ABI Sustaining: The National Center for Genome Analysis Support 2018 Annual Report

Thomas G. Doak https://orcid.org/0000-0001-5487-553X
Craig A. Stewart https://orcid.org/0000-0003-2423-9019
Scott D. Michaels

Indiana University

PTI Technical Report PTI-TR18-004

September 10, 2018

Please cite as:

Doak, T.G., Stewart, C.A., Michaels, S.D. (2018) "ABI sustaining: National center for genome analysis support 2018 annual report," Indiana University, Bloomington, IN. PTI Technical Report PTI-TR18-004. Retrieved from http://hdl.handle.net/2022/22402

PERVASIVE TECHNOLOGY INSTITUTE
INDIANA UNIVERSITY

NATIONAL CENTER FOR GENOME ANALYSIS SUPPORT
INDIANA UNIVERSITY

Table of Contents

Table of Tables

Table of Figures

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:

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.

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 just been awarded its second three year Sustaining award, DBI-1759906, to start in Sept. 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.

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 Cluster (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 third year of this sustaining award (a fourth year of no-cost extension has just started), 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 works ranging from ecology, to physiology, to economically important animals and plants.

NCGAS serves researchers in 41 states and Puerto Rico (Fig. 1), including 12 EPSCoR states, and in partnership with the Trinity development team (see Synergistic activities) has users around the world (Fig. 5).

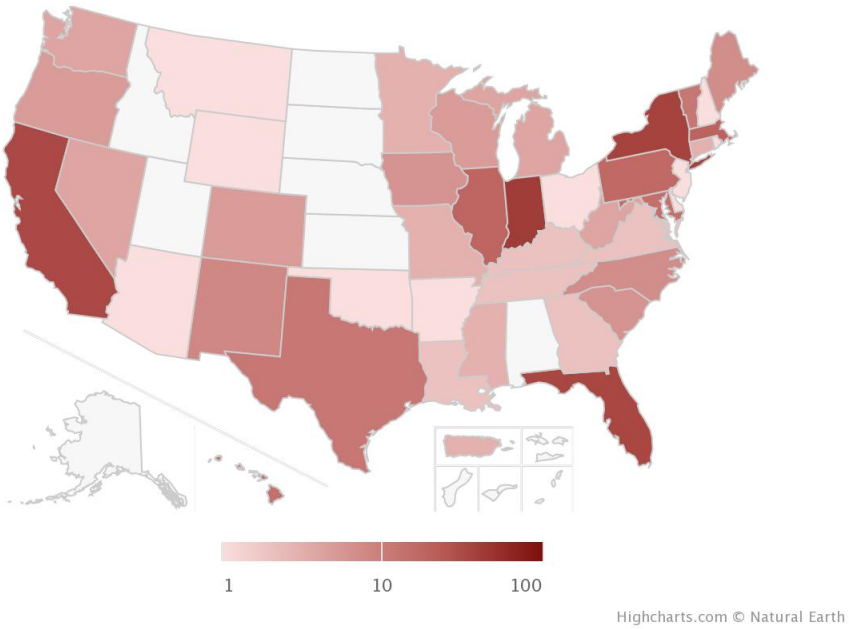

Figure 1-1. States and Territories with NCGAS Users

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, and metagenome and metatranscriptome assemblies.

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

- Southern California *Kelp Forest microbiome*
- Insights into Cyclophyllidea (tapeworms) evolution studying *N. mogurndae* (fish parasites)-genome/transcriptome analysis
- Evaluation of genomic evidence for rock-paper-scissors and arms-race dynamics in a community of pathogenic bacteria
- Coral resilience via host-microbial interactions
- Impact of hurricane Maria on the mesophotic reefs of Southwest Puerto Rico
- Transcriptomic assessment of the tree swallows in the Great Lakes Area
- Coffee, peanut, and sweet potato transcriptomes
- Daphnia genomes (population genomics and *de novo* assemblies)
- Barred tiger salamander transcriptomes
- Diatoms transcriptomes
- The diverse microbial clade Stramenopila + Alveolata + Rhizaria (SAR) (*de novo* genomes, transcriptomes)
- Heliconius butterflies (transcriptomes)
- Carrion Flies (forensic arthropods), *de novo* genome assembly/resequencing, transcriptomes
- Mussels
- Strawberry Dart Frog
- Song birds
- Polyploid salamanders (threatened)
- Bahama Giant conch (a mollusk)

- Little Brown Skate
- *D. galeata*

NCGAS supported 223 biologists using genome analysis (25 named new allocations) during the current year, including 277 short consultations and 16 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. In a recent consult, we informed a developmental ecologist that the samples they were going to bring back from SE Asia were inappropriate to their research interests, and redesigned their research direction. 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). For example, this year we installed a large number of tools to facilitate the use of Oxford Nanopore's new sequencing platform, the MinION. Tools such as base-callers (Albacore, nanopack, Canu) and other support software (poretools and poreduck) were installed to allow Nanopore sequencing support.

This year we installed or updated five packages (spades, hisat2, salmon, STAR, kallisto, ballgown) for assembly and analysis of PacBio long-read sequencing data. Hybrid assemblies are quickly becoming popular, and we installed 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 large memory *Mason* IU cluster (now retired)
- 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, etc.

Overall, we have installed and maintained 82 software packages (not including versions and Galaxy VMs) across the systems described above (see Appendix 4) in the past year.

1.2.1.3. Disseminate tools for genome assembly and analysis

Jetstream VMs: NCGAS has added 13 preconfigured virtual machines to the Jetstream library to help researchers with their visualization (genome browsers, Anvi'o, MEGAN), for training (R for Biologists, Arkansas workshop image) and in setting up their private images. 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).

Jetstream image name	About the image	Number of Instances Launched
Bcbio-nextgen toolkit v.2.0	Validated, scalable, community developed variant calling, RNA-seq, and small RNA analysis	1
BioLinux 8	Scientists handling and analyzing biological data with more than 250 bioinformatics packages	561
Web server with wiki	Hosts the LAMP (LINUX, Apache, MySQL, PhP) stack to make web service easier	5
Pop Up Browser with Tripal, JBrowse, BLAST	Hosts a LAMP stack, Tripal (build on Drupal 7), JBrowse, BLAST, and custom tools for quick starting a genome browser	17
MEGAN (MEtaGenome ANalyzer) v6	Analysis of large metagenomic datasets	1
Anvi'o v5	Open-source, community-driven analysis and visualization platform for 'omics data.	13
Drupal v7.59	Content management software used to make many of the websites and applications you use every day	3
Arkansas Workshop	The focus of this workshop was to apply bioinformatic analysis in Jetstream environment.	35
Ubuntu16_04 Rstudio for NCGAS	The goal of the workshop is to help biologists get acquainted with R, which will, in turn, help them with their analysis.	6
PEARC18 Hackathon-Developing and Applying Best Practise Protocols	The theme of this hackathon is reproducible and portable computing.	11
Tripal and Jbrowse on Ubuntu	Private image	6
Junco Browser	Private genome browser image for a research group	1
Unbuntu16_04 Rstudio OpenRefine	Private image with software installed for a research group	11

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

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 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. In addition, the NCGAS team spearheaded an effort to teach, use, and support Singularity on HPC during the PEARC 2018 conference by organizing and running a Hackathon on "Developing and Applying Best-Practice Protocols" during the event.

GitHub repository: NCGAS has a GitHub account with four active projects, where NCGAS developed scripts and pipelines are shared with the community. Two active projects include the developed *de novo* transcriptome pipeline with all the scripts to run the pipeline, and another repository specific to teach at workshops with example data and more documentation. The other project includes scripts and material taught at another workshop hosted in University of Arkansas. The fourth project is under active development to connect galaxy gateway at IU to PSC Bridges.

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 an active globus endpoint allowing users to transfer terabytes of their data quickly to and from SDA, and IU/XSEDE clusters.

