View email as a webpage | SCINET website

USDA-ARS SCINet Newsletter: October 2020

Contents

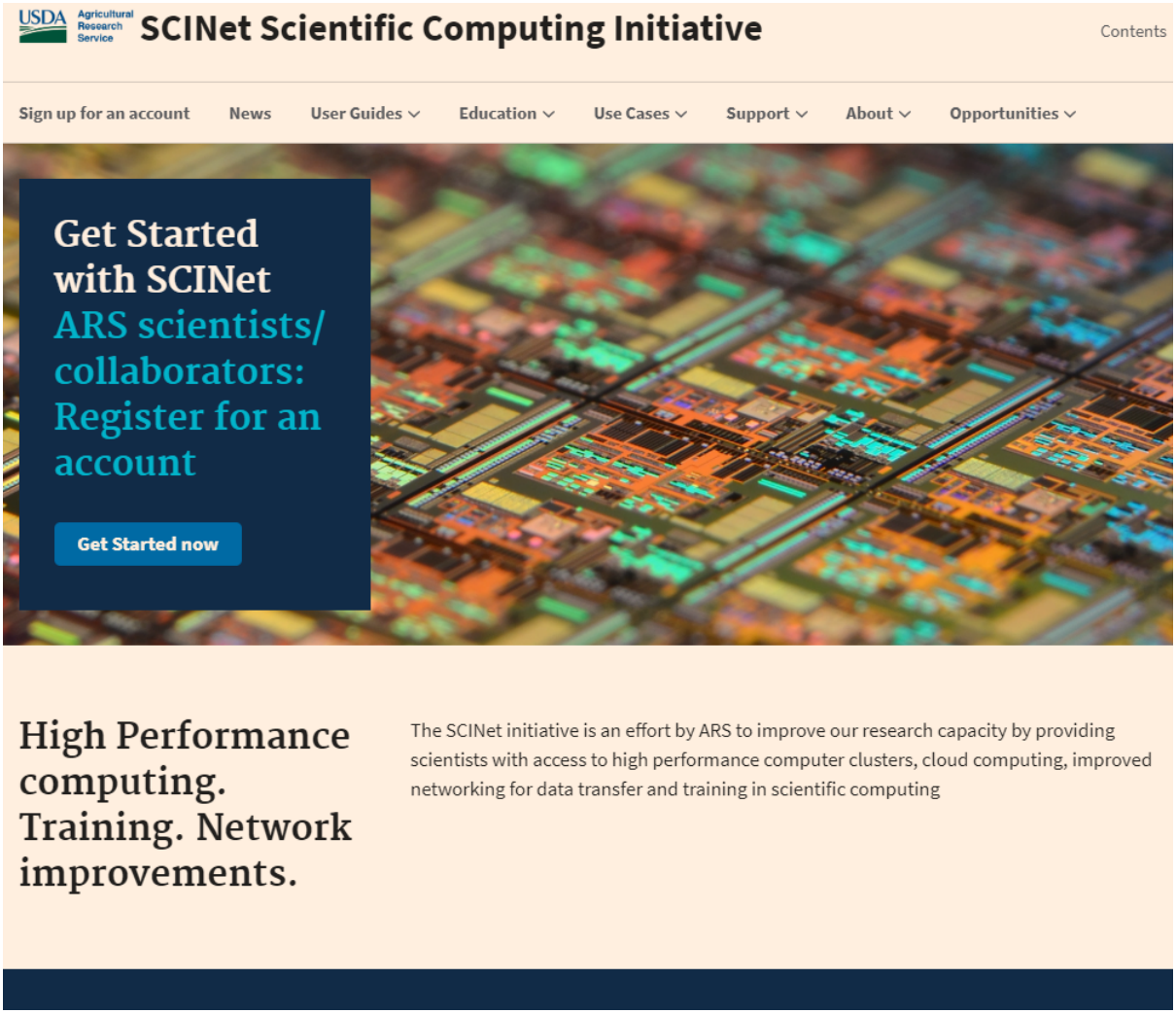
How to Get Started

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 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



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

SCINet User Tips

Large short-term data storage:

- If your analysis requires a lot of temporary disk space that is above your project directory quota, you can use unlimited short-term storage **/90daydata/your_project_name**. Files older than 90 days will be automatically deleted.
- You can also use **/90daydata/shared** to share data with other users that don't have access to your project directory. Note that anyone on the system will be able to read that data.

