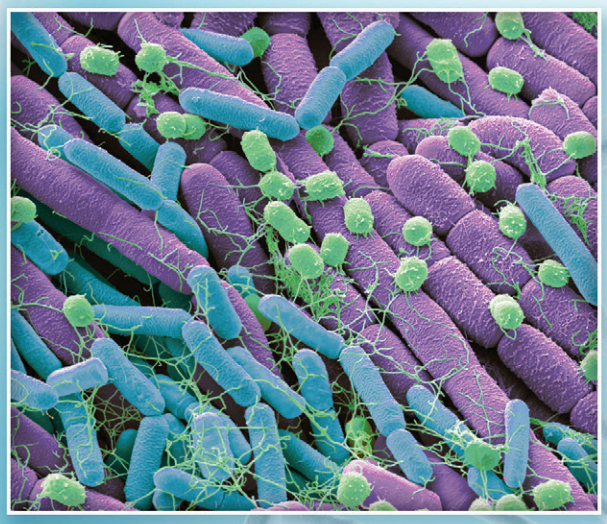
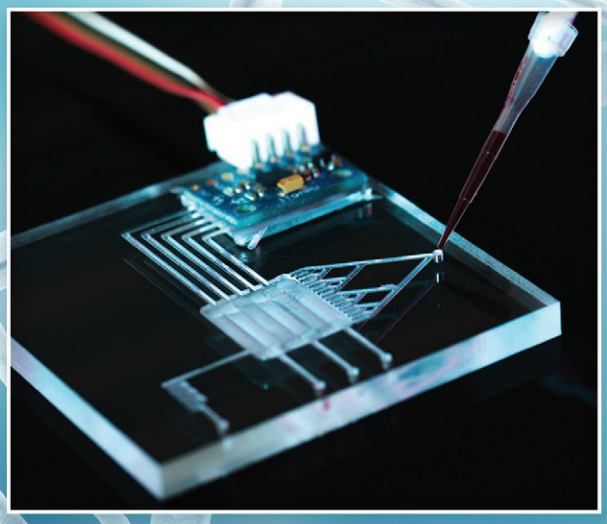
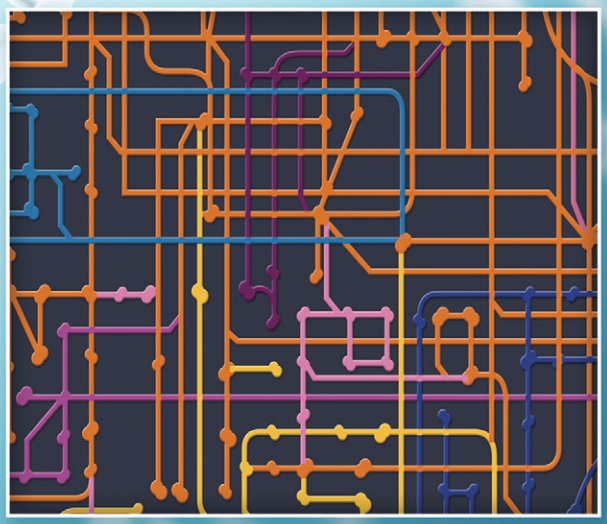
DOE/SC-0199

Breaking the Bottleneck of Genomes: Understanding Gene Function Across Taxa

Workshop Report







U.S. DEPARTMENT OF ENERGY | Office of Science

Office of Biological and Environmental Research

Breaking the Bottleneck of Genomes: Understanding Gene Function Across Taxa Workshop

November 1–2, 2018
Bethesda, Maryland

Convened by

U.S. Department of Energy
Office of Science
Office of Biological and Environmental Research

Co-Chairs

Robin Buell
Michigan State University

Adam Deutschbauer
Lawrence Berkeley National Laboratory

Organizers
Biological Systems Science Division

Dawn Adin
dawn.adin@science.doe.gov

Catherine Ronning
catherine.ronning@science.doe.gov

This report is available at genomicscience.energy.gov/genefunction/.

About BER
The Biological and Environmental Research (BER) program advances fundamental research and scientific user facilities to support Department of Energy missions in scientific discovery and innovation, energy security, and environmental responsibility. BER seeks to understand biological, biogeochemical, and physical principles needed to predict a continuum of processes occurring across scales, from molecular and genomics-controlled mechanisms to environmental and Earth system change. BER advances understanding of how Earth's dynamic, physical, and biogeochemical systems (atmosphere, land, oceans, sea ice, and subsurface) interact and affect future Earth system and environmental change. This research improves Earth system model predictions and provides valuable information for energy and resource planning.

Cover Credits
Clockwise from top left: Colored scanning electron micrograph of soil bacteria (Steve Gschmeissner, Science Photo Library). Sorghum (Center for Advanced Bioenergy and Bioproducts Innovation). Conceptual metabolic map drawing. Lab on a chip (Shutterstock).

Suggested citation for this report: U.S. DOE. 2019. *Breaking the Bottleneck of Genomes: Understanding Gene Function Across Taxa Workshop Report*, DOE/SC-0199, U.S. Department of Energy Office of Science. genomicscience.energy.gov/genefunction/.

DOE/SC-0199

Breaking the Bottleneck of Genomes: Understanding Gene Function Across Taxa

Workshop Report

Published September 2019

U.S. DEPARTMENT OF ENERGY | Office of Science

Office of Biological and Environmental Research

Contents

Executive Summary v
- Technology Innovations v
- Computational Advancements vi
- Focal Biological Systems vi
- Breaking the Genome Bottleneck viii

1. Introduction 1
- BER Genomic Research 1
- Sequencing Outpaces Gene Function Determination 2
- Gene Function Across Multiple Levels 3
- Current Gene Annotation Strategies and Resources 3
- Challenges in Gene Function Discovery and Annotation 4
- Opportunities for Rapid Progress 5

2. Technology Innovations 7
- Need for Scalable Experimental Technologies 7
- Reducing Technology Barriers 9
- Improving Gene Manipulation Efficiencies and Phenotyping 9
- Capturing Molecular Processes at the Single-Cell Level 11
- Targeting Classes of Proteins 12
- Advancing Molecular Measurements of Proteins 16
- Extending High-Throughput Genetic Approaches to Relevant Ecological Contexts 18
- Integrating Technologies to Scale Gene Function Determination 18

3. Computational Advancements 21
- Computationally Driven Gene Function Discovery 21
- Databases and Knowledgebases of Gene Annotations 21
- Computational Framework for Discovery of New Gene Functions and Accurate Annotation 23
- Infrastructure Requirements for Integrating Diverse Omic Data 24
- Gaps in Experimental Data 25
- Strategies and Data Sources for Evaluating Confidence in Gene Functional Annotation 25
- Community Engagement 27
- Potential for Gene Function Discovery by High-Performance Computing and New Algorithms 28

4. Microorganisms 29
- Target Microorganisms 30
- Moving Experimental Tools from Model to Nonmodel Microorganisms 34
- Determining Gene Function in Natural Contexts 36
- Genetic Redundancy and Functionally Distinguishing Paralogs 38

5. Plants 41
- Plant Systems: Unique Challenges 41
- Focal Species to Accelerate Gene Function Discoveries 43
- Well-Annotated Genomes and Associated Datasets 44
- Prioritization of Gene Sets for Functional Experimentation 45
- Perturbation of Genes via Gene Editing 45
- Modeling of Relevant Plant Processes 45
- G × E: Role of Environment 46
- Minimal Plant Genome Platform for Gene Function Discovery 46

6. Conclusions and Outlook 47

Appendices 49
- Appendix A. Workshop Agenda 49
- Appendix B. Breakout Session Charge Questions 50
- Appendix C. Workshop Participants 53
- Appendix D. References 54
- Appendix E. Acronyms and Abbreviations 60

Executive Summary

In the last few decades, high-throughput technologies using various "omics" have enabled unprecedented views of biological systems at the molecular level. In parallel, the integration of omic datasets using computational modeling has provided new understanding of biological processes in organisms relevant to the U.S. Department of Energy's (DOE) missions in energy and the environment. Collectively, these developments have spearheaded the advancement of systems biology, which can be viewed as a holistic approach for deciphering the complexity of biological systems. However, as high-throughput omic technologies and integrative systems biology efforts have improved our understanding of some biological systems, analyzing and finding meaningful answers within these massive datasets remain extremely challenging, in large part due to the lack of fundamental knowledge of gene function. Indeed, all sequenced genomes, both microbes and plants, contain large numbers of genes of "unknown function" that significantly limit scientists' ability to model, predict, and engineer organisms with enhanced functions relevant to DOE. Current methodologies can be employed to decipher gene function, but they are typically slow, laborious, inefficient, and not scalable. This "bottleneck" in genome understanding could be broken with new, innovative, and transformative experimental tools, datasets, and computation that can define gene function on a massive and high-throughput scale compatible with the pace of DNA sequencing.

In light of this grand challenge, DOE's Office of Biological and Environmental Research (BER) convened the Breaking the Bottleneck of Genomes: Understanding Gene Function Across Taxa workshop on November 1–2, 2018. This workshop brought together leaders in microbiology, plant sciences, technology, and computation, who collectively identified the experimental and data analysis gaps preventing large-scale gene function determination as well as opportunities for overcoming these gaps.

The workshop was organized around breakout sessions in which participants discussed research challenges associated with gene function discovery and accurate annotation across taxa. The discussions included the breadth of diverse, high-throughput technologies needed for characterizing genes of unknown function and how the diverse data from these technologies could be integrated with new and existing computational platforms to accurately propagate these annotations to newly sequenced genomes. While many of these technological and computational challenges are universal across taxa, the workshop organizers recognized that organism-specific biology and experimental limitations would prevent the development of unified solutions. Thus, separate breakout sessions were held for plants and microorganisms, which represent the most significant BER investment in genomic sciences. This report presents the challenges, knowledge gaps, and opportunities for accelerated gene function discovery and accurate gene annotation in four areas: technology, computation, microorganisms, and plants. Each of these areas is framed by the charge questions posed to all the participants and discussed at the workshop.

Technology Innovations

Multiple technology innovations will be required to advance understanding of gene function at multiple levels of organization, including the biochemical function of proteins and their roles in organismal physiology and, more broadly, in ecosystem processes. A subset of these technology gaps potentially can be addressed by greatly scaling existing methodologies; other gaps will require new research and innovation to enable the application of existing methods to diverse taxa at low cost and the development of new, groundbreaking methodologies adequate for high-throughput function determination at the protein, metabolite, organism, and ecosystem scales.

Opportunities for closing the technology innovation gap include the application of mature omic approaches to DOE-relevant species to systematically inform (1) gene function, (2) development of inexpensive and high-throughput gene manipulation

across taxa, (3) systematic linkage of genotype to phenotype, (4) expansion of work at the single-cell level, and (5) elevation of experimental studies to ecosystems. These opportunities are addressable at the single-investigator level; however, due to the efficiencies of scale and access to specialized equipment, organizing a subset of strategic omic technologies at a consortium of discovery centers could facilitate rapid method optimization and dataset generation, reduce costs and duplication of efforts, promote method standardization, enable seamless integration of knowledge such as orthogonal validation of findings, and establish a framework for robust and accurate public dissemination of "gold-standard" functional annotations and inferences. Nevertheless, transformative technology innovations to break the gene function bottleneck will require future novel approaches developed most likely through single investigators and teams of researchers focused on breaking technology barriers.

Computational Advancements

To achieve the ultimate goal of accurate gene function annotation and inference across taxa, a number of computational hurdles must be overcome. Foremost, nearly all genes from newly sequenced genomes are annotated for function based on sequence similarity to characterized proteins. However, many proteins are too distant from a characterized protein to be accurately annotated by this approach. As a result, genome databases are filled with gene annotations that are either uninformative (i.e., "hypothetical protein") or incorrect (i.e., the wrong substrate for a paralogous enzyme). In addition, although a number of established protein databases such as UniProt and RefSeq exist, along with multiple gene annotation pipelines, these resources are not always in agreement. Furthermore, updating erroneous gene annotations within these resources is not straightforward. Biocuration by experts is a proven method for accurately correcting gene annotations in databases, but this approach is costly and does not scale with the current pace of genome sequencing.

Opportunities for improving the computational inference of gene function include the integration of new and established computational resources and databases, outreach to biocurators as well as the scientific community at large, and inference of gene function from diverse omic data using new analytical approaches such as machine learning. For the community to realize these goals, computational interfaces could be developed that would seamlessly integrate with (or be a part of) existing DOE computational resources including the Systems Biology Knowledgebase (KBase), Joint Genome Institute (JGI), Environmental Molecular Sciences Laboratory (EMSL), and National Energy Research Scientific Computing Center (NERSC), as well as the National Center for Biotechnology Information (NCBI) supported by the National Institutes of Health and Protein Data Bank (PDB) managed by the Research Collaboratory for Structural Bioinformatics. Another opportunity for improving gene function knowledge would be to integrate diverse omic datasets to confidently infer gene function and to accurately transfer these annotations across newly sequenced, homologous genes using these new computational interfaces. For maximal use of these computational innovations by the community, accurate and dynamic gene function annotation across taxa will require precise versioning of data, algorithms, protocols, and annotations associated with inference provenance. Positive outcomes could be the automated (or semiautomated) and accurate inference of gene function from any sequenced genome, confidence scores and evidence for each annotation, and a community-accessible computational infrastructure for rapidly inferring new gene functions that could be validated by targeted experimentation.

Focal Biological Systems

BER-relevant organisms include microorganisms, algae, and plants because of their potential for producing sustainable biomass, synthesizing biofuels and related bioproducts, sequestering carbon, and transforming environmental contaminants. However, evolution has resulted in diverse taxa across the tree of life, and consequently there exists a very wide range of organisms whose genomes require substantially better annotation to ultimately enable biology-based solutions to the nation's energy and environmental challenges. In some instances, restricting efforts to a subset of taxa may be required to develop enabling

technologies and advanced computational approaches to discover gene function, with the important aim that these approaches would be quickly applied more broadly across taxa. Conversely, some experimental technologies and organisms are more amenable to high-throughput and low-cost experimentation and potentially can be scaled immediately.

Microorganisms

Microorganisms, including bacteria, archaea, fungi, and protists, have a profound effect on global nutrient cycles, plant health, and environmental remediation of toxins. In addition, microorganisms can be harnessed as cellular factories for metabolic engineering applications, including the production of bioproducts from plant-derived biomass. Given their massive diversity and the low cost of working with many of these systems, currently scalable technologies could be applied systematically across a representative selection of the microbial tree of life in an effort to discover new gene functions and to comprehensively refine existing annotations. Efforts could also be targeted to particular biological questions and relevance to existing DOE-funded efforts such as the microbiome of a biomass plant, a pan-genome, or hosts for metabolic engineering. A concerted effort to characterize highly conserved but poorly understood proteins, such as those that contain domains of unknown function, could offer maximal knowledge gain. In parallel, new disruptive approaches are urgently needed to determine gene function in single-cell organisms, especially for unculturable organisms and those that are culturable but currently genetically inaccessible.

Streamlining development of genetic methods across taxa, including insertional mutagenesis, targeted mutagenesis, recombineering, and CRISPR/Cas-based genome editing, could lead to understanding of the cellular roles of genes in diverse BER-relevant microorganisms. In parallel, phenotyping platforms that interrogate cellular phenotypes, including cell morphology, intracellular metabolite abundance, protein localization, and secondary metabolite production, would provide improved knowledge of gene function. Genetics-based methods are currently the most high throughput via coupling to next-generation sequencing, but the continued development of ultrasmall-volume biochemical assays (e.g., those that can be encapsulated within droplets at a massive scale) offers great promise for characterizing function from diverse protein families, including from uncultivated microorganisms. Lastly, microorganisms and their genes have evolved within the context of other organisms. Hence, it will be important to complement efforts characterizing gene function(s) in pure culture with studies using laboratory ecosystems relevant to DOE interests and missions such as abiotic factors (e.g., soil-mineral) and biotic interactions (e.g., plant-microbe).

Plants

The characterization of gene function in land plants presents unique challenges relative to single-celled microorganisms. These challenges are attributable to the size of plant genomes, complexity of organs composed of multiple cell types, heterogeneity of environments that plants live in and respond to, barriers to genetic manipulation, and logistical infrastructure required for plant experimental systems. Thus, targeting efforts and resources on one or several key species or clades of closely related species relevant to DOE's missions could enable the development of improved, paradigm-changing methods, tools, and resources to assess gene function that, once developed for a core set of species, can be applied across the plant kingdom. Opportunities for focused plant research systems include sorghum (annual C4 biofuel feedstock), switchgrass (perennial C4 biofuel feedstock), *Camelina* (nonfood oilseed), poplar (perennial C3 biofuel feedstock), and model species such as *Arabidopsis thaliana* and *Chlamydomonas reinhardtii*.

Central to all downstream efforts is access to well-annotated genomes with associated large-scale datasets that are user friendly. The example focal species, listed in the previous paragraph, are amenable to genetic transformation and already have sequenced and annotated genomes along with a subset of omic data, but these data are neither uniform in breadth and depth nor well integrated with existing datasets. To maximize knowledge within the overall scientific community, data could be integrated into a centralized repository

with appropriate standards and tools for data analysis and interpretation as well as incentives for data deposition and curation by diverse research groups. Cataloging relevant omic datasets and generating additional datasets to ensure equivalent breadth and depth across the focal species would enable the generation of priority lists for functional validation using existing gene-editing technologies. Machine-learning approaches and emerging high-throughput macro and molecular phenotyping methods provide an opportunity for exploring phenotypic plasticity of plants in diverse environments and increasing the resolution of biological process knowledge.

Synthetic organisms that represent the minimal gene complement for life provide a chassis to rapidly test gene function, and, in microbial systems, scientists have been able to fabricate minimal genomes. In plants, a similar approach might be considered in which genes and/or gene cassettes are combinatorially added or disrupted, providing a means to assess gene function as well as a platform for applying synthetic biology and engineering novel biochemical function in plants.

Breaking the Genome Bottleneck

BER-relevant microorganisms and plants contain an amazing diversity of discovered DNA sequence resources, and determining the functions of these genes would have a tremendous impact on all aspects of biology and environmental research. Although a daunting challenge, understanding gene and genome function across taxa can be significantly improved. Needed technological advancements to overcome this challenge include diverse experimental approaches that inform gene function and that can be flexibly applied across species and at low cost.

Annotating gene function is as much a computational challenge as an experimental one. Multiple annotation tools, protein sequence databases, and decades of molecular biology research already exist. Determining the best ways to leverage and couple these valuable resources to new, transformative datasets and data analysis tools will be important for breaking the genome bottleneck. DOE has a long history of solving grand scientific challenges through long-term visioning and investment, which provide an avenue for successfully characterizing the millions of genes of unknown function that currently reside in sequence databases. This endeavor's success will affect BER's mission by accelerating the development of genomics-enabled solutions to global challenges in sustainable energy development and environmental management.

1. Introduction

BER Genomic Research

The U.S. Department of Energy's (DOE) Biological and Environmental Research program (BER) has a long-standing mission to solve critical challenges in energy security and environmental stewardship. As part of this mission, BER has invested in a number of crosscutting technologies and programs, enabling multiscale, systems-level investigations into individual organisms, environments, and communities to ultimately achieve a predictive understanding of organismal function under changing conditions. Long term, this vision's successful implementation will have far-reaching impacts, including the development of (1) resilient biomass plants, (2) a range of microbial hosts for producing high-value molecules from renewable sources, and (3) an accurate model of global nutrient cycles and the impacts on them by organisms and anthropogenic inputs.

In this endeavor, genomics is fundamentally critical to BER's interdisciplinary research portfolio. Genomes of organisms and communities (metagenomes) provide the genetic "parts list" for engineering applications, understanding and manipulating nutrient cycles, and improving plant productivity (see sidebar, Research Portfolio for BER's Genomic Science Program, this page). Consequently, BER has invested significantly in the sequencing of numerous mission-relevant plants, microorganisms, and microbial communities. In parallel, a number of complementary omic approaches have been developed to characterize the genes, transcripts, proteins, and metabolites of these organisms and communities and achieve a systems-wide understanding of organismal function across scales ranging from the molecular to the ecosystem (U.S. DOE 2017).

Research Portfolio for BER's Genomic Science Program

BER's Genomic Science program supports fundamental research to understand the systems biology of plants and microbes as they respond to and modify their environment. This research builds on the foundation of sequenced genomes and metagenomes, focusing on a tightly coupled approach that combines experimental physiology, omics-driven analytical techniques, and computational modeling of functional biological networks.

The Genomic Science program research portfolio comprises:

Bioenergy Research Centers. Provide technologies and scientific insights across four multipartnership centers laying the groundwork for sustainable, cost-effective advanced biofuels and bioproducts from lignocellulosic biomass.

Systems Biology for Bioenergy. Improves fundamental understanding of microbes with bioenergy-relevant traits for deconstructing biomass and synthesizing biofuels and bioproducts.

Plant Science for Bioenergy. Elucidates and validates the functional roles of genes, gene families, and associated pathways to enhance understanding of critical processes in DOE-relevant plant systems.

Sustainability Research for Bioenergy. Investigates plant-soil-microbe interactions in laboratory and field settings to enhance biomass productivity under changing biotic and abiotic conditions.

Biosystems Design. Develops knowledge for engineering useful traits into plants and microbes to produce biofuels and bioproducts and to advance biotechnology.

Environmental Microbiome Science. Links structure and function of microbial communities in the field with key environmental or ecosystem processes.

Computational Biology. Provides new computational approaches and hypothesis-generating analysis techniques, data, and simulation capabilities such as the DOE Systems Biology Knowledgebase (KBase) to accelerate collaborative, reproducible systems science.

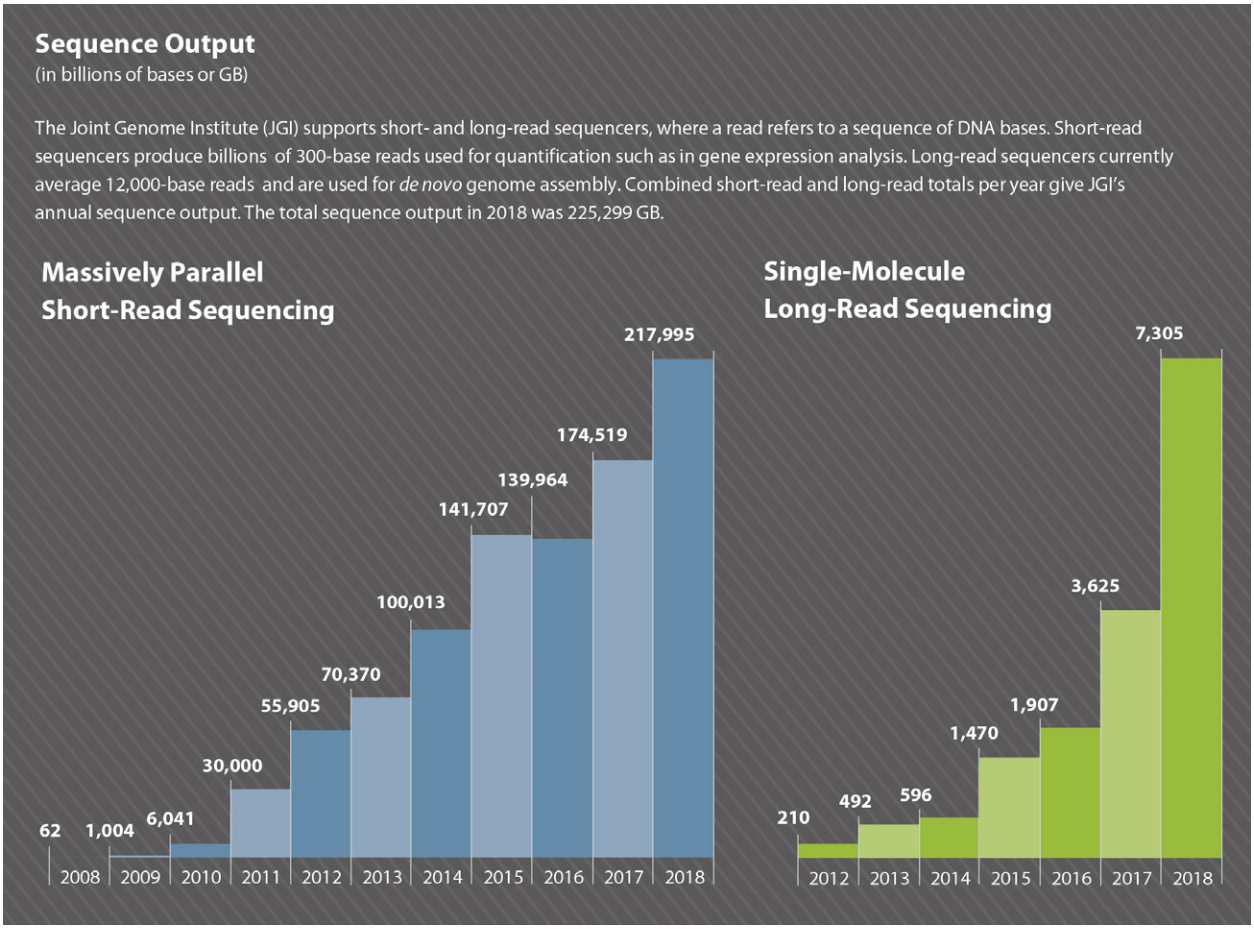


Fig. 1.1. Increase in Sequencing Output at the Joint Genome Institute. Revolutionary advances in DNA sequencing technologies have led to an explosion of data in the billions of bases (GB), helping to democratize genomics. [U.S. DOE Joint Genome Institute 2018]

Sequencing Outpaces Gene Function Determination

The rapid acceleration of DNA sequencing has revolutionized biology and democratized genomics. Sequencing the entire genomes of multiple organisms, sometimes in a matter of days, is now commonplace for individual laboratories. These advances in DNA sequencing have led to an exponential increase in the quantity of sequence data in publicly available repositories (see Fig. 1.1, this page). Databases now contain the genomes of hundreds of plants and tens of thousands of microorganisms including bacteria, fungi, and protists. Despite the availability of these data, the promise of genomics to solve global environmental and energy challenges remains unmet. A significant challenge lies in the interpretation of genome sequences and a predictive understanding of how not only individual genes but also the whole complement of genes and their encoded pathways contribute to their physiology and function in environmental and industrial contexts.

Underlying this challenge in harnessing the biological potential of organisms is the limited understanding of the function of genes within a genome. As described in more detail in subsequent sections, the overwhelming majority of genomes are annotated in an automated, transitive fashion based solely on sequence similarity or motif or domain presence, resulting in coarse,

inaccurate estimations of gene function. Few, if any, genes are characterized for molecular, biochemical, or biological function. When coupled with gene duplication and neo- or subfunctionalization, transitive annotations can be even more erroneous. Additionally, a substantial subset of genes in a genome can lack any similarity to other sequences and thus are annotated as hypothetical proteins. Such genes cannot be included in predictive models because they are not useful for describing the behavior of organisms under changing environments. Narrowing this gap between genome sequencing and functional understanding of the encoded genes would enable truly predictive biology and genome-enabled solutions to global environmental and energy challenges. In light of this significant knowledge gap, BER's advisory committee highlighted improved gene function annotation as a recommendation for further investment in a 2017 report (BERAC 2017).

BER subsequently convened the Breaking the Bottleneck of Genomes: Understanding Gene Function Across Taxa workshop on November 1–2, 2018 (see Appendices A–C, pp. 49–53). This workshop brought together leaders in microbiology, plant sciences, technology, and computation, who collectively identified the experimental and data analysis gaps preventing large-scale gene function determination across taxa, as well as current and future opportunities for overcoming these gaps. This report presents these challenges, knowledge gaps, and opportunities for accelerated gene function discovery and accurate gene annotation in four areas: technology, computation, microorganisms, and plants.

Gene Function Across Multiple Levels

There is no single definition of "gene function," but a greatly increased understanding of gene function across multiple levels of biological organization is necessary to enable predictive biology. For instance, genetic approaches can identify a gene's physiological role through the observable impact (phenotype) of a mutation in that gene (genotype). While genotype-phenotype relationships are powerful, they often do not provide mechanistic insight into the gene product's activity. Indeed, most genes encode for proteins, which transform a plethora of metabolites, serve as structural components of cells, and perform a range of diverse activities within the cell. Therefore, it is also vital to consider protein structure, biochemistry, localization, and post-translational modifications to fully understand the function of a gene. Moreover, genes and their products do not function in isolation, but rather are part of a complex network of interacting components that contribute to higher-order processes including regulation, metabolic networks, protein-protein interactomes, development, tissue differentiation, organismal interactions, and ecosystem processes. Consequently, considering and understanding gene function across these higher-order scales are necessary, particularly for predictably engineering organisms with desired traits and robust performance (see Fig. 1.2, p. 4).

