

## 1. Overview

The National Center for Genome Analysis Support (NCGAS) exists at the intersection of genomic research, the software that drives it, and the hardware that powers it. Our mission is to make the genomic data revolution more accessible by providing tools and training that allow researchers to perform cutting-edge analyses and tackle the challenges of the twenty-first century. To accomplish this mission, we provide curated, biology-specific support that reduces ambiguity in software choice, computational resource limitations, community-wide skill deficiencies, and day-to-day roadblocks. This **1) accelerates progress in genomic research by lowering the barrier to complex computational analyses** and **2) improves the quality of that research by facilitating the transfer of knowledge** between and within the biological, bioinformatic, and cyberinfrastructure communities.

NCGAS was founded in 2008 in response to the gap between life scientists' increasing need for advanced computational power and their limited utilization of nationally available cyberinfrastructure. Since then, life scientists have comprised a significant portion of users on various national systems, i.e. 24.6% on XSEDE resources [1]. As more research moves to computation-heavy analyses in response to streamlined sequencing platforms and an increase in publicly available research data, this trend will persist. Through the following four services, we aim to increase the pace at which research can be completed while also improving the community's computational skills:

- **Easy access to software within the context of genomic analysis.** NCGAS provides a curated software library of common, community-vetted genomics packages in the form of pre-installed modules on various systems, purpose-built virtual machines (VMs) for use on cloud platforms, and containers for use on local machines. The VMs and containers package workflows provide context to the software, ensure all steps are considered, and facilitate software selection.
- **Streamlined access to cloud, cluster, and scientific gateway infrastructure appropriate for diverse bioinformatic analyses.** As an XSEDE Service Provider, we maintain community and education allocations for easy onboarding of new users into the XSEDE community, and assist in acquiring project-specific research allocations. Indiana University (IU) systems supplement XSEDE resources. For the less command-line savvy, our gateways pair convenient web server-based interfaces and powerful high-memory high performance computing (HPC) hardware.
- **Free training events to optimize software, analytics, and infrastructure use.** NCGAS integrates necessary HPC skills into workflow-based course offerings to increase engagement in learning new computational skills, enable efficient completion of tasks on national resources, and put the dozens of software packages required to complete an analysis in an understandable, biological context. Training is available in-person, online, and through self-guided curricula. Online materials allow for visual and audio versions, on-demand usage, travel restrictions, and, eventually, multi-language support.
- **One-on-one consulting to supplement our other services.** Genomics and computing experts deliver consulting services to troubleshoot software use and installation, identify the best resources for different applications, and identify pertinent biases in planning and executing genomic analyses. Consulting reduces issues that persist even with accessible software, computational resources, and training while keeping us informed about community needs.

## 2. Justification

A steady average of 725 genomic and transcriptomic projects are funded yearly by the NSF [2]. These genomic researchers are faced with an array of hardware and software options, and an ever-increasing list of computational skills to learn, slowing the pace of research. NCGAS is uniquely positioned to accelerate biological research by ameliorating the computational barriers to genomic analysis. We have established a large, heterogeneous community of users who trust us to centralize knowledge about software, hardware, and computational skills. This community consistently rates our services highly and recommends us to new researchers, in part because we have successfully demonstrated that our tools and training lead to **faster production of more robust scientific products**. The purpose of this proposal is to sustain the critical, highly-refined, well-utilized services that NCGAS has provided to the

national research community since 2011. We propose further evolution of existing services to expand our reach, address changing needs, and build toward sustainable service for genomics research.

**2.1 NCGAS has built a large, heterogeneous community that relies on our services and recommends them to others.** NCGAS obtains community demographics and service evaluations from several sources: our confidential, independently managed annual user survey [3]; our resource request forms; and our workshop participant surveys [4]. These sources show that our user community spans all 50 states, as well as Puerto Rico and the District of Columbia. To date, 382 researchers use our computational resources directly on the command line, more than 1600 researchers use our gateway services, and 6519 have participated in NCGAS events. Approximately 22% of NCGAS researchers are from EPSCoR states, 6% are from Minority Serving Institutions, and 12% are from universities with fewer than 3,000 students [3].

Many researchers use our services to access computational resources, completing an average of 36% of their computation on NCGAS-supported resources (N=217), with 24% completing two-thirds or more of their computation using these resources. Transcriptome analysis and assembly are the most common analyses performed, with a growing number of population biology and genome assembly and annotation projects reported as these analyses mature and become accessible to non-model systems. These analyses are supported by our workshops, which we offer at least twice a year. In the last two years, 970 participants have taken our workshops, originating from 40 states, the District of Columbia, and Puerto Rico (13% from EPSCoR). Approximately 34% of these attendees were graduate students.

NCGAS is a vital support center for this community, with 51% of researchers reporting that NCGAS services were very or extremely important to completing their research (N=306). NCGAS is also an important amplifier for projects already funded at the national level, with 58% nationally funded (27% NSF, 17% NIH, 14% other, N=216). In addition to being vital, NCGAS is well-liked. Last year, 90% of clients reported being satisfied or very satisfied (N=236), and rated our RNAseq workflow 4.4/5 (N=475), NCGAS-supported computational resources 4.4/5 (N=133), NCGAS educational materials 4.5/5 (N=202), and NCGAS consulting services at 4.4/5 (N=158). These high ratings result in a high number of word-of-mouth recommendations (51% of acquisition, N=311), further underscoring the service's positive reputation in the community.

**2.2 NCGAS has successfully built scientifically productive tools and services.** Below we describe select projects to which NCGAS has contributed, and which NCGAS leveraged to produce tools centered around a particular analysis. They illustrate how NCGAS's expertise spans a range of client questions and technical barriers. However, these three projects hardly show the breadth of biological research projects benefiting from NCGAS services. In total, NCGAS supports 10 domain areas using genomic data, including 39 subdomains of Life Sciences, Allied Health, Biomedical Research.

**2.2.1 Agricultural Science—transcriptomics. *Background:*** Fungal disease, insect parasitism, and climate change threaten food security: economically important crops (e.g. coffee, peanuts, potatoes) experience shortages due to droughts, floods, pestilence, and abnormally high temperatures. Selecting cultivars to withstand regional environmental challenges and for disease resistance is significantly more effective with genomic resources, starting with well-annotated transcriptomes. Transcriptomes are particularly useful when species lack high-quality genome assemblies, and can be used in genome annotation once the genome is available. ***Scientific Questions Addressed:*** 1) Genetic panels for three species of coffee were created in conjunction with Cenicafe [5], allowing Colombian growers to select productive cultivars for the climate. 2) In the production peanut species, we recovered all five confirmed enzymes that drive the production of an antioxidant (resveratrol), which required the combined output of multiple transcriptome assemblers. 3) NCGAS helped define the cultivar-specific transcripts in cultivated potatoes, a major genetic resource for a polyploid crop that only had a diploid genome available. Using our transcriptome workflow, NCGAS staff generated high-quality transcriptomes for coffee [6], peanut [7, 8], and potato [9] using the best practice of leveraging several parameters across several assemblers. ***NCGAS Contribution:*** Developing a transcriptome requires bioinformatic skill in a variety of analyses, expertise with genome biology, and an understanding of the biases inherent to sequencing technology.

