



USDA-ARS SCINet Newsletter: January 2021

Contents

- **How to Get Started**
- **SCINet Website Update**
- **SCINet User Tips**
- **SCINet Training Program**
- **Research Highlights**
- **Atlas Corner**
- **Meet our SCINet Fellows**
- **Resources**
- **Contribute / Contact**

How to Get Started

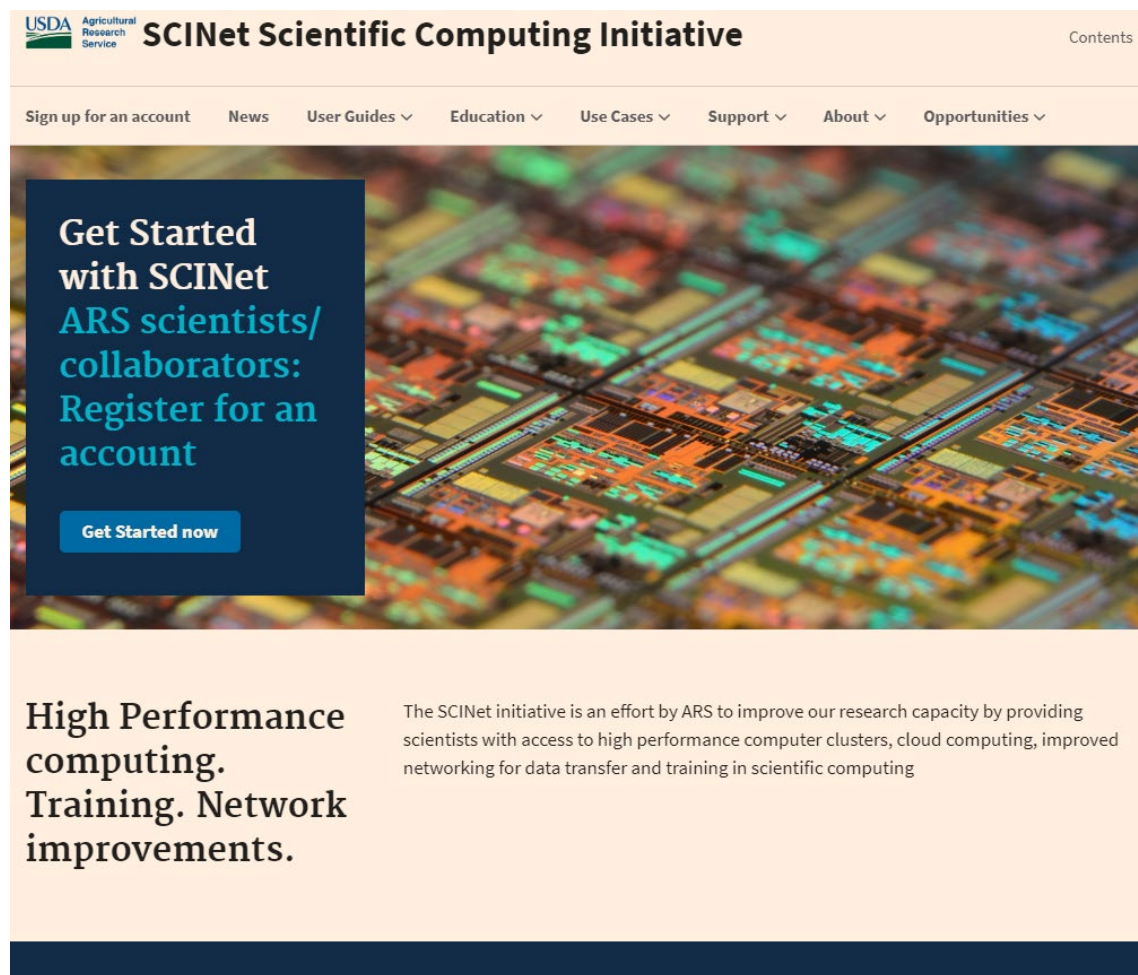


Simply [request a SCINet account](#) (eAuthentication required) to get started. Upon approval, you will receive instructions for logging into SCINet and accessing Basecamp.

Check out the [new SCINet website](#) for more info on how SCINet can enable your research.

