NCGAS NSF Annual Report

September 1, 2019 – August 31, 2020

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 usable 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 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 offers services to institutions of higher education throughout the United States and its protectorates. All states, and Puerto Rico, are home to clients of NCGAS services. Additionally, residents of every state and Puerto Rico use software that NCGAS supports and helps make available through additional gateway services like Trinity and IU Galaxy.

NCGAS clients perform various types of genomic analyses, though transcriptome assembly/analysis and genome assembly/analysis are the most popular. According to the 2020 Annual User Survey, 45% of clients are from institutions with fewer than 10,000 students, 45% of our clients report non-white ethnicity, and approximately 38% of our services go to graduate students (determined from allocation numbers and workshop seats). Of NCGAS clients requiring computing access, 22% are located in the 28 EPSCoR states.

Specific metrics on software and infrastructure (Table 1); education/outreach and dissemination (Table 2), and consulting (Table 3) services are available in the Supporting Files.

Table 1. Software and infrastructure support for current project year with previous year as reference and context.

Software and Infrastructure Support Metrics		
	Sept. 2018 - Aug 2019	**Sept 2019 - Aug 2020**
Software Support		
Packages Supported	430	**351***
Jetstream VMs	17	**19**
Jetstream instances launched	333	**166**
Infrastructure Provided		
Jobs - Carbonate	491296	**161540**
CPU - Carbonate	1,326,127.91	**1,014,242.64**
GitHub repository		
Number of repositories	10	**15**

Table 2. Metrics for outreach and training activities as well as dissemination metrics for current project year with previous year as reference and context.

Education/outreach and Dissemination Metrics		
	Sept. 2018 - Aug 2019	**Sept 2019 - Aug 2020**
Education, Outreach, Training		
Number of events	15	**13***
Number of attendees	482	**584**
NCGAS Blog /Website		
Number of blogs	22	**10**
Total page views	60,742	**32,960**
Total number of unique users	26,265	**29,289**
Science Highlights/News Articles		
Number of articles	4	**4**
Twitter		
Number of followers	201	**347**

Number of tweets posted	391	**539**
Engagements for posts	2,048	**2,750**
Facebook		
Number of follower	73	**85**
Number of posts (total)	177	**194**
Engagements for posts (engaged users) (total)	405	**449**
YouTube IU_PTI channel		
Subscribers (total)	--	**387**
Playlists (total)	2	**3**
Total views (total)	173	**5479**

*Lower than anticipated due to COVID-19 related cancellations

Table 3. Consulting metrics for current project year with previous year as reference and context.

Consulting Metrics		
	Sept. 2018 - Aug 2019	**Sept 2019 - Aug 2020**
Short Consults (<4 hr)	261	**287**
4-80 hr Consults*	24	**11**
80-160 hr Consults*	4	**1**
160+ Consults*	7	**3**
Projects Account Requests	39	**32**
Projects with grants reported	18	**12**
Total grant dollars supported	$22,693,123	**$7,572,709**

* Numbers reflect new consults started within the project year. Projects from previous years may continue, but are not included in this number.

1.2.2. Specific Objectives:

Accomplishments relevant to the achievement of goals for this project are described below:

Software and infrastructure support

Software: The National Center for Genome Analysis Support (NCGAS) provides support genome analysis software packages available on XSEDE/PSC Bridges, XSEDE/ Jetstream, IU's Karst, and IU's Carbonate 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). A list of the 351 versions of 219 software packages currently supported, as well as the 19 virtual machines publicly available, can be found at https://ncgas.org/services/software/index.html.

Note: Genomics Toolkit image available on Jetstream hosts a set of genomics toolkit that is not administered by NCGAS, but by the Jetstream team, but NCGAS offers consulting help on the included software packages.

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 by ITCR, but is now an NCGAS project. 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 clients, 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).

GenePattern: As an Information Technology in Cancer Research (ITCR) funded project, NCGAS hosts the Broad Institute's GenePattern genomics gateway that submits jobs to a high memory cluster (IU's Carbonate). The Broad hosts a cloud based gateway, but is limited in supporting high memory jobs on the cloud, causing certain analyses to be largely completed on NCGAS's high memory version. 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).

Outreach/training and dissemination

We have been offering national workshops since May 2018. In this Project Year, we offered three courses: Metagenomic Analysis (October 2019, in-person), Introduction to R for Biologists (November 2019, March 2020, online), HPC On-boarding for Biologists (January 2020, in-person). Our second Metagenomics Analysis course in March was canceled due to COVID-19.

