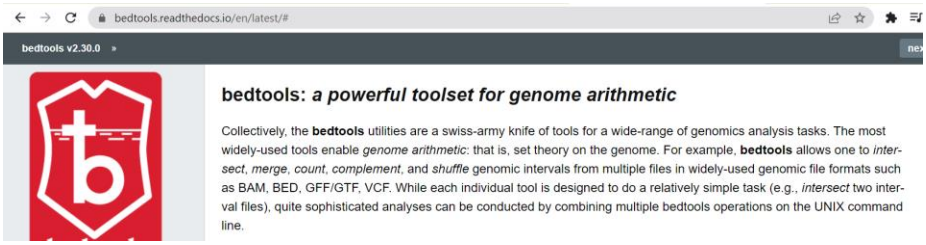
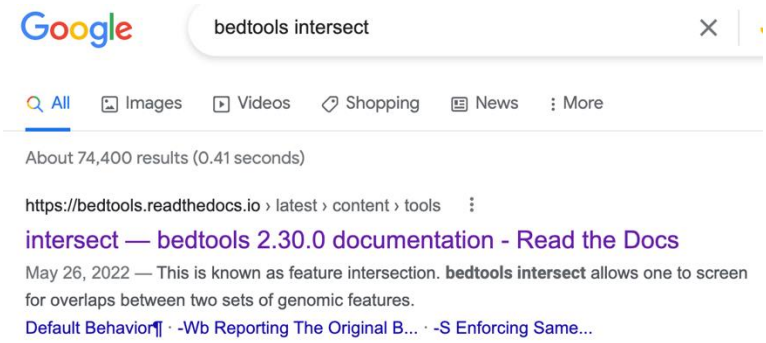
Bedtools commands Worksheet (DOWNLOAD IGV)

Author: Ariel Eraso, 2022

https://bedtools.readthedocs.io/en/latest/



Bedtools is great for doing genomic arithmetic between data sets (it even says so on their website, see below). It is also great because its very well documented. It even has a table of contents for you to peruse through. Still if you need to look for specific tools the easiest thing to do, like with most computational work, is to google.



See how easy that was! My google search took me to the website with the first link. So that's how you find any tool you need. Here I have googles bedtools intersect, one of the easiest to use So lets get started.

Important directories

Your bed files are located at **/scratch/Shares/dowell/public/sread2021/data_files/day9/**

Or are they?

How do you make sure this directory is not empty?

There are a couple of files here some end in summits.bed and some end in .narrowPeak. What is the difference? Will either set work for bedtools?

Genome annotations files are found at

/scratch/Shares/public/genomes/Homo_sapiens/NBI/RCh3/Anno/Jeans/Jeannes.gtfo
Maybe?

Loading bed tools

This is again using module load. If you don't know what version to use you can always type part of the name then hit tab to see your options.

```
[fiji-1:~$ module load bedtools
 bedtools          bedtools/2.23.0  bedtools/2.25.0  bedtools/2.28.0
[fiji-1:~$ module load bedtools/2.28.0
```

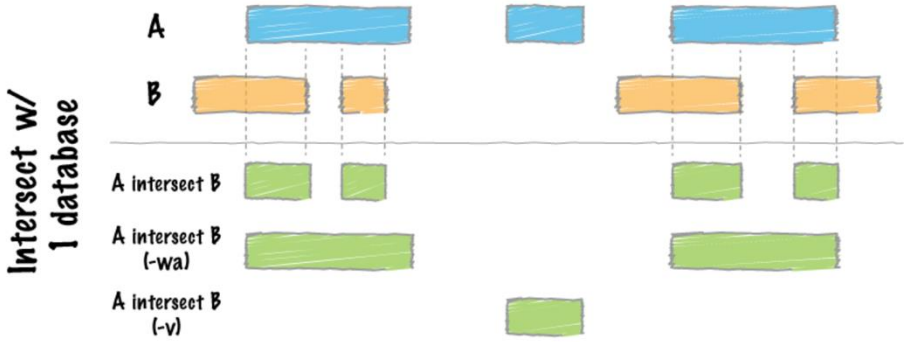
The general format for inputting bedtools commands is

```
fiji-1:~$ bedtools toolname -a file1 -b file2
```

Bedtools to call the program, followed by the tool, then you designate each file as either a or b so bedtools knows what files to use

Bedtools intersect

The most common bed-tool is bedtools intersect linked here:



https://bedtools.readthedocs.io/en/latest/content/tools/intersect.html Intersect will take file A (in blue below) with file B (orange below) and show you what parts of them overlap (green).

