Creating a Microbial Genome Analysis Project Using Publically Available Genome Sequences and Online Bioinformatics Programs

Lori R. Scott[1], Todd C. Nickle[2], Stephen Koury[3], Brad Goodner[4], Katherine Houmiel[5], Ben McFarland[5], Jennifer Tenlen[5], Andrew Lumpe[5], Kimberly Murphy[1], D. Rhoads[6], D. Wood[5] and Daihong Chen[5]

[1]Augustana College Biology Department, Rock Island IL 61201 USA
[2]Mount Royal University, Biology Department, Calgary Alberta CA T3E 6K61
[3]University at Buffalo, Buffalo NY 14214
[4]Hiram College, Biology Department, Hiram OH 44234 USA
[5]Seattle Pacific University, Biology Department, Seattle WA 98119 USA
[6]California State University, Department of Biology, San Bernardino CA 92407 USA
(**loriscott@augustana.edu; tnickle@mtroyal.ca; stvkoury@buffalo.edu**)

The Department of Energy's Joint Genome Institute (DOE JGI) has sequenced the genomes of more than 200 microbes from across the Tree of Life as part of the Genome Encyclopedia of Bacteria and Archaea (GEBA) project. The original GEBA project was completed in collaboration with the Leibniz-Institut DSMZ (Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH). In the pilot study, the DOE JGI sequenced 53 bacterial and three archaeal novel and highly diverse genomes, representing a first step towards a phylogenetically balanced set of sequenced genomes across the microbial tree of life (Wu *et al.*, 2009). An additional approximately 200 GEBA genomes have since been sequenced (DOE JGI, 2015). Most of these organisms are poorly represented in the scientific literature, thereby creating the opportunity for undergraduate and high school students to make significant contributions to our understanding of microbial genome structure and function. Students make these contributions by: 1) manually annotating open reading frames (ORFs) predicted by automated bioinformatics programs for the purpose of expanding the original annotation and identifying errors; and/or 2) performing functional genomics studies to confirm the computer-based function predictions.

Keywords: Bioinformatics, web-based project, microbial genomics

Introduction

Gene annotation is the process of using computer technology to predict a biological function for a gene or region of a chromosome. There are many different bioinformatics tools currently available, which makes choosing among them a daunting task for a novice annotator. Functional genomics is the process of performing wet-lab experimentation to confirm the function predictions. This workshop described the means whereby instructors can create microbial genome analysis projects, emphasizing gene annotation strategies, using a GEBA organism and online bioinformatics tools.

The Guiding Education through Novel Investigation (GENI) project (Figure 1; GENI-Science, 2015) provides detailed instructions for instructors and students on select methods in genome analysis, including both gene annotation and functional genomics studies. One of the projects available through the GENI-Science site is the *Meiothermus ruber* Genome Analysis Project (Figure 2; Scott, 2015). This project uses the bioinformatics platform called GENI-Annotation Collaboration Toolkit (Figure 3; GENI-ACT, 2015) as an undergraduate research tool. GENI-ACT is an adaptation of the original Joint Genome Institute's IMG-ACT system (Ditty *et al.*, 2010). This platform offers instructions on how to use publicly available bioinformatics tools for making

