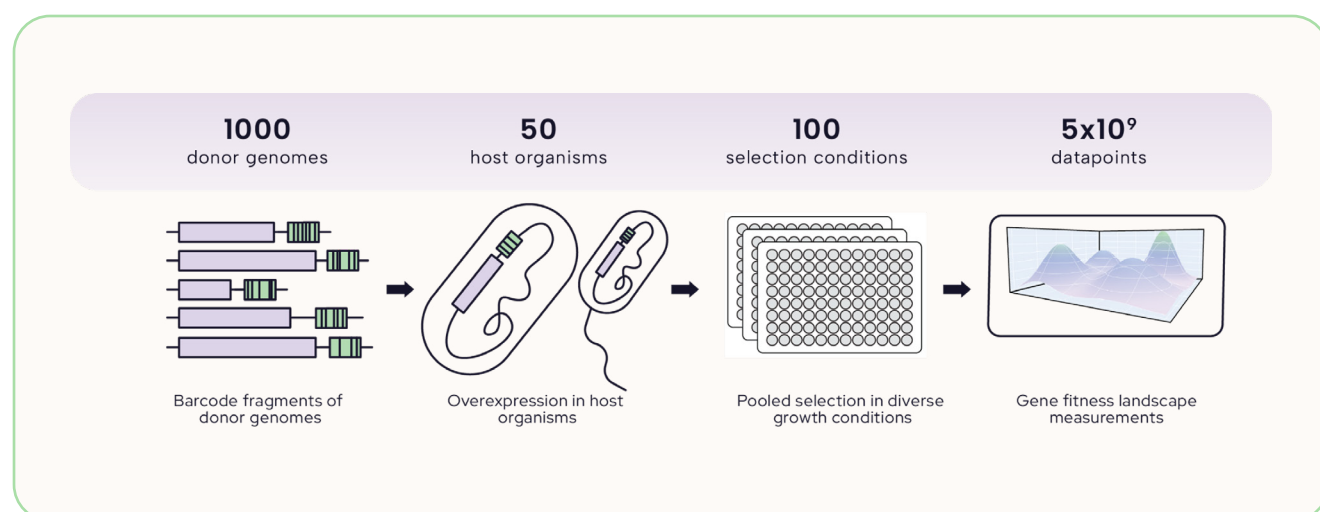


# Phenotyping uncharacterized microbial genes with a large-scale bacterial functional genomics dataset



A scalable experimental proposal outlining:

- 5x10<sup>9</sup> fitness measurements across diverse genes, hosts, and environmental conditions.
- A phased, de-risked strategy for scaling up data generation.
- Closed-loop, AI-guided experiment selection to reduce data collection costs by ~10×.

# Project Leadership

## The Align Foundation

Erika Alden DeBenedictis

Shantal Al Habib

Peter Kelly

## Pioneer Labs Proposal Lead

Jonathan Liu

## Pioneer Labs

Erika Alden DeBenedictis

Harley Greene

Nathan Hicks

Karin Isaev

Jordan Mancuso

Fatima Martin

Una Nattermann

Max Schubert

Devon Stork

Edward Sukarto

## Reviewers

Adam Deutschbauer

Adrienne Hoarfrost

Yunha Hwang

---

## About the Align Foundation

The proposal outlined in this document is under consideration for funding by The Align Foundation in partnership with Pioneer Labs. The Align Foundation advances predictive modeling in biology by creating reproducible, open biological datasets through global collaboration and automation. Founded in 2021, Align develops experimental platforms and coordinates large-scale data generation across academic and automation partners. Its end-to-end process includes community roadmapping, protocol standardization, reproducible data collection, and benchmarking predictive models. Align's mission is to create a world where biological data is more reproducible, scalable, and shareable. Learn more at [alignbio.org](https://alignbio.org) and follow us on [LinkedIn](#), [X](#), and [BlueSky](#).

## One Page Summary

Characterizing the function of microbial genes has wide-ranging importance, from basic biology research to applied strain engineering. However, this characterization is incomplete and heavily skewed towards genes from culturable microbes that can be grown in the lab, leaving many genes largely unannotated. To tackle this challenge, we have developed a platform to insert microbial genes into common host strains in high throughput. We then characterize these genes by subjecting the engineered strains to stressful conditions, linking function to improved fitness in a selective pressure.

This type of microbial functional genomics has been fruitful in the past. However, application of functional genomics to model training is currently limited by dataset size. We propose to develop a scalable platform to conduct these heterologous overexpression screens at scale. Starting with small proof of concept experiments, we propose a stepwise framework for eventually building a dataset consisting of 5 million inserted genes across 50 host strains and 100 selection conditions. Notably, our approach will utilize computational analysis in the loop of experimental design to streamline the final dataset acquisition, reducing costs by an order of magnitude while optimizing for data quality. This “Tesseract” dataset will yield diverse phenotypic fingerprints that will accelerate characterization of diverse protein function.

The Tesseract dataset, beyond immediate relevance for biological research, presents an immense opportunity to leverage AI modeling for a range of computational tasks. Across the three Tesseract dimensions of donor genes, host strains, and selection conditions, we will build models that: 1) leverage shared homology between genes to predict protein variant effects, 2) generalize measurements between various host strains to predict gene transfer success between donor genes and recipient hosts or organisms, and 3) functionally annotate classes of measured genes to predict the function of unobserved genes from unculturable and unmeasured microbes.

All data and related computational code will be publicly available via the Align Foundation website, supporting community access and reuse. We also point out several directions of future inquiry that will expand the dataset into further use cases, both experimentally and computationally. Overall, the Tesseract will enable great progress in phenotyping diverse microbial genes and present opportunities for AI to tie genetic sequence to downstream function.

# Context, Significance, and Impact

## Introduction

A microbial phenomics atlas would be an invaluable resource for applications ranging from basic science (e.g., characterizing genes of unknown function) to applied engineering (e.g. optimizing strains to thrive on low-cost feedstocks that are plentiful on Earth or Mars). Functional genomics data allows us to understand how regions of microbial genomes contribute to the growth and survival of their hosts in multiple environmental stressors. When collected at large scale, functional datasets are a rich resource that can further be used to create predictive AI models that accelerate science and engineering throughout microbiology.

Previous work has shown that genome-wide loss-of-function screens in culturable microbes (e.g. TnSeq) can produce broadly valuable insights on genes of unknown function<sup>1</sup>. One key 2018 study conducted these genome-wide knockout screens across many bacteria and in many conditions, and used this data to annotate genes of unknown function and improve existing functional annotations<sup>1</sup>.

While useful, loss-of-function screens miss out on functions that may only be revealed via overexpression in heterologous hosts, such as gene functions from hard-to-culture microbes. More recent attempts have used overexpression screens to gather data in tens of environmental conditions to obtain a phenotypic fingerprint for each added gene<sup>2</sup>, or explored using non-model microbes as the source of DNA<sup>3</sup>. However, overexpression screens have thus far not been used to systematically map gene phenotypic fingerprints across diverse conditions, forming the motivation for this proposal.

Recently, Pioneer Labs has created a version of overexpression screen methods to be higher-throughput, applicable to more host contexts, and agnostic to the DNA source. Data generated by this modified technique, as well as the data of others<sup>4,5</sup>, has demonstrated that genes can be successfully expressed across a much wider phylogenetic distance than was explored in the

2024 paper, suggesting that this technique can be used to map the enormous microbial diversity that remains unexplored<sup>6,7</sup>. This makes it uniquely possible to quantify functional properties of genes from unculturable organisms or complex metagenomes, opening up vast swathes of the tree of life to functional characterization.

In addition, we are committed to massively scaling up our overexpression screen method to leverage the exponential potential of AI trained on big data. History has repeatedly shown that breakthroughs in machine learning – from image pattern recognition to protein structure prediction – are typically preceded by the collection of large, high-quality datasets. Our envisioned dataset (the “Full Scale Tesseract”) comprises  $5 \times 10^9$  unique datapoints spanning 1000 donor genomes, 50 host strains, and 100 conditions, a scale much larger than any previous overexpression screen dataset. With this dramatic increase in scale comes numerous opportunities for utilizing AI to predict the function of unknown genes.

Rigorous scale-up is a difficult engineering challenge. We have designed a checkpointed scale-up procedure to take us from a proof-of-concept overexpression screen (~5000 genes, 1 host strain, 1 condition) to an eventual Full Scale dataset (~5M genes, 50 host strains, 100 conditions). Crucially, our strategy incorporates closed-loop computational biology to streamline experimental design and reduce the cost of such a large dataset effort by an order of magnitude.

The Full Scale Tesseract represents a new kind of rich dataset that unlocks prediction of gene function across the tree of life. By scaling to  $5 \times 10^9$  datapoints, we will bridge the gap between the microbial genotype and phenotype and unlock novel opportunities for AI to understand currently uncharacterized biology. A publicly accessible dataset of this size will catalyze the creation of new kinds of prediction and foundational models to accelerate basic research, applied engineering, and collaborative discovery across the life sciences.



Figure 1: Phased scale up of the Tesseract dataset.

## Existing datasets

The field of microbial genome phenotyping assays is well developed, and contains existing datasets that indicate the utility of additional data. Here, we survey the broader landscape of existing datasets. In general, a number of existing datasets conduct loss-of-function screens of several microbial genomes, or conduct overexpression screens of novel DNA in a single host, but none to date holistically sample a significant space of donor species, host species, and selective conditions as proposed in the Tesseract dataset.

