https://doi.org/10.1038/s41592-024-02331-5

Unlocking gene regulation with sequence-to-function models

Alexander Sasse, Maria Chikina & Sara Mostafavi

By exploiting recent advances in modern artificial intelligence and large-scale functional genomic datasets, sequence-to-function models learn the relationship between genomic DNA and its multilayer gene regulatory functions. These models are poised to uncover mechanistic relationships across layers of cellular biology, which will transform our understanding of *cis* gene regulation and open new avenues for discovering disease mechanisms.

A fundamental goal of modern biology is building models that can infer phenotype from genotype. Such models are also key to a mechanistic understanding of disease heterogeneity and to tailoring medicines to individuals. This is particularly relevant in the context of complex disease. Even though complex diseases are a result of both genetics and environment, and the genetic component may even be relatively small, genetic analysis offers a unique perspective because – unlike any other disease biomarker – the causal direction is controlled as there is no way for a phenotype to go back in time and change the genotype.

Genetic influence on phenotypes can manifest either through modifications of the protein structure itself or by affecting regulatory processes that influence the temporal and spatial dynamics of protein expression (Fig. 1a, left). In this Comment, we focus on the 'when and where' of protein expression. More specifically, we focus on the regulation of mRNA abundance – an important, but not the only, determinant of protein levels.

What would it take to have a model that interprets a personal genome in terms of gene regulatory effects and does so in a way that is relevant for disease? Historically, this question has been tackled in pieces through the lens of a statistical genetics association framework[1]. However, a fundamentally different approach is emerging at the intersection of functional genomics and deep learning, and culminates in sequence-to-function models (Fig. 1a, right). Sequence-to-function models use variants of deep convolutional neural networks to learn the mapping between DNA sequence and functional readouts (chromatin accessibility, histone modification, gene expression and so on) in one or several cellular contexts[2] (Fig. 1b). Formulating the genotype–phenotype question as a prediction problem enables computation approaches to integrate large collections of functional genomic assays in a single unifying framework.

In practice, such models are trained on large collections of functional genomic datasets across a variety of cellular contexts that are assembled by consortia such as ENCODE (Encyclopedia of DNA Elements)[3]. Most state-of-the-art models use the same training recipe and evaluation strategy. They take as input short subsequences of genomic DNA (thousands of base pairs to hundreds of kilobases) from the reference genome to predict functional outputs, including gene expression from a particular cellular context[4–6]. During training, model generalization is assessed on left-out regions of the genome (for example, random chromosomes left out during training) – a task that current models perform very well. Importantly, their ability to generalize to unseen genomic DNA can be attributed to their ability to learn 'sequence grammar'; that is, the rules for how the interactions between proteins and DNA, as well as higher-order complexes, govern functional output. Many studies have shown that for a variety of models, the learned sequence grammar recapitulates detailed mechanistic biological knowledge that has been acquired through years of genomics experimentation and analysis. For example, the models recapitulate transcription factor binding motifs[5], transcription factor complexes[5] and relationships between CTCF sites and 3D chromatin organization[7].

Importantly, because sequence-to-function models operate on the DNA sequence, they offer the ability to predict the effect of arbitrary genetic variation on functional outputs in a cell-type-specific manner (Fig. 1b) – a process referred to as in silico mutagenesis. An exciting application of this ability is to predict differences in gene expression (or other regulatory processes) among alleles within a population. In some cases, these models can already accurately predict variant effects, and approach the precision of experimental mutagenesis[4]. However, it is increasingly evident that effectively applying sequence-to-function models to interpret the complete spectrum of genetic variation on gene expression, in a way that is relevant to disease outcomes, requires enhancing their prediction resolution[8,9].

Predicting across loci of the same individual's genome and predicting across different genomes at the same locus (that is, allelic effects) differ in two important ways: the amount of variance between the DNA sequence inputs is markedly reduced, and the variance among the expression output of a single gene is relatively small compared to the expression differences between different genes. Yet, high-quality intergenome predictions are precisely what are needed to leverage this framework in the context of population genetics and disease susceptibility.

