Exercises

Consider the five regions listed in *short_list.bed* and the ChIP-seq peaks in *K562_CTCF_CTCF_ENCFF002CEL_chr15.bed*.

- Which of the five regions in *short_list.bed* overlaps with the *least* number of ChIP-seq peaks?
- What percentage of the first region in *short_list.bed* overlaps with more than one ChIP-seq peak? What percentage of the second region overlaps with more than one ChIP-seq peak?
- At what basepair does the first region in *short_list.bed* transition from overlapping two ChIP-seq peaks to overlapping only one?

Exercise

Consider the ChIP-seq peaks in
K562_CTCF_CTCF_ENCFF002CEL_chr15.bed and
K562_CTCF_CTCF_ENCFF002DBD_chr15.bed

- What percentage of chromosome 15 overlaps at least one ChIP-seq peak for each file?
- How many basepairs of chromosome 15 overlap exactly one ChIP-seq peak for each file? What percentage of chr15 is this for each file?
- Do any of the first 20 basepairs of chr15 overlap with any ChIP-seq peaks in either file?

Exercise

“We used all 711 VISTA [mouse mm10] enhancers as positive training data, and for negative training data, we created a set of 711 random regions matched to the length and chromosome distribution of the positives to represent the genomic background.”

- Given the 711 VISTA positive regions (vista.bed) and the mouse mm10 assembly genome (mm10.genome), how would you generate the list of negatives described in this methods section excerpt?
- How would you generate the same list of negatives if you wanted to make sure none overlapped with the list of known coding genes in mm10.coding.bed?

Pollard et al. *Integrating Diverse Datasets Improves Developmental Enhancer Prediction*. DOI: 10.1371/journal.pcbi.1003677