ABI Sustaining: The National Center for Genome Analysis Support 2019 Annual Report

Thomas G. Doak https://orcid.org/0000-0001-5487-553X
Craig A. Stewart https://orcid.org/0000-0003-2423-9019
Robert Henschel https://orcid.org/0000-0003-2289-9398
Matthew Hahn https://orcid.org/0000-0002-5731-8808
Yuzhen Ye https://orcid.org/0000-0003-3707-3185

Indiana University

PTI Technical Report PTI-TR19-003

September 10, 2019

Please cite as:

Doak, T. G., Stewart, C. A., Henschel, R., Hahn, M., Yuzhen, Y. (2019). ABI Sustaining: The National Center for Genome Analysis Support 2019 Annual Report (PTI Technical Report). Indiana University, Bloomington, IN. Retrieved from http://hdl.handle.net/2022/25133.

PERVASIVE TECHNOLOGY INSTITUTE
INDIANA UNIVERSITY

NATIONAL CENTER FOR GENOME ANALYSIS SUPPORT
INDIANA UNIVERSITY

Table of Contents

Table of Tables

Table of Figures

Executive Summary

The major goal of the NSF ABI Sustaining Award remains to support the continuing and expanding activities of the National Center for Genome Analysis (NCGAS).

1) Providing excellent bioinformatics consulting services to all NSF-funded researchers in need.
2) Maintaining, supporting, and delivering genome assembly and analysis software on national cyberinfrastructure (CI) systems.
3) Providing education and outreach programs on genome analysis and assembly, including designing genomics experiments, using best-of-breed software and hardware tools, and interpreting data.
4) Disseminating tools for genome assembly and analysis in forms useable by biologists.
5) Providing long-term archival storage for genome biologists.

NCGAS emphasizes genome, transcriptome, and microbiome assembly at the technically challenging end of the spectrum of current bioinformatics—for example *de novo* assembly—where both specialized computational resources and applications are needed. NCGAS has just been awarded its second three-year Sustaining award, DBI-1759906, which started in September 2018 (PIs Doak, Henschel, Stewart, Hahn, Ye). Pittsburgh Supercomputing Center (PSC) and PI Blood are again on a collaborative award. The current award has just started a year of no-cost extension.

Accomplishments:
NCGAS serves researchers in all 50 states and Puerto Rico, including 17 EPSCoR states, and has users around the world. NCGAS supported 250 biologists using genome analysis (50 named new allocations) during the current year, including 266 short consultations and 38 extended consultations.

Highlights detailed within the report include:

- The maintenance of available suites of software and addition of new ones as new methods of sequencing and analysis become available
- Use of the Jetstream cloud environment to disseminate applications
- Outreach, direct consultation, and training activities
- Very positive results from the 2018 user survey
- Professional growth via professional development opportunities among NCGAS personnel
- Increased publicity through national presentations, a website, and several social media venues
- Providing assistance for field stations

Products: The report enumerates the titles of 9 peer-reviewed papers, 11 presentations and reports, 23 blog posts, and 5 press releases produced within the specified reporting period.

Participants: This section lists the individuals who have worked on the project, and the contributions of partner organizations and collaborators.

Impacts: NCGAS's impact on its primary disciplines – biology and genomics – have been disseminated to communities of interest through a variety of means, including the products detailed in Section 2.

Changes/Problems: The only issue raised in this section is the degree to which NCGAS is limited by its low personnel numbers, especially with regard to outreach, education, and training efforts.

1. Accomplishments

1.1. *What are the major goals of the project?*

The major goal of the NSF ABI Sustaining Award remains to support the continuing and expanding activities of the National Center for Genome Analysis (NCGAS), including:

6) Providing excellent bioinformatics consulting services to all NSF-funded researchers in need.
7) Maintaining, supporting, and delivering genome assembly and analysis software on national cyberinfrastructure (CI) systems.
8) Providing education and outreach programs on genome analysis and assembly, including designing genomics experiments, using best-of-breed software and hardware tools, and interpreting data.
9) Disseminating tools for genome assembly and analysis in forms useable by biologists.
10) Providing long-term archival storage for genome biologists.

Emphasis is placed on genome, transcriptome, and microbiome assembly at the technically challenging end of the spectrum of current bioinformatics—for example *de novo* assembly—where both specialized computational resources and applications are needed.

NCGAS has been awarded its second three-year Sustaining Award, DBI-1759906, which started in Sept. 2018 (PIs Doak, Henschel, Stewart, Hahn, Ye). Pittsburgh Supercomputing Center (PSC) and PI Blood are again on a collaborative award.

1.2. *What was accomplished under these goals?*

1.2.1. Major Activities

NCGAS is a collaboration between the lead institution, Indiana University (IU), and the Pittsburgh Supercomputing Center (PSC) at Carnegie Mellon University. At IU, NCGAS is a Center of the Indiana University Pervasive Technology Institute (PTI) and a management unit in Research Technologies (RT). This placement allows NCGAS ready access to significant High Performance Computing (HPC) facilities, specialized cyberinfrastructure expertise, and administrative support. Likewise, collaborator institution PSC provides extensive HPC resources and supporting services as a national HPC center. During the first year of this sustaining award (a fourth year of no-cost extension has just started on the last award), NCGAS has continued to use NSF funding—with additional funding and facilities from IU and PSC—to aid discovery and innovation in biological sciences that use genomic methods. Through collaborative efforts with PSC, NCGAS has aided work ranging from ecology, to physiology, to economically important animals and plants.

NCGAS serves researchers in all 50 states and Puerto Rico (Fig. 1), including 17 EPSCoR states/1 territory, by providing Galaxy, GenePattern, and HPC systems for the community to use in their analyses.

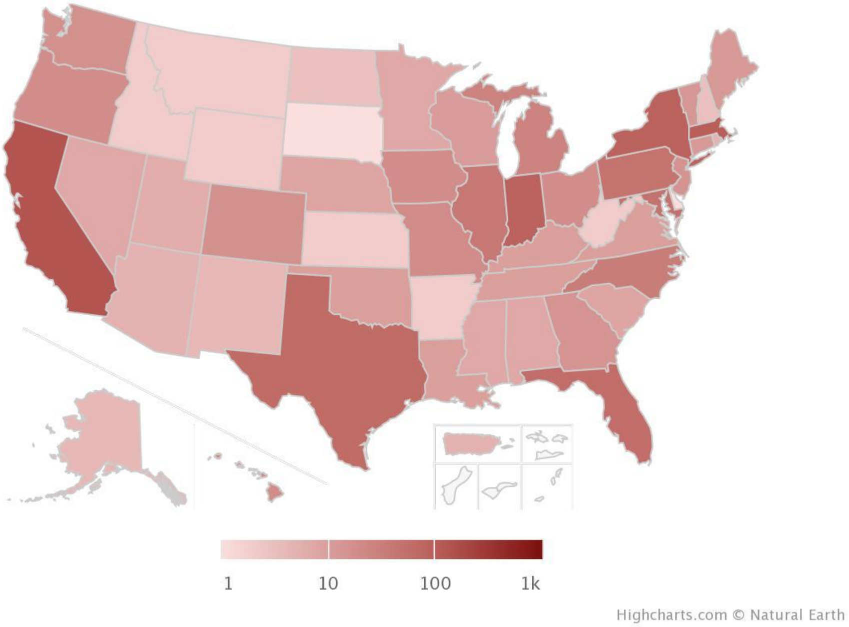

Figure 1. States and Territories with NCGAS Users, all services

1.2.1.1. Provide excellent bioinformatics consulting services.

NCGAS' most significant accomplishments continue to be those of the researchers assisted: discoveries from RNA and DNA transcriptome, metagenome and metatranscriptome assemblies.

As illustration, this past year NCGAS-aided researchers conducted work that included the following organisms (completed and on-going):

- Responses to modern diet in human populations - GWAS
- Insights into Cyclophyllidea (tapeworms) evolution studying *N. mogurndae* (fish parasites)-genome/transcriptome analysis
- Tetrahymena mating type recognition using RNAseq data
- Killer whale metatranscriptome analysis to study how toxicant and nutritional stress may impact killer whales
- Effect of fecal and cecal microbiome in murine models with chronic kidney disease-mineral and bone disorder
- Impact of hurricane Maria on the mesophotic reefs of Southwest Puerto Rico
- Transcriptomic assessment of the tree swallows in the Great Lakes Area
- Daphnia genomes (population genomics and *de novo* assemblies), incl. *D. galeata*
- Diatoms transcriptomes
- Bacteriophage
- The diverse microbial clade Stramenopila + Alveolata + Rhizaria (SAR) (*de novo* genomes, transcriptomes)
- Butterfly morphology (Vanessa cardui, Spodoptera, Zerene cesonia)
- Purple sea urchin (*Strongylocentrotus purpuratus*)
- Ciliates such as *Oxytricha trifallax*
- Grapes (*Vitis* species RNAseq)
- Interplay of honeybee mites (*Varroa jacobsoni*) and deformed wing virus
- Freshwater crustaceans (*Asellus aquaticus*)
- Aquatic ferns (*Ceratopteris* sp.)
- Deep sea crustaceans

- Largemouth bass
- Drain flies and black soldier fly (basal diptera)

NCGAS supported 372 biologists using genome analysis (39 current year, including 266 short consultations and 38 extended consultations. Details regarding many of the extended consultations are provided in an Appendix 2.

1.2.1.2. Maintain, support, and deliver genome assembly and analysis software on national CI systems

NCGAS continues to help NSF researchers navigate successful genomics work, including: understanding current-, next-, and third-generation sequencing methods; identifying applications appropriate to the data; finding and using HPC resources capable of the analysis; and finally, interpretation of the results (including whether they make sense). In this supply chain, some researchers only need access to large memory clusters, which we can provide in a number of ways. Others need instruction in basic HPC use and genomic analysis, or in how to design experiments, before ever collecting samples. We continue to attract new users from a number of sources, often by word of mouth, but also importantly from workshops we teach. We serve the important function of keeping our users current, primarily by attending international meetings (Plant and Animal Genomes (PAG) is the most important, in our estimation, for the diversity of genomes addressed). This year we maintained and updated packages for long read technologies assembly and analysis. Hybrid assemblies are quickly becoming popular, and we continue to support Canu, MaSuRCA and PacBio's SMRT analysis software. New scaffolding methods such as Hi-C serve the functions that BAC libraries used to fill, and once again enable reference-quality assemblies. PSC also makes many of these packages available on their systems, with a focus on metagenome assembly and analysis (see PSC report)—Hi-C can serve very interesting functions in metagenomics.

NCGAS at IU provides accounts to multiple HPC resources:

- The new *Carbonate* IU cluster (Mason's replacement)
- The IU/TACC *Jetstream* cloud environment, via XSEDE allocations
- PSC's *Bridges* large memory cluster and *Pylon* storage facility
- Additional XSEDE resources, including the IU/TACC *Wrangler* storage system, through NCGAS XSEDE Community and Teaching Allocations

NCGAS also provides bioinformatics software through web gateways (menu-driven):

- NCGAS Galaxy web portal: providing access to the widely used Galaxy workflow system, loaded with genomic-specific applications
- Trinity RNAseq Galaxy portal: running on IU's Carbonate cluster
- GenePattern gateway: providing access to more genomic applications specific to biomedical research, runs on IU's Carbonate

We support public and private genome browsers for several commercial tropical crops, daphnia, junco, Tetrahymena, etc.

Overall, we have installed 43 new software packages and maintain a total of 430 (not including versions) across the systems described above (see Appendix 3) in the past year.

1.2.1.3. Disseminate tools for genome assembly and analysis

Jetstream VMs: NCGAS has added 19 (15 public images and 4 private images) preconfigured virtual machines to the Jetstream library to help researchers with their visualization (genome browsers, Anvi'o, MEGAN, QIIME2), for training (R for Biologists, genome assembly workshop, and Intro to Unix), setting up private images for a research groups, and for developing workflows. These VMs are regularly updated by NCGAS team, and we have several blogposts to help our users access Jetstream resources and use them effectively (Jetstream blogposts).

Workshop images are launched by only one or two NCGAS team members, and user accounts are added to these VMs to giving the participants access. For this reason, even though these images are launched a greater number of times than the other images, they likely have only one or two unique users (in the case of Mining SRA to identify datasets, only one unique user). In cases where there more than two unique users (in the case of NCGAS_16_04 RStudio for NCGAS image, has four unique users), some participants request NCGAS allocation to continue using the instance for their own datasets. Similarly, workshop images remain public and maintained since users may use these images for the bioinformatics tools that were installed, for instance Arkansas Workshop image was taught in Fall 2017 and Docker-Singualrity-Conda image which was used for a hackathon in July 2018, continue to be used to date.

The visualization images such as genome browsers were set up on an NCGAS allocation, but available publicly for any Jetstream user to setup their own genome browser. Currently NCGAS is also helping support genome browser images on other allocations apart from NCGAS, supporting one base browser, and two other genome browsers for other research groups. NCGAS hosts additional services, such as a web server with wiki and Drupal Base—base images that can be used to host websites (public/private) to support data access, collaboration, among several other use cases. QIIME2, Anvi'o and MEGAN are all commonly used metagenomics toolkits with visualization and the option for a menu driven web front end, and cannot be currently supported on most traditional HPC infrastructure.

The two images, mining SRA to identify similar datasets and harvesting field station data are two images that include workflows developed by NCGAS to help support the research community. The workflows were developed with the help of REU students and made public with documentation.

Table 1. List of all the NCGAS preconfigured virtual machines available in the Jetstream library.

Jetstream image name	**Category**	**About the image**	**Number of instances launched**	**unique users**
NCGAS 16_04 Rstudio for NCGAS	Workshop	The goal of the workshop is to help biologists get acquainted with R, which will in turn help them with their analysis.	131	4
genome assembly workshop	Workshop	This image was set up for teaching a workshop on genome assembly using PacBio and 10X reads.	60	2
Anvio	Visualization	Open source community driven analysis and visualization platform for 'omics data\	42	5
Mining SRA to identify datasets	Workshop/ Developing workflow	The goal of this workshop is to help researchers working with, or interested in working with, SRA to be able to run their bioinformatics workflows efficiently using computational resources available through NCGAS/XSEDE.	27	1

NCGAS pop up genome browser	Visualization	Hosts ta LAMP stack, Tripal (build on Drupal 7), JBrowse, BLAST, and custom tools for quickly starting a genome browser	19	3
QIIME 2 toolkit	Visualization	Extensible and decentralized microbiome analysis package with a focus on data and analysis transparency.	10	3
Docker-Singularity-conda	Workshop	This image was set up to teach PEARC18 Hackathon-Developing and Applying Best Practices Protocols, and contains docker, singularity and conda installation	9	5
Harvesting Field Station Data	Developing workflow	This Jetstream image contains a workflow that makes allows collecting atmospheric sensor data from Raspberry Pi to a Drupal website for the community.	6	2
Web server with wiki	Visualization	Hosts the LAMP (LINUX, Apache, MySQL, PhP) stack to make web service easier	3	2
Arkansas Workshop	Workshop	The focus of this workshop was to apply bioinformatic analysis in Jetstream environment	3	2
NCGAS_18_04 RStudio for NCGAS	Workshop	Updated image - The goal of the workshop is to help biologists get acquainted with R, which will in turn help them with their analysis.	2	1
MEGAN community edition v6	Visualization	Analyzing large metagenomics datasets	1	1
Drupal Base	Visualization	Content management software used to make many of the websites and applications to use everyday	1	1
PonyLinux	Workshop	This Jetstream image is to train beginners to command line with a little game for learning UNIX	1	1
bcbio-nextgen toolkit v.2.0		Validated, scalable, community developed variant calling, RNA, and small RNA analysis	0	0
Tripal and JBrowse on Ubuntu	Visualization	Private image		
Junco Browser	Visualization	Private genome-browser for a research group		
Ubuntu 16_04 RStudio OpenRefine	Workshop	Private image - with software installed for a research group		

EML workshop	Workshop	Private image - The image was prepared for use in the Environmental Data Initiative's hands-on workshop “Creating EML with R and publishing data packages in the EDI repository” in 2017.		

Docker and Singularity containers: NCGAS has worked to take advantage of the growing popularity and functionality of containerization on Linux systems. Docker is a popular tool for insulating software and operating system environments, but it requires root permissions to run on systems. This makes it an ill fit for a shared HPC environment, but this is not an issue on cloud resources such as Jetstream. Singularity takes advantage of Docker’s widespread support but runs in a much less risky way and is compatible with HPC systems. Singularity has been tested and used by NCGAS for the GenePattern server, in order to ensure a reliable, consistent environment no matter where the tools are running. Docker has been supported by NCGAS as a means of setting up and disseminating the Trinity Galaxy server

Jupyter Notebook: Jupyter Notebook platform has also gaining a lot of popularity with features to easily share workflows, rerun specific steps of the workflow as necessary, help with knowledge transfer and reproducibility. NCGAS is working towards incorporating this service as well, starting with presenting a birds of a feather session at PEARC19 to hear the community’s feedback on this service. In general, the session saw positive feedback for the use of Jupyter, with some caveats about the software’s flaws for things like version control of notebooks and security. Many saw it as an entry into line command use, and not as an endpoint in itself. In the last year, GenePattern server has been updated to support Jupyter Notebooks as a additional feature. GenePattern users, can now use an interactive Notebook interface to develop and share reproducible workflows with the community.

GitHub repository: NCGAS has a GitHub account with 14 repositories, of which 12 have been active in this reporting period. NCGAS continues to develop scripts and pipelines with documentation which are shared with the community through GitHub. The active projects include the developed *de novo* transcriptome pipeline, an REU project on searching the SRA, three Galaxy development repositories, and four workshop repositories with script for NCGAS taught workshops. One repository called Useful scripts include a set of scripts that are commonly used among the NCGAS team—useful for the community.

Long-term archival storage: NCGAS and IU provide access to IUScholarWorks, a digital repository provided by the IU Libraries for showcasing and preserving research findings, and the Scholarly Data Archive (SDA) (~42 PB of tape) for storing and accessing research data. SDA also has a Globus endpoint allowing users to transfer terabytes of their data quickly to and from the SDA, and the IU/XSEDE clusters.

In an effort to extend more tools to our users, we maintain XSEDE allocations - one for extending our services to include biological research stations (valued at $25,157.77), another for supporting NCGAS users to use Jetstream and Bridges for their research (valued at $111,675.00), and one for serving our workshops and any supported workshops via Jetstream (valued at $33,380.00). Thus far, these allocations have served our field station outreach activities via REU tool development and our workshops, as well as external workshops.

Education and outreach programs on genome analysis and assembly, including designing genomics experiments, using best-of-breed tools, and interpreting data (see 1.3 and 5.2 include:

- NCGAS conducted a two RNAseq analysis to a national group of participants (see below).

- NCGAS had four presentations at Plant and Animal Genome Conference (PAG), showcasing NCGAS resources, pipelines and research. (Ref – NCGAS blogpost on PAG)
- Carrie Ganote and Bhavya Papudeshi participated in the Virus Discovery Hackathon at San Diego State University in January, developing a bioinformatics tool to search for virus families in SRA.
- NCGAS taught a one-day workshop at American Society for Microbiology (ASM) on Using High-Performance Computing (HPC) to mine the NCBI Short Read Archive (SRA).
- NCGAS created materials for a high-level introduction to HPC and Unix for Biologists. This was debuted in the IU Bloomington and IUPUI Indianapolis campuses. More advanced courses are being planned.
- NCGAS in the last year attended nine conferences –Plant and Animal Genome conference (PAG), American Association for Cancer Research (AACR), American Society for Microbiology (ASM) and Practice and Experience in Advanced Research Computing (PEARC), Galaxy Community Conference (GCC), Evolution meeting, Organization of Biological Field Stations, the Joint Meeting of Ichthylogists and Herpetologists and the ITCR 2019 Annual Meeting. NCGAS presented four posters, two presentations, and one workshop at these meetings, for a total reach of 304 participants.

