

ABI Sustaining: The National Center for Genome Analysis Support 2017 Annual Report

*Thomas G. Doak
Craig A. Stewart
Scott D. Michaels*

Indiana University
PTI Technical Report PTI-TR17-003

June 30, 2017

Citation:

Doak, T.G., Stewart, C.A., Michaels, S.D. (2017) "ABI development: National center for genome analysis support 2017 annual report", Indiana University, Bloomington, IN. PTI Technical Report PTI-TR17-003. Retrieved from <http://hdl.handle.net/2022/21613>



**PERVASIVE TECHNOLOGY
INSTITUTE**

INDIANA UNIVERSITY



**RESEARCH
TECHNOLOGIES**

INDIANA UNIVERSITY

University Information Technology Services
Pervasive Technology Institute

Table of Contents

ABI Development: National Center for Genome Analysis Support	i
1. Accomplishments	3
1.1. What are the major goals of the project?.....	3
1.2. What was accomplished under these goals?	3
1.2.1. Major Activities.....	3
1.2.2. Specific Objectives	6
1.2.3. Significant results.....	6
1.2.4. Key outcomes or other achievements	7
1.3. What opportunities for training and professional development has the project provided?	10
1.4. How have the results been disseminated to communities of interest?	10
1.5. What do you plan to do during the next reporting period to accomplish the goals?	11
1.5.1	??
1.5.2 Synergistic activities	
2. Products	13
2.1. Products resulting from this project during the specified reporting period	14
2.1.1. (Peer-reviewed) Journal Articles.....	14
2.1.2. Conference Papers and Presentations	Error! Bookmark not defined.
2.1.3. Other Publications.....	Error! Bookmark not defined.
3. Participants	17
3.1. Individuals.....	17
3.1.1. Full details of individuals who have worked on the project	17
3.2. Partner organizations	19
3.2.1. Full details of partner organizations.....	20
3.3. Have other collaborators or contacts been involved?	21
4. Impact.....	21
4.1. What is the impact on the development of the principal discipline(s) of the project?	21
4.2. What is the impact on other disciplines?	21
4.3. What is the impact on the development of human resources?	22
4.4. What is the impact on physical resources that form infrastructure?	22
4.5. What is the impact on institutional resources that form infrastructure?	22
4.6. What is the impact on information resources that form infrastructure?	22
4.7. What is the impact on technology transfer?	22
4.8. What is the impact on society beyond science and technology?	22
5. Changes/ Problems	22
5.1. Changes in approach and reasons for change	23
5.2. Actual or Anticipated problems or delays and actions or plans to resolve them	23
5.3. Changes that have significant impact on expenditures	23

5.4. Significant changes in use or care of human subjects23
5.5. Significant changes in the use or care of vertebrate animals23
5.6. Significant changes in the use or care of biohazards.....23

1. Accomplishments

1.1. What are the major goals of the project?

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

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

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?

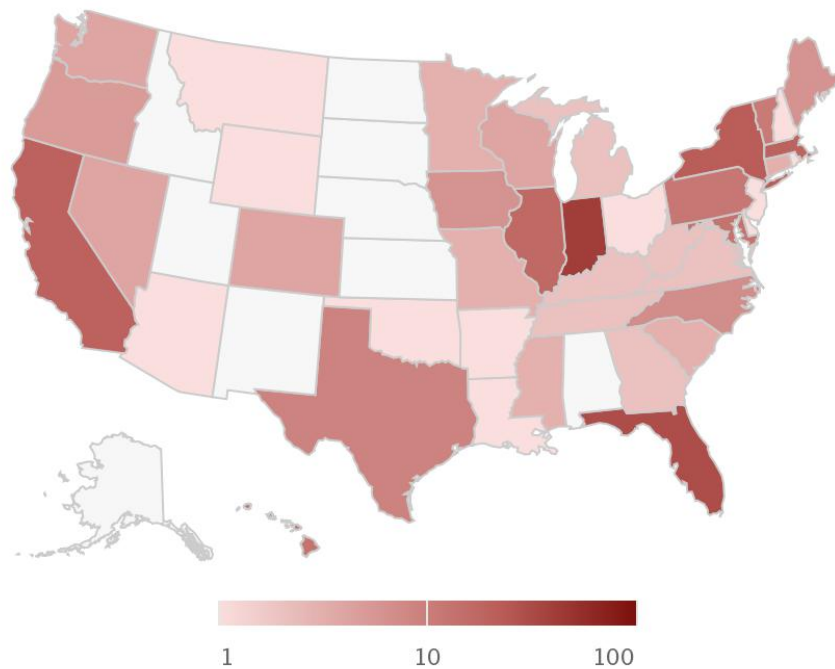
1.2.1. Major Activities

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 second 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 41 states (Fig. 1a), including 12 EPSCoR states, and in partnership with the Trinity development team (see Synergistic activities) have users around the world (Fig. 5).

National Center for Genome Analysis Support Users 2017

Representing 117 institutions in 41 states



Highcharts.com © Natural Earth

Figure 1. States with NCGAS Users

1.2.1.1. Provide excellent bioinformatics consulting services.

NCGAS' most significant accomplishments in support of biological and bioinformatics research continue to be in researchers' discoveries from RNA and DNA transcriptome, metatranscriptome, genome, and metagenome assemblies.

In year 2, NCGAS aided researchers in completing many *de novo* assemblies and genomic analysis, including the following organisms (completed and on-going):

- Coffee, peanut and sweet potato transcriptomes
- *Daphnia* genomes (population genomics and *de novo* assemblies)
- Barred tiger salamander transcriptomes
- Diatoms transcriptomes
- The diverse microbial clade Stramenopila + Alveolata + Rhizaria (SAR) (*de novo* genomes, transcriptomes)
- Heliconius butterflies transcriptomes
- Carrion Flies (forensic arthropods), *de novo* genome assembly/resequencing, transcriptomes
- Mussels,
- Crawfish Frog (endangered)
- Bahama Giant conch (a mollusk)
- Little Brown Skate,
- *D. galeata*

In addition, NCGAS supported 223 biologists doing research in the general area of genome analysis (15 named new allocations) during the current year's funding. Assistance has been provided through 386 short consultations and 29 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 the researchers' specific data and questions to HPC hardware, specialized applications and knowledge. Some researchers only need access to large memory clusters, which we can provide in a number of ways; others need instruction in basic HPC use and genomic analysis. We have been successful in this and continue to attract new users, often by word-of-mouth (documented in our just closed survey of users). One of our on-going tasks is to stay current: new hardware becomes available, 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 installed or updated five packages (spades, hisat2, salmon, STAR, kallisto) for assembly and analysis of PacBio long-read sequencing data. Hybrid assemblies are quickly becoming popular, and we installed Canu, MaSuRCA and PacBio's SMRT analysis software. Funded collaborator PSC also makes many of these packages available on PSC systems, and also focuses on 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 new *Carbonate* IU cluster (will replace Mason shortly)
- The *Jetstream* cloud environment
- Additional XSEDE resources, including 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

We now support public and private genome browsers for: XXX, XXX, etc

Overall, we have installed or updated 44 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 IUScholarWorks, 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.
- We started a collaboration with Raphael Isokpehi at Bethune-Cookman University, providing student access to Mason and teaching classes remotely. This relationship will expand in the future.

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:

- 200 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
 - Plant and Animal Genomes 2017. San Diego.
 - MDI Biological Laboratory's 2016 Environmental Genomics course / Bangor ME
 - Cyber-lectures to bioinformatics students at Bethune-Cookman.

1.2.2. *Specific Objectives*

The software supported by NCGAS as of the end of the first four years of NSF funding includes 44 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 386 short term consulting engagements (those taking less than 4 hours of staff time to resolve) and 29 long term consulting engagements (taking more than 4 hours of staff time to resolve). Many long term consultations are research collaborations that last months or years, with NCGAS staff becoming partners, playing a critical role in discoveries by scientists receiving NCGAS help.
- NCGAS completed tutorials and training and outreach activities attended by hundreds of attendees (see 1.3).
- The past year has included the Jetstream cloud opening. NCGAS has worked to bring biologists to this NSF resource, including: 1) 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; 2) Established genome browsers to, that link to Wrangler for storage; 3) Adding the research group of Tim Grimes to establish proteomic galaxy instances for both research and teaching purposes.
 - 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 15 new long-term projects, described in the attached file on consulting and significant results.

