Cell Systems

CellPress

Voices

Emerging questions in transcriptional regulation

What new questions can we ask about transcriptional regulation given recent developments in large-scale approaches?

Elphège P. Nora
Cardiovascular Research Institute, University of California, San Francisco

Will proving mechanisms always need experiments?

While many are rightfully excited about the new questions we can ask with large-scale approaches, I am equally fascinated by the concomitant shift in what molecular biologists are willing to take for answers.

As a student in molecular genetics interested in transcriptional regulation in the mid-2000s, a primary goal of my first experiments with emerging high-throughput genomic approaches had remained to investigate how the molecular mechanisms established by dissecting one locus applied (or not) genome-wide. By the mid-2010s things had already changed completely, as new protocols and software allowed younger trainees to start their career by quickly generating a deluge of data, enabling them to "shoot first and ask questions later."

Today, the democratization of artificial intelligence (AI) approaches is causing an even more profound paradigm shift for those pursuing mechanistic insight. For example, after training on reference data, a single computational student can now probe entirely *in silico* how hundreds of thousands of mutations may affect various genomic processes, such as transcription, transcription factor binding, enhancer activity, or chromatin folding—with the need for experimental validation only coming very late in such projects.

As their experimental validation rates rise, AI-based predictions may start becoming acceptable alternatives to experimental measurements when validating mechanistic models, such as biophysical simulations. All this considered, molecular biologists may therefore have to start asking themselves: ultimately, will proving a novel molecular mechanism always require experiments?

Stein Aerts
VIB Center for Brain & Disease Research, Leuven and KU Leuven

(Deep) learning enhancer codes

Large-scale single-cell profiling of gene expression and chromatin accessibility provides unprecedented amounts of training data to model and decipher gene regulation across tissues, organisms, development, and disease. Recent modeling approaches thrive on these data and allow researchers to ask ever more detailed questions about transcriptional regulation. Firstly, new types of gene regulatory network (GRN) models aim to better address an old question of "who regulates whom." Chromatin accessibility data facilitate the integration of genomic enhancers as nodes into the GRN, thereby connecting upstream transcription factors (TFs) to their target genes, forming *enhancer-GRNs* (eGRNs). As their accuracy increases, GRN models become more predictive and can be utilized to answer new questions, including "what will be the effect of a TF perturbation" and "how will a cell's transcriptome change from one state to the next, in a single-cell trajectory?"

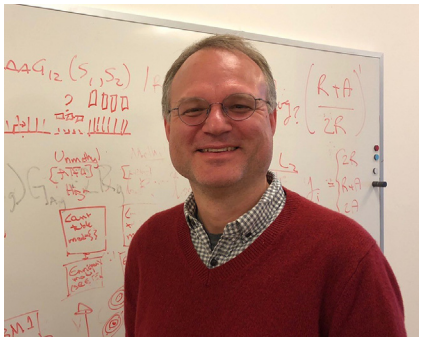
Secondly, convolutional neural network (CNN) models are trained on the DNA sequence of enhancers or entire gene loci, to predict chromatin accessibility, TF binding, and gene expression. Through "explainability" techniques, these models are scrutinizing *cis*-regulatory logic at a remarkable pace and finally provide answers to key questions such as "what is the effect of genomic variation on enhancer function," "which TFs cooperate and what is each TF's contribution (e.g., activation, repression, nucleosome displacement)," and "how do enhancers and promoters cooperate?" Interestingly, CNNs are also being used to generate synthetic enhancers with altered properties, adding a powerful synthetic biology dimension to the toolbox of single-cell regulatory genomics. Finally, a plethora of other large-scale approaches further fuel the AI-empowered computational dissection of the genomic regulatory code,

including single-molecule sequencing, massively parallel reporter assays, spatial transcriptomics, CRISPR screens, proteomics, and TF binding assays. The current era is marked by excitement, as machine learning and technology steadily lead us toward the resolution of the idiosyncratic *cis*-regulatory logic of each gene.

Patricia J. Wittkopp
University of Michigan

Treasure your exceptions

Our contemporary understanding of eukaryotic gene regulation is built upon a foundation of detailed case studies of individual genes. This mechanistic work has identified key roles for different types of DNA sequences, chromatin configurations, and regulatory proteins that motivated development of high-throughput techniques enabling these mechanisms of transcriptional regulation to be studied on a genomic scale. Data resulting from these techniques have provided a multifaceted look at gene regulatory networks and identified new components of regulatory systems (e.g., enhancer RNAs, topologically associating domains). As an evolutionary biologist, I am particularly excited about the potential for these tools to help us understand how changes in DNA sequences impact different layers of gene regulation, alter gene expression, and impact organism-level phenotypes. Some studies have now used these tools in a comparative framework, looking at the relationships among DNA sequences, chromatin structure, transcription factor binding, and gene expression between strains and species. These studies often find the predicted functional relationships (e.g., between binding of chromatin remodelers and chromatin structure) more often than expected by chance, but these relationships are rarely absolute, leaving much regulatory variation unexplained. I believe we should heed the advice of William Bateson to treasure our exceptions, looking not only at where such datasets can explain regulatory variation but also where they fail to do so. In this way, large-scale studies of transcriptional regulation can be used to point us toward specific loci that might harbor mechanisms of gene regulation that we are yet to discover.



