SAMtools and BEDtools

Folder: gencommand_proj2_data

Module 2 Exam – Command Line Tools for Genomic DS

Module 2 Exam Instructions **IMPORTANT**

For this project, it is recommended that you use the VMBox virtual environment (***Docker instead for MacOS with M1 or M2 chips) provided with the Course package and the tools therein. You may also use your own system and software, however make sure that appropriate versions are installed. The answers are compatible with the following versions of the software: samtools v.1.2, bedtools v.2.24.0.

As part of a larger project cataloging genetic variation in the plant Arabidopsis thaliana, you sequenced and assembled the genome of one strain ('wu_0_A'), then mapped back the reads to the assembled genome. The resulting BAM file is included in the package 'gencommand_proj2_data.tar.gz'. Using SAMtools and BEDtools as well as other Unix commands introduced in this course, examine the files and answer the following questions. NOTE: Input data have been obtained and modified from those generated by the 1001 Genomes Project, accession 'Wu_0_A'.

Click here to download the Project 2 Data Files

Apply these rules and steps to the questions marked above each rule.

Questions 1 - 5:

For the original set of alignments (file 'athal_wu_0_A.bam'):

Questions 6 - 10:

Extract only the alignments in the range "Chr3:11,777,000-11,794,000", corresponding to a locus of interest. For this alignment set:

Questions 11 - 15:

Determine general information about the alignment process from the original BAM file.

Questions 16 - 20:

Using BEDtools, examine how many of the alignments at point 2 overlap exons at the locus of interest. Use the BEDtools '-wo' option to only report non-zero overlaps. The list of exons is given in the included 'athal_wu_0_A_annot.gtf' GTF file.