Best practices require many software packages to be installed and executed on large computational resources, presenting a steep learning curve to those new to HPC and/or genomics. Compounding the issue, the analysis also requires many repeated analyses on data sets, especially for novices. Each final assembly combines over a dozen separate assemblies per sample, per project. To do this efficiently, NCGAS used its expertise in transcriptome assembly to streamline the analysis while providing options for tuning parameters and mechanisms to reduce software bias. The team combined this with its knowledge of LINUX job schedulers to produce a workflow that dramatically cuts the time to complete a high-quality transcriptome. **Community Benefit:** The semi-automated workflow inspired a national workshop attended by 82 participants from 43 institutions in 26 states and 1 territory. Since 2018, this workflow has led to 39 authors producing 6 publications, 2 dissertations, and 12 presentations [10].

**2.2.2 Ocean Microbiomes—metagenomics. Background:** Climate change impacts the ocean thermal profile, endangering biogeographical hot spots like coral reefs and kelp forests, which provide habitat and are important for conservation, recreation, and commercial fisheries. Prior research has shown that environmental disturbances cause major shifts in microbial community composition. Coral and kelp species depend on their microbiomes for health, but scientists do not yet understand the specifics of how these ocean microbiomes change in response to thermal change, both seasonal and climate-driven.

**Scientific Questions Addressed:** The Dinsdale Lab, then at San Diego State University, studies the impact of microbial communities on foundational species in marine environments: 1) In collaboration with NCGAS and the Bermuda Institute of Ocean Sciences, PhD student Lais Lima used data from Bermuda to develop a model that determines microbial community structure based on temperature variation. This model adapts to predict the effect of rising seawater temperature on coral reef health around the world, and could help mitigate rapid coral reef decline. 2) Additional microbiome data from Catalina Island, California showed that microbial communities were not seasonally variable and were not impacted by environmental variables like dissolved organic carbon, pH, temperature, etc. While normally stable, microbial community structure can be transformed by invasive algae, stymying the recovery of the economically and ecologically important kelp forests. **NCGAS Contribution:** Senior staff Papudeshi and various REU students assembled and annotated metagenomes on IU's Carbonate cluster, XSEDE Jetstream, and Jupyter notebooks. NCGAS produced taxa lists and functional annotations for over 50 microbiome samples using NCGAS's curated software. The team developed workflows to run the taxa identification, functional annotation, and reconstructions of metagenome-assembled genomes. NCGAS developed Jupyter support for this project in response to growing interest in the community. **Community Benefit:** Both projects supported dissertations [11, 12], and led to two co-authored publications [11, 13]. Two workflows on metagenome analysis are now available on NCGAS GitHub, an Anvi'o image to visualize metagenomes is available on Jetstream, and three blog posts about the metagenome analysis, which have been visited 646 times, are up on the NCGAS website. NCGAS also developed a metagenomics workshop around this workflow, which was taught to 28 participants in Fall 2019.

**2.2.3 Island Biogeography - Phylogenomics. Background:** Island systems have a complex biogeographical history that influences the distribution of resident species. These areas are often hotspots of biodiversity, but understanding the rules governing the patterns of genetic diversity on the islands requires nuance due to frequent introgression and recent divergence. Genomic-level analysis is becoming a popular way to address these questions, and robust phylogenetic analyses are actively developing. However, many of these analyses require millions of repetitions to complete, resulting in computational jobs that can take months. **Scientific Questions Addressed:** The Richards-Zawacki Lab at the University of Pittsburgh has obtained genomic-level data from the strawberry poison-dart frog, an amphibian native to the Bocas del Toro province of Panama. NCGAS leverages this data to elucidate the driving factors behind the frog's extraordinary color variation. This amphibian offers a model of island evolution dynamics that will help identify larger evolutionary patterns in other species, especially other amphibians experiencing decline in this biogeographic hotspot. **NCGAS Contribution:** Questions of geographic distributions of genetic variation usually require two steps - sequencing data preprocessing and phylogenetic analysis. Both steps vary greatly depending on experimental design and sequencing technology. Preprocessing reduces the technical bias, if done correctly, but introduces bias if done

incorrectly. Phylogenetic analysis is highly sensitive to methodology and can also skew results if not executed carefully. In these cases, interdisciplinary NCGAS staff provide guidance in separating the technical noise from the complex genetic processes occurring in these systems. Our dual expertise in both HPC and biology leads to successful identification of real, meaningful patterns in the data in much less time. These analyses also require more advanced cyberinfrastructure and software support. NCGAS PI Sanders worked with Jetstream to test and implement a GPU-based software acceleration of the workflow (BEAGLE) that accelerates phylogenetic analyses through the use of GPUs. **Community Benefit:** In addition to offering direct support to peer-reviewed publications and two PhD dissertations on island phylogeography [14, 15], NCGAS created a phylogenetic analysis guide (accessed 3,609 times on our website) and is developing a workshop on population genomics analysis for non-model organisms. We expected to hold it at the now-canceled Joint Meeting of Ichthyologists and Herpetologists; it was subsequently requested for the American Fisheries Society by the Douglas Lab at University of Arkansas, indicating an increasing interest in taxonomy-level domains in phylogenomics. Additionally, we drafted a GPU-based phylogenetic analysis VM and tutorial. This initial image will serve as the basis for future GPU image development, including GPU accelerated variant-calling (via software such as parabricks).

**2.2.4 Other research products enabled by NCGAS support and past NSF funding.** Since 2011, NCGAS has enabled efficient creation of scientific products, including 197 peer-reviewed journal articles (including two Science papers)[7, 9, 11, 13, 14, 16–207] and 9 dissertations[15, 208–215]. During that time, NCGAS staff has authored 48 peer-reviewed journal articles, 85 conference presentations, and 1 educational resource (e-textbook). NCGAS clients have produced more than 80 bioprojects on NCBI, including 40 transcriptomes and 16 genomes on public repositories.

**2.3 Client usage statistics for software and computing services.** Table 1 summarizes the key software curation, distribution, and provision services provided by NCGAS, routinely offered in collaboration with other NSF awardees. This table and the next provide data subdivided by the two prior NSF awards supporting NCGAS and the current award.

