USDA United States Department of Agriculture
Agricultural Research Service

SCINet Newsletter: January 2022

Research Spotlight | News | Training | Support | Connect

RESEARCH SPOTLIGHT

SCINet accelerates an inter-disciplinary technology platform for soybean post-genomic research

Yong-qiang Charles An, Research Molecular Biologist

USDA-ARS Plant Genetics Research Unit at Danforth Plant Science Center, Saint Louis, MO.

Soybean (Glycine max (L.) Merr.) is a versatile, nutrients-laden, economically invaluable crop with capacity to restore soil fertility through atmospheric nitrogen fixation. It holds great significance in ensuring adequate global nutritional food security and environmentally friendly sustainable agriculture. Consequently, soybean has been the target of extensive, worldwide research efforts which have generated a massive amount of data concerning soybean genetics and traits. These data provide an unprecedented opportunity to apply modern genomics and analytical methods to the challenge of soybean improvement, but data aggregation and analysis remain challenging. My laboratory has been building a technology platform that brings together both internal and external data resources and provides a suite of data analysis and mining tools for translating rapidly growing datasets into improved soybean cultivars.

Since its inception, SCINet has been an integral part of our large-scale data driven research and has helped us produce more than 50 TB of soybean -omics data for mining. Using SCINet's high performance computing clusters (HPCs; Ceres and/or Atlas), we have analyzed more than 5,000 whole-genome sequences and nearly 4,000 transcriptomes of soybean. These analyses have generated a wealth of genetic information that is of high potential value for soybean improvement. Collectively, these datasets are of immense utility to research community and thus, a complete annotated SNPs dataset has been released in Ag Data Commons (An et al. 2020) and Soybase. A comprehensive description of the dataset and its versatile use will be published (Zhang et al. accept with minor revisions).

SCINet also provides us with a computing environment that has enabled us to develop a suite of new genomic data analysis tools. For example, SCINet resources facilitated the development of a robust pipeline for identifying specific regions of interest within soybean genomes. Additionally, SCINet was instrumental in our discovery of two genes controlling soybean's seed protein and oil content. We found that these genes played important roles in the process of soybean domestication (Zhang et al. 2020).

Read more about this research

References:

An Y, Zhang H, Meryer R. 2020. Data from: Development of a versatile resource from 1500 diverse genomes for post-genomics research. Ag Data Commons. doi:10.15482/USDA.ADC/1519167

Zhang H, Goettel W, Song Q, Jiang H, Hu Z, Wang ML, An, Y-qC. 2020. Selection of GmSWEET39 for oil and protein improvement in soybean. PLOS Genetics 16(11): e1009114. doi:10.1371/journal.pgen.1009114

Zhang H, Jiang H, Hu Z, Song Q, An Y-qC. Accepted with Minor Revisions. Development of a versatile resource from 1500 diverse genomes for post-genomics research. BMC Genomic.

Do you use SCINet for your research? Contact SCINet-Newsletter@usda.gov for a chance to be featured in the newsletter!

NEWS

SCINet Fellows Conference

SCINet and AI-COE Fellows hosted a conference on 9-10 November, 2021 showcasing their innovative research that fostered discussions with ARS National Program Leaders, Research Leaders and scientists on timely topics related to the use of the HPCs and analysis of big data. A total of 10 presentations by current Fellows was followed by an invited speaker on collaborative research Dr. Daniel Ferguson, Director, Climate Assessment for the Southwest, University of Arizona. The Fellows led a series of discussions on topics of joint interest to the Fellows and ARS leadership. Learn more at https://scinet.usda.gov/opportunities/fellowsconference.

Protein Folding Conference

On December 1 and 2, 2021, approximately 150 ARS researchers participated in a virtual, two-day conference focused on recent computational advances in protein science. These new software tools and methods, built upon modern artificial intelligence techniques and computing hardware, will allow researchers to study protein structure and function at scales far beyond what was possible only a few years ago. Because proteins are a component of virtually all agriculturally relevant organismal traits, these new research capabilities are expected to significantly impact US agriculture.