Increasing the resolution of models fundamentally requires augmenting the training data, which raises the question of what kind of training data would be most effective. Several orthogonal strategies merit investigation. First, current models have been trained on a range of epigenomic and gene expression datasets, yet there is a notable absence of models that combine assays from all regulatory process stages (particularly post-transcriptional regulation). For example, variants could affect the abundances of isoforms or influence post-transcriptional processes (such as polyadenylation, translation

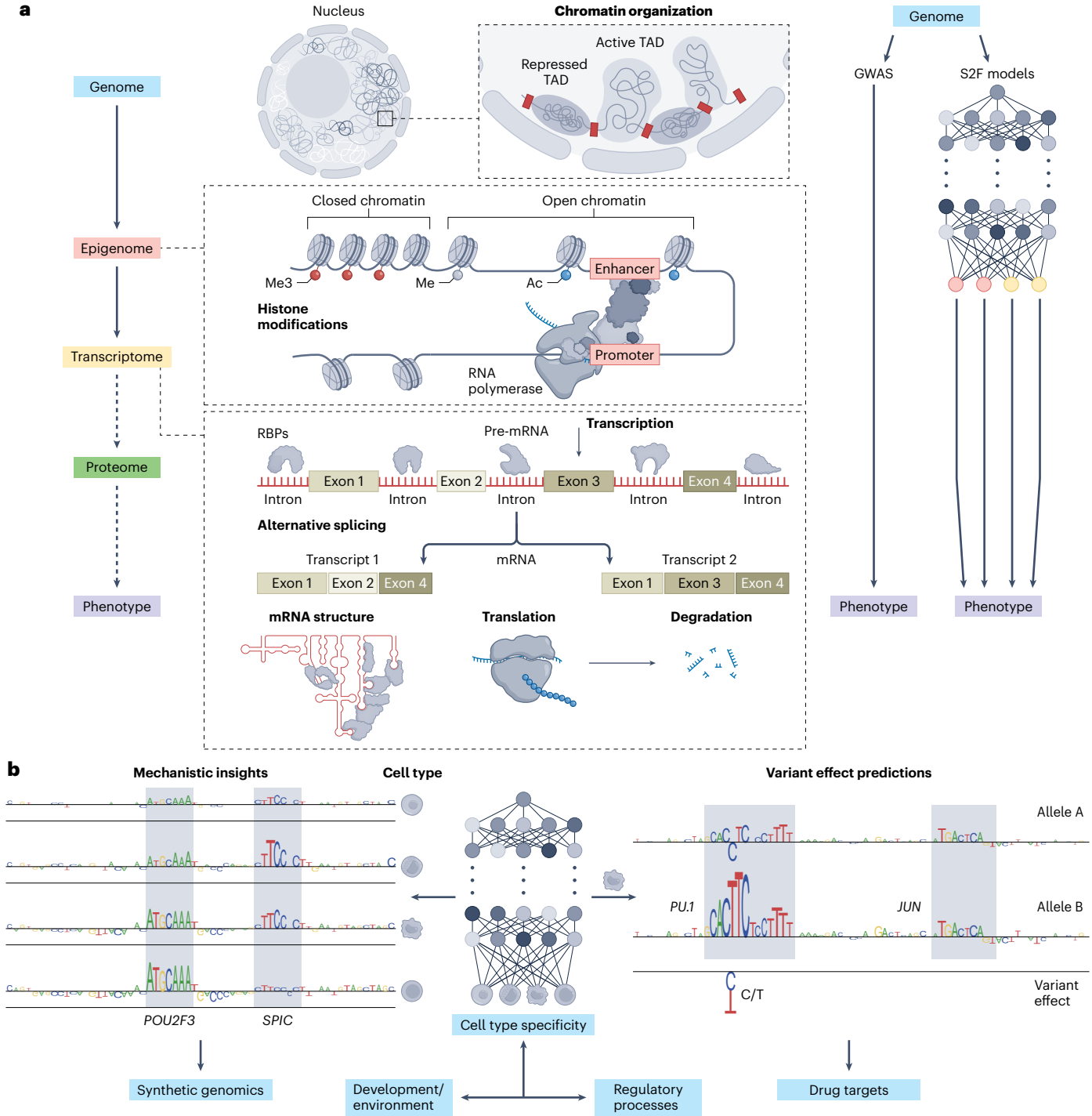

Fig. 1 | The complexity of genotype-to-phenotype relationship. a, Left, interpreting personal genomes requires a mechanistic understanding of the different layers of gene regulation and how intermediate processes (chromatin organization, epigenomic modifications, transcriptional regulation, post-transcriptional regulation and so on) are affected by genetic variation. Right, two approaches to genome interpretation, through statistical association and cell-type-specific sequence-to-function (S2F) models. **b**, Sequence-to-function models take as input genomic DNA and learn to predict its functional properties such as gene expression in a cell-type-dependent and cell-state-dependent manner. Once trained, these models can be used to predict the impact of arbitrary genetic variations (right) and to derive biological insights into the sequence grammar that determines context-dependent gene regulation (left). Ac, acetylation; Me, methylation; Me3, trimethylation; RBPs, RNA-binding proteins; TAD, topologically associating domain.

or mRNA stability), leading to changes in steady-state gene expression without affecting transcriptional rates.

