PNNL-34897

Optimizing Cell-based Antimicrobials through Pooled Genomic Libraries

September 2023

Robert G Egbert
Joshua R Elmore
Carlos Gonzalez Rivera
KJ Dorow
Sarah Akers
Andrew Frank
Andrew Wilson
Will Chrisler

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor Battelle Memorial Institute, nor any of their employees, **makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights**. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or Battelle Memorial Institute. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

**Available to DOE and DOE contractors from
the Office of Scientific and Technical
Information,
P.O. Box 62, Oak Ridge, TN 37831-0062**
www.osti.gov
**ph: (865) 576-8401
fox: (865) 576-5728
email: reports@osti.gov**

**Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
or (703) 605-6000
email: info@ntis.gov
Online ordering: http://www.ntis.gov**

PNNL-34897

Optimizing Cell-based Antimicrobials through Pooled Genomic Libraries

September 2023

Robert G Egbert
Joshua R Elmore
Carlos Gonzalez Rivera
KJ Dorow
Sarah Akers
Andrew Frank
Andrew Wilson
Will Chrisler

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99354

Abstract

DNA synthesis and assembly technologies ushered in through synthetic biology have great promise for biomanufacturing, bioremediation, and the development of living therapeutics. Unfortunately, predicting sequence to function relationships, including for biosynthetic pathways expressed in a new host organism, is difficult and often requires many iterative cycles of design, construction, and testing. We are working to develop data-driven approaches to identify the genetic determinants of growth defects and productivity for the expression of a cell-based antimicrobial. We assayed the growth, pigment production, and antimicrobial activity of a collection of over 10,000 genetic mutants of the violacein biosynthetic pathway and sequenced the genetic variation of these mutants. Through this project, we have developed an innovative codebase to automate the determination of pigmentation and antimicrobial clearing diameter for tens of thousands of genetic mutants cultivated on agar dishes. Further, we have written DNA sequence analysis code to demultiplex & provide consensus sequences from high-throughput PacBio long-read circular consensus sequencing (CCS) datasets. From this foundation, we plan to map DNA sequence to function to predict an optimal genetic design to maximize antimicrobial activity while minimizing deleterious growth effects. The workflows and algorithms developed through this project can be broadly applied to other engineered functions in microbes, uncovering sequence to function relationships for complex phenotypes where function impacts fitness.

Acknowledgments

This research was supported by Earth and Biological Sciences Directorate Mission seed funding, under the Laboratory Directed Research and Development (LDRD) Program at Pacific Northwest National Laboratory (PNNL). PNNL is a multi-program national laboratory operated for the U.S. Department of Energy (DOE) by Battelle Memorial Institute under Contract No. DE-AC05-76RL01830.

Contents

Figures

1.0 Introduction

DNA synthesis and assembly technologies ushered in by synthetic biology have great promise to revolutionize microbial engineering for biomanufacturing, bioremediation, and the development of living therapeutics. Unfortunately, predicting sequence to function relationships for complex circuits and pathways expressed in a host organisms is difficult and often requires costly iterative cycles of design, strain construction, and testing. There is no established method to predict which transcription and translation control elements in a target host will optimize fitness and function phenotypes. These challenges are compounded exponentially as the number of genes in a pathway increase. Overexpression of individual genes can lead to the build-up of toxic intermediates of a metabolic pathway, the depletion of a substrate critical to cell growth, such as an amino acid, or the toxic accumulation of a protein that is part of the function. Data-driven machine learning approaches that map genotype to complex phenotypes (e.g. growth or antimicrobial activity) for thousands of genetic variants has promise to revolutionize and accelerate genetic design.

