ABI Sustaining: The National Center for Genome Analysis Support 2016 Annual Report

Thomas G Doak
Craig A. Stewart
Scott D Michaels

Indiana University

PTI Technical Report PTI-TR16-004

August 15, 2016

Citation:

Doak, T.G., Stewart, C.A. & Michaels, S.D. (2016). ABI sustaining: National center for genome analysis support 2016 annual report (PTI Technical Report PTI-TR16-004). Bloomington, IN: Indiana University. Retrieved from http://hdl.handle.net/2022/20957

PERVASIVE TECHNOLOGY INSTITUTE
INDIANA UNIVERSITY

RESEARCH TECHNOLOGIES
INDIANA UNIVERSITY
University Information Technology Services
Pervasive Technology Institute

Table of Contents

1. Accomplishments

1.1. What are the major goals of the project?

The major goals of the NSF ABI Sustaining Award is to support National Center for Genome Analysis' (NCGAS) continuing and expanding activities during this award's duration, including:

1) Provide excellent bioinformatics consulting services.
2) Maintain, support, and deliver genome assembly and analysis software on national CI systems.
3) Disseminate tools for genome assembly and analysis.
4) Provide long-term archival storage for genome biologists.
5) Provide education and outreach programs on genome analysis and assembly, including designing genomics experiments, using best-of-breed tools, and interpreting data

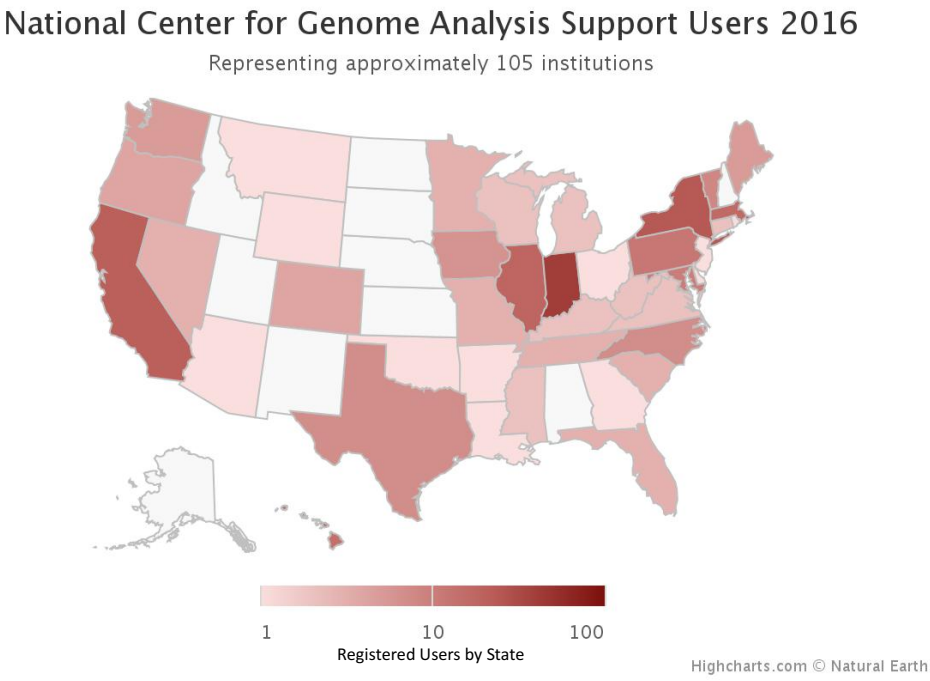
The focus of NCGAS' activities is on "whole-genome" and metagenomics research. Emphasis is placed on genome and transcriptome assembly at the technically challenging end of the spectrum of current bioinformatics—for example *de novo* genome and transcriptome assembly—where both specialized computational resources and applications are needed.

1.2. What was accomplished under these goals?

1.2.1. Major Activities

NCGAS is a collaborative project between the lead institution, Indiana University (IU), and the Pittsburgh Supercomputing Center (PSC) at Carnegie Mellon University. At IU, NCGAS is part of the Indiana University Pervasive Technology Institute and has significant HPC facilities, human resources, and administrative support from IUPTI and IU. Likewise, NCGAS funded collaborator PSC maintains extensive HPC resources and supporting services. During the first year of this sustaining award, NCGAS has continued to make significant strides in using NSF funding—with additional funding and facilities from IU and PSC—to aid discovery and innovation in the biological sciences in the US. Under the direction of IU, NCGAS has developed new opportunities through collaborative efforts between IU and PSC and has continued to aid in discoveries that range from a better understanding of basic biological processes, to discoveries that will aid management of economically and ecologically important animals and plants.

How national is the National Center for Genome Analysis Support? We now server researchers in 40 states (see map below), including 12 EPSCoR states.



1.2.1.1. Provide excellent bioinformatics consulting services.

NCGAS' most significant accomplishments in support of biological and bioinformatics research continue to be in aiding RNA and DNA transcriptome, metatranscriptome, genome, and metagenome assemblies:

In year 1, NCGAS aided researchers in completing many *de novo* assemblies, including the following organisms (completed and on-going):

- Coffee transcriptomes
- Daphnia genomes (population genomics and *de novo* assemblies)
- North Atlantic and Pacific Neocalanus copepods
- Filarial Nematode and Wolbachia endosymbiont
- Paramecium species (*de novo* genomes, transcriptomes, populations)
- Rotifer Species transcriptomes
- Barred tiger salamander transcriptomes
- Diatoms transcriptomes
- The diverse microbial clade Stramenopila + Alveolata + Rhizaria (SAR) (*de novo* genomes, transcriptomes)
- Heliconius butterflies transcriptomes
- Stalk-eyed Flies, transcriptomes
- Carrion Flies (forensic arthropods), *de novo* genome assembly/resequencing, transcriptomes
- See new projects added this year as well (attached doc)

In addition, NCGAS supported 377 biologists doing research in the general area of genome analysis (22 named allocations) during the current year's funding. Assistance has been provided through 281 short consultations and 37 extended consultations. Details regarding many of the extended consultations are provided in an attachment.

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

NCGAS continues to assist NSF researchers in genomics research. From the beginning we have accomplished this by forming a "supply line" from HPC hardware, to specialized applications and knowledge, to the researchers' specific data and questions. While some researchers only need access to large memory clusters, others need instruction in basic HPC use and genomic analysis. We have been successful in this and continue to attract new users, often by word-of-mouth. One of our on-going tasks must be to stay current: new hardware becomes available (see below), state-of-the-art applications change, new data types become available (we are starting to see a significant number of PacBio data sets), and researchers change. For example, this year we have installed or updated six packages(spades, hisat, sailfish, salmon, celera, gmap) for assembly and analysis of PacBio long-read sequencing data. Funded collaborator PSC also makes many of these packages available on PSC systems, and is also focusing on providing the best software tools for enabling high-quality metagenome assembly and analysis (see PSC report).

NCGAS at IU provides accounts to multiple clusters for direct command-line access:

- The large memory Mason IU cluster
- The newer Karst IU cluster (which will replace Mason within a year)
- The Jetstream cloud environment
- Additional XSEDE resources, including resources on PSC's *Bridges*, through an NCGAS XSEDE Community Allocation

NCGAS at IU also provides access to bioinformatics software through online web (graphic) portals:

- NCGAS Galaxy web portal: providing access to the widely used Galaxy workflow system on Mason and other XSEDE-supported resources
- Trinity RNA-Seq Galaxy portal: running on IU's Karst cluster
- GenePattern Analysis Package, running on IU's Karst

Overall, we have installed or updated 72 software packages across the systems described above (see attachment describing significant software activities).

1.2.1.3. Disseminate tools for genome assembly and analysis

NCGAS has supported the creation of the "XSEDE National Integration Toolkit" (XNIT). XNIT is a suite of software available for download and installation on computational clusters. NCGAS has added whole suits of bioinformatics software supported by NCGAS on XSEDE in the past, but this activity has lapsed in the last year due to personnel shortage and more pressing priorities. We hope to return to this important activity in the second year.

Provide long-term archival storage for genome biologists

NCGAS and IU continue to provide access for all NCGAS users to ScholarWorks, a digital repository provided by the IU Libraries for showcasing and preserving research findings, and the Scholarly Data Archive (SDA), which provides extensive capacity (approximately 42 PB of tape) for storing and accessing research data.

Provide education and outreach programs on genome analysis and assembly, including designing genomics experiments, using best-of-breed tools, and interpreting data.
While the bulk of NCGAS personnel's time is devoted to one-on-one consultation and assistance, we also do out-reach and trainings (see 1.3 and 5.2). For example, Sheri Sanders has just finished teaching at the MDI Biological Laboratory's 2016 Environmental Genomics course and we are starting to plan for presentations at the 2017 PAG meeting.
The most important training and professional development activities have been presentations and tutorials provided at national and international conferences. A summary of these tutorials is provided below:

- 290 participants in tutorials (~500 contacts at events). In the last year, some of the events attended were:
    - Galaxy Community Conference 2016/Bloomington IN (we hosted);
    - GMOD User Community 2016 / Bloomington IN (we hosted);
    - Extreme Science and Engineering Discovery Environment (XSEDE) 2016 / Miami FL
    - MDI Biological Laboratory's 2016 Environmental Genomics course / Bangor ME

1.2.2. Specific Objectives

The software supported by NCGAS as of the end of the first four years of NSF funding includes 72 packages described in detail in the attached file on significant software activities.

1.2.3. Significant results

Provide online help, consulting, and tutorials related to genome analysis.

Key highlights of NCGAS support include:

- Consulting. NCGAS in the current year completed a total of 281 short term consulting engagements (those taking less than 4 hours of staff time to resolve) and 37 long term consulting engagements (taking more than 4 hours of staff time to resolve). Many the long term consultations were research collaborations that last months or even years. In such collaborations NCGAS staff became partners and play a critical role in scientific discoveries accomplished by scientists receiving NCGAS help.
- NCGAS completed tutorials and training and outreach activities attended by hundreds of attendees (see 1.3).
- Exploring Jetstream applications to bioinformatics. As part of the Jetstream acceptance tests NCGAS helped researchers at the University of Arkansas Fayetteville to establish, provision, and use Jetstream VMs, to complete analysis of both northwest endangered river fish species, and the distribution of rattlesnake species, using ddRAD. This was done in partnership with Jeff Pummell at the Strategic Initiatives and User Services, Arkansas High Performance Computing Center. We expect there will be future opportunities to aid researchers, using Jetstream. In another capacity, NCGAS is working

with groups funded by Information Technologies in Cancer Research (ITCR) to use Jetstream in their work.