Genome-wide characterization of microbial gene function generally falls into two categories: 1) loss-of-function screens and 2) overexpression screens. Loss-of-function screens include techniques such as TnSeq, a classic technique that has been adopted broadly throughout the microbiology community and used extensively. In TnSeq, genes are disrupted with transposon insertions to study bacterial growth in certain conditions, allowing one to annotate genes of unknown function by linking disrupted genes with patterns of their mutant phenotypes. Later techniques

such as RB-TnSeq<sup>8</sup> and CRISPRi-seq<sup>9</sup> iterated on this method by increasing throughput or replacing harsh gene disruption with inhibition. Such whole-genome scale loss-of-function screens have been successful at classifying gene function<sup>1</sup>. However, the major limitation of loss-of-function screens is that they only work in culturable, genetically tractable microbes, which comprise only a small fraction of the entire microbial tree of life. They are also only capable of characterizing the phenotypes of endogenous genes within those microbial genomes.

In contrast, overexpression screens involve transferring genes from other organisms in culturable microbes, enabling characterization of genes originating in diverse, non-model microbes. Here, gene function is assessed via phenotypes induced from overexpression. These screens can also be combined with auxotrophic strains to produce a complementation assay, where an introduced gene can directly rescue the function of a knocked out gene<sup>4</sup>, providing a clean assignment of gene function. Methods such as Dub-seq<sup>2</sup> or BobaSeq<sup>3</sup> demonstrate high-throughput approaches for overexpression screens that this proposal takes direct inspiration from.

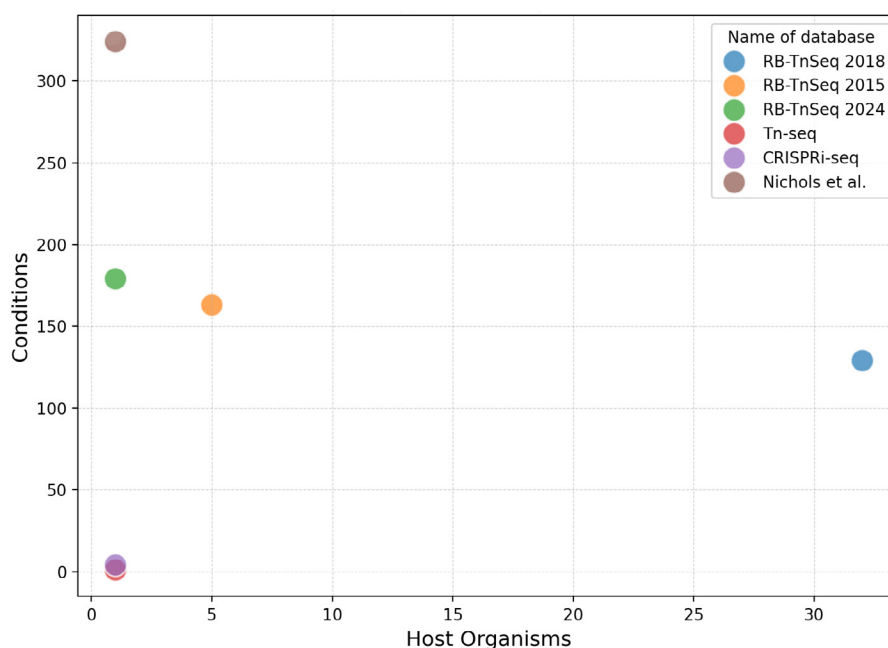


Figure 2: Summary of existing loss-of-function screen datasets. Citations and data can be found in [Supplemental Table 1](#).

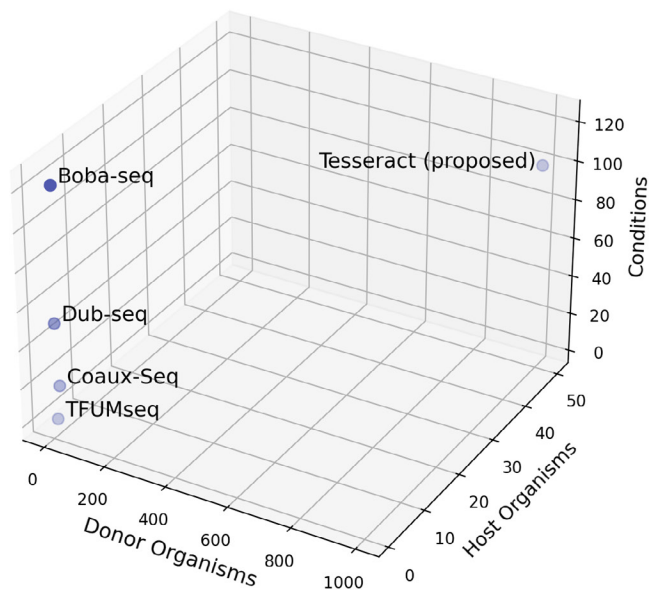


Figure 3: Summary of existing overexpression screen datasets. Citations and data can be found in [Supplemental Table 1](#).

## Method

Pioneer Labs is developing a functional genomics data acquisition protocol that maximizes the potential to characterize gene function at scale across many DNA donor organisms, expressed within diverse host microbes, and across many stressor conditions. Current protocols pull from many related functional genomics screening methods in the literature, and are most similar to Boba-Seq (2024)<sup>9</sup> and CIFR (2024)<sup>10</sup>. Additionally, we have recently de-risked methods for creating libraries containing pooled genomic DNA from many donor organisms in multiple hosts. You can read more about our results to date in our experimental<sup>11</sup> and computational<sup>12</sup> functional genomics technical essays.

Briefly, donor genomes or metagenomes are fragmented, bar-coded, and inserted into host strains in high throughput. The

fragments are between 3–10 kb, making them large enough to contain entire operons, and tiled at ~10x coverage. The pools are then passaged in a chosen environmental condition, and DNA sequencing of barcodes is used to characterize enrichment of library members. Depending on the functional characteristics of the inserted fragments, each variant's survival behavior (i.e. evolutionary fitness) in the environment may increase, decrease, or stay unchanged. Computational analysis can then compute quantitative evolutionary fitness scores for each DNA fragment. An inference model can then be used to extract the fitness contributions of individual genes, due to the high coverage of inserts. By measuring the contribution of each gene across a variety of environments, expressed in the context of diverse microbial host strains, we can obtain a phenotypic fingerprint of each gene that holistically captures its function in a wide range of contexts.

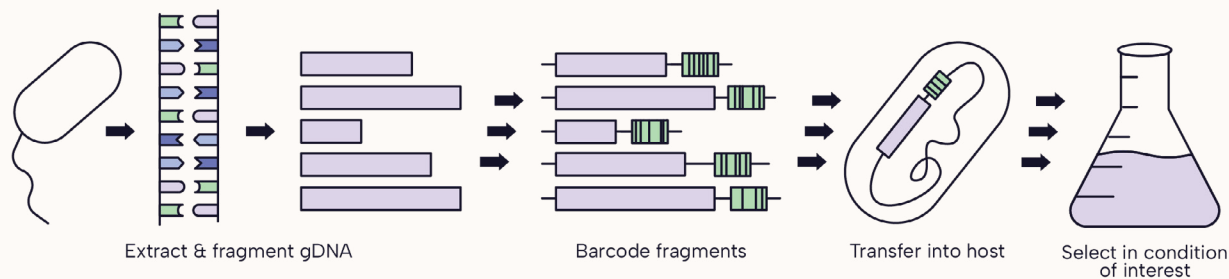


Figure 4: The Tesseract workflow for high-throughput functional screens.

## Scale Up Approach

The field of transposon and insertion screens is well explored, with numerous studies successfully gathering data and demonstrating its utility. The challenge in gathering data at a larger scale is to determine the details of data collection and sample selection that will result in high-quality, valuable data.

**Our proposed scale-up strategy accomplishes two goals:**

1. Distinguish amongst the many variations of these protocols to determine the optimal data collection strategy
2. Gather preliminary data that is necessary to make data-driven decisions about the samples that should optimally be tested next

These goals are encompassed in a three-stage strategy that allows us to expand data collection several orders of magnitude beyond any existing study in a series of steps. Each step builds conviction that we have identified the best methodology and chosen the best samples for measurement at each new stage.

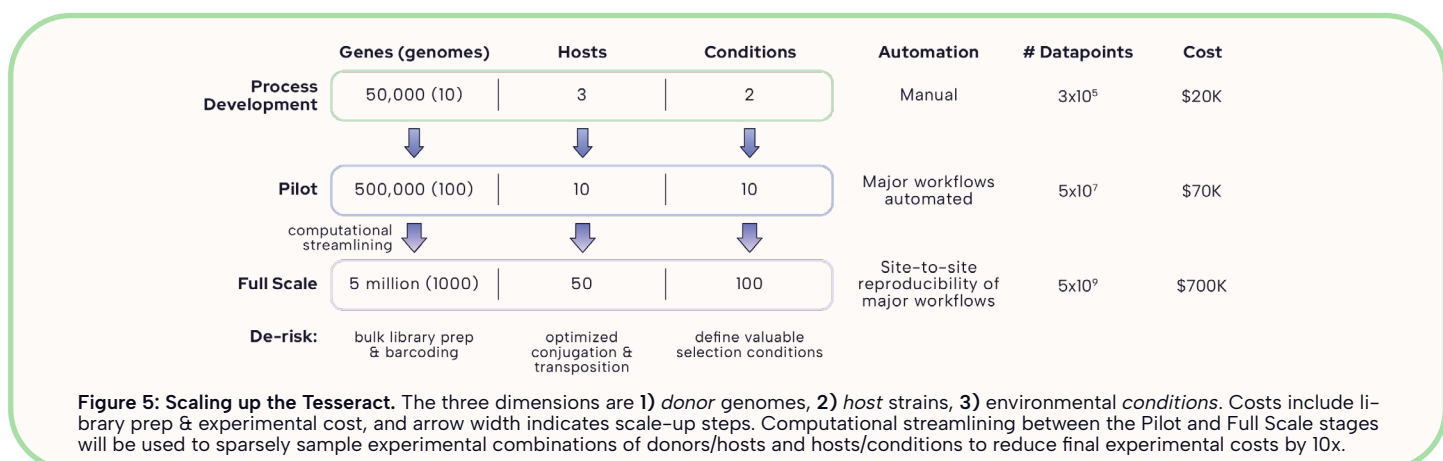
Within the stages are key technical questions that determine the best way to gather valuable, high-quality data. We outline a strategy for resolving each technical question over the three stages of scale-up. This allows us to determine the optimal strategy for each data collection methodology question prior to full-scale data collection. Details of the individual technical questions and associated de-risking experiments are provided in the table below.