(base) Chaus-MacBook-Air:Downloads aaaz$ cd gencommand_proj2_data 3

```
(base) Chaus-MacBook-Air:gencommand_proj2_data aaaz$ ls
athal_wu_0_A.bam            athal_wu_0_A_3.bam
athal_wu_0_A.bam.bai        athal_wu_0_A_annot.gtf
athal_wu_0_A.sorted.bam         athal_wu_0_a.sorted.bam.bai
athal_wu_0_A_0.bam          data.bed
athal_wu_0_A_1.bam          nohup.out
athal_wu_0_A_2.bam
(base) Chaus-MacBook-Air:gencommand_proj2_data aaaz$ samtools view athal_wu_0_A.bam |
cut -f7 | grep "*" |wc -l
   65521
(base) Chaus-MacBook-Air:gencommand_proj2_data aaaz$ samtools view athal_wu_0_A.bam |
head
GAII05_0002:1:113:7822:3886#0  1187 Chr3 11699950  60     51M  11700332    433
        AAAAAAAATGTAAAACTGCTAAATCTCTCCTCTCTAAAGAACTCGTCCCCG
        CCCCCCBBBCCCCCCCCCCCCCCCCCCCCCCCCCCCCCBAAB??@ACBBCCCD
        PQ:i:21      SM:i:37      UQ:i:0       MQ:i:37XQ:i:0     RG:Z:H100223_GAII05_0002
GAII05_0002:1:40:13457:15230#0         163    Chr3 11699950  60     51M 11700332       433
        AAAAAAAATGTAAAACTGCTAAATCTCTCCTCTCTAAAGAACTCGTCCCCG
        CCCCCCBBBCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
        PQ:i:21      SM:i:37      UQ:i:0       MQ:i:37XQ:i:0     RG:Z:H100223_GAII05_0002
GAII05_0002:1:109:7632:9781#0  147    Chr3 11699952  60     51M  11699616    -387
        AAAAAATGTAAAACTGCTAAATCTCTCCNCTCTAAAGAACTCGTCCCCGTC
        CCCCCCACCCCCCCCCCCCCC3;:;7??&AACCCCCCCCCCCCCCCCCCC
        PQ:i:33      SM:i:37      UQ:i:0       MQ:i:37XQ:i:0     RG:Z:H100223_GAII05_0002
GAII05_0002:1:19:18679:10485#0         163    Chr3 11699953  60     51M 11700321       419
        AAAAATGTAAAACTGCTAAATCTCTCCTCTCTAAAGAACTCGTCCCCGTCT
        CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCDCCCCCBCB
        PQ:i:22      SM:i:37      UQ:i:0       MQ:i:37XQ:i:0     RG:Z:H100223_GAII05_0002
GAII05_0002:1:40:3052:5465#0   147    Chr3 11699968  60     51M  11699649    -370
        CTAAATCTCTCCTCTCTAAAGAACTCGTCCCCGTCTGCACGATACTCATGA
```

```
	C?BBA:C=CACB@CBCCCCCCC@CCCCCCCCCCCCCCCCCCCCCCCCCCC
	PQ:i:40	SM:i:37	UQ:i:0	MQ:i:37XQ:i:0	RG:Z:H100223_GAII05_0002
GAII05_0002:1:67:19144:17862#0	83	Chr3	11699969	60	51M	11699644	-376
	TAAATCTCTCCTCTCTAAAGAACTCGTCCCCGTCTGCACGATACTCATGAA
	######################ACCCBCCCCCDCCCCCCCCCCCCCCCCC	PQ:i:40
	SM:i:37	UQ:i:0	MQ:i:37XQ:i:0	RG:Z:H100223_GAII05_0002
unknown:2:30:1050:1300#0/1	16	Chr3	11699969	37	36M
	TAAATCTCTCCTCTCTAAAGAACTCGTCCCCGTCTG
	:?@<8>4=9BA>B@;:@@B@B5?3BB=BBABBB@5@	UQ:i:0	RG:Z:Wii_SR03
GAII02:4:13:207:1907#0	99	Chr3	11699970	60	36M	=	11700133	199
	AAATCTCTCCTCTCTAAAGAACTCGTCCCCGTCTGC
	BCAABBCBCBBBBBA:@<<>BBBB?;BB@@8>@>7>	PQ:i:37	SM:i:37	UQ:i:0
	MQ:i:37	XQ:i:0	RG:Z:Wii_PER02
unknown:1:72:1010:1212#0/1	1024	Chr3	11699970	37	36M
	AAATCTCTCCTCTCTAAAGAACTCGTCCCCGTCTGC
	@>>BB@@B<CCA??9@BBB@3AAB@>@;A7B3B?>B	UQ:i:0	RG:Z:Wii_SR03
GAII02:4:6:1286:1492#0	147	Chr3	11699971	60	36M	=	11699819	-188
	AACCTCTCCTCTCTAAAGAACTCGTCCCCGTCTGCA	A7'9/=3;8/63)5?BBAA6=-
AB?BBBB@AB@CBB	PQ:i:46	SM:i:37	UQ:i:6	MQ:i:37	XQ:i:0
	RG:Z:Wii_PER02
(base) Chaus-MacBook-Air:gencommand_proj2_data aaaz$ samtools view athal_wu_0_A.bam |
wc -l
  221372
(base) Chaus-MacBook-Air:gencommand_proj2_data aaaz$ samtools view athal_wu_0_A.bam |
cut -f6 | grep "D" | wc -l
    2451
(base) Chaus-MacBook-Air:gencommand_proj2_data aaaz$ samtools view athal_wu_0_A.bam |
cut -f7 | grep "=" | wc -l
  150913
```

```
(base) Chaus-MacBook-Air:gencommand_proj2_data aaaz$ samtools view athal_wu_0_A.bam |
cut -f6 | grep "N" | wc -l
      0
(base) Chaus-MacBook-Air:gencommand_proj2_data aaaz$ ls
athal_wu_0_A.bam              athal_wu_0_A_3.bam
athal_wu_0_A.bam.bai          athal_wu_0_A_annot.gtf
athal_wu_0_A.sorted.bam               athal_wu_0_a.sorted.bam.bai
athal_wu_0_A_0.bam            data.bed
athal_wu_0_A_1.bam            nohup.out
athal_wu_0_A_2.bam
(base) Chaus-MacBook-Air:gencommand_proj2_data aaaz$ nohup samtools sort
athal_wu_0_A.bam athal_wu_0_A.sorted&
[1] 11716
(base) Chaus-MacBook-Air:gencommand_proj2_data aaaz$ appending output to nohup.out
samtools index athal_wu_0_A.sorted.bam
samtools index: use -M to enable indexing more than one alignment file
[1]+  Exit 1                  nohup samtools sort athal_wu_0_A.bam athal_wu_0_A.sorted
(base) Chaus-MacBook-Air:gencommand_proj2_data aaaz$ samtools view
athal_wu_0_A.sorted.bam
(base) Chaus-MacBook-Air:gencommand_proj2_data aaaz$ samtools view
athal_wu_0_A.sorted.bam "Chr3: 11777000-11794000" |head
unknown:3:76:339:1714#0/1 0       Chr3  11777470  9       36M
        CTTAATTTGAAATTGTTAGTCTATTGTGTGATATGA
        BAABB:BB@88@@<?>B@>:BAA@8>>@06B?=2?>      UQ:i:69     RG:Z:Wii_SR03
GAII05_0002:1:61:4766:5403#0    101   Chr3  11777578  0       *       11777578    0
        AATACCGGTTCCACCACTCTTACGGAAATAGCAGCGCAAATGCTTTGTCTG
        CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC@@CCBCACBBCC
        MQ:i:1        XQ:i:301    RG:Z:H100223_GAII05_0002
GAII05_0002:1:61:4766:5403#0    153   Chr3  11777578  1       9M8I20M7I7M     =
        11777578  0
```