**Table 1. Summary of software and compute resources provided over time (time intervals vary)**

|                                                    | NCGAS Development:<br>Sept 2011 - Sept 2015 | NCGAS Sustaining:<br>Sept 2015 - Sept 2018 | NCGAS Sustaining:<br>Sept 2018 - Mar 2020 |
|----------------------------------------------------|---------------------------------------------|--------------------------------------------|-------------------------------------------|
| <b>Software curated and distributed per period</b> |                                             |                                            |                                           |
| Total packages                                     | 77                                          | 303                                        | 351                                       |
| Total Cloud Images                                 | 0                                           | 13                                         | 19                                        |
| Total available GitHub repositories                | 0                                           | 0                                          | 16                                        |
| <b>Computational resources</b>                     |                                             |                                            |                                           |
| Virtual core-hours on Jetstream                    | NA - Jetstream was not yet in production    | 2,400,000                                  | 1,000,000                                 |
| Distinct genome analysis cloud images launched     | N/A                                         | 675                                        | 799                                       |
| Core-hours on Bridges (storage, GB)                | 0<br>0                                      | 37,748<br>18,383                           | 47,700<br>31,000                          |
| Core-hours provided by IU compute clusters*        | 3,827,291                                   | 5,317,501                                  | 1,885,293                                 |
| <b>Storage resources provided by IU</b>            |                                             |                                            |                                           |
| IU disk storage (GB)                               | 1,082,909                                   | 2,150,154                                  | 2,901,624                                 |

|                      |         |         |           |
|----------------------|---------|---------|-----------|
| IU tape storage (GB) | 240,569 | 629,303 | 1,141,392 |
|----------------------|---------|---------|-----------|

*\*See Facilities and Other Resources Document for description of computing resources*

**2.4 Client usage statistics for education and training services.** NCGAS has offered education and training opportunities since 2011; however, the nature of these offerings has changed over time. During the Development award (September 2011 - September 2015), we offered mostly 1-2 day training sessions on RNA-seq or Galaxy [189]. As demand for NCGAS services increased, NCGAS began to provide more robust education and training services using scalable educational platforms and developing online versions of course material. Table 3 below the education and training NCGAS has offered, and reflects this shift in strategy. The move to online instruction helped NCGAS adapt to COVID-19 closings, allowing us to scale our offerings for students and researchers who had moved to computational projects while labs are at limited capacity. Note that class and session numbers only include hands-on training opportunities and not outreach activities, presentations, or guest lectures. Education, Outreach, and Training (EOT) events refer to all outreach.

**Table 2. Summary of NCGAS education and training services (note that time intervals vary)**

|                                    | NCGAS Development: Sept 2011 - Sept 2015 | NCGAS Sustaining: Sept 2015 - Sept 2018 | NCGAS Sustaining: Sept 2018 - Mar 2020 | Total: Sept 2011 - Sept 2020 |
|------------------------------------|------------------------------------------|-----------------------------------------|----------------------------------------|------------------------------|
| Class titles offered               | 5                                        | 5                                       | 4                                      | 11                           |
| Class sessions offered / attendees | 9/262                                    | 9/375                                   | 10/1072                                | 28/1709                      |
| EOT (events/attendees)             | 91/4498                                  | 46/955                                  | 29/1556                                | 166/7009                     |

**2.3 Client usage statistics for consulting services.** Table 3 describes the quantitative history of NCGAS consulting activities. Here, the term “client” refers to any user requesting access to NCGAS computational resources. Consultation focuses on general support, including help troubleshooting software errors, installing software, and setting up VMs. Longer consultations (80+ hr) usually include intellectual contribution and co-authorship on research products.

**Table 3. Summary of NCGAS consulting services (note that time intervals vary)**

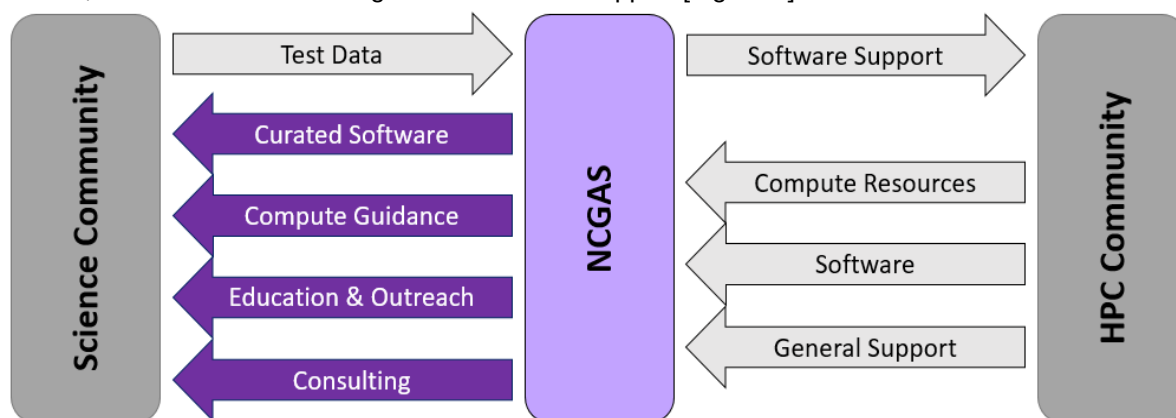
|                      | NCGAS Development: Sept 2011 - Sept 2015 | NCGAS Sustaining: Sept 2015 - Sept 2018 | NCGAS Sustaining: Sept 2018 - Sept 2020 | Total: Sept 2011 - Sept 2020 |
|----------------------|------------------------------------------|-----------------------------------------|-----------------------------------------|------------------------------|
| New Clients          | 95                                       | 196                                     | 91                                      | 382                          |
| N States Represented | 30                                       | 36                                      | 26                                      | 46                           |
| Cumulative Clients   | 95                                       | 291                                     | 382                                     | 382                          |
| <4hr Consults        | 781                                      | 1335                                    | 548                                     | 2664                         |
| 4 - 80hr Consults    | 203                                      | 137                                     | 35                                      | 375                          |
| 80+ hr Consults      | 7                                        | 10                                      | 15                                      | 32                           |

### **3. Description of product: software, cyberinfrastructure, training, and consulting.**

The functional goal of NCGAS is to reduce the “bioinformatics bottleneck,” increasing the efficiency with which researchers perform genomic analyses [216, 217]. The cost of sequencing has dropped markedly, creating an opportunity to use vast quantities of data to answer previously unapproachable, multi-level questions. However, though data generation has become easier, it requires complex analyses on large computational resources, adding time to research projects when researchers lack necessary training.

NCGAS eliminates this bottleneck by reducing the time researchers spend on software management, learning new analyses, adapting analyses to specific questions, and inefficient hardware

use. NCGAS facilitates the adoption of already-existing hardware and software by providing support and curated content for the genomic community. We use the computational resources of HPC centers (i.e. XSEDE and IU), and layer them with biological context, software, education, and consulting to bolster the HPC offerings. In return, we get test data from researchers from which we build workflows, test software, and offer HPC centers genomic software support [Figure 1].



**Figure 1: Structure of Service.** NCGAS consumes computational resources and uses them to produce software support, curated computational tools, consulting, and targeted training for the national science community. It also provides a return to the HPC community through software support, testing, and more efficient use of the systems through user training.

