EMSL

EMSL Science Drivers and Scope Workshop for:

Microbial Molecular Phenotyping Capability (M2PC)

December 2021

Scott E. Baker
Douglas Mans

PNNL-32461

U.S. DEPARTMENT OF ENERGY

DISCLAIMER

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;
ph: (865) 576-8401
fax: (865) 576-5728
email: reports@adonis.osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
email: orders@ntis.gov <https://www.ntis.gov/about>
Online ordering: http://www.ntis.gov

EMSL Science Drivers and Scope Workshop for:

Microbial Molecular Phenotyping Capability (M2PC)

December 2021

Scott E. Baker
Douglas Mans

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

Pacific Northwest National Laboratory
Richland, Washington 99354

Summary

The Environmental Molecular Sciences Laboratory (EMSL) workshop on a Microbial Molecular Phenotyping Capability was held virtually by Pacific Northwest National Laboratory (PNNL) Sep 21-22, 2021. Approximately 60 registered attendees from academia, industry, and national laboratories convened to discuss the science drivers, potential design, priorities, and vision for a microbial molecular phenotyping capability.

While researchers have made significant advances in understanding microbial genomics, phenotyping technologies to translate these discoveries to understand and harness biological functions continue to lag both in pace and scale. This disparity hinders progress toward goals such as the development of improved biofuels and bioproducts, and understanding, modeling, and simulating the transformation of materials and nutrients in the environment that drive hot spot and hot moment events. The Microbial Molecular Phenotyping Capability, envisioned as a highly automated, high throughput suite of modular workflows, aims to reduce this discrepancy, and bring functional assays on par with the current scale of next-generation genomic sequencing technologies.

At the workshop, speakers made brief presentations on each of four topical themes necessary for such a capability, namely: (i) Experimental systems (ii) Functional assays (iii) Phenotyping and (iv) Data coordination. Each presentation was followed by breakout sessions where attendees split into four groups and discussed the priorities, challenges, and potential impact a phenotyping capability might have in that area. Groups used Mural, an electronic whiteboard system, to capture the breadth of ideas brought up in each breakout room and reported salient points from their discussions at the end of each breakout session.

The discussions underscored the immense value of a high throughput phenotyping capability to advance discovery relevant to BER's mission. Attendees pointed out several areas where existing strengths and resources at EMSL could be leveraged into such a capability. A phenotyping capability as envisioned would require advanced synthetic biology and organic chemistry teams to design and manufacture probes for various workflows. Thus, it would serve to advance more than biological research, alone. Several considerations for the design of automated workflows and pipelines for phenotypic assays were laid out over the course of the discussions. Workshop attendees emphasized the prioritization of data coordination and standardization from the earliest stages, and the curation of metadata, metabolic models, probes, and other phenotyping technologies in a collective knowledgebase. These efforts to develop collaborative, collective resources would minimize redundancies in scientific efforts and improve reproducibility.

Ultimately, attendees saw a microbial molecular phenotyping capability as having the potential to do more than just create a new resource. By generating high-quality functional annotations and other data to update public databases, this new capability would reach researchers who may never be users of this resource but would still greatly benefit from its output. There is great potential for this to be an especially valuable resource in support of research at historically black colleges or universities and minority-serving institutions, thus advancing efforts for diversity, equity, and inclusion.

Acronyms and Abbreviations

API	Application programming interface
BER	Biological and Environmental Research
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
DOE	Department of Energy
EMSL	Environmental Molecular Sciences Laboratory
ENVO	Environmental ontology
ESS-Dive	Environmental Systems Science Data Infrastructure for a Virtual Ecosystem
FAIR	Findable, accessible, interoperable, and reusable
GFP	Green fluorescent protein
IGSN	International Geo Sample Number
JGI	Joint Genome Institute
KBase	Department of Energy Systems Biology Knowledgebase (KBase)
M2PC	Microbial Molecular Phenotyping Capability
MIxS	Genome Standards Consortium
MS	Mass Spectrometry
NMDC	National Microbiome Data Collective
OBO	Open biological and ontologies foundry
OSTI	Office of Scientific and Technical Information
PNNL	Pacific Northwest National Laboratory
UUID	Universally unique identifiers

Contents

1.0 Introduction

Microbial communities that sustain ecosystems are the single largest body of biological activity on Earth in terms of mass. They're critically important to the plants, crop systems, and other macroscopic organisms fundamental to sustainable ecosystems and bioenergy production. They are especially valuable to BER's mission because of their potential to serve as platforms for biofuel production, contaminant remediation, as cellular factories creating oils such as lipids or fatty acids, or as commensals that promote bioenergy resilience and production from crops. In addition, microbial/microbiome functions, such as the creation of redox gradients, biofilm production, and aggregate formation, have direct effects on the pore to core scale flow and transport properties. Understanding these functions at the microbiome, single microbe, and protein levels are critical for increasing the accuracy and fidelity of fine-grain hyperlocal and hierarchical fate and transport models that can be merged with larger climate, land-atmosphere, and Earth system models.

Several attributes of microbes, including their short lifespans, easily manipulated genomes, and small sizes. make them particularly attractive for scalable manufacturing technologies. Researchers can directly measure increases in biomass and energy production of crop plants in the presence or absence of commensal microbes, or as they genetically manipulate microbial systems to increase desirable metabolic processes. These efforts, however, are crippled by a long-standing—and continually growing—discrepancy: that between researchers' understanding of the microbial genome, and their ability to harness genomic data accurately, reliably to predict and efficiently engineer useful traits into microbial systems.