Second, it stands to reason that more examples of sequence and gene expression pairs would enhance the model's internal representation of the genotype-to-function map. This can be achieved in several ways. Cohort studies (in which both whole-genome sequencing and gene expression are measured for the same individuals) offer a substantial increase in training data points, as each individual contributes all of their alleles. However, enhancements to current model architectures are needed so that training performance does not become dominated by strong variance in the data across genes; instead, models should focus on the subtle variance across individuals. Another method to boost sequence diversity is through the incorporation of evolutionary information by, for example, jointly training sequence-to-function models on data from multiple species[10], as many gene regulatory mechanisms are conserved between distal descendants of the same ancestor. Moreover, nonhuman model organisms provide researchers with opportunities for additional experiment (such as various types of perturbations) that would be ethically unfeasible otherwise.

However, it is important to acknowledge that native genomic DNA represents only a small part of the space of possible genomic DNA composition, and that certain variations will never be observed in nature owing to their deleterious effects. Therefore, it is compelling to assume that incorporation of genomic readouts from synthetic sequences might be needed to learn a comprehensive map from DNA subsequences to their regulatory functions[11]. For example, massively parallel reporter assays enable measurements from >100,000 sequence variants for multiple regulatory processes, such as expression, splicing or degradation. However, although it is clear that these synthetic systems capture many of the gene regulatory mechanisms used by cells, how to translate what has been learned in these systems to the context of cell-type-specific sequence grammar remains an open problem.

Third, concurrent with advancements in sequence-to-function models, genomic language models present an approach inspired by the success of large language models such as ChatGPT. These models use self-supervised learning, drawing on vast quantities of unlabeled genomic DNA across different species to learn the statistical relationships between DNA sequence composition within and across genomic positions. As such, genomic language models provide a general foundation for various downstream applications, including the prediction of functional elements, gene expression and sequence design[12]. They have been widely benchmarked on a variety of classification tasks for functional noncoding elements; however, recent work has shown that this representation lacks cell-type-specific information, and predictors of cell-type-specific functional elements do not benefit from this resource-intensive pretraining step[13]. *Cis*-regulatory elements evolve rapidly and, furthermore, the functional consequence of their sequence divergence is cell-type-dependent. Current genomic language models learn their representations from a couple of hundred to a few thousand genomes, whereas large language models such as ChatGPT use data corpora that are orders of magnitude larger for training. It is unclear whether such a complex cell-type-specific DNA language — developing over the course of 1.5–2.3 billion years of evolution of eukaryotes — can be learned from genomes that are currently sequenced. Thus, further research is needed to determine how to combine genomic language models and supervised sequence-to-function models that predict cell-type-specific events[13].

To effectively apply sequence-to-function models to organism-level phenotypes (including disease), it is critical to enhance their cellular resolution to incorporate data from diverse cell types and states. With the exponential growth in single-cell-resolution measurements, we can now envision such models trained on thousands of cellular states. In the future, incorporating additional dimensions of variation — such as age, sex and environmental exposures that affect genotype-to-gene expression relationships — may require new model architectures that are capable of effectively learning at scale from diverse measurements, at cell-type resolution. As the biotechnology industry becomes increasingly interested in these causal models of cellular biology, traditional academic institutions may face challenges in competing with them on scale. If these models are developed outside of academia, we hope they will be shared openly with the community. 'Open' benchmarking datasets and models will be essential for future progress, by enabling other researchers to apply and evaluate these models, identify limitations, and offer valuable feedback to the developers.

With maturing research in the directions discussed above, it is reasonable to assume that in the near future the community will be able to develop sequence-to-function models that excel at predicting gene regulatory effects from personal genomes. However, it remains unclear whether such models are sufficient to characterize missing phenotype-causing variants. It is critical to assess our ability to identify variants that do not merely alter gene expression but have direct implications for the organism-level phenotype. This entails a deeper understanding of the genotype–phenotype relationship beyond transcriptional regulation, as many variants that alter mRNA levels have no effect on phenotype (that is, are not hits in a genome-wide association study (GWAS)). Conversely, many noncoding variants that do affect phenotypes — presumably through regulatory mechanisms — do not produce expression alterations in currently available expression quantitative trait locus (QTL) datasets[14].

