National Center for Genome Analysis Program Year 3 Report – September 15, 2013–September 14, 2014

William K. Barnett, Ph.D.
Thomas G. Doak, Ph.D.
Craig A. Stewart, Ph.D.

Indiana University

PTI Technical Report PTI-TR14-005

Revision: October 17, 2014

Citation:

Barnett, W.K., Doak, T.G., and Stewart, C.A. "National Center for Genome Analysis Program Year 3 Report – September 15, 2013 – September 14, 2014" Indiana University, Bloomington, IN. PTI Technical Report PTI-TR14-005, 2014. Available at http://hdl.handle.net/2022/18513

The facilities supported by the Research Technologies division at Indiana University are supported by a number of grants. The authors would like to acknowledge that the National Center for Genome Analysis Support is funded primarily by NSF 1062432, but our work would not be possible without the generous support of the following awards received by our parent organization, the Pervasive Technology Institute at Indiana University.

- The Indiana University Pervasive Technology Institute was supported in part by two grants from the Lilly Endowment, Inc.
- NCGAS has also been supported directly by the Indiana METACyt Initiative. The Indiana METACyt Initiative of Indiana University is supported in part by the Lilly Endowment, Inc.
- This material is based in part upon work supported by the National Science Foundation under Grant No. CNS-0521433.

Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation (NSF).

Table of Contents

Table of Tables

Table of Figures

1. Executive Summary

On September 15, 2011, Indiana University (IU) received a grant award from the National Science Foundation, through the Advances in Biological Infrastructure, to establish the National Center for Genome Analysis Support (NCGAS). This technical report describes the activities of the third 12 months of NCGAS, from September 15, 2013 through September 14, 2014.

The mission of the NCGAS is to enable the biological research community of the US to analyze, understand, and make use of the vast amount of genomic information now available. NCGAS focuses particularly on transcriptome- and genome-level assembly, phylogenetics, metagenomics/transcriptomics and community genomics.

NCGAS addresses critical problems genomics researchers face today. Next-generation sequencers generate much more data, straining laboratory and departmental cyberinfrastructure. As well, understanding the most useful software and interpreting resulting data requires special expertise. Arriving at biologically relevant conclusions often entails analytical processes that are highly detailed, complex, and fraught with technical challenges. These factors have erected technical and expertise barriers to conducting genomics research. NCGAS was created to overcome these barriers.

NCGAS provides services to biologists in the following areas:

- Consulting for biologists undertaking genome analysis, including research design, assistance creating and executing workflows, file transfer and transformation, and data interpretation.
- Optimization, hardening, and enhancement of genome analysis software and support of the use of that software, including popular assembly software and the Galaxy web-based workflow composer.
- Providing support to biologists' delivery and assistance in use of supercomputer clusters for genome analysis, including many supercomputers and supercomputer clusters that are XD (eXtreme Digital) national resources provided through XSEDE (the eXtreme Science and Engineering Discovery Environment), and high throughput computing resources delivered through the Open Science Grid (OSG). In particular, we support software on and use of:
    - Mason, a large memory supercomputer cluster at Indiana University (IU)
    - Stampede – the largest supercomputer accessible via XSEDE. Stampede is operated by the Texas Advanced Computing Center (TACC)
    - Blacklight – a shared memory supercomputer run by the Pittsburgh Supercomputing Center (PSC).
    - Gordon – a data-intensive supercomputer at the San Diego Supercomputer Center (SDSC), accessible also through NCGAS' partner program CIPRES (Cyberinfrastructure for Phylogenetic Research).
- Archiving of public data sets, with Dublin Core metadata so that they can be published and discovered through web searchers, in IU's persistent digital archive (based on DSPACE)
- Providing education and outreach programs on genome analysis and assembly, including designing genomics experiments, using best-of-breed tools, and interpreting data
- Aiding researchers preparing grant proposals, including providing Letters of Collaboration for their project's NSF grant submission and partnership agreements that commit NCGAS to aiding their research.

In order to disseminate information about the services offered by NCGAS, and about other bioinformatics and cyberinfrastructure services supported by the National Science Foundation, and to interest the scientists of tomorrow in biology, bioinformatics, and computational science in general, NCGAS engages in a vigorous program in outreach.

The primary intellectual merit of the NCGAS lies in the value of the science done by other researchers and enabled by NCGAS services. NCGAS has supported high-impact science, including help with the assembly of the pine and mango transcriptomes, the ecologically important Atlantic zooplankton Calanus finmarchus, and the assembly of the fruit fly (Drosphila melanogaster) transcriptome. As important, it supported biological research at a number (11 to date) of projects led from EPSCoR states.

During PY3, NCGAS supported new 12 research projects, for a total of 82 projects during its 3-year tenure to date. These new projects represented $25,334,081 of grant funding. NCGAS supported another $36,179,520 in grant-funded projects that carried over from previous years, for a total of $61,513,601 in funded projects supported by NCGAS in PY3. Those projects ran a total of 41,679 compute jobs on IU's Mason system. NCGAS also provided up to 50 Terabytes of long-term storage for data archiving and publishing for each project and now has archived data from 6 projects. It has provided cost effective consulting service with an average cost of $17,805 per project. During PY 3, NCGAS engaged in 51 education, outreach, and training events, which served 691 individuals. Of these 21% (146) were from underserved populations. NCGAS currently maintains 62 software tools on its systems for use by genome scientists. There is additional intellectual merit in the quality of the software implementation by the NCGAS, including use of best-practice software engineering techniques, support for high-performance computing techniques, and computational workflows to advance new biological research, and in effective leverage of national shared infrastructure services.

The broader impacts of the NCGAS are in the creation of an enhanced national research infrastructure, societal benefits that result from use of that infrastructure, and a 21st-century biology research community skilled in computational and data-intensive science and engineering. The research supported by the NCGAS will have significant societal impact as it will improve our knowledge of such important topics as understanding basic biological mechanisms, the effects of climate and environmental change, assessing the affects of substances in the environment on organisms, and understanding population dynamics. The NCGAS will also help advance US scientific competitiveness and grow a technologically literate workforce. NCGAS services will be particularly valuable to researchers at small schools, including minority-serving institutions, where biologists often lack local facilities that provide bioinformatics consulting, software, or data storage.

2. Introduction

On September 15, 2011, Indiana University (IU) received a grant award from the National Science Foundation, through the Advances in Biological Infrastructure, to establish the National Center for Genome Analysis Support (NCGAS). This technical report describes the activities of the third 12 months of NCGAS, from September 15, 2013 through September 14, 2014. The NCGAS is a partnership among Indiana University as the lead institution, the Texas Advanced Computing Center (TACC), the Pittsburgh Supercomputer Center (PSC), and the San Diego Supercomputing Center (SDSC).

The mission of the NCGAS is to enable the biological research community of the US to analyze, understand, and make use of the vast amount of genomic information now available. NCGAS focuses particularly on transcriptome- and genome-level assembly, phylogenetics, metagenomics/transcriptomics and community genomics.

NCGAS addresses critical problems genomics researchers face today. Next-generation sequencers generate much more data, straining laboratory and departmental cyberinfrastructure. As well, understanding the most useful software and interpreting resulting data requires special expertise. Arriving at biologically relevant conclusions often entails analytical processes that are highly detailed, complex, and fraught with technical challenges. These factors have erected technical and expertise barriers to conducting genomics research. NCGAS was created to counter these barriers.

Evidence of the need for and importance of proposed work and disciplinary breadth

Current research programs and NSF funding highlight genome assembly and analysis as a vital and growing dimension of biological research. Our experience and surveys of the needs of genome scientists show a surge in the need for bioinformatics, data management, and high performance systems, especially among biologists who are not bioinformaticians. Many large genome projects aim to complete sequencing during the next 4 years, demonstrating the growing emphasis on genome science and need for bioinformatics. Projects include the 1000 Genomes Project [1], the Joint Genome Institute 1000 Fungal Genomes Project [2], the BGI 5,000 insect genomes project [3], the BGI Genome 10K program (a "Noah's Ark") [4], the Avian Phylogenomic Project (48 species) [5], and the BGI Fish T1K project (1,000 fish transcriptomes) [6]. Microbial genomes continue apace. Together these projects plan the assembly of ~2,0000 eukaryotic genomes, the majority de novo. The Earth Microbiome Project [7], the USA HMP Human Microbiome project [8], the MetaHIT European human microbiome project [9], and the Canadian Microbiome Initiative [10] represent important metagenomics initiatives.

The number of eukaryotic assemblies researchers plan to complete between now and the proposed award's end is about 20x times today's number. The number of prokaryotic genome assemblies planned may exceed this increase, given current microbial single-cell genomics. Azvolinsky [11] estimated early in 2014 the number of published full-genome assemblies at 17,782 for prokaryote species and 1,017 for eukaryotes. While some of these projects have their own CI resources and full-time bioinformaticians, in our experience, many researchers said they greatly value NCGAS services and/or would not have been able to complete their research without them.

NSF funding for genomics research is significant. A search of the NSF award database on 1 August 2014 showed 284 awards involving "genomics" or "transcriptomics" totaling over $170,348,488. Adding "genomes" and searching across all NSF awards yields 2,094 awards and over $1 billion in award money. Many are for dissertation research or career development. The 284 grants may approximate the amount of direct NSF-funded genomic research; the 2,094 awards indicate how far the concept of genomes has permeated research in biology, including that funded by the NSF.

NCGAS provides services to biologists in the following areas:

1) Provide excellent bioinformatics consulting services. We provide national consulting services, support DNA genome and RNA sequence assembly support, and support other bioinformatics tools for a changing life sciences research community. Services include consulting on research design, selecting bioinformatics tools and workflows, using these tools (including data transfer and file transformation tools), and aiding data management, interpretation and visualization, and data storage and publication.

2) Maintain, support, and deliver genome assembly and analysis software on national CI systems. We coordinate the installation, version control, and support for genome assembly and analysis on NCGAS partner systems, including Stampede at TACC, Gordon (and in future Comet) at SDSC, Blacklight at PSC, and university-funded systems, such as IU's Mason large-memory cluster. These will be integrated with XSEDE and OSG submission environments for executing analyses via those programs. NCGAS focuses on providing software for RNA sequence assembly (Trinity) [NCGAS1], [NCGAS82]); DNA sequence assembly, particularly de novo assembly. NCGAS supports population genomics analyses, including mlRho [NCGAS83] and metagenomics. NCGAS provides a Galaxy workflow composition environment that supports web-based genome analysis workflows including metagenomics, which lowers the barrier for creating, executing, documenting, and sharing genomics analyses. This portal supports genome assembly and analysis on Mason and provides a tool for running BLAST (Basic Local Alignment Search Tool) on the OSG [NCGAS12, NCGAS63]. The gateway is being expanded this year to include support for metagenomics software on Blacklight. The value of these environments is:

- Biologists, not technologists, select software to be installed.
- Research goals, ease of use, and uniformity drive the delivery software environments across NCGAS-provided supercomputers through tools such as Galaxy.

We maintain and support gateways for genome analysis software on supercomputers coordinated by XSEDE, funded through the NSF XD program [12] and the OSG [13], which operates the world's largest grid of 124 sites for high-throughput computing, and university-operated supercomputers integrated with NCGAS such as Mason and Blacklight.

3) Disseminate tools for genome assembly and analysis. We enhance access to genome analysis and assembly software described above. We currently support and disseminate 62 software tools [14] and we review and modify this toolset at least once every 6 months in response to evolving community needs, incorporating software enhancements as needed. NCGAS distributes new versions of software to supercomputing facilities.

4) Provide long-term archival storage for genome biologists. We provide long-term archival storage services. These services were added to NCGAS' portfolio in PY3 in response to biology researchers' requests. This service provides two types of storage:

Tape archival storage private to the researcher, with 50 TB default storage on the IU Scholarly Data Archive (SDA) [15], which will be available for a minimum of 3 years after the end of NSF funding for NCGAS. IU staff will help researchers migrate data among storage resources.

Publication and persistent availability (into the foreseeable future) of data sets through IUScholarWorks, using the DSPACE digital library software with Dublin Core metadata and a unique digital object identifier (DOI) [NCGAS84], [NCGAS85], [NCGAS86].

5) Provide education and outreach programs on genome analysis and assembly, including designing genomics experiments, using best-of-breed tools, and interpreting data. We deliver annual summer clinics with training in research design, using bioinformatics tools such as Trinity and Galaxy, and in interpreting and managing data. We participate in scientific conferences, provide workshops, and collaborate with biologists on presentations on their research. Training materials are posted for reuse on the NCGAS web site [16]. We will also continue to apply for REU supplements to train CI professionals, and mentor bioinformatics professionals. As part of NCGAS services IU and PSC:

- Sustain and provide access to specialized computational resources supporting large-memory genome and transcriptome assembly, particularly Mason and Blacklight
- Provide tape storage through the SDA, with dual copies in Bloomington and Indianapolis, IN, managed with the secure HPSS (High Performance Storage System [17])
- Continue to develop and enhance genomics software to ensure optimized, hardened versions of best-of-breed critical software.

The services NCGAS provides make it easier for genomics researchers to conduct their science. NCGAS bioinformaticians who understand the biological problems and the relevant technologies help scientists use cyberinfrastructure tools and aid in upstream study design and downstream data interpretation, ensuring the veracity and integrity of the science, especially for those new to genomic analyses. Hardened and optimized bioinformatics software and web interfaces such as Galaxy provided by NCGAS make it easier for investigators to create, manage, and execute their own workflows. Through NCGAS, scientists have access to resources not usually available to local labs or department, such as large-memory systems for assembling and storing large data sets.

In today's environment of very inexpensive and very large raw genome data sets, NCGAS provides life science researchers with bioinformatics support, hardened and optimized software, low-barrier web workflow and analysis interfaces, and computation and storage infrastructure, all at no cost. In so doing, NCGAS is helping to accelerate biological research and dissolving bottlenecks to scientific productivity.

During the first project year, NCGAS set up shop, establishing policies, practices, and resources to meet the needs of genomics researchers. It supported a significant number of research projects (see PTI Technical Report PTI-TR13-002, http://hdl.handle.net/2022/15340). IU provided the Mason large-memory cluster and installed genome analysis applications. Each of Mason's 16 nodes has a 32-core processor and 512 gigabytes (GB) of Random Access Memory (RAM), and is architected for memory-intensive genome assembly. The low barrier system for access to Mason software and bioinformatics support quickly began serving genome scientists, while building its online presence and outreach to the research community.

NCGAS staffed up quickly. Le Shin Wu was transferred from IU's Research Technologies (RT) division of University Information Technology Services (UITS) to provide computer science support. Dr. Thomas Doak, a genomics scientist in IU's Department of Biology, was retained part-time to provide genomics consulting and outreach. NCGAS contracted with TACC to support genome projects at their site. In March of PY1, NCGAS hired Dr. Richard LeDuc to provide management leadership. In early PY3 Carrie Ganote was hired to provide bioinformatics support. By the end of PY1, NCGAS staff had installed 45 bioinformatics software packages on Mason, supported 25 NSF-funded genomics research projects, engaged in 10 outreach events, made 22 peer-reviewed or invited presentations, and exhibited at two major conferences attended by NSF-funded genomics researchers.