**Table 1: Tesseract scale up de-risking strategy**

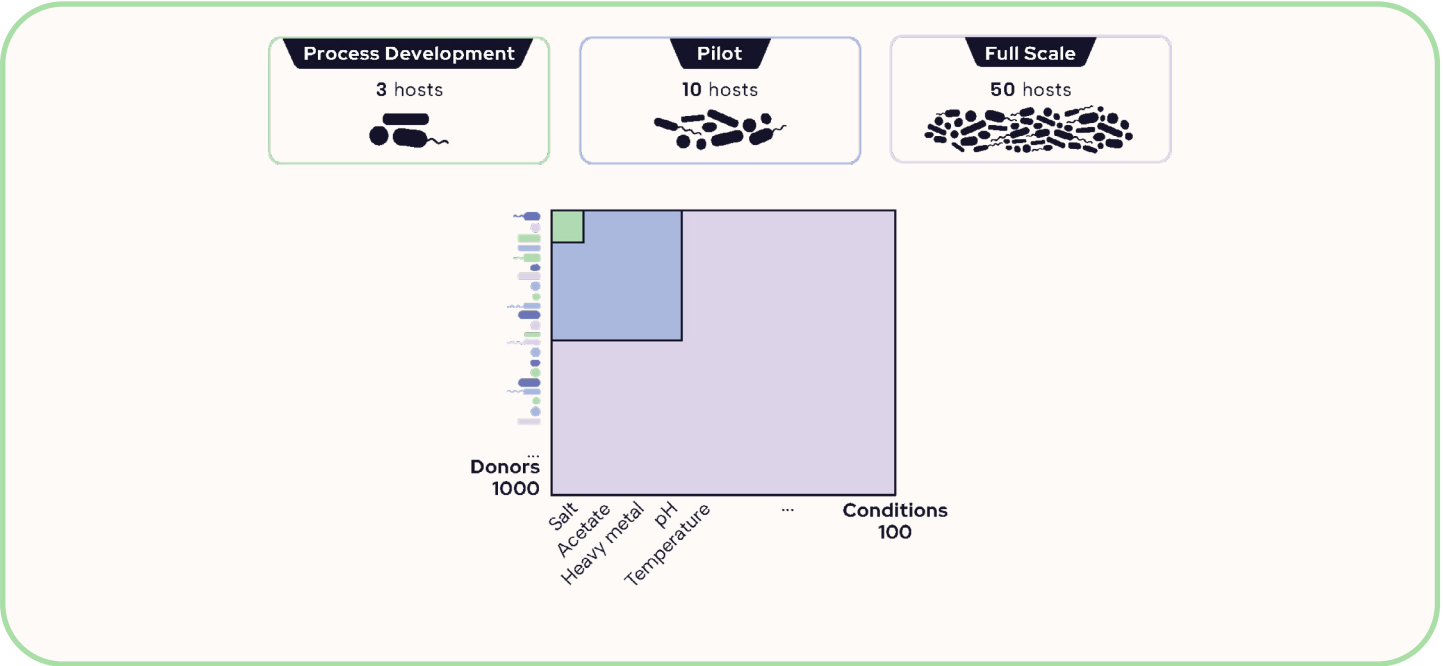
Data collection methodology question	De-risking experiment (experimental description)	Scale-up stage when de-risked
Optimal expression cassette	High-resolution <i>H. elongata</i> → <i>E. coli</i> experiment in various plasmid cassettes	Process Development
Optimal expression levels	High-resolution <i>H. elongata</i> → <i>E. coli</i> experiment with titrated expression levels with a suite of promoters	Process Development
Optimal tradeoff between library size versus data quality	Comparison of high-resolution Initial Tests vs. pooled donor genome library in Process Development	Process Development
Optimal number of technical replicates (i.e., selecting the same library multiple times in the same condition), and samples of various stringencies or over a timecourse	Multistringency experiment	Process Development
Relationship between gene transfer success and phylogenetic distance	Auxotrophy data (internal and public)	Pilot
Computational methods to choose optimal experiments	How do we prioritize which 10th of the full-scale we gather? Which donor/host and host/condition combination are most likely to yield non-zero fitness data?	Pilot
Ability to automate and standardize major components of data production	How do we automate selection and sample preparation?	Pilot
Site-to-site reproducibility of data acquisition	How do we scale up data collection with multiple facilities to ensure the robustness, scalability, and reproducibility of the dataset?	Full-scale

The above questions can be addressed using the scale-up plan outlined below. The stages are separated by the number of donor genomes that are used in the experiment, starting with Process Development (10 pooled donor genomes) and Pilot (100

pooled donor genomes). By the conclusion of the pilot, we will be prepared to collect the Full Scale dataset, which pulls from 1000 donor genomes.



# Process Development



	Donors	Hosts	Conditions
Scaleup number	10	3	2
Choice criteria	Experimental convenience, genotypic diversity, likelihood of containing genes that confer advantages in the two chosen selective conditions.	<i>E. coli</i> , <i>C. necator</i> , and <i>P. putida</i> . Chosen to explore the feasibility of scaling to gram negative hosts in well-characterized microbes.	Salt stress, acetate stress. Chosen as two conditions already well studied in the literature that work readily in liquid culture.

In Process Development, we focus on building the tools necessary to pool multiple donors into a single barcoded library. The first goal is to pool 10 donor genomes for a selection experiment and demonstrate the ability with three host strains and 1-2 selective conditions.

These 10 donor genomes will largely be chosen from a combination of experimental convenience, genotypic diversity, and the likelihood of containing useful extremophilic genes. The primary goal is to develop the engineering infrastructure to create successful multi-genome libraries. Our scheme for pooling donor genomes uses a short 8bp second barcode, such that each genomic fragment can be uniquely mapped to its parent donor. The primary technical concern is the evenness of the final library across donors, influenced by which stage of experimentation to pool the individual parent donor libraries. This data will also provide a first indication of the relative value of comparison between multiple homologues of the same protein versus the value of the impact of a single protein sequence in multiple expression hosts.

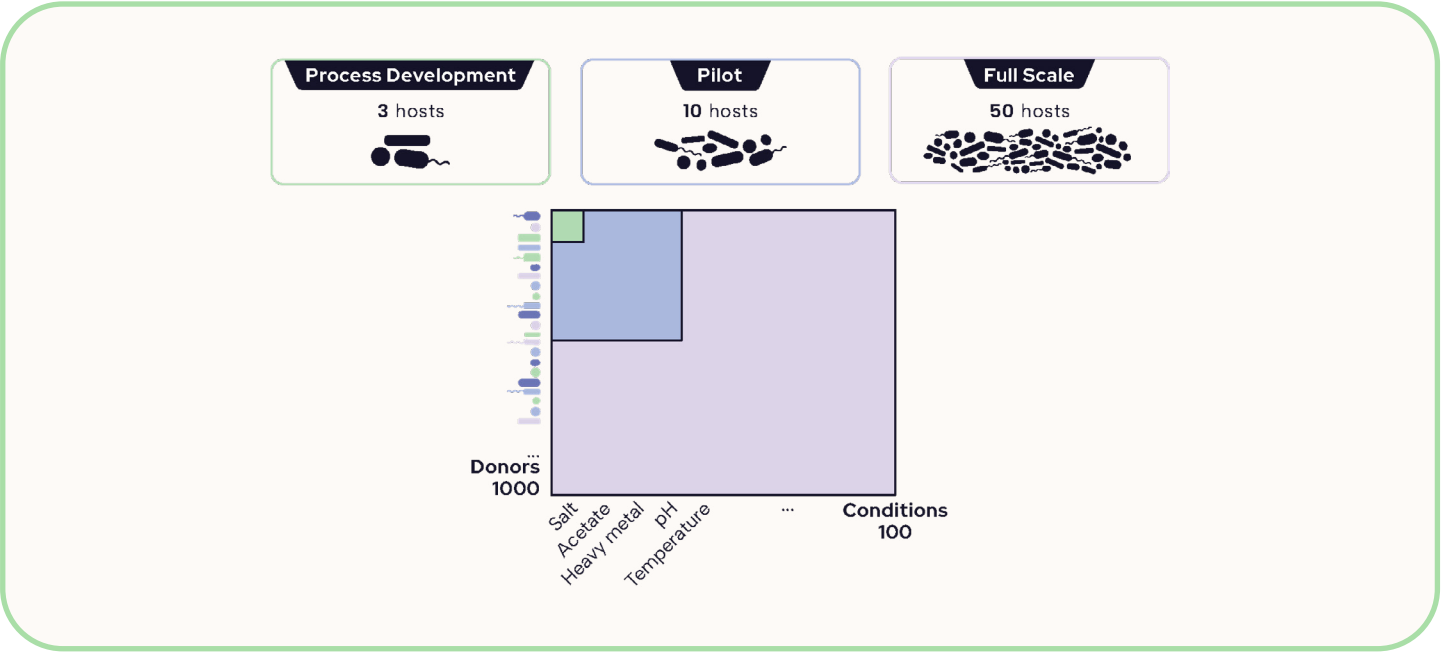
In addition, we will prioritize gathering data in a few hosts beyond just *E. coli*, in order to determine the feasibility of experimentation beyond common model microbes. For example, other studies have used organisms such as *C. necator*, but to our knowledge these do not necessarily use a consistent library creation protocol or an expression architecture that is applicable to all organisms. The proposed method will in principle allow for arbitrary donor-host combinations. The first few hosts in the Process Development stage will be selected by both ease of engineering (we require that they can be conjugated with high efficiency) and phylogenetic distance (we would like to demonstrate broad host range utility). We expect to gather data in *E. coli*, *C. necator*, and *P. putida* during process development.