Consultant services are provided by telephone, email (a ticketing system tracks requests), and in-person consultations. Consulting hours are typically 8 am to 5 pm weekdays, but support activities often extend beyond local business hours when there is time pressure on a researcher. In the last year NCGAS engaged in a total of 22 new long-term projects, described in the attached file on consulting and significant results.

Another significant result is the strengthening of our partnership with PSC (see 1.5). During the initial ABI grant, PSC was an unfunded partner, and we regularly exchanged information on Galaxy implementation, software installation, and the GenePattern suite. Now that Philip Blood at PSC is a funded member of the NCGAS team, we are able to escalate this relationship. Beyond general coordination of software suites (and computational resources), we are beginning to build a center for metagenomic analysis centered at PSC. PI Doak attended a NEON metagenomics advisory meeting two years ago, and hopes to offer that project bioinformatic support—partnership with PSC will allow us to pursue this.

1.2.4. Key outcomes or other achievements

The key outcome during the first year of this sustaining award is the continued success of NCGAS in delivering an effective consulting service focused on accelerating the research of biologists and bioinformaticians, and in so doing accelerated biological discoveries in the US. NCGAS provides a robust "supply chain" from NSF-funded and other supercomputers, through specialist applications and knowledge, to bench and field scientists across the country. NCGAS' ongoing efforts have helped enable 16 peer-reviewed scientific publications that have been published in 2015-2016 (beyond those reported in the final report for the original NCGAS award).

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

Dr. Thomas Doak was, during the first Development award, a postdoctoral fellow. Dr. Doak was promoted to the rank of Assistant Scientist at Indiana University, and is now PI and manager of this NSF development, award to continue and grow NCGAS services.

Staff member Carrie Ganote is continuing her PhD program in bioinformatics, while in the employ of NCGAS. We have promoted her as a player in the international Galaxy community and she was on the organizing committee for the 2016 Galaxy conference, hosted at IU. Ms. Ganote oversees many projects, and is mentor for new member Sheri Sanders.

Staff member Sheri Sanders began as a NCGAS team member only a few months ago, immediately after finishing her PhD at Purdue, where she used transcriptomics to characterize salamander species, some endangered. Her goal in joining NCGAS was to grow her understanding of IT and HPC, and how they impacted the biological community. She is now taking advantage of her placement in the IU HPC community to grow professionally.

Staff member Le-Shin left IU at the end of 2015, to take an attractive job at Eli Lilly and Company Pharmaceuticals.

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

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

- Publications in scientific journals
- Presentations
- Birds of a feather sessions at technical conferences
- Displays and booths at national and international technical conferences
- Articles in the lay press, most notably in Science Node, https://sciencenode.org
- NCGAS web site at ncgas.org
- In-person contacts
- Email list distribution
- Newsletters

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

1.5.1 Goals for the next year

We have initiated several new directions to carry into the next year: 1) expand our offering of genome browsers, for searchers to organize and distribute their results: 2) aggressively pursue metagenomic projects/researchers; with this grant Phil Blood of PSC was added to our team, and he has considerable experience in metagenomics researcher; 3) explore how to use the new Jetstream environment at IU to aid genomics researchers.

NCGAS infrastructure was inaugurated 5 yrs. ago with the IU purchase of the large-memory cluster Mason, each node having half a terabyte of RAM (Random Access Memory), specifically to support DNA genome assembly software. In addition to serving NCGAS allocations, Mason is an XSEDE-allocated resource, again primarily for genomics research. Mason is now antiquated, and we are in the process of replacing its capacity, within the next 6 months. This will be accomplished in three ways: 1) The IU cluster Karst will be expanded with large memory nodes, for the use of Mason users (NCGAS and XSEDE). While details are not finalized, the new nodes will have considerably faster processors and a higher memory-to-core ratio than Mason. 2) NCGAS will take advantage of the new NSF-funded PSC cluster Bridges (see PSC annual report). Bridges is already in use for metagenomic assemblies through PSC and NCGAS has an XSEDE allocation to enable our users to utilize Bridges at PSC along with other XSEDE

resources. 3) The NSF-funded cloud environment Jetstream—a partnership between IU, TACC, and others—will going into production phase Sept. 1 of this year. While not providing large memory, NCGAS has already had success helping researchers accomplish genomics science (*ex*. ecological-genomics projects from Ark.) on Jetstream, and we will continue to explore its uses.

- We are currently in the process of hiring another NCGAS team member. Our current focus candidate has a background in metagenomics, which is an area we hope to grow in (see below).
- Finalize transition from Mason to Karst nodes (see 1.2.1.2).
- We have just started our involvement in docker container technology, and IU and NCGAS plan to continue this effort to further our dedication to distributing current, stable genomic software builds to XSEDE partners and individual researchers. Researchers are increasingly using commercial cloud resources, and we can provide these users with professional applications, via containers, as well.
- The IU/TACC Jetstream cloud environment opens up a range of possibilities, and while we have gained some experience—and helped biologists successfully use it in their research—Jetstream is not even in production yet, and there is a lot to learn. The use of Jetstream goes hand-in-hand with the use of containers.
- Researchers have requested genome browsers for their assemblies since the inception of NCGAS, but we've been limited in our ability to provide them. We are now actively working to provide browsers to our users, starting with investigating both the GMOD and UCSC models, and using local Daphnia assembles as a test case.
- Having PSC and Phil Blood as a funded partner allows us to expand our services: we will further develop our support for metagenomics/metatranscriptomics. Even though PI Doak is published in metagenomics, and NCGAS has provided applications, it has never been an emphasis—while PSC has had an emphasis in metagenomics and Blood has considerable experience working with researchers and developers. PSC will emphasize metagenomics, and we will work to grow that community. The first step is ongoing: assembling a comprehensive tool set and proving this to researchers. We will then move to a metagenomics Galaxy instance We are particularly interested to see if we can serve a metagenomic component of the NEON project.

1.5.2 *Synergistic activities*

NCGAS is both a specific NSF-funded service provider in genomics, and a management group in IU's Pervasive Technology Institute. In this second guise, the NCGAS takes part in other projects that we feel augment our NSF-funded services.

- Active engagement with the Galaxy development community. NCGAS and IU hosted this year's Galaxy conference (~300 participants) and NCGAS member Carrie Ganote was on the organizing committee and a presenter.
- Active engagement with the Generic Model Organism Database (GMOD) community. This is a work in progress, but as we invest effort in genome browsers we hope to play a role in GMOD activities. The GMOD user group had a satellite meeting after the Galaxy meeting in Bloomington.
- NIH/NSF/ITCR-funded Trinity development and Galaxy hosting. The involvement of the NCGAS and the IU Scientific Applications and Performance Tuning group has both

improved Trinity and made it far more availible. While aimed at cancer research, Trinity is extensively used by our non-medical clients, esp. where obtaining a genome assembly is not feasible (e.g. marine copepods and polyploid salamanders). Thus, Trinity *de novo* assemblies are most useful for the least "model" of our users' organisms.

- NIH/NSF/ITCR-funded GenePattern hosting. Similar to Trinity, hosting GenePattern gives us an understanding of alternative software, and makes them availible to our users.
- Common Workflow Language (CWL). NCGAS and IU Scientific Applications and Performance Tuning was included in a awarded R44 grant to develop CWL. Our role will be to explore the use of CWL in Galaxy instances—again a way to improve software functionality for our users.

2. Products

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

2.1.1. (Peer-reviewed) Journal Articles

Kucukyildirim, S., Long, H., Sung, W., Miller, S. F., Doak, T. G., & Lynch, M. (2016). The Rate and Spectrum of Spontaneous Mutations in Mycobacterium smegmatis, a Bacterium Naturally Devoid of the Postreplicative Mismatch Repair Pathway. *G3 (Bethesda, Md.)*, *6*(7), 2157–2163. Retrieved from http://view.ncbi.nlm.nih.gov/pubmed/27194804

Lee, H., Doak, T. G., Popodi, E., Foster, P. L., & Tang, H. (2016). Insertion sequence-caused large-scale rearrangements in the genome of Escherichia coli. *Nucleic Acids Research*. Retrieved from http://view.ncbi.nlm.nih.gov/pubmed/27431326

Nimkulrat, S., Lee, H., Doak, T. G., & Ye, Y. (2016). Genomic and Metagenomic Analysis of Diversity-Generating Retroelements Associated with Treponema denticola. *Frontiers in Microbiology*, *7*. Retrieved from http://view.ncbi.nlm.nih.gov/pubmed/27375574

Wang, M., Doak, T. G., & Ye, Y. (2015). Subtractive assembly for comparative metagenomics, and its application to type 2 diabetes metagenomes. *Genome Biology*, *16*. Retrieved from http://view.ncbi.nlm.nih.gov/pubmed/26527161

2.1.2. Conference Papers and Presentations

Neeman, H., Bergstrom, A., Brunson, D., Ganote, C., Gray, Z., Guilfoos, B., … Voss, D. (2016). The Advanced Cyberinfrastructure Research and Education Facilitators Virtual Residency: Toward a National Cyberinfrastructure Workforce. *Proceedings of the Conference of the eXtreme Science and Engineering Discovery Environment 2016*.

2.1.3. Other Products

Title: NCGAS Home Page
URL: http://ncgas.org/
Description: The mission of the National Center for Genome Analysis Support is to enable the biological research community of the US to analyze, understand, and make use of the vast

amount of genomic information now available. NCGAS focuses particularly on transcriptome- and genome-level assembly, phylogenetics, metagenomics/transcriptomics and community genomics.

Title: Trinity Galaxy
URL: https://galaxy.ncgas-trinity.indiana.edu/root
Description: Galaxy Public, hosted by Indiana University and the Broad Institute, is a free-to-use public interface for Trinity users.

Title: GenePattern
URL: http://gp.indiana.edu/gp/pages/login.jsf
Description: GenePattern is a freely available computational biology open-source software package developed at the Broad Institute of MIT and Harvard, for the analysis of genomic data. NCGAS now hosts a free GenePattern server, with increased computational resources.

