How to get linkage information from SAM mappings using the SAM flag information

Umer Zeeshan Ijaz

I have a SAM file that I generated by aligning paired-end reads against the reference database:

```
bwa mem Mercier_New.fasta ../ID_1884/sample1/1_AGGCAGAA-CTCTCTAT_L001_R1_trim_001.fastq
../ID_1884/sample1/1_AGGCAGAA-CTCTCTAT_L001_R2_trim_001.fastq > NexteraXT_even_1ng_HISEQ_AGGCAGAA-CTCTCTAT.sam
```

The resulting SAM file has 4806207 records (4 million records) and it took 2 minutes and 7.15 seconds to get the linkage information (i.e. reads that match multiple contigs/genomes) using the following one-liner:

```
awk '$2~/^99$|^147$|^83$|^163$|^67$|^131$|^115$|^179$|^81$|^161$|^97$|^145$|^65$|^129$|^113$|^177$/ &&
$2!~/^SN/{print $1"\t"$3"\t"$7}' NexteraXT_even_1ng_HISEQ_AGGCAGAA-CTCTCTAT.sam | sort -n -k1,1 | uniq | awk 'BEGIN
{ FS="\t" } { c[$1]++; l[$1,c[$1]]=$0 } END { for (i in c) { if (c[i] > 1) for (j = 1; j <= c[i]; j++) print l[i,j]
} }'
```

