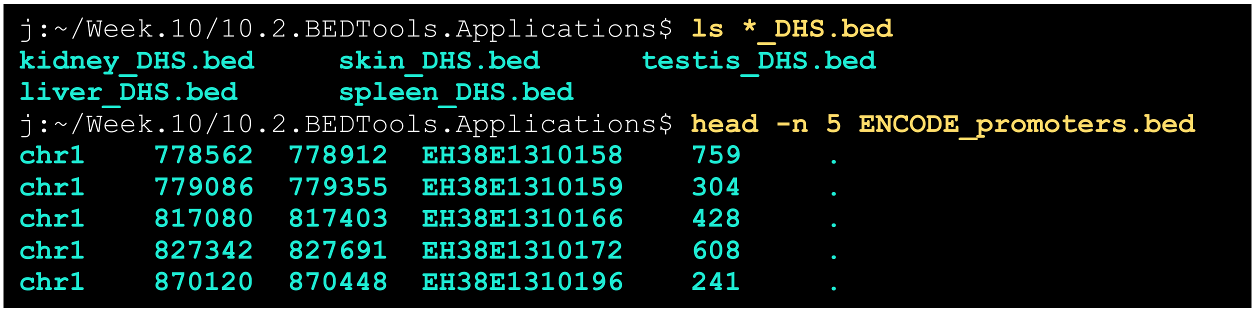
10.2 BEDTools Applications

10.2.1 BEDTools Subtract

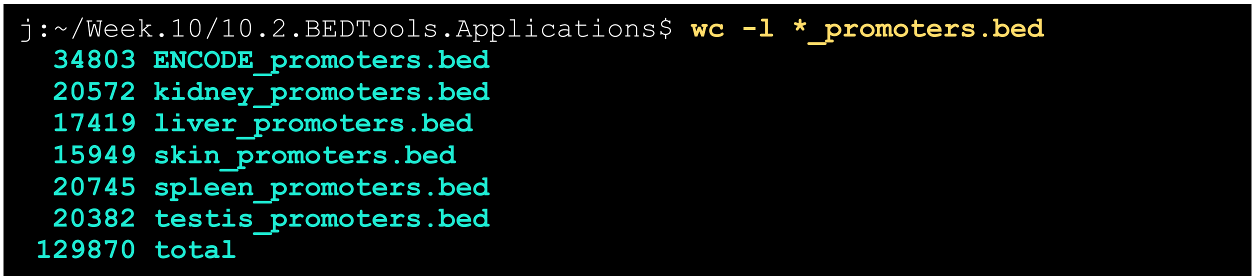
Subtract

In 10.1, DHSs in kidney, liver, skin, spleen, testis tissue samples were intersected with promoter regions identified by ENCODE.



```
j:~/Week.10/10.2.BEDTools.Applications$ ls *_DHS.bed
kidney_DHS.bed      skin_DHS.bed      testis_DHS.bed
liver_DHS.bed       spleen_DHS.bed
j:~/Week.10/10.2.BEDTools.Applications$ head -n 5 ENCODE_promoters.bed
chr1    778562  778912  EH38E1310158    759     .
chr1    779086  779355  EH38E1310159    304     .
chr1    817080  817403  EH38E1310166    428     .
chr1    827342  827691  EH38E1310172    608     .
chr1    870120  870448  EH38E1310196    241     .
```

Intersections were performed using the `bedtools intersect -u` option. This gave us 5 files, each containing the full sequences of the promoters that are accessible in each sample.

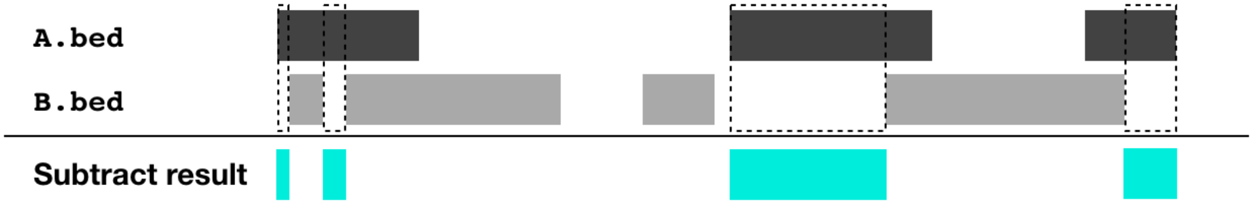


```
j:~/Week.10/10.2.BEDTools.Applications$ wc -l *_promoters.bed
  34803 ENCODE_promoters.bed
  20572 kidney_promoters.bed
  17419 liver_promoters.bed
  15949 skin_promoters.bed
  20745 spleen_promoters.bed
  20382 testis_promoters.bed
 129870 total
```

Another BEDTools subcommand is `subtract`. The `subtract` command finds regions in one BED file that are NOT overlapped by regions in another BED file, specifically it finds regions in `A.bed` that are NOT overlapped by regions in `B.bed`.

