# SCINet Newsletter: October 2024

**Research Spotlight  l  News  l  Working Groups
Training  l  Support  l  Connect**

## Research Spotlight

## PanEffect: a tool for analysis of genetic variants in Maize and Fusarium

*Carson M. Andorf[1] and Hye-Seon Kim[2]*

[1]*USDA-ARS, Corn Insects and Crop Genetics Research Unit, Ames, IA 50011.*

[2]*USDA-ARS, Mycotoxin Prevention and Applied Microbiology Research Unit, National Center for Agricultural Utilization Research, Peoria, IL 61604.*
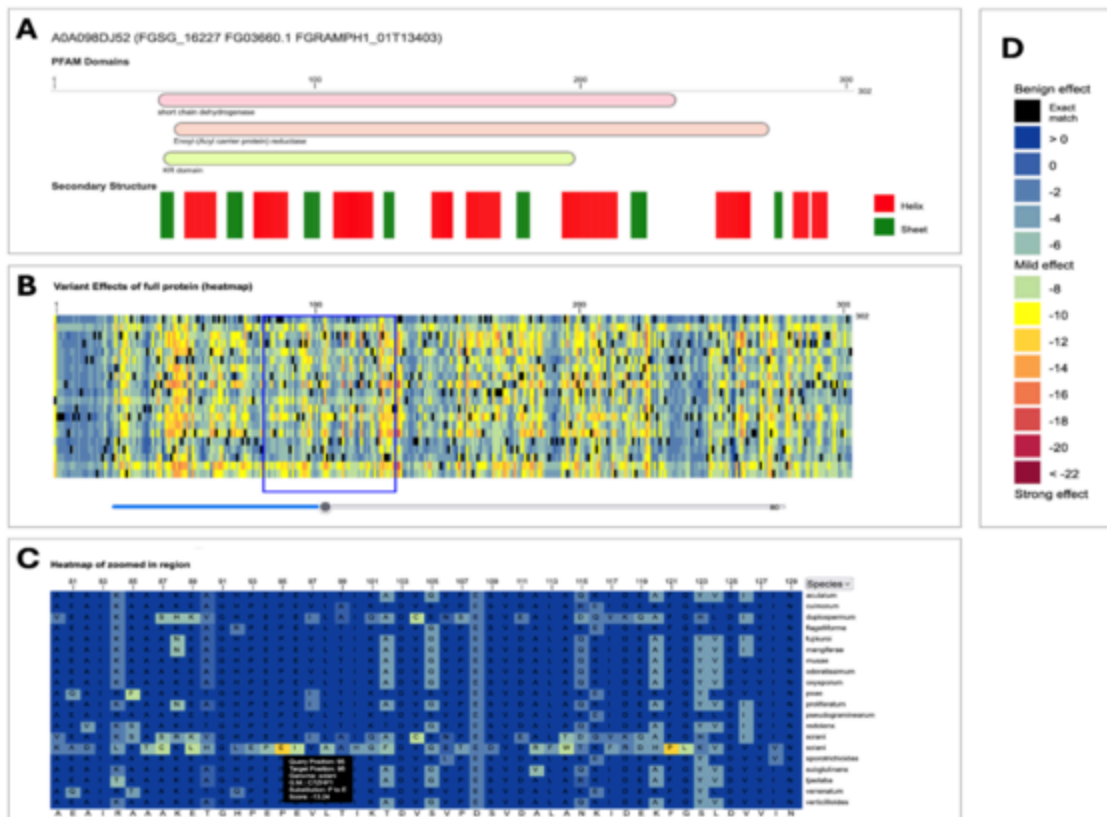
Recent advances in analyzing the structure and function of proteins, driven by deep learning, have given scientists powerful new tools for understanding how the amino acid sequence of a protein affects the phenotype, or observable characteristics, of an organism. These new capabilities have the potential to revolutionize agricultural research by unlocking the connections between proteins and a multitude of agriculturally relevant traits, such as disease resistance, drought resistance, growth potential, and much more. Nevertheless, these tools are not always easy to use and require substantial computing resources and technical know-how. We have developed new software called "PanEffect" that provides a user-friendly toolset for exploring how variations in protein sequence are likely to affect an organism's phenotype.

