SCINet Newsletter: July 2023

Research Spotlight | News | Training | Support | Connect

RESEARCH SPOTLIGHT

Research Spotlight: Cattle Genome, Pangenome, Annotation, and FarmGTEx

By George Liu, Research Biologist (Bioinformatics) and Zhenbin Hu, Research Computational Biologist/ SCINet Postdoctoral Fellow

Animal Genomics and Improvement Laboratory, BARC, ARS, USDA, Beltsville, MD

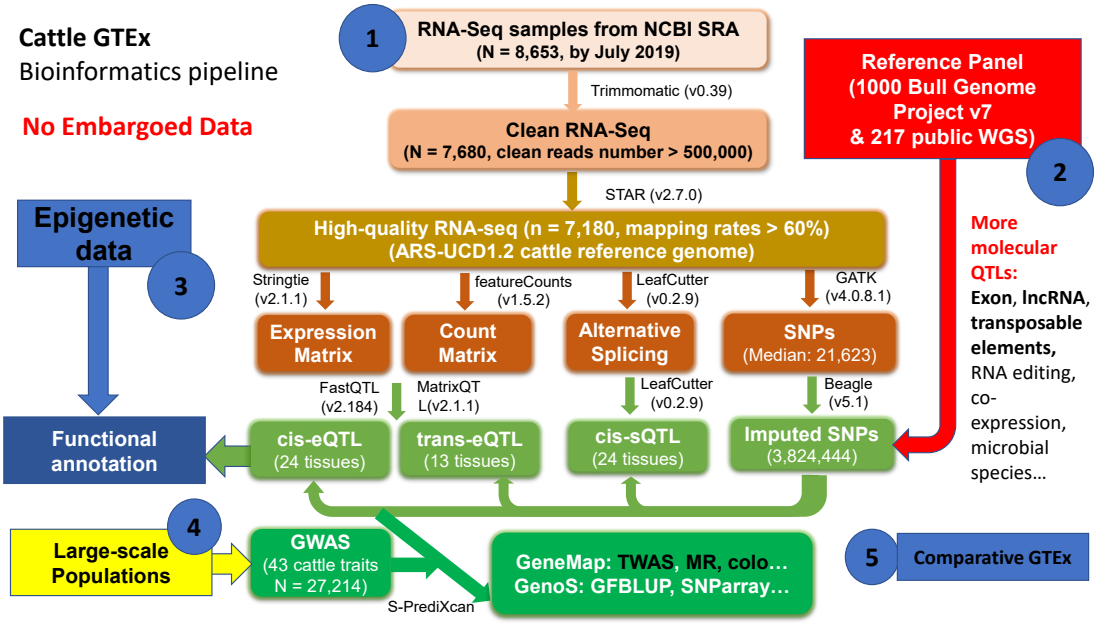


Figure 1. An overview of FarmGTEx-cattle data analysis and mining framework detailed in **A multi-tissue atlas of regulatory variants in cattle**. *Nature Genetics* 2022

The Animal Genomics and Improvement Laboratory (AGIL) has a primary goal of discovering and developing improved methods for genetic and genomic evaluation of economically important traits in dairy animals and small ruminants. It also conducts

fundamental genomics-based research to enhance animal health and productive efficiency. Dr. George Liu and his team have been diligently building comprehensive data resources and providing data analysis and mining tools for translational omics research.

Dr. Liu is also a co-founder of the Farm Animal Genotype-Tissue Expression (FarmGTEx) Consortium, which aims to create a comprehensive public resource for studying tissue-specific gene expression and regulation in major livestock species, including cattle, pigs, sheep, goats, and chickens. Since its official launch, the FarmGTEx Consortium has garnered interest from nearly 400 researchers in 48 countries. As a part of FarmGTEx, Dr. Liu co-led the development of the FarmGTEx databases for cattle, pigs, and chickens, with the CattleGTEx paper featured as a Nature Genetics cover story last year. FarmGTEx-related research has produced complete, open-access catalogs of regulatory elements for cattle, pigs, and chickens which are available on a public portal (https://www.farmgtex.org/), providing immense utility to the livestock community and industry. These resources collectively serve as primary references for animal genomics, breeding, adaptive evolution, comparative genomics, and veterinary medicine.