The first half of the conference featured two keynote speakers, Dr. John Moult (Fellow at the Institute for Bioscience and Biotechnology Research and a Professor at the University of Maryland) and Dr. Darrell Hurt (Chief of the Bionformatics and Computational Biosciences Branch at the National Institute of Allergy and Infectious Diseases at the NIH), and 13 lightning talks by ARS scientists discussing their protein-related research. Dr. Moult's and Dr. Hurt's keynote presentations and all lightning talks may now be viewed online.

During the second half of the conference, SCINet and AI COE led discussions on the opportunities and strategies for leveraging new protein science tools to advance ARS research. Six topics were identified by participants as potential areas for SCINet and AI COE working groups:

- Links from protein sequence and structure to biological function and phenotype
- Plant and animal breeding
- Pest, pathogen, and disease control
- Protein unfolding and misfolding
- Molecular farming (using plants to grow useful proteins)
- Quantum biology applied to protein science

If you are interested in participating in one or more of these working groups, please complete our short post-conference survey by Feb. 1, 2022. Dr. Brian Stucky with the SCINet Office can be contacted for more information.

Transferring data to and from the HPCs

SCINet recently launched a new data storage system, called "Juno", to complement our existing high-performance computing (HPC) infrastructure. Juno provides ARS researchers and collaborators with new long-term data storage capabilities that go beyond the short-term storage on our HPC clusters, Ceres and Atlas. Because Juno is located in Beltsville, MD, researchers will need to move data between Juno and Ceres (in Ames, IA) or Atlas (in

Starkville, MS). Here, we provide recommendations to help users transfer files as quickly as possible. These initial recommendations will be updated through time as more data are obtained.

- First, we recommend using Globus for all large data transfers between Juno and Ceres or Atlas. Globus can typically deliver faster performance than alternative file transfer options.
- Second, when transferring files, the simplest and fastest overall strategy for maximizing data transfer speeds is to move a small number of large files. For example, if a user has many small files, we recommend using archiving utility, such as tar, to combine the small files into a single large file prior to transferring the dataset. When the mean file size is very small (e.g., a few kilobytes), this process will result in a substantial increase in transfer speed compared with transferring each file individually (Figure 1).
- Third, if the mean size of the files to be transferred is around 70 megabytes or larger, we recommend transferring the files directly rather than combining them together beforehand. With files of this size (and larger), the time required to combine the files before transferring will likely not be worth the relatively small gain in transfer speed (Figure 1).

We will continue to investigate data transfer performance on SCINet infrastructure. Check the SCINet website in the future for more comprehensive instructions and recommendations.