A significant activity is the strengthening of our partnership with PSC (see 1.5). With Philip Blood and PSC as a funded member of the NCGAS team in the current grant, we have been able to escalate our relationship. Activities include: 1) coordination of software suites; 2) increasing use of Bridges when very large memory nodes are needed; 3) establishing a shared Luster file system, allowing increased interoperability (under way). 4) beginning to build a center for metagenomic analysis centered at PSC. While offering a full metagenomics service is beyond the capacity of the current grant, we are establishing foundations. We plan to submit an ABI Development grant this fall to further this work: Blood would server as PI, with IU playing a secondary role, but it will be an NCGAS-branded service. The division of labor between the two center is still to be established.

1.2.4. *Key outcomes or other achievements*

The key outcome during the second 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 8 peer-reviewed scientific publications that have been published in 2016-2017 (beyond those reported in the first year's report).

1.2.5. Results of the 2017 user survey

We have just completed a user satisfaction survey, with 56 users responding. Much like our last survey (about 2 years ago), we found there was broad satisfaction with our services, and that NCGAS had been essential to many users' research (Fig. 2). If there were complaints, they seemed to be focused on areas where we are limited by personnel—a limitation we are aware of. In the last survey, we didn't ask to correct question, to understand what funding sources our researchers relied on. We find that NCGAS is now used primarily by NSF-funded researchers (Fig. 3. After further analysis, we will generate and make available a white paper reporting the results of this survey.

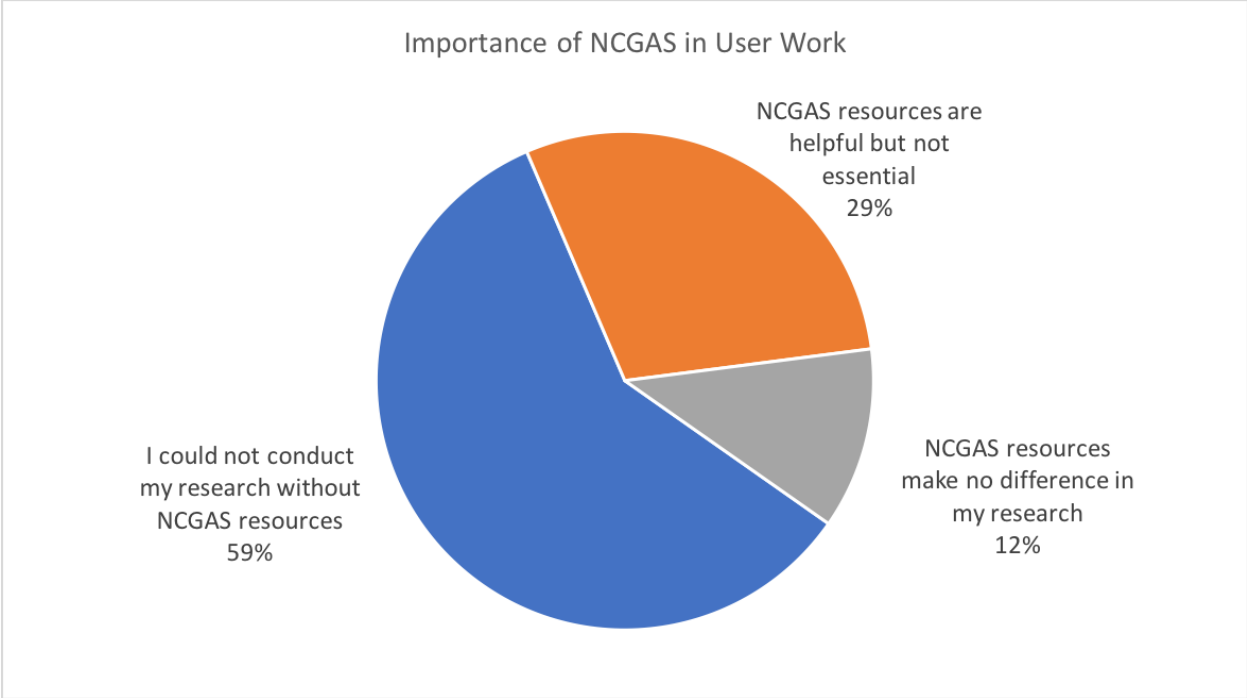


Figure 2. Usefulness

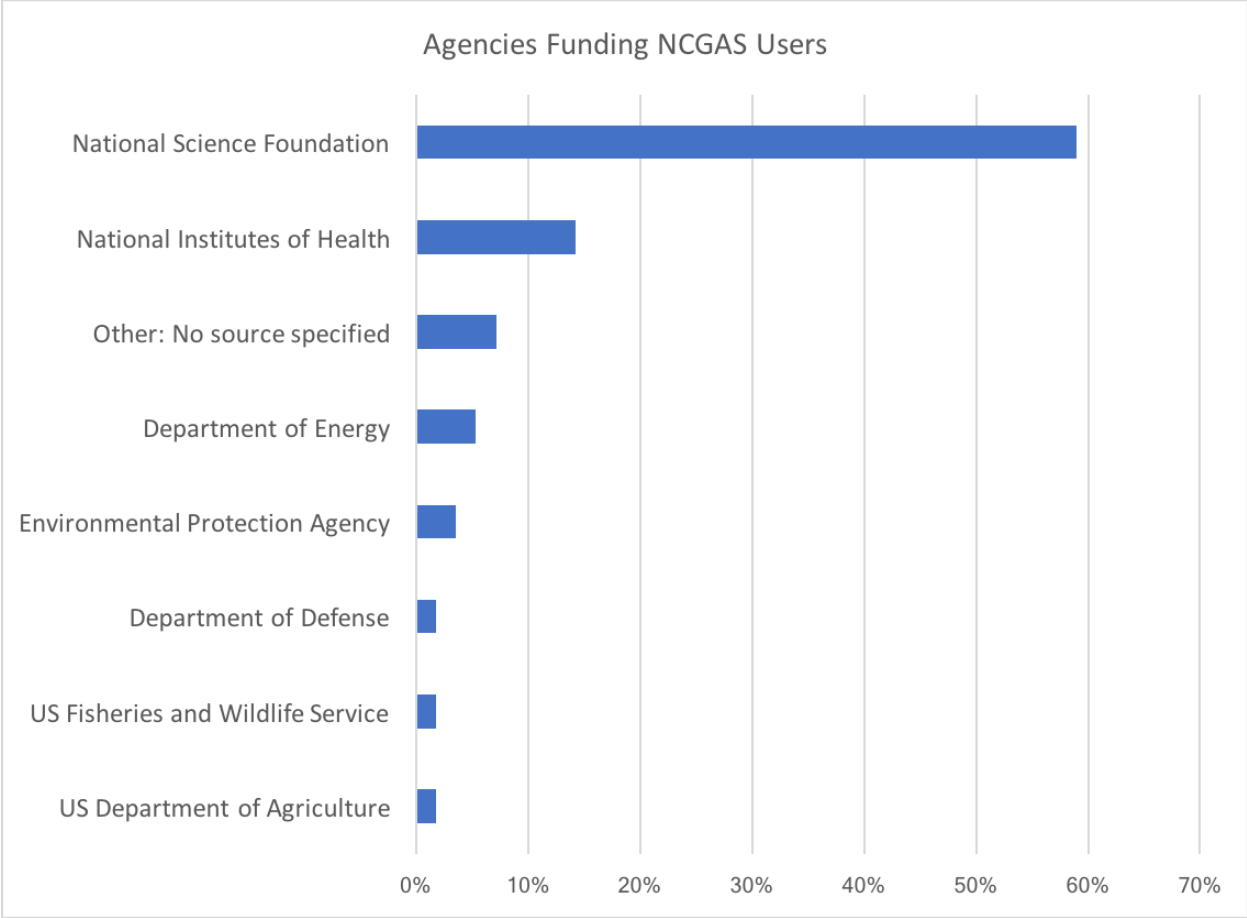


Figure 3 Funding sources reported by NCGAS to users

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 sustaining award to continue and grow NCGAS services. Dr. Doak will be the primary author on next year's sustaining renewal.