function predictions of putative ORFs, as well as an online lab notebook for data collection. In this workshop, participants were shown how the *M. ruber* Genome Analysis Project uses the GENI-Science and the GENI-ACT platforms in an authentic research experience. More importantly, however, resources are available that allow instructors to adapt the instructional scaffold provided by the *M. ruber* project to a new genome analysis project for a microorganism of their choosing.
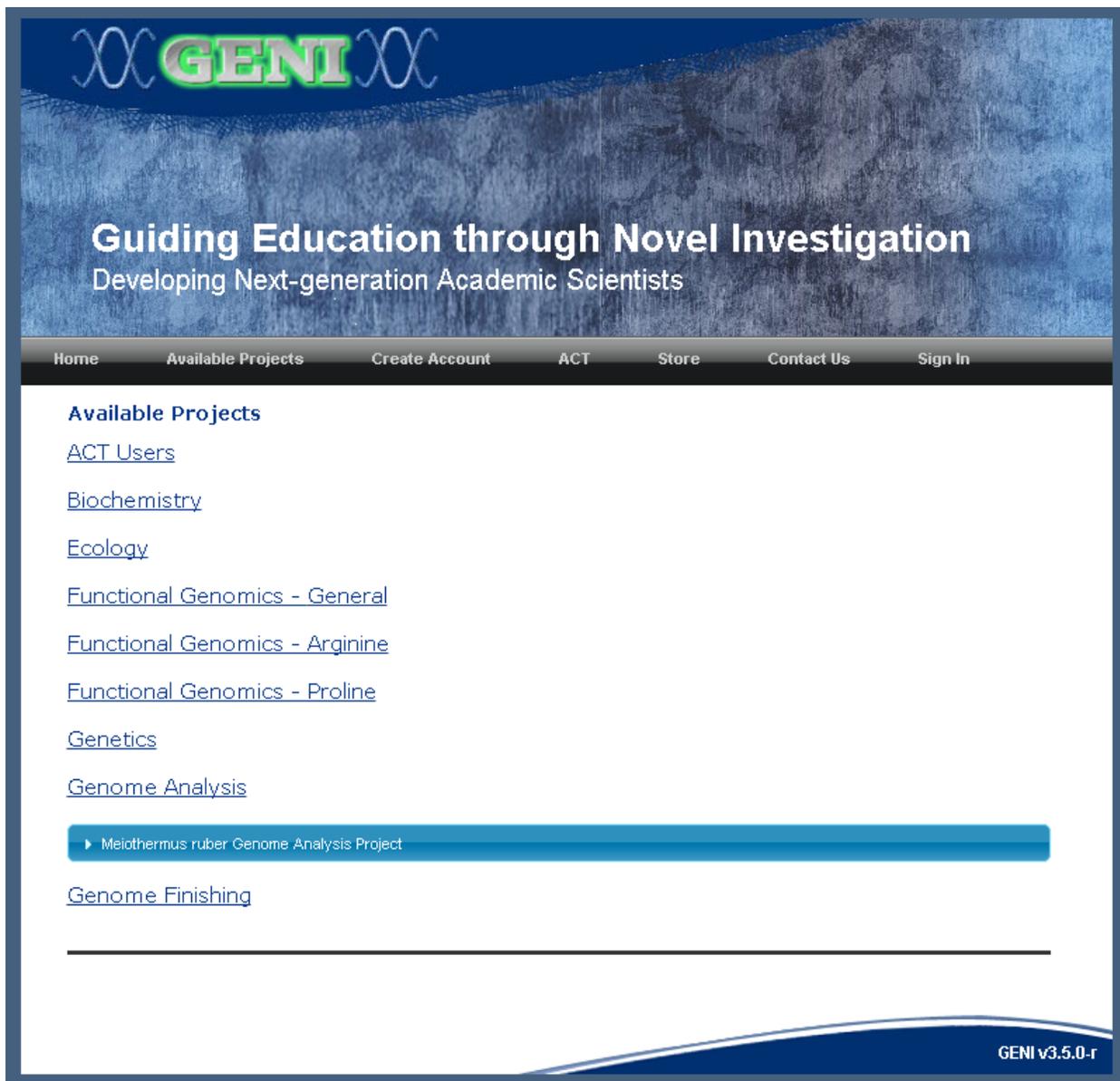
Figure 1. The GENI-Science website is an online resource for detailed instructor and student instructions in select genome analysis projects (taken from http://geni-science.org/availableprojects/ on 1/23/2016)

Meiothermus ruber is one of the GEBA organisms. It belongs to the Deinococcus-Thermus phylum. It is moderately thermophilic, red-pigmented, obligately aerobic and typically isolated from natural hot springs and artificial thermal environments (Nobre *et al.,* 1996). It has a BSL 1 designation (U.S. Public Health Service Guidelines, 2009), which means no extraordinary precautions are required in working with this organism in the lab setting. Its genome was sequenced in 2010 by Tindall *et al.,* (2010), and predicted to be 3,097,457 base pairs in length, with 3,052 protein-coding genes, 55 RNA genes, and 38 pseudogenes.

Figure 2. The *Meiothermus ruber* Genome Analysis Project is a collaboration between multiple institutions, instructors and students, all of whom are contributing to the understanding of the structure and function of genes within the *M. ruber* genome. taken from: https://sites.google.com/a/augustana.edu/meiothermus_ruber/home on 1/23/2016)

Figure 2 shows the homepage for the *Meiothermus ruber* Genome Analysis Project (Scott and Nickel, 2009), which provides an updated description of the progress made in this authentic research experience. The *M. ruber* project provides a genuine research opportunity for students to study biological processes in *M. ruber* that have never been studied in this organism using contemporary bioinformatics and wet-lab tools. It is anticipated that much of the work accomplished by students will eventually be published in the primary scientific literature. As of this date, students have studied amino acid and nucleotide biosynthesis, carbohydrate metabolism, oxidative phosphorylation, to name a few systems. Table 1 contains a sampling of the bioinformatics data collected by students through the GENI-ACT lab notebook. In this study, students compared a putative *M. ruber* gene (locus tag Mrub_1080) to a potential homolog in the model organism *E. coli* K12 MG1655 (locus tag b0242). In addition, students analyze hydropathy plots, pathway maps, and amino acid sequence logos, plus build and analyze phylogenetic trees.