In an effort to extend more tools to our users, we obtained two additional XSEDE allocations - one for extending our services to include biological research stations (valued at $25,157.77) 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). For example:

- Sheri Sanders taught at the MDI Biological Laboratory's Environmental Genomics course (now her third year there).
- NCGAS conducted a two-day course in RNAseq analysis to a national group of participants (see below).
- Jetstream and NCGAS personnel traveled to University. of Arkansas, Fayettville. They taught two days of Jetstream use and RNAseq assembly, organized by collaborator Jeff Pummill.
- NCGAS had 10 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 NCBI SoCal Bioinformatics Hackthon at San Diego State University in January developing a bioinformatics tool to search for closely related viruses in SRA.
- 197 participants in tutorials (~983 contacts at events). In the last year, some of the events attended included: Practice & Experience in Advanced Research Computing Conference Series (PEARC) 18, Pittsburgh Pennsylvania; Plant and Animal Genomes 2018, San Diego; MDI Biological Laboratory's 2017 Environmental Genomics course, Bangor, Maine; 15th Annual Rocky Mountain Bioinformatics Conference, Snowmass, Colorado, 2017; American Association for

Cancer Research, Chicago, 2018; Galaxy Community Conference, Portland Oregon 2018; Evolution, Portland 2017; Partners in Amphibian and Reptile Conservation Meeting, Michigan, 2018; Informatics Technology for Cancer Research (ITCR), Maryland, 2018.

1.2.2. Specific Objectives

There are currently 206 packages supported by NCGAS (between Carbonate, Karst, and Bridges), with 13 Jetstream images (Appendix 4).

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 277 short term consulting engagements (those taking less than 4 hours of staff time to resolve) and 16 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 (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 weekdays, 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. The 30 attendees came from across the country and represented a spectrum from graduate students to professors. Pre- and post-course questionnaires indicate that the workshop was very well received, and we will repeat it in Fall 2018 (it is again oversubscribed). We will also use this experience to develop more courses of this form (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 third 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 22 peer-reviewed scientific publications published in 2017-2018 (beyond those reported for the first two years).

1.2.5. Results of the 2018 user survey

Our third user survey was conducted during June 2018, with 115 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 Assistant 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 meeting; she also led one of the admin training sessions for the conference. Ms. Ganote oversees many projects, mentored Sheri Sanders, and is now mentoring Bhavya Papudeshi. She has been promoted from IT3 to IT4, reflecting her essential role in our organization.

Staff member Sheri Sanders has been a NCGAS team member for two 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 been working with Keithanne Mockaitis on polyploid crop species, expanding her analytical skills in these systems. 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. She has been working on gaining system administration skills, attending the Linux/Container Con in 2016 and Linux Cluster Institute's Intermediate System Admin Training in 2018. She has been developing cloud images to lead our efforts to support genome browsers and contribute to the GMOD development community. NCGAS is happily providing 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 workshops, developing skills in orchestration of events, curriculum development, and reporting.

Staff member Bhavya Papudeshi has been an NCGAS employee for one year, 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; Elizabeth Dinsdale at SDSU on analyzing Kelp Forest microbiomes, coral microbiomes; and with Marla and Michael Douglas (U. of Arkansas) on a Cyclophyllidea tapeworm transcriptome (of an invasive fish). In the last year, Papudeshi has been trained to improve her understanding of High Performance Clusters: job schedulers (TORQUE, SLURM), maintaining gateways (galaxy and Genepattern), setup preconfigured VMs on cloud computing resource, and learning R to better help the NCGAS userbase.

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:

- 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, https://sciencenode.org
- NCGAS web site at ncgas.org
- NCGAS Twitter and Blog accounts
- In-person contacts
- Email list distribution
- Newsletters
- Github-NCGAS

1.4.2. Social Media Efforts

NCGAS has a Twitter page with 79 followers and a total of 207 tweets currently. On average, we were able to reach out to approximately 120 profile visits and 1.13 % engagement rate with an average of 30 link clicks and 43 likes per month. In addition to Twitter, NCGAS has a Facebook presence as well. At current, our Facebook page has 41 followers and 40 likes. Our Facebook and Twitter pages are linked so that all tweets are posted on the facebook page.

Our website (ncgas.org) is our main online content engine. Between Aug 1, 2017 and July 31, 2018, we had 22,920 page views (10,398 unique users). 19.45% of this was on our main page, followed by our blog (5% of page views), our bioinformatics services page (4% of pageviews), our R workshop page (4% of

pageviews), and our software list (4% of pageviews). Just under 50% of people landing on our website are finding it organically through Google and 50% are from the United States.

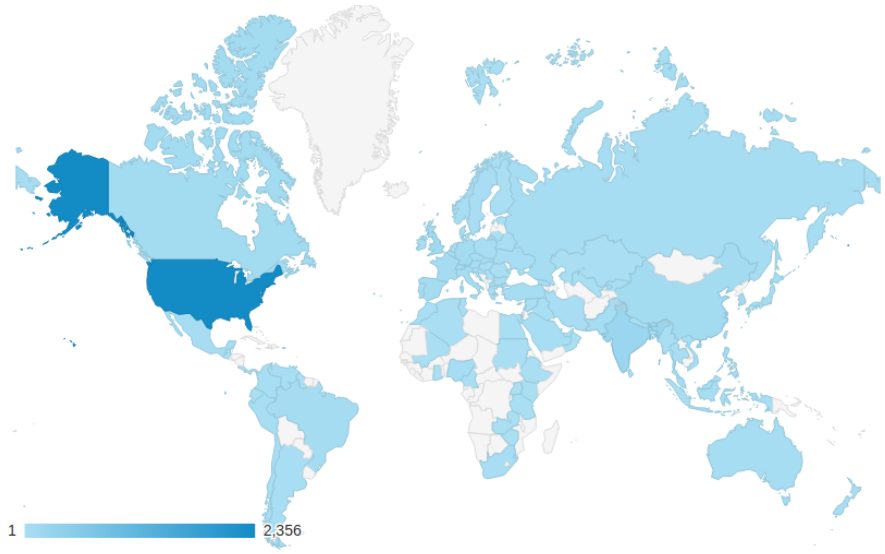

Figure 1-2. Sources of hits to the NCGAS web page.

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

1.5.1. 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. Our newest employee Bhavya Papudeshi has a background in metagenomics, and we are collaborative researchers on four ABI metagenomics proposals.
3) Continueing 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, and have a second planned for this fall. We also hope to take it on the road at least a couple of times in the next year. We have also developed a successful R workshop for biologists. We next plan to develop a bash/python/R workshop for

genomicists (Appendix 1). We are now ready to video these presentations for posting to YouTube, but the hands-on support during the RNAseq workshop is 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 2018 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 2019.
- At least one member will attend the AACR national conference.
- We are invited to visit the U. of Nebraska to introduce national HPC resources and NCGAS services, similar to last fall's trip to Arkansas. This will probably be another collaboration with Jetstream.
- 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.

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

In the last year, we worked to engage field stations and field biologists. This is motivated by our focus on biology, and our relationship with the Jetstream team. Last 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. Since then we 1) have provided VMs for groups to present workshops with (they didn't actually use them, but they carry all the tools the courses used, and are available to graduates of the courses); 2) are organizing a one day satellite workshop to the 2018 OBFS "Data generation, management, and storage to meet the needs of field stations" designed to bring service providers and station representatives into conversation. While we needed to cancel this workshop for financial constraints, we will have three representatives at the 2018 meeting: Sheri Sanders (NCGAS), Jennifer Laherty (IU Science Librarian, Craig Stewart (PTI Exec. Director and original Jetstream PI); 3) are using a summer REU program to develop a pipeline linking environmental sensors to Jetstream and to Wrangler.