1.2.2. Specific Objectives

There are currently 430 packages supported by NCGAS (between Carbonate, Karst, and Bridges), with 19 Jetstream images (Appendix 3).

1.2.3. Significant results

NCGAS provides online help, consulting, and tutorials related to genome analysis.

Key highlights of NCGAS support include:

- In the current year, NCGAS completed 266 short term consulting engagements (those taking less than 4 hours of staff time to resolve) and 38 long term consulting engagements (taking more than 4 hours of staff time to resolve). Many long-term consultations are research collaborations that last for months or years, with NCGAS staff playing a critical role in discoveries by the scientists receiving NCGAS help.
- NCGAS completed tutorials, training, and outreach activities attended by hundreds of participants (see 1.3).
- The past year included the Jetstream cloud's second year of operation. NCGAS has worked closely with the Jetstream team to allow biologists to use Jetstream in the following ways: 1) blogposts; 2); preconfigured bioinformatics virtual machines, especially pop-up genome browsers; 3) presentations at conferences: PAG, OBFS; 4) using Jetstream VMs for workshops: R workshops, and Arkansas workshop.

Consultation is provided by telephone, teleconference, email (a ticketing system tracks requests), and in person. Consulting hours are 8:00 a.m. to 5:00 p.m. on week-days, but support often extends beyond local business hours and includes weekends.

In the last year NCGAS joined 15 new long-term projects (Appendix 2).

In our continued partnership with PSC (see 1.5) and PI Philip Blood we have 1) coordinated software suites, 2) increased use of Bridges when very large memory nodes are needed, 3) initiated a shared Luster file system (DC-WAN) allowing increased interoperability, and 4) built a center for metagenomic analysis centered at PSC.

We synthesized our experience in *de novo* RNAseq assembly—gained in particular through our collaborations with Keithanne Mockaitis on large plant transcriptome projects—and our years of experience helping users—to develop a two-day intensive workshop in *de novo* transcriptome assembly. The course starts with an introduction to the HPC environment, and then presents our in-house pipeline for *de novo* assembly, which uses four different assemblers and compares their results to validate transcripts. This course was offered in Spring 2018, Fall of 2018, and again in Spring of 2019, serving a total of 78 participants (see Appendix 1).

A significant activity was our participation in four ABI submissions as collaborators or co-PIs: two innovation and two development proposals, all four with a metagenomics focus. These proposals were all declined, although three garnered strong reviews. While not awarded, these activities furthered NCGAS' relationship with a number of metagenomic labs, indicating the presence and respect NCGAS has in the community, which we expect will lead to future collaborations, independent of funding.

1759871: Collaborative Research: ABI Development: Accessible resources for at-scale multi-omic informatics. Lead institution University of Minnesota, PI Tim Griffin

1759875: Collaborative Research: ABI Development: Extending the National Center for Genome Analysis Support to Foster Reproducible, Portable, Community-validated Metagenomics Analysis. Lead institution University of Maryland, College Park, Mihai Pop.

1759764: ABI Innovation: Computational methods for integrative analyses of multi-omics data in microbiome research. Lead institution Indiana University, PI Yuzhen Ye.

1759851: Collaborative Research: ABI Innovation: Needles in Haystacks: Enabling Scientists to Search the Sequence Read Archive. Lead institution San Diego State University, PI Rob Edwards.

1.2.4. Key outcomes or other achievements

The key outcome during the first year of this sustaining award is the continued success of NCGAS in delivering effective consulting services focused on accelerating the research of biologists and bioinformaticians, and so, accelerated biological discoveries in the US NSF-funded community. NCGAS provides a robust "supply chain" from NSF-funded and other supercomputers, through specialist applications and knowledge, to bench and field scientists across the country. NCGAS' ongoing efforts have assisted 9peer-reviewed scientific publications published in 2018-2019 (beyond those reported for the first two years).

1.2.5. Results of the 2019 user survey

Our fourth user survey was conducted during June 2019, with 240 of 1601 users responding. We again find broad satisfaction with our services, and that NCGAS had been essential to many users' research. If there are complaints, they focus on areas where we are limited by personnel—a limitation we are aware of—and are mostly suggested by comments. NCGAS is used primarily by NSF-funded researchers, although exceptions are made for projects that would be considered fundable by the NSF for our target directorate, and short one-off email consultations that are not expected to take as much time as even a short consultation. A white paper on our survey results is available at http://hdl.handle.net/2022/22401and will be submitted as an Interim Report.

1.3. What opportunities for training and professional development has the project provided?

Dr. Thomas Doak is now an Sn. Scientist at Indiana University, PI and manager of this NSF sustaining award, and PI on the current NCGAS sustaining submission (in review). Extensions of the NCGAS brand have allowed him to be a PI on two collaborative NIH NCI ITCR (Information Technologies in Cancer

Research) grants (see Synergistic activities). At IU, NCGAS is a Center within the Pervasive Technologies Institute, and Dr. Doak is playing a more active role in the Institute.

Staff member Carrie Ganote is continuing her PhD program in bioinformatics while in the employ of NCGAS—she is an original NCGAS member. She is a player in the international Galaxy community and was on the 2016 Galaxy conference organizing committee, hosted at IU; is on the Galaxy Funding Committee; and served on the Scientific Board for the 2018 and 2019 meetings. Ms. Ganote oversees many projects, mentored Sheri Sanders, and continues to mentor Bhavya Papudeshi.

Staff member Sheri Sanders has been a NCGAS team member for three years, starting after finishing her PhD at the University of Notre Dame, where she used transcriptomics to characterize transcriptomic response to disease in a polyploid salamander complex. Since then, Dr. Sanders has worked on polyploid crop species, amphibian population research, and insect genome assembly/annotation. Dr. Sanders' goal in joining NCGAS was to grow her understanding of IT and HPC and how they impacted the biological community as well as to continue teaching and writing. She has enrolled in IU's Online Graduate Certificate Program for Data Science and will be taking machine learning courses in the Fall Semester, which will help increase NCGAS's GPU support. She will also be taking IoT courses to further expand on her efforts working with automating frog call identification for field stations (see section 1.5.2). NCGAS has also provided her with the opportunity to further her development of training materials, which she showed an aptitude for in graduate school. She has successfully run several national workshops, developing skills in orchestration of events, curriculum development, and reporting. Finally, she has worked with science editors to write lay articles for the community and contributes heavily to the NCGAS Blog, for which she was invited to speak and mentor at ComSciCon 2019 in Chicago.

Staff member Bhavya Papudeshi has been an NCGAS employee for two years, since completing her master's degree in bioinformatics at SDSU working in the lab of Elizabeth Dinsdale on marine metagenomics. Ms. Papudeshi's work includes optimization of metagenome assembly and binning tools to reconstruct population genomes; she strengthens our metagenomics support and our collaboration with PSC. She is working with our collaborator Rob Edwards at SDSU on mining SRA data; Marla and Michael Douglas (U. of Arkansas) on a Cyclophyllidea tapeworm transcriptome (of an invasive fish), and with Marcella Cervantes to identify genes that are related to the mating type identification in Tetrahymena. In the last year, Papudeshi has continues to improve her understanding of High Performance Clusters: job schedulers (TORQUE, SLURM), maintaining gateways (galaxy and GenePattern), setup preconfigured VMs on cloud computing resource, to better help the NCGAS userbase. She just started the PhD program in Bioinformatics at Indiana University and has taught at workshops and mentored REU students in the last year.

1.4. How have the results been disseminated to communities of interest?

1.4.1. Sources of Contact

Results have been disseminated to communities of interest in a variety of ways, including:

- NCGAS website at ncgas.org
- NCGAS Blog
- NCGAS Twitter at @ncgasiu
- NCGAS Facebook
- IU_PTI YouTube NCGAS Channel
- Public code repository at github.com/ncgas
- Articles in the lay press, in both IU Research Highlights and ScienceNode
- Birds-of-a-feather sessions at technical conferences

- Displays and booths at national and international technical conferences
- In-person contacts
- Email list distribution

We have made significant strides toward increasing our outreach and communication, with success. Our website has seen almost three times the traffic, we have more than doubled our twitter followers, and almost doubled our Facebook following.

1.4.2. Results of Contact

Our website (ncgas.org) is our main online content engine. Between Sept 1, 2018 and Aug 12, 2019, we had 60,742 page vies (up from 22,920 in 2018) with 26,265 unique users (up from 10,398 in 2018). Social Media drove 383 views from 196 users (0.8%). Of that 235 (61.36%) were from Twitter and 123 (32.11%) were from Facebook. Other sources of traffic included 3,108 (12.8%) users from direct entry, 696 (2.9%) users from referrals, and 20,304 (83.5%) users from organic search (largely for conda installations).

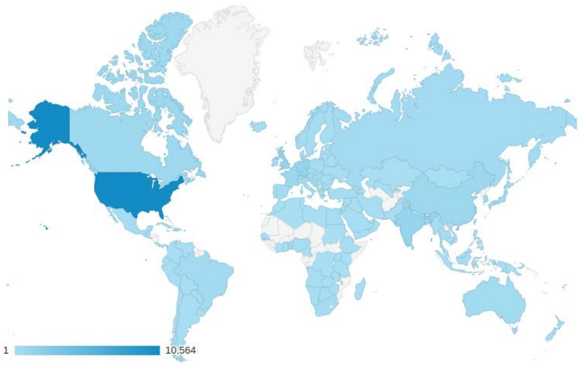

Figure 2. World sources of hits to the NCGAS web page.

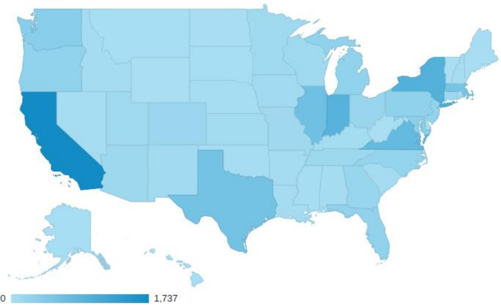

Figure 3. US sources of hits to the NCGAS web page.

NCGAS Blog: While our Blog only generated 5% of our page views in 2018, it is now generating the majority of our website traffic. We generated 21 blogs between September 1, 2018 and August 12, 2019 for a total of 27,339 (45.17% of traffic to the site) views, 24,680 (66.58%) unique views, and accounted for 22,963 (77.37%) entrances to our website. Traffic was pretty consistent to the blog (see Figure X).

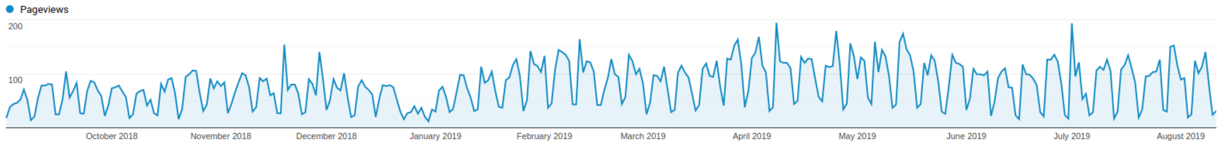

Figure 4. Pageviews for the NCGAS Blog from September 1, 2018 through August 12, 2019.

Below is a list of the unique users per blog entry generated this year:

- 7 Long read technologies potential to study the microbial world
- 17 Sharing data to help debug errors
- 45 Installing Software - R Packages
- 43 Installing perl modules locally
- 322 Third Generation Sequencing Update
- 51 Installing Software: Library errors
- 121 Reconstructing genomes from metagenomes
- 77 Installing Software: Makefiles and the make command
- 177 Visualizing KEGG Pathways
- 65 Summary of our favorite PAG 2019 talks
- 65 Software install- How to save an environment
- 85 Software Installation
- 104 NCGAS Spring 2019 Workshops
- 36 NCGAS Returns from PAG Victorious
- 161 Recent Tegu genome paper serves as a great primer for non-model genome assembly
- 86 Join NCGAS at PAG XXVII!
- 247 How EviGene Works
- 86 Setting up SSH keys
- 66 Research Desktop (RED) for new users
- 33 Manipulating Tar Files
- 81 Debugging job errors

We were also invited to write the DNA Day Blog by GigaBlog after they saw our Tegu Genome coverage. (DNA Day 2019: How to sequence the genomes of the weird and wonderful http://gigasciencejournal.com/blog/dna-day-2019-how-to-sequence-non-model-organism/). This blog generated 19 new users in 46 sessions, with an average of 4.43 pages viewed before leaving our site (total of 203 page views on our site).

There are several older that are still in the top 10 pages of our site:
• Installing Conda Packages Locally - 18,935 views, #1 page on our site. Of the 3.76% of reported search terms leading to our site, "install conda", "conda install package", "conda install local package", "conda install" were all in the top ten search terms.
• Tree making - 1,467 views, #6 page on our site
• Third Generation Sequencing - 1,378 views, #8 on our site

Twitter/Facebook: NCGAS has a twitter page with 201 followers (up from 79 in 2018) and a total of 391 tweets in total as of August 15, 2019. NCGAs uses the twitter account to reach out to a wider community with educational information (from NCGAS blog, research articles), workshop/internship opportunities, NCGAS attended conferences, cluster/software updates, and NCGAS outreach highlights. Since September 2019 to August 15, 2019, NCGAS posted 127 tweets that reached a total of 90,118 twitter profiles (709 profiles per tweet in average) of which 1,251 accounts (10 accounts per tweet in average) engaged (retweets, follows, replies, favorites, clicks, etc) with the tweets posted, leading to 127 new followers. From the twitter analytics page, our organic audience on twitter is mostly from United States

(46%), followed by United Kingdom (7%), Germany (6%), France (3%), India (3%), Canada (3%), Japan (2%), Indonesia (2%), Spain (2%), and Australia (1%), which corresponds with our user base (Fig NCGAS world map).

Overall, the activity on twitter remains consistent throughout the year, with relatively higher activity when the team is at a conference –November SC18, January PAG 2019, April AACR 2019, June ASM 2019, and July PEARC 2019 conference. Similarly, the peaks in the graphs is when we are advertising and teaching workshops the below figure– September R for Biologists, October 2018 Fall de novo transcriptome assembly workshop, February 2019 advertisement for upcoming workshops, April 2019 de novo transcriptome assembly workshop, and two blogposts –January blog on Tegu genome, March on reconstructing genomes from metagenomes, April DNA Day blog on Third generation sequencing update and GigaBlog.

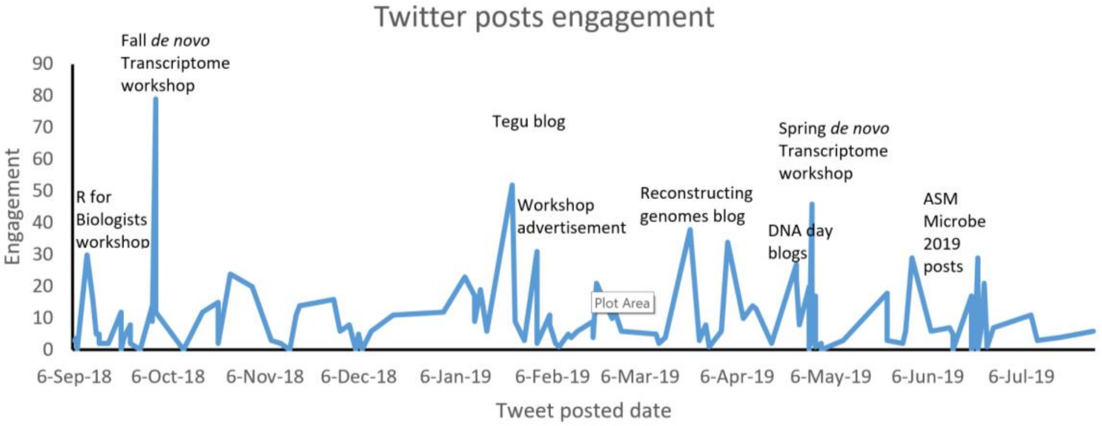

Figure 5. NCGAS twitter posts from September 2018 to August 15, 2019 with their engagement rate.

Facebook: NCGAS Facebook page currently has 73 followers (up from 41 in 2018) with 67 total likes as of August 15th, 2019. Most posts from Twitter are posted on Facebook with the same goal to reach out to a wider community. Since September 2018 to August15th, 2019 we had 87 posts that reached 3764 accounts (43.3 accounts per post in average), of which 417 accounts (5 accounts per post in average) interacted with the post (clicked the links, comment, or liked the post). This activity led to 32 new followers and 27 new likes in the last reporting year. NCGAS Facebook posts mostly reached to United States (303 accounts), mostly in Bloomington, IN (151), Indianapolis, IN (18).

Facebook posts showed a similar trend as Twitter posts, with more activity during conferences and workshops. However, the IU research highlights and ScienceNode articles gained more page views and reach on Facebook compared to Twitter which could be due to the articles containing more information about IU students.

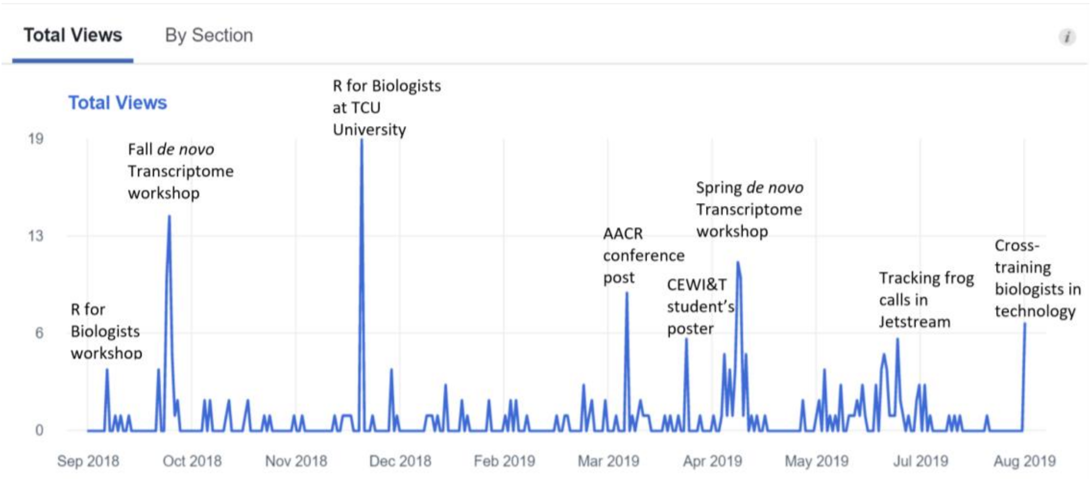

Figure 6. Facebook NCGAS page views from September 2018 to August 2019.

YouTube: Two new playlists added this reporting period:

- Playlist: 2018-2019 de Novo Assembly of Transcriptomes - 13 videos, Published 8/7/2019, 71 view total (all videos) as of 8/12/2019
- Playlist: 2018-2019 Intro to R for Biologists - 25 videos: Published 7/16/2019, 102 view total (all videos) as of 8/12/2019

Git: 10 Repositories (7 active during reporting period), 6 Contributors, and 327 Commits were reported by August 12, 2018.

IU Research Highlights:

Tracking frog calls in Jetstream (Jul 20, 2019)
Student work shines at Research Services Expo (May 14, 2019)
How frogs are cracking the evolutionary code (April 29, 2019)
IU Genome center helps shed light on insect evolution (Feb 14, 2019)
Slipping into the Jetstream (Sept 5 2018)

ScienceNode Articles:

Cross-training biologists in technology (Aug 11 2019)

BOF:

During the Practice and Experience in Advanced Research Computing (PEARC, July 28-Aug 1, 2019) conference, NCGAS organized and led a birds-of-a-feather session to discuss the challenges faced by researchers, educators, and administrators when deploying and using Jupyter Notebooks on HPC systems. The topics discussed ranged from shortfalls of Jupyter version control to security implications and scale.