Table 1. Comparison of *E. coli* K12 MG1655 γ-glutamyl kinase (proB, locus tag b0242) and Mrub_1080, the first enzyme in the proline biosynthesis pathway.

Evidence	*E. coli*	*M. ruber*
NCBI BLAST[1] *E. coli* K12 MG1655 proB against *M. ruber*	Score = 302 E-value =7e-100 % identity = 47%	
KEGG pathway[2]	Arginine and proline metabolism Map 00330	
E.C. number[3]	E.C.2.7.2.11 Glutamate 5-kinase	
Cellular location (as predicted by TMHMM[4], LipoP[5], SignalP[6], and Probius[7])	Cytoplasm	
PSORT-B[8]	Cytoplasm (score: 10.00)	Cytoplasm (score: 9.97)
Conserved Domain Database (CDD)[9]	COG0263 E-value = 0 proB (Glutamate 5-kinase)	COG0263 E value = 1.30e-168 proB (Glutamate 5-kinase)
TIGRFAM[10]	TIGR01027 Score = 816.0, E = 3.1e-242 proB: glutamate 5-kinase	TIGR01027 Score = 802.5, E = 3.5e-238 proB: glutamate 5-kinase
Pfam[11]	PF00696 Amino Acid Kinase Family Score = 150.8, E = 4.1e-44	PF00696 Amino Acid Kinase Family Score = 142.1, E = 1.9e-41
Protein Database (PDB)[12]	2J5T E value = 0.0 Glutamate 5-kinase	2J5T E value = 5.39672e-79 Glutamate 5-kinase

Figure 3. The GENI-ACT bioinformatics platform is a collection of publically available bioinformatics tools and online lab notebook for use in the annotation of genes from thousands of sequenced bacterial genomes. (taken from http://www.geni-act.org/ on 1/23/2016)

Materials

Bioinformatics programs used by L. Scott's 2015 Molecular Genetics/BIOL375 students:

[1]NCBI BLAST at http://www.ncbi.nlm.nih.gov/blast

[2]KEGG at http://www.kegg.jp/kegg/kegg2.html

[3]ExPASy   at   http://www.expasy.ch/enzyme/enzyme-search-ec.html

[4]TMHMM at http://www.cbs.dtu.dk/services/TMHMM

[5]LipoP at http://www.cbs.dtu.dk/services/LipoP/

[6]SignalP at http://www.cbs.dtu.dk/services/SignalP

[7]Phobius at http://phobius.sbc.su.se

[8]PSORTb at http://www.psort.org/psortb

[9]CDD at
http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml

[10]TIGRfam at http://tigrblast.tigr.org/web-hmm

[11]PFam at http://pfam.xfam.org/search

[12]PDB at http://www.rcsb.org/pdb/home/home.do

Notes for the Instructor

As a final step in the annotation process, instructors have the discretion of making a student's lab notebook publically available.

The following websites contain instructor resources to aid in the development of a personalized annotation project:

- GENI at GENI-Science.org provides access to detailed descriptions of both gene annotation and functional genomics projects.
- GENI-ACT at GENI-ACT.org provides detailed instructions in the use of selected bioinformatics programs and an online lab notebook to be used in microbial gene annotation projects. The various bioinformatics programs used on GENI-ACT are described on this companion site http://www.geni-act.org/education/main/.

- Microbial Genome Annotation Network (MGAN), an NSF RCN-UBE-funded project, at http://mgan-network.org/index.html provides support in the use of the GENI-ACT platform. The MGAN Networking website at http://mgan-network.org/networking.html contains links to a variety of microbial genome analysis projects and instructor resources, including a GENI-ACT lab manual, videos, and sample projects

Literature Cited

Department of Energy's Joint Genome Institute (DOE JGI). 1997-2015. Phylogenetic Diversity http://jgi.doe.gov/our-science/science-programs/microbial-genomics/phylogenetic-diversity/ (accessed 1/23/2016)

Ditty, J.L., K.M. Williams, M. Keller, G.Y. Chen, X. Liu, and R.E. Parales. 2013. Integrating Grant-funded Research into the Undergraduate Biology Curriculum Using IMG-ACT. Biochemistry and Molecular Biology Education, 41: 16–23.

