Arthropod Genomics Research Workshop II

Research team and scientific community building summary document

George Washington Carver Center
Beltsville, MD
July 26 to 28, 2016

Background

Thirty three members of the Arthropod Genomics Research (AGR) community representing all five Areas of the Agricultural Research Service (ARS) convened a workshop charged with building collaborative research communities and improving communication channels in the Agency. Participants assessed how the current knowledge, skills and abilities of the AGR scientists impact the effectiveness of genomic data analyses, and affect the capacity of the Agency to address arthropod-specific challenge areas; 1) *Controlling Vectored Diseases,* 2) *Managing Herbivorous Insects*, and 3) *Implementing Biocontrol*, as well as meet the Grand Challenge of "reducing agricultural inputs and environmental impact by 25% while increasing output by 25% by 2025". Participants were charged with forming long-term scientific teams that will support the exchange of scientific knowledge and expertise. The workshop was structured to evaluate gaps in knowledge and communication that currently exist, and to subsequently propose solutions that would increase the capacity of ARS to pursue genomics-based scientific inquiry. This task is aligned with the ARS Big Data Vision aimed to "develop a well-trained scientific and technical staff" and "enhance knowledge, skills and abilities within the ARS workforce" (USDA-ARS, 2013).

Arthropod research is highly diverse with respect the scope and diversity of species it encompasses, and thus presents unique challenges for data collection, analysis, and management. In contrast to genomic research within other National Program Areas where a relatively rich set of resources are being developed for a small number of model species, the AGR community is tasked with managing varying depths of data from among a larger number of more evolutionarily diverse species. This complexity often results in research teams composed of a small number of individuals devoted to the study of a species, wherein the breath of expertise may be insufficient to address the complexity of genomic data analyses. The distribution of resources that empower AGR scientists to conduct genomic research is crucial for achieving goals within arthropod-specific challenge areas using the high performance compute (HPC) investments made by ARS.

Furthermore, the impact of arthropods on agriculture transcends ARS program areas including crop production and animal health, which logically positions the AGR for multidisciplinary research. Open access to information and seamless communication is essential for conducting multidisciplinary research. Building teams within and across National Programs will require improved communication systems that interconnect scientists, database managers, application specialists, and administration. Changing the paradigm by which research is conducted at the ARS necessitates the formation of a partnership between the scientists and administration to adapt policies in order to facilitate a capable workforce through the distribution of expertise and explore novel means of communication.

Outcomes of the workshop are presented in the following summary document, wherein the current status of knowledge and mechanisms of communication among scientists are critically evaluated. The result of a solutions-based gap analysis are described along with the formulation of possible paths that will enhance the capacity of ARS to more effectively perform genomic-scale data analyses.

Bridging knowledge gaps with proposing solutions

Four Scientific Teams were built among AGR workshop participants; 1. *Assembly and Annotation*, 2. *Genotyping*, 3. *Gene Expression and Proteomics*, and 4. *Gene Silencing and Editing* (See **Outcome 1**). Each team performed a gap analysis which entailed the critical comparison of current status of knowledge with the level of performance needed to effectively analyze genomic data. Team evaluations also included a set of potential solutions that could lead to the effective distribution of knowledge, resources, and expertise across the Agency. The salient outcomes of these analyses are as follows:

***Table 1:** Addressing gaps in research design, and data collection, analyses, and distribution*

Gap 1a:	Lack of familiarity with research project designs
Rational 1a:	Aligning methodology with the biological question will increase the efficiency of research, and structure experimental designs and data collections to address research goals.
Solution 1a:	Provide mechanism for peer-to-peer discussions in real-time via message boards or instant messaging, as well as peer-built documentation describing best practices and workflows.
Gap 1b:	Unfamiliar with requirements for input materials among different sequencing technologies
Rational 1b:	Inappropriate extraction or sequence data collection is costly; time and fiscal resources
Solution 1b:	Provide mechanism for peer-to-peer discussions in real-time via message boards or instant messaging, as well as peer-built documentation describing best practices and workflows.
Gap1c:	Difficulty identifying and contracting with sequence providers and core facilities
Rational 1c:	Outsourcing large-scale sequencing is the most fiscally responsible option, but requires excessive time to establish contracts according to procedures within ARS Procurement and hinders the progress of research
Solution 1c:	Identify sequencing facilities and work with Procurement to provide an ARS-wide template for establishing contracts, or develop a mechanism for establishing blanket purchase agreements
Gap 1d:	Insufficient ability to develop or apply computational methods and pipelines
Rational 1d:	Analyses require several computer applications and scripts. Appropriate parameters and sequential use of programs and reformatting of data is needed successfully complete tasks
Solution 1d:	Facilitate collaborative work by ARS scientific teams to seek external input on best practices. Develop peer-written documentation or training resources. Hold data Carpentry Workshops.
Gap 1e:	Data submissions to public databases are non-intuitive
Rational 1e:	Data collected but never published or deposited in public databases have no value to stakeholders. Datasets that are atypical have no available database resources
Solution 1e:	Work with GenBank to develop a submission portal for common datasets. Utilize the Ag Data Commons at the National Ag Library (NAL) to deposit analyzed datasets and obtain digital object identifiers (DOIs). Report these DOIs as accomplishment(s) in Annual Reports
Gap 1f:	Lack of tools for metabolic systems or pathway analyses as part of annotation; downstream analysis of gene expression experiments are lacking.
Rational 1f:	Tools for post-analysis are crucial in order to derive biological relevance & guide additional experiments, or develop novel insect control measures.
Solution 1f:	Develop a software pipeline for mapping gene expression data onto pathways, and provide training to scientists.

Developing effective mechanisms of communication and peer-to-peer learning

During the workshop, researchers in the AGR formed breakout groups that discussed needs for communication strategies within ARS and proposed likely solutions that would facilitate the advancement of research programs. This evaluation took into account the previous assessment of gaps in research knowledge, and was charged with formulating mechanisms to increase the ease of communication and exchange of knowledge and expertise among scientists. The following synthesis of common gaps across the AGR community and proposed solutions to be implemented are below:

Table 2: *Addressing gaps in communication and peer-to-peer knowledge sharing*

Gap 2a:	Difficulties in finding collaborators within ARS
Rational 2a:	Need a mechanism to finding scientists with common research goals or complementary skills facilitate discoveries, novel output, and addressing Stakeholder needs
Solution 2a:	Provide enhanced employee searchable personnel pages with description of research project(s) & skills, including computation programs and method, and keywords; Develop an ARS Arthropod Genomics page or group at NAL.
Gap 2b:	Direct peer-to-peer interactions as learning mechanisms are lacking
Rational 2b:	Consultation and communication are effective means for the transfer of knowledge, especially for novices; needed in order to avoid common mistakes & resolve impasses
Solution 2b:	Forums, direct mentoring, and intensive laboratory visitations (ARS-sponsored sabbaticals within and outside ARS) need to be encouraged through recognition of technology transfer in reports and travel funding. Develop a listserv, virtual webinar series and in-person workshops to facilitate learning
Gap 2c:	Mechanism needed that allows communication of computational needs
Rational 2c:	HPC contains a list of pre-compiled programs, but new and improved computational tools need to be effectively identified and added in manor that avoids redundant requests
Solution 2c:	Two phase process 1. Discuss topic within scientific teams, 2. Agreed upon best practices communicated by team leaders to the HPC via monthly teleconferences
Gap 2d:	Need better ways to communicate achievements to customers and stakeholder
Rational 2d:	Feedback from end-users is needed in order to direct future research direction(s)
Solution 2d:	Direct communication with stakeholder though newsletters, web links and emails highlighting accomplishments in ARS genomics research, in language that emphasizes impact.
Gap 2e:	No ARS-condoned methods exist to publicly distribute best practices & methods
Rational 2e:	Interaction with the greater scientific community is needed in order to receive input to improvements for data analysis methods within ARS, and distribute ARS-based protocols
Solution 2e:	The National Ag Library (NAL) Ag Data Commons can be used as a "preferred" site for depositing supplementary data such as methods when accompanied by a dataset. Message boards and blogs are needed in order to receive real-time feedback; Hold workshops.

