A data guided approach to building an ML ready protein expression dataset

Catherine Baranowski, Aviv Spinner, Pete Kelly
The Align Foundation

Abstract

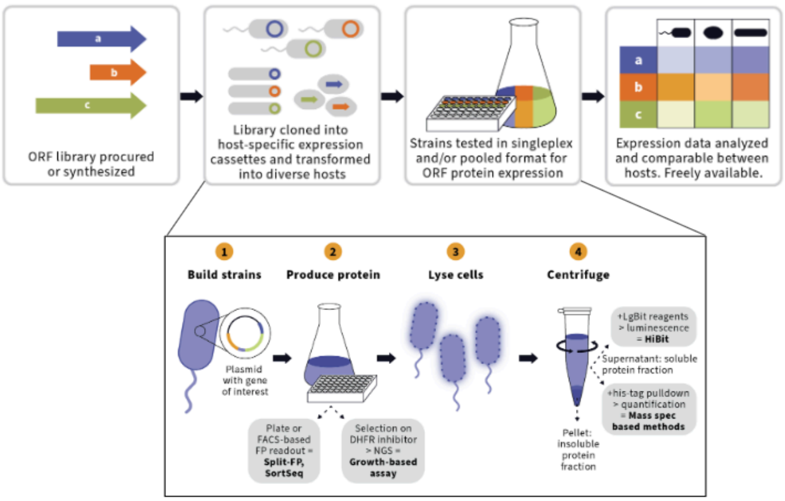
Recombinant protein expression is central to academic exploration as well as biotechnology's advancement of human health, climate applications and the bioeconomy in general. However, not all proteins can be expressed in all organisms, and the field lacks a predictive model of soluble protein expression that could replace laborious experimental trial-and-error. This project aims to design and test an openly available and extensible experimental platform and standardized data ontology for collecting soluble recombinant protein expression data across organisms. The resulting public dataset will be used in building predictive models of protein expression. Here we share preliminary assay feasibility data in our first expression host organism, *Escherichia coli*.

Introduction

Soluble protein expression impacts all corners of the scientific community. However, most models of protein expression focus on one organism[1], thus there is a need for a generalized model of protein expression in common expression hosts. However, building such a model requires a dataset of large-scale protein production across a variety of proteins in a variety of organisms. There is a lack of freely-available, large-scale expression datasets that span organisms and have been collected in consistent experimental settings. [1]

In this work, we aim to design and generate an expression host- and sequence-diverse dataset, soluble protein expression dataset specifically to feed into predictive models (Figure 1, top). "We define soluble protein expression as measurable recombinant protein in the soluble fraction of cell lysate produced in the context of a microbial expression system, represented as proteins/cell. Broadly the factors that impact expression can be divided into extrinsic and intrinsic factors[2,3]. Extrinsic factors that represent experimental choices in how to express the protein - host genotype, expression cassette architecture, and experimental details of culturing, lysis etc - can be modified for better results. Intrinsic factors that are driven by the specific amino acid sequence - foldability, stability and solubility - cannot be altered without changing the identity of the protein. Soluble protein expression is the compound outcome of both the intrinsic and extrinsic factors. For this reason, expression is a valuable single readout that summarizes many relevant inputs. Our dataset proposes to explore both *extrinsic* (different host organisms) and *intrinsic factors* (different amino acid sequences) to produce reproducible, quantitative measurements of soluble expression. Ideally, the interaction of of both these intrinsic and extrinsic factors can be predicted using predictive models that are trained on the resulting dataset." [1]

Figure 1: Dataset workflow and sample preparation summary


a
b
c
ORF library procured or synthesized
Library cloned into host-specific expression cassettes and transformed into diverse hosts
Strains tested in singleplex and/or pooled format for ORF protein expression
Expression data analyzed and comparable between hosts. Freely available.
a
b
c
1
Build strains
2
Produce protein
3
Lyse cells
4
Centrifuge
Plasmid with gene of interest
Plate or FACS-based FP readout = Split-FP, SortSeq
Selection on DHFR inhibitor > NGS = Growth-based assay
+LgBit reagents > luminescence = HiBit
Supernatant: soluble protein fraction
+his-tag pulldown > quantification = Mass spec based methods
Pellet: insoluble protein fraction