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, and is now on the Galaxy Financial Committee. Ms. Ganote oversees many projects, and mentored Sheri Sanders. She has just been promoted from IT3 to IT4, reflecting her essential role in our organization.

Staff member Sheri Sanders has now been a NCGAS team member for a year, since finishing her PhD at Notre Dame, where she used transcriptomics to characterize salamander species, some endangered. Dr. Sanders' goal in joining NCGAS was to grow her understanding of IT and HPC, and how they impacted the biological community. She now leads our efforts to support genome browsers and play a role in the GMOD development community.

Staff member Bhavya Nalagampalli Papudeshi will start as an NCGAS employee June 15th, 2017. Ms. Nalagampalli Papudeshi has just completed her master's degree in bioinformatics at SDSU, working in the lab of Liz Dinsdale. Her work includes optimization of metagenome assembly and binning tools to reconstruct population genomes, and she will strengthen our metagenomics support and our collaboration with PSC. She is already working with a collaborator of ours, Rob Edwards at SDSU, on shared metagenomics projects.

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
- NCGAS Twitter and Blog accounts
- 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 initiated several new directions in the sustaining grant's first year, which we have carried forward: 1) continue to deepen our expertise in—and offerings of—genome browsers, for searchers to organize and distribute their results; 2) aggressively pursue metagenomic projects/researchers in collaboration with Phil Blood at PSC. He has considerable experience in metagenomics research, and our newest employee Bhavya Nalagampalli Papudeshi also supports this effort; 3) continue our use of the new Jetstream environment to aid genomics researchers.

NCGAS infrastructure was inaugurated 6 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 and as an XSEDE-allocated resource, again primarily for genomics research. Mason is now antiquated, and is being replaced. This will be accomplished in three ways: 1) Mason is being replaced with the Carbonate cluster. Carbonate is considerably faster, with a higher memory-to-core ratio than Mason; Carbonate is onsite, and will open to early users June 1st. 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. 3) The NSF-funded cloud environment Jetstream: while not providing very large memory, NCGAS has already helping researchers accomplish genomics science on Jetstream (*ex.* ecological-genomics projects from AR), and we will continue to expand its uses. We have just started to use Wrangler to provide storage for Jetstream VMs.

Other goals for the year:

- Incorporate our newest hire into the team, and use this as further opportunity to expand our metagenomics expertise in collaboration with PSC.
- The IU/TACC Jetstream cloud environment opens up a range of possibilities, which we will continue to take advantage of.
- We are now actively providing genome browsers to users, starting with the GMOD-base G and JBrowsers, and will continue this.
- Having PSC and Phil Blood as a funded partner, we continue to develop our support for metagenomics/metatranscriptomics. PSC has had an emphasis in metagenomics and Blood has considerable experience working with researchers and developers. 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 Pursue Field and Marine Stations as NCGAS and Jetstream clients

In conjunction with the Jetstream development team, we have just completed a survey of field and marine station directors and managers, as represented by the membership of the Organization of Biological Field Stations (OBFS; <http://www.obfs.org/>), an association of more than 200 field stations and professionals concerned with field facilities for biological research

and education. The survey's intent was to ascertain stations' general cyberinfrastructure needs, and specifically how the Jetstream cloud could server their needs. The results of this survey are being analyzed, but an initial result is illustrated in Fig 4. After more analysis, we will generate and make available a white paper reporting the results of this survey.