**3.1 Software Support and Packaging.** NCGAS provides a curated software library of genomics packages pre-installed on nationally accessible XSEDE resources, such as Bridges at Pittsburgh Supercomputing Center (PSC), Jetstream at IU, and the Texas Advanced Computing Center (TACC). Currently, NCGAS supports 351 software packages on various clusters, 19 images on Jetstream, and 16 GitHub repositories [Table 1]. Software is available in several formats including command line (on XSEDE and IU clusters), graphical (on Jetstream), and as gateways (Galaxy, GenePattern, Genome Browsers). NCGAS has moved toward packaging software in ways that work for national repositories (i.e. VMs, biocontainers, conda, XSEDE software repository) as these distribution platforms have matured. Individual packages are easier for users to install on national and local hardware.

However, simply offering individual software packages does not accelerate research; researchers must be able to choose the software they want and run that software easily. Technology adoption is influenced by perception of its usefulness and ease of use [218, 219]. Researchers are more likely to see tools as useful when those tools are accepted by the community, and when they trust that the output is useful to their research [219]. NCGAS began packaging software into workflows to address this tendency; putting software packages into the context of full analyses clarifies the utility of the whole. Workflows also allow us to highlight the choices and biases at any step (i.e. different choices of differential expression algorithms), further bolstering trust in useful, flexible output. Researchers perceive tools to be easier to use (and therefore better choices) when tasks seem accessible and support is available [219]. Workflows are easier to package with demo data that can be run quickly to provide relevant final results, and the focus on populating the national repository with required software allows for build-on-the-fly dependencies to be incorporated into workflows. The demo data and workflows are also the basis of our associated workshops. Workshop materials are available online via git and youtube with links to multiple tutorials, videos, activities, and journal articles; these resources build real support for the workflow, even if researchers aren't able to attend the workshops.

Currently available workflows include microbiome analysis with visualization and comparison to national databases, phylogenetic analysis of population structure in non-model species, and transcriptome assembly and annotation for de novo and genome-guided systems. The popular transcriptome-based workflow is the most developed and has been readily adopted: in two years, 39 authors have produced six publications using the workflow (more are in development); two dissertations have used the workflow for a substantial portion of the research, and twelve conference presentations

have been given on research resulting from its use [10]. We are developing and testing more workflows as new analyses mature in the field; preliminary evaluation of potential methylation sites in publicly available genomes and genome annotation is one such example.

**3.2 Providing and Facilitating Efficient Use of CyberInfrastructure.** Synthesizing software into analysis-based workflows in containers and VMs also increases our ability to support university- and college-funded cyberinfrastructure (CI), and thus helps meet demand for computational biology at a low incremental cost to the federal government. However, effectively completing large-scale genomics research requires more than simply providing software in approachable ways. Researchers must decide when cloud, cluster, or gateway platforms are most appropriate for their needs; they must be able to find, access, and use those systems; and they must be able to get their data and the appropriate software on whatever system best fits their needs. NCGAS provides guidance on selecting compute resources, streamlines access on HPC resources, and facilitates data movement and management between them. When combined with the easy-to-deploy or ready-to-use software, workflows, and gateways described above, NCGAS addresses all major roadblocks to appropriate infrastructure choices.

Similar to our approach to improving software adoption, NCGAS boosts the adoption of diverse cyberinfrastructure by addressing ease-of-use and perceived utility [218, 219]. NCGAS demonstrates the ease of using appropriate resources by including workshop activities on selecting computational resources to fit analysis, archiving, collaboration, and visualization needs and demonstrating the ease of transferring data using modern protocols. We make cyberinfrastructure use more approachable by constructing interactive games for exploring basic command line skills. We maintain XSEDE allocations for the NCGAS community, which allows researchers to explore options first hand before we assist them in acquiring their own research allocations.

Since NCGAS began, 86 researchers have used our community research allocations, accounting for 116,000 core-hours on Bridges and over 4.7 million (virtual) core-hours on Jetstream; eight have gone on to get their own allocations. NCGAS also provides access to IU's compute resources (Carbonate [220], BigRed [221], and SDA [222], and Big Red 200, when available). NCGAS has provided 382 accounts on IU systems to clients across 46 states and Puerto Rico. This includes short-term spinning-disk storage on XSEDE Bridges Pylon and IU Slate during active analysis.

For the less command-line savvy, NCGAS has deployed customized versions of common gateways (Galaxy, GenePattern, externally funded) to leverage the convenience of web-server-based graphical interfaces and the high-memory HPC back-ends to provide sufficient power for advanced research. These scientific gateways lower the barrier of entry to computational analysis. Over 1400 researchers in 48 countries have been served by these gateways, building a solid perception of community acceptance and utility.

**3.3 Education and Training in HPC and Computational Analysis.** NCGAS education and training is designed to support software and cyberinfrastructure services by adding to their perceived ease-of-use. Skill building through games and demos increases experience and reduces anxiety; full courses demonstrate community acceptance; hands-on activities demonstrate utility first hand; and the volume of content showcases the support available through NCGAS.

Our training materials increase engagement in learning necessary computational skills by relating them directly to biological analyses. Computation training focuses on essential skills, curates information from academic CI groups such as XSEDE Campus Champions, IU's University Information Technology Services (IU UITS), Information Technologies for Cancer Research (ITCR), and Pittsburgh Supercomputing Center (PSC). HPC skills are taught in concert with genome analysis workflows as multi-day national workshops (both in-person and online), as tutorials on the NCGAS blog, and as one-on-one client training (via extended consulting to develop or extend workflows). This improves the community's ability to run jobs, select appropriate resources, and explore biases in software while maintaining engagement with the functional examples of our established workflows. HPC and bioinformatic skills are also taught in undergraduate education programs such as IU's Jim Holland Program, Research Experiences for Undergraduates (REU) with Jetstream, and the Center of Excellence for Women & Technology (CEWiT), which build students' computational and biological skills.

These training events are popular and effective. NCGAS has held 165 training events, serving over 6500 researchers. In the last three years, we held seven nationally-available workshops on our mature workflows (de novo transcriptome assembly and analysis, metagenomic analysis) and specific skills (introduction to R for biologists) with a total of 785 national applicants for 661 available seats. For each of these workshops, we employ anonymous, independent pre- and post-workshop surveys that include a self-assessment of skills. Reliably, we see an average increase of 20% in confidence (1 point on a 5-point scale) for skills taught in both online and in-person workshops [4, 10].

Our training materials must be accessible to as wide an audience as possible, thus, we are focused on improving this content to increase our reach. In the development award, we had offered participant support for in-person training, which increased applications by underserved members of the community (i.e. EPSCoR institutions, MSI, HBCU, under-represented minorities). We will offer this travel support again to people attending our workshops, or presenting on research resulting from our training. Moving our content online allows for larger reach, and the asynchronous content has been popular as it allows for flexibility in working around other obligations. Additionally, video and print content increase accessibility for hearing- and sight-impaired researchers, and it facilitates easy translation of captions and text to Spanish. We recognize the importance of reaching Spanish-speaking constituencies, as the U.S. has the fifth-largest Spanish speaking population in the world, with over 13 million adult Americans who speak primarily Spanish, and we will port our tutorials and online resources to Spanish [223].