NCGAS also created and improved innovative cyberinfrastructure for genome science. In December 2011, it implemented the Galaxy bioinformatic web portal. In June 2012, IU and the Broad Institute completed an optimization of the Trinity software package for RNA sequencing assembly, providing a fourfold improvement in speed with no loss in accuracy. This effort led to a five-year grant from the National Cancer Institute (NCI) through its Informatics Technology for Cancer Research (ITCR) program to the Broad Institute and IU in PY3 that will:

- Update the Trinity software to provide additional functionality (Broad Institute)
- Optimize Trinity software to improve its performance, particularly on large systems (IU), and
- Deliver Trinity through a Galaxy web portal on dedicated compute resources to the cancer research community (IU).

In PY3, NCGAS continued its science support, outreach, and cyberinfrastructure development. The following report details NCGAS efforts in PY3 towards meeting its goals.

3. Consulting and support for biological research in the US

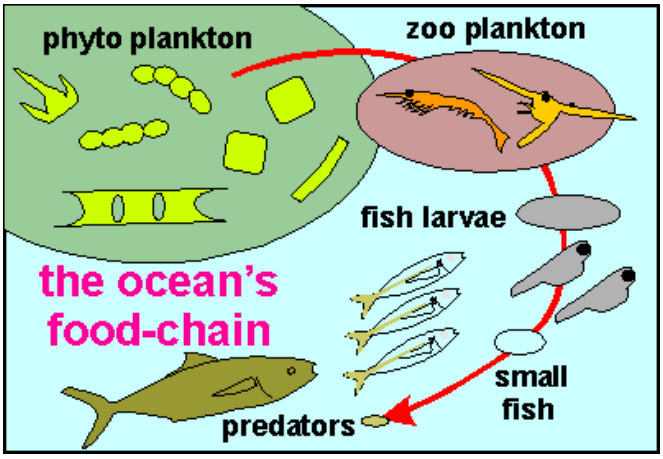
NCGAS consulting enables a broad spectrum of genomics research, listed in Appendix 1. During PY3, NCGAS supported new 12 research projects, for a total of 82 projects during its 3-year tenure to date. These new projects represented $25,334,081 of grant funding. NCGAS supported another $36,179,520 in grant-funded projects that carried over from previous years, for a total of $61,513,601 in funded projects supported by NCGAS in PY3.

Some of the highest-impact allocated projects resulting in completed contributions to scientific knowledge are described below:

Project title:	**Loblolly Pine Genome Project**
PI or project lead:	Dr. David Neale, U. California Davis, California
Funding agency:	United States Department of Agriculture, National Institute of Food and Agriculture (USDA-NIFA)
Award number(s):	#2011-67009-30030
Outcome:	The annotation of the first whole-genome shotgun assembly of loblolly pine (*Pinus taeda* L.), which comprises 20.1 Gb of sequence.
Impact / benefits:	Conifer genomes present challenges for successful sequencing, mainly due to their large size and complexity. Development of a high-quality reference genome sequence for loblolly pine can serve as a model approach for sequencing other large, complex genomes and empower the forest tree biology research community and the broader biological research community in the practical use and application of this resource.

Explanation:	NCGAS has contributed essential computing and data management to building the first comprehensive conifer gene expression reference, the loblolly pine reference transcriptome. At Indiana University (Department of Biology), co-PI Keithanne Mockaitis leads the sequencing, assembly and analysis of transcriptome references for PineRefSeq. NCGAS began participating shortly after the PineRefSeq project launched in February 2011. Mockaitis began working with the NCGAS staff almost immediately to manage the massive NGS data her group was generating. Loblolly pine represents the largest genome sequence reference to date and our work has contributed essential evidence—the annotation of transcribed regions and gene functional annotation. We have performed extensive RNA assembly with inputs of up to 1.3 billion reads, using 4 different software installations. Downstream analyses have included extensive BLAST-based protein homology annotation, analysis of differential expression among sample sources, and RNA-based variant analyses. Our ongoing work is improving gene annotation further, and assisting other stakeholders in the conifer research community with expanding the pine genetic map with markers based on expressed sequence variation and functional genomics experiments. These unprecedented references will finally enable both functional genomics in conifers and deep evolutionary studies across the plant kingdom.  Wegrzyn, J.L., Liechty, J.D., Stevens, K.A., Wu, L-S., Loopstra, C. A., Vasquez-Gross, H. A., Dougherty, W. M., Lin, B. Y., Zieve, J. J., Martínez-García, P. J., Holt, C., Yandell, M., Zimin, A. V., Yorke, J. A., Crepeau, M. W., Puiu, D., Salzberg, S. L., de Jong, P. J., Mockaitis, K., Main, D., Langley, C. H., Neale, D. B. (2014). Unique Features of the Loblolly Pine (Pinus taeda L.) Megagenome Revealed Through Sequence Annotation. Genetics, 196(3), 891-909. PMCID: PMC3948814. http://www.genetics.org/content/196/3/891.full.pdf Neale, D., Wegrzyn, J., Stevens, K., Zimin, A., Puiu, D., Crepeau, M., Cardeno, C., Koriabine, M., Morris, A. H., Liechty, J., Garcia, P. M., Gross, H. V., Lin, B., Zieve, J., Dougherty, W., Soriano, S. F., Wu, L. S., Gilbert, D., Marcais, G., Roberts, M., Holt, C., Yandell, M., Davis, J., Smith, K., Dean, J., Lorenz, W., Whetten, R., Sederoff, R., Wheeler, N., McGuire, P., Main, D., Loopstra, C., Mockaitis, K., deJong, P., Yorke, J., Salzberg, S., and Langley, C. (2014). Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biology, 15(3):R59+. doi:10.1186/gb-2014-15-3-r59

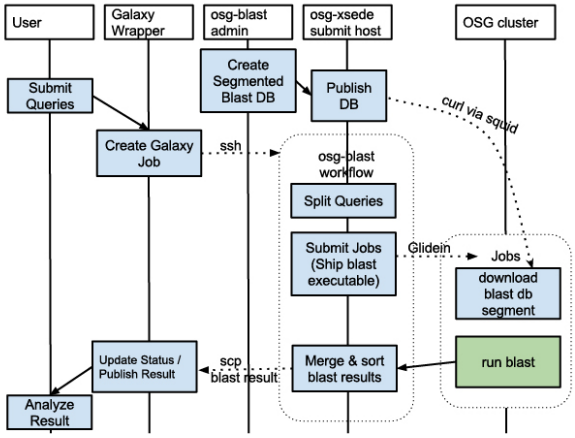
Project title:	**North Atlantic Fisheries Decline: Zooplankton transcriptomes**
PI or project lead:	Sonya Dyhrman, Lamont Doherty Earth Observatory, Columbia University, New York
Funding agency:	NSF
Award number(s):	CCF-424599, OCE-1316036, OCE-0925284
Outcome:	"De Novo Assembly of a Transcriptome for *Calanus finmarchicus* (Crustacea, Copepoda)" in the February 2014 issue of the journal *PLOS ONE.*
Impact / benefits:	Zooplankton are a key link in ocean food webs, consuming phytoplankton and in turn being consumed by many larger ocean inhabitants. Thus, any decline in zooplankton populations will result in fisheries decline. We hope to learn more about zooplankton biology by cataloging their genes and the genes' expression.

Explanation:	 Researchers from the Pacific Biosciences Research Center at the University of Hawaii Mānoa and Ohio University asked NCGAS to provide bioinformatics skills and computational resources to study the transcriptome of zooplankton, the small, multicellular organisms that form the basis of the marine food chain. In recent years, scientists have seen a corresponding decrease in zooplankton along with cod fisheries in the North Atlantic Ocean. Theorizing that global climate change might be to blame, the scientists needed data to fully understand the phenomenon. Deciphering the messages or "transcripts" that the organisms' cells produced allows researchers to pinpoint the causes of population changes. For the project, NCGAS technologists leveraged their prior expertise in optimizing Trinity software and took the lead in moving the data, running the key analyses, and transporting the results back to the researchers. Trinity was developed by researchers at the Broad Institute of MIT and Hebrew University. It produces high-quality RNA sequence assemblies used by scientists studying gene expression. These RNA sequence assemblies allow scientists to know which genes are active within a living creature. Christie, A. E., Roncalli, V., Wu, L.-S. S., Ganote, C. L., Doak, T., and Lenz, P. H. (2013). Peptidergic signaling in calanus finmarchicus (crustacea, copepoda): in silico identification of putative peptide hormones and their receptors using a de novo assembled transcriptome. General and comparative endocrinology, 187:117-135. doi: 10.1016/j.ygcen.2013.03.018

Project title:	**Study of complete RNA collection of fruit fly uncovers unprecedented complexity**
PI or project lead:	Susan E. Celniker, Department of Genome Dynamics, Lawrence Berkeley National Laboratory, California
Funding agency:	National Human Genome Research Institute, NIH, Lilly Endowment
Award number(s):	National Human Genome Research Institute modENCODE Project, contract U01 HG004271 and U54 HG006944; R01 GM076655; NHGRI K99 HG006698; modENCODE DAC sub-award 5710003102, 1U01HG007031-01 and the ENCODE DAC 5U01HG004695-04; Indiana METACyt Initiative of Indiana University (funded by an award from the Lilly Endowment); U01-HG004261 and RC2-HG005639
Outcome:	This work shows that the *Drosophila* genome is far more complex than previously suspected and suggests that the same will be true of the genomes of other higher organisms. The paper also reports a number of novel, particular results: that a small set of genes used in the nervous system are responsible for a disproportionate level of complexity; that long regulatory and so-called "antisense" RNAs are especially prominent during gonadal development; that "splicing factors" are themselves spliced in complex ways; and that the *Drosophila* transcriptome undergoes large and interesting changes in response to environmental stresses.
Impact / benefits:	The importance of *Drosophila melanogaster* as a model system cannot be overstated. Using it, the mechanisms of heredity were worked out about 100 years ago. Today, as biologists have developed increasing appreciation of how well genes and critical life processes are conserved over long evolutionary distances, flies have emerged as critical tools for understanding human biology and disease. Drosophila research is an area that has long had associations with IU, beginning with Nobel Laureate Herman J. Muller.

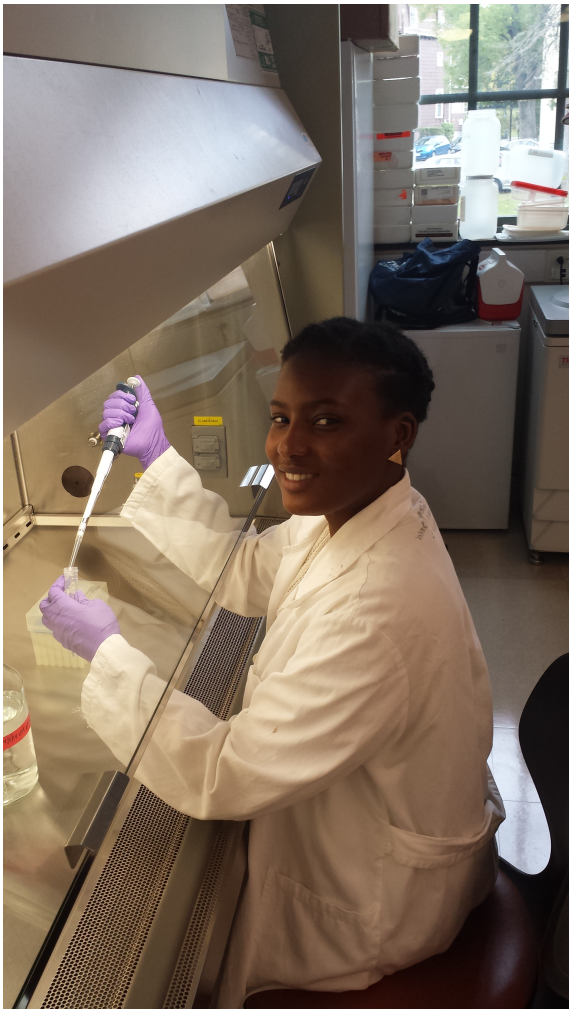
Explanation:	 In the new work, published March 16, 2014 in the journal *Nature*, scientists studied the transcriptome—the complete collection of RNAs produced by a genome—at different stages of development, in diverse tissues, in cells growing in culture, and in flies stressed by environmental contaminants. To do so, they used contemporary sequencing technology to sequence all of the expressed RNAs in greater detail than ever before possible. NCGAS provided software and computational resources for this project. Brown, J. B., Boley, N., Eisman, R., May, G. E., Stoiber, M. H., Duff, M. O., Booth, B. W., Wen, J., Park, S., Suzuki, A. M., Wan, K. H., Yu, C., Zhang, D., Carlson, J. W., Cherbas, L., Eads, B. D., Miller, D., Mockaitis, K., Roberts, J., Davis, C. A., Frise, E., Hammonds, A. S., Olson, S., Shenker, S., Sturgill, D., Samsonova, A. A., Weiszmann, R., Robinson, G., Hernandez, J., Andrews, J., Bickel, P. J., Carninci, P., Cherbas, P., Gingeras, T. R., Hoskins, R. A., Kaufman, T. C., Lai, E. C., Oliver, B., Perrimon, N., Graveley, B. R., and Celniker, S. E. (2014). Diversity and dynamics of the drosophila transcriptome. Nature, advance online publication. doi:10.1038/nature12962

Project title:	**BLAST on the Open Science Grid provides a time-saving alternative.**
PI or project lead:	Miron Livny, Open Science Grid. University of Wisconsin at Madison, Wisconsin
Funding agency:	National Science Foundation, Dept. of Energy
Award number(s):	1062432 ($1,493,200.00), 1148698 ($11,250,000.00)
Outcome:	The Basic Local Alignment Search Tool (BLAST), an algorithm for comparing primary biological sequence information, is one of the most widely used tools in bioinformatics. The National Center for Genome Analysis Support (NCGAS) and the Indiana University (IU) High Throughput Computing (HTC) group have been experimenting with using the Galaxy web-based user interface to submit BLAST jobs on the Open Science Grid (OSG). The outcome of this experimentation is the development of a faster method of deploying BLAST jobs and getting results back than one could achieve by using the supercomputers at IU.
Impact / benefits:	Galaxy at IU provides a web-based platform for data-intensive genome analysis research. It employs IU's Mason cluster for compute services and the IU Data Capacitor for project storage, and is hosted on IU's Quarry Gateway Web Services Hosting System. Galaxy is a scientific workflow platform that makes computational biology easier for research scientists who do not know computer programming. NCGAS has created Galaxy portals for IU investigators and NSF-funded life science researchers nationally. These provide ready access to the full suite of genome assembly, annotation, alignment, and other applications—as well as the file transfer and transformation utilities necessary to build genome science workflows. Today's technologies for genome sequencing are faster and cheaper, and create more sequence data than ever before. With limited local computing resources, analyzing, understanding, and using these vast amounts of genomic information becomes challenging in terms of efficiency. One solution is to split a single, sizeable analysis task into many independent, smaller tasks and then distribute them to multiple computing resources in parallel. The combination of Galaxy's easy-to-use interface and the BLAST on OSG's splitting and distributing functionality makes it easier for the researcher to get more done in less time.



