The i5k Workspace@NAL—enabling genomic data access, visualization and curation of arthropod genomes

Monica Poelchau1*, Christopher Childers1*, Gary Moore1, Vijaya Tsavatapalli1, Jay Evans2, Chien-Yueh Lee1,3, Han Lin1,3, Jun-Wei Lin1,4 and Kevin Hackett5

1National Agricultural Library, Beltsville, MD 20705, USA, 2Bee Research Laboratory, U.S. Department of Agriculture–Agricultural Research Service, Beltsville, MD 20705, USA, 3Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei 10617, Taiwan, 4Graduate Institute of Electrical Engineering, National Taiwan University, Taipei 10617, Taiwan and 5Crop Production and Protection, U.S. Department of Agriculture–Agricultural Research Service, Beltsville, MD 20705, USA

Received August 21, 2014; Revised October 2, 2014; Accepted October 03, 2014

ABSTRACT

The 5000 arthropod genomes initiative (i5k) has tasked itself with coordinating the sequencing of 5000 insect or related arthropod genomes. The resulting influx of data, mostly from small research groups or communities with little bioinformatics experience, will require visualization, dissemination and curation, preferably from a centralized platform. The National Agricultural Library (NAL) has implemented the i5k Workspace@NAL (http://i5k.nal.usda.gov/) to help meet the i5k initiative’s genome hosting needs. Any i5k member is encouraged to contact the i5k Workspace with their genome project details. Once submitted, new content will be accessible via organism pages, genome browsers and BLAST search engines, which are implemented via the open-source Tripal framework, a web interface for the underlying Chado database schema. We also implement the Web Apollo software for groups that choose to curate gene models. New content will add to the existing body of 35 arthropod species, which include species relevant for many aspects of arthropod genomic research, including agriculture, invasion biology, systematics, ecology and evolution, and developmental research.

INTRODUCTION

Insects are an incredibly diverse class, with over 1 million described species. They provide essential pollination services for agriculture (1,2), yet cause substantial damage to crops (3), and are vectors of devastating diseases (4). Further, they are important ecological, evolutionary, developmental and medical models. The genomes of insects sequenced to date have already provided important insights into genome architecture and evolution (5–7), immune response pathways (8), eusociality (9,10) and speciation (11,12). Sequencing further genomes brings great promise to answer pending questions for basic and applied biology.

Decreasing whole-genome sequencing costs, coupled with moderate genome sizes for insects and their relatives, favor new genome projects for this group. One downside for arthropod comparative genomic research is the great divergence times between arthropod groups, often in the hundreds of millions of years. Comparative analyses would therefore benefit tremendously from sequence information across the breadth of Arthropoda. The 5000 arthropod genomes initiative (i5k) has therefore set the goal to coordinate the sequencing of 5000 insect or related arthropod species (13,14). As such, the i5k initiative should galvanize the generation of large amounts of data of exceptional comparative value. While the i5k initiative will provide guidance on the sequencing of genomes, the onus is still on individual labs with a specific interest in these genomes to organize the sequencing, analysis and curation of their genome projects. To ensure wide re-use of genomic data resulting from the i5k project, it is important that this data is hosted in a centralized location for data sharing, dissemination, visualization and curation. However, developing, deploying and maintaining databases and web servers for genome access and curation (‘genome portals’) is often beyond the financial and technical reach of smaller genome projects.

Here, we introduce the i5k Workspace@NAL (https://i5k.nal.usda.gov), which the National Agricultural Library (NAL) designed to meet the genome hosting needs of the
Figure 1. View of the JBrowse genome browser for the copepod *Eurytemora affinis*. G + C content and gap tracks, MAKER-predicted protein-coding gene set, forward-strand RNA-Seq coverage and junction read alignments are shown.

