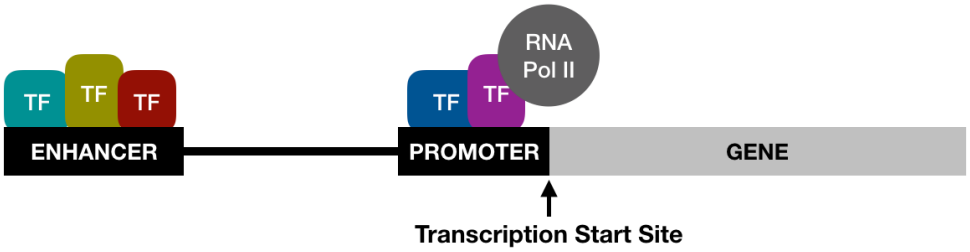
10.1 BEDTools

10.1.1 Functional Genomics Assays

Transcriptional Regulation

In humans, protein-coding genes are transcribed by RNA polymerase II (RNA pol II). RNA pol II is recruited to the transcription start site by **transcription factors** that bind to the **promoter sequence**. The promoter sequence is found just upstream of the transcription start site for a gene. **Transcription factors** (or TFs) are proteins that bind to DNA to alter the rate of expression of a gene. TFs also bind to **enhancers**. An enhancer is a region of the genome where TFs can bind to increase the transcription of a target gene. The target gene must be on the same chromosome as the enhancer, but it can be up to 1 Mb away.

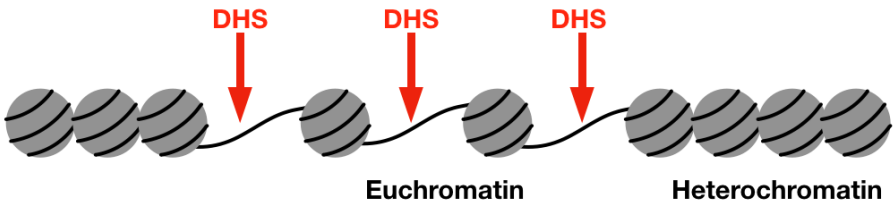


The ability of TFs to bind DNA is affected by the chromatin state: whether the chromatin is accessible or inaccessible. **Chromatin accessibility** refers to packing of chromatin: loosely packed vs. tightly packed. Chromatin that is loosely packed is more accessible. The tightening and loosening of chromatin is important to the regulation of gene expression; loosely packed (accessible) regions can be more easily bound by TFs, meaning that active transcription can occur at these regions.

Heterochromatin (or closed chromatin) is condensed (tightly packed) and therefore not accessible for transcription. **Euchromatin** (or open chromatin), is more loosely packed and is accessible for transcription.

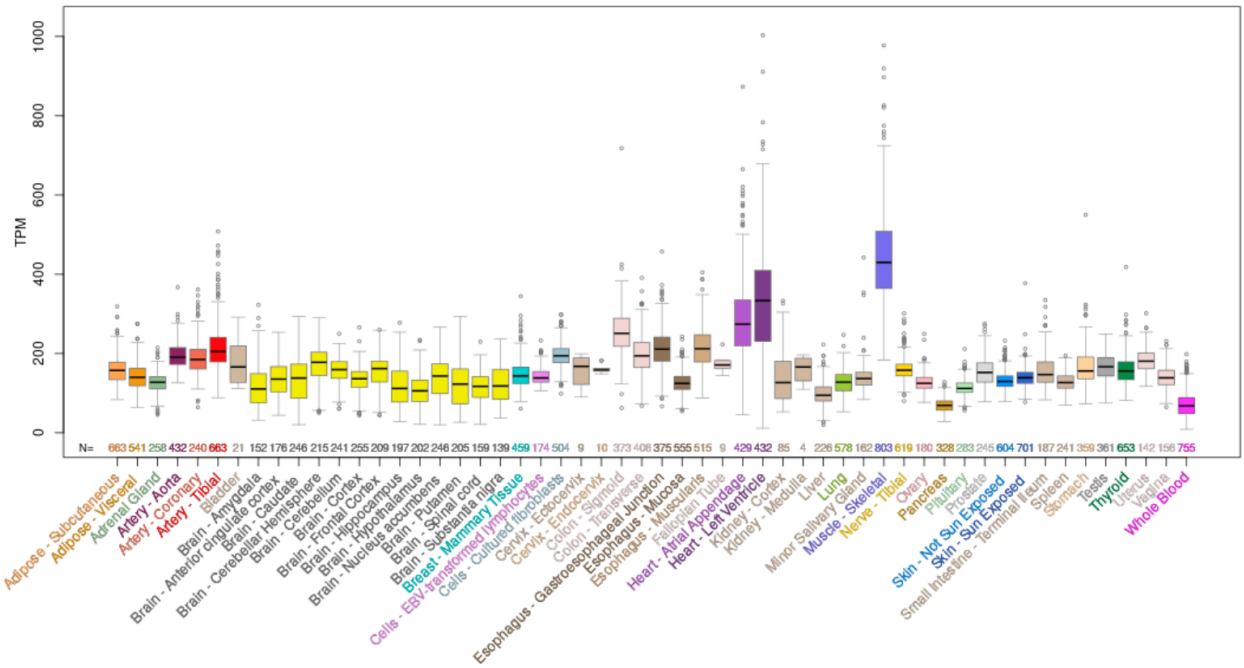
The parts of the genome contained in euchromatin and heterochromatin are **different** in different cell types.

