Discovering the autonomous rules by which sequence drives transcription using a novel massively parallel barcoded reporter assay

Vincent D FitzPatrick[1,2,3], Joris van Arensbergen[4], Ludo Pagie[4], Marcel M de Haas[4], Bas van Steensel[4], and Harmen J Bussemaker[1,2]

[1]*Columbia University, Department of Biological Sciences, New York, NY*
[2]*Columbia University Medical Center, Department of Systems Biology, New York, NY*
[3]*Columbia University, Graduate Program in Molecular Biophysics, New York, NY*
[4]*Netherlands Cancer Institute, Division of Gene Regulation, Amsterdam, Netherlands*

DNA-binding proteins regulate expression through sequence-specific interactions with gene promoters. These interactions are modulated by local chromatin context features extrinsic to the promoter sequence, making it difficult to separate sequence-dependent direct regulatory mechanisms from indirect contextual factors. To address this problem in a comprehensive and unbiased manner, we developed SuRE (Survey of Regulatory Elements). SuRE is an ultra-high-throughput barcoded reporter assay that quantifies the autonomous ability to drive expression for millions of genomic elements in parallel. The unique nature of these data allows us to dissect transcriptional regulation in novel ways. Having SuRE readouts for fragments that partially and randomly overlap with the region around annotated transcriptional start sites (TSS) allowed us to uncover detailed spatial rules. We find that sequence signals that autonomously drive expression in the sense direction are largely contained within the first 150bp upstream of the TSS. Reporter expression driven by antisense promoter fragments is about half as intense on average, and driven by signals in the same 150bp sequence window. We also asked whether the known dependence of expression on CpG content still holds in the SuRE context, when proximal promoter sequence is isolated from its genomic context. When we categorize SuRE fragments in terms of (i) the observed CpG density and (ii) the expected CpG density given the G and C density in the same fragment, we observe a striking trend of decreasing expression with increasing CpG depletion. Our results extend and refine existing ideas about the role of CpG 'islands' in transcriptional regulation. To perform the above analyses, we developed a flexible framework based on generalized linear models (GLM) and Poisson counting statistics, which we use to quantify the contribution of various (spatial or sequence) features associated with each SuRE fragment to the expression level. The same approach can be used to perform motif-based analyses of how transcription factor binding sites contribute to expression in a context-specific manner.