Program developed by Rachael Bashford-Rogers (2020)
At the Wellcome Centre for Human Genetics, University of Oxford
(rbr1@well.ox.ac.uk)

1. Introduction

This manual provides an outline of the processing pipeline for NGS data based on the sUMI protocols developed in the lab. This pipeline performs the following steps, split into stages:

Stage 1:

a) Join forward and reverse reads (merging)
b) Split sequences according to sample barcode

Stage 2:

c) Identify RNA barcode and collapse/error correct based on groups of sequences sharing same barcode

If you find the methods in sUMI_PROCESSING_PIPELINE, please cite the following reference: Ahmed et al. Ultrasensitive amplicon barcoding for next-generation sequencing facilitating sequence error and amplification-bias correction. 2020

2. User’s guide

a) Installation

Install CD-HIT(*3*):

- Download current CD-HIT at http://bioinformatics.org/cd-hit
- Unpack the file with “tar xvf cd-hit-XXX.tar.gz --gunzip”
- Compile the programs by “make”

Ensure that the following python module dependencies are installed:

- re, math, sys, collections, os, operator, commands

b) Usage

```
python sUMI_PROCESSING_PIPELINE.py <input directory> <sample
ID> <sample ID of pooled library> <sUMI primer file> <sUMI
sample barcode id> <sample barcode file> <cd-hit location>
```

Where:

Input command	**Description**
input directory	Directory where source and output files will be located. Note that a TMP directory will be created here.
sample ID	ID of the sample to be analysed.
sample ID of pooled library	ID of the sequence library from which the sample sequences will be filtered. These must be in the following format for forward and reverse paired end reads respectively: /input directory/Sequences_ID_1.fasta /input directory/Sequences_ID_2.fasta
sUMI primer file	File containing information about the sUMI primers.
sUMI sample barcode id	ID of the sUMI primer used in for this sample
cd-hit location	Directory where CD-hit is installed.

For example:

```
python sUMI_PROCESSING_PIPELINE.py EXAMPLE_DIRECTORY/
GR3_CLL_37 MB_CLL_1 Primers_MALBAC_barcoded_FR2_table.txt
MALBAC_UNI_Ind8 Barcode_groups_FR2_MALBAC.txt /local/cd-hit-
v4.5.7-2011-12-16/
```