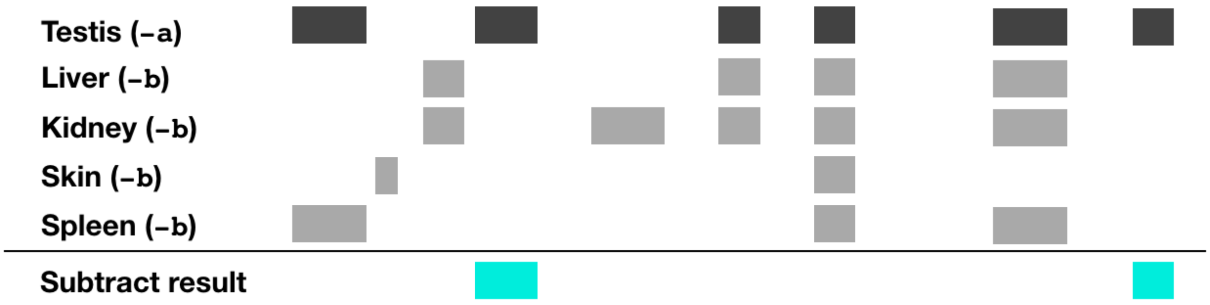
```
bedtools subtract -a A.bed -b B.bed
```

The blocks in the figure below represent regions in `A.bed` and `B.bed` as well as the result of the `subtract` subcommand.



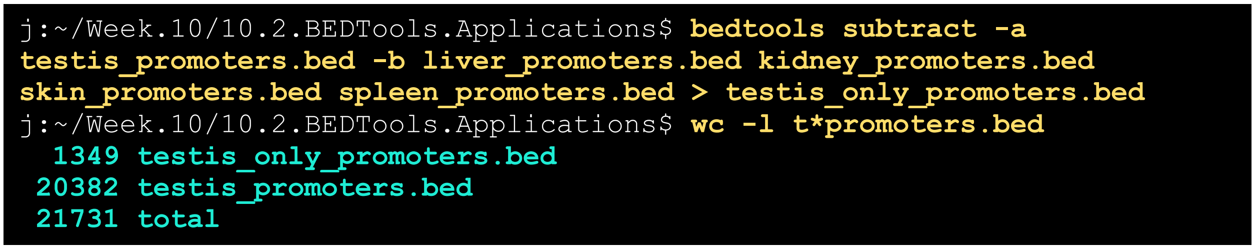
The output is a BED file containing the coordinates for each interval in `A.bed` that is not overlapped by any intervals in `B.bed`. This is the opposite of the `intersect` command.

To learn which promoters are accessible in the human testis sample, but are not accessible in the liver, kidney, skin, or spleen samples, BEDTools `subtract` can be applied. The `-b` option in `subtract` can be provided with multiple files.



Notice the `subtract` result only contains intervals in the testis file that are not in ANY of the other files.

Using `subtract`, the promoters in `testis_promoters.bed` that are not in `liver_promoters.bed`, `kidney_promoters.bed`, `skin_promoters.bed`, or `spleen_promoters.bed`, are output to the file `testis_only_promoters.bed`.



```
j:~/Week.10/10.2.BEDTools.Applications$ bedtools subtract -a
testis_promoters.bed -b liver_promoters.bed kidney_promoters.bed
skin_promoters.bed spleen_promoters.bed > testis_only_promoters.bed
j:~/Week.10/10.2.BEDTools.Applications$ wc -l t*promoters.bed
   1349 testis_only_promoters.bed
  20382 testis_promoters.bed
  21731 total
```

While there are 20,382 lines in the file `testis_promoters.bed`, only 1,239 lines are in the file `testis_only_promoters.bed`. This means that most accessible promoters in testis are also accessible in at least one of the other cell types.

10.2.2 Ensembl IDs & Transcript Isoforms

Ensembl Identifiers

"Ensembl is a project to develop a software system which produces and maintains automatic annotation on selected eukaryotic genomes." - https://ensembl.org

Ensembl is project to maintain a database of all the genes in eukaryotic genomes. Ensembl keeps track of gene annotations, transcript isoforms, homologues, and other information about eukaryotic genomes.
Importantly, Ensembl keeps track of genes by assigning each gene in each genome a unique identifier. Each human gene has a unique "ENSG" identifier. These identifiers start with the letters ENSG, which is followed by 11 numbers (**ENSG00000000001**).

For example, the Ensembl gene identifier for the human gene DDA1 is ENSG00000130311, and the Ensembl gene identifier for the human gene ANO8 is ENSG00000074855.

Sometimes identifiers end with a version number after a period, "ENSG00000000001.1". As more is learned about genes the sequences may be updated, and the version number will change. In this course, version numbers will not be considered in analyses.

Transcripts also have Ensembl identifiers. These identifiers start with the letters ENST, which is followed by 11 numbers (**ENST00000000001**).

Transcript Isoforms

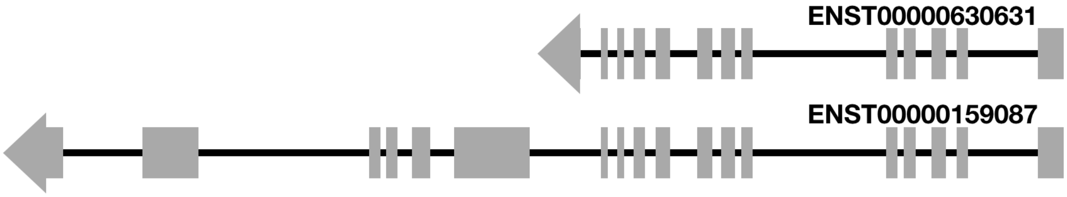
The genes DDA1 and ANO8 each have one unique gene identifier. The DDA1 gene has the transcript ID ENST00000359866, but the ANO8 gene has 2 transcript IDs: ENST00000630631 and ENST00000159087.