PanEffect's capabilities are based in part on a state-of-the-art deep learning protein sequence analysis model called ESM-2 (https://doi.org/10.1126/science.ade2574). Using tools built on ESM-2, PanEffect predicts the functional consequences of all known genetic variants of key proteins in maize and *Fusarium*, a damaging fungal pathogen of maize. PanEffect also calculates an "Effect Score" for each variant of a protein that represents how much that variant is expected to affect the organism's phenotype. Effect Scores are displayed with heatmaps that illustrate predicted phenotypic outcomes, ranging from benign to functionally significant.

The maize component of PanEffect visualizes over 550 million potential genetic variants in the well-studied maize line B73, as well as 2.3 million naturally occurring genetic variants in the maize pan-genome. The *Fusarium* component of PanEffect predicts impacts of over 270 million genetic variants in two of the most agriculturally important species of the fungus, *Fusarium graminearum* and *F. verticillioides*. Both fungi cause ear rot of maize but produce different types of toxins. To facilitate the assessment of naturally occurring genetic variation, the tool provides variant effect scores for proteins in a pan-genome constructed from 22 diverse species of *Fusarium*.

PanEffect development relied on the A100 GPU (graphics processing unit) nodes on SCINet's Atlas supercomputer. The ESM-2 model can be run most efficiently on GPU nodes. By utilizing multiple GPUs simultaneously, the project was able to predict hundreds of millions of variant scores across nearly 2.5 million proteins from 54 maize accessions and 22 *Fusarium* species.

PanEffect was developed to aid identification of genetic targets to improve maize quality and to enhance understanding of molecular mechanisms of pathogenesis in *Fusarium*, and thereby, aid efforts to control disease and mycotoxin contamination caused by the fungus. These works were published in the journal Bioinformatics (https://doi.org/10.1093/bioinformatics/btae073) and BMC Microbiology (https://doi.org/10.1186/s12866-024-03480-5). The PanEffect code is available on GitHub at https://github.com/Maize-Genetics-and-Genomics-Database/PanEffect. PanEffect data for maize is available at https://www.maizegdb.org/effect/maize/, and data for Fusarium is available at https://fusarium.maizegdb.org/.



**Figure 1:** Example views of the PanEffect tool for *Fusarium graminearum*.

- Panel A: Displays the Pfam domains and predicted secondary structures for the selected protein.

- Panel B: Shows a heatmap of all possible coding variants, color-coded based on their predicted impact on protein function.

- Panel C: Illustrates a heatmap of naturally occurring variants within a gene family across 22 diverse Fusarium genomes. Variants identical to the reference genome are shown in dark blue, while other variants are color-coded based on their likelihood of affecting protein function.

- Panel D: Provides a legend for the variant effect scores, ranging from scores above -7, indicating benign effects, to scores below -7, suggesting potential phenotypic impacts

# SCINet and AI COE Fellows

Please welcome our newest SCINet and ARS Artificial Intelligence Center of Excellence (AI-COE) fellows!

**Dr. Noelymar Gonzalez-Maldonado** is a USDA-ARS SCINet/AI-COE fellow under the mentorship of Kerri Steenwerth and Amisha Poret-Peterson (USDA-ARS, Davis, CA). Dr. Gonzalez-Maldonado is from Puerto Rico and moved to the United States mainland in 2017 to attend graduate school. She earned a B.S. in Agronomy at the University of Puerto Rico-Mayagüez, an M.S. in Environment and Natural Resources at The Ohio State University, and a Ph.D. in Soils and Biogeochemistry at the University of California - Davis. Her expertise includes soil biogeochemistry, rural sociology, and soil microbial ecology. Previously, her research focused on using a multidisciplinary and participatory approach with growers to assess the variability of soil health indicators and the soil microbiome in vineyards across California. As a USDA-ARS SCINet/AI-COE fellow, Dr. Gonzalez-Maldonado will contribute to a research project focused on using AI and machine learning to predict agricultural soil health outcomes with metagenomic data.

