Integrative Computational Tools for Systems Biology Research

Summary of projects awarded in 2023 under Funding Opportunity Announcement DE-FOA-0002878

Genomic Science Program genomicscience.energy.gov

Projects

- Integration of Computational Tools to Explore the Diversity of Temporal Regulation in Plant-Specialized Metabolism
- Expanding Python Library scikit-bio™ for Efficient Multiomic Data Integration and Complex Community Modeling
- A Deep-Learning Predictive Analytics Platform for Plant Genomics
- Toward Metagenome-Scale Metabolic Flux and Free-Energy Analysis via Deep Learning
- Developing Software Tools for the Integration of Genotype-Specific RNA-Splicing Variants, Microexon Alternative Splicing, and Phenotypic Variation in Plant Populations

Contact

BER Program Manager
Ramana Madupu
301.366.2916
Ramana.Madupu@science.doe.gov

Websites

BER Genomic Science Program
genomicscience.energy.gov

DOE Biological and Environmental Research Program
science.osti.gov/ber

DOE Office of Science
energy.gov/science

GSP Computational Biology
genomicscience.energy.gov/compbio

KBase
kbase.us

National Microbiome Data Collaborative
microbiomedata.org

The Biological and Environmental Research program (BER), within the U.S. Department of Energy (DOE) Office of Science, supports basic research to understand the fundamental nature of biological processes relevant to DOE energy and environmental mission goals. Within BER, the Genomic Science Program (GSP) supports systems biology research on microbes, plants, plant-microbe interactions, and environmental microbial communities to address DOE's mission in sustainable bioenergy development. Understanding and harnessing the metabolic and regulatory networks of plants and microbes will enable their design and re-engineering for improved energy resilience and sustainability, including advanced biofuels and bioproducts.

The widespread adoption of high-throughput, multiomic techniques has revolutionized biological research, enabling a broader view and deeper understanding of cellular processes and the biological systems they drive. In pursuit of predictive modeling and genome-scale engineering of complex biological systems important for bioenergy, GSP-supported research generates vast amounts of complex omics and other data from a wide range of analytical technologies and experimental approaches. These data span multiple spatiotemporal scales, reflecting the organizational complexities of biological systems, and present significant computational challenges for identifying causal variants that influence phenotype. Accurate modeling of the underlying systems biology depends on surmounting those challenges.

Collective characterization and quantification of pools of biological molecules (genomics, transcriptomics, proteomics, metabolomics) and their systems processes are essential to constructing coherent knowledge of systems underpinning and governing the diverse phenomics and functioning of plants, microbes, and their communities. Such characterizations necessitate the ability to combine heterogeneous datasets, integrated over time and space, and to represent emergent relationships in a coherent framework.

The breadth of data types and the complexities inherent in the integration of different data layers present significant conceptual and implementation challenges. New algorithms for incorporating data derived from innovations in genomics, molecular imaging, structural biology, and spectroscopy are needed to work effectively with, and glean useful insights from, complex integrated molecular omics data. Computational simulation and rigorous hypothesis testing depend on the ability to incorporate multiple experimental and environmental conditions as well as associated sets of metadata.

In fiscal year 2023, BER solicited applications proposing innovative computational solutions that integrate large, disparate data types from multiple and varied sources and/or the integration of data to achieve coordinated knowledge or integration of knowledge to decipher relationships of biological systems of relevance to DOE. The program sought novel computational tools and analytical approaches to large-scale, multimodal, and multiscale data that will lead to scalable solutions for omics analysis, data mining, and knowledge extraction from complex experimental and calculated datasets. Also sought were interoperable bioinformatics tools or computational applications effective for computationally intensive data processing and analyses for systems-level investigations. To aid the interpretation of multimodal data for environmental sciences, BER encouraged research focused on the enhancement of existing software or approaches already broadly used by the genomics community.

