Project Plan

Research Management Unit
National Agricultural Library

Location
Beltsville, MD
Fort Collins, CO

Project Title
The i5k Workspace@NAL: An information portal for arthropod genomes and genes

Investigator(s)
Monica F. Poelchau (Geneticist) 1.00
Christopher P. Childers (Geneticist) 1.00

Scientific Staff Years
2.00

Planned Duration
60 months

Table of Contents

PROJECT SUMMARY

Arthropods, an incredibly species-rich phylum, cause negative impacts on virtually every plant and animal, but also provide substantial services to US agriculture, impacting US agricultural by billions of US dollars annually[1]. Genomes and genomic techniques are key resources in agricultural entomology, and have the power to advance arthropod pest management and pollinator health through novel methods[2]. Genome databases serve as the authoritative resource for genomic information in their community, to enable the transformation of data into scientific discovery and knowledge. The i5k Workspace@NAL strives to become the access point for up-to-date, complete gene and genome information for orphaned arthropods, ultimately improving scientific outcomes for arthropod research. The i5k Workspace@NAL serves two main roles – 1) as a data access point to find and retrieve arthropod genomic data, and 2) as a data curation portal to manually improve existing genome annotations. Although the i5k Workspace is only 5 years old, it has been cited 44 times; referenced 127 times in google scholar; serves hundreds of registered users each year, with thousands of unique users and pageviews; and has expanded to genomes from 70 organisms – suggesting that it is becoming an established resource within the arthropod genomics community (see Appendix 2 for specific highlights). In the next 5 years, i5k Workspace@NAL will 1) improve the quality and depth of data and metadata; 2) facilitate and improve community annotation; 3) analyze and improve platform architecture; 4) provide training and advice on manual annotation and data management; and 5) continue collaborations on joint software and standards development and implementation. These goals should increase i5k Workspace data and metadata quality, while improving software development and maintenance practices, resulting in an improved user experience and increased adoption of this platform.

OBJECTIVES

The i5k Workspace aims to become a key access point for arthropod genomic data, providing up-to-date, user-submitted content for any arthropod genome project that needs our services. The two pillars of the i5k Workspace's content are 1) user-submitted genomic data and metadata, and 2) community-curated genome annotations (Figure 1). Our users serve two functions: as data producers, who submit and curate data; and data consumers, who access data from our website. The foundation of the i5k Workspace is the software platform that provides access to content and services to our users, so they can use this information to advance scientific discovery in agriculture and other domains. We have identified several challenges to these services that motivate our development in the next 5 years. We anticipate a large increase in genome submissions over the next five years due to the Ag100Pest project, which expects to generate high-quality genome assemblies and annotations of 100 agricultural pest species. We have identified challenges in the community curation services that we facilitate. Data resulting from these two sources need to cater to the FAIR data principles[4], to make the data more Findable, Interoperable, Reusable, and Accessible. Therefore, the overarching goals of the i5k Workspace's next five years will be developments towards improving our existing workflows and systems so we can handle ingesting and updating content, and improving the quality and consistency of our community-curated content. These challenges motivated the following objectives:

- **Objective 1.** Acquisition and stewardship of arthropod genomes and genome annotations
- **Objective 2.** Services to improve gene annotation quality
- **Objective 3.** Systems development and maintenance: analyze and improve our platform architecture
- **Objective 4.** Outreach and communications with stakeholders to improve awareness of the i5k Workspace
- **Objective 5.** Collaboration with other database partners on joint standards development, software development, and their implementation

Objectives 1 and 2 concern the i5k Workspace content – proper ingest and stewardship of user-submitted data and metadata, and curation and other services performed or facilitated by the i5k Workspace to increase the value of the user-submitted content for the entire community. The foundation of the i5k Workspace the software used to receive and transmit data; Objective 3 therefore addresses the foundation of the i5k Workspace, namely the development and maintenance of the software applications that allow users to access and curate the content addressed in Objectives 1 and 2. As a user-focused platform, communication with stakeholders is key to our success – stakeholders define our requirements and provide our data, and therefore regular and effective communications are imperative for the i5k Workspace's longevity. Finally, as a small project with only two ARS FTEs, the i5k Workspace has to work with other databases on shared goals in order to sustain ourselves longer-term, as all databases can provide better services if we work together on common problems. Figure 1 outlines the Project objectives and their relationships among each other.

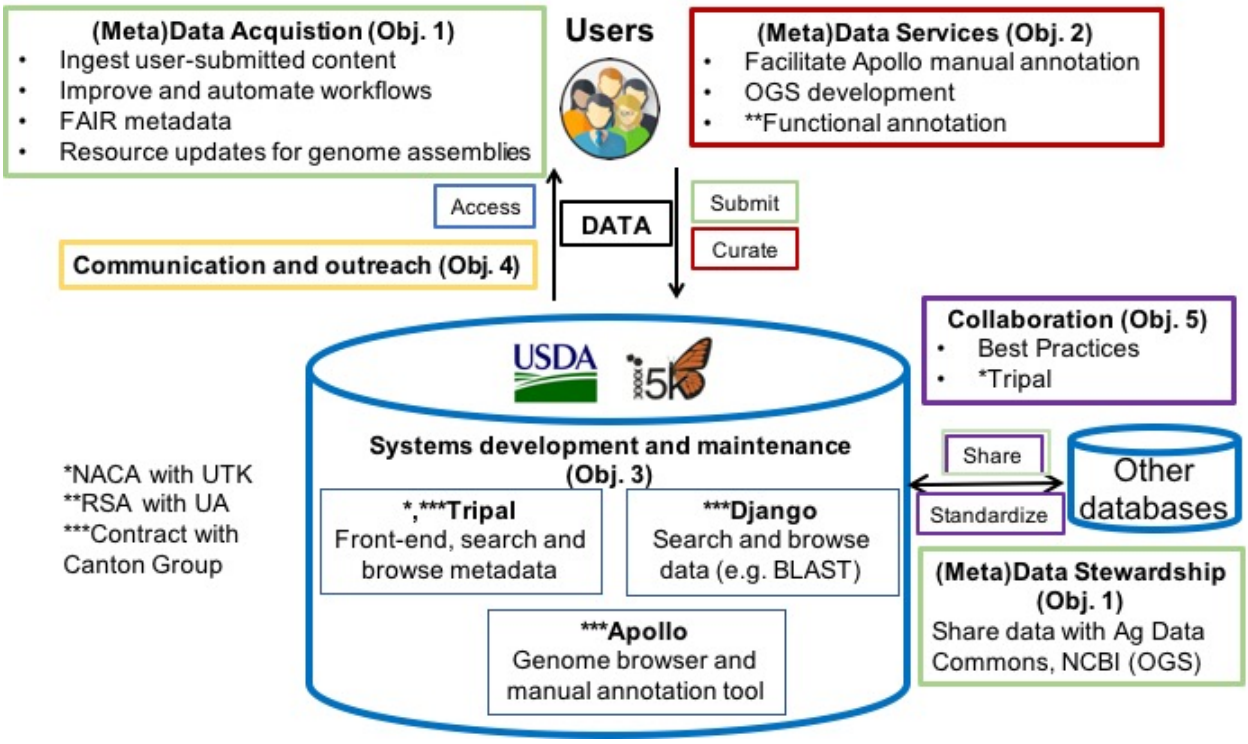

Figure 1. Diagram of i5k Workspace@NAL objectives, stakeholders, services and functions.

NEED FOR RESEARCH

Arthropods are an incredibly species-rich phylum, leading to significant interactions with humans and human activities[5]. As a result, arthropods, including insects, ticks, and mites, cause negative impacts on virtually every plant and animal, resulting in losses of billions of US dollars annually [1]. Beneficial insects – for example, pollinators or biocontrol agents – also provide substantial services to US agriculture[1]. Genomes and genomic techniques are key resources in agricultural entomology[2], and proper access to these 'big data' is essential for productive research outcomes. Genome databases have traditionally served the role of key access points to genomic data for scientific research[6], and strive to become the central authoritative resource for genomic information in their community, to enable the transformation of data into scientific discovery and knowledge. The i5k Workspace@NAL was founded in 2013 to serve as an access and curation portal for 'orphaned' insect genomes not fostered by existing genome databases[3]. Prior to the i5k Workspace@NAL, only clade- or topic-specific genome databases existed for some arthropod species (e.g. FlyBase[7], VectorBase[8], Hymenoptera Genome Database[9], AphidBase[10]), and genome database support for arthropods important for ARS research was not available, other than regular database services provided by GenBank[11], RefSeq, and in some cases Ensembl Metazoa. Without the i5k Workspace, the genomes of these species – which are often agriculturally relevant – will not have 1) database and data management services that will allow them to establish reference gene sets for the community; 2) software and tools to improve gene annotations; and 3) a central location to easily find gene and genome sequences for orphaned arthropods. Appendix 2 highlights several cases where the i5k Workspace directly facilitated research outcomes with an agricultural impact.

The i5k Workspace provides up-to-date, relevant, and curated information on arthropod genes and genomes, to accelerate basic and applied research on arthropod biology. The i5k Workspace@NAL serves two main roles for arthropod researchers – 1) as a data access point (to find and retrieve insect genomic data), and 2) as a data curation portal (to manually improve existing genome annotations). Currently, all of our content is submitted by users. Genome annotation curation is the manual improvement of structural and/or functional annotations predicted from genomes, using external evidence from experimental datasets or published literature (Figure 2). Traditional genome databases, focused primarily on one or several model organisms, perform data and metadata curation in-house. Because of the large number of nascent, non-model genomes housed at the i5k Workspace@NAL, a different approach is necessary – facilitation of curation performed by the scientific community. We contribute to open-source software projects such as Tripal[12] and Apollo[13], and also develop in-house bioinformatics software and tools that are shared with (and used by) the community (https://github.com/NAL-i5K/). To support community use of the i5k Workspace, we offer training and webinars, and workshops at the annual Arthropod Genomics symposium (https://i5k.nal.usda.gov/talks-and-presentations). We also work within and beyond the arthropod genomics community to promote standards in genome assembly and annotation management[6]. Even though the i5k Workspace is only 5 years old, as of November 2018, it has been cited 47 times; referenced 127 times in google scholar (https://scholar.google.com/scholar?hl=en&as_sdt=0%2C21&q=i5k.nal.usda.gov&btnG=&oq=i5k); and has continuously grown its content, currently hosting genomes from 70 organisms – suggesting that it is becoming an established resource within the arthropod genomics community (see metrics reports: https://i5k.nal.usda.gov/i5k-workspacenal-reporting-metrics).

I5k Workspace primary stakeholders are the i5k consortium[14], a broad, international group of researchers working on all aspects of arthropod genomics. These stakeholders use arthropod genomes to perform basic and applied studies across the spectrum of biological research. The ARS Arthropod Genomics Research group is a key subset of these stakeholders, as these are the subset of the i5k consortium that are ARS scientists. Other key stakeholders include other database providers, including those of the AgBioData consortium (https://www.agbiodata.org/databases), VectorBase[8], AphidBase[10], and Hymenoptera Genome Database[9]. These database stakeholders work with us to share data and collaborate on standards development and implementation. Input from stakeholders is gathered via email,

telephone conversations, in-person feedback at conferences and workshops, and surveys. An initial advisory committee was formed and maintained for the first year of the i5k Workspace, but not thereafter.

