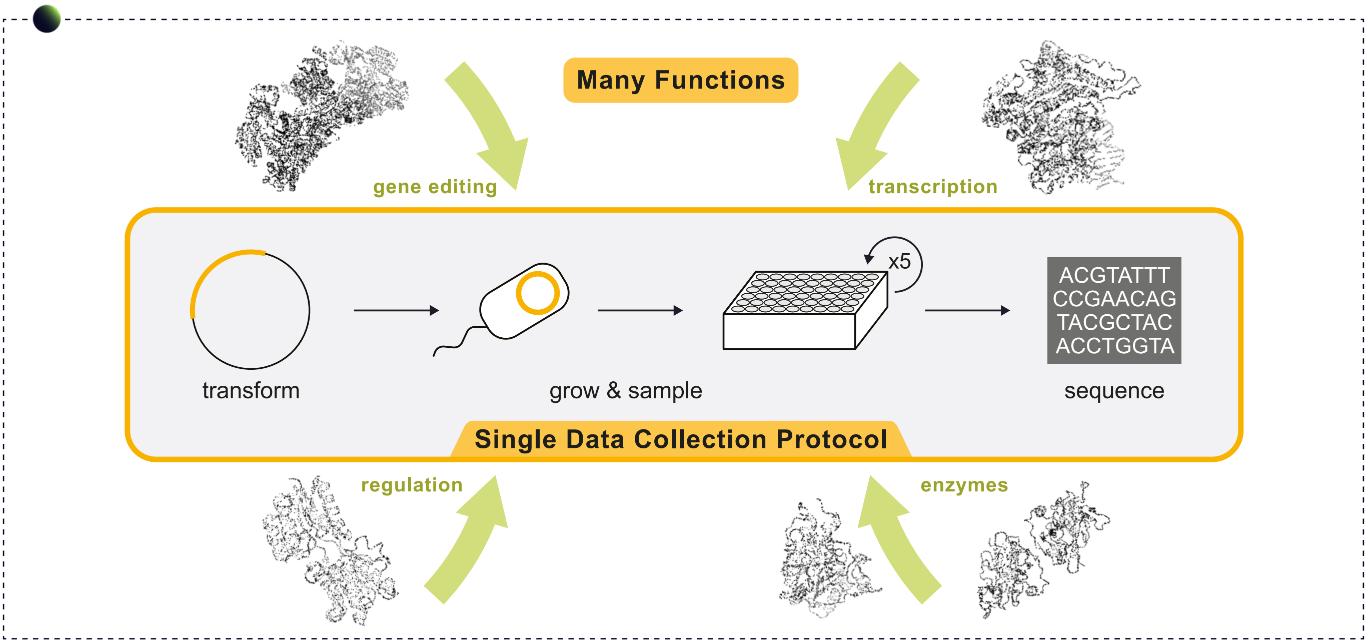
ALIGN TO INNOVATE

Design of a generalized platform for gathering protein sequence ➔ function datasets at scale



A proposed platform for gathering large protein sequence-to-function datasets.

- Uses a pooled growth-based assay to quantify protein function for < $0.05/sequence.

- Applicable to a wide variety of protein functions.

- New functions can be onboarded by validating a gene circuit and establishing a set of calibration variants.

DOI: 10.5281/zenodo.12521641

Contributors

Align to Innovate:

Peter Kelly - Program Director, Open Datasets Initiative

Dana Cortade - Technical Project Manager, Open Datasets Initiative

Proposal Co-Leaders:

David Ross - Living Measurement Systems Foundry, National Institute of Standards and Technology (NIST)

Simon d'Oelsnitz - Harvard Medical School, Harvard University

Erika DeBenedictis - Biodesign Lab, The Francis Crick Institute

Additional Proposal Authors:

Anjali Chadha - Biodesign Lab, The Francis Crick Institute

Oliver Hayes - Biodesign Lab, The Francis Crick Institute

Geoffrey Taghon - Living Measurement Systems Foundry, National Institute of Standards and Technology (NIST)

Mark Dörr - University of Greifswald

Stefan Born - Technische Universität Berlin

Reviewers:

Hassan Kane - Medium Biosciences

Han Spinner - Harvard Medical School Department of Systems Biology

Stephan Lane - Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA

Chloe Hsu - University of California Berkeley

Ben Lehner FRS FMedSci - Head of Generative and Synthetic Genomics, Wellcome Sanger Institute, Cambridge, UK; ICREA Professor, Systems and Synthetic Biology, CRG, Barcelona, ES; Honorary Professor of Biochemistry, University of Cambridge

Kevin K. Yang - Microsoft

Benjamin Scott - Global Institute for Food Security, University of Saskatchewan, Saskatoon, SK, Canada

Additional Acknowledgments:

We would also like to thank the following people for contributing their subject matter expertise to select sections of the proposal.

- **Craig Markin** - University of Manchester
- **Henning Redestig** - International Flavors & Fragrances
- **Tianhao Yu** - University of Illinois at Urbana-Champaign
- **Janet Matsen** - Benchling
- **Amelia Taylor**

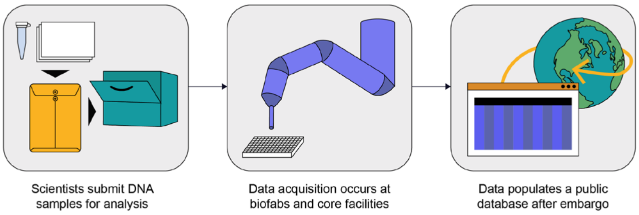
We also acknowledge the following people for their work on science communcation.

- **Rachel Sevey**
- **Olesia Bushkova**

DOI: 10.5281/zenodo.12521641

Overview

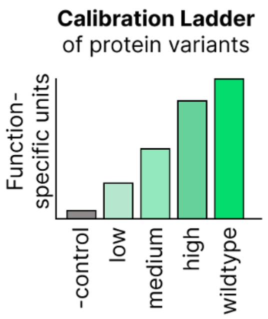
Align's protein function platform is designed to enable high-throughput collection of protein function datasets for a wide variety of functions. When complete, users will be able to send DNA samples to a collection facility for measurement, with costs that are subsidized or fully covered by Align. Users will receive their data as soon as it is available, and the data will also populate an open dataset after an embargo period.

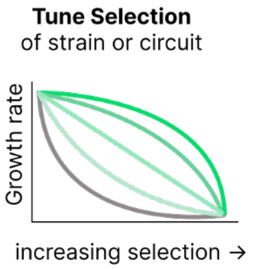

Scientists submit DNA samples for analysis
Data acquisition occurs at biofabs and core facilities
Data populates a public database after embargo

The protein function platform uses a pooled, growth-based assay to quantitatively measure protein function in high-throughput. It is compatible with measuring a wide variety of protein functions.

Onboarding measurements of a new protein function requires:

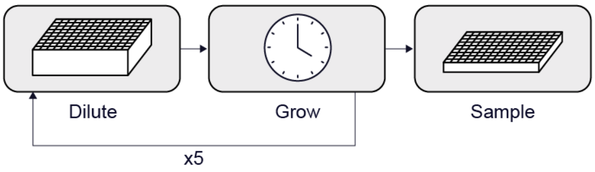
1. Identifying a calibration ladder of around 20 protein variants with known function
2. Tuning a selection circuit or selection strain so that the doubling time of bacterial cells is tied to the function of the protein variant they express.



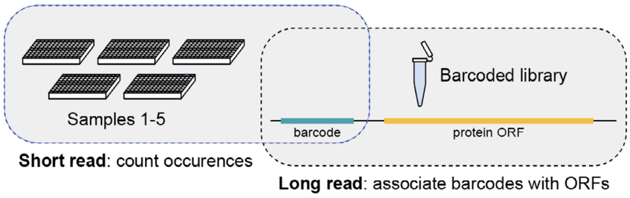


Once a function is onboarded, pools of up to 500k protein variants can be measured by creating a barcoded library of proteins, transforming the library into bacteria, growing the library in selective conditions (e.g. with antibiotic), and sequencing barcodes before and after growth to quantify differential growth rates. The calibration ladder is included in every pool, enabling quantitative measurement between batches.

To conduct the method, the pool is diluted into 96-well plates with varying selection strength (i.e. antibiotic concentration), grown, and samples are reserved for later sequencing. This process is executed five times to expand the operational range of the assay, with earlier time points capturing data on low-fitness variants, and later ones distinguishing high-fitness variants. The method can be executed by hand, but automation enables hands-off data collection with precise timings.



Long read sequencing associates the short barcodes with the full-length protein sequences, enabling functional measurements of complete ORFs.


Samples 1-5
Barcoded library
barcode
protein ORF
Short read: count occurences
Long read: associate barcodes with ORFs

DOI: 10.5281/zenodo.12521641

Table of contents

DOI: 10.5281/zenodo.12521641

Growth-Based Assays for Measuring Protein Function

Introduction

Protein functions, such as enzymatic activities, binding interactions, and membrane transport, exist as islands in the “archipelago” of the protein function landscape, which has loose and yet not fully understood relations to the corresponding sequence landscape. Machine learning (ML) algorithms have tried to bridge this gap, but today’s ML methods are still unable to find a general solution for predicting any protein’s function from its DNA sequence. This project proposes to develop an experimental platform and unified data ontology for collecting datasets from different functional ‘islands’ to build predictive models for individual protein functions. The experimental strategy uses a pooled, growth-based assay measured with DNA sequencing to create a simple, yet adaptable system that can be easily expanded to encompass new functions.

Models trained on this data will first succeed at predicting protein function within a single ‘island’, an individual family of proteins with a single function. As the dataset grows and more islands are sampled, the models will become more generalized and capable of predicting the function of protein sequences that are increasingly distant from those that have been directly measured. This will likely require many millions of data points, but a general solution for predicting any protein function from sequence would catalyze a transformation in the field of biology.

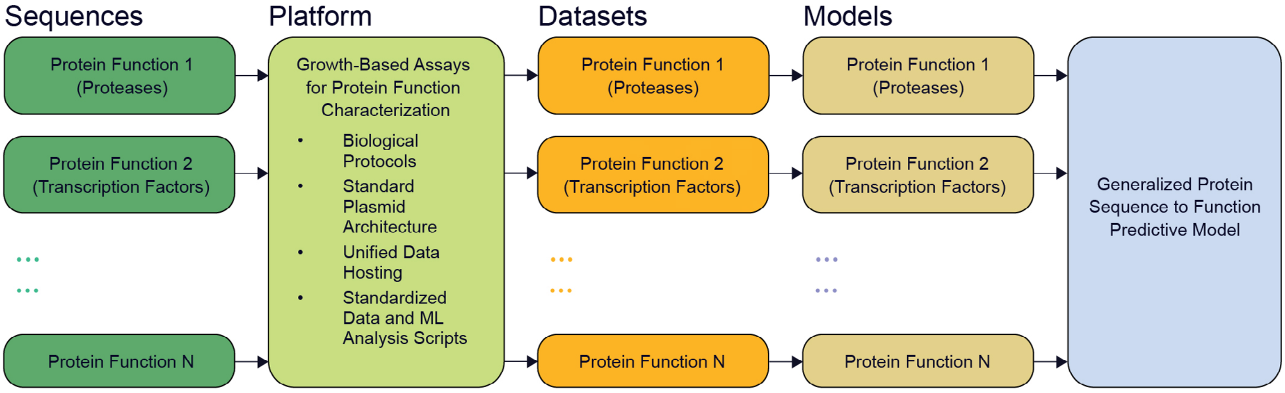

Sequences
Platform
Datasets
Models
Protein Function 1 (Proteases)
Protein Function 2 (Transcription Factors)
...
Protein Function N
Growth-Based Assays for Protein Function Characterization
• Biological Protocols
• Standard Plasmid Architecture
• Unified Data Hosting
• Standardized Data and ML Analysis Scripts
Protein Function 1 (Proteases)
Protein Function 2 (Transcription Factors)
Protein Function N
Protein Function 1 (Proteases)
Protein Function 2 (Transcription Factors)
Protein Function N
Generalized Protein Sequence to Function Predictive Model

Figure A1: Graphical abstract illustrating the goals of the Growth-Based Assays for Measuring Protein Function proposal.

DOI: 10.5281/zenodo.12521641

1. Proposal Summary

The goal of this proposal is to develop a platform for collecting protein function datasets to build models that can predict a given protein's function from its sequence. As the datasets grow and more islands are sampled, these models will become more generalized and capable of predicting the function of protein sequences that are increasingly distant from those that have been directly measured; eventually, models will be able to predict any protein function from its DNA sequence.

Building predictive models for protein function will require massive amounts of data. This proposal approaches the challenge of collecting this data with pooled, growth-based assays. We will link the activity of a gene to the ability of a cell to grow under selective pressure (e.g., antibiotic resistance), create hundreds of thousands of variants of that gene in different cells, culture the altered cells and challenge them under selective pressure (e.g., an antibiotic), and then sequence the pools of cells to measure abundance and quantify a protein's function. These high-throughput assays can produce quantitative functional characterization for hundreds of thousands of proteins per experiment at a cost of approximately $0.05 per sequence.

Crucially, growth-based assays can be used to quantify a wide variety of different protein functions, making this methodology a general-purpose platform for gathering large protein sequence-to-function datasets. The proposed dataset infrastructure will be developed with scalability in mind, and is designed to be flexible enough to accommodate different protein families and labs with different instruments, yet standardized enough to be easily parsable for ML.

Throughout this proposal, details are provided on the design for an experimental platform that leverages growth-based assays to collect quantitative data on individual protein functions. Broadly, this includes:

1. Onboarding a new protein function
   - Designing and tuning a plasmid system
   - Selecting assay controls
   - Optimizing the growth-based assay
2. Running a pilot-scale experiment (~100K variants)
3. Running full-scale experiments (>100K variants)
4. Conducting data analysis and model benchmarking

This platform is extensible to many protein functions simply by changing elements of the plasmid. This is demonstrated by applying the framework to the first two protein function targets: a DNA-binding dataset using transcription factors (at the National Institute of Standards and Technology, USA) and a protease specificity dataset (at The Francis Crick Institute, UK). The possible future expansion of this platform to additional protein functions at additional sites is also discussed.

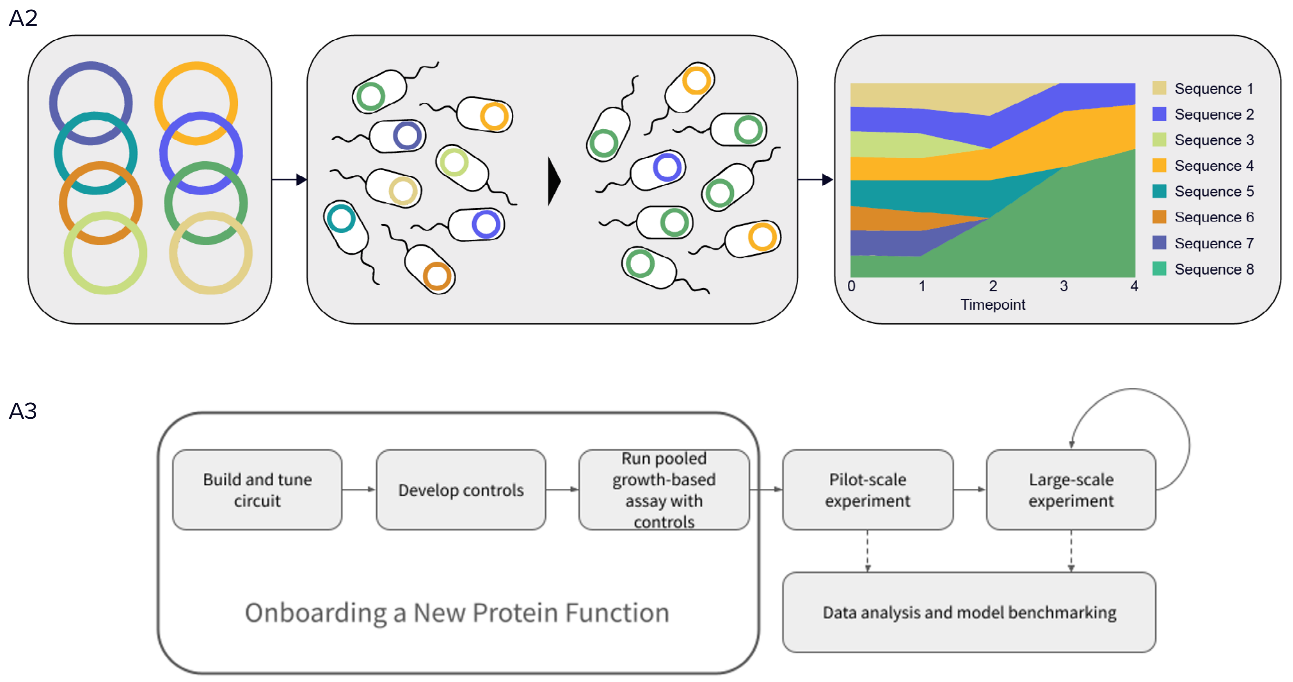


Figure A2, A3: Overview of growth based assay workflow and onboarding. A2) High-level overview of the pooled, growth-based assays pipeline. The plasmid library of barcoded proteins is created and transformed into host cells. Then the cells are pooled, grown and challenged. The resulting cells are then sequenced, counting the number of barcodes present, which can be translated back into a measurement of quantitative function. A3) Pipeline for onboarding, deploying, and analyzing of data for a new pooled, growth-based assay to measure protein function.