Current Gene Annotation Strategies and Resources

The computational annotation of a newly sequenced genome involves two stages. First, genomic features such as genes, introns and exons, and regulatory regions are identified. Second, gene functions are predicted based on sequence similarity to previously characterized genes. For the latter, a number of resources have been developed to warehouse known and predicted gene functions including protein databases such as UniProt, protein family or domain databases such as Pfam and COG, and pathway-level annotation resources such as KEGG and MetaCyc. To streamline annotation of gene function, a number of pipelines have been developed to accelerate genome annotation by bundling multiple steps of the process including SEED-RAST (Overbeek et al. 2014), the Integrated Microbial Genomes and Microbiomes (IMG/M) prokaryotic gene annotation system (Chen et al. 2019), the National Center for Biotechnology Information's annotation service, and the MAKER-P program (Campbell et al. 2014). For each of these pipelines, an assembled genome sequence and, for plants, transcript evidence are required as input, facilitating the computational annotation of many thousands of sequenced genomes.

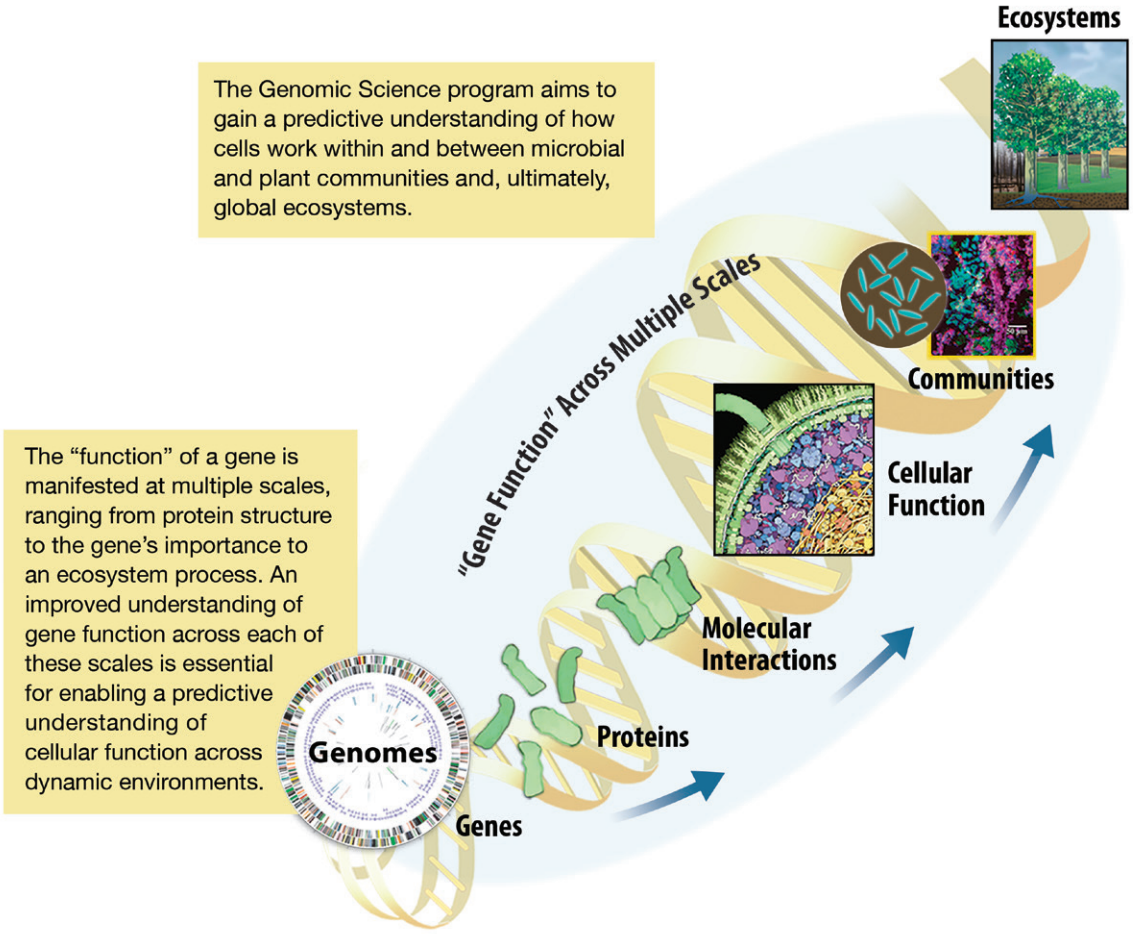


Fig. 1.2. "Gene Function" Across Multiple Scales. Understanding gene function across multiple levels of biological organization is a key objective of the Department of Energy's Genomic Science program.

Challenges in Gene Function Discovery and Annotation

The automated assignment of predicted gene functions in newly sequenced genomes is necessary to keep pace with the increased rate of DNA sequencing. However, this approach suffers from limitations. Foremost, genes that are not similar to a previously characterized gene cannot be accurately annotated with function using homology-based approaches. Even in instances where a homologous protein has been studied experimentally, accurately predicting gene function is not straightforward because closely related proteins can have different functions and distantly related proteins can have the same function. As a consequence, protein databases are filled with gene annotations that are uninformative (i.e., "hypothetical" protein), vague ("transporter" with no specific substrate), or simply wrong (Schnoes et al. 2009).

A further challenge is the propagation of existing and newly acquired knowledge, which the gene function annotations for newly sequenced genes often do not take into account. Traditionally, biocurators mine scientific literature and transfer this information into protein databases. This approach, while powerful, is currently mostly confined to databases dedicated to very well studied model organisms (e.g., the bacterium *Escherichia coli*, the yeast *Saccharomyces cerevisiae*, and the plant *Arabidopsis thaliana*), and it is not scalable to the number

of organisms and genes now under investigation. For individual investigators, the process of updating and improving gene annotations in protein databases is often neither trivial nor incentivized, and there are no platforms or forums to readily update annotations outside of these dedicated organism-specific databases.

Opportunities for Rapid Progress

Better capturing and propagating existing knowledge will improve some gene annotations, but millions of genes in sequence databases will remain vaguely or incorrectly annotated, despite decades of research. The current throughput of gene function discovery is simply too slow to keep pace with DNA sequencing. To bridge this sequence-to-function gap, new high-throughput experimental strategies and data are needed to systematically determine gene function. In parallel, improved computational approaches are required to accurately and automatically infer gene function from diverse data types and prior knowledge.

The challenge of broadly improving gene function assignments across taxa is daunting, but recent technological and computational advances offer paths forward for rapid progress. A number of measurement technologies for sequencing, proteomics, metabolomics, imaging, and structural biology have improved substantially. Many assays informative for gene function determination, such as large-scale genetics, have been effectively coupled to next-generation DNA sequencing, enabling the cost-effective generation of large datasets. Further, approaches such as gene editing that have been pioneered in model organisms are increasingly being applied to nonmodel species, thus opening the door to comparative functional genomics. Rapid reductions in the cost of DNA synthesis, protein production, and biochemical assays from using laboratory automation and miniaturization can be used to characterize millions of protein variants from diverse organisms, including those that have yet to be cultivated in the laboratory. Computational approaches such as protein structure prediction, ligand-binding prediction, and machine learning can be applied to infer function and to propagate this information across the vast space of sequenced genes.

Lastly, DOE has invested in a number of research programs, user facilities, and enabling capabilities for the scientific community that can be leveraged for the systematic discovery and accurate annotation of gene function across taxa (see sidebar, DOE User Facilities and Enabling Capabilities for Gene Function Discovery and Annotation, this page). Other new and upcoming data resources such as the National Microbiome Data Collaborative (Kyrpides et al. 2016) will also be available in the near future.

DOE User Facilities and Enabling Capabilities for Gene Function Discovery and Annotation

Joint Genome Institute (JGI). Provides access to high-throughput sequencing, DNA design and synthesis, metabolomics, and integrated computational analysis. These cutting-edge genomic capabilities enable systems-level investigations of plant and microbial metabolism and interactions, engineering of diverse organisms, integrative analyses of complementary datasets, and establishing experimentally validated links between genotype and phenotype (https://jgi.doe.gov).

DOE Systems Biology Knowledgebase (Kbase). Provides an integrated computational platform for large-scale analyses, combining multiple lines of evidence to model plant and microbial physiology and community dynamics by integrating data and tools in a unified, user-friendly graphical interface. KBase is an open-source, extensible system enabling users to bring their own data and tools together to analyze and share workflows, results, and conclusions (https://kbase.us/).

Environmental Molecular Sciences Laboratory (EMSL). Provides access to advanced proteomic, metabolomic, and transcriptomic capabilities, as well as imaging capabilities. These tools can be used to characterize the structure and dynamics of molecules (e.g., proteins,

Continued on next page

Continued from previous page

metabolites, and natural organic matter); intra- and extracellular components; whole cells (e.g., archaea, bacteria, fungi, and plants); microbial communities (e.g., soil microbiome); and physical, chemical, and biological interactions with the rhizosphere (https://www.emsl.pnl.gov).

BER Structural Biology Beamline Resources. Provide beamline-based experimental capabilities at the DOE synchrotron and neutron facilities, along with user training and support, for probing the structural and functional properties of biological samples ranging from atomic and molecular scales through cellular and tissue scales. These user facilities are operated by the Basic Energy Sciences (BES) program within DOE's Office of Science (https://berstructuralbioportal.org).

National Microbiome Data Collaborative (NMDC). Aims to make data findable, accessible, interoperable, and reusable (FAIR) through the development of platform technologies and a user-friendly, robust, integrated system with expert curation and supporting access to open and transparent data for microbiome data exploration and discovery.

High-Performance Computing and Data Storage Infrastructure. Provides the computational science community with world-class computing and networking capabilities dedicated to breakthrough science and engineering. The Advanced Scientific Computing Research (ASCR) program within DOE's Office of Science supports the following national scientific user facilities:

- **Argonne Leadership Computing Facility** (https://www.alcf.anl.gov)
- **National Energy Research Scientific Computing Center** (https://www.nersc.gov)
- **Oak Ridge Leadership Computing Facility** (https://www.olcf.ornl.gov)
- **EMSL, Molecular Science Computing** (https://www.emsl.pnl.gov/emslweb/capabilities/computing/)
- **Energy Sciences Network** (es.net)

2. Technology Innovations

The last two decades have seen the development and widespread adoption of multiple, systems-level experimental approaches for interrogating biological systems. These technological advancements have increased our understanding of gene function and associated biological processes through the measurement of molecular phenotypes (e.g., proteins, transcripts, metabolites, and molecular interactions) and macrophenotypes (e.g., growth conditions, morphology, and response to stress). However, these approaches typically have been applied only to a limited number of isolated organisms and are not at the throughput necessary to systematically uncover gene function in diverse species across multiple spatial and temporal scales (e.g., protein, activity or pathway, single-cell, and ecosystem levels). Furthermore, current approaches often lack the resolution needed for examining molecular processes at single-cell resolution within heterogeneous environments. While experimental approaches to characterize gene function may never scale equally with genome sequencing, technologies in their current form and application clearly are not sufficient to address the critical questions arising from even the best-studied systems. Therefore, numerous technological gaps must be overcome to enable large-scale gene and protein function determination across taxa (see Table 2.1, p. 8), including the development of novel, groundbreaking technologies. Additionally, concurrent advances in computational tools will be required for integrating these data to derive new insights into gene function (see Chapter 3. Computational Advancements, p. 21).

Need for Scalable Experimental Technologies

For the millions of sequenced genes that do not have a close homolog with an experimentally determined function, limited to no information is available for accurately predicting function. Given the scale of the challenge and relative lack of informative data for most genes, the development of experimental approaches and workflows that (1) provide interpretable insights into gene function at multiple scales and (2) can be applied massively across diverse organisms and gene and protein families is crucial. Some existing and new technologies have high potential for informing gene function at a large scale. Thus, expanding these technologies to U.S. Department of Energy (DOE) Office of Biological and Environmental Research (BER)–relevant taxa at scale and throughput level while complementing them with emerging technologies will provide a foundation for rapidly improving understanding of gene functions.

An illustrative example of a scalable technology that is informative for gene function understanding is phenotypic measurements of genetically modified microorganism libraries. Given the ease of working with some of these systems, large gene-knockout or gene-disruption libraries can be assayed in a single-pot assay with next-generation sequencing. By tagging each genetic modification with a DNA barcode sequence, the abundance or fitness of each mutant can be monitored simply by sequencing the DNA barcodes. One such approach, randomly barcoded transposon site mutagenesis (RB Tn-seq), was applied to 32 bacteria and used to generate millions of gene-phenotype measurements (Price et al. 2018a). The resulting dataset led to the refined annotation of hundreds of misannotated transporters and enzymes and enabled functional classifications for proteins with domains of unknown function (DUFs). Furthermore, many of the phenotypes and inferred functions were conserved among the investigated bacteria, increasing confidence in the results. This example highlights how a single technology can be applied on a massive scale to inform gene function; at the same time, it also illustrates how any single experimental approach is insufficient to completely bridge the gene-to-function gap. In the RB Tn-seq example, most investigated genes did not have a phenotype under the laboratory conditions analyzed. Furthermore, inferring the specific biochemical activity of a protein from a mutant phenotype is often not straightforward.

Table 2.1. Overview of Existing Technologies for Interrogating Biological Systems and Remaining Barriers

Current Technologies	Technological Gaps	Potential Tools
Phenomics		
Growth in culture or artificial environments	• How to do it faster? • How to work with communities?	• Imaging technologies coupled to machine or deep learning • Laboratory ecosystems • Miniaturized growth in droplets
Genomics		
Tn-seq, gene editing, genome-wide association studies (GWAS), heterologous expression, transgenic expression	• How to work with more challenging microbes (e.g., polyploidy, multinucleate, and filamentous)? • How to increase the throughput of heterologous expression and phenotyping?	• Single-cell epigenomics • High-throughput gene assembly • Massively parallel reporter screens
Transcriptomics		
RNA-seq, microarray	• How to work with communities?	• Single-cell RNA-seq • Advances in metatranscriptomics
Proteomics		
Liquid chromatography–mass spectrometry (LC-MS)/MS	• How to dramatically increase sample preparation? • How to dramatically increase the speed, sensitivity, and dynamic range of measurement technologies? • How to obtain more complete characterization of actual cellular state (intact and post-translational modifications)? • How to obtain better quantification without predesigned assays?	• Single-cell proteomics • Ion mobility • Nanopore • Probes • Top-down proteomics • Computational or machine-learned algorithms to better estimate absolute abundance from data
Metabolomics		
Gas chromatography–MS, LC-MS/MS, laser-desorption ionization MS	• How to rapidly determine false discovery rates? How to develop community-accepted standard methods? • How to identify metabolites not in reference libraries, including novel metabolites? • How to increase sensitivity and achieve massive throughput? • How to prepare samples much faster? • How to determine biological activities in high throughput?	• Single-cell metabolomics • Chemical probes • Ion mobility gas-phase separations • Coupling of MS with microfluidics to perform biochemical assays at massive scales

Continued on next page

Continued from previous page

Table 2.1. Overview of Existing Technologies for Interrogating Biological Systems and Remaining Barriers

Current Technologies	Technological Gaps	Potential Tools
Interactomics		
Protein affinity pulldown, yeast two-hybrid	• How to move past exogenous protein expression and cloning to find protein-protein or metabolite interactions? • How to measure classes of protein binding, activity, and pathways?	• Single-cell immunoprecipitations • Faster functional protein expression for structural characterization • Native MS • Activity-based probes • Cryo-electron microscopy • Massively parallel reporter screens

Reducing Technology Barriers

While many technologies exist for determining gene function (e.g., genetic manipulation, *in situ* analyses, and molecular and biochemical assays), significant barriers hinder their application across the multitude of taxa of interest to DOE's mission. As detailed in Chapter 4. Microorganisms, p. 29, and Chapter 5. Plants, p. 41, not all taxa (microbial or plant) are currently amenable to genetic manipulation, greatly hampering functional analyses. Generally, experimental approaches are developed for a specific organism (typically a model organism) or a readily assayable class of proteins (e.g., hydrolases). Unclear is how robust and applicable these approaches are for other organisms or classes of proteins. Thus, existing approaches may require refinement if they are to be applied more broadly. Dedicated research, development, and implementation to reduce these technological barriers would enable leveraging of existing technologies to a larger set of taxa, their biology, and their gene functions (see Table 2.1).

Relatedly, standardization of experimental methods and development of measurement assurance materials such as spike-in controls (for controlling variability in starting materials) will be essential for comparative analyses. For instance, the vast diversity of metabolites, coupled with experimental artifacts, limits interpretation of untargeted metabolomic datasets; as a consequence, the majority of nonredundant compounds remain unannotated (Bowen and Northen 2010; Gertsman and Barshop 2018). Developing standardized methods and expanding curated metabolomic datasets would enable deeper curation of metabolomic datasets and gene function, regardless of where the data are generated. Such datasets would provide important metrics for developing and benchmarking cheminformatic tools. Further, the development and implementation of community–agreed upon standards for data, protocols, and metadata will be crucial for the success of any large-scale determination of gene function (Burgoon 2006; Sumner et al. 2007).

Improving Gene Manipulation Efficiencies and Phenotyping

Gene manipulation technologies such as those enabled by the CRISPR/Cas system have the potential to generate a wealth of targeted modifications in diverse organisms. By measuring the phenotypes of these genetically modified organisms, researchers can gain key insights into a gene's importance to an organism's physiology and ecological interactions (Simo et al. 2014). Indeed, coupling high-throughput genetic approaches with diverse phenotyping capabilities holds great potential for systematically discovering new gene functions across species. The term phenomics encompasses molecular phenotypes (e.g., proteomics and metabolomics) and broader phenotypes at the organismal and ecosystem level (e.g., conditional growth, size, and shape; see Table 2.2, p. 10;

Table 2.2. High-Throughput Genomics and Phenomics for Determining Gene Function

Technology Stage	High-Throughput Genomics	High-Throughput Phenomics
Existing technology	• Mutagenesis approaches • Barcode or amplicon sequencing-based assays • CRISPR-based genome editing • DNA synthesis or cloning coupled with heterologous protein expression • Synthetic biology pathway editing	• Fitness • Gene expression (transcriptomics) • Protein abundance (proteomics) • Metabolomics • Metabolic flux • Reporter constructs • Fluorescent probes • Morphology and macrolevel phenotyping • Protein structure determination • Activity-based proteomics • Plant growth and response in field and natural environments
High-potential new technologies	• Droplet-based transformation screening • Miniaturized cell-free expression pipelines for structural or functional screens • Microfluidics (e.g., flow-through electroporation)	• Droplet-based protein assays • Nanodroplet processing in one pot for trace samples (NanoPOTS) single-cell omics • Microfluidics coupled to mass spectrometry (MS) • Ion-mobility spectrometry • Native (structured) MS • Cryo-electron microscopy (cryo-EM) and micrography • Small-molecule cryo-EM • Small-molecule X-ray diffraction (XRD) • Small-molecule nuclear magnetic resonance (NMR) spectroscopy • New activity and phenotype chemical and imaging probes • Sensors and imaging devices to measure difficult-to-access plant parts, such as roots, meristems, and developing floral structures

Araus et al. 2018). Therefore, increasing the efficiency and accuracy of gene manipulation is of central importance. However, for many species, genetic manipulation is difficult, labor intensive, slow, and genotype dependent. For example, in many microorganisms, including bacteria and fungi, barriers to DNA transformation including recalcitrant cell walls and restriction modification systems limit the application of molecular genetics. In plants, except for two species for which the facile floral dip method has been applied successfully, genetic modification has low-efficiency, genotype-specificity, and large-infrastructure requirements.

For microorganisms, new strategies that broadly increase the efficiency of transformation across taxa, including host-agnostic transformation systems, would enable more efficient gene function determination. For example, developing transformation methods such as microfluidic electroporation or sonoporation, in combination with targeted DNA modifications that

prevent within-host degradation, might reduce the need for species-specific optimizations. These optimized transformation approaches can be coupled with the rapid development of vector systems, including those with broad host ranges, to streamline genetic tool development in diverse microorganisms.

Except for *Arabidopsis thaliana* and *Camelina sativa*, there are two major challenges in generating genetically modified plants: transformation and regeneration of plants from calli. For plants, the morphogenic *Baby boom* and *Wuschel* technology to enhance transformation of grass species (Lowe et al. 2016) has the potential to accelerate plant transformation rates and efficiency. However, although this technology was reported in 2016, as of 2019 it is not yet broadly available in the public domain. Optimizing high-efficiency methodologies not only for gene editing but also for plant transformation of several BER focal species would greatly accelerate gene function determination.

An additional benefit of improved genetic manipulation of diverse species is the increased ability to modify multiple loci within a single organism to examine genetic interactions (epistasis). Given the gene- and pathway-level redundancy encoded in the genomes of many species (plants in particular), the ability to knock out or knock down multiple genes efficiently can reveal shared and unique roles for homologous genes within a single genome. In addition, the analysis of genetic interactions provides important insights into the robustness and connectivity of cellular metabolic and regulatory networks.

High-throughput phenomics (or phenotyping) is the much-needed complement to high-throughput genomics. Elevating phenomic studies so that they are high throughput, taxa independent, and suitable for diverse phenotypes would enable systematic studies of gene function. A major remaining challenge in phenomics is to make the technologies robust, scalable, and affordable so they can be deployed at scale. As with other new technologies, development and use of controlled vocabularies, ontologies, and standard operating procedures will be critical to success, allowing data to be leveraged across studies.

Capturing Molecular Processes at the Single-Cell Level

Studies of transcripts, proteins, and metabolites can be used to dissect interactions between single cells in a tissue or within a microbial community. Regardless, the spatial resolution of such studies is low. At best, investigations on a single plant tissue represent multiple types of cells, thereby producing a coarse-grained view of interactions between cells and the relative heterogeneity of cells of the same type. For example, plant organs such as leaves, roots, and seeds are composed of multiple cell types, while current models of gene regulation, metabolism, and physiology are derived from whole-plant or organ studies (see Fig. 2.1, p. 12), and thus are a chimera of individual cell types. Similarly, many studies on microbial biofilms reflect an amalgamation of the underlying microbes' gene expression and metabolism.

Improving spatial resolution to deconvolute the complexity of the transcriptome, proteome, and metabolome at the single-cell level in a plant organ or complex microbial environment would enable the characterization of genes and metabolites involved in organ-level and organismal interactions (Ryu et al. 2019). For example, knowing that a gene or protein is expressed or a metabolite is present in only a single cell type would greatly improve understanding of the greater function of those biomolecules in that organ or microbial community. Though initial single-cell studies in plants have focused primarily on the transcriptome, the application of single-cell resolution approaches to the epigenome and metabolome would enable more robust, spatially defined modeling of gene regulation. In microorganisms, single-cell omics has the potential to provide critical information about temporal expression patterns and abundance under a variety of environmental conditions, including for poorly culturable or unculturable species. Similarly, measuring the dynamics of genomic and transcriptomic dynamics at the single-cell level in a complex environment such as the plant-microbe interface is essential for uncovering the molecular and genetic bases for these interactions.

Application of methods such as laser capture microdissection and single-cell transcriptomics and epigenomics is already improving the understanding of function

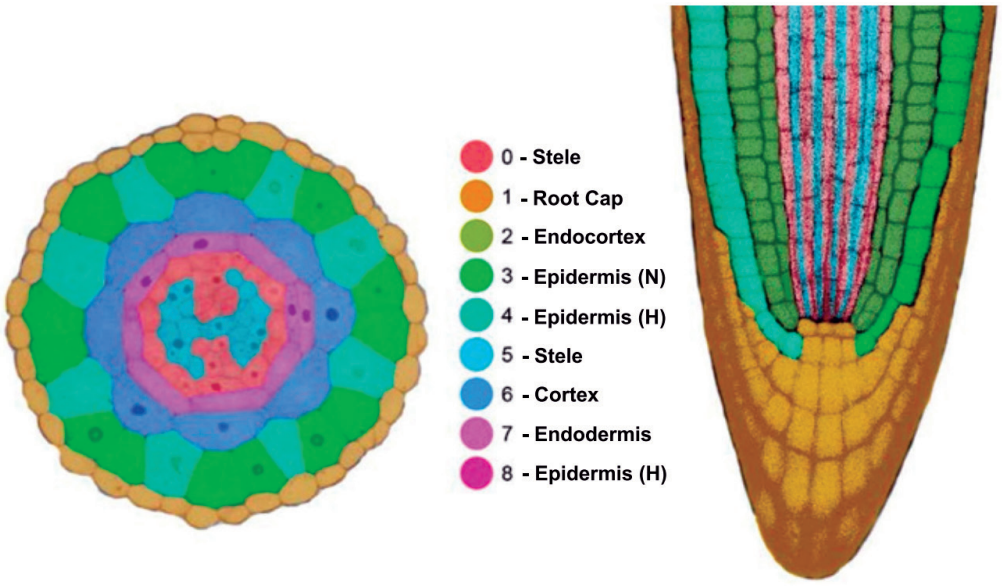


Fig. 2.1. Specific Cell Types in *Arabidopsis* Roots as Revealed Through Single-Cell RNA Sequencing. [Republished with permission of the American Society of Plant Biologists, from Ryu, K. H., et al. 2019. "Single-Cell RNA Sequencing Resolves Molecular Relationships Among Individual Plant Cells," *Plant Physiology* **179**(4), 1444–56. DOI:10.1104/pp.18.01482. Copyright 2019; permission conveyed through Copyright Clearance Center, Inc.]

in individual cells (Clair et al. 2016; Mukherjee et al. 2013). Nonetheless, single-cell transcriptomics has been demonstrated only in *A. thaliana* roots (Ryu et al. 2019) and in simplistic *Physcomitrella* leaves (Yu et al. 2018). Overall, expanding these existing capabilities to multiple taxa, as well as advancing single-cell and spatially resolved proteomics for tissues and metabolomics, would enable more comprehensive and informative insights for understanding gene function at an unprecedented resolution. Recent reports have demonstrated unbiased single eukaryotic cell proteomic analyses (Budnik et al. 2018; Zhu et al. 2018a; Zhu et al. 2018b) and spatially resolved cell types within tissues (Liang et al. 2018). For plants, accurately probing the functions of genes and proteins in their proper physiological context requires methods adapted to isolate single cells from complex plant organs without perturbing the cellular state (e.g., transcriptome, proteome, metabolome, membrane, and chromatin). For microbial systems, improvements are needed in current methodologies such as microscopy, sample manipulation tools, methods to limit sample losses, and highly sensitive multianalyte measurement capabilities to enable high-resolution interrogation of single cells or very small, organized regions. Concomitantly, computational tools, microfluidics, and robotics will be needed to automate the selection and manipulation of single cells to enable sufficient throughput with these analyses. Finally, enabling multiple different omic measurements on the same cell will significantly improve understanding of the interplay of genes, proteins, and metabolites within their cellular context and exponentially amplify knowledge derived from single-cell experiments.

Additionally, experiments with careful sampling can be used with advanced computational means to deconvolute and interpolate results, thereby improving resolution. This approach can be particularly powerful using multimodal measurement capabilities with different resolutions (Buchberger et al. 2018). Hence, improvements in resolution for all modalities that fall short of the ideal may yet move toward the larger goal of enabling the use of other technologies for verification studies.

Targeting Classes of Proteins

In many instances, genes are annotated with a general biochemical function, but they lack information

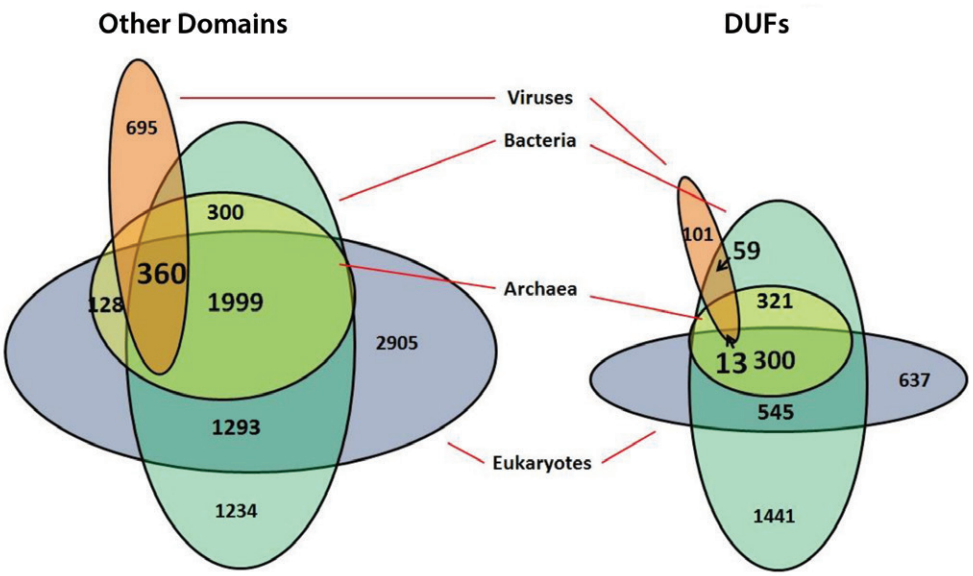


Fig. 2.2. Prevalence of Domains of Unknown Function (DUFs) in Different Organisms. Bacteria typically have more DUFs **(right)** relative to other domains **(left)**. [Reprinted under a Creative Commons license (CC-BY-NC-SA-3.0) from Goodacre, N. F., et al. 2014. "Protein Domains of Unknown Function Are Essential in Bacteria," *mBio* **5**(1), e00744-13. DOI:10.1128/mBio.00744-13.]

about substrate specificity. This limited annotation includes proteins such as transporters, enzymes, and transcriptional regulators, all of which are of keen interest to systems biologists as these proteins can be rapidly incorporated into metabolic and gene regulatory models. Given the known but broad functional categorization of these proteins, their detailed characterization is more amenable to larger-scale screening assays, for example ligand-binding assays for enzymes or transcription factor–DNA binding assays. Many such assays can be scaled with a combination of high-throughput protein production, *in vitro* translation approaches, cell-free biochemical assays, and miniaturized enzyme assays. Importantly, due to the rapid reduction in cost of DNA synthesis, approaches to characterize the specificity of these protein classes can be scaled without growing the plants and microorganisms that encode them, thus opening the door to more accurate annotations of genes from uncultivated microorganisms and for large paralogous protein families in plants.