**Key milestones during Initial Process Development:**

- De-risk a protocol for pooling donor genomes
- Establish success in host organisms beyond *E. coli*
- Obtain initial data to guide strain selection in the Pilot (both donors and recipients)



Pilot



	Donors	Hosts	Conditions
Scaleup number	100	10	10
Choice criteria	Four criteria include: 1) Wide extremophilic diversity. 2) Sufficient phylogenetic similarity to the hosts to yield results. 3) Likely to contain genes that confer fitness benefits in the conditions tested. 4) Of interest to the broader scientific community as part of specific biological studies.	<i>E. coli</i> , <i>C. necator</i> , and <i>P. putida</i> , plus seven additional hosts to be determined based on the results of Process Development.	Salt and acetate stress, together with broader industrially relevant conditions such as oxidative stress, heavy metal tolerance, extreme pH tolerance, and temperature sensitivity.

The Pilot stage constitutes the first major scale up in this project. It will be a proof of concept of a repeatable protocol for obtaining 5 million data points from 100 donor organisms in a single pooled experiment that only costs ~\$70k. This will lay down the technical foundation for further expansion to the Full Scale experiment, which will utilize the same basic protocols (see Figure 6). In addition, we plan out substantial computational analysis of the Pilot dataset to streamline Full Scale design and reduce final costs by 10x by focusing on the most information-rich datapoints to collect in the future (i.e., prioritizing on 1) gene transfer success probability and 2) host viability in a given condition).

Achieving the pilot requires engineering improvements on automation and experimental design as well as further inquiry into technical unknowns regarding the science. The primary chal-

lenge lies in selecting a broad range of donors, hosts, and environments that will yield biologically useful and meaningful results, in a fashion that can be achieved with a well-automated and streamlined protocol. Several of the de-risking technical factors outlined above will be extremely relevant at this stage, such as the relationship between phylogenetic distance and gene transfer viability, as well as the optimal number of datapoints to measure for each donor/host/condition combination.

Success at this stage entails a straightforward protocol for a pilot tesseract experiment that produces 50 million datapoints and costs approximately \$70k. Further scale-up is primarily limited by resources, rather than technical details.

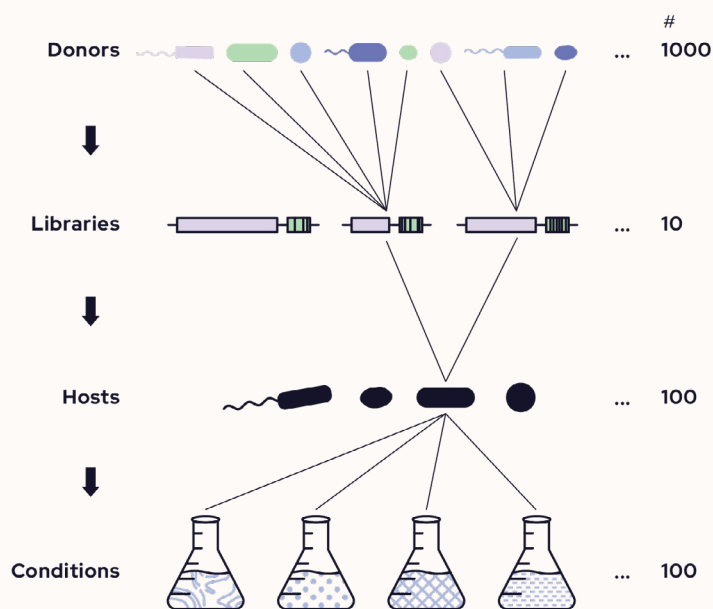


Figure 6: Pilot and full scale stages will collect data for many donor organisms in a single pooled experiment.

## Automation of key workflows

The Pilot is designed to conduct an experiment at the maximum size and obtain a final cost/datapoint estimate. The future collection of data during the Full Scale stage will involve collecting data exactly as it is done in the Pilot many times, rather than further scaling the throughput of a single run. As a result, a key goal of the Pilot stage is to demonstrate automation of key workflows and use this automation to reduce cost. See [Cost Model and Equipment Requirements](#) for more information.

Key opportunities to automate Tesseract data collection include:

- **Automated growth-based selection and sampling.** Once libraries are created and conjugated, plate-based automation can be used to conduct selection and sampling in many stressor conditions at once. This task will require automation very similar to that used in GROQ-Seq, an existing automated method for Barseq data collection that has been onboarded at several facilities. We suggest adjusting GROQ-Seq protocols to suit tesseract data generation.
- **Automated NGS sample preparation.** Once samples are acquired, DNA extraction and NGS preparation are needed to prepare samples for sequencing. NGS sample preparation is a commonly automated process.

## Donor Genomes

The Pilot will scale up the number of donor genomes in a pooled library from 10 to 100. This poses a major question in how to choose a large number of donor genomes. Broadly, we aim to choose donor genomes with four considerations in mind:

- Wide extremophilic diversity
- Sufficient phylogenetic similarity to the recipient to yield results
- Likely to contain genes that confer fitness benefits in the conditions tested
- Of interest to the broader scientific community

Computational analysis of the Process Development dataset and related de-risking experiments will inform how we balance extremophilic diversity with phylogenetic similarity. In sum, the Pilot will contain half a million genes, sourced from ~100 diverse organisms. A preliminary list of nominated organisms has been developed. To choose these genomes, we used two primary methods 1) we nominated organisms based on their likelihood of containing genes that are likely to have interesting phenotypes in this assay and 2) we selected organisms to broadly tile the tree of life. Additional input from the community will be used to nominate additional genomes and choose a final 100. The preliminary list can be viewed [here](#), and nominations can be submitted through [this form](#). The list will be continuously updated as new nominations are received.

## Hosts

The Pilot will express genes across 10 host organisms, each of which must be culturable in liquid media, capable of conjugation, and have at least one selection marker. Process development will take place in *E. coli*, *P. putida*, and *C. necator*. Additional hosts will be driven by experimental results and data analysis of the Process Development stage, and by suggestions from the scientific community.

## Conditions

The Pilot will select strains in 10 cultivation conditions. As with Process Development, we will include salt and acetate stress. We will expand this set with more industrially-relevant conditions, likely including oxidative stress, heavy metal tolerance, low phosphorus tolerance, low nitrogen tolerance, extreme pH tolerance, temperature sensitivity, and additional suggestions from the scientific community.

How we will incorporate the various conditions into the experimental design will be largely informed from our de-risking experiment on multistringency. Briefly, we are considering testing each condition across a range of concentrations or intensities to see if we can read out additional selection properties beyond that which could be gained from a single concentration.

## Computational analysis for further scale-up

One primary goal of the Pilot is to provide a sufficiently large dataset for computational analysis to loop into further experimental design and dramatically reduce the cost of the subsequent Full Scale experiment. After factoring in all possible combinations of donors, hosts, and conditions, the Full Scale will contain ~500x as many datapoints as the Pilot and would cost over \$7M on paper. However, many of these combinations likely would not result in informative data, due to inefficient gene transfer or lack of genes that appreciably improve fitness in a given stressor. Instead of naively measuring all of these combinations, we will construct an active learning scheme that learns from Pilot data to prioritize which combinations to follow up on for the Full Scale experiment. Similar computation-in-the-loop work has been demonstrated to efficiently prioritize future experimental design<sup>13</sup>.

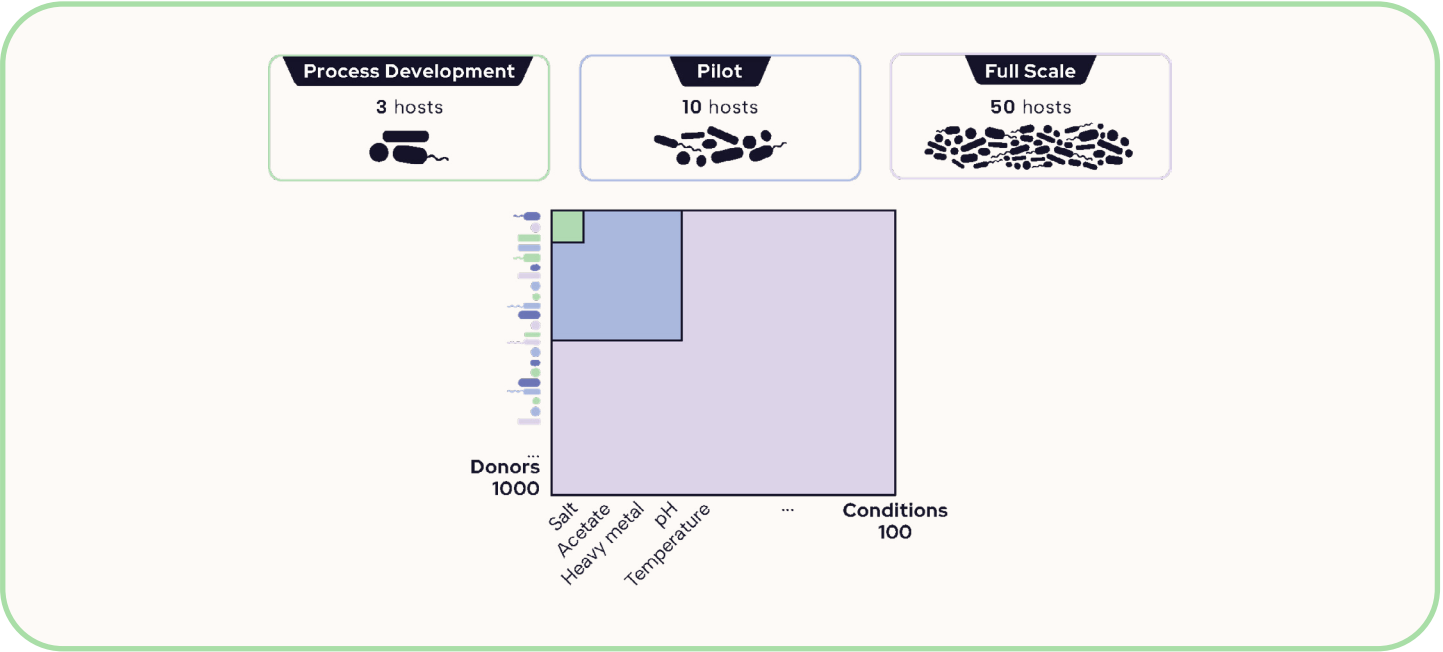
Specifically, we will create a modeling framework to analyze the Pilot data that predicts which of the combinations of donors/hosts/conditions in the proposed Full Scale experiment will be likely to yield informative results, and how to optimally pool donor organisms into sets of 100. Due to latent shared biology such as homology and convergent phenotypes between donor genomes and genes, not every possible donor/host/condition will yield equally informative results. Many donor gene/host combinations are unlikely to result in successful gene transfer due to e.g. divergent phylogeny, a pattern reported in previous overexpression screen studies<sup>4</sup> as well as in our own collected complementation assay data.