As a first pass to see how this works, try intersecting a bed file with a gene annotation file (found at). The result is a bedfile that reflects the regions in which there is overlap.

```
[fiji-1:~/bedtools_ALE$ bedtools intersect -a GABPA_summits.bed -b /scratch/Share]
 s/dowell/genomes/hg38/hg38_refseq.bed
 chr13   18212153        18212154        GABPA_peak_2    11.87998
 chr21   8208342 8208343 GABPA_peak_6    14.46615
 chr21   8208342 8208343 GABPA_peak_6    14.46615
 chr21   8209936 8209937 GABPA_peak_7    16.16573
 chr21   8209936 8209937 GABPA_peak_7    16.16573
 chr21   8209936 8209937 GABPA_peak_7    16.16573
 chr21   8209936 8209937 GABPA_peak_7    16.16573
 chr21   8209936 8209937 GABPA_peak_7    16.16573
 chr21   8209936 8209937 GABPA_peak_7    16.16573
 chr21   8209936 8209937 GABPA_peak_7    16.16573
 chr21   8254018 8254019 GABPA_peak_8    19.40786
 chr21   8392705 8392706 GABPA_peak_9    23.28121
 chr21   8392705 8392706 GABPA_peak_9    23.28121
 chr21   8392705 8392706 GABPA_peak_9    23.28121
 chr21   8392705 8392706 GABPA_peak_9    23.28121
 chr21   8392705 8392706 GABPA_peak_9    23.28121
```

The intersect between our bedfile, with the gene annotations should be extensive since these regions correspond to this genome. This gives quite a cumbersome output that is also hard to call if needed for downstream applications.

You can save your output using the > flag as so that it is more weildly

```
[fiji-1:~/bedtools_ALE$ bedtools intersect -a GABPA_summits.bed -b /scratch/Share]
 s/dowell/genomes/hg38/hg38_refseq.bed > GABPA_GENOME_int.bed
[fiji-1:~/bedtools_ALE$ lt                                                      ]
 total 24M
 -rw-rw-r-- 1 arer2562 4.6K Jun 28 15:43 GABPA_GENOME_int.bed
```

The file size is not zero this is a pretty good indication that intersect worked. To make sure file is not empty you can head your new file.

```
[fiji-1:~/bedtools_ALE$ head GABPA_GENOME_int.bed
 chr13   18212153        18212154        GABPA_peak_2    11.87998
 chr21   8208342 8208343 GABPA_peak_6    14.46615
 chr21   8208342 8208343 GABPA_peak_6    14.46615
 chr21   8209936 8209937 GABPA_peak_7    16.16573
 chr21   8209936 8209937 GABPA_peak_7    16.16573
 chr21   8209936 8209937 GABPA_peak_7    16.16573
 chr21   8209936 8209937 GABPA_peak_7    16.16573
 chr21   8209936 8209937 GABPA_peak_7    16.16573
 chr21   8209936 8209937 GABPA_peak_7    16.16573
 chr21   8209936 8209937 GABPA_peak_7    16.16573
```

Bedtools intersect -wa

The -wa flag in bedtools intersect produces an output that keeps the entire overlapping regions of your -A file as opposed to just the overlapping pieces.