Other EOT: NCGAS in the last year attended eight conferences –Plant and Animal Genome conference (PAG), American Association for Cancer Research (AACR), American Society for Microbiology (ASM) and Practice and Experience in Advanced Research Computing (PEARC), Galaxy Community Conference (GCC), Evolution meeting, and ITCR 2019 Annual Meeting. NCGAS presented four posters, two presentations, and one workshop at these meetings, which was attended by close to 304 participants. During the Galaxy Community Conference, NCGAS members participated in a 'code-fest' activity that resulted in a successful code product incorporated into the Galaxy code base.

1.5. What do you plan to do during the next reporting period to accomplish the goals? Goals for the next year

Accomplishments of the last year that we will carry forward include:

1) Offering genome browsers for searchers to organize and distribute their results. We are working with the Lynch and Mangone groups at ASU to develop effective ways of presenting large population data sets.
2) Pursuing metagenomic projects/researchers in collaboration at PSC –Taught a one-day workshop at ASM Microbe 2019 conference, and will be teaching a workshop on Metagenomic analysis in October 2019. NCGAS is also working on updating the current transcriptomics pipeline for metatranscriptome datasets. Finally, we have also added three metagenomics analysis tools (QIIME2, Anvi'o, and MEGAN) to Jetstream to help visualize metagenomics datasets.
3) Continuing our use of Jetstream to aid genomics researchers and field stations.

NCGAS was inaugurated with the IU purchase of the large-memory cluster Mason:

1) Mason has been replaced with the Carbonate cluster, which is considerably faster, with a higher memory-to-core ratio.
2) NCGAS takes advantage of the NSF-funded PSC cluster Bridges (see PSC annual report). Bridges is already in use for metagenomic assemblies through PSC and NCGAS uses an XSEDE allocation to enable users to utilize Bridges.
3) The NSF-funded cloud environment Jetstream, while not providing large memory, is being used for genomics science, esp. during workshops. We are using Wrangler to provide storage for Jetstream VMs.

Other goals for the year:

- We recently held a successful RNAseq workshop three times, and have a Metagenomics workshop planned for this fall. We also hope to take both of these on the road at least a couple of times in the next year, such as to the American Society of Microbiologists (ASM) annual meeting again. We have also developed a successful R workshop for biologists, which will be offered twice this year. We next plan to develop a bash/python/R workshop for genomicists (Appendix 1). We will continue to record these presentations for posting to YouTube, but the hands-on support during the NCGAS workshops are one of its strengths.
- The IU/TACC Jetstream cloud has been immediately useful in staging workshops, in that it allows instructors to provide each student with a provisioned VM, eliminating the need for students to install software on their own machine (frequently a disaster). We hope to participate in more such efforts.
- We now actively provide genome browsers, starting with the GMOD-base G and JBrowsers, and are working with the Tripal development group. We hope to have more time to work in this community, especially in the development of appropriate tools for population genomic data sets.

Planned travel:

- We just attended PEARC 2019 and will have two members at SC18. This aligns with our increased emphasis on explicitly teaching HPC to biologists.
- The entire NCGAS team will attend and present at PAG 2020.
- At least one member will attend the AACR national conference.
- Sanders will again teach at the MDI Biological Laboratory's Environmental Genomics course; given their interest in metagenomics, Papudeshi may also attend as an instructor.

Proposed grant submissions:

- We will resubmit a version of Collaborative Research: ABI Development: Accessible resources for at-scale multi-omic informatics. Lead institution University of Minnesota, PI Tim Griffin.
- We will resubmit a version of Collaborative Research: ABI Innovation: Needles in Haystacks: Enabling Scientists to Search the Sequence Read Archive. Lead institution San Diego State University, PI Rob Edwards.
- We hope to be included on the GenePattern ITCR renewal.
- We hope to be included on a USDA proposal with Keithanne Mockaitis.

1.5.1. Pursuing Field and Marine Stations as NCGAS and Jetstream clients

In the last two years, we worked to engage field stations and field biologists. This is motivated by our focus on biology, and our relationship with the Jetstream team. The first year we conducted a survey of field station directors to determine what they felt their cyberinfrastructure needs were. In short, they saw a need for help with data management. This was reiterated in conversations we had while attending the 2017 OBFS yearly meeting. We conducted a follow-up survey this year, but it will not be complete until early September 2019.

Highlights for the work this year:

- We have completed two more rounds of REU work (total of three rounds, 6 students, and 15 months of work) on a flagship field station project, using raspberry pi sensors to collect data, aggregate it on Jetstream-based webservers, and then use machine learning to identify frog calls. All materials are or will be available on Jetstream (later AWS and Google Cloud), and are fully documented. This endeavor has seen much interest from the community and has generated severl press releases from NCGAS (See section 1.4.2).
- We are holding a one-day satellite workshop to the 2019 OBFS "Determining the data and cyberinfrastructure needs of modern FSMLs", a reworking of our plan for the 2018 workshop. This workshop is designed to identify open problems in field station data management and cyberinfrastructure, connect groups with working solutions with those currently facing similar problems, and provide information on available resources.
- We are presenting three posters at the same 2019 OBFS annual meeting. We will be distributing our 2019 OBFS survey results at the annual meeting and making them available via IUScholarWorks.

1.5.2. 1.5.3 Galaxy gateway

NCGAS continues to support two galaxy instances, IU Galaxy instance for the local community (IU and IU affiliate users), and another instance for international users on Trinity Galaxy. Trinity Galaxy was previously funded ITCR ….

While Trinity Galaxy was initially developed for cancer research, they are also useful for other disciplines. Trinity in particular is extensively used by our non-medical client, especially where obtaining a genome assembly is not feasible, either because the genomes are too large, or the project would be too expensive for smaller labs working on non-model organisms. Currently, IU Trinity Galaxy has 979 registered users (62 countries) (Fig 5).

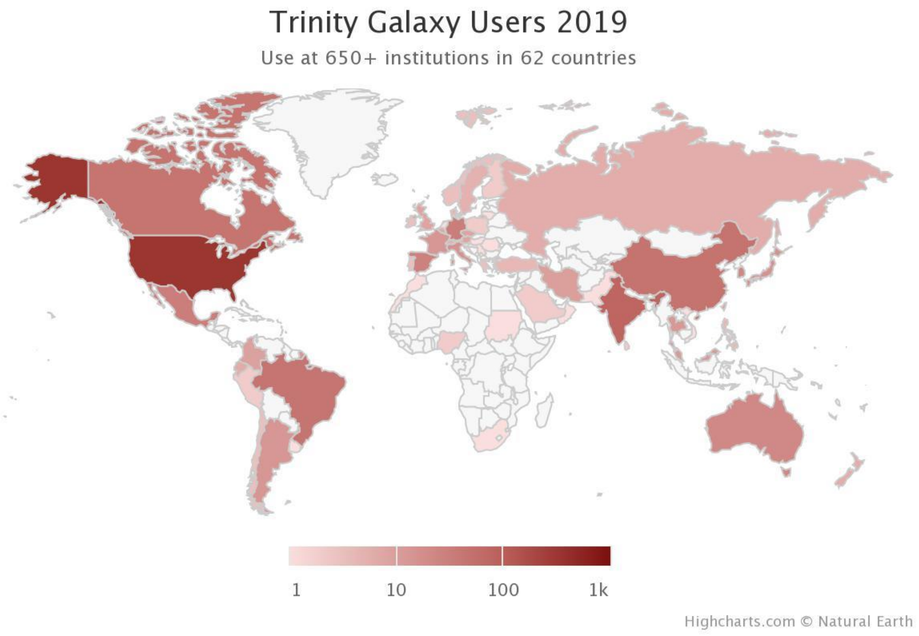
Trinity Galaxy Users 2019
Use at 650+ institutions in 62 countries
1
10
100
1k
Highcharts.com © Natural Earth

Figure 7. Users of the NCGAS-supported Trinity Galaxy instance worldwide

1.5.3. Synergistic activities

NCGAS is an NSF-funded service provider, but also a management group in IU's Pervasive Technology Institute. In this second role, NCGAS personnel participate in projects which strengthen our position as a genomics center. Here are some examples of their activities:

- Active engagement with the Generic Model Organism Database (GMOD) community. As we invest effort in genome browsers we also to play a role in browser development. Sheri Sanders leads this effort. We provide several browsers for Keithanne Mockaitis in support of her international collaborations in the genomics of commercial tropical plants (coffee, cacao, sweet potatoes, etc.). We maintain a Daphnia browser for Michael Lynch and his collaborators, and a junco browser for Ellen Ketterson and her collaborators. We also work with professor Naomi Stover at Bradley University, who maintains several ciliate (protist) browsers.
- NCGAS is a Domain Champion in the Campus Champion program and a level 2 XSEDE service provider. Tom Doak is currently the SP board representative for level 2 providers.
- NCGAS is a collaborator on NIH Information Technology in Cancer Research (ITCR) funded project, hosting the GenePattern genomics gateway along with Broad Institute. While this gateway was developed for cancer research, the gateway is useful for other disciplines as well. GenePattern is used mostly by medical clients to run genomic analysis focusing on gene expression, single nucleotide polymorphism, flow cytometry analysis. Currently, IU GenePattern has 747 registered users (42 countries) (Fig. 6) giving NCGAS a global reach.

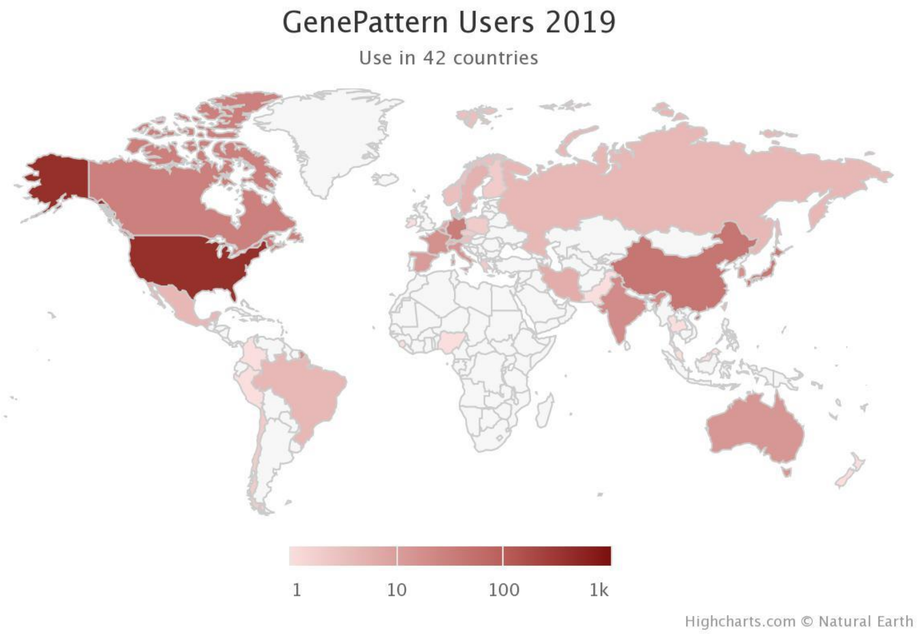

Figure 8. Users of the NCGAS-supported IU GenePattern instance worldwide.

We receive IU base funding to serve the IU genomics community. The NCGAS management group maintains all genomics software on IU clusters and responds to trouble tickets concerning this software and genomics issues. In the last reporting year, NCGAS has participated in the following IU related activities:

- R for Biologists workshop was taught on Nov 2018 (hybrid online/IU, 110), at IU/IUPUI/online with a total of 110 participants.
- NCGAS participated in three REU programs, CEW&T REU, Jim Holland Summer Research Program, and Jetstream REU program training a total of 14 students in the last year,
  - CEW&T REU program- NCGAS sponsored three REU students in biology related majors over the last year, training these students in computing and working towards developing two workflows, bioacoustics recording devices and frog call analysis, and mining the really large Sequence Reach Archive (SRA) to identify other datasets that contain similar genomes. The students presented their workflow at the annual CEW&T poster session at Indiana University. Eliza Foran who worked on bioacoustics recording devices and frog call analysis also presented her work at Research Services Expo and won the best poster award for Best visualization and was accepted into the Summer Jetstream REU program to continue working on this project. The other two students Haley Leffler and Sruthi Gananpaneni's abstract on the workflow to mine SRA database was accepted for poster presentation at ASM Microbe 2019 in San Francisco. The students also won travel awards up to $2,500 from three travel grants to attend the conference, the remaining expenses was supported through NCGAS funds.
  - Jim Holland Summer Research Program- This is a one-week research program offered to high school students. Through this program we had two who worked with NCGAS on improving the previously developed mining the really large Sequence Reach Archive (SRA) to identify other datasets that contain similar genomes project. At the end of the week, the students presented their research at a poster presentation.
  - Summer Jetstream REU program where nine students were participated and worked on three machine learning based projects. NCGAS pitched project on automatic recognition of frog calls which was an extension to the previous developed CEW&T student developed

workflow created automatic recording devices and bioacoustics analysis. In this project the REUs worked on extending this workflow with machine learning algorithms to identify frog species from frog calls. The results from this project was presented at PEARC 19 conference in Chicago.

- We participate in local projects, such as Catalina Fernandez's project on the changes in the genome of indigenous South Americans in recent times as a possible adaptation to modern diets.

This mix of activities provides a diversified funding base to the NCGAS management group and expands our reach in the genomics field.

1.5.4. We have requested a no cost extension for the coming year, and an ABI Sustaining proposal (our second) has just been awarded (1759906).

2. Products

2.1. Products resulting from this project during the specified reporting period

2.1.1. Journals or juried conference papers

1. Johri, P., Marinov, K. G., Doak. G.T., Lynch, M. (2019) 'Population Genetics of Paramecium Mitochondrial Genomes: Recombination, Mutation Spectrum, and Efficacy of Selection.', *Genome biology and evolution*, 11(5), pp. 1398–1416. doi: 10.1093/gbe/evz081.
2. Smythe, A. B., Holovachov, O. and Kocot, K. M. (2019) 'Improved phylogenomic sampling of free-living nematodes enhances resolution of higher-level nematode phylogeny', *BMC Evolutionary Biology*, 19(121). doi: 10.1186/s12862-019-1444-x.
3. Bui, L. T. and Ragsdale, E. J. (2019) 'Multiple plasticity regulators reveal targets specifying an induced predatory form in nematodes', *Molecular Biology and Evolution*, msz171. doi: 10.1093/molbev/msz171/5541796.
4. Choudhary, S. Thakur, S., Jaitak, V., Bhardwaj, P. (2019) 'Gene and metabolite profiling reveals flowering and survival strategies in Himalayan Rhododendron arboreum', *Gene*, 690, pp. 1–10. doi: 10.1016/j.gene.2018.12.035.
5. Wurch, L. L., Alexander, H., Frischkorn, K. R., Haley, S. T., Gobler, C. J., Dyhrman, S. T. (2019) 'Transcriptional Shifts Highlight the Role of Nutrients in Harmful Brown Tide Dynamics', *Frontiers in Microbiology*. Frontiers, 10, p. 136. doi: 10.3389/fmicb.2019.00136.
6. Choudhary, S., Thakur, S., Jaitak, V., Bhardwaj, P. (2019) 'Gene and metabolite profiling reveals flowering and survival strategies in Himalayan Rhododendron arboreum', *Gene*. Elsevier B.V., 690, pp. 1–10. doi: 10.1016/j.gene.2018.12.035.

7. Winker, K., Glenn, T., Withrow, J., Sealy, S., Faircloth, B. (2019) 'Speciation despite gene flow in two owls (Aegolius ssp.): Evidence from 2,517 ultraconserved element loci', *The Auk: Ornithological Advances*, 136(2), p. ukz012. doi: 10.1093/auk/ukz012.
8. Villemur, R., Payette, G., Geoffroy, V., Mauffrey, F., Martineau, C. (2019) 'Dynamics of a methanol-fed marine denitrifying biofilm: 2-impact of environmental changes on the microbial community', *Peer J- Life and Enviornment*. doi: 10.7717/peerj.7467.
9. Hennon, G. M., Morris, J. J., Haley, S.T., Zinser, E. R., Durrant, A. R., Entwistle, E., Dokland, T., Dyhrman, S. T. (2018) 'The impact of elevated CO2 on Prochlorococcus and microbial interactions with "helper" bacterium Alteromonas', *The ISME Journal*. Nature Publishing Group, 12(2), pp. 520–531. doi: 10.1038/ismej.2017.189

2.1.2. Licenses

2.1.3. Other Conference Presentations / Papers/Reports

1. Papudeshi, B., Sanders, S., Ganote, C., Doak, G.T. (2019) 'Compute Resources Available to the Research Community for Microbiome Analysis', in Plant and Animal Genome Conference 2019. Available at https://plan.core-apps.com/pag_2019/abstract/eb8d2b76e25358daf4c927eba46581f9 (Accessed: 28 January 2019).
2. Sanders, S., Papudeshi, B., Ganote, C., Doak, G.T. (2019) *NCGAS Makes Robust Transcriptome Assembly even easier with added features to an accessible de no transcriptome assembly workflow*, in Plant and Animal Genome Conference 2019. Available at: https://plan.core-apps.com/pag_2019/abstract/6e070c0ebde6a5b908f9f08fb2eecd05 (Accessed: 28 January 2019).
3. Foran, E. G.; Suggs, E. D.; Underwood, T. A.; Snapp-Childs, W. ; Sanders, S. A. (2019) 'Automatic recognition of frog calls'. doi: 10.5967/PS4S-D421.
4. Leffler, H., Ganapaneni, S., Papudeshi, B., Sanders, S., Doak, G. T. (2019) *Mining the Sequence Read Archive to identify crAssphage, a ubiquitous inhabitant of the human microbiome, in dog and pig samples*.
5. Ganapaneni, S., Leffler, H., Papudeshi, B., Sanders, S., Doak, G. T. (2019) *Coupling metagenomics with high-performance computing to mine the Sequence Read Archive (SRA) to analyze Pseudomonas phage PAK-P1*.
6. Cai, J. X. Weathers, J. G. Leffler, H., Ganapaneni, S., Papudeshi, B., Sanders, S., Doak, T. G. (2019) 'Navigating the Sequence Read Archive to identify crAssphage, an ubiquitous inhabitant of the human microbiome', in. Bloomington, Indiana: im Holland Summer Science Research Program Poster Session.
7. Foran, E., Anderson, J., Slayton, T., Guido, E., Doak, T., Sanders, S. (2019) 'Developing a workflow for bioacoustic recording devices and frog call analysis within Jetstream'. Center of Excellence for Women & Technology.
8. Papudeshi, B., Ganote, C., Sander, S., Doak, T.G. (2019) 'National cyberinfrastructure and bioinformatic analysis support available to the cancer research community', in *Bioinformatics, Convergence Science, and Systems Biology*. American Association for Cancer Research, pp. 5109–5109. doi: 10.1158/1538-7445.sabcs18-5109.
9. Sanders, S., Papudeshi, B., Ganote, C., Doak, G.T. Mansfield, C., Tseng, C. Y., Custer, T., Custer, C., Matson, C. (2019) 'Population Genetics of Tree Swallows, in Collaboration with NCGAS'. Plant and Animal Genome XXVII.
10. Papudeshi, B., Sander, S., Ganote, C., Doak, T., Chafin, T., Reshetnikov, A., Sokolov, S., Pummil, J., Douglas, M., Douglas, M. (2019) 'The Genome of Fish Tapeworm Nippotaenia percotti as a Potential Bookmark for Gene Loci that Facilitates Anthropogenic Infection.' Plant and Animal Genome XXVII.
11. Leffler, H., Ganapaneni, S., Papudeshi, B., Sanders, S., Doak, G. T. (2019) 'A workflow to identify genomes in the Sequence Read Archive for phylogenomic analysis', in. San Franscisco: ASM Microbe 2019.

