Using Unique Molecular IDs with with NEBNext Direct™ - Data Usage Guideline Page Data

In order to reap the benefits of the Unique Molecular IDs (UMIs) included in NEBNext Direct libraries it is necessary to:

1. Insert the UMI sequences into your SAM/BAM files
2. Mark or remove duplicate reads with a program that is UMI-aware

The following two sections contain recommendations for both steps.

Inserting UMIs into SAM/BAM Files

The simplest way to introduce UMIs into an existing pipeline is to use a tool that will insert UMIs into pre-existing SAM/BAM files.   For this purpose we recommend using the *AnnotateBamWithUmis* tool from the *fgbio* package, available from:

https://github.com/fulcrumgenomics/fgbio

Follow the instructions on the website to download and unpack the package. Following that, UMIs are added to a BAM using:

```
java -Xmx4g -jar fgbio.jar AnnotateBamWithUmis \
    -i in.bam -f umi.fastq -o out.bam
```

By default this will create a new file, *out.bam*, identical the the file *in.bam*, but with an additional tag RX that contains the UMI sequence read for each record. To override the tag name used to store the UMI, use the -t option:

```
java -Xmx4g -jar fgbio.jar AnnotateBamWithUmis \
    -i in.bam -f umi.fastq -o out.bam -t XX
```

Finally full help and usage information is avaiable by runing:

```
java -Xmx4g -jar fgbio.jar AnnotateBamWithUmis --help
```

Note: if your pipeline starts with Illumina BCL files instead of Fastq files you may wish to explore using *IlluminaBasecallsToSam* from the Picard package which is capable of converting from BCL to BAM, demultiplexing, and inserting UMI tags in a single pass. For more information see:
http://broadinstitute.github.io/picard/command-line-overview.html#IlluminaBasecallsToSam

Marking Duplicates with UMIs

For marking PCR duplicates within a library we recommend using *MarkDuplicates* from the *Picard* suite of tools. Duplicating marking with UMIs is an optional feature, and must be turned on explicitly. To mark duplicates in a BAM file containing UMIs run:

```
java -Xmx4g -jar picard.jar MarkDuplicates \
  INPUT=in.bam \
  OUTPUT=out.bam \
  METRICS_FILE=metrics.txt \
  BARCODE_TAG=RX
```

If a different tag name was provided earlier, substitute that tag name for *RX* here.

For further details and the full set of available options to *MarkDuplicates* refer to the online documentation:
http://broadinstitute.github.io/picard/command-line-overview.html#MarkDuplicates