This is because the ANO8 gene has two **transcript isoforms**. A transcript isoform is a version of a transcribed gene, thus a single gene can have multiple transcript isoforms. Transcript isoforms can be regulated differently and/or produce different protein products. For a given gene, transcript isoforms may have different:

- Transcription start sites
- 5'UTRs
- 3' UTRs
- Splicing patterns

Thus, transcript isoforms are different lengths and have different coordinates in the genome.

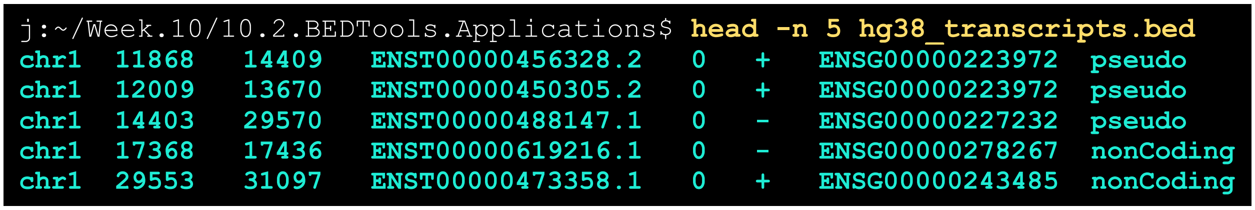
The ANO8 gene is on the negative strand of chromosome 19. The two ANO8 transcript isoforms have very close transcription start sites, but one is much longer than the other due to transcript ENST00000159087 containing more exons than the shorter transcript, ENST00000630631. The image below depicts the two ANO8 transcripts, each gray block represents one exon (note that UTRs are not differentiated).



Note that transcript isoform coordinates will overlap with one another.

Transcripts in the Human Genome

As new data is accrued, genomes are improved upon, and new versions of genomes are released. Human genome assemblies start with the letters "hg" and are followed by the version number. Human genome hg38 is the 38th version of the genome. The file `hg38_transcripts.bed` contains the locations of all transcripts in the hg38 genome.



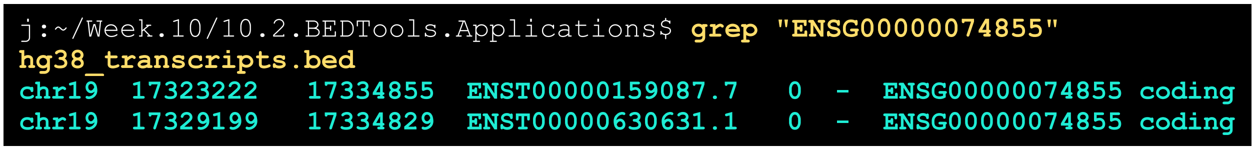
```
j:~/Week.10/10.2.BEDTools.Applications$ head -n 5 hg38_transcripts.bed
chr1  11868   14409   ENST00000456328.2   0   +   ENSG00000223972  pseudo
chr1  12009   13670   ENST00000450305.2   0   +   ENSG00000223972  pseudo
chr1  14403   29570   ENST00000488147.1   0   -   ENSG00000227232  pseudo
chr1  17368   17436   ENST00000619216.1   0   -   ENSG00000278267  nonCoding
chr1  29553   31097   ENST00000473358.1   0   +   ENSG00000243485  nonCoding
```

Each line in `hg38_transcripts.bed` contains the coordinates of one transcript, followed by the Ensembl transcript ID in the name column (4th column), a 0 placeholder in the score column (5th column), and the strand the transcript is on (6th column). This BED file also contains two more columns: the 7th column contains the Ensembl gene ID and the 8th column contains the gene type.

There are three possible categories in the gene type column:

- `coding`: encodes a protein product
- `nonCoding`: encodes an RNA transcript that not translated, for example transfer RNA, ribosomal RNA, long non-coding RNA
- `pseudo`: a pseudogene, a sequence that resembles a coding gene but is non-functional

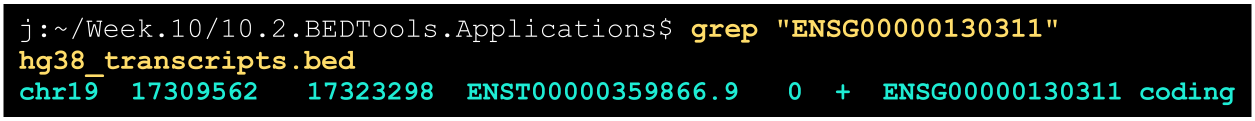
Remember that the ANO8 gene has two transcript isoforms. The result of searching for the ANO8 Ensembl gene ID (ENSG00000074855) in `hg38_transcripts.bed` produces two results.



```
j:~/Week.10/10.2.BEDTools.Applications$ grep "ENSG00000074855"
hg38_transcripts.bed
chr19  17323222   17334855  ENST00000159087.7   0   -  ENSG00000074855 coding
chr19  17329199   17334829  ENST00000630631.1   0   -  ENSG00000074855 coding
```