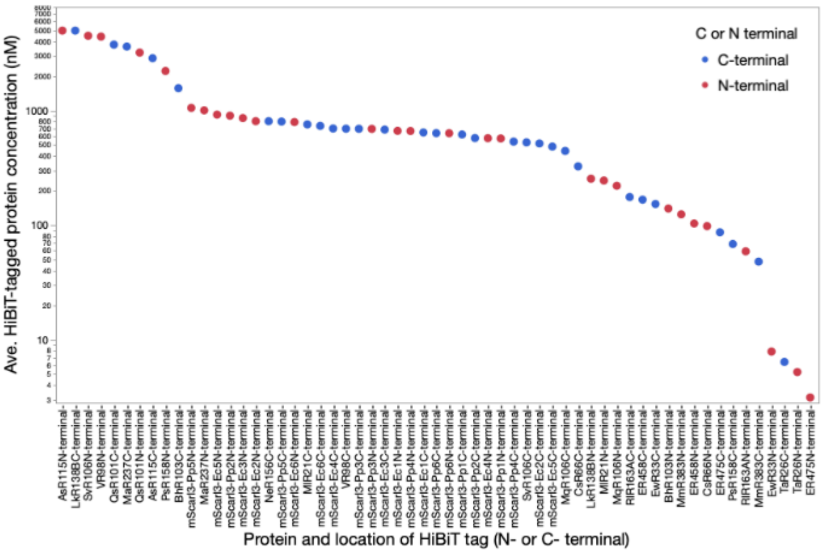
When designing a dataset for ML, two major questions arise: what types of data and how much of it do you need? Our approach tests assays of varying resolution and throughput. In order to answer the questions about data types and amounts, we need to collect data of different types. First-pass modeling of this data will be used to understand data scaling laws and guide us in larger scale pilot experiments. We plan to first generate arrayed data using HiBiT and mass spectrometry. This data will support onboarding of pooled methods like proteomics and a pooled growth based assay. To date, we have generated HiBiT assay data, discussed below.

Results

Our first round of experimentation focused on singleplex (SPX) assay feasibility in *Escherichia coli* (Ec) and *Pichia pastoris* (Pp). The goal of assay feasibility is to identify an assay data that shows quantitative protein expression, as supported by an orthogonal readout. Subsequent work to scale the assay, or use it as a way to monitor quality of a high-throughput method would proceed. We will focus on Ec data for a single assay below. "Due to its popularity for protein expression, an *E.coli* BL21(DE3) related strain, BL21-AI, was used[4,5]. Similarly, the widely used pET28a vector was used as the backbone for expression[6].

We wanted to first test two SPX assays: HiBiT and mass spectrometry. HiBiT measures protein abundance using a split NanoLuc luciferase system. Proteins of interest (20 diverse ORFs and 10 re-coded mScarlet-I3) were N- or C-terminally

Figure 2: Ladder of *E.coli* HiBiT signal



tagged with the 11 amino acid HiBiT peptide tag (a portion of NanoLuc). The HiBiT complementing LgBit polypeptide is added after cell lysis. When HiBiT and LgBiT interact, they reconstitute the luminescent enzyme, NanoLuc. Luminescence intensity is proportional to the abundance of the HiBiT tagged protein of interest[7]. HiBiT and mass spectrometry were chosen for SPX assay feasibility because they have small tags, the readouts are proportional to protein expression, they require limited optimization, and they are easily extended to other organisms (important for future dataset growth).

The experimental workflow for feasibility of HiBiT and mass spectrometry is assay independent (Figure 1, bottom panel). In fact, the same basic steps for sample generation would be used across most assays. Since we are interrogating only soluble protein expression, the soluble fraction would be used to measure target protein expression.

