# A098 · RegScan: A Statistical Framework for Identifying Functional Transcription Factor Binding Sites from Per-Nucleotide Importance Scores

**Presenter:** Zain M. Patel — Mass General Hospital, Harvard Medical School, Broad Institute

**Authors:** Zain M. Patel, Faheem A. Azeemi, Timothy Barry, Basheer Becerra, Daniel E. Bauer, Luca Pinello

**Session:** Day 2 · Fri Oct 2, 2026 · 2:15–4:15 PM

High-throughput genomics approaches, including massively parallel reporter assays (MPRAs) and CRISPR-based perturbation screens, have expanded dissection of cis-regulatory elements at nucleotide resolution. Analyses of these experiments, together with attribution and in-silico mutagenesis applied to deep-learning models, can generate per-nucleotide importance scores that quantify how sequence perturbations affect regulatory activity. However, identifying the transcription factors (TFs) underlying these effects still predominantly relies on scanning reference DNA with TF position weight matrices (PWMs), an approach that does not incorporate functional nucleotide preferences revealed by perturbation experiments. We developed RegScan, a statistical framework that uses TF PWMs to scan nucleotide-resolution importance scores and prioritize TFs whose sequence preferences are consistent with observed regulatory effects. At each candidate position, RegScan transforms the PWM relative to the reference sequence and quantifies agreement with importance scores using cosine similarity. Statistical significance is assessed against a null distribution derived from shuffled importance scores, followed by multiple-testing correction. We evaluated RegScan across three sources of nucleotide-resolution functional information: saturation-mutagenesis MPRAs, CRISPR-based regulatory tiling screens, and deep-learning-derived importance scores. In MPRA benchmarks with validated TF binding sites, RegScan improved TF rank in four of five whole-element analyses relative to sequence-only PWM scanning. When restricted to mutation-sensitive regions, RegScan matched or outperformed sequence-only scanning for all five TFs, recovering four within the top three candidates and all five within the top ten. RegScan also identified candidate TF binding sites from CRISPR-derived and deep-learning-derived importance scores. RegScan provides a framework for linking nucleotide-resolution functional measurements to TFs underlying cis-regulatory activity.

# RegScan: A Statistical Framework for Identifying Functional Transcription Factor Binding Sites from Per-Nucleotide Importance Scores

Zain M. Patel[1–3]*, Faheem A. Azeemi[1, 2, 4], Timothy Barry[1–3, 6–9], Basheer Becerra[1–3, 5–9], Daniel E. Bauer[1, 6–9], Luca Pinello[1–3]

1. Gene Regulation Observatory, Broad Institute of MIT and Harvard; 2. Molecular Pathology Unit, Krantz Family Center for Cancer Research, Massachusetts General Hospital Research Institute; 3. Department of Pathology, Harvard Medical School; 4. Massachusetts Institute of Technology; 5. Bioinformatics and Integrative Genomics PhD Program, Harvard Medical School; 6. Division of Hematology/Oncology, Boston Children's Hospital; 7. Department of Pediatric Oncology, Dana–Farber Cancer Institute; 8. Harvard Stem Cell Institute; 9. Department of Pediatrics, Harvard Medical School.

**Introduction:** High-throughput functional genomics approaches, including massively parallel reporter assays (MPRAs) and CRISPR-based perturbation screens, have expanded our ability to dissect cis-regulatory elements at nucleotide resolution. Computational analysis of these experiments, together with attribution or in-silico mutagenesis applied to deep-learning models of regulatory sequence, can generate per-nucleotide importance scores that quantify how individual sequence perturbations affect regulatory activity. However, identifying the transcription factors (TFs) underlying these effects still predominantly relies on scanning the reference DNA sequence with TF position weight matrices (PWMs). This sequence-based paradigm identifies regions resembling canonical TF motifs but does not incorporate the functional nucleotide preferences revealed by perturbation experiments.

**Methods:** We developed RegScan, a statistical framework that uses TF PWMs to directly scan nucleotide-resolution importance scores and prioritize TFs whose nucleotide preferences are consistent with observed regulatory effects. Rather than asking whether a sequence resembles a TF motif, RegScan asks whether the functional consequences of perturbing that sequence agree with the nucleotide preferences encoded by the PWM. For each candidate TF and position, RegScan transforms the aligned PWM relative to the wild-type nucleotide, yielding a representation of the TF's predicted preference for each substitution that directly parallels the observed importance scores. It then quantifies agreement using cosine similarity and assesses significance against a null distribution derived from shuffled importance scores, followed by multiple-testing correction.

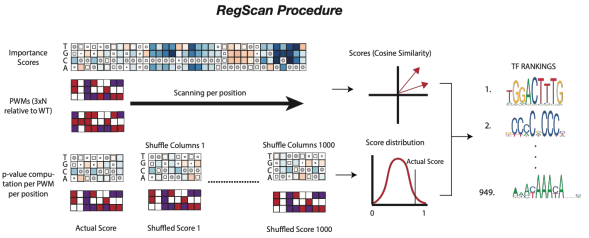


*Figure 1. RegScan workflow. TF PWMs are transformed relative to the wild-type sequence and compared with per-nucleotide importance scores using cosine similarity; significance is evaluated against a shuffled null distribution.*

**Results:** We evaluated RegScan across three complementary sources of nucleotide-resolution functional information: saturation-mutagenesis MPRAs, CRISPR-based regulatory tiling screens, and importance scores derived from deep-learning models. We first benchmarked RegScan on saturation-mutagenesis MPRA regulatory elements with experimentally validated TF binding sites [1]. When scanning the complete regulatory element without prior knowledge of the functional site, RegScan improved the rank of the validated TF in four of five elements compared with sequence-only PWM scanning. When analysis was restricted to experimentally defined single nucleotide variants, RegScan matched or outperformed sequence-only scanning for all five validated TFs. Four of five were recovered among the top three candidates and all five within the top ten, with particularly large gains when sequence-only motif matching performed poorly. For example, ELF1 improved from rank 25 with sequence-only scanning to the top-ranked candidate with RegScan. Beyond MPRAs, we applied RegScan to nucleotide-resolution importance scores from CRISPR-based regulatory screens and deep-learning models, including ChromBPNet [2], demonstrating that it operates independently of any source of functional scores.

**Conclusions:** RegScan integrates TF binding preferences with nucleotide-level functional effects to prioritize functional TF binding sites. The framework is applicable to importance scores derived from MPRA, CRISPR, and deep-learning approaches, enabling a unified strategy for linking nucleotide-resolution functional measurements to the TFs underlying cis-regulatory activity.

**References**

[1] Siraj L, Castro RI, Dewey HB, et al. Functional dissection of complex trait variants at single-nucleotide resolution. Nature. 2026.

[2] Pampari A, Shcherbina A, Kvon EZ, et al. ChromBPNet: bias factorized, base-resolution deep learning models of chromatin accessibility reveal cis-regulatory sequence syntax, transcription factor footprints and regulatory variants. bioRxiv. 2025.