The R course was converted into a Massive Open Online Course (MOOC), which allowed the course to scale from 30-60 people in live versions from the previous project year, to 100 and 400 participants in the current project year. All of our courses include a pre- and post-survey (IRB approved) with self evaluation of skill level in all learning objectives. The increase in these skills before and after courses has been consistently a full point on a 1-5 scale, regardless of live, hybrid, or online course.

Consulting

NCGAS saw 275 short (<4 hr) tickets this year, and 10 longer term consultations, which is on par with previous project years (Table X). Longer consultations are the primary source of co-authored publications. There are 28 research projects current this reporting period of which 18 reported grant support. A list of projects supported can be found here: https://drive.google.com/file/d/1xgvPruIrq-wy3l8VcyOQWD1JiI_EMFVY/view?usp=sharing.

1.2.3. Significant results

Products: We have supported 12 peer-reviewed publications, 9 conference presentations, 2 technical reports, 4 press releases, 3 video playlists from our workshops, 1 electronic textbook, and supported one dissertation.

COVID Response: We were able to move more of our educational content online and scale course sizes to compensate for educational opportunities lost by the closing of campuses throughout the country. Three hundred more students enrolled in our online R course and we have moved our HPC content online, to be nationally released in September 2020. These basic domain-centric computational skills courses are critical to helping researchers adapt to more computational work as lab time is decreased.

Community Building: NCGAS offers services to institutions of higher education throughout the United States and its protectorates. All states and Puerto Rico are home to clients of NCGAS services. The NCGAS user community performs various types of genomic analyses, with transcriptome assembly/analysis and genome assembly/analysis being the most popular (Figure 1). Forty-five percent of clients are from institutions with fewer than 10,000 students, 45% of our clients report non-white ethnicity, and approximately 38% of our services go to graduate students (determined from allocation numbers and workshop seats). Of NCGAS clients requiring computing access, 22% are located in the 28 EPSCoR states (Figure 1).

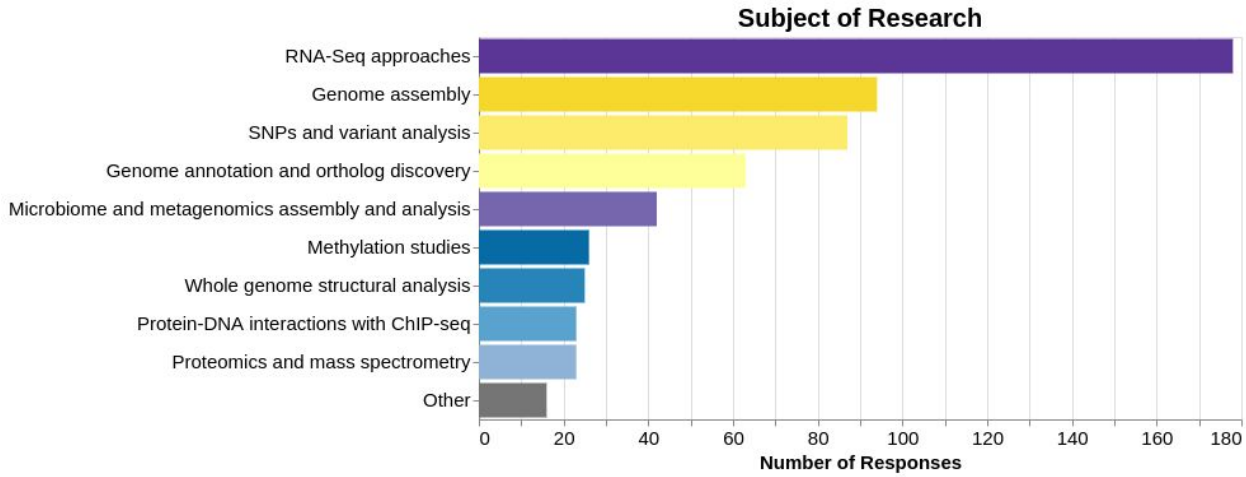

Figure 1: Breakdown of Annual User Survey response to subject of research. Analyses that we have developed and deployed national workshops for are in purple and analyses with workshops and workflows in development are in yellow. Analyses that we have no planned workflows or workshops are in blue and indeterminate analyses are in grey.