Requested research topics focused on developing toolkits, software, and novel computational, bioinformatic, statistical, algorithmic, or analytical approaches for:

- Derivation of a systems-level understanding of microbial cultures and communities from orthogonal datasets via development of integrated networks and computational models.
- Data mining and comparative analysis across large-scale datasets to infer microbial community composition and interactions or microbial community analysis to handle a wide range of functional genomics data types.
- Data mining and comparative plant genomics or multiomics to facilitate gene function discovery, investigate evolutionary relationships at the genome scale, and/or identify candidate gene and regulatory networks that influence plant adaptability to the environment.
- Development of innovative computational strategies to enhance, scale, and optimize management and processing throughput of large, complex, and heterogeneous biological data generated across scales for integration and interpretation.
- Data integration approaches and new software frameworks for management and analysis of large-scale, multimodal, and multiscale data that enhance transparency of approach, effectiveness, and efficiency of data processing.
- Integration of data across two or more biochemical and biophysical measurements, omics data, and image data to provide insights into fundamental biological processes and to identify novel biological paradigms. For example, converting information from images to enable integration with other data types or integrating genomic, biophysical, and biochemical data.

BER places high priority on ensuring that any scientific software developed under its research awards is made freely available, easy-to-use, open-access, and user-friendly. These objectives may be met through integrating software into the DOE Systems Biology Knowledgebase computational platform (KBase; kbase.us). The National Microbiome Data Collaborative (NMDC; microbiomedata.org) makes data streams freely available to test software. Applications could include plans to integrate software into the KBase computational platform or test the software on data streams available on the NMDC portal.

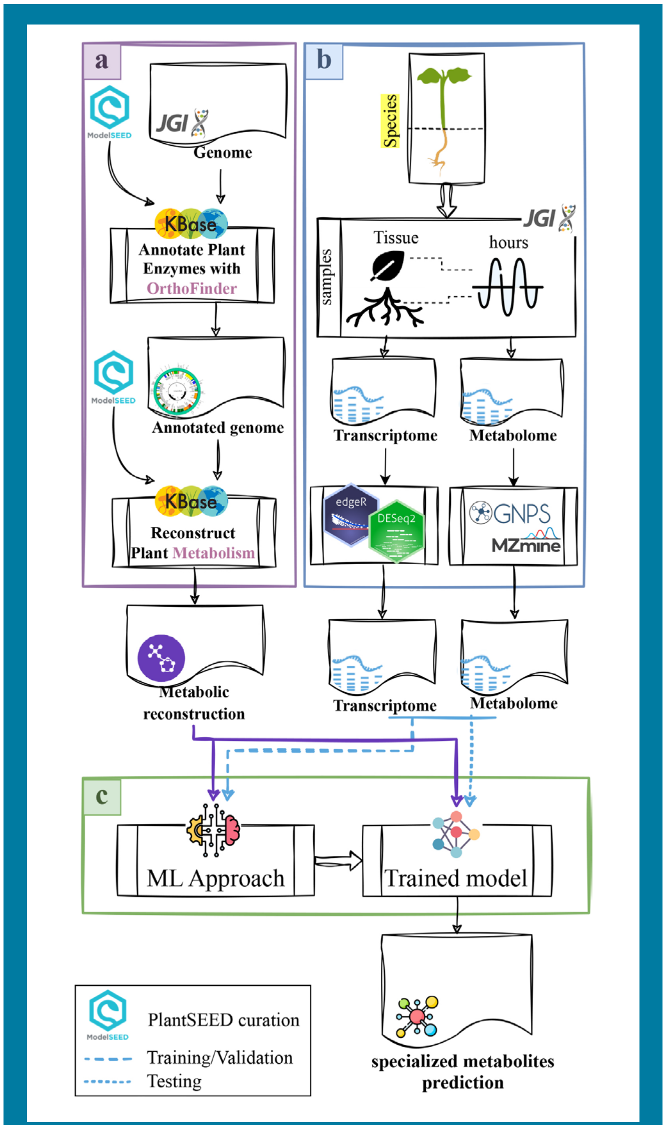
Integration of Computational Tools to Explore the Diversity of Temporal Regulation in Plant-Specialized Metabolism

Principal Investigators: Kathleen Greenham (University of Minnesota); Samuel Seaver (Argonne National Laboratory)

Plants produce an amazing diversity of specialized metabolites (SM) that offer many benefits to human society. SMs are essential for pharmaceutical products, non-medicinal applications in the chemical industry, food additives, dyes, perfumes, cosmetics, and nutraceuticals. These products offer the potential to increase the return on investment of current biofuel crops by providing high-value coproducts.

While many specialized metabolic enzymes have been characterized, their spatial and temporal regulation is less understood, creating a challenge for engineering and optimizing metabolite levels. Understanding how diverse plants differentially regulate production of products from the same SM pathway will enable researchers to engineer plants with greater reliability.