Efforts to determine gene function can also be focused on a variety of other gene properties. For example, targeted approaches can be applied to dissect the key regulatory functions of small regulatory RNAs in prokaryotes and eukaryotes. Genes involved in the production of secondary metabolites can be readily identified by a number of software tools in microorganisms and plants (Medema et al. 2011); however, the precise structure and biological function(s) of these metabolites are rarely known. Additional prioritization can be based on the available experimental data. For example, genes of unknown function have been identified as essential for viability in bacteria under standard laboratory growth conditions through genome-wide approaches like Tn-seq (Rubin et al. 2015). Likewise, comparative genomics have identified thousands of poorly characterized yet evolutionarily conserved proteins in microbial genomes. Many of these proteins contain conserved DUFs, which can be prioritized for further investigation (see Fig. 2.2, this page; Goodacre et al. 2014).

Improved annotations of transporters, enzymes, and transcriptional regulators would provide a cache of useful functions for strengthening models of metabolism and regulation. Though challenging, achieving

annotation clarity potentially can enable discovery of signatures in transcription factors, transporters, and enzymes that can be used to predict substrate specificity more accurately in newly sequenced genomes. The improved annotation of catabolic and biosynthetic genes will lead to more informed models of microbiome structure and function such as predictions of nutrient cross-feeding and other interspecies dependencies (Price et al. 2018b; Zengler and Zaramela 2018). Furthermore, genes encoding specific classes of proteins can be readily mined for biotechnological purposes; for instance, microbial secondary metabolites include many bioactive compounds such as antibiotics. Additionally, for metabolic engineering, novel enzymes and transporters may be applied to produce desirable molecules of interest in diverse microbial and plant hosts.

Screening proteins for specific biochemical activities is still challenging for a number of reasons: (1) large-scale production of many different proteins is not readily available, (2) large compound libraries for binding assays typically are focused on bioactive compounds and not on many of the ligands of interest to environmental microbiologists and plant biologists, and (3) not all enzymes or proteins are readily amenable to high-throughput assays. In addition, the cost of DNA synthesis has not yet decreased to the point where millions of different genes can be synthesized and their encoded proteins biochemically investigated; as a result, the ability to exploit the wealth of protein sequences derived from metagenomic studies is limited. For secondary metabolites, simply expressing the encoding gene clusters within the native host or in a heterologous host requires optimization (Clevenger et al. 2017). Moreover, determining the structure and biological role(s) of secondary metabolites is largely done on a case-by-case basis, despite the fact that tens of thousands of these gene clusters have been computationally identified in genomes and metagenomes [e.g., SMURF and antiSMASH (Khaldi et al. 2010; Medema et al. 2011)]. Even more challenging is the systematic functional dissection of uncharacterized microbial essential genes and/or DUFs, because there is no overriding biochemical assay that is relevant or applicable to these diverse proteins. For prokaryote proteins, at least, investigators often can use a combination of genomic context and comparative genomics to provide clues to a protein's function, although experimental validation of these predictions often remains slow and laborious. In light of these considerations, opportunities to interrogate specific gene functional categories include enzyme function and transporter screening pipelines, high-throughput approaches for characterizing transcription factors, an integrated computational-experimental approach for characterizing DUFs, and a paradigm for structure-based functional genomics.

Enzyme Function and Transporter Screening Pipelines

Assays of enzymes and transporters with known biochemical function but unknown substrate specificity are generally low throughput and highly tailored relative to the needs of a broader effort to discover and characterize these proteins. Additionally, most enzymatic assays used to determine kinetic parameters are performed under conditions that do not represent cellular conditions (Zotter et al. 2017), such as physiologically relevant concentrations of substrates and allosteric effectors, concentrations of products, and presence of protein subunits.

Many technological improvements and capabilities provide real opportunities for performing assays across a wide variety of enzyme and transporter families with much greater throughput and under more functionally relevant conditions, including (1) reduced cost of DNA synthesis; (2) miniaturized and high-throughput protein production such as for enzyme complexes, enzymes from obligate anaerobic microorganisms, and enzymes with unknown cofactor requirements; (3) technologies that can either measure activities *in vivo* (e.g., activity-based probes) or measure analytes (e.g., mass spectrometry) such as metabolic endproducts and intermediates; (4) development and community access to larger, BER-relevant chemical libraries for rapid ligand screening and as a resource for metabolite standards; and (5) high-throughput enzyme assays for diverse biochemical activities such as those based on microfluidics, droplets, nanopore technologies, and high-throughput phenotyping.

For transporters and enzymes, the integration of these aforementioned capabilities into novel expression and assay platforms will enable the high-throughput discovery of substrates and products, which could revolutionize the precise functional characterization of these proteins. For example, advances in cell-free, *in vitro* protein production (Carlson et al. 2012), coupled with assaying millions of proteins simultaneously in microfluidic devices or in encapsulated droplets (Baccouche et al. 2017), could open the door to the massive-scale characterization of entire protein families in a single experiment. Activity-based protein profiling (ABPP), which uses chemical probes to target enzymes based on their activity toward specific substrates, can be used for functional evaluation of proteins and whole cells in complex mixtures or communities without pre-existing knowledge of the protein identity or structure. A recent ABPP application using a probe selective for β-glucuronidase activity enabled the detection, isolation, and identification of microbes performing this critical function within complex microbiome samples (Whidbey et al. 2019). Importantly, the data generated by these workflows would greatly increase understanding of how small and large sequence differences among related proteins affect their function and lead to improved computational approaches for more precise annotation of these proteins in newly sequenced genomes. Ultimately, these insights will provide a framework for the rational engineering of these proteins for new functionality.

High-Throughput Approaches for Characterizing Transcription Factors

Putative transcription factors can be readily identified in microbial and plant genomes; rarely known, however, are the DNA binding motif(s) of the regulators or their impact on their target genes. The development and application of rapid and scalable approaches to characterize transcriptional factors in diverse microorganisms and plants would greatly improve predictive gene regulatory models, with a number of BER applications. For example, improved plant gene regulatory models can guide new breeding and engineering strategies to improve biomass crops with desired characteristics, such as resilience to environmental perturbations. A number of established and developing technologies can be applied to characterize transcription factors and derive regulatory networks and models. Simply decreasing the cost and increasing the throughput of transcriptomic experiments (RNA-seq) for protein-coding transcripts, antisense transcripts, and small RNAs would be welcome advances. Newer approaches that characterize transcription factor binding sites *in vitro* hold great potential, as these strategies potentially can be automated and systematically applied at a massive scale. One such approach, DNA affinity purification sequencing (DAP-seq; Bartlett et al. 2017), was used to identify binding motifs for 529 transcription factors in *A. thaliana* (O'Malley et al. 2016).

Integrated Computational-Experimental Approach for Characterizing DUFs

Given the conservation of DUF proteins and their demonstrated importance for fitness in microorganisms (Goodacre et al. 2014), a concerted effort to characterize these proteins would have a large impact on gene function understanding across diverse taxa. However, because of their diverse functions, an integrated approach using large-scale experimentation (including genetic or phenotypic data), comparative genomics, and targeted assays to validate predicted functions would likely be necessary. For example, a combination of ligand-binding assays, comparative genomics, enzymology, and genetics was used to characterize members of the DUF1537 protein family as kinases required for the catabolism of four-carbon sugar acids (Zhang et al. 2016).

Paradigm for Structure-Based Functional Genomics

Establishment of a focused paradigm for structure-based functional genomics, using new tools [cryo-electron microscopy (cryo-EM) and *in silico* structure prediction] that have been developed or matured since the end of the Protein Structure Initiative in 2015 (Montelione 2012), together with established structure-function tools, could not only provide important molecular insights into protein function but also an improved reference dataset to accurately predict structures for newly sequenced proteins.

Advancing Molecular Measurements of Proteins

The function of a protein is more than just the translated product of its encoding gene. Proteins often function in transient and stable complexes with other proteins, nucleic acids, and other macromolecules to perform their cellular activities. Many proteins are post-translationally modified, with these modifications providing key functionality. In addition, many enzymes contain one of a number of important cofactors, including metals, which are vital to activities such as redox reactions. Many proteins are also specifically localized based on their functional roles, for example, to the outer membrane or to intracellular membrane-bound organelles (in eukaryotes).

A systematic effort to measure and characterize the impact of protein modifications, cofactors, interactions, and localization would have a transformative impact on the understanding of gene, protein, and organismal function. In addition, such an effort could lead to the development of innovative new proteomic, imaging, structural, and metabolomic approaches. However, a number of challenges must be overcome: (1) low-abundance proteins are difficult to detect; (2) transient-complex formations (such as histidine kinase and response regulator phosphorylation) are difficult to detect; (3) isolating proteins from native species is challenging; (4) measuring protein-protein interactions on a large scale is typically laborious, with these datasets possibly having high rates of false positives; (5) many post-translational modifications are hard to measure including acetylation, serine or tyrosine phosphorylation, and cysteine oxidation; and (6) functional relationships regarding post-translational modifications are often lost with many proteomic technologies.

Methods with Greater Throughput, Sensitivity, and Quantification Accuracy

Improved methods for assaying proteins, including mass spectrometry, crystallography, nuclear magnetic resonance, affinity chromatography or pulldowns, and other measurement capabilities, will facilitate studies within more realistic contexts. For example, assaying proteins purified from the native organism (rather than a heterologous host) more accurately reflects the protein's true nature (including post-translational modifications and cofactors). In addition, recent noteworthy analytical developments have enabled profiling of native protein complexes, including metalloproteins, at the proteome level by mass spectrometry (Skinner et al. 2018).

Development of technologies such as molecular probes, ion mobility separations, and droplet microfluidic assays would provide either disruptive changes or new approaches for parallel measurements on a small scale and in high throughput. This research would address current technology limitations that are low throughput and require large sample sizes (see Fig. 2.3, p. 17). Perhaps the most transformative improvement in technology would be the ability to parallelize multiple omic measurements because, currently, omic datasets are typically measured in separate experiments. By integrating measurements, not only will throughput be increased but the quality of information obtained would be amplified. Integrating these technologies with high-sensitivity, high-resolution mass spectrometry of large, native proteins can accurately define and discover protein-protein and protein-ligand or metal complexes, potentially at the proteome level. Label-free, whole-cell imaging at the nanoscale with cryo-electron tomography to visualize the interactome would provide new insights into protein interactions. Taken together, these advances are anticipated to lead to new functional assignments for genes of unknown function.

Scalable Approaches for Identifying Protein-Protein Interactions

Understanding how protein-protein interactions and their dynamics are affected by localization, post-translational modifications, and allosteric effectors is critical for holistic insights into molecular mechanisms. However, high-throughput protein-protein interaction methods have been fraught with challenges related to methods that produce high rates of false positives and false negatives, in part due to stable and transient interactions requiring different technologies for detection. While two-hybrid and affinity-based methods are still the primary high-throughput methods for

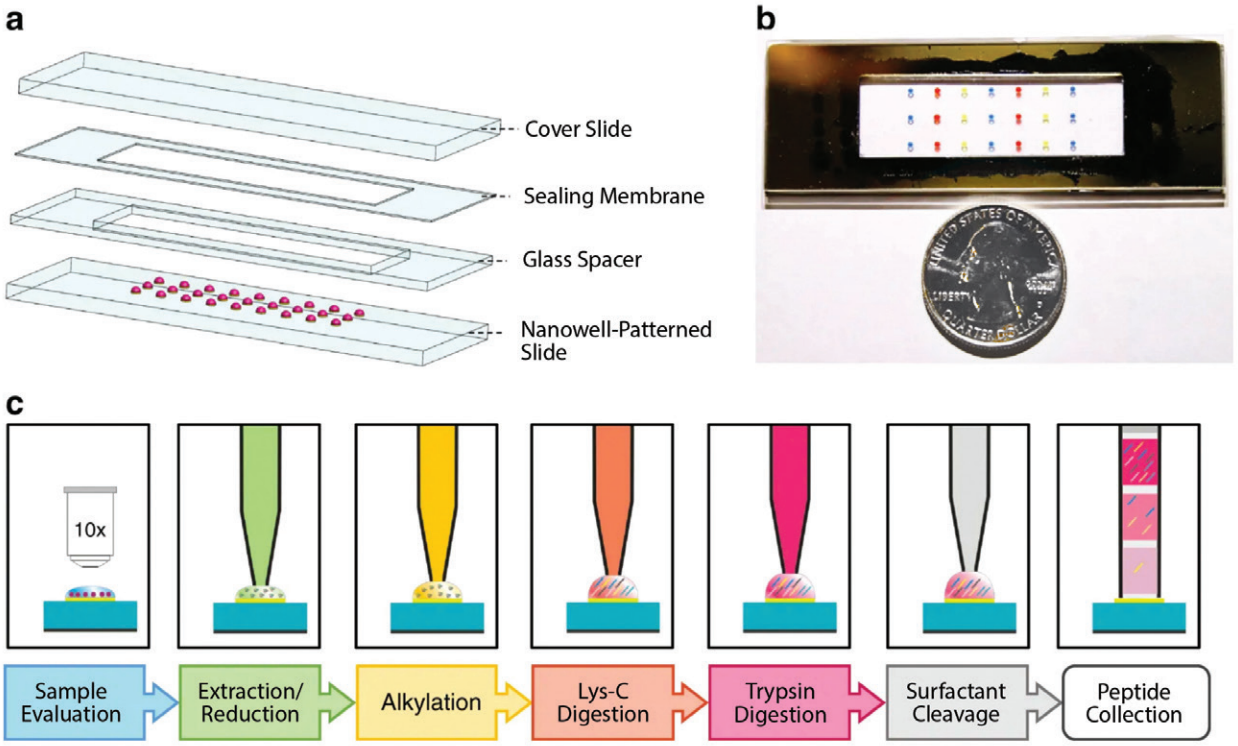


Fig. 2.3. Proteomic Sample Preparation with Nanodroplet Processing in One Pot for Trace Samples (NanoPOTS). **(a)** Schematic drawing and **(b)** photograph showing the nanoPOTS chip with each nanowell filled with 200 nanoliters of colored dye. The cover slide can be removed and resealed for dispensing and incubation. **(c)** One-pot protocol for proteomic sample preparation and capillary-based sample collection. [Reprinted under a Creative Commons license (CC-BY-4.0) from Zhu, Y., et al. 2018b. "Nanodroplet Processing Platform for Deep and Quantitative Proteome Profiling of 10–100 Mammalian Cells," *Nature Communications* **9**(1), 882. DOI:10.1038/s41467-018-03367-w.]

discovery, they both require manipulation of the proteins of interest or complex contexts where expression and regulation may be critical for discovery. Moreover, these approaches can be time consuming, and they also lack the throughput to screen for interactions on a multiproteome scale.

Development of technologies that enable low-cost and high-throughput multiplexed screening for protein-protein interactions would enable improved functional annotation through identification of novel binding partners. One promising avenue is the development of pooled, massively parallel reporter screens that use changes in DNA barcode abundance as a readout for protein-protein interactions (Diaz-Mejia et al. 2018; Diss and Lehner 2018; Schlecht et al. 2017). Because throughput of these assays is tied to exponentially declining DNA sequencing costs, performing one-pot screens between large collections of unannotated proteins and their potential interaction partners may soon become feasible. However, reaching the required scales that test tens to hundreds of millions of protein interactions in parallel requires further improvements in gene assembly, combinatorial library construction, and barcode sequencing data analysis.

For more detailed, pair-wise studies of protein-protein interactions, biolayer interferometry and other similar approaches can be used to provide a robust, label-free method for measuring binding kinetics on the medium-throughput scale. In the case of protein complexes, these studies could also identify the role of the different partners related to function. Another promising, but technically difficult, approach for

identifying interacting proteins is the application of native mass spectrometry (Schachner et al. 2019; Zhou et al. 2018). These methods use the improved mass range of intact proteomics and computational deconvolution of the spectra. Currently, application to the proteome scale has proven to be elusive, but advanced methods of gas separations (Ben-Nissan and Sharon 2018) with mass spectrometry hold promise for scaling these capabilities.

Extending High-Throughput Genetic Approaches to Relevant Ecological Contexts

Both high-throughput and classical "single-gene" investigations of environmental organisms have typically been performed under laboratory conditions that do not accurately reflect the native ecologies of these organisms. Any significant effort to discover gene function across taxa needs to consider these more natural conditions, since it is likely that the functions of a sizable fraction of the uncharacterized genes of environmental organisms can be revealed only through ecologically relevant experimentation. For example, bacterial genes specifically required for colonizing the roots of a biomass plant likely cannot be characterized using standard monoculture laboratory approaches. Using high-throughput genetic approaches within relevant environments and communities will enable the discovery of the functions of genes and their associated metabolites that mediate interspecific interactions and adaptations to environmental conditions. For both microbes and plants, this coupling of ecologically relevant experimental conditions with high-throughput omic approaches has the potential to revolutionize the understanding of genes that mediate organismal interactions and biotic-abiotic interactions *in situ*.

Identifying, creating, and controlling relevant environmental and microbial interactions for the application of high-throughput genetic, phenotypic, and other measurement approaches would be central to tackling this challenge. While initial high-throughput screens are designed to associate genes to a physiological process, validation of these results at the single-gene level using a multitude of approaches will also need to be high throughput to enable rapid determination of gene function. For example, plants are typically grown in diverse environments, and reductionist methods to determine gene function in the laboratory may fail to report gene function if the diversity of these environmental conditions is not assayed.

Integrating Technologies to Scale Gene Function Determination

Certain technologies and methodologies such as (1) time-series proteomics and metabolomics, (2) mutant libraries and metabolomics or microfluidics, and (3) stable isotope probing and proteomics and metabolomics can be synergistic when paired or grouped together as they provide rich datasets that inform the relationships between multiple parameters on a large scale and provide orthogonal validation of findings. Generating and analyzing large-scale datasets in parallel can also promote standardization of data formats. Several existing high-throughput functional genomic technologies can be immediately integrated for functional characterization of unknown genes or proteins (see Fig. 2.4, p. 19). These technologies include a broad array of high-throughput genetic approaches coupled to other functional measures such as targeted CRISPR-based genome editing, DNA synthesis coupled with heterologous protein expression, and synthetic pathway construction. Additional high-throughput approaches that can provide functional information include growth assays, measures of enzyme activities, imaging, and phenotyping technologies.

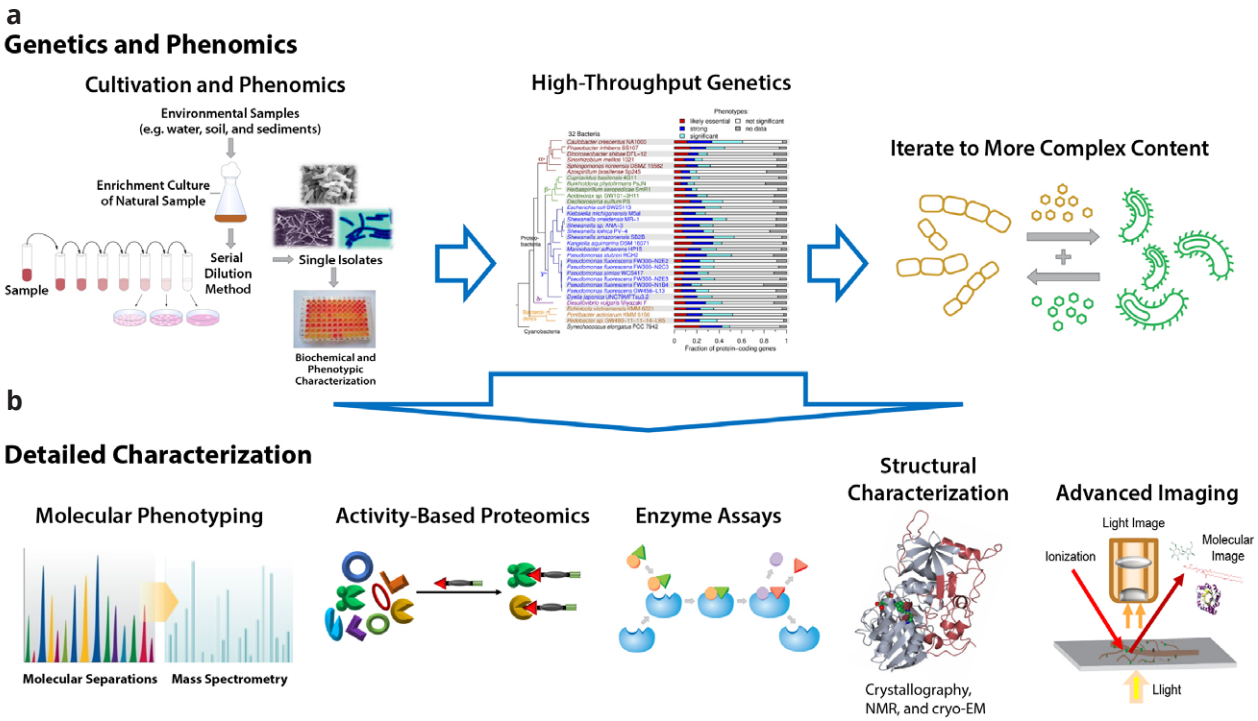


Fig. 2.4. Genetics and Phenomics Integrated with Detailed Characterization Techniques to Rapidly Advance Understanding of Gene Function in Plants and Microbes. A number of existing technologies are ready (or near ready) for immediate deployment for high-throughput determination of gene function, including **(a)** genetic and phenomic approaches from either complex environmental samples, isolated microbial samples, or individual plant accessions (not shown), and **(b)** biochemical and structural approaches to gain deep insight into protein function and activity. Importantly, the genetic and phenomic approaches are increasingly moving away from artificial conditions to conditions more relevant to Department of Energy missions (e.g., natural ecology). In the illustrated workflow, large-scale genetic and phenomic datasets lead to a wealth of phenotypes and hypotheses that are investigated in more depth with enzymatic, biochemical, structural, and proteomic methods. **Key:** cryo-EM, cryo-electron microscopy; NMR, nuclear magnetic resonance. [**Credits: Cultivation and phenomics.** Adapted and reprinted under a Creative Commons license (CC-BY-4.0) from Di Donato, P., et al. 2019. "Exploring Marine Environments for the Identification of Extremophiles and Their Enzymes for Sustainable and Green Bioprocesses," *Sustainability* **11**(1), 149. DOI:10.3390/su11010149. **High-throughput genetics.** Reprinted by permission from Springer Nature from Price, M. N., et al. 2018a. "Mutant Phenotypes for Thousands of Bacterial Genes of Unknown Function," *Nature* **557**(7706), 503–509. Copyright 2018. DOI:10.1038/s41586-018-0124-0. **Molecular phenotyping.** Courtesy Joshua Adkins, Pacific Northwest National Laboratory (PNNL). **Activity-based proteomics.** Courtesy Joshua Adkins and Aaron Wright, PNNL. **Structural characterization.** Reprinted with permission from Kohler, A. C., et al. 2017. "Structure of Aryl *O*-Demethylase Offers Molecular Insight into a Catalytic Tyrosine-Dependent Mechanism," *Proceedings of the National Academy of Sciences USA* **114**(16), E3205–214. DOI:10.1073/pnas.1619263114. **Advanced imaging.** Courtesy Joshua Adkins and Chris Anderton, PNNL.]

3. Computational Advancements

Discovering new gene functions and accurately transferring these annotations across taxa are both experimental and computational challenges. Though the development of new technologies for generating high-throughput data is vital, it is impractical for individual researchers to manually examine data to infer gene function on a case-by-case basis for all genes. Additionally, because experimentally investigating all sequenced genes is not feasible in the near term, transferring annotations from experimentally characterized genes to newly sequenced genomes will remain necessary. Consequently, advances in computational tools are urgently required to automate the inference of gene function from diverse data and interactive databases that maintain and propagate accurate gene annotations across taxa. Moreover, comparative genomics and phylogenetics can be used to generate hypotheses about gene function that can be tested with targeted experimentation.

Computationally Driven Gene Function Discovery

A number of approaches have been developed to infer gene function from sequence data alone. Classically, the basic local alignment search tool (BLAST) can be used to identify homologs of a query protein sequence against a reference database. However, the output of a BLAST search often does not direct insight into the function of the query protein. A slightly different use of BLAST, termed PaperBLAST, mitigates this limitation by identifying homologs for a query gene that have been discussed in published papers and putatively have experimental data regarding their function (Price and Arkin 2017). Phylogenomic approaches examine the evolutionary relationships among proteins to more accurately infer function, including at a resolution to differentiate neofunctionalization events (Eisen 1998). Another approach, termed sequence similarity networks (SSNs), can be used to generate clusters of related proteins at different user-defined thresholds; SSNs are useful for splitting protein families into subfamilies with conserved functions or to identify subfamilies with novel functions (see Fig. 3.1, p. 22; Atkinson et al. 2009). More broadly, a number of comparative genomic approaches that leverage the wealth of sequence data have been developed to predict gene function, including those that take into account gene co-occurrence across species or the chromosomal clustering of functionally related genes in prokaryotes to aid in gene function prediction. To demonstrate the utility of these approaches, SSNs and genome context have been coupled successfully to downstream experimental pipelines, including targeted genetics and biochemistry, to validate a number of previously unknown enzymes (Atkinson et al. 2009).

Solving the structure of a protein, identifying its ligand(s), and elucidating its dynamics all provide invaluable insights into understanding protein function. While solving the structures of many proteins may soon be feasible, for example with advances in cryo-electron microscopy (cryo-EM; Bai et al. 2015), some aspects of protein structure and activity can be predicted from sequence alone. Foremost, a number of tools are available to predict protein structure from sequence, including for protein families for which no structures are currently available (Ovchinnikov et al. 2015). Additional methods can be used to predict ligands for proteins based on analysis of protein structures (Calhoun et al. 2018). All these approaches highlight existing computational tools and methods that can be immediately applied to generate specific gene function hypotheses for individual proteins and protein families, which then can be tested by targeted experimentation. Nevertheless, to date these tools have not proven to be sufficiently predictive across enough protein families that they can be successfully implemented at scale across diverse taxa.