The i5k Workspace@NAL strives to become the central, authoritative resource on high-quality gene and genome information for 'orphaned' arthropods. Our goals are to efficiently facilitate expert manual gene annotation (for arthropod genomes that require this service); to provide meaningful access to gene and genome information to both humans and machines (cf. the FAIR data principles[4]); and to communicate with scientists and other database providers on developing and implementing best practices in genome and gene data management. These goals are challenging, given the growing amount of content that the i5k Workspace stewards. Therefore, we will 1) continuously improve the quality and depth of our data and metadata; 2) continue to facilitate and improve community annotation; 3) analyze and improve our platform architecture; 4) provide training and advice on manual annotation and genome data management to our stakeholders; and 5) continue collaborations with other databases on joint software and standards development and implementation. These goals should increase the quality of the i5k Workspace data and metadata, while improving our software development and maintenance practices, resulting in an improved user experience and increased adoption of our platform.

The work described in this project plan will support research conducted in several ARS National Programs, including NP104, NP304, and NP305, as well as USDA Strategic Goal 1, Objective 1.4: Improve Stewardship of Resources and Utilize Data-Driven Analyses To Maximize the Return on Investment (USDA Strategic Plan 2018-2022). Research on arthropods and arthropod genomics has traditionally occurred across NPs, and has focused on beneficial or negative effects of arthropods on agricultural crops, livestock, humans and structures[1]. Research components utilizing arthropod genomes include:

- All three components of NP104's action plan (Veterinary Entomology; Medical Entomology; Fire Ants and other Invasive Ants; https://www.ars.usda.gov/ARSUserFiles/np104/NP%20104%20Action%20Plan%202019%20-2024%20Final%2011Sept2018.pdf),
- Components 3 (insects and mites) and 4 (Protection of Post-Harvest Commodities, Quarantine, and Methyl Bromide Alternatives) of NP304's action plan (https://www.ars.usda.gov/ARSUserFiles/np304/NP304ActionPlan2015-2020/NP%20304%20Action%20Plan%202015-2020_FINAL%20rev.%2007.17.14.pdf)
- Component 2 of NP305's 2018 action plan (https://www.ars.usda.gov/ARSUserFiles/np305/NP305%20Action%20Plan%202018-2023%20FINAL.pdf)

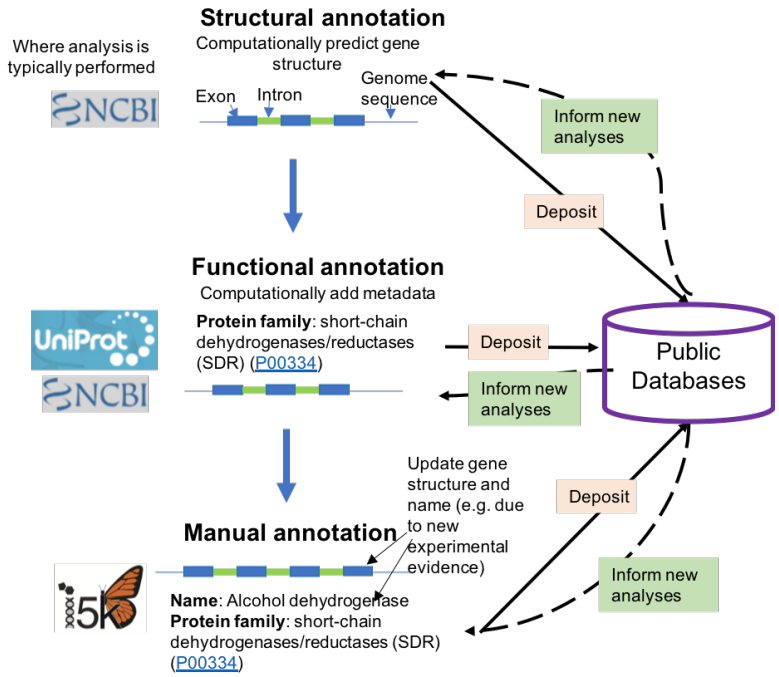

Figure 2. Figure of genome annotation types; where and in what order they are performed; and example results. Public databases (e.g. NCBI, UniProt, i5k Workspace) play an important role in the generation of annotations, as they provide access to previous results from other species, which will inform new annotations on new genome assemblies.

SCIENTIFIC BACKGROUND

Objective 1. Acquisition and stewardship of arthropod genomes and genome annotations.

(Meta)data acquisition

The decreasing costs of next-generation sequencing have made genome sequencing of non-model organisms an attainable reality for many scientists. As a result, many new genome assemblies of non-model arthropod species have recently been generated - 366 arthropod species had at least one genome assembly available in GenBank as of October 2018 (Anna Childers, pers. comm.). NCBI via GenBank provides essential data preservation services for genome sequence data, and should be considered the primary archive of most sequence-based data types associated with genome assemblies. However, submission of gene annotations of these genome assemblies to NCBI has lagged, making it the responsibility of taxon-specific genome databases to 1) serve as the main access point for users to genome annotation data, and 2) perform submission of these datasets to NCBI as a service to the scientific community. This means that genome databases will likely always have some content and functionality overlap with NCBI. Data acquisition for the genome database should be driven by the scientific community's need for the unique services that the genome database provides, which is why i5k Workspace content submission is currently entirely user-driven.

The Earth BioGenome project[15] is an international initiative that aims to sequence the genomes of all eukaryotes over a 10-year period. The "100 Ag Pest Project (Ag100Pest)" was proposed by ARS to contribute to the Earth BioGenome project, by generating high-quality genome assemblies, structural

annotations, and functional annotations of 100 pest genomes. Pending funding for this initiative, the i5k Workspace@NAL will face increased content submission from ARS stakeholders within 1-3 years after the funding is initiated. As an ARS-funded genome database, the i5k Workspace is attuned towards the needs of genome sequence data generated by ARS, and is willing to work with these data providers to ensure that these genomes are hosted in a timely fashion.

The FAIR data principles were recently developed in order to provide best practices for scientific (meta)data quality[4]. These principles aim to make (meta)data more Findable, Accessible, Interoperable, and Reusable by both humans and machines. Many scientific databases and repositories are now aiming to adopt these principles, which may require modifications to both content and software. The broad implication is that scientific (meta)data should and will become accessible between databases – one database may be the source of primary data generated by their stakeholder community, but this data and metadata should be sharable between databases.

Metadata standards are also being promoted by other groups. The AgBioData consortium (https://www.agbiodata.org/) is working towards best practices for genomic, genetic and breeding (GGB) databases[6]. The i5k Workspace is a member database, and MFP is on the steering committee. AgBioData actively endorses the FAIR data principles, and is developing metadata best practices for GGB databases that the i5k Workspace will aim to adhere to once they become available.

The i5k Workspace@NAL collects metadata directly from submitters, as well as from metadata records generated by NCBI (Figure 1). Current metadata ingest procedures are inefficient, as the three software applications that the i5k Workspace uses currently have separate metadata requirements. In addition, the i5k Workspace needs to work towards changing its content and metadata ingest procedures in order to adhere to FAIR data principles, and in the future towards AgBioData best practices. If the i5k Workspace is to efficiently ingest content from 100 new arthropod genomes, in addition to its regular rate of increase, then internal efforts need to focus on automating and streamlining metadata ingest within the next 2-3 years. Prioritization of content increases would require additional personnel (data/metadata wrangler).

(Meta)Data stewardship

Genome projects are not usually static, and some i5k genomes are being updated with new data. The i5k Workspace strives to keep content up-to-date. Integrating genome assembly updates into the i5k Workspace requires a complete re-build of the content for that genome project, because the underlying data and metadata have changed. A larger problem is the re-mapping of gene models predicted from an old genome assembly to a new genome assembly – coordinates of the gene models need to be recalculated, and depending on the changes between assemblies, many gene models may change. Some programs exist that can perform the coordinate updates, but these programs fare poorly with the format that i5k Workspace models are stored in (GFF3; https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md). We have generated two sets of programs that can perform the coordinate remapping in most use cases (https://github.com/NAL-i5K/coordinates_conversion, https://github.com/NAL-i5K/remap-gff3).

The i5k Workspace does not preserve or archive data (https://i5k.nal.usda.gov/data-management-policy), as these functions are readily available from other primary repositories. GenBank is the most suitable repository for sequence data. We encourage our users to submit all sequence data to GenBank, and only host genome assemblies that are also accessioned by GenBank. However, certain data types that the i5k Workspace hosts are unsuitable for NCBI. The National Agricultural Library's Ag Data Commons (https://data.nal.usda.gov/) can also perform data preservation (currently for up to 10 years) for most data products resulting from USDA-funded research. As of September 2018, there are 45 i5k datasets at the Ag Data Commons (https://data.nal.usda.gov/i5k), and the i5k Workspace facilitated the submission of most of these. However, (meta)data submission is 100% manual on the part of the i5k Workspace. In order for this process to be sustainable, the i5k Workspace and Ag Data Commons need to collaborate on a harvesting procedure for metadata and data files.

Objective 2. Services to improve gene annotation quality

Generating and maintaining curated, up-to-date content is one of the main objectives of genome databases. Users of model-organism genome databases have come to expect accurate gene annotations with rich functional information, as this information is key for basic and applied scientific research. For non-model organisms, such as most arthropods relevant to agriculture, automated procedures provide a first step in generating both structural and functional annotations (Figure 2). These automated annotations are not always accurate[16], and additional manual curation of gene structure and function by subject experts, and/or tuning of computational annotation pipelines towards individual organisms, can vastly improve results, and thus improve outcomes from scientific research downstream.

Most model-organism genome databases have dedicated, professional curators to manually curate gene functional annotations based on information from the scientific literature. Known as 'literature curation', these dedicated curators identify genes and their function from scientific experiments documented in the scientific literature, and enter this information into a database using the appropriate ontologies and controlled vocabularies. This allows other scientists to rapidly and easily learn known functions of a gene, simply by looking up the gene page at the appropriate database – rather than culling through hundreds of papers themselves. Experimental evidence of gene function in non-model organisms historically has been almost impossible to acquire. Recent technological advances in molecular genetics, such as RNAi or CRISPR/Cas9, now allow scientists, including ARS entomologists, to identify gene function in organisms without 'traditional' gene knockout systems (e.g. [17,18]). As a result, gene function in agriculturally relevant arthropod species has the potential to be curated, allowing for enhanced, rapid knowledge discovery in the realm of arthropod biology. While typically trained professionals are necessary for proper execution of this work, PomBase[19] (https://www.pombase.org/) is a precedent for effective community-based literature curation. The i5k Workspace@NAL does not have funds for this type of work, but can lay the groundwork for functional literature curation by: implementing and enforcing naming standards and global identifiers for arthropod genes and proteins (which are critical for associating literature with the correct gene name); and working with collaborators to identify or create appropriate ontologies for insect gene function curation.