Since its discovery, an understanding of DNA has been considered essential to studying biological functions. Thus, more than 150 years of biological research has focused on the genome: from revealing the structure and coding sequence of DNA, to the human genome project, to now CRISPR and various gene editing technologies. But the genome is not enough.

Biology is ultimately the study of function, manifest in the phenotypes of an organism. High throughput workflows and tools to understand phenotypes lag behind the accelerating pace of genomics. Consider the numbers: Scientists have estimated the existence of approximately 2.2 to 4.5 million species of microbes, sequenced nearly 200,000 complete bacterial genomes, and identified 154 million unique proteins within these sequences. But 121 million—almost 70% of those proteins—have had their functions assigned by their sequence similarity to other known proteins rather than actual experimental analyses.

Homology-based predictions of function result in incorrect or missing annotations. These errors are often propagated across databases and severely complicate the proper assignment of protein function. As a result, the accuracy of annotation in these genomes is estimated to be as low as 14%. In addition to this existing backlog of potentially misidentified proteins, approximately 3.2 million new protein sequences are added to global databases every month.

Genomics-based methods have proven insufficient even for *Escherichia coli*, the mainstay of microbial research for more than two decades. Only 58% of its genome has assigned, properly validated functions; 37% remains functionally un-annotated. These yawning gaps in the *E. coli* genome and among microbes, in general, prevent researchers from engineering metabolic pathways because data on these pathways are incomplete and riddled with proteins with no known function. Deficits in the capacity and pace of phenotyping studies prevent researchers from being able to close these gaps. Biologists typically use a design-build-test-learn framework—designing experiments to question specific hypotheses, building systems, and then testing them to learn and inform the next iteration of experimentation. But lack of accurate data on protein functions and phenotypes restricts the utility of this framework for microbial engineering efforts essential to BER's mission.

To address this gap, EMSL has proposed the creation of a highly automated, high throughput microbial molecular phenotyping capability (M2PC) for the BER user community. The capability is envisioned as a suite of modular, highly automated workflows across microbial cultivation, functional assays, and deep molecular phenotyping. This modular approach will enable the pairing of standardized molecular assays with high throughput analytical approaches, as well as next-generation variable experimental cultivation systems that can be remotely controlled by artificial intelligence-based applications to discover and rapidly characterize wholly new functions within a given microbe.

Users will be able to discover and assign new microbial functions at a scale and pace matching that of next-generation gene sequencing technologies. They would also be able to interrogate functions across species, taxa, and microbial communities to understand functions at the various levels needed to model nutrient cycles and environmental processes in ecosystems.

Such a capability would provide critical functional knowledge in microbial sciences, impact production of biomaterials and biofuels, and enhance bioenergy crop productivity.

Discussions over the next two days will help inform the design concept for this capability, identify scientific challenges, high-priority needs, focus areas for research, opportunities for collaboration, and how to best leverage EMSL technology and expertise to bear on bringing phenotyping up to par with genomics.

2.0 Workshop attendees

Workshop steering committee:

- Scott Baker (Chair), Functional and Systems Biology Science Area Lead, EMSL
- Kristin Burnum-Johnson (Co-Chair), Biomedical Scientist
- Michelle O'Malley, Professor, University of California, Santa Barbara
- Aindrila Mukhopadhyay, Staff Scientist, Lawrence Berkeley National Laboratory
- Tom Rush, Chief R&D Officer, Variant Bio
- Elizabeth Shank, Associate Professor, University of Massachusetts Chan Medical School
- Ranjan Srivastava, Professor, University of Connecticut
- Huimin Zhao, Professor, University of Illinois at Urbana-Champaign

DOE program managers, researchers from government, academia, and industry:

- Adam Abate, Professor, University of California, San Francisco
- Tyler Backman, Research Scientist, Lawrence Berkeley National Laboratory
- Constance Bailey, Assistant Professor, University of Tennessee, Knoxville
- Emily Balskus, Professor, Harvard Medical School
- Paul Bayer, Program Manager, DOE
- Chris Beecroft, Data Management, Joint Genome Institute
- Crysten Blaby, Biologist, Brookhaven National Laboratory
- Romy Chakraborty, Staff Scientist, Lawrence Berkeley National Laboratory
- Estelle Couradeau, Assistant Professor, Pennsylvania State University
- Maude David, Assistant Professor, Oregon State University
- Kevin Dean, Assistant Professor, University of Texas Southwestern Medical Center
- Brent Dorr, Team Leader, GlaxoSmithKline
- Carrie Eckert, Group Leader, Oak Ridge National Laboratory
- Emiley Eloe-Fadrosh, Program Head, Joint Genome Institute
- Michael Fero, President and CEO, TeselaGen Biotechnology
- Erick Flieger, Lead Engineer, DOE Office of Science
- Ron Gallagher, Federal Project Director, DOE
- Igor Grigoriev, Senior Staff Scientist, Joint Genome Institute
- Kimberly Harper, Pacific Northwest Site Officer, DOE
- Michael Koepke, Vice President, LanzaTech
- Resham Kulkarni, Program Manager, DOE
- Adi Lavy, Scientist, Riffyn
- Ramana Madupu, Program Manager, Joint Genome Institute
- Costas Maranas, Professor, Pennsylvania State University
- Katherine McMahon, Professor, University of Wisconsin-Madison
- Trent Northern, Deputy Directory, Lawrence Berkeley National Laboratory
- Kent Peters, Program Manager, DOE
- Mircea Podar, Distinguished Staff Scientist, Oak Ridge National Laboratory
- Pablo Rabinowicz, Program Manager, DOE
- David Ross, Physicist, National Institute of Standards and Technology
- Davinia Salvachua Rodriguez, Staff Scientist, National Renewable Energy Laboratory
- James Sethian, Professor, University of California, Berkeley
- Prem Srivastava, Program Manager, DOE
- Ryan Tappel, Manager, LanzaTech