DNase I is a DNA endonuclease that digests DNA by cutting the DNA backbone. A **DNase I hypersensitive site (DHS)** is a genomic region that is sensitive to cleavage by DNase I, because the chromatin is accessible. Therefore, euchromatin can be detected by assaying for DHSs.



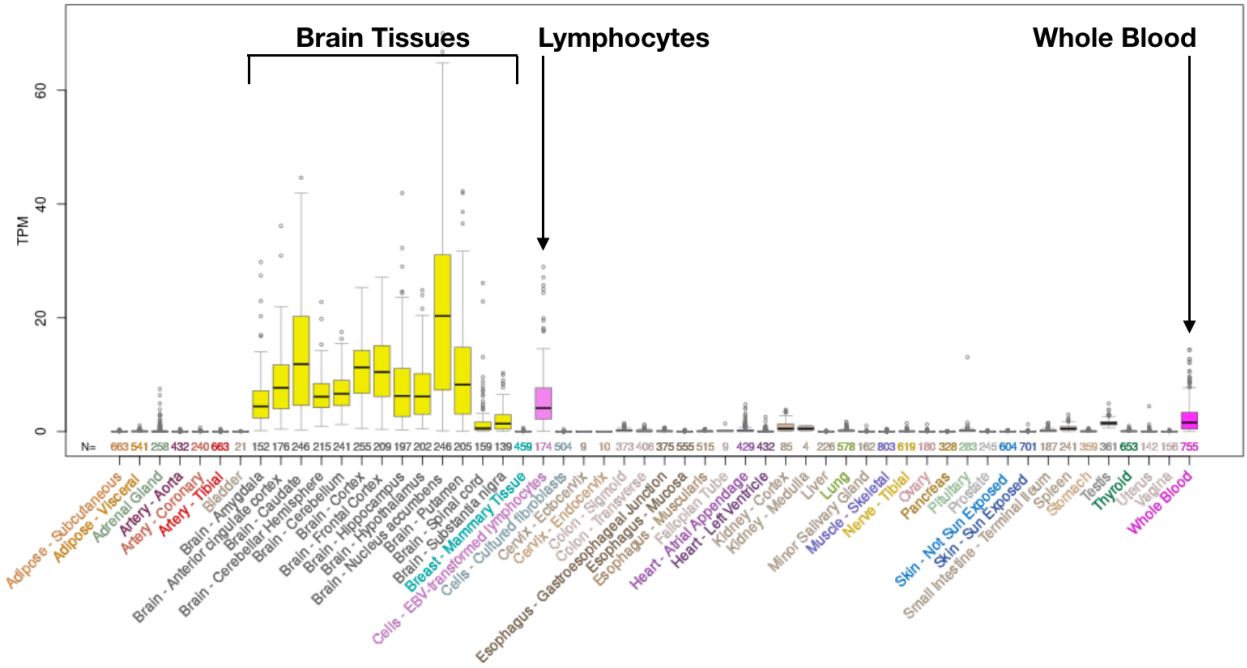
Because regions of active transcription are euchromatic, active protomers and enhancers are also marked by DHSs. (Note: regions of transcription are euchromatic, but not all euchromatic regions are transcribed).