Explanation:	The OSG can support large amounts of central processing unit (CPU) hours simultaneously. Soichi Hayashi in HTC has been researching a way to run BLAST in parallel by splitting up the target database into many chunks and making it run in a distributed high-throughput computing (DHTC) environment, namely the OSG. In turn, Carrie Ganote (NCGAS) has enabled OSG BLAST on IU's Galaxy interface. Ganote says that the interface for running BLAST on OSG will provide an alternative to the National Center for Biotechnology Information BLAST servers, which are wonderful for small jobs and parameter tinkering, but prohibitively slow for large jobs. Moore, Greg (2014). BLAST on OSG provides a timesaving alternative for large-scale analysis. Web. Accessed 16 Oct 2014. Retrieved from http://www.opensciencegrid.org/blast-on-osg-provides-a-timesaving-alternative-for-large-scale-analysis-2 Hayashi, S., Gesing, S., Quick, R., Teige, S., Ganote, C., Wu, Le-S., Prout, E. (2014). Galaxy based BLAST submission to distributed national high throughput computing resources. Presentation. Presented at the International Symposium on Grids and Clouds (ISGC) 2014 March 23-28. Academia Sinica,Taipei, Taiwan. http://hdl.handle.net/2022/18609

Project title:	**Host-Endosymbiont Relationships between Filarial Nematode and Wolbachia**
PI or project lead:	Steven Williams, University of Massachusetts – Amherst, Massachusetts
Funding agency:	NSF, NIH
Award number(s):	1248096 for GCAT-Seek project, 1R15AI039721-01 NIH "SPLICED LEADER (SL) AND NON-SL MRNA IN BRUGIA PARASITES"
Outcome:	Gene expression patterns in closely related organisms can shed some light on the underlying reasons for the differences between these organisms. Weam Zaky and Marie Jacques Seignon are searching for clues to the nature of host-endosymbiont relationship by studying the differences in gene expression between Brugia malayi and nematode species that do not cause disease and are not hosts for Wolbachia.
Impact / benefits:	One feature of the filarial parasitic nematodes is their reliance on a bacteria species, Wolbachi, to function normally. The Wolbachia bacteria infect the nematode, but provide with benefits in return. Wolbachia is spread to the offspring by the parent nematodes. If the bacteria are killed off by antibiotics, the worms will eventually die, making this relationship a potential target for controlling the disease.

Explanation:	The vector responsible for lymphatic filariasis, a disease that manifests in humans as elephantiasis, is a tiny, threadlike worm known as a nematode. Three species of nematode cause this disease, one of which is extensively studied in the lab of Steven Williams at Smith College. This parasite, Brugia malayi, is endemic in regions of China and Central Africa, where the disease it causes leads to disfigurement and incapacitation. The nematode relies on mosquitos to carry its larvae from host to host, making control of spread of the disease very difficult. Zaky, Weam. "Gene Expression in the Bacterial Endosymbiont of the Filarial Parasite, Brugia malayi". Celebrating Collaborations. Smith College. Northampton, MA. 17 Apr. 2010. Conference Presentation.	

4. Consulting interactions – summary

The interaction between NCGAS staff and the user community is classified into three categories: Short-term consultations, extended consultations, and supported projects. Short-term consultations take less than four hours of staff time and typically center on resolving a simple technical question, or advising a user on how to proceed. Extended consultations require more than four hours of effort and can be either technical or scientific. Technical consultations usually involve complex technical issues that exceed the reasonable understanding of a domain scientist. Supported projects are not documented through the ticketing system, but instead are research projects that may be supported by a combination of short-term and extended consultations, depending on the needs of the project. They are listed in Appendix 1.

In PY3 NCGAS staff reported 267 short-term and 124 extended consultations. Figure 1 shows the breakdown of the time spent on consultations.

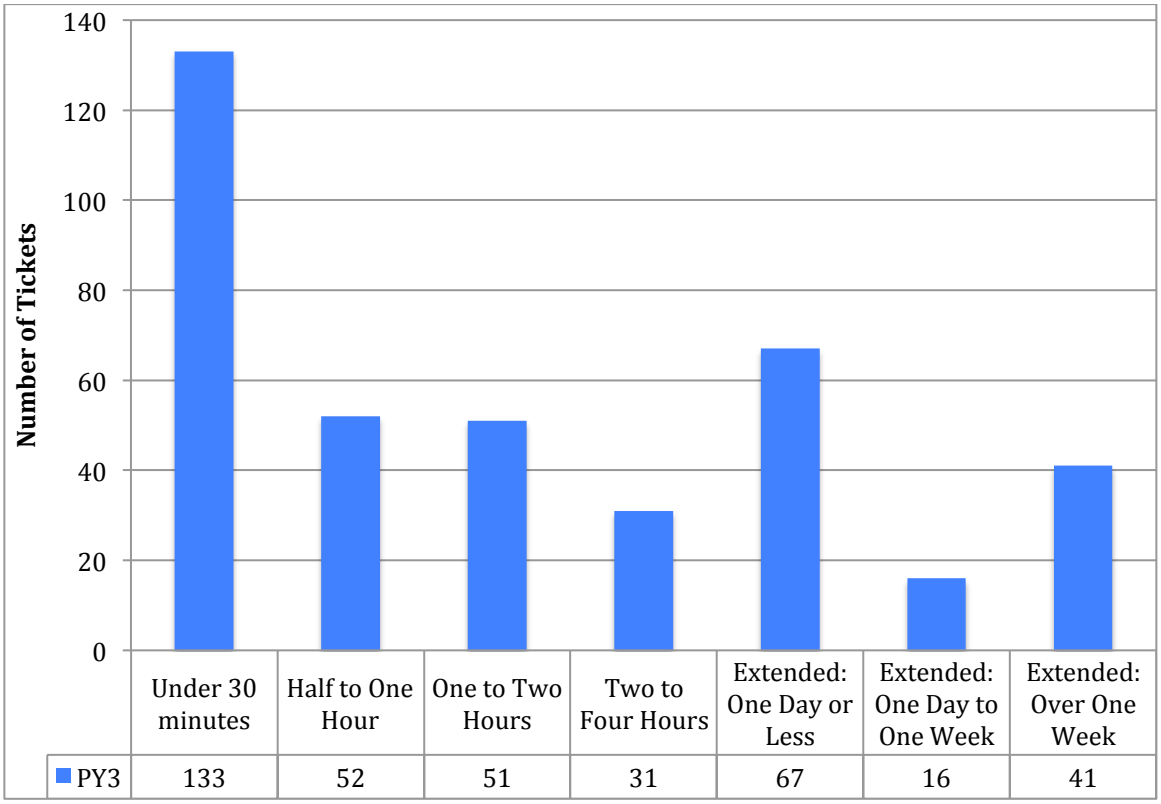


Figure 1. Number of tickets reported by NCGAS staff as a function of the time needed to complete the ticket.

4.1 Projects receiving significant support from NCGAS

One of our major services is support for biological research projects or "allocated projects," committing at least 1 person-month of consulting and allocated time on Mason. During PY3, NCGAS added 12 new research projects to the list of supported (allocated) projects (11 from non-IU institutions).

To obtain this service, researchers submit brief online applications via the NCGAS web site, which NCGAS staff review under direction of the NCGAS Science Advisory Board (SAB). Reviews match requested resources to scientific researcher needs. NCGAS assumes research done under NSF support has been favourably peer reviewed, so scientific merit is not re-reviewed. Scientific merit of requests from researchers without NSF funding is given an initial staff review. Questions about merit are referred to the

SAB. So far NCGAS has never declined a request for assistance from a US-based researcher. To date, NCGAS has approved 82 allocated projects.

Growth in the number of supported projects is shown in Figure 2 below.

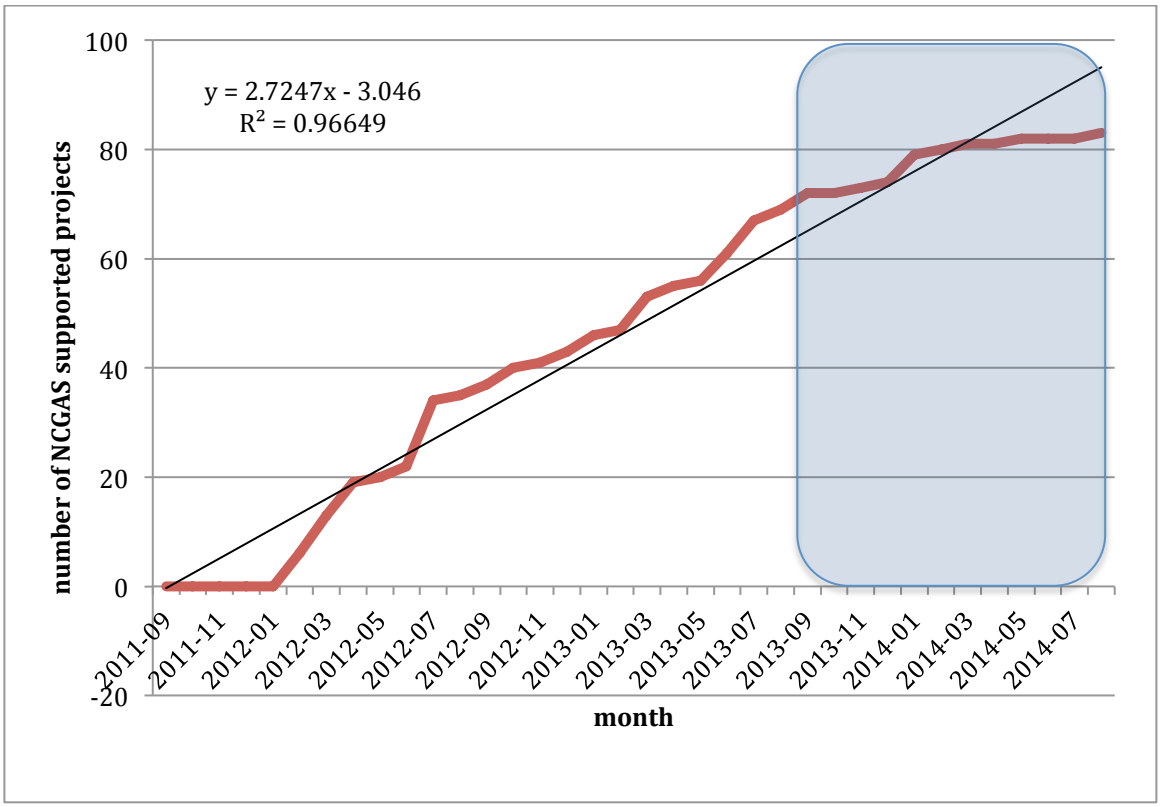


Figure 2. NCGAS-supported projects, per month, showing an addition of 2.7 new projects per month.

5. Scientific products

NCGAS services are critical to biologists without access to targeted consulting / support or required genomics tools. The proof of NCGAS success is in the outcomes. The rise in the number of NCGAS interactions with the scientific community, shown as "projects" in the figure above, reflects increasing demand. Those interactions figure into advances in research—improved tools, completed projects, publications, and so on. In the past 3 years NCGAS has enabled the following intellectual and technical outputs:

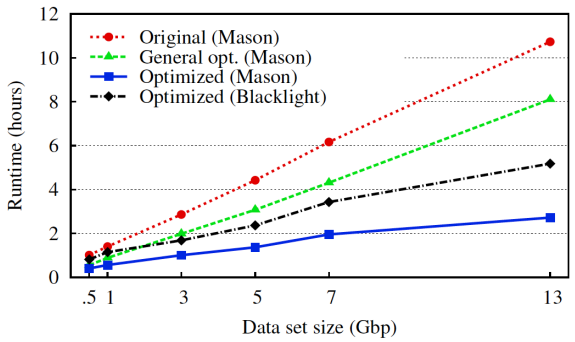


Figure 3. Trinity run times before (top, red line) and after (lowest, blue line) optimization by Indiana University/NCGAS team.

- High-impact science, including help with key assembly projects completed in the past 3 years: Assembly of the loblolly pine transcriptome (*Pinus taeda* [NCGAS16], [NCGAS69], [NCGAS 70]), cacao (*Theobroma cacao* [NCGAS8]), mango [NCGAS29], [NCGAS 68]), the ecologically important Atlantic zooplankton *Calanus finmarchicus* [NCGAS2], [NCGAS15], and assembly of the transcriptome of the fruit fly (*Drosophila melanogaster*) [NCGAS14]. Topics of supported research include biodiversity, functional genomics, population genomics, climate change and environmental impact, and identify the biological basis for disease.
- Support for junior researchers developing research results to compete for the first NSF grant award.
- Faster execution of RNA sequence assembly software Trinity (Figure 3). The June 2012 production release executes in 25% of the time of the prior version. Since then Trinity's downloads total over 16,000. Citations in publications total 421.
- Multiple high-impact, high-value genome assembly projects on Blacklight including:
    - Assembly of a 5-gigabase wheat species genome [NCGAS21]
    - Assembly of 20 primate transcriptomes, creating a non-human primate resource for insight into evolutionary processes and human disease [NCGAS19]
    - Creation of a massive soil metagenome assembly for discovery of enzymes relevant to biofuel production. This research also led to the assembly of a fungal species important in the rumen ecology of herbivores [NCGAS20],
    - Assembly of the Pacific whiteleg shrimp (Litopenaeus vannamei, [NCGAS24]) transcriptome, an important species in ecological food chains and a human food source.
- Implemented the Galaxy web-based front end [18] for software supporting genome assembly and analysis workflows on Mason (Figure 5). We enabled researchers to execute BLAST jobs on the OSG through Galaxy [NCGAS12], [NCGAS63].

NCGAS has enabled the creation of 115 products including 22 peer-reviewed scientific articles, training materials, and news releases that build public appreciation of NSF-funded basic science.

Table 1. Summary of NCGAS services and products created by NCGAS staff or researchers with NCGAS help. Current NSF program years run from September 15 to September 14. The statistics below represent IU fiscal years 1 July to 30 June, so there is some discrepancy with Project Year statistics.

Category	FY 2012	FY 2013	FY 2014	Total to date
Peer-reviewed papers published with NCGAS aid, led by NCGAS-supported researchers	0	8	9	17
Peer-reviewed scientific papers led by NCGAS staff	1	2	2	5
Major research projects allocated at least 1 person-month of NCGAS time	37	32	13	82
Extended consultations (4 hours to 1 month of staff effort)	1	110	124	235
Short-term consultations (< 4 hours of staff time)	84	390	267	741

Distinct users of Mason cluster	239	243	252	734
Compute jobs run on Mason	25,976	29,650	41,679	97,305
CPU core hours used via NCGAS gateway to OSG	0	0	589,792	589,792
Galaxy gateway users	0	65	84	149
Galaxy gateway jobs	0	405	2,349	2,754
SDA users of long-term storage	1	3	6	10
Products published via IUScholarWorks by NCGAS-supported researchers	0	0	3	3
Software versions published w/NCGAS enhancements	1	1	2	4
Total products produced by NCGAS staff or with help	25	27	63	115
Total education/outreach events	10	25	16	51
Total attendees at education/outreach events	332	1,708	691	2,731

An important question is "How national is the National Center for Genome Analysis Support?" We have now served clients in more than half of the states of the nation. Figure 1 and Table 1 show it took time to attract a national client base and for results of NCGAS-supported research to appear in peer-reviewed journals. Table 2 shows the number of states with NCGAS clients, including users of Mason, outreach/education events, and short-term consulting.

