SCINet Newsletter: July 2024

Research Spotlight | **News** | **Working Groups**

Training | **Support** | **Connect**

Research Spotlight

SCINet as a Resource for Safeguarding and Advancing ARS's Biological Collections

Harlan Svoboda, Herbarium Curator
Floral and Nursery Plants Research Unit, U.S. National Arboretum, Washington, D.C.

Across the USDA's Agricultural Research Service (ARS) there are nealy 100 biological collections containing millions of preserved and viable specimens including animal tissues, seeds, fungal cultures, plant accessions, pinned insects, and viral isolates. These specimens and the data about them document and support ARS research efforts and are an integral part of delivering on the Agency's mission.

One of these collections is the U.S. National Arboretum Herbarium, an assemblage of nearly 700,000 preserved plant specimens housed in Washington, D.C. The Herbarium dates back to the earliest days of the USDA and is still actively growing, with expeditions, research projects, and collaborations all adding substantial new material annually. The collections not only provide a unique snapshot of what plants are growing in a particular place at a given time, but also contribute valuable data and biological material for a range of research that includes describing new species, genetic analyses, and tracking invasive taxa.

Recently, the National Arboretum committed to making its priceless preserved specimens, and their data, more widely accessible to researchers and the American public through digitization. In 2021, the Herbarium' s entire collection was fully digitized using a high-throughput conveyor system to image all specimens. In total, this venture took only 15 months to complete, thanks to the high-throughput imaging system—a remarkable achievement considering that manually digitizing the collection would have taken approximately 419 years.

The question became, though: what to do with the nearly 300 TB worth of files created during the project? Specifically, the original, unprocessed image files needed special consideration to ensure their permanent safeguarding. This project highlights a sometimes-overlooked function of SCINet: long-term, permanent storage of critical USDA digital assets. Much like the physical/analog specimens themselves, which must be preserved and stored in perpetuity, so too must the digital files that represent them.

SCINet's "Juno" storage system, located in Beltsville, Maryland, is designed for long-term storage of research data and is an ideal resource for this use case. Juno is a multi-petabyte storage device that is regularly backed up to tape drives, located off-site at Mississippi State University, and can be securely accessed via command line or the Globus application. Like SCINet's supercomputers (Atlas and Ceres), Juno is operated and maintained by the SCINet Virtual Research Support Core (VRSC). Knowing that the complete digital complement of the Herbarium is safely stored on Juno means the National Arboretum can focus on better serving its stakeholders.

Led by the National Arboretum, ARS will be launching the ARS Biocollections Portal in the coming months, an exciting online endeavor which will allow users from across the globe to access and use USDA collections like never before. This Agency-wide initiative already contains four large collections with additional members slated for inclusion in the years to come. None of this could be possible, though, without first ensuring the original files are protected long-term.

This work showcases how SCINet can be incorporated into various types of projects throughout USDA, even ones that do not fall into the "typical" supercomputing or big data categories. SCINet as a partner in collections-based programs can support (and catalyze) how ARS's treasured specimens are safeguarded, managed, and utilized.
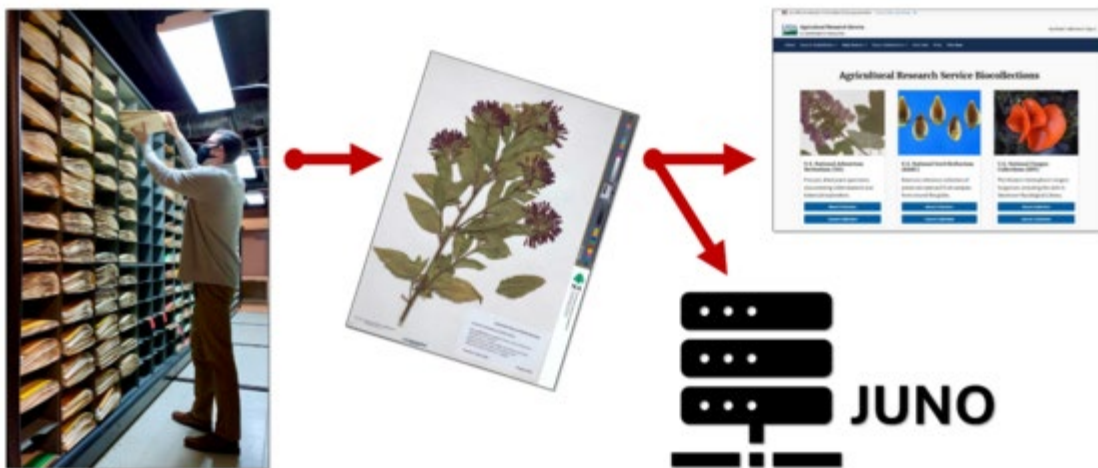

Figure 1. Preserved plant specimens in the U.S. National Arboretum Herbarium (left) were digitized in 2021, producing almost 300 TB of data and image files (center). Those digital assets will soon be accessible through the ARS Biocollections Portal, with the permanent versions being stored on Juno for safekeeping (right).

