# Depuper Pseudocode

Logan Lewis

10/15/2021

## The problem

PCR is often required for the library preps of sequencing experiments. However, the process of amplification during PCR is often biased due to several reasons, such as GC content or 3-D confirmation of the molecule. This is a problem, especially for RNA-seq data where an accurate count of reads is needed to measure expression level. Therefore, a program able to filter out these duplicates is needed.

## Pseudocode

open input SAM file
open file of known UMIs
open an output file for writing

sort the SAM file by chromosome using Pysam package

read in the known UMIs as a list

initiate an empty dictionary (called Deduped_lines)

for line in SAM file:/ if line starts with "@", skip to the next line #to remove headers
Else:
Save the QNAME to a variable using splitline #note: this will not overwrite the orignal line string
Split the QNAME by ":" anddd grab the last one #This is the UMI
If the UMI in not is list of given UMIs, skip to the next line
Else:
Save the starting postion to a variable using splitline
Save the CIGAR string to a variable using splitline
Save the Bitwise FLAG to a variable using splitline
Save the chromosome to a variable using splitline
If the chromosome is different than the previous line's chromosome:
write all values of Deduped_lines dictorary to a file
empty the dictionary
continue in this loop
Call clip_check() function to adjust the starting position
If Bitwise FLAG shows the stand is Reverse Compliment, set a varible(called Rev_comp_status) to True
Make a tuple with (UMI, Starting position, Rev_comp_status)
If this tuple is already in the Deduped_lines dictonary, skip to the next line
Else:
Make this tuple the key to a dictonary, the value is the original line string

## Function

clip_check(CIGAR_string, starting_POS):
’‘’Takes a CIGAR string and a starting postion and adjusts the starting position to account for soft clipping by addind the number before the S to the starting position”’
If the second character of a cigar string is “S”:
split the string by “S” and grab what came before it #this should be a number
cast that character as an int (called adjust_amount)
return starting_POS + adjust_amount
Else:
return starting_POS

Example input: (CIGAR_string = “2S48M”, starting_POS = 300)
Example output: 302

Example input: (CIGAR_string = “50M”, starting_POS = 100)
Example output: 100