De novo decipherment of genetic architecture with language models

Salil Bhate [1]

Abstract

A challenge in biology is to identify the genetic architecture of molecular phenotypes like single-cell gene expression and tissue organization: which genomic loci influence which phenotypic quantities? We present Phenotype Syntax Alignment (PSA), an unsupervised translation algorithm that uses language models to translate phenotypic data into the reference genome. Applying PSA to single-cell gene expression and tissue organization, we recover causal genetic associations *de novo*: from just phenotypic data and the reference genome sequence (with no external data, annotations or genetic variation). This extended abstract motivates the task and outlines our approach and results.

1. *De novo* decipherment as an ML task

A fundamental challenge in biology is to identify the genetic architecture of molecular phenotypes like single-cell gene expression and tissue organization(Palla et al., 2022). Specifically, identifying which genomic loci control which cell types and cellular neighborhoods will yield understanding of the mechanisms of biological tissues, as well as strategies for their precision-engineering.

The current strategy for identifying the genetic architecture of such phenotypes has two general stages. First, genotype and phenotype are correlated by measuring phenotypic responses to naturally occurring or experimentally introduced genetic variation. Second, machine-learning models predict effects of changes to genomic sequences using as supervision such observations, along with other genomic profiling technologies (such as ATACseq) and perturbation experiments.f

We introduce the problem of *de novo* decipherment. The goal of *de novo* decipherment is to recover the genetic architecture of a phenotype given only a phenotypic measurement (single-cell gene expression matrix or tissue imaging data) and the reference genome sequence. No biological annotations, genetic variation or prior knowledge (or indeed any other external data) can be utilized. Why consider the seemingly difficult task of *de novo* decipherment?

1. Many phenotypes cannot be modelled experimentally or measured at scale in patient cohorts (for example, tissue organization in patients with rare diseases). Thus, *de novo* decipherment provides a way to investigate genetics when otherwise difficult or impossible.
2. The inductive biases that enable *de novo* decipherment across different phenotypes may reveal broad principles of information encoding by the genome.
3. Is *de novo* decipherment even possible? Imagine if finding the genomic loci controlling complex molecular phenotypes were, with the right models, as easy as aligning an amino acid sequence with the genome!

2. Previous work

Previous machine learning approaches have successfully predicted genetic associations with molecular phenotypes. The main contexts for such models include:

- Predicting functional genomic properties (e.g. gene expression/chromatin) from sequence. E.g. (Avsec et al., 2021)
- Predicting the effects of genetic perturbations. (Ji et al., 2021)
- Variant effect prediction (including using language models) (Benegas et al., 2022)

Contrastive learning approaches have successfully performed cross-modality translation, including from genotype to phenotype (e.g. (Radhakrishnan et al., 2023))

However, such models require measurements of genetic variation, annotations or correspondences with genomic loci as supervisory signal. To our knowledge, approaches have not previously attempted *de novo* decipherment.

A preprint of a previous version of this work is available (Bhate et al., 2022).

[1]Broad Institute of MIT and Harvard, Cambridge, MA. Correspondence to: Salil Bhate <sbhate@broadinstitute.org>.

The 2023 ICML Workshop on Computational Biology. Baltimore, Maryland, USA, 2023.

3. Motivating observations

Why would we expect *de novo* of genetic architecture to be possible? Our unsupervised translation approach is motivated by the following two observations from in cell-type specific expression quantitative trait loci (eQTLs) from (Yazar et al., 2022).

1. 'Omics studies have shown that cell types that are similar in gene expression space or image patches that have similar cellular composition have similar biological functions. Therefore, nearby cells or image patches in their respective high-dimensional feature spaces should similar associations with genomic loci. Indeed, investigating the eQTL dataset, we observed that similar cell types (in terms of their gene expression) had similar eQTLs.

2. The syntactic context of a genomic locus - which motifs are present - determines its functional activity (e.g. determining transcription factor binding, protein domains or splicing). Therefore, syntactically similar genomic loci should be associated with similar phenotypic structures. But how to quantitatively parameterize the syntax at a genomic locus without any supervision or biological annotations? Self-supervised language models (trained by masked prediction) have been show to learn syntactic features useful for downstream functional predictions (Benegas et al., 2022). We therefore used embeddings from a DNA language model (pretrained on the reference genome) to extract syntactic features of genomic loci. We observed that syntactically similar loci indeed were associated with similar sets of cell types.