SCINet and AI COE Fellows

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

Dr. Taejung Chung is a USDA-ARS SCINet/AI-COE fellow under the mentorship of Dr. Ade Oladeinde (USDA-ARS, Athens, GA) and Dr. Zaid Abdo (Colorado State University, Fort Collins, CO). Dr. Chung is from South Korea and moved to the United States in 2017 to attend graduate school. He earned his M.S. and Ph.D. degrees from the Department of Food Science at Penn State. His expertise includes microbiology, microbial genomics, metagenomics, and bioinformatics. Previously, his research focused on microbial communities in diverse environments such as agricultural water, as well as whole genome analysis of foodborne pathogens like *Bacillus cereus*..

As a USDA-ARS SCINet/AI-COE fellow, Dr. Chung will contribute to a research project focused on developing a supervised machine learning tool to predict parameters that are correlated with a reduction in antimicrobial resistance and probiotic administration.

Dr. Mahesh Koirala is a SCINet/AI-COE Postdoctoral Fellow under the mentorship of Dr. Keith Fagerquist at the USDA-ARS in Albany, CA. He earned his Ph.D. in Physics, with a focus on Computational Biophysics, from Clemson University in 2022. He investigated the molecular mechanisms of disease-associated variants, utilizing tools such as Delphi for electrostatic interactions and various computational methodologies for molecular mechanics. Following his graduation, Dr. Koirala worked at Akttyva Therapeutics for 1.5 years, where he contributed to various stages of computational drug discovery.

In his current role with Dr. Fagerquist, Dr. Koirala is conducting computational simulations on the folding and prediction of 3D structures of proteins from microorganisms using AlphaFold2 and examining their dynamics through molecular dynamics tools such as GROMACS.

Dr. Adama Tukuli, advised by Dr. Lindsey Perkin at the Southern Plains Agricultural Research Center in College Station, TX, is developing AI models for precise weevil classification through an ARS-SCINet project. He is developing image-based models to quickly identify species and molecular models to pinpoint genes unique to pest species. This dual approach will enhance boll weevil eradication capabilities, evolving into a user-friendly mobile application and employing genes in Lateral Flow Biosensors for rapid, in-field diagnostics.

Dr. Tukuli earned his B.S. in Plant Sciences at Jimma University in Ethiopia and his Ph.D. in Plant, Insect, and Microbial Sciences at the University of Missouri, under the guidance of Professors Gary Stacey and Bing Stacey. There, he used CRISPR to modify genes and developed AI models to understand amino acid dynamics. He holds additional postgraduate certifications in AI and data engineering. A Ridgel fellowship recipient and former student union representative, Dr. Tukuli is eager to apply his expertise to agriculture-related problems through his SCINet/AI-COE fellowship.

News

Ceres Upgrades

SCINet's Ceres supercomputer had a maintenance period during June which included the following upgrades and changes:

- New storage hardware was installed to replace the older hardware supporting /project and /90daydata. This new storage system will be fully operational after the next Ceres maintenance period in October. Benefits of this upgrade include:

  - **Almost 75% more usable storage space (increased from ~5.5 PB to ~9.6 PB)**;

  - An approximately 80% reduction in power and cooling requirements and a nearly 90% reduction in physical space requirements;

  - Solid-state storage that provides read/write performance superior to the previous storage systems; and

  - Greater flexibility in managing storage space between /project and /90daydata as overall user storage needs evolve.

- Legacy authentication using SCINet usernames, passwords, and authentication codes is no longer supported to ensure that SCINet systems comply with federal IT security requirements. Access to SCINet is now limited to authentication with a LincPass or a Login.gov account, with the exception of some sponsored accounts still using YubiKey devices. These accounts will be transitioned to Login.gov over the next few months. For help with SCINet authentication, please see the overview provided in the User Tip section of this newsletter below and [detailed login instructions provided on the SCINet website](#).