R users:

- If you have functions that you always want to have when you start a new session in

R then you can place those functions into an .Rprofile file in your home directory: **~/.Rprofile**.

- One useful function if you use R from the terminal and not in RStudio is the wideScreen function. This function will set the text wrapping to the width of your terminal:

```
wideScreen <- function(howWide=Sys.getenv("COLUMNS")) {
  options(width=as.integer(howWide))
}
```

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:

- The **SCINet Geospatial Research Working Group held a series of workshops and training sessions** in August and September 2020 to make progress on working group technical projects, provide hands-on learning experience using the ARS Ceres high-performance computing (HPC) system, and inspire new research ideas. The sessions included 1) an annual meeting of the working group, 2) HPC and linux basics tutorial, 3) Python Dask for distributed computing tutorial, 4) computational reproducibility and collaboration with Git, Conda, and containers tutorial, 5) machine learning using gradient boosting from scikit-learn tutorial and 6) a symposium on the use of AI techniques in agricultural research. Almost 90 ARS scientists, scientific staff, and University collaborators representing many different disciplines and ARS National Programs were in attendance. Detailed information on all the sessions, including the tutorials (which anyone can work through at their own pace), can be found on the session tabs of the SCINet Geospatial Research Workshop 2020 website. Recordings will be posted soon.
- **SCINet supported online Carpentries Workshops** were offered in July and August 2020. These were a series of 4, 2-day workshops, which provided scientists and support staff with hands-on instruction to become more comfortable at the command line, introduce them to document version control with Git, and start them on the path to building their coding skills, in either R or Python, so that they can better utilize the USDA computing resources available to them for their research. Almost 80 scientists, post-docs, and collaborators attended these workshops. Please look for further Carpentry trainings in FY21 to be featured on the Upcoming Events page.
- **Coursera.org certified courses update:** The SCINet initiative has purchased 75 licenses and the AI Center for Excellence has purchased 50 licenses for ARS researchers to take Coursera courses with certification. Information about how to

obtain a license is expected to be emailed soon and will be posted on the Free Online Training page of the SCINet web site.

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

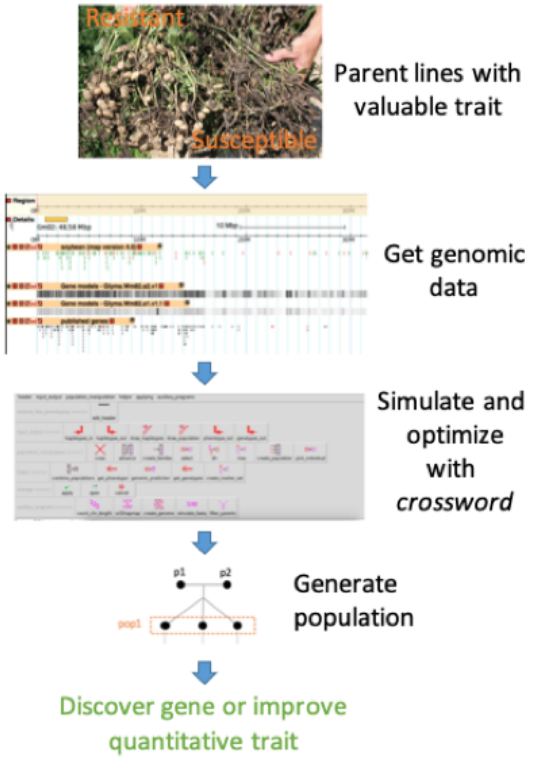
Asian giant hornet genome quickly sequenced by SCINet's Ag100Pest Initiative working group

Image from Brian Scheffler

The Asian giant hornet is invasive to North America and could potentially threaten native bee populations and pollination pathways. The first Asian Giant hornet nest was found in North America in September 2019. After receiving DNA from a hornet in that nest in May 2020, researchers in the SCINet Ag100Pest Initiative working group (a subgroup of the Arthropod Genomics Research (AGR) working group) used SCINet computing resources to quickly produce a reference genome assembly by the end of June.

