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
 - o 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)
 - o If tuple(1) == tuple(2)
 - Match = Duplicate found

- Run function to determine which read should be kept
- Continue until no match found
- o 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-ST35	TCCCTCACAGGGTCT	5:1115:20112:ACCTGG GCCTCGGCTCTGCTCGC JJJJJJJJJJJJJJJJ	ACGGGAAA			CCDDDDDCI		e@BBDDDI		375 GATCTGCGA
HWI-ST35		5:2301:2151:AAACCC ATGTCCAAGGGGATCC JJJJJJJJJJJJJJGHIJJJJ MD:Z:100 NM:i:0	าาายาาาาเหเ		DDDDDDBD	DDEDEEDD		9@DD@CC0		207 STGAGATGC
HWI-ST354R:351:COUPMACXX:5:1110:16080:AAACCC 99 1 861308 60 1S99M = 861360 152 CGTCTGAAGACGCTTATGTCCAAGGGGATCCTGCAGGTGCATCCTCCGATCTGCGACTGCCCGGGCTGCCGAATATCCTCCCCGGTGGTGAGATGC										
GGGG	@CCFFFFFHHHHHJJJ MC:Z:100М	JIJJJIGIJJIGIJJJJJJGGIJJJ MD:Z:100 NM:i:0		HHFFFFDBD AS:i:100			BDDB59>< PC0017_5	_	DBD7	
HWI-ST35		5:1110:16080:GGGCAA GCCGAATATCCTCCCCG		1 ATGCGGGG	861360 CTCGGTTG	60 GGGCTGG	100M GAGTTACT	= CTCCCCTGC	861308 GGAGCTTG	-152 TCCCTGCG
GTTTT	0?3@B@>09DB@>< MC:Z:100M	:> <ccedca<bb<d@b9i MD:Z:100 NM:i:0</ccedca<bb<d@b9i 		DDDDDDDD AS:i:100			GHFGAFA PC0017_5		IHHHHHFFI	FFFCCB
HWI-ST35		5:1308:15815:GGGCAA GCCGAATATCCTCCCCG		1 ATGCGGGG	861360 SCTCGGTTG	60 GGGCTGGG	100M GAGTTACT	= CTCCCCTGC	861308 GGAGCTTG	-152 TCCCTGCG
GTTTT	0?3@B@>09DB@>< MC:Z:100M	:> <ccedca<bb<d@b9i MD:Z:100 NM:i:0</ccedca<bb<d@b9i 		AS:i:100			GHFGAFA PC0017_5		IHHHHHFFI	FFFCCB
HWI-ST35		6:2210:3183:AAAAAA CGAATATCCTCCCCGGTC		1 GCGGGGCT	861362 CGGTTGGG	60 GCTGGGA	100M GTTACTCTO	= CCCCTGCGG	861401 AGCTTGTC	139 CCTGCGGTT
TTCA	@CCFFEFFHFFFHIIIG MC:Z:100M	IIJIIIJJJIJI6@G@GHEGI MD:Z:100 NM:i:0	_	PBDD?5BDB AS:i:100			>? <a<9-<& PC0017_5</a<9-<& 		15<55 3:</td <td></td>	
HWI-ST35		6:2103:12417:AAAGGG		1	861397	60	100M	=	861420	123
СССТ		GTTGGGGCTGGGAGTT/ JJJJJJJJHHCHFFFFFEE MD:Z:100 NM:i:0	DCBDDB;@		CACDDD05	@BCCCDD+		CC>@DDB4		
HWI-ST35		5:2103:12417:AAAGGG GTTGGGGCTGGGAGTTA		1 TGCGGAGC	861399	60 SCGGTTTC	2S98M	= [CAGGATCG	861420	123
CCCT		JJJJJJJJJHHCHFFFFFEE MD:Z:100 NM:i:0			CACDDD05	@BCCCDD+		CC>@DDB4		
HWI-ST35		6:2301:2151:CCCAAA		1	861414	60	100M	=	861307	-207
AGGG		CTCTCCCCTGCGGAGCT DBDDDBBDDDDCDBD MD:Z:100 NM:i:0	DB?DDDDD		FEHHHJIIJJJ.		IHD?1)JJJI	IJIJНННННF		TGTGCTGG
HWI-ST354R:351:COUPMACXX:6:2103:12417:CCCAAA 83 1 861420 60 100M = 861397 -123										
сстс		CTGCGGAGCTTGTCCCT DDDDDDDDDDDDDBDD MD:Z:100 NM:i:0	BDBB=DDD		FFНННННJJ	าาาาเดิเาาาHG		шшнн		