GTTCCATATCTGATTAACAATATACTCATTGATAAAATGGTAATTTTTAAA

CC>CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC

UQ:i:301 RG:Z:H100223_GAII05_0002

GAII02:4:84:1639:1875#0 99 Chr3 11778207 25 2I34M = 11778371 200

ACATTTTTGGGCAACTTGTGACCGGAATTTGTTATG

@BCB<A>)A74BABBBB8A6>BA?@;@@B@09C:<# PQ:i:235 SM:i:25

UQ:i:142 MQ:i:25 XQ:i:54 RG:Z:Wii_PER02

GAII02:4:84:483:947#0 99 Chr3 11778235 37 36M = 11778361 162

GTTATGACATAGTTTGATCATTCTTCTCCATATTAG

AAB@@A@A>B>A>>B>>AB?@BAB?/=>?8?8BB## PQ:i:225 SM:i:37

UQ:i:185 MQ:i:37 XQ:i:0 RG:Z:Wii_PER02

unknown:2:81:1339:779#0/1 16 Chr3 11778288 4 36M

ATATTCAGTTTTGGTAGAAGAGTTTCGTAAGCTTTA

=B@CBB:/<>BB?7?A=<>AAB>>BBABA@A;>>AB UQ:i:110 RG:Z:Wii_SR03

GAII01:4:68:1281:1769#0 69 Chr3 11778289 0 * = 11778289 0

CTTATAACACTGAAGTTGGAATTGATCAAAATCTGA

5=7@:A?+<4:BBBBA@BABCABBBA6@BBB@BBBB MQ:i:21 XQ:i:100

RG:Z:Wii_PER01

GAII01:4:68:1281:1769#0 137 Chr3 11778289 21 36M = 11778289 0

TATTCAGTTTTGGTAGAAGAGTTTCGTAAGCTTTAT

BCCCBB@BCCC@A?A7AA@:>BCCBAAA87>CCB8B UQ:i:100 RG:Z:Wii_PER01

GAII05_0002:1:47:10861:15828#0 145 Chr3 11778290 38 1M5I45Chr4

4508019 0

GTTATATTCAGTTTTGGTAGAAGAGTTTCGTAAGCTTTATCTTTTTCATTT

CCCCCCCCCCCCACCCCCCCCCCCCCCCCCCCCCCCCCBBCCCCCCCCCCC

PQ:i:473 SM:i:3UQ:i:184 MQ:i:45 XQ:i:204 RG:Z:H100223_GAII05_0002

GAII05_0002:1:84:7626:15880#0 81 Chr3 11778290 45 51M Chr44508017 0

ATTCAGTTTTGGTAGAAGAGTTTCGTAAGCTTTATCTTTTTCATTTTTCAA

?AAAAAABA?AAAAAAAAAAAAAAAAAAAAA77AAAAAAAAAAA::6:3 PQ:i:391

SM:i:8UQ:i:126 MQ:i:45 XQ:i:170 RG:Z:H100223_GAII05_0002

```
(base) Chaus-MacBook-Air:gencommand_proj2_data aaaz$ samtools view
athal_wu_0_A.sorted.bam "Chr3:11777000-11794000" | wc -l
   7081
(base) Chaus-MacBook-Air:gencommand_proj2_data aaaz$ samtools view
athal_wu_0_A.sorted.bam "Chr3:11777000-11794000" |cut -f7 |grep "*" | wc -l
   1983
(base) Chaus-MacBook-Air:gencommand_proj2_data aaaz$ samtools view
athal_wu_0_A.sorted.bam "Chr3:11777000-11794000" |cut -f6 | grep "D" | wc -l
     31
(base) Chaus-MacBook-Air:gencommand_proj2_data aaaz$ samtools view
athal_wu_0_A.sorted.bam "Chr3:11777000-11794000" |cut -f7 | grep "=" | wc -l
   4670
(base) Chaus-MacBook-Air:gencommand_proj2_data aaaz$ samtools view
athal_wu_0_A.sorted.bam "Chr3:11777000-11794000" |cut -f6 | grep "N" | wc -l
      0
(base) Chaus-MacBook-Air:gencommand_proj2_data aaaz$ samtools view -H
athal_wu_0_A.bam
@HDVN:1.0       GO:none    SO:coordinate
@SQSN:Chr1    LN:29923332      AS:wu_0.v7.fas   SP:wu_0.v7.fas
@SQSN:Chr2    LN:19386101      AS:wu_0.v7.fas   SP:wu_0.v7.fas
@SQSN:Chr3    LN:23042017      AS:wu_0.v7.fas   SP:wu_0.v7.fas
@SQSN:Chr4    LN:18307997      AS:wu_0.v7.fas   SP:wu_0.v7.fas
@SQSN:Chr5    LN:26567293      AS:wu_0.v7.fas   SP:wu_0.v7.fas
@SQSN:chloroplast    LN:154478 AS:wu_0.v7.fas   SP:wu_0.v7.fas
@SQSN:mitochondria LN:366924 AS:wu_0.v7.fas   SP:wu_0.v7.fas
@RGID:H100223_GAII05_0002    PL:SLX      LB:wu_PII  PI:400
        DS:wu_0_GenomeSM:wu_0
@RGID:Wii_PER01     PL:SLX      LB:wu_phaseI     PI:400       DS:wu_0_Genome
        SM:wu_0
```