3. Participants

3.1. Individuals

Table 1. Individuals that have worked on the project

Name	Most Senior Project Role	Nearest Person Month Worked
Doak, Thomas	PhD/PI	6
Stewart, Craig	PhD/co-PI	1
Michaels, Scott	PhD/co-PI	1
Henschel, Robert	Other Professional	1
Blood, Phillip	Other Professional	3
Miller, Therese	Other Professional	1
Wu, Le-Shin	Other Professional	4
Ganote, Carrie	Staff Scientist (doctoral level)	6
Sanders, Sheri	Staff Scientist (doctoral level)	3*

- Sanders has only been employed for 3 months

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

Thomas Doak
Email: tdoak@iu.edu
Most Senior Project Role: PI (doctoral level)
Nearest Person Month Worked: 6
Contribution to the Project: PI and operational management
Funding Support: NSF, NIH, IU
International Collaboration: Yes: Italy, Germany, Japan
International Travel: Yes, Italy - 0 years, 0 months, 7 days

Craig A Stewart
Email: stewart@iu.edu
Most Senior Project Role: PhD/co-PI
Nearest Person Month Worked: 1
Contribution to the Project: Co-PI responsible for oversight and outreach to new groups to generate users/projects for NCGAS services
Funding Support: Indiana University
International Collaboration: No
International Travel: Yes, Germany - 0 years, 0 months, 7 days

Scott Michaels
Email: michaels@indiana.edu
Most Senior Project Role: PhD/co-PI
Nearest Person Month Worked: 0
Contribution to the Project: Funded PSC collaborator
Funding Support: Co-PI responsible for oversight
International Collaboration: No
International Travel: No

Phillip Blood
Email: blood@psc.edu
Most Senior Project Role: Other Professional
Nearest Person Month Worked: 4
Contribution to the Project: Carnegie Mellon University and University of Pittsburgh
Funding Support: Indiana University
International Collaboration: No
International Travel: Yes

Robert Henschel
Email: henschel@iu.edu
Most Senior Project Role: Other Professional
Nearest Person Month Worked: 1
Contribution to the Project: Director over NCGAS management group and software optimization
Funding Support: Indiana University
International Collaboration: No
International Travel: Yes

Therese Miller
Email: millertm@iu.edu
Most Senior Project Role: Other Professional
Nearest Person Month Worked: 1
Contribution to the Project: Financial and reporting management
Funding Support: IU
International Collaboration: No
International Travel: No

Le-Shin Wu
Email: lewu@iu.edu
Most Senior Project Role: Other Professional
Nearest Person Month Worked: 3
Contribution to the Project: bioinformatician consultant / programmer
Funding Support: NSF, IU
International Collaboration: No
International Travel: No

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

Sheri Sanders
Email: ss93@iu.edu
Most Senior Project Role: Staff Scientist (doctoral level)
Nearest Person Month Worked: 3
Contribution to the Project: bioinformatician consultant / programmer
Funding Support: NSF
International Collaboration: No
International Travel: No

3.2. Partner organizations

Table 2. Partner organizations

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

3.2.1. Full details of partner organizations

3.2.1.1. Pittsburgh Supercomputing Center, Carnegie Mellon University

Partner's Contribution to the Project

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

More Detail on Partner and Contribution: PSC is a funded collaborator on the NCGAS sustaining award. Philip Blood, the 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 the Greenfield and Bridges supercomputers in support of NGCAS activities and in support of biological researchers who have used NCGAS services. Staff of this institution have also engaged use of NCGAS staff and facilities, and have made available resources at their site to NCGAS staff. Some of the support provided by this institution has been provided in-kind, and this institution has engaged in collaborative research on genome analysis software, particularly as regards use of Galaxy and software that requires the large shared memory architecture of PSC supercomputers. PSC also participates in education, outreach, and dissemination efforts of NCGAS.

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

Partner's Contribution to the Project

- Collaborative Research

More Detail on Partner and Contribution: Jeff Pummel, Director of the Arkansas High Performance Computing Center, has worked with us to get UofA researchers using genomic tools on Jetstream. Pummel is also an XSEDE Campus Champion and has aided in XSEDE Jetstream allocations.

3.2.1.3. Texas Advanced Computing Center, University of Texas

Partner's Contribution to the Project

- Collaborative research
- Facilities

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

software, as well as participating in education, outreach, and dissemination efforts of NCGAS.

3.2.1.4. XSEDE

Partner's Contribution to the Project

- Collaborative research
- Facilities
-

More Detail on Partner and Contribution: Staff of the NSF-funded XSEDE project have engaged use of NCGAS staff and facilities, and have made available resources at their site to NCGAS staff. Some of the support provided by XSEDE has been provided in-kind, and this institution has engaged in collaborative research on genome analysis software. XSEDE has particularly played a strong role in education, outreach, and dissemination efforts of NCGAS. NCGAS is now a Level 3 XSEDE Service Providers.

3.2.1.5. Open Science Grid

Partner's Contribution to the Project

- Collaborative Research

More Detail on Partner and Contribution: This institution has engaged in collaborative research on genome analysis software in the OSG environment.

3.3. Have other collaborators or contacts been involved?

No

4. Impact

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

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

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

4.2. What is the impact on other disciplines?

The primary other discipline on which NCGAS has had an impact is computational science and cyberinfrastructure. The largest impact that NCGAS has had in computational science has been to establish a model of a "scientific service center" focused on a particular subdiscipline, independent of federally-funded cyberinfrastructure computational resources. That is, we have decoupled federal funding for supercomputers and federal funding for supercomputer application support. This is tremendously important, as it ensures that a community that is relatively newe to use of supercomputers—biology for example—has a support funded by the BIO directorate of NSF and attuned to the needs of the current most important research.

We have also established new models for distribution of software relevant to biological research, which improves the nation's ability to use its aggregate cyberinfrastructure resources.

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

See sec. 1.3

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

Nothing to report.

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

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

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

NCGAS has facilitated the publication of several data sets important to basic biological research and to management of important plant and animal stocks. In the future, NCGAS will place a greater emphasis on genome browsers, and important entry into genomics results.

4.7. What is the impact on technology transfer?

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

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

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

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

With the departure of an original and important NCGAS team member, Le-Shin WU, we were short-handed for ~6 months. While we could maintain our level of service to existing clients, and take on new researchers who found us, we were limited in our ability to recruit and do outreach. New member Sheri Sanders will both start to fill out the team (we are hiring another member as well), and brings extensive presentation experience to the NCGAS team.

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.

6. Appendix 1. List of NSF-Funded Projects Using NCGAS Resources

PI: Frank Anderson
State: IL
Funding Organization: NSF
Award #: DEB-1036516
Title: Wormnet II: Assembling the annelid tree of life
Date initiated: 2011-1-1
Date completed: 202016-12-31
One of my lab's roles in the Wormnet II project is production and analysis of transcriptomes from ~80 clitellate annelids. We have generated these transcriptomes and are now at the assembly and phylogenetic analysis stage for various projects. I have a Linux workstation in my lab, but it is not up to the task of assembling large, paired-end data sets.

PI: Seth Bordenstein
State: TN
Funding Organization: NSF
Award #: 1456778
Title: THE GENETIC ARCHITECTURE OF MATERNAL SUPPRESSION OF SYMBIONTS
Date initiated: 2015-5-1
Date completed: 2019-4-30
The majority of animal species harbor maternally-transmitted bacteria, yet little is known about the genetic and molecular mechanisms that the animal and bacteria use to achieve maternal transmission. The proposed research begins the first forward-genetic investigation of host animal genes that regulate densities and composition of bacterial symbionts. The goal of this project (and need for NCGAS rescources) is to utilize an animal model system coupled with multi-omic technologies for host and bacteria to identify the numbers and types of host genes that regulate symbionts, their additive and epistatic interactions, and their effects on symbiont localization.

PI: Sara Cahan
State: VM
Funding Organization: NSF
Award #: BCS-1216193
Title: Modeling disease transmission using spatial mapping of vector-parasite genetics and vector feeding patterns
Date initiated: 2012-6-15
Date completed: 2017-5-31
Note: my colleagues and I already have an NCGAS allocation for NSF grant DEB-1136644. We just received notice that this allocation is ending, but the grant was given a one-year no-cost extension through December 2016. We also have a second funded grant for which we would like to request an extension of our NCGAS allocation, summary below. I can provide a list of the colleagues associated with my current allocation if it is necessary. We are taking a multidisciplinary approach to the development of spatially explicit models of vector-borne disease transmission, using Chagas disease as our model. We are (1) collecting ecological, genomic (vector and parasite), landscape, and socioeconomic data/information to: (2) develop

and parameterize models of disease transmission, (3) identify those factors most influencing transmission at local and regional spatial scales, and (4) provide next generation genomic and spatial analysis tools applicable to the study of other vector-borne diseases worldwide. We require NCGAS computing resources to analyze our next-generation sequencing and genotyping data.

PI: Christopher Chandler
State: NY
Funding Organization: NSF
Award #: 1453298
Title: CAREER: Evolution of allosomes and dosage compensation in terrestrial isopods
Date initiated: 2015-5-1
Date completed: 2020-4-30
Allosomes are chromosomes that determine sex, like the X and Y chromosomes in mammals, and they play crucial roles in the evolutionary biology of species. They have evolved independently in a wide array of species, and while these chromosomes in different groups share many similarities, they also exhibit important differences. Explaining these differences, however, remains difficult. This project will address this problem by examining allosomes in terrestrial isopod crustaceans, an ideal study system because they exhibit considerable variation in sex-determining mechanisms. However, surprisingly, their genomes have received little attention. The proposed experiments are designed to help explain why these chromosomes are so unique, and how they contribute to vital biological processes. Specifically, this research will (i) identify where changes in sex-determining chromosomes have occurred on the isopod evolutionary tree; (ii) test whether genes on these chromosomes are affected more strongly than autosomal genes by natural selection and genetic drift; and (iii) test whether these species exhibit dosage compensation. This work will also generate a large volume of genome sequencing data, providing some of the first draft genome assemblies for these under-studied organisms, requiring the use of powerful computing resources. These assembled genome sequences will create bioinformatics training opportunities for undergraduate students through the development of a new genomics course to be taught at SUNY Oswego. By studying a unique taxonomic group, this research will also help examine the generality of patterns suggested by earlier work on allosomes, providing significant insights into these influential components of so many organisms' genomes.

PI: Robert Cooper
State: CA
Funding Organization: NSF
Award #: 1257648
Title: Tracking genes in real time as they traverse a hybrid invasion landscape.
Date initiated: 2016-4-1
Date completed: 2018-7-1
Hybrid zones represent valuable opportunities to observe evolution in systems that are unusually dynamic and where the potential for the origin of novelty and rapi adaptation co-occur with the potential for dysfunction. Recently initiated hybrid zones are particularly exciting evolutionary experiments because ongoing natural selection on novel genetic combinations can be studied in ecological time. Moreover, when hybrid zones involve native and introduced species, complex

genetic patterns present important challenges for conservation policy. This project uses cutting-edge genomic tools to characterize historical and contemporary salamander populations, and to track the spread of thousands of non-native genes as they invaded the native range of the California tiger salamander over the last quarter century. Specifically I seek to identify genes that are differentially expressed in native and invasive salamanders in response to temperature stress. Analysis will consist of removing low quality reads and removing adaptor and barcode reads using FlexBar. Sequences will then be aligned using Bowtie2 and Tophat to a reference transcriptome of Ambystoma mexicanum (Sal-Site, www.ambystoma.org). Then transcripts will be counted using the program Cufflinks. I will then use WGCNA (Weighted Gene Co-Expression Network Analysis) to determine gene families that correlate with CTmax values. I will also use EdgeR to determine genes that are differentially expressed between salamanders under thermal stress. All of these analyses will require considerable computer resources that we are unable to provide in our current lab.