Table 2. The number of states with NCGAS clients

Aggregate number of states using NCGAS services			
Category	PY1 (2011-12)	PY2 (2012-13)	PY3 (2013-14)
Extended consults/major projects	13	22	25
Users of Mason cluster	18	31	33
Education & outreach events	14	17	20

Figure 4 below shows the national distribution of 82 major consulting allocations, PY1 —PY3. Many in PY1 involved IU faculty and grad students, due to proximity and the fact that two NCGAS staff were IU employees who collaborated with IU research groups. Overall, 46% of the extended consultations and major projects are with non-IU researchers. For PY2 and PY3 82% of major consulting projects served non-IU researchers. Figure 4 (where gray areas represent EPSCoR states) also shows that expanding scope of NCGAS service aids researchers in those states. Of the 82 projects 11, or 13.4%, were led by a researcher in an EPSCoR state.

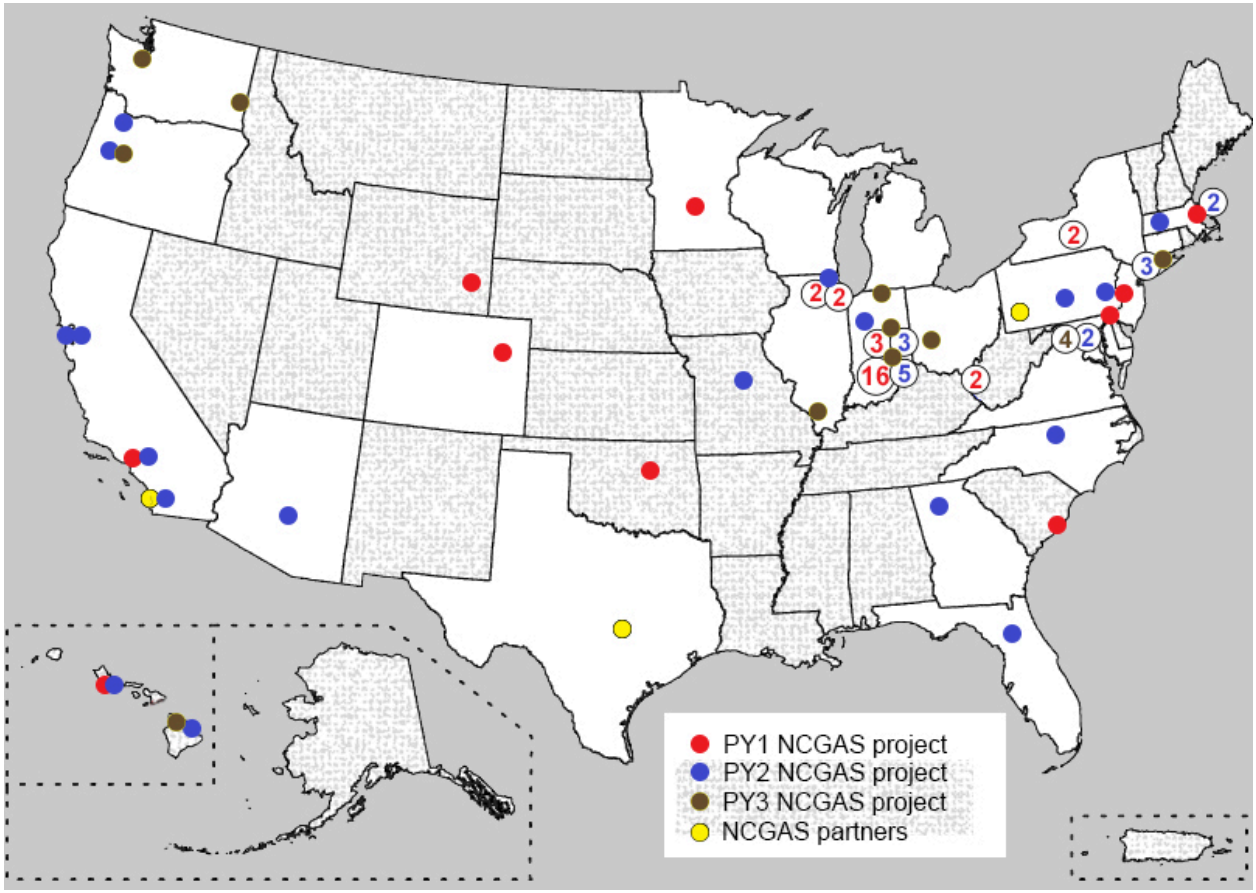


Figure 4. The national distribution of 82 major consulting allocations, PY1—PY3

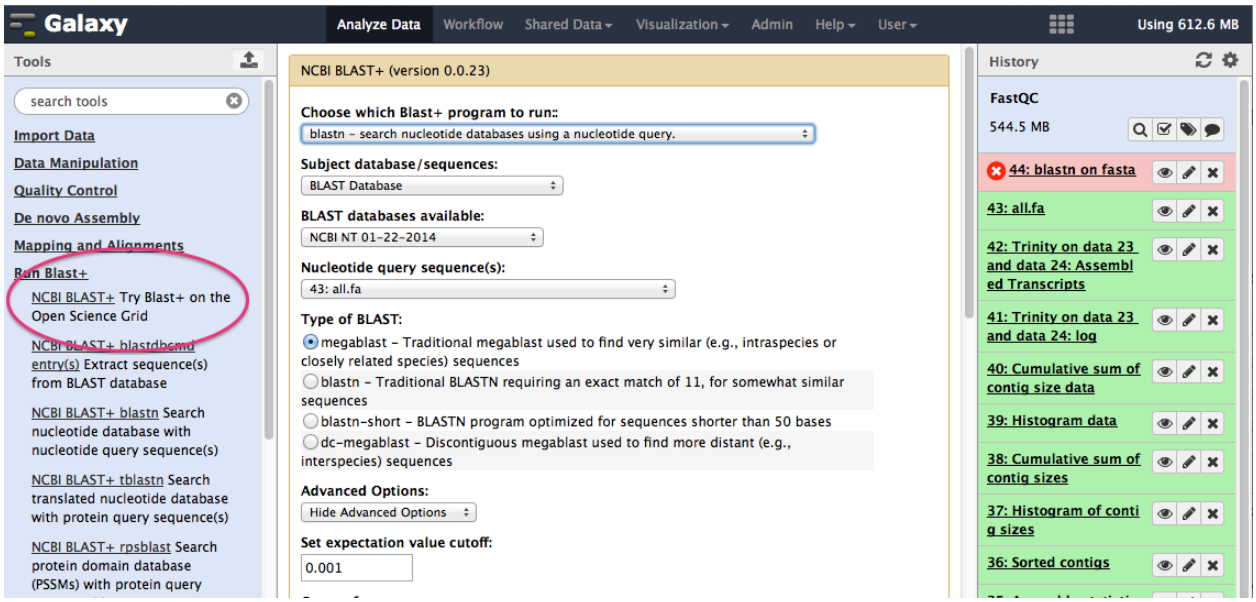


Figure 5. This image shows the web page researchers see when the use the NCGAS instance of the Galaxy web portal to analyze their next-generation DNA or RNA sequence data. The small addition of the option to run BLAST on the Open Science Grid represents a major enhancement in functionality.

6. Providing support to biologists: Delivery and assistance in use of supercomputer clusters

NCGAS provides support to biologists in delivery and assistance in use of supercomputer clusters for genome analysis, including many supercomputers and supercomputer clusters that are provided through XSEDE (the eXtreme Science and Engineering Discovery Environment) [12]. In particular, we support software on and use of:

- Mason, a large memory supercomputer cluster at IU
- Stampede – the largest supercomputer accessible as part of XSEDE. Stampede is operated by the Texas Advanced Computing Center.
- Blacklight – a shared memory supercomputer run by the Pittsburgh Supercomputing Center.

6.1 Mason

NCGAS and the Mason cluster grew significantly in 2014. Mason users grew linearly with time, from about 400 to more than 650 (Figure 6).

Across PY3, we added over 20 users per month and more than doubled our number of registered users. We started PY3 with 431 total users, and ended with well over 650.

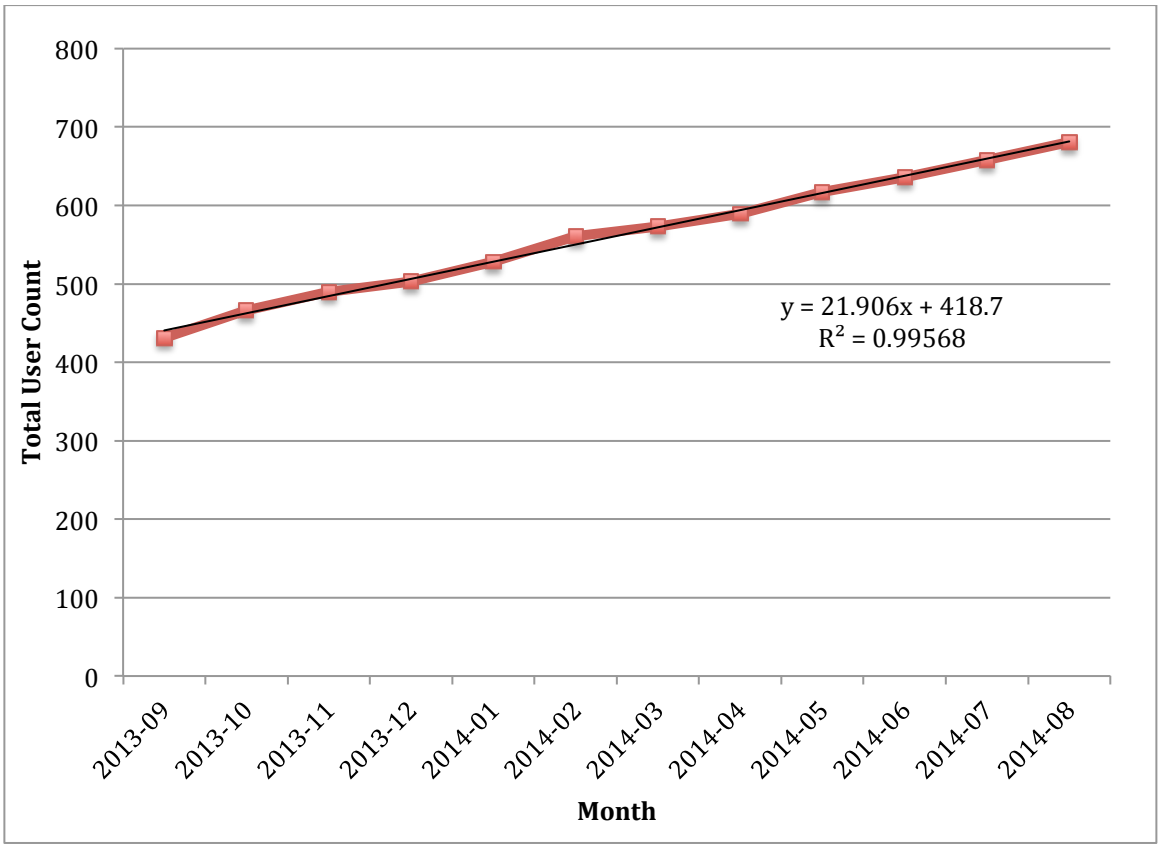


Figure 6. Total users on the Mason system across PY3.

6.2 Contributions of NCGAS partners to date (other than IU)

NCGAS is a partnership and virtual organization: one partner funded through a subcontract (TACC) and two other unfunded partners, PSC and SDSC. As we committed in our development proposal, 12 Mason nodes are dedicated to the US research community. Through the modestly funded subcontract from IU, TACC installs and maintains software distributed by NCGAS on its supercomputers, including Stampede [19]. TACC has funded a graduate assistant for local first-tier user support. SDSC and PSC have worked with NCGAS staff to maintain and update NCGAS-supported software on their supercomputers. NCGAS leverages effort by staff funded through XSEDE and NSF, funding to TACC, SDSC, and PSC, and for the CI resources funded by the NSF XD program, and the OSG. We assert:

- The genomes assembled and scientific articles published in plant sciences reflect the work of the funded graduate assistant at TACC, who is collocated with iPlant-funded staff.
- For the XD program, biology applications (other than molecular dynamics, protein folding, and docking) consumed some 6% of the computing resources used since the 2011 start of XSEDE. Job mix analysis suggests the majority involved genome assembly and analysis.
- NCGAS improved Trinity to run 4 times faster than the old version. From the beginning of PSC's partnership with NCGAS [NCGAS90], Blacklight users from more than 50 research groups have run more than 3,000 Trinity jobs, using more than 3.9 million CPU core hours.

7. Disseminating results

To make research products available, NCGAS conveys information to the scientific research community; publishes in peer-reviewed journals on our software enhancements; co-authors papers with biologists; presents at biological conferences that build awareness of NCGAS services; and builds general awareness of the scientific outcomes we have helped enable via news releases and in the semi-technical press (e.g. [20]), including *International science grid this week*, *Scientific American*, and the like. NCGAS contributions to science publications and other scientific products is shown in Table 3. This approach to dissemination aids other researchers and educators in creating curriculum materials and promotes awareness of NCGAS services. We also provide our training materials on the NCGAS web site and through a CiteULike group page [21]. Software updates are available as submissions to their respective open source repositories. We make software available as RPMs via the XSEDE Yum Repository and the XSEDE Rocks Roll.

Table 3. Summary of scientific products created by scientists with the benefit of NCGAS support during PY3.

Scientific contributions	NCGAS client-led biology and bioinformatics research	NCGAS staff-led methods and bioinformatics papers
Peer-reviewed journal technical papers – published or in press	9	2
Peer reviewed journal papers in-review	3	0
Journal papers in-preparation	4	0
Posters (science content)	1	0
Presentations, including outreach-oriented presentations	4	8

A full listing of citations for these products is presented in Appendix 2.

Appendix 3 provides further information about dissemination of NCGAS products, specifically tables that outline:

- Information about versions and characteristics of the software supported by NCGAS
- Software provided to the US research community on the IU Mason cluster
- Software provided to the US research community on XSEDE-provided resources.

8. Education, outreach, and training

Staff of the NCGAS communicate vigorously with today's science community and with students who will comprise the science community of tomorrow.