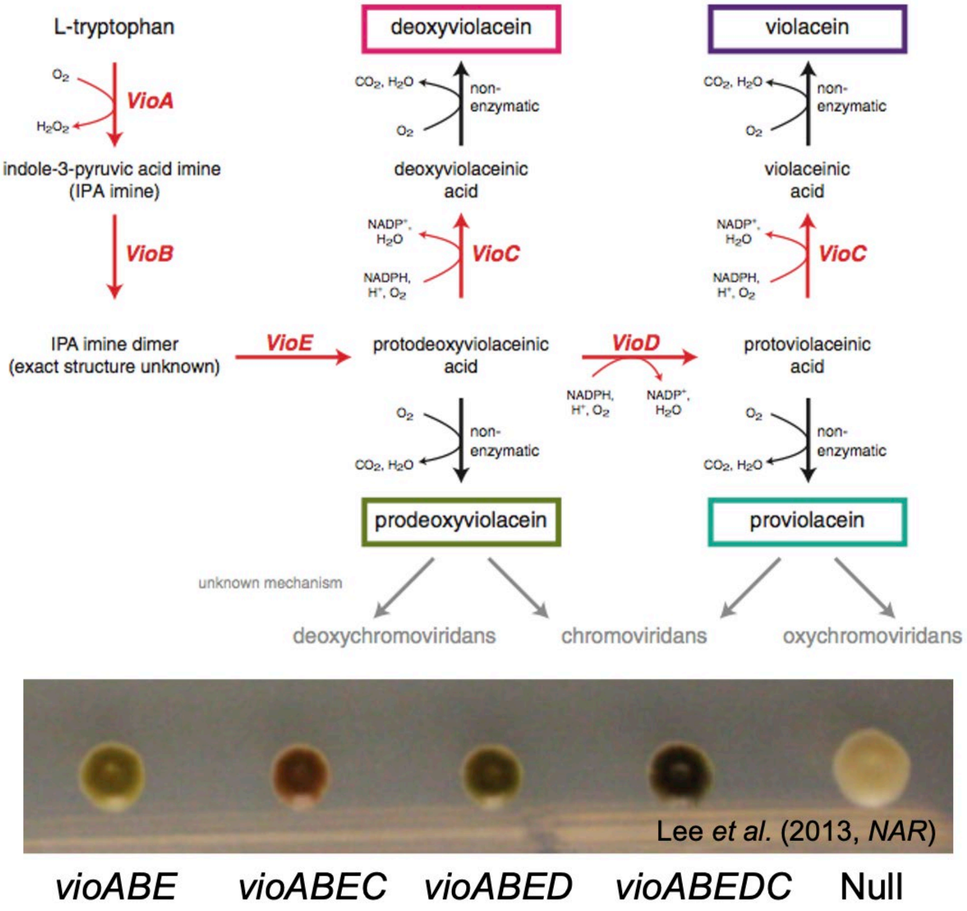


Figure 1. **Violacein biosynthetic pathway.** Using the amino acid tryptophan as a substrate, the five-gene violacein pathway produces multiple secondary metabolite intermediates with distinct pigment profiles. Following the first three genes of the pathway *vioABE*, the balanced expression of *vioD* and *vioC* determine the ratios of the four primary products of the branched pathway: prodeoxyviolacein, deoxyviolacein, proviolacein, and violacein.

2.0 Research Design and Methodology

In this project, we have identified a model genetic system as a testbed to create and optimize machine learning methods for high-throughput evaluation of genetic variants. The target function is a biosynthetic pathway for the pigment violacein (Lee et al. 2015) (**Figure 1**). We have access to a genetic variant library that parameterizes expression of the five biosynthesis genes over multiple orders of magnitude (**Figure 2A-B**). The generated library oversamples by ten-fold the 262,144 potential genetic variants. We have observed significant variation in pigmentation and antimicrobial activity for these genetic variants when the variants are challenged with competitive growth of *Bacillus subtilis*, which is sensitive to compounds from the violacein pathway (**Figure 2C**). The primary objective of the funded project was to generate a high-throughput dataset for genotype and phenotype information on tens of thousands of variants of the engineered pathway (**Figure 2D**). By collecting data for these variants and developing algorithms to extract features from the data (e.g., gene expression variants, growth fitness defect, pigmentation, competitive suppression of *Bacillus* growth), we can link genotype to phenotype in a way to predict the source of & suggest genetic designs that optimally balance gene expression, cell fitness, and engineered function (**Figure 3**).