In contrast to literature curation, typical manual gene annotation of non-model organisms edits both structure and function of the gene model[20], and is typically performed voluntarily by domain experts, from undergraduates to seasoned scientists. Evidence from other sources, such as RNA-Seq data, is used to corroborate or change gene structural annotation. Functional information from orthologous genes, if direct experimental evidence is not available, can be used to guide name choice or other functional annotations (cf. https://i5k.nal.usda.gov/i5k-workspace-gene-and-protein-naming-guidelines). These structural edits can be critical to improving the outcome of wet-lab experiments using these gene sequences[16,21]. Additionally, if these improvements are made accessible via genome databases and submitted back into primary repositories, they can be used as higher-quality reference (or training) data for automated annotations of other, newly assembled genomes. As such, manual annotation – if performed properly – can have benefits that cascade beyond the genome they were generated from. The i5k Workspace@NAL facilitates manual annotation of i5k Workspace genomes by our stakeholders via the Apollo software[13]. We have published programs for QC/QA of the resulting annotations, and for merging these manual annotations into 'Official Gene Sets' (https://github.com/NAL-i5K/GFF3toolkit); and we are committed to submitting these annotations to NCBI (Figure 2).

Proper manual annotation also requires training. Training so far, in particular regarding naming and re-use of data, has not been developed with an end goal of subsequent literature curation in mind, and could benefit from additional rigor. Additionally, multiple academic stakeholders have expressed interest in training undergraduates in manual curation for their genome assemblies. Better training and documentation tailored towards undergraduate curators could be developed in collaboration with

academic stakeholders and other databases, or the i5k Workspace could become involved in existing training efforts.

Computational procedures for functional annotation can also be refined to complement manual annotation efforts by the stakeholder community. While automated pipelines that perform functional annotation exist (e.g. Blast2GO[22]), these are general pipelines that usually do not include sufficient information relevant to non-model species. Similarly, proteins archived by NCBI's GenBank database (https://www.ncbi.nlm.nih.gov/genbank/) should automatically be annotated with Gene Ontology[23] (GO) terms by the UniProt protein database (https://www.uniprot.org/), but annotation can be delayed, or not occur at all, if 1) protein submission to GenBank is delayed; or 2) the annotations are generated by RefSeq (https://www.ncbi.nlm.nih.gov/refseq/), which is not automatically imported into GenBank. In order for ARS scientists to perform translational research on pest control and rapid gene function discovery, superior functional annotations are needed in a timely fashion. The NAL/ARS has established a Research Support Agreement with Dr. Fiona McCarthy (University of Arizona) to develop a functional annotation pipeline tuned towards arthropod genomes that will greatly enhance ARS' translational research on pest control. Dr. Anna Childers (ARS) is a key partner in this effort.

The community annotations generated by i5k Workspace annotators are only valuable if they are accessible to the broader research community. The usual mode of distribution of these annotations are as an 'Official Gene Set' (OGS), which we define as the current best representation of all a genome's gene models. Generating an OGS at the i5k Workspace entails merging computationally predicted gene models with the community's manual annotations to generate a non-redundant set of models. The i5k Workspace has developed a set of tools for this process (https://github.com/NAL-i5K/GFF3toolkit). While the GFF3toolkit programs have seen great progress since development began (358 github commits as of September 27, 2018), running the programs is still a bottleneck in the OGS generation workflow: 1) the output of the QC program still requires manual review, which can usually be performed by MFP, but occasionally requires input from the original data submitter; and 2) we still need to solidify what QC is exactly necessary for both ingest into Chado[24] (our database schema for biological data) and NCBI, due to the highly variable nature of the errors that we detect from manual annotations. We would like to get the QC process to the point where users can perform this themselves. We need to explore methods to expedite the QC process.

An additional bottleneck in this workflow is the submission of gene annotations to NCBI. NCBI's GenBank is currently the de facto repository for all gene annotations, and should occur for all i5k Workspace Official Gene Sets. The availability of an annotation in GenBank, while currently challenging, has many benefits – greater visibility and permanence; automated accessioning of gene models; and downstream automated functional annotation of gene models by sister repositories such as UniProt. We will work to adapt our own data processing scripts to better handle submission to NCBI. In addition, we will work on the backlog of OGS submissions that we already have. In the process, we will develop a new set of scripts that will reformat most OGS's into NCBI-compatible format. The resulting product will be released on github, either as part of the GFF3toolkit, or in a new repository.

Objective 3. Systems development and maintenance: analyze and improve our platform architecture

The i5k Workspace offers multiple services and uses a variety of tools to meet user expectations. Where possible, open source software (OSS) is integrated into the i5k Workspace. We use OSS for the i5k Workspace, building on and contributing to the successes of existing projects, and attempt to avoid duplicating effort when possible. Our current system has three major software components: a graphical genome browser and annotation tool[13], a platform for various types of sequence alignments (https://github.com/NAL-i5K/genomics-workspace), and a website for making content and information available to our users[12].

The main website and database is based on Tripal[12], which combines Drupal, a popular content management system, with the Chado schema for biological data storage[24]. We are aligning our project requirements with the priorities of the Tripal development community, as part of a non-assistance cooperative agreement with a Tripal development group. This agreement has already provided valuable feedback back to the Tripal community in the form of new functionality and modules (https://github.com/NAL-i5K/tripal_apollo; https://github.com/NAL-i5K/tripal_eutils).

Users view a graphical representation of the genomic data that we host via JBrowse[25], a graphical genome browser, along with Apollo[13] for genome feature community annotation. The i5k Workspace has had a long relationship with the Apollo development team, and we have dramatically shifted from a highly customized implementation of their code to a mostly standard implementation with only a few minor customizations. We actively engage with the Apollo developer team, and continue to offer feedback on functionality, testing and code contributions. Our project currently uses both Apollo 1.0.4, and Apollo 2, and we are actively migrating projects to the newer version as possible.

Sequence alignment is a key step in many bioinformatics workflows, and we internally developed a platform we named the Genomics Workspace to provide a web-based interface for BLAST, Clustal and HMMer. The Genomics Workspace is an internal project written in Python using the Django framework. This project includes several external libraries, multiple javascript frameworks in addition to the Django framework. This project will need updates to the installed version of several components, and each update will require tests to ensure we maintain full functionality.

Developments in technology, shifting departmental requirements and changing project needs require us to regularly reassess our platforms against other available solutions so we can continue better serving our users. These include testing to ensure 508 compliance and regular security tests to search for software vulnerabilities. Additional requirements, including migration to a cloud-based environment, also affect project choices moving forward.

Ongoing internal development efforts are always needed to ensure that our resources are up to date with regards to security, stability and incorporating new features and technologies as appropriate. Our current policy is to use OSS and customize it as little as possible, instead relying on contributing changes back, working with project owners or adding customizations in less invasive ways, such as through plugins.

One key need is the ability to support a regular update cycle. This is greatly benefitted by identifying issues with the deployment process, defining and prioritizing solutions and generating a roadmap for future work. Some of the issues already identified are:

- Improving testing and continuous testing/integration;
- Incorporating updates from generic repositories into our custom ones;
- Reducing the need for custom forks of repositories;
- Developing and refine best practices and standard operating procedures (SOPs).

We also need to give developers, cooperators, contractors and others access to use and develop our projects without clearing the logistical hurdles needed to access federal servers. To this end, we need to document and expand SOPs to reduce the learning curve for new developers to test and develop our software stack. We need to develop protocols and systems to assist people in accessing and developing our software products. In addition to documentation of how to set up and use tools, we are also working on building out standard development environments using Ansible, Vagrant and Docker.

Objective 4. Outreach and communications with stakeholders to improve awareness of the i5k Workspace

Outreach and communications with stakeholders are critical activities of a genome database. i5k Workspace stakeholders include the i5k consortium[14]; the ARS Arthropod Genomics Research group; and other database providers. The i5k Workspace has recently initiated a working group, comprised of key

stakeholders, that are invited to contribute feedback on i5k Workspace progress and goals in quarterly meetings. We need to establish a productive routine for the working group, and rotate members as needed.

The primary means of communication and outreach for the i5k Workspace are 1) webinars; 2) posters and presentations; and 3) direct communication via email or in-person meetings. The i5k Workspace offered regular webinars on i5k Workspace functions in the last FY, and has contributed to bioinformatics workshops at the annual Arthropod Genomics Symposium. We also regularly present on new developments at international meetings (https://i5k.nal.usda.gov/talks-and-presentations). While additional forms of communication may be desirable (e.g. social media posts, regularly scheduled blog posts), the i5k Workspace currently does not have sufficient personnel to do this.

Objective 5. Collaboration with other database partners on joint standards development, software development, and their implementation.

The new paradigm for genomic databases is collaboration. Genomic data produced by scientists is growing at a rate that is difficult to keep up with. Genome databases need more efficient approaches to accommodate this rate increase. From model organism databases in medicine[26] to bespoke genome databases in agriculture[6], the consensus is that better communication and collaboration on standards, software and data sharing is critical for longer-term database sustainability and improved scientific outcomes for users. The FAIR data principles[4] were established to provide a framework for better stewardship of data to make them findable, accessible, interoperable and reusable by humans and machines – the latter becoming increasingly important as the quantity and complexity of biological data increases. The AgBioData consortium (https://www.agbiodata.org/) was founded in 2015 to provide a venue for communication and collaboration among genomics, genetics and breeding databases in agriculture, and to promote the use of the FAIR data principles by the 30 AgBioData member databases. Many of the member databases are tailored towards smaller communities, and therefore have fewer resources to solve problems common to most databases. Similarly, the i5k Workspace can gain much in efficiency by continuing our participation in AgBioData in order to leverage the work that this group performs. Objectives 1, 2, and 4 of this project plan all refer to AgBioData as part of their approach – this is because collaboration within AgBioData has the potential to facilitate and streamline many of the i5k Workspace functions, in terms of decisions on standards and planning.

The AgBioData group's recommendations for platform development specifically advocate for shared development of open-source platforms. The Tripal software project is a prime example of collaborative software development for genome databases. Tripal is a software package for the construction of online databases, primarily geared towards genomic, genetic and breeding data. It is becoming increasingly adopted by more agricultural databases, which host data necessary for scientists to perform cutting-edge agricultural research. As Tripal's user base expands, Tripal development needs to be attuned towards community trends in scientific data re-use and interoperability – otherwise, it will become much more difficult for scientists to obtain data from these diverse resources. Tripal development is currently working towards supporting FAIR data, but individual improvements, specifically regarding genome-centric databases, are desirable. There are several use cases shared between the i5k Workspace and the Hardwood Genomics Database at the University of Tennessee-Knoxville, that, if appropriately developed and integrated into the Tripal framework as extension or core modules, could help bring the Tripal community towards greater FAIR compliance for genomic data types. Therefore, the i5k Workspace is collaborating with UTK on FAIR Tripal development for genomic data. We have a non-assistance cooperative agreement with the University of Tennessee-Knoxville to make the Tripal genome database software (http://tripal.info/) more FAIR for genome-centric data. Updates to the software are already taking place, and we hope to have updated the entire i5k Workspace to the new version of Tripal (Tripal3) by 2020.