Table 12 shows a summary of the NCGAS education, outreach and training events that reached at least 2,731 individuals. The NCGAS Research Experience for Undergraduates (REU) supplement led to 2 Associate Degree candidates at Clark State University, OH, earning full-time jobs at Wright Patterson Air Force Base. The current award includes funding for postdoctoral fellow Thomas G. Doak and his postdoctoral mentoring plan. Its success is noted by Dr. Doak's moving to a permanent position as an associate research scientist and NCGAS manager. In PY2 and PY3 we focused on reaching a diverse user group (see Table 4 for demographics and diversity). Training / outreach events in PY3 served 691. Events include attendance at meetings of minority-serving institutions, including AIHEC (American Indian Education Commission) and SACNAS (Society for the Advancement of Chicanos and Native Americans in Science).

Table 4. Demographics/diversity of attendees at NCGAS training, education, or outreach events.

	Total Attendees	African American	Hispanic American	Native American
Female	197	39	30	3
Male	494	37	33	4
Total	691	76	63	7
% of total		11%	9%	1%

Outreach to the biology research community:

Outreach to the biological research community focused on attending professional conferences where NCGAS staff manned booths to promote services and/or gave presentations in scientific sessions. Additionally, NCGAS delivered 2 workshops to biologists that were focused on training in the use of commonly used tools such as Galaxy and Trinity.

Support for students:

Through a Research Experiences for Undergraduates (REU) award, NCGAS trained two students in systems administration skills for high performance computing systems important for genomics research. Both of these students were offered jobs in this specialty within 6 months of completing their internships. NCGAS also supported 5 undergraduate biology students to attend the 2014 Genome in July workshop hosted at Indiana University.

Outreach to underserved populations and in EPSCOR states:

Of the 82 projects 11, or 13.4%, were led by a researcher in an EPSCoR state. Table 5 shows the summary of the outreach and education activities conducted by NCGAS staff in PY3. Appendix 4 lists the 16 events where NCGAS logged 204 contact hours, where audiences totaled 691. Of these participants, a total of 343 represented members of traditionally underserved groups as defined by the NSF.

Table 5. Summary of outreach and education activities by NCGAS staff for PY 3.

Outreach and education activities	
Posters (outreach and about services)	0

Press releases	2
Presentations, including outreach-oriented presentations	12
Outreach events	16
Total attendees at events	691
Total attendees at events who were members of traditionally underserved groups as defined by NSF	343
Total contact hours in presentations and EOT events	204

9. Plans for PY4

NCGAS PY4 plans are detailed online [NCGAS39]. During this time NCGAS will:

- Continue existing services as defined in Section 1 and on the NCGAS web page [22], including supporting the 40 allocated projects and accepting requests for new allocated projects
- Expand genome analysis services, linking the NCGAS Galaxy front end to metagenomics software on Blacklight, and linking the Galaxy front end to the CIPRES phylogenetics portal
- Expand software support to include GenePattern [22]
- Expand the number of genome analysis and assembly packages available as RPMs and available through a YUM (Yellowdog Updater Modified) repository [23]
- Continue current education, outreach, and training activities, including preparation for the 2nd annual IU Bioinformatics Clinic planned for July 2015 (PY1 of ABI sustaining award)
- Distribute software as RPMs (RPM Package Manager modules) to optimize installation on researchers' local resources. We will help system administrators more easily install, maintain, and run genome analysis software such as Linuxbrew [24] and bcbio-nextgen [25].

9.1 Management and Operations

William Barnett (Director, NCGAS) will be responsible for oversight and coordinating the work plan. He will directly oversee the IU NCGAS staff (Doak, Ganote, Wu), and coordinate collaboration across NCGAS partners, and specifically with OSG. He will be responsible for overseeing community feedback including biannual Advisory Committee meetings and surveys to the NCGAS user community. Tom Doak (Manager, NCGAS) will be responsible for managing NCGAS efforts at IU, including bioinformatics services, outreach events, software maintenance at IU and partner facilities, and tactical coordination among NCGAS partners. Barnett will be responsible for submitting all required reports. The IU Office of Research Administration will review and approve all financial reports. Craig Stewart (PI on the original NCGAS proposal and Associate Vice President of Research Technologies at Indiana University) will be responsible for CI guidance and for coordinating efforts with XSEDE and XD resources, including Campus Bridging efforts. Stewart will be the designated alternate for submitting reports and communicating with funding agencies. As Executive Director of IU's Pervasive Technology Institute (PTI), Stewart will also coordinate involvement and representation of NCGAS activities within the PTI education, outreach, and training program. Philip Blood (PSC) will be responsible for managing services at PSC, for maintaining metagenomics and other very large-memory applications on Blacklight, and for coordinating with XSEDE's Genomics NIP and Extended Collaborative Support Service (ECSS) activities. Scott Michaels (Director, Center for Genomics and Bioinformatics at IU) will have a leadership role in guiding service strategies, and will work with Barnett to implement NCGAS sustainability strategies. Matthew Hahn (IU) and Aviv Regev (Broad Institute) will contribute guidance on provisioning of bioinformatics tools. Gwen Jacobs (U. Hawaii) leads the NCGAS SAB, which includes James Taylor

(Johns Hopkins University), Julie Dickerson (Iowa State University), and IU's Volker Brendel, Tatiana Foroud, Haixu Tang, and Michael Lynch.

Existing NCGAS partners TACC (Matthew Vaughn) and SDSC (Robert Sinkovits) will work with NCGAS to maintain software distributed by NCGAS on their NSF-funded supercomputers (systems administration staff at both sites). TACC will help coordinate activities of NCGAS and iPlant (Vaughn). SDSC will collaborate with NCGAS to expand interfaces between the CIPRES [26] phylogenetics gateway and the NCGAS Galaxy gateway (Mark Miller, SDSC). XSEDE will work with NCGAS to coordinate installation of NCGAS-disseminated software on systems it coordinates and process requests for allocations of time on Mason under the leadership of XSEDE PI John Towns of the University of Illinois Urbana Champaign. The OSG will work with NCGAS to support BLAST running on the OSG and implement additional genome high-throughput analysis tools under the leadership of OSG PI Miron Livny (U. Wisconsin). The Broad Institute (responsible for definitive releases of Trinity) under Aviv Regev will collaborate with other IU staff on improving Trinity performance.

Risks will be managed using risk management software Brinqa Risk Analytics [27] and collaboration and project management software RedBooth under the auspices of PTI [28], which reports to the IU Office of the Vice President for Information Technology. NCGAS services are stable, and as at least 6 of PTI's 100 staff are technically capable of leading NCGAS, we anticipate no leadership risks. New team members will be recruited in accordance with institutional human resources policies, assisted by the IU Human Resources office. All staff will be professionally trained to take on assigned duties and will be matched with a mentor to promote professional growth and rapid transition to the team.

For document management and collaboration, NCGAS uses a Box site on IU's enterprise version of Box.com, which supports authenticated collaborators from multiple institutions. IU maintains a listserv (ncgas-mgmt-l@iu.edu) for email communications among project team members, and will regularly update public information on the NCGAS portal [22]. Usage data on all NCGAS systems are regularly collected and stored on the Box site. It also maintains a listserv (ncgas-advisors-l@iu.edu) for communicating with the Advisory Committee, and a Box site for sharing documents with that group.

9.2 User Survey

NCGAS services are evaluated several ways on an ongoing basis. Annual reports for PYs 1–2, available online [NCGAS 36, 38] detail program evaluations. Summaries of the most recent client survey, feedback from our Science Advisory Board, and suspended services appear below.

Survey summary. Results from our 2014 user survey show:

- On a 1-5 Likert scale, where 1 is "very dissatisfied" and 5 is "very satisfied," the average overall score for NCGAS services was 4.4 ± 1.4 (95% confidence interval).
- 63% of respondents indicated, "I could not have done my research without NCGAS," while another 30% indicated NCGAS was helpful but not essential to completing their research.
- On a 1-5 Likert scale, satisfaction with Mason was 3.9± 2.6.
- 30% of respondents indicated they use the IU or NCGAS Galaxy portal, a front end for the Mason cluster. This is consistent with the high percentage (70%) indicating use of one or more programing languages (Perl, Python, Java).
- The most heavily used software tools were Trinity, various genome assemblers, and Bowtie; bitseq was the single unused package.
- 67% of respondents indicated a scientific paper was in preparation.
- Just over half of the respondents have current funding from the NSF.

These results are based on responses from 27 clients to date. The survey remains open for additional responses as of the time this proposal is submitted.

Input from Science Advisory Board. The NCGAS Science Advisory Board (SAB) comprises the following individuals:

- Gwen Jacobs (U. Hawaii), Chair
- Michael Lynch (Indiana U.)
- Julie Dickerson (Iowa State U.)
- James Taylor (Johns Hopkins U.)
- Tatiana Foroud (IUPUI)
- Volker Brendel (Indiana U.)
- Haixu Tang (Indiana U.)

The SAB provided the following feedback in their PY3 4th-quarter review. Cyberinfrastructure resources and bioinformatics consulting services, particularly for large memory applications like assembly, were very valuable to biological researchers, notably those at smaller institutions without such resources. They recommended more effort in reaching underserved communities, more user-friendly services, and more online training materials, and suggested that NCGAS should examine cloud computing models for future service delivery. They considered the software comprehensive for existing users, but recommended installing software for population genomics and metagenomics research.

Changes from original plans and services since proposal writing. NCGAS has been quick to react to changing community needs. The following new services implemented under the development award were unanticipated at the time of writing:

- Intensive support for and code improvements in the Trinity RNA-seq software
- Providing archival storage via the IUScholarWorks digital repository
- Implementation and Support for Galaxy, including data transfer and file transformation
- Support for execution of BLAST jobs to the OSG.

We have eliminated (or plan to eliminate) a few services from the initial proposal, including a source code repository. Existing repositories sufficiently met this need for our users. At the end of 2014, we will end our partnership with Penguin Computing Inc. as Mason and Blacklight resources (at no cost to the user) have resulted in a lack of user interest in the Penguin computing system.

The services used by survey respondents are shown in Table 6.

Table 6. Summary of NCGAS user survey results.

Service	% utilization
Consulting (short-term and extended)	39%
Project support	32%
Mason use	86%
Storage	36%

Table 7. Comments made in free-text entry sections of NCGAS user survey and NCGAS responses to those comments.

User comment	NCGAS response
Data handling. Moving data to the system and file format handling is difficult.	Providing training materials for use of data moving software and hosting workshops dealing with specific file issues could lower barriers. Automated solutions can be explored through Galaxy.
Better documentation. The Knowledge Base is difficult to use or contains insufficient documentation, such as how to install Perl modules as a user.	Genomics-specific tutorials and documentation could provide a feasible solution to this issue. Presentations, outreach and training will increase awareness.
Extend functionality of our resources. Install more modules on Mason and add more built-in genomic databases to the Galaxy instance. Provide a location to upload files for use in genome browsers.	NCGAS is happy to take requests for software to be installed and functionality to be added. Users are urged to contact help@ncgas.org to resolve these issues.
Responsiveness of NCGAS staff is too low. Increase staffing.	NCGAS staff strives to provide prompt responses to all requests and to solve issues in a timely fashion. Proposals to increase staffing are currently entertained.
Better tools are needed for accessing Mason resources efficiently and effectively. Users have difficulty knowing the appropriate resources to request and waste time and compute resources.	A comprehensive guide to using specific software on our systems could be proposed. Benchmarking with popular tools and working with the SciAPT Group to thoroughly understand software usage patterns.
Policy alterations. Time limits on login nodes are too strict. Java cannot be run on login nodes.	NCGAS works closely with system administrators to find the best compromise between user needs and system health. We will reopen the issue on policy.

9.3 Sustainability

The NCGAS sustainability plan assumes a transition, over the next year, from a development to a sustained activity. Anticipating a move to sustaining status, NCGAS has reduced staffing from 2.75 to 1.35 NSF-funded FTEs. It is also pursuing additional grant funds for sustaining efforts, and is now in partnership with the Broad Institute for the development and provisioning of Trinity in service to the National Cancer Institute (NIH Grant U24 CA180922). The Center for Genomics and Bioinformatics (CGB) has an open, IU-funded position for system administration and bioinformatics. CGB and NCGAS intend to realign services to eliminate duplication. We will enable sustainability with better use of existing IU general funds. We will use existing IU funding for bioinformatics services (ongoing innovation and infrastructure), adding small subcontracts to NSF and NIH grant proposals led by researchers at IU and elsewhere for dedicated services for other large projects. The current internal cost per NCGAS allocated project is $17,805. We will begin selling our “allocated projects” services at a fixed cost of $25,000 per project beginning July 2018, based on incorporating IU’s 32% facilities and administration rate for contracts. On average, we expect to recoup just slightly more than our actual per-project cost, allowing

for a small amount of consulting at no cost to help young faculty become competitive for NSF funding. NCGAS education and outreach services will be integrated with overall PTI efforts, which already have a secure funding base. IU commits that NCGAS will be sustained after the end of the current proposed ABI sustaining grant award. IU can credibly make such a commitment, having recently transitioned the Pervasive Technology Institute to a sustainable financial basis after the end of its start-up funding

9.4 Progress on Program Year 3 (PY3) milestones

Progress on milestones for NCGAS are shown in Table 8.

Table 8. Accomplishment of NCGAS milestones in PY3.

NCGAS Progress on Year 3 Milestones

Quarter	Task	Notes
Q1 (Oct-Dec)	Assess Usage (2 new allocations requests, 99 short-term consults, 31 extended consults, 2 outreach events with 62 attendees)	
Q2 (Jan-Mar)	Assess Usage (8 new allocations requests, 213 short-term consults, 14 extended consults, 2 outreach events with 100 attendees)	
Q3 (Apr-Jun)	Assess Usage (1 new allocation request, 53 short-term consults, 7 extended consults, 7 outreach events with 320 attendees)	
	Genomics in June Workshop	
Q4 (Jul-Aug)	Assess Usage (4 new allocation requests, 51 short-term consults, 5 extended consults, 4 outreach events with 69 attendees)	
	PY 3 Satisfaction and needs survey sent out	
	Annual SAB meeting	New SAB established

10. Citations

Citations from the text are listed below, except for those designated as [NCGAS] references, which are found in Appendix 2: Scientific Products.

