Deduper Part I

Maddy Griswold

October 22, 2018

Problem: PCR is known to be biased towards which reads are duplicated. This means that there is an uneven spread of reads based soley on how well they replicate in PCR. PCR is needed in Illumina sequencing to get enough material for the sequencing to actually work. Having repeat reads will mess up downstream analysis like gene expression levels. Duplicates from PCR can look like high level of gene expression when it really is just a by product of the process.

Examples:

INPUT

header:CTGTTCAC 0 2 76814284 36 71M * 0 0 sequence QS rand1 rand2 rand3 rand4 rand5 rand6 rand7 rand8

header:CTGTGATG 0 2 76814284 36 2S90M * 0 0 sequence QS rand1 rand2 rand3 rand4 rand5 rand6 rand7 rand8

header:GATACAGT 0 3 19836572 36 3S71M * 0 0 sequence QS rand1 rand2 rand3 rand4 rand5 rand6 rand7 rand8

header:GATACAGT 0 3 19836575 36 71M * 0 0 sequence QS rand1 rand2 rand3 rand4 rand5 rand6 rand7 rand8

header:AGATCAGA 0 3 19836575 36 50M21N71M * 0 0 sequence QS rand1 rand2 rand3 rand4 rand5 rand6 rand7 rand8

header:CGGTTACC 0 4 13465256 36 7S68M * 0 0 sequence QS rand1 rand2 rand3 rand4 rand5 rand6 rand7 rand8

header:CGGTTACC 0 4 13465263 36 71M * 0 0 sequence QS rand1 rand2 rand3 rand4 rand5 rand6 rand7 rand8

OUTPUT

header:CTGTTCAC 0 2 76814284 36 71M * 0 0 sequence QS rand1 rand2 rand3 rand4 rand5 rand6 rand7 rand8

header:CTGTGATG 0 2 76814284 36 2S90M * 0 0 sequence QS rand1 rand2 rand3 rand4 rand5 rand6 rand7 rand8

header:GATACAGT 0 3 19836575 36 71M * 0 0 sequence QS rand1 rand2 rand3 rand4 rand5 rand6 rand7 rand8

header:CGGTTACC 0 4 13465263 36 71M * 0 0 sequence QS rand1 rand2 rand3 rand4 rand5 rand6 rand7 rand8

Psuedocode:

1. remove header lines {bash}
    - add to output file
2. sort samfile using samtools on left most position {bash}
3. use a sliding window of 50 (variable, might need more rows) {rest in python}
    - add to numpy array
    - columns for each column in the sam file (19) - plus 2 for UMI and adjusted position (21 total)
    - rows for each read (50)
4. add a read to each row
5. check for errors in UMIs and add column at end with UMI (col 20)
    - throw read out if Ns
    - repeat step 3 and 4 until good read is added to array or end of file
6. check CIGAR string (col 5) for insertions, deletions, and Ns (I,D,N)
    - adjust or throw out if needed
7. adjust for soft clipping (col 5) as soon as you put read into the array
    - add number of 5' soft clipping to the start position and put into new column at end (col 21)
8. check top read against rest of numpy array
    - check uniqueness against position (col 3), chromosome (col 2), UMI (col 20)
    - if duplicate, stop checking, pop off top read, and throw out
    - if unique, pop off top read and write to file
9. add next read, repeat steps 5-8 until file is finshed
    - array at end will not have 50 reads