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 Galaxy development community. NCGAS member Carrie Ganote is on the funding committee and the scientific board for the 2018 conference and will lead one of the admin training sessions for the conference.
- 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 3 XSEDE service provider. Tom Doak is currently the SP board representative for level 3 providers.
- NCGAS is a collaborator on two NIH Information Technology in Cancer Research (ITCR) funded projects: 1) development and Galaxy hosting of the Trinity *de novo* RNAseq assembly tools, and 2) hosting of the GenePattern genomics tool set. While both these tool sets are developed for cancer research, they are also useful to other disciplines. Trinity in particular is extensively used by our non-medical clients, esp. 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. GenePattern is used mostly by medical clients to run genomic analysis focusing on gene expression, single nucleotide polymorphism, flow cytometry analysis. The IU Trinity Galaxy has 861 registered users (58 countries) (Fig. 5) and IU GenePattern has 538 registered users (37 countries) (Fig. 6) giving NCGAS a global reach. While the Trinity development team will move away from the Galaxy platform, NCGAS intends to continue providing this service.

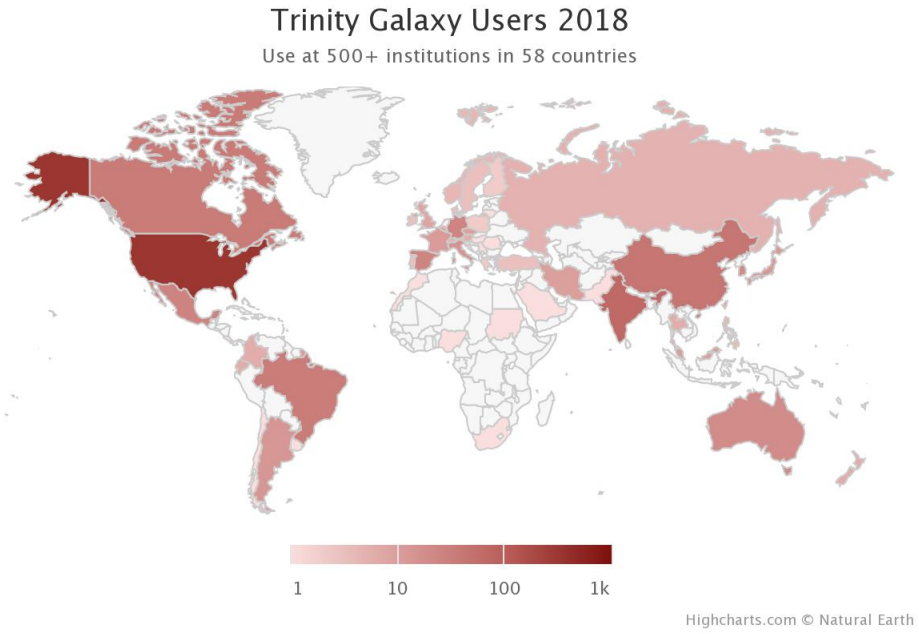

Figure 1-3. Users of the NCGAS-supported Trinity Galaxy instance worldwide

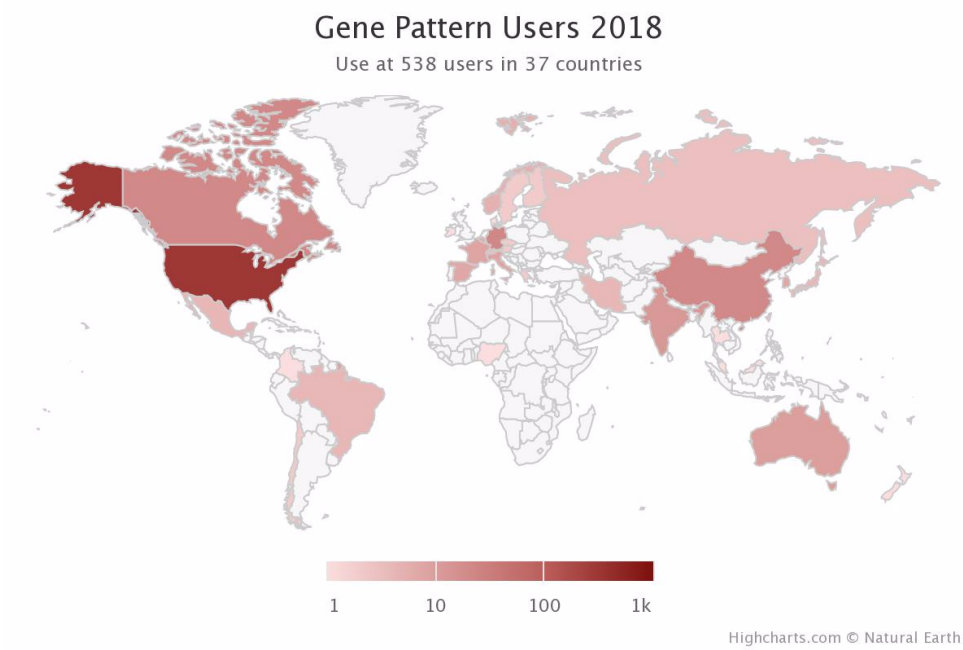

Figure 1-4. Users of the NCGAS-supported IU GenePattern instance worldwide

- Keithanne Mockaitis specializes in plant transcriptomics, and as such has been and is a member of many national and international consortiums characterizing the transcriptomes of commercial plants, including mango, cocoa, and loblolly pine. Current projects include coffee, peanut, and sweet potato. Both Ganote and Sanders are paid from NSF and USDA grants Mockaitis is a co-PI on. The NCGAS RNAseq pipeline was developed in collaboration with Mockaitis and Don Gilbert (IU).

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 conserving this software and genomics issues. NCGAS personnel teach R workshops (R for biologists, CeWIT R class), deliver guest lectures (Bioinformatics2Go, EcoLunch, Reproductive Diversity), and participate in REU programs. We participate in a local project to identify bacteriocin regions in *Xenorhabdus sps* role in Xenorhabdus-nematode-insect interaction. NCGAS personnel will mentor three Center of Excellence for Women in Technology (CEWiT) freshman and sophomore students this year.

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

- Arafat, A. *et al.* Unexplained Early Infantile Epileptic Encephalopathy in Han Chinese Children: Next-Generation Sequencing and Phenotype Enriching. *Sci. Rep.* 7, 46227 (2017).
- Boyd, B. M. *et al.* Phylogenomics using Target-Restricted Assembly Resolves Intrageneric Relationships of Parasitic Lice (Phthiraptera: Columbicola). *Syst. Biol.* 66, 896–911 (2017).
- Bradshaw, W. E. *et al.* Evolutionary transition from blood feeding to obligate nonbiting in a mosquito. *Proc. Natl. Acad. Sci.* 115, 201717502 (2017).
- Choudhary, S. *et al.* Transcriptome characterization and screening of molecular markers in ecologically important Himalayan species (Rhododendron arboreum). *Genome* 61, 417–428 (2018).
- Christie, A. E. *et al.* Prediction of a peptidome for the ecotoxicological model Hyalella azteca (Crustacea; Amphipoda) using a de novo assembled transcriptome. *Mar. Genomics* 38, 67–88 (2018).
- Harke, M. J. *et al.* Periodic and coordinated gene expression between a diazotroph and its diatom host. *ISME J.* (2018). doi:10.1038/s41396-018-0262-2
- Jilling, T. *et al.* Surgical necrotizing enterocolitis in extremely premature neonates is associated with genetic variations in an intergenic region of chromosome 8. *Pediatr. Res.* 83, 943–953 (2018).
- Maralit, B. A., Jaree, P., Boonchuen, P., Tassanakajon, A. & Somboonwiwat, K. Differentially expressed genes in hemocytes of Litopenaeus vannamei challenged with Vibrio parahaemolyticus AHPND (VPAHPND) and VPAHPNDtoxin. *Fish Shellfish Immunol.* 81, 284–296 (2018).
- Roncalli, V., Cieslak, M. C., Sommer, S. A., Hopcroft, R. R. & Lenz, P. H. De novo transcriptome assembly of the calanoid copepod *Neocalanus flemingeri*: A new resource for emergence from diapause. *Mar. Genomics* 37, 114–119 (2018).
- Roncalli, V. *et al.* A deep transcriptomic resource for the copepod crustacean *Labidocera madurae*: A potential indicator species for assessing near shore ecosystem health. *PLoS One* 12, e0186794 (2017).
- Winker, K., Glenn, T. C. & Faircloth, B. C. Ultraconserved elements (UCEs) illuminate the population genomics of a recent, high-latitude avian speciation event. *PeerJ Prepr.* (2018). doi:https://doi.org/10.7287/peerj.preprints.27035v1

