Bedtools commands HW

Author: Ariel Eraso, 2022

Intersect

I showed you how to intersect the GABPA and BACH summits and we saw 0 overlap. What happens if you intersect the GABPA Narrow peaks to the BACH Narrow Peaks?

What is different about the summits and Narrow Peaks that leads to differences in intersects?

How does your output change if you use the -wa flag?

How would you keep only the sections of intersect B?

Does this flag change the number of overlaps?

How would you figure this out?

What is the third region of intersection between these two files, what is the 10th?

How could you use bedtools intersect to determine what regions in your GAPBA narrow peak files DON'T overlap with the BACH peaks?

What does this look like on IGV

How many of these entries are unique intersections (hint it involves the unique command)

Closest

We tried bedtools closest with the summmits Now what happens if you use the narrow peak files?

How would you the closest non-overlapping regions?

What is the 3rd non-overlapping region

How would you repeat this if you had to use multiple files?

What does this look like on IGV?

Window

For this you will get different results with either set of files (summits or narrow peaks) so choose your favorite.

Using bedtools intersect we learned to use -wa to get an output reporting which regions in A overlap with B and then used -v to determine which regions of A don't overlap with B. What corresponding flags in bedtools window do the same thing?

How do this flags change the output using the default output?

What flag would you use to add 500 bp only to the upstream side of the window?

What flag would you use to add 500bp downstream to the window?

Put these output files on IGV. How do these flags change your outputs?

Shuffle

Sometimes you want to shuffle your data before testing it in jaccard to see if shuffled data will give the same results. How do you sort shuffled data, so it fits into jaccard?

Most of our shuffling was performed on the summit files. Since each region is one nucleotide long, there isn’t a large chance of your regions getting shuffled onto each other.

Does the same hold true for narrow peak files why or why not?

What flag could you use to avoid this possibility?

Perform seeded shuffles of you narrow peak files: one with the flag that excludes overlaps (same as q above) and one without. Does this change your output?

What is the flag to allow your shuffle to place data beyond the end of your chromosome?

Does jaccard accept input files created using this flag? If you got sequencing data that mapped to regions beyond the ends of the chromosomes, would you trust it? Why or why not? What does this highlight about the process of getting good results from computational work?

Jaccard

You now have made many shuffled files to go along with your narrow peaks. For this exercise feel free to use any of the narrow peaks files.

The default overlap require for a positive hit with jaccard is 1bp.

What flag would you use if you wanted to get back only positive hits if they overlap with at least 20% of your A file?

How does this affect your false positive rate? How does it affect your false negative rate?

What flag would you use to get back only positive hits if they overlap with at least 20% of your B file?