2.1.4. Other Products

2.1.4.1. *Blog Posts*

1. Debugging job errors. Available at: https://ncgas.org/Blog_Posts/Job%20errors.php. (Accessed: September 14 2018)
2. Manipulating Tar Files. Available at: https://ncgas.org/Blog_Posts/Manipulating%20Tar%20Files.php. (Accessed: October 22 2018)
3. Research Desktop (RED) for new users. Available at: https://ncgas.org/Blog_Posts/Research%20Desktop.php. (Accessed: November 7 2018)

4. Setting up SSH keys. Available at: https://ncgas.org/Blog_Posts/Generating%20ssh-keys.php. (Accessed: November 30 2018)
5. How EviGene Works. Available at: https://ncgas.org/Blog_Posts/EviGene.php. (Accessed: December 17 2018)
6. Join NCGAS at PAG XXVII. Available at: https://ncgas.org/Blog_Posts/PAG2019%20Abstracts.php. (Accessed: January 2 2019)
7. Recent tegu genome paper serves as a great primer for non-model genome assembly. Available at: https://ncgas.org/Blog_Posts/Tegu_Genome.php. (Accessed: January 8 2019)
8. NCGAS returns from PAG victorious. Available at: https://ncgas.org/Blog_Posts/PAG_Results.php. (Accessed: January 24 2019)
9. NCGAS Spring 2019 workshops. Available at: https://ncgas.org/Blog_Posts/NCGAS%20Spring%202019%20Workshops.php . (Accessed: January 31 2019)
10. Software installation. Available at: https://ncgas.org/Blog_Posts/Software%20Installation.php. (Accessed: February 5 2019)
11. Software install- How to save an environment. Available at: https://ncgas.org/Blog_Posts/Understanding%20environment%20variables.php. (Accessed: February 8 2019)
12. Summary of favorite PAG 2019 talks. Available at: https://ncgas.org/Blog_Posts/PAG2019%20favorite%20talks.php. (Accessed: February 13 2019)
13. Visualizing KEGG pathways. Available at: https://ncgas.org/Blog_Posts/Ghost%20Koala.php. (Accessed: February 21 2019)
14. Installing software- Makefiles and the make command. Available at: https://ncgas.org/Blog_Posts/makefiles%20and%20make%20command.php. (Accessed: March 11 2019)
15. Reconstructing genomes from metagenomes. Available at: https://ncgas.org/Blog_Posts/Reconstructing%20genomes.php (Accessed: March 22 2019)
16. Installing Software- Library errors. Available at: https://ncgas.org/Blog_Posts/Library%20errors.php (Accessed: April 8 2019)
17. Third generation sequencing update. Available at https://ncgas.org/Blog_Posts/Third%20Generation%20Sequencing%20Update.php (Accessed: April 25 2019)
18. Installing perl modules locally. Available at: https://ncgas.org/Blog_Posts/perl%20installs.php (Accessed: May 30 2019)
19. Installing software- R packages. Available at https://ncgas.org/Blog_Posts/Installing%20Programs%20-%20R.php (Accessed: June 7 2019)
20. Sharing data to help debug errors. Available at: https://ncgas.org/Blog_Posts/Sharing%20data.php (Accessed: June 27 2019)
21. Long read technologies potential to study the microbial world. Available at https://ncgas.org/Blog_Posts/Long-reads%20for%20microbes.php (Accessed: July 31 2019)

2.1.4.2. *Press Releases*

1. *IU Genome center helps shed light on insect evolution* (2019) *IT news and Events*. Available at: https://itnews.iu.edu/articles/2019/IU-genome-center-helps-shed-light-on-insect-evolution-.php.
2. *Tracking frog calls in the Jetstream* (2019) *IT News and Events*. Available at: https://itnews.iu.edu/articles/2019/Tracking-frog-calls-in-the-Jetstream-.php.
3. *Student work shines at Research Services Expo* (2019) *IT News and Events*. Available at: https://itnews.iu.edu/articles/2019/Student-work-shines-at-Research-Services-Expo-.php.
4. *How frogs are cracking the evolutionary code* (2019) *IT news and Events2*. Available at: https://itnews.iu.edu/articles/2019/How-frogs-are-cracking-the-evolutionary-code-.php.

5. *Cross-training biologists in technology* (2019) *ScienceNode*. Available at: https://sciencenode.org/feature/Cross-training biologists in technology.php.

3. Participants

3.1. Individuals

Table 2. Individuals who have worked on the NCGAS project.

Name	Most Senior Project Role	Nearest Person Month Worked
Doak, Thomas	PhD/PI	8
Stewart, Craig	PhD/co-PI	1
Henschel, Robert	PhD, director, management group	1
Yuzhen Ye	PhD/co-PI	
Blood, Phillip	PhD/co-PI (collaborative grant)	1
Miller, Therese	Other Professional	3
Nalagampalli Papudeshi, Bhavya	Staff Scientist, PhD candidate	1
Ganote, Carrie	Staff Scientist, PhD candidate	6
Sanders, Sheri	Staff Scientist, PhD	6
		6

3.1.1. Full details of individuals who have worked on the project

Thomas Doak
Email: tdoak@iu.edu
Most Senior Project Role: PI (doctoral level)
Nearest Person Month Worked: 8
Contribution to the Project: PI and operational management
Funding Support: NSF, NIH, Indiana University
International Collaboration: Yes: Italy, Germany, Japan
International Travel: No

Craig A. Stewart
Email: stewart@iu.edu
Most Senior Project Role: PhD/co-PI
Nearest Person Month Worked: 1
Contribution to the Project: Co-PI responsible for oversight and outreach to new groups to generate users/projects for NCGAS services
Funding Support: Indiana University

International Collaboration: No
International Travel: Yes: Germany - 0 years, 0 months, 7 days

Yuzhen Ye Scott Michaels
Email: yye@indiana.eduu
Most Senior Project Role: PhD/co-PI
Nearest Person Month Worked: 1
Contribution to the Project: Funded PSC collaborator
Funding Support: Co-PI responsible technical oversight
International Collaboration: No
International Travel: No

Phillip Blood
Email: blood@psc.edu
Most Senior Project Role: Other Professional
Nearest Person Month Worked: 4
Contribution to the Project: Carnegie Mellon University and University of Pittsburgh, collaborative grant
Funding Support: NSF, Carnegie Mellon University
International Collaboration: Yes
International Travel: Yes

Robert Henschel
Email: henschel@iu.edu
Most Senior Project Role: Co-PI
Nearest Person Month Worked: 1
Contribution to the Project: software optimization
Funding Support: Indiana University, NIH
International Collaboration: Yes
International Travel: Yes

Therese Miller
Email: millertm@iu.edu
Most Senior Project Role: Other Professional
Nearest Person Month Worked: 1
Contribution to the Project: Director over NCGAS management group; financial and reporting management
Funding Support: Indiana University, NSF
International Collaboration: No
International Travel: No

Carrie Ganote
Email: cgannot@iu.edu
Most Senior Project Role: Staff Scientist (doctoral level)
Nearest Person Month Worked: 6

Contribution to the Project: bioinformatician consultant / programmer
Funding Support: NSF, NIH
International Collaboration: No
International Travel: Yes

Bhavya Nalagampalli Papudeshi
Email: bhnala@iu.edu
Most Senior Project Role: Staff Scientist (Masters level)
Nearest Person Month Worked: 6
Contribution to the Project: bioinformatician consultant / programmer
Funding Support: NSF
International Collaboration: No
International Travel: NO

Sheri Sanders
Email: ss93@iu.edu
Most Senior Project Role: Staff Scientist (doctoral level)
Nearest Person Month Worked: 6
Contribution to the Project: bioinformatician consultant / programmer
Funding Support: NSF
International Collaboration: No
International Travel: No

3.2. *Partner organizations*

3.2.1. *Full details of partner organizations*

Table 3. Partner organizations

Name	Type of Partner Organization	Location
Pittsburgh Supercomputing Center, Carnegie Mellon University	Academic Institution	Pittsburgh, PA
Texas Advanced Computing Center, University of Texas	Academic Institution	Austin, TX
XSEDE	Other Nonprofits	United States
Arkansas High Performance Computing Center, University of Arkansas, Fayetteville	Academic Institution	Fayetteville, AR

3.2.1.1. Pittsburgh Supercomputing Center, Carnegie Mellon University

Partner's Contribution to the Project

- Directly supports NCGAS activities through Collaborative Award
- In-Kind Support
- Facilities
- Collaborative Research
- Personnel Exchanges

More Detail on Partner and Contribution: PSC is a funded collaborator on the NCGAS sustaining award. Philip Blood, PI of the NCGAS collaborative award at PSC, manages NCGAS genomics support activities at PSC, installs and maintains NCGAS software on PSC systems, coordinates NCGAS activities with those of XSEDE, and works with genomics researchers to enable large scale sequence assembly and analysis on PSC systems. In addition, PSC has provided facilities, computer time, and storage space on Bridges in support of NCGAS activities and in support of biological researchers who use NCGAS services. Staff of PSC have made resources available at their site to NCGAS staff. Some of the support provided by this institution has been in-kind, and this institution has engaged in collaborative research on genome analysis software, particularly as regards use of Galaxy and software that requires the large shared memory architecture of PSC supercomputers. PSC also participates in the education, outreach, and dissemination efforts of NCGAS.

3.2.1.2. Texas Advanced Computing Center, University of Texas

Partner's Contribution to the Project

- Collaborative research
- Facilities

More Detail on Partner and Contribution: TACC is an awardee on the Jetstream and Wrangler grants, as well as the CyVerse center, which provides many opportunities for collaboration. It has provided facilities, computer time, and storage space in support of NGCAS activities and in support of biological researchers who have used NCGAS services. Staff of this institution have also engaged use of NCGAS staff and facilities and made available resources at their site to NCGAS staff.

3.2.1.3. XSEDE

Partner's Contribution to the Project

- Collaborative research
- Facilities

More Detail on Partner and Contribution: Staff of the NSF-funded XSEDE project have engaged use of NCGAS staff and facilities and have made resources available at their site to NCGAS staff. Some of the support provided by XSEDE has been provided in-kind, and this institution has engaged in collaborative research on genome analysis software. XSEDE has played a particularly strong role in education, outreach, and dissemination efforts of NCGAS. NCGAS is a Level 2 XSEDE Service Providers and an XSEDE Domain Champion.

3.2.1.4. Arkansas High Performance Computing Center, University of Arkansas, Fayetteville

Partner's Contribution to the Project

- Collaborative Research

More Detail on Partner and Contribution: Jeff Pummill, Director of the Arkansas High Performance Computing Center, has worked with NCGAS to facilitate UofA researchers using genomic tools on Jetstream and Mason. Pummill is an XSEDE Campus Champion and has aided in XSEDE Jetstream allocations. Pummill and biology faculty invited Jetstream and NCGAS to give a workshop at U of A in fall 2017, which was very successful. NCGAS continues to work on collaborative projects with Pummill and U of A researchers.

3.3. Have other collaborators or contacts been involved?

No

4. Impact

4.1. What is the impact on the development of the principal discipline(s) of the project?

Results have been disseminated to communities of interest in a variety of ways, including:

- Publications in scientific journals
- Presentations
- Birds of a feather sessions at technical conferences
- Displays and booths at national and international technical conferences
- Articles in the lay press, most notably in Science Node (formerly International Science Grid This Week) at https://sciencenode.org
- NCGAS web site at ncgas.org
- In-person contacts
- Email list distribution
- Newsletter
- Blog, Facebook and Twitter posts

4.2. What is the impact on other disciplines?

The primary discipline on which NCGAS has had an impact—beyond biology and genomics—is computational science and the use of cyberinfrastructure. NCGAS serves as a model for a "domain-specific scientific service center," independent of federally-funded cyberinfrastructure; we have decoupled federal funding for supercomputers and funding for supercomputer application support. This ensures that a community relatively new to supercomputers—biology, for example—has support funded by the BIO directorate of NSF and is attuned to the needs of current research in the field.

NCGAS personnel are all XSEDE Domain Champions, a new category supplementing Campus Champions, who are domain agnostic. As a very recent XSEDE program, we will see how Domain Champions work out, but in essence, NCGAS has always been a Domain Champion for biologists and genomics and serves as a model for how to be a DC.

We also work to distribute software relevant to biological research, improving the nation's ability to use its aggregate cyberinfrastructure resources.

4.3. What is the impact on the development of human resources?

See 1.3.

4.4. What is the impact on physical resources that form infrastructure?

Nothing to report.

4.5. What is the impact on institutional resources that form infrastructure?

The software distributed by NCGAS has improved the effectiveness and ease-of-use of cyberinfrastructure resources throughout the nation.

4.6. What is the impact on information resources that form infrastructure?

NCGAS has facilitated the publication of several data sets important to basic biological research and to the management of important plant and animal stocks. In the future, NCGAS will place a greater emphasis on genome browsers, an important product of 'omic research.

4.7. What is the impact on technology transfer?

The primary impact of NCGAS on technology transfer is in providing a collection of genomics applications easily available to any researcher. In the case of Trinity and GenePattern, NCGAS stands at the interface of developers and users.

4.8. What is the impact on society beyond science and technology?

The societal impact of genomic characterization is gradual but tremendous over time. Even the human genome's impact was muted at first and is still being explored. Understanding the genomes of pine tree, cacao, and mango will allow these important crop plants to be better managed over coming decades. The potential impact of science supported by NCGAS on society through improved management of food supplies and increased understanding of how organisms adapt to global climate change could be of fundamental importance to US and global populations. The speed with which human microbiome characterization has both begun to inform medical decisions (in nearly every field of medicine, including cancer) and swept through popular media is amazing.

5. Changes/Problems

5.1. Changes in approach and reasons for change

As time and personnel permit, we will grow our educational mission. See current and planned offerings Appendix 1.

5.2. Actual or anticipated problems or delays and actions or plans to resolve them

As a fundamentally service-oriented organization, our limitations are nearly always personnel. As our blog offerings continue to expand, and as we capture workshops on YouTube offerings, we hope to help researchers without direct involvement, but this can only go so far. For example, in our first offering of the two day RNAseq workshop, we had all three team members on the floor (plus a nonspecialist), helping students as they got "stuck" and attempting not to leave any of them behind. In the post-workshop survey, participants reported that the "in-person" aspect of the training was important to the experience, and so we will add a fourth instructor for the next iteration of the course.

5.3. Changes that have significant impact on expenditures

Both Doak and Ganote have had significant pay raises in the last 2 year, and a salary increase for Sanders in the last year.

5.4. Significant changes in use or care of human subjects

Nothing to report.

5.5. Significant changes in the use or care of vertebrate animals

Nothing to report.

5.6. Significant changes in the use or care of biohazards

Nothing to report.

6. Appendices

Appendix 1. NCGAS National Workshop Results

6.1.1. *Overview of 2018-2019 Workshops*

Workshops run as originally planned (numbers in parentheses are number of participants):

- de novo transcriptome assembly and analysis - Ran May 2018 (28), Oct 2018 (27), May 2019 (28)
  - replaced with Metagenomic Analysis for Oct 2019.
  - all slides, code, and videos are available on github and YouTube
- HPC for Biologists - Running Aug 2019 (30 x 2 sessions)
  - will be recorded and made publicly available.
  - github and slides will be made available.
  - includes Linux training image for Jetstream Cloud
- R workshop - Ran Nov 2017 (IU, 20), Feb 2018 (IU, 30), Nov 2018 (hybrid online/IU, 110), Nov 2019 (as MOOC, 100)
  - 70-page textbook produced Jan 2019
  - publicly available, free online course version produced March 2019
  - videos available on YouTube and in the online course

Added to plan this year:

- Metagenomic Analysis - Mining the SRA - Jun 2019 (14)
  - Ran through American Society of Microbiology (ASM) Microbe Conference.
  - github of materials and lectures available
  - Metagenomic Analysis - Running Oct 2019 (25)

Delayed:

- PAG - Collaborating on HPC (not accepted, reapplying for 2020)

Removed from plan this year:

- Biology for HPC Managers - postponed for more planning
- MDIBL - workshop was not funded this year, canceled

6.1.2. *Workshop Participation*

We have offered six nationally available workshops since May 2018. These include the de novo transcriptome assembly and analysis workshop (three times), Introduction to R for Biologists (once nationally), Metagenomic Analysis - Mining the SRA (once nationally), and one Metagenomic Analysis (applications complete, to run in Oct 2019). We have added a workshop page (https://ncgas.org/Workshops.php) to our website to help ease registration and finding previous materials.

For these national workshops, we have had 295 applicants for 231 seats (28% over-subscribed). These applicants come from 33 states (including 17 EPSCoR states), 1 US territory, and 8 additional countries. The applicants were at various points in their careers with 3 undergraduates, 15 Masters students, 75 Ph.D. students, 45 post-doctoral fellows, 32 faculty, 32 staff, 35 other, and 57 unreported applying.

6.1.3. *Workshop Attendees:*

We select based on EPSCoR status and merit for the de novo Transcriptome Assembly and Analysis workshop, as we will for the October run of Metagenomic Analysis. The R workshop is first-come-first serve. We have had 157 total attendees for five workshops (This does not include the October run of Metagenomic Analysis as we are currently in the selection process). Participants came from 18 states (including 10 EPSCoR states) and 1 territory, and 3 countries (Fig X). A total of 2 undergraduates, 8

Masters students, 37 PhD students, 21 post-doctoral fellows, 19 faculty, 13 staff, 24 other positions attended these five workshops. This is roughly representative of the application pool, but these numbers will be re-evaluated upon selection of Metagenomic Analysis applicants.

Three participants attended two workshops, and several have been referred by lab mates, advisors, and coworkers, an indication of success.

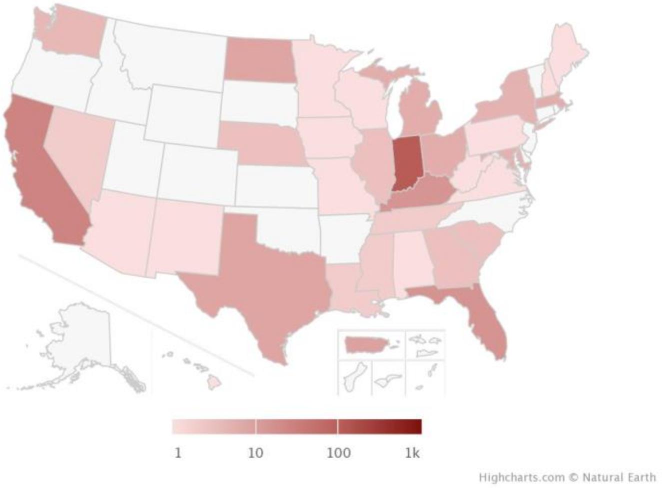

Figure 9. Origin of NCGAS National Workshops applicants in US.