We expect that after the Pilot, there should be enough data to infer the likelihood of successful gene expression given donor organism/host strain combinations. At this stage we will develop models to predict the top 10% of experimental combinations likely to yield the most informative data, in order to streamline data collection for the Full Scale Tesseract. See [“Computational Analysis and Modeling” in the “Data Engineering and Computational Analysis”](#) section for a more detailed discussion of the gene transfer modeling that will be utilized here. By incorporating computation in the loop, we’ll focus data collection on donor/hosts and host/condition combinations that are likely to produce useful signal, thereby cutting down the size of the Full Scale experiment by a factor of ten.

### Key milestones during Pilot:

- Determine optimal number of replicates, samples, conditions, etc for best data quality
- Demonstration of an automation protocol for conducting data acquisition at scale, and a final cost-per-datapoint
- Proof point that computational analysis can be effectively used to guide sample choice, reducing the amount of data acquisition needed by an order of magnitude

# Full Scale



	Donors	Hosts	Conditions
Scaleup number	1000	50	100
Choice criteria	Pilot hosts, plus 900 additional genomes chosen computationally based on results of Pilot.	Pilot hosts, plus 40 additional hosts to be determined from community input.	Conditions will be chosen from community input and for overlap with existing datasets such as The Align Foundation's 1000 Microbes database <sup>14</sup> .

Progressing from the Pilot experiment to the Full Scale primarily involves a linear and dramatic expansion in resources to produce a dataset with vastly more libraries, hosts, and conditions. The primary constraint lies in automation, throughput, and cost – the size of the Full Scale necessitates many large sequencing runs and batches of data. At this point, we require a robust experimental protocol to hand off the operations to The Align Foundation and associated CROs.

Data collection at this scale may be best conducted by multiple sites, see [Future Dataset Expansion: Adding New Data Collection](#) sites for additional information about requirements for additional data collection facilities.

## Donor Genomes

The Pilot previously incorporated 100 donor genomes. We will include the previously used genomes for continuity, and expand with selection of 900 additional genomes. We will incorporate a strong computational component to select the additional genomes, including factors such as phylogenetic and phenotypic diversity from models developed during the Pilot stage. We will prioritize donor/host combinations that are expected to yield informative data, which should dramatically reduce the number of possible experimental combinations.

To source the genomes, we will use industry relevant strains from Align's 1000 Microbes database<sup>14</sup>, collaborate with academic labs to incorporate and test their strains-of-interest, and order any additional strains or genomic DNA from strain databanks like ATCC and DSMZ

## Hosts

We will expand the number of host organisms from 10 to 50. At this point, we will begin to solicit increased feedback from the community on host selection, with the goal of representative host diversity across a range of application areas.

## Conditions

We will scale up the number of conditions from 10 to 100 at this stage. As a result, there will be many engineering constraints on doing many condition/host combinations.

## Automation and site-to-site reproducibility

Scale up to Full Scale data acquisition will require a shift toward multiple facilities with expanded capacity to process samples. Additionally, the ability to gather quantitative data at multiple facilities and obtain quantitatively similar results is a powerful proof point for the quality of the data acquisition method. During the pilot phase, we will define the criteria and metrics for assessing reproducibility across sites. Once established, these criteria and the resulting reproducibility outcomes will be shared with the community through open science reports.

# Data Engineering and Computational Analysis

## Data Processing and Infrastructure

The bulk of the data processing pipelines are already written and hosted in a Github [repository](#). While they will be updated with slight improvements over the execution of this proposal, the primary pipelines are already functional. The overall scheme borrows heavily from the computational methodology outlined in the Boba-Seq<sup>3</sup> method.

Briefly, the data workflow takes the PacBio HiFi and Illumina sequencing of the library barcodes and donor genome fragments to generate an aggregated data table consisting of individual barcodes and their associated donor genome fragments (Figure 7). Next, we calculate quantitative fitness values for each barcode based on its change in population frequency over the timepoints of the experiment. From individual barcode fitnesses, we then infer gene-level fitnesses via a Bayesian linear mixed model, resulting in a unique fitness value for each unique gene/host/condition combination.

## Computational Analysis and Modeling

There are three categories of computational tasks where tesseract data could be applied.

- 1. Protein variant effect prediction.** Tesseract data contains homologues of proteins from many donors. Comparison between the fitness of these homologues presents a new type of dataset for modeling protein sequence → function.
- 2. Gene transfer success prediction.** Tesseract data contains many instances of the same genomic fragment expressed in multiple hosts. Comparison between fitness in multiple contexts creates an opportunity to model the determining

factors underlying protein expression, folding, and function in multiple cellular contexts.

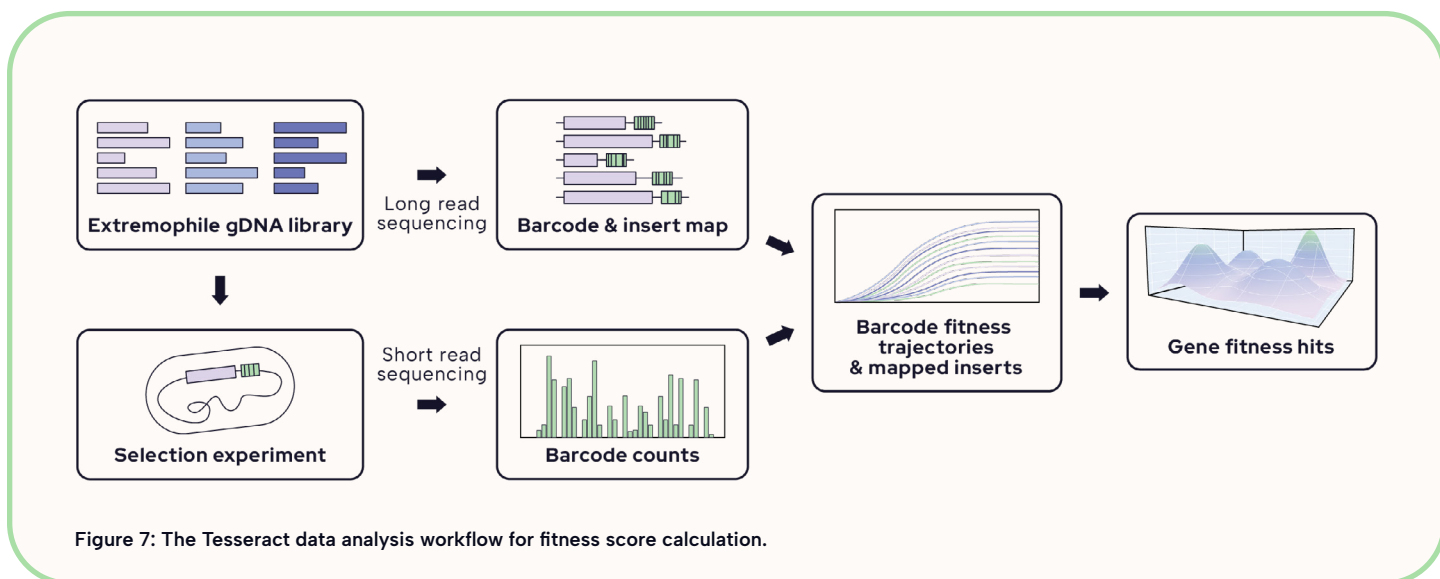
- 3. Functional class prediction.** Tesseract data allows us to measure a phenotypic fingerprint of each gene's fitness contribution across many stressor conditions. This creates an opportunity to build predictive models of novel gene sequences or genes of unknown function.

Below we describe the precedent for each of these modeling tasks and the potential applications of successful predictive models.

### Protein variant effect prediction.

Numerous machine learning approaches exist in the protein variant effect prediction problem space<sup>15</sup>, generally incorporating some mixture of supervised and unsupervised learning. However, previous works typically focus on predicting the effects of single or few amino acid mutations<sup>16</sup>, rather than the variation found between natural homologous sequences. Recent work suggests that the information found in naturally occurring homology can improve prediction of protein fitness<sup>17</sup>.

The many donor genomes in the Tesseract present an opportunity to build models to predict fitness from gene sequence. Specifically, many genes will be homologues of each other, sharing functional similarity and likely possessing similar (but not identical) measured fitnesses. To start with, we will consider each unique host and condition combination and query the set of high-fitness genes that emerge from that experiment. By leveraging the natural homology between these genes, we will build an ML model to predict fitness from gene sequence, either with classical k-mer based methods or more contemporary DNA language model



embeddings<sup>18</sup>. Success here is indicated by a model that can accurately predict fitnesses of held-out sequences from a subsetted training dataset. In an ideal case, the model can also predict fitnesses of unobserved sequence (whether natural or synthetic), which we could validate with a follow-up experiment.