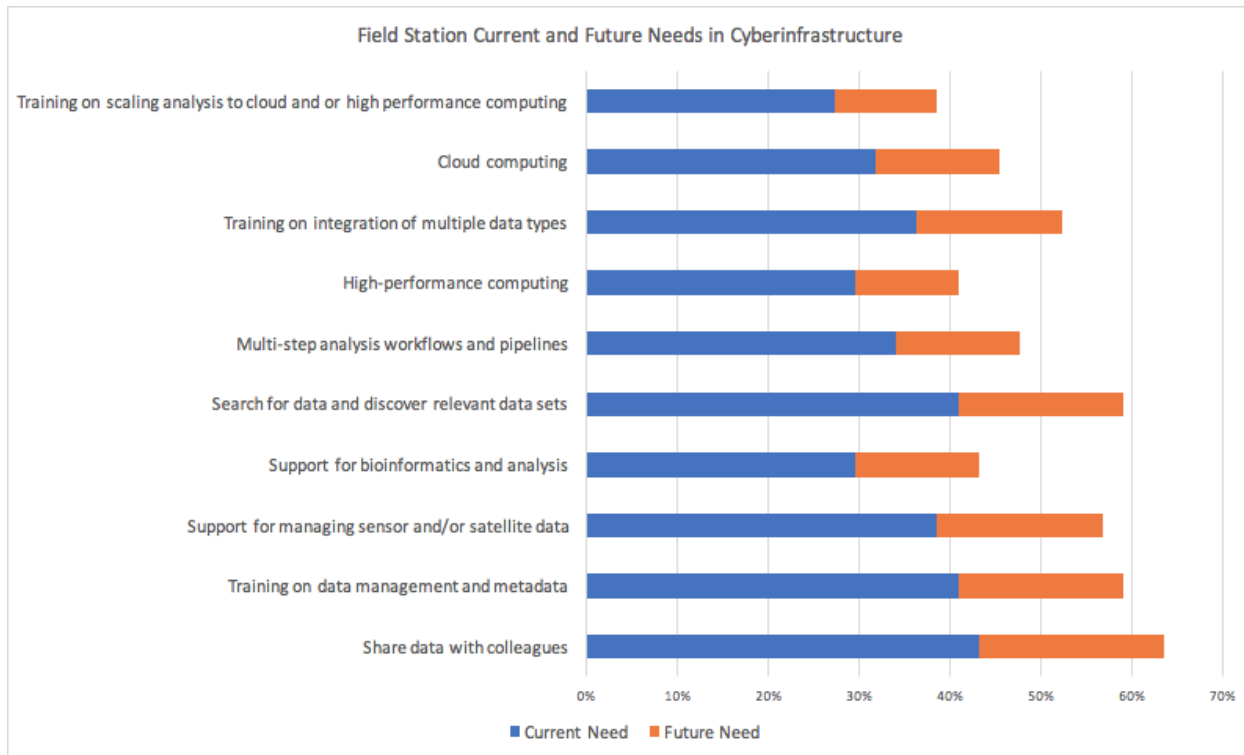


Figure 4 Cyberinfrastructure needs of field and marine stations

1.5.3 Synergistic activities

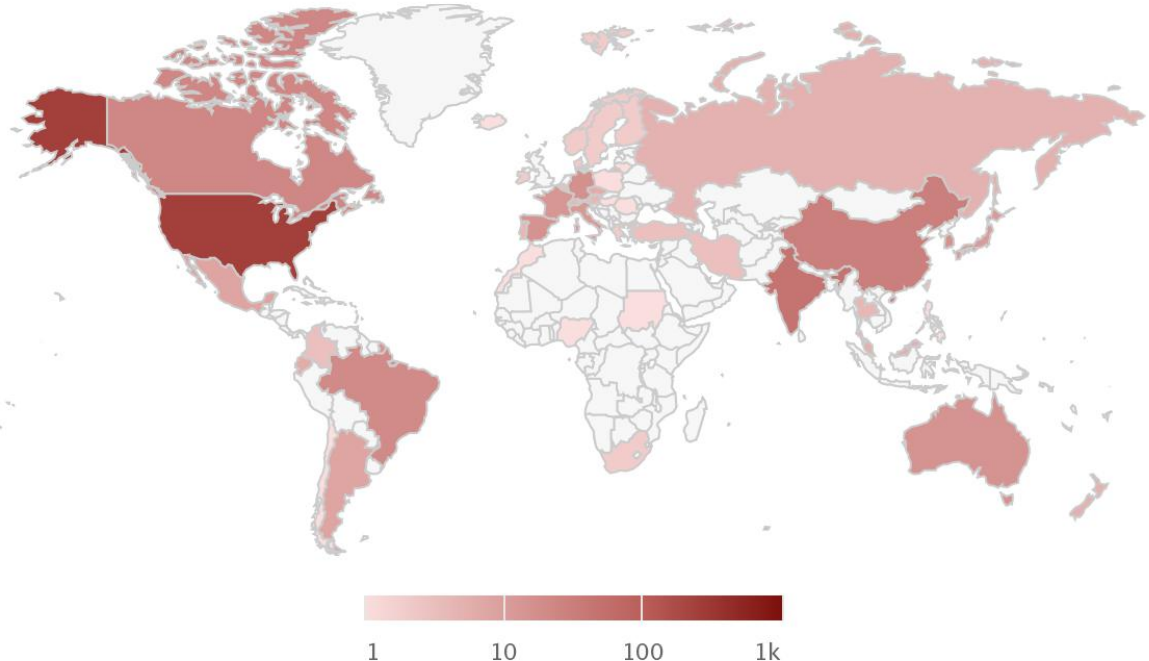
NCGAS is both a specific NSF-funded service provider in genomics, and a management group in IU's Pervasive Technology Institute. In this second guise, the NCGAS takes part in other projects that we feel augment our NSF-funded services.

- Active engagement with the Galaxy development community. NCGAS and IU hosted last year's Galaxy conference (~300 participants) and NCGAS member Carrie Ganote was on the organizing committee and a presenter, and is now on the finance committee.
- 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. We are collaborating with professor Naomi Stover at Bradley University, who maintains several ciliate browsers.
- NIH ITCR-funded Trinity development and Galaxy hosting. Involvement of NCGAS and the IU Scientific Applications and Performance Tuning group has both improved Trinity and made it far more available. 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. The IU Trinity Galaxy has 775 registered users, and gives NCGAS a global reach (Fig. 3).

Trinity Galaxy Users 2017

Use at 486 institutions in 51 countries



• **Figure 5 Users of the NCGAS-supported Trinity Galaxy instance worldwide**

- NIH/NSF/ITCR-funded GenePattern hosting. Similar to Trinity, hosting GenePattern gives us an understanding of alternative software, and makes them available to our users.
- Keithanne Mockaitis specializes in plant transcriptomics, and as such has been and is a member of many national and international consortiums characterizing commercial plants, including mango, cocoa, and loblolly pine. Current projects are coffee, peanut, and sweet potato. Both Genote and Sanders are paid from NSF and USDA grants Mockaitis is a co-PI on.

This mix of activities provides a diversified funding base to the NCGAS management group.

Figure 4. In addition to the ABI funding to support NSF researchers, the IU NCGAS management group undertakes a number of other synergistic projects, supporting by other funding sources.

1.5.4 ***We will submit our second sustaining proposal Aug. 2017***

2. Products

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

2.1.1. Journals or Juried Conference Papers

- Almada, Amalia A. and Tarrant, Ann M. and Olson, Julie (2016). {Vibrio elicits targeted transcriptional responses from copepod hosts}. *FEMS Microbiology Ecology*. 92 (6), fiw072+. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1093/femsec/fiw072
- Darris, Carl E. and Tyus, James E. and Kelley, Gary and Ropelewski, Alexander J. and Nicholas, Hugh B. and Wang, Xiaofei and Nahashon, Samuel (2015). Molecular tools to support metabolic and immune function research in the Guinea Fowl (*Numida meleagris*). *BMC Genomics*. 16 (1), . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1186/s12864-015-1520-6
- Duncan, Rebecca P. and Feng, Honglin and Nguyen, Douglas M. and Wilson, Alex C. C. (2016). Gene Family Expansions in Aphids Maintained by Endosymbiotic and Nonsymbiotic Traits. *Genome Biology and Evolution*. 8 (3), 753--764. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1093/gbe/evw020
- Feng, Ni Y. and Fergus, Daniel J. and Bass, Andrew H. (2015). Neural transcriptome reveals molecular mechanisms for temporal control of vocalization across multiple timescales. *BMC Genomics*. 16 (1), 408+. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1186/s12864-015-1577-2
- Fergus, Daniel J. and Feng, Ni Y. and Bass, Andrew H. (2015). Gene expression underlying enhanced, steroid-dependent auditory sensitivity of hair cell epithelium in a vocal fish. *BMC Genomics*. 16 (1), . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1186/s12864-015-1940-3
- Gulia-Nuss, Monika and Nuss, Andrew B. and Meyer, Jason M. and Sonenshine, Daniel E. and Roe, R. Michael and Waterhouse, Robert M. and Sattelle, David B. and de la Fuente, Jos'e and Ribeiro, Jose M. and Megy, Karine and Thimmapuram, Jyothi and Miller, Jason R. and Walenz, Brian P. and Koren, Sergey and Hostetler, Jessica B. and Thiagarajan, Mathangi and Joardar, Vinita S. and Hannick, Linda I. and Bidwell, Shelby and Hammond, Martin P. and Young, Sarah and Zeng, Qiandong and Abrudan, Jenica L. and Almeida, Francisca C. and Ayl'n, Nieves and Bhide, Ketaki and Bissinger, Brooke W. and Bonzon-Kulichenko, Elena and Buckingham, Steven D. and Caffrey, Daniel R. and Caimano, Melissa J. and Croset, Vincent and Driscoll, Timothy and Gilbert, Don and Gillespie, Joseph J. and Giraldo-Calder'n, Gloria I. and Grabowski, Jeffrey M. and Jiang, David and Khalil, Sayed M. and Kim, Donghun and Kocan, Katherine M. and Kov'ci, Juraj and Kuhn, Richard J. and Kurtti, Timothy J. and Lees, Kristin and Lang, Emma G. and Kennedy, Ryan C. and Kwon, Hyeogsun and Perera, Rushika and Qi, Yumin and Radolf, Justin D. and Sakamoto, Joyce M. and S'anchez-Gracia, Alejandro and Severo, Maiara S. and Silverman, Neal and v'Simo, Ladislav and Tojo, Marta and Tornador, Cristian and Van Zee, Janice P. and V'zquez, Jes'us and Vieira, Filipe G. and Villar, Margarita and Wespiser, Adam R. and Yang, Yunlong and Zhu, Jiwei and Arensburger, Peter and Pietrantonio, Patricia V. and Barker, Stephen C. and Shao, Renfu and Zdobnov, Evgeny M. and Hauser, Frank and Grimmelikhuijzen, Cornelis J. and Park, Yoonseong and Rozas, Julio and Benton, Richard and Pedra, Joao H. and Nelson, David R. and Unger, Maria F. and Tubio, Jose M. and Tu, Zhijian and Robertson, Hugh M. and Shumway, Martin and Sutton, Granger and Wortman, Jennifer R. and Lawson, Daniel and Wikel, Stephen K. and Nene, Vishvanath M. and Fraser, Claire M. and Collins, Frank H. and Birren, Bruce and Nelson, Karen E. and Caler, Elisabe (2016). Genomic insights into the Ixodes scapularis tick vector of Lyme disease.. *Nature communications*. 7 . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; ISSN: 2041-1723
- Horton, Melissa A. and Oliver, Randy and Newton, Irene L. (2015). No apparent correlation between honey bee forager gut microbiota and honey production.. *PeerJ*. 3 . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; ISSN: 2167-8359
- Huth, Troy J. and Place, Sean P. (2016). Transcriptome wide analyses reveal a sustained cellular stress response in the gill tissue of *Trematomus bernacchii* after acclimation to multiple stressors. *BMC Genomics*. 17 (1), . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1186/s12864-016-2454-3