**Dr. Minho Kim** is a SCINet/AI-COE fellow under the mentorship of Dr. Michael Rothrock (USDA-ARS, Athens, GA). His research focuses on leveraging data science tools, such as predictive modeling, to enhance food safety. In his previous work at the University of Illinois, he developed a simulation tool to optimize sampling plans for detecting foodborne pathogens in powdered infant formula. In another project, he conducted a risk assessment to evaluate the risk of specific serotypes and levels of *Salmonella* in raw poultry products.

Currently, Dr. Kim is working to identify important predictors that may affect pathogen loads in the preharvest environment of pasture-raised broilers. Dr. Kim is interested in how differences in virulence among genes or strains of *Salmonella* could be used to inform public health efforts, as well as exploring how reducing overall pathogen loads in the preharvest environment can contribute to lowering public health risks.

# News

## Apply Now to Serve as an AI-COE/SCINet Graduate Student Internship Mentor in 2025!

We are excited to announce that we are accepting applications to serve as an ARS AI-COE/SCINet graduate student internship mentor in 2025!

These remote internships allow graduate students with strong data science and computational skills to spend either a summer or a semester working full time with an ARS mentor (or mentors) on an ARS research project. Each 10-week internship includes a competitive stipend and travel funding for the participant to spend time onsite with their ARS mentor(s). We are working with five partner universities to recruit student participants.

We want each of our interns to have an outstanding experience, which means we need outstanding ARS mentors and research projects! If you are interested in participating, please visit the mentor application page for more information about how to prepare and submit your application.

Applications are due by COB Friday, December 6, 2024.

## FY25 SCINet/AI-COE Postdoctoral Fellowships Call for Proposals

SCINet and the ARS AI-COE are again offering postdoctoral fellowship funding to ARS scientists who wish to mentor SCINet/AI-COE fellows working in their labs.

These fellowships provide an exciting opportunity for participants to address agricultural problems by developing and applying new and emerging scientific computing technologies, including big data analytics, artificial intelligence, and machine learning. Fellows will be able to conduct research in collaboration with ARS scientists, use SCINet's high-performance computing clusters and other computational resources, and access the numerous training opportunities available through SCINet and the AI-COE.

For information and to submit a proposal, please visit the proposal application page. Please note that FY24 awardees are not eligible for an FY25 award.

The deadline for applications is COB on Friday, December 6, 2024.

## FY25 AI Innovation Fund Awards Call for Proposals

The ARS AI-COE is again sponsoring AI Innovation Fund awards to support research projects that apply AI and machine learning (ML) methods to agricultural research or that develop new software tools or data products that use AI or ML techniques. We expect to fund 4 to 6 projects of up to $100,000 each. Funds will need to be spent this fiscal year, so projects should have a short budget timeline or involve partnerships that can be funded through collaborative agreements.

Please visit the AI Innovation Fund page for more information and for application instructions. Please note that FY24 awardees are not eligible for an FY25 award.

The deadline for applications is COB on Friday, December 6, 2024.

## SCINet/AI-COE Fellow Receives ORISE Future of Science Award

Congratulations to Dr. Keo Corak, a former SCINet/AI-COE fellow and now ARS SY, on winning the Oak Ridge Institute for Science and Education (ORISE) Future of Science (Post Doc) Award for their outstanding sugarcane research endeavors. Dr. Keo Corak was nominated by their mentor, ARS scientist Dr. Amanda Hulse-Kemp.

