Comprehensive, high-resolution binding energy landscapes of transcription factor binding

Daniel D. Le[1,*], Tyler C. Shimko[1,*], Arjun K. Aditham[2,3], Allison M. Keys[3,4], Scott A. Longwell[2], Yaron Orenstein[5], and Polly M. Fordyce[1,2,3,6#]

Gene expression *in vivo* is regulated primarily by transcription factor (TF) proteins that bind regulatory sequences within genomic DNA, thereby recruiting or blocking the transcriptional machinery to either activate or repress transcription. However, it remains difficult to predict TF occupancy *in vivo* based on regulatory sequence. Thermodynamic models that explicitly consider TF concentrations and the change in Gibbs free energy upon binding have shown the greatest success; however, a scarcity of energetic data has precluded their broad implementation. Although high-throughput *in vitro* transcription factor (TF) binding site characterization techniques have greatly increasing the speed of TF target site discovery, these techniques sacrifice the ability to measure binding energies in order to query a large sequence space. Here, we present a novel high-throughput experimental assay and analysis pipeline capable of estimating changes in binding energies for > 1 million sequences in parallel at high resolution. To demonstrate the capabilities of the assay, we took advantage of existing specificity information to refine the binding motifs of two model TFs from S. cerevisiae, Pho4 and Cbf1. By coupling the existing MITOMI microfluidic platform to a DNA sequencing-based readout and high-capacity neural network models, we generated comprehensive thermodynamic landscapes for an exhaustive library of all 10-mer sequences flanking the shared Pho4 and Cbf1 consensus motif. These measurements reveal that sequence specificity extends far beyond the known consensus to distal flanking positions and that the extended binding specificity of Cbf1, in particular, is surprisingly epistatic. When combined with existing in vivo datasets, we find that sites occupied by TFs in vivo are both energetically and mutationally distant from the lowest-energy sequences, providing evidence that even small differences in binding energies can provide a basis for evolutionary selection.

*,# These authors contributed equally to this work.
[1]Department of Genetics, Stanford University, Stanford, CA 94305.
[2]Department of Bioengineering, Stanford University, Stanford, CA 94305.
[3]ChEM-H Institute, Stanford University, Stanford, CA 94305
[4]Department of Chemistry, Stanford University, Stanford, CA 94305.
[5]Department of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer-Sheva, Israel PM 653.
[6]Chan Zuckerberg Biohub, San Francisco, CA 94158.

Acknowledgements: This work was funded by NIH grant XX00000. This footnote is optional.
[1]Department of Systems Biology, University of Quantitative Biology, Other address information. E-mail: author@place.edu
[2]Mathematical Biology Center, BioPark USA. E-mail: two@place.gov
[3]Another Dept, Another Institution, Address, E-mail: three@place.com