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

SCINet Website Update



USDA Agricultural Research Service **SCINet Scientific Computing Initiative** Contents

Sign up for an account News User Guides Education Use Cases Support About Opportunities

Get Started with SCINet
ARS scientists/
collaborators:
Register for an
account

Get Started now

High Performance computing. Training. Network improvements.

The SCINet initiative is an effort by ARS to improve our research capacity by providing scientists with access to high performance computer clusters, cloud computing, improved networking for data transfer and training in scientific computing

New content is constantly being added to the [SCINet website](#). Please send any website feedback to SCINet-Newsletter@usda.gov.

SCINet User Tips

Optimize your allocation requests: the `seff` command can show the user how efficient the job was; this is important to keep in mind the next time you estimate what resources you request.

Do you have tips to share?

Email them to SCINet-Newsletter@usda.gov to be included in future newsletters.

SCINet Training Program

SCINet-funded Training

- **Image processing workshops:** In October 2020 the USDA-ARS and New Mexico State University held two back-to-back workshops on image processing using classical machine learning (ML) and deep learning (DL). 25 participants learned to pre-process image data for ML and DL research

techniques, apply common ML classifiers to image data and assess performance, train and test a convolutional neural network (CNN) for image classification, modify CNN architecture for application to new data, and visualize characteristics of a CNN to help interpret performance. The workshops are expected to be offered again in [February or March 2021 \(dates TBD\)](#) for ARS scientists, scientific staff, and University collaborators with at least intermediate level computer programming skills. Fill out the [prerequisite skills assessment survey](#) to be added to the waiting list for the next offering. The workshop materials are also available to work through at your own pace on the [workshop website](#). Materials include how to prepare your computer, scripts (Python code provided in Jupyter notebook format), image datasets, and video recordings from the October workshop sessions.

- [SCINet Geospatial Research Workshop 2020](#) and training session recordings are **now available** on each session's tab, including tutorials which anyone can work through at their own pace. Tutorials from this workshop can also be found in the new [Geospatial Workbook](#).
- **The Carpentries R and Python workshops:** Multi-day online workshops covering Unix, Git, and either R or Python from The [Carpentries](#) will be offered in February or March 2021, details forthcoming on the [Upcoming Events page](#). Prior workshops in July and August were well attended, with waiting lists. If you are on a waiting list, you will receive priority to the next round of Carpentries workshops. If you would like to be added to the Carpentries e-mail distribution, contact Kathy Yeater.
 - **The Carpentries instructor training** is available to qualified applicants who desire to become certified Carpentries workshop instructors. [This training](#) teaches participants The Carpentries pedagogy and provides the resources to instruct Carpentries workshops in Unix, Git, SQL, OpenRefine, R, Python, and more. Contact Kathy Yeater for more information about instructor training.
- **Coursera.org certified courses:** The SCINet initiative and the AI Center of Excellence are excited to provide a new training opportunity through [Coursera](#). This program provides a limited number of free Coursera licenses to ARS employees to complete courses and specializations focused on scientific computing and artificial intelligence. Completion of courses and specializations will result in widely recognized certificates and credentials for top-tier institutions. Additional information, including the application process, can now be found on the [Coursera Training](#) page of the SCINet website.

Free Online Computational Training (Self-paced)

- Make use of your work-from-home time with computational training! A large list of free tutorials and courses has been compiled on the [Free Online Training page](#). Training topic areas include Python, R, SAS, and MATLAB programming; statistics; data science concepts; AI and machine learning; GIS; Google Earth Engine; Git and GitHub; reproducibility, productivity, and integration management tools; and bioinformatics and ecology domain learning. Know of additional free training opportunities? Send them to SCINet-Newsletter@usda.gov.

SCINet Online Science Tutorials

- Browse our growing set of SCINet science tutorials created by ARS scientists and the SCINet Virtual Research Support Core. Our [ARS Science Tutorials page](#) includes Ceres Onboarding and Intro to Unix for new HPC users, two geospatial computing tutorials, a QTL Analysis tutorial for sequencing in R, and machine learning training material.