@RGID:Wii_PER02 PL:SLX LB:wu_phaseI PI:400 DS:wu_0_Genome SM:wu_0

@RGID:Wii_SR03 PL:SLX LB:wu_phaseI PI:400 DS:wu_0_Genome SM:wu_0

@PGID:stampy VN:1.0.3_(r627) CL:-g /lustre/scratch103/sanger/xcg/tmp/tmp.zYfXz26246 -h /lustre/scratch103/sanger/xcg/tmp/tmp.zYfXz26246 --readgroup=ID:Wii_PER01,LB:wu_phaseI,SM:wu_0,PI:400,PL:SLX,DS:wu_0_Genome --bwaoptions=-q10 /lustre/scratch103/sanger/xcg/wu_0.v7.fas -o /lustre/scratch103/sanger/xcg/wu_0/A/aln_A1.sp0.sam -M /lustre/scratch103/sanger/xcg/wu_0/read_1_1.sp0.fq.gz /lustre/scratch103/sanger/xcg/wu_0/read_1_2.sp0.fq.gz

@PGID:samtools PN:samtools PP:stampy VN:1.17 CL:samtools view -H athal_wu_0_A.bam

@COTM:Fri, 17 Sep 2010 12:20:13 BST WD:/lustre/scratch103/sanger/xcg/wu_0/self HN:bc-16-1-07.internal.sanger.ac.uk UN:xcg

(base) Chaus-MacBook-Air:gencommand_proj2_data aaaz$ samtools view athal_wu_0_A.bam | head

GAII05_0002:1:113:7822:3886#0 1187 Chr3 11699950 60 51M 11700332 433 AAAAAAAATGTAAAACTGCTAAATCTCTCCTCTCTAAAGAACTCGTCCCCG CCCCCCBBBCCCCCCCCCCCCCCCCCCCCCCCCCCCCBAAB??@ACBBCCCD PQ:i:21 SM:i:37 UQ:i:0 MQ:i:3XQ:i:0 RG:Z:H100223_GAII05_0002

GAII05_0002:1:40:13457:15230#0 163 Chr3 11699950 60 51M 11700332 433 AAAAAAAATGTAAAACTGCTAAATCTCTCCTCTCTAAAGAACTCGTCCCCG CCCCCCBBBCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC PQ:i:21 SM:i:37 UQ:i:0 MQ:i:3XQ:i:0 RG:Z:H100223_GAII05_0002