Databases and Knowledgebases of Gene Annotations

A number of resources have been developed for generating and storing gene annotations. For some model organisms, such as *Escherichia coli* (Keseler et al. 2017),

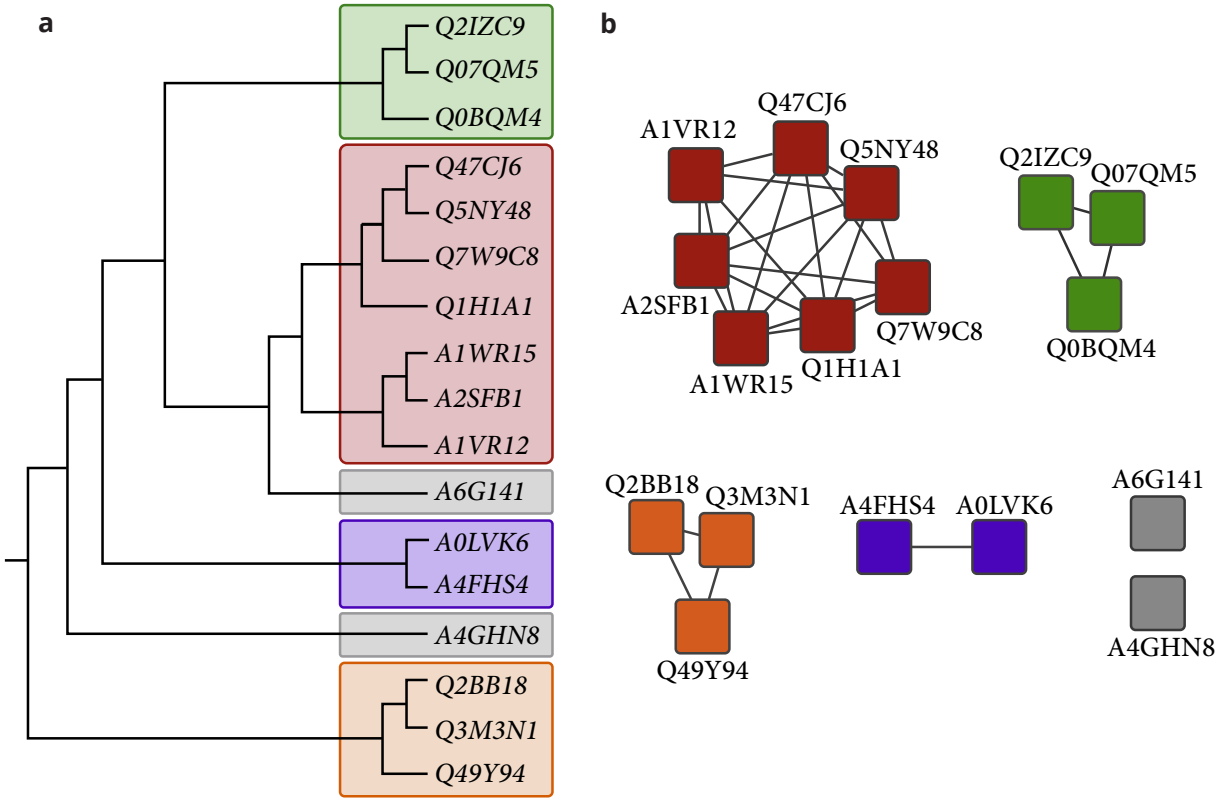


Fig. 3.1. Comparison of a Phylogenetic Tree and a Sequence Similarity Network (SSN) for the Same Proteins. Both **(a)** phylogenetic trees and **(b)** SSNs can be used to infer sequence-function relationships among proteins, although SSNs can be computationally easier to compute. [Reprinted from Gerlt, J. A., et al. 2015. "Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A Web Tool for Generating Protein Sequence Similarity Networks," *Biochimica et Biophysica Acta–Proteins and Proteomics* **1854**(8), 1019–37. DOI:10.1016/j.bbapap.2015.04.015. Copyright 2015, with permission from Elsevier.]

Saccharomyces cerevisiae (Cherry et al. 2012), and *Arabidopsis thaliana* (Berardini et al. 2015), high-quality databases are available with manual curation by experts of the literature. However, the development of databases of equal quality for the wealth of nonmodel species under investigation is not practical. Similarly, the Swiss-Prot database within UniProt contains many manually curated entries for protein function, but these curated entries represent only a small fraction of the total number of sequenced proteins contained in UniProt. Moreover, nearly all of these proteins are annotated using purely computational approaches. As described in the previous section, many of these computational annotations of gene function should be viewed skeptically, especially in the absence of experimental evidence for a close homolog. Also unclear is whether these resources will be able to incorporate the rapid influx of sequenced genes (e.g., from bacterial isolates and metagenomes), many of which ultimately may have experimental data regarding their functions (e.g., by heterologous protein expression and *in vitro* biochemical assays).

Including these existing databases and annotation pipelines in any future efforts is important to elucidate gene function on a large scale across taxa, but in their current form these resources may not be ideal for this grand task. First, gene annotations derived from major annotation pipelines often do not agree, although combining evidence from multiple sources can be used to improve the process (Griesemer et al. 2018). Second, though model organism knowledgebases are

invaluable, extending these annotations across both related and distant taxa is not easy. Third, these databases do not do an effective job of rapidly updating recent advances in gene function understanding into their annotations, largely a result of the scale of data being generated. Relatedly, updating and improving gene function annotations within existing databases are not trivial for researchers, for example when the researcher was not the original depositor of the genome sequence.

A potential strategy for overcoming these limitations is establishing an improved database of proteins with experimentally determined and known functions. Work in this area could include basic text mining to guide curators to papers and more advanced text mining to automatically extract gene annotations from the literature, possibly still requiring manual review. In addition, crowd-sourcing approaches could be used to engage the community to input experimentally verified functions into such a database, particularly if the process for including a new annotation is straightforward. Even in such a system though, how gene function is defined remains an issue. For metabolic enzymes, there is a clear definition, but for other classes of proteins, the definition is more complicated. For example, considering a transcriptional regulator, the annotation could include the effector molecule, the DNA binding motif(s), and the effect on expression of target genes.

Computational Framework for Discovery of New Gene Functions and Accurate Annotation

One opportunity for discovering new gene functions and rapidly increasing the quality of genome annotations is a proper computational infrastructure, with community coordination and appropriate experimental data (see Fig. 3.2, this page). This platform could integrate seamlessly with (or be a part of) existing U.S. Department of Energy (DOE) computational resources, including the Systems Biology Knowledgebase (KBase), Joint Genome Institute (JGI), Environmental Molecular Sciences Laboratory (EMSL), and National Energy Research Scientific Computing Center (NERSC), as well as the National Center for Biotechnology Information (NCBI) supported by the National Institutes of Health, Protein Data Bank (PDB) managed by the Research Collaboratory for Structural Bioinformatics, and the UniProt database.

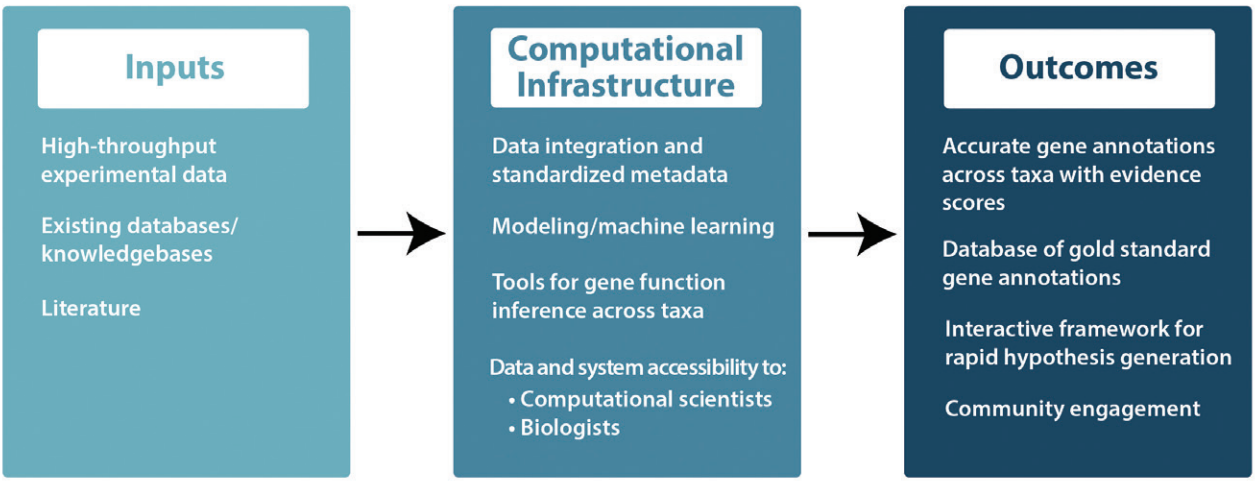


Fig. 3.2. Computational Infrastructure for Gene Function Discovery. An integrated computational system for accurate gene function annotation, propagation, and discovery could incorporate inputs including diverse experimental data, existing databases and knowledgebases, and prior literature. These data would feed into a single (or multiple connected) system(s) readily accessible by both computational researchers and scientists without a strong computational background. Within this framework, data and prior information would be used to accurately annotate genes across taxa, and the community could use experimental data and comparative genomic tools to rapidly develop hypotheses into the functions of individual genes, possibly testable by targeted experimentation. System outcomes include active community engagement and high-confidence gene annotations across taxa. [Courtesy Adam Deutschbauer, Lawrence Berkeley National Laboratory]

However, most of these existing resources cannot incorporate diverse data types into their infrastructure nor partially automate the inference of gene function from experimental data, as likely will be required to systematically discover gene functions across many species. Some databases such as the Pathosystems Resource Integration Center (PATRIC) have tools for analyzing omic data (Wattam et al. 2017), but these are primarily stand-alone tools, not designed to automatically infer new gene functions and improve annotations. Other resources such as KBase offer increasing capability for integrating diverse omic data (Arkin et al. 2018), but the tools for inferring gene function based on these measurements are not fully developed.

An integration between computation and experimentation would enable new gene functions to be discovered, and erroneous annotations would be identified and corrected across taxa. Additionally, a high-quality, up-to-date database of gene annotations and supporting evidence would become available, as well as a set of fully integrated computational tools and models to support the inference and evaluation of new gene annotations. Fully integrated annotation evidence sources would include sequence similarity; protein domain or motif information; expression levels linked to condition; co-expression; transcription start sites; ortholog, paralog, and homolog mappings; synteny and conserved synteny relationships; pan-genome information; variant information [e.g., single-nucleotide polymorphisms (SNPs) and insertion or deletion of bases (indel)]; mutant resources; gene family trees; and chromatin states. Further, such computational systems would offer the ability to traverse different versions of the annotations, with a unified set of ontologies and identifiers to traverse species.

Computational platforms for discovering new gene functions and accurately inferring gene annotations from new experimental data and prior knowledge face a number of challenges that include: (1) maintaining homology maps for all proteins in real time; (2) maintaining taxonomy maps for all species in real time; (3) updating and maintaining versioning on all functional annotations over time; (4) automatically mining the literature for new annotations; (5) integrating diverse omic data to discover new annotations in a semiautomated manner; (6) identifying and correcting erroneous annotations and generally reconciling conflicting annotations; (7) mapping between the many existing disparate annotation ontologies; and (8) engaging and accommodating biocurators and the scientific community in the gene function discovery process. Although these are all significant challenges, technology, methods, algorithms, and computational approaches potentially can be applied to overcome them.

Infrastructure Requirements for Integrating Diverse Omic Data

As already discussed, multiple omic technologies are required to accurately determine gene function across scales. As such, large-scale efforts to improve gene annotations across taxa will result in the generation of numerous diverse data types, which will include, but not be limited to, all the classic omic types used today. For new data types, standard formats and repositories are likely not currently available and would need to be developed. Furthermore, new computational resources would need to support the rapid storage, query, and cross-connection of all these data types, as well as embody a wide range of tools and algorithms for integrating these data.

Data Accessibility

First and foremost, computational infrastructures that support gene function discovery and the propagation of accurate gene annotations across taxa should be user friendly for both biologists and data scientists. The data should be intuitive and easy to query, visualize, browse, upload, and download. Data also should be available in a standardized and easily accessible format for computational algorithms and models to consume. All data should be accessible (and searchable) via programmatic application programming interfaces (APIs) for both onsite and offsite use.

Data Requirements

Data should be comparable. Ideally, a common set of unique identifiers would be available for all entities (e.g., genes and proteins) referenced in the data. For example, all proteins could include UniRef identifiers whenever possible (Suzek et al. 2007). In addition,

protein sequences can provide a universal link between resources, because the sequence can be coupled to tools to find identical or nearly identical sequences in other databases. Moreover, the experimental conditions related to all data should be fully specified in sufficient detail to inform the user whether different samples come from the same or similar experiments. Data should be accountable, meaning all experimental protocols, algorithms, tools, and transformations that generated the data should be fully documented. All protocols should be stored in centralized public repositories, such as protocols.io (https://www.protocols.io; Kindler et al. 2016). In many cases, data with temporal and spatial distributions will be required, and the computational infrastructures should support this requirement. Generally, data should conform to the findable, accessible, interoperable, and reusable (FAIR) principles (Wilkinson et al. 2016).

Extensibility Requirements

Experimentalists will continue to produce new data types due to the emergence of novel experimental methodologies and technologies, and the analysis and integration of these data types will require the accelerated development and assimilation of innovative tools and algorithms. Thus, the computational infrastructures should support the facile creation of new data types and integration of new tools. The KBase platform (Arkin et al. 2018) provides a roadmap for this principle, as this system is designed to support the management, organization, sharing, discovery, and interconnection of diverse data, with emerging support for rapid extension to new data types.

Database Links

Numerous databases already exist that contain gene annotations and raw datasets that will be useful for future efforts in gene function discovery [e.g., UniProt (Magrane and UniProt Consortium 2010), RefSeq (Magrane and UniProt Consortium 2010; Maglott 2000), PATRIC (Wattam et al. 2018), KBase (Arkin et al. 2018), Integrated Microbial Genomes and Microbiomes (IMG/M; Chen et al. 2019), ProteomeXchange (www.proteomexchange.org) and Pride for proteomics (https://www.ebi.ac.uk/pride/archive/), and MetabolomeXchange for metabolites (www.metabolomexchange.org/site/)]. As much as possible, data from these sites should be imported and displayed directly for all users in any new computational system, in addition to providing linkouts to external data sources.

Gaps in Experimental Data

Some computational teams are already developing algorithms, tools, and data integrations to support genome annotation, such as PathwayTools (Karp et al. 2016) and RAST (Aziz et al. 2008). However, these efforts are currently limited by the availability of data to support their computational inferences. An increase in experimental data of many diverse types is necessary to improve annotations across taxa. Several particularly pertinent examples include: (1) biochemical functional assays of proteins with kinetic parameters; (2) gene-specific, or genotype-specific, population-scale phenotypic data, especially for plants; (3) new protein structures that fill gaps in current structural data; (4) knowledge of regulatory network machinery including transcription factors and their binding sites; (5) increased functional characterization of protein families with no characterized relatives; and (6) identification of functional residues in proteins, which can be used to distinguish enzyme families and correct misannotated proteins, as recently illustrated for the MetA and MetX families involved in methionine biosynthesis (Bastard et al. 2017).

Strategies and Data Sources for Evaluating Confidence in Gene Functional Annotation

Multiple computational methods are available to aid in discovering and testing the veracity of new gene annotations, including pathway and network modeling. Models can help (1) identify a specific subset of gene annotations that should be re-evaluated by a human expert (Reed et al. 2006) and (2) identify and fill gaps in annotated metabolic networks (Kumar et al. 2007). Metabolic models are also a useful tool for studying the consistency of annotations and providing feedback on annotations (e.g., false positives or false negatives; Aziz et al. 2008; Kumar and Maranas 2009). Thus, these models could be a key part of the initial

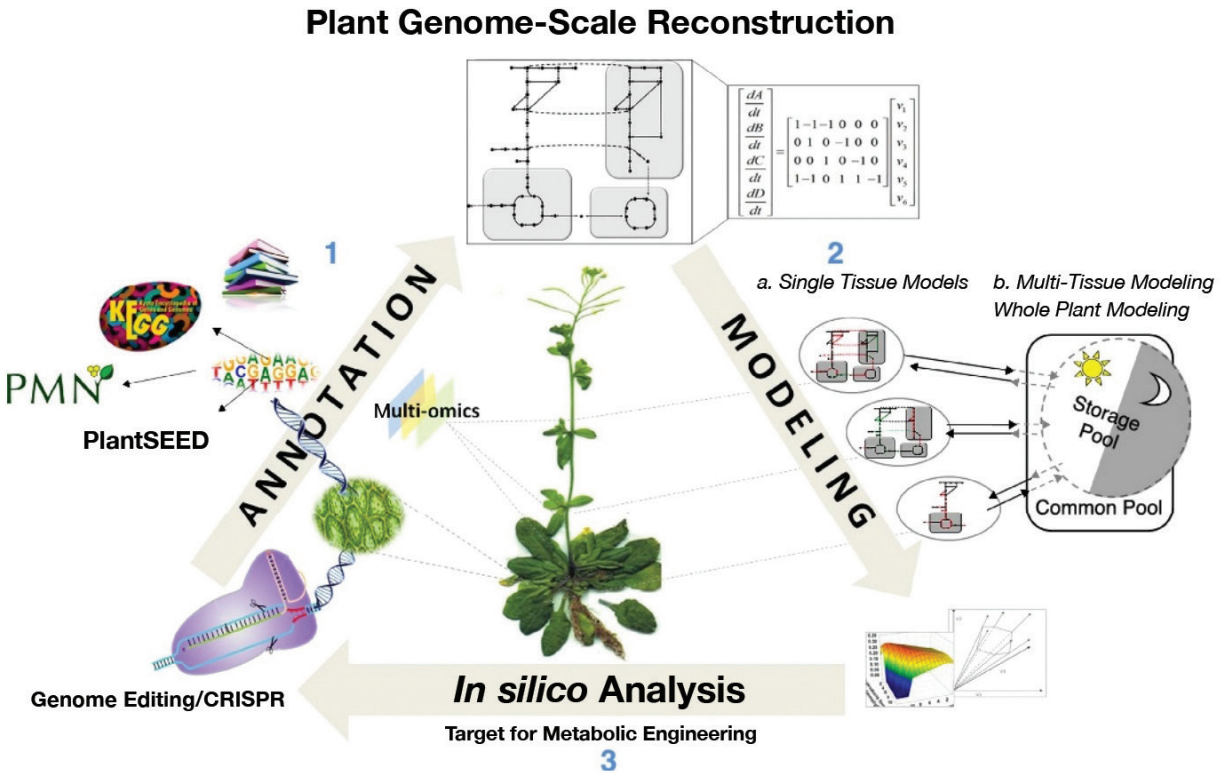


Fig. 3.3. Use of Plant Genome-Scale Reconstructions in the Concept of Plant Systems Biology. Step 1: A plant genome-scale reconstruction is developed using genome annotation. The network is extracted from public metabolic databases and curated based on experimental evidences and best biochemical knowledge. **Step 2:** Model implementation and multi-omic integration; **(a)** the model is implemented and validated to simulate key physiological scenarios of single tissues or **(b)** the multi-tissue model framework is implemented to simulate metabolism at the multi-tissue or whole-plant level, considering the diurnal cycle. **Step 3:** *In silico* analyses are used to build biological knowledge or to test a new experimental hypothesis, including targets for metabolic engineering studies. [Reprinted from Dal'Molin, C. G. O., and L. K. Nielsen. 2018. "Plant Genome-Scale Reconstruction: From Single Cell to Multi-Tissue Modelling and Omics Analyses," *Current Opinion in Biotechnology* **49**, 42–48. DOI:10.1016/j.copbio.2017.07.009, with permission from Elsevier.]

ground-truthing process for genome annotations. Likewise, gene regulatory network (GRN) models are useful tools for identifying putative regulatory factors and their target genes (Gaudinier et al. 2018). The outputs of complex analytical workflows, including explainable artificial intelligence (AI)–based approaches, also can be represented as GRNs (Telenti et al. 2018).

Metabolic and regulatory models can also be used to integrate data across biological scales to reveal emergent phenotype predictions (e.g., gene essentiality), which can be compared with experimental data (Monk et al. 2017; see Fig. 3.3, this page). Integrative and multiscale modeling can be achieved by combining existing data and models to build virtual crops (Marshall-Colon et al. 2017). More broadly, integrated networks derived from multiple, independent datasets across many omic layers can be constructed to represent a complex biological system. Such networks can then be mined in various ways to determine biological narratives centered on specific topics of interest, including gene function discovery. Such a computational infrastructure can also be used to identify multiple lines of evidence for annotation inference or hypothesis generation for targeted experimental confirmation.

In addition to modeling, the accuracy of gene function annotations can be inferred from a number of other strategies. First, the agreement of multiple annotation sources can be leveraged to build more-accurate predictions (Griesemer et al. 2018). In another approach, entire protein families can be constructed, refined,

and tested for functional and taxonomic consistency (Meyer et al. 2009; Seaver et al. 2018; Suzek et al. 2007). Diverse omic data can be leveraged to increase the confidence of gene annotations, for example, when a mutant phenotype is available for a gene of interest (Price et al. 2018a). Existing literature can be mined for annotation evidence from the gene of interest or a homolog (Price and Arkin 2017). However, functional annotations, and the experimental evidence supporting these annotations, are not always available in databases in neatly digitized formats. Often these data can be found only in publications or their supplemental information. The process of manually mining the data and translating the information to digital format is laborious and error prone. Fortunately, text mining through the use of natural language processing techniques is a potentially powerful tool for extracting data from the literature (Krallinger et al. 2010). It is also an essential tool, given the vast scale of current unmined literature content that cannot be handled through human curation alone. However, the success of text-mining methods relies to a substantial degree on standards for journal article content. Authors must be required to provide information on the proteins examined, methods employed, and confidence of results using a controlled, but extendable vocabulary.

All these aforementioned approaches can be used to compute an aggregate confidence in each distinct functional assignment for a gene. A key observation is that if confidence levels can be computed and provided, then even "low-confidence" information can be retained in the database to the community's benefit (Griesemer et al. 2018). Although confidence scores can be computed based on the available experimental data, supporting literature, consistency analysis, and model-based assessment, other potential sources of confidence also exist. Extremely unlikely is the development of a single set of confidence inputs, weights, and criteria that will serve all genes, genomes, and applications. Each individual user seeking annotations will value data types differently. Thus, ensuring that annotation confidence schemes are highly extensible and user customizable is important. Furthermore, if users contribute an annotation directly, then those users should be able to indicate the confidence levels on these annotations. The implementation of this confidence system could start with a simple goal, such as enabling users to indicate which datatypes they want to include (high confidence) or exclude (no confidence) when building annotations.

Community Engagement

Given the scale of the gene-to-function challenge and its impact on all of biology, community engagement is vital in any systematic initiative to decipher gene function and accurately annotate genes across taxa. Also important is ensuring that an integrated master set of annotations is maintained for each protein in a database, in particular for experimentally validated gene functions. This database could be synced with external data sources, performing routine and documented version updates that the community can reference through a version number. Further, text mining could be employed as a strategy to update annotations from emerging literature. Finally, users should be able to curate functions without requiring that a consensus first be reached with the broad research community before alternative potential functions of a gene can be explored by an individual user. To satisfy all these requirements, systems that support gene function discovery and gene annotation across taxa may need to embody many of the features of version control systems: (1) a version identifier assigned to every set of functional assignments as they evolve over time; (2) ability of individual users to branch and eventually merge different annotation sets; (3) ability of users to submit "pull requests" to merge proposed annotations into existing databases; and (4) a mechanism to resolve conflicting annotations when they arise. The community should be able to rapidly explore all branches of the annotation repository to quickly view all alternative annotations being proposed for a given protein or protein family. Users should be able to receive credit or recognition for their gene function discovery and annotation efforts, and this credit system requires careful thought; experimental biologists and curators must be incentivized to load data and annotations into the system. Finally, having fields for "the function is not" and "negative results" from experiments (linked to a user for cross-validation) would augment knowledge of gene function. Most of

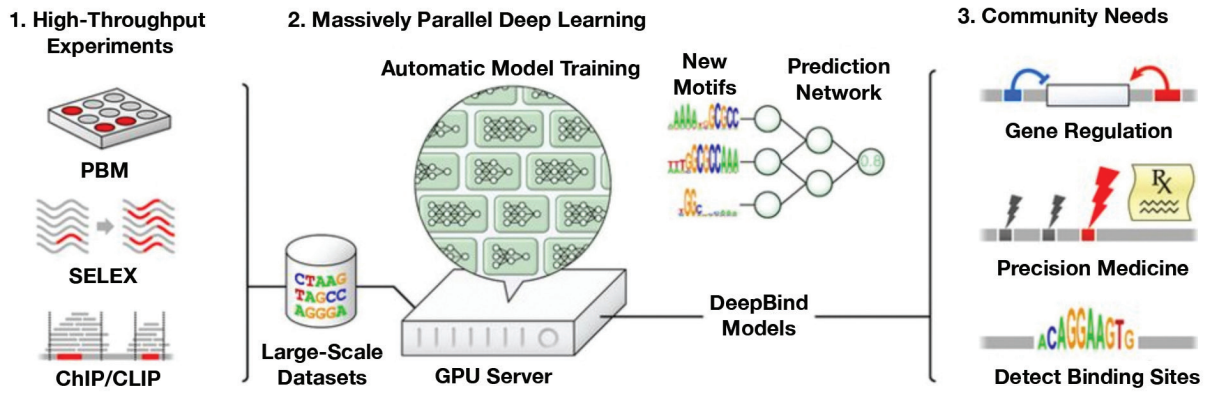


Fig. 3.4. Machine Learning to Understand Protein Function. (1) The sequence specificities of DNA- and RNA-binding proteins can now be measured by several types of high-throughput assay, including protein binding microarray (PBM), systematic evolution of ligands by exponential enrichment (SELEX), and chromatin immunoprecipitation (ChIP)-seq and crosslinking and immunoprecipitation (CLIP)-seq techniques. **(2)** DeepBind models capture these binding specificities from raw sequence data by jointly discovering new sequence motifs along with rules for combining them into a predictive binding score. Graphics processing units (GPUs) are used to automatically train high-quality models, with expert tuning allowed but not required. **(3)** The resulting DeepBind models can then be used to identify binding sites in test sequences and to score the effects of novel mutations. [Reprinted by permission from Springer Nature from Alipanahi, B., et al. 2015. "Predicting the Sequence Specificities of DNA- and RNA-Binding Proteins by Deep Learning," *Nature Biotechnology* **33**(8), 831–38. DOI:10.1038/nbt.3300.]

these described features are not currently supported by existing annotation platforms.

Potential for Gene Function Discovery by High-Performance Computing and New Algorithms

New algorithms are constantly emerging in the area of gene function discovery and genome annotation, and the ever-increasing availability of new sequence and experimental omic data ensures that these algorithms must remain dynamic to this rapidly changing landscape. Since many of these new analysis approaches will be benchmarked against well-studied protein families or model organisms, these "gold standard" datasets must remain updated with the latest literature and omic data, because often new experimental approaches are piloted on model species. Additionally, engaging and potentially incentivizing computational scientists to develop new algorithms for accurate gene annotations are important. One example is the Critical Assessment of Function effort, a community-wide competition to develop algorithms for predicting gene function against a benchmarked dataset (Jiang et al. 2016).

Machine learning is one potential approach for inferring gene function from diverse omic data and propagating gene annotations across taxa (see Fig. 3.4, this page). Machine-learning approaches are already useful where appropriate datasets are present and will continue to offer increasingly powerful opportunities to make novel discoveries and improve annotations into the future. For example, explainable-AI approaches will enable capture of higher-order interactions (including epistatic and pleiotropic relationships) for a better view of the combinatorial interactions responsible for cellular function and their role in complex, emergent properties and phenotypes of an organism. In some model organisms and gene families, enough data are already present to start making this an effective strategy; however, researchers familiar with applied machine-learning methods (including, but not limited to, the wider computational biology community) typically are not experts in the biology of these species or gene families, especially for currently available plant-related datasets. Bringing experimental biologists and computational data scientists together is important to bridge these knowledge and expertise gaps. At a minimum, experimental data need to be made available in a consistent, easily-accessible format to facilitate their use with machine-learning algorithms.

4. Microorganisms

Achieving sustainable solutions to energy and environmental challenges requires a mechanistic understanding of microbial physiology and ecology (Alivisatos et al. 2015; Blaser et al. 2016). The incredible genetic diversity of microorganisms—defined herein as bacteria, archaea, fungi, protists, and other single-cell eukaryotes—provides a vast array of metabolic and ecological activities that can be discovered and used in a wide range of applications. For example, they are primary drivers in global nutrient cycles and contribute significantly to the transformation of environmental toxins. Microorganisms also interact with higher eukaryotes such as plants, providing fitness advantages to these hosts. Additionally, they can be exploited for use as cellular factories, turning organic and inorganic carbon into a plethora of bioproducts (see Fig. 4.1, this page), including fuels, fine chemicals, and possibly even industrial detoxification agents.

However, this vast genetic diversity in microorganisms also presents a huge challenge for harnessing these activities because only a fraction of their massive functional diversity has been studied experimentally in the laboratory. For this reason, there are millions of sequenced genes from microorganisms in public databases with vague or uninformative annotations.