Functions:

```
#*****************************************************************************
#Function     : addRead
#Description: add sam file read to numpy array as stringsas row with 2 columns
#             for UMI and updated starting position
#Parameters : arry - numpy array contianing sliding window
#             read - read from sam file
#             file - OR FILE TO READ FROM??
#Returned     : none (Passed by reference)
#Test Case  : header:CTGTTCAC   0   2   76814284    36  71M *   0   0
#             sequence  QS  rand1   rand2   rand3   rand4   rand5   rand6   rand7   rand8

#             header:AATTCGAG   0   2   82938523    36  4S68M   *   0   0
#             sequence  QS  rand1   rand2   rand3   rand4   rand5   rand6   rand7   rand8

#             header:AATTCGAG   0   2   82938527    36  68M *   0   0
#             sequence  QS  rand1   rand2   rand3   rand4   rand5   rand6   rand7   rand8

#Results    : | header:CTGTTCAC |   0   | 2 | 76814284 |    36 |    71M | * | 0 | 0 |
#             sequence  | QS    | rand1 | rand2 | rand3 | rand4 |   rand5   |
#             rand6 | rand7 | rand8 | UMI | adjPos |

#             | header:AATTCGAG | 0 | 2 | 82938523  | 36    | 4S68M | * | 0 | 0 |
#             sequence  | QS    | rand1 | rand2 | rand3 | rand4 | rand5 | rand6 |
#             rand7 | rand8 | UMI | adjPos |

#             | header:AATTCGAG | 0 | 2 | 82938527  | 36    | 68M   | * | 0 | 0 |
#             sequence  | QS | rand1    | rand2 | rand3 | rand4 | rand5 | rand6 |
#             rand7 | rand8 | UMI | adjPos |
#*****************************************************************************
```

```
#*****************************************************************************
#Function     : popRead
#Description: pop the top read for either trash (duplicate)
#             or writing to file (unique) removing last 2 colums
#Parameters : arry - numpy array contianing sliding window
#Returned     : read - top read in the numpy array
#Test Case  : [0] | header:CTGTTCAC |   0   | 2 | 76814284 |    36 |    71M | * | 0 | 0 |
#             sequence  | QS    | rand1 | rand2 | rand3 | rand4 |   rand5   |
#             rand6 | rand7 | rand8 | UMI | adjPos |
#             [1] | header:AATTCGAG | 0 | 2 | 82938523  | 36    | 4S68M | * | 0 | 0 |
#             sequence  | QS    | rand1 | rand2 | rand3 | rand4 | rand5 | rand6 |
#             rand7 | rand8 | UMI | adjPos |
#             [2] | header:AATTCGAG | 0 | 2 | 82938527  | 36    | 68M   | * | 0 | 0 |
#             sequence  | QS | rand1    | rand2 | rand3 | rand4 | rand5 | rand6 |
#             rand7 | rand8 | UMI | adjPos |
#Results    : header:CTGTTCAC   0   2   76814284    36  71M *   0   0
#             sequence  QS  rand1   rand2   rand3   rand4   rand5   rand6   rand7   rand8
#*****************************************************************************
```

```
#*****************************************************************************
#Function     : writeToFile
#Description: write unique reads to a file
#Parameters : read - unique read to be written to a file
#             file - output file
#Returned     : none
#Test Case  : header:CTGTTCAC   0   2   76814284    36  71M *   0   0
#             sequence  QS  rand1   rand2   rand3   rand4   rand5   rand6   rand7   rand8
#Results    : same as above in file
#*****************************************************************************
```

```
#*****************************************************************************
#Function     : adjustPos
#Description: adjust the left most position taking into account soft clipping
#Parameters : read - sam file read with information on soft clipping and position
#Returned     : read - updated read with adjusted start position
#Test Case  : | header:CTGTTCAC |   0   | 2 | 76814284 |    36 |    71M | * | 0 | 0 |
#             sequence  | QS    | rand1 | rand2 | rand3 | rand4 |   rand5   |
#             rand6 | rand7 | rand8 | UMI | adjPos |

#             | header:AATTCGAG | 0 | 2 | 82938523  | 36    | 4S68M | * | 0 | 0 |
#             sequence  | QS    | rand1 | rand2 | rand3 | rand4 | rand5 | rand6 |
#             rand7 | rand8 | UMI | adjPos |

#             | header:AATTCGAG | 0 | 2 | 82938527  | 36    | 68M   | * | 0 | 0 |
#             sequence  | QS | rand1    | rand2 | rand3 | rand4 | rand5 | rand6 |
#             rand7 | rand8 | UMI | adjPos |

#Results    : | header:CTGTTCAC |   0   | 2 | 76814284 |    36 |    71M | * | 0 | 0 |