- The Linux distribution has been changed from AlmaLinux to Red Hat Enterprise Linux 9.2. There are no expected complications with previously installed software or workflows.

Internships Update

The second year of the AI-COE/SCINet Graduate Student Internships Program is coming to a close as our summer interns wrap up their research projects with their ARS mentors. In total, we had 29 graduate students participate in summer or spring internships with ARS researchers this year. This year's interns were affiliated with the New College of Florida; the University of Florida; North Carolina State University; Alabama A&M University; Virginia State University; or the AI Institute for Food Systems, a multi-university partnership headquartered at the University of California, Davis. They were primarily from data science and computer science degree programs and were paired with ARS researchers for projects that could benefit from the intern's computational skills. To recognize the accomplishments of these students, we are

holding an internships research symposium on Friday, August 16, 2024 from 11 AM – 5 PM ET. Anyone interested in attending this virtual event can join using this Zoom link.

We will continue this AI-COE-funded internship program next year and a call for mentors will be sent out in the fall. Many thanks to the students who dedicated their time and hard work to these internships, the ARS scientists who volunteered to serve as internship mentors, and the universities who have partnered with us for this internships program.

SCINet Strategic Planning Update

Beginning in January, the SCINet Office has been conducting a new SCINet strategic planning effort in which we are meeting with ARS research units across the country either virtually or in person to learn more about their scientific computing needs and to increase awareness of SCINet. Our ultimate goal is to assess scientific computing needs across the agency and develop a five-year plan for how we can best meet those needs. Along the way, we are also identifying and resolving more immediate needs or user experience issues that we learn about during these visits, including:

- **Improving the user-friendliness of the *Welcome to SCINet* email for new users**: We have overhauled the email content sent to new users with streamlined access instructions plus helpful links for available interfaces, user support, and training resources.

- **Improving documentation for logging on to SCINet:** In response to concerns about difficulties with logging on to SCINet systems, we have completely overhauled the documentation on our website for accessing SCINet.

- **Announcing important SCINet changes on our website:** We have now created a *What's New* page on the SCINet website for sharing short announcements about SCINet developments such as added software features, user experience improvements, login procedure updates, policy changes, and new working groups.

- **Identifying proprietary software to offer on SCINet supercomputers**: We have distributed a survey to help identify proprietary software desired by multiple SCINet users. Based on the survey results, we will be exploring which software are compatible with licensed use on our computing systems. (If you have not completed the survey, please do so!)

- **Providing training on software package and environment management on SCINet supercomputers**: We have designed a Software Package and Environment Management Workshop that was first delivered on July 19 (see the post-workshop report below) and will be offered again on October 3-4 (see workshop announcement below).

Working Group Updates

Translational Omics Working Group

The Translational Omics Working Group is hosting an upcoming webinar:

The genomic and metabolic making of yeast ecological diversity
Thursday, August 8, 2024, 11 AM - 12 PM ET

Dr. Chris Hittinger, Professor of Genetics, University of Wisconsin-Madison, 1552 University Avenue, Madison, WI 53726, 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)

Geospatial Research Working Group

The Geospatial Research Working Group will be holding its annual workshop virtually during the week of November 4, 2024. This is a community event to learn more about each other's research and gain awareness of more tools for streamlining geospatial workflows on SCINet computing resources. Sessions will include lightning talks and hands-on tutorials provided by GRWG members. To join the GRWG, please contact the working group leads, Heather Savoy and Amy Hudson.

Training

The Carpentries Unix, Git and R Workshop

August 19-22, 2024, 2024, 1 PM – 5 PM ET

In this interactive, hands-on workshop, participants will learn basic programming tools and concepts for the Unix Shell, version control with Git, and R. Here is the daily schedule:

Date	Session
August 19, 2024	The Unix Shell
August 20, 2024	Version control with Git
August 21, 2024	R for Reproducible Scientific Analysis Session 1
August 22, 2024	R for Reproducible Scientific Analysis Session 2

To register for the Unix, Git, and R Workshop, please fill out this form.

Practicum AI Computer Vision Workshop

September 9,11, & 13, 2024, 3 PM – 5 PM ET

Practicum AI's Computer Vision Course builds on a basic understanding of Python and deep neural networks to demystify the technology that allows machines to interpret and make decisions based on visual data. We will explore image classification, object detection, and image segmentation tasks. Each module in this course will provide hands-on experiences to create your own working models. By the end, you'll be ready to instantiate and employ a computer vision model to make predictions.