User survey: Since 2017, NCGAS has contracted an independent evaluation group to conduct a survey of its clients. This survey is confidential and has advance approval of the Indiana University IRB (Institutional Review Board). In 2020, 28% of clients reported that NCGAS services were very important to completing their work, with another 19% stating that NCGAS services were helpful (N=308; Figure 2). Clients complete an average of 36% of their

computation on NCGAS resources (N=165). Further, 34% of clients report using NCGAS services as a graduate student, seeking assistance from NCGAS before they have grants to cover the costs of external consultation, computing, or other services. NCGAS clients are happy with the service, with 76% of clients reporting being satisfied or extremely satisfied with the service in general (N=237), and 69% report being very satisfied with the available clusters (N=134) and 67% are very satisfied with NCGAS consulting services (N=159). Word-of-mouth recommendations of NCGAS services drive 38% of client acquisition (N=119), further underscoring the service' positive reputation in the community.

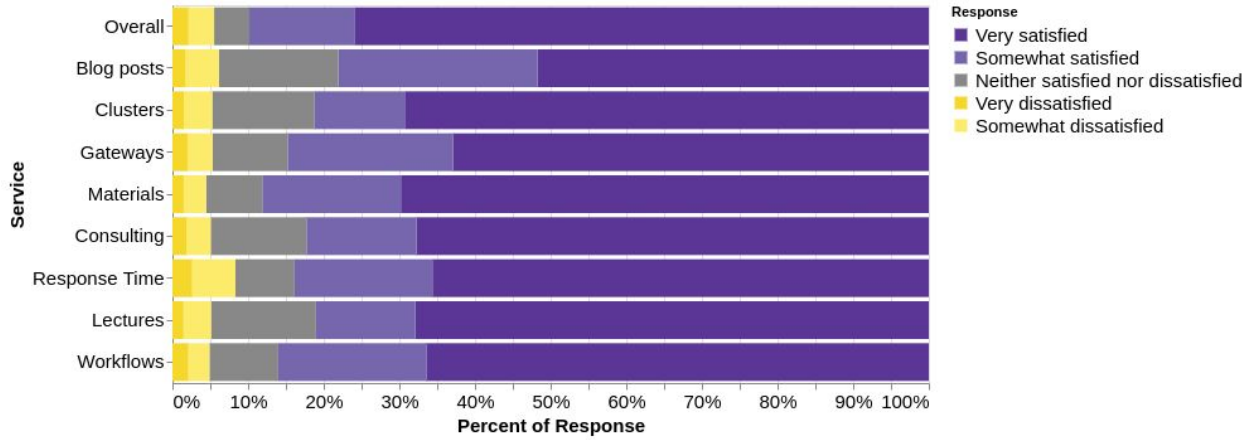

Figure 2: Annual User Survey results on satisfaction ratings for various NCGAS services. Percentages are based on response per question, with the number of responses range from 134 to 308. Yellow responses are dissatisfactory, neutral are grey, and all positive responses are purple.

1.2.4. Key outcomes or Other achievements

NCGAS has successfully refreshed our web presence with a fully new website with increased accessibility consideration, an advertisement video for easy outreach, and a migrated blog to a more fully featured platform.

NCGAS collaborations have added projects with IU Health, working on extending our machine learning materials to help a physician modify her work to include machine learning to predict the possibility of radiation-induced leukopenia within the first week of treatment. This work is in progress and will result in potentially further ITCR grant involvement and publications.

Additionally, the Jetstream REU project predicting frog calls from recordings using neural networks and decision trees has attracted the attention of the National Parks Service. This year, they provided us with data for more frog species, and invited our undergraduates to present their work to the Great Lakes Inventory & Monitoring Network, with the potential for further interaction and funding.

The Stakeholder Advisory Board (SAB) was also reorganized this year, to include new members throughout the country, and increase diversity in leadership (* indicate members of under-represented groups in information technology):

- Dr. Lydia Bright, Assistant Professor at State University of New York at New Paltz*
  Dave Clements of the Galaxy Project Team at Johns Hopkins
- Dr. Raphael Isokpehi, Professor at Bethune-Cookman University*
- Dr. Micheal Lynch, Center Director and Professor of Arizona State University's Biodesign Center for Mechanisms of Evolution
- Dr. Mihai Pop, Director of the University of Maryland Institute for Advanced Computer Studies (UMIACS)
- Jeff Pumill, Director of Strategic Initiatives & User Services at the University of Arkansas
- Dr. Rachel Schwartz, Assistant Professor at the University of Rhode Island Coastal Institute*

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

National Workshops: The NCGAS team has provided numerous training and development opportunities to domain scientists and students. This includes courses:

- Intro to HPC for Biologists (in-person)
- R for Biologists (online, twice)
- Metagenomic Analysis (in-person)

Undergraduate training programs: NCGAS Staff members served as mentors for seven undergraduates from under-represented groups, through the XSEDE Jetstream REU program, IU's Center for Excellence and Women in Technology REU program, IU's Bepko Learning Center Fellowship, and through hourly work. These students participated in the following projects to gain valuable skills in computation and biological analysis:

- Automatic recognition of frog calls using Machine Learning using Jyputer
- Mining the Sequence Read Archive (SRA) for Metagenomic Analysis
- Visualization of Metagenomic Analyses
- Transcriptome assembly of Metagenomic Reads
- Mining the NCBI Database for Sequence-based Epigenetic Predictions and Phylogenetic analysis
- Gamification of Biological Training Materials to Increase Engagement

Several of these students have presented their data at virtual conferences, or are slotted to in the near future. All participants had additional instruction on production of posters and documentation for their work.

Graduate Student Training: NCGAS Staff members also served as mentors for three computational biology graduate student projects:

- Genome-guided Transcriptome Assembly of Ciliates in Relation to Mating Types
- Software Installation and HPC Skills
- Genome Annotation and Genome Browser Construction

One of these students presented her data at the (virtual) 2020 Young Investigator Ciliate Molecular Biology Conference, hosted virtually in Portugal this year.

Staff Continuing Education: Carrie Ganote, Bhavya Papudeshi, and Sheri Sanders were all provided subsidized tuition through IU employment to take coursework in genomics, bioinformatics, machine learning, and data visualization. These courses directly led to two machine learning projects led by Sheri Sanders, and supported both Carrie Ganote and Bhavya Papudeshi's doctoral progress.

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

Online resources: Results have been communicated to communities of interest through published papers in peer-reviewed journals and conference proceedings (listed in the Products section). In addition, NCGAS has a strong social media presence (Twitter, Facebook, YouTube, Git), a highly visited blog and associated website:

Twitter: NCGAS has a twitter page with 347 followers (up from 201 in 2019) and a total of 539 tweets in total as of July 1, 2020. 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. This resulted in 449 engagements.

Facebook: NCGAS Facebook page currently has 85 followers (up from 73 in 2019) with 449 total engagements. Facebook posts showed a similar trend as Twitter posts, with more activity during conferences and workshops.

YouTube: Two main playlists were used during this reporting period:

- Playlist: 2018-2019 de Novo Assembly of Transcriptomes - 13 videos, Published 8/7/2019, 265 view total (all videos) as of 8/26/2020
- Playlist: 2018-2019 Intro to R for Biologists - 25 videos: Published 7/16/2019, 5,168 view total (all videos) as of 8/26/2020

Github: We have 16 Repositories (6 active during reporting period), with 6 Contributors as of August 26, 2020.

Blog: We produced 10 posts, generating 32,960 page views from 39,289 unique users. Our blog generates the majority of our website traffic. Our blog was also migrated to a new platform this year, resulting in an extended downtime and an associated lower number of posts this project year.

Website: Our website was rebuilt by a graphic design team during this year. As a result, our analytics tracking was interrupted, and we do not have accurate numbers for traffic for this project year. However, we can report that over 16,000 page views were recorded for our website before the revision, in keeping with the ~40,000 hits we have gotten in the last two years (largely driven by our blog).

Conferences: NCGAS attended several national conferences this project year, before many were canceled or postponed due to the pandemic. Starred conferences included presentations by NCGAS Staff (see products for citations):

- Organization of Biological Field Stations, Belgium*
- SuperComputing 19, Denver Colorado*
- Plant and Animal Genome XXVII, San Diego, California*
- Bioinformatics Community Conference 2020, virtual
- Practice & Experience in Advanced Research Computing Conference 2020, virtual
- 2020 Young Investigator Ciliate Molecular Biology Conference, virtual*

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

Software and Infrastructure Support:

- We will continue to review software and update versions as software comes available. Part of Thomas Doak's Chief Scientist role will be to review current workflows in publication, determine candidates for testing with public data, and handing them off to Carrie Ganote to implement in Galaxy and on national compute clusters.
- Existing workflows, such as our *de novo* transcriptome assembly workflow and metagenomics analysis workflow, will be containerized in a collaborative effort with Rich Knepper's group at Cornell. They have produced a template for containerizing workflows, and are testing it's implementation with our transcriptome workflow on git. Upon successful completion, we will test their documentation of building containers with the second workflow. The hope is that this will also produce mature educational materials to share with biological software developers to ease distribution of products in the community.

