CLASSIC: A platform for high throughput mapping of genetic design spaces in mammalian cells and ML guided prediction of gene circuit behavior

ABSTRACT

Massively parallel genetic screens have been used to map sequence-to-function relationships for a variety of genetic elements. However, because these approaches only interrogate short sequences, it remains challenging to perform high throughput (HT) assays on constructs containing combinations of multiple sequence elements arranged across multi-kb length scales. Overcoming this barrier could accelerate genetic design. For example, by screening diverse gene circuit designs, "composition-to-function" mappings could be created that provide insight into genetic part composability. Here, we introduce CLASSIC, a novel genetic screening platform that combines long- and short-read next-generation sequencing (NGS) modalities to quantitatively assess pools of constructs of arbitrary length containing diverse part compositions. We show that CLASSIC can measure expression profiles of >10^5 drug-inducible gene circuit designs (from 6-9 kb) in a single experiment in human cells. As we show, with a dataset of this size, it is possible to train machine learning (ML) models that not only predict the behavior of circuits from unmeasured regions of circuit design space, but also can be used as based models to expand design space mapping through an iterative active learning process. Furthermore, we show that by mapping entire circuit design landscapes, we gain critical insight into underlying circuit design and part composability principles that extend our understanding beyond standard biophysical models. Overall, our work shows that the expanded experimental throughput offered by CLASSIC dramatically augments the pace and scale of genetic design and establishes an experimental basis for AI-driven design of complex genetic systems.

INTRODUCTION. While cellular regulation is understood as a collection of interacting modules, a predictive understanding of how these modules interact within and across organizational scales to quantitatively specify a functional output remains elusive. Incomplete understanding of functional composability also presents a challenge for designed genetic systems. For example, in gene circuits, where DNA-encoded "parts" are combined together to specify novel regulatory relationships, *a priori* design of precise input/output behavior is non-trivial. Regulatory interactions within a gene circuit must be carefully tuned before the right composition parts that support a desired circuit behavior are identified[1]. Further, as parts must work within a crowded intracellular environment[2], incidental molecular coupling can occur between parts and with host cell machinery. Because these context-dependent interactions are difficult to predict, they can confound circuit design, making it a challenge to achieve a desired behavior.

Profiling an entire circuit design landscape in a single experiment with HT screening based approaches could enable rapid identification of circuit variants with desired behaviors. This could also facilitate the development of ML/AI models that are capable of inferring context-specific part function or forward predicting circuit behavior. HT screening approaches that utilize short-read NGS as a readout[3,4] have been used to generate detailed sequence-to-function mappings for multiple genetic part classes, including promoters[4], terminators[5], transcription factors (TFs)[6], nucleic acid switches[7], and receptors[8]. However, high-depth functional profiling of libraries of DNA constructs long enough (>1 kb) to encode entire circuits can be costly and technically challenging.

To overcome this challenge, we devised an approach[9] that combines long-read nanopore (ONT) and short-read NGS (Illumina) (**Fig. 1**). In this approach, libraries are generated via pooled part assemblies that incorporate semi-random barcodes. Nanopore sequencing is used to rapidly and inexpensively assess part composition and create a composition-to-barcode index. The library is then introduced into cells, binned based on expression phenotype, and analyzed by short-read sequencing to produce a barcode-to-phenotype index. A map matching construct composition to phenotype can then be revealed by comparing the two indices. Using this technique, which we refer to as CLASSIC[9] (combining long- and short-range sequencing to investigate genetic complexity), it is possible to obtain high-depth phenotypic expression data for large libraries (>10^6) of part compositions of arbitrary length using standard phenotypic selection or flow sorting experiments. To demonstrate the performance capabilities of CLASSIC, we created a library of >10^5 drug-inducible circuit designs in a single experiment (**Fig. 2**), simultaneously varying multiple categories of parts within the 2-gene circuit. *This represents the largest gene circuit library analyzed to date*. The abundance of data we gathered with this experiment enabled us to train an ML model (deep neural network) that could not only accurately predict unmeasured circuit configurations, but could be used as a base model to map additional categories of genetic parts through iterative fine-tuning experiments. Using this active learning process, it was possible for CLASSIC to traverse through an expansive genetic design space (~10^7), enabling us to rapidly converge on regions of desirable circuit function.