Defining interaction within and among scientific communities

Interaction among scientists: Facilitating research & identifying solutions to knowledge gaps

During the AGR Workshop, four collaborative scientific teams were formed based on application areas; 1. *Assembly and Annotation*, 2. *Genotyping*, 3. *Gene Expression and Proteomics*, and 4. *Gene Silencing and Editing*. These scientific teams are designed to form the foundation of the ARS research communities by providing a long-term and sustainable mechanism for group interactions and collaboration, and also will serve as an advisory group to continually devise solutions to novel gaps that arise as scientific methods and technologies change. Adapting to changing scientific practices will require implementing or revising best practices, training and documentation, and data standards to support the transition of ARS scientists towards use of the most effective genomic technologies. Thus gap analyses analogous to that performed at the workshop will likely be integrated into normal functioning of scientific teams, where gaps in available scripts or computational applications are identified and exchanged with other scientific teams experiencing analogous impasses. Scientific teams are the point of contact and will facilitate collaboration among scientists with common research goals and will serve as the vehicle by which peer-to-peer knowledge is exchanged. Each scientific team will have a lead or group of leaders that will coordinate the communication strategy via scheduled teleconferences, or devising an electronic system to submit questions and communicate research impasses within the team. Each scientific team will not act in isolation, but instead function as part of an integrated community that works collectively to advance genomic research within ARS. Thus, engagement across scientific core teams will aim to identify instances when common impasses exist within more than one scientific team and work collaboratively to devise solutions, where facilitation and coordination of these needs is to be fulfilled by the virtual research support core (VRSC; see below).

Mechanisms for virtual support of applications and user resources

The virtual research support core (VRSC) is currently defined as one or more external contracted entities that provide support for the operational infrastructure and maintenance of bioinformatics applications on the HPC, as well as teams of ARS scientists that generate data management frameworks and knowledge resources that facilitate genomic data distribution and analyses. Challenges arise due to the varying degrees of computational background and available resources among ARS scientists, and empowering the workforce will necessitate devising a system that defines best practices for data management and performing analytical processes, as well as mechanisms to distribute these outputs in the form of user training, protocols, support documentation, and other resources. The VRSC is currently composed of outside contractors that support the HPC hardware and applications (group 1. *Tools and Workflows*,) as well as two additional working groups populated by ARS scientists (2. *Data Management*, and 3. *Training and Documentation*). Groups within the VRSC acquire and elicit input from the scientific teams defined earlier in order to maintain function of current applications and incorporate new applications as needed, as well as communicate the needs of the user community with the Office of National Programs when modification of ARS administrative policy is justified (**Fig. 1**). The VRSC also holds a pivotal position within the ARS Big Data initiative, with each component operating in the following capacity:

1. *Tools and Workflows*: ARS scientists in this group will act as the initial interface with the VRSC. The role of this group is to provide the science drivers and early adopters of SCINet. Earlier adopters of the platform will be prioritized for access to "Ceres". Responsibilities of this group will be to provide feedback on the initial training and documentation for remote operation of the HPC for all ARS scientists requesting access, and initial beta testers of the ScienceGateway ("onboarding"; **Outcome 2**). A second target of this group will be to provide "prototype" projects. The projects will serve as exemplars for prioritized applications, standard data sets, developing training and standards (group 2), and benchmarking HPC applications (**Outcome 3**). We anticipate 2-3 exemplar projects over the course of a year. The outcomes of this group will serve as the foundation for objectives and milestones for VRSC groups 2 and 3 below.

2. *Data Management*: ARS scientists in this group will focus on best practices for data management from the beginning to end of the experiment. The plan will encompass recommendations of best practices for experimental design, analyses, and long-term archiving. One of the key activities of this group will be to make recommendations for standards needed to describe and exchange the data, including ontologies, and file formats. This group will interact with scientific teams to define needs, reviewing those which are currently established as well specialized requirements (for example, ontologies that are general compared to those specific to one species group). The group will initially focus on the exemplar projects that have been defined by the Group 1 (**Outcome 3**). Data management is a crucial but often overlooked component of the Big Data infrastructure, and supports the long-term applicability of data investments made by ARS to support a unified set of descriptive terms used for data submissions. See **Outcome 4**.