The i5k Workspace@NAL has two main goals. First, it aims to help the i5k ‘data producers’, in particular ‘orphaned’ groups without the technical or financial means for genome hosting, at the interface of sequence retrieval and analysis—i.e. how do you access, visualize, curate and disseminate your data once you have received it from the sequencing center? Second, we aim to provide a unified framework for ‘data consumers’ to retrieve relevant genomic information from our data providers. We outline the i5k Workspace, and explain the steps that we take to help i5k data producers disseminate and curate their genome assemblies, and how we present this content to the i5k data consumers. Finally, we explore the future directions the i5k Workspace@NAL will take to continue improving its services for the i5k community, and the arthropod genomics community at large.

**USING THE i5k WORKSPACE@NAL**

**Data producers**

*Setting up your genome portal.* The i5k@NAL hosts genome assemblies for any arthropod genome project that requires our services. There is no hosting preference for a particular taxonomic group or application, but we ask that no agreements with other genome portals be in place, to avoid redundant hosting and curation efforts. Our only requirement is a genome assembly—preferably approved by NCBI (the U.S. National Institutes of Health National Center for Biotechnology Information) or other INSDC (International Nucleotide Sequence Database Collaboration) members—in FASTA format (scaffold, contig and .agp mapping files). Any sequence features that have been mapped to this assembly can also be submitted, including but not limited to official gene sets (OGS), other consensus gene sets or gene predictions, homology alignments, RNA-Seq mappings and transcriptomes (Figure 1). We provide a tutorial on how to map RNA-Seq reads to a genome in iPlant (https://i5k.nal.usda.gov/content/performing-rnaseq-alignments-ipi], and have developed an extension of the exonerate alignment program (15) that generates gff3-formatted output (https://github.com/hotdogee/exonerate-gff3), allowing users to map moderately-sized transcriptomes against a genome assembly. We will also transform BAM files to BigWig format in-house if requested. We can advise on what files would be useful to visualize, given the genome community’s specific needs. Finally, we ask for genome communities to provide us with information to populate the organism’s ‘landing’ page, as well as metadata about each file given to us to communicate to other users of the data. Data can be transferred to us via ftp or iPlant (https://i5k.nal.usda.gov/content/sharing-files-us). We recommend that each genome community designate a ‘community contact’ to serve as the main contact for data files, as well as mediate the manual curation process if this is to be a part of the genome project.

*Data processing—what we do with your data.* For each new organism and genome assembly, we generate customized organism ‘landing pages’ (e.g. https://i5k.nal.usda.gov).
The i5k initiative strives to coordinate the sequencing of 5000 insect or related arthropod genomes. The i5k Workspace is committed to hosting content for projects under the i5k umbrella that require our services. Therefore, our content is dynamic and grows with the i5k community itself. We currently host data from 35 arthropod species and counting (Table 1), including two non-insect classes: four Arachnidian and two Maxillopodan species. Within class Insecta, we host data on 2 Palaeopteran, 4 Hymenopteran, 4 Coleopteran, 1 Trichopteran, 10 Dipteran, 6 Hemipteran, 1 Thyasanopteran and 1 Lepidopteran species. The main source for this data is the Baylor College of Medicine’s i5k pilot project (https://www.hgsc.bcm.edu/arthropods/i5k-pilot-project-summary), which jump-started the i5k initiative with the coordinated sequencing and annotation of 29 arthropod genomes. Among the species are several of interest to invasion biology (the Asian long-horned beetle (Anoplophora glabripennis)), the Colorado potato beetle (Leptinotarsa decemlineata), the copepod (Eurytemora affinis) and the infamous bed bug (Cimex lectularius). The Baylor College of Medicine’s modENCODE project sequenced eight Drosophila species, which are available (https://www.hgsc.bcm.edu/arthropods/drosophila-modencode-project). We have also started hosting organisms that were not part of the i5k pilot species: the Asian citrus psyllid (Diaphorina citri) (http://psyllid.org/), the copepod (Tigriopus californicus), the tobacco hornworm (Manduca sexta) and the Hessian fly (Mayetiola destructor). In addition to genome assemblies, all species that we host have at least one set of consensus gene models (.gff3, .faa, .fna file formats), and component ‘evidence’ data files of the consensus gene models. Each assembly and annotation set has a separate analysis page that describes how the analysis was performed. Two recently added species, M. sexta and M. destructor, have an OGS, and we are in the process of developing more rigorous search options via the
The interactive BLAST+ result viewer displaying a TBLASTN search of one query sequence on two genomes with 63 HSPs. The focused HSP is highlighted in orange.