- Michael Washburn, Professor, University of Kansas Medical Center
- Boris Wawrik, Program Manager, DOE
- Janet Westpheling, Professor, University of Georgia
- Tanja Woyke, Staff Scientist, Joint Genome Institute
- Yasuo Yoshikuni, Staff Scientist, Joint Genome Institute
- Petrus Zwart, Staff Scientist, Lawrence Berkeley National Laboratory

Staff at PNNL:

- Douglas Mans, EMSL Director
- Angus Bampton, Project Manager
- Jay Bardhan, Computer Scientist
- Alexander Beliaev, Biologist
- Robert Egbert, Engineer
- Scott Lea, Chemist
- Mary Lipton, Chemist
- Lee Ann McCue, EMSL Chief Data and Analytics Officer
- Lili Pasa-Tolic, Lab Fellow
- Gert Patello, EMSL Chief Operations Officer
- Sam Paulson, Project Manager
- Nikki Powell, EMSL Project Manager
- Weijun Qian, Lab Fellow
- Theva Thevuthasan, Project Manager
- Steven Wiley, Lab Fellow
- Aaron Wright, Chemist

3.0 Presentations

3.1 Presentation: Experimental systems: *Scott Baker*

Microbial activities in the environment, such as the breakdown of woody plant materials, are both ubiquitous and fundamental—yet they remain mysterious at the molecular level. That's in part because of their complexity and partly because of the disparity between genomics and phenotypic characterization. These hurdles prevent researchers from applying existing genomic data to further understand, engineer, and harness these environmental processes for sustainable biotechnology strategies.

Microbes associated with plants, soils, and other niches are a promising platform to (i) bridge this gap and forge the missing links between genomic information and the behaviors of complex systems and (ii) advance biological understanding of complex environmental processes. A Microbial Molecular Phenotyping Capability (M2PC) will encompass the experimental systems to achieve these twofold objectives. While many environmental systems are complex, a high-throughput M2PC capability can help tame these complex systems and develop tractable ones that can be harnessed and engineered for various purposes.

The goal with phenotypic measurements is to accomplish them at a pace that exceeds current genome-based data generation to catch up with the existing glut of poorly annotated sequences, as well as usefully interpret data from ongoing large-scale genomics efforts.

Thus far, biologists have relied on descriptive, observational approaches to understand microbial complexity. With the ability to generate high throughput, quantitative data on phenotypes and microbial functions, biological research in the BER space could progress to predictive studies. Much like computational chemists or physicists, biologists would be able to apply data in an informed fashion to predictive, hypothesis-driven studies. For example, once the phenotypic functions of a lignin-digesting microbial community are identified, they may be replicated by a defined synthetic community with limited members that can be remixed in different stoichiometries, depending on users' needs.

Studying the molecular machinery involved in environmental processes and the microbial communities that deploy this machinery will advance discovery science and hypothesis-driven studies and lead to the sustainable production of fuels, chemicals, and materials—a goal that is central to BER's mission.

3.1.1 Discussion

Questions to launch the discussion:

- What types of microbes and microbial consortia should be studied?
- What throughput is necessary?
- What growth conditions or variables should be included?
- What questions are being addressed during growth?

Group discussions in this session focused on experimental priorities and approaches to (i) close the throughput gap between genomics and tools for phenotyping and functional screening and (ii) the need to translate phenotypes studied in laboratory experiments to real-world environmental relevance, since removing an organism from its natural environment often changes its phenome.

Across the four groups, discussions highlighted the need to grow all organisms that can reasonably be cultured, including fungi, archaea, anaerobes, and others. Since growth itself is a phenotype, the goal of

microbial culture in a phenotyping capability is not merely to grow organisms for downstream analysis. Culture is crucial to, and part of, the experimental perturbation process.

Flexibility in culture conditions was deemed a critical aspect; molecular phenotyping requires the ability to culture individual organisms, but also co-culture consortia and communities and engineered synthetic communities. Growth conditions should also encompass a variety of different conditions so experiments can span both generalist and specialist microbes to understand their niche requirements and adaptability. The discussions envisioned a high throughput system that could study a matrix of growth conditions and microbes—either one culture under many different conditions or many microbial cultures under the same conditions to identify those that best adapted.

Cultivation systems should also enable the study of organisms that grow in minimally supportive conditions—literally "where nothing should grow"—so researchers can identify the phenotypes that enable such colonization.

