Homework 3: Deepika Wali

Question #1

Create a BED file representing all of the intervals in the genome that are NOT exonic.

Answer1:

walid@20cb4a715e4b:~/Desktop/data2$ bedtools complement -i exons.bed -g genome.txt > Non\_exonic2.bed

walid@20cb4a715e4b:~/Desktop/data2$ cat Non\_exonic2.bed | head

chr1 0 11873

chr1 12227 12612

chr1 12721 13220

chr1 14829 14969

chr1 15038 15795

chr1 15947 16606

chr1 16765 16857

chr1 17055 17232

chr1 17368 17605

chr117742 17914

Question #2

What is the average distance from GWAS SNPs to the closest exon? (Hint - have a look at the `closest` tool.)

Answer2:

First we used bedtools closest and -d feature to get the distance of each SNP to the closest exon and sorted at the same time.

walid@20cb4a715e4b:~/Desktop/data2$ bedtools closest -a gwas.bed -b exons.bed -d -sorted > Avg\_distance.bed | head

walid@20cb4a715e4b:~/Desktop/data2$ cat Avg\_distance.bed | head -20

chr1 1005805 1005806 rs3934834 chr1 1007125 1007955 NM\_001205252\_exon\_0\_0\_chr1\_1007126\_r 0 - 1320 (11th Column)

chr1 1079197 1079198 rs11260603 chr1 1078118 1079434 NR\_038869\_exon\_2\_0\_chr1\_1078119\_f 0 + 0

Column 11 give us the distance measurement so we now will count the average distance using column 11.

walid@20cb4a715e4b:~/Desktop/data2$ awk '{tot\_len += $11}END{print tot\_len/NR}' Avg\_distance.bed

46713.1

Question #3

Count how many exons occur in each 500kb interval (“window”) in the human genome, what is the average value? (Hint - have a look at the `makewindows` tool.)

Answer: (500kb= 500000 bp) Here we use makewindow tool create window in genome.txt file and used -i srcwinnum for the source name and the window number information.

walid@20cb4a715e4b:~/Desktop/data2$ bedtools makewindows -g genome.txt -w 500000 -i srcwinnum | head -10

chr1 0 500000 chr1\_1

chr1 500000 1000000 chr1\_2

chr1 1000000 1500000 chr1\_3

chr1 1500000 2000000 chr1\_4

chr1 2000000 2500000 chr1\_5

chr1 2500000 3000000 chr1\_6

chr1 3000000 3500000 chr1\_7

chr1 3500000 4000000 chr1\_8

chr1 4000000 4500000 chr1\_9

chr14500000 5000000 chr1\_10

walid@20cb4a715e4b:~/Desktop/data2$ bedtools makewindows -g genome.txt -w 500000 -i srcwinnum > exon\_count.bed | head -20

walid@20cb4a715e4b:~/Desktop/data2$ bedtools intersect -a exon\_count.bed -b exons.bed -c | head

chr1 0 500000 chr1\_1 37

chr1 500000 1000000 chr1\_2 197

chr1 1000000 1500000 chr1\_3 477

chr1 1500000 2000000 chr1\_4 445

chr1 2000000 2500000 chr1\_5 209

chr1 2500000 3000000 chr1\_6 96

chr1 3000000 3500000 chr1\_7 83

chr1 3500000 4000000 chr1\_8 271

chr1 4000000 4500000 chr1\_9 9

chr14500000 5000000 chr1\_10 12

walid@20cb4a715e4b:~/Desktop/data2$ bedtools intersect -a exon\_count.bed -b exons.bed -c | awk '{tot\_len += $5}END{print tot\_len/NR}'

72.5188 exons occurs in each 500kb intervals.

Question #4

Are there any exons that are completely overlapped by an enhancer? If so, how many?

Answer:

Yes, there are overlapping enhancer on the exons. So, to look for enhancer we will use chromhmm file and look for enhancer in it.

walid@20cb4a715e4b:~/Desktop/data2$ grep 'Enhancer' hesc.chromHmm.bed > overlap\_enhancers.bed

walid@20cb4a715e4b:~/Desktop/data2$ cat overlap\_enhancers.bed | head -10

chr1 27537 27737 6\_Weak\_Enhancer

chr1 30337 30537 6\_Weak\_Enhancer

chr1 34737 34937 7\_Weak\_Enhancer

chr1 35737 35937 7\_Weak\_Enhancer

chr1 35937 36137 6\_Weak\_Enhancer

chr1 36137 36337 7\_Weak\_Enhancer

chr1 36337 36537 5\_Strong\_Enhancer

chr1 36537 37537 6\_Weak\_Enhancer

chr1 37537 37737 7\_Weak\_Enhancer

chr156337 56537 6\_Weak\_Enhancer

walid@20cb4a715e4b:~/Desktop/data2$ bedtools intersect -a exons.bed -b overlap\_enhancers.bed -wo -f 1 | wc -l

13746

As mentioned, that how many are completely overlapped so we used (-f 1).

Question #5

What fraction of the GWAS SNPs are exonic?

Answer: Even with 50 % fraction as a cut off we get the same value for SNP that are exonic.

walid@20cb4a715e4b:~/Desktop/data2$ bedtools intersect -a gwas.bed -b exons.bed -wo -f 1 | wc -l

3439

walid@20cb4a715e4b:~/Desktop/data2$ bedtools intersect -a gwas.bed -b exons.bed -wo -f 0.5 | wc -l

3439

Question #6

What fraction of the GWAS SNPs lie in either enhancers or promoters in the hESC data we have?