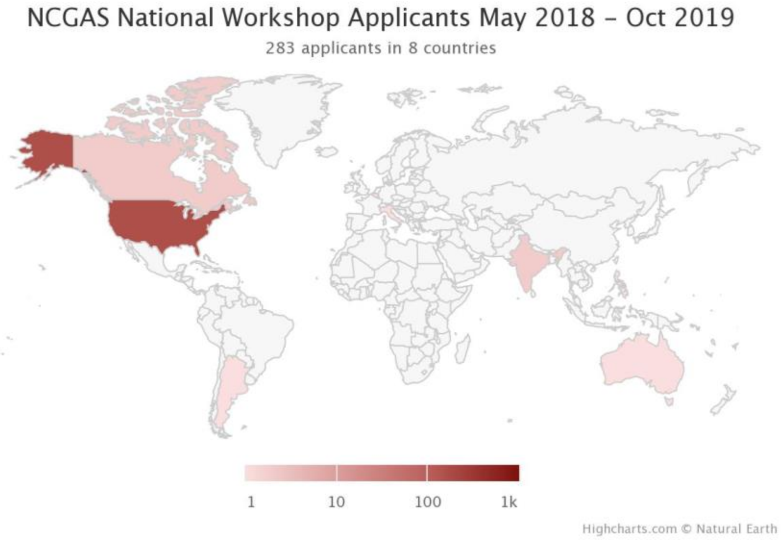
NCGAS National Workshop Applicants May 2018 – Oct 2019
283 applicants in 8 countries
1
10
100
1k
Highcharts.com © Natural Earth

Figure 10. Origin of NCGAS National Workshops applicants in the World.

6.1.4. Workshop Outcomes:

Before and after our national workshops, we conduct a survey which includes a self-reported confidence level in skills taught in the course.

Likert scale of self-reported confidence in skills taught at workshop:

1 No previous experience or knowledge
2 Knowledge of its function, but no hands-on experience
3 Ability to run very limited examples, such as small data sets and tutorials

4 Ability to run more realistic examples, such as real data
5 Ability to troubleshoot tasks for myself and others

For our national run of Introduction to R for Biologists, we taught in a hybrid format. Students took the course either at IUPUI with the main instructor, IUB telecast but with instructors to assist with labs/questions/etc., and fully online with telecast lecture and online office hours/help. The results of 11 skills ranging from "Using RStudio" to "Writing a custom function" are as follows:

Table 4. Self-assessed improvement of 11 skills before and after Introduction to R for Biologists Workshop

	All Participants	IUB (hybrid)	IUPUI (in-person)	Online only
Pre-assessment of skill level	1.91	1.89	1.38	2.05
Post-assessment of skill level	2.99	3.04	2.87	3.02
Improvement	1.08	1.15	1.49	0.97

For the de novo transcriptome workshop, we had participants self-report confidence level on 10 skills, ranging from "Using Unix" to "Downstream Analysis of Transcriptomes" (Table X).

Table 5. Self-assessed improvement of 11 skills before and after de novo Transcriptome Assembly and Analysis Workshop, averaged across first two iterations

	Average of 10 skills self-assessed
Pre-assessment of skill level	2.76
Post assessment of skill level	3.70
Improvement	0.94

A couple IUB participants expressed that they would have preferred the instructor be on site, and the in-person workshop had the largest improvement in R. However, IUPUI also the lowest initial self-evaluation and this trend could be the result of general confidence increase than actual efficacy of the delivery method. The preference becomes muddied further when looking at other survey questions. When asked in surveys, 100% of the transcriptome workshop participants requested in-person workshops. However, a common request on our user survey is to have a webinar or video tutorials. As a result, we have been working to offer a mix of material delivery while still maintaining our ability to scale to the national audience. This effort has included YouTube videos, online office hours, a full online course, in addition to in-person workshops at different locations in the country, often corresponding to conferences our clientele would already be attending.

6.1.5. Plans for Next Year:

HPC Onboarding for Biologists (local only, Aug 2019)
Metagenomic Analysis (Oct 2019)
Introduction to R for Biologist (Nov 2019)
Collaborating on National HPC (Jan 2020)
Introduction to R for Biologists (Feb 2020)
de novo Transcriptome Assembly and Analysis (April 2020)
Metagenomic Analysis – Search SRA (Jun 2020)
HPC Onboarding for Biologists (local only, Aug 2020)

Metagenomic Analysis (Oct 2020)
*Introduction to Python for Biologists (Nov 2020)
*Genome Assembly and Annotation (April 2021)

* indicates workshops in development

Appendix 2. List of NSF Funded Projects Using NCGAS Resources

Petra H. Lenz, University of Hawaii at Manoa

5/22/15

This is a proposal to investigate winter recruitment in the copepod Neocalanus flemingeri in the Gulf of Alaska. Calanid copepods like N. flemingeri are characterized by rapid population growth during the spring coincident with the annual phytoplankton bloom. Recruitment to this spring population is dependent on the successful emergence from diapause followed by reproduction, and survival and growth of this next generation. However, an apparent mismatch between the presence of nauplii (youngest stages) in December/January and the occurrence and unpredictability of the spring phytoplankton bloom have raised questions regarding the timing of female reproduction, and subsequent survival of nauplii. Here, we propose to combine laboratory and field approaches to determine whether female reproduction is synchronized and the strategies for nauplius survival during low food conditions. Gene expression studies using RNA-Seq technology will be used to develop molecular markers for female dormancy and reproductive readiness and for naupliar growth, which in turn will be used to evaluate field collected individuals. As a first step, we will generate a de novo transcriptome for Neocalanus flemingeri.

Charles Delwiche, University of Maryland, College Park

6/1/16

The bulk of eukaryotic diversity is microbial and, when compared to plants, animals and fungi, much of this microbial diversity has been under-sampled from the standpoint of morphological, phylogenetic and genomic data. This skew in data not only has consequences for our understanding of the biodiversity of eukaryotic life on Earth, but also how we interpret cellular and evolutionary biology in the broadest sense. One of the most diverse major clades of eukaryotes is a relatively recently recognized clade that unites the Stramenopila, Alveolata and Rhizaria into the 'SAR' group. Initially this clade was controversial because it forced a re-evaluation of the evolution of several characters, most notably the spread of photosynthesis across eukaryotes. However, additional data have robustly supported SAR as an independent clade. Despite the abundance, economic importance, and diversity of SAR taxa, genomic-scale data are rare and concentrated in only a few areas, Apicomplexa (e.g. malarial parasites), oomycetes (e.g. parasitic water molds) and diatoms (e.g. ecologically important phytoplankton). Here we propose to use a three-tier approach to both increase the number of taxa available for phylogenomics and massively expand the representation of genomic data from SAR taxa. This will include: 1) diversity discovery using targeted environmental sequence surveys coupled with high-throughput FlowCam imaging; 2) high-throughput transcriptomic sequencing; and 3) single cell genomics of unculturable taxa. Organisms used for genomic data will be imaged and both novel and classical images will be added to the Encyclopedia of Life (EOL). Additionally, data will be analyzed using both phylogenomics and a non-phylogenetic similarity network approach to capture the genetic mosiacism of the photosynthetic lineages within SAR.

Seth Bordenstein, Vanderbilt University

7/6/16

The majority of animal species harbor maternally-transmitted bacteria, yet little is known about the genetic and molecular mechanisms that the animal and bacteria use to achieve maternal transmission. The proposed research begins the first forward-genetic investigation of host animal genes that regulate densities and composition of bacterial symbionts. The goal of this project (and need for NCGAS rescources) is to utilize an animal model system coupled with multi-omic technologies for host and bacteria to identify the numbers and types of host genes that regulate symbionts, their additive and epistatic interactions, and their effects on symbiont localization.

Jill Wegrzyn, University of Connecticut

12/6/16

Connecting high quality, curated, phenotypic and genotypic data with geo-location and environmental data will enable fundamental questions in tree biology to be elucidated. Providing access to these integrated datasets and the tools to interrogate them in a fully targeted manner, is best achieved through community databases where the crop curation expertise resides. The usage of standard ontologies, cross-site querying functionality and web-services driven interoperability with other database and resources will expand the utility of data from community databases in an unprecedented way. Tripal is an open-source, customizable, scalable, modular database platform designed to address the constraints and resource inefficiencies of legacy database systems. This project will both leverage and coordinate funded efforts to enhance or update tree crop databases (Genome Database for Rosaceae, Citrus Genome Database, TreeGene and Hardwood Genomics Web) to Tripal that will support cross-site communication, adoption of existing standards, and "big data" integration and analysis. In addition to database related development and testing, we will work on developing workflows related to genome assembly, transcriptome assembly, annotation, and related pipelines.

Douglas R. Cook, University of California at Davis

4/17/17

Legume species are key components of both natural and agricultural ecosystems, and for human nutrition. Their importance derives in large part from their capacity for symbiotic nitrogen fixation with soil bacteria, enabling them to return vital nitrogen to the soil environment and to create seed and forage of high protein content. Two decades of molecular and genomic studies in model systems have revealed the presence of exquisite genetic pathways that initiate symbiosis, but despite these advances we have essentially no understanding of genes that regulate symbiotic performance in the natural environment. Our research aims to understand the evolution an important legume crop species - chickpea - Cicer arietinum, by elucidating the changes in its capacity for symbiotic association with Rhizobial strains, and resistance to pathogens such as Fusarium, compared in to its wild progenitors - C. reticulatum and C. echinospermum. Using a combination of ecology, population genomics, classical molecular genetics and functional assays, we are poised to explain how human selection has reshaped these and other biological processes during domestication. This study of gene function in natural versus human-built environments and its outcomes will have relevance to both basic science and agriculture. Data and biological resources generated under this project will be available through public repositories, including the NCBI and USDA-ARS's GRIN. Training and mentoring of under-represented minorities, and students at various academic levels - high school, undergraduate and graduate sis geared towards their professional development.

Peter Waddell, Ronin Institute

5/26/17

DeNovo short--read sequencing to ~10X costs ~$120 for a 200mb genome; ideal for investigating biodiversity and/or undergraduate projects. However, a typical bioinformatics/genomics/phylogenetics work--flow will cost thousands more due to biases, paralogy, misassembly, etc. Here we propose implementing and extending robust phylogenetic methods to take advantage of evolutionary distances derived using Assembly--Free--Alignment--Free (AFAF) methods. These provide rapid evaluation of the quality of evolutionary trees and a range of hybridization--introgression scenarios. Generating such distances can require a few TB of storage of raw reads and up to 1/2 TB of memory for some methods/implementations.

Jeffrey Blanchard, University of Massachusetts-Amherst

12/14/17

The acceleration of global warming due to terrestrial carbon-cycle feedbacks may be an important component of future climate change. Both the sign and magnitude of these feedbacks in the real Earth system are still highly uncertain because of gaps in basic understanding of terrestrial ecosystem processes. The Harvard Forest Long Term Ecological Research (LTER) site in Petersham, MA is home to three experimental soil warming sites heated continuously to 5° C above ambient, a temperature increase which falls within the range of the Intergovernmental Panel on Climate Change (IPCC) worst-case-scenario projections for increased global air temperature by the year 2100. To test changes in community composition and activity soil samples were collected from the experimental plots at Prospect Hill in 2015 at six time points between April and November (6 time points x 6 plots x 2 treatments x 2 soil horizons = 144 samples). RNA was extracted and the resulting libraries sequenced using the Illumina NextSeq platform. Analysis of RNA will provide insights into the taxonomic structure of soil microbial communities, and how this community might be shifting in response to climatic drivers. It is well known that factors involved in seasonality (ie: temperature, moisture, carbon availability) are drivers of microbial community structure. We expect that a different subset of the microbial community will be more highly represented in the rRNA data at different points during the growing season. We are in particularly need of computational resources for running assembly programs

Nathan G. Swenson, University of Maryland

1/20/18

Ecologists have increasingly highlighted how the phylogenetic and functional dimensions of biodiversity might be necessary to elucidate the mechanisms underlying the diversity and dynamics of species in assemblages. However, phylogenetic and functional trait analyses have several key limitations. Specifically, we argue the phylogenetic investigations are limited because they rely on the assumption of phylogenetic conservatism and because a phylogenetic result can never indicate which individual aspects of function are most important. Functional trait studies are limited to a small set of traits and are therefore ignorant of many other known or unknown important axes of function. Further, most functional trait analyses fail to closely investigate individual- and population-level responses within species to local-scale gradients in environmental conditions. Here we argue these limitations may be overcome through the integration of genetic diversity and transcriptomics into community ecology. The present research will quantify the phylogenetic, functional, genetic and transcriptomic diversity within a series of four long-term forest plots in the USA and China to ask when, where and why are levels of phylogenetic, functional trait or transcriptomic similarity the best predictors of species co-occurrence and dynamics. NCGAS resources would be used for transcriptome assembly. We have previously used Mason for this project.

Christopher Chandler, SUNY Oswego

3/26/18

Allosomes are chromosomes that determine sex, like the X and Y chromosomes in mammals, and they play crucial roles in the evolutionary biology of species. They have evolved independently in a wide array of species, and while these chromosomes in different groups share many similarities, they also exhibit important differences. Explaining these differences, however, remains difficult. This project will address this problem by examining allosomes in terrestrial isopod crustaceans, an ideal study system because they exhibit considerable variation in sex-determining mechanisms. However, surprisingly, their genomes have received little attention. The proposed experiments are designed to help explain why these chromosomes are so unique, and how they contribute to vital biological processes. Specifically, this research will (i) identify where changes in sex-determining chromosomes have occurred on the isopod evolutionary tree; (ii) test whether genes on these chromosomes are affected more strongly than autosomal genes by natural selection and genetic drift; and (iii) test whether these species exhibit dosage compensation. This work will also generate a large volume of genome sequencing data, providing some of the first draft genome assemblies for these under-studied organisms, requiring the use of powerful computing resources. These assembled genome sequences will create bioinformatics training opportunities for undergraduate students through the development of a new genomics course to be taught at SUNY Oswego. By studying a unique taxonomic group, this research will also help examine the generality of patterns suggested by earlier work on allosomes, providing significant insights into these influential components of so many organisms' genomes.

Laurel Yohe, Stony Brook University

3/27/18

There are some protein-coding genes that occur in multiple copies, and their evolution can lead to new function or complete degradation over time, depending on the selective pressures from the environment. However, comparing these genes among species is a challenging problem for geneticists and evolutionary biologists interested in understanding the functional implications of mutations and copy number variation. The mammalian olfactory receptor gene family is one of the most extreme examples of evolution by gene duplication and loss. Olfactory receptors bind to the chemical cues released into the environment. The number of functional receptors is highly variable, with some species possessing over 1,000 functional receptors and some only a few hundred. While it is hypothesized that receptor repertoires reflect the degree of ecological reliance on the sense of smell, measuring how natural selection has shaped these gene families is difficult because current models fail to simultaneously account for gene duplication, loss and mutation, and the gene copies in each species are similar. This NSF DDIG project will generate quantitative methods to test molecular adaptation in olfactory genes of bat species with divergent specialized diets. The match between dietary specialization and olfactory receptors provides an ideal system to test if these newly evolved receptors are related to the reliance on plant resources. This DDIG fosters inter-institutional collaboration and addresses a fundamental gap between micro- and macroevolutionary processes. The innovative methods developed in the project will be applicable to all gene families, including those involved in immune function and pathogen recognition. The results will uncover the evolutionary forces that shape one of the least understood mammalian senses.

Nikolaos Schizas, University of Puerto Rico Mayguez

4/30/18

Major Hurricane Maria, on 20 September 2017, delivered a devastating blow to the island of Puerto Rico. This powerful and rare weather event provides an opportunity to examine the effects of extreme physical forces on coral reefs. The research team will concentrate on the mesophotic reefs or "twilight zone reefs", which rival shallow water coral reefs in diversity and beauty. Mesophotic coral ecosystems are reefs found between 30 and 100 m depth, and are thought to serve as refugia for the declining shallow water coral reefs because they are further removed from anthropogenic and natural disturbances. The investigators are positioned with pre-deployed oceanographic instruments and previously-collected ecological data to measure the effects of Hurricane Maria on the presumably sheltered, mesophotic reefs. The investigators will document the hurricane effects on the reef communities by comparing photographic data from pre-established transects at shallow and mesophotic reefs. Misplaced corals will be used to test the capacity of extreme weather events to shape the population genomic connectivity of the coral host and their microbial communities.   Our research group is requesting NCGAS resources to analyze transcritpome, SNP and metagenomic bacterial and viromic data. We have limited computing resources and our infrastructure has been substantially weakened by the Hurricane. The investigators will document the hurricane effects on the reef communities by comparing photographic data from pre-established transects at shallow and mesophotic reefs. Misplaced corals will be used to test the capacity of extreme weather events to shape the population genomic connectivity of the coral host and their microbial communities. The project will involve the collaboration of an interdisciplinary team of five researchers and a graduate student from a Hispanic-serving Institution. The investigators will disseminate the results to Natural Resources Management Agencies and will use local media outlets and organize outreach events to inform local communities about the effects of hurricanes on coral reefs.

Scott D Cinel, University of Florida

5/8/18

I aim to describe the effects of prolonged exposure to bat predation risk on the demography, localized developmental and reproductive physiology of the corn earworm moth (Helicoverpa zea; CEW) across a natural landscape. For my Master's thesis, I carried out a differential gene expression (DGE) experiment using RNA sequencing followed by transcriptome assembly, annotation, DGE analysis, and Gene Ontology (GO) term enrichment analysis on the proto- and deuto-cerebral tissues of bat-call exposed vs. unexposed adult male fall armyworm (Spodoptera frugiperda) moths to identify specific genes and functional gene sets that are induced upon prolonged predator-cue exposure. Notably, GO analysis revealed that 290 transcripts related to glutamate synthesis and metabolism, several cellular signaling cascades, and macromolecular complex binding were significantly overrepresented in bat-call-exposed relative to non-exposed males. Further, several transcripts possibly involved in neuronal formation, axonal guidance, neural plasticity, gene regulation, calcium ion release from the endoplasmic reticulum, transportation of calcium to the mitochondria, and further cellular metabolic processes were significantly upregulated, suggesting that exposure to bat ultrasound enhanced the cellular metabolic state and significantly impacted gene regulation. I am now requesting resources from NCGAS to re-analyze some of the aforementioned datasets in the Galaxy web portal to ensure their accuracy while also planning my future use of it for a tissue-specific RNA-seq study of stress physiology thoughout the noctuid moth nervous system during and after predator exposure.

Rebecca de Wardt, San Diego State University

5/9/18

The central dogma of molecular biology is: DNA is transcribed into RNA, which is then translated into proteins. These proteins drive the cell's activities, yielding the organism's phenotype. Genotypes offer many clues about phenotype, but 1/2 to 1/3 of genes remain unknown. Phenotypic assessments of organisms are limited to only a few traits. Therefore, our research is linking genotype to phenotype, by high-throughput approaches to sequencing and trait assessment. Here, we start the analysis by comparing gene content of 15 novel marine bacteria and their phenotypic profiles for 71 different carbon sources.

Isolated marine bacteria were individually cultured; whole genome sequencing was completed on each isolated bacterium using Illumina MiSeq. Genomes were annotated with Rapid Annotation using Subsystem Technology. The analysis of genes within the bacteria, using Bray Curtis clustering, showed the genomes divided into 3 major clades with 10 genomes sharing greater than 90 similarity index. Bray Curtis clustering of phenotypes showed only 80 similarity index for 11 bacteria with 2 major clades. Spearman Correlation between the genotypic and phenotypic clustering analysis demonstrated that the genomes and phenotypes tested were not correlated (Rho = 0.029, p = 0.385). The discontinuity between phenotype and genotype, suggests that some hypothetical genes could be contributing to the phenotypes we describe. These annotated genomes have, on average, 1,419 hypothetical proteins. In order to aid in identification of hypothetical proteins being utilized during specific substrate metabolism, transcriptomes are being employed. NCGAS' resources will aide in the assembly, annotation and analysis of these transcriptomes.