To bring these concepts up to scale, discussions centered on the necessary upgrades to throughput while recognizing that throughput is often limited by the diversity of systems being studied. For instance, a system that's high throughput for axenic *E. coli* cultures might not be high throughput for other organisms or microbial communities. To account for both, growth facilities would need to be highly scalable, most likely beginning with varied, less-defined conditions, and scaling up to several fine-tuned, highly specific conditions as researchers start to engineer traits and enter the "test" phase of the design-build-test cycle. Nonetheless, not all microbes may be amenable to such high throughput platforms so a phenotyping capability should also support the cultivation of niche specialists that require boutique growth facilities.

The ideal capability would encompass a growth facility that could literally do thousands of different growth conditions with lots of different mutated forms of the organism, and then monitor not only the growth of the organism, but also specific tagged gene products and related phenotypes.

Spatial systems biology approaches, such as spatially defined metabolomics, can help researchers study gene expression activity and protein functions in specific growth conditions. Experimental cultivation within an M2PC will need to capture metabolism and other phenotypes during growth in a high throughput manner, for example by using specialized probes for a flow cytometry-based approach. The ability to capture patterns of growth and changes in metabolites or organisms' behavior should be built into cultivation as a real-time process to capture the phenotypes pertinent to growth.

While many studies so far have focused on so-called "model and chassis" organisms, the group discussions highlighted the need for more—and more diverse—model organisms that represent a range of metabolic metrics beyond just the laboratory workhorse *E .coli*. They highlighted the need for more platforms that are as well understood as *E .coli* currently is, so as to understand complex metabolic functions that cannot be replicated or captured in currently available models. At the same time, they emphasized the need to improve annotation and functional studies in *E. coli* and existing model and chassis systems.

To capture metabolic functions during growth, researchers recommended a multi-assay strategy. This can be done by combining approaches such as (i) gene knockouts, (ii) CRISPR-based screens to identify functions critical to growth under specific conditions and (iii) reductive microbiome approaches to understand the units of a community, their interactions, and which individual functions are critical to maintaining community structure. Other strategies, such as adapting computational and machine learning-based tools to predict protein structure and function, were also discussed.

Capturing and curating real-world environmental conditions is also essential for experimental cultivation in an M2PC. A database of growth conditions and real-world distribution, as well as environment, could prove a valuable resource to the research community by offering context and a range of conditions to explore in high-throughput initial studies of growth.

Microbial growth could be studied in soil-like, pore-scale textured experimental platforms such as EcoFab, or in unsaturated porous media or synthetic soil communities that replicate real-world conditions. These assays will enhance and support efforts for in situ phenotyping because culturable, reduced-complexity communities cannot fully recapitulate environmental phenotypes. Some efforts, such as GeoChip, are already underway to increase the throughput of in situ phenotyping.

Such information, along with genomic and metagenomic data, could reveal the growth conditions necessary for organisms currently thought to be unculturable. These organisms carry immense potential for bioengineering and industrial applications, but they have long been impractical to study in laboratory contexts. Defining protein functions in the biological space represented by unculturable organisms could accelerate discovery and potential applications of their unique functions.

3.2 Presentation: Functional assays: *Aaron Wright*

Chemical probes enable a wide range of experiments to functionally characterize proteins or microbial communities of interest. As experimental tools, chemical probes can help characterize protein-protein interactions, protein binding to small molecules such as substrates or cofactors and regulatory mechanisms such as post-translational modification. In addition to detecting a protein's presence or binding partners, probes can also elucidate enzyme kinetics or the rate of a protein's activity.

Broadly speaking, modular probes have three components: a reactive group, binding core, and a reporting group. The reactive group enables a probe to link to the target of interest, while the binding core is usually a selectivity factor that helps direct a probe to the correct target, such as a sugar moiety that helps a probe home in on a specific glycosyl hydrolase enzyme. Reporting groups span myriad tools to detect a protein or its activity and may range from fluorescent reporters for cell sorting or imaging, to electron-rich groups that can be visualized with cryo-EM.

They can also be placed using a click chemistry approach. Instead of adding a reporter group, probe designers leave a "handle" where different reporter groups can be placed once the reactive core has bound to its target. Designed this way, the possibilities for reporter groups are a veritable cottage industry in their own right. Researchers can buy various kinds to click onto probes for mass spectrometry-based identification, FISH, BONCAT, microarray-based approaches such as GeoChip, or other imaging approaches.

This sort of modular design enables scientists to make targeted, metabolite-based probes that can home in on a single protein or the activities of an entire microbial community. Several different decision points crop up during the design, synthesis, and application of probes; expertise in chemical biology and synthetic chemistry are hence crucial to effective probe development.

Newer probe technologies can not only be applied to whole microbial communities and deployed in cell-sorting technologies, they can also be used to sequence regions of interest while keeping the cells alive for downstream cultivation and experiments. This ability expands researchers' repertoire. Instead of focusing on specific enzymes or proteins, researchers can isolate fractions with the function of interest and follow up with detailed genomic sequencing and annotation, phylogenetic analysis, or interactome mapping.

The subsequent group discussion used these possibilities as a launching point for a discussion on how to automate and perform large-scale functional assays in microbial systems relevant to BER's mission. One of the goals they focused on was the critical need to target unannotated proteins and prioritize functions of significance. For example, high-throughput functional assays to study lignocellulose degradation would identify and map proteins that are as yet unannotated, but relevant to that metabolic process.

Currently, many probe-based measurements are one-off experiments. Groups discussed how probes could be designed to profile functions of proteins individually or within microbial consortia, how probes might be deployed across myriad biological systems and data analyses scaled and made shareable. Existing platforms and capabilities have yet to address the challenge of scaling these technologies. Doing so could provide immense benefits to the user community.