**3.4 Consulting.** NCGAS supplements the above services with rapid-response consulting to troubleshoot software use and installation, to identify the best resources for different applications, and to identify pertinent biases in planned genomic analyses or sequencing platforms. Rapid-response consulting reduces the progress-slowness effect of day-to-day issues present even with accessible software, computational resources, and training. When we notice questions have become common, we convert their answers into tutorials available on our blog, reducing staff time-spend. We are seeking a platform for bi-directional community interaction (such as a google group), making it possible for researchers to interact, and for vetted clients to answer questions, while also adding a search feature.

Consulting is initiated through the help ticket system and through conversations at conferences, workshops, and presentations. NCGAS sees an average of about 300 short (<4 hr) tickets per year and an additional 40 or so consults over 4 hours long [Table 3]. Consultations longer than 40 hours require applications for extended consulting with NCGAS staff, which includes defining project goals, deliverables, timelines, and authorship of publications. Accepted consultations must either lead to the development of community tools and workflows, or they must be externally funded.

### **3.5 NCGAS in the context of other services funded by the federal government or individual institutions of higher education.**

**3.5.1 Support for NSF-funded software creators and creations.** First, our software curation and distribution efforts complement and support access to and use of many NSF-funded software products. Of the 382 packages in our current curated library, most were created with NSF support or were NIH-funded applications that have been repurposed for NSF-type research (e.g. Trinity). In particular, NCGAS supports and complements major federal investments in projects such as Galaxy, Trinity, and GenePattern. One of NCGAS's distinguishing characteristics is that we do not generate our own major bioinformatics software. We are not competing with existing software producers, although we have made important contributions to software like Trinity [33, 55] and earlier phylogenetic tools like fastDNAmI [224]. Rather, we facilitate the use of software created with NSF and NIH funding in a larger context and with increased ease of use, multiplying the value of federal investments used to create these software.

**3.5.2 Support of NSF-funded software on NSF-funded CI systems.** NCGAS adds value to major NSF investments in systems like Jetstream and Bridges (and soon, Jetstream2 and Bridges2) by installing and maintaining vetted software to support full analytical workflows on these systems, and by facilitating installation of workflows on other NSF-funded systems supported via XSEDE. Thus, leading-edge national facilities have support for cutting-edge genomic analyses, and the domain support necessary to ensure their functionality. NCGAS is a member of the XSEDE Service Provider Forum [225] and through its activities, creates consistency in software and version availability across XSEDE-supported CI



systems. NCGAS depends on the national HPC environment, and contributes to its function.

3.5.3 Support and on-boarding of NSF-funded resource platforms. Computing time on NSF-funded national CI resources coordinated through XSEDE is allocated through a competitive national process, with reviews for allocations made on a quarterly basis. The resources allotted are significant, and the application process requires correspondingly significant effort. NCGAS assists in selecting systems and requesting resources, decreasing the barrier to entry for researchers.

3.5.4 Participation in NSF-funded outreach efforts to traditionally under-represented groups in computing. NCGAS utilizes XSEDE's extensive resources to engage a diverse community. The XSEDE Broadening Participation program will include NCGAS information in their visits to schools and conferences, which serve large groups of students from traditionally underrepresented groups, such as Tapia Celebration of Diversity in Computing. NCGAS participates in Campus Champions and Domain Champions - a community of 680 research computing facilitators in 325 U.S. institutions - increasing our reach and interaction with other XSEDE domain and computing groups.

3.5.5 NCGAS provides computational facilities for bioinformatics research that complements major NSF investments. NCGAS differentiates itself from gateways such as GenePattern Cloud, Galaxy Main, CIPRES, and CyVerse. GenePattern and Galaxy Main provide easy-to-use gateways, which can house our workflows and leverage our software packaging on machines that support their work (i.e. Bridges, Jetstream). CIPRES supports a small selection of software for phylogenetic analysis, for which we provide educational materials. CyVerse offers tools and platforms, and NCGAS provides complementary training, consulting, and support for researchers using their institution's hardware or XSEDE resources.

3.5.6 Archiving of data with NSF-funded infrastructure. NCGAS will use the current award to obtain membership with Dryad, the NSF-funded data archiving service (NSF #1933812), to provide NCGAS researchers with data archiving to support publications that are supported by our service.

**3.6. Projections of usage during proposed award period.** NCGAS is focused on scaling services to serve a growing community without increasing cost. Our effort is shifting from individual assistance to more focus on common problems in the community, with the following projections [Table 4].

- **Software:** Efforts are underway to improve software distribution on more diverse cyberinfrastructure, with genomic containers being deployed on Bridges-2. Four containers have been built and will be converted to Biocontainers before the granting period in addition to 29 Biocontainers implemented from the repository, with the expectation that streamlining the process will lead to an additional 20 containers contributed each year.
- **Cyberinfrastructure:** We expect a continued 10% growth in our core hours usage, as has been the trend for the last three years. We also predict that client numbers will continue to grow at the typical 10-15% we have experienced in the last four years. We expect our consulting numbers to remain consistent as they have for the last four years.
- **Education, outreach, and training:** In order to meet the ever-increasing demand for workshops, we will continue to move workshops online, allowing our contact hours to scale with community need. This is particularly important in response to increased demand for online education during COVID-19. As a result of this migration, we expect to add 200-300 more participants each year as we scale course size without adding more offerings (due to staff size).
- **Consulting:** Consulting will continue to balance day-to-day community support through 1) ~250 short consultations we provide every year; 2) two extensive collaborations per staff member per year that serve to keep NCGAS current and respected in the field.

NCGAS will use the time gained through increased efficiency to increase time allotted to EPSCoR states, MSI, HBCU, and smaller institutions. Researchers at these institutions may need more support with deploying software, accessing cyberinfrastructure, and completing analyses effectively, due to limited local HPC resources. As such, an individual-level focus will remain for these institutions.