Guiding Investigation Though Novel Investigation (GENI-Science). 2015. http://www.geni-science.org (accessed 1/23/2016)

Guiding Investigation Though Novel Investigation – Annotation Collaboration Toolkit (GENI-ACT). 2015. http://www.geni-act.org (accessed 1/23/2016

Microbial Genome Annotation Network (MGAN). 2009. http://mgan-network.org/index.html (accessed 1/23/2016)

Scott, L. 2015. Under "Available Projects" at Guiding Investigation though novel investigation (GENI). 2015. http://www.geni-science.org (accessed 1/23/2016)

Scott, L. and T. Nickel. 2009. Meiothermus ruber genome analysis project. https://sites.google.com/a/augustana.edu/meiothermus_ruber/home (accessed 1/23/2016)

Tindall, B., J. Sikorski, S. Lucas, E. Goltsman, A. Copeland, T. Glavina Del Rio, M. Nolan, H. Tice, J. Cheng, C. Han, S. Pitluck, K. Liolios, N. Ivanova, K. Mavromatis, G. Ovchinnikova, A. Pati, R. Fähnrich, L. Goodwin, A. Chen, K. Palaniappan, M. Land, L. Hauser, Y. Chang, C. Jeffries, M. Rohde, M. Göker, T. Woyke, J. Bristow, J. Eisen, V. Markowitz, P. Hugenholtz, N. Kyrpides, H. Klenk, A. Lapidus. 2010. Standards in Genomic Science. August 30; 3(1): 26–36. Published online 2010 July 29. doi: 10.4056/sigs.1032748

U.S. Department of Health and Human Services. Public Health Service Centers for Disease Control and Prevention National Institutes of Health. HHS Publication No. (CDC) 21-1112. Revised December 2009. http://www.cdc.gov/biosafety/publications/bmbl5/index.htm (accessed 1/23/2016)

Wu, D., P. Hugenholtz, K. Mavromatis, R. Pukall, E. Dalin, N. Ivanova, V. Kunin, L. Goodwin, M. Wu, B. Tindall, S. Hooper, A. Pati, A. Lykidis, S. Spring, I. Anderson, P. D'haeseleer, A. Zemla, M. Singer, A. Lapidus, M. Nolan, A. Copeland, C. Han, F. Chen, J. Cheng, S. Lucas, C. Kerfeld, E. Lang, S. Gronow, P. Chain, D. Bruce, E. Rubin, N. Kyrpides, H. Klenk, J. Eisen. (2009). A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature*, *462*(7276), 1056–1060. http://doi.org/10.1038/nature08656.

About the Authors

Lori Scott received her MS and PhD from Illinois State University (Normal, IL) in the biological sciences. She currently holds the position of Professor of Biology at Augustana College (Rock Island, IL). She is the PI of an NSF RCN-UBE award (title: Microbial Genome Annotation Network) and an NSF S-STEM award (title: Promoting Diversity and Community College Transfer Success in the Biology and Biochemistry Curriculum). In addition, she is a contributor to an NSF LSAMP and an NSF TUES awards. The theme underlying these awards is promoting authentic undergraduate research. Her teaching interests are in cell biology and genetics. She is particularly interested developing a program that integrates undergraduate research in microbial genome analysis throughout the undergraduate curriculum.

Mission, Review Process & Disclaimer

The Association for Biology Laboratory Education (ABLE) was founded in 1979 to promote information exchange among university and college educators actively concerned with teaching biology in a laboratory setting. The focus of ABLE is to improve the undergraduate biology laboratory experience by promoting the development and dissemination of interesting, innovative, and reliable laboratory exercises. For more information about ABLE, please visit **http://www.ableweb.org/.**

Papers published in *Tested Studies for Laboratory Teaching: Peer-Reviewed Proceedings of the Conference of the Association for Biology Laboratory Education* are evaluated and selected by a committee prior to presentation at the conference, peer-reviewed by participants at the conference, and edited by members of the ABLE Editorial Board.

Citing This Article

Scott, L. R., T. C. Nickle, S. Koury, B. Goodner, K. Houmiel, B. McFarland, J. Tenlen, A. Lumpe, K. Murphy, D. Rhoads, D. Wood, and D. Chen. 2016. Creating a Microbial Genome Analysis Project Using Publically Available Genome Sequences and Online Bioinformatics Programs. Article 21 in *Tested Studies for Laboratory Teaching,* Volume 37 (K. McMahon, Editor). Proceedings of the 37th Conference of the Association for Biology Laboratory Education (ABLE), http://www.ableweb.org/volumes/vol-37/?art=21