Lydia Bright, SUNY New Paltz

5/29/18

Certain Paramecium caudatum strains are commonly infected by Holospora undulata, a bacterial symbiont that is able to specifically infect the micronucleus. Our overall goal is to determine the precise cellular factors that determine susceptibility and resistance to H. undulata in P. caudatum cells. The ways in which H. undulata moves into the cell and infects the nucleus suggest that proteins involved with signaling and membrane trafficking mediate this infection. We collected and sequenced P. caudatum RNA at three timepoints early in the infection process, in order to focus on early signaling and trafficking events in infection. We would like to use NCGAS resources to analyze the full sets of genes, both evolutionarily across P. caudatum strains, and by functional prediction as well.

Sierra Desiree Baney, Millersville University

Significant differences have been observed in the contractile activities of various muscles in the squid Doryteuthis pealeii, but molecular mechanisms responsible for these differences are not understood. In many animals, large muscle proteins, like twitchin and kettin, play a role in controlling muscle stiffness and elasticity. Preliminary research indicates that different forms (isoforms) of both twitchin and kettin are expressed in two squid retractor muscles with distinct physiological characteristics. Differential processing during gene expression appears to be responsible for the production of these isoforms, which appear to be impacting muscle contraction. To understand the role of these protein isoforms and the mechanisms by which they arise, we plan to analyze RNA sequences for kettin and twitchin that are produced by six diverse muscle types. Identifying correlations that exist between the type of muscle activity observed and the specific isoforms present will lead to a better understanding of how invertebrate muscle activity is influenced by twitchin and kettin. Comparing the DNA sequence of each gene to the sequences of mature RNA expressed in specific cells will verify if alternative intron splicing or other types of RNA editing are responsible for the production of kettin and twitchin isoforms.

Peter Dunn, UW-Milwaukee

7/24/18

This study will use genomic analyses to test several recent hypotheses for the genetic basis of ornament expression in a warbler, the common yellowthroat. This warbler is one of a few species of birds in which females are known to choose mates based on ornamental traits that are associated with indices of fitness. Interestingly, sexual selection targets different ornaments in different populations of the common yellowthroat. In Wisconsin, females prefer males with larger black masks, whereas in New York females prefer males with larger, more colorful yellow bibs. The PIs have already shown that mask size in WI and bib color in NY are honest signals of similar aspects of male quality in each population. Based on these results, the PIs plan to use a genomics approach to determine if: 1) elaborate ornaments are related to the expression of genes in growth, immunity and oxidative stress pathways, as predicted by recent hypotheses, and 2) these traits exhibit molecular parallelism where the same genetic pathways are most closely associated with the ornament preferred by females in each location. To maximize the probability of finding differences and to expand the tests of parallelism, the study will also incorporate geographic comparisons with other subspecies and a closely-related species in Belize, which have larger differences in ornament size and color. As part of the project we will conduct workshops on genomics at local colleges, an international behavior meeting and, in Belize, for a local zoo and conservationists.

Frank Edward Anderson, Southern Illinois University

9/19/18

Relationships among the six major extant lineages of Decapodiformes-Idiosepiida, Myopsida, Oegopsida (including Bathyteuthoidea), Sepiida, Sepiolida and Spirulida-have puzzled cephalopod biologists for over a century. High-throughput sequencing methods and genome-scale data have successfully resolved several recalcitrant phylogenetic questions, but recent studies of mitochondrial proteome and nuclear genome/transcriptome data have supported differing patterns of relationships within Decapodiformes. The standard response to this dilemma has been that broader taxon sampling and even more data will be needed to clarify relationships, but can we do better than that? To squeeze as much phylogenetic insight as possible from currently available genome-scale data, we will infer relationships within Decapodiformes using all publicly available decapodiform transcriptome and genome data as of September 2018 and evaluate the impact of numerous factors known to influence phylogenetic inference, including taxon/data sampling, orthology inference, outgroup choice, missing data, conflicting signals among loci and compositional and substitution rate heterogeneity among lineages. Our findings should clarify which aspects of decapodiform phylogeny are robustly supported, eliminate some proposed hypotheses from serious consideration and highlight the remaining problematic nodes. NCGAS resources are critical for the transcriptome assembly stage of this project; I am running Trinity in my lab and on SIU's research computing cluster, but these resources are not enough to handle the largest data sets I am using.

Eve Syrkin Wurtele, Iowa State University

9/28/18

Orphan genes are protein-coding regions of the genome that do not have any recognizable homology in other species. Previously, genes were thought to only arise from pre-exiting genes, by duplication, and gradually acquiring new functionality. But, recent evidences show protein coding sequences continuously arise de novo, from noncoding sequences. These orphan genes are a major source of evolutionary novelty and can confer novel beneficial traits. We hypothesize that orphan genes are a source of new genetic variability that enhances adaptability to changing environmental pressures and can be harnessed to improve agronomic traits. Thus, orphan genes have huge implications for basic and translational research in agriculture and plant sciences are huge. The vast functional genomic data will enable us to systematically discover and characterize orphan genes, their evolution, and their impact on fundamental processes. Large memory machines and storage is required to download and process big public data sets, testing multiple parameters to enable us to move these goals forward. This space is unavailable to us at Iowa State University. Therefore we are requesting this allocation from NCGAS. We propose following objectives in order to further explore the orphan genes and to fully tap their potential as a novel source of variation and for crop improvement: Goal 1: Systematically uncover orphan genes in the sequenced genomes (plants and animals), and annotate their origin and function. Goal 2: Predict potential functions for orphan genes for the species with rich expression data in public databases using evidence-based approaches.

Daniel P Tonge, Keele University

10/5/18

Following discussion with NCGAS team on Wednesday; We have sampled 20 horse chestnut trees for the purpose of understanding the impacts of "bleeding canker", on the horse chestnut transcriptome. As no transcriptome is currently available, we have generated in excess of 400 million RNA-seq reads (paired end, 75nt) using the Illumina HiSeq system. We have attempted de novo assembly locally, however do not have the RAM available to perform this task. In order to develop our workflow, we have highly subsamples the original data to 20,000 reads and performed an end-to-end assembly, annotation and differential expression workflow, identifying a range of interesting genes in the process. Now optimised, we wish to repeat this workflow using the whole of the initial data pool to ensure we are not missing any interesting but lowly expressed genes. :-)

Nikolaos V Schizas, University of Puerto Rico Mayaguez

10/15/18

The octocoral Briareum asbestinum (Pallas 1766) is distributed widely across the shallow coral reefs of the Caribbean. Unlike most octocorals, B. asbestinum has a wide depth range from 1 meter to 40 meters (Bayer 1961). Its richness in secondary metabolites and defense compounds enables the species to widely disperse and outcompete other benthic cnidarians for space. Briareum asbestinum exhibits two distinct colony morphs in the Caribbean: a digitate morph and an encrusted morph. The encrusted morph overgrows hard substrates and has been inconsistently referred to either as B. asbestinum or B. polyanthes. Although these two morphotypes have undergone morphological characterization, their taxonomic status remains unresolved. We will present the first transcriptome of the non-scleractinian reef building Briareum asbestinum of the digitate morphotype from southwest Puerto Rico. With this first transcriptome we hope to start resolving a long-term debate on the se morphotype status. Using next generation pair end sequencing (Illumina NextSeq2500) we obtained 159,754,702 raw paired-end reads and with the NCGAS tools and pipeline we plan to assemble the remaining high quality reads.

Marcella D Cervantes, Albion College

10/15/18

We have previously identified the mating type locus of the ciliate Tetrahymena thermophila (Cervantes et al. 2015). The somatic mating type locus contains two genes that proved to be essential for progeny production. Knockouts of either gene did not produce progeny. Further characterization of these knockouts suggests the two genes have different functions, rather than working cooperatively. One knockout can pair with a cell of a different mating type, but pair formation is not stable. The other knockout does not show any characteristics of mating type recognition. We have performed RNA isolation to determine the gene activation differences between individual knockouts compared to each other and to wild type cells.

Matthew Darwin Robbins, USDA ARS Forage and Range Research Laboratory

10/16/18

We plan to identify genes associated with salt tolerance through gene expression analysis. Salt tolerant and salt susceptible cultivated varieties of two common turf grasses, Kentucky bluegrass (Poa pratensis L.) and perennial ryegrass (Lolium perenne L.) will be treated under control and salt stress. Differentially expressed genes in tolerant cultivars with respect to susceptible cultivars will be considered for further validation and DNA marker assisted selection. Additionally, the response of an alkaligrass (Puccinellia distans (Jacq.) Parl) cultivar will be included as a positive control. In order to determine differential gene expression, we need the resources at NCGAS to create a de novo transcriptome assembly, the map the reads to this assembly, and run differential expression analysis. We will utilize the IU Carnbonate cluster and the de novo transcriptome assembly pipeline developed by NCGAS.

Yash Sondhi, Florida International University

10/21/18

The project examines sensory system evolution in Bombycoid moths. My role in the project includes looking at the evolution of visual pigments and the functional characterisation of different visual genes in diurnal and nocturnal Bombycoid moths. NCGAS resources will help greatly in assembling the RNAseq data we are generating in order to perform comparative transcriptome analysis on moths that are active during the day and night. We will also be using computing resources to build phylogenetic trees and annotate the de-novo transcriptomes we get.

Laura Aline Katz, Smith College

11/16/18

Description of Research Testate amoebae have a rich history of classification based on the shape and composition of their tests (shells), with the unit of diversity being the 'morphospecies'. Yet preliminary work in the PI's lab and elsewhere suggest that morphology can be misleading as individuals of some morphospecies are genetically quite distinct. Such data suggest a disconnect between morphological and molecular evolution among these microorganisms. Moreover, few studies have used molecular tools to explore the biodiversity of these beautiful organisms, particularly in the United States. Such studies are important given that testate amoebae serve as bioindicators of climate change and are abundant in threatened habitats like bogs and fens. The PI's lab are working on addressing the following interrelated hypotheses on the biodiversity of testate amoebae: H1: There is considerable discordance between morphospecies designations and molecules as many morphospecies are underlain by multiple, non-monophyletic phylogenetic species. H2: Additional phylogenetic species will be discovered in both the relatively isolated fens of the Rocky Mountains and in the widespread bogs of Alaska. H3: Abundant community members, as assessed by both single cell PCR and community DGGE, vary across time and space. Need for NCGAS To test our hypotheses, we are using single-cell omics as well as other methods for community works. We amplify single-cell genomes and transcriptomes of testate amoebae, perform illumina Hiseq sequencing. For analyzing the large amount of resulting data including from assemblies to genome mapping, and tree building etc., it would be really helpful and much more efficient to be able to access the great and diverse resources on NCGAS.

Barbara Ambrose, New York Botanical Garden

11/19/18

Ceratopteris is a fern model species, it is the first Fern genome to be sequence and is used widely to look at fertilization, but little is known about its development. Its leaves are dimorphic, meaning that the leaves that bear sporangia have a different morphology in comparison to the sterile leaves. In order to better understand the genetic basis of the two morphologies found during leaf development we are generating transcriptome data from various stages of sterile and fertile leaves of two Ceratopteris varieties, rn3 and hnn, and developing a differentially expressed gene analysis.

Sara Cahan, University of Vermont

11/28/18

Phenoypic plasticity is ubiquitous across the tree of life and plays a fundamental role in allowing organisms to optimize their phenotypes in the face of variable environmental conditions. Plasticity is mediated is via shifts in gene expression, which may be regulated by epigenetic mechanisms, including posttranslational histone modification, regulatory non-coding RNA and chromatin structure. In the proposed project, the PIs will test the hypothesis that epigenetic regulators act as an intermediary between environmental sensors and protein production, altering the set of genes available for transcription at the level of chromatin accessibility and then fine-tuning expression through the action of epitranscriptomic molecules. The project will focus on a well-characterized, experimentally tractable, and evolutionarily important plastic response, thermal acclimation in Drosophila melanogaster. We will be conducting considerable epigenetic (ATACseq, CHiPseq) and transcriptomic (RNAseq) profiling, generating 10-20TB of data requiring significant computational power for bioinformatic analyses. In the past, NCGAS has been an exceptional resource to me for such analyses.

Heather Bracken-Grissom, Florida International University

1/8/19

Bioluminescence, which is rare on land, is extremely common in the deep sea, being found in 80% of the animals living between 200 and 1000 m. These animals rely on bioluminescence for communication, feeding, and/or defense, so the generation and detection of light is essential to their survival. Our present knowledge of this phenomenon has been limited due to the difficulty in bringing up live deep-sea animals to the surface, and the lack of proper techniques needed to study this complex system. However, new genomic techniques are now available, and a team with extensive experience in deep-sea biology, vision, and genomics has been assembled to lead this project. This project is aimed to study three questions 1) What are the evolutionary patterns of different types of bioluminescence in deep-sea shrimp? 2) How are deep-sea organisms' eyes adapted to detect bioluminescence? 3) Can bioluminescent organs (called photophores) detect light in addition to emitting light? Findings from this study will provide valuable insight into a complex system vital to communication, defense, camouflage, and species recognition. The first objective is to use phylogenomic methods to build a robust phylogeny. This will allow us to trace the evolutionary origins of the two bioluminescence modes (secretion and photophore) within oplophorid shrimp. Secondly, this project will use RNAseq data to characterize the visual systems of deep-sea shrimp to better understand how shrimp distinguish between different wavelengths of emitted bioluminescence. Lastly, integrative methods will be used to examine photosensitivity in several non-bacterial (autogenic) light organs - the photophore and organs of Pesta (light organ of Sergestidae).

Ross Whetten, North Carolina State University

1/13/19

This project uses transcriptome data and ATAC-seq data to identify regions of accessible chromatin in conifer genomes. The goal is to identify regions of accessible chromatin not in or near expressed genes, as these are likely to harbor regulatory elements important for controlling gene expression. Genetic variation in these elements will be tested, in conjunction with available SNP data from coding regions, to determine the extent of phenotypic variation that can be accounted for in a range-wide collection of P. taeda.

Yee Ming Leslie Beh, Columbia University

2/1/19

Oxytricha trifallax is a ciliated protozoan with two genomes - a germline micronucleus and a somatic macronucleus. The macronuclear genome exhibits a distinct architecture, with 16,000 chromosomes averaging ~3.2kb in length with compact 5' and 3' UTRs. It remains unknown how transcriptional regulation is achieved in this genome, given the paucity of non-coding DNA. It may be possible that distinct chromosomes are spatially localized within the macronucleus, allowing co-expression of particular sets of genes (or, conversely, silencing). This project aims to elucidate the 3-dimensional organization of the Oxytricha macronucleus through Hi-C, a well-established genomic technique for assaying such interactions. The IU computing cluster will be used for this analysis.

Sen Xu, University of Texas at Arlington

2/19/19

We are interested in understanding the origin of obligate asexual reproduction and investigating the genetic basis of the variation of recombination rate. We intensively use high-throughput genomic, transcriptomic, and epigenomic data to address these questions. NCGAS resources have been critical for our data analysis and data storage.

Laura Landweber, Columbia University

3/7/19

Three-dimensional organization of chromatin is essential for regulating gene expression. Nuclear architecture is especially crucial during development and particularly dynamic in the germline. Ciliates, as single-celled eukaryotes exhibiting nuclear dimorphism, possess highly unique germline genomes, which are transcriptionally silent yet perform sexual functions. These include the generation of a structurally distinct somatic genome that produces all gene expression in the daughter cell. The development of a somatic genome from a germline genome is an RNA-guided epigenetic process, requiring elimination of over 90% of germline DNA, rearrangement of remaining fragments, and finally massive amplification and telomere addition. The scrambled germline is specialized for this distinct role, divorced from transcriptional functions that typically drive genome architecture. Short sequences flanking somatic-destined regions help join these regions in the proper order, but little beyond the germline's 1D sequence is understood, largely due to the overwhelming DNA content of the somatic nucleus. I have developed a novel FACS-based method for high-purity, high-throughput separation of Oxytricha nuclei, and will apply this protocol to the thorough characterization the germline. Using 3C and microscopy methods, I will determine the 3D organization of the germline genome and test how this recapitulates developmental function, as well as improve the germline genome assembly, whose total chromosomes are yet unknown. To this end, I have prepared and sequenced germline Hi-C Illumina libraries. Analysis of Hi-C data using the well-attested Juicer software requires the use of a computational cluster and ample storage space, and for this reason I seek use of NCGAS resources.

John Tarpey, Loyola University New Orleans

4/10/19

CDTA of a newly described Chagas Disease Vector, Triatoma Mopan. We will assemble the transcriptomes using SOAPdenovo, Transabyss, and Trinity, and the output from each will be combined using Evigene. Significant progress has already been made with SOAPdenovo, Transabyss, and Trinity, but the last step, Evigene is still to be completed.

Maurine Neiman, University of Iowa

4/22/19

We require NCGAS resources for analysis of the P. antipodarum genome, including gene prediction and genome annotation. The available software and computing power of the IU large memory cluster will enable us to complete these critical steps for generating the P. antipodarum reference genome information that we will use in addressing the genomic consequences of asexuality

Shivam Bhardwaj, Mississippi State University

5/8/19

Butterfly wings are exquisite mosaic displays of individual scales with distinct colors and patterns, which help them deflect predators and attract mates. These patterns are critical to their evolutionary success and are determined by interactions of conserved genetic pathways with their developmental environment. I am interested in spatiotemporal regulation of butterfly wing pattern development during late larval and early pupal development through sex-determination and non-canonical Wnt signaling pathways. To answer these questions, I recruited cell-specific reporters to identify cellular and subcellular localization of gene products during wing development. I have established working protocols for immunostainings to visualize JNK pathway gene products and in-situ hybridization for doublesex as well as other Wnt-PCP candidate genes. We have generated guide RNA for CRISPR-Cas9 knockouts for candidate genes in the non-canonical Wnt signaling pathways. We are currently in the process of accumulating more data for each of these projects.

Catalina Fernandez, Indiana University

5/8/19

During the course of human evolution diet has had identifiable evolutionary consequences in our biology as a species, and contributed to genetic and cultural variation among populations. Additionally, this cultural and genetic variation may contribute to epidemiological patterns of chronic non-communicable diseases (NCDs). Some scholars maintain that many NCDs today are the result of an "evolutionary mismatch" between our genetically-based physiology and metabolism, viewed as well-adapted to a hunting-gathering lifestyle, and contemporary dietary and physical activity patterns. However, there is growing evidence indicating that relatively recent migrations into new ecological regions and dietary shifts (plant and animal domestication) resulted in genetic adaptations (e.g. gene variants contributing to high altitude adaptation, or lactase persistence). The question arises as to whether relatively recent dietary shifts have had evolutionary consequences for our species, and to what extent these are related to the rise in NCDs in post-nutrition transition populations. Thus, this study aims to evaluate whether genomic regions involved in nutrient metabolism have been differentially targeted by selection in two post-nutrition transition Amerindian populations with contrasting histories of subsistence strategies (hunting-gathering vs. agropastoralism) and evaluate whether these signatures are subsistence-specific and correspond to the prevalence of several NCD risk biomarkers. This study will include two Amerindian groups of Chile: the Mapuche, former hunter-gatherers from south-central Chile, and the Atacamenos, former agropastoralists who inhabit the Atacama Desert. This research design will comprise ethnographic fieldwork including oral history interviews, dietary and physical activity surveys, measurements of blood and anthropometric NCD risk biomarkers, and genome-wide scans for selection. We hypothesized that the distinctive diets historically consumed by these populations have constituted selective pressures that resulted in subsistence-specific genomic signatures of selection on metabolic processes. Alternatively, under the evolutionary mismatch framework, it is predicted that the former hunter-gatherer population will show stronger selection signatures on metabolic regions associated with lipid and carbohydrate metabolism and higher prevalence of NCD risk biomarkers than the former agropastoralist population.