Harmen J. Bussemaker
Columbia University

How transcription factors interact!

One of the paradoxes of regulatory genomics is how it is possible for transcription factors to control only a subset of genes: the fold-difference in equilibrium constant (K_D) between optimal and non-specific DNA binding is $< 10^5$ for most human TFs, but the genome consists of $> 10^9$ base pairs. TFs can increase their target specificity by forming complexes, but dissecting this molecular complexity in a way that allows us to, for example, predict the impact of non-coding genetic variants on TF function, has proven challenging.

Variation is key to learning. The classic example is linear regression: variation in X allows us to quantify the relationship between X and Y. The recent explosion in single-cell and spatial assays, along with perturbative screens, has given us access to functional readouts in the context of natural and synthetic variation in cellular state. In a parallel technological advance, massively parallel reporter assays have given us a way to comprehensively explore sequence space through vast libraries of natural or synthetic *cis*-regulatory DNA.

All this multiplexing across cells, genes, and variants comes at the cost of *sparsity* of the read counts that constitute the data in the era of massively parallel sequencing. A potentially powerful approach to dealing with this sparsity is to summarize the data in a biophysically interpretable way, in terms of cell-state-specific nuclear TF protein concentrations, along with binding energy models that precisely define how TFs interact with DNA and with one another.

Martha Bulyk
Brigham & Women's Hospital and Harvard Medical School

Encodings and outputs of *cis*-regulatory elements

A fundamental question in biology is "how are instructions for gene regulation encoded in the genome?" The development of highly parallel reporter assays has precipitated a multitude of studies that have assayed a wide range of DNA sequences for their transcriptional regulatory effects on reporter gene expression. These studies have determined the enhancer activities of natural or synthetic DNA sequences and the effects of non-coding variants and have also investigated the contributions of core promoter elements to expression output. Investigators have probed the "grammar" of how TF binding site arrangement in enhancers produces quantitative transcriptional output. For which TFs, in which *cis*-regulatory elements (CREs), and in what cellular contexts are precise gene regulatory outputs, such as in development, critically dependent on lower-affinity sites? How well do results from these compact reporter constructs capture the activity of endogenous elements located far upstream or downstream of promoters? (How) are these elements' activities modulated in their native chromosomal context?

Some highly parallel CRE assays have investigated the activities of silencers—negatively acting regulatory elements—about which far less is known than enhancers. Nearly all the silencers my lab identified in this way in *Drosophila* embryonic mesoderm acted as enhancers in a different cellular context. Such dual readout of CREs raises questions about how different regulatory encodings coincide and highlights the need to test CREs in multiple cell types, including developmental contexts. Advances in profiling physical interactions among genomic regions (e.g., Hi-C and related methods) allow one to identify chromosomal contacts made by silencers. What do those interactions reveal about mechanisms of silencer activity? Are other types of elements (e.g., insulators, tethering elements) bifunctional? What are the effects of non-coding variants in bifunctional elements? Ultimately, we need not just a "catalog" of enhancers versus silencers but a multidimensional matrix of CRE quantitative outputs across cell states and to understand how that regulatory output is encoded and readout.

Saurabh Sinha
Georgia Institute of Technology

Gene regulation and animal behavior

Single cell multi-omics technologies are rapidly changing the study of cell populations, revealing diverse cell types, intercellular signaling, gene regulatory networks (GRNs), etc. in heterogeneous tissues. These recent developments can be game-changing for mechanistic studies of animal behaviors, which are frequently studied in terms of associated activities of neuronal networks (NNs) but also induce large changes in brain transcriptome and epigenome. These changes are coordinated by GRNs. Charting "behavior-related GRNs" and understanding their interplay with NNs, developmental GRNs and environmental stimuli is a grand challenge, and recent breakthroughs in single-cell -omics might just be the catalyst for solving it. These technologies are already being used in mapping brain GRNs, and their potential for deciphering cell-cell communication may reveal how NNs shape GRN dynamics and epigenomic states. Emerging technologies for spatial omics at single-molecule resolution can provide detailed views of subcellular events involving RNAs and proteins, including localization, complex formation, translation, and transport, all of which may underlie systems-level regulation in the polarized cells of the brain. Drawing out such rich views of intra- and intercellular regulation at the scale of brain regions, even whole brains, at multiple time scales and under carefully designed behavioral conditions, potentially in parallel with state-of-the-art techniques for mapping NN activity and connectivity, can revolutionize the study of animal behavior.