**Table 4. Projections of future NCGAS usage**

|              | Software                | Cyberinfrastructure    |                     |                 |             | Education, Outreach, Training |              |               | Consulting |         |                |
|--------------|-------------------------|------------------------|---------------------|-----------------|-------------|-------------------------------|--------------|---------------|------------|---------|----------------|
| Project Year | Software added to repos | Core hours XSEDE (mil) | Core hours IU (mil) | Gateway Clients | HPC Clients | Events                        | Participants | Contact Hours | < 4 hr     | 4-80 hr | Collaborations |
| PY9*         | 0                       | 1.0                    | 1.15                |                 | 375         | 13                            | 1074         | 46            | 287        | 11      | 2              |
| PY10**       | 5                       | 1.1                    | 1.42                |                 | 425         | 13                            | 1000         | 92            | 250        | 10      | 4              |
| PY11         | 20                      | 1.21                   | 1.56                |                 | 489         | 13                            | 1200         | 92            | 250        | 10      | 4              |
| PY12         | 20                      | 1.33                   | 1.72                |                 | 563         | 13                            | 1500         | 92            | 250        | 10      | 4              |
| PY13         | 20                      | 1.46                   | 1.89                |                 | 648         | 13                            | 1800         | 92            | 250        | 10      | 4              |

\* indicates project years from current grant, \*\* indicates current year predictions

## 4. Technical implementation

**4.1 Universally-accessible software distribution across institutions and platforms.** NCGAS will build and maintain all of our software in containers that will be hosted in community container registries and incorporated into workflows that can be easily and reproducibly executed across any institution or compute platform. To accomplish this, NCGAS will make all software available as containers through the BioContainers community project [226], which already has dozens of contributors providing thousands of containerized bioinformatics packages in a community container registry [227]. Biocontainers also automatically builds containers for all bioconda packages, allowing a single source for multiple repositories we aim to support. NCGAS will either update existing BioContainer software or add containers currently missing from the BioContainer registry, which can then be pulled and run using either Docker [228] or Singularity [229]. NCGAS also plans to update current workflows and build new workflows using Nextflow [230], a powerful workflow language that works natively with Docker and Singularity containers. Nextflow workflows can be readily deployed on local HPC (e.g. via SGE, LSF, SLURM, PBS, HTCondor), cloud (Kubernetes, AWS, Google Cloud), or hybrid systems, and can be centrally launched and monitored via Nextflow Tower [231]. Many community-curated bioinformatics Nextflow workflows are already available through the nf-core project [232], which plans to collaborate with the BioContainers project to standardize these community workflows. These efforts will enhance the uptake and impact of NCGAS software and workflows. In these efforts, NCGAS will build on our methodology developed over the past several years for converting analytical best-practices from the community into workflows well-suited to HPC use. By moving individual packages to these platforms, distributed workflows can pull appropriate versions, eliminating the need to install software before running analyses, and freeing users to run workflows on any system to which they have access, including NSF resources like the upcoming Jetstream2 and Bridges-2 systems.

Bridges-2 and Jetstream2 will enable early testing of new deployment strategies, with further testing on other NSF systems, such as SDSC's Expanse. Deployment on AWS and Google Cloud systems using Nextflow is already robustly tested and documented. Upon hardening this distribution structure, we will develop training materials to enable biologists to run these community workflows, and to build and distribute their own workflows to increase replicability without computational roadblocks.

**4.2 Facilitating use of NSF-funded computational resources and supplementing with IU cyberinfrastructure.** NCGAS obtains compute and storage resources from various XSEDE systems through community and educational allocations through which NCGAS clients are able to access these resources. The primary resources are Bridges, which will be replaced by Bridges-2, and the Jetstream cloud computing system, which will be replaced by Jetstream-2. IU has provided compute clusters with sufficient memory and computational power to enable most genomics workflows. NCGAS provides access to these resources through an internal allocation request system.

XSEDE Bridges supports analyses that require more memory than can be supported by researcher's local clusters or IU's Carbonate. Both of these systems provide users with storage hardware,

either in the form of community allocated Pylon storage on Bridges, or Slate project space on IU systems. Jetstream supports the webserver images that host our gateways, which (in the case of Galaxy) then generates jobs on IU's Carbonate system. These three systems are therefore critical to our provisioning researchers with cyberinfrastructure upon which to complete their work. These systems all also provide Globus end-points for fast, reliable data transfer, which we leverage in helping researchers design their projects' use of different compute systems. To a lesser degree, NCGAS relies on the national repositories such as NCBI, EDI, and genome browsers to support our researchers' work.

Policies for providing access to these platforms through our community allocations and IU's affiliate system include requiring annual confidential surveys to assess the project's intellectual merit, its associated funding, its fit within NCGAS scope, and a commitment to share supported projects with NCGAS. Annual usage statistics are tracked per user on XSEDE's allocation portal and through communication with individual system administrators. XSEDE clients using excessive amounts of the allocations are assisted in acquiring their own XSEDE allocations. Researcher utilization of all of these resources will be greatly enhanced by our move to deploy containers in community registries with Nextflow workflows that enable execution anywhere (see Section 4.1).

**4.3 Providing training on a scalable, nationally-accessible platform.** Education and training operates through IU PTI's Research Technology Outreach infrastructure, which handles registration, workshop orchestration, and facilities management [233]. Further training requests come in through IU PTI and XSEDE's ticketing systems, and offerings are advertised through XSEDE Campus Champions, Evolution Directory, local campus mailing lists, NCGAS's community mailing list, and social media (twitter/facebook/website). We will collaborate with XSEDE Broadening Participation to increase our visibility among MSIs and traditionally underserved groups (TUGs) at outreach events and conferences.

The workshops often use Jetstream or IU's computational resources (including designated allocations). Materials are made available via external, free platforms such as GitHub [234] and YouTube [235] to allow for persistent availability. Workshop topics are determined by recommendations made in our Annual User Survey and by our Stakeholder Advisory Board (SAB), on post-workshop surveys, and by workflows developed in response to common analyses. Participants are selected based on applications, which are evaluated for project merit, EPSCoR/MSI status, funding status, and career level. All materials are available online within a week of the workshop via github, allowing for access to videos, slides, and scripts. More polished versions are available within one year as online courses on the Canvas platform.

We will use the template of the successful *Cena y Ciencias* outreach program, which began at the University of Illinois, Urbana-Champaign to provide high-quality Spanish content in collaboration with the Genomics and Eco-evolution of Multi-Scale Symbiosis (GEMS) Biology Integration Institute, and with the support of SACNAS at IU. Providing Spanish-speaking role-model scientists to students, and increasing the audience for our online resources, will strengthen outreach efforts and broader impacts.

**4.4 Making consulting services sustainable.** NCGAS consulting services rely on the Pervasive Technology Institute ticketing system and IU-supported email and communication tools. All tickets are tracked within the ticketing system for time dedicated to that ticket. Once 40 hours per year of time has been dedicated to a project (80 hours for EPSCoR institutions, MSIs, and institutions under 3000 students), we require a meeting with leadership to discuss goals, deliverables, and timeline for our consultation, and an application for extended collaboration. All collaborations must pass evaluation of scientific merit, fall within NCGAS activities, and either be funded or lead to a community product (see use cases in Section 2.2). A maximum of two active collaborations per staff member and at most 15% of each employee's time can be spent on approved collaborations. All applications that we cannot support will receive help finding other resources, such as XSEDE's Extended Collaborative Support Service.

## **5. Project Management Operations**

Proposal PI Dr. Sanders will be responsible for oversight and management of NCGAS daily activities, and will lead strategic planning activities with NCGAS leadership. Project management will be done via Teamwork project planning software, now the standard for the IU Pervasive Technology Institute. Software packaging, bug fixing, risk tracking and management, and tracking of long term consultations

will be managed in the IU PTI ticketing system. Table 5 lists named personnel and staff functions that will enable effective execution of the activities proposed here. “Real time” communication is and will remain in Slack, with live meetings in Zoom. Team notices, Standard Operating Procedure documents, etc., will be created and maintained in a private Confluence wiki.