Tripal framework for species with an OGS. For the genome browsers, our data processing pipeline generates two tracks based on the reference assembly per project: G+C content as a heatmap and the location of gaps in the assembly. Many groups have also provided additional user contributed data files, such as mapped RNA-Seq reads and transcriptomes to aid manual curation (Figure 1). We can make any of the data hosted in the genome browsers available on our ‘downloads’ page upon request. With the exception of the manually curated genes, which require registration to access, all data is publicly available. After the curation process is completed for each genome, the manually curated genes will be made freely available.

Accessing and visualizing i5k Workspace content. Users can navigate through the available species’ organism or genome browser pages via drop-down menus on the i5k Workspace home page. Data downloads are available via a central downloads page (https://i5k.nal.usda.gov/content/data-downloads). All organisms’ genomes, predicted peptide and transcript sequences can be accessed via our BLAST page (https://i5k.nal.usda.gov/blast), and multiple species can be searched against simultaneously if the same residue type is selected (e.g. either nucleotide or protein). This allows for straightforward retrieval of individual sequences of interest. Performing BLAST searches against a genome generates a link to the JBrowse genome browser, in which the resulting HSPs are visualized as a track. The user can then log in to their Web Apollo account straight from the browser window to begin annotating. For example, an annotator can use BLAST to align a sequence from a well-annotated reference species, such as Drosophila melanogaster, to multiple genome assemblies at once, open the results in multiple genome browsers and begin annotating if they notice that the computationally predicted gene feature needs editing. We also provide a tutorial for using BLAST in the i5k Workspace (https://i5k.nal.usda.gov/content/blast-tutorial). We use Google Analytics to identify user metrics such as duration of stay, aggregated user behavior and entry/exit points. Identification of resources that are in high demand, such as BLAST, is critical for determining what may need refactoring to increase capacity. Alternatively, low traffic resources need to be assessed to determine why they are under utilized and whether steps may be taken to improve usage.

FUTURE DIRECTIONS

The i5k Workspace is available for any arthropod genome project. We therefore anticipate that our content will expand in the future. To prepare for future expansions, we are developing in two directions: automation and delegation. First, we strive to automate as many tasks as possible. We are currently working on an improved and streamlined data submission system that will harvest metadata from submitters in a user-friendly way, and are also developing tools for more consistent processing of gff3 files. Second, not all tasks can be automated; for those that cannot, we will develop tools to delegate most of the responsibility to the community contacts. Finally, we intend to streamline the tools that we offer, and only expand in a few more limited directions. For example, we strive to improve our organism browser to accommodate more species; therefore, we are identifying other options to search our content, for example via tree-based viewers. We will implement tools to visualize pre-computed orthology and paralogy data of official...
Table 1. Species hosted at the i5k Workspace@NAL, their common name and source