PI: Liliana M Davalos
State: NY
Funding Organization: NSF
Award #: 1442142
Title: Dimensions: Collaborative Research: Discovering genomic and developmental mechanisms that underlie sensory innovations critical to adaptive diversification
Date initiated: 2016-9-15
Date completed: 2019-9-15
This project will uncover the evolution of genes and structures of the auditory, visual, and olfactory and vomeronasal systems in a large superfamily of bats characterized by diverse sensory adaptations associated with specific diets. Analyses of gene evolution will be used to test the hypothesis that sensory innovations arise through gene duplication and positive selection. Measurements of gene expression from tissues collected in the field, and experiments to express key bat genes in developing embryos will be used to elucidate how genes shape adaptive sensory structures. Comparative analyses of detailed measurements of the size of sensory structures will evaluate trade-offs between sensory systems and the way these may limit diversity of diets or species. State-of-the art methods to quantify relationships between gene and trait evolution and species diversity will be used to discover the impact of sensory adaptation on species diversity through time. This research will illuminate the main biological forces in the genome, during embryonic development, and in anatomical structures that contribute to the success of species in adapting to their ever-challenging environment. We have RNA seq data for 5 tissue types for >40 species, making great computational demands for analyses.

PI: Charles Delwiche
State: MD
Funding Organization: NSF
Award #: 1541510
Title: GoLife: Collaborative Research: Bringing the diverse microbial clade Stramenopia + Alveolata + Rhizaria (SAR) into a modern genomic context.
Date initiated: 2016-1-1
Date completed: 2021-1-1

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

PI: Frank Jones
State: OR
Funding Organization: NSF
Award #: 1257976
Title: Collaborative Research: Intraspecific variation in drought responses of tropical tree seedlings - consequences for species distributions under climate change
Date initiated: 2013-5-1
Date completed: 2017-4-30
Tropical forests harbor the majority of the Earth's terrestrial biological diversity and provide humans with valuable products and ecosystem services. Rainfall is generally high in tropical forests, but at the same time, most tropical forests experience one or two dry seasons per year. As a result of climate change, pronounced shifts in dry season length and intensity are predicted for tropical forests. During dry periods tropical trees and their seedlings generally have lower growth and survival, but species vary in how strongly they are impacted by drought. From temperate forests, we know that drought responses vary not only among tree species, but also within species, with populations growing at different sites showing different responses. Such differences within species may be due to genetic factors after long-term evolutionary adaptation to different sites, or it can be due to short-term plasticity in responses to environmental factors. How strongly populations within tree species vary in their drought responses, and to what extent that variation is influenced by genetic or environmental factors, will play a large role in determining how species will respond to climate change. However, to date we know virtually nothing about differences in drought responses within and among populations of tropical tree species

PI: Neil Kelleher

State: IL
Funding Organization: NSF
Award #: DMS-0800631
Title: Statistical Approaches to Integration of Mass Spectral and Genomic Data of Yeast Histone Modifications
Date initiated: 2012-12-1
Date completed: 2013-11-30
New statistical and analytical methods will be developed to study regulatory role of histone modifications in Saccharomyces cerevisiae. Gene activities in eukaryotic cells are concertedly regulated by transcription factors and chromatin structure. The basic repeating unit of chromatin is the nucleosome, an octamer containing two copis each of four core histone proteins. While nucleosome occupancy in promoter regions typially occludes transcription factor binding, thereby repressing global gene expression, the role of histone modification is more complex. Histone tails can be modified in various ways, including acetylation, methylation, phosphorylation, and ubiquitination. Even the regulatory role of histone acetylation, the best characterized modification to date, is still not fully understood. Mass spectral and genome-wide microarray data from Saccharomyces cerevisiae have offered new opportunities for investigators to evaluate the regulatory effects of histone modifications. The investigators will develop statistical methods for identifying target genes of histone modifications and associated DNA sequence features of histone modifications. The investigators will also develop computational and statistical methods for predicting histone modifications and their interactions.

PI: Petra H. Lenz
State: HA
Funding Organization: NSF
Award #: OCE 1459235
Title: Collaborative Proposal: Optimizing Recruitment of Neocalanus copepods through Strategic Timing of Reproduction and Growth in the Gulf of Alaska
Date initiated: 2015-3-1
Date completed: 2019-2-28
This is a proposal to investigate winter recruitment in the copepod Neocalanus flemingeri in the Gulf of Alaska. Calanid copepods like N. flemingeri are characterized by rapi population growth during the spring coincident with the annual phytoplankton bloom. Recruitment to this spring population is dependent on the successful emergence from diapause followed by reproduction, and survival and growth of this next generation. However, an apparent mismatch between the presence of nauplii (youngest stages) in December/January and the occurrence and unpredictability of the spring phytoplankton bloom have raised questions regarding the timing of female reproduction, and subsequent survival of nauplii. Here, we propose to combine laboratory and field approaches to determine whether female reproduction is synchronized and the strategies for nauplius survival during low food conditions. Gene expression studies using RNA-Seq technology will be used to develop molecular markers for female dormancy and reproductive readiness and for naupliar growth, which in turn will be used to evaluate field collected individuals. As a first step, we will generate a de novo transcriptome for Neocalanus flemingeri.

PI: Leonie Moyle

State: IN
Funding Organization: NSF
Award #: MCB-1127059
Title: Deciphering Mechanisms of Prezygotic Reproductive Isolation in Solanum
Date initiated: 2011-12-1
Date completed: 2016-11-30
Reproductive isolation is an essential element of the biological species concept. However, we are only beginning to understand the mechanistic nature of reproductive barriers between closely related species of higher plants. The genus Solanum, with its rich diversity of mating systems, offers a unique system in which these mechanisms may be revealed. A broad range of molecular and genomic resources is available for these studies, including rapily expanding reference genomes, and comprehensive genetic maps and EST collections. Solanum species exhibit a prezygotic IRB known as unilateral interspecific pollen rejection, or UI. We have already made significant headway in understanding UI in tomato by characterizing pollen tube growth in interspecific crosses, determining the relationship of interspecies pollen rejection to SI, and identifying several candidate IRB genes. Here, we will further mine the genomes of wild tomato species for IRB genes, use transgenic approaches to live-image pollen during rejection and test function of newly identified IRB genes and examine structural differences in chromosome sets. While tomato will be the best system to use for uncovering IRB mechanisms, we will ultimately apply what we learn to potato, a close relative and one of the world's most important food crops, one with a limited gene pool but a rich family of wild relatives that could provide genetic diversity for crop improvement if IRB can be overcome.

PI: Anja Schulze
State: TX
Funding Organization: NSF
Award #: DEB-1036186
Title: Gene expression profiles in response to low dissolved oxygen in the bearded fireworm, Hermodice carunculata
Date initiated: 2016-3-1
Date completed: 2017-12-31
This project will examine gene expression in response to low dissolved oxygen in a common and and ecologically important marine invertebrate, the bearded fireworm, Hermodice carunculata. Low oxygen, or hypoxia, is an emergent stressor in the marine environment which, in extreme cases ('dead zones') can lead to widespread mortality. The cellular, physiological and organismal responses to low oxygen are poorly studied in invertebrates but could have far-reaching implications for entire ecosystems. Changes in gene expression precede physiological stress responses can therefore aid in identifying physiological pathways that may be affected. We propose to study differential gene expression under controlled laboratory conditions as well as between low oxygen and normoxic field sites on the coast of Brazil and the Gulf of Mexico. We anticipate that this project will lead to peer-reviewed publications, conference presentations, contribute to the dissertation work of two students and lead to future inter-institutional collaborations.

PI: H. Bradley Shaffer
State: CA

Funding Organization: NSF
Award #: DEB-1257648
Title: Tracking genes in real time as they traverse a hybrid invasion landscape.
Date initiated: 2013-3-1
Date completed: 2017-2-28
This project characterizes the genetic landscape of the California tiger salamander with regards to introgression with the invasive barred tiger salamander. Computational requirements for which we request NCGAS resources include mainly CPUs for large amounts of read mapping, as well as some time with high-RAM nodes for sequence assembly.

PI: Joseph Vitti
State: MA
Funding Organization: NSF
Award #: NSF GRFP 2014155775
Title: Characterizing natural selection and adaptation in diverse human populations
Date initiated: 2016-8-3
Date completed: 2017-8-3
Natural selection was instrumental not only in the genesis of our species, but also in its diversification. By examining patterns of genomic variation within and among populations, we can identify and characterize genetic variants that have been subject to selection, bringing instances of local adaptation to light. This project -- the culmination of my PhD research as an NSF GRFP fellow, takes a new suite of computational tools that I have designed and implemented and applies them to explore a rich new dataset (1000 Genomes Phase 3) which includes full sequence data for individuals from previously uncharacterized populations in South Asia, West Africa, and East Asia. Computational capacity represents the major limitation on my ability to bridge the gap from tool-building to empirical analysis, as this step necessitates the iterative generation of simulated data en masse.

PI: Laurel Yohe
State: NY
Funding Organization: NSF
Award #: NSF Graduate Research Fellowship
Title: Wake Up and Smell the pier: Olfactory Receptor Repertoires Reflect Specialization in Carollia
Date initiated: 2016-5-1
Date completed: 2017-8-30
My dissertation focuses on the comparative analysis of olfactory receptors across species with different dietary preferences. Specifically, I am testing if bats that feed on fruit have larger or more specialized olfactory receptor profiles compared to bats that eat insects. The proposed project focuses on a narrower question that falls within the breadth of my dissertation: do generalist species have a larger olfactory repertoire compared to specialist species? In order to test this research question, I have sequenced the olfactory receptors of 3 closely related species of the genus Carollia that demonstrate different levels of specialization on pier plants. One species primarily feeds on pier and nothing else, while one is a generalist frugivore, and the other falls in between on this spectrum. Additionally, pier plants rely on this group of bats for seed dispersal and we expect pier odorants have evolved to attract bats. This project will illuminate

how sensory biology evolves with resource preferences, and will improve our understanding of mutualistic interactions in the environment. I originally requested (through XSEDE) access to Mason to take advantage of the SMRT-analysis pieline designed for PacBio data. I do not have any other access to a Linux server, which is required for SMRT. As my project has advanced, I am also interested in comparing methods for sequencing olfactory genes, including PacBio and Illumina transcriptome data, both of which I have already generated. The space and computational demands of using both the SMRT-toolkit and Trinity have inspied my request for NCGAS. Expanding my access to the Mason server will allow my project reach its full potential.