A critical factor contributing to our success has been the utilization of SCINet's high-performance computing clusters (HPCs), including Ceres and Atlas. From the very beginning, SCINet has been an integral part of our large-scale data-driven research, enabling the processing of over 500 TB of omics data for mining. Our analyses have encompassed more than 5,000 whole-genome sequences and over 40,000 transcriptome datasets, yielding a wealth of genetic information critical for animal improvement. Additionally, SCINet resources facilitated the development of the Cattle Gene Atlas and enabled the first-ever transcriptome comparison between humans and cattle.

SCINet also provides us with a computing environment that empowers us to develop new data analysis tools using artificial intelligence (AI) and machine learning (ML). Dr. Zhenbin Hu, a SCINet/AI-COE Postdoctoral Fellow working alongside Dr. Liu, is currently employing ML to train a Convolutional Neural Network (CNN)-based model to construct high-quality databases of structural variations. He plans to customize and apply the Sparse Conditional Gaussian Graphical Model and other frameworks to the CattleGTEx data. Furthermore, Dr. Hu is also actively processing large-scale genomics data for the Sheep/GoatGTEx project as a core member. Looking ahead, we are confident that the large, distributed computing and GPU resources of SCINet will be essential for developing new pipelines, enabling more efficient execution of massive and complex data analysis jobs.

Dr. Liu and Dr. Hu are pleased to announce the formation of a new SCINet working group, the Translational Omics Working Group, to help facilitate omics-related research in ARS. Please see the article in the “News” section, below, for more information about this working group, including how to join!

SCINet and AI-COE Fellows

Welcome **Dr. Olivia Haley**! Dr. Haley is currently working on the Maize Genetics and Genomics Database (MaizeGDB) project under the mentorship of Dr. Carson Andorf at the Corn Insects and Crop Genetics Research Unit in Ames, Iowa. She obtained a B.S. in Biological Sciences from the University of South Carolina and an M.Sc. in Plant Sciences from McGill University. Prior to her doctorate, Dr. Haley worked in a food microbiology testing lab where she gained critical insight into the food safety needs of the produce industry. For her doctoral dissertation at Kansas State University, she studied applications of UV-C light to inactivate pathogens in the fresh produce supply chain and earned her Ph.D. in Horticulture and Natural Resources in 2023.

She joined MaizeGDB to pursue her curiosity in machine learning. Her research focus is the use of artificial intelligence for protein functional annotation and structural prediction in *Zea mays* (maize), and she is fascinated by protein-protein interactions between maize and toxigenic fungi.

Welcome **Dr. Aaron Yerke**! Dr. Yerke received his bachelor’s degree in biology from East Carolina State University in 2011, and his master’s degree in biomedical sciences from East Carolina’s Brody School of Medicine in 2013. He wrote his master’s thesis under the supervision of Dr. M.D. Motaleb, which focused on the lifecycle of *Borrelia burgdorferi*, the causative agent of Lyme disease. He then joined Dr. Shengmin Sang’s lab at

North Carolina Agricultural and Technical State University at the campus in Kannapolis, NC, as a lab manager and technician. Here he gained wide exposure to various laboratory techniques, including cell culturing, mouse studies, and gut microbiome studies, and data analysis.

In 2017, he began his studies in the PhD program of the Bioinformatics and Genomics department at the University of North Carolina at Charlotte. Here, he worked on various projects such as open-source software development, metabolic and sequencing data analysis, and compositional data analysis. His dissertation research, under the supervision of Dr. Anthony Fodor, focused on assessing the impacts of various data transformations on machine learning accuracy. He graduated in 2022. In 2023, he joined the Food Components and Health Lab at the Beltsville Human Nutrition Research Center as a research fellow under the supervision of Dr. David Baer and Dr. Lauren O'Conner. He is interested in using his skills to explore machine learning applications in biomarker discovery.

In addition to his research, Dr. Yerke enjoys solar cooking, gardening, and exploring local trails and greenways with his wife and son.

NEWS

SCINet/AI-COE Graduate Student Internships Update

The first year of the AI Center of Excellence (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 26 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; 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 efforts of these students, we are holding an internships research symposium on August 7, 2023 from 11am-5pm EDT. 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 pilot internships program.

AlphaFold Workshop

On June 27 and 29, the Protein Function and Phenotype Prediction Working Group hosted a two-day, hands-on training workshop focused on AI-based protein structure prediction and functional annotation. A total of 34 attendees registered for and participated in the workshop.

The first day of the workshop focused on how to run Alphafold with GPU nodes on both HPC clusters, Ceres and Atlas. During this session, Dr. Hye-Seon Kim demonstrated how to log into SCINet and transfer data to/from the HPC clusters, run SLURM batch jobs, and visualize the Alphafold results via publicly accessible web search tools (e.g, VAST+ for comparing 3D structures and Foldseek for fast/accurate protein search). This session helped scientists without prior HPC experience become familiar with the command line and batch job submission for Alphafold.