| Species name                      | Common name                   | Source                        |
|-----------------------------------|-------------------------------|-------------------------------|
| Agrius convolvuli                 | Emerald ash borer             | BCM i5k pilot                 |
| Anoplophora glabripennis          | Asian long-horned beetle      | BCM i5k pilot                 |
| Athalia rosae                     | Turnip sawfly                 | BCM i5k pilot                 |
| Centruroides exilicauda           | Bark scorpion                 | BCM i5k pilot                 |
| Ceratitis capitata                | Mediterranean fruit fly       | BCM i5k pilot                 |
| Cymex lectularius                 | Bed bug                       | BCM i5k pilot                 |
| Copidosoma floridana              | Parasitic wasp                | BCM i5k pilot                 |
| Diaphorina citri                  | Asian citrus psyllid          | International Psyllid Consort |
| Drosophila biarmipes              | NA                            | BCM modENCODE project         |
| Drosophila bipectinata            | NA                            | BCM modENCODE project         |
| Drosophila elegans                | NA                            | BCM modENCODE project         |
| Drosophila euagruscicornis        | NA                            | BCM modENCODE project         |
| Drosophila fuscipalpita           | NA                            | BCM modENCODE project         |
| Drosophila hippoglobin            | NA                            | BCM modENCODE project         |
| Drosophila rhopaloeae             | NA                            | BCM modENCODE project         |
| Drosophila takahashii             | NA                            | BCM modENCODE project         |
| Ephemerella danica                | Mayfly                        | BCM i5k pilot                 |
| Eurytemora affinis                | Common copepod                | BCM i5k pilot                 |
| Frankliniella occidentalis        | Western flower thrips         | BCM i5k pilot                 |
| Gerris bueoni                     | Water strider                 | BCM i5k pilot                 |
| Hapalodiscia vitripennis          | Glassy-winged sharpshooter    | BCM i5k pilot                 |
| Ladona fulva                      | Scare chaser                  | BCM i5k pilot                 |
| Latrodectus hesperus              | Western black widow spider    | BCM i5k pilot                 |
| Leptinotarsa decemlineata         | Colorado potato beetle        | BCM i5k pilot                 |
| Limnephilus lunatus               | Cadis fly                     | BCM i5k pilot                 |
| Loxosceles reclusa                | Brown recluse spider          | BCM i5k pilot                 |
| Manduca sexta                     | Tobacco hornworm              | Kansas State University       |
| Mayettola destructor              | Hessian fly                   | Purdue University            |
| Oncopeltus fasciatus              | Milkweed bug                  | BCM i5k pilot                 |
| Onthophagus taurus                | Bull-headed dung beetle       | BCM i5k pilot                 |
| Orussus abietinus                 | Parasitic wood wasp           | BCM i5k pilot                 |
| Pachyopsylla venustella           | Hackberry petiole gall psyllid| BCM i5k pilot                 |
| Parasteatoda tepidariorum         | Common house spider           | BCM i5k pilot                 |
| Tityopus californicus             | Harpacticoid copepod          | UC San Diego                  |
| Trichogramma pretiosum            | Parasitic wasp                | BCM i5k pilot                 |

The Baylor College of Medicine is abbreviated as BCM.

Conflicts of interest statement. None declared.

REFERENCES

1. Committee on the Status of Pollinators in North America, N.R.C. (2007) Status of Pollinators in North America. The National Academies Press, Washington, D.C.
2. Losey, J.E. and Vaughan, M. (2006) The economic value of ecological services provided by insects, Bioscience, 56, 311–323.
3. Pimentel, D., Lach, L., Zuniga, R. and Morrison, D. (2000) Environmental and economic costs of nonindigenous species in the United States. Bioscience, 50, 53–65.
4. Lounibos, L.P. (2002) Invasions by insect vectors of human disease. Annu. Rev. Entomol., 47, 233–266.
5. Wyler, S., Kriventseva, E.Y., Schröder, R., Kadowaki, T. and Zdobnov, E.M. (2007) Quantification of ortholog losses in insects and vertebrates. Genome Biol., 8, R242.
6. Drosophila 12 Genomes Consortium, Clark, A.G., Eisen, M.B., Smith, D.R., Bergman, C.M., Oliver, B., Markow, T.A., Kaufman, T.C., Keffer, M., Gelbart, W. et al. (2007) Evolution of genes and genomes on the Drosophila phylogeny. Nature, 450, 203–218.
7. Zdobnov, E.M., von Mering, C., Letunic, I., Torres, D., Suyama, M., Copley, R.R., Christophides, G.K., Thomasova, D., Holt, R.A., Subramanian, G.M. et al. (2002) Comparative genome and proteome analysis of Anopheles gambiae and Drosophila melanogaster. Science, 298, 149–159.
8. Evans, J.D., Aonostein, K., Chen, Y.P., Hetru, C., Imler, J.-L., Jiang, H., Kanost, M., Thompson, G.J., Zou, Z. and Hultmark, D. (2006) Immune pathways and defense mechanisms in honey bees Apis mellifera. Insect Mol. Biol., 15, 645–656.
9. Simola, D.F., Wissler, L., Donahue, G., Waterhouse, R.M., Helmkampf, M., Roux, J., Nygaard, S., Glastad, K.M., Hagen, D.E., Viljakainen, L. et al. (2013) Social insect genomes exhibit dramatic evolution in gene composition and regulation while preserving regulatory features linked to sociality. *Genome Res.*, 23, 1235–1247.