The following plot shows the expression level of the gene NDUFB10 in 54 tissue types (sourced from GTEx: https://gtexportal.org). The expression varies a slightly between tissue types, however, the gene is expressed in all 54 tissue types.



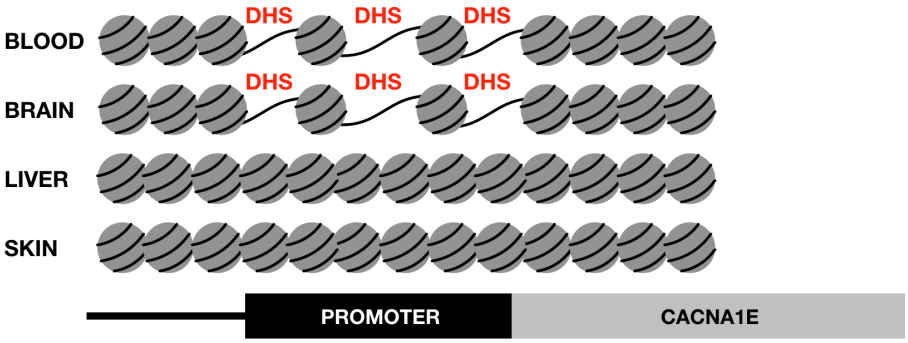
A gene that is expressed across all tissues is referred to as a **ubiquitously expressed gene** (also sometimes referred to as a **housekeeping gene**). Based on this data, the promoter for NDUFB10 is active in all tissue types, meaning that the promoter is accessible in all tissue types. Therefore, one would expect the NUFB10 promoter to be a DHS in all tissues.

Not all genes are ubiquitously expressed, there are also **tissue specific genes**: genes that are expressed in some tissues and not in others. For example, the gene CACNA1E. The following plot shows the expression level of the gene CACNA1E in 54 tissue types (sourced from GTEx: https://gtexportal.org).



CACNA1E is expressed in brain tissues and blood cells (lymphocytes & whole blood), it is not expressed in the other tissues. Therefore, the promoter for CACNA1E is only active (and accessible) in brain tissues and blood cells.

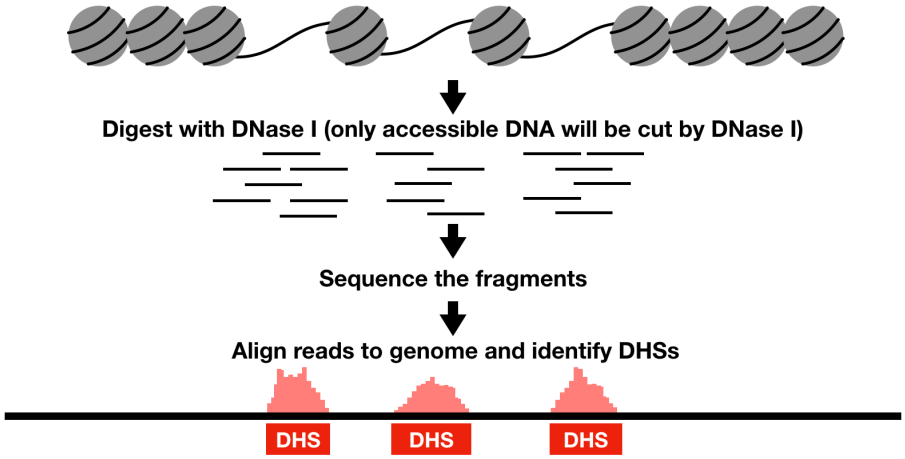
The tissue specific expression of CACNA1E, indicates that the promoter would only be a DHS in brain tissues and blood cells. Inaccessibility of the promoter in other tissues will prevent the gene from being expressed.



Therefore, identification of the locations of DHSs in different tissue types can aid in the understanding of gene regulation.