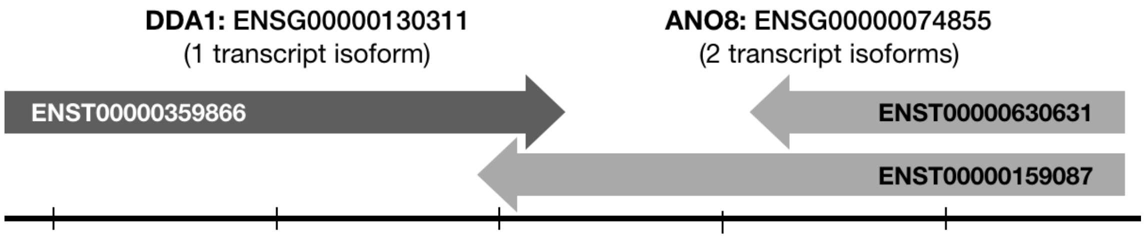
Both entries have in the BED file have the same gene ID, but different transcript IDs. The first transcript returned is much longer than the second.

The DDA1 gene only has one transcript isoform. The result of searching for the DDA1 Ensembl gene ID (ENST00000359866) in `hg38_transcripts.bed` produces one result.



```
j:~/Week.10/10.2.BEDTools.Applications$ grep "ENSG00000130311"
hg38_transcripts.bed
chr19  17309562   17323298  ENST00000359866.9   0   +  ENSG00000130311 coding
```

Illustrated below are the coordinates for the DDA1 transcript isoform and the two ANO8 transcript isoforms. Both genes are on chromosome 19, DDA1 is on the positive strand and ANO8 is on the negative strand.



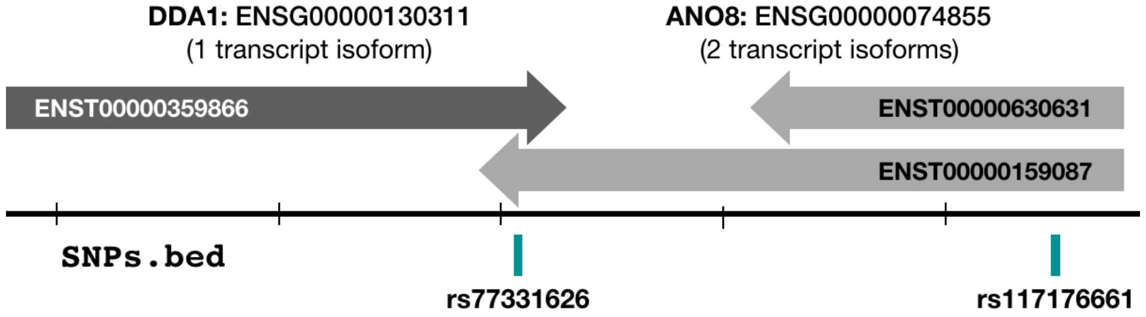
The two ANO8 transcript isoforms overlap one another. The DDA1 transcript also overlaps one of the ANO8 transcript isoforms. Many genes in the genome overlap with one another.

BEDTools Results with Transcripts

It is important to consider overlapping genes and transcript isoforms when using BEDTools. For example, the file SNPs.bed contains coordinates for two single nucleotide polymorphisms (SNPs) with identifiers.

```
j:~/Week.10/10.2.BEDTools.Applications$ cat SNPs.bed
chr19   17323283        17323284        rs77331626    0     .
chr19   17334398        17334399        rs117176661   0     .
```

In the genome, the SNP rs77331626 overlaps with both DDA1 and one ANO8 transcript. The SNP rs117176661 overlaps with both ANO8 transcript isoforms.



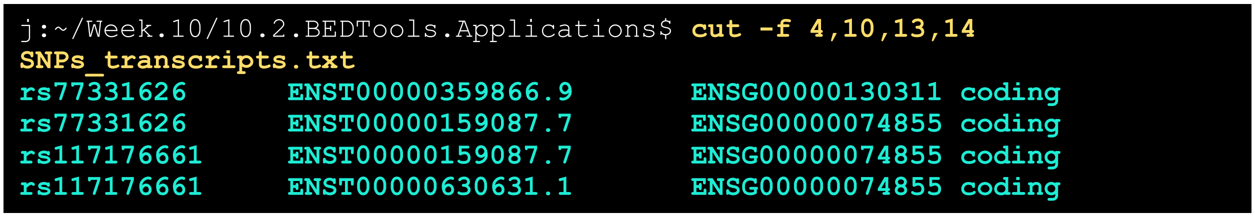
Intersecting SNPs.bed with hg38_transcripts.bed can return multiple results for each SNP. Note that the –wb option is used, which returns each item in SNPs.bed and information about the region it intersects in hg38_transcripts.bed.

```
j:~/Week.10/10.2.BEDTools.Applications$ bedtools intersect -wb -a
SNPs.bed -b hg38_transcripts.bed > SNPs_transcripts.txt
j:~/Week.10/10.2.BEDTools.Applications$ cat SNPs_transcripts.txt
chr19  17323283   17323284  rs77331626    0   .  chr19   17309562   17323298
ENST00000359866.9    0    +     ENSG00000130311 coding
chr19  17323283   17323284  rs77331626    0   .  chr19   17323222   17334855
ENST00000159087.7    0    -     ENSG00000074855 coding
chr19  17334398   17334399  rs117176661   0   .  chr19   17323222   17334855
ENST00000159087.7    0    -     ENSG00000074855 coding
chr19  17334398   17334399  rs117176661   0   .  chr19   17329199   17334829
ENST00000630631.1    0    -     ENSG00000074855 coding
```