One unknown parameter is the number of homologous sequences needed to train a successful model. While we may observe numerous homologous genes from the Tesseract, it may be insufficient for the task of fitness prediction from gene sequence. This motivates the opportunity for follow-up experiments with a targeted library of many homologous sequences, discussed more in “Future Dataset Expansion.”

A successful model here would greatly improve the efficiency of protein engineering. Given a host strain and target application (i.e. stressor condition), the model could predict DNA sequences that are likely to result in proteins that improve biomass production in the condition. Instead of having to experimentally screen large sequence libraries – a slow and expensive process – the model can be leveraged to conduct an *in silico* screen yielding small sequence libraries with high probability of success. Such a regime would dramatically decrease experimental costs and increase speed of protein engineering.

### Gene transfer success prediction.

The shared homology between donor genes can also be used to model the likelihood of successful gene expression + function when we examine cross-host behavior. While predicting protein variant effects is a highly explored scientific problem, modeling gene transfer between donor and host organisms is relatively unexplored. We will build models to examine and predict fitness for a range of homologous genes across the variety of host strains used in the Tesseract. For a given donor gene/host combination, the underlying gene transfer probability dictates the measured fitness.

This approach, while not new, benefits greatly from a dataset possessing the nature and scale of the Tesseract. Existing similar studies of gene transfer success – both using similar<sup>4</sup> and different<sup>5</sup> approaches – must be substantially expanded to create data for modeling. Here, the large number of donor gene/host combinations allow for an unprecedented investigation into gene transfer success across many different condition contexts.

By querying the fitnesses across all donor genes and hosts for a single condition, we can model the underlying gene transfer matrix from the variation in observed fitness. These models will incorporate different biology, from organism-level phylogenetic distance to protein-level similarity scores. Performance of these models will then inform us on the best predictors of gene transfer success. Eventually, we can investigate combining all the conditions into a single model of gene transfer success, pending results from the single-condition models.

Predicting gene transfer success is important both for basic science and for rapid natural product engineering. For basic science, successfully predicting gene transfer is the first step towards building a whole-cell systems model, as it involves modeling numerous key cellular functions beyond the immediate context of inserted gene sequence, e.g. transcription, translation, protein folding. For product engineering, there is a constant struggle to successfully express pathways from non-model microbes in lab strains<sup>19</sup>. Building models that can predict gene transfer success *in silico* would demystify the engineering process and allow scientists to target protein/host combinations that are more likely to work in the first place.

### Functional class prediction.

Finally, we can utilize the cross-condition data to build models to functionally annotate the genes in the Tesseract. From gene sequence similarity between genes, we can cluster our high-fitness genes into different groups for each condition, resulting in clusters of putative functional mechanisms. Examining these clusters across all conditions in the Full Scale Tesseract then yields a set of functional labels for each gene hit corresponding to each condition. Similar work has successfully been achieved in TnSeq deletion screens<sup>1</sup>, but this type of modeling approach has yet to be demonstrated in unculturable microbes.

The first piece of this modeling task is mainly descriptive, where we assign functional labels to the genes measured in the Tesseract. The next step is predictive – for unobserved genes across the tree of life, we can compute gene similarities with genes in the Tesseract dataset and predict functional labels for unobserved genes that lie within the functional clusters we observe. This is particularly impactful for genes from unmeasured or unculturable microbes. These can then be tested by targeted synthesis of metagenome-derived sequences for screening. As a follow-up study, we can then construct models that go beyond the functional labels defined by the conditions in the Tesseract to models that examine pathways and mechanisms in the host strains, arriving at a more systems-level ontology of functional annotations.

Importantly, Tesseract data allows protein function prediction to scale beyond manually curated functional classes like gene ontology (GO) terms. Numerous existing works have successfully used machine learning approaches to predict GO terms<sup>20–23</sup>. Our high-throughput measurement of genes across conditions allows us to generate direct annotations of functional class without relying on pre-existing GO terms, providing data for building models to predict protein function that falls outside the pre-existing GO framework. These datasets may enable a new type of protein function models for inferring molecular mechanism from phenotypic label.



The following table summarizes the key proposed modeling tasks for each category outlined above:

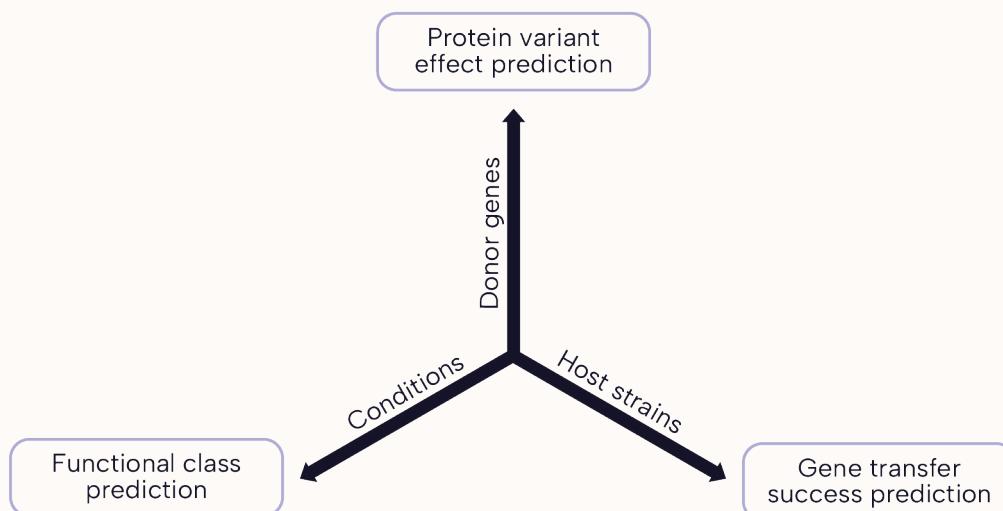


Figure 8: Computational tasks enabled by the Tesseract dataset.

Model category	Input variables	Output prediction	Training corpus
Protein variant effect prediction	Gene sequences	Fitness of the input sequence (in a fixed host and condition)	Many gene homologues for each host/condition combination, across all common bacterial gene families
Gene transfer success prediction	Gene sequences, host organisms	Fitness of the input sequence in the input host (in a fixed condition)	Each gene sequence/host combination measured across conditions
Functional class prediction	Gene sequence, conditions	Functional annotation of gene	Many genes across all conditions, measured across hosts

## Data Storage

Data will be made available in an Align-hosted database accessible to the public through a REST API. For longevity and redundancy of data, snapshots of this database at certain time points (e.g., for the initial data set, or for large data increases) will be duplicated to a third-party repository (e.g., Nature Scientific Data or Zenodo).

Data from Initial Results are already available in the Align-hosted database [here](#). Note that you'll be required to make a free login to access the data. The dataset is hosted under the tab "Tesseract." The portal contains the dataset, a link to our public [Github](#), and a README describing the data and code.



## Cost model of data acquisition

Building a simple cost model of the Tesseract shows that building libraries and creating selection-ready libraries is a relatively small cost in comparison to data acquisition through sequencing. This suggests that judicious choice of which libraries to select in which conditions is important, instead of running every possible combination of libraries, hosts & conditions.

The below model takes into account two primary costs: FTE time & reagents, with some information on automation CapEx requirements. We break down the main protocols by step, FTE cost, and automation requirements.

Assumptions:

- **To fully cover a donor organism genome, we need to sample at least 50 fragments per position.** For an average bacterial genome size of 5 million bases, and average fragment size of 5kb, we need 50,000 inserts/organism.
- **Each library contains fragmented genomes from 100 unique donor organisms,** making each library 5 million variants in size.

- **To fully associate full-length inserts with their barcodes in that library, 3 PacBio sequencing runs are needed.** One PacBio run produces 15M reads, meaning three PacBio runs are sufficient for ~10x coverage of a 5 million variant library. A single PacBio run costs \$3075, including library prep kits.
- **To fully quantify library frequency despite unevenness, we need 50x sequencing coverage per barcode per sample.** Therefore, each sample from a 5 million member library must be sequenced with 250 million depth. A Novaseq X Plus 25B 100-cycle Flowcell generates  $2.5 \times 10^{10}$  reads for \$15,000.
- **We will sequence 5 unique samples per library/host/condition combination** during selection in Process Development and 4 unique samples in Pilot and Full Scale.

Protocol step	Automation required	FTE time	Reagent cost	Total cost and time
Library construction	<u>Minimal.</u> Requires tapestation, could use a small liquid handler for tagmentation pooling.	<u>~20 days per library,</u> including genomic DNA prep/purification, tagmentation, barcoding, transformation, QC.	<u>\$115 per library.</u>  Each genome needs \$10 for genomic preps, \$7/genome for RCA (if needed), and \$8/genome tagmentation. USER assembly mix & comp cells are \$90/library.	<b>\$150 &amp; 20 FTE days</b> per library.
PacBio sequencing	<u>None.</u> Low-throughput sequencing is sufficient.	<u>1 day per library.</u>  One FTE can prep and submit one PacBio library in one day.	<u>\$9225 per library.</u>  Each PacBio run is \$3000 + \$75 in library prep kits. Three runs are needed for a 5-million-member library.	<b>\$10k &amp; 1 FTE day</b> per library.
Library conjugation	<u>None.</u> Manual plating procedure.	<u>1 day/conjugation.</u>  One FTE can conjugate five pools per week.	Minimal – media, antibiotics.	<b>\$50 &amp; 1 FTE day</b> per library/host combination.
Selection	Basic liquid handling robot for setting up conditions & transfers.	<u>500 samples/day.</u>  One FTE could run 2500 selections in a week, using 96-well plates & automation.	Minimal – media & condition additives.	<b>\$2 &amp; 0.002 FTE days</b> per combination of library, host, & condition.
Sequencing	NGS sequencing prep automation.	<u>2000 samples/day.</u>  Using sample prep automation, one FTE can prep 2000 NGS samples in one day.	<u>\$152/sample.</u>  Sequencing prep cost is ~\$2.22 per sample for polymerase, and sequencing cost is \$150.	<b>\$154 &amp; 0.0005 FTE days</b> per sample generated from selection.