GAII05_0002:1:109:7632:9781#0 147 Chr3 11699952 60 51M 11699616 -387 AAAAAATGTAAAACTGCTAAATCTCTCCNCTCTAAAGAACTCGTCCCCGTC CCCCCCACCCCCCCCCCCCCCC3;:;7??&AACCCCCCCCCCCCCCCCCCC PQ:i:33 SM:i:37 UQ:i:0 MQ:i:3XQ:i:0 RG:Z:H100223_GAII05_0002

```
GAII05_0002:1:19:18679:10485#0  163  Chr3  11699953  60  51M  11700321  419
    AAAAATGTAAAACTGCTAAATCTCTCCTCTCTAAAGAACTCGTCCCCGTCT
    CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCDCCCCCBCB
    PQ:i:22  SM:i:37  UQ:i:0  MQ:i:3XQ:i:0  RG:Z:H100223_GAII05_0002
GAII05_0002:1:40:3052:5465#0  147  Chr3  11699968  60  51M  11699649  -370
    CTAAATCTCTCCTCTCTAAAGAACTCGTCCCCGTCTGCACGATACTCATGA
    C?BBA:C=CACB@CBCCCCCCC@CCCCCCCCCCCCCCCCCCCCCCCCCCC
    PQ:i:40  SM:i:37  UQ:i:0  MQ:i:3XQ:i:0  RG:Z:H100223_GAII05_0002
GAII05_0002:1:67:19144:17862#0  83  Chr3  11699969  60  51M  11699644  -376
    TAAATCTCTCCTCTCTAAAGAACTCGTCCCCGTCTGCACGATACTCATGAA
    ######################ACCCBCCCCCDCCCCCCCCCCCCCCCCC  PQ:i:40
    SM:i:37  UQ:i:0  MQ:i:3XQ:i:0  RG:Z:H100223_GAII05_0002
unknown:2:30:1050:1300#0/1  16  Chr3  11699969  37  36M
    TAAATCTCTCCTCTCTAAAGAACTCGTCCCCGTCTG
    :?@<8>4=9BA>B@;:@@B@B5?3BB=BBABBB@5@  UQ:i:0  RG:Z:Wii_SR03
GAII02:4:13:207:1907#0  99  Chr3  11699970  60  36M  =  11700133  199
    AAATCTCTCCTCTCTAAAGAACTCGTCCCCGTCTGC
    BCAABBCBCBBBBBA:@<<>BBBB?;BB@@8>@>7>  PQ:i:37  SM:i:37  UQ:i:0
    MQ:i:37  XQ:i:0  RG:Z:Wii_PER02
unknown:1:72:1010:1212#0/1  1024  Chr3  11699970  37  36M
    AAATCTCTCCTCTCTAAAGAACTCGTCCCCGTCTGC
    @>>BB@@B<CCA??9@BBB@3AAB@>@;A7B3B?>B  UQ:i:0  RG:Z:Wii_SR03
GAII02:4:6:1286:1492#0  147  Chr3  11699971  60  36M  =  11699819  -188
    AACCTCTCCTCTCTAAAGAACTCGTCCCCGTCTGCA  A7'9/=3;8/63)5?BBAA6=-
AB?BBBB@AB@CBB PQ:i:46  SM:i:37  UQ:i:6  MQ:i:37  XQ:i:0
    RG:Z:Wii_PER02
(base) Chaus-MacBook-Air:gencommand_proj2_data aaaz$ ls
athal_wu_0_A.bam            athal_wu_0_A_3.bam
athal_wu_0_A.bam.bai        athal_wu_0_A_annot.gtf
athal_wu_0_A.sorted.bam     athal_wu_0_a.sorted.bam.bai
```

```
athal_wu_0_A_0.bam          data.bed
athal_wu_0_A_1.bam          nohup.out
athal_wu_0_A_2.bam
(base) Chaus-MacBook-Air:gencommand_proj2_data aaaz$ bedtools bamtobed -i
athal_wu_0_A.bam > data.bed
-bash: bedtools: command not found
(base) Chaus-MacBook-Air:gencommand_proj2_data aaaz$ bedtools2 bamtobed -i
athal_wu_0_A.bam > data.bed
-bash: bedtools2: command not found
(base) Chaus-MacBook-Air:gencommand_proj2_data aaaz$ cd
(base) Chaus-MacBook-Air:~ aaaz$ conda install -c bioconda bedtools
(base) Chaus-MacBook-Air:~ aaaz$ cd /Users/aaaz/Downloads/gencommand_proj2_data
(base) Chaus-MacBook-Air:gencommand_proj2_data aaaz$ ls
athal_wu_0_A.bam            athal_wu_0_A_3.bam
athal_wu_0_A.bam.bai        athal_wu_0_A_annot.gtf
athal_wu_0_A.sorted.bam             athal_wu_0_a.sorted.bam.bai
athal_wu_0_A_0.bam          data.bed
athal_wu_0_A_1.bam          nohup.out
athal_wu_0_A_2.bam
(base) Chaus-MacBook-Air:gencommand_proj2_data aaaz$ bedtools bamtobed -i
athal_wu_0_A.bam > data.bed
(base) Chaus-MacBook-Air:gencommand_proj2_data aaaz$ head data.bed
Chr3 11699949  11700000  GAII05_0002:1:113:7822:3886#0/2       60   +
Chr3 11699949  11700000  GAII05_0002:1:40:13457:15230#0/2      60   +
Chr3 11699951  11700002  GAII05_0002:1:109:7632:9781#0/2       60   -
Chr3 11699952  11700003  GAII05_0002:1:19:18679:10485#0/2      60   +
Chr3 11699967  11700018  GAII05_0002:1:40:3052:5465#0/2 60     -
Chr3 11699968  11700019  GAII05_0002:1:67:19144:17862#0/1      60   -
Chr3 11699968  11700004  unknown:2:30:1050:1300#0/1       37   -
Chr3 11699969  11700005  GAII02:4:13:207:1907#0/1    60   +
```

```
Chr3 11699969  11700005  unknown:1:72:1010:1212#0/1      37    +
Chr3 11699970  11700006  GAII02:4:6:1286:1492#0/2  60    -
(base) Chaus-MacBook-Air:gencommand_proj2_data aaaz$ ls
athal_wu_0_A.bam              athal_wu_0_A_3.bam
athal_wu_0_A.bam.bai          athal_wu_0_A_annot.gtf
athal_wu_0_A.sorted.bam             athal_wu_0_a.sorted.bam.bai
athal_wu_0_A_0.bam            count_base.txt
athal_wu_0_A_1.bam            data.bed
athal_wu_0_A_2.bam            nohup.out