Attendees envisioned various ways the EMSL user community might build functional assays into M2PC capability: by individual contributors developing and incorporating probes into workflows, via high throughput fluxomics approaches that could be carried out at a central facility, or other possibilities. This input could be used to prioritize functions of interest and workflows for high throughput probe design.

3.2.1 Discussion

Questions to launch the discussion:

- What chemical probes and enzyme kinetic assays are needed?
- What sorts of screening assays for microbes are necessary?

Attendees highlighted the need for a wide range of assays to understand molecular interactions at the scale of proteins down to downstream metabolites, both within cells and microbial communities. In addition to assays for specific protein functions, phenotyping microbes will also require the analytical and imaging facilities to measure myriad macroscopic functions, such as the spatiotemporal distribution of strains within biofilms or in planktonic polymicrobial communities. They emphasized the need to probe proteins not just within cell-free extracts, but in their native physiological and environmental contexts. Discussions highlighted the immense value of assays, such as C-13 labeling to understand metabolic flux, which can elucidate enzyme kinetics, rate-limiting steps, and opportunities and targets for protein engineering.

Improvements in fluorescent reporters were identified as another key component. Currently available fluorescent reporters, such as GFP, perform inefficiently in anaerobes, extremophiles, and other microbes that grow in unusual environmental niches. Better fluorescent probes, as well as potential alternatives to fluorescence-activated cell sorting (FACS)—such as Raman-activated cell sorting—were discussed as potentially valuable assays. New techniques from materials science, such as the ability to resolve the nanoscale distribution of elements, could also inform the development of alternative probes.

A large-scale phenotyping capability offers the potential to accelerate discovery via the development of standardized, high throughput pipelines. These could reveal synergistic protein-protein interactions or enzyme functions in large libraries of gene knockouts and help prioritize targets for downstream experiments. Developers could identify points of entry into these standard pipelines, as well as branch points where protocols might diverge from a high throughput workflow into more tailored experiments. These pipelines would also aid in the accelerated analysis of engineered protein functions.

Complex assays can prove challenging to scale up, but group discussions highlighted the value of increasing the throughput of simpler assays, as well. Attendees emphasized the importance of "quick and dirty" assays that provide qualitative data, rather than precise, quantitative results. Many simple screens can be improved to

quick, yet massive, scales using robotics, automated plate readers, and analyses software. These high-throughput workflows could help quickly prioritize proteins for subsequent screens. Attendees envisioned throughput as a staged, hierarchical model of simple and complex assays ordered to maximize efficient functional annotation.

Multiple orthogonal assays are critical to establish and validate observed enzymatic activity or protein function, but they are often challenging to do in routine academic laboratories. In a central phenotyping capability, complementary assays will help assure the protein being assayed is properly folded, is in its real-world active form, and correctly represents the phenotype of interest. These factors are especially critical when studying proteins from organisms that have unusual codon bias or unusual folding environments.

When tested for bioproduction applications, proteins often fail to reproduce their native functions in heterologous hosts or chassis organisms. These issues may stem from bottlenecks in transcription and translation—not knowing crucial binding partners in a protein complex or whether other enzymes are necessary to produce intermediate metabolites. High throughput phenotyping assays could systematically identify the sources of these errors and optimize design-build-test cycles for plug-and-play synthetic biology applications.

Attendees recommended a knowledge base of phenomics approaches that would systematically curate the results of all assays, including guidance to future users on optimizing protocols and feedback on what worked and what didn't, since the latter could also offer valuable insight into protein functions. As proteins of unknown function go through these varied assays and are correctly annotated, this annotation would be passed on to appropriate databases to minimize the perpetuation of incorrect or flawed annotation.

Attendees highlighted that "throughput" in such high throughput facilities refers not just to assays, but also to the development and synthesis of probes, themselves. While users have scientific subject matter expertise, identifying the best probe for a research question, as well as designing and synthesizing it, requires chemical biology expertise. A dedicated synthetic chemistry unit within the microbial phenotyping capability would be vastly useful to develop a multitude of probes for protein substrates or products.

Two potential approaches were discussed: one where a dedicated unit within the phenotyping capability could develop probes, and another where the effort could be spread across collaborators at synthetic biology laboratories. Both approaches, simultaneously applied as a shared resource, could serve as a powerful tool to researchers. Common probes could be synthesized in larger facilities, while the synthesis of bespoke probes for specialized experiments could be driven by individual user needs. One example highlighted was the use of lanthanides to tag antibodies to study enzymes or microbial surface proteins. A knowledge base describing the applications and limitations of each probe would help share data and optimize protocols within the user community.

3.3 Presentation: Phenotyping: *Kristine Burnum-Johnson and Lili Pasa-Tolic*

3.3.1 Phenotyping: *Kristin Burnum-Johnson*

When characterizing a microbe's phenotype, research studies often begin with genomic analysis, followed by transcriptomics or RNA sequencing and an analysis of proteins and metabolites relevant to the function under study.

Genomics provides a static snapshot of all possible ways a given cell might use its genes. Analyses of proteins, carbohydrates, lipids, and metabolites capture the dynamic processes by which a cell is constantly reacting to

its environment. But cellular information doesn't always flow in a linear manner from the genome, to RNA, to proteins, to metabolites.