2. Definitions

Genetic Components:

Gene of interest (GOI): The protein sequence of which one is trying to determine the function.

Action site: The sequence in the selection cassette with which the GOI interacts. In this proposal, the action site for transcription factors (TFs) is an operator and the action site for proteases is a substrate. For future protein function analyses, the action site could be large or small sequences or complex formations of several proteins, depending on the protein function being onboarded.

Selection cassette: A region of the plasmid that is specific to each function, containing all variable components needed to form a circuit to interact with the GOI and report out its function (via antibiotic resistance, fluorescence, etc.) (described in section 4.2.4 Selection Cassette (Function-Specific)) .

Measurements:

Fitness: The calculated exponential growth rate of all the cells containing a given genotype. There are multiple methods used to measure fitness consistent with this definition, but all of those methods may not give comparable results. So, for additional clarity, the following definitions for methods to measure fitness are included:

Barcode-counting fitness: Fitness measured using normalized barcode counts in a pooled assay.

End-point-density fitness: Fitness measured using optical density (OD) measurements at the end point of sequential cell cultures grown in 96-well plates. This method is meant to approximate the barcode-counting method using plate reader measurements.

Growth-curve fitness: Fitness estimated from a fit to a growth curve (i.e., OD vs. time).

Singleplex function: A low-throughput, high-fidelity measurement of the activity of the GOI measured in cells using, for example, calibrated fluorescent measurements on a cytometry/plate reader. Results from singleplex function measurements are used to calibrate the large-scale, pooled function measurements.

Pooled function: A high-throughput measurement of the activity of the GOI is measured in cells using the pooled, growth-based assay.

Sequences and Plasmids:

Selection plasmid: The final plasmid that will be used in the pooled assay with a set combination of antibiotic resistance (AbR) gene and a ribosome binding site (RBS).

Normalization controls (aka “always-on” genotypes): Pairs of sequences of the GOI and action site that result in constitutively high expression of the resistance gene used for selection (e.g., TetA). When incorporated into the selection plasmid, they are referred to as normalization plasmids.

Calibration controls: Pairs of sequences of the GOI and action site with measured function (e.g., transcriptional output, protease activity) that varies over the full range of the pooled assay. When incorporated into the selection plasmid, they are referred to as calibration plasmids.

Blank plasmid: A plasmid similar to the selection plasmid, but without fluorescent control. Used for background subtraction in fluorescent measurements of function on a flow cytometry or plate reader.

3. Context, Significance, and Impact

Estimating a protein's structure used to be an intractable problem that required long and tedious wet-lab experiments. A large leap forward occurred in 2021 with the advent of AlphaFold, a model able to predict the structure of a protein solely from the DNA sequence encoding it with accuracy meeting or exceeding experimental approaches[1].

From a practical engineering standpoint, however, it is more valuable to know the function of a protein than its structure. The inevitable evolution of predictive models in protein engineering will be to craft a model that can predict a protein's function, not just its structure, from a DNA sequence.

The ability to predict the function of natural and designed proteins would accelerate both basic science and R&D by enabling researchers to focus attention on protein targets most likely to work for a given application, thus reducing the number of experiments to be performed. This is already proving to be true for predicting some properties, like stability, of natural proteins[2]. However, to generate these powerful predictive models for more protein functions and for designed proteins, high-fidelity datasets on protein must be designed from the start for machine learning.

Expressing and measuring the function of proteins is expensive, time-consuming, and challenging; thus, the largest, consistent datasets tend to be collected by industry and kept private. Publicly-available datasets on protein function are overwhelmingly small in size, few in number, and reported in different functional units. Until recently, there were approximately 10-15 open-source, high-quality datasets that relate protein sequence-to-function, all of which contain fewer than 100k data points (see Hsu, C. et al[3]: Figure 4; see Supplementary Table 1). These datasets address different protein functions, and were collected using different assays with different units. This poses problems for obtaining a consistent type of measurement that can be used to build generalizable models between multiple protein functions. To combat these issues, there has been an increase in efforts to curate[2] or create larger datasets[4]. In order to move towards a generalizable model for protein function, there needs to be a continual push to align incentives and to increase the effort for generating large, consistent datasets in the public domain.

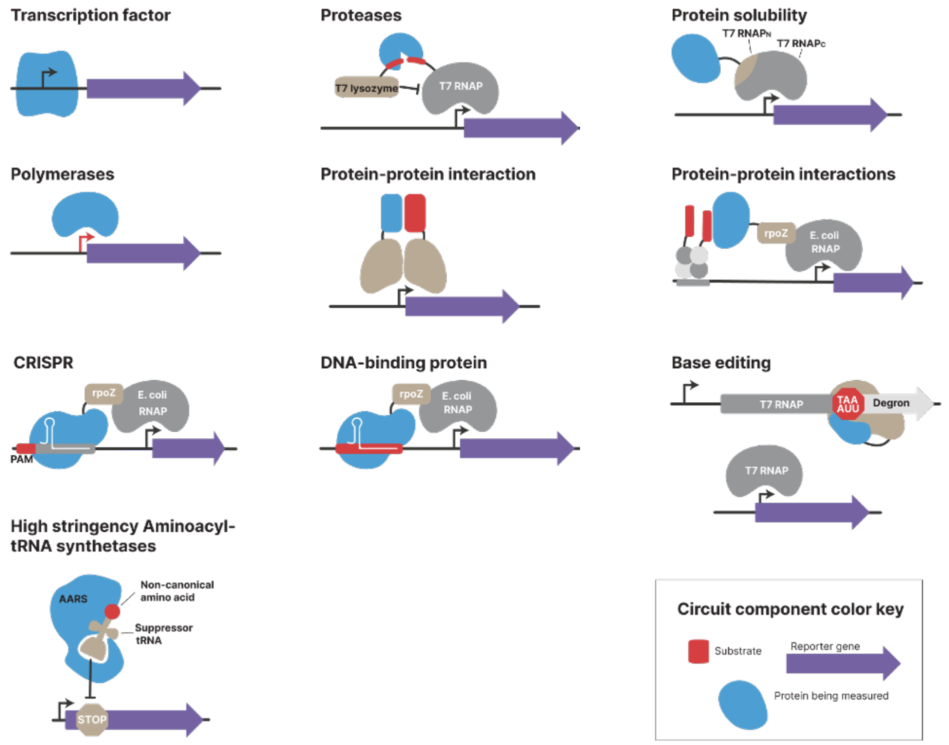


Figure A4: Examples of the wealth of circuits that can be onboarded from the existing literature. These include transcription factors; proteases[8,9]; two different circuits for measuring protein solubility[10,11]; polymerase promoter specificity[12]; two different circuits for measuring protein-protein interactions[13–15]; CRISPR specificity[16]; DNA binding proteins[17]; base editors[18,19]; and aminoacyl tRNA synthetases[20].

DOI: 10.5281/zenodo.12521641

Several shortcomings with existing protein function datasets have prompted us to develop a platform for gathering a large sequence-to-function protein dataset. Our proposal will address these deficiencies as follows:

1. Datasets are small in size.

The proposed assay can gather hundreds of thousands of data points at once, making it approximately one order of magnitude larger than the largest existing dataset each time an experiment is run.

2. Function datasets are few in number.

While this proposal demonstrates the utility of a platform with protease and DNA-binding proteins, growth-based assays are also applicable to a wide variety of other protein functions (Fig. A4). The automation and analysis platform presented here need only be engineered once before it can be applied to a wide variety of protein functions by using different plasmids.

3. Datasets are reported in different functional units.

All growth-based assays (such as the ones presented here) measure sequencing read counts, and using calibration controls generates quantitative function scores with real units. Additionally, the data acquisition methods and data ontology we present here are consistent across functions, so bioinformatics and machine learning can be applied to correct common failure modes and iteratively improve the platform.

4. Function data is challenging to collect.

Growth-based assays are amongst the simplest ways to collect function data at this magnitude and are compatible with many institutions' equipment capabilities. Additionally, many parts of the workflow, including sequencing, can be readily outsourced.

The proposed pooled, growth-based assay format was chosen because of its extensibility. Growth-based assays are compatible with a number of existing selection cassettes for additional protein functions and biophysical properties (Fig. A4).

We have chosen two selection cassettes to demonstrate the versatility of this platform: DNA-binding proteins and proteases. These initial targets were selected because they provide a combination of low technical risk with high-value applications. DNA binding proteins are important regulators, with the proposed transcription factor (TF) protein families having varied but widespread clinical relevance (e.g., antibiotic resistance[5]) and importance in synthetic biology (e.g., chemical sensors[6] and logic gates[7]). Similarly, proteases have been used as therapeutics for many diseases[8,9]. Both targets present a low engineering risk because the initial collaborating labs have previously worked with the function-specific circuits in their respective selection cassettes.

One limitation of growth-based assays worth mentioning is their inability to differentiate between factors impacting a protein's functional measurement. For example, a poorly expressing protein with high activity could have a similar functional score to a highly expressing protein with low activity. As a future step, we plan to onboard measurements of protein expression, folding, stability, and other to help deconvolute these unknowns.

4. Experimental Design Choices

4.1 Host

This platform will be created for use in *E. coli*, but the exact strain will be depend on each protein function and its associated target proteins. It is ideal to have a strain with relatively high transformation efficiency, which is required for screening large protein libraries. As an example, MG-1655, a strain with the lac operon fully deleted, was one of the strains chosen for use in the Transcription Factor circuit due to the fact that LacI is a target protein for development.

Any strain used for these experiments will have its genotype verified via whole genome sequencing before proceeding. The strain used for a particular growth-based assay experiment will be recorded in the experimental data.

In future experiments, the host selection would ideally be expanded into multiple industrially and academically relevant hosts, like yeasts.

4.2 Plasmid Design

A single-plasmid system will be used for data generation in growth-based assays. Broadly, the plasmid has four sections: a barcode region, a plasmid backbone, a GOI region, and a function-specific selection cassette. The single-plasmid system allows for variations in the protein open reading frame (ORF) and/or circuit to be covered by the same nanopore sequencing read and eliminates the need for co-transformations or custom competent cells to support a two-plasmid system.

4.2.1 Barcode Region

The barcoding system involves tagging both the backbone and insert with "half" barcodes to allow the library to be assembled with a unique barcode for each variant. The priming locations for PCR-amplifying the barcode should be kept constant amongst different protein function circuits so that this set of primers only needs one validation. These barcodes will be ordered as primers (DNA oligos) that add the random bases onto the ends of the GOI or plasmid backbone fragment, plus the other sequences needed for cloning to adjacent segments of the plasmid. Click here to see the Benchling (Benchling, San Francisco, USA) design of the complete barcode region.

Component	# of base pairs	Explanation
Primer binding regions	24-27	Flank the barcodes that have been used in the past with success. Have GC content between 50-60%.
Spacer sequences	9	Sandwiched between the primer binding regions and the barcodes, making the entire barcode region 160 bp. A length for getting good coverage from an NGS read, while also producing a PCR product that can be easily distinguished from the long oligos used for the PCR.
Two "half" barcodes	33 each	Design avoids multiple barcodes being assigned to the same GO/action site pair. Barcodes incorporate interspersed S and W nucleotides to aid analysis.
Homology region	25	For Gibson assembly. Has 52% GC content and a Tm of ~65oC with hairpin structures avoided.
Overall length	160	Long enough to allow for robust magnetic-bead-based PCR cleanup, while being short enough so both half barcodes can be read within a 150 bp read from either direction.

Table A1: Barcode region design descriptions and rationale.

4.2.2 Plasmid Backbone

Component	Description
Barcode region	This is the half of the barcode sequence in the plasmid backbone region (see Table A1).
KanR	Kanamycin resistance gene marker for plasmid maintenance.
Ori	p15A low copy origin of replication.
Terminators	Used to insulate the plasmid backbone from the protein variant and circuit regions.

Table A2: Plasmid backbone component descriptions.

4.2.3 Gene of Interest (GOI) Region

The gene of interest region contains an operon with the protein sequence to be expressed in the opposite orientation to the genes in the circuit region. This prevents unintended expression of genes in the circuit. It also contains barcode and circuit region homology for ease of cloning.

Component	Description
Half-Barcode	This is the half of the barcode sequence in the gene of interest region (see Table A1).
Gene of Interest (GOI)	A promoter (constitutive or inducible) and medium-strength RBS drive the expression of the GOI. Tuning can be performed at the transcriptional level (different strength promoters) and at the translational level (variable RBS design). It is best to use a bicistronic leader peptide for context-independent tuning of translation strength.
Homology	Region that is consistent across all plasmids and circuit designs. Provides a standard sequence for priming and/or scarless assembly.

Table A3: Gene of Interest (GOI) component descriptions.

4.2.4 Selection Cassette (Function-Specific)

The selection cassette is a region of the plasmid specific to each function. Figure A9 and Table A4 illustrate and describe the generalized parts that are relevant to the two protein functions being onboarded in this proposal; however, each newly onboarded function may or may not utilize all of these components. The selection cassette contains the function-specific circuit components with which the GOI will (e.g., operators or substrates) and all of the necessary reporters produced by the interaction (e.g., antibiotic resistance gene (AbR) for fitness measurements and a fluorescent protein of choice used for function measurements during development). For protein functions that are onboarded in the future, the inclusion and positions of these components can be changed to suit the circuit and reporters for the function of interest.

Component	Description
Antibiotic resistance gene (AbR)	Either tetracycline resistance or zeocin resistance, depending on the protein function.
Fluorescent Protein (FP)	Used for quantitative measurements of protein function. This could be any number of a variety of proteins dependent upon a particular circuit design.
Function-specific circuit component	A circuit-specific region of the plasmid that interacts with the GOI. This functional element can repress or activate the expression of the downstream genes (AbR and FP). May contain the action site and the promoter for the downstream genes.
Homology	Region that is consistent across all plasmids and circuit designs. Provides a standard sequence for priming and/or scarless assembly.
RiboJ element	Used to give a more reproducible transcript start sequence. It self-cleaves at a defined position, causing the variable starting portion of the transcript to be cleaved off.
Ribosome binding site (RBS)	Upstream of AbR and upstream of FP. Must be changeable to facilitate tuning for different proteins/functions.

Table A4: Selection cassette component descriptions.