When we consider intermediate levels of regulation such as chromatin accessibility or histone modifications (for example, chromatin accessibility QTLs and histone acetylation QTLs), the coverage of GWAS hits by variants that are molecular QTLs increases[15]. As chromatin features affect phenotypes via gene expression, the statistical disconnection between the two is intriguing. Several explanations are possible. Because gene expression control is distributed across many regulatory elements, the final output may be well buffered, which implies that a relatively large effect size in the chromatin state of a single regulatory element may translate to a much smaller one when we measure the expression of the target gene. Another possibility is that chromatin features reflect either the history or the future potential for regulation in some highly specific cell-type, developmental or signaling context that is not captured by current assays. Finally, it is possible that phenotype-associated variants alter a property of gene regulation other than mean expression at steady state. For example, they could affect the kinetics or the variation in expression instead. Regardless of the specific mechanisms, it is likely that we will need to measure additional outputs beyond standard gene-expression measurements across multiple individual genomes to build a comprehensive model of regulatory variant effects.

Ultimately, the model of the effects of sequence changes on gene regulation will need to be reconciled with the protein-centric coding-variant perspective. Both views bring something to the table. Rare variant studies often uncover coding mutations that pinpoint specific cellular processes. However, mutations in broadly expressed

genes are also observed, which indicates that a subtle loss-of-function or gain-of-function effect is adequately compensated in all but a few select conditions. The critical contribution of regulatory variants is that they can precisely identify the disease-relevant cellular contexts.

Finally, current models are often designed on principles that were developed for language models. Although our genome can be represented as a sequence of strings, its 'function' is manifested through intricate layers of biology that span many millions of interactions across regulatory layers. This expansive space is exponential in nature, and high-throughput experimental data alone may not provide adequate training material to comprehensively learn this space. Therefore, new model architectures that leverage known biology to incorporate meaningful inductive biases will be crucial components to effectively learn from diverse data modalities.

In summary, with maturing of sequence-to-function models and their continuous improvements, we are entering unprecedented times for understanding cellular biology in a tractable and causal manner. These models are appealing because they capture the causal relationship between regulatory layers, revealing the encoding of function in our genomes. There is lots to be done here to build models that enable the community to simulate how genome function unfolds in different cellular and organismal contexts through the lens of *cis* gene regulation. It seems plausible that we can leverage the controlled causal direction of genotype analysis to begin to shed light on the elusive contributions of environmental factors to phenotypes. Although establishing causality in environmental influences remains a daunting challenge, a genetically anchored framework may provide new avenues to untangle this complexity.

Alexander Sasse[1], Maria Chikina[2,3] ✉ & Sara Mostafavi[1,3] ✉
[1]Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA, USA. [2]Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, PA, USA. [3]These authors jointly supervised this work: Maria Chikina, Sara Mostafavi.
✉e-mail: mchikina@gmail.com; saramos@cs.washington.edu

Published online: 9 August 2024

References

1. Uffelmann, E. et al. *Nat. Rev. Methods Primers* **1**, 59 (2021).
2. Li, Z. et al. *Cell Rep. Methods* **3**, 100384 (2023).
3. Luo, Y. et al. *Nucleic Acids Res* **48**, D882–D889 (2020).
4. Avsec, Ž. et al. *Nat. Methods* **18**, 1196–1203 (2021).
5. Avsec, Ž. et al. *Nat. Genet.* **53**, 354–366 (2021).
6. Zhou, J. et al. *Nat. Genet.* **50**, 1171–1179 (2018).
7. Zhou, J. *Nat. Genet* **54**, 725–734 (2022).
8. Sasse, A. et al. *Nat. Genet.* **55**, 2060–2064 (2023).
9. Huang, C. et al. *Nat. Genet.* **55**, 2056–2059 (2023).
10. Kelley, D. R. *PLOS Comput. Biol.* **16**, e1008050 (2020).
11. de Boer, C. G. & Taipale, J. *Nature* **625**, 41–50 (2024).
12. Dalla-Torre, H. et al. Preprint at *bioRxiv* https://doi.org/10.1101/2023.01.11.523679 (2023).
13. Tang, Z. & Koo, P. K. Preprint at *bioRxiv* https://doi.org/10.1101/2024.02.29.582810 (2024).
14. Mostafavi, H., Spence, J. P., Naqvi, S. & Pritchard, J. K. *Nat. Genet.* **55**, 1866–1875 (2023).
15. Arthur, T. D. et al. Preprint at *bioRxiv* https://doi.org/10.1101/2024.04.10.588874 (2024).

Acknowledgements

We thank C. de Boer and X. Tu for helpful comments. ChatGPT was used to refine some of the sentences.

Competing interests

The authors declare no competing interests.