RELATED RESEARCH

A search of the USDA CRIS database for all active USDA in-house projects with the keyword "database" revealed no other genome databases focused on arthropods. There are other projects outside of USDA in-house projects with overlapping interests, however – for arthropods, Hymenoptera Genome Database (Dr. Chris Elsik); Fourmidable (Dr. Yannick Wurm); VectorBase (Drs. Mary Ann McDowell, Daniel Lawson); AphidBase (Dr. Fabrice Legeais); InsectBase (Dr. Fei Li); WaspBase (Dr. Fei Li). In addition, there are several other USDA genome databases that focus on crops, and therefore may be interested in their insect pests: GrainGenes (Dr. Taner Sen), Gramene (Dr. Doreen Ware), MaizeGDB (Dr. Carson Andorf), and SoyBase (Dr. Steven Cannon). Many of these databases also are members of the AgBioData consortium, and collaboration and communication between these databases will be important for sustainability.

APPROACH AND RESEARCH PROCEDURES

Objective 1. Acquisition and stewardship of arthropod genomes and genome annotations.

Goal 1a. Continue user-submitted content ingest.
Approach:

- **Add user-submitted content in a timely fashion.** Users submit content to the i5k Workspace on a regular (but unpredictable) basis. The i5k Workspace currently uses an internal document with SOPs for (meta)data ingest. We will continue to use this document in year 1 and 2 of this project plan, and will adjust our procedure when outcomes from Goal 1b render an improved workflow. User-submitted content should become live at the i5k Workspace within 4 weeks after submission or less, unless there are circumstances beyond our control (e.g. improper formatting of the submitted file).

Contingencies.

- Addition of user-submitted content to the i5k Workspace@NAL depends on users submitting content – in the unlikely event that there are no submissions, then no new content will be added.
- The AgPest100 project may significantly increase our numbers. In this case, we may need to modify our goal of ingesting a new assembly within 4 weeks.

Goal 1b. Improve and automate workflows for data and metadata ingest to the extent possible.

Approach:

- **Evaluate and, if possible, implement best methods for automating metadata and data ingest for new organism/genome onboarding**. The i5k Workspace@NAL (meta)data ingest procedure is a workflow with many steps, and uses many different tools. A new workflow approach is needed to automate this process as much as possible. We will initially evaluate the Common Workflow Language (CWL) as a method for automating (meta)data ingest to the extent possible. CWL is an open standard for workflows grouped in in YAML structured text files[27]. Once a pilot implementation has been developed, we will test the effectiveness of the workflow by timing content ingest manually vs. automated for several test datasets.

Contingencies. After evaluation of CWL, we may need to change the workflow language or platform for this project. There are many other options available, and another language may prove more useful after a full evaluation of our requirements.

Goal 1c. Align analysis-level metadata with FAIR data principles to the extent possible.

Implementation of the FAIR data principles at the i5k Workspace will require some time. Data entities at the i5k Workspace are currently at the sequence (protein or nucleotide) and dataset level (e.g. an entire genome assembly, or gene prediction set) - users will want to retrieve information at each level. Implementing FAIR data principles at the sequence level is desirable because this is the unit that researchers perform their work on; however, (meta)data should accommodate FAIR principles at both the dataset and the feature level.

Approach:

- **Ensure that as much i5k Workspace metadata as possible is derived from an existing standard.** Identify standards to review**;** for example DATS https://biocaddie.org/group/working-group/working-group-3-descriptive-metadata-datasets**.** Identify which metadata elements to use from selected standards.

- **Develop a metadata map across all i5k Workspace applications and primary repository destinations**. Currently, the i5k Workspace, Ag Data Commons, and NCBI collect different sets of metadata for similar data types. These discrepancies make metadata collection and ingest cumbersome for i5k Workspace personnel, and users submitting metadata to us. If needed, NAL metadata librarians will be consulted on how best to perform the mapping.
- **Implement collection of all updated metadata in Tripal.**
    - Modify dataset submission forms to derive as much as possible from one or more existing metadata standards.
    - Identify whether values in submission form can be pre-populated from existing standard**.** If so, figure out how to load the controlled vocabulary into chado, and then show value choices to user via drop-down
    - Consider including data license choice in submission form for later display (could be Ft. Lauderdale vs. publication vs. DOI)
    - Identify how metadata can be stored in the chado database schema, and displayed to user.
    - Ensure that analysis-level metadata is available on gene page display.
    - Ensure that analysis-level metadata is available for all file downloads.

Contingencies. It is possible that the metadata that we will need to collect over the next 5 years will change – either due to new standards, new user requirements, or a change in our primary repository partners. In this case, the metadata map, and metadata collection will need to be updated. Additionally, the metadata collection implementation assumes that the i5k Workspace production site will be running Tripalv3 at the time (rather than Tripal v2 as currently). We anticipate that we will update our production site to Tripal3 by early 2020. Finally, the dataset submission form changes may require developer time, which may not be available.

Goal 1d. Collaborate with the Ag Data Commons on harvesting of i5k Workspace (meta)data to ensure data preservation of appropriate datasets.

Approach:

- **Develop requirements and implementation plan jointly with Ag Data Commons personnel**.
- If personnel to implement the plan are available, **implement pilot phase of harvest.** Once a pilot implementation has been developed, we will test the effectiveness of the workflow by timing content ingest manually vs. automated for several test datasets.
- If the pilot is successful in increasing efficiency, adopt process during regular content ingest, or on a regular schedule.

Contingencies.

- This goal has the following prerequisites: 1) a metadata map between i5k Workspace and AgDataCommons needs to be complete (Goal 1c); 2) all content on the i5k Workspace needs to be updated with this new analysis-level metadata (Goal 1c); 3) Tripal3 needs to be implemented on our production site (Goal 5b), and 4) Tripal3's web services module needs to be enabled and tested (Goal 5b).
- Metadata standards may change, and will need to be re-evaluated. We are assuming that the software applications we will use, and the primary repositories where metadata will be deposited, will stay the same – this also may change (although unlikely).
- It is possible that our efforts to identify the best way to automate 1) do not identify a feasible method, or 2) that efforts for automation do not actually result in higher efficiency/less time spent per ingest (e.g. if a significant amount of manual review cannot be avoided).
- It is possible that this will require software developer time. As we currently don't have a permanent developer, we can't guarantee that the appropriate personnel will be in place to help implement this. In this case, this Goal may stall until personnel is available.

- The Ag Data Commons may change their collection policy in the next five years, meaning that we may not have the budget to submit i5k Workspace data to them. In this case, we will try to identify alternative methods for data preservation and management at the dataset level.

Goal 1e. Update i5k Workspace resources when updated genome assemblies become available.

Approach:

- **Develop SOPs for genome assembly updates**. The i5k Workspace is still identifying what procedures need to be put in place when a genome assembly is updated. The procedures include, but are not limited to: Communication on updates with the research community; re-mapping important gene prediction and RNA-Seq files to the new genome assembly; updating existing i5k Workspace resources with the new assembly; and retiring old resources.
- **Update i5k Workspace resources when updated genome assemblies become available.** Decisions on when and how to update will follow the developed SOPs.

Contingencies. Updating i5k Workspace resources to new genome assemblies is quite labor-intensive- in fact, more so than adding a new genome assembly. It is unclear how well we can perform this depending on demands from our users. Additionally, if genome assembly updates are extensive, gene models may not remap well. We may need to consider retiring/deprecating some i5k Workspace organisms depending on the individual use case, and the SOPs will need to accommodate this.

Objective 2. Services to improve gene annotation quality

Goal 2a. Facilitate community annotation of i5k genomes via documentation and training.

Approach.

- **Develop or implement improved training on community annotation via Apollo, in collaboration with other academic and government stakeholders.** Training modules for manual annotation via Apollo exist in abundance (cf. https://www.slideshare.net/MonicaMunozTorres). The i5k Workspace already holds 1-hour webinars explaining the principles of manual annotation via the Apollo software. However, we need to improve our annotation training strategy in order to improve the quality and regularity of user-submitted annotations. The i5k Workspace will continue to develop two sets of training modules specific to i5k needs: improved 1-hour webinars, and multi-hour workshops. We will identify specific problem genes that represent good working models for beginning, intermediate, and advanced annotators to learn from. With permission from original sources, we will also share templates or SOPs for best reporting of manual annotations 1) within Apollo and 2) in subsequent genome publications. We will iteratively improve the one-hour training webinars, holding one every other month. We will attend training workshops from other collaborators as part of the development process.
- **Work with faculty on developing or facilitating undergraduate training courses.** Many i5k Workspace users have commented that manual annotation via Apollo is an effective way to teach undergraduate students about fundamental principles of molecular biology and genomics. The i5k Workspace would benefit from additional curation via undergraduates. However, undergraduate students benefit from more structured training than the i5k workspace is capable of providing; e.g. weekly group meetings. The i5k Workspace will identify already existing resources for undergraduate training. If these are sufficient, the i5k Workspace will advertise these, creating resource pages on the i5k Workspace website for others to access. If existing training materials are insufficient, the i5k Workspace will identify faculty collaborators to develop and promote better training materials for undergraduate manual annotation training. We will explore the

AgBioData consortium as a facilitator for these training types, as it is possible that there is enough shared expertise (or need) within AgBioData for undergraduate training for Apollo.

- **Implement and enforce naming standards and globally unique identifiers for arthropod genes and proteins.** Naming standards and unique identifiers for genes and proteins are critical for associating literature with the correct gene or protein. The i5k Workspace developed naming standards for the i5k community based on existing recommendations from UniProt and the INSDC (https://i5k.nal.usda.gov/i5k-workspace-gene-and-protein-naming-guidelines). Apollo training will need to emphasize these guidelines. In addition, our QC procedures of manual annotations will need to incorporate a review of gene and protein names.
- **Work with collaborators to identify or create appropriate ontologies for insect gene function curation.** Ontologies are structured controlled vocabularies that represent knowledge from a specific domain. Ontologies are important for automated retrieval and access of qualitative and quantitative information, and allow humans and computers to retrieve and integrate disparate data types. For example, the Sequence Ontology (SO) describes the features and attributes of biological sequence, and their relationships with each other (https://doi.org/10.1186/gb-2005-6-5-r44); the Drosophila gross anatomy ontology describes the Drosophila melanogaster anatomy, and FlyBase curates its data types with this ontology, such that scientists can look up whether any genes, alleles, or other datatypes have been associated with a part of the fly anatomy in the literature. Several arthropod ontologies already exist: (Hymenoptera anatomy ontology: https://www.ebi.ac.uk/ols/ontologies/hao; Spider ontology: https://www.ebi.ac.uk/ols/ontologies/spd; Drosophila melanogaster development ontology: https://www.ebi.ac.uk/ols/ontologies/fbdv; Drosophila gross anatomy ontology: https://www.ebi.ac.uk/ols/ontologies/fbbt; Mosquito gross anatomy ontology: https://www.ebi.ac.uk/ols/ontologies/tgma; Tick gross anatomy ontology: https://www.ebi.ac.uk/ols/ontologies/tads). The Phenotype RCN made initial strides towards developing an arthropod anatomy ontology (http://www.phenotypercn.org/?page_id=48), but it is not clear where the effort stands now. This sub-goal is an exploratory effort to identify whether the i5k Workspace can work with ontology developers and more mature genome databases, possibly facilitated under the umbrella of the AgBioData consortium, to identify or contribute to ontologies for insect anatomical phenotype curation.