The annual ORISE Future of Science Awards celebrate excellence performed by ORISE participants in the areas of scientific achievement, professional growth, project contributions, and leadership.

---

# Working Group Updates

---

## Breeding AI and ML Working Group

We are pleased to announce the formation of a new SCINet working group, the Breeding AI and ML Working Group!

The focus of this group is to create a space where researchers working on addressing problems in breeding using artificial intelligence (AI) and machine learning (ML) methods can exchange knowledge and build community support. Specifically, this working group will develop and share learning resources for ARS scientists to promote the development and use of AI/ML data and models in breeding research.

The group is intended both for scientists already using SCINet resources and those who would like to expand their use of SCINet in their research.

Please see the working group's web page for more information and contact Dr. Amanda Hulse-Kemp (amanda.hulse-kemp@usda.gov) if you have any questions.

## Arthropod Genomics Research (AGR) Working Group

The October teleconference for the Arthropod Genomics Research (AGR) working group coincided with a National Agriculture Library (NAL)-led training on the preparation of a Data Management Plan (DMP) on October 8, 2024.

This training was tailored for National Program 304 (NP304), Plant Protection and Quarantine. A recording of this event is at NP304 DMP training trimmed V2 20241008.mp4 and the slide deck from this presentation is at 202408_NP304 Data Management Plan Training v1.2.pptx. Additional DMP resources from NAL and outlined in this presentation are available at https://www.nal.usda.gov/data/data-management-planning. These resources will be helpful to all those involved in the writing and review of NP304 Project Plans this fall and winter, and of interest to all ARS scientists and administrators.

If you would like NAL experts to provide feedback on your DMP, please submit your DMP draft and a copy of the project proposal to https://www.nal.usda.gov/ask-question. Please allow at least 5 business days for this service.

## Geospatial Research Working Group

Several group members have leveraged SCINet resources in recent workflows and manuscripts.

Dr. Erika Pierce and team compared ground and satellite estimates of grassland productivity across shortgrass-steppe vegetation types and grazing managements in *Remote Sensing:* https://doi.org/10.3390/rs16152780.  Nikolas Santamaria was an AI-COE/SCINet graduate student intern with the group in 2023 and contributed to this publication.

Dr. Mahesh Maskey contributed to work led by Dr. Amanda Nelson published in *Hydrology* on calibrating a agro-hydrological model under different grazing activities: https://doi.org/10.3390/hydrology11040042.

Drs. Efrain Duarte and Alexander Hernandez reviewed the monitoring of soil moisture dynamics across scales in semi-arid ecosystems in *Applied Sciences:* https://doi.org/10.3390/app14177677. Dr. Duarte is a current SCINet/AI-COE fellow.


## Translational Omics Working Group

The Translational Omics Working Group is hosting an upcoming webinar:

**Harnessing Transposons for Precise and Efficient Genome Editing in Plants**
Thursday, November 14, 2024, 11 AM - 12 PM ET

Dr. Peng Liu, Senior Research Scientist, Donald Danforth Plant Science Center, St Louis, MO, USA

For more information or to join the working group's email list, please contact George Liu (George.Liu@usda.gov), Zhenbin Hu (Zhenbin.Hu@usda.gov), or Wenli Li (Wenli.Li@usda.gov)

# Training

## 2024 Annual GRWG Workshop

On November 6-7, 2024, the SCINet Geospatial Research Working Group (GRWG) will be hosting their 2024 Annual Workshop. The workshop is split over two half-days that will have sessions including:

- Lightning talks on geospatial analyses and workflows that use SCINet resources.

- Hands-on tutorials that cover a variety of topics:

  - Applying computational methods such as environmental niche modeling and graph convolutional neural networks.

  - Using Python packages for multidimensional arrays to process geospatial data.

  - Using VS Code on SCINet and an example of developing Python scripts with VS Code.

- Discussions on GRWG activities for the new year, including desired training topics and collaboration opportunities.