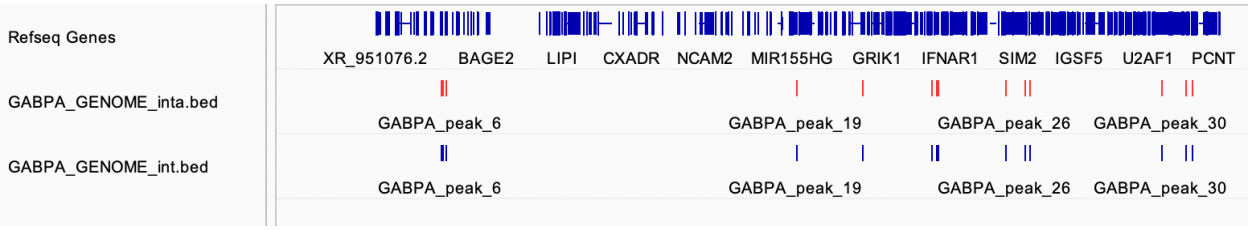
```
[fiji-2:~/bedtools_ALE$ bedtools intersect -wa -a GABPA_summits.bed -b /scratch/Shares/dowell/genomes/hg38/hg38_
refseq.bed > GABPA_GENOME_inta.bed
[fiji-2:~/bedtools_ALE$ head GABPA_GENOME_inta.bed
chr13   18212153        18212154        GABPA_peak_2    11.87998
chr21   8208342 8208343 GABPA_peak_6    14.46615
chr21   8208342 8208343 GABPA_peak_6    14.46615
chr21   8209936 8209937 GABPA_peak_7    16.16573
chr21   8209936 8209937 GABPA_peak_7    16.16573
chr21   8209936 8209937 GABPA_peak_7    16.16573
chr21   8209936 8209937 GABPA_peak_7    16.16573
chr21   8209936 8209937 GABPA_peak_7    16.16573
chr21   8209936 8209937 GABPA_peak_7    16.16573
chr21   8209936 8209937_GABPA_peak_7    16.16573
```

A simple head won't tell you if there is actually a difference between these two outputs.

For this you should use igv. If you haven't already got it on your computer it can be downloaded here: https://software.broadinstitute.org/software/igv/download

You can copy your files from remote to your computer. Obviously change your information for your serve but here is the command for fiji, a server which you are not using.

```
arieleraso@Ariels-MacBook-Pro ~ % rsync arer2562@fiji.colorado.edu:/Users/arer25
62/bedtools_ALE/GABPA_GENOME_int.bed /Users/arieleraso/Desktop/
```



What you see is that the genome is so expansive, and the GABA regions so small in comparison, that you don't really see a difference if you add the -wa flag (in red).

What happens if your two files don't have any overlap?

Bedtools intersect -v

Here I have intersected two summit files. As you remember, if you don't tell bedtools to save your output, it should print. Yet it didn't. Does this mean it does not work?

```
[fiji-1:~/bedtools_ALE$ bedtools intersect -a GABPA_summits.bed -b BACH1_summits.
bed
fiji-1:~/bedtools_ALE$
```

One easy way to check: plug this into a script and check your err files.

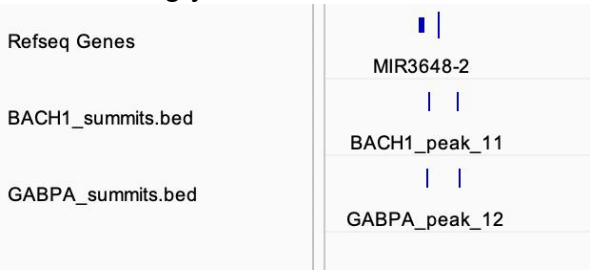
```
0 Jun 28 11:12 _Bedtools_commands_8405749.err
0 Jun 28 11:12 _Bedtools_commands_8405749.out
```

As you can see my files are empty. A quick head of my error file confirms its empty.