Joseph Bisesi, University of Florida

5/22/19

Single-walled carbon nanotubes are small man-made materials that are currently being used in a number of consumer industries. With this increased use, there is increased likelihood that these materials will make their way into aquatic environments. The materials have properties which make them highly likely to be consumed by aquatic organisms such as fish where they may interact with the fish intestinal system. The overall objective of this project is to understand how single-walled carbon nanotubes influence the intestinal environment by making fish more susceptible to chemicals (i.e. pollutants) that are already present in our environment and are associated with health concerns. To achieve this objective the PI will examine how single-walled carbon nanotubes may change the normal structure and function of the intestines of exposed fish and determine whether these changes contribute to increasing their susceptibility to other chemicals that are already present in our aquatic environments. Single-walled nanotubes are lipophilic in nature and are likely to sorb other aquatic contaminants already present in our environment as well as alter the intestinal tract of fish through dietary exposure routes. The objective of this research is to better understand the specific interactions of single-walled carbon nanotubes with fish intestines by probing their interaction with lipids in the gastrointestinal system which may result in structural and functional changes in the intestinal lipidome. The PI will also examine whether such changes in gastrointestinal lipid profiles result in release of chemicals that may be adsorbed to single-walled carbon nanotubes, increasing bioaccumulation of these contaminants. The PI will accomplish these objectives by examining gastrointestinal lipid profiles during single-walled carbon nanotube dietary exposure, determining the effects of lipids on sorption of chemicals to these materials, and examining the effects of nano-lipid interactions on chemical bioaccumulation. Activities performed in this project are expected to have far reaching impacts on an international, national, and regional scale. Results from the nano-bio interaction studies will contribute to the design of next generation nanomaterials that take advantage of, or avoid, identified downstream behavior. In this way the PI can continue to utilize the amazing properties of nanomaterials without detrimental consequences. On a local level, the research team plans to include students from underrepresented groups in this project as well as increasing public knowledge on the benefits of nanomaterials through the use of videos. The scientific community will be involved in this project on a national scale through teaching of short courses at regional meetings as well as leading sessions at national meetings. Finally, the research team will expand education by integrating information about nanomaterial uses for combating health issues in resource poor countries through existing teaching programs in international communities.

Meredith Protas, Dominican University of California

5/24/19

We work on the species Asellus aquaticus which is a freshwater crustacean that has cave (eyeless, pigmentless) and surface (eyed, pigmented) forms. The two forms can interbreeding allowing for the investigation of the genetic basis of these traits. We have sequenced the transcriptomes of adult F2 animals of different color phenotypes and would like to look at differential expression of the different color morphs. To investigate this question, we would greatly benefit from use of XSEDE resources as we are at a small school with little computing resources.

Tess Leuthner, Duke University

6/5/19

I am collaborating with Joe Shaw on the following project: Mutations are the ultimate source of all genetic variation. How organisms interact with the environment to influence the processes that generate mutations is therefore, critical for understanding the origin of all evolutionary change. Scientists are rapidly discovering the integrative role the epigenome, defined as chemical compounds that interact with DNA to modify gene function, plays in mediating responses between the environment and the expression of individual traits. However, the molecular events that underlie these environmentally governed controls remain virtually unknown. This research is designed to characterize the influence of environment induced epigenetic state on the frequency, types and genetic location of mutations to better understand their influence on the health and fitness of organisms, populations, and evolution. The waterflea, Daphnia pulex, offers a unique opportunity for these studies, because it is well established as a model of environmental stress, its genome is well characterized, and its reproductive strategy that includes clonal reproduction allows for control of both genetic and epigenetic background. This research through outreach programs at both Indiana University and Mount Desert Island Biological Laboratory will benefit the development of young scientist, especially related to the novel bioinformatics and computational approaches that will be developed in this proposal. The knowledge gained through this research stands to have profound implications for society and the long-term health of populations, which are living longer in the presence of a large diversity of chemicals that can modify DNA.

Anna Rix, University of Alaska Fairbanks

6/10/19

The long evolution of Antarctic notothenioid fishes in the cold, oxygen-rich waters of the Southern Ocean has reduced their thermal tolerance, heat shock response, the oxygen-carrying capacity of their blood, and may have reduced their capacity to respond to hypoxia. The master regulator of hypoxia response in metazoans is the transcription factor hypoxia-inducible factor-1 (HIF-1). HIF-1α in notothenioids has a polyglutamine/glutamic acid (polyQ/E) insert that varies in length with phylogeny. HIF-1 may also play a role in temperature tolerance as well as in other fishes HIF-1 DNA binding changes with temperature. This project seeks to determine if HIF-1 regulated genes are responding to temperature using existing transcriptomics data. Additionally, we hope to determine the prevalence of polyQ/E inserts throughout the transcriptome of various notothenioids as well as characterizing the hypoxia response of Notothenia coriiceps using RNA-seq. NCGAS resources will be used to enable assembly of transcriptomes and analysis of these transcriptomes.

Ryland Young, Texas A&M University

6/27/19

To annotate newly isolated and sequenced phage genomes using the best prediction software available, we will employ the best structure prediction algorithm, HHPred, on whole predicted proteomes. This is an important problem because phage are being explored and applied in medicine, food safety, and the agricultural fields. Using supercomputing resources is absolutely necessary to maintain pace with the rate of new genome discovery and annotation.

Gretchen Hofmann, University of California, Santa Barbara

6/28/19

This project aims to understand the role of DNA methylation in transgenerational plasticity and acclimation to global change in the purple urchin Strongylocentrotus purpuratus, mothers of which provision their offspring differently depending on whether they experience high pCO2 upwelling conditions during gametogenesis. Larvae from upwelling-conditioned parents exhibit greater performance under high pCO2 and show marked differences in gene expression relative to offspring from non-upwelling parents. Preliminary data shows that genome-wide DNA methylation also varies between larvae from parents that experienced upwelling and non-upwelling conditions. The requested NCGAS computing resources will be used to analyze reduced-representation bisulfite sequencing and RNA-seq data in order to understand how transgenerationally plastic changes in DNA methylation affect global transcription and the regulation of physiologically-relevant pathways. Ultimately, this project will uncover epigenetic mechanisms that control rapid acclimation to global change in an ecologically-critical model system.

Jonathan A. Karty, Indiana University, Bloomington

Direct Injection Mass Spectrometry for Metabolomics (DIMS-M) uses high resolution mass spectrometry to compare the metabolic states of biological systems1. In short, an extract of living cells or tissue is ionized by electrospray mass spectrometry without chromatographic separation. Sample data are recorded as 11-17 narrow-range sub-spectra to minimize space charge effects and extend the dynamic range of the analysis. We intend to use DIMS-M to infer the metabolic pathways affected by exposure to a series of potential toxins for five model organisms and to create dosage-response curves for each experiment. Galaxy-M is needed for stitching together DIMS-M sub-spectra, feature detection, multivariate statistical analysis of the detected feature, compound identification, and generating Figures and Tables. The workflows and scripts have already been written and tested by the Viant Laboratory; they can be found here: https://github.com/Viant-Metabolomics/Galaxy-M. We intend to record data from at least 1200 samples in our initial experiments. Our hope is to make this platform robust enough for making it a service to offer other researchers and as an integral tool in obtaining future external funding for projects involving toxicology and metabolism. We would like to adapt Galaxy-M to compare full-scan, accurate mass GC-MS data acquired in other studies. 1. Southam, A. D., Weber, R. J. M., Engel, J., Jones, M. R. & Viant, M. R. A complete workflow for high-resolution spectral-stitching nanoelectrospray direct-infusion mass-spectrometry-based metabolomics and lipidomics. Nat. Protoc. 12, 310 (2017).

Bethany Jenkins, University of Rhode Island

8/13/19

Diatoms generate an estimated 40% of primary production in marine habitats, exerting profound control over global carbon cycling. Despite their importance, there are fundamental gaps in our understanding of diatom eco-physiology and thus, how diatom-mediated biogeochemical transformations are controlled. In laboratory studies with isolates, there are profound differences among diatom species' responses to nutrient limitation that imply different species likely contribute differently to nutrient uptake, carbon flux and burial. Yet, understanding diatom physiology in the field is restricted because the approaches to assess in situ physiology in a species-specific manner is lacking and translating data from cultured isolates to species in a mixed field community is challenging. This proposal focuses on the application of a powerful new approach, called Quantitative Metabolic Fingerprinting (QMF), to address this knowledge gap and examine species-specific physiological responses in the field. The proposed work will provide transformative insights into how ocean geochemistry controls the response of individual species, and how metabolic potential is partitioned between diatom species, thus providing new insights into the structure and function of marine systems. The overall goal is to examine how diatom species respond to changes in biogeochemistry across marine provinces, from the coast to the open ocean, by following shifts in diatom metabolism using QMF. NCGAS resources will be essential to analyze the large metatranscriptomic datasets generated during this project. Carbonate will be used to trim the data, map reads and analyze differential expression in order to understand the variations in physiology and community composition between contrasting marine environments.

Appendix 3. Software Supported by NCGAS

The National Center for Genome Analysis Support (NCGAS) provides support for the following genome analysis software packages available on Indiana University's Karst, Carbonate cluster, PSC Bridges, and as public pre-configured instances on Jetstream (Table 1). Access to NCGAS computational and consulting services is awarded through an allocation process to genomics research projects funded by the National Science Foundation (NSF). For more information, https://ncgas.org/software-list.phpand https://ncgas.org/Software%20list%20Jetstream.php.

Note: Genomics Toolkit image available on Jetstream hosts a set of genomics toolkit that is not supported by NCGAS but Jetstream team.

PSC Bridges cluster, and XSEDE Jetstream computing.

Table 6. Genome analysis software packages available on Indiana University's Carbonate, Karst clusters.

Package	**Description**	**Version**	**Carbonate**	**Karst**	**Bridges**	**Jetstream**
abyss	Parallel assembler for short read sequence data	1.5.2	x		x	Genomics Toolkit
		1.9.0			x	
		2.0.2	x		x	
admixture	Maximum likelihood estimation	1.3.0		x		
		5.1		x		
AllPaths-LG	Whole-genome shotgun assembler using Illumina long and short read library	524888			x	
Annovar	Functionally annotates genetic variants detected from diverse genomes	2016.02.01			x	

AFNI	Processing, analyzing, and displaying functional MRI (FMRI) data	1 .0.13	x			
		1 .3.03	x			
		16.3.00		x		
Anvi'o	Analysis and visualization of 'omics data	2.0.2			x	Anvio, Mining SRA to identify datasets
		2.2.2			x	
		2.3.1			x	
		2.3.2			x	
		2.4.0			x	
		3.0.0			x	
ARAGORN	Detects tRNA, mtRNA and tm RNA genes	1.2.3			x	
ATLAS	Automatically Tuned Linear Algebra Software	3.101.2			x	
augustus	Predicts genes in eukaryotic genomic sequences	3.3	x	x		genome assembly workshop
		3.2			x	

autodocksuite	Predicts how small molecules, e.g., substrates or drug candidates, bind to a receptor of known 3D structure	4.2.3		x		
		4.2.6			x	
bamtools	C++ API and toolkit for analyzing and managing BAM files	2.4.0			x	bcbio-nextgen toolkit v.2.0, Genomics Toolkit, genome assembly workshop
		2.4.1	x	x		
bamutil	several programs that perform operations on SAM/BAM files	1.0.13		x		bcbio-nextgen toolkit v.2.0
		1.0.14	x			
Barrnap	Predicts location of ribosomal RNA genes in genomes	0.6			x	
bbmap	Short read aligner for DNA and RNA-seq data	38.6				Genomics Toolkit

bcftools	Discovery of correlated genomic features, such as ESTs, polymorphisms, and mobile elements	1.3.1		x	x	Genomics Toolkit
		0.1.19			x	
		1.5	x			
Bedops	Scalable boolean and other set operations, statistical calculations, archiving, conversion and other management of genomic data of arbitrary scale	2.4.19			x	
		2.4.35			x	
BLASR	Mapping single molecule sequencing reads using Basic Local Alignment with Successive Refinement	1.3.1			x	
bedtools	Discovery of correlated genomic features, such as ESTs, polymorphisms, and mobile elements	2.20.1		x		bcbio-nextgen toolkit v.2.0, Genomics Toolkit
		2.26.0	x	x		
		2.25.0			x	

bfast	Blat-like Fast Accurate Search Tool (BFAST) facilitates the fast and accurate mapping of short reads to reference sequences	0.7.0a		x		
bioconductor	R packages for analysis and comprehension of high-throughput genomic sequence data	2.12		x		R for Biologists, Harvesting Field Station Data
		3.1		x		
		3.3	x	x		
		3.6	x			
		3.9	x	x		
biopython	Collection of python packages to support bioinformatics	1.59		x		bcbio-nextgen toolkit v.2.0
		1.7.0	x	x		
		1.7.3	x			
Bismark	program to map bisulfite treated sequencing reads to a genome of interest	0.19.0	x		x	
blast	Search tool that finds regions of local similarity between nucleotide or protein sequences	2.6.0+	x		x	Genomics Toolkit, genome assembly workshop
		2.2.27		x		
		2.2.2		x		
		2.2.31		x	x	

		2.7.1			x	
blat	Alignment tool like BLAST, but structured differently	35	x	x	x	
blasr	align PacBio long reads to reference genomes	1	x			
bowtie	Ultrafast, memory-efficient short read aligner. See related tool Bowtie2.	1.1.2		x	x	bcbio-nextgen toolkit v.2.0, Genomics Toolkit
		1.1.1			x	
		1.2.2	x			
bowtie2	Ultrafast, memory-efficient tool for aligning sequencing reads to long reference sequences. See related tool Bowtie.	2.1.0		x		bcbio-nextgen toolkit v.2.0, Arkansas Workshop, Genomics Toolkit
		2.2.3		x		
		2.2.6		x		
		2.2.7			x	
		2.3.2	x			
		2.3.4.1			x	
		2.3.5		x		
busco	Assesses genome assembly and annotation completeness	3.0.2	x			Arkansas Workshop, genome assembly workshop
		1.22			x	

breakdancer	Provides genome-wide detection of structural variants from next generation paired-end sequencing reads	1.1		x		
		1.3.6		x		
breseq	Finds mutations relative to a reference sequence in short-read DNA re-sequencing data for haploid microbial-sized genomes	0.23		x		
		0.30.0		x		
		0.30.2	x			
		0.32.0	x			
bwa	Fast light-weight tool that aligns relatively short sequences to a sequence database	0.6.2				bcbio-nextgen toolkit v.2.0, Genomics Toolkit
		0.7.10		x		
		0.7.12	x			
		0.7.15		x		
café	analyze changes in gene family size for evolutionary inferences	4.2	x			

canu	Fork of the Celera Assembler designed for high-noise single-molecule sequencing	1.3		x	x	
		1.4		x		
		1.5			x	
		1.6	x		x	
		1	x			
		1.7		x		
cd-hit	Tool to cluster and compare protein or nucleotide sequences	4.6.	x			
		2016.06.21			x	
centrifuge	Rapid and efficient classification of DNA sequences from microbial samples	1.0.3	x		x	
		1.0.4			x	
cfsan	Python based SNP pipelines	1.0.1	x			
checkm	Tools for assessing genome quality recovered from isolates, single cells or metagenomes	1.0.7			x	Arkansas Workshop
Circos	Produces visualizations in a circular layout, commonly used for genomics data	0.69.2			x	

clustalw2	Multiple alignment of nucleic acid and protein sequences	2.0.12		x		
		2.1			x	
colormap	ColorMap is a hybrid method for correcting noisy long reads	1		x		
crisprdo	Evaluate the goodness of an sgRNA in both sensitivity and specificity	1		x		bcbio-nextgen toolkit v.2.0
		0.7.15		x		
cufflinks	Assembles transcripts, estimates their abundances, tests for differential expression and regulation in RNA-Seq samples	2.0.2		x		Genomics Toolkit
		2.1.1		x		
		2.2.0		x		
		2.2.1	x		x	
cutadapt	Reads a FASTA or FASTQ file, finds and removes adapters, writes the changed sequence to standard output	1.11		x		Genomics Toolkit
		1.16	x		x	

		1.1	x			
		1.5			x	
		1.9.1	x			
cytoscape	Platform for visualizing complex-networks and integrating these with attribute data	3.6.1	x	x		
Dammit	Simple de novo transcriptome annotator	0.3			x	
deseq2	Estimate variance-mean dependence in count data from high-throughput sequencing assays and test for differential expression	3.0.0		x		
deepTools	Suite of python tools for the efficient analysis of high-throughput sequencing data	2.3.5			x	
		3.1.3	x			
Detonate	DE novo TranscriptOme r Na-seq Assembly with or without the Truth Evaluation	1.1			x	

dose2geno	Generate Best Guess Genotypes from Dosage and Info File	4		x		
Diamond	Alignment tool for aligning short DNA sequencing reads to a protein reference database	0.7.11			x	
		0.8.31			x	
		0.8.3	x	x		
		0.9.13	x			
Dimspy	processing of direct-infusion mass spectrometry-based metabolomics and lipidomics data	1.03	x			
Discovar	Variant caller and small genome assembler	52488			x	
Discovar de novo	Large (and small) de novo genome assembler	52488			x	
discasm	Extract reads that map to reference genomes in a discordant fashion and optionally include reads that do not map to the genome at all, and perform a de novo transcriptome assembly of these reads	0.1.2	x			

Drupal7	Drupal provides a back-end framework for at least 2.3% of all websites worldwide					Pop-up genome browser, Harvesting Field Station Data, Drupal Base
ECTools	Long read correction, plus other correction tools	12/1/2014			x	
E-utilities	Table interface into the Entrez query and database system at the National Center for Biotechnology Information (NCBI)	6/24/2015				Mining SRA to identify datasets
EPA-ng	perform maximum likelihood-based phylogenetic placement of genetic sequences	0.1.0-beta	x			
edgeR	Differential expression analysis of RNA_seq expression profiles with biological replication	3		x		
edge-pro	EDGE-pro (Estimated Degree of Gene Expression in PROkaryotes) is an efficient software system to estimate gene expression levels in prokaryotic genomes from RNA-seq data.					Genomics Toolkit

eigensoft	Uses principal components analysis to explicitly model ancestry differences between cases and controls along continuous axes of variation	6.1.3		x		
		7.2.1		x		
emboss	A high-quality package of free, Open Source software for molecular biology	6.5.7		x		
		6.6.0			x	
EricScript	Computational framework for the discovery of gene fusions in paired end RNA-seq data.	0.5.5			x	
ensembl	Various tools to assist in use and analysis of Ensembl data	95.1			x	
evigene	Pipeline script for processing large piles of transcript assemblies, from several methods into the most biologically useful set of mRNA, classified into primary and alternate transcripts	2013.07.27	x			
exonerate	Exonerate is a general purpose tool for biological sequence comparison	2.4	x		x	

Fastortho	reimplementation of the orthomcl program that does not require the use of databases or perl.	1	x			
Falcon	Collection of genome assembly tools	0.4.1			x	
fasta3	The FASTA package - protein and DNA sequence similarity searching and alignment programs					Genomics Toolkit
fastqc	Quality control for high-throughput sequence data	0.10.1		x		Arkansas Workshop, Genomics Toolkit
		0.11.3			x	
		0.11.5	x	x		
FASTA-Splitter	Divides a single FASTA file into multiple files	0.2.4			x	
		0.1.2			x	
fastx	A collection of command line tools for Short-Reads FASTA/FASTQ files preprocessing	0.0.13		x		
		0.0.14			x	
flash	Fast and accurate tool to merge paired-end reads from NGS experiments	1.2.11		x	x	
flexbar	Flexbar preprocesses high-throughput	2.4		x		

	sequencing data efficiently					
FragGeneScan	Finds fragmented genes in short reads and predicts prokaryotic genes in incomplete assemblies or complete genomes	1.2			x	
Freebayes	Bayesian genetic variant detector designed to find small polymorphisms, specifically SNPs, indels, MNPs and complex events	1.2.0	x			
gatk	Genome Analysis Toolkit, for high-throughput sequencing data	3	x			bcbio-nextgen toolkit v.2.0, Genomics Toolkit
		3.5			x	
		3.6			x	
		3.7			x	
		4.0.1.2			x	
		4.beta.5			x	
Genome MuSiC	Tools to discover the significance of somatic mutations found in a cohort of cancer samples, and with respect to various external data sources	0.4.1			x	
GFFRead	GFF/GTF parsing utility providing format conversions, region filtering, and FASTA sequence extraction	0.9.8 c			x	

