# Hands-On Exercises #1

1. **How many binding sites are there for transcription factor SREBP1?**

awk '$4=="SREBP1"' tfbs.bed | wc -l

7 binding sites for SREBP1

1. **On how many different chromosomes is there an NFKB binding site with score greater than 300?**

awk '$4=="NFKB" && $5>300{print $1}' tfbs.bed   
| sort -u | wc -l

13 different chromosomes

1. **What is the longest binding site in this file? The length of each binding site is the distance between its start and stop coordinates.**

awk '{print $3-$2,$0}' tfbs.bed | sort -nr | head -1

The binding site for STAT1 on chr10 is 4432 nt long

1. **What are the names of all the transcription factors with binding sites overlapping the 1Mb interval chr3:57,000,000-58,000,000?**

awk '$1=="chr3" && $3>57000000 && $2<58000000{print $4}'  
tfbs.bed | sort -u

BAF170

CTCF

Ini1

JunD

STAT1

# Hands-On Exercises #2

1. **How many columns does the file have? Can we use whitespace as the delimiter?**

awk '{print NF}' refFlat.txt | uniq

All rows have 11 columns, even with the default whitespace delimiter. So yes, whitespace can be used as a delimiter.

1. **Which columns have the start and end position of each transcript?**

head -1 refFlat.txt | tr "\t" "\n" | awk '{print NR, $0}'

5 txStart

6 txEnd

1. **By convention, non-coding transcripts have cdsStart equal to cdsEnd. How many non‑coding transcripts are there?**

awk '$7==$8' refFlat.txt | wc -l

5567

1. **How many coding genes have transcripts on more than one chromosome? Note that the genes are named by the gene symbol in column 1 of refFlat.txt, while the transcripts are named by the accession in column 2.**

awk '$7!=$8{print $1,$3}' refFlat.txt | sort -u   
| awk '{print $1}' | uniq -c | awk '$1>1' | wc -l

221

1. **If you exclude transcripts on alternate haplotypes (e.g. chr6\_apd\_hap1, chr6\_cox\_hap2, etc.), how many coding genes have transcripts on more than one chromosome?**

awk '$3!~/hap/ && $7!=$8{print $1,$3}' refFlat.txt | sort -u   
| awk '{print $1}' | uniq -c | awk '$1>1' | wc -l

36

1. **What is the distribution of the number of transcripts for each gene, i.e. how many genes have a single transcript, how many have two, etc.?**

awk 'NR>1{print $1,$2}' refFlat.txt | sort -u | awk '{print $1}' | uniq -c | awk '{print $1}' | sort -n | uniq -c

# Hands-On Exercises #3

1. **Identify the transcripts (accession) that appear more than once (i.e. mapped to more than one location). Separate the RefSeq transcripts into two files: those that are mapped to only a single location, and those that are mapped to multiple locations.**

awk '{print $2}' refFlat.txt | sort | uniq -c   
| awk '$1>1{print $2}' > dup\_acc.list

sort -k2,2 refFlat.txt > refFlat.by\_acc

join -t $'\t' -1 1 -2 2 dup\_acc.list refFlat.by\_acc   
> refFlat.multimap

join -t $'\t' -1 1 -2 2 -v 2 dup\_acc.list refFlat.by\_acc   
> refFlat.singlemap

1. **a) Attach the intensity from the intensity.txt file to the corresponding RefSeq transcript and save the output to a file called refFlat\_with\_intensity.txt.**

First, sort both files on join column:

sort -k2,2 refFlat.txt > refFlat.by\_acc

sort -k1,1 intensity.txt > intensity.by\_acc

Then, join all transcripts that have an intensity value:

join -1 2 -2 1 refFlat.by\_acc intensity.by\_acc   
 > refFlat\_with\_intensity.txt

**b) Where there is no intensity data for a given accession, set the intensity to zero and save the output to a file called refFlat\_wo\_intensity.txt.**

Join all transcripts that do NOT have an intensity value, taking care to reverse the order of the first two fields (to match output of the join above), and add a zero for the missing intensity value:

join -1 2 -2 1 -v 1 refFlat.by\_acc intensity.by\_acc   
 | awk '{print $2,$1,$3,$4,$5,$6,$7,$8,$9,$10,$11,0}'   
 > refFlat\_wo\_intensity.txt

**c) Combine the two files together using the append operator “>>”.**

cat refFlat\_with\_intensity.txt refFlat\_wo\_intensity.txt >> refFlat\_plus\_intensity.txt

# Extra Credit Hands-On Exercises

1. **What is the average number of exons per transcript? What is the average number of exons for non-coding transcripts?**

awk 'NR>1{e=e+$9;c=c+1} END{print e/c}' refFlat.txt

average number of exons per transcript: 9.8

awk '$7==$8 && NR>1{e=e+$9;c=c+1} END{print e/c}' refFlat.txt

average number of exons for non-coding transcripts: 4.6

1. **Generate a file that retains only the longest transcript (txStart to txEnd) for each gene.**

awk 'NR>1{print $0,$6-$5}' refFlat.txt | sort -k1,1 -k12,12nr   
| awk -v OFS='\t' '$1!=g{print;g=$1}' > refFlat.longest\_txt

1. **A BED6-formatted file has six columns of interest: chrom, start, end, name, score, strand. Produce a BED6 file that contains 1000bp upstream of each transcript. The “name” should be “gene:prom:accession” (e.g.,** WASH7P:prom:NR\_024540***)* and the score for each promoter is 0*. More challenging variation:* take account of the orientation of the transcript: for those on '-' strand, "upstream" will be to the right of the end of the transcript (larger chromosome positions), rather than to the left of the start position.**

awk -v OFS="\t" '$4=="+"{print $3,$5-1000,$5,$1":prom:"$2,0,$4}  
$4=="-"{print $3,$6,$6+1000,$1":prom:"$2,0,$4}' refFlat.txt  
> refseq\_promoters.bed