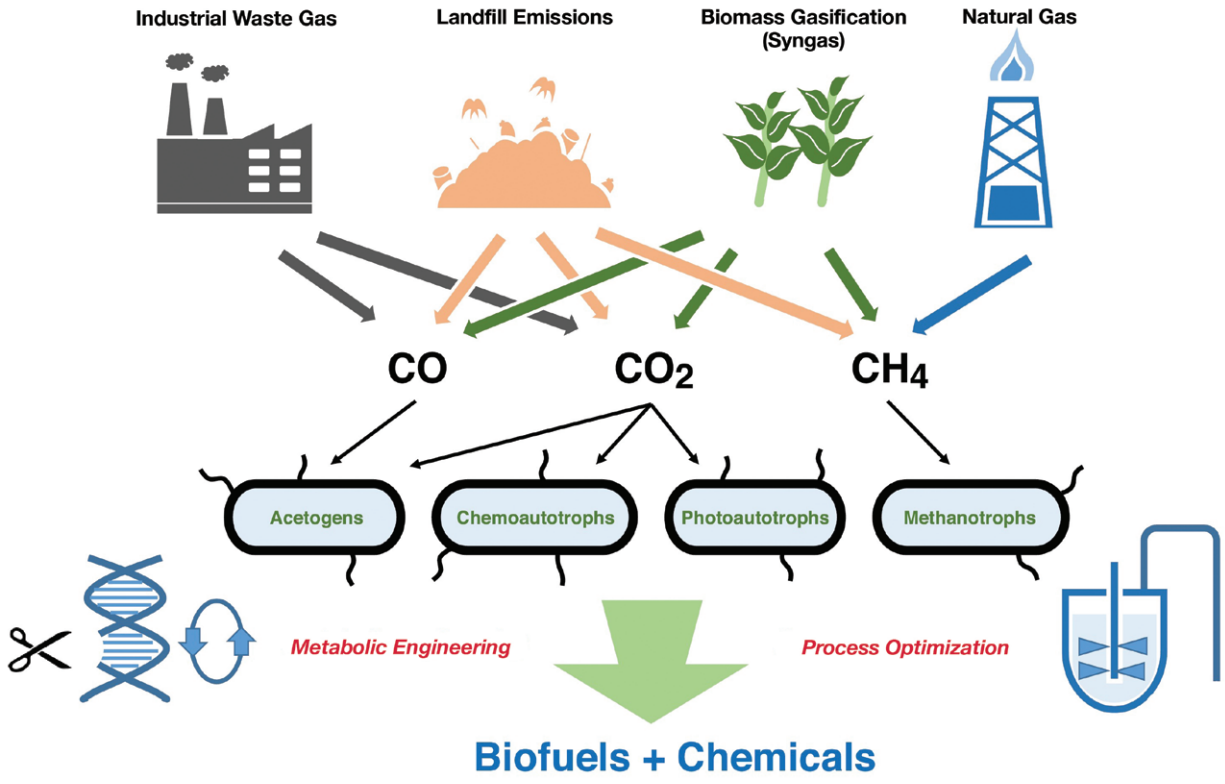


Fig. 4.1. Sustainable Conversion of Waste Gases to Biofuels and Chemicals Through a Combination of Industrial Process Optimization and Genetic Engineering Approaches. Key: CO, carbon monoxide; CO_2, carbon dioxide; CH_4, methane. [Reprinted under a Creative Commons license (CC-BY-NC-ND 4.0) from Humphreys, C. M., and N. P. Minton. 2018. "Advances in Metabolic Engineering in the Microbial Production of Fuels and Chemicals from C1 Gas," *Current Opinion in Biotechnology* **50**, 174–81. DOI: 10.1016/j.copbio.2017.12.023.]

Even in model organisms such as *Escherichia coli*, *Bacillus subtilis*, and *Saccharomyces cerevisiae*, there remain hundreds to thousands of poorly understood genes. Predictably, harnessing the beneficial properties of microorganisms, as well as microbiomes (communities of microorganisms), requires a foundational understanding of the "parts" (i.e., genes, metabolites, and proteins) that make up these systems with sufficient detail to accurately predict overall processes.

Additional challenges complicate the elucidation of gene function in microorganisms. First, sequencing of DNA from diverse environments including soil has revealed a large number of microorganisms that have not been cultivated successfully in isolation in the laboratory (see Fig. 4.2, p. 31). Therefore, investigating the physiology of most environmental bacteria as monocultures is probably not possible. Second, even among culturable microorganisms, there are no universal experimental tools or molecular genetic approaches that can be applied readily across all (or even most) organisms. Indeed, many of the advanced experimental genomic tools developed in microorganisms like *E. coli* and *S. cerevisiae* are not readily portable to other microorganisms.

Despite these challenges, microorganisms by their very nature offer unique advantages for systematically discovering gene function relative to multicellular eukaryotes, including plants. First, given their small size, experimentation with microorganisms is less costly and more readily amenable to the laboratory automation that is increasingly available to individual investigators. Second, many functional assays with microorganisms can be multiplexed, for example with libraries of genetic variants. Third, the genomic co-localization of functionally related genes in prokaryotes facilitates comparative genomic approaches for gene function prediction that can be validated with experimentation. Lastly, the determination of high-quality genome sequences, even from fungi and uncultivated prokaryotes, is less of an issue compared to that with plants.

Target Microorganisms

Microorganisms are often easier to study compared to multicellular eukaryotes, but their great diversity poses unique challenges that may require some degree of prioritization for gene function determination. On the other hand, for a larger-scale effort to characterize genes of unknown function, exploring a diversity of microorganisms can ensure an abundant source of variety to capture novel genes. To maximize the utility of generated data and improved gene annotations, efforts could be coupled to other projects funded by the U.S. Department of Energy's (DOE) Office of Biological and Environmental Research (BER), such as by prioritizing potentially new hosts for metabolic engineering or a specific microbiome. At the same time, model microorganisms will continue to play a key role, particularly for new experimental tool development and advanced systems biology, as well as for developing computational strategies to infer gene function based on diverse omic data.

The impact of gene function discovery from a diverse phylogenetic range of microorganisms would (1) ensure the development of standardized workflows that can be applied to new microbes, (2) discover gene functions that are too distant in sequence to be accurately and entirely predicted by principles of homology, and (3) identify and correct a large fraction of misannotations within public sequence databases. A systematic effort to characterize the genomes of a single microbiome, particularly one associated with a host, would have the added benefits of providing an ecological context to the findings.

Numerous challenges must be overcome to enable systematic investigation of gene function in diverse microorganisms. Foremost, assessing, cultivating, and dissecting gene function in a phylogenetically diverse range of microorganisms are not trivial. Public stock centers contain large numbers of type strains, but moving hundreds to thousands of these isolates into an experiment-based gene function annotation pipeline would be costly and labor intensive. In addition, there are not many examples of extensive culture collections (i.e., bacteria, archaea, and fungi) from a single microbiome or environment. Relatedly, many microorganisms, such as obligate anaerobes and autotrophs, require specialized expertise and equipment for cultivation, and currently developed experimental

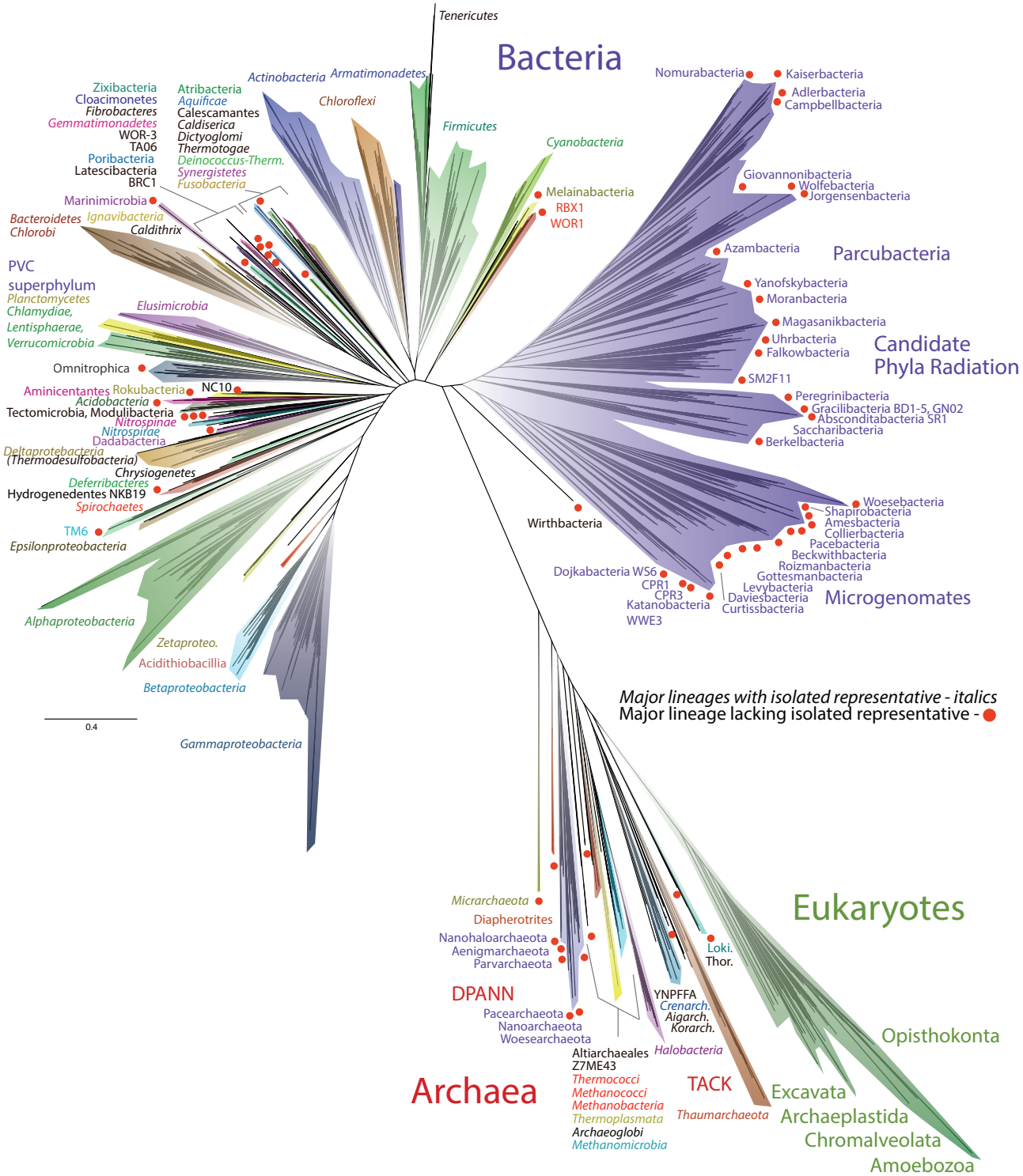


Fig. 4.2. Tree of Life from Unbiased Metagenomic Sequencing. Many of these clades have no cultured representatives. [Reprinted under a Creative Commons license (CC-BY-4.0) from Hug, L. A., et al. 2016. "A New View of the Tree of Life," *Nature Microbiology* **1**(16048). DOI:10.1038/NMICROBIOL.2016.48.]

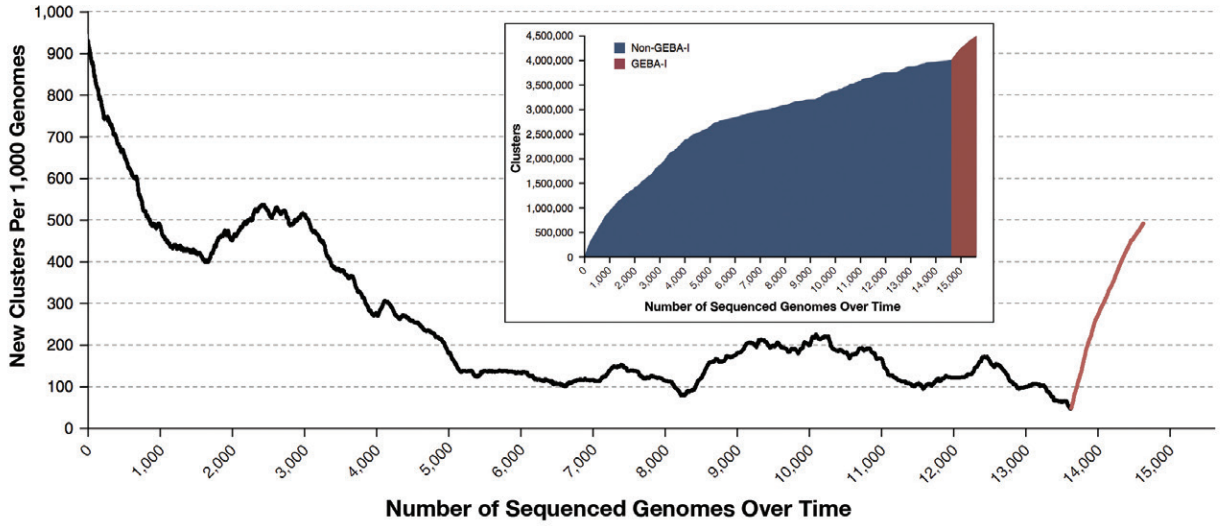


Fig. 4.3. Increase in New Bacterial Protein Families. The increase at the end resulted from the sequencing of new clades of bacteria from the *Genomic Encyclopedia of Bacteria and Archaea* (GEBA-I) project. [Reprinted under a Creative Commons license (CC-BY-4.0) from Mukherjee, S., et al. 2018. "1,003 Reference Genomes of Bacterial and Archaeal Isolates Expand Coverage of the Tree of Life," *Nature Biotechnology* **35**(7), 676–83. DOI:10.1038/nbt.3886.]

tools (e.g., gene manipulation approaches) are unlikely to work in most microorganisms. Despite these challenges, opportunities for gene function determination in microorganisms include diversity, gene function discovery from a microbiome, microbial abundance and activity in DOE-relevant environments, a pan-genome focus, and new tool development.

Microbial Diversity

To capture the greatest novelty of protein sequence resources, a systematic assessment of gene function could be performed in the broadest range of culturable microorganisms, including bacteria, archaea, and fungi. Similar themed efforts have been undertaken by the *Genomic Encyclopedia of Bacteria and Archaea* (GEBA; Wu et al. 2009) and 1,000 Fungal Genomes projects (Grigoriev et al. 2014). In these projects, whole-genome sequencing of cultured microorganisms from across the microbial tree of life led to a substantial increase in the number of sequenced protein families (see Fig. 4.3, this page). A similar broad-scope effort to elucidate the functions of newly sequenced protein families would offer added benefits, such as the availability of sequenced genomes, strain distribution through public repositories, and fostering broad community interest.

Gene Function Discovery from a Microbiome

Many microbiomes of BER relevance (e.g., soil, groundwater, sediment, and rhizosphere) consist of a wide range of bacteria, archaea, phages, and fungi. A systematic investigation of gene function from members of a single microbiome would, therefore, meet the diversity criteria outlined in the previous section, but also provide additional key benefits: (1) investigations into microbial physiology and interactions, including those with a eukaryotic host, can be performed under laboratory conditions that closely mimic native environments (see next section, Microbial Abundance and Activity in DOE-Relevant Environments, p. 33); and (2) the resulting data can be used to demonstrate the utility of improved gene annotations on an ecosystem-level process of BER interest, such as predicting carbon flow in the rhizosphere under changing conditions. Potential microbiomes for functional dissection include the soil and rhizosphere microbiomes of a bioenergy crop (see Fig. 4.4, p. 33). For example, there exists a large culture

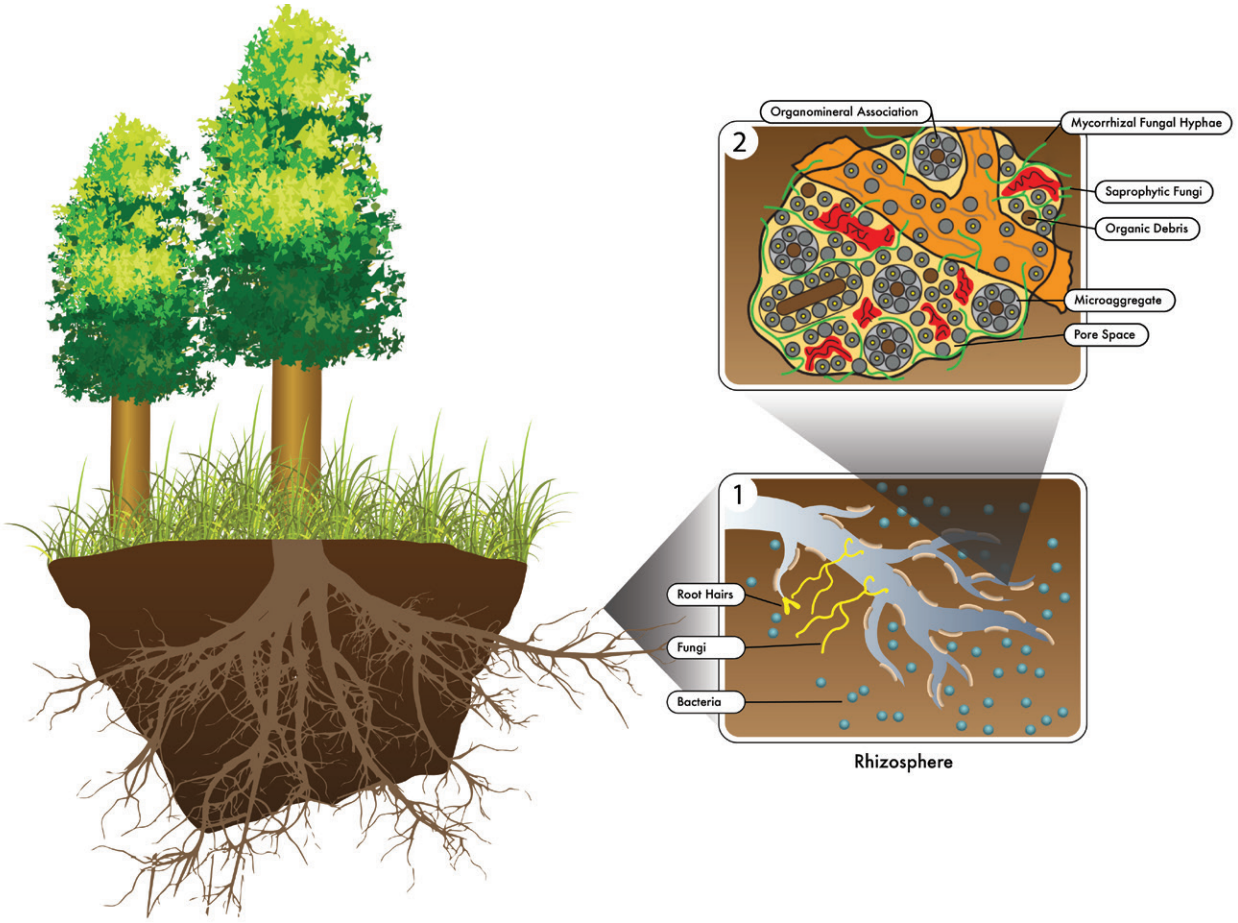


Fig. 4.4. Soil and Rhizosphere Microbiomes. (1) Plant roots and their associated microbial communities (in the rhizosphere) interact for mutual benefit. Microbial communities acquire sugars and other materials from plant roots and assist plant hosts with nutrient and water acquisition, provide pathogen defense, and mediate transformation of root tissues into soil organic matter. **(2)** From this plant-microbe association, soils are further developed into groups of soil particles (i.e., aggregates) that consist of plant and fungal debris; microbially derived organomineral associations; mycorrhizal fungal hyphae; organic matter colonized by saprophytic fungi, clay, and polysaccharides; and other pore-space chemicals. [Image modified from U.S. DOE 2017. Soil aggregate illustration **(2)** modified from Jastrow and Miller 1998.]

collection of bacteria (~2,700 taxa) and fungi (~1,400 taxa) for the poplar microbiome (Blair et al. 2018) as part of a DOE-funded effort.

Microbial Abundance and Activity in DOE-Relevant Environments

Unbiased shotgun metagenomic sequencing has been performed in a number of diverse environments, and, increasingly, studies are identifying the active fraction of microorganisms (e.g., those that respond to external stimuli). The availability of these datasets through comparative analysis portals such as those at DOE's Joint Genome Institute enables meta-analyses of microbiomes across multiple environments. To maximize the utility of gene function discovery and improved annotations relevant to multiple DOE-funded efforts, prioritization based on highly abundant or active microorganisms in diverse environments could be performed. Ideally, representative species of these commonly abundant or active microorganisms would be available in cultivation.

Pan-Genome Focus

The pan-genome represents all the genes from strains of a particular species (Mira et al. 2010). The genes shared

by all strains are referred to as the "core" genome, while those present only in a subset of strains are referred to as the "accessory" genome. In many species of microorganisms, especially bacteria, the accessory genome can be substantially larger than the core genome. A systematic assessment of a pan-genome's gene function can generate important insights into the specialized metabolic and physiological adaptations of closely related microorganisms to different environmental niches.

New Tool Development

New experimental and computational tools for gene function assignment are needed. Typically, in microorganisms, new tools are developed in species such as *E. coli* and *S. cerevisiae* that already have an experimental toolkit and some degree of prior investigation. One primary advantage of tool development in an established system is the ability to benchmark new datasets and analyses rapidly against other omic datasets and literature. As a side benefit, the development of new approaches in classic models can help close the gap of unannotated genes within these exemplar species (Sévin et al. 2016). Importantly, development of new tools likely will continue in model microorganisms, so these approaches should be developed to enable their rapid application, at low cost, to a multitude of other microorganisms.

Moving Experimental Tools from Model to Nonmodel Microorganisms

A number of approaches have been developed for large-scale gene function characterization in microorganisms. The most established involves the systematic genetic modification of a microorganism and subsequent analysis of the phenotypic properties of the mutant strains under different conditions. Genes with similar phenotypes in an organism tend to have related functions (Deutschbauer et al. 2011; Nichols et al. 2011). In addition, the identification of a phenotype for a gene in a specific condition(s) can provide important information about the gene's physiological role (Price et al. 2018a). For some microorganisms, ordered collections of genetically modified organisms such as gene-deletion strains (i.e., in *E. coli*, *B. subtilis*, *S. cerevisiae*, and *Neurospora crassa*), ordered transposon mutant libraries (i.e., in *Shewanella oneidensis* and *Desulfovibrio* spp), or overexpression strains enable genome-wide screens for diverse activities. For example, single mutants can be assayed in high throughput on microplates by metabolomics to identify novel transporters and enzymes (Baran et al. 2013). In another example, the overexpression of each protein enables large-scale, untargeted metabolomics to link specific genes to previously unknown enzymatic functions (Sévin et al. 2016).

Large mixtures of genetically modified strains also can be assayed in parallel for novel phenotypes by next-generation sequencing methods. The primary advantage of these pooled approaches is the throughput enabled by the measurement of thousands of phenotypes in a single-pot reaction. The primary disadvantage is the difficulty in assaying "in trans" phenotypes in a mixed population, such as secreted enzymes and interspecies interactions. In bacteria and yeast, transposon site sequencing (Tn-seq) enables the simultaneous tracking of mutants in a pooled assay (van Opijnen et al. 2009). By adding randomized DNA barcodes to the transposon, these libraries can be readily assayed across many hundreds of conditions (Wetmore et al. 2015), thus enabling gene function inference through cross-species inference. Another example involves high-efficiency recombineering approaches that enable the assessment of precise, user-defined genetic modifications across entire genomes, including gain-of and loss-of-function alleles and combinations of mutants (Warner et al. 2010). More recently, strategies based on gene editing technology have been applied to analyze the consequences of reducing expression in hundreds to thousands of bacterial genes in parallel by a sequence-specific approach, including in *E. coli* (Rousset et al. 2018) and *B. subtilis* (Peters et al. 2016). Similar strategies to interrogate large pools of genetically defined mutants are also well established in multiple yeast species (Coradetti et al. 2018; Giaever et al. 2002; see Fig. 4.5, p. 35).

Although these genetic strategies are powerful, they generally have been developed in well-established model microorganisms and are not often readily applicable to diverse microbial species. Indeed, many

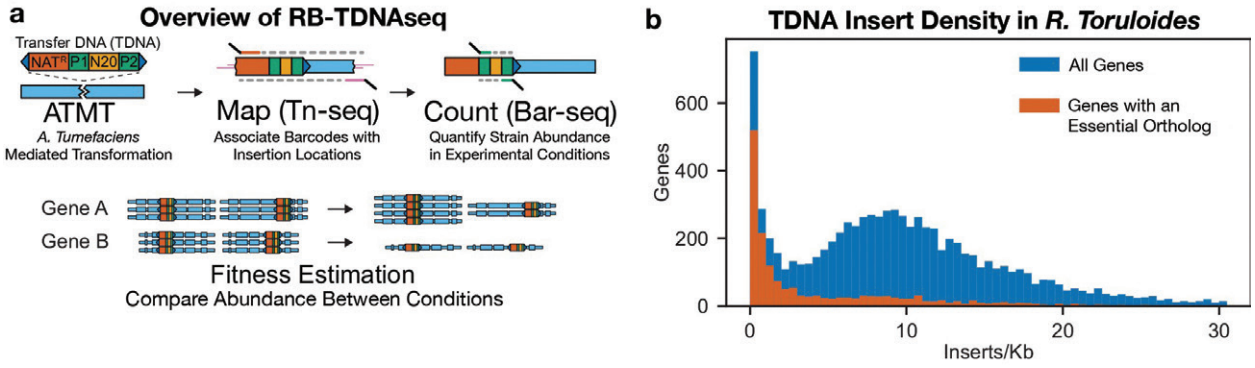


Fig. 4.5. Assaying Mutant Populations of *Rhodosporidium toruloides*. (a) General strategy of RB-TDNAseq, a method for high-throughput genetics. **(b)** Histogram of insert density in coding regions for all genes and genes with essential orthologs. **Key:** ATMT, *Agrobacterium tumefaciens*-mediated transformation; Bar-seq, barcode analysis by sequencing; Tn-seq, transposon sequencing. [Reprinted under a Creative Commons license (CC0) from Coradetti, S. T., et al. 2018. "Functional Genomics of Lipid Metabolism in the Oleaginous Yeast *Rhodosporidium toruloides*," *eLife* **7**, e32110. DOI:10.7554/eLife.32110.]

genetic features such as promoters, drug resistance markers, and plasmids are highly specific to a particular microbial clade. However, some recent work suggests that broad-range or even universal genetic parts may be possible (Rantasalo et al. 2018).

Other omic approaches relevant to gene function inference can be performed systematically in the absence of genetic modification, although the utility of these approaches for characterizing the functions of unknown genes is less established compared to those using genetics. These methods include large-scale transcriptomics, under the realization that genes with similar expression patterns across all or a subset of conditions are more likely to have similar functions. Additionally, in some instances, the genes induced in a particular condition are important to the adaptation of a microorganism to that condition (Price et al. 2013). A similar rationale can be applied to shotgun proteomics. Untargeted metabolomics can be performed systematically in diverse microorganisms, although turning these data into inferences of gene function remains largely nascent.

The rapid adoption of existing high-throughput genetic strategies to new microorganisms will enable the construction of large datasets for gene function inference in diverse taxa (Price et al. 2018a). However, a number of challenges need to be overcome. For instance, genetic tools developed in model organisms are rarely applicable in their current form to new species. Obstacles often exist when applying molecular genetic tools in a new microorganism, including (1) low-transformation efficiency; (2) host-specific barriers to foreign DNA such as restriction-modification and CRISPR systems; (3) lack of functional genetic parts including promoters, replicons, and reporters; (4) protein toxicity, such as in recombineering and CRISPR/Cas-based approaches; (5) lack of selectable and counterselectable markers; and (6) challenges with organismal physiology, for example the multinucleate nature of many filamentous fungi. Computationally, turning large-scale phenotypic data from one or more microorganisms into improved or new gene annotations is not automated, but rather done largely on a case-by-case basis. Furthermore, the process by which improved or new gene annotations based on automated or manual inferences from experimental data are displayed and propagated in existing databases is unclear (see Chapter 3. Computational Advancements, p. 21). Possible research directions in this area include accelerated genetic tool development, advances in transformation, and nongrowth-based assays.

Accelerated Genetic Tool Development

Strategies are needed to rapidly construct and test genetic tools against a large panel of nonmodel microorganisms. Given the low cost of DNA synthesis,

approaches that use *in silico* design and combinatorial assembly of custom plasmids are attractive. In addition, approaches that simultaneously test the efficiency of multiple genetic systems in parallel can accelerate the identification of a working system (Liu et al. 2018; Peters et al. 2019). Finally, there is a need to develop conjugative systems for transferring DNA into a multitude of bacteria and fungi. For example, the range of DNA transfer from *Agrobacterium* species is extremely broad, and the transfer can be harnessed to mutagenize a diverse set of eukaryotic microorganisms (Coradetti et al. 2018). Overall, the microbiology community would greatly benefit from rapid approaches to bring "on board" a multitude of new microorganisms, including the development of one or more of the following: targeted gene deletions or insertions, CRISPR-based gene editing, CRISPR interference, random insertional mutagenesis, stable replicable vectors, and characterized genetic parts for controlled heterologous gene expression.

Advances in Transformation

Simply getting DNA into a microorganism is often a significant barrier to genetic manipulation. Innovative approaches to introduce nucleic acids into microorganisms are needed, including the rapid assessment of electroporation parameters (Garcia et al. 2016), overcoming host defense systems that degrade foreign DNA (Weigele and Raleigh 2016; Zhang et al. 2012), new nanotechnology-based methods (Cunningham et al. 2018), new chemical delivery methods, and new conjugation strategies.