2.1.2. Licenses

2.1.3. Other Conference Presentations / Papers/Reports

- Anderson, J., Slayton, T., Guido, E., Doak, T. & Sanders, S. Harvesting Field Station Data; Raspberry Pi Sensors to Jetstream Databases. (*Practice & Experience in Advanced Research Computing* (PEARC) 18 Pittsburgh 2018).
- Doak, T., Ganote, C., Sanders, S. & Hallock, B. NCGAS: National Resources for Computationally Intensive Bioinformatics. in *Practice & Experience in Advanced Research Computing* (PEARC 2017 New Orleans, 2017).
- Ganote, C., Mendes, F., Henschel, R., Hahn, M. & Fulton, B. Introducing CAFE: Computational Analysis of (gene) Family Evolution. (Plant and Animal Genomes (PAG) 2018, San Diego 2018).

- Ganote, C., Sanders, S., Wu, L.-S., Doak, T. & Mockaitis, K. Solving the challenges of complex genome analysis collaborations on-line using XSEDE resources Plant and Animal Genomes (PAG) 2018 San Diego Poster. (2018).
- Doak, T. G., Stewart, C. A., Wernert, J., Hancock, D. Y. & Miller, T. Summary of the Survey of Field/Marine Station Directors and Managers. (Report 2017). http://hdl.handle.net/2022/22009
- Doak, T. G., Stewart, C. A. & Michaels, S. D. Summary of the 2017 NCGAS User Survey. (Report, 2017). http://hdl.handle.net/2022/21614
- Stewart, C. *et al.* Pervasive Technology Institute Annual Report: Research Innovations and Advanced Cyberinfrastructure Services in Support of IU Strategic Goals During FY 2017. (Report 2017). http://hdl.handle.net/2022/21809
- Papudeshi, B. Reconstruction of Metagenome-Assembled Microbial Genomes from a Micro-biome Plant and Animal Genomes (PAG) San Diego 2018 Poster. (2018).
- Papudeshi, B., Sanders, S., Ganote, C., Fischer, J. & Doak, T. Bioinformatic analysis using Jetstream, a cloud computing environment. Plant and Animal Genomes (PAG) San Diego (2018).
- Sanders, Sheri; Podicheti, Ram; Yang, Tianhong; Fang, Lingling; Jayanthi, Srinivas; Rajan, Gayathri; Kumar, Thallapuranam Krishnaswarmy Suresh; Medina-Bolivar, Fabricio; Mockaitis, Keithanne. Stilbenoid prenylation pathway discovery in peanut using targeted transcriptomics Plant and Animal Genomes (PAG) San Diego 2018 Poster). (2018).

2.1.4. Other Products

2.1.4.1. *Blog Posts*

- Adding Users to Jetstream. Available at: https://ncgas.org/Blog_Posts/Adding Users to Jetstream VMs.php. (Accessed: 6th September 2018)
- Getting started on Bridges. Available at: https://ncgas.org/Blog_Posts/Getting started on Bridges.php. (Accessed: 6th September 2018)
- Getting Started on Globus. Available at: https://ncgas.org/Blog_Posts/Getting Started with Globus.php. (Accessed: 6th September 2018)
- Getting Started on Jetstream. Available at: https://ncgas.org/Blog_Posts/Getting Started on Jetstream.php. (Accessed: 6th September 2018)
- Getting Started with Ansible. Available at: https://ncgas.org/Blog_Posts/Ansible.php. (Accessed: 6th September 2018)
- Installing conda packages locally. Available at: https://ncgas.org/Blog_Posts/Installing conda packages locally.php. (Accessed: 6th September 2018)
- Join NCGAS at PAG XXVI! Available at: https://ncgas.org/Blog_Posts/PAG2018 Abstracts.php. (Accessed: 6th September 2018)
- Mason gets a replacement. Available at: https://ncgas.org/Blog_Posts/Mason gets a replacement.php. (Accessed: 6th September 2018)
- NCGAS has now added XSEDE resources to our field station support! Available at: https://ncgas.org/Blog_Posts/Field Station Allocations.php. (Accessed: 6th September 2018)
- NCGAS now has resource allocation specifically for workshop support for clients! Available at: https://ncgas.org/Blog_Posts/Workshop Allocation.php. (Accessed: 6th September 2018)
- Running Anvio on Jetstream. Available at: https://ncgas.org/Blog_Posts/Running Anvio on Jetstream.php. (Accessed: 6th September 2018)
- Running Canu on Carbonate. Available at: https://ncgas.org/Blog_Posts/Running canu on co3.php. (Accessed: 6th September 2018)

- Setting up Globus on Jetstream. Available at: https://ncgas.org/Blog_Posts/Setting up Globus on Jetstream.php. (Accessed: 6th September 2018)
- Setting up Volumes on Jetstream. Available at: https://ncgas.org/Blog_Posts/Setting up Volumes on Jetstream.php. (Accessed: 6th September 2018)
- Workflow Management. Available at: https://ncgas.org/Blog_Posts/Workflow Management.php. (Accessed: 6th September 2018)
- Workshop- Introduction to R for Biologists. Available at: https://ncgas.org/Blog_Posts/R Workshop 2018.php. (Accessed: 6th September 2018)

2.1.4.2. *Press Releases*

- Biological Sciences and Arkansas High Performance Computing Center Host Bioinformatics Workshop. *https://news.uark.edu* (2017)
- Cracking the coffee code. Available at: https://sciencenode.org/feature/cracking-the-coffee-code.php. (Accessed: 10th August 2018)
- IU projects shine at 2017 Plant and Animal Genome Conference | IT News & Events. Available at: https://itnews.iu.edu/articles/2017/iu-projects-shine-at-2017-plant-and-animal-genome-conference-.php. (Accessed: 1st September 2017)
- Teaching HPC and genomics side-by-side | IT News & Events. Available at: https://itnews.iu.edu/articles/2018/Teaching HPC and genomics side-by-side.php. (Accessed: 16th August 2018)
- Teaching HPC and genomics side-by-side. Available at: https://sciencenode.org/feature/Teaching HPC and Genomics Side-by-Side.php. (Accessed: 10th August 2018)

3. Participants

3.1. Individuals

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

Table 3-1. Individuals who have worked on the NCGAS project

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

Scott Michaels
Email: michaels@indiana.edu
Most Senior Project Role: PhD/co-PI
Nearest Person Month Worked: 1
Contribution to the Project: Funded PSC collaborator
Funding Support: Co-PI responsible for 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: Other Professional
Nearest Person Month Worked: 1
Contribution to the Project: Director over NCGAS management group and 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: 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

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

Table 3-2. Partner organizations

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 3 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 year, and a salary increase is anticipated for Sanders

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. List of NCGAS Workshops Planned

6.1.1. General Plan

We currently help run the Environmental Genomics Workshop once a year (now three years); we are on our second run of the transcriptome workshop; and we are about to run the first instance of the R workshop.

A general scheme for building on these workshops to a full yearly workshop schedule could be:

January PAG:	Collaborating on HPC	Beginning 2019
March/Spring Workshop:	Transcriptome Assembly	Began 2018
June/Local Workshop:	R or Python Alternating	Began 2018
JulyAssist in Workshop:	MDIBL	Began 2017
August/Local Workshop:	HPC for Biologists	Beginning 2019
October/Fall Workshop:	Transcriptome or Genome Assembly	Beginning 2020?
November SC:	Biology for HPC Managers	Beginning 2019?