Prerequisites: This course assumes familiarity with Jupyter notebooks, basic Python programming, and a general understanding of neural networks and deep learning. The Practicum AI Beginner Series provides this foundation.

To register for the Practicum AI Computer Vision Workshop, please fill out this form.

Software Package/Environment Management Workshop

October 3, 2024, 1:00 PM – 4:30 PM ET

October 4, 2024, 1:00 PM – 4:00 PM ET

Package and computing environment management systems for R, Python, and Anaconda make it easy to install the software you need for your research projects. They also make it

easy to accidentally use up all of the space in your home directory or end up with frustrating software conflicts! In this workshop presented by the SCINet Office, we will cover best practices for managing software packages and computing environments on SCINet's supercomputers. This will be a hands-on workshop that will provide you with the practical knowledge and skills you need to spend less time worrying about package management and more time focusing on your research! At least some experience with the command line will be helpful for working through the exercises. If you did not have a chance to attend this workshop when we first offered it in July, please join us in October!

[To register for the Software Package and Environment Management Workshop, please fill out this form.](#)

Multispectral UAV Imagery Workshop

October 9-10, 2024, 1 PM – 5 PM ET

This workshop will focus on processing multispectral imagery from unoccupied aerial vehicles (UAVs) and extracting zonal statistics for geospatial modeling using OpenDroneMap on SCINet systems. Participants will be introduced to the different stages of processing multispectral imagery collected using UAVs, and the development of tools for the extraction of zonal statistics in a format that can be directly used in geospatial modeling will be shared. Topics covered in the workshop will include:

- Installation and set up of the miniconda environment

- Pre-processing: Tools to filter out redundant imagery by altitude and location

- Pre-processing: Conversion of multispectral imagery digital numbers to reflectance

- Processing: OpenDroneMap to generate multispectral orthophotos

- Tools to generate vegetation indices from multispectral imagery, volume estimates from digital surface models, and extraction of zonal statistics

Workshop participants should have some familiarity with Python, using the command line, and basic knowledge about remote sensing and geospatial concepts. For more information about the prerequisites, please see the registration form.

[To register for the Multispectral UAV Imagery Workshop, please fill out this form.](#)

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

Genome Assembly Workshop Report

SCINet, in collaboration with the Genome Informatics Facility at Iowa State University, recently held a 3-day workshop on Genome Assembly from June 25-27, 2024. The first day of the workshop covered assembly basics such as quality control, assembly, scaffolding and polishing for prokaryotic genomes. The hands-on exercises focused on downloading and assembling Illumina, PacBio, and Nanopore reads. During the second day, participants were engaged in exercises tailored to eukaryotic genome assembly, such as assembling the yeast genome with nanopore reads and HiFi reads. On the final day of the workshop, participants were introduced to Hi-C and learned about scaffolding with Hi-C reads, haplotype phasing, and assembly validation and visualization. The workshop's maximum capacity was 40 participants, but we had many more people register. Due to high interest in this workshop, we will be offering it again in the future.

The recordings for the workshop can be found here.

Software Package and Environment Management Workshop Report

The first offering of SCINet's Software Package and Environment Management Workshop was held on July 19, 2024. Sessions included a 30-minute introduction to software package and environment management concepts and terminology; a 2-hour, hands-on session that focused on Python and conda; and another 2-hour, hands-on session focused on R. More people were interested in registering for this workshop event than we could effectively include, so the workshop will be repeated in October and again regularly in the future. In the meantime, there is a recording of the July workshop event.

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: Streamlined login procedures

SCINet login procedures have been evolving over the last year or so, both in response to changing federal IT security requirements and to make it easier for researchers to use SCINet systems. After using passwords and 6-digit, time-based codes for multi-factor authentication (MFA) for several years, passwords and time-based codes have been deprecated and replaced with MFA options that are simpler and faster for most users. There are now two supported methods for MFA, depending on whether a SCINet user has a LincPass or AltLinc card:

- For SCINet users with a LincPass or AltLinc card, the LincPass or AltLinc card should be used for MFA and SCINet access.

- For SCINet users without a LincPass or AltLinc card, Login.gov should be used for MFA and SCINet access.

These authentication methods are used for logging on to all SCINet systems and interfaces, whether web-based interfaces such as Open OnDemand or command-line access via SSH. To go along with these changes, we've overhauled, simplified, and streamlined the documentation for SCINet access on our website, and we encourage you to take a look for more information.

Do you have tips to share? Email them to 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 August 15, 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/.

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 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
Marlen Eve, 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