Research Highlights

SCINet moves ARS to the forefront of genomics research

By: Brian Scheffler

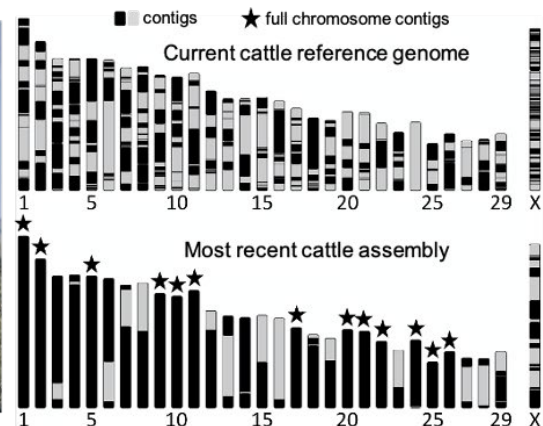
In this SCINet newsletter, we bring you fascinating advancements from the field of genomics, and how ARS is using new technology and SCINet resources to move this field forward. Did you know that the first human genome was released in 2001 at an estimated cost of \$2.7 B in FY 1991 dollars? Actually, \$0.3 B less than the original 1991 prediction. How often do you see a government-funded project come in under budget? Today, ARS plays a critical role in genomic research of important agriculture species - from plant to animal production systems and even their nasty pests. This role is possible because of ARS initiatives to develop important infrastructure like the computational resources offered under SCINet combined with sequencing technology and scientific expertise. And all of this under \$2.7 B per genome! Today, ARS can (and does) produce quality genomes for almost any species for a reasonable cost and as you will see below, from the large (cows) to the almost invisible (little bitty insects).

Why are genomes so difficult to generate? Imagine you have a set of encyclopedias (26 volumes of A-Z) and you put them through a paper shredder. Now put it all back together in correct order by volume while also identifying all the sections and chapters. That is the basis of producing a genome where a volume is equal to a chromosome. As you can imagine, it is hard to put all the little pieces back together and in the right place. How many pieces for the human genome? Using short-read Illumina technology, this is a puzzle consisting of 1,344,000,000 pieces with each piece being 300 base pairs. Even using new high-quality, long-read PacBio HiFi reads, there are 8,960,000 pieces at 15,000 base pairs each. Did I mention that wheat is almost 6 times bigger than the human genome???

Why are quality genomes important? Well, think about the different levels of road maps we have. The better the quality of a map then the easier it is to find things and to understand the relationship of one location to another. The same for the map of a genome: the better the genome map, the easier it is to find genes of interest for further studies or manipulation. In plants, disease resistance can be the result of gene duplication, and duplicated genes can be very hard to find if a quality genome is not available.

Read below for how ARS leads in creating reference genomes for livestock and how the AG100Pest initiative has put ARS at the cutting edge of insect genomes - where genomes are being developed from a single insect sample.

SCINet resources for creating livestock reference genomes



By: Ben Rosen

As scientists in [USDA-ARS's Animal Genomics and Improvement Lab](#), one of our missions is to generate scientific resources used to improve our ability to raise cattle, sheep and goats to their fullest potential. We have sequencing technologies that can be used to sequence the DNA of these animals and generate a blueprint of their genomes. A genome consists of long strings of DNA called chromosomes. The problem we face is that the sequencing technologies can't sequence full chromosomes, but rather they sequence millions of pieces of DNA or fragments of the chromosome that have to be pieced together like a puzzle using HPCs (high performance computer [HPC] clusters).

My main focus is to improve the accuracy and completeness of reference genomes that are then used to, for example, search for disease-causing mutations or breed for a beneficial trait like milk production. Our most recent cattle genome that we pieced together had more unbroken chromosomes compared to the more highly fragmented (contigs) previous reference genome for cattle (see above image; [Rosen et al. 2020, GigaScience](#)). This more complete genome might have DNA that was missed in earlier assembled genomes. This more complete genome is important to researchers studying large-scale chromosome biology and its role in gene regulation.

SCINet note: Not all genomes are assembled equally. Genome size and complexity have an enormous impact on the computational time required for assembly. The HPCs available through SCINet, named Ceres and Atlas, allow our research group to put together the pieces of DNA using different parameters of the program we use to determine the best possible solution. SCINet is a shared computing resource where users can purchase access to the HPC, but are able to use way more than they could afford for short periods of time. This approach has enabled us to triple our compute capacity without increasing the burden on cost or local IT resources. SCINet also has amazing staff that support the machines, software installation and domain experts as part of the Virtual Research Support Core (VRSC). This in-house ARS capacity is essential because we can now accomplish objectives that would never have been possible before.