Outreach/training:

- Proposed courses and workshops for the upcoming year are:

    - 9/2020: Introduction to HPC for Biologists, full national online launch
    - 10/2020: R for Biologists, online MOOC
    - 11/2020: NCBI Workshop: Mining the Sequence Read Archive (SRA), online
    - 1/2021: Introduction to HPC for Biologists
    - 3/2021: R for Biologists, online MOOC
    - 5/2021: Population Analysis of Non-model Organisms with RADseq, new course
    - 6/2021: Machine Learning using NEON Data, Ecological Society of America, new course
    - 7/2021: Population Analysis of Non-model Organisms with RADseq at the Joint Meeting of Ichthyologists and Herpetologists 2021
- Purposed undergraduate and graduate training programs are:
    - Fall 2020: Bepko Internship for underrepresented minority undergraduate
    - Spring 2021: Graduate Assistantship for Biology or Computer Science student
    - Summer 2021: Jetstream REU program

Consulting:

- Consulting will continue with minimal change. We expect there to be approximately 250-300 requests for short-term consults, as this has been our average for several years, despite a growing community. We will also identify 5-10 common issues from these tickets to be addressed in blog content, allowing users to find answers through web searches rather than direct contact.

User service and administrative elements of this work will include:

- Updating user documentation and streamlining our user database to be in a tidy data form, allowing for much easier analysis of the database in the future.
- Conducting the NCGAS Stakeholder Advisory Board Meeting, to be conducted by November 2020.
- Conducting the annual NCGAS user survey during the summer of 2021.

2. Products (resulting from this project during the specified reporting period)

Journal Articles

NCGAS project team generated (in bold). User-generated in plain text.

1. Lima, L., F., O.; Weissman, M.; Reed, M.; **Papudeshi, B.**; Alker, A., T.; Morris, M., M.; Edwards, R., A.; de Putron, S., J.; Vaidya, N., K.; and Dinsdale, E., A. (2020) Modeling

of the Coral Microbiome: the Influence of Temperature and Microbial Network. mBio, 11(2).

2. Petek, M., M. Zagorščak, Ž. Ramšak, **S. Sanders**, Š, Tomaž, E. Tseng, M. Zouine, A. Coll, K. Gruden (2020). Cultivar-specific transcriptome and pan-transcriptome reconstruction of tetraploid potato. Scientific Data. doi: https://doi.org/10.1101/845818.

3. Cinel, S.D., Taylor, S.J. (2019) Prolonged bat call exposure induces a broad transcriptional response in the male fall armworm (Spodoptera frugiperda; lepidoptera: Noctuidae) brain. Frontiers in Behavior Neuroscience. https://doi.org/10.3389/fnbeh.2019.00036

4. Wong, J.M., Gaitan-Espitia, JD, Hofmann, GE. (2019) Transcriptional profiles of early stage red sea urchins (Mesocentrotus franciacanus) reveal differential regulation of gene expression across development. Marine Genomics. https://doi.org/10.1016/j.margen.2019.05.007

5. Rivera-García L, R Rivera-Vicéns, A Veglia, NV Schizas (2019) De novo transcriptome assembly of the digitate morphotype of Briareum asbestinum (Octocorallia: Alcyonacea) from the southwest shelf of Puerto Rico. Marine Genomics https://doi.org/10.1016/j.margen.2019.04.001

6. Roncalli, V., Cieslak, M. C., Hopcroft, R. R., & Lenz, P. H. (2020). Capital Breeding in a Diapausing Copepod: A Transcriptomics Analysis. *Frontiers in Marine Science*, *7*, 56. https://doi.org/10.3389/fmars.2020.00056

7. 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*, *10*, 136. https://doi.org/10.3389/fmicb.2019.00136

8. Choudhary, S., Thakur, S., Jaitak, V., & Bhardwaj, P. (2019). Gene and metabolite profiling reveals flowering and survival strategies in Himalayan Rhododendron arboreum. *Gene*,*690*, 1–10. https://doi.org/10.1016/j.gene.2018.12.035

9. 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), ukz012. https://doi.org/10.1093/auk/ukz012

10. Smythe, A. B., Holovachov, O., & Kocot, K. M. (2019). Improved phylogenomic sampling of free-living nematodes enhances resolution of higher-level nematode phylogeny. *BMC Evolutionary Biology*, *19*(121). https://doi.org/10.1186/s12862-019-1444-x