**Table 5. Roles of key project team members**

| Key Personnel          | Role                                                                                                                                                                                                                            |
|------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Sheri Sanders: IU PI   | Principal Investigator: provides overall project leadership and NCGAS group management, develops flexible community workflows and associated EOT.                                                                               |
| Matthew Hahn: IU Co-PI | Collaborates with NCGAS through training offered collaboratively with the Center for Genomics and Bioinformatics, which Hahn leads. Hahn also advises on biology community needs and software to support those needs.           |
| Yuzhen Ye: IU Co-PI    | Advises NCGAS on new algorithms for genome analysis and performance characteristics of genome analysis software.                                                                                                                |
| Irene Newton: IU Co-PI | Current biology professor; has been invited.                                                                                                                                                                                    |
| NCGAS staff            | NCGAS staff will maintain gateway efforts, software installations, consult on research projects, and help develop educational materials. Coordinates software repository activities and manages other relationships with XSEDE. |
| Phil Blood: PSC PI     | PI on Linked Collaborative Proposal from Pittsburgh Supercomputing Center. Personally involved in consulting on community genomics. Will coordinate software installation of NCGAS-curated software on PSC Bridges-2 system.    |

Service standards. Service standards for our key activities are and will be as follows (these apply to NCGAS staff directly and to partner institutions):

- Software library maintenance. New versions of software supported by NCGAS will be made available in NCGAS-managed software repositories within 1 month of new stable releases.
- New workflows will be implemented across hardware systems supported by NCGAS within 2 months of new stable releases.
- Training needs and class offerings will be evaluated and planned three times per year in preparation for traditional university fall, spring, and summer semesters.
- Requests for consultations will be acknowledged by the next business day and resolved within 4 hours of staff time and 7 business days. Consultations longer than 4 hours will necessitate a meeting and agreement of goals of consultation/collaboration, end products, and a timeline.
- Access to services: Consultations almost always begin via email to the ticketing system; we will verify that email addresses of those requesting help are from accredited US colleges or universities. Requests for access to computational resources include a survey allowing for the evaluation of the project’s merit, appropriateness of our support, and US affiliation.

### **5.3 Mechanisms for interacting with the user community.**

**5.3.1 Stakeholder Advisory Board.** The NCGAS SAB includes experts in biology, bioinformatics, and cyberinfrastructure, and is chaired by Jeff Pummill of the University of Arkansas. For current members, see Letters of Collaboration from Bright, Clements, Isokpehi, Lynch, Pummill, Schwartz. The SAB holds annual meetings to provide advice on community needs, practices, and opportunities for synergistic activities with regional and national groups, and files an annual report.

**5.3.2 Consultation and Help Desk.** Researchers can ask questions and initiate consulting interactions via [ncgas.org](http://ncgas.org) or direct email. Our lightweight application for services is designed to verify that applicants have appointments at appropriate organizations (colleges, universities, or federally-funded research centers), and to ensure that services requested match with those we offer. We understand the presence of federal funding for a research project to indicate that it has passed peer review; we require “sanity check” by staff before accepting non-federally funded projects to ensure they fit with NCGAS goals.

**5.3.3 Online Media.** NCGAS conveys information, persistently, via online media. NCGAS' redesigned

website [236] and searchable blog allow for community-wide information distribution when we receive an interesting or common query or have information on a new resource of interest. Social media allows for in-the-moment communication about events, maintenance schedules, and software updates, driving 78% of our website's initial entry traffic [237]. YouTube and self-guided online courses allow participants to review workshop materials. We also underwrite Science Node, a free online publication with a worldwide readership of over 140,000, and contribute "underwriter content" about the research NCGAS supports.

**5.3.4 Education, Outreach and Training (EOT) events.** NCGAS has built a strong educational program, disseminating information through several formats. A brief summary of participation is in Table 2.

## 6. Outcomes Assessment

Below we identify the metrics we will use to measure success toward project goals (see Table 6). We use an independent evaluator to deploy anonymous annual user surveys containing information on client satisfaction and demographics, as well as pre- and post-workshop surveys to evaluate learning outcomes. We collect information on the other metrics as part of our usual activities from the systems we support, the communication systems we use, and our web analytics.

**Table 6. Outcomes assessment metrics and goals**

| Project Goal                                  | Metric                                                   | Goal                                                                           |
|-----------------------------------------------|----------------------------------------------------------|--------------------------------------------------------------------------------|
| General                                       | Client satisfaction                                      | >90% client satisfaction                                                       |
| Intellectual merit:<br>Software Support       | Software packages supported                              | Support 90% of software suggested by SAB, in reports, as VMs and/or containers |
|                                               | Cloud images launched                                    | ~300 per year                                                                  |
| Intellectual merit:<br>Consulting             | <4 hr / 4-80 hr / 80 + hrs / projects per year           | 250 / 10 / 4 / 45                                                              |
| Intellectual merit:<br>Training and Education | Workshops given; Workshop applications and participation | 6 workshops/year; 100% participation; 125% oversubscription of applications    |
|                                               | Participant conversion to clients                        | 20% conversion                                                                 |
|                                               | Self-assessment of learning                              | 1-point increase in skill levels on 1-5 likert scale                           |
| Science impacts                               | Publications NCGAS team / clients                        | 2 publications / 25 publications, per year                                     |
| Community reach                               | Social media metrics                                     | Add 50 followers/year (Twitter); 10% increase website hits and blog hits       |
| Broader impacts                               | Students mentored                                        | 2 pre-college, 4 undergraduate per year                                        |
|                                               | Student clients                                          | > 30% student participation                                                    |
|                                               | Diversity in clients                                     | > 20% clients underrepresented minorities                                      |

## 7. Business planning

**7.1 Short-term support mechanisms.** Currently, NCGAS staff are collaborating with the Genomics and Eco-evolution of Multi-scale Symbioses [GEMS, NSF #2022049] Biological Integration Institute, for which NCGAS will provide training and software support for researchers associated with the project. Workshops developed and run for this group will be supported by this grant, but the material will become part of the larger repository of educational materials available to the full community. Similarly, time to manage, maintain, and troubleshoot software for the group will be funded by GEMS, but the packages and any associated tutorials and blogs will be available to the full NCGAS community.

**7.2 Potential long-term support mechanisms.** In addition to pursuing awards similar to GEMS and GenePattern, we will seek supplemental support for NCGAS activities by remodelling our extended collaborations. After the initial threshold of 40 or 80 consulting hours (depending on institution type),

collaborations must either generate a community tool (to continue at no cost), be funded externally through charge-back of consulting fees, or by direct funding of staff on grant subawards. We will use this award period to establish these funding streams, generate testimonials, and establish a recruitment plan for further work. To date, NCGAS has established long-term funded collaborations with the USDA, Cornell University, and University of Arizona.