By staggering start dates (three going now, adding two to three next year that are already mostly written, then adding new ones to swap in as others are more seasoned), the development portion is also staggered, making this doable. Most of this material already exists in some form, but a few things would need to be

filled in. This is with the exception of the genome assembly and annotation workshop, which is still in early development as we work with collaborators. We could also easily alternate the PAG and SC workshops to reduce load. Finally, the R, Python, or Collaborating on HPC workshops would be easy to adapt to field station examples to be run at the Organization of Biological Field Station meeting.

6.1.2. Large (national level):

6.1.2.1. More involvement in the Environmental Genomics Workshop—Extension of current involvement

Description: Joe Shaw, et al. want to write NCGAS into the next Environmental Genomics (MDIBL) grant. This would be a R25 through NIH and through IU. They should budget for FTE, storage space on slate-condo, and likely a Jetstream allocation. As they want more year-round contact (with the one week in Maine being an in-person anchor for the course), they would want full allocations for the students—NCGAS to set up infrastructure for them to be able to host webinars, shared workspace/website, and some additional training. Regarding additional training, we see this as a way to build HPC skills into the curriculum and test run the basic bash, R, and python mini workshops listed below. Much of this material already exists from other tutorials/workshops, and remaining material would be useful for the below listed "collaborating on HPC" workshop.

Goals: Train biologists on analyses of environmental RNAseq data, round out NCGAS budget with NIH funding, fund development of other workshop materials.

Staffing needs: 1 staff member for website/resource management and occasional tutorials (web based). Estimated 20% FTE.

6.1.2.2. *de novo* transcriptome workshop (3 days) – currently running

Description: The workshop consists of discussions, lectures, and hands-on tutorials, covering topics important to constructing and analyzing transcriptomes without the use of a genome. Material covers both the availability and use of high-performance computing (HPC) resources and the task of assembling a new transcriptome, in order to provide a comprehensive preparation for this and future bioinformatic tasks. Transcriptome assembly will consist of using four separate assemblers (Trinity, SOAP de novo, Velvet Oases, and TransABySS), with multiple kmers, to be combined and curated with Evigenes. This combined-assembly with multiple parameters is more robust than simply using one assembler, and the NCGAS pipeline streamlines the process and allows for customization if desired. We will also discuss downstream analyses such as differential expression and annotation. While material will make heavy use of XSEDE and IU machines, the material is transferable to any cluster.

Goals: Improve HPC skills and participant comfort in assembling a transcriptome with a more complex analysis workflow.

Staffing Needs: Three instructors and 1 additional assistant would be sufficient for a 25 person workshop.

See https://github.com/NCGAS/Transcriptome-assembly-workshop-2018 for schedule.

6.1.2.3. 3rd generation assembly and annotation (3 days)—early development

Description: This workshop consists of discussion, lectures, and hands-on tutorials covering topics important to constructing new genomes from third generation data, using high performance computing (HPC) resources. We will discuss choice in sequencing technologies (used in combine), considerations based on genome architecture (size, repetitive nature, *etc*.), selecting the correct computational resources to complete the assembly, how to perform an assembly in a best-practices workflow, and finally how to assess the quality of an assembly. By using multiple assemblers, the assembly is more robust than

individual platforms provide. The workflows provided will be flexible to allow for user-specific adaptation. While material will make use of XSEDE and IU machines, the material is transferable to any cluster.

Goals: Improve HPC skills and participant comfort in planning a genome project and assembling a genome with current sequencing technologies.

Staffing needs: Two instructors could probably handle this, as it would be largely hands-on. A third as assistance would be ideal. Keeping this workshop to ~12 people would be ideal with 1) our limited experience at this time and 2) how diverse this analysis can be.

Day 1:

Experimental design

Different considerations design activity

Resource needs

Day2:

What the data looks like

Assembly best practices

Quality assessment

Day 3:

Corroboration with transcriptomes

Further annotation

Combining annotation with overlap calls

6.1.2.4. Introduction to HPC for biologists (3 day)—in design phase

Description: This is a course designed to get biologists starting HPC-style work up-to-speed using Unix and clusters. Starting from ground zero, we will introduce Unix, what it is good for and why one should use it, and how to set up a Unix environment, as a foundation for effective use of a cluster. We will talk about how to manipulate text files, submit jobs effectively, and about different resources available to researchers. Additionally, we will cover basic R and Python (including the below short courses in R and Python).

Goals: The goal is to get the participant used to using Unix and the command line so that they can perform their own bioinformatic analyses. Even if they decide they hate doing this kind of thing (it isn't for everyone!), they will still be able to clearly convey to a bioinformatician what they need done, as well as understand the assumptions of each analysis.

Staffing Needs: Thomas Doak could cover the material, but it would be better to switch off between 3 instructors for a workshop of ~25 people.

Day 1:

am Unix navigation

Permissions

Environmental variables

Setup of your environment

pm Root dir

Software installation

File movement and management

Job handling

Day 2:

am Sed/grep/awk

Regex

HPC resources

pm R I

Python I

Day 3:

am R II

pm Python II

6.1.2.5. Introduction to Biology for HPC managers (1 day) – in design

Description: This workshop is planned as a one-day workshop at a conference, such as SC or PEARC. The workshop will be problem-based, with the goal of elucidating common behaviors of biologists on computers (which changes rapidly); it will also include a discussion of possible solutions for common HPC problems biologist face. Current and foreseen future problems will be discussed, framed by case-studies.

Goals: The goal is to provide insight, but also help develop and share tools that are useful in managing biological users' scripts for compression, wrappers for common workflows, FAQs for commonly faced issues (*i.e.* "what resources do I request?").

Staffing Needs: 2 people would serve this well, with a third being useful for additional input if available. As this is mostly case study driven, and more in the style of the planned OBFS workshop, this will be more a discussion of HPC solutions (driven by participants) guided by the viewpoint of biological users (instructors).

Day 1:

am File types and compression

Storage issues and what to save

Common workflows/heterogenous workflows

pm Information we want

Scaling issues

Future problems

6.1.2.6. Biological collaboration on HPC (1 day) – in design

Description: A workshop designed with the intention of being a conference add-on, such as to PAG or Evolution.

The nature of science today is highly collaborative. Often these collaborations take place across large physical spaces, and require digital tools to organize, visualize, and analyze data. This workshop will discuss available resources through XSEDE and other national infrastructure and how to manage them, specifically discussing VM set up, teaching basic admin skills such as keeping virtual machines secure, how to move data around efficiently, and how to figure out how to run jobs on any given machine. Case studies—such as running collaborative Rstudio servers, citizen science projects, and hosting genome browsers—will be discussed and demonstrated.

Goals: Facilitate the use of national infrastructure to collaborate in science, by providing introduction to necessary HPC skills and tools developed at NCGAS.

Staffing Needs: Since we envision this as a PAG workshop, it would be possible to have the whole of NCGAS there. However, two instructors would be perfectly fine. During the demo side of things (pm), it would be useful to have someone walking through and 1-2 assistants helping troubleshoot as we go.

Day 1:

am XSEDE introduction

Allocations and Permissions

Managing Data and Transfer

XSEDE/HPC resources activity

Job handlers and their differences

pm JS Doing analyses case study: R Studio

JS/W Doing analyses case study: Citizen Science

JS Webservers genome browsers, software, galaxy

6.1.3. B. Small (local level):

6.1.3.1. R for biologists workshop – currently running

Description: This is a four-part workshop (2 hr. each session) covering the basics of R: the general syntax of the language, the basic data types and how to manipulate them, as well as how to find more information when doing novel analyses. The course does not focus on any particular analysis, but uses DNA sequences as a case study in applying the material (ecological examples could be used instead). We will also cover using RStudio on Jetstream (a research cloud) but use of personal installations on laptops is fine for the workshop.

Goals: The goal of this course is to get users started in R, to be able to read and write code, and to know where to get help when needed.

Staffing Needs: One instructor, though an assistant would be helpful.