The one-liner returns (few records) the data formatted as follows:

```
HWI-ST1242:77:C13H3ACXX:7:2206:11068:81469   Acidobacterium_capsulatum_ATCC_51196         Bordetella_bronchiseptica_strain_RB50
HWI-ST1242:77:C13H3ACXX:7:2206:11068:81469   Bordetella_bronchiseptica_strain_RB50        Acidobacterium_capsulatum_ATCC_51196
HWI-ST1242:77:C13H3ACXX:7:1306:17569:74649   Thermotoga_petrophila_RKU-1                  Thermotoga_sp._RQ2
HWI-ST1242:77:C13H3ACXX:7:1306:17569:74649   Thermotoga_sp._RQ2                           Thermotoga_petrophila_RKU-1
HWI-ST1242:77:C13H3ACXX:7:2101:13560:96472   Thermotoga_petrophila_RKU-1                  Thermotoga_sp._RQ2
HWI-ST1242:77:C13H3ACXX:7:2101:13560:96472   Thermotoga_sp._RQ2                           Thermotoga_petrophila_RKU-1
HWI-ST1242:77:C13H3ACXX:7:1209:7458:61056    Treponema_denticola_ATCC_35405               Treponema_vincentii_ATCC_35580_NZACYH00000000.1
HWI-ST1242:77:C13H3ACXX:7:1209:7458:61056    Treponema_vincentii_ATCC_35580_NZACYH00000000.1      Treponema_denticola_ATCC_35405
HWI-ST1242:77:C13H3ACXX:7:1301:15983:83574   Bordetella_bronchiseptica_strain_RB50        Rhodopirellula_baltica_SH_1_complete_genome
HWI-ST1242:77:C13H3ACXX:7:1301:15983:83574   Rhodopirellula_baltica_SH_1_complete_genome  Bordetella_bronchiseptica_strain_RB50
HWI-ST1242:77:C13H3ACXX:7:1103:12322:58936   Bordetella_bronchiseptica_strain_RB50        Desulfovibrio_piger_ATCC_29098
HWI-ST1242:77:C13H3ACXX:7:1103:12322:58936   Desulfovibrio_piger_ATCC_29098               Bordetella_bronchiseptica_strain_RB50
HWI-ST1242:77:C13H3ACXX:7:1210:12644:24976   Bacteroides_thetaiotaomicron_VPI-5482        Thermoanaerobacter_pseudethanolicus_ATCC_33223
HWI-ST1242:77:C13H3ACXX:7:1210:12644:24976   Thermoanaerobacter_pseudethanolicus_ATCC_33223      Bacteroides_thetaiotaomicron_VPI-5482
HWI-ST1242:77:C13H3ACXX:7:2108:8801:73269    Sulfurihydrogenibium_yellowstonense_SS-5     SulfuriYO3AOP1
HWI-ST1242:77:C13H3ACXX:7:2108:8801:73269    SulfuriYO3AOP1                               Sulfurihydrogenibium_yellowstonense_SS-5
HWI-ST1242:77:C13H3ACXX:7:2108:7167:90955    Salinispora_arenicola_CNS-205                Shewanella_baltica_OS223,
HWI-ST1242:77:C13H3ACXX:7:2108:7167:90955    Shewanella_baltica_OS223,                    Salinispora_arenicola_CNS-205
HWI-ST1242:77:C13H3ACXX:7:2303:16257:47096   Shewanella_baltica_OS185                     Shewanella_baltica_OS223,
HWI-ST1242:77:C13H3ACXX:7:2303:16257:47096   Shewanella_baltica_OS223,                    Shewanella_baltica_OS185
```

```
HWI-ST1242:77:C13H3ACXX:7:1305:15998:77806   Sulfurihydrogenibium_yellowstonense_SS-5     SulfuriYO3AOP1
HWI-ST1242:77:C13H3ACXX:7:1305:15998:77806   SulfuriYO3AOP1                               Sulfurihydrogenibium_yellowstonense_SS-5
HWI-ST1242:77:C13H3ACXX:7:1215:18228:5604    Rhodopirellula_baltica_SH_1_complete_genome  Salinispora_arenicola_CNS-205
HWI-ST1242:77:C13H3ACXX:7:1215:18228:5604    Salinispora_arenicola_CNS-205                Rhodopirellula_baltica_SH_1_complete_genome
HWI-ST1242:77:C13H3ACXX:7:2109:5989:11352    Bordetella_bronchiseptica_strain_RB50        Leptothrix_cholodnii_SP-6
HWI-ST1242:77:C13H3ACXX:7:2109:5989:11352    Leptothrix_cholodnii_SP-6                    Bordetella_bronchiseptica_strain_RB50
HWI-ST1242:77:C13H3ACXX:7:1203:6513:69746    Treponema_vincentii_ATCC_35580_NZACYH00000000.1     =
HWI-ST1242:77:C13H3ACXX:7:1203:6513:69746    Treponema_vincentii_ATCC_35580_NZACYH00000000.1     Treponema_denticola_ATCC_35405
HWI-ST1242:77:C13H3ACXX:7:2202:20027:26498   Treponema_denticola_ATCC_35405               Treponema_vincentii_ATCC_35580_NZACYH00000000.1
HWI-ST1242:77:C13H3ACXX:7:2202:20027:26498   Treponema_vincentii_ATCC_35580_NZACYH00000000.1     Treponema_denticola_ATCC_35405
HWI-ST1242:77:C13H3ACXX:7:2211:7978:77793    Akkermansia_muciniphila_ATCC_BAA-835         Bordetella_bronchiseptica_strain_RB50
HWI-ST1242:77:C13H3ACXX:7:2211:7978:77793    Bordetella_bronchiseptica_strain_RB50        Akkermansia_muciniphila_ATCC_BAA-835
HWI-ST1242:77:C13H3ACXX:7:2201:12614:25584   Thermotoga_petrophila_RKU-1                  Thermotoga_sp._RQ2
HWI-ST1242:77:C13H3ACXX:7:2201:12614:25584   Thermotoga_sp._RQ2                           Thermotoga_petrophila_RKU-1
```

