Gera

Deduper psuedocode

Due to the limitations of template yields from RNA extraction protocols, amplification is required to get reads from the sequencer. The amplification is done through PCR which uses primers to bind to the template strand. As a result, there is an unbalance proportion of reads for each strand. Since the number of reads is important for gene expression these duplicates need to be remove. Through the use of the Sam file, we can use the alignment section and use certain fields to remove the PCR duplicates. The fields that will use are RNAME(3) for chromosome, position (4) and the bitwise flag (2) for strand specific. For the bitwise flag, were assuming all reads are mapped, will be using the bit 16 for reverse complement. In addition, we will be using the CIGAR field to find soft clipping. Soft clipping helps determining if the read is a duplicate through finding the position on the reference genome. Finally, we will use QNAME field for known Same unique Molecular Index (UMIs).

Pseudocode:

1. Initialize argparse to take a Sam file of uniquely mapped reads.
2. Initialize argparse to output name for the Sam file.
3. Create functions:
4. Create a method to check the bitwise flag:

Def check\_bitwise(num)

“Takes an integer and returns reverse if the bitwise flag is the reverse complement,”

-converts the index one in the array into to an integer and stores it as a variable.

-Using if statements if flag equals 16 for reverse strand.

- if true return reverse, else return forward

Return reverse or forward

Input: 83

Output: reverse

1. Create a method to write files

Def write\_file(I, f)

“Takes two parameters, one the string to write and the other the file to write in on.”

-Use the write function to write the string to the specific file.

No return

Input: K00337:113:HN7KGBBXX:4:2105:11860:30767-ACGAAGGT^AGTGCTGT 99….

Output: No output

Def soft\_clip\_forw(chromosome\_location, CIGAR\_string)

“Takes the chromosome location and Cigar string, returns the new chromosome location. This is the forward strand”

Finds the first s character in the string and stores all characters prior to the s as an integer.

Subtracts the integer to the chromosome location and returns new location

Return new\_location

Input: 502, 2S50M

Output: 550

Def soft\_clip\_rev(chromosome\_location, CIGAR\_string)

“Takes the chromosome location and Cigar string, returns the new chromosome location. This is the reverse strand”

Checks if there is a first s character in the string, if true, delete anything prior to that.

Checks if there is in I in the string, if true, delete prior until you get to a character.

Split the string by characters and convert the remaining strings to integers.

Sum the integers and add to chromosome location.

Adds the integer to the chromosome location and returns new location

Returns new\_location

Input:502 , 3S50M2S

Output: 554

1. Read in the sam file and sort by chromosome and position.
2. Set a while loop to be open.

- Set a for loop to iterate two lines

-If statement to start at the header position of each read.

-Split each line each and store to a variable.

-check strand orientation between two reads, Initial and next read, by check\_bitwise().

-if both reads are the opposite strand, then strands are not pcr duplicates. Write the initial read to file and next read will be the initial read.

-Move down the following read and call that the next read.

-Nested if statements:

-if the strand orientation is the same than use the either soft\_clip\_rev() or soft\_clip\_forw() functions depending orientation to get new starting position.

-if the UMI’s in both reads are the same then move down the line for the next read to compare to the initial read. Else, if either if statements are false then write the initial read to file and next read will be the initial read.