6.1.3.2. Python for biologists workshop (designed)

Description: This is a four-part workshop (1.5 hr. each session) covering the basics of python: the general syntax of the language, the basic data types and how to manipulate them, use of packages to

extend functionality, as well as how to find more information when doing novel workflows. We will also discuss available biological tools, such as biopython and bioconda. The course does not focus on any particular analysis, but uses DNA sequences as a case study while covering class material. Students will write a sequence trimming program to practice string handling, then covert it into a function.

Goals: The goal of this course is to get users started in python, be able to read and write code, and to figure out where to get help.

Staffing Needs: Carrie and Tom would be a good set for this. Carrie uses python more than Tom does, so she would be a good one to include in development, after Tom has the first pass written. An assistant would be helpful, especially during labs.

6.1.3.3. Data Management on HPC (designed)

Description: This is a four-part workshop (1.5 hr. each session) that will cover data movement, data compression, archiving, and version control for biologists working on HPC. We will run a demo on how to manage nuanced permissions for project directories with ACLs, and provide tips on the efficient use of time and space on clusters, resources available to help manage data, and walk users through setting up a project space in a clear, organized fashion.

Goals: The goal of this course is to help biologists using HPC clusters manage their data in an organized fashion, and take the greatest advantage limiting storage resources.

Staffing Needs: One instructor

Appendix 2. List of NSF Funded Projects Using NCGAS Resources

Joseph Vitti, Harvard University, Broad Institute

8/3/16

Natural selection was instrumental not only in the genesis of our species, but also in its diversification. By examining patterns of genomic variation within and among populations, we can identify and characterize genetic variants that have been subject to selection, bringing instances of local adaptation to light. This project, the culmination of my PhD research as an NSF GRFP fellow, takes a new suite of computational tools that I have designed and implemented and applies them to explore a rich new dataset (1000 Genomes Phase 3) which includes full sequence data for individuals from previously uncharacterized populations in South Asia, West Africa, and East Asia.

Raphael D. Isokpehi, Bethune-Cookman University

8/26/16

Computational capacity represents the major limitation on my ability to bridge the gap from tool-building to empirical analysis, as this step necessitates the iterative generation of simulated data en masse.

Predicting Microbial Genome-Encoded Biomolecular Networks. According to the Integrated Microbial Genomes database (http://img.jgi.doe.gov/): “At the start of 2015, IMG had a total of 32,802 genome datasets from all domains of life and 5,234 metagenome datasets, out of which 27,341 genome datasets and 3,193 metagenome datasets are publicly available.” Functional and structural annotations for genes in microbial genomes are increasingly available as multivariate data sets in formats suitable for a variety of cognitive activities including knowledge discovery, sense making, problem-solving and planning future research.

The exponential increase in whole genome sequences of bacteria and archaea presents a source of large and complex data on functional and structural annotations of genes. The annotations for function and transcriptional direction of genes adjacent to a gene locus in genomes of bacteria and archaea can be informative on biological process that involve the gene. Therefore, there is a critical need to index of Microbial Gene Loci based on Transcriptional Direction of Adjacent Genes to facilitate cognitive activities including knowledge discovery and planning future research. The NCGAS resources will be used for performing diverse actions on the data sets including comparative analysis.

Thomas Hahn, University of Arkansas

8/29/16

We would like to analyze transcriptomic (i.e. microarray and RNA Seq.), proteomic, and epigenetic data. Specifically, in the short term (i.e. by Thursday), we'd like to know how the abundance of the 74 mother-enriched and the 64-daughter-enriched proteins identified by Yang et al (2015) (see http://www.ncbi.nlm.nih.gov/pubmed/26351681) changes across the 12 time points of the yeast's life, for which its transcriptome and proteome was taken in a study by Janssens et al (2015) (see https://elifesciences.org/content/4/e08527). Moreover, we'd like to further explore the hypothesis by Janssens, et al, according to which aging is caused by an uncoupling between transcription and translation by investigating the changes in abundance distributions of other proteins, which we believe could cause this uncoupling, e.g. proteins of the ESCRT protein sorting complex, protein degradation, ribosomal biosynthesis (e.g. changes in rRNA, ribosomal subunits, tRNAs, assemble factors, etc.) and proteins involved in chromatin modeling as suggested by Pal et al (2016) in their review article about Epigenetic and Aging (see http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4966880/). This requires thorough and detailed genomic analyses of the transcriptomic and proteomic datasets provided by Janssens et al, which can be accessed at http://www.ebi.ac.uk/pride/archive/projects/PXD001714/files and at http://www.ebi.ac.uk/pride/archive/projects/PXD001714/files respectively.

Later on, we'd like to explore how histone modification can affect lifespan as described by Sen et al. (2015) (see http://www.ncbi.nlm.nih.gov/pubmed/26159996). Therefore, we need to analyze their dataset at http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=yhylgoigxrwtfuh&acc=GSE65767 to compare their effects of aging on histone modification with those we hope to reveal by analyzing data provided by Janssens et al. Exact steps will be determined during weekly meetings.¬†

WenLe Li, USDA

8/29/16

Efficiency of feed utilization in dairy cattle is vital for the overall sustainability of dairy production since it can help reduce greenhouse gas emission and feed associated cost. Two commonly used approaches in the study of genomic determinants of feed efficiency include SNP genotyping based breeding selection and genome-wide association studies (GWAS). However, SNP-based predicting/selection has a modest heritability value of ~0.4, indicating the need for better genomic biomarkers. On the other hand, the raw data identified by GWAS alone is insufficient to elucidate the molecular mechanisms associated with targeted phenotypic traits. Consequently, genomic factors and their physiological impacts on dairy feed efficiency remain largely unaddressed.

Representing the functional part of the genome, transcriptome is known to have distinct profiles specific to cell type, developmental state and disease status. Recently developed whole-transcriptome RNA-seq provides a tremendous potential to identify transcriptome biomarkers that are directly relevant to phenotypic traits. Using a whole-transcriptome RNA-seq technology, we propose to perform a systematic transcriptomics study on various tissue types in dairy cattle under a wide variety of conditions. Targeted

phenotypes include animals' response to diets with varied protein and carbohydrate concentrations, milk protein and fatty acid concentration, and lactating cow's disease state. This study will allow us to build a comprehensive knowledge base of dairy cattle transcriptomics. Such knowledge will most likely facilitate the identification of new and reliable biomarkers, which can be applied for further improvement in dairy cattle feed management and quality of milk.

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.

William Wesley, Loyola Marymount University

1/9/17

We are investigating the causes and consequences of individual-level physiological variation in mussels inhabiting the spatially and temporally variable rocky intertidal zone. We have completed a fairly large Illumina RNAseq run for a comparative physiology project (PE 150 on 12 individuals to build reference transcriptome + PE 50 on 49 individuals to look at gene expression variation), and now need to *de novo* assemble the reference transcriptome and look at individual variation in gene expression. My institution has no resources for this sort of computational analysis.

Marc Vermulist, University of Pennsylvania

2/22/17

The genome provides a precise biological blueprint of life. To implement this blueprint correctly, the genome must be read with great precision; however, due to the constraints of biological fidelity, it is impossible for this process to be completely error-free. As a result, transcription errors can occur at any time, in any transcript, and how these random errors affect cellular health is completely unknown. To fill this gap in our knowledge, we recently monitored yeast cells that were genetically engineered to display error-prone transcription. We discovered that transcription errors give rise to misfolded proteins that induce proteotoxic stress. Thus, transcription errors represent a new molecular mechanism by which cells can acquire disease. As a result, it will be important to learn more about the mechanisms that induce or suppress transcription errors, because these mechanisms could either delay or accelerate the progression of proteotoxic diseases. To this end, we developed the first next-gen sequencing assay that is capable of measuring the fidelity of transcription in a genome-wide fashion. We now propose to use this technology on yeast and mice to identify the parameters that control the fidelity of transcription in eukaryotic cells.