- Krishnakumar, Raga and Chen, Amy F. and Pantovich, Marisol G. and Danial, Muhammad and Parchem, Ronald J. and Labosky, Patricia A. and Blleloch, Robert (2016). FOXD3 Regulates Pluripotent Stem Cell Potential by Simultaneously Initiating and Repressing Enhancer Activity. *Cell Stem Cell*. 18 (1), 104--117. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1016/j.stem.2015.10.003
- Kucukyildirim, Sibel and Long, Hongan and Sung, Way and Miller, Samuel F. and Doak, Thomas G. and Lynch, Michael (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. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; ISSN: 2160-1836
- Lee, Heewook and Doak, Thomas G. and Popodi, Ellen and Foster, Patricia L. and Tang, Haixu (2016). Insertion sequence-caused large-scale rearrangements in the genome of Escherichia coli.. *Nucleic acids research*. . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; ISSN: 1362-4962
- Newton, Irene L. and Clark, Michael E. and Kent, Bethany N. and Bordenstein, Seth R. and Qu, Jiaxin and Richards, Stephen and Kelkar, Yogeshwar D. and Werren, John H. (2016). Comparative Genomics of Two Closely Related Wolbachia with Different Reproductive Effects on Hosts.. *Genome biology and evolution*. 8 (5), 1526--1542. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1093/gbe/evw096
- Newton, Irene L. G. and Sheehan, Kathy B. (2015). Passage of Wolbachia pipientis through Mutant Drosophila melanogaster Induces Phenotypic and Genomic Changes. *Applied and Environmental Microbiology*. 81 (3), 1032--1037. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1128/aem.02987-14
- Nimkulrat, Sutichot and Lee, Heewook and Doak, Thomas G. and Ye, Yuzhen (2016). Genomic and Metagenomic Analysis of Diversity-Generating Retroelements Associated with Treponema denticola.. *Frontiers in microbiology*. 7 . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; ISSN: 1664-302X
- Orsini, Luisa and Gilbert, Donald and Podicheti, Ram and Jansen, Mieke and Brown, James B. and Solari, Omid S. and Spanier, Katina I. and Colbourne, John K. and Rush, Douglas and Decaestecker, Ellen and Asselman, Jana and De Schampelaere, Karel A. C. and Ebert, Dieter and Haag, Christoph R. and Kvist, Jouni and Laforsch, Christian and Petrusek, Adam and Beckerman, Andrew P. and Little, Tom J. and Chaturvedi, Anurag and Pfrender, Michael E. and De Meester, Luc and Frilander, Mikko J. (2016). Daphnia magna transcriptome by RNA-Seq across 12 environmental stressors. *Scientific Data*. 3 160030+. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1038/sdata.2016.30
- Raborn, R. Taylor and Spitze, Ken and Brendel, Volker P. and Lynch, Michael (2016). Promoter architecture and sex-specific gene expression in the microcrustacean Daphnia pulex revealed by large-scale profiling of 5'-mRNA ends. *bioRxiv*. 047894+. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1101/047894
- Rokop, Z. P. and 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. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1128/aem.01259-15
- Roncalli, Vittoria and Cieslak, Matthew C. and Lenz, Petra H. (2016). Transcriptomic responses of the calanoid copepod Calanus finmarchicus to the saxitoxin producing dinoflagellate Alexandrium fundyense.. *Scientific reports*. 6 . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1038/srep25708
- Sousounis, Konstantinos and Qi, Feng and Yadav, Manisha C. and Millⁿ, Jos^e L. and Toyama, Fubito and Chiba, Chikafumi and Eguchi, Yukiko and Eguchi, Goro and Tsonis, Panagiotis A. (2015). A robust transcriptional program in newts undergoing multiple events of lens regeneration throughout their lifespan. *eLife*. 4 . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.7554/elife.09594
- Suzuki, Haruo and Dapper, Amy L. and Jackson, Craig E. and Lee, Heewook and Pejaver, Vikas and Doak, Thomas G. and Lynch, Michael and Preer, John R. (2015). Draft Genome Sequence of Caedibacter varicaedens, a Kappa Killer Endosymbiont Bacterium of the Ciliate Paramecium biaurelia.. *Genome*

announcements. 3 (6), . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; ISSN: 2169-8287

- Tarpy, David R. and Mattila, Heather R. and Newton, Irene L. G. (2015). Development of the Honey Bee Gut Microbiome throughout the Queen-Rearing Process. *Applied and Environmental Microbiology*. 81 (9), 3182--3191. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1128/aem.00307-15
- Wang, Mingjie and Doak, Thomas G. and Ye, Yuzhen (2015). Subtractive assembly for comparative metagenomics, and its application to type 2 diabetes metagenomes.. *Genome biology*. 16 . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; ISSN: 1474-760X

2.1.2. Licenses

2.1.3. Other Conference Presentations / Papers

- Sanders, S., C. Ganote, T. Doak (2016). *Cluster Quick Guide*. Mount Desert Island Biological Laboratory. Bar Harbor, ME. Status = PUBLISHED; Acknowledgement of Federal Support = Yes
- Sanders, S., C. Ganote, T. Doak (2016). *Data movement and management*. Mount Desert Island Biological Laboratory. Bar Harbor, ME. Status = PUBLISHED; Acknowledgement of Federal Support = Yes
- Sanders, S., C. Ganote, T. Doak (2016). *Moving forward in bioinformatics*. Bethune-Cookman University. Daytona, FL. Status = PUBLISHED; Acknowledgement of Federal Support = Yes
- Sanders, S., C. Ganote, T. Doak (2016). *Moving off GUIs: A guide to what's next*. Bioinformatics 2 Go course, Indiana University. Bloomington, IN. Status = PUBLISHED; Acknowledgement of Federal Support = Yes
- Seetharam, Arun and Gomez, Antonio and Purcell, Catherine M. and Hyde, John R. and Blood, Philip D. and Severin, Andrew J. (2015). *NCBI-BLAST Programs Optimization on XSEDE Resources for Sustainable Aquaculture*. Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure. New York, NY, USA. Status = PUBLISHED; Acknowledgement of Federal Support = Yes
- Doak, C. Ganote, S. Sanders, T., P. Blood. (2016). *NCGAS: Providing National Cyberinfrastructure to Biologists, esp. Genomicists*. 2017 IU Research Technologies Kickoff. Bloomington, IN. Status = PUBLISHED; Acknowledgement of Federal Support = Yes
- Ganote, C., T. Doak, S. Sanders, C. Ganote, T. Doak (2016). *NCGAS: Providing National Cyberinfrastructure to Biologists, with a Focus on Genomics*. Presented at Great Plains Network Monthly Webinar. Online. Status = PUBLISHED; Acknowledgement of Federal Support = Yes
- Doak, T., C. Ganote, S. Sanders, P. Blood. (2016). *NCGAS: Providing National Cyberinfrastructure to Biologists, esp. Genomicists*. Monthly PI Meeting of the NIH NCI ITCR (Information Technologies in Cancer Research Group). Online. Status = PUBLISHED; Acknowledgement of Federal Support = Yes
- Sanders, S., C. Ganote, T. Doak (2016). *Navigating high performance computing (HPC) resources*. Guest Lecture at Bethune-Cookman University. Daytona, FL. Status = PUBLISHED; Acknowledgement of Federal Support = Yes
- Blood, P. D. (2017). *Supercomputing resources for open, accessible, reproducible genomic analysis*. Joint CAMI-M3 workshop on metagenomic software validation. College Park, MD. Status = PUBLISHED; Acknowledgement of Federal Support = Yes
- Ganote, C., T. Doak, S. Sanders, C. Ganote, T. Doak. (2017). *The National Center for Genome Analysis Support*. Great Lakes Bioinformatics Conference. Chicago, IL. Status = PUBLISHED; Acknowledgement of Federal Support = Yes
- Sanders, S., C. Ganote, T. Doak (2016). *The white coats are coming: The growth, success, and future of computing in biology*. Supercomputing 2016. Salt Lake, UT. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

2.1.4. Other Products

- *Blog Post*.