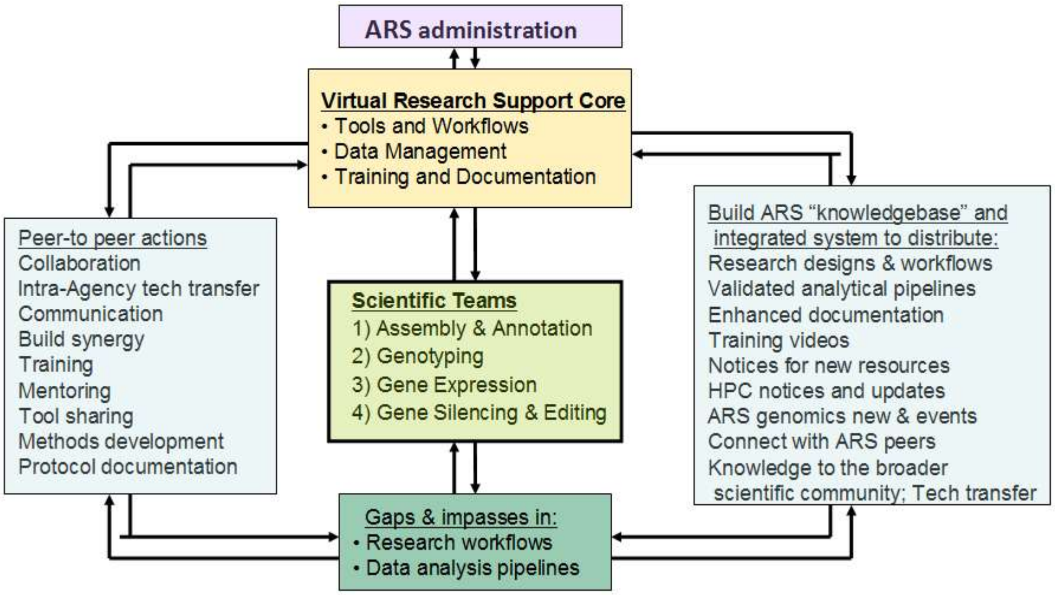
3. *Training and Documentation*: A second group of ARS scientists will coordinate and prioritize the development of training resources requested by scientific teams. Training resources may encompass video modules, suggested workflows and pipelines, workflows, and written documentation prepared in cooperation with experts within the scientific teams and the VRSC. The training materials will focus on basic software and data literacy skills, exemplar workflows based on the output from group 1, prototype targets, and recommended best practices as defined by the community (group 2.) The Training group will also work with VRSC to provide direct access to training documents within the HPC environments, and also with NAL to distribute resources to the greater scientific community through assignment of digital object identifiers (DOIs) that can be cited in research publications and reported in the accomplishments of individual ARS scientists. See **Outcome 5**.

Coordinated efforts within the VRSC and interaction with scientific teams will lead to the community building, focused on communication, training, and access to data storage and high performance computes resource, and expansion of the Big Data resources with ARS. The primary focus of this group will be to support standards development and implementation, integration of tools to improve the "Ceres" HPC resource, and the development of best practices to support genomic data analysis and data management by the greater ARS community. Communication between the three components of the VRSC will be needed in order to coordinate efforts and provide a complete set of resources needed by scientists in order to utilize new applications or tools. For example, the *Training and Documentation* group may be called upon to provide written documentation for new best practices for a new application installed by *Tools and Workflows*.. Thus, regular meetings between members of the VRSC will be conducted in order to define paths for effectively and efficiently addressing using needs.

Transmitting user needs and feedback within a virtual ARS community

Interactions between scientific core teams and the VRSC are important for effective analysis of genomic data using the HPC. Coordinated efforts within the VRSC and within scientific teams will lead to the continual integration of new tools and the development of resources that support genomic data analysis and data management by the greater ARS community. The VRSC serves as the hub at which scientific teams interact with one another, transmit needs to appropriate components of the VRSC, and also work with the VRSC to provide technical expertise which can be captured and distributed ARS-wide (**Fig. 1**). For instance, work among members of each scientific team may address needs of that group, but when activities are integrated into the VRSC input from across ARS can be leveraged in order to solve larger common solutions. These interactions will require seamless communication between all participants, and transparency of actions undertaken in the VRSC, among scientific teams, and within corresponding administrative oversight. In addition to regularly scheduled teleconferences wherein plans and timelines are developed in response to requested user needs, the VRSC will work with community members, to provide a path for communication of system updates, releases of new data standards, and training resources in order to facilitate immediate use by ARS scientists.

Figure 1: Interaction, communication and outputs among the cooperative action between scientific teams, the Virtual Research Support Core (VRSC) and ARS administration.



References

USDA-ARS. 2013. Big Data and Computing: Building a vision for ARS information management. Workshop Summary. Beltsville, MD, Feb 5-6.

Arthropod Genomics Research Workshop Outcomes

Outcome 1: Collaborative Scientific Teams.

Collaborative research teams built during the workshop performed gap analyses and provided guidance with respect to proposed solutions, but also will for the core units upon which knowledge, skills, abilities, and expertise will be distributed through constant interaction and exchange of ideas. These scientific teams form the foundation to the VRSC by identifying applications, documentation or training needs by one or more group of scientists, and provide guidance and direct input regarding the construction of subsequent solutions.