Conducted as part of the Ag100Pest Initiative, the quick turnaround time from obtaining the sample to producing the genome assembly is promising for invasive insect management-such as pinpointing where the nest in Canada originated from. The genome is publicly available in the AgDataCommons and more information can be found here in a recent USDA ARS Press Release.

Simulations to help constrain experiments: identifying genes with high potential to increase crop yields



By Justin Vaughn

Many technological developments made the Green Revolution possible, including selective breeding to promote specific plant genes that increased crop yields. For example, substituting the common version of the oxidase gene with a broken version (*sd-1)* in rice led to plants of shorter stature, which facilitated mechanized farming and a resultant step change increase in food yield (Spielmeyer et al., 2002). Today, opportunities to use plant genetics to maximize crop yields are greater than ever. Geneticists can generate large suites of different versions of any gene using CRISPR-Cas9 technology. The history of *sd-*

1 suggests that such capacity will have a profound impact on agriculture and global food production. Unfortunately, there remains a critical gap in our ability to discover which genes we should target for such exploration.

The USDA's Genomics and Bioinformatics Unit uses genomics and phenomics to improve the efficiency of identifying genes with high potential to increase crop yields. Computer simulations of controlled crosses allow researchers to explore experimental scenarios without the time, expense, and manual labor associated with physical experiments. We developed software, *QTLsurge*, to simulate specific kinds of experiments. *QTLsurge* runs in the R statistical programming environment, so it can be used on a local computer or a remote cluster like Ceres. We also developed an open source simulation platform, *crossword*, that can be used for a broad range of crops and breeding systems (Korani and Vaughn, 2019). Unlike many available tools, *crossword* accepts empirical genomics data as a starting point and thus gives an accurate and directly applicable reflection of likely experimental outcomes.

Our simulations supported the commonsense prediction that larger populations offer much more capacity to finely resolve genes. Importantly, this signal is not masked by experimental noise. Less intuitively, for large populations, phenotypic replication was unnecessary and the number of extreme individuals needed for sampling could vary from 5% to 20% without affecting resolution. These results have a substantial impact on reducing the manual labor required for experiments of this scale.

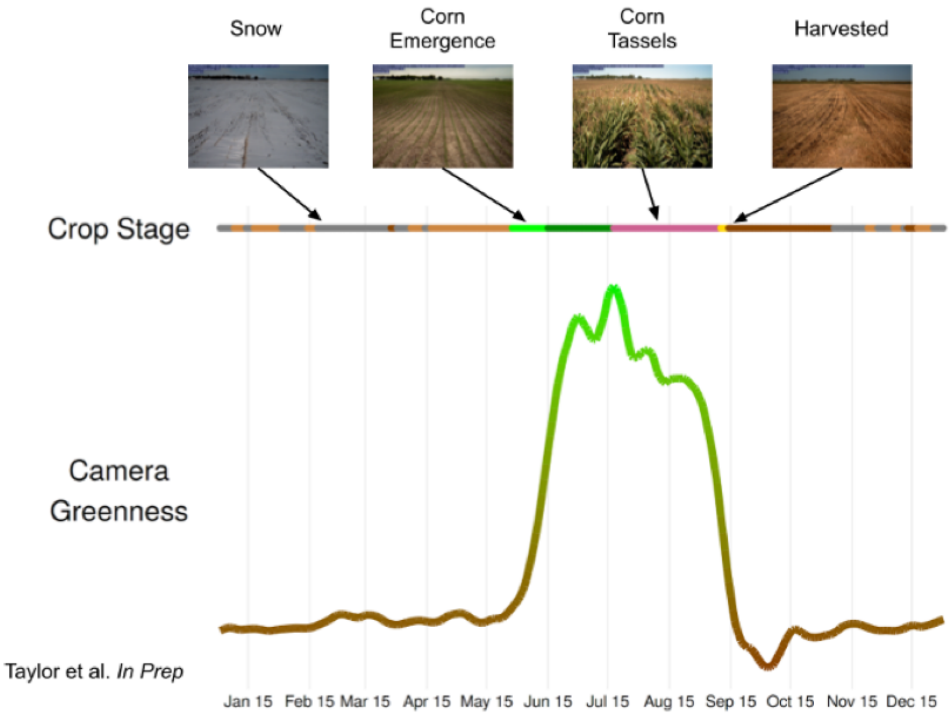
SCINet Note: This study required substantial computational resources to simulate huge populations across a range of parameter sets. The full study assessed ~5,000 parameter combinations and required thousands of CPU hours. While this effort would have taken years on a personal computer, we were able to get results in three weeks by distributing jobs across a subset of Ceres CPUs. For more information about using SCINet for plant breeding research, visit the SCINet website.