Blog post available at http://ncgas.org/Blog_Posts/Third%20Generation%20Sequencing.php

2.1.5. Other Publications

- Doak, T., C. Ganote, S. Sanders, P. Blood. (2016). *NCGAS: Providing National Cyberinfrastructure to Biologists, esp. Genomicists*. Research Meeting of the Michael Lynch Research Group. Status = PUBLISHED; Acknowledgement of Federal Support = Yes
-

3. Participants

3.1. Individuals

Table 1. Individuals that have worked on the project

Name	Most Senior Project Role	Nearest Person Month Worked
<u>Doak, Thomas</u>	PhD/PI	8
<u>Stewart, Craig</u>	<u>PhD/co-PI</u>	<u>1</u>
<u>Michaels, Scott</u>	<u>PhD/co-PI</u>	<u>1</u>
<u>Henschel, Robert</u>	PhD, director, management group	1
<u>Blood, Phillip</u>	PhD/co-PI (collaborative grant)	3
<u>Miller, Therese</u>	Other Professional	1
<u>Nalagampalli Papudeshi , Bhavya</u>	<u>Staff Scientist, Masters</u>	<u>0*</u>
<u>Ganote, Carrie</u>	Staff Scientist, PhD	6
<u>Sanders, Sheri</u>	Staff Scientist, PhD	6

*Nalagampalli Papudeshi starts employment June 2017

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

Thomas Doak

Email: tdoak@iu.edu

Most Senior Project Role: PI (doctoral level)

Nearest Person Month Worked: 8

Contribution to the Project: PI and operational management

Funding Support: NSF, NIH, Indiana University

International Collaboration: Yes: Italy, Germany, Japan

International Travel: No

Craig A Stewart

Email: stewart@iu.edu

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

Scott Michaels

Email: michaels@indiana.edu
Most Senior Project Role: PhD/co-PI
Nearest Person Month Worked: 1
Contribution to the Project: Funded PSC collaborator
Funding Support: Co-PI responsible for oversight
International Collaboration: No
International Travel: No

Phillip Blood

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

Robert Henschel

Email: henschel@iu.edu
Most Senior Project Role: Other Professional
Nearest Person Month Worked: 1
Contribution to the Project: Director over NCGAS management group and software optimization
Funding Support: Indiana University, NIH
International Collaboration: Yes
International Travel: Yes

Therese Miller

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

Carrie Ganote

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

Bhavya Nalagampalli Papudeshi

Email: cgannot@iu.edu

Most Senior Project Role: Staff Scientist (Masters level)

Nearest Person Month Worked: 6

Contribution to the Project: bioinformatician consultant / programmer

Funding Support: NSF

International Collaboration: No

International Travel: nNO

Sheri Sanders

Email: ss93@iu.edu

Most Senior Project Role: Staff Scientist (doctoral level)

Nearest Person Month Worked: 6

Contribution to the Project: bioinformatician consultant / programmer

Funding Support: NSF

International Collaboration: No

International Travel: No

3.2. Partner organizations

Table 2. Partner organizations

Name	Type of Partner Organization	Location
<u>Pittsburgh Supercomputing Center, Carnegie Mellon University</u>	Academic Institution	Pittsburgh, PA
<u>Texas Advanced Computing Center, University of Texas</u>	Academic Institution	Austin, TX
<u>XSEDE</u>	Other Nonprofits	United States
<u>Arkansas High Performance Computing Center, University of Arkansas, Fayetteville</u>	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 Bridges supercomputers in support of NCGAS 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 and Mason. Pummel is 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 and Wrangler grants, as well as center for CyVerse, which provides many opportunities for collaboration. It has provided facilities, computer time, and storage space in support of NCGAS 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 and an XSEDE Domain Champion.

3.3. *Have other collaborators or contacts been involved?*

No

4. Impact

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

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 “domain-specific scientific service center”, independent of federally-funded cyberinfrastructure computational resources; we have decoupled federal funding for supercomputers and funding for supercomputer application support. This ensures that a

community relatively new to supercomputers use —biology for example—has support funded by the BIO directorate of NSF and is attuned to the needs of current research in the field.

NCGAS has just become an XSEDE Domain Champion, a new category supplementing Campus Champions, who are domain agnostic. As a very recent XSEDE program, we will see how Campus Champions works out.

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, an important product of ‘omic research.

4.7. *What is the impact on technology transfer?*

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

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

The societal impact of genomic characterization is gradual, but 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 (in nearly every field of medicine, including cancer), and swept through popular media, is amazing.

5. Changes/ Problems

5.1. Changes in approach and reasons for change

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 has helped to fill out the team, 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. Appendices

Appendix 1. List of NSF Funded Projects Using NCGAS Resources

Joseph Vitti, Harvard University, Broad Institute

8/3/16

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

Raphael D. Isokpehi, Bethune Cookman University

8/26/16

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

Predicting Microbial Genome-Encoded Biomolecular Networks. According to the Integrated Microbial Genomes database (<http://img.jgi.doe.gov/>): "At the start of 2015, IMG had a total of

32,802 genome datasets from all domains of life and 5,234 metagenome datasets, out of which 27,341 genome datasets and 3,193 metagenome datasets are publicly available."

Functional and structural annotations for genes in microbial genomes are increasingly available as multivariate data sets in formats suitable for a variety of cognitive activities including knowledge discovery, sense making, problem-solving and planning future research.

The exponential increase in whole genome sequences of bacteria and archaea presents a source of large and complex data on functional and structural annotations of genes. The annotations for function and transcriptional direction of genes adjacent to a gene locus in genomes of bacteria and archaea can be informative on biological process that involve the gene.

Therefore there is a critical need to index of Microbial Gene Loci based on Transcriptional Direction of Adjacent Genes to facilitate cognitive activities including knowledge discovery and planning future research. The NCGAS resources will be used for performing diverse actions on the data sets including comparative analysis. -†

Thomas Hahn, University of Arkansas
8/29/16

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

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

Wendi Li, USDA
8/29/16

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

Representing the functional part of the genome, transcriptome is known to have distinct profiles specific to cell type, developmental state and disease status. Recently developed whole-transcriptome RNA-seq provides a tremendous potential to identify transcriptome biomarkers that are directly relevant to phenotypic traits. Using a whole-transcriptome RNA-seq technology, we propose to perform a systematic transcriptomics study on various tissue types in dairy cattle under a wide variety of conditions. Targeted phenotypes include animal's response to diets with varied protein and carbohydrate concentrations, milk protein and fatty acid concentration, and lactating cow's disease state. This study will allow us to build a comprehensive knowledge base of dairy cattle transcriptomics. Such knowledge will most likely facilitate the identification of new and reliable biomarkers, which can be applied for further improvement in dairy cattle feed management and quality of milk.

Jill Wegrzyn, University of Connecticut
12/6/16

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

William Wesley, Loyola Marymount University
1/9/17

We are investigating the causes and consequences of individual-level physiological variation in mussels inhabiting the spatially and temporally variable rocky intertidal zone. We have completed a fairly large Illumina RNAseq run for a comparative physiology project (PE 150 on 12 individuals to build reference transcriptome + PE 50 on 49 individuals to look at gene

expression variation), and now need to de novo assemble the reference transcriptome and look at individual variation in gene expression. My institution has no resources for this sort of computational analysis.

Marc Vermulst, University of Pennsylvania

2/22/17

The genome provides a precise biological blueprint of life. To implement this blueprint correctly, the genome must be read with great precision; however, due to the constraints of biological fidelity, it is impossible for this process to be completely error-free. As a result, transcription errors can occur at any time, in any transcript, and how these random errors affect cellular health is completely unknown. To fill this gap in our knowledge, we recently monitored yeast cells that were genetically engineered to display error-prone transcription. We discovered that transcription errors give rise to misfolded proteins that induce proteotoxic stress. Thus, transcription errors represent a new molecular mechanism by which cells can acquire disease. As a result, it will be important to learn more about the mechanisms that induce or suppress transcription errors, because these mechanisms could either delay or accelerate the progression of proteotoxic diseases. To this end, we developed the first next-gen sequencing assay that is capable of measuring the fidelity of transcription in a genome-wide fashion. We now propose to use this technology on yeast and mice to identify the parameters that control the fidelity of transcription in eukaryotic cells. For our experiments, we will generate large datasets to fully characterize the transcription error spectrum caused by RNA polymerase under various conditions and in response to various DNA damaging compounds, which we expect will significantly increase the transcription error rate in treated cells. We will therefore need NCGAS resources to properly store and access our data sets in the future.