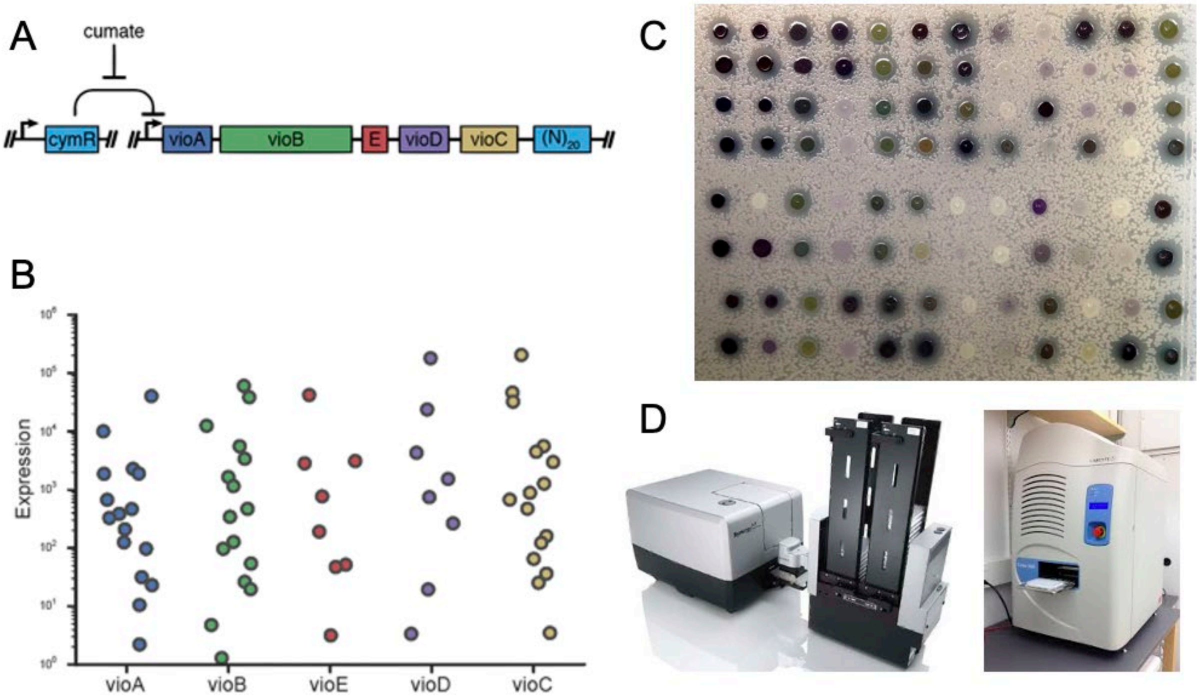
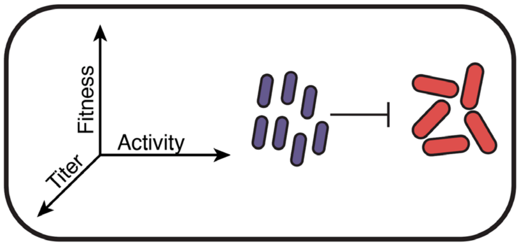


Figure 2. **Design and evaluation of violacein expression variant library.** Workflow to map genotype to complex growth and antimicrobial production phenotypes. (A) Base cumate-inducible violacein biosynthesis pathway for mutant library, integrated on *E. coli* genome. (B) Expected expression levels of translational control variants for each gene of the violacein pathway. (C) Random sampling of pigment production variants (96 spotted samples) in presence of a competitor *Bacillus subtilis* that is sensitive to secondary metabolites of the branched violacein pathway. (D) High-throughput instruments (plate stacker, left; acoustic liquid handler, right) to measure growth and antimicrobial production at a scale of tens of thousands of individual genetic variants.



Competitive fitness against gram-positive

Figure 3. **Multi-factor optimization of cellular function.** The outcome of machine learning algorithms built on our high-throughput Design-Build-Test methodology should inform genetic designs that optimally balance competing cellular resources dedicated to gene expression (Titer), cellular fitness (Fitness), and engineered function (Activity).

3.0 Results and Discussion

We have developed multiple novel experimental workflows and computational algorithms to enable a future machine learning approach to assess gene to function for the violacein biosynthetic pathway and other engineered functions in bacterial systems. The workflow consists of (1) single-cell sorting by fluorescence-activated cell sorting to generate 30 384-well plates of individual genetic variants, (2) high-throughput growth assays to assess growth defects induced by expression of the variant pathway, (3) high-throughput polymerase chain amplification (PCR) of the pathway variant from the genome as well as a pooled nucleic acid barcoding approach to combine and demultiplex all samples for PacBio Circular Consensus Sequencing (CCS), (4) leveraging an acoustic printer to spot pathway variants and a competitor species on agar plates to assess pigmentation and antimicrobial phenotypes, (5) developing image processing workflows to automate the assessment of colony size, pigmentation profile, and clearing size for arrayed variant plates, and (6) establishment of a DNA sequence analysis pipeline to map demultiplexed high-throughput sequencing reads to growth, pigmentation, and antimicrobial clearing phenotypes. Once integrated, these workflows and algorithms will provide a comprehensive platform to predict sequence-function relationships in bacteria to enable rapid engineering and optimization of engineered functions.