References:

Korani, W., Vaughn, J.N. 2019. *Crossword*: A data-driven simulation language for the design of genetic-mapping experiments and breeding strategies. *Scientific Reports* 9, 4386. https://doi.org/10.1038/s41598-018-38348-y

Spielmeyer, W., Ellis, M.H., Chandler, P.M., 2002. Semidwarf (sd-1),"green revolution" rice, contains a defective gibberellin 20-oxidase gene. *Proceedings of the National Academy of Sciences* 99 (13), 9043-9048. https://doi.org/10.1073/pnas.132266399

Identifying crop phenology with deep learning



By Shawn Taylor* and Dawn Browning

**Shawn 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 Shawn and other featured fellows Jennifer Chang and Yanghui Kang.*

There is an enormous amount of data currently being generated to research best practices for agriculture. Satellite data, imagery from drones and near-surface cameras, sensor data from meteorological stations and eddy covariance towers, and traditional ground observations are among the data continuously collected across the 18 sites in the Long Term Agricultural Research (LTAR) Network. Combining different data sources can potentially provide insights which are not possible using any single source, but harmonizing these diverse sources is a challenge due to differences in temporal and spatial resolution. The SCINet LTAR Phenology Working Group is exploring methods and limitations to large-scale data integration from multiple sensor arrays available to ARS scientists.

One sensor system, the PhenoCam network, is a global array of near-surface cameras used to track vegetation processes by looking down on the canopy. With hundreds of cameras taking up to 48 photos a day, this data stream has provided novel insights into many biological processes, yet also presents a big data challenge (Kosmala et al. 2018). The primary output from PhenoCam imagery is a greenness metric which correlates well with above-ground plant processes in most areas, but can provide confounding signals in agricultural fields due to practices like harvest timing and crop rotations.

One under-used aspect of PhenoCam images is their contextual information which can include crop type and timing of management activities. With millions of PhenoCam images, identifying these features by hand is impossible. We combined PhenoCam images with deep learning methods to automatically identify crop stages such as timing of emergence, flowering, and harvest. Crop phenology, or the timing and duration of discrete crop stages, can be used to complement data or metrics from other sensors. For example, daily carbon fluxes estimated via eddy covariance measurements could potentially be partitioned by crop stage, which before would have required destructive sampling throughout the growing season. Croplands cover a large portion of the land surface, and so crop phenology is vital to understanding large-scale greening patterns across the globe. The ability to identify crop phenology in a scalable, repeatable, and automated way using millions of images represents an agricultural innovation that can be used to develop a global database of crop phenology observations (Hufkins et al. 2019). This database could be used to verify model outputs in regional to global remote sensing studies or large-scale research of carbon dynamics.

SCINet Note: We used the Ceres HPC to train a deep learning image classification model to identify the crop stage from PhenoCam images. Ceres was then used to classify 140,000 individual images from LTAR locations totalling 54GB of data. Our model was derived from open source deep learning models implemented in the Python TensorFlow package.