Generating whole-genome data from a single insect: PacBio HiFi genome assembly and HiC scaffolding pipelines for reference quality genomes using SCINet

By: Scott Geib

The damage that insect pests cost to human health and agriculture is enormous. For instance, invasive insects can cost a minimum of \$70.0 billion per year globally, and associated health costs can exceed \$6.9 billion per year. One method that can help control insect pests is the study of pest insect genomes, because insect genomic data are useful resources for developing alternative and eco-friendly pest control policies. Such analyses allow entomologists to discover molecular diversity in insect populations that underlie the causes of pest population outbreaks. But many of the existing insect genome sequences are not of high quality, which means that critical functional data may not be present in these genomes. Until recently, sequencing costs made it impractical to sequence large numbers of genomes, so that the genomes of many insect pests have yet to be sequenced.

However, technologies for cheaper, more high-quality genome assemblies are emerging, including the application of PacBio HiFi reads. These single-molecule long reads are processed as circular, consensus sequencing (CCS) reads to generate relatively long single molecule reads (5-20kb+) with exceptionally high accuracy. This technology improves the quality and completeness of genome sequences to capture more contiguous (i.e., not separated by gaps) DNA sequences than older sequencing technologies. Through performing DNA extraction and library preparation in my lab in Hilo, HI, and collaborating with PacBio Sequel II sequencing platforms available within the agency (Stoneville, MS & Clay Center, NE), we generated HiFi data for a number of insect pests as part of the [Ag100Pest](#) sequencing initiative.

SCINet Note: This project utilized computational resources on the Ceres HPC. Pipelines for performing pre-assembly filtering, HiFi assembly, post-assembly scaffolding with HiC and final assembly filtering (to remove microbial and mitochondrial components) have been developed by the Ag100Pest assembly team, and now due to the resources available on Ceres, we can go from raw data to assembly in less than a day, and curation to final genome in less than a week (requiring some manual review of the final assembly). This is revolutionary to non-model genomics, and allows expansion of the scope of projects into pan-genome studies, strain level characterizations, and more.

Selecting Robust Climate Change Projections for Agricultural Systems



By: Kerrie Geil*

**Kerrie is one of 10 postdocs in SCINet's first postdoc cohort. Scroll down to our new "Meet our SCINet Fellows" section below for a short introduction to Kerrie and other featured fellows Alicia Foxx and Melanie Kammerer.*

Scientists of many disciplines, including agricultural fields, often use climate change projections in their research. For example, these projections have been used to estimate how crops or ecological systems may be impacted by future climate conditions, or to predict the spread of diseases and pests. Many sources of climate change projections exist for these research applications: over 100 different global general circulations models (GCMs) in the [Coupled Model Intercomparison Project \(CMIP\) archives](#); large ensembles generated from a single GCM (such as [NCAR's CESM large ensemble](#)); dynamically downscaled GCM products that generate higher spatial resolution information using regional climate models (such as [CORDEX](#)); and statistically downscaled GCM products that generate higher resolution information using empirical equations (such as the [MACA product](#)). Unfortunately, there are no quality standards or model performance thresholds implemented for any of this data. How should a scientist choose the most appropriate and robust source of projections for their particular research application considering the many sources and varied quality of available data?

The best practice is to spend time evaluating the performance of climate model simulations using metrics that are relevant to each particular research application and to then avoid the use of climate projections from any model that doesn't perform well. In other words, a scientist should ensure that a model can simulate, fairly realistically, the phenomenon of interest/impact before using its projections. This applies to "bias-corrected"

downscaled products as well as GCMs because 1) a poorly performing model shouldn't be downscaled in the first place and 2) many model biases exist but only a couple are corrected in the downscaling process.