Each row has the following 14 fields:

1. SNP chromosome
2. SNP start
3. SNP end
4. SNP name
5. SNP score
6. SNP strand
7. Transcript chromosome
8. Transcript start
9. Transcript end
10. Transcript name
11. Transcript score
12. Transcript strand
13. Gene name
14. Gene type

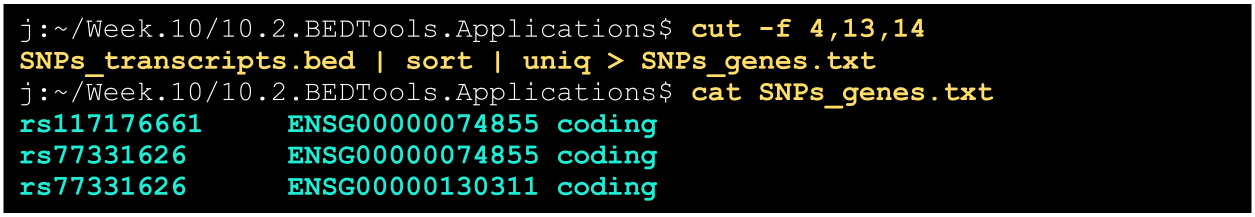
To view the results more easily, the fields can be reduced to SNP name, transcript name, gene name, and gene type (fields 4, 10, 13, & 14).



```
j:~/Week.10/10.2.BEDTools.Applications$ cut -f 4,10,13,14
SNPs_transcripts.txt
rs77331626      ENST00000359866.9       ENSG00000130311 coding
rs77331626      ENST00000159087.7       ENSG00000074855 coding
rs117176661     ENST00000159087.7       ENSG00000074855 coding
rs117176661     ENST00000630631.1       ENSG00000074855 coding
```

There are two results for SNP rs77331626. It overlaps a transcript from the gene ENSG00000130311 and a transcript from the gene ENSG00000074855. There are also two results for SNP rs117176661. It overlaps two transcripts from the gene ENSG00000074855.

To reduce the results to identify which genes are overlapped by each SNP, the SNP name, the gene name, and the gene type can be extracted with the `cut` command and then piped to `sort` and `uniq`.



```
j:~/Week.10/10.2.BEDTools.Applications$ cut -f 4,13,14
SNPs_transcripts.bed | sort | uniq > SNPs_genes.txt
j:~/Week.10/10.2.BEDTools.Applications$ cat SNPs_genes.txt
rs117176661     ENSG00000074855 coding
rs77331626      ENSG00000074855 coding
rs77331626      ENSG00000130311 coding
```

The two transcript overlaps for rs117176661 are collapsed into a single row for the gene.

10.2.3 BEDTools Closest

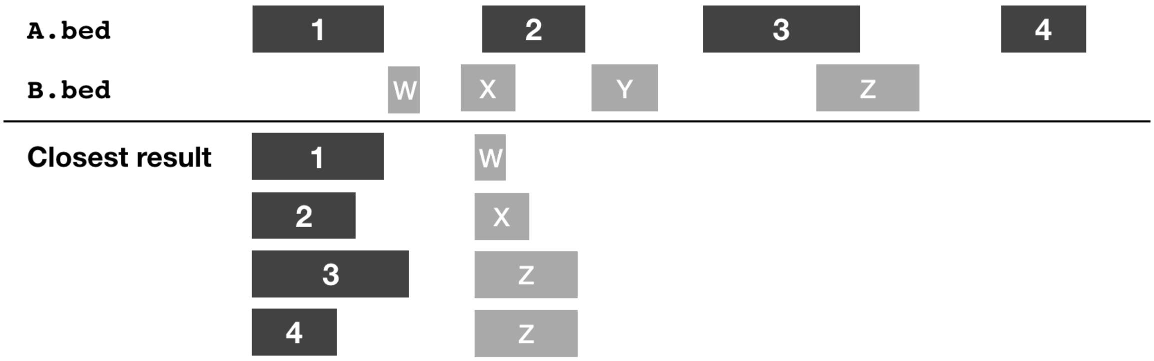
Closest

With BEDTools subcommands `intersect` and `subtract`, the set of promoters accessible in the testis tissue sample, but not in the liver, kidney, skin, or spleen tissue samples were identified. To find the genes that are likely regulated by each of the accessible promoters, the closest gene to each promoter can be determined.

Another BEDTools subcommand is `closest`. The `closest` command finds the region in one BED file that is closest to each region in another BED file. Specifically, for each region in `A.bed` the closest region in `B.bed` is found.