Julia Zeitlinger
Stowers Institute for Medical Research

What are the sequence rules driving gene regulation?

The exponential growth of large-scale genomics datasets in different organisms, tissues, and cell types creates both a need and unique opportunity to harvest the underlying information in a new learning paradigm. After decades of focusing on mechanisms of gene expression, it is now time to come back to a concept that has its origins before the rise of modern molecular biology and biochemistry: much of biology has a DNA sequence basis. With the development of neural networks that predict genomics data from sequence, learning how gene regulation is encoded in DNA is now feasible. It does however require a drastic departure from previous computational approaches and biological reasoning. Traditionally, we take genomics datasets apart in a hypothesis-driven fashion and extract sequence rules one at a time. In the new paradigm, we initially set aside our biological assumptions and let neural networks learn highly complex combinatorial sequence rules inside a black box. Only after having achieved high prediction accuracy are the relevant sequences and rules extracted from the model. With this learning paradigm, we can now put DNA sequence back into the driver's seat. What are the sequence rules of gene regulation when we learn them in an unbiased way? What are the unifying rules across cell types and organisms? How are they connected to the mechanisms of gene regulation? How do regulatory mutations affect an organism? Ultimately, this will lead to knowledge in biology that is both fundamental in nature and directly applicable to understanding human health and disease.

Justin Crocker
European Molecular Biology Laboratory, Heidelberg

Regulatory networks in a natural context

Gene regulatory networks are complex across every level of organization—enhancers are templates for transient protein interactions; regulatory networks are dense webs of interacting transcription factors; networks themselves are modified by the environment and epigenetic landscape. Furthermore, we know that vast numbers of genes contribute to trait variation and the heritability of complex diseases—the bulk of this variation is in transcriptional regulatory regions. All of these interactions are products of evolution and subject to continual change. This complexity at every level of biological organization creates an intimidating task in understanding transcriptional regulation.

A further challenge is that living systems do not exist in isolation or idealized laboratory environments. Instead, organisms' habitats are complex and dynamic, which include other species. Even in cases where an environment's impact on phenotypes is well described for an individual organism or across its population, the underlying molecular processes and mechanisms are not. It is clear that by only examining systems isolated from their natural environments, we will fall short of understanding the intricacies of the regulatory networks that shape phenotypic variation.

Developmental biology is uniquely poised to address these challenges, offering a powerful lens to explore regulatory networks. Development biology systems can be used to explore how transcriptional regulation is integrated over organismal development and subsequent life cycles. Importantly, we can use controlled laboratory conditions to mimic the varied and varying natural environments in which regulatory networks have evolved and continue to evolve. It will be essential to continue to develop precise, high-throughput techniques for mapping networks across diverse cell types. While difficult, embracing such a research program provides the substantial advantage of the ability to focus on the function of regulatory networks closer to natural contexts.

Juan Ignacio Fuxman Bass
Boston University

High-throughput beyond correlation

The high-throughput studies that emerged in the early 2000s brought significant advancements in our understanding of gene regulation, expanding our knowledge beyond cherry-picked examples of model genes. However, these studies had limited cell-type resolution, relied on correlation between different parameters, and focused on statistical associations with diseases, rather than causal links. Recent advances have enhanced our ability to perform more mechanistic studies, surpassing the limitations of early works. Large-scale -omics projects and consortia have enabled integration between datasets and have produced vast amounts of data for training machine learning models that can predict transcriptional activity and the impact of non-coding genetic variants. Additionally, high-throughput reporter assays have been instrumental in identifying the causal variants among the tens or hundreds of variants statistically associated with complex diseases or traits in genome-wide association studies. Single-cell multi-omics, perturbation, and single-molecule approaches are allowing us to determine the impact of chromatin states, transcription factors, and DNA methylation on gene expression by comparing several parameters in each cell. Further, cryo-EM coupled with AI-driven structure predictions, as well as dynamics studies of transcriptional bursting, are significantly increasing our understanding of gene regulation at high molecular and temporal resolution. Future studies leveraging these and other technologies will enable end-to-end pipelines, from predictions of causal genetic variants to the development of variant-specific therapeutics, and will generate high-resolution models of gene expression that integrate chromatin states, transcription factor and cofactor recruitment, molecular compartments, and dynamics.

DECLARATION OF INTERESTS

J.Z. is an investigator at the Stowers Institute for Medical Research and also a professor (affiliate track) at the Department of Pathology and Laboratory Medicine at the University of Kansas Medical Center. H.J.B. is a co-founder and shareholder of Metric Biotechnologies, Inc. Columbia University has filed a patent on a technology tangentially related, on which H.J.B. is one of the inventors, to the topic of the Voices piece that H.J.B. contributed.