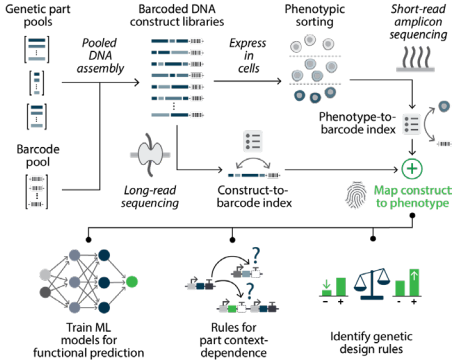
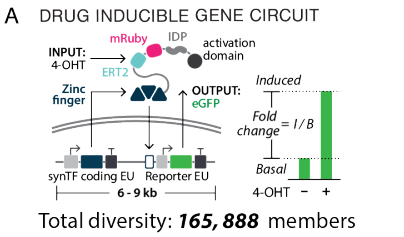


Fig 1: Overview of the CLASSIC workflow for high throughput screening of design spaces

RESULTS. The circuit we analyzed consisted of two genes: one that encoded a synthetic zinc-finger (ZF)-based transcription factor (synTF) with appended activation domains (ADs), and reporter gene harboring synTF binding motifs (BMs) upstream of a minimal promoter driving GFP expression (**Fig. 2A**). The inclusion of an ERT2 domain in the synTF renders it responsive to induction with 4-hydroxy tamoxifen (4-OHT). Diversification of the circuit design across 10 independent part categories resulted in an overall circuit design space of 165,888 compositions. Long-read nanopore (ONT) sequencing of the pooled library to index barcode sequences to circuit identity yielded assignments for 95.3% of total compositions, with no observable bias from the library construction process (**Fig. 2B, left**). We integrated this library into



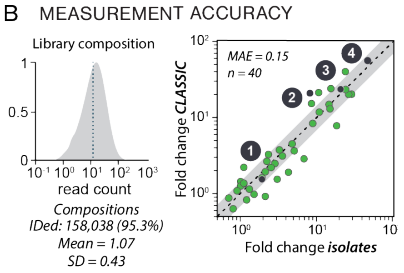


Fig 2: 10^5-member library construction and measurement with CLASSIC

HEK293T cells and sorted un-induced and 4-OHT-induced populations separately into 8 bins based on their GFP fluorescence levels. Following Illumina NGS analysis of each bin, a total of 121,292 (73% of design space) compositions were identified, and basal, induced, and fold-change expression values for each variant were computed. Fold-change values derived from CLASSIC measurements demonstrated excellent overall agreement (*MAE*=0.15) (**Fig. 2B, right**) with randomly isolated library members directly measured by flow cytometry (n = 40). These results argue that our CLASSIC workflow can be used to make large-scale experimental measurements of circuit behavior with a degree of quantitative accuracy comparable to the measurement of individual circuits.

To capture the behavior of compositions that were unmeasured by CLASSIC, we encoded circuit part categories as features and trained a deep neural network model to predict the basal and induced expression of all circuits, and observed excellent predictive power (r^2 = 0.86 and 0.88 respectively, **Fig. 3 left**). This allowed us to complete the design space with predictions for all 166k library members (**Fig. 3, middle**). Using the trained model, we made functional predictions for 96 circuits from across of fold-change behavior space, constructing and testing each individual to validate model accuracy. The model showed excellent concordance with individual variants, particularly for those in the (HFC) region (>25 fold induction in gene expression, **Fig 3, middle inset**). Statistical analysis of the completed design space revealed molecular rules underpinning circuit design, including the non-intuitive coupling between different part categories to produce HFC behavior. Additionally, clustering and UMAP projection[10] revealed several distinct families of part compositions were capable of supporting HFC circuit behavior (**Fig 3, right**). These results show that CLASSIC can furnish data capable of training an ML model with excellent predicative power. We further validated the ability to add novel parts to the design space and fine tune our model to capture the behavior of these new parts and expand the design space to >10^6 circuits (results not shown).

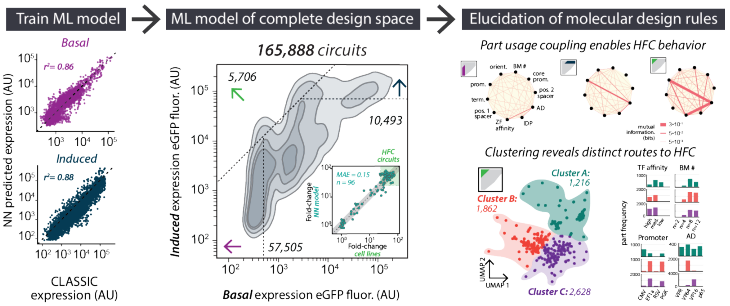


Fig 3: ML model trained on CLASSIC data to model the entire genetic design landscape