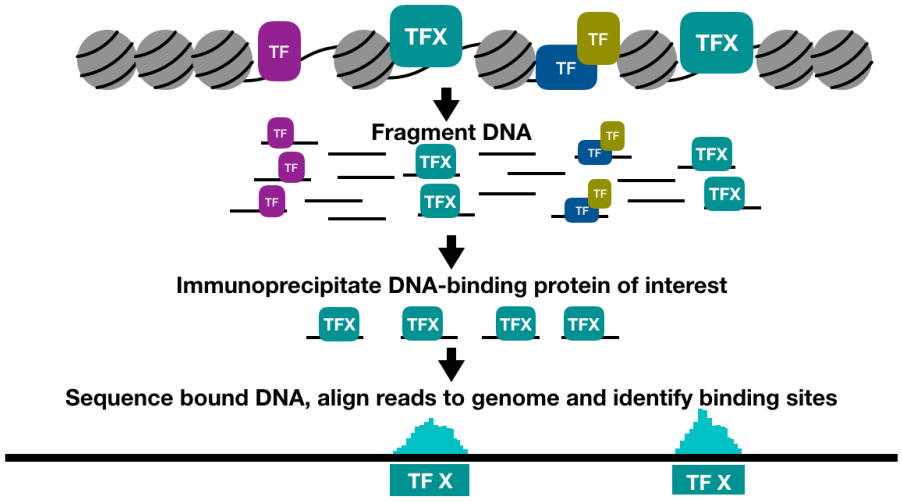
Genomic Assays

DNase-seq is an assay for the identification of DHS locations in the genome. The assay is performed by isolating DNA from a large sample of cells, and digesting it with DNase I. Only DNA that is accessible will be cut by DNase I. This means that accessible regions will be cut into short fragments. The short DNA fragments resulting from the digestion are then sequenced and aligned to the genome to determine where the DHSs are located.



The alignment of these reads produces **peaks**. Because many cells were used in the assay, the same genomic region can be represented in many reads. When these reads are aligned to the genome, fragments from the same region will "stack" on top of one another creating peaks. The height of the peak is the count of the number of stacked fragments. Statistical methods are used to identify the locations of peaks from the sequencing results. Each peak represents a DHS.

ChIP-seq (chromatin immunoprecipitation and sequencing) is an assay for the identification of DNA-binding protein binding site locations in the genome. ChIP-seq is performed for a specific DNA-binding protein, for example, a specific TF. In this example, the protein being assayed will be referred to as TFX. The assay is performed by crosslinking proteins and their bound DNA to stick them together. Then, DNA is extracted from a sample of millions of cells and fragmented. The protein of interest is then immunoprecipitated to isolate the DNA fragments that are bound by it. The protein is then degraded, and the DNA fragments are sequenced and aligned to the genome. ChIP-seq data produces peaks in the genome where the protein was bound to DNA, providing the genomic locations of the protein's binding sites.



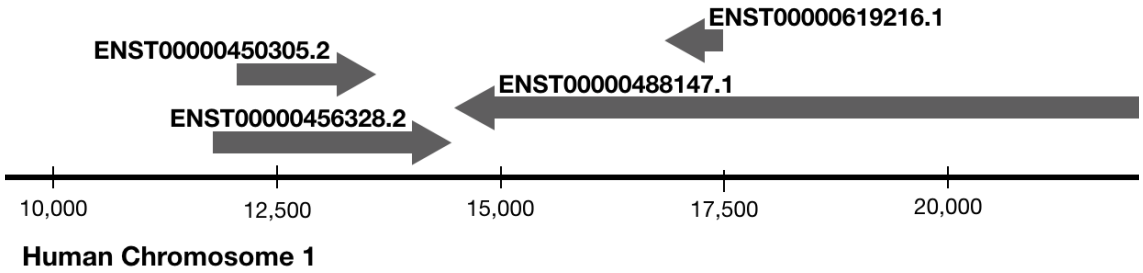
The ENCODE consortium is an effort to annotate the function DNA elements in the genome (https://www.encodeproject.org). This includes the identification of transcription factor

binding sites and DNase I hypersensitive sites. Currently, ENCODE has over 1,400 publicly available DNase-seq datasets across hundreds of tissue types and over 7,500 ChIP-seq datasets. Additionally, ENCODE hosts thousands publicly available datasets from other types of genomic assay (CLIP-seq, ATAC-seq, etc.).