The goal of this project is to build a computational tool in the DOE Systems Biology Knowledgebase (KBase) that will enable researchers to integrate transcriptome data with metabolic networks of general and specialized metabolism for different plant species. With this tool, researchers will be able to explore different combinations of SM precursors and identify key enzyme engineering targets. The project aims to build a set of classifiers by applying machine-learning techniques that would enable prediction. The project will focus on the glucosinolate (GSL) class of specialized metabolites within the plant order Brassicales. Project objectives are to (1) experimentally design and benchmark the biosynthesis of multiple GSLs in eight phylogenetically distinct species from diverse families within Brassicales using high-resolution time-series datasets; (2) reconstruct the general and specialized metabolic networks for GSL biosynthesis, enabling the integration of omics data; (3) train and test the model to predict GSL biosynthesis; and (4) use the KBase platform to encode this approach in a series of apps so other researchers may apply it to pathways of interest. To demonstrate the tool's utility for target identification for SM production, virtual and onsite training workshops will be hosted. This will help spur research into engineering plants as platforms for coproduction of biofuels and coproducts and also increase the plant user community on KBase.



Parallel Computational and Experimental Processes for Building a Prediction Tool for Specialized Metabolites. (a) The computational process functionally annotates eight Brassicales genomes from the DOE Joint Genome Institute (JGI) and reconstructs metabolic networks using DOE Systems Biology Knowledgebase (KBase) applications. **(b)** The experimental process samples different species over time to generate transcriptomics and metabolomics data. **(c)** Machine learning integrates experimental data with metabolic reconstructions to generate a trained model predicting the abundance of a consensus set of desired metabolites. [Courtesy Argonne National Laboratory]

Expanding Python Library scikit-bio™ for Efficient Multiomic Data Integration and Complex Community Modeling

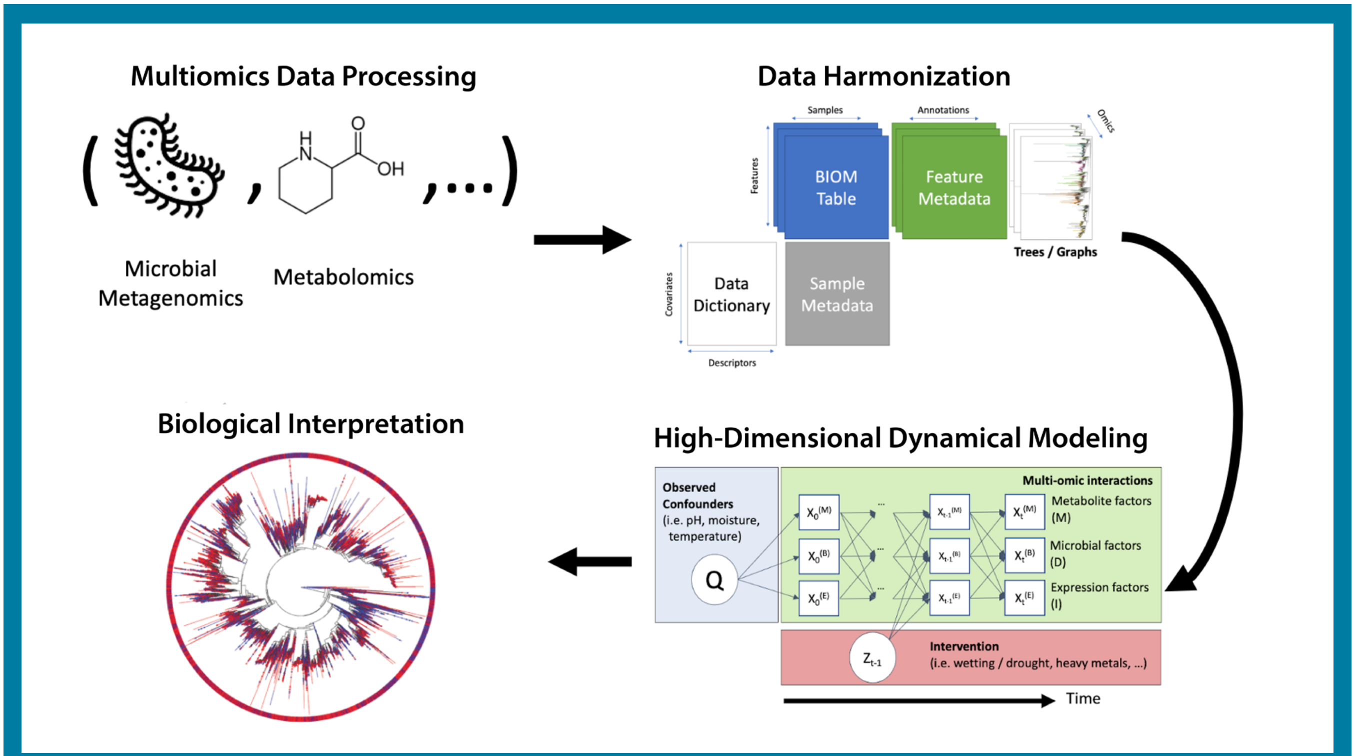
Principal Investigator: Qiyun Zhu (Arizona State University)