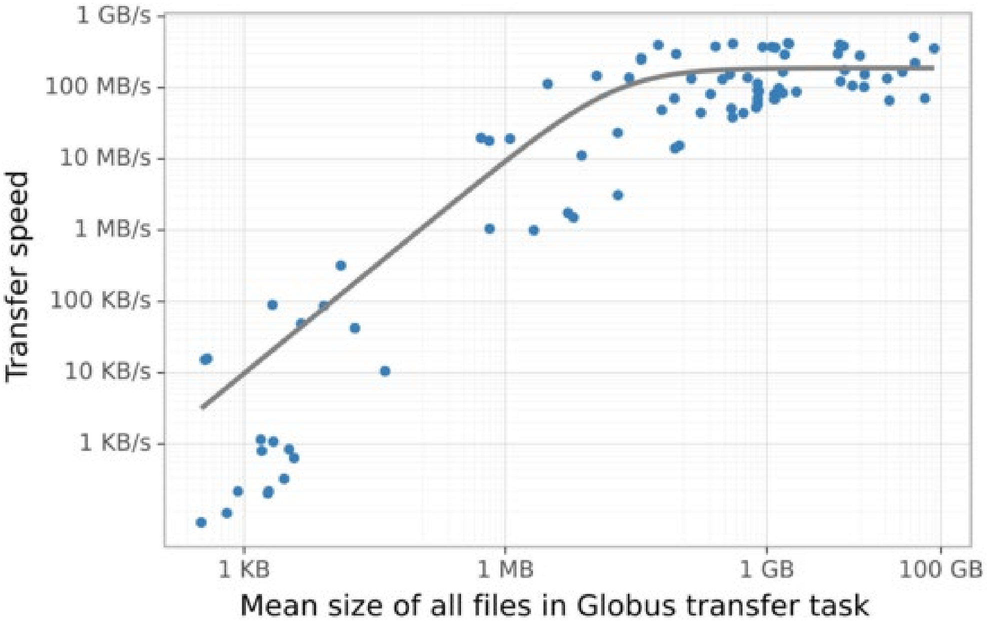


Figure 1. The relationship between the mean file size of a Globus data transfer task and the observed transfer speed when moving files between Juno and Ceres or Atlas. The x-axis represents the mean size of all files included in a Globus file transfer task. Each blue dot represents a single transfer task. The solid, gray line depicts the expected mean transfer speed for a given mean file size, as estimated by a statistical model of data throughput. Note that both the x-and y-axis scales are logarithmic.

Open OnDemand for Virtual Desktops and Web Apps on Ceres and Atlas

Open OnDemand is now available. This software provides web-browser access to high-performance computing resources on Ceres and Atlas, including virtual desktop environments and scientific web applications. Available apps include:

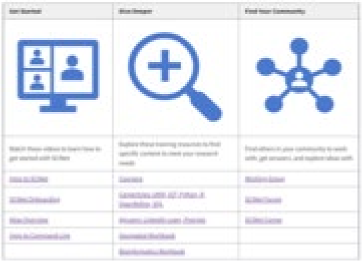
- A collection of Open OnDemand core apps, including a File Manager for browser-based file system access, a lightweight File Editor, a Shell App for in-browser command-line access, and a Job Composer App for creation and management of Slurm batch jobs
- Virtual Desktop Environment (CentOS Linux)
- JupyterLab and RStudio Server (with more web apps coming soon)

Open OnDemand allows user App Development, enabling SCINet users to develop and deploy private custom web apps on Ceres or Atlas from their home directory. Custom Interactive Apps launch apps Slurm jobs on compute nodes, run with the SCINet user's privileges, and allow access to Ceres/Atlas parallel file systems (see the Open OnDemand Interactive Apps development documentation for more details). Open OnDemand app development is currently opt-in; to enable for your account, please contact the SCINet VRSC (scinet_vrsc@usda.gov), specify the system (Ceres or Atlas), and provide a brief justification.

To get started with Open OnDemand on Atlas, see the Atlas Documentation. Preliminary documentation for accessing the Ceres Open OnDemand is available in the SCINet RStudio Server guide, with additional documentation updates to follow.

TRAINING

Getting Started Learning Path


Get Started
Dive Deeper
Find Your Community

With the expansive list of free training available online, finding the right training to meet your learning needs can be daunting. Take the first steps in getting started with the SCINet Introductory Learning Pathway. Learn about SCINet, how to sign up for an account, and what is possible when supported by SCINet infrastructure. Then dive in with hands-on tutorials available across multiple searchable platforms to find the information you need for just in time learning.

Training Opportunities

The Carpentries Workshops: Looking to further your learning journey? Sign up for one of the Carpentries courses that start next month.

- Software Carpentry (UNIX, GIT, Python); 3/8 and 3/15; 10-6pm ET
- Software Carpentries (UNIX, GIT, R); 2/24 and 3/31; 10-6 ET
- Data Carpentry (OpenRefine, SQL, Python); 2/16, 2/17, 2/23, 2/24; 1-5 pm ET
- Data Carpentry (OpenRefine, SQL, R) 3/7, 3/8, 3/14, 3/15; 1-5pm ET

If you are interested in helping the instructors during one or more of these workshops, please contact scinet-training@usda.gov.

Certified Carpentries Instructors: This training will teach you Carpentries pedagogy and result in the ability to lead Carpentries workshops in Unix, git, R, Python, and more. If you are interested in joining the ARS Carpentries instructor team, please reach out to scinet-training@usda.gov.

Courses by Mississippi State University: There are still seats available for upcoming training events offered by our collaborators including Introduction to Atlas and Data Wrangling. Learn more and register at https://forms.office.com/g/rYdpFLxvFV. There are waiting lists available for several other courses, including Moving from SAS to R for Statistical Analysis. Sign up to get notified when the courses are offered again.