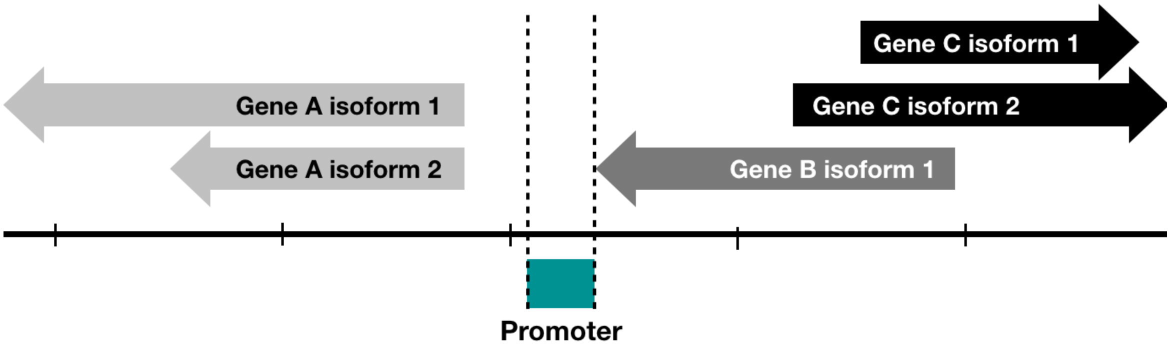
```
bedtools closest -a A.bed -b B.bed
```

The blocks in the figure below represent regions in `A.bed` and `B.bed` as well as the result of the `closest` subcommand.



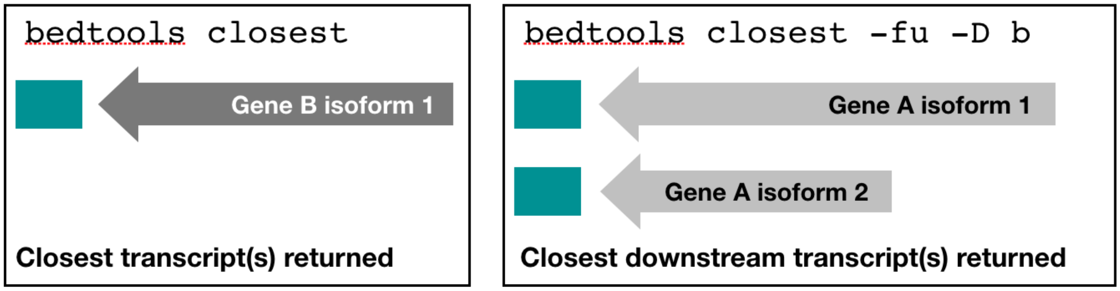
The output is a BED file containing the coordinates for each interval in `A.bed` followed by the coordinates for interval in `B.bed` that is closest.

In the case of finding the closest gene to each promoter, it is not sufficient to find only the closest gene. Instead, the closest gene *downstream* of the promoter is needed. The visual below depicts a promoter and its proximity to three genes, two of which have multiple isoforms:



The closest gene to the promoter is gene B (isoform 1), however, this gene is on the negative strand and therefore is upstream of the promoter. The promoter is upstream of gene A (isoforms 1 & 2), therefore the promoter gene A is likely to be regulated by this promoter.

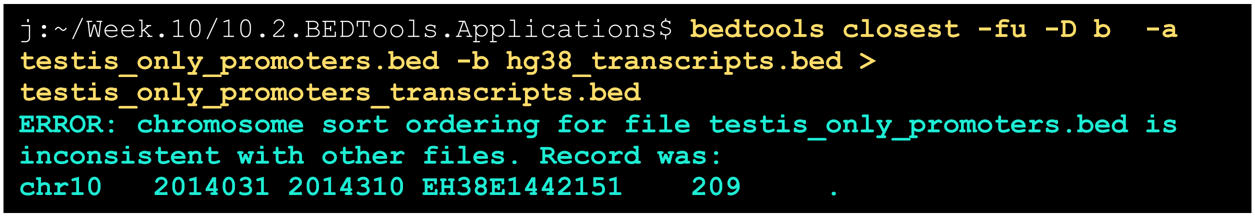
The default version of BEDTools `closest` would return gene B. To instead return the gene A transcript isoforms, `closest` can be run with the options `-fu -D b`, which instruct the command to return the closest downstream region in file `-b` to each region in file `-a`.



bedtools closest
Gene B isoform 1
Closest transcript(s) returned
bedtools closest -fu -D b
Gene A isoform 1
Gene A isoform 2
Closest downstream transcript(s) returned

Sort

To find the closest transcript(s) in `hg38_transcripts.bed` to each promoter in `testis_only_promoters.bed`, BEDTools `closest` can be run with the options `-fu -D b`.