Nongrowth-Based Assays

Most large-scale genetic assays, especially those performed with pools of genetically modified strains, measure growth (or fitness) under different laboratory-based conditions. However, growth is only one of a number of phenotypes to which a gene can contribute. Consequently, developing large-scale approaches is important for phenotyping genetic variants for cellular morphology, intracellular metabolite abundance, protein localization, secondary metabolite production, and metabolite transformations. One example of expanding high-throughput genetics to a nongrowth phenotype is lipid accumulation using buoyant density and fluorescence-activated cell sorting (FACS) coupled with next-generation sequencing (Coradetti et al. 2018).

Determining Gene Function in Natural Contexts

Microorganisms and their genes have evolved within the context of their natural environments, including associations with other organisms and interactions with a range of dynamically changing environmental parameters. Therefore, simply limiting the characterization of gene function to pure culture, laboratory-based assays in flasks and microplates will limit the functional dissection of genes important for ecosystem-relevant processes of BER interest. These genes include those that participate in microbial interactions, microbe-abiotic (e.g., soil) interactions, and microbe-plant interactions (see Fig. 4.6, p. 37; Zhalnina et al. 2018).

Large-scale genetic surveys of mutants in diverse bacteria have revealed that many gene deletions or disruptions do not have a measurable phenotype under standard laboratory conditions (Price et al. 2018a). In some instances, this lack may be due to genetic redundancy (see next section, Genetic Redundancy and Functionally Distinguishing Paralogs, p. 38), while in many others it could be due to the unnatural conditions in which the assays were performed. Assaying microorganisms and their genes under more realistic, ecologically relevant conditions will not only identify phenotypes (and functions) for these genes (Cole et al. 2017), but also provide key insights into the ecosystem-level activities and functions of diverse microorganisms within the environment.

Precise recapitulations of natural environments within the laboratory are not only challenging, but also may result in inherently more complicated and lower-throughput experimentation. For example, a more realistic experiment may include, among other things, many more microorganisms, low levels of substrates, slow growth, and low temperatures. Other aspects of the environment, such as nonuniform and dynamic geochemical and physical conditions, are also difficult to recreate in the laboratory. To accelerate

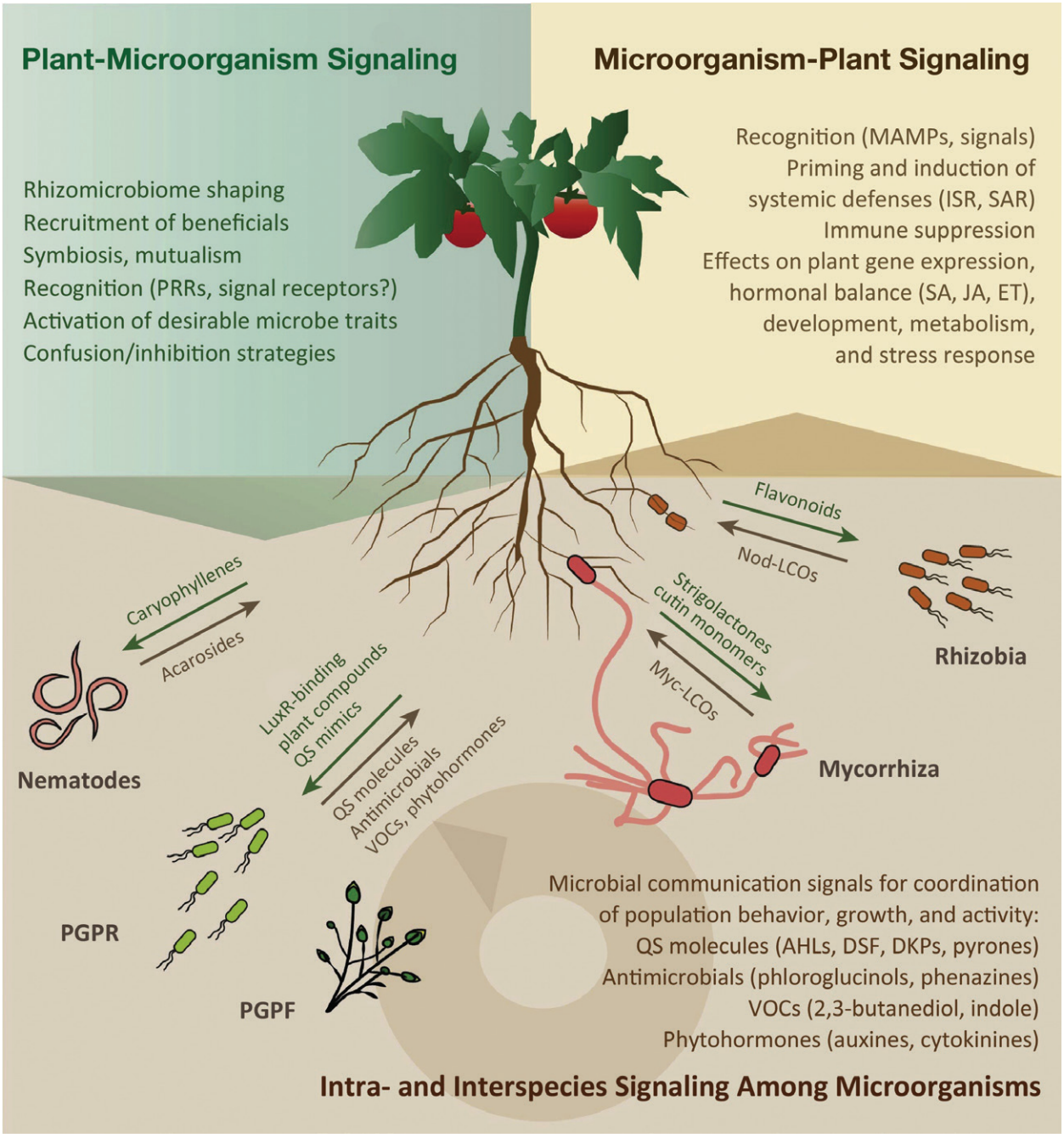


Fig. 4.6. Known Molecules and Events Involved in Intra- and Interspecies Signaling Among Microorganisms and Interkingdom Signaling Between Microorganisms and Plants in the Rhizosphere. Key: AHL, acyl homoserine lactone; DKP, diketopiperazine; DSF, diffusible signal factor; ET, ethylene; ISR, induced systemic resistance; JA, jasmonate; LuxR, response regulator protein; MAMP, microbe-associated molecular pattern; Myc-LCOs, mycorhizzal lipochitinoligosaccharides; Nod-LCOs, nodular LCOs; PGPF, plant growth-promoting fungus; PGPR, plant growth-promoting rhizobacteria; PRR, pattern recognition receptor; QS, quorum sensing; SA, salicylic acid; SAR, system-acquired resistance; VOC, volatile organic compound. [Reprinted from Venturi, V., and C. Keel. 2016. "Signaling in the Rhizosphere," *Trends in Plant Science* **21**(3), 187–98. DOI:10.1016/j.tplants.2016.01.005, with permission from Elsevier.]

gene function discovery within natural ecological contexts, a number of research avenues can be explored, such as constructing realistic ecosystems in the laboratory and effectively using *in situ* omic data.

Construction of Realistic Ecosystems in the Laboratory

The scale of the gene function determination challenge gives importance to rapidly recreating aspects of diverse environments in high throughput. One effort under way is the development of fabricated ecosystems, or EcoFABs (Gao et al. 2018; Sasse et al. 2019), using microfluidics and three-dimensional printing technologies to rapidly and at low cost construct soil and rhizosphere environments in the laboratory. EcoFABs in conjunction with bacteria, phage, fungi, and plants can be used to investigate genes underlying key biotic and abiotic interactions.

Effective Use of In Situ *Omic Data*

Genes for functional characterization can be prioritized using *in situ* information being gained from metagenomic analysis coupled with temporal monitoring. For instance, stable isotope incorporation can provide powerful insights into *in situ* activities coupling gene or protein expression and metabolites. These studies can lead to targeted screens and drive science questions regarding the importance of genes and their expression level to the geophysicochemical responses of microorganisms within natural environments.

Genetic Redundancy and Functionally Distinguishing Paralogs

Microbial genomes often contain multiple genes or entire pathways that perform the same biochemical or physiological function. This genetic redundancy is prevalent among different protein classes including enzymes, transporters, and efflux systems. In other instances, the interconnections and complexity of biological networks (e.g., metabolic and gene regulatory) enable a degree of robustness within biological systems, whereby the knockout of any single gene can be compensated for via alternative paths. Additionally, microbial genomes contain paralogs, or a pair of evolutionarily related genes, which may or may not have undergone functional differentiation. In all these instances, simple single-gene genetics is often insufficient to functionally untangle these complexities, because achieving a strong phenotype resulting from the loss of a single gene is challenging (or impossible). Nonetheless, more comprehensive characterization of gene function in microorganisms is necessary to tackle this complexity.

While most microorganisms likely display at least some level of genetic redundancy, the extent of this redundancy remains to be determined. Even if a single bacterial genome encodes two proteins with identical biochemical properties, these genes could be under the control of different regulatory systems, and thus each paralog is expressed under a particular condition. Understanding which proteins and pathways are redundant for a given microorganism can offer insight into the selective pressures relevant to the organism's evolution, under the hypothesis that a microorganism is more likely to evolve redundancy for important functions. A more detailed dissection of the differing functionality of paralogs can shed important insights into the evolution of protein families and provide a framework for more accurate gene annotations in newly sequenced genomes.

Classic genetic approaches for investigating redundancy involve the construction of strains carrying multiple mutations and the identification of epistatic (genetic) interactions between the pair of mutations, an approach that is currently difficult to achieve in a high-throughput manner in diverse microorganisms. Biochemical approaches to differentiate paralogs are subject to the same challenges highlighted earlier, including the development of a suitable assay to detect potentially subtle differences in the protein's activities. These current limitations potentially can be overcome with multiple-gene knockouts and conditional genetic interaction screening, gain-of-function genetics, and gene characterization in related species.

Multiple-Gene Knockouts and Conditional Genetic Interaction Screening

Pioneered in yeast, systematically constructing and phenotyping all pairs of knockouts are possible

(Tong et al. 2004). However, the approach used in this large effort is probably not suitable for cost-effective application to a host of nonmodel microorganisms. One attractive approach that is more readily transferable to diverse species is the use of CRISPR/Cas methods, whereby two or more guide RNAs can be used to edit or reduce the expression of multiple genes in a single strain (Adames et al. 2019). Another advantage of this approach is that it should be more readily amenable to phenotyping multigene mutated strains under different experimental conditions, thus launching the systematic discovery of condition-specific genetic interactions (Jaffe et al. 2019).

Gain-of-Function Genetics

Measuring phenotypes in strains overexpressing one or more genes is potentially less prone to issues related to genetic redundancy. A number of options exist for the systematic overexpression of genes, including random shotgun expression from plasmids (Mutalik et al. 2019), precise open reading frame (ORF) libraries (Kitagawa et al. 2005), and activation modes of CRISPR (Chavez et al. 2016).

Gene Characterization in Related Species

Any particular microbial strain may carry two or more copies of a gene or pathway, making genetic approaches for investigating gene function challenging. However, close homologs of these genes may be a single copy in other genomes, thereby simplifying phenotypic characterization of mutations and gene function.

5. Plants

Relative to microorganisms, the characterization of gene function in land plants presents unique challenges that are attributable to the size and complexity of plant genomes, multicellularity and complexity of plant organs, broad technical barriers that restrict genetic manipulation, complexity of heterogeneous environments that plants inhabit, and the need for substantial infrastructure for experimentation. Multiple opportunities exist to advance the knowledge of gene function in plants. They center on improving efficiencies in research discovery by (1) targeting research in focal species, (2) incorporating existing technologies into U.S. Department of Energy (DOE) Office of Biological and Environmental Research (BER) programs and focal plants, (3) adopting modeling approaches to better define components of biological processes, (4) expanding the environments in which gene function is determined, and (5) advancing paradigm-changing experimental and computational platforms for gene function characterization in plants.

Plant Systems: Unique Challenges

Large and Redundant Genomes

The average plant genome encodes more than 30,000 genes; polyploid species have even larger genomes and corresponding gene complements (see Fig. 5.1, this page). Furthermore, plant genomes contain substantial functional redundancy due to the extensive history of both segmental and whole-genome duplications found in nearly all angiosperms (i.e., flowering plants), resulting in large numbers of expanded gene families (see Fig. 5.2, p. 42). Within these large gene families, individual paralogs can undergo sub- or neofunctionalization, and thus the functional annotation of paralogs based on the activity of a single gene within the family can result in false or misleading annotations (Panchy et al. 2016). As a consequence, this high degree of genetic redundancy presents multiple challenges in determining gene function through single-gene knockout experiments (Simon et al. 2013).

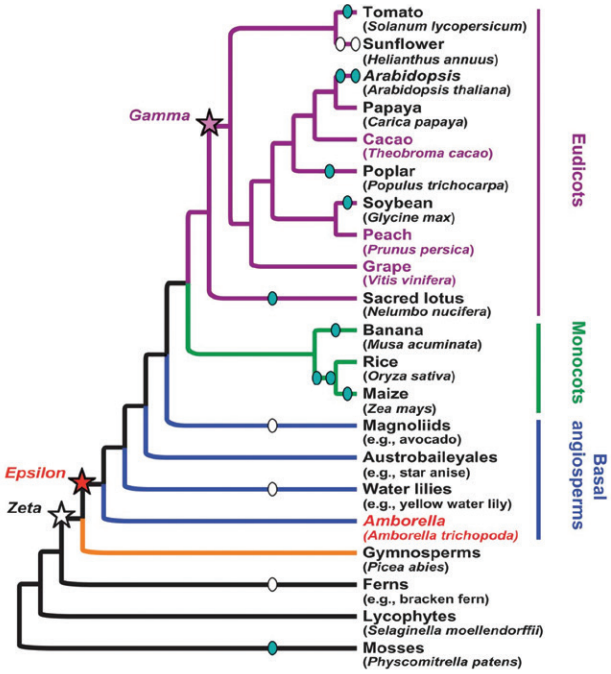
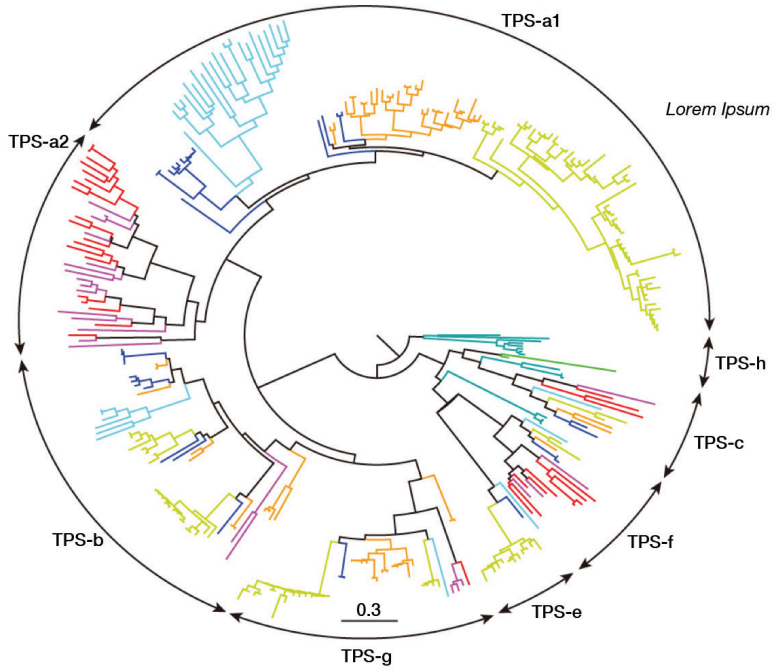


Fig. 5.1. Tree of Life. Phylogeny of land plants with whole-genome duplication and polyploidy events highlighted by stars and ellipses, respectively. Numbers of annotated protein-coding genes include tomato (34,727), sunflower (52,243), *Arabidopsis* (27,655), papaya (27,332), cacao (29,452), poplar (42,950), soybean (56,044), peach (26,873), grape (26,346), sacred lotus (36,385), banana (36,528), rice (39,045), maize (39,591), *Amborella* (26,846), gymnosperms (28,354), lycophytes (22,273), and mosses (32,926). These annotation numbers were taken from Phytozome 12.1 (https://phytozome.jgi.doe.gov/pz/portal.html), with the exception of *Nelumbo nucifera* (Wang et al. 2013), *Oryza sativa* (rice.plantbiology.msu.edu), *Picea abies* (Nystedt et al. 2013), and *Zea mays* (Ensembl 18; ensemblgenomes.org). [From *Amborella* Genome Project. 2013. "The *Amborella* Genome and the Evolution of Flowering Plants," *Science* **342**(6165), 1241089. DOI:10.1126/science.1241089. Reprinted with permission from AAAS.]

Genomes and genes within a species can diverge, resulting in single-nucleotide polymorphisms, small insertions or deletions, and structural variation including copy number variation, presence or absence variation, inversions, and translocations. The extent of genome diversity within a species is attributable to various factors such as modes of reproduction



Species	Number of TPS Genes
Chlamydomonas	0
Physcomitrella patens	2
Selaginella moellendorffii	13
Solanum tuberosum	32
Vitis vinifera	83
Eucalyptus grandis	113
Glycine max	30
Medicago truncatula	34
Arabidopsis thaliana	34
Populus trichocarpa	59
Fragaria vesca	58
Brachypodium distachyon	16
Sorghum bicolor	47
Zea mays	36
Oryza sativa	51

Fig. 5.2. Terpene Synthase Gene Family Expansion Across the Plant Kingdom. [Reprinted under a Creative Commons license (CC-BY-NC-SA-3.0) from Myburg, A. A., et al. 2014. "The Genome of *Eucalyptus grandis*," *Nature* **510**, 356–62. DOI:10.1038/nature13308.]

(e.g., inbreeding versus outcrossing), selection events throughout its life history, and genetic bottlenecks during domestication and subsequent improvement by artificial selection (Hardigan et al. 2017; Evans et al. 2018). Pan-genomes, the collective set of genes present within a species, have been constructed for a number of plant species including crops (Gordon et al. 2017; Hardigan et al. 2017; Hirsch et al. 2014; Schatz et al. 2014). Interestingly, the functions of some structural variants have been shown experimentally to be involved in adaptive responses such as flowering time, tolerance to abiotic and biotic stressors, and synthesis of specialized metabolites (Tao et al. 2019). Thus, use of a single reference accession or genotype will fail to provide representation of all the genes and their function within a single species. An additional challenge to gene function determination in plants will be understanding the function of the dispensable genes within the pan-genome.

This collective complexity in gene complement results in a highly complex transcriptome, proteome, and metabolome, but it also has led to the great adaptability of plant species to diverse environments. Indeed, gene duplication and adaptability to changing environmental conditions have resulted in highly complex gene regulatory networks, as indicated by the greater diversity of transcription factors in plants than in any other clade of organisms (Riechmann et al. 2000). As a consequence, there is limited knowledge of the "regulome" of DOE mission–relevant plant species because detailed studies on transcription factors, cis-regulatory regions, enhancers, chromatin state, and alternative splicing have not been performed at a level sufficient to construct robust, predictive models of plant metabolic and other processes.

Multicellular, Sessile Organisms

In contrast to microorganisms, plants have complex organs consisting of multiple cell types, each with distinct transcriptomes, proteomes, and metabolomes. Organs of BER interest are leaves (the primary site of photosynthesis and carbon fixation); stems, stalks, and trunks (water- and nutrient-conducting suite of

tissues); and roots (site of nutrient acquisition and carbon sequestration) (see Fig. 5.3, this page). Each of these organs has a complex set of cell types with dynamic spatiotemporal gene expression profiles. While large-scale omic datasets are available for a number of BER-relevant species, there is no single-cell resolution for even a single species' organs, with the exception of *Arabidopsis thaliana* roots (Ryu et al. 2019). As a result, the overwhelming majority of omics-based approaches provide a mosaic representation of the epigenome, transcriptome, proteome, and metabolome at the organ level.

Restrictive and Cumbersome Genetic Manipulation

With few exceptions, genetic manipulation methods for all angiosperms are rudimentary and labor intensive. Indeed, only two species of Brassicaceae, *A. thaliana* and *Camelina sativa*, can be transformed with the facile floral dip method (Clough and Bent 1998). All other species require transformation via the plant pathogen *Agrobacterium tumefaciens* or biolistic gun followed by regeneration of plants via tissue culture, a lengthy, labor-intensive method that is highly genotype dependent. A major breakthrough was reported recently in the transformation efficiency of monocots using the morphogenic regulators *Baby boom* and *Wuschel* (Lowe et al. 2016), but this technology is not yet readily dispersed in the public sector. Thus, a major bottleneck in assessing gene function in plants is the inability to rapidly generate and screen transgenic lines that are either deficient (i.e., knockout or knockdown) or overexpressing genes of interest, a standard approach in microbial systems.

Gene Function in Diverse Environments

Plants exist within complex environments and as sessile organisms; as a consequence, they require phenotypic plasticity to adapt to changing environmental conditions. Thus, one component in determining gene function is understanding the genotype-by-environment (G × E) effects on phenotype. The logistics and costs of phenotyping experiments in diverse environments typically are prohibitive, so most gene function assays are performed in a single or limited number of environments. The Genomes to Fields (G2F) Initiative, a group of maize researchers phenotyping a common set of genotypes in diverse environments using standardized methods, highlights one mechanism by which G × E can be assessed (https://www.genomes2fields.org; Gage et al. 2017; Alkhalifah et al. 2018).

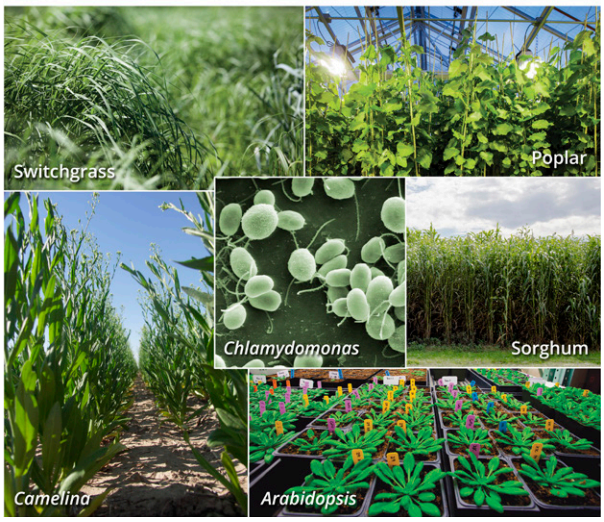


Fig. 5.3. Example Species for Gene Function Determination in Biological and Environmental Research Studies. Plant tissues and organs used for biofuel and bioproduct purposes represent a mosaic of cell types. [Switchgrass courtesy Great Lakes Bioenergy Research Center. Poplar courtesy Center for Bioenergy Innovation. Sorghum courtesy Center for Advanced Bioenergy and Bioproducts Innovation. *Arabidopsis* courtesy Lawrence Berkeley National Laboratory. *Camelina* courtesy Jean-Nicolas Enjalbert, Colorado State University. *Chlamydomonas*, Wikimedia Commons, Dartmouth Electron Microscope Facility, Dartmouth College.]

Focal Species to Accelerate Gene Function Discoveries

Multiple photosynthetic land plant and algal species that span angiosperm phylogeny and phenology have been explored for biofuel potential over the last few decades. Complementing these "biofuel" species, including biomass and oilseed species, are model plant species that provide a platform for more rapid, efficient gene function discovery as well as a deep knowledgebase of core gene function in the plant kingdom. Identifying a set of focal biofuel species (see Fig. 5.3) for determination of gene function that are coupled to efforts in model species can provide synergistic

opportunities not only to maximize knowledge gain but also to harness the novel evolution of traits present in biofuel species.

Biomass Species

Three BER target angiosperm species for biomass production (i.e., poplar, switchgrass, and sorghum) provide opportunities for high rates of gene function determination as all three species have substantial phenomic and genomic resources that can be readily leveraged. As a woody perennial C3 dicot, poplar is a well-established biofuel feedstock, has excellent genomic and diversity datasets and resources (Lin et al. 2018; Tuskan et al. 2006), and is well characterized for biofuel feedstock properties (Sannigrahi et al. 2010). Poplar can be grown over a large geographic distribution (Stanton et al. 2010), but the logistics of storing and growing modified poplar germplasm remain a challenge. Switchgrass and sorghum, both C4 grasses, are comparatively close relatives, and their genomes share synteny (Sharma et al. 2012). However, switchgrass, a native North American prairie grass, is a polyploid (i.e., tetraploid and octoploid) perennial that presents substantial challenges with respect to forward and reverse genetics, as well as the inherent logistical challenges present in field studies. In contrast to switchgrass, sorghum is an annual inbred diploid. Given the rapidly maturing reverse genetic and phenomic resources in sorghum (Brenton et al. 2016), it is a strong candidate as a key plant species for improving understanding of the functional roles of plant genes in determining phenotypes across diverse environments.

Oilseed Species

Among BER target angiosperms for oilseed and biobased products, *Camelina* produces oil suitable for jet fuel that can be directly extracted from seed and is a close relative of the model species *A. thaliana*, sharing significant gene identity and synteny (Kagale et al. 2014; Malik et al. 2018). Strategically, *Camelina*, which can be transformed with the floral dip method, is a recent hexaploid having substantial genetic redundancy (Kagale et al. 2014). This challenge can be overcome through gene editing to make a pseudo-allelic series (Aznar-Moreno and Durrett 2017; Morineau et al. 2017). To date, limited studies suggest that *Camelina* is tolerant to cold and drought and is able to grow with minimal amendments (Jiang et al. 2014; Obour et al. 2017; Vollmann et al. 2007). Maximally exploiting *Camelina* as a biofuel crop requires understanding the genetic and physiological basis of its resilience, providing an opportunity to advance ongoing and future synthetic biology efforts focused on engineering enhanced and modified oil content in *Camelina*.

Model Species

Other BER-relevant species include the model species *A. thaliana* and *Chlamydomonas reinhardtii*, both of which have well-developed tools and extensive knowledgebases (Berardini et al. 2015; Krishnakumar et al. 2017; Waese et al. 2017) and provide powerful synergies for characterizing genes of unknown function. Collecting gene function evidence through experimentation with model systems and contextualizing that information within an evolutionary framework can be a powerful approach for discovering the function of conserved plant protein families. This functional information could serve as a catalyst for targeted (and often more resource-intensive) experimentation as needed to understand aspects of gene function potentially unique to a specific bioenergy crop of BER interest.

A number of the species described in this section are part of DOE's Plant Flagship Genomes project (https://jgi.doe.gov/our-science/science-programs/plant-genomics/plant-flagship-genomes/). For this project, Joint Genome Institute resources are centrally vested in developing genome sequence data relevant to DOE's missions, thereby providing an opportunity to capitalize on previous investments in plant-based systems.

Well-Annotated Genomes and Associated Datasets

Central to all downstream efforts, user-friendly access to well-annotated genomes with associated large-scale omic datasets can enable greater discoveries and enhanced efficiencies of gene function determination. Although genome sequence data and a subset of omic data exist for BER-relevant plant species, these data are

neither uniform in breadth and depth nor integrated either within the species or with other species, thereby limiting researchers in their ability to extract information from existing datasets. Thus, placing publicly available data into an integrated repository with appropriate standards and tools for data analysis and interpretation would be a powerful action for advancing the field. Generating transcriptome compendia that collate data from a wide range of experimental conditions would provide an atlas of potential gene function via "guilt by association," as well as facilitate identification of gene family members with redundant function. Chromosomal context, which has been shown to be effective in subsetting genes within the same taxonomic clade, could be incorporated into functional annotation. Cataloging focal species for relevant omics and associated datasets including developing multiple reference genomes, characterizing the pan-genome, constructing diversity panels with associated genetic diversity data, and cataloging the phytobiome would enable enhanced resolution of putative gene function and generation of priority lists for functional validation.

Prioritization of Gene Sets for Functional Experimentation

As the functions of uncharacterized genes are, by definition, unknown, focusing solely on uncharacterized genes that are DOE-mission relevant is not possible. However, a number of prioritization strategies can be employed to select genes that are more likely to play significant roles in determining desirable plant phenotypes in one or multiple environments. For example, orthology analyses with the predicted proteomes of photosynthetic and nonphotosynthetic species have yielded a set of proteins (termed GreenCut) that are restricted to photosynthetic organisms (Merchant et al. 2007), thereby honing the list of genes that may have relevant functions key to plants. With deeper and broader sampling of the tree of life, comparative evolutionary analyses can have increased precision and resolution. Indeed, for BER focal species, comparative genomic analyses can be performed to identify conserved protein families in the plant lineage as well as plant-specific protein families and a phylogenomic framework, forming the basis for functional inference associations between taxa. In the case of extremely large yet sub- or neofunctionalized gene families, higher-resolution molecular structures and computational modeling can be used to identify and prioritize residues likely to affect specificity or active-site function. This structure-guided approach is likely to aid in identifying sets of homologous genes where appropriate for sharing detailed enzymatic functional data versus those where detailed enzymatic function is likely to have changed. This approach also can be used to facilitate and improve predictions of alternatively spliced genes and identify key features of DNA-binding proteins (e.g., transcription factors) that govern binding specificity.