11. Bui, L. T., & Ragsdale, E. J. (2019). Multiple plasticity regulators reveal targets specifying an induced predatory form in nematodes. *Molecular Biology and Evolution*, *msz171*. https://doi.org/10.1093/molbev/msz171

12. Gross, J. B., Sun, D. A., Carlson, B. M., Brodo-Abo, S., & Protas, M. E. (2019). Developmental Transcriptomic Analysis of the Cave-Dwelling Crustacean, Asellus

aquaticus. *Genes*,*11*(1), 42. https://doi.org/10.3390/genes11010042

Publications under review

None

Other Conference Presentations / Papers

1. **Doak TG, Sanders SA, Ganote C, Papudeshi B**, Fischer J, Hancock DY. (2020). National Center for Genome Analysis Support (NCGAS): Genomics and other Science in the NSF-Funded Jetstream Cloud. Plant and Animal Genome 2020, San Diego, California. Available at http://hdl.handle.net/2022/25301.

2. **Papudeshi B**, Leffler H, Ganapaneni S, **Sanders SA**, Ganote C, and Doak TG. (2020). Mining Microbial Genomes from Datasets on the Sequence Read Archive. Plant and Animal Genome 2020, San Diego, California. Available at http://hdl.handle.net/2022/25300.

3. **Sanders, S**, (2019). Teaching Machine Learning to Domain Scientists: Supporting Newcomers to AI on HPC Systems. Indiana University Booth presentation at SuperComputing 19.

4. Mansfield, C., Tseng, C., **Sanders, S.**, Custer, TW, Custer, CM, Matson, CW. (2019) Genetic diversity comparison of tree swallow populations in the Great Lakes region using RNA-sequencing. SETAC North America 40th Annual Meeting.

5. Song, J., Brill, R.W, McDowell, J. (2019) Investigating local adaptation and plasticity of an estuarine-dependent teleost, Spotted Seatrout (Cyanoscion nebulosus). In American Fisheries Society and The Wildlife Society 2019 Join Annual Conference. Retrieved from https://afs.confex.com/afs/2019/meetingapp.cgi/Paper/40622.

6. **Papudeshi, B.**, Chafin, T., **Sanders, S.**, Ganote, C., Reshetnikov, A., Sokolov, S., Doak, T., Pummil, J.F., Douglas, M.R., Douglas, M. (2019) Genome and transcriptome analysis of fish tapeworm Nippotaenia percotti through scientific collaboration between research labs and national cyberinfrastructure. In American Fisheries Society and The Wildlife Society 2019 Join Annual Conference. Retrieved from https://afs.confex.com/afs/2019/meetingapp.cgi/Paper/39888.

7. Hannah Erickson. Mapping the Mating Type Recognition Pathway of Tetrahymena thermophila. Smith College. 2020 Young Investigator Ciliate Molecular Biology Conference.

8. **S. Sanders**, E. Foran, E. Guido, J. Anderson, T. Slayton, T.G. Doak. (2019). Automatically Survey Frogs Using Raspberry Pis, Jetstream Cloud, and Machine Learning. In Organization of Biological Field Stations Annual Meeting 2020.

9. H. Leffler, S. Ganapaneni, **B. Papudeshi,C. Ganote, S.A. Sanders, T.G. Doak**. (2019). Mining Microbial Genomes from Datasets on the Sequence Read Archive. In Organization of Biological Field Stations Annual Meeting 2020.

Other Publications

Technical reports

1. **Sanders, S., C. Ganote, B. Papudeshi, C. Stewart. T. Doak**. (2019) "Summary Report on Scaling the Introduction to R for Biologists Workshop by National Center for Genome Analysis Support (NCGAS) to a Massive Open Online Course (MOOC)", Indiana University, Bloomington, IN. PTI Technical Report. Retrieved from http://hdl.handle.net/2022/24888
2. **Sanders, S., C. Ganote, B. Papudeshi, C. Stewart. T. Doak**. (2019) "Summary of the National Center for Genome Analysis Support (NCGAS) 2018-2019 de Novo Transcriptome Workflow and Workshops", Indiana University, Bloomington, IN. PTI Technical Report. Retrieved from http://hdl.handle.net/2022/24887

News Articles

1. https://itnews.iu.edu/articles/2020/Jetstream%20REU%20student%20Tenecious%20Underwood%20awarded%20prize%20at%20ERN%20conference.php
2. https://itnews.iu.edu/articles/2020/Outstanding%20opportunities%20for%20undergrads%20interested%20in%20cyberinfrastructure.php
3. https://itnews.iu.edu/articles/2020/Taking-data-science-skills-to-the-people-.php
4. https://eventfund.codeforscience.org/scaling-up-online-r-courses/