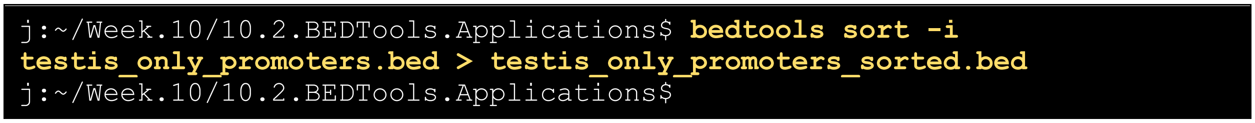
```
j:~/Week.10/10.2.BEDTools.Applications$ bedtools closest -fu -D b  -a
testis_only_promoters.bed -b hg38_transcripts.bed >
testis_only_promoters_transcripts.bed
ERROR: chromosome sort ordering for file testis_only_promoters.bed is
inconsistent with other files. Record was:
chr10   2014031 2014310 EH38E1442151    209     .
```

An error occurred stating that the “chromosome sort ordering for `testis_only_promoters.bed` is inconsistent with other files”. Many BEDTools commands require BED files to be sorted in the same order (although `intersect` and `subtract` do not).

To sort `testis_only_promoters.bed`, BEDTools `sort` can be used:

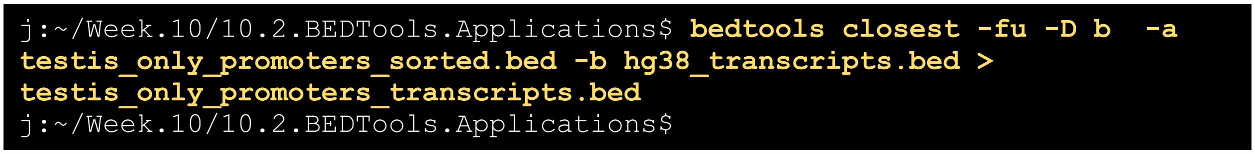
```
bedtools sort -i A.bed
```

A new file containing the sorted contents of `testis_only_promoters.bed` can be generated using BEDTools `sort`.



```
j:~/Week.10/10.2.BEDTools.Applications$ bedtools sort -i
testis_only_promoters.bed > testis_only_promoters_sorted.bed
j:~/Week.10/10.2.BEDTools.Applications$
```

Now BEDTools `closest` can be run again, this time using the sorted file `testis_only_promoters_sorted.bed`.

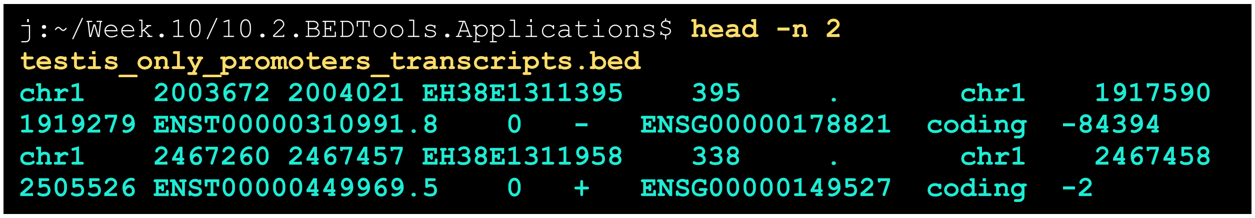


```
j:~/Week.10/10.2.BEDTools.Applications$ bedtools closest -fu -D b  -a
testis_only_promoters_sorted.bed -b hg38_transcripts.bed >
testis_only_promoters_transcripts.bed
j:~/Week.10/10.2.BEDTools.Applications$
```

This time there was no error and the command was successful. In the case that another error was thrown, `hg38_transcripts.bed` would also need to be sorted.

Curating Results

The output of BEDTools `closest` will look much like the result of the intersection of `SNPs.bed` with `hg38_transcripts.bed` performed in 10.2.2, when the `-wb` option was used.



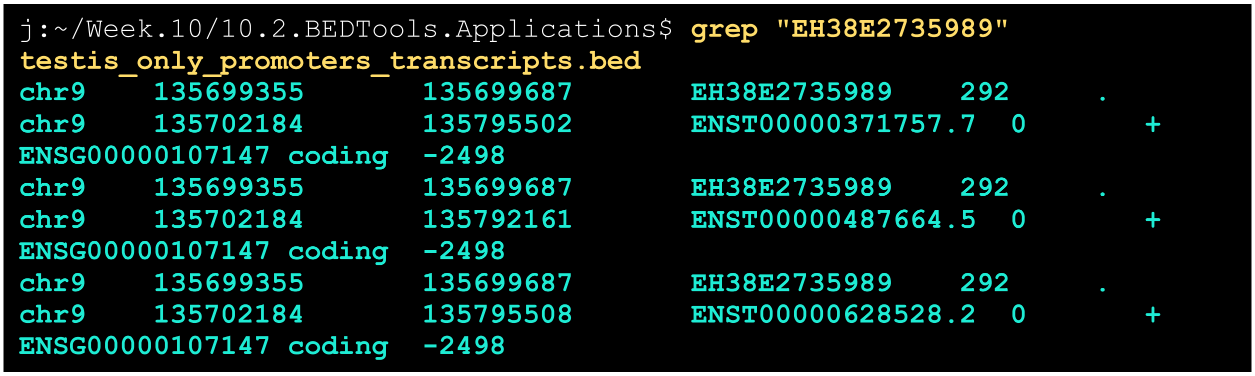
```
j:~/Week.10/10.2.BEDTools.Applications$ head -n 2
testis_only_promoters_transcripts.bed
chr1    2003672 2004021 EH38E1311395    395     .       chr1    1917590
1919279 ENST00000310991.8    0    -    ENSG00000178821  coding  -84394
chr1    2467260 2467457 EH38E1311958    338     .       chr1    2467458
2505526 ENST00000449969.5    0    +    ENSG00000149527  coding  -2
```