Perturbation of Genes via Gene Editing

Currently, genome editing approaches can be used to generate knockout and knockdown lines that are coupled with phenotyping to facilitate gene function identification. Increasing the scale and pace of gene editing along with higher-throughput phenotyping would enable researchers to more rapidly associate genes with function. With the high degree of redundancy in plant genomes, gene editing of subsets or entire gene families, as recently demonstrated for the sorghum kafirin gene family (Li et al. 2018), could generate a pseudo-allelic series of edited lines with various numbers of paralogs disrupted, thereby enabling discovery of gene function in these paralogous families.

Modeling of Relevant Plant Processes

Biomass production across diverse environments, variation in carbon allocation between organs, and composition of biomass are key traits of interest to DOE missions. Current knowledge of plant biological processes is limited to simplistic models, and, yet, analogous to fitness growth curves in single-celled organisms, biomass production integrates the output of numerous pathways and biological processes. Consequently, the effect of functional variation in any single gene is likely to represent a small proportion of total phenotypic variation (Boyle et al. 2017; Simon et al. 2013). Likewise, allocation of photosynthetic carbon between roots and shoots, determining subsequent

resource capture below and above ground, respectively, and eventually new growth are highly plastic and determined by multiple genes and development pathways (Drouet and Pagès 2003; Weiner 2004; Yu et al. 2014). One strategy that has been successfully employed for other complex traits is to identify the component traits that each feed into the variability in the final output trait (York 2019). New computational approaches such as machine learning can be applied to plant systems to improve understanding of gene function not only by discovering new interactions, but also by increasing the knowledge resolution of biological processes (Moore et al. 2019). Expanded datasets such as transcriptomic, proteomic, metabolic, and whole-plant traits collected in a time-series manner from common experiments, including diversity panels or mutant populations that capture environmental perturbations, would facilitate model development (Bechtold et al. 2016; Caldana et al. 2011; Marshall-Colon et al. 2017).

G × E: Role of Environment

Important to note is that sustainable production of biomass and biofuels will require improved cultivars with high productivity in diverse environments with minimal amendments. A robust understanding of the genetic and molecular basis of biomass and biofuel feedstock production in such environments, including stress or marginal conditions, could facilitate strategic breeding and biotechnology innovations in these crops. The high-throughput phenotyping methods currently coming online in plants provide an opportunity to explore gene function determination in the phenotypic plasticity of focal crop species in diverse environments and identify natural variation associated with traits of interest.

Minimal Plant Genome Platform for Gene Function Discovery

Genomic diversity, which has resulted in phenotypic diversity that can be exploited for energy production, complicates determination of gene function across diverse taxa. This complication results from logistical barriers in transformation, an inability to readily traverse gene function through sequence similarity, and reduced efficiencies. Synthetic organisms that represent the minimal gene complement for life provide a platform to rapidly test gene function. In microbial systems, scientists have been able to fabricate minimal genomes using a bottom-up approach with a combination of gene synthesis and molecular biology (Hutchison et al. 2016). For yeast, scientists have constructed synthetic chromosomes that permit rapid determination of gene function via gene or allele replacement, deletion, and rearrangement using the SCRaMbLE system (Synthetic Yeast 2.0; syntheticyeast.org). For plants, a bottom-up approach in which minimal chromosomes are synthesized would be challenging due to not knowing which genes are essential. Likewise, while the use of an approach similar to the SCRaMbLE system would potentiate discovery of gene function in plants, as a first step, knowledge of all essential genes would be required to prevent inadvertent lethality. Thus, the generation of a minimal plant genome using a top-down approach in which genes are sequentially removed from a wild-type plant with an amenable transformation system would serve two functions: (1) provide a platform for gene function determination for all taxa, and (2) provide foundational information on genes essential to plant function. Engineering a minimal plant genome could be either model based or empirically based, both of which would define the minimal set of genes to produce a functioning flowering plant. A suite of computational approaches can be used to identify which genes or genome regions can and cannot be removed, their order, and their level of redundancy. Although new methods such as somatic embryogenesis and protoplast transformation or regeneration may make construction of a minimal plant genome possible in other species, it is only likely to be feasible at this time in *A. thaliana* due to the ease of gene editing and floral dip transformation (Clough and Bent 1998). Access to a minimal plant genome would also provide a framework to then combinatorially add genes and gene cassettes and assess function in a transformative way, greatly facilitating synthetic biology in plant systems.

6. Conclusions and Outlook

Only a tiny fraction of sequenced genes has been studied experimentally, creating a massive knowledge gap in genome understanding. This gene-to-function challenge affects all biological research and is particularly pronounced for Office of Biological and Environmental Research (BER)–funded investigations into energy and the environment, given the great diversity of plant and microbial taxa being studied. As BER strives for a genomics-enabled, predictive understanding of engineered and natural environmental systems, this vision is significantly and negatively impacted by the millions of genes of unknown function that have been identified to date. In light of this grand challenge and its importance to U.S. Department of Energy (DOE) missions, BER convened the Breaking the Bottleneck of Genomes: Understanding Gene Function Across Taxa workshop on November 1–2, 2018. Workshop participants identified challenges associated with large-scale gene function determination across taxa and current and future opportunities for overcoming these challenges.

The workshop focused on four key areas: (1) experimental technologies and (2) computational systems necessary for large-scale gene function discovery and accurate gene annotation that are applicable across all taxa, as well as (3) microbe- and (4) plant-specific biology and experimental limitations hindering the development of unified solutions.

Technologically, new breakthrough approaches are needed that can be systematically applied at low cost to inform gene function across many organisms. Though some of these scalable approaches are already available in some form, most have not yet been applied systematically and at scale across a multitude of diverse organisms and protein families. Computationally, new resources and modification of existing platforms are crucial not only for storing and propagating high-quality, experimentally determined gene functions, but also for providing a data analysis framework for the inference of new gene functions from diverse experimental data.

BER-relevant organisms include microorganisms, algae, and plants that can produce sustainable biomass, synthesize biofuels and bioproducts, sequester carbon, and transform environmental contaminants. Given the relative ease and low cost of working with many microorganisms, opportunities are available to functionally characterize the genes of a diversity of organisms including bacteria, fungi, and archaea, or to focus on the comprehensive characterization of a targeted microbiome. In contrast, for plants, detailed and systematic interrogation of gene function may be feasible at this time for only a few targeted species, with the expectation that the data and experimental approaches could ultimately be extended more broadly.

The gene-to-function challenge is daunting, but workshop participants enthusiastically agreed that the "genome bottleneck" can be broken. In particular, measurement technologies are rapidly progressing across multiple scales including metabolites, proteins, cellular networks, and fabricated and natural ecosystems. The data generated by these technologies can be coupled to computational advances in automated inference, including those enabled by machine learning, to infer new functions and accurately annotate newly sequenced genes when experimental data for homologs exist. Solving the gene-to-function challenge has the potential to transform the understanding of biology, with long-lasting implications for improving capabilities to predictably engineer and harness organisms for sustainable energy and environmental outcomes.

Appendix A: Workshop Agenda

Breaking the Bottleneck of Genomes: Understanding Gene Function Across Taxa

November 1–2, 2018

Bethesda North Marriott, Maryland

Thursday, November 1

8:00 a.m.	Breakfast
8:30 a.m. – 9:00 a.m.	Welcome, Introduction, and Overview U.S. Department of Energy (DOE) Biological and Environmental Research (BER) program representatives and co-chairs
9:00 a.m. – 10:40 a.m.	**Plenary Sessions: Differences and Commonalities** 9:00 a.m. – 9:25 a.m. Valerie de Crecy-Lagard 9:25 a.m. – 9:50 a.m. Jeffrey Skerker 9:50 a.m. – 10:15 a.m. Shawn Kaeppler 10:15 a.m. – 10:40 a.m. Geoffrey Chang
10:40 a.m. – 11:00 a.m.	Break
11:00 a.m. – 11:45 a.m.	**Brainstorming: What Is Function, What Are the Real Bottlenecks, and What Is the Definition of Success?** Robin Buell and Adam Deutschbauer
11:45 a.m. – 12:15 p.m.	General Discussion
12:15 p.m. – 1:00 p.m.	Working Lunch
1:00 p.m. – 3:00 p.m.	**Breakout Session I:** Plants – Led by James Schnable Microbes – Led by Judy Wall
3:00 p.m. – 3:15 p.m.	Break
3:15 p.m. – 5:15 p.m.	**Breakout Session II:** Computation – Co-led by Molly Megraw and Chris Henry Technologies – Co-led by Martin Jonikas and Trent Northen
5:15 p.m. – 6:00 p.m.	Reports from Breakout Groups – Quick 10-minute survey, no slides
6:00 p.m.	Dinner on your own

Friday, November 2

8:00 a.m.	Breakfast
8:30 a.m. – 9:15 a.m.	**Breakout Session I** – Work on slides
9:15 a.m. – 10:00 a.m.	**Breakout Session II** – Work on slides
10:00 a.m. – 10:15 a.m.	Break
10:15 a.m. – 12:15 p.m.	Presentations from Breakout Groups (25 minutes, 5-minute discussion)
12:15 p.m. – 1:15 p.m.	Working Lunch (boxed lunch) Discussion and Wrap-Up
1:15 p.m.	Participants Adjourn
1:15 p.m.	Discussion and Writing Session Co-chairs, Breakout Leads, and DOE BER staff

Appendix B: Breakout Session Charge Questions

U.S. Department of Energy (DOE)
Office of Biological and Environmental Research (BER)
Breaking the Bottleneck of Genomes:
Understanding Gene Function Across Taxa Workshop

Focus: To use advanced experimental and computational tools to rapidly determine the functions of the millions of uncharacterized microbial and plant genes identified by DNA sequencing **AND/OR** to greatly narrow the gap between our ability to sequence microbial and plant genes and our capacity to accurately annotate them.

Goals:

- To identify the roadblocks and research gaps that currently restrict our ability to systematically elucidate gene function in DOE-relevant microorganisms and plants.
- To develop a long-term blueprint for the functional annotation of genes of unknown function.
- To determine how new experimental tools can be coupled with advanced computation to greatly increase the confidence associated with genome annotations.

Expected Outcome: State-of-the-art community scientific recommendations and critical research priorities that will help inform BER program direction in large-scale plant and microbial gene function determination.

Charge Questions for Breakout Sessions:

Brainstorming: Identify Topic Areas

- Identify the most compelling questions that can be addressed by large-scale determination of plant and microbial gene function.
- What does success look like in 5 years? In 20 years? In the future, by what process will a newly sequenced microbial genome, metagenome, or plant genome be annotated for gene function?
- What is gene function with respect to its biochemical versus physiological/environmental role? Should one be prioritized over the other?
- How can improved gene annotations be integrated into and improve metabolic and gene regulatory models? How can improved gene and genome annotation be incorporated into understanding biological processes such as metabolism, gene regulation, and adaptation to environment?
- What is the role of existing databases/knowledgebases in large-scale gene function annotation? Do they provide quality resources for the community? Can they be re-aligned with emerging needs of the community? How can new experimental and computational approaches synergize and enhance prior data?
- How will improved microbial and plant gene annotations improve BER mission needs? And also complement existing DOE research areas?
- Given the scale of the challenge, what is the role for large-team projects versus smaller single-investigator projects? Are there funding barriers to enabling success?
- What roles can existing DOE research facilities (such as user facilities) play in such an endeavor? How does one weigh the benefits of identifying functions for many proteins in a species of interest versus a deep dive into a single function?
- What roles can nonbiology disciplines (e.g., engineering, robotics, physics, high-performance computing, and nanotechnology) perform in this effort?
- Should gene annotation be crowdsourced? If so, how do you engage the broader community to participate in such an effort?

- Given the scale of the challenge and the eventual democratization of large-scale data, should large-scale gene annotation efforts also be incorporated into education and outreach programs? Each class gets data for a microorganism/plant that is generated by a single center that can produce large amounts of data?

Identify Key Knowledge Gaps and Opportunities

- Identify research priorities for addressing gaps and opportunities.

A. Breakout Session I: Plants

a. What is the value in identifying a small subset of species for targeted investigation of gene function and technology development? Would a C3, C4, and oilseed species be sufficient?

b. As phenotyping is expensive and logistically challenging, what phenotypes should be prioritized, and what technologies for phenotyping should be developed?

c. Gene and genome duplication cloud interpretation of gene knockout experiments; what new approaches can be taken to develop a better approach for understanding gene function?

d. What types of high-throughput *in vitro* assessments of gene function would be a priority for plant biology?

B. Breakout Session I: Microbes

a. Should microorganisms be prioritized for investigation and technology development? If so, which ones and why?

b. What is the value in a systematic effort to characterize gene function in a phylogenetically diverse range of microorganisms?

c. Should class(es) of proteins be targeted for investigation [enzymes, transporters, domains of unknown function (DUFs), and other conserved hypotheticals]? If so, which ones and why?

d. What strategies can be employed to determine microbial gene function within a natural ecological context, including microbe-microbe interactions?

e. How can we streamline molecular genetic tool development in nonmodel microorganisms?

f. How do we overcome the limitations of functional assignments to specific proteins when there are compensatory systems within microbes? Is crosstalk of regulatory systems a useful feature or simply unavoidable? Can the compensatory systems span members of a community, and is that important only in environments of high microbial densities?

g. Can protein complexes with channeling of metabolic intermediates be examined with current high-throughput techniques?

h. How do we overlay our phenotypic analysis of proteins with post-translational modification processes?

i. Because nearly 50% of proteins are metalloproteins, how do we establish specificity of the metals used or flexibility of function with different metals?

j. Is phenotypic understanding of protein function limited by the necessity to perform analyses under conditions of convenience to the researcher?

C. Breakout Session II: Computation

a. How can large-scale experimental data of diverse types (metabolomics, proteomics, biochemistry, transcriptomics, genetics, physiology, imaging, and structural biology) be effectively integrated with existing annotation strategies to greatly improve the accuracy and resolution of these annotations? Are existing annotation strategies sufficient for this task?

b. What experimental data not available now are needed to enable predictive biology?

c. What is the role for pathway and network modeling in these efforts?

d. How can we test the veracity of new gene annotations? Provide a confidence level, provide specific evidence (data/literature/associations) to support the annotation? Can/should parts of this process be automated, potentially through the use of text-mining/natural-language-processing techniques?

e. Gene annotations are largely static and often are not updated with current biological knowledge. In light of this, how can new information (i.e., data) be effectively propagated to existing gene annotations in databases?

f. Can high-performance computing and new algorithms (such as machine learning) be applied to improve annotations? Do we currently have the data to make this an effective strategy?

D. Breakout Session II: Technologies

a. What technologies do we have in hand that can be applied to determine gene function? Can these technologies be scaled massively and inexpensively?

b. What are the major technological gaps to enable large-scale gene function determination in microorganisms and plants?

c. What early stage, but potentially groundbreaking technologies, should be prioritized? Why?

d. How should technologies be integrated to scale gene function determination?

Prepare Summary Presentations

- For the given topic area, what are the top obstacles to accurate gene function annotation?
- For the given topic area, what specific recommendations do you have for research areas that should be prioritized and implemented by BER?
- What does success look like?

Appendix C: Workshop Participants

Co-Chairs

Robin Buell
Michigan State University

Adam Deutschbauer
Lawrence Berkeley National Laboratory

Participants

Adam Abate
University of California, San Francisco

Joshua Adkins
Pacific Northwest National Laboratory

Crysten Blaby-Haas
Brookhaven National Laboratory

Geoffrey Chang
University of California, San Diego

Valerie de Crecy-Lagard
University of Florida

John Gerlt
University of Illinois

Chris Henry
Argonne National Laboratory

Dan Jacobson
Oak Ridge National Laboratory

Joseph Jez
Washington University in St. Louis

Martin Jonikas
Princeton University

Shawn Kaeppler
University of Wisconsin;
Wisconsin Crop Innovation Center

Sasha Levy
Stanford University;
Joint Institute for Metrology in Biology

Amy Marshall-Colon
University of Illinois

Molly Megraw
Oregon State University

Trent Northen
Lawrence Berkeley National Laboratory;
Joint Genome Institute

Andrei Osterman
Sanford Burnham Prebys Medical Discovery Institute

James Schnable
University of Nebraska

Jeffrey Skerker
University of California, Berkeley;
Lawrence Berkeley National Laboratory

Michael Udvardi
Noble Research Institute

Dan Voytas
University of Minnesota

Judy Wall
University of Missouri

Dave Weston
Oak Ridge National Laboratory

Appendix D: References

Adames, N. R., et al. 2019. "Yeast Genetic Interaction Screens in the Age of CRISPR/Cas," *Current Genetics* **65**(2), 307–27. DOI:10.1007/s00294-018-0887-8.

Alivisatos, A. P., et al. 2015. "Microbiome. A Unified Initiative to Harness Earth's Microbiomes," *Science* **350**(6260), 507–08. DOI:10.1126/science.aac8480.

Alkhalifah, N., et al. 2018. "Maize Genomes to Fields: 2014 and 2015 Field Season Genotype, Phenotype, Environment, and Inbred Ear Image Datasets," *BMC Research Notes* **11**(1), 452. DOI:10.1186/s13104-018-3508-1.

Araus, J. L., et al. 2018. "Translating High-Throughput Phenotyping into Genetic Gain," *Trends in Plant Science* **23**(5), 451–66. DOI:10.1016/j.tplants.2018.02.001.

Arkin, A. P., et al. 2018. "KBase: The United States Department of Energy Systems Biology Knowledgebase," *Nature Biotechnology* **36**(7), 566–69. DOI:10.1038/nbt.4163.

Atkinson, H. J., et al. 2009. "Using Sequence Similarity Networks for Visualization of Relationships Across Diverse Protein Superfamilies," *PLOS One* **4**(2), e4345. DOI:10.1371/journal.pone.0004345.

Aziz, R. K., et al. 2008. "The RAST Server: Rapid Annotations Using Subsystems Technology," *BMC Genomics* **9**, 75. DOI:10.1186/1471-2164-9-75.

Aznar-Moreno, J. A., and T. P. Durrett. 2017. "Simultaneous Targeting of Multiple Gene Homeologs to Alter Seed Oil Production in *Camelina sativa*," *Plant and Cell Physiology* **58**(7), 1260–67. DOI:10.1093/pcp/pcx058.

Baccouche, A., et al. 2017. "Massively Parallel and Multiparameter Titration of Biochemical Assays with Droplet Microfluidics," *Nature Protocols* **12**(9), 1912–32. DOI:10.1038/nprot.2017.092.

Bai, X. C., et al. 2015. "How Cryo-EM Is Revolutionizing Structural Biology," *Trends in Biochemical Sciences* **40**(1), 49–57. DOI:10.1016/j.tibs.2014.10.005.

Baran, R., et al. 2013. "Metabolic Footprinting of Mutant Libraries to Map Metabolite Utilization to Genotype," *ACS Chemical Biology* **8**(1), 189–99. DOI:10.1021/cb300477w.

Bartlett, A., et al. 2017. "Mapping Genome-Wide Transcription-Factor Binding Sites Using DAP-seq," *Nature Protocols* **12**(8), 1659–72. DOI:10.1038/nprot.2017.055.

Bastard, K., et al. 2017. "Parallel Evolution of Non-Homologous Isofunctional Enzymes in Methionine Biosynthesis," *Nature Chemical Biology* **13**(8), 858–66. DOI:10.1038/nchembio.2397 [https://www.nature.com/articles/nchembio.2397#supplementary-information].

Bechtold, U., et al. 2016. "Time-Series Transcriptomics Reveals That *AGAMOUS-LIKE22* Affects Primary Metabolism and Developmental Processes in Drought-Stressed *Arabidopsis*," *The Plant Cell* **28**(2), 345–66. DOI:10.1105/tpc.15.00910.

Ben-Nissan, G., and M. Sharon. 2018. "The Application of Ion-Mobility Mass Spectrometry for Structure/Function Investigation of Protein Complexes," *Current Opinion in Chemical Biology* **42**, 25–33. DOI:10.1016/j.cbpa.2017.10.026.

BERAC. 2017. *Grand Challenges for Biological and Environmental Research: Progress and Future Vision*, DOE/SC–0190, Biological and Environmental Research Advisory Committee, BERAC Subcommittee on Grand Research Challenges for Biological and Environmental Research [https://science.energy.gov/~/media/ber/berac/pdf/Reports/%20BERAC-2017-Grand-Challenges-Report.pdf].

Berardini, T. Z., et al. 2015. "The *Arabidopsis* Information Resource: Making and Mining the "Gold Standard" Annotated Reference Plant Genome," *Genesis* **53**(8), 474–85. DOI:10.1002/dvg.22877.

Blair, P. M., et al. 2018. "Exploration of the Biosynthetic Potential of the *Populus* Microbiome," *mSystems* **3**(5), e00045-18. DOI:10.1128/mSystems.00045-18.

Blaser, M. J., et al. 2016. "Toward a Predictive Understanding of Earth's Microbiomes to Address 21st Century Challenges," *mBio* **7**(3), e00714-16. DOI:10.1128/mBio.00714-16.

Bowen, B. P., and T. R. Northen. 2010. "Dealing with the Unknown: Metabolomics and Metabolite Atlases," *Journal of the American Society for Mass Spectrometry* **21**(9), 1471–76. DOI:10.1016/j.jasms.2010.04.003.

Boyle, E. A., et al. 2017. "An Expanded View of Complex Traits: From Polygenic to Omnigenic," *Cell* **169**(7), 1177–86. DOI:10.1016/j.cell.2017.05.038.

Brenton, Z. W., et al. 2016. "A Genomic Resource for the Development, Improvement, and Exploitation of Sorghum for Bioenergy," *Genetics* **204**(1), 21–33. DOI:10.1534/genetics.115.183947.

Buchberger, A. R., et al. 2018. "Mass Spectrometry Imaging: A Review of Emerging Advancements and Future Insights," *Analytical Chemistry* **90**(1), 240–65. DOI:10.1021/acs.analchem.7b04733.

Budnik, B., et al. 2018. "SCoPE-MS: Mass Spectrometry of Single Mammalian Cells Quantifies Proteome Heterogeneity During Cell Differentiation," *Genome Biology* **19**(1), 161. DOI:10.1186/s13059-018-1547-5.

Burgoon, L. D. 2006. "The Need for Standards, Not Guidelines, in Biological Data Reporting and Sharing," *Nature Biotechnology* **24**(11), 1369–73. DOI:10.1038/nbt1106-1369.

Caldana, C., et al. 2011. "High-Density Kinetic Analysis of the Metabolomic and Transcriptomic Response of *Arabidopsis* to Eight Environmental Conditions," *The Plant Journal* **67**(5), 869–84. DOI:10.1111/j.1365-313X.2011.04640.x.

Calhoun, S., et al. 2018. "Prediction of Enzymatic Pathways by Integrative Pathway Mapping," *eLife* **7**, e31097. DOI:10.7554/eLife.31097.

Campbell, M. S., et al. 2014. "MAKER-P: A Tool Kit for the Rapid Creation, Management, and Quality Control of Plant Genome Annotations," *Plant Physiology* **164**(2), 513–24. DOI:10.1104/pp.113.230144.

Carlson, E. D., et al. 2012. "Cell-Free Protein Synthesis: Applications Come of Age," *Biotechnology Advances* **30**(5), 1185–94. DOI:10.1016/j.biotechadv.2011.09.016.

Chavez, A., et al. 2016. "Comparison of Cas9 Activators in Multiple Species," *Nature Methods* **13**(7), 563–67. DOI:10.1038/nmeth.3871.

Chen, I. A., et al. 2019. "IMG/M v.5.0: An Integrated Data Management and Comparative Analysis System for Microbial Genomes and Microbiomes," *Nucleic Acids Research* **47**(D1), D666–77. DOI:10.1093/nar/gky901.

Cherry, J. M., et al. 2012. "*Saccharomyces* Genome Database: The Genomics Resource of Budding Yeast," *Nucleic Acids Research* **40**(Database issue), D700–05. DOI:10.1093/nar/gkr1029.

Clair, G., et al. 2016. "Spatially-Resolved Proteomics: Rapid Quantitative Analysis of Laser Capture Microdissected Alveolar Tissue Samples," *Scientific Reports* **6**, 39223. DOI:10.1038/srep39223.

Clevenger, K. D., et al. 2017. "A Scalable Platform to Identify Fungal Secondary Metabolites and Their Gene Clusters," *Nature Chemical Biology* **13**(8), 895–901. DOI:10.1038/nchembio.2408.

Clough, S. J., and A. F. Bent. 1998. "Floral Dip: A Simplified Method for Agrobacterium-Mediated Transformation of *Arabidopsis thaliana*," *The Plant Journal* **16**(6), 735–43. DOI:10.1046/j.1365-313x.1998.00343.x.

Cole, B. J., et al. 2017. "Genome-Wide Identification of Bacterial Plant Colonization Genes," *PLOS Biology* **15**(9), e2002860. DOI:10.1371/journal.pbio.2002860.

Coradetti, S. T., et al. 2018. "Functional Genomics of Lipid Metabolism in the Oleaginous Yeast *Rhodosporidium toruloides*," *eLife* **7**, e32110. DOI:10.7554/eLife.32110.

Cunningham, F. J., et al. 2018. "Nanoparticle-Mediated Delivery Towards Advancing Plant Genetic Engineering," *Trends in Biotechnology* **36**(9), 882–97. DOI:10.1016/j.tibtech.2018.03.009.

Deutschbauer, A., et al. 2011. "Evidence-Based Annotation of Gene Function in *Shewanella oneidensis* MR-1 Using Genome-Wide Fitness Profiling Across 121 Conditions," *PLOS Genetics* **7**(11), e1002385. DOI:10.1371/journal.pgen.1002385.

Diaz-Mejia, J. J., et al. 2018. "Mapping DNA Damage-Dependent Genetic Interactions in Yeast Via Party Mating and Barcode Fusion Genetics," *Molecular Systems Biology* **14**(5), e7985. DOI:10.15252/msb.20177985.

Diss, G., and B. Lehner. 2018. "The Genetic Landscape of a Physical Interaction," *eLife* **7**, e3247. DOI:10.7554/eLife.32472.

Drouet, J. L., and L. Pagès. 2003. "GRAAL: A Model of GRowth, Architecture and carbon ALlocation During the Vegetative Phase of the Whole Maize Plant: Model Description and Parameterisation," *Ecological Modelling* **165**(2–3), 147–73. DOI:10.1016/s0304-3800(03)00072-3.

Eisen, J. A. 1998. "Phylogenomics: Improving Functional Predictions for Uncharacterized Genes by Evolutionary Analysis," *Genome Research* **8**(3), 163–67. DOI:10.1101/gr.8.3.163.

Evans, J., et al. 2018. "Extensive Genetic Diversity is Present Within North American Switchgrass Germplasm," *The Plant Genome* **11**(1), 170055. DOI:10.3835/plantgenome2017.06.0055.

Gage, J. L., et al. 2017. "The Effect of Artificial Selection on Phenotypic Plasticity in Maize," *Nature Communications* **8**(1), 1348. DOI:10.1038/s41467-017-01450-2.

Gao, J., et al. 2018. "Ecosystem Fabrication (EcoFAB) Protocols for the Construction of Laboratory Ecosystems Designed to Study Plant-Microbe Interactions," *Journal of Visualized Experiments* **134**, e57170. DOI:10.3791/57170.

Garcia, P. A., et al. 2016. "Microfluidic Screening of Electric Fields for Electroporation," *Scientific Reports* **6**, 21238. DOI:10.1038/srep21238.

Gaudinier, A., et al. 2018. "Transcriptional Regulation of Nitrogen-Associated Metabolism and Growth," *Nature* **563**(7730), 259–64. DOI:10.1038/s41586-018-0656-3.

Gertsman, I., and B. A. Barshop. 2018. "Promises and Pitfalls of Untargeted Metabolomics," *Journal of Inherited Metabolic Disease* **41**(3), 355–66. DOI:10.1007/s10545-017-0130-7.

Giaever, G., et al. 2002. "Functional Profiling of the *Saccharomyces cerevisiae* Genome," *Nature* **418**(6896), 387–91. DOI:10.1038/nature00935.

Goodacre, N. F., et al. 2014. "Protein Domains of Unknown Function are Essential in Bacteria," *mBio* **5**(1), e00744-00713. DOI:10.1128/mBio.00744-13.