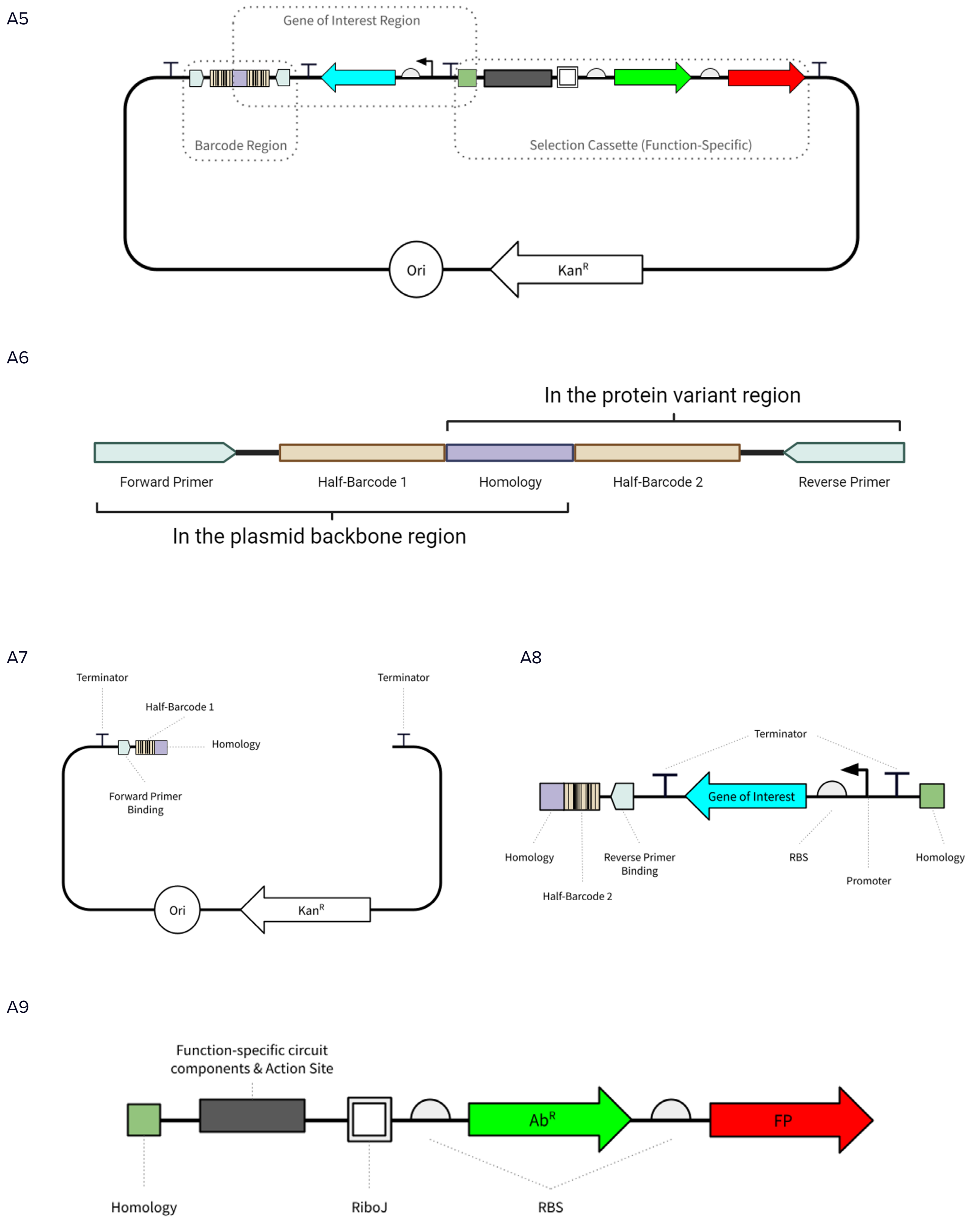


Figure A5-A9: Plasmid Design. A5) The generalized plasmid map including the protein backbone, barcode region, gene of interest region, and selection cassette. A6) Scheme of the barcoded region. A7) Scheme of the plasmid backbone. A8) Scheme for the gene of interest region. A9) Scheme for the selection cassette region.

4.3 Protein Target Selection Strategies

Under the archipelago of protein function metaphor, each protein function could be seen as an "island" composed of all of the protein families that exhibit the function of interest. To begin mapping out each functional island, an assay must be developed using well-characterized or prototypical members of a protein family. This allows for an assessment of whether or not the assay is performing as expected when compared to published methods (e.g., the wild-type RamR (Uniprot ID: A0A0F6AY66) will be used to develop the TF assay). During the pilot stage of development, much more of the local sequence space surrounding the initial development targets is sampled (e.g., more variants of RamR). Finally, during the large-scale data collection phase, an expanded set of the local sequence space variants can be explored, as well as divergent sequences and new protein families (e.g., adding both well-characterized and uncharacterized proteins in RamR's family (the TetR family) and additional TF families).

4.3.1 Development and Control Targets

These targets are pairs of sequences for the GOI and action site, depending on the protein function being onboarded, that are used during assay development and as ongoing controls. These sequences should be previously validated or be well-characterized in the literature. This group is composed of 15-25 sequences outlined in Section 5: Onboarding a New Protein Function. For example, these sequences could include a wildtype (WT) protein sequence and sequences of its characterized point mutant variants spanning a range of activity, coupled with the same action site sequence (e.g., WT LacI and a few of its characterized point mutant variants, coupled with the same operator sequence).

4.3.2 Pilot-Scale Targets

These targets should result from a deeper sampling of the types of sequences used for development and control targets, such as point mutations in a WT TF protein, or additional action site sequences, such as more operators. In general, the pilot-scale pooled assay will be bottle-necked to a plasmid library of 100,000 plasmids. These plasmids can contain combinations of variations in the sequence of the GOI and of the action site (e.g., TF and operator site or protease and cleavage site). Depending on the protein function being onboarded and the desired final predictive model, the resulting split between variation in the GOI and the region it is acting upon will change (i.e., the amount of TF diversity versus operator site diversity).

4.3.3 Large-Scale Targets

Depending on the protein function being onboarded, the large-scale collection can increase diversity of proteins used as GOIs (e.g., additional proteases) and/or increase the diversity of sequences used in the action site (e.g., additional substrates). To increase GOI diversity, new proteins within the pilot experiment family can be analyzed, as can proteins from additional families within the same functional island. Ideally, each new protein family in a functional island will have the same distribution of well-characterized and uncharacterized protein sequences to be explored as GOIs. These include:

1. Proteins that are well-characterized to cross-validate results with existing literature.
2. Uncharacterized proteins of two types: those that can be predicted by some other means (e.g., using homology data) and those that cannot. This will allow us to see if the results align with those predicted by homology data and to extend the resulting model's predictive power by adding proteins that cannot currently be predicted by other means.

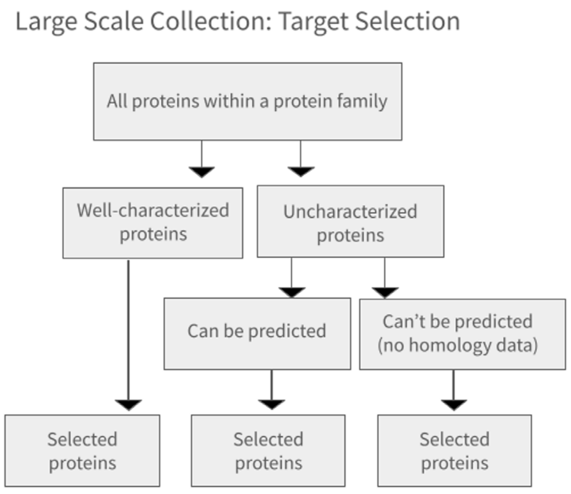

Large Scale Collection: Target Selection
All proteins within a protein family
Well-characterized proteins
Uncharacterized proteins
Can be predicted
Can't be predicted (no homology data)
Selected proteins
Selected proteins
Selected proteins

Figure A10: How to choose targets for large-scale data collection.

DOI: 10.5281/zenodo.12521641

5. Onboarding a New Protein Function

Each new protein function only needs one onboarding. The process for doing this is broken down into the following stages (Figure A11):

Stage 1:

Tune selection stringency and optimize the selection plasmid.

- Identify tuning sequences to optimize the selection plasmid and select which antibiotic marker to use.

Stage 2:

Choose controls.

- Choose normalization controls that have high fitness in every assay condition.
- Choose calibration controls that span the assay's desired dynamic range.

Stage 3:

Conduct BarSeq validation and the first pooled assay.

- Validate the BarSeq protocol.
- Run a first pooled assay with just the normalization and calibration plasmids. This step is designed to be quick to sequence and analyze, allowing for rapid troubleshooting and iteration of the first pooled assay.

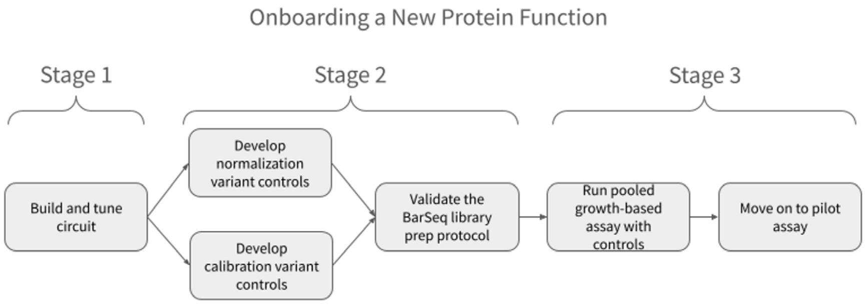

Onboarding a New Protein Function
Stage 1
Stage 2
Stage 3
Build and tune circuit
Develop normalization variant controls
Develop calibration variant controls
Validate the BarSeq library prep protocol
Run pooled growth-based assay with controls
Move on to pilot assay

Figure A11: Process diagram for onboarding a new protein function.

5.1 Stage 1 - Tune Selection Stringency and Optimize the Selection Plasmid

The first stage of onboarding a new protein function is to detail the selection plasmid. To simplify method development and onboarding, the plasmid should be designed to use the same overall layout and many of the same specific components/parts for all protein functions. However, each protein function will require some unique plasmid components and optimization of the control elements (e.g., promoters and RBSs). The components that need to be optimized are in the GOI region and the selection cassette (function-specific). To avoid the expense and effort of a full combinatorial test of the plasmid design space, the plasmid optimization is separated into two steps: tuning the regulation components related to function and tuning the components used for the measurements (Fig. A12).

The function regulation components that may require tuning include the promoter and RBS controlling the GOI and any other regulatory components within the function-specific circuit region. When tuning the function components, a fixed set of measurement components should be used: a medium-strength RBS controlling the zeocin resistance gene (AbR) and a strong RBS controlling the fluorescence protein.

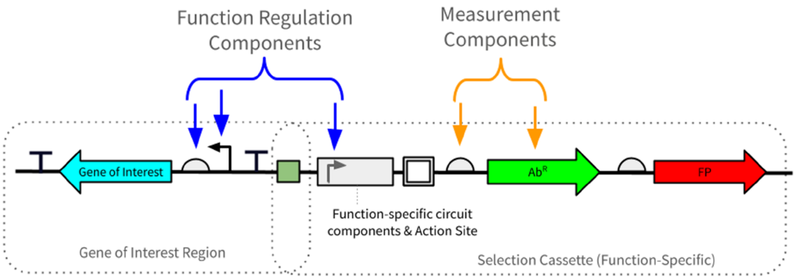

Function Regulation Components
Measurement Components
Gene of Interest
Ab^R
FP
Function-specific circuit components & Action Site
Gene of Interest Region
Selection Cassette (Function-Specific)

Figure A12: Components in the GOI region and selection cassette that can be varied to tune the selection plasmid.

Tune the function regulation components in the following manner:

1. Use previous experience with similar protein functions/ circuits or examples from the literature to choose the initial set of function regulation components.
2. Measure the fluorescence response in at least two conditions that are expected to give very low and very high activity/function for the gene of interest using the Singleplex Assay for Function protocol[21].

   Note: The two conditions can be achieved using different sequence variants for the GOI and/or by using additives that modulate the function of the GOI. For example, the TF dataset will use the WT sequence of the RamR transcription factor both with and without induction with tetrahydropapaverine (THP). Based on past experience with RamR, the ratio of fluorescence measured with versus without induction is expected to be over 100.

   Analyze the resulting data from the function measurements according to the steps outlined in the data analysis section of the Singleplex Assay for Function Measurements protocol[21].

Stage Gate 1

The ratio of fluorescence measured between the high and low activity/function conditions must be large enough to cover the expected dynamic range. If it is significantly lower than expected, the strength of the function components must be adjusted and re-tested. As an example, the ratio for the transcription factor plasmids is expected to be greater than 100.

Once the appropriate function components have been identified, the measurement components can be tuned to determine the best combination of AbR gene and RBS strength to create the final selection plasmid. Different sequences for the GOI and/or different additives to vary the function will be used, and the influence of using different combinations of AbR gene and RBS strength on the measurable output range will be monitored using several different concentrations of the selection antibiotic. Ideally, in at least one of the tuning plasmids, the accessible output range should span two-three orders of magnitude when measured with different concentrations of the selection antibiotic. Whichever plasmid and RBS combination provides the best output range will be selected as the final selection plasmid.

The following steps are used to tune the measurement components and select the final plasmid:

1. Determine the function of the tuning plasmids in each condition by measuring the signal of the fluorescent protein expressed in series with the AbR gene. Cytometry is the best measurement method, but using a plate reader (FL/OD) is also acceptable if there is minimal change in the growth phenotype (i.e., OD per cell density) over all assay conditions. These measurements should be calibrated using standards. To perform this measurement, follow the procedure outlined in the Singleplex Assay for Function Measurements protocol[21].
   a. Check that this function measurement has sufficient dynamic range for each protein type. In particular, ensure that the dynamic range of the fluorescent protein measurement is not significantly affected by the choice of RBS used to control the selection resistance gene. To do this, after measuring the function in plasmids with different RBS strengths, check that the measured dynamic range (ratio of the highest to lowest mean fluorescence) is comparable for each choice of RBS. If it is not, measure fitness and function with different versions of the plasmid (i.e., measure fitness with one choice of RBS and function with a different RBS).
2. Measure the fitness of each variant in each condition using a growth-based assay described in the Singleplex Assay for Fitness Measurements protocol[22]. This is similar to the final pooled assay but all plasmids are measured individually. Use the combinations of the following conditions with a minimum of two replicates (n=2) each:
   a. Suggested antibiotic concentrations (for initial testing):
      - For the Tet plasmids: 0, 0.625, 1.25, 2.5, 5, and 10 µg/mL.
      - For the Zeo plasmids: 0, 25, 50, 100, 200, and 400 µg/mL.

   b. Additives: depending on the assay.
3. Analyze the resulting data from the function and fitness measurements according to the data analysis section of the Singleplex Assay for Fitness Measurements protocol[22] to see if any of the antibiotic resistance x RBS strength combinations resulted in an acceptable circuit.

Stage Gate 2

The final selection plasmid (i.e., the proper resistance gene and RBS strength) must satisfy the following criteria:

- Fitness has a smooth, measurable change over the relevant range of function for one or more antibiotic concentrations.
- Fitness with zero antibiotic is approximately constant over the relevant range of function (i.e., the resistance gene does not cause a significant fitness defect).
- The function measurement should have a dynamic range that is not limited by the choice of the selection cassette components.

If none of the initial circuit/plasmid designs satisfy these criteria, build additional designs (i.e., re-tune the system).

DOI: 10.5281/zenodo.12521641

5.2 Stage 2 - Choose Controls

This stage aims to identify normalization and calibration controls using the optimized circuit. These controls will be used to further validate the circuit and will be included in all future experimental pools. By the end of this stage, two normalization controls will be developed that have a high fitness at every antibiotic concentration (and with every additive). In addition, 10-20 different calibration controls will be developed that densely sample the functional range and that can be used to calibrate and quantify function for other proteins in the library when the assay is run at scale. Each of the final controls will be assigned a permanent barcode sequence that matches the pattern of the library barcodes, but with the W and S positions switched. Selected controls will carry the permanent barcodes into the pooled assay context.