PI: Scott Cinel
State: IL
Funding Organization: NSF
Award #:
Title: Transcriptomic Stress Responses to Bat Ultrasound in an Agricultural Pest, the Fall Armyworm (Spodoptera frugiperda)
Date initiated: 2015-12-10
Date completed: 2016-5-20
Though used to exemplify the concept of stress, predator-induced responses in prey organisms have, until recently, been viewed as acute and transitory, with minimal lasting impact on population demographics. However, fear factors significantly into ecological processes, as predator-induced behavioral changes have been acknowledged in driving patterns of optimal foraging and game theory. Recently, several studies have confirmed that predatory stress responses in prey are similar to chronic stress responses in humans, conferring significant impacts on fitness, activity, and survival. Further, prey undergo adaptive responses simply in the presence of predators by keying in on auditory, visual, or chemosensory cues. These cues can trigger the up-regulation of genes associated with stress responses, such as adipokinetic hormone genes involved in fat mobilization to provide enhanced energy in the face of a predator. I am investigating patterns of gene expression in prey exposed to high-intensity indirect predator cues. Specifically, I am exposing an ultrasonic-sensitive pest moth, the fall armyworm (Spodoptera frugiperda), to insectivorous bat call recordings. Post-exposure, I will utilize RNA-seq to compare whole brain gene expression of these exposed individuals to that of moths that have never been exposed to predator cues. I am requesting access to NCGAS services to carry out the computational tasks of de novo transcriptome assembly, quality control, normalization, annotation, and differential gene expression analyses. NCGAS services suit my project ideally as I am an NSF IGERT Fellow, and I require a cheap alternative to my university's bioinformatics cluster to accommodate my restricted budget.

PI: Matthew L. Niemiller
State: IL
Funding Organization: NSF
Award #:
Title: Systematics and evolution of subterranean amphipods (Amphipoda: Crangonyctidae) using full-genome sequencing
Date initiated: 2015-5-1
Date completed: 2016-12-31

Amphipods in the family Crangonyctidae offer an excellent opportunity to study the factors that promote or constrain speciation and adaptation, as extensive variation exists in a wide range of ecologically-relevant traits among taxa in the family. In particular, several putatively independent lineages have successfully colonized subterranean habitats and convergently evolved to life in perpetual darkness. However, evolutionary relationships among lineages and populations of most groups are in great need of study, as previous hypotheses are based entirely on morphological characters, including some that may be homoplasious. Next-generation sequencing (NGS) now allows the generation multilocus data sets consisting of hundreds to thousands of unlinked loci for phylogenetic inference. Here, we have generated NGS genomic dataset for six surface and subterranean taxa in the family to develop markers for phylogenetic inference at different time depths and to search for loci related to subterranean phenotypes, such as genes associated vision and pimentation. De novo genome assembly and annotation using NCGAS resources would provide preliminary genomic resources for a proposal for NSF funding through the Division of Environmental Biology to advance our understanding of diversity, systematics and evolutionary history of subterranean organisms. We will compare several recently developed approaches of species tree estimation using genome-scale datasets to develop a phylogenetic framework for future comparative studies examining morphological and genetic convergence of candidate genes underlying phenotypi traits that are the target of adaptive and regressive evolution in crangonyctid amphipods.

PI: Christine Picard
State: IN
Funding Organization: NSF
Award #: IU
Title: De novo genome assembly of Phormia regina
Date initiated: 2013-6-19
Date completed: 2017-6-19
We have obtained ~100Gb of genomic sequence data (paired-end Illumina reads) for which we are assembling the genome de novo (no reference genome). We have a preliminary assembly done using CLC Genomics Workbench, on a local machine, but we'd like to get some additional assemblies for comparison sake. We'd like to use Abyss and SOAPdenovo (and perhaps a 3rd option). In addition, we have collected an additional ~30Gb of RAD-seq data on the same species and would like to use Stacks (open source) software for the analysis of this data in order to do some SNP discovery.

PI: Josephine Reinhardt
State: NY
Funding Organization: NSF
Award #:
Title: Genomic impacts of meiotic drive sex chromosomes
Date initiated: 2016-7-5
Date completed: 2017-7-5
The primary research goal of this project is to determine the impact that meiotic drive - a powerful form of genomic conflict - has on the emergence of novel variation, with a particular focus on novel genes and structural variants. I will be employing a comparative approach, studying two independently evolved sex-ratio meiotic drive systems that are found at similar,

stable frequencies (10-30%) in natural populations. In both systems, the drive locus is associated with near-zero recombination rates between the standard (ST) and sex-ratio (SR) chromosomes, likely due to multiple inversions in both cases. These are ideal conditions to maximize the impact of sex-ratio drive on the gene content and molecular evolution of the X chromosome.

PI: Megan Porter
State: HA
Funding Organization: NSF
Award #: DEB/IOS
Title: The evolution and development of the complex stomatopod visual system
Date initiated: 2015-9-1
Date completed: 2016-12-30
Stomatopod adult visual systems contain the most complex array of photoreceptor classes known so far, with up to 20 photoreceptor types specialized to sample spectral, spatial, and polarization information. Developmentally, stomatopod species transition from a larval phase with a simple compound eye generally containing two types of photoreceptor to a 20 channel adult eye by constructing an entirely new retina over the course of a few hours or days. I am actively studying this unique ontogenetic transition and will investigate how the stomatopod visual system achieves a 2 to 20-channel increase in sensory input in a matter of hours. In particular, my lab is using RNAseq methods to look at the ontogentic change in expression of the gene networks involved in eye development, as well as the genes involved in phototransduction. I have recently moved to the University of Hawai'i at Manoa where I am able to actively pursue this line of research. I am currently generating transcriptomes across a developmental series for several stomatopod species. These datasets will be used as preliminary data for NSF IOS pre-proposal submission in January 2016, as well as for a subsequent NSF CAREER award proposal in July 2016. However, my institution does not currently have an easily accessible, low-cost option for RNAseq data analysis. De novo assembly, annotation, and expression profiles generated using NCGAS resources would provide the means of analyzing my data in a timely manner in order to use the results for upcoming grant target deadlines.

PI: Kazem Taghva
State: NV
Funding Organization: NSF
Award #: Information and computing
Title: Sea Cucumber Genome Mapping
Date initiated: 2016-1-20
Date completed: 2017-1-20
Sea cucumber genome mapping: sea cucumber has extraordinary power of wound healing and tissue regeneration. We are primarily interested in what genes may be supporting that capability. The first step we are taking is construct a genomic map. We have generated the whole genome sequencing data from the particular species of our interest and need to figure out how to construct a whole genome map. The genome map of sea urchin, which is a close relative of sea cucumber, is available, and this would help in the alignment. This is a project in cooperation with Keck School of Medicine's Professor Chih-lin Hsieh.

7. Appendix 2. IU Publications by NCGAS Users

Gulia-Nuss, M., Nuss, A. B., Meyer, J. M., Sonenshine, D. E., Roe, R. M., Waterhouse, R. M., Sattelle, D. B., de la Fuente, J., Ribeiro, J. M., Megy, K., Thimmapuram, J., Miller, J. R., Walenz, B. P., Koren, S., Hostetler, J. B., Thiagarajan, M., Joardar, V. S., Hannick, L. I., Bidwell, S., Hammond, M. P., Young, S., Zeng, Q., Abrudan, J. L., Almeida,

F. C., Ayllo´n, N., Bhide, K., Bissinger, B. W., Bonzon-Kulichenko, E., Buckingham, S. D., Caffrey, D. R., Caimano, M. J., Croset, V., Driscoll, T., Gilbert, D., Gillespie, J. J., Giraldo-Caldero´n, G. I., Grabowski, J. M., Jiang, D., Khalil, S. M., Kim, D., Kocan, K. M., Ko˘ci, J., Kuhn, R. J., Kurtti, T. J., Lees, K., Lang, E. G., Kennedy, R. C., Kwon, H., Perera, R., Qi, Y., Radolf, J. D., Sakamoto, J. M., Sa´nchez-Gracia, A., Severo, M. S., Silverman, N., S˘imo, L., Tojo, M., Tornador, C., Van Zee, J. P., V´azquez, J., Vieira, F. G., Villar, M., Wespiser, A. R., Yang, Y., Zhu, J., Arensburger, P., Pietrantonio, P. V., Barker, S. C., Shao, R., Zdobnov, E. M., Hauser, F., Grimmelikhuijzen, C. J., Park, Y., Rozas, J., Benton, R., Pedra, J. H., Nelson, D. R., Unger, M. F., Tubio, J. M., Tu, Z., Robertson, H. M., Shumway, M., Sutton, G., Wortman, J. R., Lawson, D., Wikel, S. K., Nene, V. M., Fraser, C. M., Collins, F. H., Birren, B., Nelson, K. E., Caler, E., and Hill, C. A. (2016). Genomic insights into the ixodes scapularis tick vector of lyme disease. Nature communications, 7.

Horton, M. A., Oliver, R., and Newton, I. L. (2015). No apparent correlation between honey bee forager gut microbiota and honey production. PeerJ, 3.

Krishnakumar, R., Chen, A. F., Pantovich, M. G., Danial, M., Parchem, R. J., Labosky, P. A., and Blelloch, R. (2016). FOXD3 regulates pluripotent stem cell potential by simultaneously initiating and repressing enhancer activity. Cell Stem Cell, 18(1):104–117.

Newton, I. L., Clark, M. E., Kent, B. N., Bordenstein, S. R., Qu, J., Richards, S., Kelkar, Y. D., and Werren, J. H. (2016). Comparative genomics of two closely related wolbachia with different reproductive effects on hosts. Genome biology and evolution, 8(5):1526–1542.

Newton, I. L. G. and Sheehan, K. B. (2015). Passage of wolbachia pipientis through mutant drosophila melanogaster induces phenotypic and genomic changes. Applied and Environmental Microbiology, 81(3):1032–1037.

Orsini, L., Gilbert, D., Podicheti, R., Jansen, M., Brown, J. B., Solari, O. S., Spanier, K. I., Colbourne, J. K., Rush, D., Decaestecker, E., Asselman, J., De Schamphelaere, K. A. C., Ebert, D., Haag, C. R., Kvist, J., Laforsch, C., Petrusek, A., Beckerman, A. P., Little, T. J., Chaturvedi, A., Pfrender, M. E., De Meester, L., and Frilander, M. J. (2016). Daphnia magna transcriptome by RNA-seq across 12 environmental stressors. Scientific Data, 3:160030+.