Recordings will be made available for those who cannot participate during the event. If you are interested in joining the geospatial research working group, please reach out to Heather Savoy (heather.savoy@usda.gov) or Amy Hudson (amy.hudson@usda.gov).

## Coursera

The SCINet Office and the AI-COE are excited to provide training opportunities through Coursera. Coursera licenses are available to ARS scientists and support staff for training focused on scientific computing, data science, artificial intelligence, and related topics. Successful completion of courses and specializations result in widely recognized certificates and credentials.

Please visit the SCINet Coursera Training Page to request a license. Licenses will be assigned on a rolling basis and are active for three months. Users may be able to extend their licenses upon request.

## Workshop Reports

### The Carpentries Unix, Git and R Workshop Report

SCINet in collaboration with Carpentries-certified ARS instructors David Molik, Chen Dong, and Aaron Dickey held a 4-day workshop on The Carpentries: Unix, Git and R from August 19-22, 1-5 PM ET. This hands-on workshop provided participants with the foundational skills in Unix command line, version control with Git, and data analysis and visualization with R. There was a 30-participant cap for the workshop, and this will be a recurring workshop.

## Computer Vision Workshop Report

The Computer Vision course, part of the Practicum AI series offered in collaboration with the Research Computing team at the University of Florida, was held on September 9, 11, and 13, 2024. This intermediate-level course continued developing the concepts and skills learned in the foundational Deep Learning course offered in January. Through hands-on modules, participants learned about computer vision tasks, including image classification, object detection, and image segmentation.

## Software Package and Environment Management Workshop Report

The second offering of SCINet's Software Package and Environment Management Workshop was held on October 3 and 4, 2024. Based on feedback from the first offering of this workshop in July, the workshop was expanded to two afternoon sessions. The first afternoon included a 30-minute introduction to software package and environment management concepts and terminology followed by a hands-on tutorial focused on Python and conda. The second afternoon featured a hands-on tutorial focused on R. Given the popularity of this workshop, we expect to offer it regularly in the future.

## Multispectral UAV Imagery Workshop Report

The SCINet Geospatial Research Working Group hosted a workshop on October 9 and 10, 2024 that focused on processing multispectral UAV imagery and extracting zonal statistics for geospatial modeling using OpenDroneMap on SCINet systems. This workshop was developed and instructed by Alexander Hernandez, Efrain Duarte (a SCINet/AI-COE fellow), and Kaden Patten of the Forage and Range Research Unit.

# Please help us improve our training offerings!

What scientific computing training do you need? The SCINet Office's goal is to provide training opportunities and resources that meet the needs of ARS researchers, so we would be grateful if you could complete our short training request form and let us know how we can best help you learn the computing skills you need. Your feedback will help us decide where we should focus our efforts over the next year and beyond.

Training opportunities are continually being updated on the SCINet Upcoming Events webpage. For more information on any of the above trainings, registration questions, or suggestions, please email SCINet-training@usda.gov.

# Support

## Getting Started with SCINet Is As Easy as 1,2,3

If you do not already have a SCINet account, we hope you will consider joining the 2,300+ researchers who do. Follow the steps below to get your SCINet account.

1. Request a SCINet account to get started.

2. Read the SCINet FAQs covering general info, accounts/login, software, storage, data transfer, support/policy/O&M, parallel computing, and technical issues.

3. Register for a SCINet Forum account to connect to other users, ask questions, and learn how SCINet can enable your research.

P.S. Don't forget to complete your annual security training! This is required to maintain your account.

**For technical assistance with your SCINet account, please email scinet_vrsc@usda.gov.**


## Support email addresses

All requests for help with user accounts, login problems, resource requests, or support for the Ceres HPC cluster should be sent to the SCINet Virtual Research Support Core (VRSC) at scinet_vrsc@usda.gov. Help requests specific to the Atlas HPC cluster should be sent to help-usda@hpc.msstate.edu.