These observations imply that the genetic architecture of a phenotype is consistent with the *geometry* of the phenotype (similarities between observations), as well as with the syntax of the reference genome.

4. The Phenotype Syntax Alignment algorithm

Given how much structure there is in the geometry of molecular phenotypes, as well as in the syntax of the genome, is consistency of the genetic architecture of a phenotype is consistent with phenotypic geometry and with genomic syntax a sufficiently strong constraint to enable *de novo* decipherment?

There are two algorithmic ingredients to Phenotype Syntax Alignment: representing possible genetic architectures with a language model, and optimizing the parameters of the language model using loss functions that enforce consistency with phenotypic geometry and genomic syntax.

4.1. Architecture

First, we parameterize, using a language model, the space of *possible genetic architectures*. Briefly, a possible genetic architecture is represented by a function:

$$T_\theta : \mathbb{R}^n \times \{A, T, C, G\}^{2000} \to [-1, 1],$$

where:

- θ indicates the parameters of the neural network T_θ
- $\mathbb{R}^n$ is high dimensional phenotype space (e.g. gene expression space or image patch feature space).
- $\{A, T, C, G\}^{2000}$ indicates the space of 2000mers.

A useful notion of genetic architecture that is modelled by such a function is where $T_\theta(x, s)$ represents the correlation (across a population) of the density of the phenotype at x (e.g. the frequency of cells with gene expression x) with mutations in a genomic 2000-mer s. The architecture of T_θ just adds a cross-attention layer onto a frozen, pre-trained DNA language model.

4.2. Training

We construct differentiable loss functions capturing the two forms of consistency. These loss functions are efficiently estimated by batch-sampling.

- Consistency with phenotypic geometry:

  $$L_P(\theta) = \sum_{x,y \in X} |d(x, y) - \mathbb{E}_{s \sim \text{genome}} |T_\theta(x, s) - T_\theta(y, s)||$$

  where $X \subset \mathbb{R}^n$ is the set of observed cells or image patches.

- Consistency with genomic syntax:

  $$L_G(\theta) = \sum_x \left[\frac{(T(x, s) - \frac{1}{K} \sum_{t \in N(s)} T(x, t))^2}{\sum_{x \in X} T(x, s)^2} \right]$$

  where $N(s)$ denotes the set of K-nearest syntactic neighbors (precomputed beforehand).

In practice, we sample 100k 2000-mers from the genome in each forward pass, and we use dimensionality reduction on the phenotype matrix with the SEACells algorithm (Persad et al., 2023), so that we can enforce consistency with the overall geometry of a phenotype in each forward pass (instead of sampling batches).

5. Evaluation

After performing PSA on a phenotype, we use genetic architecture T_θ to make predictions about which loci are associated with which quantitative traits. Specifically, we score loci from the genome according to

$$T_\theta(x_c, s) - T_\theta(x_c + \delta_g, s),$$

where δ_g indicates a small shift in the direction of some feature (either gene or image feature), and x_c indicates the mean feature vector for a cluster (cell type or cellular neighborhood). Thus, our hypothesis is that these scores (referred to as quantitative trait scores, QT-scores) are associated with the QTLs for the feature g in the cells or cellular neighborhood corresponding to c.

If the QT-scores recover genetic associations established in experimental studies or patient cohorts, we can conclude that the PSA algorithm has succeeded at *de novo* decipherment.

5.1. PSA recovers eQTLs

Do the QT-scores recapitulate genetic architecture as established in patient scores? We generated QT-scores from PSA for cell-type specific gene expression, evaluated their ROC-AUCs for distinguishing cis-eQTLs and trans-eQTLs from loci that were not associated with expression of the correct genes in the correct cell types.

Our ROC-AUCs for such tasks were in the range of 0.55-0.6. . We performed a battery of permutation tests, negative controls and ablations to evaluate the specificity of eQTL recovery. These tests (with p-values in the range 0.0001-0.01)) confirmed that PSA was recovering the right eQTLs, *de novo*, specifically for the right genes and the right cell-types in a manner that depended on the high-dimensional geometry of the input phenotype.