Other Products

Audio or Video Products

1. IU PTI. (2020). National Center for Genome Analysis Support (NCGAS) [YouTube Playlist]. Retrieved September 1, 2020, from https://www.youtube.com/playlist?list=PLqi-7yMgvZy_qGhVYev1waN00-GJabSG9.
2. IU PTI. (2020). De Novo Assembly of Transcriptomes [YouTube Playlist]. Retrieved September 1, 2020, from https://www.youtube.com/playlist?list=PLqi-7yMgvZy_IaAiPG89AX2cQH2JY4Ifo.
3. IU PTI. (2020).Intro to R for Biologists [YouTube Playlist]. Retrieved September 1, 2020, from https://www.youtube.com/playlist?list=PLqi-7yMgvZy-1vFDC7dlQB7hfTrcH5Qh7.

Educational aids or curricula

1. **Sanders, S**. Introduction to R for Biologists. E-book, National Center for Genome Analysis Support, Second Edition, 2020. Retrieved from https://ncgas.org/training/r_textbook_full.pdf.

Thesis / Dissertations

Using our transcriptome pipeline:

1. Wong, JM. (2019) Investigating the response of sea urchin early developmental stages to multiple stressors related to climate change. University of California, Santa Barbara. https://search.proquest.com/docview/2311653028?pq-origsite=gscholar

2. Rivera-Garcia, L. (2019) Comparative transcriptomics of the two distinct morphologies of the Carribean octocoral Briareum asbestinum. University of Puerto Rico Mayaguez. https://scholar.uprm.edu/handle/20.500.11801/2447

3. Participants

3.1. Individuals

First Name	Last Name	Most Senior Project Role	Nearest Person Month Worked	Email (if new to project)	Affil-iation	Contribution
Thomas	Doak	PD/PI	8		IU	PI
Sheri	Sanders	Co-PD/PI	8		IU	Co-PI
Matthew	Hahn	Co-PD/PI	1		IU	Co-PI
Yuzhen	Ye	Co-PD/PI	1		IU	Co-PI
Craig	Stewart	Co-PD/PI	1		IU	Co-PI
Carrie	Ganote	Other	6		IU	Software support
Therese	Miller	Other Professional	3		IU	Administration
Bhavya	Papudeshi	Other	6		IU	Consulting and software support
Winona	Snapp-Childs	Other Professional	1	wsnappch@iu.edu	IU	Administration
Dyuti	Pant	Undergraduate	3	dyupant@iu.edu	IU	Undergraduate research
Lyric	Cooper	Undergraduate	3	lycoop@iu.edu	IU	Undergraduate research
Christine	Campbell	Undergraduate	4	chmacamp@iu.edu	IU	Undergraduate research
Sarah	Washington	Undergraduate	5	saewashi@iu.edu	IU	Undergraduate programming assistant
Kate	Mortensen	Graduate Student	2	kmorten@iu.edu	IU	Graduate programming assistant
Ashley	Brooks	Graduate Student	2	brooksa@iu.edu	IU	Graduate HPC assistant
Eliza	Foran	Undergraduate	4	egforan@iu.edu	IU	Undergraduate research
Tenecious	Underwood	Undergraduate	2	tciousunderwood@gmail.com	Livingstone College	Undergraduate research
Haley	Leffler	Undergraduate	2	hleffler@iu.edu	IU	Undergraduate research

3.2. Partner organizations

Name: 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. 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.

Name: XSEDE
Partner's Contribution to the Project: Collaborative research
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 sites 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 Provider and an XSEDE Domain Champion.

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?

NCGAS was founded in 2008 in response to the gap between the heavy need of life scientists for advanced computational power and their limited utilization of nationally available cyberinfrastructure funded by the National Science Foundation. Since then, life scientists have become a significant portion of the users on various national systems (i.e. XSEDE Jetstream Cloud) and NCGAS has kept pace by diversifying training, outreach, and supported software as

the field changes - continuing to provide critical skills to the growing field of computational biology and bioinformatics. NCGAS helps smooth the transition to HPC computing for genome scientists while tempering their impact on systems by providing computational resources, consulting, and training. This impact will only grow as more and more students and researchers are pushed to computational work while labs, facilities, field stations, and campus are at reduced capacity.