Glimmer	Gene Locator and Interpolated Markov ModelER to identify coding regions	3.0.2			x	
		3.0.4			x	
GraPhlan	Tool for producing high quality circular representations of taxonomic and phylogenetic trees	0.9.7			x	
gmap	Align cDNA to reference to determine gene structure and structural variants	5/15/2014		x		
		4/4/2016		x		
		6/20/2017	x			
		5/11/201	x			
gmapfusion	identifying candidate fusion transcripts based on transcript sequences	0.3.0	x			
guppy	Oxford Nanopore Technologies' basecalling algorithms	3.1.5-cpu	x			

hisat2	Fast, sensitive alignment, mapping NGS reads (DNA & RNA) to a population of human genomes or against a single reference genome	0.1.6-beta		x		bcbio-nextgen toolkit v.2.0, Genomics Toolkit
		2.0.4		x	x	
		2.1.0	x			
hmmer	Biosequence analysis u sing profile hidden Markov models	3		x		genome assembly workshop
		2.3.2				
		3.1.b2	x		x	
HUMAnN2	Pipeline for efficiently and accurately profiling the presence/absence and abundance of microbial pathways in a community from metagenomic or metatranscriptomati c sequencing data	0.10.0			x	
htseq	Analysis of high-throughput sequencing data using Python	0.9.1	x		x	bcbio-nextgen toolkit v.2.0

htslib	C library for accessing common file formats, such as SAM, CRAM and VCF, used for high-throughput sequencing data	1.5	x			
IDBA	Iterative De Bruijn Graph De Novo Assembler short read assembler for transcriptomes	1.1.1			x	
IGV	The Integrative Genomics Viewer (IGV) is a high-performance visualization tool					Genomics Toolkit
ipython/Jupyter	Interactive python shell				x	bcbio-nextgen toolkit v.2.0
Ipyrad	Interactive toolkit for assembly and analysis of restriction-site associated genomic data sets – RAD, RADseq, GBS	0.7.2	x			
impute2	Estimating (imputing) unobserved genotypes in SNP association studies	2.2.2		x		
Interproscan	Overview of the families that a protein belongs to and the domains sites in the sequences	5.29-6	x			
Jellyfish/Jellyfish2	Tool for fast, memory efficient counting of k-mers in DNA	1.1.11			x	bcbio-nextgen toolkit v.2.0,
		2.2.6			x	
		2.2.9	x			

		2.2.10		x		
Jbrowse	JBrowse is a fast, scalable genome browser built completely with JavaScript and HTML5.					Genomics Tookit, Pop-up genome broswer
kallisto	Quantifies abundances of transcripts from RNA-Seq data, or more generally of target sequences using high-throughput sequencing reads	0.42.3		x		
		0.43.0		x	x	
		0.43.1	x			
khmer	Tools to work with DNA shotgun sequencing data from genomes, transcriptomes, metagenomes, and single cells	2.0.0			x	
		2.1.1	x			

Kraken	System for assigning taxonomic labels to short DNA sequences	0.10.5			x	bcbio-nextgen toolkit v.2.0
		1			x	
		2.0.7			x	
ldhat	LDhat is a package of programs for the analysis of recombination from population genetic data	2.2		x		
ldhot	LDhot is a package for inferring the location of recombination hotspots from patterns of linkage disequilibrium within samples	2014		x		
limma	Data analysis, linear models and differential expression for expression data	3.0.0		x		
lincrna	Identification, Characterization, and Quantification of Long Intergenic Non-Coding RNAs	1.0.0		x		
Longranger	analysis pipelines that processes Chromium sequencing (10X)output	2.1.6	x			genome assembly workshop
Loupe	Loupe Cell Browser gene expression tutoria					genome assembly workshop
mach	Markov Chain based haplotyper	1.0.1		x		

macs	Model-based Analysis of ChIP-Seq, for identifying transcript factor binding sites	1.4.2		x		bcbio-nextgen toolkit v.2.0,
		1.4.3			x	
		2.1.0	x			
		2.1.2	x			
MACSE	Aligns coding NT sequences with respect to their AA translation while allowing NT sequences to contain multiple frameshifts and/or stop codons	1.2			x	
MAFFT	Multiple sequence alignment program	7.3			x	
manta	Structural variant and indel caller for mapped sequencing data	1.4.0				Genomics Toolkit
maker	ab initio gene predictions	2.31.10			x	
		2.31.9	x			
mash	MinHash dimensionality-reduction technique for clustering abd searching massive sequence collections	2.1.1	x			
masurca	assembler combines the benefits of deBruijn graph and Overlap-Layout-Consensus assembly approaches.	3.2.0	x			
		3.1.3			x	
		3.2.2			x	
		3.2.6			x	

		3.2.7			x	
mcl	Scalable unsupervised clustering algorithm	14-137	x			
megahit	A single node assembler for large and complex metagenomics NGS reads, such as soil	1.1.2	x			Arkansas Workshop
		1.1.1			x	
MALT	MEGAN Alignment Tool, and extension of MEGAN	0.3.8			x	
Mapsembler 2	Targeted assembly software	2.2.4			x	
MaxBin	Software for binning assembled metagenomic sequences based on an Expectation-Maximization algorithm	2.1.1			x	
MEGAN	MEta Genome ANalyzer	5.11.3			x	MEGAN community edition
		6	x			
metabat	Metagenome binning tool	2.11.3	x			Arkansas Workshop
Meraculous	Whole genome assembler for Next Generation Sequencing data geared for large genomes	2.2.4			x	

MetaPhlAn	Computational tool for profiling the composition of microbial communities from metagenomic shotgun sequencing data	1.7.7			x	
		2.6.0			x	
MetaVelvet	Extension of Velvet assembler to de novo metagenome assembly from short sequence reads	1.2.10			x	
Metapalette	k-mer based bacterial community reconstruction technique	1	x			
migrate	Estimates effective population sizes and past migration rates between n population	3.3.2		x		
mimar	Markov chain Monte Carlo method to estimate parameters of an isolation-migration model	121720 10		x		
MinCED	Program to find CRISPRs in full genomes or environmental datasets such as metagenomes	0.2.0			x	
minia	Minia is a short-read assembler based on a de Bruijn graph	2.0.7	x			

minimac	A low memory, computationally efficient implementation of the MaCH algorithm for genotype imputation	11162012		x		
mira	assembler	4.0.2			x	
mirdeep2	discovers active known or novel miRNAs from deep sequencing data	0.1.2		x		
miso	Mixture-of-Isoforms is a probabilistic framework that quantitates the expression level of alternatively spliced genes from RNA-Seq data	0.4.6		x		
mothur	Bioinformatics tool for analyzing 16S rRNA gene sequences	1.31.2		x		
		1.34.2		x		
		1.38.1			x	
		1.40.5	x			
		1.41.3	x			
mrbayes	MrBayes is a program for Bayesian inference and model choice across a wide range of phylogenetic and evolutionary models	3.2.1 mpi		x		Genomics Tookit

		3.2.6		x		
mrsfast	mrsFAST is designed to map short reads to reference genome assemblies in a fast and memory-efficient manner	3.3.7		x		
MUctect	Identification of somatic point mutations in next generation sequencing data of cancer genomes.					bcbio-nextgen toolkit v.2.0
MultiQC	Aggregate results from bioinformatics analyses across many samples into a single report					Genomics toolkit
mummer	System for rapidly aligning entire genomes, whether in complete or draft form	3.23	x	x	x	
muscle	MUSCLE is one of the best-performing multiple alignment programs according to published benchmark tests	3.8.31	x	x	x	
MyCC	Automated binning tool that visualizes metagenomes and and identifies reconstructed genomic fragments	42341			x	
nanopack	long read processing and analysis	0.1.14	x			
Nasp	pipleine to identify SNP's in the data.	1.1.2	x			

ngsutils	NGSUtils is a suite of software tools for working with next-generation sequencing datasets	0.5.2a		x		
NGSCheck Mate	Identifies NGS data files from the same individual	2016.1 0.12			x	
ninja	Infers phylogeny using neighbor-joining tree	1.2.2		x		
novoalign	Aligns short reads to reference genome for resequencing experiments	3. .0	x			bcbio-nextgen toolkit v.2.0
oases	De novo transcriptome assembler for very short reads	0.2.09	x			
OpenRefine *	open source, power tool for working with messy data.					EML Workshop
paml	Phylogenetic analyses of DNA or protein sequences using maximum likelihood	4		x		
		4.9		x	x	
paup	Phylogenetic Analysis Using Parsimony is a computational phylogenetics program for inferring evolutionary trees	4	x	x		

PBJELLY	Highly automated pipeline that aligns long sequencing reads (such as PacBio RS reads or long 454 reads in fasta format) to high-confidence draft assembles	15.8.24			x	genome assembly workshop
perl	Programming language				x	Arkanasas Workshop,
phenix	For reference mapping, VCF genearation, VCF filtering, and SNP identification	4-Jan	x			
phylip	PHYLogeny Inference Package	3.69	x	x	x	
Phylosift	Tool suite to conduct phylogenetic analysis of genomes and metagenomes	1.0.1			x	
picard	manipulating high-throughput sequencing (HTS) data and formats.	2.14.0	x			Genomics Toolkit
		2.17.0			x	
		2.1.1			x	
		2.18.0			x	
		2.20.2			x	

Pilon	Automatically improves draft assemblies and finds variation among strains, including large event detection	1.16			x	
		1.23	x			
Platanus	De novo sequence assembler for NGS data	1.2.4			x	
plink	Open-source C/C++ library for working with human genetic variation data	1.07		x		
		2	x	x		
		1.9		x		
polyphen	automatic tool for prediction of possible impact of an amino acid substitution on the structure and function of a human protein	2.2.2	x			

Pplacer	Places query sequences on a fixed reference phylogenetic tree to maximize phylogenetic likelihood or posterior probability according to a reference alignment	1.1.7			x	
Primer3	PCR design tool	1.1.4				
		2.2.3				
		2.3.7				
Prinseq	Quality control for high-throughput sequence data	1.2	x			Arkansas Workshop
Prokka	Software tool for rapid annotation of prokaryotic genomes	1.11			x	
Prodigal	Fast, reliable protein-coding gene prediction for prokaryotic genomes	2.6.2			x	
		2.6.3			x	
psi4	PSI4 provides a wide variety of quantum chemical methods using state-of-the-art numerical methods and algorithms	1.1	x		x	
python	Programming language				x	Arkansas Workshop,
QIIME2	Quantitative Insights Into Microbial Ecology	2.0.0				QIIME2
QUAST	Assembly evaluation tool	4.6.3	x			Arkansas Workshop, genome assembly workshop

		5.0.0	x			
R	statistical computing and graphics	3.2.3			x	EML Workshop, R for Biologists, Harvesting Field Station data
Rstudio	open-source integrated development environment (IDE) for R, a programming language for statistical computing and graphics.	3.3.1			x	
		3.3.3			x	
		3.4.1			x	
		3.5.1			x	
		3.5.2			x	
		3.6.0			x	
r8s	Estimates absolute rates ('r s') of molecular evolution and divergence times on a phylogenetic tree	1		x		
raxml	Maximum likelihood phylogeny estimation for interpreting relationships between sets of data	8.0.26		x		Genomics Toolkit
		8.2.11	x			
		8.2.9			x	

Raxml-ng	phylogenetic tree inference tool which uses maximum-likelihood (ML) optimality criterion.	0.9.0	x			
		0.9.0-pthreads	x			
Ray	Ray is a parallel software that computes de novo genome assemblies with next-generation sequencing data	2.3.1	x		x	
reap	estimating kinship coefficients and identify-by-descent (IBD) sharing probabilities in samples with admixed ancestry	1.2		x		
repeatmasker	Screens DNA sequences for interspersed repeats and low complexity DNA sequences	4.0.7	x	x		
		4.0.6			x	

repeatmodeler	Screens DNA sequences for interspersed repeats and low complexity DNA sequences	4.0.6				
		1.0.10	x			
		1.0.		x		
rmats	computational tool to detect differential alternative splicing events from RNA-Seq data	3.2.5	x			
		4.0.2	x			
RNAmmer	Predicts 5s/ s, 16s/1 s, and 23s/2 s ribosomal RNA in full genome sequences	1.2			x	
rosetta	Software suite including algorithms for computational modeling and analysis of protein structures	3.5		x		
		3.7			x	
		3.8		x		

		3.1	x			
rsem	Estimates gene and isoform expression levels from RNA-Seq data	1.2.19		x		
		1.2.21			x	
		1.3.0	x			
sailfish	Alignment-free isoform quantification from RNA-seq reads using lightweight algorithms	1.2.21				
		0.7.3		x		
		0.8.0		x		
		0.9.2			x	
salmon	Tool for fast transcript quantification from RNA-seq data	0.4.2		x		bcbio-nextgen toolkit v.2.0
		0.6.0			x	bcbio-nextgen toolkit v.2.0, Arkansas Workshop,
		0.7.2			x	
		0.8.1			x	
		0.11.0	x	x		
		0.9.1	x		x	
samtools	Utilities for manipulating	1.3		x	x	Mining SRA to identify datasets, Genomics

	alignments in SAM format	1.3.1			x	Toolkit, genome assembly workshop
		1.5	x			
		0.1.1		x		
		1.9	x		x	
		0.1.19				
		1.2		x		
		1.7			x	
Seqtk	Toolkit for processing sequences in FASTA/Q formats	1.2-r94			x	
SignalP	Predicts the presence and location of signal peptide cleavage sites in amino acid sequences from different organisms: Gram-positive prokaryotes, Gram-negative prokaryotes and eukaryotes	4.1c			x	
skmer	estimating distances between genomes from low-coverage sequencing reads	3.0.1	x			

SNVMIX	Detects single nucleotide variants from next generation sequencing data	0.11.8-r5			x	
smrt	Automated and distributed secondary analysis of sequencing data generated by the PacBio single-molecule, real-time (SMRT) sequencing system	2.2.0		x		PacBio SMRT analaysis
		5.1.0			x	
		7.0.1			x	
somaticsniper	SomaticSniper is a program to identify single nucleotide positions that are different between tumor and norma	1.0.5			x	
soapdenovo	De novo assembler for next generation sequencing reads	r240		x		
		9/10/2015			x	
		2.04	x			
soapdenovotrans	A de novo transcriptome assembler inherited from the SOAPdenovo2 framework, designed for assembling transcriptome with alternative splicing and	1.03	x			

	different expression level					
solar eclipse	SOLAR-Eclipse is an extensive, flexible software package for genetic variance components analysis, including linkage analysis, quantitative genetic analysis, SNP association analysis (QTN and QTLD), and covariate screening	.3.1		x		bcbio-nextgen toolkit v.2.0
spades	St. Petersburg genome assembler - is intended for both standard isolates and single-cell MDA bacteria assemblies	3.10.1	x	x	x	Genomics Toolkit
		3.11.1	x		x	
		3.8.1			x	
		3.8.2		x		
		3.9.0		x		
spm	Statistical Parametric Mapping		x	x		
		12	x	x		

sra-toolkit	Collection of tools and libraries for using data in the INSDC Sequence Read Archives	2.3.5-2		x		Mining SRA to identify datasets
		2.5.4		x		
		2.8.1			x	
		2.8.2	x			
		2.9.1			x	
stacks	A modular pipeline for building loci from short-read sequences	1.4	x			
star	Spliced Transcripts Alignment to a Reference	2.5.3	x			Genomics Toolkit
		2.6.1a	x			
		2.4.1		x		
		2.5.2		x	x	
		1.3.3b				
		2.5.4			x	
		2.7.0			x	
Strelka	Somatic variant calling workflow for matched tumor-normal samples	1.0.01			x	
starfusion	Spliced Transcripts Alignment to a Reference	1.1.0	x		x	
		1.3.1			x	

		1.5.0			x	
stringtie	Fast and highly efficient assembler of RNA sequence alignments into potential transcripts	1.3.3b	x		x	Genomics Toolkit
Supernova	Supernova is a software package from 10X Genomics for de novo assembly from Chromium Linked-Reads that are made from a single whole-genome library from an individual DNA source	0.2.6				genome assembly workshop
tabix	Tabix works on generic tabular data formats for genomic information, quickly retrieving features overlapping specified areas on the genome	0.26	x	x		Genomics Toolkit
Tcoffee	Multiple Sequence Alignment Server					Genomics Toolkit

Theano	Python library that allows you to define, optimize, and evaluate mathematical expressions involving multi-dimensional arrays efficiently	0.8.0			x	
		0.8.2			x	
TMHMM	Predicts transmembrane helices in proteins	2			x	bcbio-nextgen toolkit v.2.0,
tophat	Fast splice junction mapper for RNA-Seq reads	2.0.7		x		Genomics Toolkit
		2.1.0		x	x	
		2.1.1	x		x	
tophat2	Splice junction mapper for RNA-Seq reads	4.6.2				bcbio-nextgen toolkit v.2.0,
		3.14				
		2.1.1	x			
transabyss	Analysis for ABySS-assembled contigs from shotgun transcriptome data for finding splice sites and variants	1.0.3				
		1.5.5	x			
		2.0.1	x		x	
transdecoder	TransDecoder identifie s candidate coding regions within transcript sequences	0.5.1				

		0.4.5				
		3.0.1			x	
Transrate	Software for de-novo transcriptome assembly quality analysis	1.0.3	x		x	
trim galore	A wrapper tool around Cutadapt and F astQC to consistently apply quality and adapter trimming to FastQ files	2.1.1				Arkansas Workshop
		0.4.5	x		x	
		0.6.3			x	
trimmomatic	Fast, multithreaded command line tool that can be used to trim and crop Illumina (FASTQ) data as well as to remove adapters.	0.36	x	x	x	Genomics Toolkit
		0.35		x		
trinity	Package which enables the efficient and robust de novo reconstruction of transcriptomes from RNA-Seq data	2.3.2			x	Arkansas Workshop,
		2.4.0	x	x	x	
		2.0.1				
		2.6.6	x			
		2.8.4	x	x	x	
		2.1.1		x	x	
		2.0.6			x	
		2.2.0			x	

Trinotate	Annotation suite designed for automatic functional annotation of transcriptomes, particularly de novo assembled transcriptomes, from model or non-model organisms	3.1.1	x		x	
		2.0.2			x	
Tombo	modified nucleotides from nanopore sequencing data	1.5	x			
vcftools	Tool providing easily accessible methods for working with complex genetic variation data in the form of VCF files	0.1.14		x		bcbio-nextgen toolkit v.2.0, Genomics Toolkit
		0.1.13	x	x		
		1.2.10				
		0.1.10		x		
		0.1.17			x	
		0.1.15			x	
velvet	De novo genomic assembler specially designed for short read sequencing technologies	1.21	x		x	
warbleR	Bioacoustics research encompasses a wide range of questions, study systems and methods, including the software used for analyses.					Harvesting Field Station Data

7. Acknowledgements

This document was developed with support from National Science Foundation (NSF) grant ABI-1759906 (2018). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation, Indiana University, or other funding agencies.