5.2.1 Choose Normalization Controls

Normalized controls will be selected to be "always-on", regardless of conditions, and will be used to normalize signals during the pooled assay. These protein sequences will be used in the pooled assay and have a permanent barcode. The sequences can be sourced in two ways:

1. After transforming a plasmid library, grow a diluted culture of the library with the selection antibiotic, then plate the resulting culture onto agar that also contains the selection antibiotic. Pick colonies to assay for verifying high fitness across conditions.
2. Design or source from the literature protein sequences that are expected to be "always-on".

Start by testing five normalization control candidates, because some of them may not have the desired fitness and/or may have unexpected sequences. Select the best two controls for use in the final pooled assay. The goal is to generate graphs similar to the examples below for the normalization controls (Fig. A14).

Follow this general procedure to assess the normalization control candidates:

1. Verify the entire plasmid sequence for each plasmid individually using sequencing. Check each of the plasmids for unexpected mutations (i.e., to sections of the plasmid that should be constant) that might impact the measured function or barcode sequencing.
   a. After transformation, plate culture and pick three colonies for each plasmid, grow up to the stationary phase in liquid culture and make a glycerol stock for each colony/clone.
   b. Then, use a scraping from that glycerol stock to start a new culture, grow up enough volume to stationary phase, then mini-prep plasmid and send it for whole plasmid sequencing.
   c. Those glycerol stocks should then last the duration of the project (across multiple library measurements) requiring whole-plasmid sequencing of it once.
2. Measure the fitness of each normalization plasmid in each condition following the Singleplex Assay for Fitness Measurements protocol[22]. Use combinations of the following conditions with a minimum of two replicates (n=2) each:
   a. Antibiotic concentrations:
   - If the Tet plasmid was chosen in stage 1: 0, 0.625, 1.25, 2.5, 5, and 10 µg/mL.
   - If the Zeo plasmid was chosen in stage 1: 0, 25, 50, 100, 200, and 400 µg/mL.
   - These concentrations may need to be adjusted after Stage 1 to match the best set of concentrations for the chosen resistance gene and RBS.

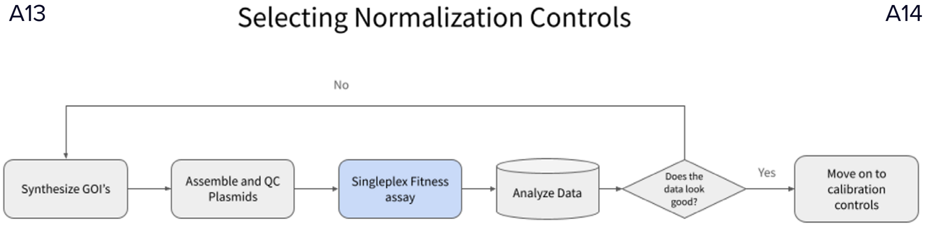
   b. Additives: Use a similar set of additives and additive concentrations used in the pooled assay.
3. Analyze the resulting data from the fitness measurements according to the pipeline described in the data analysis section of the Singleplex Assay for Fitness Measurements protocol[22] to determine which two of the variants have the highest and most stable fitness across all conditions (e.g., antibiotic concentrations or additives).

Stage Gate 3

The two selected normalization plasmids should have the following:

- No mutations in the plasmid barcode region and no homopolymer repeats longer than three bases in the barcode sequences (4.2.1 Barcode Region).
  - If none of the initial set of five candidate plasmids satisfy these criteria, troubleshoot the plasmid assembly protocol or reagents (e.g., the oligos used as primers to add the barcodes).
- Constant, high fitness at all antibiotic concentrations and all additive conditions, ideally within 20% of the fitness with zero antibiotic.
  - If none of the initial set of five candidate plasmids satisfy this criterion, re-evaluate the design or selection processes used to create the candidate normalization plasmids.

A13



A14

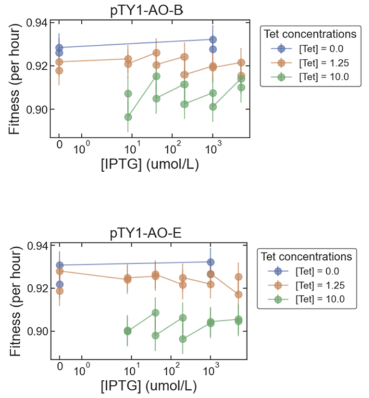


Figure A13-A14: Normalization controls. A13) Process flow and decision tree diagram for developing normalization controls. A14) Example graphs from previous NIST experiments illustrating the fitness measurements of two different normalization controls. Note: the graphs for both of the normalization plasmids above are always constant and have high fitness.

5.2.2 Choose Calibration Controls

Select calibration controls to densely cover the desired dynamic range of the assay and use them to turn the pooled assay measurements into quantitative measurements of protein function. These protein sequences will be used in the pooled assay and need permanent barcodes. These protein sequences can be sourced in two ways:

1. Plate a diluted portion of a library that has been transformed into cells and then pick colonies. This is the simplest option for generating calibration sequences.
2. Design calibration sequences using prior knowledge from the literature.

Choose at least 10-20 calibration controls to have data density of 30 points spread across two to three orders of magnitude in function.

To assess the calibration control candidates, do the following:

1. Construct a 'blank plasmid'. For all function measurements, measure the fluorescence background with a non-fluorescent control. Modify an existing normalization plasmid to remove the fluorescent protein gene.
   a. The blank plasmid will not be used in the pooled assay, so it does not need to be barcoded.
2. Verify the entire plasmid sequence for each plasmid individually using sequencing. Check each of the plasmids for unexpected mutations (i.e., to sections of the plasmid that should be constant) that might impact the measured function or barcode sequencing.
   a. After transformation, plate culture and pick three colonies for each plasmid, grow up to the stationary phase in liquid culture and make a glycerol stock for each colony/clone.
   b. Then, use a scraping from that glycerol stock to start a new culture, grow up enough volume to stationary phase, then mini-prep plasmid and send it for whole plasmid sequencing.
   c. Those glycerol stocks should then last the duration of the project (across multiple library measurements) requiring whole-plasmid sequencing of it once.
3. Select a subset of the calibration sequences that sparsely span the range of function desired for the pooled assay.
   a. Use at least three calibration sequences (with induction, this should result in~ 10 points).
4. For the selected subset, measure both fitness and function, using the Singleplex Assay for Fitness Measurements protocol[22] and the Singleplex Assay for Function Measurements protocol[21] respectively. The goal here is to verify that the pooled assay calibration procedure is likely to work using the singleplex fitness assay in place of barcode sequence counting.
   a. For function measurements, do not use the selection antibiotic.
   b. For both fitness and function, use the same set of additives and additive concentrations as planned for the pooled assay.
5. Analyze the function and fitness data following the data analysis section of the Singleplex Assay for Fitness Measurements protocol[22] to produce a plot similar to that in Figure A16.
6. Use the fitness data to determine the length of incubation periods. For repeated growth cycles, the length of culture time must be adjusted so that the end-point cell density is either constant or slightly declining over the course of the repeated time points.
7. Use the fitness versus function data to determine the optimal antibiotic concentrations following the procedure described in the data analysis section of the Singleplex Assay for Fitness Measurements protocol[22]. Choose multiple antibiotic concentrations to provide good sensitivity over the entire range of functional output.

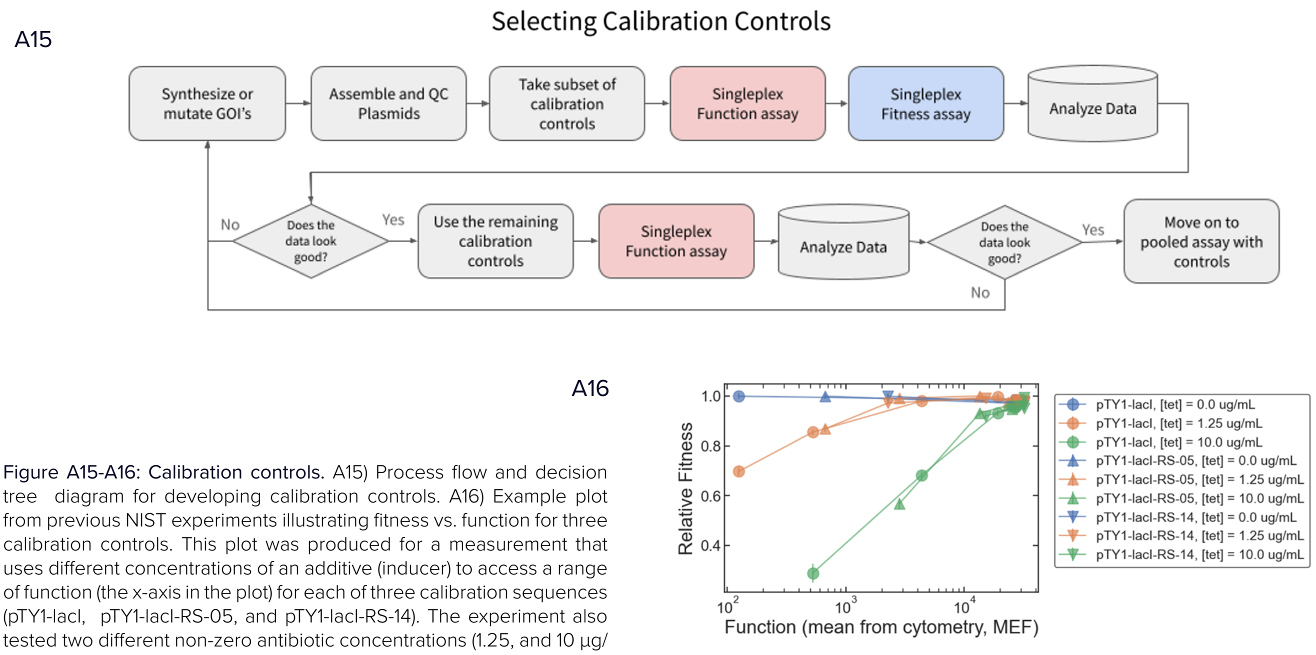

A15
Selecting Calibration Controls
Synthesize or mutate GOI's
Assemble and QC Plasmids
Take subset of calibration controls
Singleplex Function assay
Singleplex Fitness assay
Analyze Data
No
Does the data look good?
Yes
Use the remaining calibration controls
Singleplex Function assay
Analyze Data
Does the data look good?
Yes
Move on to pooled assay with controls
No
A16
Relative Fitness
1.0
0.8
0.6
0.4
10^2
10^3
10^4
Function (mean from cytometry, MEF)
pTY1-lacI, [tet] = 0.0 ug/mL
pTY1-lacI, [tet] = 1.25 ug/mL
pTY1-lacI, [tet] = 10.0 ug/mL
pTY1-lacI-RS-05, [tet] = 0.0 ug/mL
pTY1-lacI-RS-05, [tet] = 1.25 ug/mL
pTY1-lacI-RS-05, [tet] = 10.0 ug/mL
pTY1-lacI-RS-14, [tet] = 0.0 ug/mL
pTY1-lacI-RS-14, [tet] = 1.25 ug/mL
pTY1-lacI-RS-14, [tet] = 10.0 ug/mL

Figure A15-A16: Calibration controls. A15) Process flow and decision tree diagram for developing calibration controls. A16) Example plot from previous NIST experiments illustrating fitness vs. function for three calibration controls. This plot was produced for a measurement that uses different concentrations of an additive (inducer) to access a range of function (the x-axis in the plot) for each of three calibration sequences (pTY1-lacI, pTY1-lacI-RS-05, and pTY1-lacI-RS-14). The experiment also tested two different non-zero antibiotic concentrations (1.25, and 10 µg/mL, plotted with different colors).

Stage Gate 4

For each antibiotic concentration, the fitness versus function values need to lie along a consistent curve; curves for different antibiotic concentrations must have systematic variation, with higher antibiotic concentration resulting in lower fitness. In the example plot in Fig. A16, all of the data shown in orange fall along a consistent curve (for three calibration sequences measured with 1.25 µg/mL tet), and all of the data shown in green fall along a different curve (for three calibration sequences measured with 10 µg/mL tet). The green curve shows lower fitness than the orange curve, as is expected.

If the fitness versus function values do not lie along consistent curves for each antibiotic concentration, try the following troubleshooting steps:

- Check the reproducibility of the singleplex fitness and function measurements.
- Select and measure a different subset of calibration sequences. Do just one or two sequences not follow a consistent curve?
- If the plates were pipetted with an automated liquid handler, check the log files to verify that each well was pipetted correctly.

After passing the stage gate, continue to do the following:

1. Measure the function of each of the remaining calibration sequences following the Singleplex Assay for Function Measurements protocol[21], with a minimum of two replicates (n=2) each:
   a. For function measurements, do not use the selection antibiotic.
   b. Use the same set of additives and additive concentrations as planned for the pooled assay.
2. Analyze the resulting data from the function measurements according to the data analysis section of the Singleplex Assay for Function Measurements protocol[21] to determine which of the 10-20 variants best cover the dynamic range.
   a. This data will also be used to validate that the first pooled assay was successful in order to pass stage gate 6.

5.3 Stage 3 - Conduct BarSeq Validation and the First Pooled Assay

5.3.1 Validate the BarSeq Library Prep Protocol

This validation needs to be performed initially once for each site and then repeated only whenever new batches of reagents are used (e.g., magnetic beads, reagents, or primers). Use a culture grown from normalization and calibration variant mixture to test the BarSeq sequencing library prep protocol outlined in the Automated Bar-Seq Library Preparation and Pooling protocol[23]:

- Grow 100 mL culture up to the stationary phase, then mini-prep or midi-prep plasmid.
  - Use one aliquot of cells transformed with the library.
- Test the BarSeq library prep and magnetic-bead-based cleanup protocol with the plasmid.
  - First do a manual test, pipetting manually instead of using a liquid handler, with a single forward sample multiplexing tag (i.e., primer), and a single reverse sample multiplexing tag. Test with 100 pmol/L and 250 pmol/L input plasmid concentration. Quantify the resulting cleaned-up product and run on a gel.
  - Next, do an automated test using the liquid handler, with 24 different combinations of eight forward and 12 reverse sample multiplexing tags. Test with 250 pmol/L input plasmid in each sample. Quantify the resulting cleaned-up product and run on a gel.

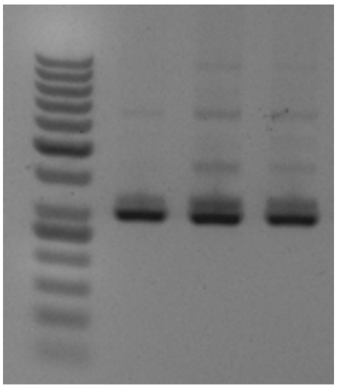


Figure A17: Example gel image from previous NIST experiments. Ladder is on the left; three different BarSeq product samples are in the other lanes. The darkest band is at the expected size. Other bands (longer DNA) are normal for BarSeq prep. Gel extraction should NOT be used to remove the other bands.

DOI: 10.5281/zenodo.12521641

Stage Gate 5

If the following QC checks are passed, proceed to performing the first pooled assay using the control plasmids. If not, you will need to troubleshoot the BarSeq method.

- The Barseq product yield should be 15-30 ng/uL (eluted into 45 uL).
- If there are bands in the gel at shorter length than the expected product (not present in Fig. 16), troubleshoot the PCR or suspect incomplete removal of BarSeq primers.
  - If there is incomplete removal of BarSeq primers, test the method with different bead-to- sample ratios.

The automated test should give results similar to the manual test, and consistent results for all 24 samples. If not, troubleshoot the automation protocol by verifying volumes at each pipetting step.