10.1.2 BED Files

Genomic Coordinates

The following diagram shows the first four transcripts on human chromosome 1. Each arrow represents one gene.



The first transcript on chromosome 1 (ID: ENST00000456328.2) spans the chromosome from base 11,868 to base 14,409. The arrow representing this gene points from the 5' end to the 3' end, indicating that the transcript is on the sense strand (plus strand). Transcripts that are on the anti-sense strand (negative strand) are indicated with arrows pointing in the opposite direction.

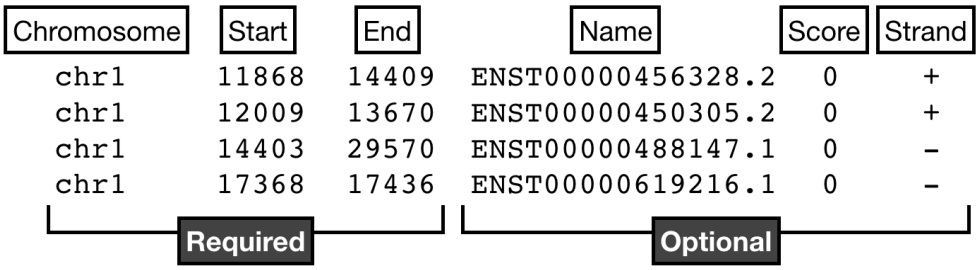
The locations, or coordinates, of a transcript in the genome can be described by indicating the chromosome it is one, the start position (5' end), the end position (3' end), and the strand it is on. These are the coordinates of the first 4 transcripts:

Name	Chromosome	Start position	End position	Strand
ENST00000456328.2	chr1	11868	14409	+
ENST00000450305.2	chr1	12009	13670	+
ENST00000488147.1	chr1	14403	29570	-
ENST00000619216.1	chr1	17368	17436	-

Genomic coordinates are stored in **BED format**. BED format is a tab-delimited format for storing the locations of genomic elements. Files storing data in BED format are called BED files and have the extension `.bed`.

A BED file is required to have at least 3 columns, in the following order: chromosome, start position, and end position. The start position is the base at the furthermost 5' end of the element, and the end position is the base at the furthermost 3' end of the element. Each line

in the file provides the coordinates for a specific region or element in the genome. Below the coordinates of the first 4 transcripts in on human chromosome 1 are shown in BED format.



Often there will be six columns in a BED file. The three optional columns contain the name of the element, a score for the element, and the strand the element is on. In this example, the name is a unique identifier for each transcript, the score is 0, and the strand indicates the strand the transcript is on in the genome.

The score column is included in the BED file, despite no scores being relevant to the information as the columns must be included in order. To include strand information, a score column must also be included. If a column is included but does not have information associated with it, the column will have a 0 or a period for each line as a place holder.

Genomic Assay BED Files

BED files are often used to store the results of genomic assays like the locations of DNase I hypersensitive sites, or transcription factor binding sites. This section of the course will explore working with BED files using DHSs from five human tissue samples: kidney, liver, skin, spleen, and testis. The coordinates of the DHSs are in the BED files: `kidney_DHS.bed`, `liver_DHS.bed`, `skin_DHS.bed`, `spleen_DHS.bed`, and `testis_DHS.bed`.