In reality though, this best practice is almost never followed due to the time, computational resources, and evaluation knowledge required for a model performance analysis. Instead, scientists often select one of the many available downscaled GCM products and use a multi-model average of projections, without considering the implications of model bias. For any domain scientist (outside of climate science) this is completely understandable. What ecologist, hydrologist, or rangeland scientist wants to spend months assessing the quality of climate models when they could grab a single climate projection product and instead spend that time focusing on cropland, hydrology, or rangeland science questions? We must work toward a model evaluation solution that produces more robust science while also being much more convenient and understandable for scientists to implement.

As part of my SCINet postdoctoral fellowship, I am working with members of the [USDA-ARS Vesicular Stomatitis \(VS\) grand challenge project](#) to determine the most robust climate model projections for predicting changes in the geographic range of the livestock disease VS under future climate conditions (forthcoming article in the journal *Climate*). As part of this Grand Challenge Project, ARS scientists Debra Peters, Luis Rodriguez, Lee Cohnstaedt, Barbara Drolet, Justin Derner, and Emile Elias, along with our collaborator from USDA APHIS, Angela Pelzel-McCluskey, are developing process-based early warning strategies to predict the spread of vector-borne disease across the US. My research will improve those predictions through a more objective approach to selecting climate data driving the spread of disease. Our experience with this project will have application to other agricultural problems where scientists need to select the climate change model projections for their research.

I am a climate scientist trained in climate model evaluation and selection for research and decision-making applications. During my time as a postdoc at USDA, I plan to work on a range of projects to evaluate climate model performance and to assist in selection of climate projections for specific research applications. Eventually, I plan to develop a web-hosted tool for selecting robust model projections using the results and knowledge gained from these analyses. For scientists who are currently evaluating model performance, this tool will save countless research hours. For scientists who are not looking at model performance before selecting climate projections, this tool will provide more robust results.

I am currently working with the VS Grand Challenge group but am looking for additional collaborations. Please don't hesitate to contact me if you are interested in collaborating!

Do you use SCINet for your research?

Contact SCINet-Newsletter@usda.gov for a chance to be featured in the newsletter!

Atlas Corner

ARS increases computational capabilities through collaboration with Mississippi State University:

Through a collaboration with Mississippi State University, [SCINet Resources](#) were recently expanded to a second HPC cluster (Atlas) housed and maintained by [Mississippi State University's High Performance Computing Collaboratory \(HPC2\)](#). Atlas - now available to those with valid SCINet/Ceres credentials - is a Cray CS500 Linux Cluster with a peak performance of 565 TeraFLOPS utilizing 240 computational nodes, 23,040 2.40GHz Xeon Platinum 8260 processor cores, 101 terabytes of RAM, 8 NVIDIA V100 GPUs, and a Mellanox HDR100 InfiniBand interconnect.

As part of this collaboration, MSU's [Geosystem Research Institute \(GRI\)](#), [Department of Comparative Biomedical Sciences](#), and [Department of Wildlife Fisheries, and Aquaculture](#) are contributing training and postdoc support in a wide range of scientific research methods and technologies (e.g., unmanned aerial systems, surveillance and analysis of plant and animal pests, disease ecology). Collaborative research between ARS and MSU will focus on spatial epidemiology, disease ecology, molecular epidemiology, landscape analysis, and agroecology. All of these research areas are expected to take advantage of the computational power of Atlas.

Meet our SCINet Fellows

Here are introductions to a few of SCINet's first cohort of postdoctoral fellows. SCINet postdocs are tasked with developing cross-ARS unit collaborative research projects leading to a SCINet working group, and working on their own research projects that utilize the ARS SCINet high-performance computing resources (Ceres, Atlas). They will also contribute to non-research projects that further the SCINet Computing Initiative, such as the SCINet website, newsletter, and various computational trainings. We will continue introductions in upcoming issues of the SCINet newsletter. For introductions to Jennifer Chang, Yanghui Kang, and Shawn Taylor, check out the [October 2020 newsletter](#). We're now introducing: **Alicia Foxx, Kerrie Geil, and Melanie Kammerer**.

