A Comparative Genomics Approach to Identifying Candidate Enhancers Associated with Phenotypes

Daniel E. Schaffer

Advisors: Dr. Irene Kaplow and Prof. Andreas Pfenning

Computational Biology Department, Carnegie Mellon University

0. Abstract

Mammals are substantially different from each other across a large number of phenotypes. Many observed phenotypic differences are likely due to differences in the activity of enhancers, short regions of the genome involved in regulating the expression of nearby genes. Studying enhancer evolution is difficult because enhancer activity tends to be cell type-specific, and we cannot obtain most tissues and cell types from most species. We can obtain experimentally assayed open chromatin regions (OCRs), which are a proxy for enhancers, in tissues and cell types of a few species and use them to train machine learning models for predicting open chromatin. Predicting open chromatin enables us to evaluate whether the activity or inactivity of an OCR across species may be associated with the presence or absence of a phenotype. We can then identify orthologs of OCRs in a large, diverse set of mammalian genomes recently generated by the Zoonomia Project and predict whether those orthologs have open chromatin. I used the predicted open chromatin to develop a method for associating OCRs with specific phenotypes. Applying this method to neurological phenotypes using motor cortex and parvalbumin neuron open chromatin revealed dozens of new OCR-phenotype associations. Many associated OCRs were close, both in the linear genome and in the 3D conformation of chromatin, to relevant genes, including brain size-associated OCRs near genes mutated in microcephaly or macrocephaly.