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

Non-coding DNA sequences regulate gene expression by integrating transcription factor (TF) binding sites with broader sequence context, collectively forming the cis-regulatory code that governs cellular function. Deciphering this code is essential for understanding how genetic variation rewires regulatory programs, contributes to disease, and drives evolutionary innovation. Deep learning models trained on sequence-function relationships have advanced the prediction of regulatory activity by inherently learning biological mechanisms, such as the importance of TF binding sites, that attribution methods capture at the level of individual sequences.

However, attribution methods fall short of distilling the sequence rules that encode specific mechanisms across mutational landscapes. This creates a critical gap: while we can predict how individual mutations affect regulatory activity and mechanism, we lack a comprehensive framework to decipher sequence-mechanism relationships. Understanding these relationships would allow us to uncover the underlying sequence features and fundamental attribution signals that define a given mechanism and reveal the pathways by which genetic variation influences gene regulation.

Here, we present SEAM, an explainable AI framework that systematically probes mutational effects on attribution maps to decipher sequence-mechanism relationships. SEAM reveals the remarkable flexibility of regulatory sequences, demonstrating how specific mutations can reprogram TF binding, reshape context-dependent interactions, and drive functional diversity. By additionally disentangling motif-driven mechanisms from broader contextual features, SEAM offers unprecedented insights into the modular and adaptive architecture of the cis-regulatory code. Beyond prediction, SEAM provides a versatile framework for prioritizing disease-associated variants, mapping regulatory evolution, and engineering synthetic DNA sequences with tailored functions, establishing a new paradigm for understanding how genetic variation shapes regulatory complexity.