5.3.2 Run a First Pooled Assay Using Control Plasmids

The control plasmids must now contain unique barcodes, and each barcode-control sequence combination should be recorded. If these plasmids do not have barcodes, reclone and sequence them individually before running the first pooled assay. For the pooled assay, run a measurement with a small library composed of the mixture of the selected normalization and calibration variants. Run this assay the same way as the large-scale pooled assay will be run; for example, use the same plate layout, antibiotic concentrations, and timings. Step-by-step details of the pooled assay can be found in the Pooled, Growth-Based Assays protocol using the Control Variants branch[24].

For this test pooled assay, only use Illumina (Illumina, San Diego, USA) sequencing to count barcodes, because each of the calibration variants will be sequenced individually; long-read plasmid sequencing is not needed.

> Use NovaSeq, as that will be used for the large-scale assays. It is possible to use iSeq or MiSeq instead, but these assays may conflict with some of the experimental decisions (e.g., how the BarSeq primers are designed).

Use the steps outlined in the data analysis pipeline to turn sequencing data of barcode counts into measures of fitness and use the function measurements from the calibration data to generate fitness versus function graphs. Fig. A20 shows plots that indicate whether or not the pooled assay works.

Stage Gate 6

- For each antibiotic concentration, the barcode-counting fitness versus. function (measure in section 2.2.2) values must lie along a consistent curve. Curves for different antibiotic concentrations should have systematic variation, with higher antibiotic concentration showing lower fitness.
- Verify that the measurable range of function (x-axis) spans the desired range and that the antibiotic concentrations are correct.
- Verify that the barcode-counting fitness vs. function curves approximately match the singleplex fitness vs. function curves generated for stage gate 4.

If the Barcode-counting fitness versus function values do not lie along consistent curves for each antibiotic concentration, try the following troubleshooting steps:

- Check the reproducibility of the singleplex function measurements.
- Ignore one or two calibration plasmids with outlying data if there are no other outliers. Check the full plasmid sequence for the outlier calibration plasmid(s) to determine if there are any off-target mutations. Those calibration plasmids can be ignored or left out of the pilot-scale experiment.
- Check log files when pipetting plates with an automated liquid handler to verify that each well was pipetted correctly.

DOI: 10.5281/zenodo.12521641

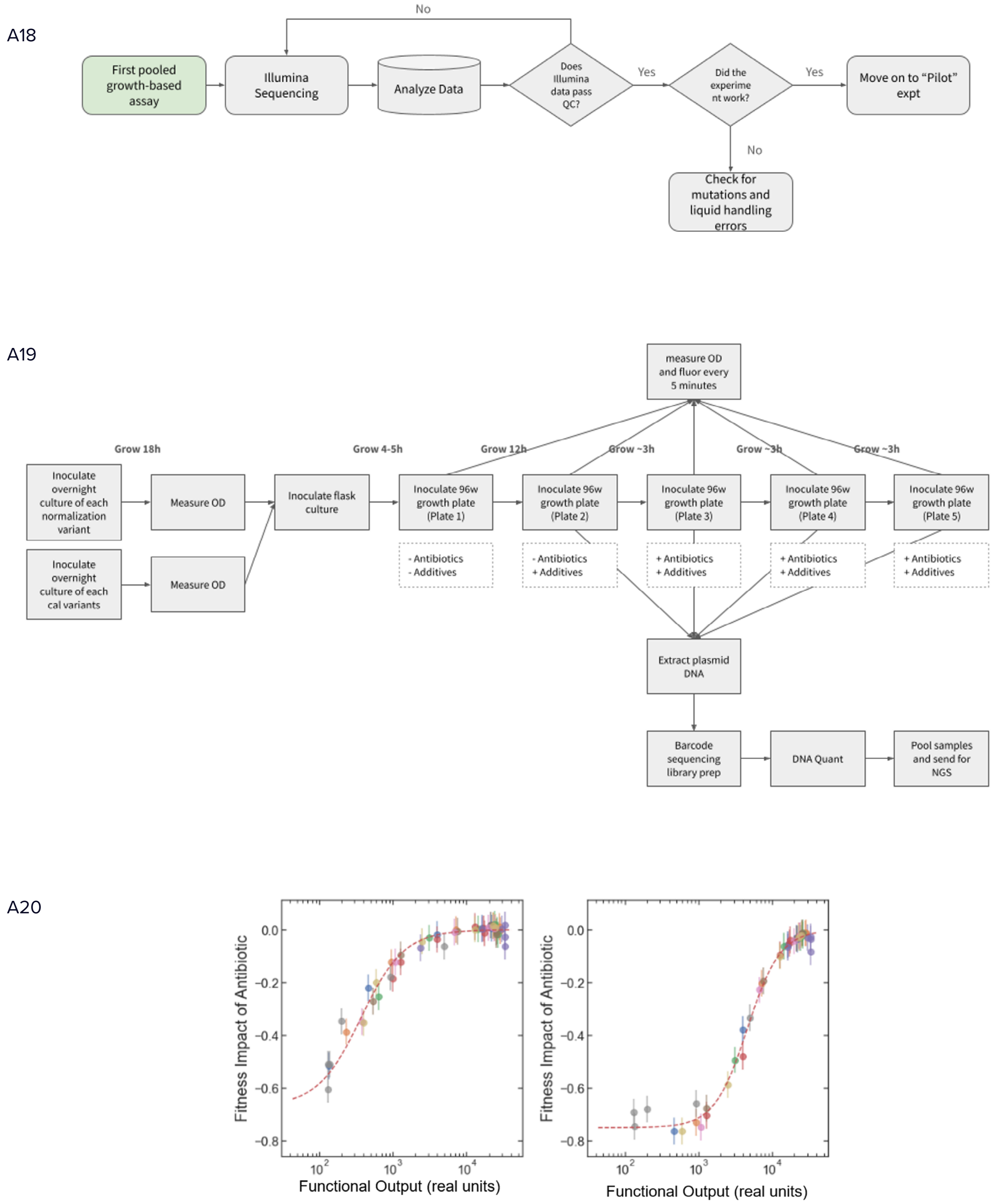

A18
First pooled growth-based assay
Illumina Sequencing
Analyze Data
Does Illumina data pass QC?
No
Yes
Did the experime nt work?
Yes
Move on to “Pilot” expt
No
Check for mutations and liquid handling errors
A19
Grow 18h
Grow 4-5h
Grow 12h
Grow ~3h
Grow ~3h
Grow ~3h
measure OD and fluor every 5 minutes
Inoculate overnight culture of each normalization variant
Measure OD
Inoculate overnight culture of each cal variants
Measure OD
Inoculate flask culture
Inoculate 96w growth plate (Plate 1)
Inoculate 96w growth plate (Plate 2)
Inoculate 96w growth plate (Plate 3)
Inoculate 96w growth plate (Plate 4)
Inoculate 96w growth plate (Plate 5)
- Antibiotics
- Additives
- Antibiotics
+ Additives
+ Antibiotics
+ Additives
+ Antibiotics
+ Additives
+ Antibiotics
+ Additives
Extract plasmid DNA
Barcode sequencing library prep
DNA Quant
Pool samples and send for NGS
A20
Fitness Impact of Antibiotic
0.0
-0.2
-0.4
-0.6
-0.8
10^2
10^3
10^4
Functional Output (real units)
Fitness Impact of Antibiotic
0.0
-0.2
-0.4
-0.6
-0.8
10^2
10^3
10^4
Functional Output (real units)

Figure A18-A20: First pooled assay with control plasmids. A18) Process flow and decision tree diagram for running and evaluating the first pooled assay using controls. 19) Process flow diagram for the first pooled growth-based assay protocol. 20) Example of barcode-counting fitness vs. function output for all controls from previous NIST experiments. Note: The y-axis on those plots is from the barcode-counting fitness in a pooled assay. The x-axis is from the independent measurement of protein function.

DOI: 10.5281/zenodo.12521641

6. Plasmid Library Construction

As long as plasmids adhere to the above plasmid architecture diagram and meet the QC metrics outlined below, plasmid construction can proceed in various ways depending on the experimental goal. Options for generating sequence diversity in the GOI or action site may include: ordering mutated library synthesis; generating diversity via PCR; and/or using combinatorial approaches. Regardless of the diversity generation method, the GOI region and the plasmid backbone must be appended with appropriate half-barcodes (see Barcode region). Each team should use a scarless assembly method (such as Gibson) to insert the sequence of interest into the plasmid backbone.

After generating the library, transform the cloned library into E. coli following the manufacturer's instructions. After transformation, dilute and plate a fraction of the library culture on a plate to be evaluated during Stage Gate 7. Grow the rest in 50mL liquid culture and freeze it in 1mL glycerol stock aliquots.

Stage Gate 7

If the library passes the QC steps listed below, move on to the pilot-scale pooled assay. If not, troubleshoot the library building methods and try again.

- Compute the number of transformants. Count the number of colonies on the plated fraction and use this to compute the total number of transformants in the entire pool (i.e., initial library diversity). Ensure that the number of transformants is at least as large as the desired library size before proceeding.
- Perform clonal sequencing validation of individual library members. Pick 20 colonies for whole-plasmid sequencing. Optionally, these variants can be stocked, and possibly added as additional calibration variants. Ensure that clonal variants have expected sequences before proceeding.

DOI: 10.5281/zenodo.12521641

7. Pilot-Scale Pooled Assay (100,000 variants)

In this stage, growth manipulations are performed on a pooled library of ~100,000 members in the presence of the selective antibiotic, and samples are miniprepped and prepared for sequencing.

7.1 Bottlenecking the Library to Achieve Target Diversity (100,000 Members)

If the initial library diversity is significantly greater than 100,000, bottleneck the library to a diversity of 100,000 or less for measurement. Bottlenecking ensures good sequencing coverage of the pilot experiment by delibrarely capping the number of library members.

There are two protocol options for bottlenecking the library outlined in the Library Bottlenecking Protocols[25]. The first protocol using a flow cytometer with a volumetric, positive-displacement sample introduction to count the number of cells per unit volume in a diluted sample of the library culture is preferred. The cell count is used to prepare a bottleneck culture with a specified number of cells (~100,000). The second protocol uses basic microbial culture equipment to grow and pick a colony based on estimating colony forming units closest to the 100,000 target. It is less precise, but does not require any specialized equipment beyond what is needed for basic bacterial cell culture. Follow the protocols

7.2 Nanopore Sequencing and Pooling QC

Mutations (or assembly errors) in the assay plasmid can break the selection circuit and give a phenotype where the antibiotic resistance gene is always highly expressed. Nanopore results will be used to identify plasmid barcodes that should be ignored because of significant plasmid errors/mutations.

Culture one aliquot of library freezer stock; extract the plasmid and send for Nanopore (Oxford Nanopore Technologies, Oxford, UK) sequencing of the library (10 Gb for initial QC).

Stage Gate 8

If the bottlenecked library pool passes the QC described below, process the samples through the growth-based assay. If not, troubleshoot the plasma extraction procedure and library construction procedures. Analysis for these QC requirements are incorporated into the NIST team's Github.

- Analyze the number of distinct barcodes. This should be comparable to the expectation based on the estimated number of transformants or the size of the bottleneck applied to the library after transformation.
  - This will require barcode clustering with a tool like bartender1.1, which is normally used with the Illumina sequencing data, but can also be used directly with the nanopore data.
- Analyze the abundance distribution of each barcode (i.e., a histogram of the number of nanopore reads for each barcode). The number of distinct barcodes should be close to the expected library diversity (i.e., 100,000), and the width of the distribution, measured as the ratio between the 99th quantile and the mode, should not be greater than 100.
- Analyze the distribution of barcode nearest-neighbor edit distances (Hamming or Levenshtein). With the barcode design proposed here, there should not be any barcodes with nearest neighbors with a distance of less than 4.
  - With the relatively high error rates of nanopore sequencing, there may be some apparent barcodes with nearest neighbors with distance less than 4. We should check that all of them are plausibly explained as nanopore read errors.
- Analyze the fraction of plasmids with complete assembly. More than 80% should have no missing parts.
- Analyze the rate of off-target mutations in the library (mutations outside the gene of interest, where the design should be constant).
- Analyze the distribution of the number and location of mutations in the GOI and other variable sequence regions.
  - The number of mutations per GOI and the locations of those distributions should be consistent with the expectations based on the methods used to generate library diversity.
  - Or, if library diversity is not just due to point mutations (e.g., shuffling, chimeras), the number of distinct genotypes should be consistent with expectations based on the methods used to generate library diversity.

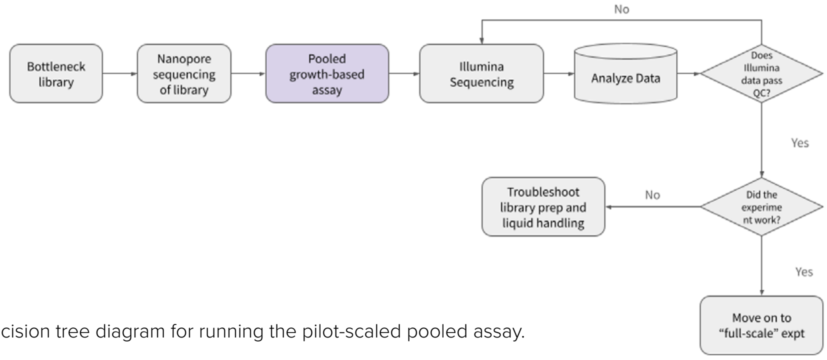

No
Bottleneck library
Nanopore sequencing of library
Pooled growth-based assay
Illumina Sequencing
Analyze Data
Does Illumina data pass QC?
Yes
Troubleshoot library prep and liquid handling
No
Did the experime nt work?
Yes
Move on to "full-scale" expt

Figure A21: Process flow and decision tree diagram for running the pilot-scaled pooled assay.

7.3 Pooled, Growth-Based Assay

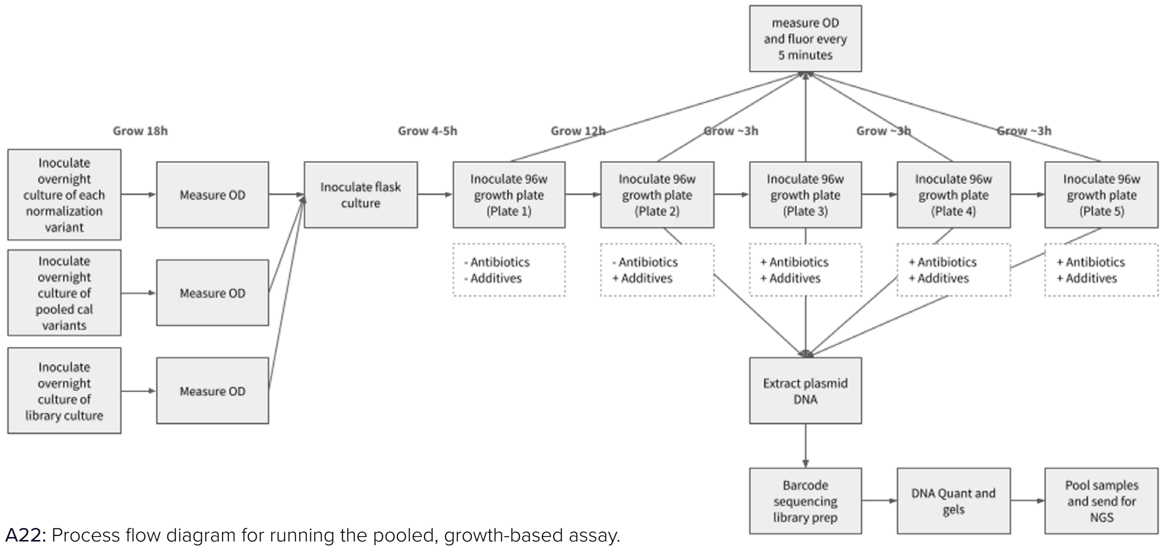

measure OD
and fluor every
5 minutes
Grow 18h
Grow 4-5h
Grow 12h
Grow ~3h
Grow ~3h
Grow ~3h
Inoculate overnight culture of each normalization variant
Inoculate overnight culture of pooled cal variants
Inoculate overnight culture of library culture
Measure OD
Measure OD
Measure OD
Inoculate flask culture
Inoculate 96w growth plate (Plate 1)
Inoculate 96w growth plate (Plate 2)
Inoculate 96w growth plate (Plate 3)
Inoculate 96w growth plate (Plate 4)
Inoculate 96w growth plate (Plate 5)
- Antibiotics
- Additives
- Antibiotics
+ Additives
+ Antibiotics
+ Additives
+ Antibiotics
+ Additives
+ Antibiotics
+ Additives
Extract plasmid DNA
Barcode sequencing library prep
DNA Quant and gels
Pool samples and send for NGS

Figure A22: Process flow diagram for running the pooled, growth-based assay.