Goal 2b. Continue and improve Official Gene Sets (OGS) generation and handling.

Approach:

- **Continue generation of OGS's as needed.** Initially, the GFF3toolkit and custom scripts will continue to be used for this process.
- **Immediate submission of Official Gene Sets to 1) NCBI and 2) the Ag Data Commons**. Previously, we only submitted our Official Gene Sets to the Ag Data Commons for preservation. All new OGS's will now also be submitted to NCBI prior to their release.
- **For newly submitted genome annotation datasets**, run NCBI's table2asn_gff software during the initial submission process. This will help prevent problems with NCBI submission after OGS generation.
- **Identify and address QC bottlenecks in the manual annotation process**. Typically, QA/QC of manual annotations is the most time-consuming aspect of OGS generation. Annotators identify new and creative ways to mark up their annotations within the Apollo software, which in turn reveals unexpected bugs in the Apollo software. We will continue to work with the Apollo software developers on identifying bugs and solutions for them; and will continue to develop documentation on the best ways to use Apollo2.
- **Maintain and update the GFF3toolkit programs.** New OGS's that we generate will often present new use cases that our original program design did not accommodate. This requires infrequent updates to the GFF3toolkit functionality.

- **Develop a new pipeline for OGS cleanup for NCBI submission.** This pipeline will be developed with the i5k Workspace's current submission backlog as test cases.
- **Update Tripal on our production site to accommodate new gene page changes.**

Contingencies. This goal relies on collaboration with multiple partners – NCBI, Apollo, Ag Data Commons, and community annotators. As such, any change in the behavior of these partners may require us to change our approach. The development of a method to automatically run QC programs on Apollo annotations depends on the recruitment of an appropriate individual to do so (NTU student intern or other developer).

Goal 2c. Computationally generate functional annotations of i5k genomes.

The NAL and ARS have set up a Research Support Agreement with the University of Arizona to generate a functional annotation pipeline tuned towards arthropods. The resulting workflows will be made available for ARS and the broader research community, and can be used for functional annotation of the AgPest100 Project, as well as any other arthropod genomes that result from the Earth BioGenome project. Functional annotations (GO and KEGG) will also be made publicly available to the research community and bioinformatics developers that utilize this information in their functional enrichment tools.

Approach.

- **Develop a functional annotation pipeline and documentation tuned towards non-model arthropod genomes.** University of Arizona will develop the pipeline based on datasets that the i5k Workspace will provide. ARS will test the pipeline on local hardware and will provide feedback for improvement. Dr. Anna Childers (ARS) is a collaborator on this effort. Both the pipeline and the resulting functional annotation datasets will be shared with the public on appropriate platforms.

Contingencies. This Goal requires collaboration with other ARS scientists for thorough testing. We are currently recruiting individual testers for the pipeline – however, if appropriate testers cannot be found, the deliverables may be delayed.

Goal 2d. Plan for additional value-added services for i5k Workspace datasets.
I5k Workspace users will benefit from additional datasets made available in the genome browsers. These include, but are not limited to, aligned RNA-Seq tracks to provide evidence for manual annotation. Additional track types mentioned by stakeholders include TE (transposable element) predictions, LGT (lateral gene transfer) predictions, and methylation analyses. These additional analyses provide more in-depth information about the genome than only gene predictions due. Users have requested that the i5k Workspace generate these tracks internally. If the i5k Workspace were to attempt this, it would have to be an automated process. The i5k Workspace needs to establish best practices regarding metadata ingest (Goal 1c), needs to install Tripal3 (Goal 5b) to accommodate the metadata, and needs to establish best practices for workflow development (Goal 1b) prior to considering implementing workflows to automate generating new track types for our organisms.

Approach.

- **Define requirements for adding additional, automated track types to the i5k Workspace** (Apollo2 for the data and metadata, and Tripal3 for the metadata). Prior to defining any steps to implement these tracks, we need to analyze what steps it would take to 1) generate the tracks, 2) add them to Apollo2 and Tripal3, and 3) automate the entire process.
- **If the requirements analysis reveals that automated addition of track types may be possible, develop and test a prototype implementation of one additional track type.** Which track type to choose should become apparent in the requirements analysis.

Contingencies. This goal has completion of Goal 1b, 1c, and 5b as prerequisites. If any of these goals are delayed, this in turn will also delay or prohibit Goal 2d. In addition, successful completion of this goal depends on recruiting a suitable intern from National Taiwan University as an internship project.

Objective 3. Systems development and maintenance: analyze and improve our platform architecture

Approach and research procedures

Goal 3a. Analyze platform architecture
Goal 3b. Codify our project best practices
- Research best practices for software project management (e.g. [28–30])

Goal 3c. Write a system implementation plan
- Handled in conjunction with existing Ops contract

Goal 3d. Develop a migration and deployment plan for Tripal 3
- Use the Tripal non-assistance co-operative agreement and the Ops contract to forward this

Goal 3e. Implement upgrade plans developed in Goal 3a
Goal 3f. Use project requirements to determine if better solutions exist for our current systems
- If so, map out plan for incorporating new features, tools, technologies, as appropriate

Contingencies. These systems are part of a highly dynamic field which changes frequently. Depending outside events, such as the discovery of a software vulnerability, we may need to reprioritize to meet unforeseen critical needs.

Objective 4. Outreach and communications with stakeholders to improve awareness of the i5k Workspace

Goal 4a. Continue and increase outreach activities to i5k Workspace stakeholders.

Approach.
- **Continue collection and evaluation of performance metrics to evaluate i5k Workspace use.** As a fairly new database, the i5k Workspace only recently began collecting a set of metrics to evaluate the utility of our tools and resources for our stakeholder community. The primary goal at present it to continue recording these metrics in order to evaluate what a healthy baseline for the i5k Workspace is. We would expect a steady increase of i5k Workspace Tripal and Blast use as we increase the content of the i5k Workspace. In contrast, use of Apollo, and subsequently the number of user-created annotations collected, will likely depend on the rate of newly submitted content, and possibly our Apollo training activities.
- **Participation in i5k coordinating committee and ARS-AGR meetings; and representation of the i5k Workspace at international meetings.** A major motivation for our participation should be to 1) increase the visibility of the i5k Workspace within these communities; 2) to identify needs that the stakeholder community may have and that the i5k Workspace should fill; and 3) to initiate or continue conversations with the stakeholder community on arthropod genome (meta)data management.
- **Continue presenting regular i5k Workspace webinars**. To focus our efforts, we will initially focus on providing Apollo training during our webinars. We may expand the webinar focus in later years based on feedback from our stakeholders.

- **Expand the i5k Workspace user base with new outreach activities**. Given sufficient NAL funds, we will set up an information booth at the Entomological Society of America meeting in 2019. We may partner with other parts of the NAL, or perhaps other insect genome databases, on this booth. If the booth is well-received, as measured by number of visits or informal feedback, and increased traffic to the i5k Workspace website during and after the ESA meetings, then we may consider repeating this activity in following years.

Contingencies. International meeting attendance, and running a booth at ESA, can only occur if sufficient NAL funds are available for these activities. If funding is not available, we will focus our outreach efforts on the other elements from this goal.

Objective 5. Collaboration with other database partners on joint standards and software development, and their implementation

Goal 5a. Collaborate with the AgBioData consortium to guide i5k Workspace best practices

Approach.

- **Continue to work with the AgBioData consortium, as SC members or participants**. This includes participation within the monthly steering committee meetings; within the monthly webinars; at the PAG annual meeting (if sufficient travel funds are available); and at AgBioData in-person workshops (if funds are available).
- **Co-chair the AgBioData data federation working group, if funded.** The consortium has submitted a grant proposal for additional workshops on best practices development, which should result in specific recommendations and solutions for AgBioData member databases on adoption of FAIR data principles and other issues (MP is a Co-PD on this proposal). MFP will co-chair the working group on data federation, which will identify best methods for data sharing via web services.
- **Invite other insect genome databases, such as VectorBase, AphidBase, and LepBase, to join the AgBioData consortium and collaborate on data federation issues**. Hymenoptera Genome Database is already a consortium member.

Goal 5b. Collaborate with University of Tennessee-Knoxville on development and implementation of FAIR Tripal modules for genome-centric data

Approach.

- Improve the overall Tripal3 deployment/setup documentation and the main Tripal3 codebase.
- Update the i5k Workspace site and data to Tripal3.
- Identify or develop genome-centric extension modules for Tripal.
- Write a peer-reviewed publication on how Tripal and/or Chado should implement FAIR data principles.

Contingencies. Some AgBioData activities may only occur if the recently submitted grant proposal is funded. If it is not funded, we will still continue our participation, but the outcomes may be more restricted.

PHYSICAL AND HUMAN RESOURCES

The i5k Workspace currently has two FTEs: Chris Childers and Monica Poelchau. Additionally, we host two-three interns each year, at the graduate student level, via a non-assistance cooperative agreement with the National Taiwan University. These interns often help us develop and refine our systems, in addition to pursuing projects of interest. Finally, we collaborate with the NAL's Information Systems Division, who maintain our servers and perform security and accessibility checks.

PROJECT MANAGEMENT AND EVALUATION

Chris Childers (CPC) and Monica Poelchau (MFP) will be jointly responsible for project plan implementation. CPC will focus on Objective 3, and MFP will focus on the remaining objectives, and there may be overlap between responsibilities on these objectives. Progress on milestones will be documented in annual reports. The i5k Workspace team (which includes student interns) meet daily in 10-minute 'scrum' style meetings, and weekly for 1-hour discussions. This ensures that remote members of the team are up-to-date on team activities.

We have been evaluating the i5k Workspace's use and utility via google analytics and other metrics (https://i5k.nal.usda.gov/i5k-workspacenal-reporting-metrics). i5k Workspace stakeholders should have access to information on activities and health of the i5k Workspace. The metrics were chosen to 1) reflect the day-to-day use of the tools and resources that we provide (e.g. overall site visits, vs. number of annotations created in Apollo), as well as 2) the longer-term impact of the i5k Workspace's efforts on the scientific community (number of citations of the i5k Workspace in peer-reviewed publications). We will collect these metrics quarterly.

DATA MANAGEMENT

Preservation of data is explained in Objective 1d. A key aspect of this project plan is to make stakeholder data FAIR – therefore, data management is a foundation of the entire project plan. The NAL performs nightly backups of servers, and the Apollo2 application is backed up via a master-slave configuration that should store new manual annotations instantaneously in a separate PostgreSQL database.

MILESTONES TABLE

Project Title	The i5k Workspace@NAL: An information portal for arthropod genomes and genes
Project No.	
National Program (Number: Name)	
Objective	Objective 1. Acquisition and stewardship of arthropod genomes and genome annotations
NP Action Plan Component	
NP Action Plan Problem Statement	
Subobjective	
Goal/Hypothesis	Goal 1a. User-submitted content ingest

SY Team	Months	Milestone	Anticipated Product	Progress/Changes
MP	12	Regular content ingest	New Content available within 4 weeks of submission	
MP	24	Begin AgPest100 content ingest; regular content ingest	New content available within 4 weeks of submission	*This column for plan management after peer review.*
MP	36	AgPest100 and regular content ingest	New content available within 4 weeks of submission	
MP	48	AgPest100 and regular content ingest	New content available within 4 weeks of submission	
MP	60	AgPest100 and regular content ingest	New content available within 4 weeks of submission	

Goal/Hypothesis	Goal 1b. Improve and automate workflows for data and metadata ingest.