(Data downloaded from ENCODE. Accessions: ENCFF873DKJ, ENCFF236MEL, ENCFF247DDB, ENCFF546BTE, ENCFF885MEF)

```
j:~/Week.10/10.1.BEDTools$ ls *_DHS.bed
kidney_DHS.bed     skin_DHS.bed      testis_DHS.bed
liver_DHS.bed      spleen_DHS.bed
j:~/Week.10/10.1.BEDTools$ head -n 5 liver_DHS.bed
chr1    10036   10043
chr1    10072   10091
chr1    10158   10171
chr1    10578   10590
chr1    10626   10633
j:~/Week.10/10.1.BEDTools$ wc -l *_DHS.bed
  732642 kidney_DHS.bed
  315632 liver_DHS.bed
  306276 skin_DHS.bed
  643159 spleen_DHS.bed
  561651 testis_DHS.bed
 2559360 total
```

Viewing the first 5 lines of the `liver_DHS.bed` (see above) reveal that the BED file only contains the three required columns: chromosome, start, and end. Each line represents the location of a single DHS identified in the liver sample. The first five lines indicate that there is a DHS in the liver sample on chromosome 1 from base 10036 to base 10043, base 10072 to base 10091, base 10158 to base 10171, etc. The other four files share the same format. The number of DHSs found in each sample varies (see above), ranging from ~300,000 to ~700,000 DHSs per tissue sample.

One of the projects completed by ENCODE is a registry of candidate ***cis*-regulatory elements** in the human genome. *Cis*-regulatory elements are non-coding DNA elements that regulate transcription. Thus, both promoters and enhancers are *cis*-regulatory elements. By performing ChIP-seq on markers of promoters and enhancers and combining the output with DNase-seq data, locations of likely promoters and enhancers in the human genome were identified.

Candidate promoter sequences identified by ENCODE are in the file `ENCODE_promoters.bed`. (Data downloaded from UCSC Table Browser: hg38 ENCODE_cCREs, filtered for "PLS" (promoter-like sequence).)

```
j:~/Week.10/10.1.BEDTools$ 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     .
```

The BED file of promoters contains their location in the first 3 columns, a unique identifier in the name column, a score in the score column, and a place holder in the strand column. BED files that have results from genomic assays often contain scores, these can be:

- Z-scores: the number of standard deviations a peak's height is from the mean height.
- Fold-change: the number of times greater a peak's height is in the experiment than it is in a control.
- P-value: significance value from some other type of statistical test. Often the score is given as –log10(p-value), meaning the higher the score, the more significant.

The ENCODE candidate promoters score column contains Z-scores.

10.1.3 BEDTools Intersect

BEDTools

The BEDTools suite contains a set of tools for analyzing and manipulating BED files (https://bedtools.readthedocs.io/). BEDTools is hosted on Bioconda and can be installed using Conda on JupyterHub.

```
j:~$ conda install -c bioconda bedtools
```

Similarly to BLAST, BEDTools is composed of multiple commands. By running the command `bedtools`, a summary of the BEDTools subcommands can be viewed. The first 18 lines of the output are shown below.

```
j:~/Week.10/10.1.BEDTools$ bedtools
bedtools is a powerful toolset for genome arithmetic.

Version:    v2.30.0
About:      developed in the quinlanlab.org and by many contributors
worldwide.
Docs:       http://bedtools.readthedocs.io/
Code:       https://github.com/arq5x/bedtools2
Mail:       https://groups.google.com/forum/#!forum/bedtools-discuss

Usage:      bedtools <subcommand> [options]

The bedtools sub-commands include:

[ Genome arithmetic ]
    intersect     Find overlapping intervals in various ways.
    window        Find overlapping intervals within a window around an
interval.
    closest       Find the closest, potentially non-overlapping
interval.
    coverage      Compute the coverage over defined intervals.
    map           Apply a function to a column for each overlapping
interval.
                                    ...
```

BEDTools subcommands are with the following syntax:

```
bedtools subcommand subcommand_arguments_and_options
```

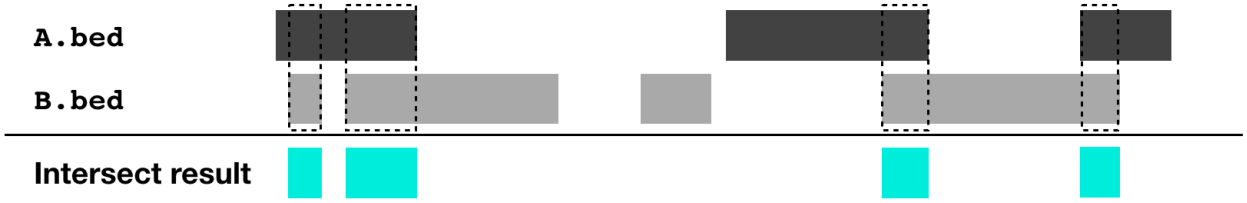
Arguments for running individual subcommands differs. A list of all BEDTools subcommands is found on the BEDTools website: (https://bedtools.readthedocs.io/en/latest/content/bedtools-suite.html). Each subcommand has a page with details on the subcommand's function, the arguments, and further information. In the command line, a help page can be pulled up by running: `bedtools subcommand -h`