References:

Hufkens K., Melaas E.K., Mann M.L., et al. 2019. Monitoring crop phenology using a smartphone based near-surface remote sensing approach. *Agricultural and Forest Meteorology*. 265:327–337. https://doi.org/10.1016/j.agrformet.2018.11.002

Kosmala M., Hufkens K., Richardson A.D. 2018. Integrating camera imagery, crowdsourcing, and deep learning to improve high-frequency automated monitoring of snow at continental-to-global scales. *PLoS One* 13:1–19. https://doi.org/10.1371/journal.pone.0209649

Do you use SCINet for your research?

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

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-site collaborative research projects that utilize the ARS SCINet high-performance computing resources. 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. First up we have **Jennifer Chang**, **Yanghui Kang**, and **Shawn Taylor**.

Jennifer Chang, Bioinformatics Scientist

Jennifer Chang grew up in Wisconsin and has been programming since 2006. At Cornell College, she double majored in Biochemistry and Computer Science graduating in 2011. In 2017, Jennifer earned a Ph.D. in bioinformatics from Iowa State University. During her Ph.D., her research software “Mango Graph Studio” led to co-founding a software company and working on two Department of Defense Small Business Innovation Research (SBIR) contracts. She eventually left the company and was a postdoc with Dr. Amy Vincent at USDA-ARS, automating swine influenza reports from 2017-2020. In June 2020, Jennifer shifted to a SCINet postdoc position with Dr. Andrew Severin and Dr. Brian Scheffler where she collaborates on the bioinformatics workbook. She’s still in Ames, Iowa, learning slurm and automating pipelines for different hardware architectures. She hopes to write flexible general-purpose pipelines that reduce tedium and increase joy of discovery. Jennifer cares about collaboration, welcoming environments, and teaching.

Yanghui Kang, Physical Scientist

Yanghui started her SCINet Postdoc position in May 2020 after receiving a Ph.D. degree in Geography from the University of Wisconsin-Madison. She works with Dr. Feng Gao and Dr. Martha Anderson at the Hydrology and Remote Sensing Laboratory in the Beltsville Agricultural Research Center, Beltsville, Maryland. Yanghui’s research projects have focused on the large-scale high-resolution monitoring of core agroecosystem variables (e.g., Leaf Area Index (LAI), crop yield), with the help of satellite remote sensing, machine learning, crop growth modeling, and data assimilation techniques. At ARS, Yanghui is currently developing a machine-learning-based approach to map LAI from Landsat and Sentinel-2 images over the entire globe. She is also interested in deriving crop phenological stages from satellite observations and monitoring agroecosystem dynamics through data assimilation. Yanghui is bringing her experience working with big data to construct a SCINet common data library for the geospatial community, allowing us to optimize storage space on our HPCs.

Shawn Taylor, Ecologist

Shawn obtained his Ph.D. from the University of Florida in 2019, where he researched best practices in ecological forecasting and implemented a continental-scale phenology forecast. He is currently a postdoc at the USDA-ARS Jornada Experimental Range in Las Cruces, NM working with Dr. Dawn Browning, and will transition to his SCINet postdoc in October. Shawn is a member of the SCINet LTAR Phenology Working Group and participated in a 2019 workshop in Las Cruces, NM focused on devising collaborative workflows on the SCINet high-performance computing (HPC) system, Ceres. He is currently integrating large streams of sensor data across the Long Term Agricultural Research (LTAR) network to help promote increased production and sustainability.

New Tools

Reads2Resistome streamlines the process of turning a bacterial culture into an informative annotated genome, performing both genome assembly and in-depth genome characterization. Users with experience in Linux basic commands can analyze bacterial genomes sequenced using either short and/or long read sequencing technologies. Reads2Resistome takes fastq reads as input and performs assembly, annotation and genome characterization with the goal of producing an accurate and comprehensive description of the bacterial genome and collection of all the antibiotic resistance genes, virulence genes, and other resistance elements within the chromosome, plasmids or bacteriophage.

Contribute / Contact

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.

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

SCINet Leadership Team

Deb Peters, Acting Chief Science Information Officer

Stan Kosecki, 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