Co-Investigators: James Morton (Gutz Analytics); Rob Knight (University of California–San Diego)

This project proposes to expand development of the widely used Python library, scikit-bio, to support large-scale multiomic data analysis that models complex relationships between plants, microbes, and the environment. The team will implement functionalities to enable (1) efficient analysis of heterogeneous data types, (2) multiomic data integration, and (3) biological features annotation.

Microbes play a critical role in maintaining soil health and promoting plant growth by cycling nutrients and facilitating root uptake. Understanding the intricate relationships between plants and microbial symbionts offers opportunities to develop targeted interventions for improving bioenergy productivity and sustainability, such as using microbial inoculants or manipulating soil microbial communities.

Recent studies have shown the merits of gaining insights into environmental and plant-associated microbial communities using multiomics data integration. However, this approach has limited ability to elucidate the mechanistic understanding required to guide intervention design. This shortcoming arises from the limited capability to process and integrate across omic layers and to model complex



Proposed Mulitomics Analysis Pipeline. The scikit-bio™ package provides a scalable and standardized framework for processing diverse high-throughput omics data types. This framework includes data objects for managing spare feature tables, sample and feature metadata to facilitate advanced statistical modeling, and graphs (trees and networks) to model complex relations among biological features. A suite of tools for multiomic data integration and analysis include the state-space variational autoencoder (SSVAE) model for making inferences using longitudinal, high-dimensional data. The analysis will be further enhanced using graph-driven annotation of features, permitting profound biological interpretation of identified correlations and biomarkers. [Biological interpretation portion reprinted under a Creative Commons (CC-BY 4.0) license from McDonald, D., et al. 2023. "Greengenes2 Unifies Microbial Data in a Single Reference Tree," *Nature Biotechnology*. DOI:10.1038/s41587-023-01845-1.]

relationships. Efficient computational frameworks that connect microbial profiles to functional and molecular profiles are essential for DOE to advance transformative science and innovative solutions.

This project aims to expand the development of scikit-bio, which implements various data structures, algorithms, and metrics for bioinformatics analyses. Scikit-bio powers multiple widely adopted software packages, notably QIIME 2 and Qiita. The project seeks to support efforts to improve the overall usability, efficiency, and robustness of the library and to make it the standard Python library for analyzing multimodal, multiomic biological data.

Specific project aims include:

- **Analyze large-volume and heterogeneous omic data types.** This includes efficient data structures and considers the sparse nature of omics data; transformation and normalization methods that consider compositionality and zero-inflation; and parsers and callers of omic type- and field-specific software tools and databases (e.g., KBase and NMDC). This approach will enable unifying data structures and properties for subsequent integrated analysis.
- **Integrate and analyze multiomic data, including developing a longitudinal multiomics machine-learning model that can infer the effects of environmental stresses on soil microbiomes.** This approach will incorporate phylogenetic information while accounting for high-dimensional, sparse, and compositional multiomics data including metagenomic, metatranscriptomic, and metabolomic data to enable insights into community ecology.
- **Facilitate the functional annotation of biological features.** This includes developing data structures using phylogenetic approaches, ontology-constrained annotation, and k-nearest neighbors classification. A standardized approach for functional annotation will serve as the template for ongoing efforts to annotate biological features such as microbial genomes, proteins, and metabolites. Functionalities will be tested on real-world, large-scale, and multiomic datasets involving microbiomes associated with plants and natural environments generated from previous DOE efforts. Evaluation will be centered on the ability to discover new biological signals and efficacy for modeling complex relationships among communities, organisms, and functions.