Each row has the following 15 fields:

1. Promoter chromosome
2. Promoter start
3. Promoter end
4. Promoter name
5. Promoter score
6. Promoter strand
7. Transcript chromosome
8. Transcript start
9. Transcript end
10. Transcript name
11. Transcript score
12. Transcript strand
13. Gene name
14. Gene type
15. **Distance from promoter to transcript**

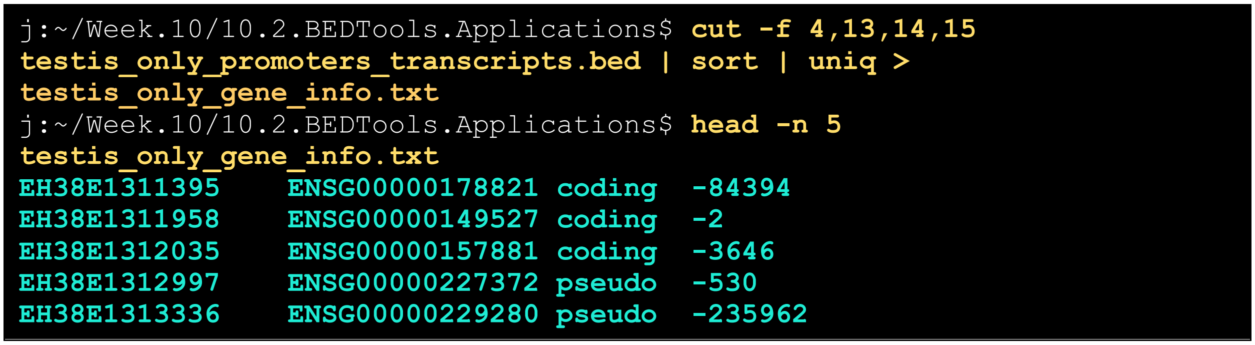
The final field in the output file is the distance between the promoter and the closest **downstream** transcript.

Because transcripts are being used, if a gene has multiple isoforms that all start at the same place in the genome, a promoter will be equally close to all the isoforms. For example, the promoter with the ID EH38E2735989 is equally close to three transcript isoforms for the gene ENSG00000107147.



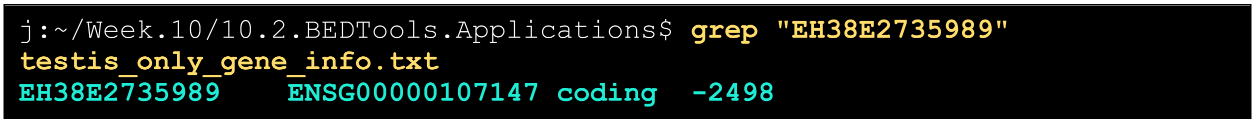
```
j:~/Week.10/10.2.BEDTools.Applications$ grep "EH38E2735989"
testis_only_promoters_transcripts.bed
chr9    135699355       135699687       EH38E2735989    292     .
chr9    135702184       135795502       ENST00000371757.7   0       +
ENSG00000107147 coding  -2498
chr9    135699355       135699687       EH38E2735989    292     .
chr9    135702184       135792161       ENST00000487664.5   0       +
ENSG00000107147 coding  -2498
chr9    135699355       135699687       EH38E2735989    292     .
chr9    135702184       135795508       ENST00000628528.2   0       +
ENSG00000107147 coding  -2498
```

For further analysis, each of these instances can be collapsed into one by retrieving each *unique* pair of promoter and gene, along with gene type and the distance from the promoter (fields 4, 13, 14, & 15).



```
j:~/Week.10/10.2.BEDTools.Applications$ cut -f 4,13,14,15
testis_only_promoters_transcripts.bed | sort | uniq >
testis_only_gene_info.txt
j:~/Week.10/10.2.BEDTools.Applications$ head -n 5
testis_only_gene_info.txt
EH38E1311395    ENSG00000178821 coding  -84394
EH38E1311958    ENSG00000149527 coding  -2
EH38E1312035    ENSG00000157881 coding  -3646
EH38E1312997    ENSG00000227372 pseudo  -530
EH38E1313336    ENSG00000229280 pseudo  -235962
```

In this file the promoter with the ID EH38E2735989 is represented by one line with the gene ENSG00000107147.



```
j:~/Week.10/10.2.BEDTools.Applications$ grep "EH38E2735989"
testis_only_gene_info.txt
EH38E2735989    ENSG00000107147 coding  -2498
```

This file now contains each promoter that is accessible in the testis sample and not in the liver, kidney, skin, or spleen samples, along with the closest gene, the type of gene, and the distance between the promoter and the gene.

10.2.4 Analysis of Results in R

The contents of the `testis_only_gene_info.txt` file generated in 10.2.3 can be further analyzed in R. Open the file `Supplement.10.2.Part.2.Rmd` in RStudio to continue.