10. Weinstock, G.M., Robinson, G.E., Gibbs, R.A., Weinstock, G.M., Weinstock, G.M., Robinson, G.E., Worley, K.C., Evans, J.D., Maleszka, R., Robertson, H.M. et al. (2006) Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature*, 443, 931–949.

11. Lawniczak, M.K.N., Emrich, S.J., Holloway, A.K., Regier, A.P., Olson, M., White, B., Redmond, S., Fulton, L., Appelbaum, E., Godfrey, J. et al. (2010) Widespread divergence between incipient *Anopheles gambiae* species revealed by whole genome sequences. *Science*, 330, 512–514.

12. Neafsey, D.E., Lawniczak, M.K.N., Park, D.J., Redmond, S.N., Coulibaly, M.B., Traore, S.F., Sagnon, N., Costantini, C., Johnson, C., Wiegand, R.C. et al. (2010) SNP genotyping defines complex gene-flow boundaries among African malaria vector mosquitoes. *Science*, 330, 514–517.

13. Robinson, G.E., Hackett, K.J., Purcell-Miramontes, M., Brown, S.J., Evans, J.D., Goldsmith, M.R., Lawson, D., Okamuro, J., Robertson, H.M. and Schneider, D.J. (2011) Creating a buzz about insect genomes. *Science*, 331, 1386–1386.

14. i5K Consortium. (2013) The i5K initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. *J. Hered.*, 104, 595–600.

15. Slater, G.S.C. and Birney, E. (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, 6, 31.

16. Sanderson, L.-A., Ficklin, S.P., Cheng, C.-H., Jung, S., Feltus, F.A., Bett, K.E. and Main, D. (2013) Tripal v1.1: a standards-based toolkit for construction of online genetic and genomic databases. *Database*, 2013, bat075.

17. Mungall, C.J., Emmert, D.B. and The FlyBase Consortium. (2007) A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics*, 23, i337–i346.

18. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, 10, 421.

19. Skinner, M.E., Uzilov, A.V., Stein, L.D., Mungall, C.J. and Holmes, I.H. (2009) JBrowse: a next-generation genome browser. *Genome Res.*, 19, 1630–1638.

20. Lee, E., Helt, G.A., Reese, J.T., Munoz-Torres, M.C., Childers, C.P., Buels, R.M., Stein, L., Holmes, I.H., Elsik, C.G. and Lewis, S.E. (2013) Web Apollo: a web-based genomic annotation editing platform. *Genome Biol.*, 14, R93.

21. Elsik, C.G., Worley, K.C., Zhang, L., Milshina, N.V., Jiang, H., Reese, J.T., Childs, K.L., Venkatraman, A., Dickens, C.M., Weinstock, G.M. et al. (2006) Community annotation: procedures, protocols, and supporting tools. *Genome Res.*, 16, 1329–1333.