Breaking the Bottleneck of Genomes: Understanding Gene Function Across Taxa

EXECUTIVE SUMMARY

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

Dawn Adin
U.S. Department of Energy

Catherine Ronning
U.S. Department of Energy

This report will be available at genomicscience.energy.gov/genefunction/.

About BER

The Biological and Environmental Research 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.

Breaking the Bottleneck of Genomes: Understanding Gene Function Across Taxa

April 2019

U.S. DEPARTMENT OF ENERGY | Office of Science

Office of Biological and Environmental Research

Executive Summary

In the last few decades, high-throughput 'omics technologies have enabled unprecedented views of biological systems at the molecular level. In parallel, the integration of these omics 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 the systems biology era, which can be viewed as a holistic approach to decipher the complexity of biological systems. However, as high-throughput omics technologies and downstream systems biology efforts have improved 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 gene function knowledge. 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 through the use of 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 these diverse data 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 for each of these four areas: technology, computation, microorganisms, and plants. Each of these areas is framed by the charge questions posed to all participants and discussed at the workshop.

Technology Innovations

Multiple technology innovations will be required to advance gene function identification at multiple levels of "function," including a detailed biochemical understanding of protein function and the importance of a gene in an organism's physiology and in an ecosystem process. 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, organism, and ecosystem scale.

Opportunities for closing the technology innovation gap include the application of mature omics 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, centralizing a subset of strategic omics technologies at a consortium of discovery centers could facilitate rapid method optimization and dataset generation, reduce costs and duplication of efforts, promote method standardization, and enable seamless integration of knowledge such as orthogonal validation of findings. In addition, transformative technology innovations to break the genome bottleneck will require novel, yet-to-be-developed approaches 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 homology 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 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 omics 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 computational resources in the DOE space 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 omics 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. Positive outcomes could be the automated (or semi-automated) 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 genetic method development 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 ultra-small volume biochemical assays (e.g., those that can be encapsulated within droplets at a massive scale) offer great promise for characterizing function from diverse protein families, including from uncultivated microorganisms. Lastly, expanding gene function determination from standard laboratory conditions to the natural environments in which microorganisms evolved offers a new dimension for understanding gene function.

Plants

The characterization of gene function in flowering plants presents unique challenges relative to single-celled microorganisms. These challenges are attributable to the size of plant genomes, complexity of organs within plant structures, heterogeneity of environments in which plants exist, 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), camelina (nonfood oilseed), poplar (perennial, C3 biofuel feedstock), and *Arabidopsis thaliana* and *Chlamydomonas reinhardtii* (model species).

Central to all downstream efforts is access to well-annotated genomes with associated large-scale datasets that are user friendly. The focal species, listed in the previous paragraph, are amenable to genetic transformation and have sequenced and annotated genomes, along with a subset of omics 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 single repository with appropriate standards and tools for data analysis and interpretation. Cataloguing relevant omics 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 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. A key opportunity for assessing gene function in plants would be engineering a minimal plant genome to serve as a chassis for testing gene function by combinatorially adding or disrupting genes and gene cassettes and assessing function in a transformative way. This approach would greatly facilitate synthetic biology in plant systems and provide a platform for engineering novel biochemical function.

Breaking the Genome Bottleneck

BER-relevant microorganisms and plants contain an amazing diversity of discovered DNA sequence space to date, 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 of 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.

This gene function annotation challenge is as much a computational challenge as an experimental one. Multiple annotation tools, protein sequence databases, and decades of molecular biology research already exist. How to best leverage and couple these valuable resources to new, transformative datasets and data analysis tools will be important to 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 impact BER's mission by accelerating the development of genomic-enabled solutions to global challenges in sustainable energy development and environmental management.

Workshop Agenda

Breaking the Bottleneck of Genomes: Understanding Gene Function Across Taxa

November 1–2, 2018

Bethesda, Maryland/Bethesda North Marriott

Thursday, November 1

8:00 a.m.	Breakfast
8:30 a.m. – 9:00 a.m.	**Welcome, Introduction, and Overview** — DOE 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) and 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) and 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.	**Work on Slides from Breakout Session I**
9:15 a.m. – 10:00 a.m.	**Work on Slides from Breakout Session II**
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.	**Discussion and Wrap-Up** (working lunch)
1:15 p.m.	**Participants Adjourn**
	Discussion and Writing Session — Co-chairs, breakout leads, and DOE BER staff

Workshop Participants

Co-Chairs

Robin Buell
Michigan State University

Adam Deutschbauer
Lawrence Berkeley National Laboratory

Participants

Adam Abate
University of California, San Francisco

Josh 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; DOE Joint Genome Institute

Andrei Osterman
Sanford Burnham Prebys Medical Discovery Institute

James Schnable
University of Nebraska

Jeffrey Skerker
University of California, Berkeley; Joint BioEnergy Institute; Lawrence Berkeley National Laboratory

Michael Udvardi
Noble Foundation

Dan Voytas
University of Minnesota

Judy Wall
University of Missouri

Dave Weston
Oak Ridge National Laboratory