METHODS. CLASSIC gene circuit library assembly uses a custom hierarchical cloning scheme involving pooled assembly of part-containing "input" plasmids that combinatorially associate to libraries of yield diverse part combinations. Final multigene assemblies also incorporate a BFP gene containing barcode sequences. Circuit pools are then genomically integrated at single copy into HEK293T cells using a bacteriophage recombinase. Cells are then expanded, sorted into bins according to GFP intensity, and then RNA extracted and converted to cDNA for Illumina sequencing to quantitate barcode abundance across bins and compute circuit output. Nanopore sequencing performed on the library was analyzed via a custom sequence analysis pipeline. Data from both sequencing modalities were then combined to assign phenotype to circuit identities by comparing the barcode sequences from the two modes. For the ML model, data values were one-hot encoded to produce a 4 x 10 matrix for each circuit (number of parts x number of features in circuit). Using this matrix as an input, and expression in both basal and induced conditions as an output, we trained a fully connected neural network using the pytorch package in python. MSE was used as the loss function, and a randomly selected validation set was used to monitor training progress. Part-coupling was quantified by computing the pairwise mutual information between all pairs of part categories.

DISCUSSION. This work establishes the feasibility of combining long- and short-read NGS to perform massively parallel quantitative profiling of multi-kb length-scale genetic part assemblies in human cells. By enabling HT profiling of diverse combinations of genetic elements, CLASSIC holds potential as an approach for exploring the emergence of function from genetic composition across a range of organizational scales and phylogenetic contexts, including for viruses, bacterial operons, and chromatin domains. As our data demonstrate, this approach significantly expands the scope of inquiry for synthetic biology. We showed that data acquired using CLASSIC can be used to train ML models to accurately make predictions for out-of-sample circuit behavior and reveal design rules that may be non-intuitive and challenging to capture using biophysical modeling alone. While extensive recent work has used ML approaches to develop sequence-to-function models for various classes of genetic parts[11-13] our work serves as a critical starting point for developing ML/AI-based models of gene circuit function that use genetic part compositions as learned features. While our current work has focused on mapping a design space of 10^5 compositions, it may be possible to create predictive models for more complex circuits with far more expansive design spaces by using data acquired with CLASSIC to train high capacity deep-learning algorithms (e.g., transformers) which require much larger datasets than currently exist. Such approaches could work in black-box fashion, without the incorporation of regulatory or biophysical priors, or synergistically with existing mechanistic frameworks to create interpretable models that provide deeper insights into genetic design.

References

1. Cameron, D. E., Bashor, C. J. & Collins, J. J. A brief history of synthetic biology. *Nat Rev Microbiol* **12**, 381-390, doi:10.1038/nrmicro3239 (2014).
2. Yeung, E. *et al.* Biophysical Constraints Arising from Compositional Context in Synthetic Gene Networks. *Cell Syst* **5**, 11-24 e12, doi:10.1016/j.cels.2017.06.001 (2017).
3. Kinney, J. B., Murugan, A., Callan, C. G., Jr. & Cox, E. C. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc Natl Acad Sci U S A* **107**, 9158-9163, doi:10.1073/pnas.1004290107 (2010)
4. de Boer, C. G. *et al.* Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. *Nat Biotechnol* **38**, 56-65, doi:10.1038/s41587-019-0315-8 (2020)
5. Bogard, N., Linder, J., Rosenberg, A. B. & Seelig, G. A Deep Neural Network for Predicting and Engineering Alternative Polyadenylation. *Cell* **178**, 91-106 e123, doi:10.1016/j.cell.2019.04.046 (2019)
6. DelRosso, N. *et al.* Large-scale mapping and systematic mutagenesis of human transcriptional effector domains. *bioRxiv*, 2022.2008.2026.505496, doi:10.1101/2022.08.26.505496 (2022).
7. Angenent-Mari, N. M., Garruss, A. S., Soenksen, L. R., Church, G. & Collins, J. J. A deep learning approach to programmable RNA switches. *Nat Commun* **11**, 5057, doi:10.1038/s41467-020-18677-1 (2020)
8. Jones, E. M. *et al.* Structural and functional characterization of G protein-coupled receptors with deep mutational scanning. *Elife* **9**, doi:10.7554/eLife.54895 (2020)
9. Rai, K*, O'Connell, Ronan W*., et al. "Ultra-high throughput mapping of genetic design space." *bioRxiv* (2023): 2023-03.
10. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:1802.03426 (2018).
11. Voigt, C. A. Synthetic biology 2020-2030: six commercially-available products that are changing our world. Nat Commun 11, 6379, doi:10.1038/s41467-020-20122-2 (2020).
12. Valeri, J. A. et al. Sequence-to-function deep learning frameworks for engineered riboregulators. Nat Commun 11, 5058, doi:10.1038/s41467-020-18676-2 (2020).
13. Hollerer, S. et al. Large-scale DNA-based phenotypic recording and deep learning enable highly accurate sequence-function mapping. Nat Commun 11, 3551, doi:10.1038/s41467-020-17222-4 (2020).