1. 1000 Genomes: A Deep Catalog of Human Genetic Variation. [14 October, 2014]; Available from: http://www.1000genomes.org/
2. 1000 Fungal Genomes Project [14 October, 2014]; Available from: http://1000.fungalgenomes.org/home/
3. i5kInsect and other Arthropod Genome Sequencing Initiative [14 October 2014]; Available from: http://www.arthropodgenomes.org/wiki/i5K
4. Genome 10k: Unveiling Animal Diversity [14 October, 2014]; Available from: https://genome10k.soe.ucsc.edu/
5. The avian phylogenomic project [14 October, 2014]; Available from: http://gigadb.org/dataset/101000
6. Fish-T1K: Transcriptomes of 1,000 Fishes. [14 October, 2014], Available from: http://www.fisht1k.org/
7. Earth Microbiome Project. [14 October, 2014]; Available from: http://www.earthmicrobiome.org/
8. NIH Human Microbiome Project. [14 October 2014]; Available from: http://www.hmpdacc.org/
9. MetaHIT: Metagenomics of the Human Intestinal Tract. [14 October 2014]; Available from: http://www.metahit.eu/
10. Canadian Microbiome Initiative. [14 October, 2014]; Available from: http://www.cihr-irsc.gc.ca/e/39939.html
11 Azvolinsky, A., Sequencing the Tree of Life, in The Scientist.
12. National Science Foundation. High Performance Computing System Acquisition: Enhancing the Petascale Computing Environment for Science and Engineering. [31 Jan 2011]; Available from: http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503148.
13. Pordes, R., D. Petravick. B. Kramer, D. Olson , M. Livny, A. Roy, P. Avery, K. Blackburn, T. Wenaus, F. Würthwein, I. Foster, R. Gardner, M. Wilde, A. Blatecky, J. McGee, R. Quick, The Open Science Grid. Journal of Physics Conference Series, 2007. 78: p. 012057.
14. Indiana University Pervasive Technology Institute. Software supported by NCGAS. [12 Aug 2014]; Available from: http://ncgas.org/ncgas-software.php.
15. Indiana University Pervasive Technology Institute. Scholarly Data Archive. [16 Aug 2010]; Available from: http://pti.iu.edu/storage/sda.
16. Indiana University Pervasive Technology Institute. NCGAS Training Materials. [12 Aug 2014]; Available from: http://ncgas.org/training.php
17. HPSS - High Performance Storage System. Home page. [30 Apr 2013]; Available from: http://www.hpss-collaboration.org.
18. Indiana University Pervasive Technology Institute. Galaxy Services. [12 Aug 2014]; Available from: http://www.ncgas.org/galaxy.php
19. Texas Advanced Computing Center. Stampede. [21 Oct 2011]; Available from: http://www.tacc.utexas.edu/stampede.
20. Moore, G. BLAST on OSG provides a timesaving alternative for large-scale analysis. International Science Grid This Week. Available from: http://www.isgtw.org/feed-item/open-science-grid-news-blast-osg-provides-timesaving-alternative-large-scale-analysis
21. CiteULike. Group: NCGAS. 12 Aug 2014]; Available from: http://www.citeulike.org/group/19027
22. Indiana University Pervasive Technology Institute. National Center for Genome Analysis Support. [12 Aug 2014]; Available from: http://ncgas.org.

23. XSEDE. What is the XSEDE Yum Repository, and how do I use it? [12 Aug 2014]; Available from: http://www.xsede.org/web/xup/knowledge-base/-/kb/document/bdwx.
24. Homebrew. Linuxbrew. [12 Aug 2014]; Available from: http://brew.sh/linuxbrew.
25. Blue Collar Bioinformatics. bcbio-nextgen. [12 Aug 2014]; Available from: https://bcbio-nextgen.readthedocs.org/en/latest.
26. Cyberinfrastructure for Phylogentic Research. Home page. [12 Aug 2014]; Available from: http://www.phylo.org.
27. Brinqa. Home page. [12 Aug 2014]; Available from: http://brinqa.com.
28. Indiana University Pervasive Technology Institute. Home page. [12 Sept 2011]; Available from: http://www.pti.iu.edu.

11. Appendix 1. Research projects that received allocations and support from NCGAS during PY3.

The primary goal of NCGAS is to provide consulting, software, and computational support to existing genomics and life science projects. This section outlines each project supported in PY3: first, the 23 NSF-funded projects in order of their request for support, and then the remaining nine projects.

11.1. New NSF-funded Projects

PI: Carlos A Machado

Institution: University of Maryland

State: MD

Funding Organization: NSF

Award #: 1330766

Amount: $510,000

Title: Evolutionary Genomics of Long Intergenic Non-coding RNAs (lincRNAs) in Drosophila

Date initiated: 2013-09-15

Date completed: 2016-09-14

The importance of non-coding RNAs is becoming increasingly apparent for a wide range of biological processes, but they are still a relatively new topic, and many aspects of their function and evolution remain poorly understood. Long intergenic non-coding (linc) RNAs are perhaps least understood among the non coding RNAs, yet they are known to play roles in dosage compensation, embryonic development, stem cell maintenance and differentiation, and epigenetic regulation of expression. The proposed work is focused on lincRNAs across the species Drosophila pseudoobscura and Drosophila persimilis. RNA-seq will be used to catalog changes in lincRNA expression during development, and these data will be used to evaluate divergence in lincRNA sequence and expression patterns. Computer resources from NCGAS will be important to help with transcriptome assembly, read mapping and possibly data archival later in the project.

PI: Thomas M. Williams

Institution: University of Dayton

State: OH

Funding Organization: NSF

Award #: 1146373

Amount: $464,900

Title: Collaborative Research: The structure, function, and evolution of a regulatory network controlling sexually dimorphic fruit fly development

Date initiated: 2013-09-27

Date completed: 2014-12-31

Gene networks are fundamental to animal development, and though the complexity of these networks has been mapped in model organisms, the critical connection between network evolution and organismal diversity remains unclear. This project provides a unique perspective to these complex biological networks by investigating how new fruit fly pigmentation patterns were achieved through the modification of connections between pivotal members of a pigmentation gene network. Using candidate gene and genome-scale approaches it will be determined how a key regulatory protein is connected to and thereby controls the utilization of a battery of target genes necessary to make a pigmentation pattern. Furthermore, by comparing network connections between both related species and populations within a species that exhibit different pigmentation patterns it will be revealed how these connections evolved and the network effects resulting from natural variation in the production of this key regulatory protein. As an expansion to the scope of this project we propose use RNA-seq to identify differentially expressed genes between males and females for two species. Additionally, we propose to compare gene expression between these two species to determine whether similar of different genes show sex-specific patterns of expression.

PI: Thomas D. Kocher

Institution: University of Maryland

State: MD

Funding Organization: NSF

Award #: 1143920

Amount: $909,999

Title: Genomic Architecture of an Adaptive Radiation

Date initiated: 2013-11-11

Date completed: 2016-01-31

This research project considers the genetic mechanisms of vertebrate diversification from two perspectives. The first is that of population genetics. How many genes experience divergent selection during the evolution of new species? How are they distributed across the genome? The second perspective is that of the evolution of development. Which genes are modified to produce new phenotypes? The flock of some 800 closely related cichlids fishes in Lake Malawi are an ideal model system for this study. These species differ in the shape of their jaws and teeth, the color of their skin, and their mechanisms of sex determination. The genomes of 40 Metriaclima populations will be sequenced to quantify the proportion of the genome that has diverged among sister species. The genomes of 20 Labeotropheus populations will be sequenced and compared to Metriaclima to identify regions of the genome that have diverged between these genera that differ in craniofacial morphology. These data will provide the most comprehensive picture ever produced of the genomic architecture of speciation and adaptation in vertebrates. Analysis of these data will improve our understanding of the gene network controlling the development of neural crest cells, which contribute to a wide variety of pigmentation and craniofacial phenotypes unique to vertebrates. The data generated from this project is very large (over 50TB so far). NCGAS resources will provide backup of raw and analyzed data as well as providing much compute time required to analyze the data generated.

PI: Farrah Bashey-Visser

Institution: Indiana University

State: IN

Funding Organization: NSF

Award #: 0919015

Amount: $464,905

Title: Assessing the maintenance of diversity in microbial communities

Date initiated: 2013-12-16

Date completed: 2014-08-31

This work focuses on Xenorhabdus bacteria, which are insect pathogens and nematode mutualists. We have found multiple, behaviorally distinct clones of two different species existing over a relatively small spatial scale. We are interested in understanding the how these different behavioral types can persist and have evolved. One major phenotype that we have explored is bacteriocin production. Bacteriocins are antibiotic compounds that are noted for their ability to kill closely related strains. We have found that bacteriocin-based antagonisms are important in determining competitive outcomes and in altering infection dynamics (Bashey et al 2012). In addition, we see that in the absence of bacteriocin-based interactions, faster killing parasites are competitively dominant (Bashey et al. 2011). These results suggest there may be multiple strategies for success in this community (Bashey et al. 2013). The first project we would like NCGAS assistance in is in describing the bacteriocin cluster of one of our natural isolates. Based on other species and a genetic mutant of one our isolates, antagonistic behavior is due to a phage-tail like bacteriocin, encoded for by a prophage cluster of approximately 30 genes. We would like to examine the structure of this region and compare it to the other species publically available. Moving forward, we plan to sequence a number of our field isolates. We would like NCGAS assistance in comparing these genomes to understand the clonal structure and degree of recombination among our isolates, as well as, in identifying candidate loci for differences in other competitive traits and virulence.

PI: Karen Carleton

Institution: University of Maryland

State: MD

Funding Organization: NSF

Award #: IOS-0841270

Amount: $799,956

Title: Proximate and ultimate causes of sensory system evolution

Date initiated: 2014-01-15

Date completed: 2014-06-30

We study the visual system of cichlid fishes. These fish have some of the most variable visual systems known and have seven cone opsin genes that produce visual pigments sensitive from ultraviolet to red wavelengths. Species differ by which of the opsin genes they express such that closely related species can have quite different visual sensitivities. We are working to identify the genetic network that controls differential expression of these opsin genes. We have performed QTL analyses based on crosses between species which differ in opsin expression. This has used RADseq to identify regions of the genome correlated with expression of a particular opsin gene (O'Quin et al 2012). We are now fine mapping in those regions and are analyzing a number of retinal transcriptomes of parents from our genetic cross, or cichlid transcriptomes from the Broad's cichlid genome project to identify genes that differ in either sequence or expression in the QTL regions. Having access to the Galaxy tools as well as some of the variant calling programs (SNPeff) would be helpful to our efforts.

PI: James Beets

Institution: University of Hawaii

State: HI

Funding Organization: NSF

Award #: EPS-0903833

Amount: $20,000,000

Title: Transcriptomic response of the coral holobiont to growth anomaly in Montipora capitata

Date initiated: 2014-01-15

Date completed: 2014-08-31

Coral are vitally important organism to coral reefs and are currently threatened by anthropogenic and natural factors. Coral disease is a growing concern, especially in light of the unknown effects of climate change on the severity and prevalence of coral disease across the globe. In order to understand and predict the effects of disease on coral viability, we must first understand disease pathology. This project takes an RNA-seq approach to understanding the pathology of growth anomaly in a common Hawaiian coral Montipora capitata. Growth anomaly is characterized by protruding lesions with abnormal skeletal and tissue structure, along with reduced density of polyps and symbiotic algae, although the pathology is not well understood. This study looks at healthy and diseased coral colonies to determine the proteins and biological pathways that are differentially expressed among healthy and diseased tissue. As a small institution, UH Hilo is unable to support the bioinformatic needs of this project including analysis and assembly of transcripts as well as conducting BLAST on a large scale and annotation and storage of data.

PI: William Bradshaw

Institution: University of Georgia

State: GA

Funding Organization: NSF

Award #: IOS-1258063

Amount: $169,254

Title: Geographic variation and comparative gene expression: Nature's Gift to resolving the connection between the daily clock and the seasonal timer

Date initiated: 2014-01-22

Date completed: 2014-12-31

Although insects are among the most abundant and diverse organisms on Earth, the basis of the physiological mechanism that times their seasonal activities remains contentious and poorly understood. Seasonal timing in arthropods is driven primarily by response to day length (photoperiodism). Unlike the circadian clock that controls hundreds of daily metabolic and behavioral events and is well defined at the molecular level, no single photoperiod gene has been definitively identified in any insect. The real question has been and still persists: What is the evolutionary genetic connection between the two physiological processes? The components necessary to answer this question are first, the molecular resources and skills to apply them and second, clear genetic variation of photoperiodic response among natural populations that permits a comparative approach over a climatic gradient. The latter is nature's gift

to the clocks community and is exquisitely presented by the pitcher-plant mosquito, Wyeomyia smithii. We are focusing on six populations of W. smithii that represent a gradient in length of the growing season and, concomitantly, a gradient of genetically determined photoperiodic response. Importantly, we include coastal and mountain populations in North Carolina that differ in seasonality but, because they are at the same latitude, experience the exact same day length.

PI: Karst Downey

Institution: Washington State University

State: WA

Funding Organization NSF

Award #: 1119000

Amount: $370,422

Title: The genetic basis of flower color polymorphism in Leavenworthia

Date initiated: 2014-01-24

Date completed: 2014-05-15

Expression levels of genes shown to yield yellow-colored compounds (carotenoids) will be measured following de novo assembly of transcripts taken from white and yellow morphs of a plant species (Leavenworthia stylosa). Differential expression analyses will identify mutations in candidate genes of the carotenoid biosynthesis pathway that potentially cause the shift in flower color. The results will help to explain the physiological processes causing a shift in flower color, which will provide insights that can be applied to the study of evolution for an important class of plant compounds with health benefits for humans.

PI: Gerald Wilkinson

Institution: University of Maryland

State: MD

Funding Organization: NSF

Award #: DEB-0952260

Amount: $464,367

Title: Origin and Evolution of Sex Chromosomes in Stalk-eyed Flies

Date initiated: 2014-01-27

Date completed: 2015-02-28

The project will utilize next-generation sequencing of testes transcriptomes and comparison to existing fly genome sequences to identify genes for six species of stalk-eyed flies that span 40 MYA of evolution. Comparative genomic hybridization to custom oligoarrays and/or comparative genomic sequencing of males and females will be used to determine sex-linkage and gene movements. The effects of chromosomal gene translocation on gene expression and protein evolution will be assessed in a phylogenetic context. The scope and level of resolution provided by this analysis will rival what is currently available for Drosophila. Results from this work will assess the generality of the canonical theory for Y chromosome evolution and will have relevance to many additional questions in evolutionary

biology, including the genetic basis of meiotic drive, reproductive isolation, sex determination and dosage compensation.

PI: Andreas Madlung

Institution: University of Puget Sound

State: WA

Funding Organization: NSF

Award #: IOS-1339222

Amount: $625,623

Title: RUI: Comparative transcriptomic and proteomic analysis of phytochrome responses in tomato

Date initiated: 2014-02-01

Date completed: 2016-01-31

Plants respond both to exogenous and endogenous signals to optimize processes that allow them to access water and nutrients from the ground and optimally orient their bodies for photosynthesis in three-dimensional space. Many developmental decisions are taken within the first hours of emergence of the seedling from the ground. These decisions are in direct response to the seedling's environmental conditions, particularly with respect to the available light. Both light quantity and -quality are sensed using elaborate light detector mechanisms and the information is translated into various growth responses. One of these light receptors is called phytochrome, which consists of a chromophore that is attached to an apoprotein. A small gene family encodes multiple types of phytochrome apoproteins that can homo- or heterodimerize and act as transcription factors, allowing for a multitude of physiological responses in the plant. This project will use a genome wide approach to directly compare transcriptional and proteomic changes in tomato seedlings and several phytochrome mutants during early development. We plan to (a) generate mutants via RNAi in phytochrome genes, for which mutants are not yet available, (b) to use both RNAseq and mass spectrometry to identify novel interactions of genes, both on the transcript and the protein level during the first 6 hours of seedling responses to light, and (c) functionally analyze the genes of two lesser-studied members of the phytochrome gene family. NCGAS resources are requested for the RNAseq portion of this project.