The expanded Python library scikit-bio will continue to be released under an open-source license to permit wide adoption. The high-quality codebase will improve the processing and integration of large, multiomic data and enhance the modeling and decoding of complex systems involving plants, microbes, and environments. These capacities are important for understanding and improving bioenergy production, plant biomass decomposition, and environmental sustainability—all of which strengthen energy security and resilience.

A Deep-Learning Predictive Analytics Platform for Plant Genomics

Principal Investigator: Asa Ben-Hur (Colorado State University)

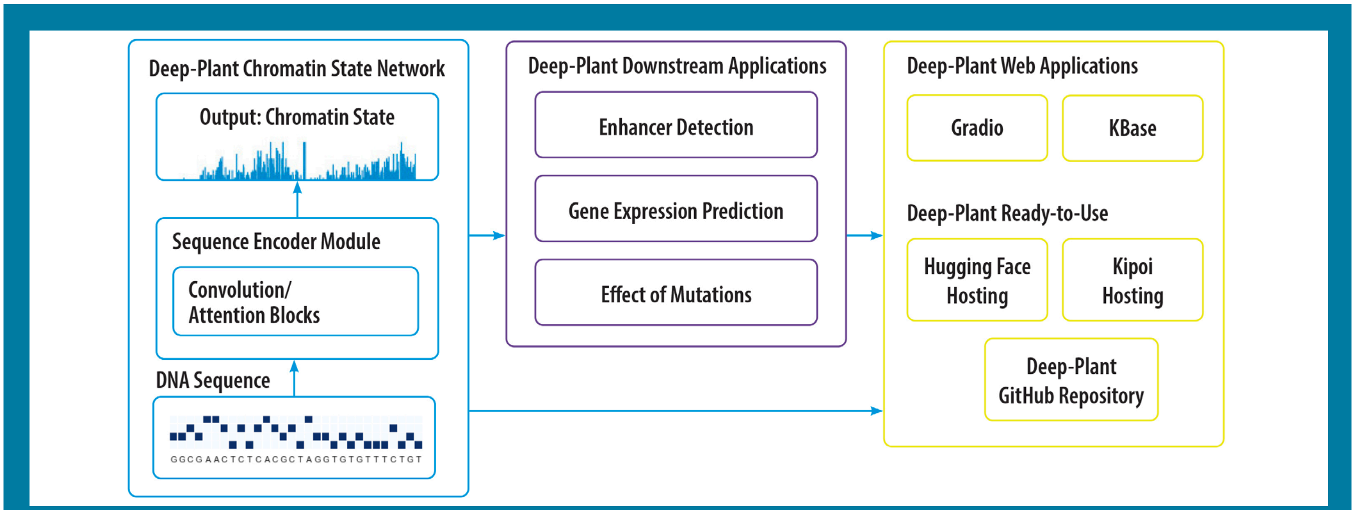
Co-Investigators: Anireddy S.N. Reddy (Colorado State University); Jie Liu (University of Michigan Medical School)

Understanding gene regulation at the molecular level is a major challenge in modern plant biology. It is governed by a multitude of proteins, RNAs, and especially transcription factors. Transcription factors control gene expression by proximal binding to genes in their promoter regions or at distal enhancers. The activity and binding of transcription factors is modulated by the state of the DNA molecule—namely whether it is accessible or wrapped around histones—and by various modifications of DNA and histones.

In recent years, databases providing vast amounts of animal and plant genomics data from various assays have been curated from thousands of published studies. These include genome-wide expression studies (e.g., RNA-seq), determination of transcription-factor binding sites (e.g., ChIP-seq), and various modifications of histones and DNA accessibility (e.g., ChIP-seq, DNase I-seq, ATAC-seq).

Deep learning has demonstrated its value in modeling large and complex genomics compendia in mammals, providing gene regulation insights in those systems. However, very little work has been carried out in plants. This project leverages the wealth of plant data available to create a deep-learning framework called DEEP-PLANT, which will model plant chromatin state and its consequences for gene regulation. Specifically, the DEEP-PLANT model will predict transcription-factor binding and chromatin state directly from sequence in the context of DNA accessibility data. Using such detailed models of chromatin state will enable researchers to model various aspects of gene regulation, including factors regulating gene expression under different conditions and in different tissues, as well as the predictions of enhancers and transcription factors. Finally, the project will demonstrate the value of modeling chromatin state in plants at base-level resolution to predict the effects of genetic variation on expression phenotypes.