On the second day of the workshop, the focus shifted to functional annotation using protein sequences, structures, and embeddings. During the session, Dr. Carson Andorf demonstrated how researchers could utilize the Diamond and FoldSeek tools on the Ceres platform to assign functional Gene Ontology terms to query proteins. This approach provides valuable insights into the functions of these proteins, aiding in various biological studies and research.

Dr. Andorf concluded the workshop by showing how embeddings from a protein language model can serve as input for a machine learning method, enabling researchers to add confidence scores to thousands of Gene Ontology terms. This approach opens up new possibilities for efficient functional annotation and further advancements in protein research.

Recordings from these workshops will soon be accessible on the Protein Function and Phenotype Prediction Working Group page of the SCINet website.

Group Licensing for ASReml

Breeding Insight OnRamp Director Amanda Hulse-Kemp, SCINet/AI-COE fellow Keo Corak, and USDA Network Administrator Rob Butler are exploring group licensing options for the popular statistics software ASReml and/or ASReml-R.

You can fill out this survey to indicate your interest in joining an ASReml group purchase: https://forms.office.com/g/BiFSuzm9zW.

For questions about this survey, please contact Amanda Hulse-Kemp (amanda.hulse-kemp@usda.gov) or Keo Corak (keo.corak@usda.gov).

Invasion Genomics Symposium

SCINet Fellow Rebecca Clement is organizing a symposium on "Invasion Genomics" at the Entomology Society of America meeting this November. If you are using some sort of genomics (or other "omics") tool to study biological invasions for an invasive group, or know someone who might be interested in giving a 15-minute talk at this session, please contact Dr. Clement (rebecca.clement@usda.gov). Early career scientists and scientists from diverse gender/ethnic backgrounds are especially encouraged to inquire.

Translational Omics Working Group

We are excited to introduce the SCINet Translational Omics Working Group, an interdisciplinary team of researchers and experts dedicated to advancing the field of translational omics and its applications in agriculture. Omics technologies will play a pivotal role in the new agricultural revolution by unraveling the complex molecular underpinnings of complex traits and helping guide future farming practices.

The concept of "omics" refers to the comprehensive study of various biological molecules, including genomics (structural genomics - sequence assembly and variation and functional genomics - gene function and annotation), epigenomics (changes in gene expression regulation without change in DNA sequence), transcriptomics (gene expression analysis), proteomics (proteins and their interactions), metabolomics (small molecule metabolites), and metagenomics (microbiota - a community of microorganisms). These approaches have already demonstrated immense potential in understanding molecular mechanisms, finding biomarkers, and developing effective selections and treatments. The primary goal of the Translational Omics Working Group is to foster collaboration, knowledge-sharing, and innovation among researchers and experts in diverse fields, including but not limited to genomics, bioinformatics,

computational biology, and artificial intelligence (AI). Together, we aim to overcome research and technical challenges, explore novel techniques, and promote omics data integration in agriculture and food research.

Key objectives of the Translational Omics Working Group are:

- *Education and Outreach:* Organize seminars, workshops, and conferences to disseminate knowledge (tools and resources) and raise awareness about the potential of omics technologies in transforming agricultural research.
- *Data Integration:* Facilitate discussions and methodologies for integrating multi-omics data to comprehensively understand biological processes and molecular mechanisms.

Other potential applied topics include:

- *Biomarker and Treatment Development:* Promote research and validation of omics-based biomarkers for crop and animal selection. Explore the utility of omics technologies in identifying potential targets and improving the efficiency of disease management.
- *Ethical and Legal Considerations:* Address the ethical, legal, and privacy challenges associated with omics data while ensuring that our research adheres to the highest standards of data security and animal welfare. Collaborate with veterinarians and animal caretakers to effectively translate omics research into animal husbandry practice, ensuring the benefits reach animals promptly and responsibly.

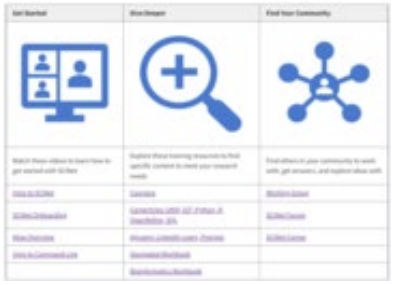
By fostering a dynamic environment for interaction and cooperation, we believe the Translational Omics Working Group will play a significant role in advancing AI applications to omics research and bringing us closer to the new agricultural revolution. We invite all ARS researchers and supporting professionals interested in omics and AI and their translational potentials to join us on this exciting journey. Together, we can make a tangible impact on ARS's research mission and revolutionize the future of farming.