Example Sam Out

HWI-ST35	4R:351:C0UPMACXX:5	:1115:2011	2:ACCTGG	99	1	861268	60	100M	=	861543	375
	TCCCTCACAGGGTCTG	CCTCGGCT	CTGCTCGC	ACGGGAAA	AGTCTGAA	AGACGCTTA	TGTCCAAG	GGGATCCT	GCAGGTGC	ATCCTCCG	ATCTGCGA
CTGCC	CCCFFFFFHHHGFHJIIJ	וורורורורוורור	JJIIJJIGCHC	HGGIGIIJIJG	HGFFFFFF	DD@BDCC	CDDDDDCD	DECC@C9	@BBDDDD	DDD59>	
	MC:Z:100M	MD:Z:100	NM:i:0	MQ:i:60	AS:i:100	XS:i:0	RG:Z:1719	PC0017_51			
HWI-ST35	4R:351:C0UPMACXX:6	:2301:2151	:AAACCC	99	1	861307	60	100M	=	861414	207
	AGTCTGAAGACGCTTA	ATGTCCAAC	GGGGATCC	FGCAGGTG	CATCCTCCC	SATCTGCGA	ACTGCCCGG	GCTGCCGA	ATATCCTC	CCCGGTGG	TGAGATG
GGGG	CCCFFFFFHHHHHJJJJJ	ווווווווווווווווווווווווווווווווווווווו	IIIIIGHIIIIII.	IIIGIIIIIIHH	HFFFFDDDD	DDDDDBD	DDEDEEDD	DDDD>BD9	@DD@CCC	DBB>	
	MC:Z:100M	MD:Z:100	NM:i:0	MQ:i:60	AS:i:100	XS:i:0	RG:Z:1719	PC0017_51			
HWI-ST35	4R:351:C0UPMACXX:5	:1110:1608	0:GGGCAA	147	1	861360	60	100M	=	861308	-152
	CGACTGCCCGGGCTG	CCGAATAT	CCTCCCCG	STGGTGAG	ATGCGGGG	CTCGGTTG	GGGCTGGG	GAGTTACTC	TCCCCTGC	GAGCTTG	TCCCTGCG
GTTTT	0?3@B@>09DB@><	> <ccedca<< td=""><td>BB<d@b9d< td=""><td>DDDDCDD</td><td>DDDDDDDI</td><td>DBDDDFFFH</td><td>ннныши</td><td>GHFGAFAIJ</td><td>JJJFJJJJJJJJH</td><td>HHHHHFFF</td><td>FFCCB</td></d@b9d<></td></ccedca<<>	BB <d@b9d< td=""><td>DDDDCDD</td><td>DDDDDDDI</td><td>DBDDDFFFH</td><td>ннныши</td><td>GHFGAFAIJ</td><td>JJJFJJJJJJJJH</td><td>HHHHHFFF</td><td>FFCCB</td></d@b9d<>	DDDDCDD	DDDDDDDI	DBDDDFFFH	ннныши	GHFGAFAIJ	JJJFJJJJJJJJH	HHHHHFFF	FFCCB
	MC:Z:100M	MD:Z:100	NM:i:0	MQ:i:60	AS:i:100	XS:i:0	RG:Z:1719	PC0017_51			
HWI-ST35	4R:351:C0UPMACXX:6	:2210:3183	:AAAAAA	99	1	861362	60	100M	=	861401	139
	ACTGCCCGGGCTGCC	GAATATCC1	CCCCGGTG	GTGAGAT	GCGGGGCT	CGGTTGGG	GCTGGGAG	STTACTCTC	CCCTGCGG	AGCTTGTC	CCTGCGGT
TTCA	@CCFFEFFHFFFHIIIGI	111111111111111111111111111111111111111	G@GHEGI	A@EHFBD?	BDD?5BDB	BDB@BD(8	+9:>::A:AC>	>? <a<9-<&(< td=""><td>>?<acc@::< td=""><td>15<55<?3:</td><td></td></td></acc@::<></td></a<9-<&(<>	>? <acc@::< td=""><td>15<55<?3:</td><td></td></td></acc@::<>	15<55 3:</td <td></td>	
	MC:Z:100M	MD:Z:100	NM:i:0	MQ:i:60	AS:i:100	XS:i:0	RG:Z:1719	PC0017_51			
HWI-ST35	4R:351:C0UPMACXX:6	:2103:1241	7:AAAGGG	163	1	861397	60	100M	=	861420	123
	AGATGCGGGGCTCGG	TTGGGGC1	GGGAGTTA	ACTCTCCCC	TGCGGAGC	TTGTCCCT	GCGGTTTTC	AGGGTTTT	CAGGATCG	AGAGTCCT	AACCTCAC
CCCT	CCCFFFFFHHHHHJJHJ	אאווווווווווווווווווווווווווווווווווווו	HFFFFFEE	DCBDDB;@	BBDDBCDC	CACDDD05@	BCCCDD+8	3?BACCD>C	C>@DDB4:	:>>CBBDCC	2?BB0
	MC:Z:100M	MD:Z:100	NM:i:0	MQ:i:60	AS:i:100	XS:i:0	RG:Z:1719	PC0017_51			
HWI-ST35	4R:351:C0UPMACXX:6	:2301:2151	:CCCAAA	147	1	861414	60	100M	=	861307	-207
	GGGGCTGGGAGTTAC	тстсссст	GCGGAGCT	TGTCCCTG	CGGTTTTCA	GGGTTTTC	AGGATCGA	GAGTCCTA	ACCTCACCO	CTGCGGG	TGTGCTGG
AGGG	BDDBCCDC:CC>:+B@	DBDDDBBD	DDDCDBDI	OB?DDDDD	DDEDFEFF	EHHHJIIJJJJ	HEILLILLILLI	IHD?1)JJJJI	IJНННННFI	FFFFCCC	
	MC:Z:100M	MD:Z:100	NM:i:0	MQ:i:60	AS:i:100	XS:i:0	RG:Z:1719	PC0017_51			
HWI-ST35	4R:351:C0UPMACXX:6	:2103:1241	7:CCCAAA	83	1	861420	60	100M	=	861397	-123
	GGGAGTTACTCTCCCC	CTGCGGAG	CTTGTCCCT	GCGGTTTT	CAGGGTTT	TCAGGATC	GAGAGTCC ⁻	ГААССТСАС	CCCTGCGG	GTGTGCTC	GAGGGAG
CCTC	?DCDC@::DBDDDDDI	DDDDDDDD	DDDDBDD	BDBB=DDD	DEEEDDFFI	FFHHHHHJJ.	JJJJIGIJJJHG.	IIGGDJJJJJJ	ннниши	HHFFFFFCC	CC
	MC:Z:100M	MD:Z:100		MQ:i:60		XS:i:0		PC0017_51			