This work, carried out in *Arabidopsis* and rice, will shed light on conserved aspects of gene regulation across dicots and monocots. It will also provide plant biologists with tools to form hypotheses on factors driving gene expression. DEEP-PLANT models will be accessible to researchers with varying levels of computational expertise. In addition to the DEEP-PLANT codebase and command-line tools for accessing its functionality, a web server will provide model predictions and impacts of genetic variation. Outreach efforts will concentrate on community education about DEEP-PLANT tools and the value of deep learning for gene-regulation research. The models developed in this project will provide a valuable resource to the plant research community and make powerful deep-learning tools accessible to a wider audience.


Deep-Plant Chromatin State Network
Output: Chromatin State
Sequence Encoder Module
Convolution/ Attention Blocks
DNA Sequence
GGCGAACTCTCACGCTAGGTGTGTTTCTGT
Deep-Plant Downstream Applications
Enhancer Detection
Gene Expression Prediction
Effect of Mutations
Deep-Plant Web Applications
Gradio
KBase
Deep-Plant Ready-to-Use
Hugging Face Hosting
Kipoi Hosting
Deep-Plant GitHub Repository

DEEP-PLANT Framework Predicts Chromatin State, Including Transcription-Factor Binding, Directly from Genetic Sequence. This deep learning framework creates genomic sequence representations for a variety of downstream applications. The approach enables quick and efficient construction of highly accurate gene regulation models and elucidates mutations affecting gene function. The resulting trained models will be widely accessible via web applications and frameworks such as Kipo and Hugging Face. [Courtesy Colorado State University]

Toward Metagenome-Scale Metabolic Flux and Free-Energy Analysis via Deep Learning

Principal Investigators: Junyoung O. Park (University of California–Los Angeles); Pin-Kuang Lai (Stevens Institute of Technology)

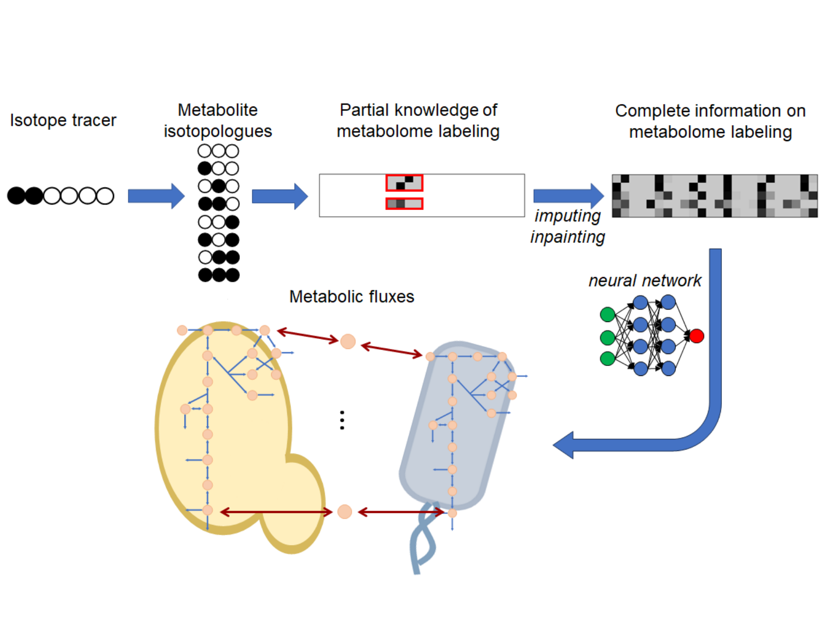
Metabolism is a dynamic network of biochemical reactions that support cell proliferation and biosynthesis. The ability to engineer metabolism for added societal benefits relies on a quantitative understanding of metabolism, which requires comprehensive measurement of its dynamic states. Precisely controlling metabolic pathways would enable efficient and sustainable production of advanced biofuels and bioproducts. However, challenges arise from insufficient ability to quantify metabolic fluxes (i.e., rates at which pathways are utilized) on a systems level.

