

## **Project Two**

### **Part One**

Write a program `countReads.py` that accepts a “SAM” format file containing short read alignments and a “BED” format file of gene models and then reports the number of alignments that overlap each gene model.

To get started, use `svn mv` to rename your `project1` directory – change the name to “`projects.`” (You’ll re-use the programs you’ve written next semester in many assignments, and so the name “`project1`” no longer makes sense!)

### **The spec:**

Your program should accept two arguments: the name of a BED file containing gene models (specified by `-b`) and the name of a SAM format file containing short read alignments (specified by `-a`, for alignments). If the SAM format file is not given, then the program should obtain the SAM format alignments data from `stdin`.

It should also accept an optional third argument that specifies the name of a file to write. If not given, it should print to `stdout`.

For example, you should be able to run the program as follows:

```
$ countReads.py -b TAIR9.bed -a sample.sam > output.txt
$ countReads.py -a sample.sam -b TAIR9.bed output.txt
$ cat sample.sam | countReads.py -b TAIR9.bed output.txt
```

Methods your program should support:

`getStartEnd`, accepts a line of SAM format text, and returns a tuple containing two ints representing the start and end position of the alignment relative to the genomic sequence in *interbase coordinates*. Note that to implement this method, you’ll need to decode the SAM CIGAR string.

`reportReads`, accepts these named arguments:

1. `bed_fh` – an opened (for reading) BED format file
2. `sam_fh` – an opened (for reading) SAM format file
3. `out_fh` – an opened (for writing) file

Your `reportReads` method should count the number of reads overlapping individual gene models in the BED file. It should report the counts by printing the original BED format line to `out_fh` and by inserting the number of overlapping reads into the Bed line’s “score” field.

### **Part Three:**

For this part of the project, you'll work with this BAM format file from an RNA-Seq experiment investigating the effects of water deprivation stress on gene expression and splicing patterns in Arabidopsis:

<http://teaching.transvar.org/data/DryBT2.sm.bam>

You can use `samtools view` to convert the BAM file to SAM format. You can download it or access on-line.

For example, the following command prints just the SAM header:

```
$ samtools view -H http://teaching.transvar.org/data/DryBT2.sm.bam
[knet_seek] SEEK_END is not supported for HTTP. Offset is unchanged.
@HD   VN:1.0       SO:sorted
@SQ   SN:chr1     LN:30427671
@SQ   SN:chr2     LN:19698289
@SQ   SN:chr3     LN:23459830
@SQ   SN:chr4     LN:18585056
@SQ   SN:chr5     LN:26975502
@SQ   SN:chrC    LN:154478
@SQ   SN:chrM    LN:366924
@PG   ID:TopHat  VN:1.1.4     CL:/common/lorainelab/sw/tophat-
1.1.4.Linux_x86_64/tophat --max-multihits 1 --solexa1.3-quals -I 1200 -
p 8 -F 0 -o tophat-1.1.4-heat-drought/DryBT2 -G
/storage/lorainelab/fastapub/tair/TAIR9_GFF3_genes-fixed.gff
/storage/lorainelab/bowtieindex/a_thaliana
/storage/lorainelab/illumina/chapel_hill_6/SE_101007_UNC4-
RDR3001561_00034_FC_706CJAAXX/s_2_sequence.txt
```

or:

```
$ samtools view -c http://teaching.transvar.org/data/DryBT2.sm.bam
chr1:2,257,286-2,260,303
[knet_seek] SEEK_END is not supported for HTTP. Offset is unchanged.
[bam_index_load] attempting to download the remote index file.
4147
```

Download [http://www.bioviz.org/quickload/A\\_thaliana\\_Jun\\_2009/TAIR9.bed.gz](http://www.bioviz.org/quickload/A_thaliana_Jun_2009/TAIR9.bed.gz).

Use your `countReads.py` program to create a new "BED" format file identical to `TAIR9.bed` but which contains the number of overlapping reads in the "score" field.

Feel free to use `binf_prog/class/test/testProject2.py` to test and debug your code.

#### Part Four: Optional

Next semester (6112) you'll add this option, but if you like, you can implement it now and get a head start.

Support and take advantage of the SO (“Sort Order”) flag in the SAM file header to speed up your program.

**References:**

For information about the SAM format and samtools, see:  
<http://samtools.sourceforge.net/>