where 1st column is the read identifier (the read identifiers will repeat on subsequent rows and may have a multiplicity of 2 or more), 2nd and 3rd column give linkage information. For example, let's grep the records from the SAM file using one of the identifiers given above:

```
[uzi@quince-srv2 ~/METAGENOMEanalysis_v0.1/test]$ grep -i "HWI-ST1242:77:C13H3ACXX:7:1203:6513:69746" *.sam
HWI-ST1242:77:C13H3ACXX:7:1203:6513:69746 65    Treponema_vincentii_ATCC_35580_NZACYH00000000.1 1506973     41
      77M24S      Treponema_denticola_ATCC_35405      57994 0
      AAAATGTACATCCCTGTACATTTTTTACCTACGTGTTTTCAAGGCAAACACACTACTGATTATAACCACCCCCATCCCCCCCCCCAAAACATACATCCCTGT
      @CCFFFDDHHHHHJJJJJJJJJJJJJJJIJIJJJD11?0*00*0)0?G;D################################################## NM:i:4
      AS:i:57     XS:i:33
HWI-ST1242:77:C13H3ACXX:7:1203:6513:69746 129   Treponema_denticola_ATCC_35405      57994 30    39S56M6S
      Treponema_vincentii_ATCC_35580_NZACYH00000000.1 1506973     0
      GGCAAAATCAGGTAATCCTTTTGAAATATGCGGGGCATCTACTGCGTCGTTACGCAAAAGCGAATGCTCAACGTACAGCAAGTACGCCTCCGCTTCGCTTT
      BCCFFFFFHHHHFHHIIIJJJJJIJJJIJJJJJJJGIGIIGIGIEE;BDDAACA@?BBD>9=8@5?::>AC?588A243(934>C79@89A9309559588 NM:i:3
      AS:i:41     XS:i:27     XP:Z:Treponema_vincentii_ATCC_35580_NZACYH00000000.1,-1507188,61S40M,30,0;
HWI-ST1242:77:C13H3ACXX:7:1203:6513:69746 145   Treponema_vincentii_ATCC_35580_NZACYH00000000.1 1507188     30
      61S40M      =     1506973     -255
      AAAGCGAAGCGGAGGCGTACTTGCTGTACGTTGAGCATTCGCTTTTGCGTAACGACGCAGTAGATGCCCCGCATATTTCAAAAGGATTACCTGATTTTGCC
      8859559039A98@97C>439(342A885?CA>::?5@8=9>DBB?@ACAADDB;EEIGIGIIGIGJJJJJJJIJJJIJJJJJIIIHHFHHHHFFFFFCCB NM:i:0
      AS:i:40     XS:i:29     XP:Z:Treponema_denticola_ATCC_35405,+57994,39S56M6S,30,3;
```

Here, the read is paired-end read 1 if it is the member of the set (99,83,67,115,81,97,65,113) and is paired-end read 2 if it is the member of the set (147,163,131,179,161,145,129,177), respectively. So the first record returned from the grep statement is paired-end read 1 and next two records are paired-end read 2. You can just get rid of redundant column $7 is "=" record and run the query as:

```
awk '$2~/^99$|^147$|^83$|^163$|^67$|^131$|^115$|^179$|^81$|^161$|^97$|^145$|^65$|^129$|^113$|^177$/ && $2!~/^SN/ && $7!~/=/{print $1"\t"$3"\t"$7}' NexteraXT_even_1ng_HISEQ_AGGCAGAA-CTCTCTAT.sam | sort -n -k1,1 | uniq | awk 'BEGIN { FS="\t" } { c[$1]++; l[$1,c[$1]]=$0 } END { for (i in c) { if (c[i] > 1) for (j = 1; j <= c[i]; j++) print l[i,j] } }'
```

which ran in **7.7 seconds** for the above SAM file ;)