If you are interested in and would like to join the group, please **fill out the following survey Translational Omics Working Group**.

For more information, please contact George Liu (George.Liu@usda.gov) or Zhenbin Hu (Zhenbin.Hu@usda.gov).

TRAINING

Training Opportunities



Getting Started: 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.

The Carpentries Workshops:

The Data Carpentries Genomics Workshop has been postponed and is projected to be rescheduled in October.

Geospatial Research Working Group:

The SCINet Geospatial Research Working Group's Annual Workshop is set to take place September 25-29, 2023. This year's theme is "Machine Learning and Deep Learning with Geospatial Data" and will feature lightning talks, interactive tutorials, and community discussions around research topics. If you're interested in attending and joining our working group, please fill out this quick survey or reach out directly to leads Heather Savoy and Amy Hudson. Recordings of previous trainings from the working group are available on the SCINet/AI-COE Training & Workshops channel, including last year's Annual Workshop sessions and the June 2023 presentation on the Geospatial Data Act of 2018.

Courses by Mississippi State University: Mississippi State regularly offers Introduction to Atlas courses. Additionally, there are waiting lists available for several other courses, including an Intensive R course to help scientists with no R experience become familiar with the programming language and start performing statistical analyses in 4 days. Sign up to get notified when these courses are offered.

Coursera.org Courses: 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.

Training opportunities are continuously being updated on the 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

There are now more than 2,000 registered SCINet users. If you do not already have a SCINet account, we hope you will consider joining the 2,000+ 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.

Reminder: SCINet Account Request Approvals

SCINet policy requires that *all* requests for sponsored SCINet accounts (that is, SCINet accounts for non-ARS users) must be reviewed and approved by someone other than the account requester/sponsor. The secondary reviewer will typically be the supervisor of the account sponsor. To ensure that SCINet remains compliant with account security requirements, we cannot create sponsored user accounts if the same person attempts to serve as both account sponsor and secondary reviewer.

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

Multi-factor authentication (MFA)

Phishing-resistant MFA is more secure than traditional MFA because it reduces the ability of attackers to trick users into providing their authentication credentials. Phishing-resistant MFA requires a user to possess and present a specific physical device or token when logging into a system. Without the token, the login cannot occur.

Most USDA staff are already familiar with using a form of phishing-resistant MFA: their USDA LincPass credential. The LincPass is now being adopted by SCINet for phishing-resistant MFA login. All USDA SCINet users will soon use their LincPass card when logging into SCINet resources. For those without a LincPass, a YubiKey is a hardware authentication device which looks like a USB thumb drive. This physical security key will function similarly to the LincPass card. The Google Authenticator code is not considered to be a phishing-resistant authentication method and will no longer be supported.

Authentication With LincPass and YubiKey

The login processes for GUI services, such as Open OnDemand, Galaxy, and the SCINet Forum, have been changed for these new methods:

If you have a LincPass, select the option of “USDA LincPass”.

After selecting this, you will be automatically directed to login using your usual eAuth-based login.

If you are accessing using YubiKey, you will enter your SCINet credentials (username and password) and click “Sign In”. Then you will be directed to a new screen showing available security keys.

You will select “Sign in with Security Key”. A pop-up will appear asking you if you would like to use your passkey. You will select “Use a different device” in the bottom left corner.

The next pop-up will have three options. You will select “USB security key”. The final pop-up will instruct you to insert your security key and touch it. You will now insert your USB YubiKey (if you haven’t already) and then touch it. This will then automatically log you into the service you were attempting to access.

The ssh login process has had similar changes. Further instructions, which are actively being updated, for all login processes can be found on the SCINet website’s login guide.

If you need assistance with this login process, please email your questions to scinet_vrsc@usda.gov.

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

SCINet Corner: First 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.

The next SCINet Corner will be held Thursday, August 10, 2023 at 1 pm EDT. 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 Team

Every newsletter highlights SCINet community members as a way to connect the ARS scientific computing community. To see all the SCINet community 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 Leadership Team

Brian Stucky, Acting Chief Science Information Officer
Rob Butler, SCINet Program Manager
Jeremy Edwards, 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