PI: Qiang Shawn Cheng

Institution: Southern Illinois University at Carbondale

State: IL

Funding Organization NSF

Award #: IIS-1218712

Amount: $254,661

Title: Pattern Learning in a Minimax Framework

Date initiated: 2014-03-11

Date completed: 2016-03-10

Large-scale or high-dimensional data are increasingly common in diverse domains such as science, engineering, medicine, and economics. Extracting useful patterns and finding pertinent features from such data are critical to building predictive models and discovering knowledge. Traditional data analytics have degraded performance or completely break down when dealing with such data due to high dimensionality as well as significant data uncertainty such as outliers, incomplete data, and distortions. Pattern learning under uncertainty thus is needed for extracting essential patterns, selecting latent features, or recognizing complex objects from large-scale data; nonetheless, necessary theory and methods for explicitly accounting for data uncertainty for pattern learning are still largely elusive. The central goal of this research is to fill this gap by establishing necessary theory and methods for reliable pattern learning that explicitly account for uncertainty.

PI: Kenneth Nephew

Institution: Indiana University

State: IN

Funding Organization: NSF

Award #: ARC-1142201

Amount: $299,994

Title: Methylation of Circulating Free DNA in Ovarian Cancer Patients

Date initiated: 2014-08-01

Date completed: 2016-08-01

Platinum-resistant ovarian cancer is uniformly fatal. Platinum resistance is associated with epigenetic anomalies including aberrant DNA methylation, a reversible epigenetic mark. We hypothesized that DNA methyltransferase inhibitors (DNMTIs) restore ovarian cancer sensitivity to platinum and our recent phase I/II trial showed that low-dose 5-aza-dC followed by carboplatin resulted in promising clinical activity in women with platinum-resistant ovarian cancer. However, current DNMTIs are rapidly degraded by hydrolytic cleavage, deaminated by cytidine deaminase and unstable during intravenous infusion, limiting their potential as cancer therapeutics. SGI-110, a dinucleotide combining 5-aza-dC and deoxyguanosine (Astex Pharmaceuticals, Inc.), is less prone to deamination and more stable. A Phase I clinical trial assessing the safety and tolerability of SGI-110 and carboplatin has recently been completed. Blood and tumor biopsy samples were collected at baseline and over the duration of the trial. The Phase II clinical trial is currently ongoing.

11.2. Projects receiving ongoing support during PY3 and initiated in PY1 and PY2

PI: Jacob Freimer
Institution: University of California, San Francisco
State: CA
Funding Organization: NSF
Award #: **1000122262**

Amount: $

Title: Mapping mRNA and RNA binding protein interactions
Date initiated: 2013-07-30
Date completed: 2015-07-30

We are using UV crosslinking and RNA sequencing to map RNA binding protein/mRNA interactions in early mouse development. We will use the NCGAS to store and analyze the sequencing data.

PI: Jeffrey Boore
Institution: Iowa State University
State: IA
Funding Organization: NSF
Award #: 1122176
Title: The Genomic Consequences of Asexuality
Amount: $ 679,068
Date initiated: 2013-7-11
Date completed: 2014-12-31

For my portion of this project, I must assemble a snail genome (420 MB). I have in hand for this 47X coverage in Illumina 2x100 and 12X coverage in PacBio. I must then create gene models using ab initio, homology, and RNA-seq data methods and reconcile this into a single gene set. I must map Illumina genome reads from four other closely related lineages onto this reference genome and characterize the variation among these.

PI: Ronald Burton
Institution: University of California San Diego, Scripps
State: CA
Funding Organization: NSF
Award #: 1155030
Amount: $412,635
Title: Collaborative Research: Ecological genomics of stress response in an intertidal copepod
Date initiated: 2013-07-01
Date completed: 2015-04-01

The marine copepod Tigriopus californicus has become a model system for studies of: 1) allopatric differentiation and the evolution of post-zygotic reproductive isolation, and 2) the physiology of response to environmental stress. Over the past few years, RNA-seq studies have rapidly advanced our understanding of transcriptional responses of the species to both interpopulation hybridization and stress response. This project involves the de novo sequencing of the T. californicus genome and resequencing of several geographic populations to address a variety of evolutionary questions. Access to NCGAS support will greatly facilitate our effort to obtain the best possible genome sequence of this organism - a key step in our proposed study.

PI: Mahdi Belcaid
Institution: University of Hawaii
State: HI
Funding Organization: NSF
Award #: 1260169
Amount: $665,062
Title: Multispecies connectivity: Comparative analysis of marine connectivity and its drivers for the coral reefs of Hawaii
Date initiated: 2013-06-20
Date completed: 2015-06-20

The exchange of individuals among populations, termed connectivity, is a central element of population persistence and maintenance of genetic diversity, and influences most ecological and evolutionary processes. To date, field studies of marine connectivity have necessarily focused on one or a few species at a time, providing little understanding of both the extent of variability in connectivity across a whole community and what factors drive that variability. This project will address these questions with population genetic datasets of a diverse marine fauna sampled across the Hawaiian Archipelago. By combining these genetic data with extensive oceanographic, ecological, and historical data, this project can potentially transform our understanding of the basis of the genetic structure of populations and the processes influencing genetic patterns. This project will provide unique, new knowledge to basic marine ecology and the science of Ecosystem-based Management, while incorporating the latest analytical and simulation approaches.

PI: J. Andrew DeWoddy
Institution: Purdue University
State: IN
Funding Organization: NSF
Amount: $3,686,667
Award #: DGE-1333468
Title: An evaluation of MHC based mate choice in captive koala (Phascolarctos cinereus)
Date initiated: 2013-07-12
Date completed: 2015-08-31

The koalas in the population at the San Diego Zoo are mating unpredictably (45% copulation success). When pairs fail to copulate, it wastes zoo resources and can lead to decreased genetic variability in the population. Mate choice based on Major Histocompatibility Complex (MHC) genotypes has been shown in multiple species including humans, mice, fish, birds, etc. Thus far, Population managers cannot account for MHC preferences when creating mating pairs for two main reasons: (1) the koala MHC has not been described in enough detail to develop an assay to genotype the individuals, and (2) MHC-based mate choice has not been studied in koala. The goal of this project is to (1) characterize the koala MHC, (2) design an MHC genotyping assay, and (3) compare genotype to copulation success. Two koala transcriptomes were sequenced (buffy coat and spleen) using next-generation sequencing. MHC transcripts (identified via Blast) were used to design PCR primers to create a marker assay for koala MHC. Once we have the assay, we will genotype all the individuals in the San Diego Zoo colony and investigate the effects of MHC genotype on copulation success. If a significant relationship is found, the results can be incorporated into the breeding scheme at the San Diego Zoo. If no significant relationship is found, this assay can provide a good foundation to investigate MHC-based mate choice in other marsupial species. We will also compare the buffy coat and spleen transcriptome data.

PI: Jeffrey Blanchard
Institution: Harvard University
State: MA
Funding Organization: NSF
Award #: 1237491
Amount: $1,960,004
Title: Microbial Forest Soil Community Dynamics
Date initiated: 2013-04-04
Date completed: 2015-04-04

Terrestrial ecosystems play a major role in controlling and steering the flow of the carbon cycle. Three quarters of the carbon in terrestrial ecosystems is found as organic matter in soils, most of which is derived from plant detritus. The complex relationships between plants and diverse soil microbes are not well understood. Soil microbial decomposers can either access and respire old C from soil (net ghg C emissions), access and respire new C from plants (no net C change), or store C in the soil through either direct or indirect methods (C granules or dead microbial bodies, resulting in net ghg C emissions). The ability to predict rates of substrate utilization, sequestration of stable organic molecules, and the release of greenhouse gases such as CO2 and CH4, which impact climate, depends on a deeper understanding of the interactions between microbial community members, their utilization of plant detritus, and subsequent feedbacks on plant growth. Our goal for this project period is to test the hypothesis that microbial community composition, and in turn function, control the response of plants to soil warming and ultimately ecosystem carbon cycling.

PI: Sean Patrick Mullen
Institution: Boston University
State: MA
Funding Organization: NSF
Award #: 1020136
Amount: $471,760
Title: Collaborative Research: The comparative genetics of wing pattern diversity in mimetic butterflies.
Date initiated: 10/16/12
Date completed: 10/30/14

Elucidating the genetic basis of adaptive phenotypic variation is central to our understanding of the origins and maintenance of biological diversity. One issue of particular importance is whether changes in homologous genes underlie the independent evolution of similar adaptive phenotypes. Butterflies display a massive array of color patterns, but much of this diversity appears to be a result of variation in the elements of a conserved wing pattern ground plan. The goal of our current grant is to greatly expand the scope of available comparisons by characterizing the genetic basis of phenotypic diversification across three of the most striking examples of wing pattern mimicry in butterflies. Specifically, we will identify the genes responsible for color pattern mimicry across Heliconius, Limenitis, and Papilio butterflies using a novel strategy that utilizes bulk segregant analyses paired with Illumina sequenced RAD tags, fine mapping, BAC contig sequencing, SNP discovery via genome resequencing, and association mapping in natural populations. We have made extraordinary progress in the last two years and have identified excellent candidate genes in all three systems. Moving forward, we need improved access to high-memory computational clusters to refine our BAC intervals using de novo assemblies based on a mixture of short-read and mate-pair libraries. In addition, we have generated a large number of resequenced data sets that need to be assembled against our BAC references, which will allow us to identify potential causal variants and will establish a core set of candidate SNPs for association mapping in natural populations.

PI: Endymion D. Cooper
Institution: University of Maryland
State: MD
Funding Organization: NSF
Award #: DEB-1036506
Amount: $420,300
Title: Collaborative Research: Assembling the Green Algal Tree of Life (GRAToL)
Date initiated: 2013-06-01
Date completed: 2015-12-01

The Delwiche lab component of the GrAToL project involves high-throughput sequencing of transcriptomic and genomic datasets. Analysis of this data involves assembly of transcriptomic and genomic datasets and high-throughput annotation using blast and similar methods to identify homologous genes.

PI: Bruce McClure
Institution: Colorado State University
State: CO
Funding Organization: NSF
Award #: MCB 1127059
Amount: $4,813,145
Title: GEPR: Deciphering Mechanisms of Prezygotic Reproductive Isolation in Solanum
Date initiated: 2013-08-09
Date completed: 2015-08-09

Intellectual Merit: Both plants and animals have evolved ways to prevent interbreeding between species. In other words, each species is reproductively isolated from other species because of reproductive barriers that prevent hybridization. The focus of this research is to understand the nature of reproductive barriers between species within the genus Solanum, which includes two important crop species: potato and tomato. In the previous funding period, the timing of reproductive barrier formation and site of barrier action in inter-species crosses was determined, and male and female genes involved in forming reproductive barriers were identified. In the current project, this information will be used to pursue the detailed molecular mechanisms that constitute inter-species recognition and rejection during mating attempts. Prior research has identified a population of S. habrochaites (a wild tomato species) with incipient reproductive barriers that isolate it from other populations. This system will now be developed as a powerful model for answering fundamental questions about how new species evolve. In the previous funding period, studies were conducted using tomato because it is an excellent model system for genetic and genomic studies. In the current project, research on reproductive barriers will be expanded into potato, an increasingly important crop worldwide.

Broader Impacts: Undergraduates, graduate students, and postdoctoral fellows in the research laboratories will participate in teaching undergraduate "Many Minds" laboratories in an Introductory Biology course; this will ensure that integration of research and teaching becomes second nature as their careers advance. Public outreach will also be a key component of the project. During this project three 90-second radio spots will be produced to air on the Public Radio Earth & Sky series (www.earthsky.org), which reaches about 15 million listeners. In addition, three podcasts will be produced for Earth & Sky and the project (www.irbtomato.org) web sites. One topic of these media efforts will be the importance of preserving wild germplasm, using tomato as an example, which should resonate with lay audiences. The project will also impact society by advancing crop improvement. The wild relatives of tomato and potato possess genes for resistance to pathogens, drought, cold, and salinity — traits that are particularly important in a time of global climate change. Unfortunately, accessing these agronomic traits is often prevented or impeded by reproductive barriers. This project will lead to understanding that will facilitate inter-species

crosses between domesticated and wild species by altering reproductive barriers, an advance that will greatly expand the genetic base for crop improvement to include resistance to disease and environmental stresses.

PI: Scott H. Harrison
Institution: North Carolina A&T University
State: NC
Funding Organization: NSF
Award #: DEB
Funding: $ none
Title: Analysis of Genomic Data
Date initiated: 2013-07-25
Date completed: 2015-07-25

I am studying ways to efficiently assemble tools and workflows for the analysis of genomic data. A guiding objective of this work is to better link the potential for discovery analysis provided by larger data sets to confirmatory, experimental research. This requires that the sensitivity and specificity of discovery algorithms be rigorously evaluated for relevance to experimental biology. This relevance is a function of the coverage of life's diversity across multiple levels of biological organization, and the predictive performance of algorithms. The cross-comparisons require that many-to-many relationships between genomic entities be calculated. The avalanche of genomic data requires significant data storage and supercomputing throughput capacity. Furthermore, based on available algorithms, some of these comparisons are very memory-intensive.

PI: Alice Barkan
Institution: University of Oregon
State: OR
Funding Organization: NSF
Award #: MCB-1243641;IOS-0922560
Amount: $4,498,196
Title: Deciphering the Code for RNA Recognition by PPR Proteins; and Macromolecular Networks Underlying Chloroplast Biogenesis
Date initiated: 2012-09-11
Date completed: 2015-09-11

This is the Abstract for the Pending Renewal of our PGRP grant. This project addresses a central problem in plant biology: the biogenesis, function, and environmental adaptation of chloroplasts. We will employ state-of-the-art methods and an extensive collection of photosynthetic mutants to discover new genes and elucidate regulatory mechanisms underlying chloroplast development, C4 differentiation, and photosynthetic homeostasis. Maize is chosen as the experimental organism because it offers a rich collection of chloroplast biogenesis mutants and the maize leaf blade is an excellent experimental system. The results will be of broad relevance, as the project employs novel approaches to fundamental questions that are not well understood in any organism. To assess the contribution of differential translation to the restructuring of the proteome that accompanies chloroplast biogenesis and C4 differentiation, this project will employ a genome-wide ribosome profiling method that has accelerated studies of translatomes in non-plant systems. Ribosome occupancy on cytosolic, plastid, and mitochondrial mRNAs will be profiled along a developmental series during leaf blade differentiation and separately in mesophyll and bundle sheath cells. The contribution of translational controls to the optimization of photosynthesis will be explored by profiling ribosomes after exposure to light regimes that trigger photosynthetic acclimation. A novel method that rapidly profiles plastid ribosomes will be used to address previously intractable questions: genome-wide effects of light on plastid translation initiation/elongation, and the coordinated

assembly and translation of plastid-encoded proteins. A large-scale forward genetic strategy that exploits IIlumina sequencing and a large collection of non-photosynthetic mutants will be used to discover new chloroplast biogenesis genes. Gene functions will be inferred with a unique phenomics pipeline that has been augmented by the ability to profile plastid ribosomes rapidly and at low cost. To elucidate chloroplast-to-nucleus signaling pathways, selected mutants in the collection will be analyzed by RNA-seq.