Raborn, R. T., Spitze, K., Brendel, V. P., and Lynch, M. (2016). Promoter architecture and sex-specific gene expression in the microcrustacean daphnia pulex revealed by large-scale profiling of 5'-mRNA ends. bioRxiv, pages 047894+.

Rokop, Z. P., Horton, M. A., and Newton, I. L. G. (2015). Interactions between cooccurring lactic acid bacteria in honey bee hives. Applied and Environmental Microbiology, 81(20):7261–7270.

Roncalli, V., Cieslak, M. C., and Lenz, P. H. (2016). Transcriptomic responses of the calanoid copepod calanus finmarchicus to the saxitoxin producing dinoflagellate alexandrium fundyense. Scientific reports, 6.

Suzuki, H., Dapper, A. L., Jackson, C. E., Lee, H., Pejaver, V., Doak, T. G., Lynch, M., and Preer, J. R. (2015). Draft genome sequence of caedibacter varicaedens, a kappa killer endosymbiont bacterium of the ciliate paramecium biaurelia. Genome announcements, 3(6).

Tarpy, D. R., Mattila, H. R., and Newton, I. L. G. (2015). Development of the honey bee gut microbiome throughout the Queen-Rearing process. Applied and Environmental Microbiology, 81(9):3182–3191.

8. Appendix 3. Software Supported by NCGAS

Software supported by NCGAS

The National Center for Genome Analysis Support (NCGAS) provides support for the following genome analysis software packages available on Indiana University's Mason cluster, and the Extreme Science and Engineering Discovery Environment (XSEDE) digital services Karst (TACC) and Greenfield (SDSC). Access to NCGAS computational and consulting services is awarded through an allocation process to genomics research projects funded by the National Science Foundation (NSF). For more, see the National Center for Genome Analysis Support site or email NCGAS.

Links to source code downloads and licensing information are provided for those who may want to install packages on local workstations or clusters.

ABySS

Assembly By Short Sequences (ABySS); de novo assembly of DNA for metagenomics, comparative genomics, and creation of draft genomes:

- Documentation: ABySS project site
- Download: ABySS releases
- License: BC Cancer Agency (BCCA) software license agreement (academic use)

Supported version(s)	Mason	Karst	Greenfield
1.3.3	x		
1.3.3-openmpi	x		
1.3.4	x		
1.3.4-openmpi	x		
1.3.6-openmpi	x		
1.5.1-openmpi	x		
1.5.2-openmpi	x		

Admixture

ADMIXTURE is a software tool for maximum likelihood estimation of individual ancestries from multilocus SNP genotype datasets. It uses the same statistical model as STRUCTURE but calculates estimates much more rapidly using a fast numerical optimization algorithm.

- Documentation: Admixture project site
- Documentation: Admixture documentation
- Download: Admixture downloads
- License: Citation

Supported version(s)	Mason	Karst
1.3.0 *		x

ALLPATHS-LG

Whole-genome shotgun assembly using Illumina long and short insert libraries for greatest accuracy:

- Documentation: ALLPATHS-LG manual (in PDF format)
- Download: Latest source code
- License: Copyright 2012 Broad Institute

Supported version(s)	Mason	Karst
41292	x	
43460	x	
45684	x	
52488*	x	

AMOS

A Modular, Open-Source (AMOS) collection of tools and class interfaces for the assembly of DNA reads, including modular assembly pipelines, and tools for overlapping, consensus generation, contigging, and assembly manipulation:

- Documentation: AMOS project page
- Documentation:AMOS wiki
- Download: Current downloads
- License: Perl Foundation Artistic License 2.0

Supported version(s)	Mason	Karst
3.0	x	
3.1	x	x

Arachne

Whole genome shotgun assembly of long Sanger reads:

- Documentation: ArachneWiki
- Download: Latest source code
- License: Copyright 2012 Broad Institute

Supported version(s)	Mason	Karst
3.2	x	

BamUtil

A repository that contains several programs that perform operations on SAM/BAM files:

- Documentation: BamUtil wiki
- Download: Github page
- License: GNU General Public License, version 3.0 (GPLv3)

Supported version(s)	Mason	Karst
1.0.13	x	x

BCFTools

Discovery of correlated genomic features, such as ESTs, polymorphisms, and mobile elements:

- Documentation: Project home page
- Documentation: BCFTools manual
- Download: Github page
- License: GNU General Public License, version 3.0 (GPLv3); MIT License

Supported version(s)	Mason	Karst
1.3*	x	x

BEDTools

Discovery of correlated genomic features, such as ESTs, polymorphisms, and mobile elements:

- Documentation: Project home page
- Documentation: BEDTools manual
- Download: Current downloads
- License: GNU General Public License, version 2 (GPLv2)

Supported version(s)	Mason	Karst
2.20.1	x	

Bio3D

R package containing utilities for processing, organizing, and exploring protein structure and sequence data:

- Documentation: Project home page
- Documentation: Bio3D manual
- Download: Current downloads
- License: GPLv2

Supported version(s)	Mason	Karst
1.1-4	x	

Bioconductor

R packages for analysis and comprehension of high-throughput genomic sequence data:

- Documentation: Project home page
- Documentation: Bioconductor packages
- Installation: Instructions
- License: GPLv2

Supported version(s)	Mason	Karst
2.10	x	
2.12		x

Bioperl

Collection of perl packages to support bioinformatics:

- Documentation: Project home page
- Documentation: Github page
- Installation: Instructions
- License: GPL; Artistic License; Perl artistic license

Supported version(s)	Mason	Karst
1.6.1	x	

Biopython

Collection of python packages to support bioinformatics:

- Documentation: Project home page
- Documentation: Github page
- Installation: Instructions
- License: Custom

Supported version(s)	Mason	Karst
1.59	x	x
1.63	x	

Bismark

A tool to map bisulfite converted sequence reads and determine cytosine methylation states:

- Documentation: Project home page
- Download: Source code
- License: GNU General Public License, version 3.0 (GPLv3)

Supported version(s)	Mason	Karst

Supported version(s)	Mason	Karst
0.12.5	x	x

BitSeq

Transcript isoform level expression and differential expression estimation for RNA-seq:

- Documentation: Project home page
- Download: Github page
- License: OSI Artistic License 2.0

Supported version(s)	Mason	Karst
0.4.1	x	

BFAST

Blat-like Fast Accurate Search Tool (BFAST) facilitates the fast and accurate mapping of short reads to reference sequences, where mapping billions of short reads with variants is of utmost importance:

- Documentation: Project home page
- Download: BFAST @ SourceForge
- License: GNU General Public License, version 2 (GPLv2)

Supported version(s)	Mason	Karst
0.7.0a		x

BLAT

Fast alignment of highly similar sequences of DNA/proteins to find ESTs or to align reads to reference:

- Documentation: BLAT FAQ
- Download: Source code
- Download: Executables
- License: Freely available for academic, nonprofit, and personal use; a license is required for commercial use

Supported version(s)	Mason	Karst
35	x	x

Bowtie

Alignment of short reads to a reference genome in order to approximate coverage, find polymorphisms, and assess assembly quality:

- Documentation: Bowtie project site
- Download: Bowtie @ SourceForge
- License: Perl Foundation Artistic License 2.0

Supported version(s)	Mason	Karst
0.12.8	x	x
1.1.1	x	
1.1.2*	x	x
2.0.6	x	x
2.1.0	x	x
2.2.3	x	x
2.2.6	x	x

Breseq

Breseq is a computational pipeline for finding mutations relative to a reference sequence in short-read DNA re-sequencing data for haploid microbial-sized genomes:

- Documentation: Project home page
- Download: Github page
- License: GPLv2

Supported version(s)	Mason	Karst
0.27	x	

BreakDancer

Provides genome-wide detection of structural variants from next generation paired-end sequencing reads:

- Documentation: Project home page
- Download: BreakDancer @ SourceForge
- License: GPL, version 3 (GPLv3)

Supported version(s)	Mason	Karst
1.1		x

Burrows-Wheeler Aligner (BWA)

Alignment of long and short reads from a variety of technologies, allows gaps, for approximating coverage, finding polymorphisms, and assessing assembly quality:

- Documentation: BWA project site
- Download: BWA @ SourceForge
- License: GPL, version 3 (GPLv3); MIT License

Supported version(s)	Mason	Karst
0.6.2	x	
0.7.2	x	
0.7.6a	x	
0.7.10		x

Cafe

Computational analysis of (gene) family evolution:

- Documentation: Cafe manual
- Download: Cafe @ SourceForge
- License: Freely available

Supported version(s)	Mason	Karst
2.1	x	
3.0	x	

CD-HIT

Clustering Database at High Identity with Tolerance (CD-HIT) is a clustering program for large sets of protein and DNA to determine relationships between many sequences:

- Documentation: CD-HIT project page
- Download: Current downloads
- License: GPLv2

Supported version(s)	Mason	Karst
4.5.6	x	

Celera

De novo whole-genome shotgun (WGS) DNA sequence assembler; reconstructs long sequences of genomic DNA from fragmentary data produced by whole-genome shotgun sequencing; developed at Celera Genomics and released to SourceForge in 2004 as the wgs-assembler:

- Documentation: Celera project page
- Download: wgs-assembler @ SourceForge
- License: GPLv3

Supported version(s)	Mason	Karst
2012_20_11	x	
7.0	x	
8.3rc2	x	

Clustal

Multiple alignment of nucleic acid and protein sequences:

- Documentation: Project home page
- Download: Source code
- License: Lesser General Public License (LGPL)

Supported version(s)	Mason	Karst
Omega-1.2.1	x	
W2-2.0.12		x

Cufflinks

Map RNA-Seq reads to reference genomes in order to annotate genes, discover splice variants, and estimate differential expression:

- Documentation: Cufflinks project page
- Download: Cufflinks downloads
- License: Boost Software License

Supported version(s)	Mason	Karst
2.0.2	x	x
2.1.1	x	x
2.2.0	x	x

Cutadapt

Trims adapter sequences from high-throughput sequencing data:

- Documentation: Project home page
- Documentation: User guide
- Download: Current downloads
- License: MIT License

Supported version(s)	Mason	Karst
1.10	x	x
1.2.1	x	
1.7.1	x	
1.9	x	

Cytoscape

Open source platform for visualizing molecular interaction networks and biological pathways:

- Documentation: Cytoscape project page
- Documentation: Cytoscape user documentation
- Download: Current downloads
- License: GNU Lesser General Public License, version 3 (LGPLv3)