For our experiments, we will generate large datasets to fully characterize the transcription error spectrum caused by RNA polymerase under various conditions and in response to various DNA damaging compounds, which we expect will significantly increase the transcription error rate in treated cells. We will therefore need NCGAS resources to properly store and access our data sets in the future.

Christopher Chandler, SUNY Oswego

4/7/17

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 understudied 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.

Elze Rackaityte, University of California, San Francisco

4/13/17

The developing human fetal immune system achieves active immune tolerance through a large population of CD25+FoxP3+ regulatory T (Treg) cells present in secondary lymphoid organs. Crucially, fetal naive T cells also preferentially differentiate into Treg cells upon T cell receptor (TCR) stimulation. This project aims to identify whether the epigenome of fetal naive T cells closely resemble that of Treg cells or exist as an intermediate between adult naive and adult Treg cells, suggesting that the fetal naive T cell exists in a state poised for Treg cell differentiation. Sequencing will be carried out downstream of ChIP (Chromatin Immunoprecipitation) for the histone mark H3K27ac which identifies regions of active enhancers and by proxy, gene expression. This will be correlated with ATAC-seq data (Assay for Transposase-Accessible Chromatin) to identify a signature defining each cell type (fetal naive, adult naive, adult Treg). As ATAC-seq is amendable to an input of 50,000 cells, we will be able to use this to examine changes in the fetal epigenome in response to different stimuli in future experiments. Using the Illumina HiSeq platform we will generate genome-wide single (ChIP-seq) and paired-end (ATAC-seq) reads. We will align sequences using Bowtie2, use MACS2 and HOMER for peak calling, and Bedops and Bedtools functionalities to determine shared and differential gene regions. We will utilize the ROSE program to call super enhancer regions for H3K27ac marks. Given the large amount of data and the computing power required, access to supercomputing resources offered by IU Mason will greatly speed up analysis.

Douglas R. Cook, University of California, 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 of 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) is geared towards their professional development.

Appendix 3. IU Publications by NCGAS Users

- Arafat, A. *et al.* Unexplained Early Infantile Epileptic Encephalopathy in Han Chinese Children: Next-Generation Sequencing and Phenotype Enriching. *Sci. Rep.* 7, 46227 (2017).
- Boyd, B. M. *et al.* Phylogenomics using Target-Restricted Assembly Resolves Intrageneric Relationships of Parasitic Lice (Phthiraptera: Columbicola). *Syst. Biol.* 66, 896–911 (2017).
- Bradshaw, W. E. *et al.* Evolutionary transition from blood feeding to obligate nonbiting in a mosquito. *Proc. Natl. Acad. Sci.* 115, 201717502 (2017).
- Choudhary, S. *et al.* Transcriptome characterization and screening of molecular markers in ecologically important Himalayan species (Rhododendron arboreum). *Genome* 61, 417–428 (2018).
- Christie, A. E. *et al.* Prediction of a peptidome for the ecotoxicological model Hyalella azteca (Crustacea; Amphipoda) using a de novo assembled transcriptome. *Mar. Genomics* 38, 67–88 (2018).
- Harke, M. J. *et al.* Periodic and coordinated gene expression between a diazotroph and its diatom host. *ISME J.* (2018). doi:10.1038/s41396-018-0262-2
- Jilling, T. *et al.* Surgical necrotizing enterocolitis in extremely premature neonates is associated with genetic variations in an intergenic region of chromosome 8. *Pediatr. Res.* 83, 943–953 (2018).
- Maralit, B. A., Jaree, P., Boonchuen, P., Tassanakajon, A. & Somboonwiwat, K. Differentially expressed genes in hemocytes of Litopenaeus vannamei challenged with Vibrio parahaemolyticus AHPND (VPAHPND) and VPAHPNDtoxin. *Fish Shellfish Immunol.* 81, 284–296 (2018).
- Roncalli, V., Cieslak, M. C., Sommer, S. A., Hopcroft, R. R. & Lenz, P. H. De novo transcriptome assembly of the calanoid copepod Neocalanus flemingeri: A new resource for emergence from diapause. *Mar. Genomics* 37, 114–119 (2018).
- Roncalli, V. *et al.* A deep transcriptomic resource for the copepod crustacean Labidocera madurae: A potential indicator species for assessing near shore ecosystem health. *PLoS One* 12, e0186794 (2017).

- Winker, K., Glenn, T. C. & Faircloth, B. C. Ultraconserved elements (UCEs) illuminate the population genomics of a recent , high-latitude avian speciation event. *PeerJ Prepr.* (2018). doi:https://doi.org/10.7287/peerj.preprints.27035v1

Appendix 4. 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 Carbonate, Karst clusters, as well as PSC Bridges cluster. 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, see the National Center for Genome Analysis Support website or email NCGAS.

Package	Version	Carbonate	Karst	Bridges
abyss	1.5.2	x		x
	1.9.0			x
	2.0.2	x		x
admixture	1.3.0		x	
AFNI	16.3.00		x	
	18.0.13	x		
allpathslg	52488			x
Annovar	2016-02-01			x
Anvi'o	2.0.2			x
	2.2.2			x
	2.3.1			x
	2.3.2			x
	2.4.0			x

	3.0.0			x
ARAGORN	1.2.38			x
Aspera	3.6.2			x
ATLAS	3.101.2			x
augustus	3.2			x
	3.3	x	x	
autodock suite	4.2.3		x	
	4.2.6			x
bamtools	2.4.0			x
	2.4.1	x	x	
bamutil	1.0.13		x	
	1.0.14	x		
barrnap	0.6			x
bcftools	0.1.19			x
	1.3.1		x	x
	1.5	x		
bedops	2.4.19			x
	2.4.35			x
bedtools	2.20.1		x	

	2.25.0			x
	2.26.0	x	x	
bfast	0.7.0a		x	
bioconductor	2.12		x	
	3.1		x	
	3.3	x	x	
	3.6	x		
biopython	1.59		x	
	1.7	x	x	
bismark	0.19.0	x		x
BLASR	1.0	x		
	1.3.1			x
blast	2.2.27		x	
	2.2.28		x	
	2.2.31		x	x
	2.6.0+	x		x
	2.7.1			x
blat	35	x	x	x
bowtie	1.1.1			x

	1.1.2		x	x
	1.2.2	x		
bowtie2	2.1.0		x	
	2.2.3		x	
	2.2.6		x	
	2.2.7			x
	2.3.2	x		
	2.3.4			x
breakdancer	1.1		x	
	1.3.6		x	
breseq	0.23		x	
	0.3		x	
	0.30.2	x		
	0.32.0	x		
busco	1.22			x
	3.0.2	x		
bwa	0.7.10		x	
	0.7.12	x		
	0.7.13			x

	0.7.15		x	
cafe	4.2	x		
canu	1.3		x	x
	1.4		x	
	1.5			x
	1.6	x		x
	1.7			x
CDBFASTA	2013			x
cd-hit	4.6.6			x
	4.6.8	x		
centrifuge	1.0.3	x		x
checkm	1.0.7			x
circos	0.69.2			x
clustalw	2.0.12		x	
CNVkit	0.9.2			x
colormap	1		x	
crisprdo	1		x	
	0.7.15		x	

cufflinks	2.0.2		x	
	2.1.1		x	
	2.2.0		x	
	2.2.1	x		x
cutadapt	1.11		x	
	1.16	x		x
	1.5.0			x
	1.9.1	x		
cytoscape	3.6.0		x	
	3.6.1	x		
dammit	0.3			x
deeptools	2.3.5			x
deseq2	3		x	
detonate	1.1			x
diamond	0.7.11			x
	0.8.31			x
	0.8.38	x		
	0.9.13	x		
discasm	0.1.2	x		

discovar	52488			x
discovar de novo	52488			x
dose2geno	4		x	
ECtools	41974			x
edgeR	3	x	x	
eigensoft	6.1.3		x	
emboss	6.5.7		x	
	6.6.0			x
ensembl	81		x	
epa-ng	0.1.1	x		
EricScript	0.5.5			x
evigene	2013.07.27	x		
exonerate	2.4	x		
falcon	0.4.1			x
fasta-splitter	0.2.4			x
fastqc	0.10.1		x	
	0.11.3			x
	0.11.5	x	x	
fastq-splitter	0.1.2			x

fastx	0.0.13		x	
	0.0.14			x
flash	1.2.11		x	x
flexbar	2.4		x	
	2.6.0			x
FragGeneScan	1.2			x
gatk	3.5			x
	3.6			x
	3.7			x
	3.8	x		x
	4.5			x
genome MuSiC	0.4.1			x
gffread	0.9.8			x
glimmer	3.0.2			x
	3.0.4			x
gmap	2014-05-15		x	
	2016-04-04		x	
	2017-06-20	x		
	2018-11-05	x		