For assay feasibility, we used 2 sets of control open reading frames (ORFs), a diverse set of ORFs[8], and a fluorescent protein (FP) sequence (mScarlet-I3) that has been codon varied multiple times to produce unique sequence variants. We hypothesized that the diverse ORFs will have a range of expression and could be used as an expression ladder for downstream experiments. We included the codon varied mScarliet-I3 (FP) set because it has been shown that a range of GFP expression can be produced by changing codon usage in Ec[9]. We will not know the expression of these mScalet-I3 variants in Ec *a priori*, however fluorescence provides a quick orthogonal readout to HiBiT and mass spectrometry therefore we feel these proteins are valuable to include." [1]

There is a ladder of HiBiT signal in *E.coli* among the diverse ORF set (Figure 2). Unsurprisingly, the placement of the HiBiT tag on N- or C-terminus has a different impact on HiBiT signal depending on the protein. Curiously, there is not a strong correlation between fluorescence of the FP variants and the observed HiBiT signal. Mass spectrometry will be used as an orthogonal assay to validate protein expression in these strains.

Methods

Input glycerol stocked strains were grown overnight at 37°C (pre-culture). These were subcultured into a fresh production plate and this was grown for 2 hours before arabinose and IPTG were added to induce protein expression. After induction, plates were grown at 25°C overnight. 250µL of the production plate was spun down, followed by a freeze/thaw cycle. Lysis buffer was added and the plate was incubated at 25°C for 30 minutes. Following this, the lysis plate was centrifuged and 100nL was used in the final HiBiT reaction following the manufacturer's protocol. See Appendix for additional details.

Discussion

Our preliminary data highlight both the feasibility and limitations of the HiBiT assay for measuring soluble recombinant protein expression in *E. coli*. The observed HiBit response among the diverse ORF set confirms that the HiBiT assay can provide a quantitative readout of protein abundance. However, the lack of strong correlation between the fluorescence signal from the codon-varied mScarlet-I3 proteins suggests additional factors are influencing the readout beyond simple protein abundance.

To address these incongruencies, we propose establishing a ground-truth for our data by incorporating mass spectrometry-based quantification as an orthogonal approach to validate protein abundance. In parallel, we will continue to explore whether the HiBiT assay poses similar challenges in *P. pastoris*.

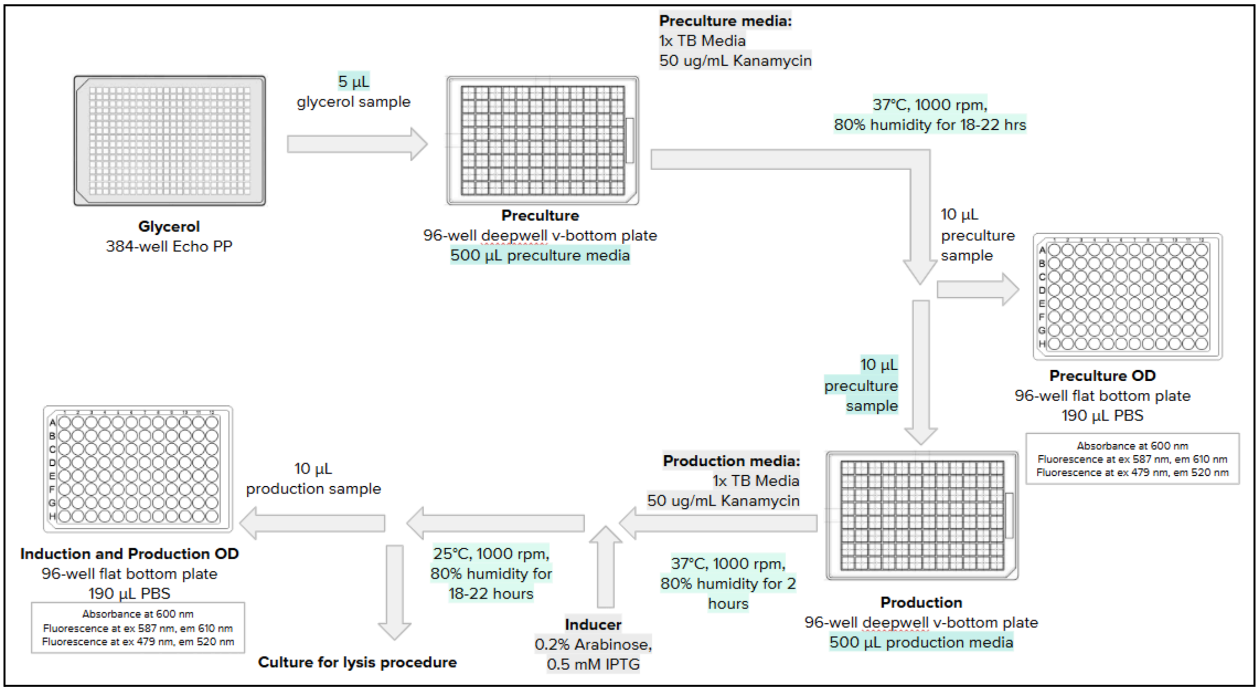
These findings emphasize the importance of a multi-modal approach in constructing a robust, organism-spanning protein expression dataset. Rather than relying solely on large, unlabeled datasets used in unsupervised learning (e.g., multiple sequence alignment-based models or protein language models) or small, labeled datasets used in supervised learning (e.g., regression or neural networks trained on bespoke proteins), we aim to strike a balance by creating high quality datasets on a broad class of proteins. This approach aims to generate a dataset that will enable more accurate and generalizable predictions of soluble protein expression. With additional validation, we envision this dataset serving as a foundation for developing machine learning models capable of predicting soluble protein expression across diverse host organisms, ultimately reducing the need for trial-and-error approaches in protein expression.