SY Team	Months	Milestone	Anticipated Product	Progress/Changes
MP	12	Workflow platforms evaluated	Decision made on Workflow platform; Documentation of evaluation process on github; Requirements of appropriate platform identified	
MP	24	Pilot workflow developed	Pilot workflow tested for genome assemblies; gene predictions; and mapped datasets. RC1 of pilot workflow on github.	*This column for plan management after peer review.*
MP	36	Testing of pilot workflow complete; Begin implementation of production workflow	Evaluation metrics for pilot workflow	
MP	48	Production implementation of workflow; 50% new content ingested using workflow	First version released on github; Content ingested faster	
MP	60	100% new content ingested using workflow; Fine-tuning and maintenance	Content ingested faster	

Goal/Hypothesis	Goal 1c. Align analysis-level metadata with FAIR data principles to the extent possible

SY Team	Months	Milestone	Anticipated Product	Progress/Changes
MP	12	Metadata standards identified	List of standards to comply with, including individual metadata elements	
MP	24	Metadata map developed; 25% of existing analysis records updated with new metadata	Metadata map; Updated metadata	*This column for plan management after peer review.*
MP	36	Update data submission forms to include new metadata; Cascade metadata to other i5k Workspace applications (Apollo, BLAST); 100% of existing analysis records updated with new metadata	Updated submission forms; Updated content; Updated metadata	
MP	48	Fine-tune data submission forms and content ingest workflows	Updated content	
MP	60	Update older content with new metadata	Updated content	
Goal/Hypothesis		Goal 1d. I5k Workspace content harvested by Ag Data Commons		
SY Team	Months	Milestone	Anticipated Product	Progress/Changes
MP	12	N/A	N/A	
MP	24	Draft implementation plan developed	Draft implementation plan available	
MP	36	Harvest tools developed	RC1 of harvest procedure available on github	*This column for plan management after peer review.*
MP	48	Pilot harvest implemented	Test content ingested at Ag Data Commons	
MP	60	Regular content harvest successfully implemented at Ag Data Commons	First release of harvest procedure available on github; Content ingest at Ag Data Commons from i5k Workspace performed via tools	
Goal/Hypothesis		Goal 1e. Update i5k Workspace resources when updated genome assemblies become available.		
SY Team	Months	Milestone	Anticipated Product	Progress/Changes
MP	12	N/A	Ad-hoc i5k Workspace content updates perfomed	
MP	24	Initiate work on SOP development	Ad-hoc i5k Workspace content updates perfomed	
MP	36	Draft SOPs developed	Draft SOPs available	*This column for plan management after peer review.*
MP	48	Implementation of SOPs initiated	Streamlined i5k Workspace content updates perfomed	
MP	60	Implementation of SOPs	Streamlined i5k Workspace content updates perfomed	

Project Title	The i5k Workspace@NAL: An information portal for arthropod genomes and genes
Project No.	
National Program (Number: Name)	
Objective	Objective 2. Services to improve gene annotation quality
NP Action Plan Component	
NP Action Plan Problem Statement	

Subobjective				
Goal/Hypothesis		Goal 2a. Facilitate community annotation of i5k genomes		
SY Team	Months	Milestone	Anticipated Product	Progress/Changes
MP	12	Annotation training resources identified and collated; Bimonthly 1-hour Apollo training webinars held	i5k Workspace web page with annotation training resources listed; Updated annotation training webinars available on i5k workspace webpage; improved annotator knowledge	
MP	24	Faculty collaborators recruited if necessary; Undergraduate annotation training material identified; Naming standards fully incorporated into training; Initiate multi-hour training module development	Collaborations with faculty for undergraduate training in manual annotation; Help page at i5k Workspace with materials to faciliate undergraduate manual annotation training; Improved naming by i5k community annotators	*This column for plan management after peer review.*
MP	36	Initial multi-hour training modules developed; Name QC integrated into OGS QC pipeline	Improved manual annotations by i5k community; More appropriate gene and protein names used in i5k annotations	
MP	48	Ontology development groundwork	Increased understanding of needs for insect ontology development	
MP	60	Ontology collaborators identified and problems defined		
Goal/Hypothesis		Goal 2b. Generate Official Gene Sets, and submit to NCBI		
SY Team	Months	Milestone	Anticipated Product	Progress/Changes
MP	12	Continue OGS generation as needed; Initiate work on OGS processing pipeline for NCBI submissions; 1/3 of OGS NCBI submission backlog processed	New OGS's available at i5k Workspace, Ag Data Commons, GenBank; Updates to GFF3toolkit as needed; At least 3 OGS's submitted to NCBI	
MP	24	OGS processing pipeline for NCBI submissions used on at least one NCBI submission; 1/3 of OGS NCBI submission backlog processed; Continue OGS generation as needed	rc1 release of OGS processing scripts; At least 3 OGS's submitted to NCBI; New OGS's available at i5k Workspace, Ag Data Commons, GenBank	*This column for plan management after peer review.*
MP	36	OGS processing pipeline for NCBI submissions used on at least one NCBI submission; Continue OGS generation as needed; Tripalv3 installation	v1 release of OGS processing scripts; New OGS's available at i5k Workspace, Ag Data Commons, GenBank; Updated gene pages are available at the i5k Workspace	
MP	48	Continue OGS generation as needed	New OGS's available at i5k Workspace, Ag Data Commons, GenBank	
MP	60	Continue OGS generation as needed	New OGS's available at i5k Workspace, Ag Data Commons, GenBank	
Goal/Hypothesis		Goal 2c. Computationally generate functional annotations of i5k genomes		
SY Team	Months	Milestone	Anticipated Product	Progress/Changes

MP	12	2-3 ARS scientists for pipeline testing recruited; Initial release of functional annotation pipeline; Testing of functional annotation pipeline complete	Functional annotation pipeline v1 released; Eight arthropod genomes functionally annotated; Functional annotations from eight genomes shared via AgBase, ADC, i5k Workspace and/or other platform	
MP	24	Continued internal testing and use of functional annotation pipeline	At least 4 arthropod genomes functionally annotated and released to public	*This column for plan management after peer review.*
MP	36	Continued internal testing and use of functional annotation pipeline	At least 4 arthropod genomes functionally annotated and released to public	
MP	48	Continued internal testing and use of functional annotation pipeline	At least 4 arthropod genomes functionally annotated and released to public	
MP	60	Continued internal testing and use of functional annotation pipeline	At least 4 arthropod genomes functionally annotated and released to public	

Project Title	The i5k Workspace@NAL: An information portal for arthropod genomes and genes
Project No.	
National Program (Number: Name)	
Objective	Objective 3. Systems development and maintenance: analyze and improve our platform architecture
NP Action Plan Component	
NP Action Plan Problem Statement	
Subobjective	
Goal/Hypothesis	Goal 3a. Analyze platform architecture

SY Team	Months	Milestone	Anticipated Product	Progress/Changes
CC	12	Define requirements and use cases for sequence search	Documented software requirements and use cases	
CC	24	Define plan for improvement for sequence search; Define requirements and use cases for manual gene curation	Product improvement plan; Documented software requirements and use cases	*This column for plan management after peer review.*
CC	36	Define plan for improvement for manual gene curation	Product improvement plan	

CC	48	Define requirements and use cases for content management system	Documented software requirements and use cases	
CC	60	Define plan for improvement for content management system	Product improvement plan	
Goal/Hypothesis		Goal 3b. Codify our project best practices		
SY Team	Months	Milestone	Anticipated Product	Progress/Changes
CC	12	Documentation for deployments and SOP for Genomics Workspace; Assess current test thoroughness, define plan for improvement for sequence search	SOP and deployment docs for every app; Report on test coverage	
CC	24	Documentation for deployments and SOP for Apollo; Assess current test thoroughness, define plan for improvement for manual gene curation	SOP and deployment docs for every app; Report on test coverage	*This column for plan management after peer review.*
CC	36	Documentation for deployments and SOP for Drupal; Assess current test thoroughness, define plan for improvement for content management system	SOP and deployment docs for every app; Report on test coverage	
CC	48	Refine SOP and deployment docs as needed	Revised deployment documents	
CC	60	Refine SOP and deployment docs as needed	Revised deployment documents	
Goal/Hypothesis		Goal 3c. Write a system implementation plan		
SY Team	Months	Milestone	Anticipated Product	Progress/Changes
CC	12	Develop plan for better meeting Technical Control Board deployment requirements for Drupal	TCB request best practices document	
CC	24	Develop plan for better meeting Technical Control Board deployment requirements for Apollo	TCB request best practices document	*This column for plan management after peer review.*
CC	36	Develop plan for better meeting Technical Control Board deployment requirements for Genomics Workspace	TCB request best practices document	
CC	48	Refine TCB request best practices documents as needed	Revised documents	
CC	60	Refine TCB request best practices documents as needed	Revised documents	
Goal/Hypothesis		Goal 3d. Develop a migration and deployment plan for Tripal 3		
SY Team	Months	Milestone	Anticipated Product	Progress/Changes
CC	12	Develop initial update roadmap; Define requirements for migration; Implement initial steps as described in roadmap	Tripal 3 migration roadmap; Requirements document for migration; Migration documentation	

CC	24	Complete migration as described in roadmap	Migration documentation	*This column for plan management after peer review.*
CC	36	Review status and confirm that migration is complete	Updated Tripal instance	
CC	48			
CC	60			
Goal/Hypothesis		Goal 3e. Implement upgrade plans developed in Goal 3a		
SY Team	Months	Milestone	Anticipated Product	Progress/Changes
CC	12	Deploy updates for sequence search tool	Updated search tool application	
CC	24	Deploy updates for Apollo	Updated Apollo application	*This column for plan management after peer review.*
CC	36	Deploy updates for content management system	Update content management system	
	48			
	60			
Goal/Hypothesis		Goal 3f. Use project requirements to determine if better solutions exist for our current systems		
SY Team	Months	Milestone	Anticipated Product	Progress/Changes
	12			
CC	24	Classify and score possible options based on requirements for sequence search	Scored and sorted list of possible alternative software	*This column for plan management after peer review.*
CC	36	Classify and score possible options based on requirements for manual gene annotation	Scored and sorted list of possible alternative software	
	48			
CC	60	Classify and score possible options based on requirements for a content management system	Scored and sorted list of possible alternative software	

Project No.	
National Program (Number: Name)	
Objective	Objective 4. Outreach and communications with stakeholders to improve awareness of the i5k Workspace
NP Action Plan Component	
NP Action Plan Problem Statement	
Subobjective	
Goal/Hypothesis	Goal 4a. Continue and increase outreach activities to i5k Workspace stakeholders.