(base) Chaus-MacBook-Air:gencommand_proj2_data aaaz$ bedtools intersect -wo -a
athal_wu_0_A_annot.gtf -b data.bed | head
Chr3 tair10_Ws_0      exon 11780787  11780828  .      +      transcript_id
"Transcript:AT3G30260.1"; gene_id "Gene:AT3G30260"; Chr3 11780744 11780795
      GAII05_0002:1:35:13043:5507#0/2      77    +      8
Chr3 tair10_Ws_0      exon 11780787  11780828  .      +      transcript_id
"Transcript:AT3G30260.1"; gene_id "Gene:AT3G30260"; Chr3 11780746 11780797
      GAII05_0002:1:110:10000:12812#0/2    31    -      10
Chr3 tair10_Ws_0      exon 11780787  11780828  .      +      transcript_id
"Transcript:AT3G30260.1"; gene_id "Gene:AT3G30260"; Chr3 11780747 11780798
      GAII05_0002:1:100:16792:20331#0/2    14    -      11
Chr3 tair10_Ws_0      exon 11780787  11780828  .      +      transcript_id
"Transcript:AT3G30260.1"; gene_id "Gene:AT3G30260"; Chr3 11780747 11780798
      GAII05_0002:1:100:16806:20347#0/2    14    -      11
Chr3 tair10_Ws_0      exon 11780787  11780828  .      +      transcript_id
"Transcript:AT3G30260.1"; gene_id "Gene:AT3G30260"; Chr3 11780748 11780799
      GAII05_0002:1:79:17089:10562#0/1     66    +      12
Chr3 tair10_Ws_0      exon 11780787  11780828  .      +      transcript_id
"Transcript:AT3G30260.1"; gene_id "Gene:AT3G30260"; Chr3 11780749 11780800
      GAII05_0002:1:14:6378:3379#0/2 49    +      13
```

```
Chr3 tair10_Ws_0 exon 11780787 11780828 . + transcript_id
"Transcript:AT3G30260.1"; gene_id "Gene:AT3G30260"; Chr3 11780750 11780789
GAII05_0002:1:76:1002:13555#0/2 36 + 2
Chr3 tair10_Ws_0 exon 11780787 11780828 . + transcript_id
"Transcript:AT3G30260.1"; gene_id "Gene:AT3G30260"; Chr3 11780751 11780802
GAII05_0002:1:59:8410:5306#0/2 6 + 15
Chr3 tair10_Ws_0 exon 11780787 11780828 . + transcript_id
"Transcript:AT3G30260.1"; gene_id "Gene:AT3G30260"; Chr3 11780755 11780806
GAII05_0002:1:7:15754:3597#0/1 96 + 19
Chr3 tair10_Ws_0 exon 11780787 11780828 . + transcript_id
"Transcript:AT3G30260.1"; gene_id "Gene:AT3G30260"; Chr3 11780756 11780792
GAII01:4:70:494:263#0/2 4 - 5
(base) Chaus-MacBook-Air:gencommand_proj2_data aaaz$ bedtools intersect -wo -a
athal_wu_0_A_annot.gtf -b data.bed | wc -l
   3101
(base) Chaus-MacBook-Air:gencommand_proj2_data aaaz$ bedtools intersect -wo -a
athal_wu_0_A_annot.gtf -b data.bed | cut -f16 > count_base.txt
(base) Chaus-MacBook-Air:gencommand_proj2_data aaaz$ bedtools intersect -wo -a
athal_wu_0_A_annot.gtf -b data.bed | wc -l
   3101
(base) Chaus-MacBook-Air:gencommand_proj2_data aaaz$ bedtools intersect -wo -a
athal_wu_0_A_annot.gtf -b data.bed | cut -f4 | sort -u | wc -l
     21
(base) Chaus-MacBook-Air:gencommand_proj2_data aaaz$ bedtools intersect -wo -a
athal_wu_0_A_annot.gtf -b data.bed | cut -f5 | sort -u | wc -l
     21
(base) Chaus-MacBook-Air:gencommand_proj2_data aaaz$ bedtools intersect -wo -a
athal_wu_0_A_annot.gtf -b data.bed | cut -f9 | cut -d "" -f4 | head
cut: bad delimiter
```

```
(base) Chaus-MacBook-Air:gencommand_proj2_data aaaz$ bedtools intersect -wo -a
athal_wu_0_A_annot.gtf -b data.bed | cut-f9 | cut-d "" -f4 | head
-bash: cut-d: command not found
-bash: cut-f9: command not found
(base) Chaus-MacBook-Air:gencommand_proj2_data aaaz$ bedtools intersect -wo -a
athal_wu_0_A_annot.gtf -b data.bed | cut -f9 | cut -d"" -f4| head
cut: bad delimiter
(base) Chaus-MacBook-Air:gencommand_proj2_data aaaz$ bedtools intersect -wo -a
athal_wu_0_A_annot.gtf -b data.bed | cut -f9 | cut -d "" -f4 | sort -u | wc -l
cut: bad delimiter
       0
(base) Chaus-MacBook-Air:gencommand_proj2_data aaaz$ bedtools intersect -wo -a
athal_wu_0_A_annot.gtf -b data.bed | cut -f9 | cut -d" " -f4 | head
"Gene:AT3G30260";
"Gene:AT3G30260";
"Gene:AT3G30260";
"Gene:AT3G30260";
"Gene:AT3G30260";
"Gene:AT3G30260";
"Gene:AT3G30260";
"Gene:AT3G30260";
"Gene:AT3G30260";
"Gene:AT3G30260";
(base) Chaus-MacBook-Air:gencommand_proj2_data aaaz$ bedtools intersect -wo -a
athal_wu_0_A_annot.gtf -b data.bed | cut -f9 | cut -d" " -f4 | sort -u | wc -l
       4
(base) Chaus-MacBook-Air:gencommand_proj2_data aaaz$
```