We are also considering support through development of subscription services to academic departments or other small research groups such as field stations or non-profits. Subscription services would cover access to novel educational content and priority support, and would be renewed/charged annually. In addition, we will pursue becoming an Extended Collaboration Support Service (ECSS) formal partner in the XSEDE project. NCGAS is currently a Level 3 partner in the XSEDE project. More details on these options will be developed after the completion of the Ecological Society of America's Sustaining Biological Infrastructure workshop, currently being attended by PI Dr. Sheri Sanders.

## 8. Intellectual Merit

NCGAS's merit lies in both the research we support and in the methods behind that support. NCGAS extends the impact of federal dollars put towards scientific research; more than half our clients are federally funded, and they are highly satisfied with our research support services. We manage software and provide access to tools that enable hundreds of researchers to use NSF-funded advanced cyberinfrastructure. We provide training and support, increasing their ability to do their work efficiently. We contribute actively to projects when expertise is needed, and have co-authored dozens of peer-reviewed publications while supporting hundreds of projects. That experience returns to the community in our national outreach, through which we have trained over 6500 researchers at no-charge workshops and seminars. Our focus on scalable distribution of computational tools and training accelerates the productivity of the researchers that NCGAS serves.

## 9. Broader Impacts

The most significant broader impacts of NCGAS and its services are the societal benefits that derive from research we support. A nuanced understanding of the genomic underpinnings of physiological response to environmental changes, chemical exposure, and treatment protocols is increasingly critical in the age of environmental change and the rise of personalized medicine. Supporting research to understand physiological responses through genomic and transcriptomic methods will deepen our understanding of how organisms respond to, and, in turn, increase the efficacy of, climate remediation, biodiversity conservation, ecotoxicology assessments, and tailored medical intervention.

An important aspect of NCGAS's services is our focus on increasing the diversity of the community using genomic techniques by increasing the computational resources, skills, and support for individuals from TUGs in STEM disciplines. We seek to include the whole community and to serve the growing demand for computationally competent biologists and genomic/bioinformatic analysts. Thus, we will tailor services to institutions in EPSCoR states and MSIs by making software more accessible and usable on smaller systems, offering access to computational resources through XSEDE community allocations and IU-provided hardware, and providing free consulting to the community. We must also continue training and consulting work that leads to full participation of women and under-represented groups. This entails offering short courses and online content to accommodate participants with limited release time for professional development; it also means considering the diversity of our participants when offering slots in over-subscribed events. We plan to offer training materials with closed captions and Spanish translations to make development opportunities more accessible.

We currently support genomics researchers in all 50 states and Puerto Rico, from high school students to faculty researchers. The following activities move individuals into the US STEM workforce:

- **Graduate Students.** Graduate students remain one of the largest subgroups of our training and consulting activities. Over 70% of these attendees for national workshops are women or members of other TUGs.
- **Undergraduate students.** NCGAS offers in-depth mentoring to undergraduate students from the national research community (via a REU project related to the Jetstream grant award) and to local IU students (through the IU Center of Excellence for Women & Technology and other

Research Experience for Undergraduates programs). Roughly 2% of our workshop participants are undergraduate students learning cutting edge analyses that contribute to their research.

- **K-12.** We mentor 1-2 underrepresented minority high school students in the Jim Holland Summer Science Research Program (SSRP) at IU [238]. NCGAS also contributes to the Jim Holland Research Intensive in STEM Education (RISE), which takes 10 SSRP students (high school seniors) for a two-week introduction to STEM disciplines. Combined, these programs create an effective pipeline from high school to graduate school, which increases minority participation in STEM.

NCGAS does not just advocate for a diverse STEM workforce; NCGAS is such a workforce in action. NCGAS is led by a woman, and the group's staff has been majority women for the last four years.

## 10. Results of Prior NSF Support

**Sheri Sanders (PI), Matthew Hahn (Co-PI), Yuzhen Ye (Co-PI).** NSF DBI 1759906, \$962,613.

9/01/2018 – 08/31/2021. “Collaborative Research: ABI Sustaining: The National Center for Genome Analysis Support”. This award is the second DBI Sustaining award (1st sustaining award: NSF DBI-1458641, \$621,218, 09/01/2015 - 08/31/2019); Development award: NSF DBI-1062432, \$1,487,162, 09/15/2011 - 08/31/2015) to continue NCGAS services to NSF-funded researchers needing assistance with genomic analysis, including access to HPC hardware, curated applications tools, training (workshops, blogs, etc.) and guidance/advice at any step in their study. Intellectual Merit: NCGAS's merit rests in the discoveries made by researchers who utilize NCGAS, the identification and support of best-of-breed genomics analysis software, our expertise in providing consulting support and training, and the high intellectual quality of our training materials and events. There is also intellectual merit in the ongoing evolution of the various services we support in service of the most important genomic research in the US today. Broader Impacts: NCGAS's most important broader impacts are the societal benefits that derive from the research we support. Our work informs water-use decisions impacting several at-risk species, and improves knowledge of ecologically and economically important species and communities. We aid the creation of a well-educated 21st century workforce by offering training, education, and outreach events designed to interest students in genomics and bioinformatics. Publications and Evidence of Research Products and their availability: [7, 8, 13, 14, 18, 20–30, 33, 34, 36–38, 53, 55, 57–63, 97, 130, 131, 144, 148, 153, 159–161, 163, 167, 174, 175, 239–316] [16, 17, 39–52, 54, 64–94, 96–130, 132–152, 154–158, 160–162, 164–173, 317–319] . As described in the Justification section, IU has contributed compute and storage facility resources and effort, including support for evaluation services and for education/training events.

**Irene Newton (Co-PI).** NSF IOS 1456545 8/1/2015-7/31/2018 “Collaborative Research: How does an intracellular symbiont manipulate host cell biology?” Intellectual merit: This project performed comparative genomics and a high-throughput screen to identify candidate-secreted substrates in the Wolbachia-Drosophila symbiosis. We have completed all of the aims, screening over 80 Wolbachia candidate effectors in yeast under 5 stress conditions and performing a test of vesicle trafficking defects in yeast. We discovered that Wolbachia effectors inducing the most significant growth defects are co-regulated and correlated in their evolution. Most interestingly, these effectors are upregulated during host pupation and L2 stages, suggesting they either function in concert or are required during the same host biological and developmental process. These promising candidates (n = 14) were moved into a Drosophila expression vector and are being expressed in insect cell lines by the Sullivan laboratory (UCSC). The Newton laboratory has moved to express one additional candidate (WD0811) in transgenic flies. Finally, the laboratory is following up results from the characterization of WalE1 (the actin bundling Wolbachia effector) to identify host targets. Broader impacts: The project has resulted in seven high-impact publications to date [44, 79, 317, 319] [141] [318] [45]. Data availability: All data have been uploaded to public repositories and the effector library generated made available through the Newton lab and advertised via ASM's MRA journal.