Multiple strands of information converge in complex feedback loops to regulate metabolic activity. For example, environmental factors, such as a change in pH, might initiate a stress response that consequently alters a microbe's rates of transcription, translation, or protein turnover. Myriad regulatory mechanisms operate within individual microbes and microbial communities. The abundance of proteins or metabolites can alter rates of transcription or translation. Proteins might mediate transcriptional repression, where a protein's expression inhibits its own transcription.

But current experimental approaches often fail to capture this complexity, often only capturing the presence or abundance of metabolites or proteins. For instance, protein-mediated transcriptional regulation might manifest in an experiment as a mismatch between the abundance of a transcript and the proteins it encodes. These drawbacks create knowledge gaps in researchers' understanding of microbial functions.

Measuring proteins, lipids, and metabolites that reveal cellular workings beyond the genome allows researchers to capture the downstream effectors and regulatory processes critical to microbial biology, and thus, microbe-mediated processes such as soil nutrient cycles or the production of biofuels and materials.

High throughput multiomics approaches offer a practical route to study metabolites and post-genomic regulatory mechanisms and apply this information in engineering/design-build-test cycles. As an example of how automated, high throughput multiomics approaches can advance functional studies, consider the question of engineering Aspergillus pseudoterrus to make high amounts of a small acid, 3-hydroxypropionic acid (3-HP). M2PC workflows could be used to design, build, and test a suite of thousands of engineered mutant strains to simultaneously identify protein abundance, as well as the levels of the metabolite 3-HP.

This information can be fed into computational or machine learning-based models of metabolic pathways that could enable precise predictions about where to delete or increase protein expression to increase 3-HP production in future design-build cycles. That could lead to the ultimate goal: a fine-tuned strain that uses all its metabolic resources to make a desired target molecule.

3.3.2 Phenotyping: *Lili Pasa-Tolic*

In traditional bottom-up proteomics approaches, researchers typically cut proteins into pieces, sequence short peptides, and then piece data back together to reconstruct the protein. But this approach doesn't encompass a protein's three-dimensional form, interactions between multi-unit complexes, glycosylation, and other post-translational modifications essential to its function. These variations are often the sources of significant variation in a protein's binding targets or enzyme's efficiency, and design-build-test cycles are often slowed down or stalled without this data.

One solution to the problem is top-down proteomics approaches that bypass protein digestion and examine all the combinatorial variations in a protein, often described as 'proteoforms.' These approaches enable researchers to characterize all combinations of post-translational modifications in proteins that govern phenotypes.

The utility of such data is perhaps best illustrated by the spike protein from SARS-CoV2—a molecule so heavily decorated by glycans that nearly all of the protein's surface is covered in a sugary shield. This glycan coat enables the virus to evade host responses and play an active role in prompting certain domains of the spike protein into conformations where it can interact with host proteins.

Understanding such glycan coats is difficult when they're attached to proteins, particularly in the current absence of precision tools that allow researchers to edit or change specific glycan sites without altering the protein conformation. The only currently available tools are essentially some combination of traditional and top-down proteomics and glycoproteomics. For example, native mass spectrometry can be used to preserve structure and non-covalent interactions in proteins and elucidate various interactions that occur in protein complexes, such as the stoichiometry of complexes with other proteins, RNA, ligands, or metal cofactors that occur within a biological context.

The approach can also help screen ligands or libraries against target proteins with reasonable throughputs. It can also be used to track the dynamics of transient protein complexes over time or in different conditions and identify the most appropriate conditions to capture in follow-up studies. Combined with approaches, such as cryo-transmission electron microscopy (cryo-TEM), native MS has been applied to studies such as understanding the dynamics of circadian clock complex proteins in cyanobacteria.

By providing higher-resolution functional details, workflows that integrate native MS and cryo-TEM approaches can increase the utility of phenotypic data available for downstream design-build-test cycles. The ultimate goal of such approaches is to develop a so-called visual proteomics landscape, where each component in this cellular landscape is defined by its spatial distribution and orientation relative to other components.

Still, heterogeneity in sample biochemistry and computational image processing remain large bottlenecks for native MS-guided cryo-TEM structure determination. Developing these and other similar technologies to maturity to allow more routine, systematic, and higher throughput applications would be extremely powerful as part of the molecular phenotyping capability within M2PC.

The group discussions following these presentations focused on the need for large-scale microbial phenotyping capabilities with high throughput approaches, workflow design, and considerations on how to incorporate lower throughput, cutting-edge technologies that are equally important to understanding protein function and the relationship between genotypes and phenotypes.

3.3.3 Discussion

Questions to launch the discussion:

- What kind of multiomics and structural biology workflows are needed?
- What approaches should be applied to capture proteins of unknown function?
- How can models be integrated?

Discussions focused on the urgent need to look beyond a protein's sequence and elucidate three-dimensional structures, post-translational modifications such as glycosylation, the formation of protein-protein complexes, and other aspects of proteins that markedly alter their structure and function—thus, microbial phenotypes.

Understanding these aspects of proteins is critical to the eventual aim of engineering microbes for improved material production or other purposes. While the research community has recognized the need to study these in a high throughput fashion for years, few have had the resources to follow through. Attendees saw the M2PC as an opportunity for EMSL to take on the grand challenge.