3.1 Single-cell sorting

The genetic variant library consists of over two million genetically barcoded variants of the violacein biosynthetic pathway. To assess the genetic and phenotypic variation, the library variants must be arrayed into single wells of bioassay plates. To accomplish the variant arraying, we utilized the EMSL cell sorting capability to generate thirty 384-well plates of individual genetic variants. The sorted cells were cultivated in rich LB growth media with glycerol as a cryoprotectant and grown to saturation for cold storage at -80 °C.

3.2 High-throughput growth assays

We employed a 20-plate plate stacker and a BioTek plate reader (**Figure 2D**) to assay the growth of variants from all plates in synthetic defined growth media M9 supplemented with casamino acids and glucose. We grew all variants in a control condition (no inducer) and an induced condition. The induced condition included cumate at 100 uM to fully express the variant violacein pathway harbored by the strain. All growth data has been collected and is ready to be analyzed once sequencing analysis and well-mapping is completed.

3.3 High-throughput PCR & PacBio sequencing

Using the growth assay plates as a source, we used PCR amplification of genomic DNA from each variant to create enough DNA for high-throughput sequencing. To pool samples for PacBio CCS sequencing, we introduced unique DNA barcodes for each variant well using the Labcyte acoustic liquid handler. Once all samples were amplified (**Figure 4**), we pooled the samples into a single tube, normalized the concentration of the DNA, and sent the samples to the University of Washington core facility for PacBio CCS sequencing. The sequencing data and all associated well-to-DNA barcode metadata is stored in a PNNL shared folder for sequence variant and mutational analysis.

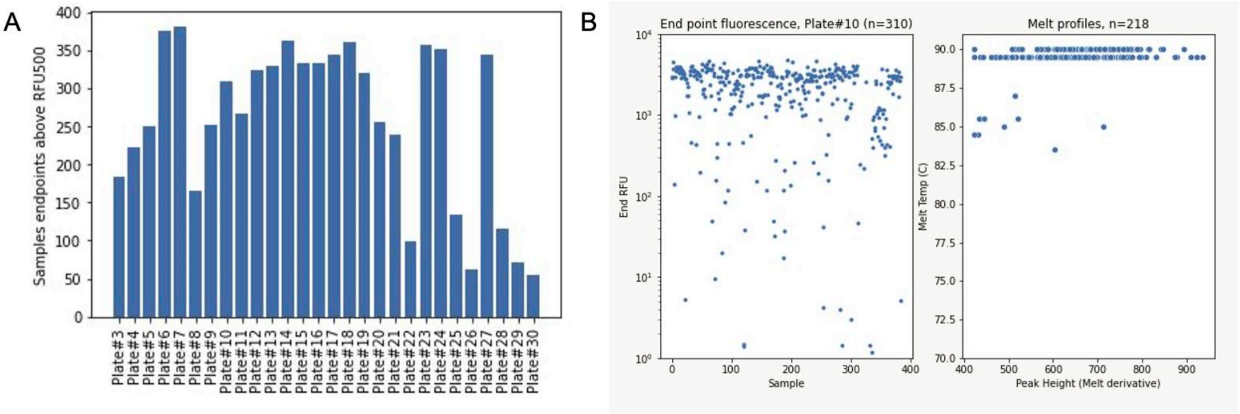


Figure 4. **Pathway amplification quality control.** Using a BioRad CFX384, variant pathways were amplified by PCR. (A) Number of samples per plate that exceeded expected fluorescence levels consistent with yields suitable for high-throughput sequencing. (B) Melt profiles for plate 10, with melts values above 88 °C indicating successful amplification of the full pathway.