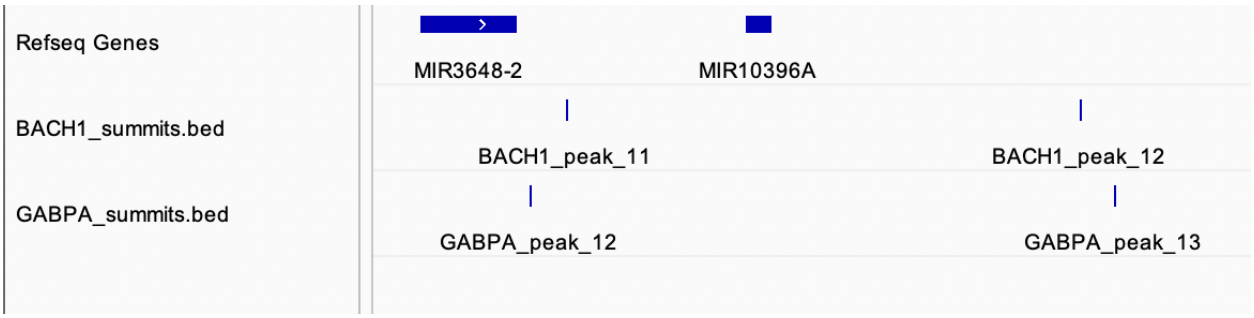
```
[fiji-2:/scratch/Users/arer2562/p53_machinelearning/eno$ head _Bedtools_commands_8405749.err
fiji-2:/scratch/Users/arer2562/p53_machinelearning/eno$
```

So what happened? Lets try looking at this in IGV. Rsync down both the input files as well as your saved output file (remember for this its just the same command as before followed by > newfile.bed

The first thing you'll notice is that the empty intersect files don't load onto igv. The second thing you'll notice is that the files seem to overlap quite a bit right?



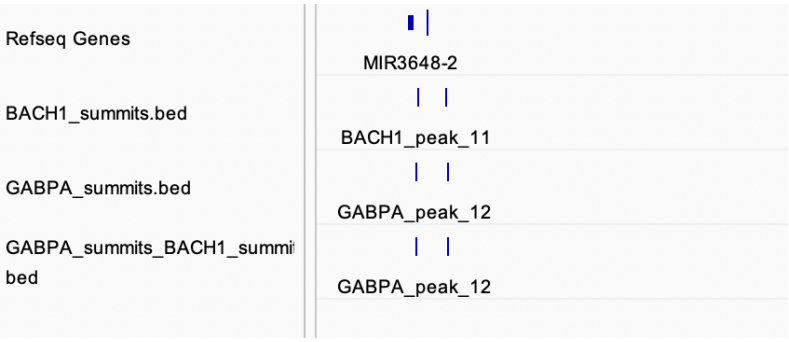
Zooming in more reveals that these two files don't overlap.



Summit files are only the single nucleotide containing the most signal from a region. That is what I have used for this. Hence the lack of overlap.

This makes for a great opportunity, however, to learn the -v flag.

This flag gives you the inverse of your output. So if you were to ask for intersects of the two summits, with the -v flag, you would get instead all of the places where they don't intersect.



As you can see this gives you all of the GABPA peaks. Based on this, do you think the empty intersect file was an err?