SY Team	Months	Milestone	Anticipated Product	Progress/Changes
MP	12	Quarterly performance metrics captured; Bi-monthly webinars	Metrics reported on i5k Workspace metrics page; Webinar slides available at i5k Workspace	
MP	24	Quarterly performance metrics captured; Bi-monthly webinars; i5k Workspace (or additional) booth at Entomological Society of America meeting	Metrics reported on i5k Workspace metrics page; Webinar slides available at i5k Workspace; Increased visibility/use of i5k Workspace site	*This column for plan management after peer review.*
MP	36	Quarterly performance metrics captured; Bi-monthly webinars	Metrics reported on i5k Workspace metrics page; Webinar slides available at i5k Workspace	
MP	48	Quarterly performance metrics captured; Bi-monthly webinars	Metrics reported on i5k Workspace metrics page; Webinar slides available at i5k Workspace	
MP	60	Quarterly performance metrics captured; Bi-monthly webinars	Metrics reported on i5k Workspace metrics page; Webinar slides available at i5k Workspace	

Project Title	The i5k Workspace@NAL: An information portal for arthropod genomes and genes
Project No.	
National Program (Number: Name)	
Objective	Objective 5. Collaboration with other database partners on joint standards development, software development, and their implementation
NP Action Plan Component	
NP Action Plan Problem Statement	
Subobjective	
Goal/Hypothesis	Goal 5a. Collaborate with the AgBioData consortium to guide i5k Workspace best practices

SY Team	Months	Milestone	Anticipated Product	Progress/Changes
MP	12	AgBioData PAG workshop	AgBioData bylaws finalized	
MP	24	AgBioData workshop 1	Improved recommendations for data federation across biological databases	*This column for plan management after peer review.*

MP	36	AgBioData workshop 2	Improved recommendations for data federation across biological databases	
MP	48	AgBioData workshop 3	Improved recommendations for data federation across biological databases	
MP	60	AgBioData workshop 4	Improved recommendations for data federation across biological databases	
Goal/Hypothesis		Goal 5b. Collaborate with University of Tennessee-Knoxville on development and implementation of FAIR Tripal modules for genome-centric data		
SY Team	Months	Milestone	Anticipated Product	Progress/Changes
MP	12	FAIR Tripal paper draft submitted to peer-reviewed journal	Paper on the Tripal software and FAIR data standards in genomics	
MP	24	Tripal3 update initiated; Content migration to Tripal3 initiated where necessary	Tripal3 code deployed on dev site; Content migration performed on dev site	*This column for plan management after peer review.*
MP	36	All Tripal modules updated to Tripal3	Tripal3 code deployed on production	
	48			
	60			

LITERATURE CITED

1. Coates, B. S. *et al.* Arthropod genomics research in the United States Department of Agriculture-Agricultural Research Service: Current impacts and future prospects. *Trends Entomol.* **11**, (2015).
2. Poelchau, M. F. *et al.* Agricultural applications of insect ecological genomics. *Curr. Opin. Insect Sci.* **13**, 61–69 (2016).
3. Poelchau, M. *et al.* The i5k Workspace@NAL--enabling genomic data access, visualization and curation of arthropod genomes. *Nucleic Acids Res.* (2014). doi:10.1093/nar/gku983
4. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
5. Thomas, G. W. C. *et al.* The Genomic Basis of Arthropod Diversity. (2018). doi:10.1101/382945
6. Harper, L. *et al.* AgBioData consortium recommendations for sustainable genomics and genetics databases for agriculture. *Database* **2018**, (2018).
7. FlyBase Consortium. The FlyBase database of the Drosophila genome projects and community literature. *Nucleic Acids Res* **30**, (2002).
8. Giraldo-Calderón, G. I. *et al.* VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. *Nucleic Acids Res.* **43**, D707–D713 (2015).
9. Elsik, C. G. *et al.* Hymenoptera Genome Database: integrating genome annotations in HymenopteraMine. *Nucleic Acids Res.* **44**, D793–D800 (2016).
10. Legeai, F. *et al.* AphidBase: a centralized bioinformatic resource for annotation of the pea aphid genome. *Insect Mol. Biol.* **19**, 5–12 (2010).
11. Benson, D. A., Boguski, M. S., Lipman, D. J., Ostell, J. & Ouellette, B. F. GenBank. *Nucleic Acids Res* **26**, (1998).
12. Sanderson, L.-A. *et al.* Tripal v1.1: a standards-based toolkit for construction of online genetic and genomic databases. *Database* **2013**, bat075–bat075 (2013).

13. Lee, E. *et al.* Web Apollo: a web-based genomic annotation editing platform. *Genome Biol.* **14**, R93 (2013).
14. i5K Consortium. The i5K Initiative: Advancing Arthropod Genomics for Knowledge, Human Health, Agriculture, and the Environment. *J. Hered.* **104**, 595–600 (2013).
15. Lewin, H. A. *et al.* Earth BioGenome Project: Sequencing life for the future of life. *Proc. Natl. Acad. Sci.* **115**, 4325–4333 (2018).
16. Yandell, M. & Ence, D. A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet.* **13**, 329–342 (2012).
17. Zattara, E. E., Macagno, A. L. M., Busey, H. A. & Moczek, A. P. Development of functional ectopic compound eyes in scarabaeid beetles by knockdown of *orthodenticle*. *Proc. Natl. Acad. Sci.* **114**, 12021–12026 (2017).
18. Gundersen-Rindal, D. E. & et al. Arthropod genomics research in the United States Department of Agriculture, Agricultural Research Service: Application of RNA interference and CRISPR gene-editing technologies in pest control. Trends in Entomology 13: 109-137. *Trends Entomol.* **13**, 109–137 (2017).
19. Wood, V. *et al.* PomBase: a comprehensive online resource for fission yeast. *Nucleic Acids Res.* **40**, D695–D699 (2012).
20. Hosmani, P. S. *et al.* A quick guide for student-driven community genome annotation. *ArXiv180503602 Q-Bio* (2018).
21. Ruzzante, L., Reijnders, M. J. M. F. & Waterhouse, R. M. Of Genes and Genomes: Mosquito Evolution and Diversity. *Trends Parasitol.* (2018). doi:10.1016/j.pt.2018.10.003
22. Conesa, A. *et al.* Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 3674–3676 (2005).
23. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).

24. Mungall, C. J., Emmert, D. B. & The FlyBase Consortium. A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics* **23**, i337–i346 (2007).
25. Buels, R. *et al.* JBrowse: a dynamic web platform for genome visualization and analysis. *GENOME Biol.* **17**, (2016).
26. Howe, D. G. *et al.* Model organism data evolving in support of translational medicine. *Lab Anim.* **47**, 277–289 (2018).
27. Amstutz, P. *et al.* Common Workflow Language, v1.0. (2016). doi:10.6084/m9.figshare.3115156.v2
28. Beyer, B. *Site Reliability Workbook: Practical Ways to Implement SRE.* (O'Reilly Media, Incorporated, 2018).
29. *Site reliability engineering: how Google runs production systems*. (Oreilly, 2016).
30. McConnell, S. *Software project survival guide*. (Microsoft Press, 1998).
31. McKenna, D. D. *et al.* Genome of the Asian longhorned beetle (Anoplophora glabripennis), a globally significant invasive species, reveals key functional and evolutionary innovations at the beetle-plant interface. *GENOME Biol.* **17**, (2016).
32. Benoit, J. B. *et al.* Unique features of a global human ectoparasite identified through sequencing of the bed bug genome. *Nat Commun* **7**, (2016).
33. Papanicolaou, A. *et al.* The whole genome sequence of the Mediterranean fruit fly, Ceratitis capitata (Wiedemann), reveals insights into the biology and adaptive evolution of a highly invasive pest species. *Genome Biol.* **17**, 192 (2016).
34. Poynton, H. C. *et al.* The Toxicogenome of *Hyalella azteca* : A Model for Sediment Ecotoxicology and Evolutionary Toxicology. *Environ. Sci. Technol.* **52**, 6009–6022 (2018).
35. Schoville, S. D. *et al.* A model species for agricultural pest genomics: the genome of the Colorado potato beetle, Leptinotarsa decemlineata (Coleoptera: Chrysomelidae). *Sci. Rep.* **8**, (2018).
36. Saha, S. Improved annotation of the insect vector of citrus greening disease: biocuration by a diverse genomics community. *Database* **2017**, (2017).

37. Harrison, M. C. *et al.* Hemimetabolous genomes reveal molecular basis of termite eusociality. *Nat. Ecol. Evol.* **2**, 557–566 (2018).

38. Armisen, D. *et al.* The genome of the water strider Gerris buenoi reveals expansions of gene repertoires associated with adaptations to life on the water. (2018). doi:10.1101/242230

39. Panfilio, K. A. *et al.* Molecular evolutionary trends and feeding ecology diversification in the Hemiptera, anchored by the milkweed bug genome. (2018). doi:10.1101/201731

40. Robertson, H. M. *et al.* Genome sequence of the wheat stem sawfly, Cephus cinctus, representing an early-branching lineage of the Hymenoptera, illuminates evolution of hymenopteran chemoreceptors. *Genome Biol. Evol.* (2018). doi:10.1093/gbe/evy232

41. Robertson, H. M., Baits, R. L., Walden, K. K. O., Wada-Katsumata, A. & Schal, C. Enormous expansion of the chemosensory gene repertoire in the omnivorous German cockroach *Blattella germanica*. *J. Exp. Zoolog. B Mol. Dev. Evol.* (2018). doi:10.1002/jez.b.22797

42. Robertson, H. M. Noncanonical GA and GG 5 ` Intron Donor Splice Sites Are Common in the Copepod Eurytemora affinis. *G3-GENES GENOMES Genet.* **7**, 3967–3969 (2017).

43. Chen, M.-J. M., Lin, H., Chiang, L.-M., Childers, C. P. & Poelchau, M. F. The GFF3toolkit: QC and Merge Pipeline for Genome Annotation. in *Insect Genomics: Methods and Protocols* (eds. Brown, S. J. & Pfrender, M. E.) 75–87 (Springer New York, 2019). doi:10.1007/978-1-4939-8775-7_7

44. Taning, C. N. T., Andrade, E. C., Hunter, W. B., Christiaens, O. & Smagghe, G. Asian Citrus Psyllid RNAi Pathway - RNAi evidence. *Sci. Rep.* **6**, (2016).

45. McKenna, D. D. *et al.* Genome of the Asian longhorned beetle (*Anoplophora glabripennis*), a globally significant invasive species, reveals key functional and evolutionary innovations at the beetle–plant interface. *Genome Biol.* **17**, 227 (2016).

46. Brand, P. *et al.* The origin of the odorant receptor gene family in insects. *eLife* **7**, (2018).

47. Thomas, G. W. C. *et al.* The Genomic Basis of Arthropod Diversity. *bioRxiv* (2018). doi:10.1101/382945

ACCOMPLISHMENTS FROM PRIOR PROJECT PERIOD

Terminating ARS Research Project Number: N/A
Title: N/A
Project Period: N/A

Project Accomplishments and Impact:
The i5k Workspace was initiated in 2013, outside of the OSQR review cycle. In the past 5 years, we have accomplished the following (See also the Appendix for highlighted success stories of the i5k Workspace):

Data acquisition and stewardship.
Accomplishments:
- Published a paper on the i5k Workspace@NAL in Nucleic Acids Research in 2014[3];
- Onboarded 70 organisms;
- Added 45 i5k datasets at the Ag Data Commons;
- Added 209 datasets to the i5k Workspace

Impact:
- As of 12/2018, we have 523 registered users;
- 7,923 users in FY2018 per google analytics;
- NAR paper has been cited 47 times as of November 2018;
- 127 Google scholar citations of the i5k Worskpace URL as of November 2018.

