# DECIPHERING THE DETERMINANTS OF MECHANISTIC VARIATION IN REGULATORY SEQUENCES

Evan Seitz, David McCandlish, Justin Kinney, Peter Koo

Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology, Cold Spring Harbor, NY

Regulatory DNA sequences encode rules that orchestrate interactions with transcription factors (TFs) to control gene expression. Deep neural networks (DNNs) trained to predict functional genomics outputs from DNA sequences have shown promise in elucidating the genetic basis of molecular functions like enhancer activity and transcription initiation. Post hoc attribution analysis helps interpret cis-regulatory mechanisms underlying DNN predictions by generating attribution maps, which assign importance scores to each nucleotide. These maps often reveal patterns, called motifs, characteristic of specific regulatory elements like TF binding sites (TFBSs). The arrangement of motifs scaled to their importance in a map suggests a distinct cis-regulatory mechanism.

While attribution analysis has aided in deciphering complex motif syntax, attribution maps can be challenging to interpret due to seemingly spurious scores that obscure motifs. Moreover, attribution maps from different regulatory sequences–even those differing by a single nucleotide–show a wide diversity of cis-regulatory mechanisms. As a result, it remains unclear what mechanisms are evolutionarily poised at a regulatory sequence and which sequence determinants control them.

Here we introduce SEAM (Systematic Explanation of Attribution-derived Mechanisms), an explainable AI framework that reveals cis-regulatory mechanistic variation within localized sequence space. SEAM generates a library of sequences through partial random mutagenesis around a sequence of interest, computes attribution maps for each, and clusters the maps to uncover the diversity of mechanisms poised in local sequence space. Averaging maps within each cluster reduces noise, leading to clearer mechanistic insights. By analyzing the sequences associated with each cluster, SEAM identifies the sequence determinants driving mechanistic variation and separates motifs from context-specific background, providing refined mechanistic explanations.

Applying SEAM to several genomic DNNs–including ChromBPNet, ProCapNet, CLIPNET, DeepSTARR, and DeepMel2–we observed that regulatory DNA sequences sample a rich repertoire of distinct mechanisms from just a few specific mutations. Many variants alter mechanisms by disrupting or optimizing existing TFBSs and creating de novo sites, revealing a spectrum of mechanisms that finely tune gene expression. To explore SEAM's extensibility, we examined different libraries to uncover mechanisms that generalize across sequence contexts, and combinatorially-complete variant effects measured experimentally to reveal alternative TF binding modes. SEAM offers a powerful framework for studying regulatory mechanisms across populations and species, and provides a new axis for mechanism-informed sequence design to precisely modulate gene regulatory networks.