#             sequence  | QS    | rand1 | rand2 | rand3 | rand4 |   rand5   |
#             rand6 | rand7 | rand8 | UMI | 76814284 |

#             | header:AATTCGAG | 0 | 2 | 82938523  | 36    | 4S68M | * | 0 | 0 |
#             sequence  | QS    | rand1 | rand2 | rand3 | rand4 | rand5 | rand6 |
#             rand7 | rand8 | UMI | 82938527 |

#             | header:AATTCGAG | 0 | 2 | 82938527  | 36    | 68M   | * | 0 | 0 |
#             sequence  | QS | rand1    | rand2 | rand3 | rand4 | rand5 | rand6 |
```

```
#             rand7 | rand8 | UMI | 82938527 |
#******************************************************************************
```

```
#******************************************************************************
#Function     : getUMI
#Description: get the UMI from the header line (last 8 characters in col 0)
#             and add column with UMI
#Parameters : read - sam file read with header column
#Returned     : read - updated read with adjusted start position
#Test Case  : | header:CTGTTCAC |   0   | 2 | 76814284 |    36 |    71M | * | 0 | 0 |
#             sequence  | QS    | rand1 | rand2 | rand3 | rand4 |   rand5   |
#             rand6 | rand7 | rand8 | UMI | adjPos |

#             | header:AATTCGAG | 0 | 2 | 82938523  | 36    | 4S68M | * | 0 | 0 |
#             sequence  | QS    | rand1 | rand2 | rand3 | rand4 | rand5 | rand6 |
#             rand7 | rand8 | UMI | adjPos |

#             | header:AATTCGAG | 0 | 2 | 82938527  | 36    | 68M   | * | 0 | 0 |
#             sequence  | QS | rand1    | rand2 | rand3 | rand4 | rand5 | rand6 |
#             rand7 | rand8 | UMI | adjPos |

#Results    : | header:CTGTTCAC |   0   | 2 | 76814284 |    36 |    71M | * | 0 | 0 |
#             sequence  | QS    | rand1 | rand2 | rand3 | rand4 |   rand5   |
#             rand6 | rand7 | rand8 | CTGTTCAC | adjPos |

#             | header:AATTCGAG | 0 | 2 | 82938523  | 36    | 4S68M | * | 0 | 0 |
#             sequence  | QS    | rand1 | rand2 | rand3 | rand4 | rand5 | rand6 |
#             rand7 | rand8 | AATTCGAG | adjPos |

#             | header:AATTCGAG | 0 | 2 | 82938527  | 36    | 68M   | * | 0 | 0 |
#             sequence  | QS | rand1    | rand2 | rand3 | rand4 | rand5 | rand6 |
#             rand7 | rand8 | AATTCGAG | adjPos |

#******************************************************************************
```

```
#******************************************************************************
#Function     : compareRows
#Description: compare the first row to the rest in terms of chromosome,
#             position, and UMI
#Parameters : arry - numpy array contianing sliding window
#Returned     : uniq - bool
#               TRUE  - unique read
#               FALSE - duplicate read
#Test Case  : [0] | header:CTGTTCAC |   0   | 2 | 76814284 |    36 |    71M | * | 0 | 0 |
#             sequence  | QS    | rand1 | rand2 | rand3 | rand4 |   rand5   |
#             rand6 | rand7 | rand8 | UMI | 76814284 |
#             [1] | header:AATTCGAG | 0 | 2 | 82938523  | 36    | 4S68M | * | 0 | 0 |
#             sequence  | QS    | rand1 | rand2 | rand3 | rand4 | rand5 | rand6 |
#             rand7 | rand8 | UMI | 82938527 |
#             [2] | header:AATTCGAG | 0 | 2 | 82938527  | 36    | 68M   | * | 0 | 0 |
#             sequence  | QS | rand1    | rand2 | rand3 | rand4 | rand5 | rand6 |
#             rand7 | rand8 | UMI | 82938527 |

#             [0] | header:AATTCGAG | 0 | 2 | 82938523  | 36    | 4S68M | * | 0 | 0 |
#             sequence  | QS    | rand1 | rand2 | rand3 | rand4 | rand5 | rand6 |
#             rand7 | rand8 | AATTCGAG | 82938527 |
#             [1] | header:AATTCGAG | 0 | 2 | 82938527  | 36    | 68M   | * | 0 | 0 |
```

```
#             sequence  | QS | rand1    | rand2 | rand3 | rand4 | rand5 | rand6 |
#             rand7 | rand8 | AATTCGAG | 82938527 |


#             [0] | header:AATTCGAG | 0 | 2 | 82938527  | 36    | 68M   | * | 0 | 0 |
#             sequence  | QS | rand1    | rand2 | rand3 | rand4 | rand5 | rand6 |
#             rand7 | rand8 | AATTCGAG | 82938527 |

#Results    : TRUE
#             FALSE
#             TRUE
#*****************************************************************************
```