Speaker: Katherine Pollard

Title:
Sequence-structure-function modeling for DNA

Abstract:
The human genome sequence folds in three dimensions (3D) into a rich variety of locus-specific contact patterns. Despite growing appreciation for the importance of 3D genome folding in evolution and disease, we lack models for relating mutations in genome sequences to changes in genome structure and function. Towards that goal, we discovered that the organization of gene regulatory domains within chromosomes and the specific sequences that sit at boundaries between domains are under strong negative selection in the human population and over primate evolution. Motivated by this signature of functional importance, we developed a deep convolutional neural network, called Akita, that accurately predicts genome folding from DNA sequence alone. Representations learned by Akita underscore the importance of the structural protein CTCF but also reveal a complex grammar beyond CTCF binding sites that underlies genome folding. Akita enabled rapid in silico predictions for effects of sequence mutagenesis on the 3D genome, including differences in genome folding across species and in disease cohorts, which we are validating with CRISPR-edited genomes. This prediction-first strategy exemplifies my vision for a more proactive, rather than reactive, role for data science in biomedical research.