Gordon, S. P., et al. 2017. "Extensive Gene Content Variation in the *Brachypodium distachyon* Pan-Genome Correlates with Population Structure," *Nature Communications* **8**(1), 2184. DOI:10.1038/s41467-017-02292-8.

Griesemer, M., et al. 2018. "Combining Multiple Functional Annotation Tools Increases Coverage of Metabolic Annotation," *BMC Genomics* **19**(1), 948. DOI:10.1186/s12864-018-5221-9.

Grigoriev, I. V., et al. 2014. "MycoCosm Portal: Gearing Up for 1000 Fungal Genomes," *Nucleic Acids Research* **42** (Database issue), D699–704. DOI:10.1093/nar/gkt1183.

Hardigan, M. A., et al. 2017. "Genome Diversity of Tuber-Bearing *Solanum* Uncovers Complex Evolutionary History and Targets of Domestication in the Cultivated Potato," *Proceedings of the National Academy of Sciences USA* **114**(46), E9999–10008. DOI:10.1073/pnas.1714380114.

Hirsch, C. N., et al. 2014. "Insights into the Maize Pan-Genome and Pan-Transcriptome," *The Plant Cell* **26**(1), 121–35. DOI:10.1105/tpc.113.119982.

Hutchison, C. A., III, et al. 2016. "Design and Synthesis of a Minimal Bacterial Genome," *Science* **351**(6280), aad6253. DOI:10.1126/science.aad6253.

Jaffe, M., et al. 2019. "Improved Discovery of Genetic Interactions Using CRISPRiSeq Across Multiple Environments," *Genome Research* **29**(4), 668–81. DOI:10.1101/gr.246603.118.

Jastrow, J. D., and R. M. Miller. 1998. "Soil Aggregate Stabilization and Carbon Sequestration: Feedbacks Through Organomineral Associations." In *Soil Processes and the Carbon Cycle*. pp. 207–23. Eds. R . Lal et al. CRC Press LLC, Boca Raton, Florida.

Jiang, Y., et al. 2016. "An Expanded Evaluation of Protein Function Prediction Methods Shows an Improvement in Accuracy," *Genome Biology* **17**(1), 184. DOI: 10.1186/s13059-016-1037-6.

Jiang, Y., et al. 2014. "*Camelina* Seed Quality in Response to Applied Nitrogen, Genotype, and Environment," *Canadian Journal of Plant Science* **94**(5), 971–80. DOI:10.4141/cjps2013-396.

Kagale, S., et al. 2014. "The Emerging Biofuel Crop *Camelina sativa* Retains a Highly Undifferentiated Hexaploid Genome Structure," *Nature Communications* **5**, 3706. DOI:10.1038/ncomms4706.

Karp, P. D., et al. 2016. "Pathway Tools Version 19.0 Update: Software for Pathway/Genome Informatics and Systems Biology," *Briefings in Bioinformatics* **17**(5), 877–90. DOI:10.1093/bib/bbv079.

Keseler, I. M., et al. 2017. "The EcoCyc Database: Reflecting New Knowledge About *Escherichia coli* K–12," *Nucleic Acids Research* **45**(D1), D543–50. DOI:10.1093/nar/gkw1003.

Khaldi, N., et al. 2010. "SMURF: Genomic Mapping of Fungal Secondary Metabolite Clusters," *Fungal Genetics and Biology* **47**(9), 736–41. DOI:10.1016/j.fgb.2010.06.003.

Kindler, L., et al. 2016. "Method-Centered Digital Communities on Protocols.io for Fast-Paced Scientific Innovation," *F1000Research* **5**, 2271. DOI:10.12688/f1000research.9453.2.

Kitagawa, M., et al. 2005. "Complete Set of ORF Clones of *Escherichia coli* ASKA Library (A Complete Set of *E. Coli* K–12 ORF Archive): Unique Resources for Biological Research," *DNA Research* **12**(5), 291–99. DOI:10.1093/dnares/dsi012.

Krallinger, M., et al. 2010. "Analysis of Biological Processes and Diseases Using Text Mining Approaches," In *Bioinformatics Methods in Clinical Research. Methods in Molecular Biology (Methods and Protocols)* **593**, pp. 341–82. Ed. R. Matthiesen. Humana Press, New York, N.Y. DOI:10.1007/978-1-60327-194-3_16.

Krishnakumar, V., et al. 2017. "ThaleMine: A Warehouse for *Arabidopsis* Data Integration and Discovery," *Plant and Cell Physiology* **58**(1), e4. DOI:10.1093/pcp/pcw200.

Kumar, V. S., and C. D. Maranas. 2009. "GrowMatch: An Automated Method for Reconciling *In Silico/In Vivo* Growth Predictions," *PLOS Computational Biology* **5**(3), e1000308. DOI:10.1371/journal.pcbi.1000308.

Kumar, V. S., et al. 2007. "Optimization Based Automated Curation of Metabolic Reconstructions," *BMC Bioinformatics* **8**, 212. DOI:10.1186/1471-2105-8-212.

Kyrpides, N. C., et al. 2016. "Microbiome Data Science: Understanding Our Microbial Planet," *Trends in Microbiology* **24**(6), 425–27. DOI:10.1016/j.tim.2016.02.011.

Li, A., et al. 2018. "Editing of an Alpha-Kafirin Gene Family Increases Digestibility and Protein Quality in Sorghum," *Plant Physiology* **177**(4), 1425–38. DOI:10.1104/pp.18.00200.

Liang, Y., et al. 2018. "Spatially Resolved Proteome Profiling of <200 Cells from Tomato Fruit Pericarp by Integrating Laser-Capture Microdissection with Nanodroplet Sample Preparation," *Analytical Chemistry* **90**(18), 11106–14. DOI:10.1021/acs.analchem.8b03005.

Lin, Y. C., et al. 2018. "Functional and Evolutionary Genomic Inferences in *Populus* through Genome and Population Sequencing of American and European Aspen," *Proceedings of the National Academy of Sciences USA* **115**(46), E10970–78. DOI:10.1073/pnas.1801437115.

Liu, H., et al. 2018. "Magic Pools: Parallel Assessment of Transposon Delivery Vectors in Bacteria," *mSystems* **3**(1), e00143-17. DOI:10.1128/mSystems.00143-17.

Lowe, K., et al. 2016. "Morphogenic Regulators *Baby boom* and *Wuschel* Improve Monocot Transformation," *The Plant Cell* **28**(9), 1998–2015. DOI:10.1105/tpc.16.00124.

Maglott, D. R. 2000. "NCBI's LocusLink and RefSeq," *Nucleic Acids Research* **28**(1), 126–28. DOI:10.1093/nar/28.1.126.

Magrane, M., and UniProt Consortium. 2010. "UniProt Knowledgebase: A Hub of Integrated Data," *Nature Precedings*. DOI:10.1038/npre.2010.5092.1.

Malik, M. R., et al. 2018. "*Camelina sativa*, an Oilseed at the Nexus Between Model System and Commercial Crop," *Plant Cell Reports* **37**(10), 1367–81. DOI:10.1007/s00299-018-2308-3.

Marshall-Colon, A., et al. 2017. "Crops *In Silico*: Generating Virtual Crops Using an Integrative and Multi-Scale Modeling Platform," *Frontiers in Plant Science* **8**, 786. DOI:10.3389/fpls.2017.00786.

Medema, M. H., et al. 2011. "antiSMASH: Rapid Identification, Annotation, and Analysis of Secondary Metabolite Biosynthesis Gene Clusters in Bacterial and Fungal Genome Sequences," *Nucleic Acids Research* **39**(Suppl 2), W339–46. DOI:10.1093/nar/gkr466.

Merchant, S. S., et al. 2007. "The *Chlamydomonas* Genome Reveals the Evolution of Key Animal and Plant Functions," *Science* **318**(5848), 245–50. DOI:10.1126/science.1143609.

Meyer, F., et al. 2009. "FIGfams: Yet Another Set of Protein Families," *Nucleic Acids Research* **37**(20), 6643–54. DOI:10.1093/nar/gkp698.

Mira, A., et al. 2010. "The Bacterial Pan-Genome: A New Paradigm in Microbiology," *International Microbiology* **13**(2), 45–57. DOI:10.2436/20.1501.01.110.

Monk, J. M., et al. 2017. "*i*ML1515, a Knowledgebase that Computes *Escherichia coli* Traits," *Nature Biotechnology* **35**(10), 904–08. DOI:10.1038/nbt.3956.

Montelione, G. T. 2012. "The Protein Structure Initiative: Achievements and Visions for the Future," *F1000 Biology Reports* **4**, 7. DOI:10.3410/B4-7.

Moore, B. M., et al. 2019. "Robust Predictions of Specialized Metabolism Genes Through Machine Learning," *Proceedings of the National Academy of Sciences USA* **116**(6), 2344–53. DOI:10.1073/pnas.1817074116.

Morineau, C., et al. 2017. "Selective Gene Dosage by CRISPR-Cas9 Genome Editing in Hexaploid *Camelina sativa*," *Plant Biotechnology Journal* **15**(6), 729–39. DOI:10.1111/pbi.12671.

Mukherjee, S., et al. 2013. "Proteomic Analysis of Frozen Tissue Samples Using Laser Capture Microdissection." In *Proteomics for Biomarker Discovery. Methods in Molecular Biology (Methods and Protocols)* **1002**, pp. 71–83. Eds. M. Zhou and T. Veenstra. Humana Press, New York, N.Y. DOI:10.1007/978-1-62703-360-2_6.

Mutalik, V. K., et al. 2019. "Dual-Barcoded Shotgun Expression Library Sequencing for High-Throughput Characterization of Functional Traits in Bacteria," *Nature Communications* **10**(1), 308. DOI:10.1038/s41467-018-08177-8.

Nichols, R. J., et al. 2011. "Phenotypic Landscape of a Bacterial Cell," *Cell* **144**(1), 143–56. DOI:10.1016/j.cell.2010.11.052.

Nystedt, B., et al. 2013. "The Norway Spruce Genome Sequence and Conifer Genome Evolution," *Nature* **497**(7451), 579–84. DOI:10.1038/nature12211.

Obour, A. K., et al. 2017. "*Camelina* Seed Yield and Fatty Acids as Influenced by Genotype and Environment," *Agronomy Journal* **109**(3), 947–56. DOI:10.2134/agronj2016.05.0256.

O'Malley, R. C., et al. 2016. "Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape," *Cell* **165**(5), 1280–92. DOI: 10.1016/j.cell.2016.04.038.

Ovchinnikov, S., et al. 2015. "Large-Scale Determination of Previously Unsolved Protein Structures Using Evolutionary Information," *eLife* **4**, e09248. DOI:10.7554/eLife.09248.

Overbeek, R., et al. 2014. "The SEED and the Rapid Annotation of Microbial Genomes Using Subsystems Technology (RAST)," *Nucleic Acids Research* **42**(Database issue), D206–14. DOI:10.1093/nar/gkt1226.

Panchy, N., et al. 2016. "Evolution of Gene Duplication in Plants," *Plant Physiology* **171**(4), 2294–2316. DOI:10.1104/pp.16.00523.

Peters, J. M., et al. 2019. "Enabling Genetic Analysis of Diverse Bacteria with Mobile-CRISPRi," *Nature Microbiology* **4**(2), 244–50. DOI:10.1038/s41564-018-0327-z.

Peters, J. M., et al. 2016. "A Comprehensive, CRISPR-Based Functional Analysis of Essential Genes in Bacteria," *Cell* **165**(6), 1493–1506. DOI:10.1016/j.cell.2016.05.003.

Price, M. N., and A. P. Arkin. 2017. "PaperBLAST: Text Mining Papers for Information About Homologs," *mSystems* **2**(4), e00039-17. DOI:10.1128/mSystems.00039-17.

Price, M. N., et al. 2018a. "Mutant Phenotypes for Thousands of Bacterial Genes of Unknown Function," *Nature* **557**(7706), 503–509. DOI:10.1038/s41586-018-0124-0.

Price, M. N., et al. 2018b. "Filling Gaps in Bacterial Amino Acid Biosynthesis Pathways with High-Throughput Genetics," *PLOS Genetics* **14**(1), e1007147. DOI:10.1371/journal.pgen.1007147.

Price, M. N., et al. 2013. "Indirect and Suboptimal Control of Gene Expression is Widespread in Bacteria," *Molecular Systems Biology* **9**, 660. DOI:10.1038/msb.2013.16.

Rantasalo, A., et al. 2018. "A Universal Gene Expression System for Fungi," *Nucleic Acids Research* **46**(18), e111. DOI:10.1093/nar/gky558.

Reed, J. L., et al. 2006. "Systems Approach to Refining Genome Annotation," *Proceedings of the National Academy of Sciences USA* **103**(46), 17480–84. DOI:10.1073/pnas.0603364103.

Riechmann, J. L., et al. 2000. "*Arabidopsis* Transcription Factors: Genome-Wide Comparative Analysis Among Eukaryotes," *Science* **290**(5499), 2105–10. DOI:10.1126/science.290.5499.2105.

Rousset, F., et al. 2018. "Genome-Wide CRISPR-dCas9 Screens in *E. coli* Identify Essential Genes and Phage Host Factors," *PLOS Genetics* **14**(11), e1007749. DOI:10.1371/journal.pgen.1007749.

Rubin, B. E., et al. 2015. "The Essential Gene Set of a Photosynthetic Organism," *Proceedings of the National Academy of Sciences USA* **112**(48), E6634–43. DOI:10.1073/pnas.1519220112.

Ryu, K. H., et al. 2019. "Single-Cell RNA Sequencing Resolves Molecular Relationships Among Individual Plant Cells," *Plant Physiology* **179**(4), 1444–56. DOI:10.1104/pp.18.01482.

Sannigrahi, P., et al. 2010. "Poplar as a Feedstock for Biofuels: A Review of Compositional Characteristics," *Biofuels, Bioproducts and Biorefining* **4**(2), 209–26. DOI:10.1002/bbb.206.

Sasse, J., et al. 2019. "Multilab EcoFAB Study Shows Highly Reproducible Physiology and Depletion of Soil Metabolites by a Model Grass," *New Phytologist* **222**(2), 1149–60. DOI:10.1111/nph.15662.

Schachner, L. F., et al. 2019. "Standard Proteoforms and Their Complexes for Native Mass Spectrometry," *Journal of the American Society for Mass Spectrometry* **30**(7), 1190–98. DOI:10.1007/s13361-019-02191-w.

Schatz, M. C., et al. 2014. "Whole Genome *De Novo* Assemblies of Three Divergent Strains of Rice, *Oryza sativa*, Document Novel Gene Space of *aus* and *indica*," *Genome Biology* **15**(11), 506. DOI:10.1186/s13059-014-0506-z.

Schlecht, U., et al. 2017. "A Scalable Double-Barcode Sequencing Platform for Characterization of Dynamic Protein-Protein Interactions," *Nature Communications* **8**, 15586. DOI:10.1038/ncomms15586.

Schnoes, A. M., et al. 2009. "Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies," *PLOS Computational Biology* **5**(12), e1000605. DOI:10.1371/journal.pcbi.1000605.

Seaver, S. M. D., et al. 2018. "PlantSEED Enables Automated Annotation and Reconstruction of Plant Primary Metabolism with Improved Compartmentalization and Comparative Consistency," *The Plant Journal* **95**(6), 1102–13. DOI:10.1111/tpj.14003.

Sévin, D. C., et al. 2016. "Nontargeted *In Vitro* Metabolomics for High-Throughput Identification of Novel Enzymes in *Escherichia coli*," *Nature Methods* **14**(2), 187–94. DOI:10.1038/nmeth.4103.

Sharma, M. K., et al. 2012. "A Genome-Wide Survey of Switchgrass Genome Structure and Organization," *PLOS One* **7**(4), e33892. DOI:10.1371/journal.pone.0033892.

Simo, C., et al. 2014. "Metabolomics of Genetically Modified Crops," *International Journal of Molecular Sciences* **15**(10), 18941–66. DOI:10.3390/ijms151018941.

Simon, M., et al. 2013. "Tissue-Specific Profiling Reveals Transcriptome Alterations in *Arabidopsis* Mutants Lacking Morphological Phenotypes," *The Plant Cell* **25**(9), 3175–85. DOI:10.1105/tpc.113.115121.

Skinner, O. S., et al. 2018. "Top-Down Characterization of Endogenous Protein Complexes with Native Proteomics," *Nature Chemical Biology* **14**(1), 36–41. DOI:10.1038/nchembio.2515.

Stanton, B. J., et al. 2010. "*Populus* Breeding: From the Classical to the Genomic Approach." In *Genetics and Genomics of Populus* **8**, pp. 309–48. Eds. S. Jansson, et al. Springer New York, N.Y. DOI:10.1007/978-1-4419-1541-2_14.

Sumner, L. W., et al. 2007. "Proposed Minimum Reporting Standards for Chemical Analysis; Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI)," *Metabolomics* **3**(3), 211–21. DOI:10.1007/s11306-007-0082-2.

Suzek, B. E., et al. 2007. "UniRef: Comprehensive and Non-Redundant UniProt Reference Clusters," *Bioinformatics* **23**(10), 1282–88. DOI:10.1093/bioinformatics/btm098.

Tao, Y., et al. 2019. "Exploring and Exploiting Pan-Genomics for Crop Improvement," *Molecular Plant* **12**(2), 156–69. DOI:10.1016/j.molp.2018.12.016.

Telenti, A., et al. 2018. "Deep Learning of Genomic Variation and Regulatory Network Data," *Human Molecular Genetics* **27**(R1), R63–71. DOI:10.1093/hmg/ddy115.

Tong, A. H., et al. 2004. "Global Mapping of the Yeast Genetic Interaction Network," *Science* **303**(5659), 808–13. DOI:10.1126/science.1091317.

Tuskan, G. A., et al. 2006. "The Genome of Black Cottonwood, *Populus trichocarpa* (Torr. & Gray)," *Science* **313**(5793), 1596–604. DOI:10.1126/science.1128691.

U.S. DOE. 2017. *Technologies for Characterizing Molecular and Cellular Systems Relevant to Bioenergy and the Environment*, DOE/SC-0189, U.S. Department of Energy Office of Science [https://genomicscience.energy.gov/technologies/].

U.S. DOE Joint Genome Institute. 2018. *Progress Report. U.S. Department of Energy Joint Genome Institute* [https://jgi.doe.gov/wp-content/uploads/2019/03/2018_Progress_Report_online.pdf].

van Opijnen, T., et al. 2009. "Tn-Seq: High-Throughput Parallel Sequencing for Fitness and Genetic Interaction Studies in Microorganisms," *Nature Methods* **6**(10), 767–72. DOI:10.1038/nmeth.1377.

Vollmann, J., et al. 2007. "Agronomic Evaluation of *Camelina* Genotypes Selected for Seed Quality Characteristics," *Industrial Crops and Products* **26**(3), 270–77. DOI:10.1016/j.indcrop.2007.03.017.

Waese, J., et al. 2017. "ePlant: Visualizing and Exploring Multiple Levels of Data for Hypothesis Generation in Plant Biology," *The Plant Cell* **29**(8), 1806–21. DOI:10.1105/tpc.17.00073.

Wang, Y., et al. 2013. "The Sacred Lotus Genome Provides Insights into the Evolution of Flowering Plants," *Plant Journal* **76**(4), 557–67. DOI:10.1111/tpj.12313.

Warner, J. R., et al. 2010. "Rapid Profiling of a Microbial Genome Using Mixtures of Barcoded Oligonucleotides," *Nature Biotechnology* **28**(8), 856–62. DOI:10.1038/nbt.1653.

Wattam, A. R., et al. 2018. "Assembly, Annotation, and Comparative Genomics in PATRIC, the All Bacterial Bioinformatics Resource Center." In *Comparative Genomics. Methods in Molecular Biology* **1704**, pp. 79–101. Eds. J. Setubal, et al. Humana Press, New York, N.Y. DOI:10.1007/978-1-4939-7463-4_4.

Wattam, A. R., et al. 2017. "Improvements to PATRIC, the All-Bacterial Bioinformatics Database and Analysis Resource Center," *Nucleic Acids Research* **45**(D1), D535–42. DOI:10.1093/nar/gkw1017.

Weigele, P., and E. A. Raleigh. 2016. "Biosynthesis and Function of Modified Bases in Bacteria and Their Viruses," *Chemical Reviews* **116**(20), 12655–87. DOI:10.1021/acs.chemrev.6b00114.

Weiner, J. 2004. "Allocation, Plasticity and Allometry in Plants," *Perspectives in Plant Ecology, Evolution and Systematics* **6**(4), 207–15. DOI:10.1078/1433-8319-00083.

Wetmore, K. M., et al. 2015. "Rapid Quantification of Mutant Fitness in Diverse Bacteria by Sequencing Randomly Bar-Coded Transposons," *mBio* **6**(3), e00306–15. DOI:10.1128/mBio.00306-15.

Whidbey, C., et al. 2019. "A Probe-Enabled Approach for the Selective Isolation and Characterization of Functionally Active Subpopulations in the Gut Microbiome," *Journal of the American Chemical Society* **141**(1), 42–47. DOI:10.1021/jacs.8b09668.

Wilkinson, M. D., et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship," *Scientific Data* **3**, 160018. DOI:10.1038/sdata.2016.18.

Wu, D., et al. 2009. "A Phylogeny-Driven Genomic Encyclopaedia of Bacteria and Archaea," *Nature* **462**(7276), 1056–60. DOI:10.1038/nature08656.

York, L. M. 2019. "Functional Phenomics: An Emerging Field Integrating High-Throughput Phenotyping, Physiology, and Bioinformatics," *Journal of Experimental Botany* **70**(2), 379–86. DOI:10.1093/jxb/ery379.

Yu, P., et al. 2014. "Phenotypic Plasticity of the Maize Root System in Response to Heterogeneous Nitrogen Availability," *Planta* **240**(4), 667–78. DOI:10.1007/s00425-014-2150-y.

Yu, Z., et al. 2018. "Droplet-Based Microfluidic Analysis and Screening of Single Plant Cells," *PLOS One* **13**(5), e0196810. DOI:10.1371/journal.pone.0196810.

Zengler, K., and L. S. Zaramela. 2018. "The Social Network of Microorganisms—How Auxotrophies Shape Complex Communities," *Nature Reviews Microbiology* **16**(6), 383–90. DOI:10.1038/s41579-018-0004-5.

Zhalnina, K., et al. 2018. "Need for Laboratory Ecosystems to Unravel the Structures and Functions of Soil Microbial Communities Mediated by Chemistry," *mBio* **9**(4), e01175-18. DOI:10.1128/mBio.01175-18.

Zhang, G., et al. 2012. "A Mimicking-of-DNA-Methylation-Patterns Pipeline for Overcoming the Restriction Barrier of Bacteria," *PLOS Genetics* **8**(9), e1002987. DOI:10.1371/journal.pgen.1002987.

Zhang, X., et al. 2016. "Assignment of Function to a Domain of Unknown Function: Duf1537 is a New Kinase Family in Catabolic Pathways for Acid Sugars," *Proceedings of the National Academy of Sciences USA* **113**(29), E4161–69. DOI: 10.1073/pnas.1605546113.

Zhou, M., et al. 2018. "Surface Induced Dissociation Coupled with High Resolution Mass Spectrometry Unveils Heterogeneity of a 211 kDa Multicopper Oxidase Protein Complex," *Journal of the American Society for Mass Spectrometry* **29**(4), 723–33. DOI:10.1007/s13361-017-1882-x.

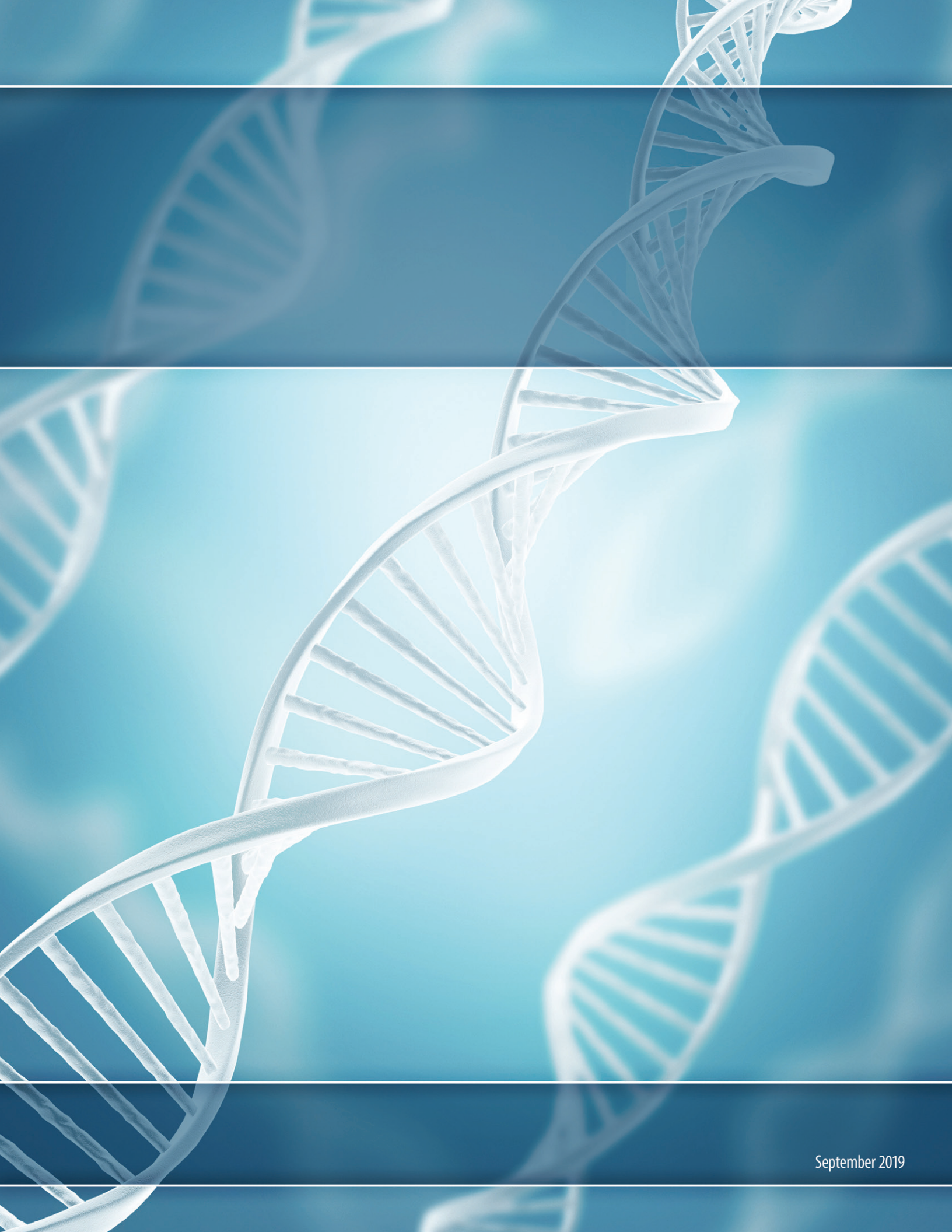
Zhu, Y., et al. 2018a. "Proteomic Analysis of Single Mammalian Cells Enabled by Microfluidic Nanodroplet Sample Preparation and Ultrasensitive NanoLC-MS," *Angewandte Chemie International Edition* **57**(38), 12370–74. DOI:10.1002/anie.201802843.

Zhu, Y., et al. 2018b. "Nanodroplet Processing Platform for Deep and Quantitative Proteome Profiling of 10–100 Mammalian Cells," *Nature Communications* **9**(1), 882. DOI:10.1038/s41467-018-03367-w.

Zotter, A., et al. 2017. "Quantifying Enzyme Activity in Living Cells," *Journal of Biological Chemistry* **292**(38), 15838–48. DOI:10.1074/jbc.M117.792119.

Appendix E: Acronyms and Abbreviations

ABPP	activity-based protein profiling
AI	artificial intelligence
BER	DOE Office of Biological and Environmental Research
BLAST	basic local alignment search tool
CRISPR	clustered regularly interspaced short palindromic repeats
CRISPR/Cas	targeted genome editing system using engineered nucleases (e.g., Cas9, dCas9)
cryo-EM	cryo-electron microscopy
DOE	U.S. Department of Energy
DUF	domain of unknown function
EcoFAB	fabricated ecosystem
EMSL	DOE Environmental Molecular Sciences Laboratory
FAIR	findable, accessible, interoperable, and reusable data principles
G × E	genotype by environment
GRN	gene regulatory network
IMG/M	JGI Integrated Microbial Genomes and Microbiomes system
JGI	DOE Joint Genome Institute
KBase	DOE Systems Biology Knowledgebase
NCBI	National Center for Biotechnology Information
NERSC	National Energy Research Scientific Computing Center
NMDC	National Microbiome Data Collaborative
PATRIC	Pathosystems Resource Integration Center
PDB	Protein Data Bank
SSN	sequence similarity network
Tn-seq	transposon sequencing



September 2019