Intersect

One BEDTools subcommand is `intersect`. It is an extremely useful command with many applications. The intersect command finds regions in one BED file that are overlapped by regions in another BED file, specifically it finds regions in `A.bed` that are overlapped by regions in `B.bed`.

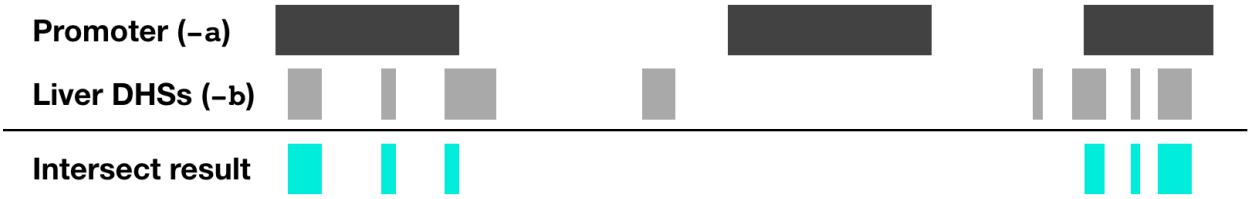
```
bedtools intersect -a A.bed -b B.bed
```

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



The output is a BED file containing the coordinates for each overlapping interval.

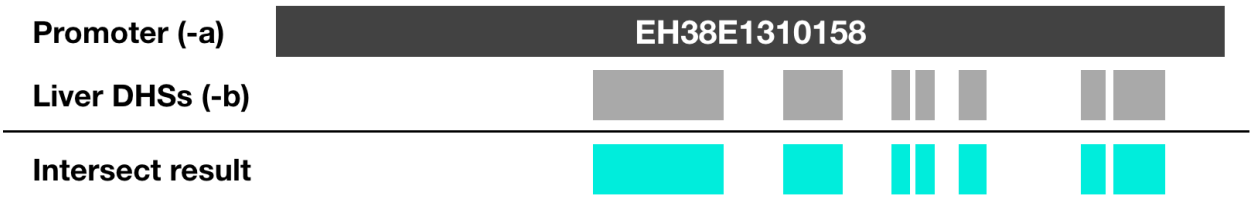
To learn which promoters are accessible in the human liver, BEDTools intersect can be applied to the promoter file (`ENCODE_promoters.bed`) and our file of liver DHSs (`liver_DHS.bed`) as illustrated below.



This command will return the coordinates of the promoter regions that are overlapped by accessible sites in the liver. When running BEDTools commands, the output should always be redirected to a new file with a name describing the output. Otherwise, the results will just output to the terminal. In this case, promoters accessible in liver will be returned, so the output is directed to a file named `liver_promoters.bed`.

```
j:~/Week.10/10.1.BEDTools$ bedtools intersect -a ENCODE_promoters.bed -b
liver_DHS.bed > liver_promoters.bed
j:~/Week.10/10.1.BEDTools$ head liver_promoters.bed
chr1    778679  778727  EH38E1310158    759     .
chr1    778749  778771  EH38E1310158    759     .
chr1    778789  778796  EH38E1310158    759     .
chr1    778798  778805  EH38E1310158    759     .
chr1    778814  778824  EH38E1310158    759     .
chr1    778859  778868  EH38E1310158    759     .
chr1    778871  778890  EH38E1310158    759     .
chr1    827466  827475  EH38E1310172    608     .
chr1    827527  827555  EH38E1310172    608     .
chr1    827580  827617  EH38E1310172    608     .
```

Each line in `liver_promoters.bed` contains one region of overlap between `ENCODE_promoters.bed` and `liver_DHS.bed`. Viewing the first 10 lines of the file, it is apparent that there are 7 liver DHSs that overlap the promoter with the ID EH38E131058 (see the name column). DHSs tend to be short as they are broken up by nucleosomes, whereas the promoter regions are much longer. Thus, there are many short overlaps within a promoter region:



The `intersect` subcommand has several options (flags) that can be applied to alter the output. One of which is the `-u` flag, which causes the command to return the **unique** elements in the `-a` file that are overlapped by `-b`. The same `intersect` command run with the `-u` option would only return the EH38E131058 promoter once, as is overlapped by at one or more liver DHS.