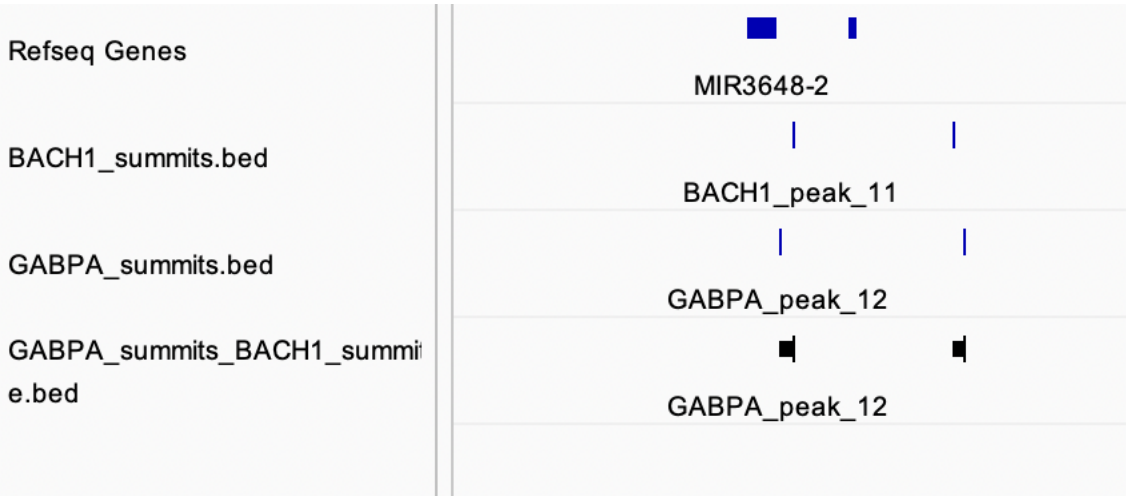
Bedtools closest

https://bedtools.readthedocs.io/en/latest/content/tools/closest.html

As you can guess, bedtools closest outputs the regions that are closest to overlapping. By default this includes overlapping regions. Closest can be a great complement to intersect, especially if your intersect file ends up empty

```
fiji-2:~/bedtools_ALE$ bedtools closest -a GABPA_summits.bed -b BACH1_summits.bed > GABPA_summits_BACH1_summits_close.bed
fiji-2:~/bedtools_ALE$ head GABPA_summits_BACH1_summits_close.bed
chr1    630936  630937  GABPA_peak_1    12.05997        chr1    630948  630949 BACH1_peak_1     39.30252
chr13   18212153        18212154        GABPA_peak_2    11.87998        .       -1      -1      .       -1
chr17   26885773        26885774        GABPA_peak_3    9.23636 .       -1      -1      .       -1
chr2    89840332        89840333        GABPA_peak_4    12.05997        .       -1      -1      .       -1
chr21   5128593 5128594 GABPA_peak_5    7.48780 chr21   8208253 8208254 BACH1_peak_3    13.19981
chr21   8208342 8208343 GABPA_peak_6    14.46615        chr21   8208253 8208254BACH1_peak_3     13.19981
chr21   8209936 8209937 GABPA_peak_7    16.16573        chr21   8209883 8209884BACH1_peak_4     35.57483
chr21   8254018 8254019 GABPA_peak_8    19.40786        chr21   8254021 8254022BACH1_peak_6     38.95876
chr21   8392705 8392706 GABPA_peak_9    23.28121        chr21   8392829 8392830BACH1_peak_8     43.21281
chr21   8435903 8435904 GABPA_peak_10   6.74743 chr21   8435623 8435624 BACH1_peak_9    10.33165
```

You can rsync this file as you would any other and throw it on IGV. The first hing you should notice, compared to the intersect file, is that it loads onto IGV.