1. Assembly & annotation

Scott Geib
Pia Olafson
Monica Poelchau
William Wintermantel
Anna Childers
Michael Sparks
Paul Shirk

2. Genotyping

Brad Coates
Erin Scully
Keith Hopper
Kai-Shu Ling
Brian Rector
Michael Simone-Finstrom
Lucy Stewart
Matthew Lewis
Sonja Scheffer
Lindsey Perkin

3. Gene expression

Brenda Oppert
John Ramsey
Steven Cook
Noble Egekwu
Dunlap Chris
George Yocum
Susan Lawrence
Chris Childers

Gene silencing & editing

Dawn Gundersen-Rindal
Man-Yeon Choi
Wayne Hunter
Meg Allen
Steve Garczynski
Bob Shatters
Robert Vander Meer

Team leader(s) in bold

Outcome 2: Users were identified for receipt of Ceres user accounts and training.

1) Michael Simone-Finstrom
2) Lindsey Perkin
3) Kenlee Friesen
4) Keith Hopper
5) Anna Childers
6) Monica Poelchau
7) Pia Olafson
8) Kai-Shu Ling
9) Steve Cook
10) Erin Scully
11) Alison Gerken
12) Brenda Oppert

Outcome 3: Defined Virtual Research Support Core (VRSC) prototype projects

A dataset was identified at the workshop generated by Brad Coates, Ph.D. (Research Geneticist, USDA-ARS, Corn Insects & Crop Genetics Research Unit, Ames, IA 50011) which will serve as an exemplar dataset for benchmarking associated applications on the HPC as well as developing a training set with protocol documentation for analogous data analysis by other ARS scientists.

Title: Differential gene expression in the Western bean cutworm, *Striacosta albicosta*, midgut feeding on corn, *Zea mays*, compared to bean, *Phaseolus vulgaris*.

Project description: The goal of this project is to investigate host plant interactions of a specialized lepidopteran herbivorous insect pest that has made a host plant shift from dry land beans to corn in regions of the Western Corn Belt. Adaptation of feeding behavior, gut metabolism, and catabolic pathways are poorly understood, but are hypothesized to involve changes in expression or function of digestive and detoxification enzymes. A transcriptome-based approach was used to predict any significant changes in midgut gene expression when larvae fed upon different host plants. The species is also of interest due to evolved high levels of resistance to transgenic maize that expresses the *Bacillus thuringiensis* (Bt) Cry1F toxin.

Data to be analyzed:
Paired end reads collected for assembly of a reference transcriptome consisting of ~400 bp and ~500 bp inserts respectively for MiSeq and HiSeq data. Reads were overlapping among MiSeq reads. Single end 50 bp reads collected from two different treatments; bean (*n* = 6 biological replicates) and corn (*n* = 6) remain as raw data that have not been processed.

Goals for data analysis:

1) Trim read data to remove residual adapter sequence and nucleotide with Phred quality scores (q) < 20.
2) Assemble a reference transcriptome using filtered paired end MiSeq and Hiseq read data.
3) Annotate the reference transcripts.
4) Map single end read data to the reference transcriptome.
5) Predict transcripts with significant differences in normalize read counts, and generate graphical outputs for publication.
6) Perform pathway analyses.

Outcome 4: Roadmap for data management

The metadata standard of *Minimum information about any sequence (MIxS)* has been established by the Genomic Standards Consortium (GSC), and is the most and broadly recognized and adopted. Analogous adoption by the USDA BigData Initiative will result in ARS datasets being closely aligned with standards already in place by the major databases of the International Nucleotide Sequence Database Collaboration (INSDC). Additional database resources are available for disseminating project metadata, including the National Center for Biotechnology Information genome project pages and the Genomes Online Database (GOLD). This will lead to the assignment of common contextual information to ARS-derived genomics information such that will facilitate transfer to INSDC repositories, as well as offer controlled vocabularies that facilitate ongoing use, retrieval, and comparison of datasets. Maximizing utility of datasets by application of standards also optimizes the return on research investments and ensures stable archiving of data, and expands downstream impacts by increasing data accessibility to Stakeholders. Within one year, workshop participants agreed to develop a proposed framework for data management within the AGR that is MIxS-compliant. The roadmap for developing this data management plan will include *1. Develop controlled vocabularies and ontologies for arthropods*, and *2. Collaborate with the GSC and INSDC members to devise a MIxS extension for arthropod species*. These actions will be done in cooperation with the Metadata and Ontologies Working Group (MOWG).