Many emails are currently being sent to other SCINet email boxes. For the most expedient response to your support requests, be sure to send them to scinet_vrsc@usda.gov or to help-usda@hpc.msstate.edu for Atlas-specific requests.


## SCINet User Tip: Accessing the shell from Open OnDemand

Ever since SCINet's first supercomputer, Ceres, became operational, SCINet users have typically accessed a bash shell via the "ssh" command from their local terminal programs (e.g., PowerShell in Windows). It is of course still possible to access SCINet's supercomputers, Ceres and Atlas, via ssh, but Open OnDemand provides an alternative that many users might find more convenient.

To access the shell via Open OnDemand, you'll first need to log on to Open OnDemand on either Ceres or Atlas. Once you have arrived at the Open OnDemand dashboard, click on the "Clusters" menu at the top, and then click on "Ceres Shell Access" or "Atlas Shell Access" (Figure 1). This will open a new tab in your web browser that will launch a bash session on the cluster's login node, just as if you had connected via SSH.
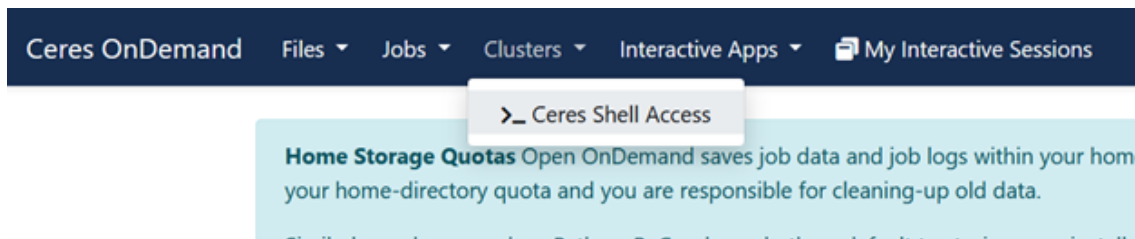
*Figure 1. Accessing the command-line shell on Open OnDemand on Ceres.*

**Do you have tips to share? Email them to [ARS-SCINet-Office@usda.gov](mailto:ARS-SCINet-Office@usda.gov) to be included in future newsletters.**

## SCINet Corner

SCINet Corner is a VRSC-moderated virtual space for people to share knowledge, discuss best practices, learn about new opportunities, and explore resources to support progress on their projects.

The next SCINet Corner will be held on December 12, 2024, 1 PM – 2 PM ET. You can register for this and future SCINet Corners here.

**Have a question that just can't wait? Want to see what other users are doing? Reach out to the ever-expanding SCINet Forum community for ideas, support, or just someone to bounce ideas off of at [https://forum.scinet.usda.gov/](https://forum.scinet.usda.gov/).**

# Connect

## The SCINet Community

Every newsletter highlights SCINet community members as a way to connect the ARS scientific computing community.  To see all the SCINet community updates and review past newsletters, visit the Newsletter Archive.

## Contribute

Do you use SCINet for your research? We would love to share your story! Email ARS-SCINet-Office@usda.gov to contribute content, ask questions, or provide feedback on the SCINet newsletter or website.

## SCINet Office

Haitao Huang, Computational Biologist
Moe Richert, Web Developer
Lavida Rogers, Training Coordinator
Heather Savoy, Computational Biologist
Brian Stucky, Computational Biologist, Acting Chief Scientific Information Officer

## SCINet Leadership Team

Brian Stucky, Acting Chief Scientific Information Officer
Rob Butler, SCINet Program Manager
Jeremy Edwards, Science Advisory Committee (SAC) Chair
Pamela Starke-Reed, Acting Associate Administrator

Note: This newsletter is edited to comply with ARS editorial standards.

**SCINet Website**

Stay Connected with the USDA Agricultural Research Service
5601 Sunnyside Avenue, Beltsville, MD 20705