Christopher Chandler, SUNY Oswego

4/7/17

Allosomes are chromosomes that determine sex, like the X and Y chromosomes in mammals, and they play crucial roles in the evolutionary biology of species. They have evolved independently in a wide array of species, and while these chromosomes in different groups share many similarities, they also exhibit important differences. Explaining these differences, however, remains difficult. This project will address this problem by examining allosomes in terrestrial isopod crustaceans, an ideal study system because they exhibit considerable variation in sex-determining mechanisms. However, surprisingly, their genomes have received little attention. The proposed experiments are designed to help explain why these chromosomes are so unique, and how they contribute to vital biological processes. Specifically, this research will (i) identify where changes in sex-determining chromosomes have occurred on the isopod evolutionary tree; (ii) test whether genes on these chromosomes are affected more strongly than autosomal genes by natural selection and genetic drift; and (iii) test whether these species exhibit dosage compensation. This work will also generate a large volume of genome sequencing data, providing some of the first draft genome assemblies for these 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.

Elze Rackaityte, University of California, San Francisco
4/13/17

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

Douglas R. Cook, University of California, Davis
4/17/17

Legume species are key components of both natural and agricultural ecosystems, and for human nutrition. Their importance derives in large part from their capacity for symbiotic nitrogen fixation with soil bacteria, enabling them to return vital nitrogen to the soil environment and to create seed and forage of high protein content. Two decades of molecular and genomic studies in model systems have revealed the presence of exquisite genetic pathways that initiate symbiosis, but despite these advances we have essentially no understanding of genes that regulate symbiotic performance in the natural environment. -†

Our research aims to understand the evolution an important legume crop species - chickpea - *Cicer arietinum*, by elucidating the changes in its capacity for symbiotic association with Rhizobial strains, and resistance to pathogens such as *Fusarium*, compared in to its wild progenitors - *C. reticulatum* and *C. echinospermum*. Using a combination of ecology, population genomics, classical molecular genetics and functional assays, we are poised to explain how human selection has reshaped these and other biological processes during domestication. This study of gene function in natural versus human-built environments and its outcomes will have relevance to both basic science and agriculture. Data and biological resources generated under this project will be available through public repositories, including the NCBI and USDA-ARS's GRIN. Training and mentoring of under-represented minorities, and students at various

academic levels - high school, undergraduate and graduate students geared towards their professional development.

Appendix 2. IU Publications by NCGAS Users

Citations Crediting NCGAS as of 05/2017

Almada, A. A., Tarrant, A. M., & Olson, J. (2016). *Vibrio* elicits targeted transcriptional responses from copepod hosts. *FEMS Microbiology Ecology*, 92(6), fiw072+.

URL <http://dx.doi.org/10.1093/femsec/fiw072>

Darris, C. E., Tyus, J. E., Kelley, G., Ropelewski, A. J., Nicholas, H. B., Wang, X., & Nahashon, S. (2015). Molecular tools to support metabolic and immune function research in the guinea fowl (*Numida meleagris*). *BMC Genomics*, 16(1).

URL <http://dx.doi.org/10.1186/s12864-015-1520-6>

Duncan, R. P., Feng, H., Nguyen, D. M., & Wilson, A. C. C. (2016). Gene family expansions in aphids maintained by endosymbiotic and nonsymbiotic traits. *Genome Biology and Evolution*, 8(3), 753-764.

URL <http://dx.doi.org/10.1093/gbe/evw020>

Feng, N. Y., Fergus, D. J., & Bass, A. H. (2015). Neural transcriptome reveals molecular mechanisms for temporal control of vocalization across multiple timescales. *BMC Genomics*, 16(1), 408+.

URL <http://dx.doi.org/10.1186/s12864-015-1577-2>

Fergus, D. J., Feng, N. Y., & Bass, A. H. (2015). Gene expression underlying enhanced, steroid-dependent auditory sensitivity of hair cell epithelium in a vocal fish. *BMC Genomics*, 16(1).

URL <http://dx.doi.org/10.1186/s12864-015-1940-3>

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., Ayllón, N., Bhidé, 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-Calderón, G. I., Grabowski, J. M., Jiang, D., Khalil, S. M., Kim, D., Kocan, K. M., Kocak, 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., Sánchez-Gracia, A.,

- Severo, M. S., Silverman, N., SÃäimo, L., Tojo, M., Tornador, C., Van Zee, J. P., VaÃÄzquez, 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., & Hill, C. A. (2016). Genomic insights into the ixodes scapularis tick vector of lyme disease. *Nature communications*, 7. URL <http://view.ncbi.nlm.nih.gov/pubmed/26856261>
- Horton, M. A., Oliver, R., & Newton, I. L. (2015). No apparent correlation between honey bee forager gut microbiota and honey production. *PeerJ*, 3. URL <http://view.ncbi.nlm.nih.gov/pubmed/26623177>
- Huth, T. J., & Place, S. P. (2016). Transcriptome wide analyses reveal a sustained cellular stress response in the gill tissue of trematomus bernacchii after acclimation to multiple stressors. *BMC Genomics*, 17(1). URL <http://dx.doi.org/10.1186/s12864-016-2454-3>
- Krishnakumar, R., Chen, A. F., Pantovich, M. G., Danial, M., Parchem, R. J., Labosky, P. A., & Billech, R. (2016). FOXD3 regulates pluripotent stem cell potential by simultaneously initiating and repressing enhancer activity. *Cell Stem Cell*, 18(1), 104-117. URL <http://dx.doi.org/10.1016/j.stem.2015.10.003>
- Newton, I. L., Clark, M. E., Kent, B. N., Bordenstein, S. R., Qu, J., Richards, S., Kelkar, Y. D., & 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. URL <http://dx.doi.org/10.1093/gbe/evw096>
- Newton, I. L. G., & 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. URL <http://dx.doi.org/10.1128/aem.02987-14>
- 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 Schampelaere, K. A. C., Ebert, D., Haag, C. R., Kvist, J., Laforsch, C., Petrussek, A., Beckerman, A. P., Little, T. J., Chaturvedi, A., Pfrender, M. E., De Meester, L., & Frilander, M. J. (2016). *Daphnia magna* transcriptome by RNA-seq across 12 environmental stressors. *Scientific Data*, 3, 160030+.

URL <http://dx.doi.org/10.1038/sdata.2016.30>

Raborn, R. T., Spitze, K., Brendel, V. P., & Lynch, M. (2016). Promoter architecture and sex-specific gene expression in the microcrustacean *daphnia pulex* revealed by large-scale profiling of 5'Äö-mRNA ends. *bioRxiv*, (pp. 047894+).
URL <http://dx.doi.org/10.1101/047894>

Rokop, Z. P., Horton, M. A., & Newton, I. L. G. (2015). Interactions between cooccurring lactic acid bacteria in honey bee hives. *Applied and Environmental Microbiology*, 81(20), 7261-Äi7270.
URL <http://dx.doi.org/10.1128/aem.01259-15>

Roncalli, V., Cieslak, M. C., & Lenz, P. H. (2016). Transcriptomic responses of the calanoid copepod *calanus finmarchicus* to the saxitoxin producing dinoflagellate *alexandrium fundyense*. *Scientific reports*, 6.
URL <http://dx.doi.org/10.1038/srep25708>

Sousounis, K., Qi, F., Yadav, M. C., MillaÄÄn, J. L., Toyama, F., Chiba, C., Eguchi, Y., Eguchi, G., & Tsonis, P. A. (2015). A robust transcriptional program in newts undergoing multiple events of lens regeneration throughout their lifespan. *eLife*, 4.
URL <http://dx.doi.org/10.7554/elife.09594>

Tarpy, D. R., Mattila, H. R., & Newton, I. L. G. (2015). Development of the honey bee gut microbiome throughout the Queen-Rearing process. *Applied and Environmental Microbiology*, 81(9), 3182-Äi3191.
URL <http://dx.doi.org/10.1128/aem.00307-15>