Manual annotations and value-added services.
Accomplishments:
- 15,409 manually annotated genes were generated in the course of the i5k pilot project; more are untracked;
- 9 Official Gene Sets were created de novo with the GFF3toolkit programs;
- 4 Official Gene Sets were updated with remap-gff3.

Impact:
- 12 published, peer-reviewed manuscripts, to our knowledge, have used our manual annotation and/or OGS generation services in their results. [5,31–42]

Software.
Accomplishments:
- Developed multiple software applications (https://github.com/NAL-i5K/), including the Genomics-Workspace (https://github.com/NAL-i5K/genomics-workspace), the sequence search application of the i5k Workspace@NAL; and the GFF3Toolkit[43], the foundation of much of our data processing.

Impact:
- 9 Official Gene Sets were created de novo with the GFF3toolkit program
- 574 github commits were perfomed in FY2018.

How past accomplishments relate to the current project plan.
The i5k Workspace will continue to build on past accomplishments in the current project cycle. The past 5 years have demonstrated to us that the scientific community is using and accepting the i5k Workspace as a resource for arthropod genomics, and that further services regarding genome and gene annotation access and curation are needed. We have also gained experience in developing, deploying, and maintaining the software packages required to give our stakeholders access and tools to interact with the data hosted at the i5k Workspace.

ISSUES OF CONCERN STATEMENT

Animal Care: Not relevant
Endangered Species: Not relevant
National Environmental Policy Act: On the basis that this Federal project is undertaken for the sole purpose of conducting research, this project is categorically excluded, in accordance with the National Environmental Policy Act.
Human Study Procedure: Not relevant
Laboratory Hazards: Not relevant
Occupational Safety and Health: The work will be conducted in the office environment.
Biosafety/Biosecurity/Quarantine: Not relevant
Intellectual Property Issues: This research will be conducted to create resources maintained in the public domain.

Non-assistance Cooperative Agreements (NACA)

Capacity building to solve complex computational challenges in agricultural research

Cooperator: National Taiwan University
Objective: Mutually build institutional capacity in computational bioscience and scientific big data management; and train a workforce to serve in these emerging science and technology fields.
Approach: Renew a five-year program to exchange graduate students from NTU. One to three students per year will serve a one-year rotation at the National Agricultural Library (NAL). While at the NAL, the students will work on computational and programming problems in the areas of genomics and scientific data management. Problems will be selected based on the computational needs of the Agricultural Research Service, NAL and the agricultural research community. Research on these problems will provide developmental opportunities for the students to work on real world computational and programming challenges while building their skills and expanding their professional networks. One NTU Faculty member – ideally, with students enrolled in the program, but possibly faculty with prospective students - will perform one visit to the NAL per year in order to 1) present a seminar on their research and mutual interests to the NAL (and a broader audience if desired); 2) work with their students on current projects; and 3) meet with individual NAL staff members to continue building connections between NTU and the NAL.
Relevance to Project Plan: Graduate student interns will work directly on i5k Workspace objectives, primarily from Objectives 1-3. The exact work performed is determined based on the fit of the intern to the work.

Tripal development for FAIR genomic data

Cooperator: University of Tennessee - Knoxville
Objective: This non-assistance cooperative agreement aims to develop and improve Tripal's resources for FAIR genomic data.
Approach: As much as possible, Tripal software will be developed such that the development effort can be shared with the larger Tripal community, e.g. as an extension module shared on the Tripal site (http://tripal.info/extensions). The Tripal software will be developed to accommodate the following: 1) Develop a module that allows Tripal users to retrieve data and metadata in a more genome-centric way; 2) Improve the organization and display of gene families or related genes; 3) Improve metadata ingest for user-submitted datasets; 4) Improve on the i5k Workspace's NCBI metadata ingest module; and 5) Improved provenance documentation for Tripal data at the feature, organism, analysis and project level. In addition, the i5k Workspace site will be initially be upgraded to Tripal 3, and later on to Drupal8/Tripal4. As part of this process, the i5k Workspace Tripal site will be maintained and supported.
Relevance to Project Plan: This NACA is directly related to work in Objective 3, Goal 3d "Develop a migration and deployment plan for Tripal 3", and Objective 5, Goal 5b "Collaborate with University of Tennessee-Knoxville on development and implementation of FAIR Tripal modules for genome-centric data"

Generation of a functional annotation pipeline for arthropod genomes

Cooperator: University of Arizona
Objective:
1. Develop a functional annotation pipeline tuned for arthropod genomes;
2. Annotate at least 8 arthropod genomes using the pipeline;

3. Make the finished pipeline available for public use;
4. Share the resulting datasets with the public.

Approach: The Cooperator will develop a functional annotation pipeline and documentation tuned towards non-model arthropod genomes. The Cooperator will test the pipeline based on datasets that ARS will provide. ARS will test the pipeline on local hardware and will provide feedback for improvement. Both the pipeline and the resulting functional annotation datasets will be shared with the public on appropriate platforms.

Relevance to Project Plan: This Research Support Agreement is directly related to Objective 2, Goal 2c: Computationally generate functional annotations of i5k genomes.

APPENDICES

Appendix 1. List of Acronyms.

Ag100Pest – The 100 agricultural pest genomes project
ARS – Agricultural Research service
BLAST – Basic Local Alignment Search Tool
Cas9 - CRISPR-associated protein 9
CRISPR - **clustered regularly interspaced short palindromic repeats**
CWL – Common Workflow Language
DATS model - DatA Tag Suite model
DOI – Digital Object Identifier
FAIR – Findable, Accessible, Interoperable, Re-usable
FTE – Full-time employee
GFF3 – Generic feature format 3
GGB database – Genomic, genetic and breeding database
GO – Gene Ontology
HMM – Hidden Markov Model
INSDC – International Nucleotide Sequence Database Collaboration
KEGG – Kyoto Encyclopedia of Genes and Genomes
NAL – National Agricultural Library
NCBI – National Center for Biotechnology Information
NP – National Program
OSS – Open-source software
PAG meeting – Plant and Animal Genomes meeting
QA – Quality Assurance
QC – Quality Control
RCN – Research Coordiation Network
SO – Sequence Ontology
SOP – Standard Operation Procedure
UniProt – The Universal Protein Resource
USDA – United States Department of Agriculture
YAML – Yet Another Markup Language

Appendix 2. Research Highlights.

Highlight 1. Genes involved in insecticide resistance.

The bed bug[32], medfly[33] and Colorado potato beetle[35] genomes – all insect pests of great agricultural or medical concern – contain catalogues of genes hypothesized to be involved in insecticide resistance. **For each of these pest species, researchers used i5k Workspace tools to improve the structural and functional annotation of these insecticide resistance genes.** Knowledge of the full repertoire and correct structure of these genes will enable further research on insect pest control.

Highlight 2. RNAi genes in the Asian Citrus Psyllid.

I5k resources were used to help identify RNAi genes in the Asian Citrus Psyllid *Diaphorina citri*, the vector of citrus greening disease, a devastating disease of citrus crops. RNAi technology allows researchers to silence genes in living organisms, a potentially powerful method to limit the impact of insect pests on agricultural crops. **The analysis facilitated by the i5k Workspace provided evidence of a functional RNAi machinery in *Diaphorina citri*, which could be further exploited to develop RNAi-based management strategies[44].**

Highlight 3: Characterization of wood-digesting genes in *Anoplophora glabripennis*.

The Asian longhorned beetle *Anoplophora glabripennis* is an invasive, wood-eating pest species, and can cause severe damage to economically important hardwood trees. The genes that allow this insect to digest wood are therefore of great interest from an evolutionary and pest control perspective. **i5k Workspace tools allowed researchers to characterize the repertoire of genes encoding plant cell wall degrading enzymes (PCWDEs), with a particular focus on the glycoside hydrolase gene family[45].** This initial characterization of glycoside hydrolase genes allowed the researchers to verify their function via follow-up experiments. Gene families encoding PCWDEs have expanded in *Anoplophora glabripennis*, and there is a remarkable diversity of families in their arsenal, suggesting that the expanded number and diversity of genes are an adaptation to wood-feeding in this beetle. Some of these genes may have been acquired via lateral gene transfer from bacterial symbionts. **This knowledge establishes the basis for further experiments to disrupt PCWDE function in this beetle.**

Highlight 4. Characterization of chemoreceptors across arthropod genomes.

Chemosensory receptors allow arthropods to detect and discriminate external chemicals – allowing them to smell and taste the physical world, recognize mates, and detect animal or plant hosts. Chemoreceptors are therefore of extreme interest to biologists as targets for insect control. Chemoreceptor gene families are exceptionally diverse, have expanded and contracted multiple times within Arthropoda, and are thus difficult to accurately predict using automated gene prediction methods. **Researchers are using the i5k Workspace's tools to identify and curate chemoreceptors in many of the i5k Workspace genomes.** Chemoreceptor annotations have been included into the i5k Workspace Official Gene Sets, and in some cases been submitted to NCBI (where all will eventually be submitted). **The research conducted using i5k Workspace resources has facilitated many findings regarding the evolutionary dynamics of chemoreceptor gene families, including in agricultural pest species:**

- The odorant receptor gene family likely originated with the evolution of terrestriality in insects[46];
- Lineage-specific chemoreceptor gene family expansions in the major pest species *Cephus cinctus* (wheat stem sawfly) might be involved in adaptations to new grasses, including wheat[40];
- The Colorado potato beetle, *Leptinotarsa decemlineata*, contains expansions in the gustatory receptor family that may be an adaption to exploiting hosts in the nightshade family[35];

- The German cockroach, *Blattella germanica*, has an enormous expansion of chemosensory genes, likely related to its extreme omnivorous diet[41];
- Bed bugs have a reduced chemosensory gene repertoire, likely due to their bloodfeeding ecology[32];
- Research into chemoreceptors in the copepod *Eurytemora affinis* uncovered that this organism has evolved a modified intron donor splice site recognition, a completely unanticipated discovery[42].

Highlight 5. Genome data management support for the i5k pilot project.

The i5k pilot project is a major undertaking to sequence, assemble and annotate the genomes of 28 arthropod species (https://www.hgsc.bcm.edu/arthropods/i5k)[47]. The i5k Workspace supported the i5k pilot project with data management services – generating Official Gene Sets, and submitting gene sets to primary repositories such as the Ag Data Commons (https://data.nal.usda.gov/i5k) and GenBank. Without the i5k Workspace, i5k pilot would have had substantial difficulties 1) improving gene annotations; 2) generating Official Gene Sets, and 3) depositing these gene annotations in primary repositories. All of these activities are necessary for a genome resource to become a sustainable community resource and approach model-organism quality.