## Calculating total cost & FTE time of Full Scale

Tesseract data collection is fundamentally composed of two activities

- 1. Library creation.** Creation and characterization of the donor libraries, a one-time cost that generates a permanent asset that can be reused.
- 2. Data collection.** Use of those donor libraries to conduct phenotypic data acquisition.

Creation of the libraries is relatively inexpensive, but time-consuming. In contrast, the cost of data collection is dominated by the cost of DNA sequencing, but requires very little hands-on time. Overall, data acquisition will vastly dominate the cost of the Full Scale dataset, making development of closed-loop algorithms for selecting data rich areas to focus on for data collection a high-leverage opportunity to reduce cost.

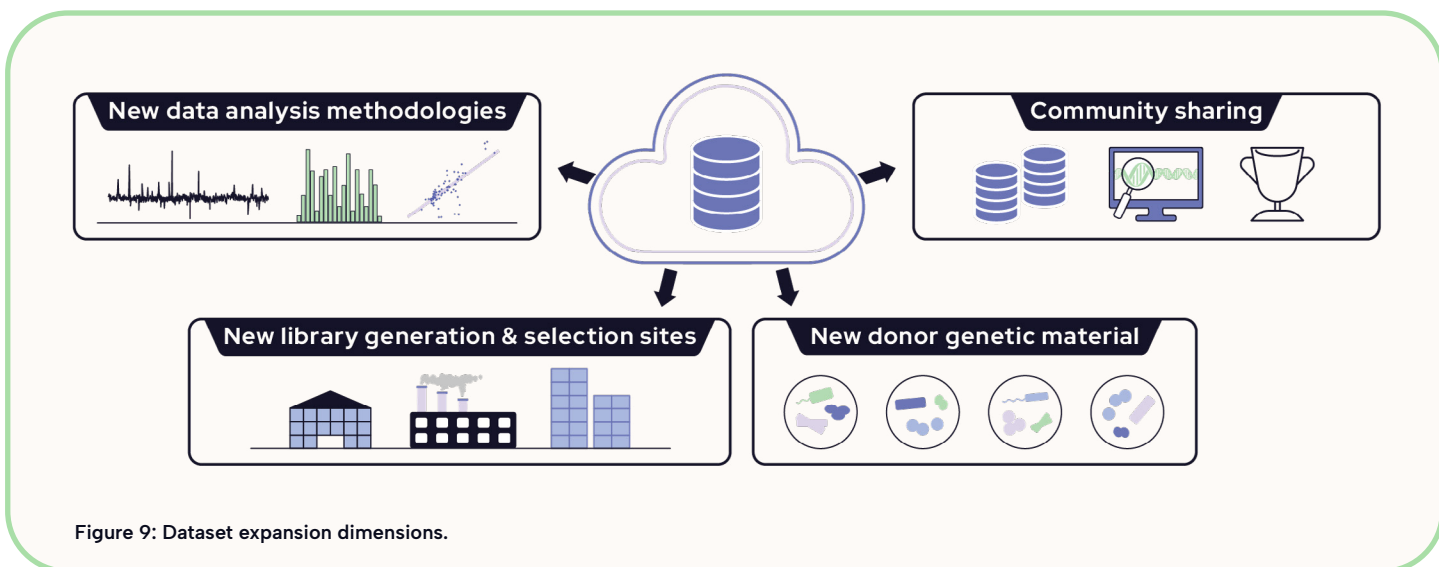
In more detail, from the above cost table we can see that there is a baseline cost of \$10,150 & 21 FTE days to build a library (composed of 100 donor genomes) and conduct PacBio sequencing to characterize its composition. Transferring that library into a new host costs roughly \$50 & 1 FTE day per library-host combination. Selecting and sequencing the resulting library-host combination costs \$616 per condition and 0.002 FTE days, assuming automation.

If we were to gather the dense matrix of all possible data with all combinations of libraries, hosts & conditions, 10 libraries x 50 hosts x 100 conditions costs a total of \$30.5 million and ~900 FTE days. Instead, we have proposed to use closed-loop data acquisition to identify the subset of these experiments that generate the richest data. Below, we show three options for how data acquisition should be paired down through judicious choice of which library/host/condition combinations to measure.

Dataset type	Donor organisms	Host organisms per library	Conditions per host	# fitness values	Library Cost	Sequencing Cost	FTE time (days)
Full sized	1000	50	100	2.50E+11	\$126,500	\$30,400,000	910
Closed-loop choice of conditions	1000	50	10	2.50E+10	\$126,500	\$3,040,000	730
Closed-loop choice of donors	100	50	100	2.50E+10	\$12,650	\$3,040,000	91
Closed-loop choice of all parameters	1000	10	10	5.00E+09	\$106,500	\$608,000	314

In total, by doing 100 host-library combinations instead of 500, and 10 selection conditions per host-library combination instead of 100, we can acquire a diverse dataset for \$700k.

## Future Dataset Expansion



This proposal lays out the baseline approach to generating a large dataset of horizontally transferred genes. However, there are many ways this could be expanded in the future to increase the breadth of DNA in the dataset, the contexts in which it is assayed and the quality of the data generated. Here we'll discuss some meaningful ways of expanding the Tesseract dataset:

### Exploring Impact of Expression Construct

The proposed dataset deliberately employs a single expression construct containing a promoter on one side and no promoter on the other, ensuring that all DNA fragments are expressed across organisms. This architecture enables detection of genes that depend on exogenous promoters versus those under native regulation, based on the directional bias by which fragments containing a gene express in one configuration versus the reverse. It also removes the need to clone libraries in a host-specific manner, allowing data acquisition to focus on measurements across many hosts and re-use of the same libraries—an area of functional genomics that has been relatively underexplored.

However, expression construct design remains a valuable and challenging area of bioengineering despite extensive prior work. An extension of the Tesseract framework could include gathering data on the same genes using multiple expression cassettes. For example, systematically varying promoters and ribosome binding sites (RBSs), or employing inducible promoters, would likely reveal different functional genes or distinct fitness effects for the same gene. One could imagine incorporating resources such as the POSSUM toolkit<sup>24</sup> to enable rapid construct variation. This represents a promising direction for future dataset expansion.

### Adding new sources of donor genetic material

All scales of the Tesseract constrain donor genomes to purchased or extracted genomic DNA from single bacterial strains. This is advantageous because it allows us to specifically couple learnings to individual strains via the donor genome barcode (discussed in Process Development), prioritize library coverage and evenness, and maximize learnings by focusing on transfers that are more likely to work. Conclusions from the Full Scale Tesseract will inform potential expansions beyond solely testing individual and bacterial genomic DNA. Two clear paths forward are expanding to

1. Microbial metagenomic communities
2. Incorporation of cDNA from eukaryotic or microbial sources.

Metagenomic communities represent a powerful source of genetic diversity, particularly from unculturable microbes. However, they are not the first choice for inclusion because achieving sufficient genome coverage to sample each gene many times within an unculturable microbe's genome may be difficult. Current Tesseract data analysis depends on sampling every gene in each donor genome multiple times to generate confident measurements of gene fitness contributions through comparison among overlapping fragments. Moving to a metagenomic context would likely involve changing the analysis pipeline to focus on the homolog-level rather than the exact identity. Potential mitigation strategies include bottlenecking followed by phi29-XT Whole-genome amplification on a subset of extracted metagenomic DNA. Despite these challenges, various workarounds exist to enable effective sampling of DNA from metag-

enomes remains a promising area for future dataset expansion.

Using cDNA would efficiently extract single reading frames from eukaryotic and archaeal organisms, allowing extension of the Tesseract across the tree of life. cDNA pulldown is well established in eukaryotes and archaea — where poly-A tails enable direct mRNA capture using poly-T oligos — and with the addition of ribosomal binding sites to the expression constructs would allow that cDNA to drop directly into the existing infrastructure, though issues of low coverage of unexpressed genes may hinder full coverage.

Applying the same cDNA approach to microbial organisms would increase the efficiency of the libraries, decreasing the number of inserts with partial genes. Developing a microbial cDNA pulldown workflow therefore represents both an opportunity and a technical risk. One approach might combine ribosome pulldown, RNA purification, and reverse transcription into a single workflow, integrating existing techniques such as ribosome profiling in *E. coli*<sup>25</sup> with downstream RNA isolation and cDNA synthesis methods that have not yet been optimized to function in series. This integration will require careful validation, as microbial cDNA libraries may be biased toward highly abundant transcripts, such as ribosomal proteins, potentially limiting proteomic diversity. If successful, however, this approach would yield high-quality, expression-weighted donor libraries focused on actively translated genes, substantially improving the efficiency and informativeness of large-scale phenotyping. By capturing microbial “translatomes” from both model and extremophilic organisms—including diverse environmental communities such as cheese rinds<sup>26</sup> and aquatic microbiomes<sup>27</sup>—cDNA-derived donor libraries could complement genomic DNA libraries and directly link sequence, expression, and phenotype, producing richer, functionally meaningful datasets for AI-driven modeling of gene and protein function.