References

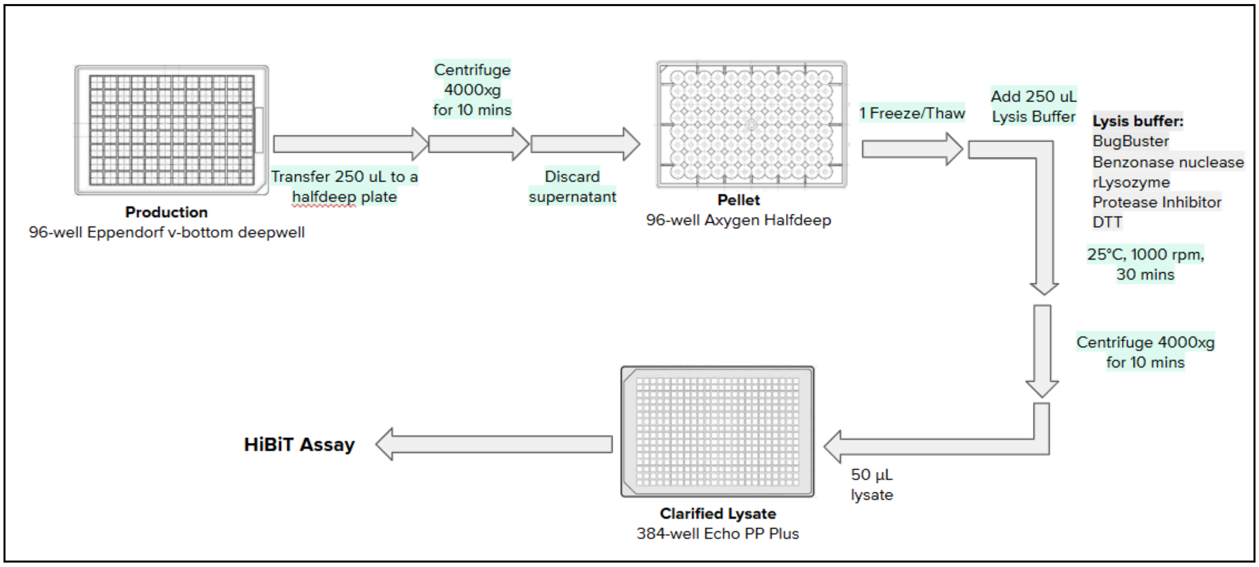
1. Gaber, A. *et al.* A strategy for scalable data collection of soluble protein expression in diverse hosts. Preprint at https://doi.org/10.5281/ZENODO.14014028 (2024).
2. Hon, J. *et al.* SoluProt: prediction of soluble protein expression in Escherichia coli. *Bioinformatics* **37**, 23–28 (2021).
3. Kramer, R. M., Shende, V. R., Motl, N., Pace, C. N. & Scholtz, J. M. Toward a molecular understanding of protein solubility: increased negative surface charge correlates with increased solubility. *Biophys. J.* **102**, 1907–1915 (2012).
4. Bhawsinghka, N., Glenn, K. F. & Schaaper, R. M. Complete Genome Sequence of Escherichia coli BL21-AI. *Microbiol Resour Announc* **9**, (2020).
5. One Shot^{TM} BL21-AI^{TM} Chemically Competent *E. coli*. https://www.thermofisher.com/order/catalog/product/C607003.
6. Shilling, P. J. *et al.* Improved designs for pET expression plasmids increase protein production yield in Escherichia coli. *Commun Biol* **3**, 214 (2020).
7. HiBiT Protein Tagging Technology. https://www.promega.com/resources/technologies/hibit-protein-tagging-system/.
8. Boël, G. *et al.* Codon influence on protein expression in E. coli correlates with mRNA levels. *Nature* **529**, 358–363 (2016).
9. Constant, D. A. *et al.* Deep learning-based codon optimization with large-scale synonymous variant datasets enables generalized tunable protein expression. *bioRxiv* 2023.02.11.528149 (2023) doi:10.1101/2023.02.11.528149.