Alicia Foxx, Ecologist

Alicia started the SCINet Postdoc position in June 2020, working out of Gainesville, FL with Dr. Adam Rivers as her supervisor. She completed her Ph.D. in plant biology and conservation from the joint program between Northwestern University and the Chicago Botanic Garden in Illinois. Her research used restoration-relevant native plants of the Colorado Plateau to inform ecological theory and to provide recommendations on native plant material based on their performance. The work she is performing in the SCINet Postdoc position is applied machine learning on microbial communities and includes interrogating applications of machine learning in bioinformatics and the



impacts of data distribution on algorithm performance. Additionally, her work includes characterizing the seed microbiome through meta-analysis combined with machine learning applications and experimental tests of the seed microbiome to understand vertical transmission of microbes in plants. During her SCINet Postdoc, Alicia will form a new working group to tackle these research topics.

Kerrie Geil, Climate Scientist

Kerrie served as a AAAS Policy Fellow at USDA ARS in Beltsville MD from 2018-2020 and has recently transitioned in July 2020 to a SCINet Postdoc position in Las Cruces, NM in the research group of Dr. Deb Peters. She received her Ph.D. in Atmospheric Sciences from the University of Arizona in 2016. Kerrie has expertise in global, continental, and regional-scale climate dynamics including teleconnections/climate oscillations (ENSO, PDO, etc.), land-atmosphere-ocean interactions, and the North American Monsoon system. In addition to research, Kerrie co-leads the [SCINet Geospatial Working Group](#), organizes computational workshops (such as the [late-summer 2020 Geospatial Workshop](#) and the USDA-ARS/NMSU [Machine Learning and Deep Learning Workshop](#)), and is developing an online archive of tutorials and other computing resources for ARS scientific staff during her postdoc.



Melanie Kammerer, Ecologist

Melanie is an agroecologist and landscape ecologist, and recently started as a SCINet postdoc in State College, PA mentored by Drs. Sarah Goslee and Deb Peters. Melanie recently completed a USDA NIFA pre-doctoral fellowship during her Ph.D. in Ecology at Penn State University. In her graduate research, Melanie studied landscape and climate drivers of wild bee communities using long-term population monitoring data, new field studies, and high-resolution landscape characterization. Melanie is currently developing a generalizable, data-driven method to estimate floral resources for pollinators at the landscape scale, which she hopes to apply to diverse questions, including evaluating USDA conservation practices and mapping floral nutrition. She is organizing a new SCINet Pollinator working group as well as serving on the SCINet software and geospatial common data library committees.



Resources

Contribute / Contact

[Workbooks! for bioinformatics and geospatial workflows:](#)

The best way to learn informatics is through examples of real-world problems. Towards this goal, we have developed a workbook for [bioinformatics](#) that provides the reader with an in-depth understanding of experimental design, data acquisition, data wrangling, data analysis and visualization. This is accomplished through worked-out example problems in each of these sections along with one or more advanced problem sets and corresponding solutions. In addition, we have started to work on a [geospatial workbook](#) to provide similar examples of real-world problems. More workbooks will likely be produced in the future so we have put [all the workbooks](#) in a single location. If you are looking for a place to start your journey in informatics, the [Introduction to Unix](#) tutorial and screen cast is highly recommended. If you have tutorials to share or tutorial suggestions you would like to see made, please send an email to scinet_vrsc@usda.gov.

The [Ag Data Commons](#), a data catalog and generalist repository maintained by the National Agricultural Library, provides public access to USDA-supported digital scientific research data. Ag Data Commons datasets consist of data resources linked or stored locally, and descriptive metadata. The infrastructure supports Digital Object Identifiers (DOIs) for data assets, links to other systems, machine-readable formats, and access via APIs. The Ag Data Commons curation team uses their expertise to ensure each dataset has metadata necessary for increased discoverability. Data in the Ag Data Commons carry an open license with minimal restrictions on access and use: you do not need to register for an account to view and access data. For information on submitting your data to the Ag Data Commons, watch [this video](#) or contact the data curation team with questions.

For questions about this newsletter, to contribute content, feedback on the SCINet website, or SCINet policy and development questions please email SCINet-Newsletter@usda.gov.

SCINet Newsletter Editors: Amy Hudson and Margaret Woodhouse

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

SCINet Leadership Team

Deb Peters, Acting Chief Science Information Officer

Rob Butler, Acting SCINet Project Manager

Adam Rivers, Science Advisory Committee (SAC) Chair

Brian Scheffler, Ex Officio

[SCINet Website | Comments](#)

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