1. How many alignments does the set contain?

221372

First, a couple of introductory comments. A BAM file contains alignments for a set of input reads. Each read can have 0 (none), 1 or multiple alignments on the genome. These questions explore the relationships between reads and alignments.

The number of alignments is the number of entries, excluding the header, contained in the BAM file, or equivalently in its SAM conversion. To find the number of alignments, we can apply ('%' denotes the terminal prompt):

```
% samtools flagstat athal_wu_0_A.bam
```

which will list the number of alignments on line 1. An alternate method would be to count the number of lines in the converted SAM file (header excluded):

```
% samtools view athal_wu_0_A.bam | wc -l
```

Note that, if the file was created with a tool that includes unmapped reads into the BAM file, we would need to exclude the lines representing unmapped reads, i.e. with a '*' in column 3 (chrom):

```
% samtools view athal_wu_0_A.bam | cut –f3 | grep –v '*' | wc -l
```

2. How many alignments show the read's mate unmapped?

65521

An alignment with an unmapped mate is marked with a '*' in column 7. Note that the question asks for alignments, not reads, so we simply count the number of lines in the SAM file with a '*' in column 7:

```
% samtools view athal_wu_0_A.bam | cut –f7 | grep –c '*'
```

3. How many alignments contain a deletion (D)?

2451

Deletions are be marked with the letter 'D' in the CIGAR string for the alignment, shown in column 6:

```
% samtools view athal_wu_0_A.bam | cut –f6 | grep –c 'D'
```

4. How many alignments show the read's mate mapped to the same chromosome?

150913

Alignments with the read's mate mapped to the same chromosome are marked with a '=' in column 7:

```
% samtools view athal_wu_0_A.bam | cut –f7 | grep –c '='
```

5. How many alignments are spliced?

0

A spliced alignment will be marked with an 'N' (intron gap) in the CIGAR field:

```
% samtools view athal_wu_0_A.bam | cut –f6 | grep –c 'N'
```

6. How many alignments does the set contain?

7081

We first need to construct the reduced set, i.e. to extract from the original set only those alignments in the specified region. For this, we need to sort and index the file:

```
% samtools sort athal_wu_0_A.bam athal_wu_0_A.sorted
```

This will create the file 'athal_wu_0_A.sorted.bam'. We then index this file:

```
% samtools index athal_wu_0_A.sorted.bam
```

This will create the index file 'athal_wu_0_A.sorted.bam.bai' in the current directory. Lastly, we extract alignments in the specified range:

```
% samtools view –b athal_wu_0_A.sorted.bam "Chr3:11777000-11794000" > athal_wu_0_A.region.bam
```

The option '-b' will generate output in BAM format. The resulting BAM file will be sorted, so it can be indexed directly if needed.Common pitfalls: make sure to specify the correct reference sequence ('Chr3', not 'chr3') and exclude ',' when representing the query coordinates. Also, make sure to use the sorted and index BAM file. To determine the number of alignments in the new (region) file, we can use the same commands as for Q1, e.g.:

```
% samtools flagstat athal_wu_0_A.region.bam
```