Metabolic flux distributions are a direct readout of dynamic metabolic state. However, metabolic fluxes are intangible deduced quantities resulting from catalytic interactions between metabolites and enzymes according to kinetic and thermodynamic laws. Because rate (metabolic flux) and energy (Gibbs free energy of reaction) are both related to levels of metabolites and enzymes, systems-level quantification of metabolic fluxes and free energies can facilitate the integration of metabolomics and proteomics. Knowledge of metabolic fluxes and free energies offers dual benefits of laying a solid foundation for metabolic engineering and integrating multiomic data.

The overarching goal of this project is to develop a computational toolset for genome-scale and metagenome-scale quantification of metabolic fluxes and free energies. To effectively achieve this computationally intensive goal, teams at the University of California–Los Angeles and Stevens Institute of Technology will combine deep learning with stable-isotope tracing and simulation techniques. Using multilayer neural networks, the teams will develop deep-learning models that predict metabolic fluxes and free energies from the isotope-labeling patterns of metabolites. With augmented flux and free-energy determination capability, the software will (1) ensure that multiomic data are coherent, (2) impart quantitative systems-level knowledge of metabolism in individual and across multiple species, and (3) reveal precise metabolic control strategies. The software will be open source and freely available to researchers at academic and nonprofit institutions.

Specific project aims include:

- Accelerate and broaden metabolic-flux and free-energy analysis by deep learning.
- Streamline the construction of nonmodel organism–metabolic networks by isotope tracing.
- Deconvolute metabolic fluxes and interactions in microbial communities.



From Isotope Labeling Patterns to Metabolic Fluxes in Microbial Communities. Quantification of metabolic fluxes in microbial communities is an unconquered quest for the metabolism research community. While isotope labeling patterns from isotope-tracing experiments encode metabolic fluxes, measuring and decoding labeling patterns are experimentally and computationally challenging. A combination of metabolic models and machine-learning models streamlines the determination of metabolic network structures and fluxes in new nonmodel organisms and complex microbial communities. The open-source software tools will expand the accessibility of fluxomics to microbiologists, systems biologists, and metabolic engineers. [Courtesy University of California–Los Angeles; Stevens Institute of Technology]

Developing Software Tools for the Integration of Genotype-Specific RNA-Splicing Variants, Microexon Alternative Splicing, and Phenotypic Variation in Plant Populations

Principal Investigator: Chi Zhang (University of Nebraska–Lincoln)

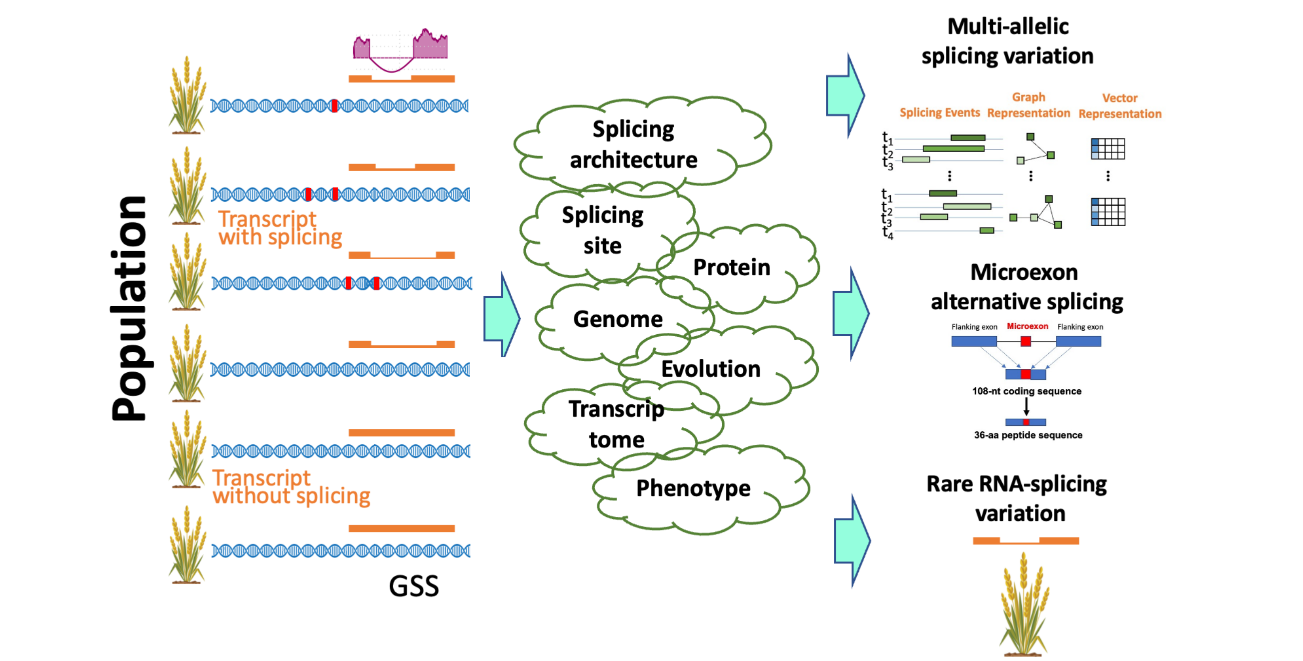
Co-Investigators: Hongfeng Yu (University of Nebraska–Lincoln); Harkamal Walia (University of Nebraska–Lincoln); Bin Yu (University of Nebraska–Lincoln)