Appendix

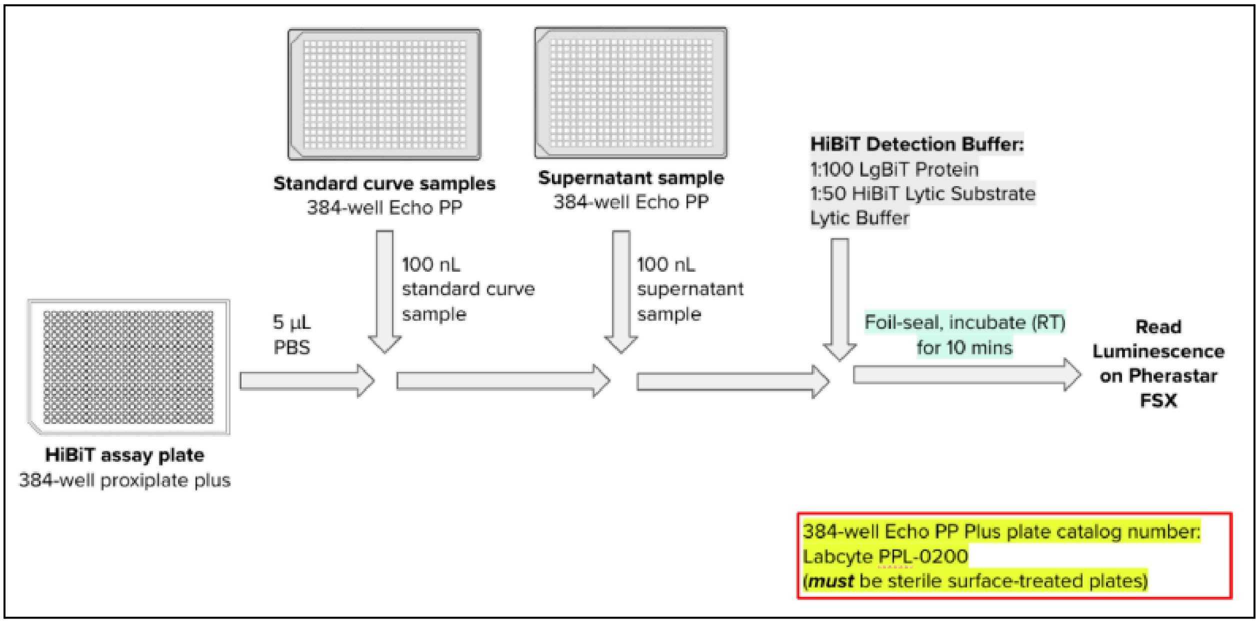
Production sample generation



Lysis of Samples



HiBiT assay plate generation



384-well Echo PP Plus plate catalog number:
Labcyte PPL-0200
(***must*** be sterile surface-treated plates)