Devin D
Bi 624
De-Duper Part1

PCR duplicates are two or more reads that come from the same original sequence.

AACCTTGG-original seq
AACCTTGG –duplicate
AACCTTGG-duplicate

Duplicates arise during the library amplification steps of preparation. If certain sequences happen to amplify really well you can get uneven amplification. Uneven amplification can lead to skewed sequence data. Removing PCR duplicates can be useful for reducing over-amplification data bias.

Algorithm design / pseudocode

- Write a function to check sequence file UMI's against all known UMI's
    - Remove any sequences with unknown UMI
    - Def UMICheck()
        - Compare sequence umi against list of known UMI's
        - return lines that pass UMI check to new .sam file
- Run UMI check function
- Write a function to search the cigar string for softclipping
    - Adjust the alignment start position if Softclipping occurs
    - Def SoftCheck()
        - Grab cigar string determine if softclipping ie(2S99M)
        - Adjust start position to reflect presence of soft clipping
            - If SPOS= 12 and 2S99M new SPOS = 10
        - Return adjusted start position
- Write a function that will compare duplicates for best possible alignment ie least softclipping
    - Def RetainRead()
        - If no soft clipping retain read
        - If soft clipping store clipping amount in variable for comparison to other reads with same start and soft clipping
        - Lowest amount is retained.
- Use Sam tools sort to align file by chromosome and alignment position
- Open the sorted file
    - for line in sortedFile
        - Split the line into parts
        - Run soft clipping function
        - Retain chromosome #, Alignment start Position and UMI as a tuple(1)
    - Read the next line
        - Split into parts
        - Run soft clipping function
        - Retain chromosome #, Alignment start Position and UMI as a tuple(2)
    - If tuple(1) == tuple(2)
        - Match = Duplicate found

- Run function to determine which read should be kept
- Continue until no match found

- Else tuple(1) != tuple(2)
  - Tuple(1) is retained if read at this start position not already retained
  - tuple(2) becomes tuple(1) and the next line of the file is read and compared

*NOTES

In order to not waste time looking over reads I have already determined are at same start Position as the previous (duplicates) Ideally I would skip lines I have already read on subsequent passes.