NCGAS provides hundreds of researchers with accounts on high end clusters they wouldn't typically have access to on an average campus (including a 6 petaFLOP Cray). We curate a software library of hundreds of life-sciences packages to reduce the time spent installing software, which is a time consuming process. We maintain Galaxy and GenePattern gateways with over 1400 total users to make using software even easier, without sacrificing the computational power necessary for cutting edge research projects. We are active contributors to the field, consistently authoring peer-reviewed publications each year, while supporting 230 research projects. That experience is returned to our community through national outreach that has trained more than 600 researchers this year through free workshops and seminars.

4.2. What is the impact on other disciplines?

The impact of NCGAS work includes contributions to computer science, medicine, and agriculture.

Computer Science: NCGAS produces several virtual machines for distribution on the XSEDE national cloud, Jetstream. The team also provides training in UNIX operating system set up, administration of users, job scheduling, data movement, setting up web-servers, and creating and monitoring system services. Additionally, NCGAS has developed content for engaging non-computational students in learning basic UNIX, creating a My Little Pony-based dungeon crawler game (Pony Linux) that has proven popular in our HPC Onboarding course.

Medicine: NCGAS produced a variety of metagenomic tools for assembling, analyzing, visualizing, and identifying microbiomes, which is of great and increasing interest to the medical community. Additionally, machine learning work has been extended to a collaboration with IU Medical School to improve predictions of radiation-induced leukopenia, unlocking the potential to predict potentially detrimental effects of dosage within the first week of treatment (on-going work).

Agriculture: Recent collaboration with a European center produced a pan-transcriptome resource for the cultivated potato. This project extended NCGAS's de novo transcriptome analysis workflow to polyploid plants and to a pan-transcriptome (transcriptome over several species). This is an important contribution to agriculture as the only potato transcriptome resources previously available were on non-cultivated varieties and were missing a large number of the gene catalog. Increased identification of genes that are differentially found in cultivars, as well as genes that are only found in the cultivated polyploids, is a requisite first step toward selecting and improving crops.

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

- NCGAS has mentored eight non-staff students this year, all of which belong to under-represented minority groups within Information Technology. Seven of the students were women, five were of minority ethnicities. Several of these students have been with us through several programs. We have helped two secure further research positions, as well as presentations to the National Parks Service. One student credits the Jetstream REU experience (for which Sheri Sanders and Winona Snapp-Childs mentored) as his reason for going into graduate school. He started a cyber security master's program in August 2020.
- The NCGAS team provides training for these undergraduate and graduate students in not only biology and bioinformatics, but system administration, improving workforce development in cyberinfrastructure. Topics include UNIX operating system set up, administration of users, job scheduling, data movement, setting up web-servers, and creating and monitoring system services. Additionally, NCGAS has developed content for engaging other non-computational students in learning basic UNIX, creating a My Little Pony-based dungeon crawler game (Pony Linux) that has proven popular in our HPC Onboarding course. This game is inclusive, is particularly attractive to female students, and dramatically increases engagement with learning basic unix commands.

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?

Nothing to report.

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

Nothing to report.

4.7. What is the impact on technology transfer?

Nothing to report.

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

- NCGAS has mentored eight non-staff students this year, all of which belong to under-represented minority groups within Information Technology. Seven of the students were women, five were of minority ethnicities. Several of these students have been with us through several programs. We have helped two secure further research positions, as well as presentations to the National Parks Service. One student credits the Jetstream REU experience (for which Sheri Sanders and Winona Snapp-Childs mentored) as his reason for going into graduate school. He started a cyber security master's program in August 2020.
- The NCGAS team provides training for these undergraduate and graduate students in not only biology and bioinformatics, but system administration, improving workforce development in cyberinfrastructure. Topics include UNIX operating system set up, administration of users, job scheduling, data movement, setting up web-servers, and creating and monitoring system services. Additionally, NCGAS has developed content for engaging other non-computational students in learning basic UNIX, creating a My Little Pony-based dungeon crawler game (Pony Linux) that has proven popular in our HPC Onboarding course. This game is inclusive, is particularly attractive to female students, and dramatically increases engagement with learning basic unix commands.

5. Changes/ Problems

5.1. Changes in approach and reasons for change

Nothing to report.

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

Staff Member Bhavya Papudeshi has terminated her employment with NCGAS on August 31, 2020. She is transitioning to a full time PhD position with a lab she has been collaborating with while working at NCGAS. We are in the process of selecting a replacement staff member, to be welcomed to the team in October of 2020.

5.3. Changes that have significant impact on expenditures

Nothing to report.

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.