This ability to recover, *de novo*, the correct eQTLs for the correct genes in the correct cell types was consistent when PSA was applied across biological replicates.

5.2. PSA recovers causal associations

We investigated causality in two ways. First, we assessed whether the QT-scores from the PSA output correctly resolved causal eQTLs (determined by fine-mapping). Next, we assessed whether the QT-scores predicted CRISPR screen hits for the corresponding traits. Our AUCs in such tasks were in the range of 0.55-0.6, and p-values (using a similar battery of tests) were in the range $0.0001 - 0.01$.

5.3. PSA recovers genetic architecture from images

We applied PSA to a dataset of tissue image patches from the colorectal cancer immune tumor microenvironment, where the features were the density of different cell types. This level of organization is very 'distant' from gene regulation, so provides a way to evaluate the generality of PSA.

The QT scores in this context recovered QTLs for the cellular composition of the tumor microenvironment.

Moreover, we performed gene-set enrichment analysis on the specific loci highlighted by the learned genetic architecture for different cellular neighborhoods. This analysis recovered causal genetic pathways involved in microenvironment organization across cancers, as well as pathways specific to the neighborhoods of the CRC dataset. Thus, *de novo* decipherment with PSA can be used to gain mechanistic insights into the genetic pathways driving tissues - from only immunofluorescence imaging.

6. Summary

We introduce *de novo* decipherment: the task of recovering the genetic architecture of molecular phenotypes from a phenotypic measurement and the reference genome alone. We design the PSA algorithm, motivated by the observation that the genetic architecture of single-cell gene expression is consistent with phenotypic geometry and genomic syntax. We show, through a range of validation tasks and statistical tests, that PSA deciphers *de novo*, partially, the causal genetic architecture of single-cell gene expression and tissue organization.

Our work suggests that attention of the computational biology community toward *de novo* decipherment, via:

- Model optimization via establishment of ground truth datasets.
- Scaling PSA to evolutionary-scale data.
- Incorporation other inductive biases (e.g. consistency with 3D-genome organization).

could yield approaches to infer the genetic architecture of complex phenotypes in clinically significant settings where alternatives are not feasible, as well as basic understanding of the genome.

Acknowledgements

This work was supported by the Eric and Wendy Schmidt Center and the Broad Institute of MIT and Harvard.

References

Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J. R., Grabska-Barwinska, A., Taylor, K. R., Assael, Y., Jumper, J., Kohli, P., and Kelley, D. R. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18(10):1196–1203, 2021.

Benegas, G., Batra, S. S., and Song, Y. S. Dna language models are powerful zero-shot predictors of non-coding variant effects. *bioRxiv*, pp. 2022–08, 2022.

Bhate, S. S., Seigal, A., and Caicedo, J. Deciphering causal genomic templates of complex molecular phenotypes. *bioRxiv*, pp. 2022–08, 2022.

Ji, Y., Lotfollahi, M., Wolf, F. A., and Theis, F. J. Machine learning for perturbational single-cell omics. *Cell Systems*, 12(6):522–537, 2021.

Palla, G., Fischer, D. S., Regev, A., and Theis, F. J. Spatial components of molecular tissue biology. *Nature Biotechnology*, 40(3):308–318, 2022.

Persad, S., Choo, Z.-N., Dien, C., Sohail, N., Masilionis, I., Chaligné, R., Nawy, T., Brown, C. C., Sharma, R., Pe'er, I., et al. Seacells infers transcriptional and epigenomic cellular states from single-cell genomics data. *Nature Biotechnology*, pp. 1–12, 2023.

Radhakrishnan, A., Friedman, S. F., Khurshid, S., Ng, K., Batra, P., Lubitz, S. A., Philippakis, A. A., and Uhler, C. Cross-modal autoencoder framework learns holistic representations of cardiovascular state. *Nature Communications*, 14(1):2436, 2023.

Yazar, S., Alquicira-Hernandez, J., Wing, K., Senabouth, A., Gordon, M. G., Andersen, S., Lu, Q., Rowson, A., Taylor, T. R., Clarke, L., et al. Single-cell eqtl mapping identifies cell type–specific genetic control of autoimmune disease. *Science*, 376(6589):eabf3041, 2022.