Example Sam input

```
HWI-ST354R:351:C0UPMACXX:5:1115:20112:ACCTGG 99 1 861268 60 100M = 861543 375
TCCCTCACAGGGTCTGCCTCGGCTCTGCTCGCACGGGAAAAGTCTGAAGACGCTTATGTCCAAGGGGATCCTGCAGGTGCATCCTCCGATCTGCGA
CTGCC CCCFFFFFHHHGFHJIIJJJJJIJJJJJJJJIIJJIIJJIGCHCHGGIGIIJIJGHGFFFFFFFDD@BDCCCDDDDDCDDECC@C9<@BBDDDDDDD59>
MC:Z:100M MD:Z:100 NM:i:0 MQ:i:60 AS:i:100 XS:i:0 RG:Z:1719PC0017_51

HWI-ST354R:351:C0UPMACXX:6:2301:2151:AAACCC 99 1 861307 60 100M = 861414 207
AGTCTGAAGACGCTTATGTCCAAGGGGATCCTGCAGGTGCATCCTCCGATCTGCGACTGCCCGGGCTGCCGAATATCCTCCCCGGTGGTGAGATGC
GGGG CCCFFFFFHHHHHHJJJJJJJJJJJJJJIIJJJJJJJGHIIJJJJJJJGJJJJIJIHHFFFFDDDDDDDDDBDDDEDEEDDDDDD>BD9@DD@CCCDBB>
MC:Z:100M MD:Z:100 NM:i:0 MQ:i:60 AS:i:100 XS:i:0 RG:Z:1719PC0017_51

HWI-ST354R:351:C0UPMACXX:5:1110:16080:AAACCC 99 1 861308 60 1S99M = 861360 152
CGTCTGAAGACGCTTATGTCCAAGGGGATCCTGCAGGTGCATCCTCCGATCTGCGACTGCCCGGGCTGCCGAATATCCTCCCCGGTGGTGAGATGC
GGGG @CCFFFFFHHHHHJJJJIJJJIGIJJIGIJJIJJJJGGIIJJJIIJIIIIJJJJJGHHFFFFDBDDDDDDDDDDEEEDDDBDDB59><@ABCDDDDDBD7
MC:Z:100M MD:Z:100 NM:i:0 MQ:i:60 AS:i:100 XS:i:0 RG:Z:1719PC0017_51

HWI-ST354R:351:C0UPMACXX:5:1110:16080:GGGCAA 147 1 861360 60 100M = 861308 -152
CGACTGCCCGGGCTGCCGAATATCCTCCCCGGTGGTGAGATGCGGGGCTCGGTTGGGGCTGGGAGTTACTCTCCCCTGCGGAGCTTGTCCCTGCG
GTTTT 0?3@B@>09DB@><><CCEDCA<BB<D@B9DDDDDCDDDDDDDDDDDBDDDFFFHHHHJIIIIIGGHFGAFAIJJJJFJJJJJJJIHHHHHHFFFFFCCB
MC:Z:100M MD:Z:100 NM:i:0 MQ:i:60 AS:i:100 XS:i:0 RG:Z:1719PC0017_51

HWI-ST354R:351:C0UPMACXX:5:1308:15815:GGGCAA 147 1 861360 60 100M = 861308 -152
CGACTGCCCGGGCTGCCGAATATCCTCCCCGGTGGTGAGATGCGGGGCTCGGTTGGGGCTGGGAGTTACTCTCCCCTGCGGAGCTTGTCCCTGCG
GTTTT 0?3@B@>09DB@><><CCEDCA<BB<D@B9DDDDDCDDDDDDDDDDDBDDDFFFHHHHJIIIIIGGHFGAFAIJJJJFJJJJJJJIHHHHHHFFFFFCCB
MC:Z:100M MD:Z:100 NM:i:0 MQ:i:60 AS:i:100 XS:i:0 RG:Z:1719PC0017_51

HWI-ST354R:351:C0UPMACXX:6:2210:3183:AAAAAA 99 1 861362 60 100M = 861401 139
ACTGCCCGGGCTGCCGAATATCCTCCCCGGTGGTGAGATGCGGGGCTCGGTTGGGGCTGGGAGTTACTCTCCCCTGCGGAGCTTGTCCCTGCGGTT
TTCA @CCFFEFFHFFFHIIIGIIJIIIIJJJIIJ6@G@GHEGIA@EHFBD?BDD?5BDBBDB@BD(8+9:>::A:AC>?<A<9-<&(>?<ACC@:15<55<?3:
MC:Z:100M MD:Z:100 NM:i:0 MQ:i:60 AS:i:100 XS:i:0 RG:Z:1719PC0017_51

HWI-ST354R:351:C0UPMACXX:6:2103:12417:AAAGGG 163 1 861397 60 100M = 861420 123
AGATGCGGGGCTCGGTTGGGGCTGGGAGTTACTCTCCCCTGCGGAGCTTGTCCCTGCGGTTTTCAGGGTTTTCAGGATCGAGAGTCCTAACCTCAC
CCCT CCCFFFFFHHHHHHJJHJJJJJJJJJJHHCHFFFFFFEEDCBDDB;@BBDDBCDCACDDD05@BCCCDD+8?BACCD>CC>@DDB4::>>CBBDCC2?BB0
MC:Z:100M MD:Z:100 NM:i:0 MQ:i:60 AS:i:100 XS:i:0 RG:Z:1719PC0017_51

HWI-ST354R:351:C0UPMACXX:6:2103:12417:AAAGGG 163 1 861399 60 2S98M = 861420 123
CCATGCGGGGCTCGGTTGGGGCTGGGAGTTACTCTCCCCTGCGGAGCTTGTCCCTGCGGTTTTCAGGGTTTTCAGGATCGAGAGTCCTAACCTCAC
CCCT CCCFFFFFHHHHHHJJHJJJJJJJJJJHHCHFFFFFFEEDCBDDB;@BBDDBCDCACDDD05@BCCCDD+8?BACCD>CC>@DDB4::>>CBBDCC2?BB0
MC:Z:100M MD:Z:100 NM:i:0 MQ:i:60 AS:i:100 XS:i:0 RG:Z:1719PC0017_51

HWI-ST354R:351:C0UPMACXX:6:2301:2151:CCCAAA 147 1 861414 60 100M = 861307 -207
GGGGCTGGGAGTTACTCTCCCCTGCGGAGCTTGTCCCTGCGGTTTTCAGGGTTTTCAGGATCGAGAGTCCTAACCTCACCCCTGCGGGTGTGCTGG
AGGG BDDBCCDC:CC>:+B@DBDDDBBDDDDCDBDDB?DDDDDDDEDFEFFFEHHHJIIJJJJIIJJJJJJJIJJIHFIHD?1)JJJIIJJJHHHHHFFFFFCCC
MC:Z:100M MD:Z:100 NM:i:0 MQ:i:60 AS:i:100 XS:i:0 RG:Z:1719PC0017_51

HWI-ST354R:351:C0UPMACXX:6:2103:12417:CCCAAA 83 1 861420 60 100M = 861397 -123
GGGAGTTACTCTCCCCTGCGGAGCTTGTCCCTGCGGTTTTCAGGGTTTTCAGGATCGAGAGTCCTAACCTCACCCCTGCGGGTGTGCTGGAGGGAG
CCTC ?DCDC@::DBDDDDDDDDDDDDDDDDDBDDBDBB=DDDDEEEDDFFFFHHHHHJJJJJJIGIJJJHGJIGGDJJJJJJJJJJJJJJHHHHHFFFFFCCC
MC:Z:100M MD:Z:100 NM:i:0 MQ:i:60 AS:i:100 XS:i:0 RG:Z:1719PC0017_51
```

