Applying machine learning prediction of regulatory elements to identify trait-associated loci within the human population and across species.

Irene M. Kaplow[1,6], Daniel E. Schäffer[1], BaDoi N. Phan[1,2], Alyssa J. Lawler[3,4], Jing He[5], Amanda Kowalczyk[1,6], Ashley R. Brown[1], Chaitanya Srinivasan[1], Morgan E. Wirthlin[1,6], William R. Stauffer[5], Andreas R. Pfenning[1,6]*

Affiliations:

[1]Computational Biology Department, School of Computer Science, Carnegie Mellon University; Pittsburgh, PA, USA.

[2]Medical Scientist Training Program, School of Medicine, University of Pittsburgh; Pittsburgh, PA, USA.

[3]Department of Biological Sciences, Mellon College of Science, Carnegie Mellon University; Pittsburgh, PA, USA.

[4]Current Affiliation: Broad Institute of Harvard and MIT; Cambridge, MA, USA.

[5]Neurobiology Department, School of Medicine, University of Pittsburgh; Pittsburgh, PA, USA.

[6]Neuroscience Institute, Carnegie Mellon University; Pittsburgh, PA, USA.

*Corresponding author. Email: apfenning@cmu.edu

Genome conservation is a powerful tool to annotate new genomes, prioritize trait-associated genetic variants within a population, and to link differences in traits across species to differences in selective pressures. The vast majority of computation methods to infer conservation rely on the alignment of individual nucleotide sequences. While these approaches work well for many protein sequences and highly conserved non-coding regions, they fail at the vast majority of enhancers. These distal regulatory elements are often conserved in their cell type- and tissue-specific function, even when nucleotide conservation is low. To overcome this limitation, we developed the TACIT (Tissue Aware Conservation Inference Toolkit) approach, in which machine learning models learn the regulatory code connecting genome sequence to tissue-specific open chromatin, allowing us to accurately predict cases where differences in genotype are associated with differences in tissue-specific open chromatin at enhancer regions. We trained machine convolutional neural network models on brain and liver open chromatin from the Pfenning laboratory and from the FAANG project to predict open chromatin across the genomes of 222 mammals. Within the human population, we demonstrate that conserved neural cell type-specific open chromatin provides a substantial boost to prioritizing and interpreting disease-associated genetic variants. Across mammals, we identify a number of cases where predicted brain open chromatin is associated with the species' brain size. These loci show a tendency to be near genes associated with microcephaly and macrocephaly within the human population.