Workflows at an M2PC should aim to image, characterize, and determine the impact of post-translational changes on protein function. Capabilities for in vitro protein expression are critical to elucidate structure and

experimentally validate functions. Enzymes expressed in vitro could also be used to synthesize substrates that can be used in downstream experiments.

These workflows will need to use a variety of host expression systems because efficient and accurate protein glycosylation outside a native host can be challenging. Some proteins and enzymes are easy to capture in standard systems, such as *E. coli* or *Saccharomyces cerevisiae*, while others require less common host systems, such as tobacco plants, or exotic microbes. In situ measurements in native hosts will also be necessary to identify physiological conditions for structure and function and, if necessary, recapitulate these in experimental systems.

These assays could also identify proteins of unknown function by using tools that infer function from structure and applying assays of metabolic flux to identify substrates being altered by a protein's activity. Characterizing and engineering ligand-binding proteins could create a powerful resource for the discovery and development of biosensors.

High throughput multiplexed workflows are essential to realize these possibilities within practical timescales. Workflows would ideally need to be completed within a month to execute design-build-test cycles a reasonable number of times.

The discussions highlighted the importance of considering throughput, not just in terms of speed, but also scale – meaning not just how quickly samples run, but how many can be processed simultaneously. Attendees proposed two possibilities for specialized pipelines: one that can run a large number of samples through commonly used assays, and another for smaller numbers of samples that need targeted experiments. This latter workflow would rapidly turn around smaller numbers of samples that require more complex, specialized assays necessary for targeted experiments. In designing these pipelines, users and capability developers will need to quantify the tradeoff between high throughput, low fidelity measurements and low throughput, information-dense experiments to make better judgments about allocating resources.

These workflows will likely span multiple techniques being run simultaneously. For example, workflows within a phenotyping capability might combine diverse methods, such as soft x-ray tomography with expansion microscopy and light sheet microscopy, to be able to determine spatial interactions and patterns in cells. Thus, the rate of data generation is limited to the pace of the slowest technique.

A specific focus on microbial communities will also prove necessary to understand functional outputs or the correlations between genotypes and phenotypes. Pipelines that include spatiotemporal measurements and systematic, combinatorial studies of microbes (mixing three species in different ratios and combinations, for example) will be essential in understanding how inter-species interactions affect protein function.

Computational tools are critical to scaling up the throughput and utility of data in these pipelines. Currently, artificial intelligence- and machine learning-based approaches, such as the AlphaFold Protein Structure Database used to predict protein structures, are optimized for human proteins and model organisms. Extending these to microbes and using them in conjunction with experimental techniques, such as cryo-EM, will provide insight into how microbial proteins assemble into complexes and their functional context. Artificial intelligence and machine learning approaches could also help parse data on post transcriptional modifications, such as glycosylation and capture interactions in metabolite binding protein complexes. Programs such as AlphaFold could prove extremely useful to prioritize proteins for structural analysis and characterization in iterative design-build-test-learn cycles.

Data from these varied pipelines should not remain siloed within projects but be broadly accessed by many users within the capability. A robust infrastructure to standardize and curate results from different workflows will prove essential. The discussions highlighted that within a single pipeline, different techniques might have different standards or resolution for data. Combining datasets and results meaningfully will require expertise handling output from a variety of techniques. Integrating data in a meaningful manner will also require a consideration of time frames, since metabolite flux occurs on the scale of seconds, transcripts change in minutes, and proteins on the scale of hours.

Metadata on sample collection, environmental conditions, and experimental details are often critical to making sense of data but remain in the hands of users. This phenotyping capability could incentivize users to provide metadata in a standardized format and use standard nomenclature across projects, and thus, maximize the output and utility of generated data.

One option explored was that of a specialized knowledgebase specific to functions of interest, e.g., biomass degradation, such that all data on microbes studied for their biomass degradation pathways, assays used, results, and metadata would be collated to provide a snapshot of all phenotypes relevant to the function of interest. Such consistent, high-quality data would prove of enormous value, not just to the users who generated the information, but others who could extend and apply it in future studies.

3.4 Presentation: Data and coordination: *Lee Ann McCue*

To set the stage for the last section of the workshop, this presentation provided an overview of data management and coordination at EMSL, including updates and recent improvements.

In phenomics, as in other disciplines, data science relies on the “FAIR” principles. Briefly, these principles state that data and digital objects, any metadata about them, and searchable resources based on these data and metadata must be: findable, accessible, interoperable, and reusable.

Previous presentations and discussions highlighted the need for robust, accessible, and shared knowledgebases of materials, strains, and data produced within this microbial phenotyping capability.

Rich metadata, curated using controlled vocabularies to describe biological or environmental samples, are critical to obtain a clear, detailed provenance of the samples, themselves, and any data generated from them. These controlled vocabularies need to conform to community standards and include globally unique, persistent identifiers.

Adhering to FAIR principles and committing to open, community-driven science is necessary so the data generated in this phenotyping capability can be used and reused by the community to build models and understand processes. At present, EMSL users have immediate and persistent access to data they generate at EMSL for their projects. Within the next year, that data will be automatically made accessible to the research community one year after it is generated, although users may make it public sooner if they wish to do so. This update aims to balance users’ research needs with guidance from the Office of Science and BER with respect to sponsors’ expectations for making data accessible.