> This assay grows up the bacteria five times, generating data for four timepoints. Step-by-step details of the pooled assay can be found in the Pooled, Growth-Based Assays protocol using the Variant Library branch[24].

The plates to include in this assay include:

Plate 1 - First growth in a 96-well plate. This acts as a starter culture for the following plates.

Plate 2 - Zero antibiotic time point (t=1), but does contain different additives (e.g., ligands or inhibitors). This time point can be used to obtain a baseline for the abundance of each barcode before selection begins.

Plate 3 - First antibiotic time point (t=2). This first time point will best capture data for low-activity variants, which will quickly drop out of the pool as selection begins.

Plate 4 - Second antibiotic time point (t=3).

Plate 5 - Third antibiotic time point (t=4). This last time point will best capture data that distinguishes fitness amongst high-activity variants, which will remain in the pool throughout selection.

Stage Gate 9

During this experiment, several QC steps are done to confirm that the assay is running as intended. If the data passes all QC checks outlined below, then proceed to sequencing. If these QC checks fail, samples generated by the method do not proceed to sequencing. In case of failure, review the manual sample loading steps and check liquid handling log files for accuracy.

- Analyze the number of distinct barcodes. This monitors bacterial growth by measuring absorbance. This protocol calls for growing bacteria in a heated shaking plate reader. If available, measuring growth curves is valuable as a way to roughly monitor the growth of the cultures. Samples might fail QC if the final optical density increases during each incubation step, indicating that the incubation time is not set properly or if the culture in one of the wells (e.g., without antibiotic) fails to grow as expected.
- Quantify DNA extraction. After extracting the plasmid from each sample, most of the extracted plasmid is used in the BarSeq library prep. The remaining portion should be used to quantify the amount of plasmid DNA extracted for each sample at each time point using Qubit or an equivalent DNA quantitative measurement. Because DNA extraction is occurring on a small sample volume of a low-copy plasmid, yields are expected to be low; for the first time point, a typical yield should be approximately 1 ng/uL (in ~50 uL). For subsequent time points, the yields will be lower, particularly for the samples grown with the highest antibiotic concentrations; yield for time points 3 and 4 may be below the detection limit for the Qubit measurement. Samples might fail QC if samples that are expected to have high DNA concentration (e.g., early time points and samples grown with zero antibiotics) have very low DNA concentrations (e.g., less than 10% compared with similar samples).
- Quantify sequencing-ready DNA. The amount of amplified DNA at the end of the BarSeq library prep should also be measured for each sample at each time point using Qubit or an equivalent DNA quantitative measurement. Expected yields are between 10 ng/uL and 30 ng/uL (in 45 uL), though some samples could have as little as 1 ng/uL (e.g., samples grown with the highest antibiotic concentration at the later time points). This DNA quantification should also be used to rebalance the DNA amount for each sample when the samples are pooled to be sent to the sequencer. The DNA product from the BarSeq library prep should also be run on gels for each time point and compared with results obtained during BarSeq method development. The gels should not have any visible bands that are shorter than the expected amplicon size (315 bp); shorter bands indicate incomplete PCR cleanup that could result in a very low sequence count from Illumina sequencing. As with the plasmid DNA, samples might fail QC if samples from early time points or samples grown with zero antibiotics have very low DNA concentrations (e.g., < 5 ng/uL).

7.4 Sequencing

In this step, the final pool is sequenced in two stages of increasing depth, interspersed with bioinformatic QC analysis.

The cells passaged in this assay only undergo an upper limit of at most ~10 doublings, so it is unlikely that any cells evolving mechanisms to evade selective pressure will take over the population and render the experimental results null. However 'cheaters' can exist in the pooled assays. In these cases, some of the sequencing data will be lost to the adapted cell (proportional to the percentage of the adapted cells in the assay), but the remaining results will still be valid.

To prepare the pool for the sequencing provider, do the following (Fig. A23):

1. Pool each of the four replicates of a given condition from a single time point. This is done for time points 1-4 in the Pooled, Growth-Based Assays protocol using the Variant Library branch[24].
   a. Run the automated mini-preps immediately after each time point to give 24 plasmid DNA samples.
   b. All four time points together give 96 plasmid samples.
2. Process those 96 plasmid samples with the BarSeq protocol to give PCR product samples.
   a. Follow the Automated Bar-Seq Library Preparation and Pooling protocol[23].
3. Pool the PCR product samples in a way to try to re-balance the amount of DNA from each. Send a single tube for sequencing.
   a. Use only some of each sample to pool for sequencing (keep a portion of each of the 96 PCR products separate) in case the balance between samples for subsequent sequencing orders must change.
   b. Follow the pooling protocols outlined in the Automated Bar-Seq Library Preparation and Pooling protocol[23].
4. Send the pooled sample to a sequencing provider to be sequenced as a "pre-made library". Order one lane of NovaSeq for initial data.

Stage Gate 10

If the sequencing data passes all of the QC checks outlined below, move on to analysis of the pooled data analysis and machine learning evaluation. If not, order additional sequencing.

1. Inspect the redicted fitness of calibration variants. The first good indication of whether or not the assay worked is in the plots of fitness versus function for the calibration variants. The measured 'fitness' of calibration variants (their doubling time in growth-based assay) should track the measured 'function' of the same variants (their measurements in lower-throughput higher fidelity assay).
2. Inspect initial well-to-well variability of barcodes. At the first time point, all the samples should have the same distribution of barcode abundance (because no antibiotic was applied at that time point). Thus, the variability of barcode abundance from well-to-well in the first time point can also be checked. The barcodes should be evenly distributed amongst samples: As a baseline, the relative standard deviation for the barcode read fraction at the first time point should be consistent with Poisson sampling of the barcode reads: std(read fraction)/mean(read fraction)≅1/sqrt(mean read count per sample).
3. Inspect barcode abundance. Plot a histogram of how often each unique barcode occurs. All barcodes should exhibit good read coverage, with a minimum of 10 barcodes per sample at each time point. The barcodes should be relatively evenly distributed at early time points and should become more long-tailed at later time points as low-activity variants drop out. Specifically, for the first time point, the barcode count distribution should satisfy a QC check similar to that applied to the nanopore sequencing (Stage Gate 8): the width of the distribution, measured as the ratio between the 99th quantile and the mode, should not be greater than 100.
4. Inspect the total number of barcodes per sample. Sequencing has pooled together four replicates of 24 conditions at several time points. In this step, use bioinformatics to inspect whether or not samples were successfully pooled, or if uneven coverage of a particular sample requires more reads. Apart from intentional under-sampling of the first time point, most samples should have the same total number of reads within a factor of 3. A few samples, particularly at later time points, can have a read count more than 3-fold lower than other samples but if any samples have a read count more than 10-fold lower than the geometric mean for other samples, that low-count sample should be re-balanced (i.e., more of that sample included in pooling) if additional sequencing is run.
5. Order additional sequencing if necessary and repeat the QC for larger library diversities (>100k barcodes). Typically, one should order another lane of sequencing, but only for a sample pooled from time points 2-5 (growth plates 3-5).

DOI: 10.5281/zenodo.12521641

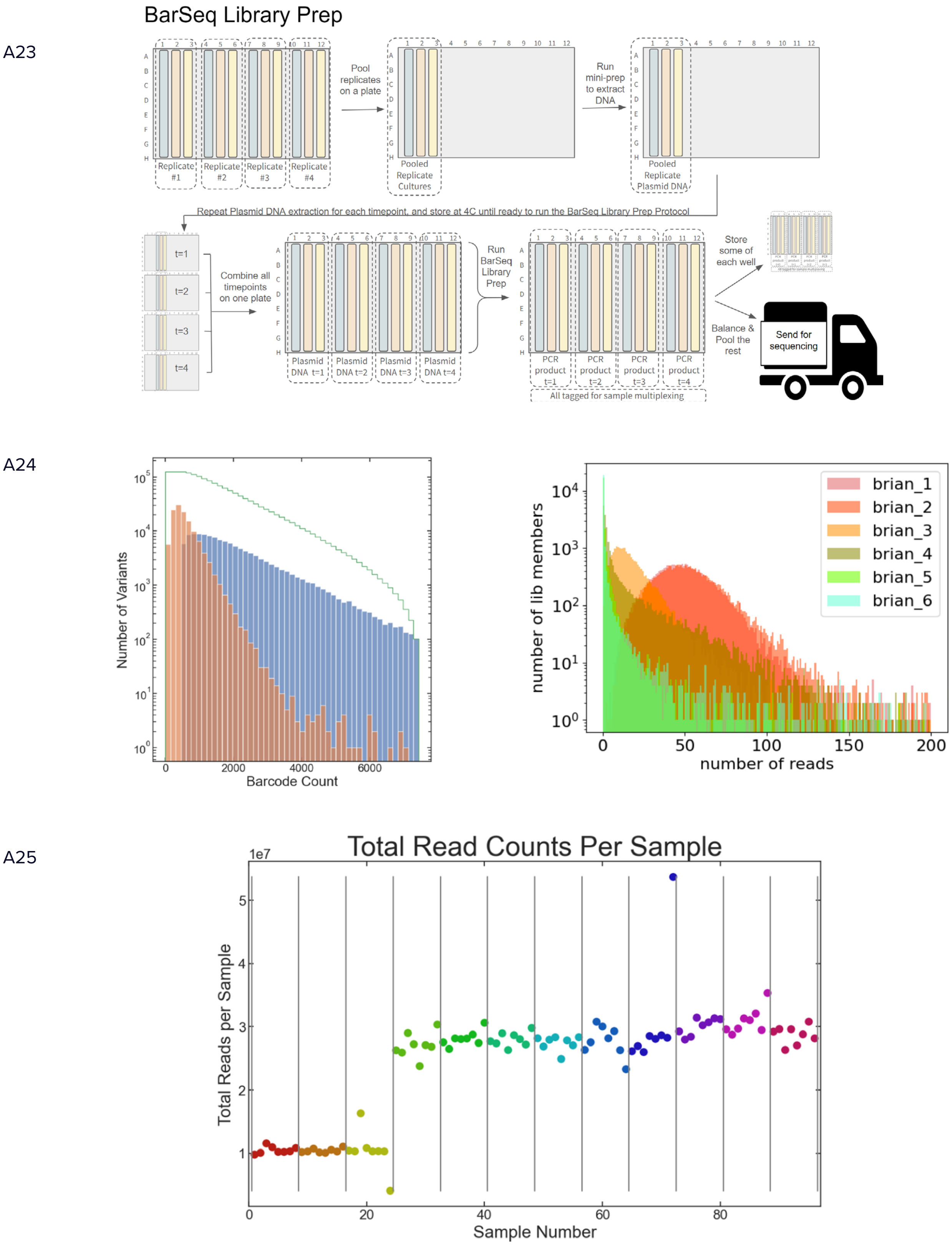


Figure A23-A25: Sequencing process. A23) Illustration for the two-step pooling and consolidation of samples coming out of the pooled, growth-based assay. The exact plate map will be dependent on the format of each new protein function being onboarded. 24) Barcode reads from previous NIST experiments. Top: A histogram plot of the total number of barcode reads for each variant (summed over all samples and time points). Bottom: Another example of a library undergoing six rounds of selection and becoming more long-tailed. 25) Barcode reads per sample from previous NIST experiments. The example of 96 samples is from 24 wells/conditions x 4 time points. Samples 1-24 are from the first time point, with intentionally less data for those samples.

DOI: 10.5281/zenodo.12521641

8. Data Analysis and Machine Learning Evaluation

8.1 Calculating Quantitative Protein Activity

The goal of these growth-based assays is to produce quantitative measurements of protein function. To achieve this, the plasmid system is designed to enable pooled assay formats using barcodes that can be compared to a standard curve created by the calibration controls. During the course of the pooled assay, cells are iteratively grown and challenged during five growth cycles. The pools of strains from cycles 3-5 are sequenced to quantify the barcodes present in the samples (in general, each barcode represents a unique protein variant/ target combination). The strategy outlined below is used to translate the barcode counts for each variant in each condition into their respective quantitative functions.

An initial draft of a complete analysis pipeline can be found at: https://github.com/djross22/nist_lacI_landscape_analysis. In the long term, Align to Innovate will host a version of this analysis pipeline to be used for dataset expansion (additional sequences and new protein functions).

At a high level, the steps of the analysis include:

1. Parse the barcodes. Here, the input is raw barcode sequencing data (Illumina), and the outputs are text files listing the barcode sequence, sample multiplexing tag sequence, and unique molecular identifier sequence for each read, with one file for the forward reads and one for the reverse reads.
2. Cluster the barcodes. Here, the inputs are the output files from the previous parsing step, and the outputs are files indicating the consensus sequence for each barcode sequence (i.e., cluster center), cluster IDs (typically an integer), and the sequencing read errors found in the dataset for each cluster ID.
   a. Depending on the clustering algorithm used, an additional step may be added to merge clusters for barcode reads with in-del errors (resulting in different length barcodes).
3. Use the long-read Nanopore sequencing and the barcode clustering results to identify the genotype corresponding to each barcode ID: the sequence of the protein of interest, the DNA sequence it interacts with (e.g., operator, peptide substrate), and the sequences for all other portions of the plasmid.
4. Sort the barcode reads by sample and count the number of reads for each barcode ID in each sample at each timepoint.
5. Use changes in the barcode counts versus time to calculate the fitness of each variant in each sample. Barcode counts are normalized by counts for the normalization variants.
   a. Depending on the method used for introducing sequence variability, it may be useful to combine the read counts for different barcodes that correspond to the same genotype and/or the same amino acid sequence for the protein of interest.
6. Use the fitness of each variant in each sample and the calibration variant data (fitness versus function) to convert fitness data to quantitative function estimates.
   a. It is important in this step to use analysis methods that provide both a point estimate and an uncertainty estimate for the function(s) of each barcode variant.

Stage Gate 11

Evaluate data quality using the following steps:

1. Check reproducibility. Typically, there are some genotypes that are present with multiple copies in the library (with different barcodes). For example, with random mutagenesis strategies for library generation, the WT sequence will often be present with hundreds or thousands of copies. Sets of duplicate genotypes should be used to evaluate the reproducibility of the results within a pooled assay and to check the uncertainty calibration for resulting function estimates (Fig. A26).
2. Compare the function estimate from the pooled assay with the singleplex function measurement for the calibration sequences. This should give an upper bound on the fidelity of the measurement (lower bound on the typical RMSE).
3. Evaluate the distribution and variance of the data. If the data has a non-reproducible, strange distribution, running ML on it probably won't be very useful nor will the data be of the quality we want to contribute to the project.

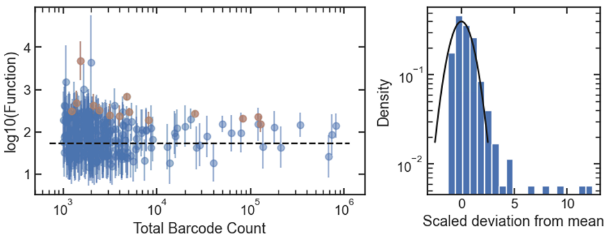


Figure A26: Function versus barcode count from previous NIST experiments. These are results for over 250 WT proteins with different barcodes.