3.4 High-throughput agar spot & clearing assays

To assess the pigmentation and antimicrobial clearing phenotypes, we developed a novel assay for the Labcyte Echo acoustic liquid handler by spotting a dense, custom array of microbial inoculations on agar plates. This required significant development time to identify the proper concentration of cells to populate the plates, but to prevent overgrowth. We spotted the pathway variants alone on plates in a 384-well array with 100 uM cumate to assess pigmentation and visual growth defects and in competition with a dense array of *Bacillus subtilis* spots to assess the variant efficacy at growth suppression. All plates were successfully generated and imaged with a Canon DSLR camera post-cultivation (**Figure 5**). Files are all stored on a shared PNNL drive.

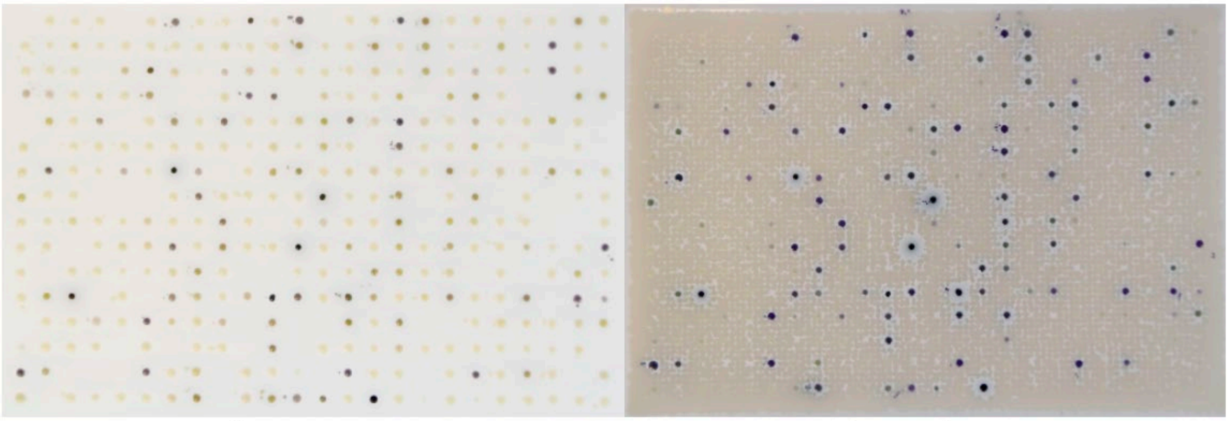


Plate 10, spotting

Plate 10, clearing

Figure 5. **High-throughput spotting and clearing assays.** Employing the Labcyte Echo acoustic liquid handler, violacein pathway variants were spotted in a 16x24 grid alone or in competition with *Bacillus subtilis* colonies arrayed in a 16,000-spot grid. Spotting assays enable assessment of pigmentation and growth defects associated with expression of the pathway. Clearing assays provide a functional output associated with the antimicrobial phenotype.

3.5 Image processing pipeline

Using the images generated from the spotting and clearing assays, we have employed image processing machine learning algorithms to assess colony growth, pigmentation, and antimicrobial activity. First, we cropped all spotting and clearing images. Next, wehave employed a recently development image segmentation algorithm (Kirillov et al. 2023) to enforce a regular grid structure to the spotting and clearing plates which can be complicated by wells with no cell growth or no pigmentation or clearing phenotypes (**Figure 6**). The code is in a GitHub repository (https://github.com/PerConSFA/poolGLASS_clearings) and we are seeking SULI interns with experience and interest in machine learning and image processing to finalize the development of the code that will ultimately provide a dataframe with plate and well position, spotting colony size, pigmentation data in RGB, and clearing size in millimeters.