GMAP-Fusion	0.3.0	x		
graphlan	0.9.7			x
hisat2	0.1.6-beta		x	
	2.0.4		x	x
	2.1.0	x	x	
hmmer	2.3.2			x
	3.0		x	
	3.1b2	x		x
htseq	0.9.1	x		x
htslib	1.5	x		
HUMAnN2	0.10.0			x
IDBA	1.1.1			x
impute2	2.2.2		x	
infernal	1.1.2			x
interproscan	5.29	x		
jellyfish	1.1.11			x
	2.2.6			x
	2.2.9	x		
kallisto	0.42.3		x	

	0.43.0		x	x
	0.43.1	x		
khmer	2.0			x
kraken	0.10.5			x
	1.0			x
ldhat	2.2		x	
ldhot	2014		x	
limma	3		x	
lincrna	1		x	
Long Ranger	2.1.6	x		
mach	1.0.18		x	
macs	1.4.2		x	
	1.4.3			x
	2.1.0	x		
	2.1.1			
macse	1.2			x
maft	7.3			x
maker	2.31.9	x		
MALT	0.3.8			x

Mapsembler 2	2.2.4			x
Marvel	2018-01-20			x
masurca	3.1.3			x
	3.2.2			x
	3.2.6			x
	3.2.7			x
MaxBin	2.1.1			x
megahit	1.1.1			x
	1.1.2	x		
MEGAN	5.11.3			x
	6.0	x		
meme	4.12.0	x		
Meraculous	2.2.4			x
metabat	2.11.3	x		
Metaphlan	1.7.7			x
	2.6.0			x
MetaVelvet	1.2.10			x
Methylpy	1.25			x
migrate	3.3.2 mpi		x	

	3.3.2 serial		x	
mimar	12172010		x	
minCED	0.2.0			x
minia	2.0.7	x		
minimac	11162012		x	
mira	4.0.2			x
miso	0.4.6		x	
	fastmiso-3682184-3			
mothur	1.31.2		x	
	1.34.2		x	
	1.38.1			x
	1.39.0			
	1.39.5	x		
	1.40.5	x		
mrbayes	3.2.1 mpi		x	
	3.2.1 serial		x	
mrsfast	3.3.7		x	
mummer	3.23	x	x	x

muscle	3.8.31	x	x	
MyCC	42341			x
nanopack	0.14	x		
NGSCheckMate	2016-12-10			x
ngsutils	0.5.0c		x	
	0.5.2a		x	
	0.5.9		x	
ninja	1.2.2		x	
novoalign	3.8.0	x		
oases	0.2.09	x		
paml	4.8		x	
	4.9			x
paup	4		x	
pbjelly	15.8.24			x
phylip	3.69		x	
phylosift	1.0.1			x
picard	2.1.1			x
	2.14.0	x		
	2.17.0			x

pilon	1.16			x
platanus	1.2.4			x
plink	1.07		x	
	0.10			x
polyphen2	2.2.2	x		
pplacer	1.1.7			x
primer3	1.1.4			x
	2.2.3			x
	2.3.7			x
prodigal	2.6.2			x
	2.6.3			x
prokka	1.11			x
psi4	1.1	x		
QoRTs	44ab10d			x
quast	4.6.3	x		
r8s	1.8		x	
raxml	7.4.2		x	
	8.0.26		x	
	8.2.9			x

	8.2.11	x		
ray	2.3.1	x		x
rmats	3.2.5	x		
repeatmasker	4.0.6			x
	4.0.7	x	x	
repeatmodeler	1.0.8		x	
	1.0.10	x		
RNAmmer	1.2			x
rosetta	3.5		x	
	3.7			x
	3.8		x	
rsem	1.2.19		x	
	1.2.21			x
	1.3.0	x		
sailfish	0.7.3		x	
	0.8.0		x	
	0.9.2			x
salmon	0.11.0	x	x	
	0.4.2		x	

	0.6.0			x
	0.7.2			x
	0.8.1			x
	0.9.1	x		
samtools	1.2		x	
	1.3		x	x
	1.3.1		x	x
	1.5	x		
	0.1.18		x	
	0.1.19			x
	1.7			x
scythe	0.992-beta			x
seqtk	1.2-r94			x
shannon	2017-05-10			x
sickle	1.33			x
SignalP	4.1c			x
smrt	2.2.0		x	
SNVMIX	0.11.8-r5			x
soapdenovo	1.03	x		

	r240		x	
	42286			x
soapdenovotrans	1.03	x		
solareclipse	8.3.1		x	
somaticsniper	1.0.5			x
spades	3.10.0		x	
	3.10.1	x		x
	3.11.1	x		x
	3.8.1			x
	3.8.2		x	
	3.9.0		x	
spm	8	x	x	
	12	x	x	
sprai	0.9.9-23		x	
sra-toolkit	2.3.5-2		x	
	2.5.4		x	
	2.8.1			x
	2.8.2	x		
stacks	1.48	x		

star	2.4.1d		x	
	2.5.2b		x	
	2.5.3	x		
staraligner	2.5.2b			x
	2.5.4b			x
starfusion	1.1.0	x		x
	1.3.1			x
strelka	1.0.14			x
stringtie	1.3.3b	x		x
subread	1.6.1			x
super-deduper	7c48db4			x
supernova	2.0.0	x		x
tabix	0.2.6	x	x	
theano	0.8.0			x
	0.8.2			x
TMHMM	2.0.0			x
tophat	1.4.1		x	
	2.1.0			x
	2.1.1	x		x

tophat2	2.0.7		x	
	2.1.0		x	
	2.1.1	x		
tpp	4.6.2		x	
transabyss	1.5.5	x		
	2.0.1	x		x
transdecoder	3.0.1			x
	4.0.0	x		
transrate	1.0.3	x		x
trim galore	0.4.5	x		x
trimmomatic	0.35		x	
	0.36	x	x	x
trinity	2.1.1		x	x
	2.0.6			x
	2.2.0			x
	2.3.2			x
	2.4.0	x	x	x
	2.6.6	x		
trinotate	2.0.1			x

	3.1.1	x		
tRNAscan-SE	1.23			x
UNCeqR	42559			x
VarScan	2.4.2			x
vcftools	0.1.10		x	
	0.1.13	x	x	
	0.1.15			x
	0.1.14		x	
velvet	1.2.10	x		x
weaver	2017-11-17			x
wgs-assembler	8.2			x
XHMM	1			x

Table 6-1. Genome analysis software packages available on Indiana University's Carbonate and Karst clusters, as well as PSC Bridges cluster.

7. Acknowledgements

This document was developed with support from National Science Foundation (NSF) grant OCI-1053575. 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.

This material is based in part upon work supported by National Science Foundation Grant 1759906. Our work would not be possible without the generous support of the following awards received by Research Technologies and the Pervasive Technology Institute at Indiana University.

- The Indiana University Pervasive Technology Institute was supported in part by two grants from the Lilly Endowment, Inc.
- NCGAS has also been supported directly by the Indiana METACyt Initiative. The Indiana METACyt Initiative is supported in part by the Lilly Endowment, Inc.

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.