DOI: 10.5281/zenodo.12521641

8.2 ML Analysis Strategy for Pooled Assay Data

Through the collection of data using these high-throughput, pooled, growth-based assays, data analysis tools and machine learning models will be utilized to evaluate the data and ensure its quality for other machine learning applications. Use the methods outlined below, beginning with the pilot data collection, and iteratively throughout the large-scale data collection to continuously observe the dataset as it grows in size.

8.2.1 Data and Metadata Overview

The proposed data schema should work for all function data generated with the growth-based assay platform (Fig. A27). During methods development, the schema will be updated to abide by common database schema recommendations and include all semantic data and linked data available. The schema should be flexible enough to incorporate diverse metadata. For example, being able to accommodate capturing all metadata relevant to a robot screening platform (e.g., gitlab.com/larasuite). We will semantically annotate the data in a standardized way. The ontologies used will be selected during methods development to best suit this particular data and metadata well enough and will be useful for logic reasoning for machine learning (e.g. be decidable). The data will be stored in a unified format (JSON-LD, XML, or YAML) for ease of consumption. The schema proposed here is one possibility (Fig. A27), during the execution of the project, open schemata (e.g. in JSON-schema format) will be made publically available to describe the data and metadata.

The goal is to create a structure that is flexible enough to accommodate different protein families and labs with different instruments, yet standardized enough to be easily parsable for ML. Along with the datasets, we will develop the protocols produced in a machine readable form, since these datasets will be later processed by artificial intelligence or ML applications advanced enough to «understand» the procedures.

8.2.2 Proposed Machine Learning Models and ML Metadata

Two main classes of ML models should be trained and deployed by users interacting with the Align datasets: models that need homologous sequences and encode specific mutations and models that are able to digest distant sequences (of different lengths.) One class of model may outperform another for a user's specific engineering task, so several different model architectures should be trained and evaluated from both classes to provide a broad demonstration of the data:

1. Class 1: Models that need homologous sequences and encode specific mutations. These models will not be able to generalize beyond this homology. For non-probabilistic regressors predictive uncertainty can be determined from cross-validation ensembles.

   a. Standard linear and nonlinear regressors on top of a simple encoding (1-hot of mutations, physico-chemical of substituted amino acids):

   - *Ridge or Kernel-Ridge: strongly-regularized, simple regression algorithms which would work to an extent, and hopefully already be useful, even with small amounts of data*[3].
   - *Gradient boosted trees: out-of-the box regressors that would be expected to give good baselines on some of the encodings. The regressor collection can also contain neural network-based models suitable for small data.*
   - *Neural networks for deep learning: such as a graph-based approach*[25] *which has the advantage of encoding ligand information.*
   - *Transformers that have been shown to powerfully predict protein fitness*[2].
   - *Lasso, Bayesian ridge, Gaussian Processes.*

   b. LANTERN (landscape interpretable nonparametric model)[26]: a fully interpretable genotype-phenotype landscape model created and used by David Ross's group at NIST.

   - *This model can turn changes in function into a linear combination of mutation effect vectors, providing insight into how each mutation individually impacts the protein's function.*

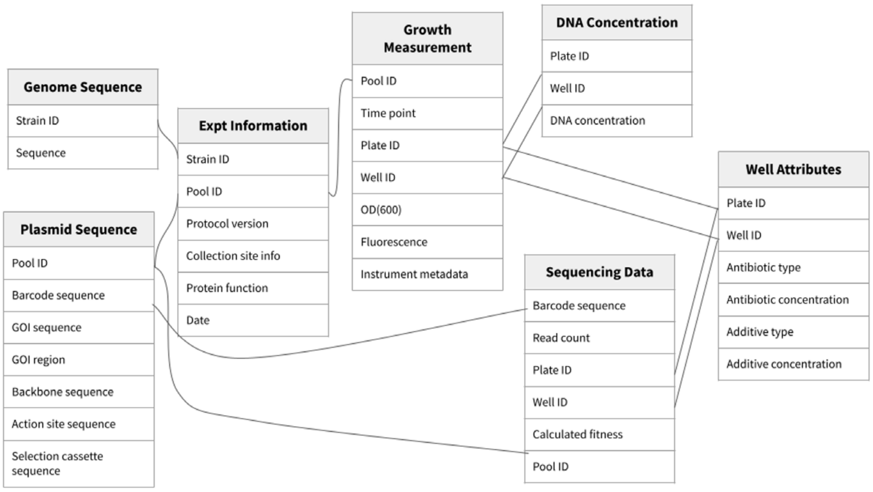


Figure A27: Data scheme for all data and metadata to be collected during the pooled, growth-based assay experiments.

DOI: 10.5281/zenodo.12521641

- *Additional models that may be tested of this type are the Mave-NN[27] and MoCHI[28] models*

2. **Class 2:** Models that are able to digest distant sequences (of different lengths.)
   a. Models like 1a), but using transfer encodings from 'large language models' like ESM, UniProt etc. All models in class 2 can also be applied in class 1, but a real advantage is only to be expected once you want to generalize to proteins that are not very close (by a number of point mutations) to the proteins already seen by the model. Ligand information can also be encoded as language.
   b. Models using graph neural networks on computed/estimated 3D-structures or automated docking for such structures.

Metadata Collected With Each Model Run

All the models will have certain metadata in common, but some will have extra metadata. The guiding philosophy of this proposal is to capture and report all possible metadata for each of the proposed models using a consistent structure. We will create an infrastructure to capture all relevant metadata and the protocols used to perform ML analysis in a manner that allows for extensibility and machine readability.

As an example, in the case of using a LANTERN model, the following metadata will be reported:

- Model parameters: Epochs, dimensions, learning rate.
- Hardware information: CPU, GPU, RAM in GB (AWS instance type, etc.).
- Elapsed time to train.
- Split strategy.
- Set number (e.g., train/dev/test1.)
- Performance metrics:
  - *Loss (neg log likelihood) vs. epoch plot.*
  - *Root Mean Squared deviation.*
  - *R2 correlation.*
  - *Spearman coefficient (weighted and unweighted).*
  - *Pearson coefficient (weighted and unweighted).*
  - *Kendall's tau*
  - *Median absolute deviation*

8.2.3 Proposed Methods for ML Evaluation

ML models require evaluation to determine if sparsity in the data or nuisance variables are affecting prediction accuracy. Initial models built on the pilot-scale pooled assay will inform decisions on the size of the large-scale assays. If lack of data is hindering accuracy, variant density mapped to protein structure can quickly reveal variant-sparse regions with 3D context, allowing for targeted library expansion. Models will be cross-validated using the same sets of shuffled data to maximize robustness.

Cross Validation Strategy

Due to the enormous combinatorial space of possible sequences, even the large-scale dataset will still be comparably small and could result in estimates with a high variance. To combat this issue, we will utilize a nested splitting strategy to form sets for cross-validation of the data. The data will be pre-split into multiple sets following the form outlined below, so that all tested models will be benchmarked using the same splits. By training/testing on the same split, any sets that have unusual distributions can be identified.

The outer split creates train and test sets:

- Train (used for model selection/hyperparameter tuning), further split into:
  - *Train-train set*
  - *Train-development set*
- Test (used to report the average results on selected models)

Additionally, the following strategies will be implemented as additional evaluations according to the number of wild-type proteins being measured within an assay:

1. For experiments on mutations of one wild type:
   a. Random splitting (yields estimates of prediction loss on new samples drawn from the same distribution)
   b. Splitting across different mutations (yields estimates of prediction loss on samples containing new mutations).
   c. Splitting across different positions (yields estimates of prediction loss on samples across positions.) Separate positions will only occur in one split thereby estimating the extrapolative power of models.
2. For experiments on mutations of several wild-types:
   a. Splitting within each group of homologous variants according to a) b) c) above (yields estimates of prediction loss on new samples homologous to samples the model has seen so one can see whether having seen the other groups improves predictions on one group).
   b. Split between groups of homologous variants according to a) b) c) above (yields estimates of prediction loss on proteins non-homologous to the ones seen in training; a hard case, but very relevant for the development of the field).

Model Evaluation Steps

1. **Determine if any of the metadata not relevant to measured fitness or function** (e.g., day of the week, barcode, site information, plate number, run-order, set number (train/dev/test), etc.) are influencing fitness predictions (i.e., are nuisance variables). This determines if there is any consistent bias in the data from experimental conditions or equipment.
2. **Evaluate if the mutational coverage is adequate ("Qualitation").**
   a. Do a structural analysis of the coverage of protein sequence/surface.
      - *Acquire structure/model of the protein and map all mutant positions observed; this provides a 3D heatmap of mutant hotspots.*
      - *Calculate percent diversity across 1D sequence string.*
      - *Calculate percent diversity across the surface area.*
      - *Consider all of these metrics as variables for predictive power: how does 1D/2D/3D mutational coverage lead to the model's predictive power?*
   b. Do QC in the initial model. Prior probability will be

examined as dimensions are added to the model.

- *If large decreases in probability occur as dimensions increase (Fig. A28, left), the training data is influencing the model predictions within these reduced dimensions.*
- *If small, steady declines in the prior probability occur as dimensions are added (Fig. A28- right), the training data does not strongly influence model prediction. This implies there is not enough mutational coverage.*

3. Determine dataset size for large-scale collection by checking how many variants are needed to build an accurate model.

   a. Do this by checking how the metrics values change as the total number of data points used for the selected models varies.

   - *Set the minimum number of data points as the amount that achieves the best metrics values (primarily the RMS deviation). If the best metrics achieved are still considered poor, estimate the relationship between total data points and model accuracy to determine how many additional points are needed.*

4. Compare the differences between good and bad models.

   a. During step 2, identify "good" and "bad" baseline models during step 2 based on their metric values and compare them as the number of data points included is increased.

   - *If the difference between good models and basic/poor models increases, the data contains a real trend.*
   - *If the difference does not increase, there is a deep flaw in the dataset.*

5. Predict an initial set of variants to evaluate and validate the models (quantitation).

   a. Determine the correlation to knowns. Predict the calibration+norm variants, plot observed fitness versus predicted, and determine if predicted and observed fitness are monotonously related.

   b. Assess predictive power. Order 20-50 predicted protein sequences with phenotypes representative of the whole observed landscape; this number can be adjusted based on individual assay conditions for running one 96-well plate. Protein function will be determined using the procedure outlined in the Singleplex Assay for Function Measurements protocol[21]. Then, observed fitness versus model-predicted fitness will be plotted.

The reported average performance on any test set will be an estimate of running the same model selection and fitting procedure on a similar training set and applying it to a similar test set. The test performance will not be used to perform a second round of model selection, because that would again bear the risk of 'conceptual overfitting'.

Stage Gate 12

If at least one of the four ML models successfully passes the evaluation steps outlined above, proceed to large-scale data collection. If not, perform the following actions to improve model accuracy (ordered by increasing effort and cost):

- Adjust the model parameters (epochs, dimensions, learning rate, etc.).
  - *Define an acceptable range of n parameters to iterate through to prevent this becoming an endless process.*
- Return to step 4.3 and repeat data collection using another random aliquot of the library.
- Create targeted variant sets by analyzing mutational diversity as described in 5.2.3.2a and generating specific variants to fill in gaps

Note: After large-scale data collection, repeat the process with any failed models using the expanded dataset to determine if a lack of data was responsible for the failure.

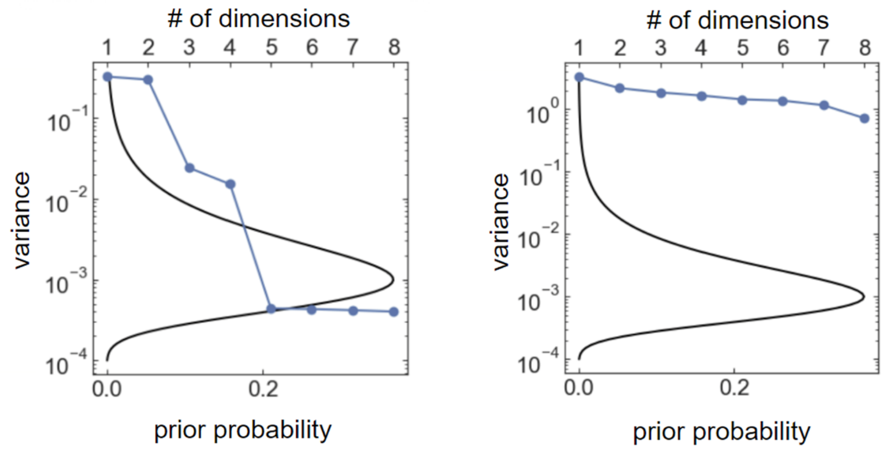


Figure A28: Left: The desired highly predictive outcome of 'large decreases' in probability as dimensions are decreased.
Right: An example of a model that is poorly predictive because there are very small decreases in variance as dimensions are decreased.

9. Large-Scale Data Collection

Now that a new protein function has been onboarded and pilot-scale data has been collected and analyzed, the assay can be routinely run at a larger scale. The main difference between the pilot-scale and large-scale assay is the library size. In between batches, data will be inspected to determine if any assay condition adjustments are needed (e.g., choice of concentrations for antibiotics and other additives). Collection differs only in two ways:

1. Use of larger pool sizes. Initially, pools are bottlenecked to 100k variants. The results of the pilot-scale assay will be used to estimate how large the pool size can be for a target level of data accuracy and with constraints on the number of lanes of sequencing and gigabytes of Nanopore data. Additionally, insights from the ML analysis will enable the evaluation of the number of data points needed to improve model accuracy. Full-scale data collection should be accomplished with larger pool sizes of 100k-500k per pool. To determine the optimal pool size after the pilot-scale pooled assay, determine how the measurement uncertainty for the pooled assay scales with the number of Illumina reads:
   a. Identify several variants from the pilot dataset. Synthesize and measure those variants (as clonal variants), using the same high-fidelity measurement used to characterize the calibration variants. Use these high-fidelity measurements to check the calibration of the uncertainty estimates from the pooled assay data. If the pooled assay uncertainty is well-calibrated (or after the data analysis is adjusted to make it well-calibrated), use the empirical scaling of uncertainty with barcode read count to consider hypothetical scenarios such as measurement of a library with larger diversity (lower barcode count per variant) and smaller diversity (higher barcode count per variant).
   b. The optimal pool size will also depend on these factors:
      - *Sequencing strategy (Nanopore and Illumina) and the return on investment for sequencing deeper versus adding additional targets.*
      - *Number of assay conditions for the pooled assay (roughly the number of antibiotic concentrations times the number of additive conditions). With more conditions, the Illumina sequencing would need to be spread across more samples.*
      - *Relative abundance distribution of variants in the library. With a more even distribution, a larger library with an equivalent amount of Illumina sequencing could be measured.*
2. Measurement of diverse libraries. Multiple libraries generated using different library diversification techniques and/or user-provided samples can be measured during full-scale data collection. These could consist of more divergent sequences, as well as new protein families (e.g., adding both well-characterized and uncharacterized proteins from the TetR family and additional TF families).

When calculating quantitative protein activity, the predictive capabilities are only reliable within the data range of the calibration controls used to develop the standard curve (i.e., prediction starts to fall off quickly outside of this calibrated range). To continuously improve the dynamic range of the assay as more diverse libraries are explored, new calibration controls of higher fitness should be selected, validated, and incorporated into the growth-based assay. Throughout large-scale collection, continuously look for better calibration control candidates and test them according to the procedures outlined in Section 5.2.2.