Re-running the same command with the `-u` option changes the output of the `intersect` command and thus alters the contents of the `liver_promoters.bed` file. Note that the `-u` option is applied directly after `intersect`.

```
j:~/Week.10/10.1.BEDTools$ bedtools intersect -u -a ENCODE_promoters.bed
-b liver_DHS.bed > liver_promoters.bed
j:~/Week.10/10.1.BEDTools$ head liver_promoters.bed
chr1    778562  778912  EH38E1310158    759     .
chr1    827342  827691  EH38E1310172    608     .
chr1    904594  904931  EH38E1310207    517     .
chr1    911145  911459  EH38E1310218    397     .
chr1    959147  959495  EH38E1310295    561     .
chr1    960392  960733  EH38E1310299    403     .
chr1    966292  966617  EH38E1310307    376     .
chr1    997988  998287  EH38E1310338    229     .
chr1    1000068 1000409 EH38E1310345    504     .
chr1    1013312 1013662 EH38E1310368    531     .
```

The file contents have changed. There is only one result for promoter region with the ID EH38E131058 and the start and end positions contain the full coordinates of the promoter.

The same command can be run for kidney, skin, spleen, and testis sample DHSs to determine the promoters that are accessible in these samples.

```
j:~/Week.10/10.1.BEDTools$ bedtools intersect -u -a ENCODE_promoters.bed
-b kidney_DHS.bed > kidney_promoters.bed
j:~/Week.10/10.1.BEDTools$ bedtools intersect -u -a ENCODE_promoters.bed
-b skin _DHS.bed > skin_promoters.bed
j:~/Week.10/10.1.BEDTools$ bedtools intersect -u -a ENCODE_promoters.bed
-b spleen_DHS.bed > spleen_promoters.bed
j:~/Week.10/10.1.BEDTools$ bedtools intersect -u -a ENCODE_promoters.bed
-b testis_DHS.bed > testis_promoters.bed
```

Looking back at the number of DHSs in each of the tissue samples, the numbers range across the tissue samples from a minimum of 306,276 in skin to 732,642 in kidney.

```
j:~/Week.10/10.1.BEDTools$ wc -l *_DHS.bed
  732642 kidney_DHS.bed
  315632 liver_DHS.bed
  306276 skin_DHS.bed
  643159 spleen_DHS.bed
  561651 testis_DHS.bed
 2559360 total
```

Checking the number of accessible promoters in each of the samples reveals a lower variance than the total number of DHSs per sample.

```
j:~/Week.10/10.1.BEDTools$ 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
```

Tissue samples range from 15,949 accessible promoters in skin to 20,572 in kidney. Note that this does not necessarily indicate that skin tissues always have fewer active promoters, and thus less transcribed genes, than kidney tissues. This data is from *individual samples* and is only representative of the sample used, not the tissue as whole.

Furthermore, experimental biases or technical effects may have altered the number of DHSs identified, meaning that the DHSs do not fully capture the biology. There may be more DHSs in the kidney sample than the skin sample for non-biological reasons; for example:

- Different scientists may have performed the experiments
- Samples can be of different quality
- One cell type may be easier to work with than the other

To learn if this is a real biological difference or if it is simply caused by the experiment, one would need to:

- perform multiple replicates of DNase-seq with each tissue sample
    - This is called a *technical* replicate
- perform multiple replicates of DNase-seq with more samples of the same tissue type
    - This is called a *biological* replicate

This does not mean that the data used in this lecture cannot be analyzed. It means that more data would be necessary to come to a strong conclusion about any results derived from these experiments.