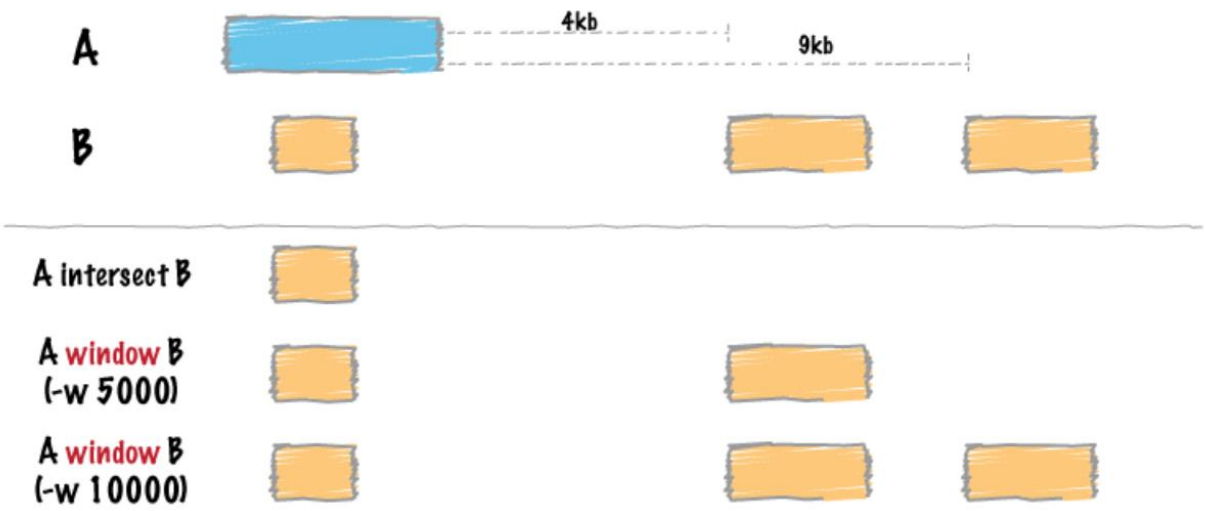


Closest results in a tophat representing the closest region. This is a helpful tool anytime you don't have overlaps.

Bedtools window

https://bedtools.readthedocs.io/en/latest/content/tools/window.html

Bedtools window searches for overlaps around regions in your files with an added window. This can be important when you want to know what is around your regions within a certain number of nucleotides. By default, the window added is 1000 bp.



This can be important when you want to know what is around your regions within a certain number of nucleotides. By default, the window added is 1000 bp.

```
[fiji-1:~/bedtools_ALE$ bedtools window  -a GABPA_summits.bed -b BACH1_summits.bed > GABPA_summits_BACH1_summits_win.bed

[fiji-1:~/bedtools_ALE$ head GABPA_summits_BACH1_summits_win.bed
chr1    630936  630937  GABPA_peak_1    12.05997        chr1    630948  630949  BACH1_peak_1    39.30252
chr21   8208342 8208343 GABPA_peak_6    14.46615        chr21   8208253 8208254 BACH1_peak_3    13.19981
chr21   8209936 8209937 GABPA_peak_7    16.16573        chr21   8209883 8209884 BACH1_peak_4    35.57483
chr21   8254018 8254019 GABPA_peak_8    19.40786        chr21   8253106 8253107 BACH1_peak_5    15.34869
chr21   8254018 8254019 GABPA_peak_8    19.40786        chr21   8254021 8254022 BACH1_peak_6    38.95876
chr21   8392705 8392706 GABPA_peak_9    23.28121        chr21   8392829 8392830 BACH1_peak_8    43.21281
chr21   8435903 8435904 GABPA_peak_10   6.74743 chr21   8435623 8435624 BACH1_peak_9    10.33165
chr21   8437112 8437113 GABPA_peak_11   17.01971        chr21   8437089 8437090 BACH1_peak_10   24.62233
chr21   8987203 8987204 GABPA_peak_12   7.41037 chr21   8987272 8987273 BACH1_peak_11   17.18604
chr21   8988316 8988317 GABPA_peak_13   24.13173        chr21   8988250 8988251 BACH1_peak_12   27.18232
```

On IGV the output looks like this.

As you can see, it looks a lot like the output for closest. The real power of window is that it allows you to edit the window in which you are looking.