Appendix 3. Software Supported by NCGAS

The National Center for Genome Analysis Support ([NCGAS](#)) provides support for the following genome analysis software packages available on Indiana University's [Mason](#) cluster. Access to NCGAS computational and consulting services is awarded through an allocation process to genomics research projects funded by the National Science Foundation ([NSF](#)). For more, see the [National Center for Genome Analysis Support 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
1.3.6-openmpi	x	
1.5.1-openmpi	x	
1.5.2-openmpi	x	
2.0.2	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	

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	

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	x
2.26.0	x	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

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
----------------------	-------	-------

Supported version(s)	Mason	Karst
2.12		x
3.1	x	x
3.3	x	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
0.12.5	x	
0.16.3	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	

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
1.1.2	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
0.23		x
0.27*	x	
0.30	x	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
1.3.6		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
0.7.15		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	

Supported version(s)	Mason	Karst
3.0	x	

Canu

Canu is a fork of the Celera Assembler designed for high-noise single-molecule sequencing (such as PacBio RSII or Oxford Nanopore MinION):

- Documentation: [Canu manual](#)
- Download: [Canu @ github](#)
- License: [GPLv2](#)

Supported version(s)	Mason	Karst
1.3	x	x
1.4		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
7.0	x	
8.3rc2	x	

Circos

Circos is a package designed to create circular graphics for genomic and other data:

- Documentation: [Circos Webpage](#)
- Download: [Download options](#)
- License: [GPLv3](#)

Supported version(s)	Mason	Karst
0.69-3	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
1.11	x	x
1.12	x	

Supported version(s)	Mason	Karst
1.2.1	x	
1.7.1	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	

DESeq2

Estimate variance-mean dependence in count data from high-throughput sequencing assays and test for differential expression based on a model using the negative binomial distribution:

- Documentation:
- Download:
- License:

Supported version(s)	Mason	Karst
3.0	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	

EdgeR

Differential expression analysis of RNA_seq expression profiles with biological replication.

- Documentation: [EdgeR at bioconductor](#)
- Download: [Source code](#)

- License: [GPLv2](#)

Supported version(s)	Mason	Karst
3.0		x

EGGLIB

EggLib is a C++/Python library and program package for evolutionary genetics and genomics.

- Documentation: [eggLib homepage](#)
- Download: [Source code](#)
- Reference: [eggLib Paper](#)
- License: [GPLv3](#)

Supported version(s)	Mason	Karst
2.1.3	x	

Eigensoft

The EIGENSTRAT method uses principal components analysis to explicitly model ancestry differences between cases and controls along continuous axes of variation; the resulting correction is specific to a candidate marker's variation in frequency across ancestral populations, minimizing spurious associations while maximizing power to detect true associations. The EIGENSOFT package has a built-in plotting script and supports multiple file formats and quantitative phenotypes.

- Documentation: [EIG @ github](#)
- Download: [Releases](#)
- License: [Broad and Harvard Medical School Copyright](#)

Supported version(s)	Mason	Karst
6.1.3		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

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
0.11.5	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

Flexbar

Flexbar preprocesses high-throughput sequencing data efficiently. It demultiplexes barcoded runs and removes adapter sequences. Moreover, trimming and filtering features are provided. Flexbar increases read mapping rates and improves genome and transcriptome assemblies. It supports next-generation sequencing data in fasta/q and csfasta/q format from Illumina, Roche 454, and the SOLiD platform:

- Documentation:
- Download:
- License:

Supported version(s)	Mason	Karst
----------------------	-------	-------

Supported version(s)	Mason	Karst
2.4		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
3.0	x	
2.0	x	
1.0	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	
3.7	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

HAMSTR

HaMStR is a profile hidden Markov model based tool for a directed ortholog search in EST or protein sequence data. The program takes a pre-defined core group of orthologous sequences (core orthologs) and a set of sequences from a search taxon as input. HaMStR then combines in a two-step strategy a pHMM based search and a reverse search via BLAST to extend the core ortholog group with novel sequences from the search taxon:

- Documentation: [HAMSTR Sourceforge site](#)
- Download: [HAMSTR versions](#)
- License: [GPLv3](#)

Supported version(s)	Mason	Karst
13.2.3		x

HISAT2

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
0.1.6-beta		x
2.0.4	x	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	x
3.1b2	x	

IMPUTE2

IMPUTE is a program for estimating (imputing) unobserved genotypes in SNP association studies. The program is designed to work seamlessly with the output of the genotype calling program CHIAMO and the population genetic simulator HAPGEN, and it produces output that can be analyzed using the program SNPTEST:

- Documentation: [IMPUTE2 project site](#)
- Download: [IMPUTE2 versions](#)
- License: [GPLv3](#)

Supported version(s)	Mason	Karst
2.2.2		x

Kallisto

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

- Documentation: [Kallisto @ github](#)
- Tutorial: [Getting Started](#)
- Download: [Kallisto versions](#)
- License: non-commercial license

Supported version(s)	Mason	Karst
0.42.3	x	
0.43.0	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	

LDhat

LDhat is a package of programs for the analysis of recombination from population genetic data. The key feature of the package is the estimation of population recombination rates using the composite likelihood method:

- Documentation: [LDhat @ github](#)
- Documentation: [LDhat manual](#)
- License: [GPLv3](#)

Supported version(s)	Mason	Karst
2.2		x

LDhot

LDhot is a package for inferring the location of recombination hotspots from patterns of linkage disequilibrium within samples of populations genetic data. Built to support OpenMP:

- Documentation: [LDhot @ github](#)
- License: [GPLv3](#)

Supported version(s)	Mason	Karst
2014		x

LIMMA

Data analysis, linear models and differential expression for expression data:

- Documentation: [limma @ bioconductor](#)
- Download: [Source code](#)
- License: [GPLv2](#)

Supported version(s)	Mason	Karst
3.0		x

lincRNA

ICQ-lincRNA (Identification, Characterization, and Quantification of Long Intergenic Non-Coding RNAs), offers an end-to-end solution to identify and annotate expressed lincRNAs in next generation RNA sequencing data.:

- Documentation: [lincRNA project page](#)
- Download: [lincRNA versions](#)

Supported version(s)	Mason	Karst
1.0		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	
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	
3.2.1	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

MIMAR

MIMAR is a Markov chain Monte Carlo method to estimate parameters of an isolation-migration model. It uses summaries of polymorphism data at multiple loci surveyed in a pair of diverging populations or closely related species and in contrast to previous methods, allows for intralocus recombination.:

- Documentation: [MIMAR project page](#)
- Download: [Source code](#)

Supported version(s)	Mason	Karst
12172010		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	

miRho

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

- Documentation: [miRho project page](#)
- Documentation: [miRho - 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 \(miRHO 2.8.tgz\)](#)
- 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
1.32.1		
1.34.2	x	x
1.36.1	x	
1.38.1	x	
1.39.0	x	
mpi/1.31.2		x
mpi/1.32.1		
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: [MrBayes Manual](#)
- Download: [Releases](#)
- License: [GPLv2](#)

Supported version(s)	Mason	Karst
mpi/3.2.1	x	x
serial/3.2.1	x	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: [Muscle Manual](#)

- Download: [MUSCLE download](#)

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
2.2.31		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
0.5.9		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	
1.2.2	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	
2.8.1	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
1.80		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	
1.3.0	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 Manual](#)
- Download: [Salmon @ github](#)
- License: [GPLv3](#)

Supported version(s)	Mason	Karst
gnu/0.4.2		x
0.8.2	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
1.3.1*	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	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
1.03	x	
1.05	x	
R240	x	X

SOAPdenovo-Trans

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

- 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	x
3.9.0	x	x
3.10.0	x	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
2.8.1	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	
1.44	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
2.5.2b	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	
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	
1.5.5	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	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
07/17/14	x	
2.0.6	x	
2.1.1		x
2.2.0*	x	
2.4.0	x	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
1.2.08-k111-openmp	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.