Example Sam Out

```
HWI-ST354R:351:C0UPMACXX:5:1115:20112:ACCTGG 99 1 861268 60 100M = 861543 375
TCCCTCACAGGGTCTGCCTCGGCTCTGCTCGCACGGGAAAAGTCTGAAGACGCTTATGTCCAAGGGGATCCTGCAGGTGCATCCTCCGATCTGCGA
CTGCC CCCFFFFFHHHGFHJIIJJJJJIJJJJJJJJIIJJIIJJIGCHCHGGIGIIJIJGHGFFFFFFFDD@BDCCCDDDDDCDDECC@C9<@BBDDDDDDD59>
MC:Z:100M MD:Z:100 NM:i:0 MQ:i:60 AS:i:100 XS:i:0 RG:Z:1719PC0017_51

HWI-ST354R:351:C0UPMACXX:6:2301:2151:AAACCC 99 1 861307 60 100M = 861414 207
AGTCTGAAGACGCTTATGTCCAAGGGGATCCTGCAGGTGCATCCTCCGATCTGCGACTGCCCGGGCTGCCGAATATCCTCCCCGGTGGTGAGATGC
GGGG CCCFFFFFHHHHHJJJJJJJJJJJJJJIIJJJJJJJJGHIIJJJJJJJGJJJJIJIHHFFFFDDDDDDDDDBDDDEDEEDDDDDD>BD9@DD@CCCDBB>
MC:Z:100M MD:Z:100 NM:i:0 MQ:i:60 AS:i:100 XS:i:0 RG:Z:1719PC0017_51

HWI-ST354R:351:C0UPMACXX:5:1110:16080:GGGCAA 147 1 861360 60 100M = 861308 -152
CGACTGCCCGGGCTGCCGAATATCCTCCCCGGTGGTGAGATGCGGGGCTCGGTTGGGGCTGGGAGTTACTCTCCCCTGCGGAGCTTGTCCCTGCG
GTTTT 0?3@B@>09DB@><><CCEDCA<BB<D@B9DDDDDCDDDDDDDDDDBDDDFFFHHHHJIIIIIGGHFGAFAIJJJJFJJJJJJJIHHHHHHFFFFFCCB
MC:Z:100M MD:Z:100 NM:i:0 MQ:i:60 AS:i:100 XS:i:0 RG:Z:1719PC0017_51

HWI-ST354R:351:C0UPMACXX:6:2210:3183:AAAAAA 99 1 861362 60 100M = 861401 139
ACTGCCCGGGCTGCCGAATATCCTCCCCGGTGGTGAGATGCGGGGCTCGGTTGGGGCTGGGAGTTACTCTCCCCTGCGGAGCTTGTCCCTGCGGTT
TTCA @CCFFEFFHFFFHIIIGIIJIIIIJJJIJIJ6@G@GHEGIA@EHFBD?BDD?5BDBBDB@BD(8+9:>::A:AC>?<A<9-<&(>?<ACC@:15<55<?3:
MC:Z:100M MD:Z:100 NM:i:0 MQ:i:60 AS:i:100 XS:i:0 RG:Z:1719PC0017_51

HWI-ST354R:351:C0UPMACXX:6:2103:12417:AAAGGG 163 1 861397 60 100M = 861420 123
AGATGCGGGGCTCGGTTGGGGCTGGGAGTTACTCTCCCCTGCGGAGCTTGTCCCTGCGGTTTTCAGGGTTTTCAGGATCGAGAGTCCTAACCTCAC
CCCT CCCFFFFFHHHHHHJJHJJJJJJJJJJHHCHFFFFFFEEDCBDDB;@BBDDBCDCACDDD05@BCCCDD+8?BACCD>CC>@DDB4::>>CBBDCC2?BB0
MC:Z:100M MD:Z:100 NM:i:0 MQ:i:60 AS:i:100 XS:i:0 RG:Z:1719PC0017_51

HWI-ST354R:351:C0UPMACXX:6:2301:2151:CCCAAA 147 1 861414 60 100M = 861307 -207
GGGGCTGGGAGTTACTCTCCCCTGCGGAGCTTGTCCCTGCGGTTTTCAGGGTTTTCAGGATCGAGAGTCCTAACCTCACCCCTGCGGGTGTGCTGG
AGGG BDDBCCDC:CC>:+B@DBDDDBBDDDDCDBDDB?DDDDDDDEDFEFFFEHHHJIIJJJJIIJJJJJJJIJJIHFIHD?1)JJJIIJJJHHHHHFFFFFCCC
MC:Z:100M MD:Z:100 NM:i:0 MQ:i:60 AS:i:100 XS:i:0 RG:Z:1719PC0017_51

HWI-ST354R:351:C0UPMACXX:6:2103:12417:CCCAAA 83 1 861420 60 100M = 861397 -123
GGGAGTTACTCTCCCCTGCGGAGCTTGTCCCTGCGGTTTTCAGGGTTTTCAGGATCGAGAGTCCTAACCTCACCCCTGCGGGTGTGCTGGAGGGAG
CCTC ?DCDC@::DBDDDDDDDDDDDDDDDDDBDDBDBB=DDDDEEEDDFFFFHHHHHJJJJJJIGIJJJHGJIGGDJJJJJJJJJJJJJJHHHHHFFFFFCCC
MC:Z:100M MD:Z:100 NM:i:0 MQ:i:60 AS:i:100 XS:i:0 RG:Z:1719PC0017_51
```