Some records are:

```
HWI-ST1242:77:C13H3ACXX:7:2207:10952:95139   Persephonella_marina_EX-H1            Thermoanaerobacter_pseudethanolicus_ATCC_33223
HWI-ST1242:77:C13H3ACXX:7:2207:10952:95139   Thermoanaerobacter_pseudethanolicus_ATCC_33223      Persephonella_marina_EX-H1
HWI-ST1242:77:C13H3ACXX:7:2216:3177:14124    Shewanella_baltica_OS185              Shewanella_baltica_OS223,
HWI-ST1242:77:C13H3ACXX:7:2216:3177:14124    Shewanella_baltica_OS223,             Shewanella_baltica_OS185
HWI-ST1242:77:C13H3ACXX:7:2205:11785:35383   Deinococcus_radiodurans_R1_chromosome_1,_complete_sequence  Leptothrix_cholodnii_SP-6
HWI-ST1242:77:C13H3ACXX:7:2205:11785:35383   Leptothrix_cholodnii_SP-6             Deinococcus_radiodurans_R1_chromosome_1,_complete_sequence
HWI-ST1242:77:C13H3ACXX:7:2104:19027:61363   Sulfurihydrogenibium_yellowstonense_SS-5     SulfuriYO3AOP1
HWI-ST1242:77:C13H3ACXX:7:2104:19027:61363   SulfuriYO3AOP1                        Sulfurihydrogenibium_yellowstonense_SS-5
HWI-ST1242:77:C13H3ACXX:7:1301:5017:76112    Sulfurihydrogenibium_yellowstonense_SS-5     SulfuriYO3AOP1
HWI-ST1242:77:C13H3ACXX:7:1301:5017:76112    SulfuriYO3AOP1                        Sulfurihydrogenibium_yellowstonense_SS-5
HWI-ST1242:77:C13H3ACXX:7:2316:4885:64809    Sulfurihydrogenibium_yellowstonense_SS-5     SulfuriYO3AOP1
HWI-ST1242:77:C13H3ACXX:7:2316:4885:64809    SulfuriYO3AOP1                        Sulfurihydrogenibium_yellowstonense_SS-5
HWI-ST1242:77:C13H3ACXX:7:2305:6004:6780     Acidobacterium_capsulatum_ATCC_51196  Bordetella_bronchiseptica_strain_RB50
HWI-ST1242:77:C13H3ACXX:7:2305:6004:6780     Bordetella_bronchiseptica_strain_RB50 Acidobacterium_capsulatum_ATCC_51196
HWI-ST1242:77:C13H3ACXX:7:2302:12911:73962   Gemmatimonas_aurantiaca_T-27_DNA      gi|83591340|ref|NC_007643.1|
HWI-ST1242:77:C13H3ACXX:7:2302:12911:73962   gi|83591340|ref|NC_007643.1|          Gemmatimonas_aurantiaca_T-27_DNA
HWI-ST1242:77:C13H3ACXX:7:1205:3889:3509     Thermotoga_neapolitana_DSM_4359       Thermotoga_sp._RQ2
HWI-ST1242:77:C13H3ACXX:7:1205:3889:3509     Thermotoga_sp._RQ2                    Thermotoga_neapolitana_DSM_4359
```

The purpose of the final awk statment is to remove singletons i.e. get rid of reads where both paired-ends match to the same genome/contig. Refer to table 1 on how I have used the flags.

FLAG	flags	pair	itself		mate		proper?	aligner?
One of the mate is unmapped								
73	1+8+64 73	1	map	+	unmap			
133	1+4+128 133	2	unmap		map	+		
89	1+8+16+64 89	1	map	+	unmap	-		
121	1+8+16+32+64	1	map	-	unmap	-		
165	1+4+32+128 165	2	unmap	+	map	-		ssaha
181	1+4+16+32+128	2	unmap	-	map	-		bwa
101	1+4+32+64 101	1	unmap	+	map	-		ssaha
117	1+4+16+32+64	1	unmap	-	map	+		bwa
153	1+8+16+128 153	2	map	-	unmap	+		
185	1+8+16+32+128	2	map	-	unmap	-		
69	1+4+64 69	1	unmap	+	map	+		
137	1+8+128 137	2	map	+	unmap	+		
Both unmapped								
77	1+4+8+64 77	1	unmap	+	unmap	+		
141	1+4+8+128 141	2	unmap	+	unmap	+		
mapped in correct orientation and within insert size								
99	1+2+32+64 99	1	map	+	map	-	y	
147	0+1+2+16+128 147	2	map	-	map	+	y	
83	1+2+16+64 83	1	map	-	map	+	y	
163	1+2+32+128 163	2	map	+	map	-	y	
mapped within the insert size but wrong orientation (++ or - -)								
67	1+2+64 67	1	map	+	map	+	y	
131	1+2+128 131	2	map	+	map	+	y	
115	1+2+16+32+64 115	1	map	-	map	-	y	
179	1+2+16+32+128 179	2	map	-	map	-	y	
mapped uniquely but wrong insert size, and could possibly reside in different contigs								
81	1+16+64 81	1	map	-	map	+		
161	1+32+128 161	2	map	+	map	-		
97	1+32+64 97	1	map	+	map	-		
145	1+16+128 145	2	map	-	map	+		
65	1+64 65	1	map	+	map	+		
129	1+128 129	2	map	+	map	+		
113	1+16+32+64 113	1	map	-	map	-		
177	1+16+32+128 177	2	map	-	map	-		

Table 1: SAM format flags (Ref: http://ppotato.files.wordpress.com/2010/08/sam_output.pdf)