At present, EMSL’s open data portal can be searched by bibliographic information, keywords, time frame, and so on, but not by metadata. To incorporate metadata using a controlled vocabulary, EMSL has begun to implement unique sample identifiers for all samples received for analysis. Some examples currently in use include IGSN identifiers (International Geo Sample Number) that are generated by an external service for physical samples from the field, and UUIDs (universally unique identifiers) that are generated using secure

random number generators. Users can tag samples with IGSNs and EMSL can also help users generate UUIDs. These identifiers are especially critical for samples shared across facilities, such as EMSL and JGI, or when a sample is split and sent to multiple types of instruments for data generation. Unique identifiers help tag the parent sample, child samples and all the data generated from each of them so users can track and have provenance of the entirety of information. Current EMSL users have already begun to implement this as they are asked to provide unique IDs and metadata spreadsheets along with samples.

In addition, EMSL has also begun associating metadata, such as latitude and longitude at collection sites, soil type, species taxonomy, etc., directly with the samples, themselves. These efforts leverage community resources, such as the Genome Standards Consortium (MIxS standards), the open biological and ontologies foundry (OBO), and environmental ontology (ENVO). Leveraging these existing standards from the research community avoids redundancies and wasted labor on generating new standards.

As incoming metadata is operationalized with a rich, controlled vocabulary, EMSL can expand the search portal to support searches for data using standard metadata fields and controlled vocabularies—thus expanding the utility and accessibility of data, as well as making research more reproducible and consistent over time.

EMSL's efforts at data coordination are supported by strong, active partnerships and many active working groups. Partnerships with other DOE-funded resources, such as NMDC, JGI, KBase, ESS-Dive, OSTI.gov, and others, have been vital to these efforts. These efforts serve as a prelude to the new microbial phenotyping capability, where updated data management and coordination efforts can be built into pipelines from the capability's inception, and thus, inform and accelerate research and discovery for users and the community.

3.4.1 Discussion

Questions to launch this discussion:

- How do we plan data standardization in conjunction with experiments?
- How do we envision coordination with other DOE BER facilities and projects?

Attendees saw immense potential for this microbial phenotyping capability to pioneer best practices for data integration across laboratories, as well as across types of samples, assays, and workflows.

Discussions highlighted the importance of building standards for data collection and curation into experimental assays at the onset, beginning with the sourcing of samples, instrumentation, and other aspects. Establishing standard, modular sets of experimental conditions for microbial growth and functional assays will also help refine the parameter space for metadata and produce cohesive information that does not become too complex to be useful.

To make data (re-)usable, it must first be searchable. Metadata categorized and tagged with relevant, detailed ontology can help researchers quickly scan existing datasets to glean the current state of knowledge and avoid repetitive experiments. An ontology that features terms relevant to DOE mission areas, such as biomass conversion or bioenergy, would be especially useful to program managers.

This standardized information should be linked to each sample through all assays and workflows, so that even samples that are split and processed in separate pipelines for different assays remain linked in the database.

Reproducibility is also a key consideration. One option suggested was to create repositories for standard organisms, reagents, or other materials that can be used as a basis of comparison for experiments. These standards would help integrate data within the knowledgebase by creating a baseline for normalization.

Curating data on successful experimental design and variables used therein as part of the metadata would also help promote reproducibility. While this applies to metabolic models, as well, further close attention is needed to establish the kinds of metadata required to repurpose models for future study.

The discussions also centered around precisely how data would be made available. They highlighted the value of releasing data within a year of its generation, rather than keeping it embargoed for years. A standard framework for retrieving data using standard APIs or other similar tools were considered as a better option than raw data tables or simple website queries in terms of enabling a variety of end use applications.

A knowledgebase equipped with machine learning tools that links related datasets, probes, and assays to enable users to perform “experiments on the fly” could also be extremely valuable. Metadata on probes could include specifics on validation, off-target effects, and other details that are often left out of published papers. Design-build-test cycles can also be integrated into a phenotyping knowledgebase so data on metabolite flux through various engineered pathways become available as a resource for future studies.

Attendees also recognized that a capability such as this one would encompass a wealth of unpublished data. A data management system that enables users to access and repurpose data without the context of a published paper would be especially important. These sorts of data integration will require computation, statistical analyses, and machine learning approaches applied from the experimental design stage across the entire phenotyping capability.

Strategies to maintain this annotated database as a long-term resource should be identified early on. Tools to integrate data from short-term and long-term experiments can feed into a common repository for the entire community.

Collaborating with existing resources to incorporate standardized data management practices will avoid redundancies and improve the reach of this phenotyping capability as a community resource—not just for EMSL users, but for others in the industry, bioenergy research centers, and various other consortia. The goal of this microbial phenotyping capability would be to not just create a new resource but generate high-quality functional annotations and other data that are used to update public databases. In doing so, this new capability would reach researchers who may never see this resource in-person but would still greatly benefit from its output.

4.0 Conclusion and next steps

- Identify how existing EMSL capabilities for spatially defined metabolomics, high-resolution imaging, and other technologies can be leveraged for the new capability
- Explore available expertise within the DOE landscape for potential collaborations and dedicated partnerships with NMDC, JGI, and other laboratories
- Expand on data coordination efforts and controlled vocabularies to build data infrastructure

EMSL

The Environmental Molecular Sciences Laboratory

A U.S. Department of Energy Office of Science user facility at
Pacific Northwest National Laboratory

3335 Innovation Boulevard
Richland, WA 99354

www.emsl.pnnl.gov

EMSL