Pre-mRNA splicing is an essential step in the regulation of gene expression. Extensive research has been conducted, and many bioinformatics tools have been developed in RNA alternative splicing. Nonetheless, gaps remain. What is lacking are efficient and accurate methods to analyze population transcriptome datasets in order to define and identify the full spectrum of RNA-splicing variants [i.e., genotype-specific RNA splicing (GSS), microexon alternative splicing, and rare RNA-splicing] and link these RNA-splicing variants to phenotype in plant populations. There are open questions, significant knowledge gaps, and requirements for bioinformatic tools regarding the landscape, regulation, and phenotypic outcomes of RNA-splicing variation in plants—especially at the population level and for microexon (exon ≤15 nucleotides) splicing.

Project goals:

- Create broadly applicable software tools to quantify GSS variants based on multiomics data in a population to facilitate gene function discovery.
- Develop frameworks to associate RNA-splicing variants to identify candidate genes that influence plant adaptability to the environment.

Research efforts will center on the hypothesis that a set of GSS variants exists—including rare variants and alternative splicing of microexons—that responds to the phenotypic variation exhibited by organisms during environmental perturbations. This hypothesis is based on a preliminary study that identified GSS in a plant



Identification of Genotype-Specific RNA-Splicing Variants. Bioinformatics tools are needed to address the landscape, regulation, and phenotypic outcomes of RNA-splicing variation in plant populations. An integrated software tool can facilitate gene function discovery by enabling quantification of genotype-specific RNA-splicing variants based on a population's multiomics data. It can also identify candidate genes influencing plant adaptability to the environment by quantifying associate RNA-splicing variants. [Courtesy University of Nebraska–Lincoln] population in response to stress. The study demonstrated that GSS can be utilized to prioritize causal RNA splicing in genome-wide association studies (Yu et al. 2021). More recently, the team accurately identified 2,398 small microexons in 10 diverse plant species using 990 RNA-seq datasets, the majority of which were not annotated in the reference genomes (Yu et al. 2022).

Goals will be accomplished through three objectives: (1) employ deep-learning approaches to enhance the genome-wide analysis of GSS variants in plant populations including both coding and noncoding RNAs; (2) develop a software tool to identify microexon alternative splicing in plant populations and predict corresponding protein functions; (3) develop a framework for identifying rare RNA-splicing variants in plant populations and conducting genome-wide association studies between rare RNA-splicing variants and phenotype.

The expected project outcome is development of integrative computational tools that can analyze GSS variants in plant populations, study microexon alternative splicing, and prioritize causal RNA-splicing variants, including rare RNA-splicing variants, for plant responses to environmental perturbations. This project will advance genomic data analysis at the population level and yield important information on molecular-level mechanisms of RNA splicing responding to the phenotypic response. The proposed software tool will provide an indispensable bridge to connect gene function and expression regulation to physiology and phenotype.

References

Yu, H., et al. 2021. “Genome-Wide Discovery of Natural Variation in pre-mRNA Splicing and Prioritising Causal Alternative Splicing to Salt Stress Response in Rice,” *New Phytologist* **230**(3), 1273–287. DOI:10.1111/nph.17189.

Yu, H., et al. 2022. “Pervasive Misannotation of Microexons that are Evolutionarily Conserved and Crucial for Gene Function in Plants,” *Nature Communications* **13**, 820. DOI:10.1038/s41467-022-28449-8.