## Adding new data analysis methodologies

There are several extensions of the Tesseract that could improve data quality or demonstrate predictive accuracy of modeling approaches. Here are a few examples that we envision:

Refined calibration of fitness measurements — currently the measured fitnesses are in arbitrary units relative to a WT control, but we are investigating methods of calibrating them to absolute units that can be compared across experiments.

Validated protein variant effect prediction — if we have conviction in our ability to model the fitness of diverse protein variants, then we could potentially synthesize novel sequences and predict their function in an ensuing selection experiment.

Validated gene transfer prediction — once we have trained a model to predict gene transfer success between donor genes and host strains, we can conduct follow-up experiments in unmeasured host strains to assess prediction accuracy.

Validated functional annotation — once we have trained a model to predict protein functional class on our data, we can predict the annotations of unobserved genes belonging to unculturable microbes. Follow-up experiments would then consist of creating libraries from those organisms in relevant stressor conditions to see if they emerge as high-fitness genes as predicted.

# Community outreach/library sharing

## Physical asset sharing

The barcoded libraries produced to obtain Tesseract data are themselves a valuable physical asset that can be repurposed for other scientific uses throughout the research community. There are additional opportunities for community outreach with the Tesseract. Of immediate impact is data sharing of the whole data corpus on The Align Foundation platform. In addition, the libraries themselves are straightforward, shareable reagents.

## Modeling hackathons and benchmarking competitions

We plan to host a hackathon once the Pilot dataset has been acquired to engage with the broader computational community. The hackathon will be structured to encourage researchers to develop novel ways of investigating the data, stimulating discourse in the broader community on potential applications.

# Supplemental Material

## 1. Existing datasets table

Name of database	Category	Paper Title	Year of publication	Donor organisms	Host organisms	Conditions	Data source
Tesseract	overexpression	Proposed dataset	2025	1,000	50	100	Experimentally generated
Dub-seq <sup>2</sup>	overexpression	Dual-barcoded shotgun expression library sequencing for high-throughput characterization of functional traits in bacteria	2019	1	1	52	Experimentally generated
Boba-seq <sup>3</sup>	overexpression	Barcoded overexpression screens in gut Bacteroidales identify genes with roles in carbon utilization and stress resistance	2024	7	1	122	Experimentally generated
Coaux-Seq <sup>4</sup>	overexpression	High-throughput protein characterization by complementation using DNA barcoded fragment libraries	2024	11	1	20	Experimentally generated
TFUMseq <sup>28</sup>	overexpression	Improving microbial fitness in the mammalian gut by in vivo temporal functional metagenomics	2015	1	1	2	Experimentally generated
RB-TnSeq 2018 <sup>1</sup>	loss of function	Mutant phenotypes for thousands of bacterial genes of unknown function	2018	-	32	26-129	Experimentally generated
RB-TnSeq 2015 <sup>8</sup>	loss of function	Rapid Quantification of Mutant Fitness in Diverse Bacteria by Sequencing Randomly Bar-Coded Transposons	2015	-	5	163	Experimentally generated
RB-TnSeq 2024 <sup>29</sup>	loss of function	Machine learning analysis of RB-TnSeq fitness data predicts functional gene modules in <i>Pseudomonas putida</i> KT2440	2024	-	1	179	Used existing data
Tn-seq <sup>30</sup>	loss of function	Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms	2009	-	1	1	Experimentally generated
CRISPRi-seq <sup>9</sup>	loss of function	CRISPRi-seq for genome-wide fitness quantification in bacteria	2022	-	1	2	Experimentally generated
Nichols et al. <sup>31</sup>	loss of function	Phenotypic Landscape of a Bacterial Cell	2011	-	1	324	Experimentally generated

# Suggested Reading



**Barcoded overexpression screens in gut Bacteroidales identify genes with roles in carbon utilization and stress resistance**

[August 2024](#)

This proposed functional genomics technique pulls elements from BOBA-Seq<sup>3</sup>, a recent paper introducing functional genomics.



**Mutant phenotypes for thousands of bacterial genes of unknown function.**

[May 2018](#)

Data from earlier genome-wide deletion screens have been used to classify genes of unknown function<sup>1</sup>.



**Laying the groundwork for data-driven evolution**

[May 2025](#)

Pioneer Labs has collected preliminary data in a simple setting with 1 donor genome, 1 host strain, and 1 selection condition<sup>32</sup>.

# Citations

- Price, M. N. *et al.* Mutant phenotypes for thousands of bacterial genes of unknown function. *Nature* **557**, 503–509 (2018).
- Mutalik, V. K. *et al.* Dual-barcoded shotgun expression library sequencing for high-throughput characterization of functional traits in bacteria. *Nat. Commun.* **10**, 308 (2019).
- Huang, Y. Y. *et al.* Barcoded overexpression screens in gut Bacteroidales identify genes with roles in carbon utilization and stress resistance. *Nat. Commun.* **15**, 6618 (2024).
- Biggs, B. W. *et al.* High-throughput protein characterization by complementation using DNA barcoded fragment libraries. *Mol. Syst. Biol.* **20**, 1207–1229 (2024).
- Sandberg, T. E., Szubin, R., Phaneuf, P. V. & Palsson, B. O. Synthetic cross-phyla gene replacement and evolutionary assimilation of major enzymes. *Nat. Ecol. Evol.* **4**, 1402–1409 (2020).
- Jensen, P. A. Ten species comprise half of the bacteriology literature, leaving most species unstudied. *Microbiology* (2025).
- Vince, O. *et al.* Breaking through biology’s data wall: Expanding the known tree of life by over 10x using a global biodiversity pipeline. *bioRxiv* (2025) doi:10.1101/2025.06.11.658620.
- Wetmore, K. M. *et al.* Rapid quantification of mutant fitness in diverse bacteria by sequencing randomly bar-coded transposons. *MBio* **6**, (2015).
- de Bakker, V., Liu, X., Bravo, A. M. & Veening, J.-W. CRIS-PRi-seq for genome-wide fitness quantification in bacteria. *Nat. Protoc.* **17**, 252–281 (2022).
- Federici, F. *et al.* CIFR (Clone-Integrate-Flip-out-Repeat): A toolset for iterative genome and pathway engineering of Gram-negative bacteria. *Metab. Eng.* **88**, 180–195 (2025).
- Pioneer Labs Reports. The first turn of our engineering crank. Preprint at <https://doi.org/10.6084/M9.FIGSHARE.29042423.V1> (2025).
- Pioneer Labs Reports. *Laying the Groundwork for Data-Driven Evolution*. (figshare, 2025). doi:10.6084/M9.FIGSHARE.28970489.V1.
- Dama, A. C. *et al.* BacterAI maps microbial metabolism without prior knowledge. *Nat. Microbiol.* **8**, 1018–1025 (2023).
- Jensen, P. A. *et al.* Genotype to phenotype: Design of an extensible experimental platform for characterizing microbes. Preprint at <https://doi.org/10.5281/ZENODO.15397940> (2025).

15. Diaz, D. J., Kulikova, A. V., Ellington, A. D. & Wilke, C. O. Using machine learning to predict the effects and consequences of mutations in proteins. *Curr. Opin. Struct. Biol.* **78**, 102518 (2023).
16. Notin, P. *et al.* ProteinGym: Large-scale benchmarks for protein fitness prediction and design. *Neural Inf Process Syst* **36**, 64331–64379 (2023).
17. Olivares-Gil, A. *et al.* Semi-supervised prediction of protein fitness for data-driven protein engineering. *J. Cheminform.* **17**, 88 (2025).
18. Simon, E., Swanson, K. & Zou, J. Language models for biological research: a primer. *Nat. Methods* **21**, 1422–1429 (2024).
19. Baranowski, C. *et al.* Can protein expression be ‘solved’? Preprint at <https://doi.org/10.5281/ZENODO.14014258> (2024).
20. Gligorijević, V. *et al.* Structure-based protein function prediction using graph convolutional networks. *Nat. Commun.* **12**, 3168 (2021).
21. Boadu, F., Lee, A. & Cheng, J. Deep learning methods for protein function prediction. *Proteomics* **25**, e2300471 (2025).
22. Kulmanov, M. & Hoehndorf, R. DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics* **36**, 422–429 (2020).
23. Kulmanov, M., Khan, M. A. & Hoehndorf, R. DeepGO: Predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *arXiv [q-bio.GN]* (2017).
24. Gilbert, C. *et al.* A scalable framework for high-throughput identification of functional origins of replication in non-model bacteria. *bioRxiv* (2023) doi:10.1101/2023.05.19.541510.
25. Mohammad, F. & Buskirk, A. R. Protocol for ribosome profiling in bacteria. *Bio Protoc.* **9**, (2019).
26. Saak, C. C. *et al.* Longitudinal, multi-platform metagenomics yields a high-quality genomic catalog and guides an in vitro model for cheese communities. *mSystems* **8**, e0070122 (2023).
27. Hinlo, R., Gleeson, D., Lintermans, M. & Furlan, E. Methods to maximise recovery of 9 12. 13. 14. 15. environmental DNA from water samples. *PLoS One* **12**, (2017).
28. Yaung, S. J. *et al.* Improving microbial fitness in the mammalian gut by in vivo temporal functional metagenomics. *Mol. Syst. Biol.* **11**, 788 (2015).
29. Borchert, A. J. *et al.* Machine learning analysis of RB-TnSeq fitness data predicts functional gene modules in *Pseudomonas putida* KT2440. *mSystems* **9**, e0094223 (2024).
30. van Opijnen, T., Bodi, K. L. & Camilli, A. Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nat. Methods* **6**, 767–772 (2009).
31. Nichols, R. J. *et al.* Phenotypic landscape of a bacterial cell. *Cell* **144**, 143–156 (2011).
32. Labs, P. & Liu, J. Laying the groundwork for data-driven evolution. *Essays from Pioneer Labs* <https://pioneerlabs.substack.com/p/laying-the-groundwork-for-data-driven> (2025).