7. How many alignments show the read's mate unmapped?

1983

% samtools view athal_wu_0_A.region.bam | cut –f7 | grep –c '*'

8. How many alignments contain a deletion (D)?

31

% samtools view athal_wu_0_A.region.bam | cut –f6 | grep –c 'D'

9. How many alignments show the read's mate mapped to the same chromosome?

0

Incorrect: 150913

% samtools view athal_wu_0_A.bam | cut –f7 | grep –c '='

10. How many alignments are spliced?

0

% samtools view athal_wu_0_A.bam | cut –f6 | grep –c 'N'

11. How many sequences are in the genome file?

7

This information can be found in the header of the BAM file. Starting with the original BAM file, we use samtools to display the header information and count the number of lines describing the sequences in the reference genome:

% samtools view –H athal_wu_0_A.bam | grep –c "SN:"

12. What is the length of the first sequence in the genome file?

29923332

The length information is stored alongside the sequence identifier in the header (pattern 'LN:seq_length'):

```
% samtools view –H athal_wu_0_A.bam | grep "SN:" | more
```

13. What alignment tool was used?

stampy

The program name is listed in the '@PG' line in the BAM header (pattern 'ID:program_name'):

```
% samtools view –H athal_wu_0_A.bam | grep "^@PG"
```

The '^' sign in the search pattern tells the grep function to match the pattern '@PG' at the start of the line.

14. What is the read identifier (name) for the first alignment?

GAII05_0002:1:113:7822:3886#0

This information is shown in column 1 of the first alignment record in the SAM file:

```
% samtools view athal_wu_0_A.bam | head -1 | cut –f1
```

15. What is the start position of this read's mate on the genome? Give this as 'chrom:pos' if the read was mapped, or '*" if unmapped.

No answer

⊗ **Incorrect**

16. How many overlaps (each overlap is reported on one line) are reported?

3101

We start by running BEDtools on the alignment set restricted to the specified region (Chr3:11777000-11794000) and the GTF annotation file listed above. To allow the input to be

read directly from the BAM file, we use the option '-abam'; in this case we will need to also specify '-bed' for the BAM alignment information to be shown in BED column format in the output:

```
% bedtools intersect –abam athal_wu_0_A.region.bam –b athal_wu_0_A_annot.gtf –bed -wo > overlaps.bed
```

This will create a file with the following format: Columns 1-12 : alignment information, converted to BED format Columns 13-21 : annotation (exon) information, from the GTF file Column 22 : length of the overlapAlternatively, we could first convert the BAM file to BED format using 'bedtools bamtobed' then use the resulting file in the 'bedtools intersect' command. To answer the question, the number of overlaps reported is precisely the number of lines in the file (because only entries in the first file that have overlaps in file B are reported, according to the option '-wo'):

```
% wc –l overlaps.bed
```

17. How many of these are 10 bases or longer?

2899

The size of the overlap is listed in column 22 of the 'overlaps.bed' file. To determine those longer than 10 bases, we extract the column, sort numerically in decreasing order, and simply determine by visual inspection of the file the number of such records. For instance, in 'vim' we search for the first line listing '9' (':/9'), then determine its line number (Ctrl+g). Alternatively, one can use grep with option '-n' to list the lines and corresponding line numbers:

```
% cut –f22 overlaps.bed | sort –nrk1 > lengths
```

Or

```
% cut –f22 overlaps.bed | sort –nrk1 | grep –n "^9" | head -1
```

For the latter, the last "10" line will be immediately above the first "9", so subtract 1 from the answer.

18. How many alignments overlap the annotations?

3101

Columns 1-12 define the alignments:

```
% cut –f1-12 overlaps.bed | sort –u | wc -l
```

Potential pitfalls: Multiple reads may map at the same coordinates, so the information in columns 1-3 is insufficient. The minimum information needed to define the alignments is contained in columns 1-5, which include the read ID and the flag, specifying whether this is read 1 or read 2 in a pair with the same read ID).

19. Conversely, how many exons have reads mapped to them?

```
21
```

Columns 13-21 define the exons:

```
% cut –f13-21 overlaps.bed | sort –u | wc -l
```

20. If you were to convert the transcript annotations in the file "athal_wu_0_A_annot.gtf" into BED format, how many BED records would be generated?

```
4
```

This question simply asks for the number of transcripts in the annotation file, since the BED format would represent each transcript on one line. This information can be obtained from column 9 in the GTF file:

```
% cut –f9 athal_wu_0_A.annot.gtf | cut –d ' ' –f1,2 | sort –u | wc -l
```