03-713: Bioinformatics Data Integration Practicum
Spring 2025
Project Description

The project for this course will enable students to begin to address the following questions:

- Is transcriptional regulatory element activity more conserved across tissues or species?
  - To what extent does transcriptional regulatory element conservation across tissues and species differ between enhancers and promoters?
  - To what extent does the transcriptional regulatory code differ between tissues and species?
- To what extent does the transcriptional regulatory code differ between enhancers and promoters?
  - Does this depend on tissue?
  - Does this depend on species?
- To what extent are the biological processes upregulated in tissue conserved across species?

While working on this project, students should gain the following skills:

- Use a computer cluster
- Use github
- Evaluate the quality of open chromatin data
- Manipulate genomics files
- Map open chromatin regions across species
- Find biological processes that are likely to be regulated by transcriptional regulatory elements
- Find sequence patterns that are over-represented in genomic sequences of interest

Specific project description: Every team will be given a human and a mouse open chromatin dataset from three different tissues. Teams should complete the following tasks with the datasets and include analyses of results for each task in the project report:

1. Evaluate the quality of the datasets. **This needs to be completed for all the datasets. The remaining tasks can be completed using only the datasets from the two tissues with the highest-quality datasets (though you are welcome to do them for all datasets and adjust your quality evaluations based on the results).**
2. Map the mouse open chromatin regions to the human genome and map the human open chromatin regions to the mouse genome.
   a. Identify the open chromatin regions in each species whose orthologs in the other species are open and those whose orthologs in the other species are closed.
3. Find the open chromatin regions in each species that are open in both tissues and those that are open in only one tissue. Then answer the following question:

a. For each species, tissue combination, are more open chromatin regions open in the other tissue or the other species?

4. Find candidate biological processes regulated by the open chromatin regions in:
    a. The open chromatin regions from each species, tissue combination.
    b. The open chromatin regions that are shared across tissues for each species.
    c. The open chromatin regions that are specific to each tissue for each species.
    d. The open chromatin regions that are shared across species for each tissue.
    e. The open chromatin regions that are specific to each tissue for each species.
5. Divide the open chromatin data into likely enhancers and likely promoters. You may have a third category of regions that could be promoters or enhancers that you can choose not to use for the remaining parts of the project. Then answer the following questions:
    a. For each species, tissue combination, how does the percentage of enhancers compare to the percentage of promoters that are shared across tissues?
    b. For each species, tissue combination, how does the percentage of enhancers compare to the percentage of promoters that are shared across species?
6. Find sequence patterns that occur more than expected by chance in each of the following sets of peaks:
    a. Full set of peaks for each species, tissue combination.
    b. Enhancers for each species, tissue combination.
    c. Promoters for each species, tissue combination.
    d. Enhancers that are shared across tissues in each species.
    e. Enhancers that are specific to each tissue in each species.
    f. Enhancers that are shared across species for each tissue.
    g. Enhancers that are specific to each species for each tissue.

The following resources and tools may be helpful for completing this project (Note that some tools have been installed as modules on Bridges-2.):

- Pittsburgh Supercomputing Center user guide and tutorial (https://www.psc.edu/resources/bridges-2/user-guide/, https://www.wavlab.org/activities/2022/psc-usage/)
- github guide (https://docs.github.com/en/get-started)
- HALPER (https://github.com/pfenninglab/halLiftover-postprocessing, https://pubmed.ncbi.nlm.nih.gov/32407523/)
- bedtools (https://bedtools.readthedocs.io/en/latest/, https://pubmed.ncbi.nlm.nih.gov/20110278/)
- GREAT (https://great.stanford.edu/great/public/html/, https://pubmed.ncbi.nlm.nih.gov/20436461/)
- MEME suite (https://meme-suite.org/meme/, https://meme-suite.org/meme/doc/meme-chip.html, https://pubmed.ncbi.nlm.nih.gov/21486936/)