Coursera.org Courses Update: The SCINet Office and the AI Center of Excellence are excited to provide training opportunities through Coursera. Coursera licenses are available to ARS scientists and support staff to complete training focused on scientific computing and artificial intelligence. Successful completion of courses and specializations will result in widely recognized certificates and credentials. Please visit the SCINet Coursera Training Page to request a license by January 15th for the quarter starting February 1st, 2022.

SCINet Fall Training Series: SCINet and AI-COE in collaboration with the VRSC teams at Iowa State University and Mississippi State University presented a series of workshops during the fall of 2021. The courses led researchers through a continuum of learning to develop an understanding of SCINet, from logging into one of the HPCs to navigating in the HPC environment, and submitting jobs. The series included the following courses: Intro to SCINet, SCINet Onboarding, Intro to Atlas and Intro to Command Line. If you missed these events, check out the recordings that are part of SCINet's Introductory Learning Path.

Training opportunities are continuously being updated on the new SCINet Upcoming Training 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

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 change your password when logging in for the first time.

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

SCINet User Tips

By Andrew Severin

Andrew manages the Genome Informatics Facility at Iowa State University. He leads a team of experts tasked with enabling USDA scientists to translate big data into informative data on their specific scientific questions. He is an interdisciplinary scientist working at the interface of genetics and bioinformatics. His academic background is in biochemistry with a Ph.D. in NMR spectroscopy.

Hate that the Terminal program on a MAC asks your permissions every time you change Directories? Fix it by giving Terminal access to all folders:

- Settings
- Security & Privacy
- Full Disk Access

Need a Markdown editor for your Github repositories? Check out Atom editor at https://atom.io. Here are some recommended packages you can install to add additional functionality: language-swift-89, language-r, markdownn-folding, markdown-pdf, minimap wordcount, drag-relative-path, markdown-scroll-sync, autocomplete-python, autocomplete-swift, autocomplete-R.

Do you have tips to share? Email them to SCINet-Newsletter@usda.gov to be included in future newsletters.

SCINet Corner: Third Thursdays each Month

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.

This reoccurring meeting occurs on the third Thursday each month. The next event is on Thursday, January 20th (1pm EST). Meeting times may change. It is recommended to join the event via Google Chrome or Firefox.

Register at **https://forms.gle/7DcBoBvbGcjQDBP38**

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 Team

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

Fellow Focus

ORISE Fellows bring a fresh perspective to the ARS scientific community, while allowing these Fellows to learn through hands-on research under our top scientists. SCINet funds a number of ORISE Fellows each year with the mission to support research using high performance computing and computational science. If you are interested in learning more about their skillset, please reach out to their mentors.

Melanie Veron, ORISE Fellow, SCINet Office- Melanie earned a bachelor’s degree in Biological Sciences at Southern Illinois University Carbondale in 2017 and a master’s degree in Quantitative Methods of Biodiversity, Conservation, and Epidemiology at the University of Glasgow in 2018. Her graduate research project involved exploring methods of ecological network reconstruction and analysis of various microbial communities. Prior to becoming an ORISE fellow with the ARS, Melanie worked as a Clinical

Research Data Manager for cancer immunotherapy studies at the University of Chicago Medicine. Her research interests are broad and diverse, spanning from biodiversity and conservation to microbiology and public health. Melanie is eager to use her background to contribute to applied ecological research, strengthen her computational skills, and assist with the support and training services of the SCINet Office under the mentorship of Dr. Brian Stucky.

Noa Mills, ORISE Fellow, SCINet Office- Noa recently completed a bachelor's in Computational Mathematics at the University of California Santa Cruz with a focus on modeling US wild fires using data science tools. As an ORISE Fellow, Noa will be using a variety of computational methods to analyze large agricultural datasets under the direction of Dr. Brian Stucky.

Do you know someone who might be interested in becoming an ORISE Fellow with SCINet or the AI Center of Excellence? If so, please share the link with them. All opportunities have been recently updated with an FY22 deadline. For more information on the positions including deadlines and potential start dates, visit: https://www.zintellect.com/Catalog and enter the keyword: SCINet.

Contribute

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

SCINet Leadership Team

Deb Peters, Acting Chief Science Information Officer
Rob Butler, Acting SCINet Project Manager
Adam Rivers, Science Advisory Committee (SAC) Chair
Steve Kappes, 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