33 biased FASTQ quality encoding (5 means (ASCII_VALUE_OF '5') - 33 = 20). Default is 20 or '5'. -qz disables the thresholding.

-IB or -IS: input file is BAM or SAM, default is SAM as it is convenient to run AgilentMBCDedup right after BWA.

-OB or -OS: output file is BAM or SAM, default is BAM.

-N number: number of read pairs to process before an intermediate bam file is written and memory is cleared. Default is 2000000. Increasing that number increases memory needs and decreases computation time.

-b amplicons.bed: amplicon bed file downloaded from Agilent SureDesign's website.

-o output_file_name: name of the file generated by AgilentMBCDedup.

input_bam_file_name: name of the input BAM or SAM file,

index2_fastq_file_1: input fastq file or files containing the barcode sequences for the read pairs in the input bam file. It's OK if some records have been filtered out during processing, i.e. if the fastq files have more records than the BAM file.

Output format

The output of AgilentMBCDedup is an annotated bam/sam file. If the duplicates are merged the output looks like the following example:

```
AmpliconName:0 99 chr1 156084500 60 149M = 156084533 182 TCAGTGTT...
AABA3D@D... c1:Z:0,5 MD:Z:149 XF:Z:AmpliconName XI:i:1 NM:i:0 XM:Z:CATATCCTAA
XQ:Z:CCCCCFFFFF AS:i:149 rd:Z:[HWI-M00168:344:000000000-
AJ1A2:1:2105:18304:28543]
```

The name of the consensus record is the name of the amplicon that was matched for that pair of set of pairs. The read quality is the consensus quality, i.e. for each base the highest quality for all teh reads in that set. The special tags used are:

XM: molecular barcode sequence

XQ: molecular barcode base qualities

XI: Number of reads that were merged

XF: name of the amplicon that was matched

rd[...]: list of read names that were merged in that record.

c1 and/or c2: this is the location in read sequence coordinates (0 based) of the beginning and or end of the amplicon capture zone. Because capture does not happen if the match between amplicon and read is not exact, especially for the first/last 5 bases, those locations will show a reference bias, or rather a bias for the sequence they were designed for.

al: alternate set of reads for that amplicon and barcode. If there are several distinct read sequences that have been seen for a amplicon and barcode set, the most likely sequences are chosen by number of read pairs and base qualities, and the other sequences are summarized in the al tag.

If the duplicates are marked but kept in the output, here's what the output looks like:

```
HWI-M00168:344:000000000-AJ1A2:1:2105:18304:28543 99 chr1 156084500 60 149M =
156084533 182 TCAGTGTT...  AABA3D@D... c1:Z:0,5 MD:Z:149 XF:Z:AmpliconName
XI:i:1 NM:i:0 XM:Z:CATATCCTAA XQ:Z:CCCCCFFFFF AS:i:149  XS:i:19 rd:Z:[HWI-
M00168:344:000000000-AJ1A2:1:2105:18304:28543] bq:Z:CCCCCFFFFF rq:Z:AABA...
```

The record is the same as the input record, decorated with the same tags as before, with the following changes:

The read base qualities and molecular base qualities are the quality for that record. The consensus qualities for the whole set are reported in the rq and bq tags for read and base qualities. The read that is chosen to be the representative read for that set is the one with the best average read1-read2-mbc base quality.

The other reads in that set are written in the output with the SAM flag PCR duplicate set. They are decorated with their barcode and individual barcode quality.

In both cases if a consensus read has a consensus molecular barcode which has one or more base qualities that are below the barcode quality threshold, the SAM flag has the "QC failed" bit set.

Example

To mark duplicates after running bwa (i.e. using a SAM file) and create an unsorted BAM output:

```
/etc/alternatives/java_sdk_1.8.0/bin/java -Xmx24G -jar
/home/my_name/bin/AgilentMBCDedup.jar -U -D -b 04818-1398248357.Amplicons.bed
-o test.bam FRLJ100715HaloS4.sam FRLJ100715HaloS4.R2.fastq.gz
```