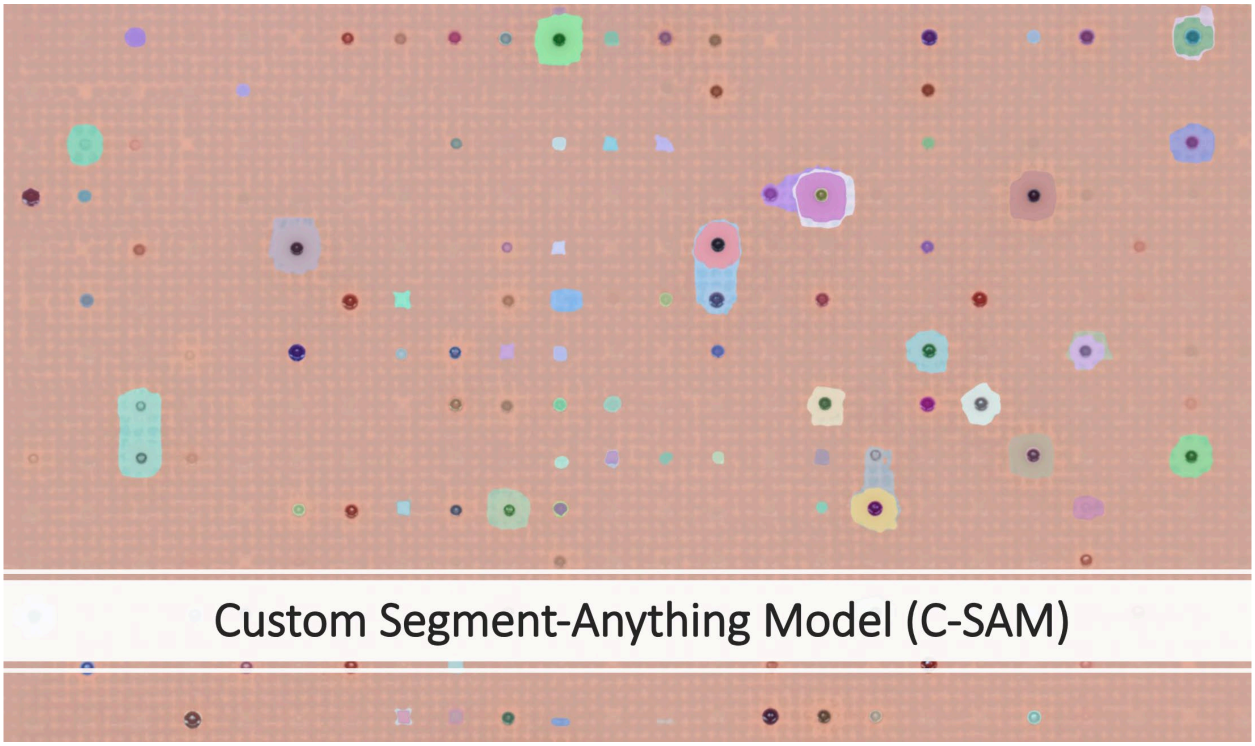


Figure 6. **Automated segmentation of colonies and clearings.** The Segment Anything algorithm automates the process of identifying individual colonies on clearing (shown) and spotting plates. Additional training on our dataset will be necessary to fully parameterize a scalable model to extract all of the expected features.

3.6 Sequence analysis pipeline

To demultiplex PacBio CCS reads and build consensus sequences for each well, we developed a custom Snakemake pipeline (https://snakemake.readthedocs.io/en/stable/). The pipeline identifies variant-unique barcodes, maps the CCS sequence reads to the reference sequence, clusters individual CCS reads to common barcodes, and builds a consensus sequence (https://github.com/PerConSFA/poolGLASS). We found a fraction of wells contained multiple pathway barcodes, suggesting the cell sorting did not sort single cells for all wells or some other form of contamination occurred. While this pipeline has been generated, we were not able to complete the generation of a dataframe that links plate and well position to DNA barcode and sequence variant for each pathway gene. We will complete the sequence analysis by engaging SULI interns with experience and research interests in computational biology.

Discussion

Through significant efforts & adaptations from our original plans, we have collected all necessary data to map sequence to function for the violacein antimicrobial activity. While some analysis pipelines are complete, we will need to recruit through the SULI program to complete generation of the dataframes that will be used for machine learning. We anticipate the collective work will result in a high-interest paper in a reputable synthetic biology journal. This work will increase PNNL’s visibility as a national leader in synthetic biology research. Further, the workflows and algorithms developed through this project can be broadly applied to other engineered functions in microbes, mapping DNA sequence to function for large collections of genetic mutants, ultimately leading to more predictive frameworks for economical and efficient design of genetic circuits and pathways.

4.0 References

Kirillov, Alexander, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, and Wan-Yen Lo. 2023. "Segment anything." *arXiv preprint arXiv:2304.02643*.

Lee, Michael E., William C. DeLoache, Bernardo Cervantes, and John E. Dueber. 2015. "A Highly Characterized Yeast Toolkit for Modular, Multipart Assembly." *ACS Synthetic Biology* 4 (9): 975-986. https://doi.org/10.1021/sb500366v.

Pacific Northwest National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354

1-888-375-PNNL (7665)

www.pnnl.gov