Bedtools window -w

To use a window differing from the 1000 bp default, simply add the -w (w stands for window) and specify a window size.

```
[fiji-1:~/bedtools_ALE$ bedtools window -a GABPA_summits.bed -b BACH1_summits.bed -w 5000 > GABPA_summits_BACH1_summits_
win5000.bed
[fiji-1:~/bedtools_ALE$ head GABPA_summits_BACH1_summits_win5000.bed
chr1    630936  630937  GABPA_peak_1    12.05997        chr1    630948  630949  BACH1_peak_1    39.30252
chr21   8208342 8208343 GABPA_peak_6    14.46615        chr21   8208253 8208254 BACH1_peak_3    13.19981
chr21   8208342 8208343 GABPA_peak_6    14.46615        chr21   8209883 8209884 BACH1_peak_4    35.57483
chr21   8209936 8209937 GABPA_peak_7    16.16573        chr21   8208253 8208254 BACH1_peak_3    13.19981
chr21   8209936 8209937 GABPA_peak_7    16.16573        chr21   8209883 8209884 BACH1_peak_4    35.57483
chr21   8254018 8254019 GABPA_peak_8    19.40786        chr21   8253106 8253107 BACH1_peak_5    15.34869
chr21   8254018 8254019 GABPA_peak_8    19.40786        chr21   8254021 8254022 BACH1_peak_6    38.95876
chr21   8392705 8392706 GABPA_peak_9    23.28121        chr21   8391286 8391287 BACH1_peak_7    13.65037
chr21   8392705 8392706 GABPA_peak_9    23.28121        chr21   8392829 8392830 BACH1_peak_8    43.21281
chr21   8435903 8435904 GABPA_peak_10   6.74743 chr21   8435623 8435624 BACH1_peak_9    10.33165
[fiji-1:~/bedtools_ALE$ wc -l GABPA_summits_BACH1_summits_win.bed
```

The overlap has now been extended up til the nearest 5000 bp. Does this appear different from the previous 1000 bp window?

Besides IGV how could you make sure there are actually more overlaps?

Bedtools shuffle

https://bedtools.readthedocs.io/en/latest/content/tools/shuffle.html

Suppose you did all of your intersects, bedtools closest, bedtools window, etc. and found that there was large overlap between the two data sets. Firstly, congrats. Second, you have to make sure this is real overlap and not just chance. For this you can rearrange your peaks with bedtools shuffle and then check for intersects. As an example lets use bedtools intersect on our summits. Given that the regions are so small you would expect this shuffle to still not yield any overlaps. Well that depends on the file.

```
fiji-1:~/bedtools_ALE$ bedtools shuffle -i GABPA_summits.bed -g /scratch/Shares/dowell/genomes/hg38/hg38_refseq.bed > G
ABPA_shuffle.bed
fiji-1:~/bedtools_ALE$ head GABPA_shuffle.bed
chr1    2237970 2237971 GABPA_peak_1    12.05997
chr1    2230921 2230922 GABPA_peak_2    11.87998
chr1    1759174 1759175 GABPA_peak_3    9.23636
chr1    974694  974695  GABPA_peak_4    12.05997
chr1    3000690 3000691 GABPA_peak_5    7.48780
chr1    552456  552457  GABPA_peak_6    14.46615
chr1    427121  427122  GABPA_peak_7    16.16573
chr1    1237128 1237129 GABPA_peak_8    19.40786
chr1    2696075 2696076 GABPA_peak_9    23.28121
chr1    1588334 1588335 GABPA_peak_10   6.74743
```

The flags are a little different here, you might notice. Instead of using minus -a and -b we use -i and -g to designate our input files. Here -i stands for input and -g is for genome. In this case, since our data is mapped onto hg38, we used hg38 for our reference genome.

As you can see, shuffling has shifted our GABPA peaks. So GABAPA peak 6 is gone! Scroll through you igv track to find it again. Here is mine try to find yours.

What you will notice is that when overlapping with the genome, even after shuffling, the number of overlaps is about the same. Well that's because the genome is so large. Your tiny summit peaks are bound to overlap with the genome just by random chance. Which is what I was getting at before.

I showed you bedtools shuffle and you saw that our shuffles were different. That's because everytime you use bedtools shuffle your results will be different. Unless you seed your results.

Bedtools shuffle -seed

Bedtools shuffles based on a randomly generated number. This number is generated once again everytime you shuffle. The vastness of math makes it difficult to recreate the same number twice via random number generation. The -seed flag allows you to manually input the number, rather than generating it. Thus, if you use the same number your shuffle with always be the same.

```
fiji-1:~/bedtools_ALE$ bedtools shuffle -i GABPA_summits.bed -g /scratch/Shares/dowell/genomes/hg38/hg38_refseq.bed -se
ed 305 > GABPA_shuffle_305.bed
```