10. Dataset Expansion

This proposal outlines a platform for collecting functional data and provides specific examples of two protein function datasets. The proposed pooled, growth-based assay format was chosen because of its extensibility. Growth-based assays are compatible with several existing selection cassettes for determining additional protein functions and biophysical properties (Fig. 3). There is already interest in expanding this platform to examine new protein functions, and leveraging common plasmid designs, protocols, and analysis pipelines will facilitate onboarding new protein functions. Several scientists have already expressed interest in onboarding new selection cassettes into the platform, such as:

- A general solubility circuit to distinguish between proteins that express poorly versus proteins with low function [idea from Ben Lehner].
- Cas protein + gRNA pairs to focus on the gRNA pairing problem [idea from Marc Güell, Dimitrije Ivančić, and Noelia Ferruz].
- Aminoacyl-tRNA synthetases to challenge the field in terms of designing enzymes that interact with small molecules and RNA [idea from Ross Thyer and Erika DeBenedictis].
- Protein-protein binding [idea from Ron Koder (CUNY) and Erika DeBenedictis].
- Bacterial two-component systems [idea from Katie Hatsat and the DeGrado lab (UCSF)].

Literature searches can also be used to identify many additional circuits. For example, any molecular biology paper that uses flow cytometry to separate functional from non-functional protein variants can be converted into a growth-based assay by exchanging the fluorescent protein for an antibiotic protein. Additionally, several studies showcase another class of growth-based selections that can be created with synthetic auxotroph strains. Using these strains in our platform would allow us to link the function of any cell metabolism enzyme to growth.

Align is also incentivizing participation in this growing effort; Align will cover experimental costs, including sequencing,

Component	NIST Equipment	Description
Liquid handling (growth-based assay)	Hamilton STAR with 8-channel and 96-channel heads*; and MPE2 positive-pressure filter press (used for automated plasmid extraction)*	Hamilton STAR with 8-channel and 96-channel heads*
Liquid handling (BarSeq library prep)	Hamilton STAR with 8-channel and 96-channel heads, with magnet base and multiple heater-shakers*	None
Plate sealer	Azenta (4titude) model a4s‡	Manual
Plate peeler/de-sealer	Azenta (Brooks) X-Peel‡	Manual
Centrifuge	Hettich Rotanta 460	Agilent V Spin🠋
Multimode plate-reader	Agilent (Biotek) Neo2🠋	BMG Labtech Spectrostar Omega
Flow cytometer	Attune flow cytometer with autosampler▲	<>

Table A5: Equivalent automation equipment capabilities at NIST and The Francis Crick Institute.

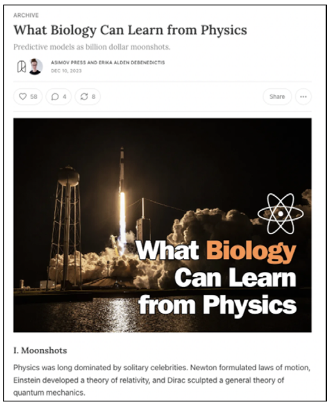
*Hamilton Company, Reno, USA; ‡Azenta Life Sciences, Burlington, USA; Hettich, Tuttlingen, Germany; 🠋Agilent Technologies, Santa Clara, USA; ▲ThermoFisher, Waltham, USA; BMG LABTECH, Ortenberg, Germany.

11. Suggested Reading

Below is a suggested curated reading list to better understand this document and its context:

This proposed dataset platform is part of Align's Open Dataset Initiative, which pioneers new ways to identify, collect, and share large datasets in life science. Read more about the Open Datasets initiative here:

"What Biology Can Learn from Physics", *December 2023, Asimov Press*



ARCHIVE

What Biology Can Learn from Physics

Predictive models as billion dollar moonshots.

ASIMOV PRESS AND ERIKA ALDEN DEBENEDICTIS
DEC 10, 2023

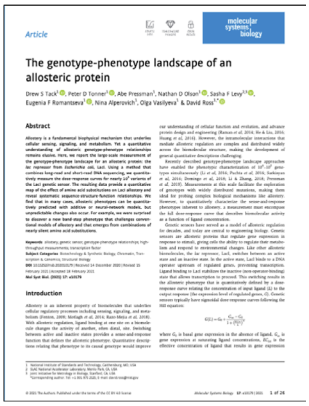
♡ 58 · 💬 4 · 🔁 8 · Share · ···

I. Moonshots

Physics was long dominated by solitary celebrities. Newton formulated laws of motion, Einstein developed a theory of relativity, and Dirac sculpted a general theory of quantum mechanics.

Growth based assays have previously been used to gather large sequence ➔ function datasets. See this paper for an example of collecting large datasets on the LacI protein:

"The genotype-phenotype landscape of an allosteric protein", *2021, Molecular Systems Biology*



Article

molecular systems biology

The genotype-phenotype landscape of an allosteric protein

Drew S Tack[1], Peter D Tonner[1], Abe Pressman[1], Nathan D Olson[1], Sasha F Levy[2,3], Eugenia F Romantseva[1], Nina Alperovich[1], Olga Vasilyeva[1] & David Ross[1,*]

Abstract

Allostery is a fundamental biophysical mechanism that underlies cellular sensing, signaling, and metabolism. Yet a quantitative understanding of allosteric genotype-phenotype relationships remains elusive. Here, we report the large-scale measurement of the genotype-phenotype landscape for an allosteric protein: the *lac repressor* from *Escherichia coli*, LacI. Using a method that combines long-read and short-read DNA sequencing, we quantitatively measure the dose-response curves for nearly 10^5 variants of the LacI genetic sensor. The resulting data provide a quantitative map of the effect of amino acid substitutions on LacI allostery and reveal systematic sequence-structure-function relationships. We find that in many cases, allosteric phenotypes can be quantitatively predicted with additive or neural-network models, but unpredictable changes also occur. For example, we were surprised to discover a new band-stop phenotype that challenges conventional models of allostery and that emerges from combinations of nearly silent amino acid substitutions.

Keywords allostery; genetic sensor; genotype-phenotype relationships; high-throughput measurements; transcription factor
Subject Categories Biotechnology & Synthetic Biology; Chromatin, Transcription & Genomics; Structural Biology
DOI 10.15252/msb.202010179 | Received 14 December 2020 | Revised 15 February 2021 | Accepted 18 February 2021
Mol Syst Biol. (2021) 17: e10179

Introduction

Allostery is an inherent property of biomolecules that underlies cellular regulatory processes including sensing, signaling, and metabolism (Fenton, 2008; Motlagh *et al*, 2014; Razo-Mejia *et al*, 2018). With allosteric regulation, ligand binding at one site on a biomolecule changes the activity of another, often distal, site. Switching between active and inactive states provides a sense-and-response function that defines the allosteric phenotype. Quantitative descriptions relating that phenotype to its causal genotype would improve our understanding of cellular function and evolution, and advance protein design and engineering (Raman *et al*, 2014; He & Liu, 2016; Huang *et al*, 2016). However, the intramolecular interactions that mediate allosteric regulation are complex and distributed widely across the biomolecular structure, making the development of general quantitative descriptions challenging.

Recently described genotype-phenotype landscape approaches have enabled the phenotypic characterization of 10^4–10^5 genotypes simultaneously (Li *et al*, 2016; Puchta *et al*, 2016; Sarkisyan *et al*, 2016; Domingo *et al*, 2018; Li & Zhang, 2018; Pressman *et al*, 2019). Measurements at this scale facilitate the exploration of genotypes with widely distributed mutations, making them ideal for probing complex biological mechanisms like allostery. However, to quantitatively characterize the sense-and-response phenotypes inherent to allostery, a measurement must encompass the full dose-response curve that describes biomolecular activity as a function of ligand concentration.

Genetic sensors have served as a model of allosteric regulation for decades, and today are central to engineering biology. Genetic sensors are allosteric proteins that regulate gene expression in response to stimuli, giving cells the ability to regulate their metabolism and respond to environmental changes. Like other allosteric biomolecules, the *lac* repressor, LacI, switches between an active state and an inactive state. In the active state, LacI binds to a DNA operator upstream of regulated genes, preventing transcription. Ligand binding to LacI stabilizes the inactive (non-operator-binding) state that allows transcription to proceed. This switching results in the allosteric phenotype that is quantitatively defined by a dose-response curve relating the concentration of input ligand (L) to the output response (the expression level of regulated genes, G). Genetic sensors typically have sigmoidal dose-response curves following the Hill equation:

$$G(L) = G_0 + \frac{G_\infty - G_0}{1 + \left(\frac{EC_{50}}{L}\right)^n}$$

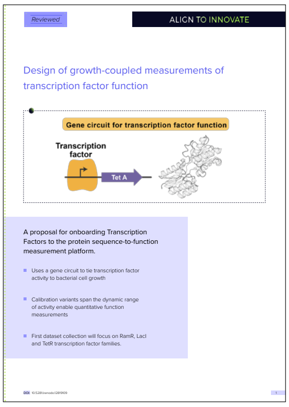
where G_0 is basal gene expression in the absence of ligand, G_∞ is gene expression at saturating ligand concentrations, EC_{50} is the effective concentration of ligand that results in gene expression

[1] National Institute of Standards and Technology, Gaithersburg, MD, USA
[2] SLAC National Accelerator Laboratory, Menlo Park, CA, USA
[3] Joint Initiative for Metrology in Biology, Stanford, CA, USA
*Corresponding author. Tel: +1 301 975 2020; E-mail: david.ross@nist.gov

© 2021 The Authors. Published under the terms of the CC BY 4.0 license · *Molecular Systems Biology* **17**: e10179 | 2021 · **1 of 26**

In addition to the proposed platform described in this document, three groups have proposed specific protein functions to onboard. See the following proposals for onboarding transcription factors, proteases, and aminoacyl tRNA synthetases.

"Design of growth-coupled measurements of transcription factor function", *2024, Align to Innovate*



Reviewed

ALIGN TO INNOVATE

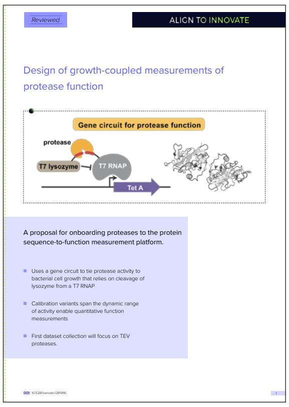
Design of growth-coupled measurements of transcription factor function

A proposal for onboarding Transcription Factors to the protein sequence-to-function measurement platform.

- Uses a gene circuit to tie transcription factor activity to bacterial cell growth
- Calibration variants span the dynamic range of activity enable quantitative function measurements
- First dataset collection will focus on RamR, LacI and TetR transcription factor families.

DOI: 10.5281/zenodo.12819109 · 1

"Design of growth-coupled measurements of protease function", *2024, Align to Innovate*



Reviewed

ALIGN TO INNOVATE

Design of growth-coupled measurements of protease function

A proposal for onboarding proteases to the protein sequence-to-function measurement platform.

- Uses a gene circuit to tie protease activity to bacterial cell growth that relies on cleavage of lysozyme from a T7 RNAP
- Calibration variants span the dynamic range of activity enable quantitative function measurements
- First dataset collection will focus on TEV proteases.

DOI: 10.5281/zenodo.12819116 · 1

12. References

1. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

2. Notin, P. et al. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. *arXiv [cs.LG]* (2022).

3. Hsu, C., Nisonoff, H., Fannjiang, C. & Listgarten, J. Learning protein fitness models from evolutionary and assay-labeled data. *Nat. Biotechnol.* **40**, 1114–1122 (2022).

4. Tsuboyama, K. et al. Mega-scale experimental analysis of protein folding stability in biology and design. *Nature* **620**, 434–444 (2023).

5. Furlan, J. P. R. & Stehling, E. G. Predicting tigecycline susceptibility in multidrug-resistant Klebsiella species and Escherichia coli strains of environmental origin. *Braz. J. Microbiol.* **54**, 1915–1921 (2023).

6. Stanton, B. C. et al. Genomic mining of prokaryotic repressors for orthogonal logic gates. *Nat. Chem. Biol.* **10**, 99–105 (2014).

7. Hersey, A. N., Kay, V. E., Lee, S., Realff, M. J. & Wilson, C. J. Engineering allosteric transcription factors guided by the LacI topology. *Cell Syst* **14**, 645–655 (2023).

8. Blum, T. R. et al. Phage-assisted evolution of botulinum neurotoxin proteases with reprogrammed specificity. *Science* **371**, 803–810 (2021).

9. Packer, M. S., Rees, H. A. & Liu, D. R. Phage-assisted continuous evolution of proteases with altered substrate specificity. *Nat. Commun.* **8**, 956 (2017).

10. Wang, T., Badran, A. H., Huang, T. P. & Liu, D. R. Continuous directed evolution of proteins with improved soluble expression. *Nat. Chem. Biol.* **14**, 972–980 (2018).

11. Levy, E. D., Kowarzyk, J. & Michnick, S. W. High-resolution mapping of protein concentration reveals principles of proteome architecture and adaptation. Cell Rep. **7**, 1333–1340 (2014).

12. Leconte, A. M. et al. A population-based experimental model for protein evolution: effects of mutation rate and selection stringency on evolutionary outcomes. *Biochemistry* **52**, 1490–1499 (2013).

13. Pu, J., Zinkus-Boltz, J. & Dickinson, B. C. Evolution of a split RNA polymerase as a versatile biosensor platform. *Nat. Chem. Biol*. **13**, 432–438 (2017).

14. Pu, J., Disare, M. & Dickinson, B. C. Evolution of C-Terminal Modification Tolerance in Full-Length and Split T7 RNA Polymerase Biosensors. *Chembiochem* **20**, 1547–1553 (2019).

15. Badran, A. H. et al. Continuous evolution of Bacillus thuringiensis toxins overcomes insect resistance. Nature **533**, 58–63 (2016).

16. Miller, S. M. et al. Continuous evolution of SpCas9 variants compatible with non-G PAMs. Nat. Biotechnol. **38**, 471–481 (2020).

17. Popa, S. C., Inamoto, I., Thuronyi, B. W. & Shin, J. A. Phage-Assisted Continuous Evolution (PACE): A Guide Focused on Evolving Protein–DNA Interactions. *ACS Omega* **5**, 26957–26966 (2020).

18. Thuronyi, B. W. et al. Continuous evolution of base editors with expanded target compatibility and improved activity. *Nat. Biotechnol.* (2019) doi:10.1038/s41587-019-0193-0.

19. Richter, M. F. et al. Phage-assisted evolution of an adenine base editor with improved Cas domain compatibility and activity. *Nat. Biotechnol.* 1–9 (2020).

20. Bryson, D. I. et al. Continuous directed evolution of aminoacyl-tRNA synthetases. *Nat. Chem. Biol.* **13**, 1253–1260 (2017).

21. Ross, D. Singleplex assay for function measurements v2. (2024) doi:10.17504/protocols.io.dm6gpzwx8lzp/v2.

22. Ross, D. Singleplex assay for fitness measurements v1. (2024) doi:10.17504/protocols.io.8epv5xye4g1b/v1.

23. Ross, D. & Alperovich, N. Automated Bar-Seq Library preparation and pooling v2. (2024) doi:10.17504/protocols.io.3byl49qdjgo5/v2.

24. Ross, D. Pooled, growth-based assays v1. (2024) doi:10.17504/protocols.io.5qpvokq1bl4o/v1.

25. Ross, D. Library Bottlenecking Protocols V.3 . (2024) doi.org/10.17504/protocols.io.x54v9227pl3e/v3

26. Lin, X. DeepGS: Deep Representation Learning of Graphs and Sequences for Drug-Target Binding Affinity Prediction. *arXiv [cs.LG]* (2020).

27. Tonner, P. D., Pressman, A. & Ross, D. Interpretable modeling of genotype-phenotype landscapes with state-of-the-art predictive power. *Proc. Natl. Acad. Sci. U. S. A.* **119**, e2114021119 (2022).

28. Tareen, A. et al. MAVE-NN: learning genotype-phenotype maps from multiplex assays of variant effect. *Genome Biol.* **23**, 98 (2022).

29. MoCHI: Neural Networks to Fit Interpretable Models and Quantify Energies, Energetic Couplings, Epistasis and Allostery from Deep Mutational Scanning Data. *(Github)*.

DOI: 10.5281/zenodo.12521641