Supported version(s)	Mason	Karst
2.8.3	x	

EMBOSS

The European Molecular Biology Open Software Suite, A high-quality package of free, Open Source software for molecular biology:

- Documentation: EMBOSSproject page
- Download: EMBOSS downloads
- License: Gnu Public License

Supported version(s)	Mason	Karst
6.5.7		x

Ensembl

Various tools to assist in use and analysis of Ensembl data:

- Documentation: EDENA project page
- Download: EDENA downloads
- License: Apache 2.0

Supported version(s)	Mason	Karst
81	x	x

EDENA

Exact De Novo Assembler (EDENA); de novo assembly of short reads for smaller genome assembly:

- Documentation: EDENA project page
- Download: EDENA downloads
- License: GPLv3

Supported version(s)	Mason	Karst
2.1.1	x	

FastQC

Quality control for high-throughput sequence data:

- Documentation: FastQC project page
- Documentation: FastQC help documentation
- Download: Current downloads
- License: GPLv3

Supported version(s)	Mason	Karst
0.10.1	x	x

FastX

A collection of command line tools for Short-Reads FASTA/FASTQ files preprocessing:

- Documentation: Project home page
- Documentation: Github page
- Download: Binaries and releases
- License: Affero GPLv3 or greater

Supported version(s)	Mason	Karst
0.0.13		x

FlashPCA

Performs fast principal component analysis (PCA) of single nucleotide polymorphism (SNP) data:

- Documentation: Github page
- License: GPLv3

Supported version(s)	Mason	Karst
1.2.6		x

Galaxy

A flexible GUI wrapper for bioinformatics tools, allowing users to manipulate genomic data and run analyses:

- Documentation: The Galaxy Project
- Download: Get Galaxy
- License: OSI Academic Free License 3.0 (AFL 3.0)

Supported version(s)	Mason	Karst
3	x	

GATK

GATK (Genome Analysis Toolkit) is a suite of genomics analysis tools with a focus on variant calling and gene finding:

- Documentation: GATK website
- Download: Download the GATK
- License: Academic, non-commercial research purposes only

Supported version(s)	Mason	Karst
1.1-33	x	
3.4-0	x	

GenomeMapper

Short read alignment, allows gaps, allows multiple references; used for estimating coverage, finding polymorphisms, variant calling, and quantitative analysis:

- Documentation: GenomeMapper project site
- Download: GenomeMapper versions
- Licensing terms not yet determined

Supported version(s)	Mason	Karst
0.4.3	x	x

GMAP

Align cDNA to reference to determine gene structure and structural variants:

- Documentation: GMAP README file
- Download: GMAP source code
- License: Free to use and modify for own purpose; copyright (2005-2011) Genentech, Inc.

Supported version(s)	Mason	Karst
04/04/16	x	x
05/15/14	x	x

HISAT

A fast and sensitive spliced alignment program for mapping RNA-seq reads:

- Documentation: Project home page
- Documentation: Github page
- Download: Source code
- License: GPLv3

Supported version(s)	Mason	Karst

Supported version(s)	Mason	Karst
0.1.6-beta		x

HMMER

Searches sequence databases for homologs of protein sequences and makes protein sequence alignments:

- Documentation: HMMER project page
- Documentation: HMMER User's Guide
- Download: Current downloads
- License: GPLv3

Supported version(s)	Mason	Karst
3.0	x	
3.1b2	x	x

Kallisto

A library and toolkit for analysis and transformations of fixed-length DNA subsequence (k-mer) datasets:

- Documentation: khmer project page
- Documentation: khmer's command-line interface
- Download: Official repository
- License: Berkeley Software Distribution (BSD) license; Copyright California Institute of Technology and Michigan State University

Supported version(s)	Mason	Karst
1.0	x	
1.3	x	
2.0	x	

Khmer

A library and toolkit for analysis and transformations of fixed-length DNA subsequence (k-mer) datasets:

- Documentation: khmer project page
- Documentation: khmer's command-line interface
- Download: Official repository
- License: Berkeley Software Distribution (BSD) license; Copyright California Institute of Technology and Michigan State University

Supported version(s)	Mason	Karst
1.0	x	

Supported version(s)	Mason	Karst
1.3	x	
2.0	x	

MACH

MACH 1.0 is a Markov Chain based haplotyper. It can resolve long haplotypes or infer missing genotypes in samples of unrelated individuals:

- Documentation: MACH project page
- Download: MACH download
- License: Custom, do not distribute

Supported version(s)	Mason	Karst
1.0.18		

MACS

Model-based Analysis of ChIP-Seq (MACS); algorithm for analyzing data from chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-Seq) that models the length of sequenced chromatin immunoprecipitation (ChIP) fragments to identify transcript factor binding sites:

- Documentation: MACS project page
- Documentation: README for MACS
- Download: Current downloads
- License: Perl Foundation Artistic License 1.0

Supported version(s)	Mason	Karst
1.4.2	x	x

MAKER

Pipeline for genome annotation that identifies and masks out repeat elements, aligns ESTs and proteins to a genome, produces ab initio gene predictions, and automatically synthesizes these data into final annotations with evidence-based quality values for downstream annotation management:

- Documentation: MAKER project page
- Documentation: MAKER wiki
- Download: Register to download
- License: Available for academic use under the Perl Foundation Artistic License 2.0 or the GPLv3

Supported version(s)	Mason	Karst
2.27-beta	x	

Supported version(s)	Mason	Karst
2.31.6	x	

MaSuRCA

MaSuRCA is whole genome assembly software. It combines the efficiency of the de Bruijn graph and Overlap-Layout-Consensus (OLC) approaches. MaSuRCA can assemble data sets containing only short reads from Illumina sequencing or a mixture of short reads and long reads (Sanger, 454, Illumina):

- Documentation: MaSuRCA project page
- Download: Email required
- License: GPLv3

Supported version(s)	Mason	Karst
2.3.2	x	

MEGAHIT

MEGAHIT is a single node assembler for large and complex metagenomics NGS reads, such as soil. It makes use of succinct de Bruijn graph (SdBG) to achieve low memory assembly:

- Documentation: Github wiki
- Download: Github page
- License: GPLv3

Supported version(s)	Mason	Karst
1.0.2	x	

MEME

The MEME Suite allows the biologist to discover novel motifs in collections of unaligned nucleotide or protein sequences, and to perform a wide variety of other motif-based analyses.

- Documentation: Project page
- Download: MEME download
- License: Custom

Supported version(s)	Mason	Karst
gcc/4.10.1_4	x	
gcc/4.11.2	x	

MetAMOS

A modular metagenomic assembly, analysis, and validation pipeline:

- Documentation: MetAMOS project page
- Documentation: MetAMOS documentation
- Download: Official repository; download the latest tagged release (tar.gz, .zip)
- License: Free for academic, non-commercial use, MetAMOS includes several third-party tools available under various open source and proprietary commercial licenses; see LICENSE.txt

Supported version(s)	Mason	Karst
1.1	x	
1.5rc3	x	

Migrate

Estimates effective population sizes and past migration rates between n population assuming a migration matrix model with asymmetric migration rates and different subpopulation sizes:

- Documentation: Project page
- Download: Downloads page
- License: MIT License

Supported version(s)	Mason	Karst
Intel/mpi/3.3.2		x
Intel/serial/3.3.2		x

Minimac

A low memory, computationally efficient implementation of the MaCH algorithm for genotype imputation:

- Documentation: Project home page
- Download: Source code
- License: MIT License

Supported version(s)	Mason	Karst
11162012		x

MISO

MISO (Mixture-of-Isoforms) is a probabilistic framework that quantitates the expression level of alternatively spliced genes from RNA-Seq data, and identifies differentially regulated isoforms or exons across sample:

- Documentation: Project home page
- Download: Github page
- License: GPLv2

Supported version(s)	Mason	Karst
0.4.6		x
fastmiso-3682184-3	x	

mlRho

Serial program for estimating the population mutation and recombination rates from shotgun-sequenced diploid genomes:

- Documentation: mlRho project page
- Documentation: mlRho - a program for estimating the population mutation and recombination rates from shotgun-sequenced diploid genomes *(Molecular Ecology, Volume 19, Issue Supplement s1, 277-284*)
- Download: Latest version (mlRHO_2.8.tgz)
- License: GPLv2

Supported version(s)	Mason	Karst
1.7	x	

mothur

Bioinformatics tool for analyzing 16S rRNA gene sequences:

- Documentation: Project home page and wiki
- Documentation: mothur manual
- Download: Download mothur
- License: GPLv3

Supported version(s)	Mason	Karst
1.31.2	x	x
1.32.1	x	
1.34.2	x	x
1.36.1	x	
1.38.1	x	
mpi/1.31.2	x	x
mpi/1.32.1	x	
mpi/1.34.1	x	x

MrBayes

MrBayes is a program for Bayesian inference and model choice across a wide range of phylogenetic and evolutionary models. MrBayes uses Markov chain Monte Carlo (MCMC) methods to estimate the posterior distribution of model parameters:

- Documentation:

- Download:
- License:

Supported version(s)	Mason	Karst
mpi/3.2.1		x
serial/3.2.1		x

mrsFAST

mrsFAST is designed to map short reads to reference genome assemblies in a fast and memory-efficient manner:

- Documentation: Project home page
- Documentation: Github page
- Download: Tarball link
- License: Custom

Supported version(s)	Mason	Karst
3.3.7		x

MUMmer

Ultra-fast alignment of large-scale DNA and protein sequences:

- Documentation: MUMmer home page
- Download: MUMmer @ SourceForge
- License: OSI Artistic License 2.0

Supported version(s)	Mason	Karst
3.22	x	
3.23	x	x

MUSCLE

MUSCLE is one of the best-performing multiple alignment programs according to published benchmark tests, with accuracy and speed that are consistently better than CLUSTALW:

- Documentation:
- Download: MUSCLE download
- License:

Supported version(s)	Mason	Karst
3.8.31		x

NCBI BLAST+

A search tool for comparing primary biological sequence information, such as the amino-acid sequences of different proteins or the nucleotides of DNA sequences:

- Documentation: Project home page
- Download: FTP
- License: Public domain

Supported version(s)	Mason	Karst
2.2.27		x
2.2.28		x

NGSUtils

NGSUtils is a suite of software tools for working with next-generation sequencing datasets:

- Documentation: Project home page
- Download: NGSUtils download
- License: BSD

Supported version(s)	Mason	Karst
0.5.0c	x	x
0.5.2a		x

NINJA

Infers phylogeny using neighbor-joining tree:

- Documentation: NINJA project site
- Download: NINJA downloads
- License: GNU LGPLv3

Supported version(s)	Mason	Karst
1.2.1	x	

Novoalign

Aligns short reads to reference genome for resequencing experiments:

- Documentation: Novoalign documentation page
- Download: Novoalign downloads
- License: License types

Supported version(s)	Mason	Karst
2.07.13	x	
3.00.02	x	

Oases

De novo transcriptome assembler for very short reads:

- Documentation: Oases project page
- Documentation: Oases manual
- Documentation: Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels *(Bioinformatics, Volume 28, Issue 8, 1086-1092)*
- Download: Current version (oases_0.2.08) (requires Velvet 1.2.08 or higher)
- License: GPLv3

Supported version(s)	Mason	Karst
0.2.08	x	

OrthoMCL

A genome-scale algorithm for grouping orthologous protein sequences:

- Documentation: OrthoMCL project page
- Download: OrthoMCL download
- License: Custom

Supported version(s)	Mason	Karst
2.0.9	x	

PAML

PAML is a package of programs for phylogenetic analyses of DNA or protein sequences using maximum likelihood:

- Documentation: Project home page
- Documentation: Manual
- Download: Source
- License: Free for academic use, copyright to author

Supported version(s)	Mason	Karst
4.8		x

Picard

Provides tools and methods for manipulating sequence alignments for assembly quality assessment, variant calling, and downstream processing:

- Documentation: Picard website
- Download: Picard @ SourceForge
- License: Apache License, Version 2.0; MIT License

Supported version(s)	Mason	Karst
1.52	x	

PHYLIP

PHYLIP (the PHYLogeny Inference Package) is a package of programs for inferring phylogenies (evolutionary trees):

- Documentation: Project home page
- Download: Downloads page
- License: Open source

Supported version(s)	Mason	Karst
3.39		x

PLINK

Whole genome association analysis toolset designed to perform a range of basic, large-scale analyses in a computationally efficient manner:

- Documentation: Project home page
- Download: Download page
- License: GPLv2

Supported version(s)	Mason	Karst
1.07		x

r8s

Estimates absolute rates ('r8s') of molecular evolution and divergence times on a phylogenetic tree):

- Documentation: r8s manual
- Tutorial: Phylogenetics: r8s lab
- Download: R8s download @ SourceForge
- License: GPLv2

Supported version(s)	Mason	Karst
1.8		x

RAxML

Maximum likelihood phylogeny estimation for interpreting relationships between sets of data:

- Documentation: Developer's website

- Documentation: Hybrid MPI/Pthreads Parallelization of the RAxML Phylogenetics Code (in PDF format)
- Documentation: Hybrid Parallelization of the MrBayes & RAxML Phylogenetics Codes (in PDF format)
- Download: Standard RAxML downloads
- License: GPLv2

Supported version(s)	Mason	Karst
7.2.6	x	
7.2.8	x	x
7.4.2		X
8.0.26	x	x

Rosetta

Algorithms for computational modeling and analysis of protein structures:

- Documentation: Rosetta project page
- Download: Rosetta download form
- License: Varies

Supported version(s)	Mason	Karst
gnu/mpi/3.5	x	x

RSEM

RSEM (RNA-Seq by Expectation-Maximization); accurate quantification of gene and isoform expression from RNA-Seq data:

- Documentation: RSEM project page
- Documentation: README for RSEM
- Download: Source code downloads; RSEM GitHub repository
- License: GPLv3

Supported version(s)	Mason	Karst
1.2.19	x	x
1.2.5	x	

Sailfish

Alignment-free algorithm for the estimation of isoform abundances directly from a set of reference sequences and RNA-seq reads:

- Documentation: Sailfish Home
- Download: Sailfish Download
- License: GPLv3

Supported version(s)	Mason	Karst
gnu/0.7.3		x
gnu/0.8.0		x

Salmon

Salmon is a tool for quantifying the expression of transcripts using RNA-seq data:

- Documentation: Salmon @ GitHub
- Download: Salmon Download @ GItHub
- License: GPLv3

Supported version(s)	Mason	Karst
gnu/0.4.2		x

sam2count

Python script for creating a counts table from reads aligned to transcripts:

- Documentation/download: sam2counts project page
- License: GPLv3

Supported version(s)	Mason	Karst
1.0	x	

SAMtools

Provides various utilities for manipulating alignments in Sequence Alignment/Map (SAM) format, including sorting, merging, indexing and generating alignments in a per-position format:

- Documentation: SAMtools website
- Download: SAMtools @ SourceForge
- License: BSD; MIT License

Supported version(s)	Mason	Karst
0.1.18	x	
0.1.19	x	x
1.2	x	x
1.3	x	x

Scythe

3'-end adapter contaminant trimmer that uses a naive Bayesian model to classify contaminant substrings in sequence reads:

- Documentation: Scythe project page
- Documentation: Scythe README
- Download: Scythe source code; Scythe relies on Heng Li's kseq.h (which is bundled with the source) and requires the zlib data-compression library
- License: MIT License

Supported version(s)	Mason	Karst
0.992	x	

SHORE

SHORE (Short Read) is a mapping and analysis pipeline for mapping short DNA reads to a reference genome to find genomic polymorphisms and structural variants, and perform quantitative analysis:

- Documentation: SHORE project site
- Documentation: SHORE manual
- Download: SHORE @ SourceForge
- License: GPLv3

Supported version(s)	Mason	Karst
0.6.1 beta	x	

SMRT Analysis

Automated and distributed secondary analysis of sequencing data generated by the PacBio single-molecule, real-time (SMRT) sequencing system:

- Documentation: Official documentation
- Download: Current downloads
- License: PacBio software end user license agreement

Supported version(s)	Mason	Karst
1.3.1	x	
2.0.1	x	
2.2.0	x	

SOAPdenovo

De novo assembly of short reads for large genomes, creating reference genomes of novel organisms:

- Documentation: SOAPdenovo home page
- Download: SOAPdenovo downloads
- License: GPL v3

Supported version(s)	Mason	Karst
1.04	x	
1.05	x	
R240	x	

SOAPdenovo-Trans

A de novo transcriptome assembler inherited from the SOAPdenovo2 framework, designed for assembling transcriptome with alternative splicing and different expression leve:

- Documentation: SOAPdenovo-Trans manual
- Download: SOAPdenovo-Trans GitHub repository
- License: GPL v3

Supported version(s)	Mason	Karst
1.03	x	

SortMeRNA

SortMeRNA is a biological sequence analysis tool for filtering, mapping and OTU-picking NGS reads:

- Documentation: SortMeRNA manual @ GitHub
- Download: SortMeRNA GitHub repository
- License: GPL v3

Supported version(s)	Mason	Karst
gnu/mpi/3.5	x	

SPAdes

SPAdes - St. Petersburg genome assembler - is intended for both standard isolates and single-cell MDA bacteria assemblies:

- Documentation: SPAdes manual
- Download: SPAdes download form
- License: free use

Supported version(s)	Mason	Karst
3.5	x	
3.6.1	x	
3.8.2	x	

SRA Toolkit

Tools and libraries for working with data files and reference sequences from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA):

- Documentation: Understanding and using SRA
- Documentation: SRA Toolkit installation and configuration, protected data usage guide, and frequently used tools
- Download: Latest release; latest source code
- License: Public domain

Supported version(s)	Mason	Karst
2.1.15	x	
2.3.5-2	x	x
2.5.4		x

Stacks

A modular pipeline for building loci from short-read sequences:

- Documentation: Stacks project page
- Documentation: Stacks Manual
- Download: Latest version (stacks-1.21.tar.gz)
- License: GPLv3

Supported version	Mason	Karst
1.0.6	x	

STAR

Spliced Transcripts Alignment to a Reference:

- Documentation: STAR @ GitHub
- Download: STAR repository
- License: GPLv3

Supported version(s)	Mason	Karst
2.4.1d	x	x

Tabix

Tabix works on generic tabular data formats for genomic information, quickly retrieving features overlapping specified areas on the genome:

- Documentation: Part of the SAMtools package
- Download: Tabix project page
- License: MIT/X11

Supported version(s)	Mason	Karst
0.2.6	x	x

TopHat

Splice junction mapper for RNA-Seq reads; aligns RNA-Seq reads to mammalian-sized genomes using the short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons:

- Documentation: TopHat manual
- Download: TopHat downloads
- License: Boost Software License

Supported version(s)	Mason	Karst
1.4.1		x
2.0.5	x	x
2.0.7	x	x
2.1.0	x	x

TPP

A software solution for MS/MS-based shotgun proteomics analysis:

- Documentation: TPP @ SourceForge
- Download: TPP download
- License: Lesser General Public License (LGPL)

Supported version(s)	Mason	Karst
4.6.2		x

Trans-ABySS

Analysis for ABySS-assembled contigs from shotgun transcriptome data for finding splice sites and variants:

- Documentation: Trans-ABySS project site
- Download: Trans-ABySS releases
- License: BCCA software license agreement (academic use)

Supported version(s)	Mason	Karst
1.3.2	x	

TransDecoder

TransDecoder identifies candidate coding regions within transcript sequences, such as those generated by de novo RNA-Seq transcript assembly using Trinity, or constructed based on RNA-Seq alignments to the genome using Tophat and Cufflinks:

- Documentation: TransDecoder @ GitHub
- Download: Most recent version download
- License: Copyright 2012

Supported version(s)	Mason	Karst
2.0.1	x	

Trimmomatic

Trimmomatic performs a variety of useful trimming tasks for Illumina paired-end and single ended data:

- Documentation: Trimmomatic Home
- Download: Download Trimmomatic
- License: GPLv3

Supported version(s)	Mason	Karst
0.35	x	

Trinity

Combines three software modules (Inchworm, Chrysalis, and Butterfly) for de novo reconstruction of transcriptomes from RNA-Seq data:

- Documentation: Trinity project page
- Download: Trinity RNA-Seq Assembly @ SourceForge
- License: BSD license

Supported version(s)	Mason	Karst
10/5/12	x	
2/25/13	x	
08/14/13	x	
11/10/13	x	
04/13/14	x	
07/17/14	x	
2.0.6	x	
2.1.1		x
2.2.0	x	

VCFtools

A program package designed for working with VCF files:

- Documentation: VCF tools manual
- Download: VCF tools @ SourceForge
- License: LGPLv3

Supported version(s)	Mason	Karst
0.1.10	x	x
0.1.13	x	x
0.1.14	x	x

Velvet

De novo assembly of short reads with paired ends for smaller genome assembly of novel organisms:

- Documentation: Velvet manual
- Download: Velvet release history
- License: GPLv2

Supported version(s)	Mason	Karst
1.2.03	x	
1.2.08	x	
1.2.10	x	
1.2.10-longseq	x	

This document was developed with support from National Science Foundation (NSF) grant OCI-1053575. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.