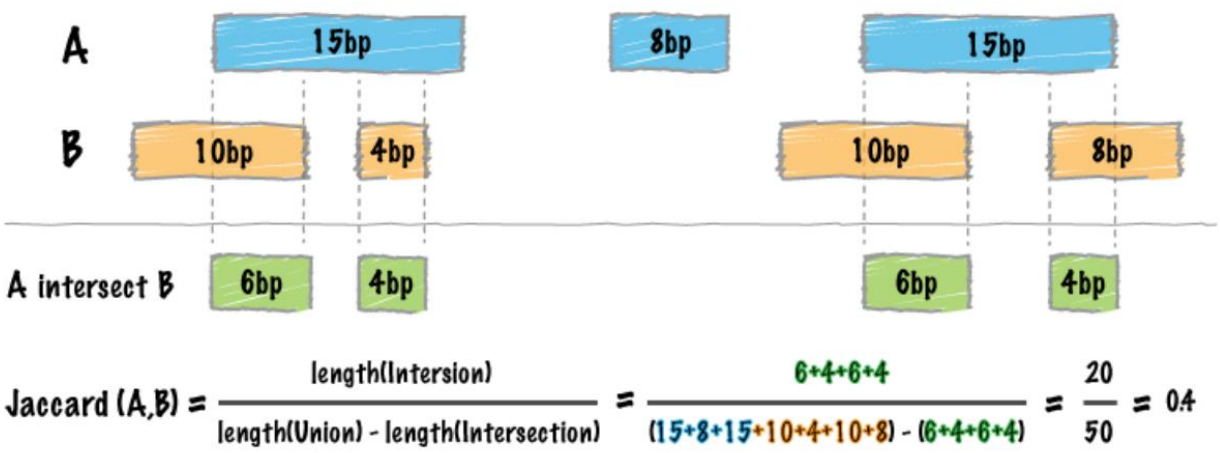
I shuffled seeding with 305 because I was raised in Miami #dale. If you line this up with your previous seed yo can see that they are different.

Now time for some partner work!

Talk to the person nearest to you or whoever you make eye contact with first. Ask them what number they seeded with. Recreate their seeded shuffle. Do your igv tracks match?

Jaccard

https://bioweb.pasteur.fr/docs/modules/bedtools/2.25.0/content/tools/jaccard.html



Jaccard is used when you want to quantify the ratio between regions of your two data sets that overlap vs regions that do.

For this we go back to -a and -b file flags and our output is a .txt file

As you can see, literally all of your GABPA peaks overlap. Likely by chance because your genome is such a loarge thing to overlap with.

```
[fiji-1:~/bedtools_ALE$ bedtools jaccard -a GABPA_summits.bed -b /scratch/Shares/dowell/genomes/hg38/hg38_refseq.bed > G
ABPA_genome_Jack.txt
[fiji-1:~/bedtools_ALE$ head GABPA_genome_Jack.txt
intersection    union   jaccard n_intersections
24      1439997401      1.66667e-08     24
```

If you jaccard with the BACH file you see the opposite.

```
[fiji-1:~/bedtools_ALE$ head GABPA_summits_BACH1_summits_jack.txt
intersection    union   jaccard n_intersections
0       74      0       0
```

There is a great way to check if jaccard results are due to chance. Shuffle your file (as you conveniently have) and then re-jaccard.

Note there is a possibility you get an error due to chromosome regions being out of order. This can be fixed by first sorting your files.

Note

The `jaccard` tool requires that your data is pre-sorted by chromosome and then by start position (e.g., `sort -k1,1 -k2,2n in.bed > in.sorted.bed` for BED files).

If you shuffle your GABPA peaks and then jaccard to your genome file you'll get

```
fiji-1:~/bedtools_ALE$ head GABPA_genome_sorted_305_Jack.txt
intersection    union   jaccard n_intersections
18      1439997407      1.25e-08        18
```

```
[fiji-1:~/bedtools_ALE$ head GABPA_genome_sorted_666_Jack.txt
intersection    union   jaccard n_intersections
22      1439997403      1.52778e-08     22
```

Numbers fairly similar to our previous experiences!