Answer:

For promoters or enhancers: 1285 SNP lie either in enhancers or promoters

walid@20cb4a715e4b:~/Desktop/data2$ grep 'Promoter' hesc.chromHmm.bed > overlap\_promoter.bed

walid@20cb4a715e4b:~/Desktop/data2$ cat overlap\_promoter.bed | head -5

chr1 27737 28537 2\_Weak\_Promoter

chr1 28537 30137 1\_Active\_Promoter

chr1 30137 30337 2\_Weak\_Promoter

chr1 30537 30737 3\_Poised\_Promoter

chr1713137 713337 2\_Weak\_Promoter

walid@20cb4a715e4b:~/Desktop/data2$ bedtools intersect -a gwas.bed -b overlap\_promoter.bed overlap\_enhancers.bed -wo | wc -l

1285

chr1 1079197 1079198 rs11260603 2 chr1 1078937 1080337 7\_Weak\_Enhancer 1

chr1 3649561 3649562 rs9662633 2 chr1 3648540 3650140 6\_Weak\_Enhancer 1

chr1 3651030 3651031 rs12562437 2 chr1 3650140 3651140 7\_Weak\_Enhancer 1

chr1 8021972 8021973 rs35675666 1 chr1 8021013 8022013 1\_Active\_Promoter 1

chr1 8021972 8021973 rs35675666 1 chr1 8021013 8022013 1\_Active\_Promoter 1

Question #7

Create intervals representing the canonical 2bp splice sites on either side of each exon (don’t worry about excluding splice sites at the first or last exon). (Hint - have a look at the flank tool.)

Answer:

Here we used flank bedtools that created flank in each direction and as mentioned was 2bp splice site. If we look in the answer 11871 and 11873 have a difference of 2 bp in the first exon. So that is the 2 bp splice site created.

walid@20cb4a715e4b:~/Desktop/data2$ bedtools flank -i exons.bed -g genome.txt -b 2 | head -20

chr1 11871 11873 NR\_046018\_exon\_0\_0\_chr1\_11874\_f 0 +

chr1 12227 12229 NR\_046018\_exon\_0\_0\_chr1\_11874\_f 0 +

chr1 12610 12612 NR\_046018\_exon\_1\_0\_chr1\_12613\_f 0 +

chr1 12721 12723 NR\_046018\_exon\_1\_0\_chr1\_12613\_f 0 +

chr1 13218 13220 NR\_046018\_exon\_2\_0\_chr1\_13221\_f 0 +

chr1 14409 14411 NR\_046018\_exon\_2\_0\_chr1\_13221\_f 0 +

chr1 14359 14361 NR\_024540\_exon\_0\_0\_chr1\_14362\_r 0 -

Question #8

Which hESC ChromHMM state (e.g., `11\_Weak\_Txn`, `10\_Txn\_Elongation`) represents the most number of base pairs in each of chromosome 19 and chromosome 8?

Answer:

To calculate number of base pairs across the chromatin state, we will first use the awk command to locate in the first columm for the chromosome 8 and chr 19 in the chromHmm file.

walid@20cb4a715e4b:~/Desktop/data2$ awk '$1 == "chr8"' hesc.chromHmm.bed | head -20

chr8 10000 20000 13\_Heterochrom/lo

chr8 20000 20200 8\_Insulator

chr8 20200 21000 13\_Heterochrom/lo

chr8 21000 21400 8\_Insulator

chr8 21400 44400 13\_Heterochrom/lo

chr8 44400 45200 8\_Insulator

chr8 45200 49600 11\_Weak\_Txn

chr8 49600 49800 7\_Weak\_Enhancer

chr8 49800 66200 13\_Heterochrom/lo

walid@20cb4a715e4b:~/Desktop/data2$ awk '$1 == "chr19"' hesc.chromHmm.bed | head -20

chr19 60000 60200 13\_Heterochrom/lo

chr19 60200 60400 8\_Insulator

chr19 60400 62000 11\_Weak\_Txn

chr19 62000 69000 9\_Txn\_Transition

chr19 69000 69200 11\_Weak\_Txn

chr19 69200 69600 6\_Weak\_Enhancer

chr19 69600 70400 2\_Weak\_Promoter

chr19 70400 71600 1\_Active\_Promoter

chr19 71600 71800 2\_Weak\_Promoter

chr19 71800 72400 6\_Weak\_Enhancer

walid@20cb4a715e4b:~/Desktop/data2$ awk '$1 == "chr8"' hesc.chromHmm.bed | awk '{a[$4] += ($3-$2)}END {for (i in a) print i, a[i]}'| sort -k2,2nr

13\_Heterochrom/lo 107002678

11\_Weak\_Txn 22299825

7\_Weak\_Enhancer 3121603

10\_Txn\_Elongation 2566001

walid@20cb4a715e4b:~/Desktop/data2$ awk '$1 == "chr19"' hesc.chromHmm.bed | awk '{a[$4] += ($3-$2)}END {for (i in a) print i, a[i]}'| sort -k2,2nr

13\_Heterochrom/lo 25202194

11\_Weak\_Txn 15828772

10\_Txn\_Elongation 4841844

9\_Txn\_Transition 1770999

12\_Repressed 1523400

\*\*\*