Intellectual Merit
The project will (i) assign molecular and physiological functions to ~100 chloroplast biogenesis genes in maize, including many novel genes whose orthologs have not been characterized; (ii) define the translatome dynamics underlying the installation of the photosynthetic apparatus and the distinct proteomes in BS and M cells; (iii) provide a comprehensive description of the progression of mitochondrial and plastid gene expression during the differentiation of photosynthetic leaf tissue; (iv) discover how regulated translation in the cytosol and chloroplast contribute to maintaining photosynthetic homeostasis under shifting light conditions.

Need for NCGAS resources:
i) We process and store a large quantity of Illumina data in our efforts to identify causal mutations underlying chloroplast biogenesis phenotypes. ii) We are developing a database resource for comparative genomics in plants that requires occasional intensive computation (e.g. computing a large number of phylogenetic trees). We had been using University of Oregon resources for these purposes, but new charges have been implemented that are prohibitive.

PI: Sarah Schaack
Institution: Reed College
State: OR
Funding Organization: NSF
Award #: 1150213
Amount: $523,966
Title: Upon Which Selection Can Act: Quantifying How Mutation and Environment Generate Genotypic & Phenotypic Variation in an Emerging Ecological & Evolutionary Genomic Model
Date initiated: 2013-07-04
Date completed: 2017-07-01

Current direct estimates of mutation rates at the genomic level are limited because: a) many types of mutations are ignored, b) mutation rates are estimated for a single genotype and extrapolated to the species level, c) not all mutations influence fitness-rendering phenotypic assays and sequence-based estimates asynchronously, and d) mutation rates and effects may vary in different environmental backgrounds. The goal of our work is to accurately quantify the rate and spectrum of spontaneous mutations in multiple genotypes from multiple populations, assess the influence of mutational variance on a range of traits, and assess this influence in multiple environments. We conduct this work using Daphnia, an emerging model system for ecological and evolutionary genomics.

PI: Vincent P Buonaccorsi
Institution: Juanita College
State: PA
Funding Organization: NSF
Award #: 1248096
Amount: $445,039
Title: RCN-UBE - GCAT-SEEK: The Genome Consortium for Active Undergraduate Research and Teaching Using Next-Generation Sequencing
Date initiated: 2013-06-07
Date completed: 2015-01-31

GCAT-SEEK is an emerging consortium of small-college faculty focusing on bringing next-gen sequencing technology to classrooms at primarily undergrad institutions. Access to these computer resources will greatly improve the speed and data size that can be used for these classroom modules. These resources will initially be used for a workshop for faculty from 9 different institutions, who will bring this training back to nearly 500 undergraduate students.

PI: James Giovannoni (Lukas Mueller)
Institution: Cornell University
State: NY
Funding Organization: NSF
Award #: 0820612
Amount: $10,407,225
Title: Petunia genome project Research and Teaching Using Next-Generation Sequencing
Date initiated: 2012-02-05
Date completed: 2014-12-31

Our NSF funded project is for the tomato genome sequence. We are collaborating with an independent Petunia sequencing project, with the goal to integrate the results in a comparative genomics approach on the SGN website (http://solgenomics.net) with the tomato data.

PI: Jeff Palmer
Institution: Indiana University
State: IN
Funding Organization: NSF
Award #: 1027529
Amount: $2,420,822
Title: The Geraniaceae genomes project: Accelerated and coordinated evolution across the three plant genomes
Date initiated: 2012-03-14
Date completed: 2016-12-31

Plant cells contain genomes in three distinct compartments, the mitochondrion, nucleus and plastid. Over time, thousands of genes transferred among these genomes and now there is extensive communication among the compartments and considerable conservation and stability of the genomes. The plant family Geraniaceae represents an important exception to this pattern because its mitochondrial and plastid genomes have experienced remarkably accelerated rates of change in gene content, gene order, and rates of nucleotide substitutions. The cause of these accelerated rates is unknown, but may be directed by genes encoded in the nucleus. This project investigates the basis for this accelerated evolution with the aim of understanding how different genomes within a cell can influence one another and co-evolve over time. The project will sequence, from 30 members of the Geraniaceae, the DNA in mitochondrial and plastid genomes and the genes expressed in the nucleus. These data will be analyzed to elucidate the mechanisms

of inter-compartmental crosstalk and co-evolution in plant cells. The goals of this project are to determine the extent of genomic upheaval in the mitochondrial and plastid genomes of the Geraniaceae and to identify the correlated changes in the nuclear genome that have driven this instability. The large and complex data sets generated in this project - 60 organelle genomes (some of them many Mb in size and full of repetitive DNA) and complete nuclear transcriptomes from 30 plants - require powerful, high-speed computing systems as available through NCGAS for efficient genome and gene assembly and analysis.

PI: Yuzhen Ye
Institution: Indiana University
State: IN
Funding Organization: NSF
Award #: DBI-0845685
Amount: $520,810
Title: Computational Protein Function Annotation for Metagenomics
Date initiated: 2012-04-14
Date completed: 2014-08-31

We are interested in the functional annotation of microbial organisms living in Human beings. To do this, we need to computationally analyze big sequencing data from different metagenomic projects. Thus, we need to use the NCGAS resources.

PI: Richard Ree
Institution: The Field Museum
State: IL
Funding Organization: NSF
Award #: 1119098
Amount: $399,424
Title: Phylogeny, biogeography, and diversification in Pedicularis (Orobanchaceae)
Date initiated: 2012-07-11
Date completed: 2015-09-01

The disproportionate abundance of species in mountains is a striking and mysterious pattern in global biodiversity. This project will unravel the evolutionary history of one of the largest genera of flowering plants, the louseworts, whose 770 species are found in mountain ranges across the Northern Hemisphere, but are especially rich in the Hengduan Mountains of China, the Altai-Tienshan of Russia, and the Himalayas. Phylogenetic relationships of a global sample of species will be reconstructed using DNA sequences. This "family tree" will then be used as an historical framework to test hypotheses about evolution and biogeography using additional data. For example, louseworts exhibit spectacular diversity in their flowers, but are pollinated only by bumblebees. Does competition between co-occurring louseworts for pollinator services cause evolutionary divergence in flower form and accelerate the splitting of ancestral species into distinct descendants? Other questions pertain to geographic origins, such as: is the Hengduan region an evolutionary "cradle" that favors new species formation, or is it a "museum" that harbors immigrants from other regions? Finally, the phylogenetic tree will be used to construct a natural classification for the lousewort genus, with taxonomic names reflecting lineages with common ancestry. This project will reconstruct a large, conspicuous, and enigmatic branch of flowering plants on the tree of life, and reveal historical patterns and evolutionary processes that have shaped the diversity and distributions of plant species across the Northern Hemisphere since the Miocene. Understanding these evolutionary dynamics is critical to conservation planning, e.g., to put future climate change in an historical context. The mountains where louseworts occur are particularly vulnerable to

climate change, raising the imperative to document these species and their evolutionary heritage. The research will also shed light on the evolution of floral form and function, and its contribution to the tempo and mode by which plant species coexist and proliferate.

PI: Gavin J.P. Naylor
Institution: College of Charleston
State: SC
Funding Organization: NSF
Award #: DEB-1132229
Amount: $2,240,878
Title: Collaborative Research: Jaws and Backbone: Chondrichthyan Phylogeny and a Spine for the Vertebrate Tree of Life
Date initiated: 2012-07-18
Date completed: 2016-09-30

The Chondrichthyans, or sharks, rays and chimaeras, are some of the best known marine animals in popular culture but poorly known in terms of their evolution. Despite being an ancient group, we know surprisingly little about the patterns and processes that gave rise to their current diversity. This project will provide an accounting of their diversity and genealogy of relationships among species based on DNA sequence comparisons. The project centers around the development of new technologies based on cross-species gene capture and next generation (Illumina) sequencing. The project has proposed to sequence about 1400 single copy exons that are shared across vertebrates for approximately 1000 different species, most of which are sharks and rays. Because there is no reference genome for any elasmobranch each of the 1400 exons will require de novo assembly of the Illumina reads. In the past we have used the Abyss platform to conduct our exon assemblies. However the ram requirements of de novo assembly are such that this aspect of the workflow has proven to be rate limiting to our progress. The facilities and help that are associated with NCGAS should make this aspect of our work flow much more efficient. We are looking forward to working with the infrastructure, staff and hardware resources offered by NCGAS.

PI: Thomas J. McGreevy, Jr.
Institution: University of Rhode Island
State: RI
Funding Organization: NSF
Award #: 1003226
Amount: $123,000
Title: A Landscape Genomics Approach to Identify the Genetic Mechanism of Adaptation in a Geographically Diverse Lizard
Date initiated: 2012-12-01
Date completed: 2014-12-01

The identification of the genetic basis of adaptation is a major objective of evolutionary biology that has broad implications for ecology and conservation biology. Recent advances in the population genomics and landscape genetics disciplines have greatly facilitated the identification of adaptive loci and allowed for the integration of environmental variables using a landscape genomics approach. This research merges spatially referenced genetic, morphological, and environmental data from a model system of Anolis lizards in the Caribbean using a Geographic Information Systems (GIS) framework. Anolis lizards are an ideal group to investigate the genetic mechanism of adaptation because they have diverged into over 400 species and are a classic example of adaptive radiation and speciation. The GIS-based analytical approach created will serve as a model framework for additional investigations of Anolis and other taxa. Our research objectives are to: 1) identify adaptive loci in A. marmoratus complex using population genomic

methods and identify their chromosomal position based on the annotated Green Anole (A. carolinensis) genome map; 2) determine if a similar set of adaptive loci are identified in the different A. marmoratus subspecies and if the adaptive loci correlate to morphological features and environmental variables; and 3) model the spatial distribution of adaptive and neutral loci on Guadeloupe to determine how the environment affects neutral and adaptive gene flow. The National Center for Genome Analysis Support facility would be used to analyze our genomic and environmental data.

PI: John R. Finnerty
Institution: Boston University
State: MA
Funding Organization: NSF
Award #: 0924749
Amount: $579,015
Title: LiT: Rel Homology Domain Signal Transduction Pathways in the Sea Anemone Nematostella vectensis
Date initiated: 2013-02-01
Date completed: 2014-12-31

Our current NSF-sponsored project explores the functional evolution of the NFÎºB signaling pathway in basal animals (sea anemones and corals; phylum Cnidaria). Using the starlet sea anemone, Nematostella vectensis (Nv), we have identified functionally important variation in the NFkB protein at both microevolutionary and macroevolutionary scales. All components of canonical NFkB signaling are present, but (1) the inhibitory domain of the NFkB protein has been split off from the ancestral protein and (2) there are two markedly different alleles of the protein found in natural populations that vary in their DNA binding activity and trans-activation. To reconstruct the functional evolution of NFkB signaling in cnidarians, we must isolate all core genes/proteins from outgroup taxa for functional testing and reconstruct the phylogenetic history of key evolutionary modifications of NFkB and its interacting proteins and cis-regulatory regions. Towards this end, we are sequencing the transcriptome and genome of three key outgroup taxa using Next Generation Sequencing: (1) a distinct genetic strain of Nv that exhibits a different variant of NFkB than the genetic variant whose genome was sequenced; (2) a closely related sea anemone, Edwardsiella lineata; and (3) the coral Astrangia poculata. The generated data amount to 50-100X coverage of the transcriptomes and genomes of these model cnidarians, but the assembly of hundreds of millions of paired 100bp reads exceeds the capacity of the blade server at Boston University (maximum 128GB RAM). To this end, we request NCGAS server access to complete assemblies using the De Brujin graph assemblers Velvet and Oases.

PI: Bastian Bentlage
Institution: University of Maryland
State: MD
Funding Organization: NSF
Award #: 1046075
Amount: $658,748
Title: Can evolutionary history predict how changes in biodiversity impact the productivity of ecosystems?
Date initiated: 2013-03-21
Date completed: 2014-08-31

The goal of this project is to predict which species extinctions will have the greatest effect on primary production, the fundamentally important process of photosynthetic capture of biomass. This project tests the novel hypothesis that evolution leads to genetic divergence and niche differences among species. In turn, niche differences lead to a “division of labor” that determines how efficiently biological

communities capture the limited resources needed to produce new biomass. To test this hypothesis, the project bridges the fields of ecology, phylogenetics, and genomics to examine how one of the most widespread and ecologically important groups of algae impacts the productivity of lakes throughout North America. The goals are to: 1) Create a new molecular phylogeny that can be used to test whether assemblages of freshwater algae are more genetically diverse than expected by chance. , 2) Experimentally manipulate the evolutionary and genetic divergence of species to assess how these aspects of biodiversity control niche differences and primary production. , 3) Conduct analyses of gene expression data to identify the genetic basis of niche differences among species, and relate these to rates of primary production by phytoplankton communities. For the latter objective we are investigating differential expression among several hundred Illumina rRNA seq libraries. To perform these analyses in a time-efficient manner we request resources at NCGAS. In particular, we would like to leverage parallel computing to cut down on the time it takes to map hundreds of short read datasets against reference transcriptomes.

PI: Alex Buerkle
Institution: University of Wyoming
State: WY
Funding Organization: NSF
Award #: 1050149
Amount: $253,756
Title: Genomic outcomes of repeated hybrid speciation
Date initiated: 2012-07-15
Date completed: 2015-07-15

We are using population genomic data to understand the extent to which the outcomes of speciation are repeated among different instances of hybrid speciation. This research is focused on Lycaeides butterflies. We are assembling a genome for the butterfly, and doing large scale resequencing. We are also doing large scale MCMC for Bayesian estimation of parameters of interest.

11.3. Research projects supported not with NSF funding, but in areas that NSF funds

PI: Sarath Chandra Janga

Institution: Indiana Univeristy
State: IN
Funding Organization: None, but in areas funded by NSF
Award #: NA
Title: Exploiting comparative genomics and metagenomics approaches for function prediction and understanding functional diversity
Date initiated: 2012-05-25
Date completed: 2017-05-25

This project will use the currently available vast amount of metagenomic data from different sources to 1) develop new methods for function prediction 2) develop rapid annotation systems for newly sequenced metagenomes (for public access) 3) understand diversity in the metagenomes for establishing criteria to identify new metagnomes to sequence.

PI: Keithanne Mockaitis

State: IN
Funding Organization: None, but in areas funded by NSF
Award #: NA
Title: Assembly and analyses of conifer transcriptomes for Pine Genome
Date initiated: 2012-06-18
Date completed: 2015-06-18

This work is sequencing and assembling the genomes of Pinus taeda (loblolly pine) as well as three additional conifers. Mockaitis is responsible for broad transcriptome generation for annotation and functional genomics. Most data are Illumina strand-specific paired sequence reads. These will be subjected to assembly and comparative analyses in progressive experiments. The first need for NCGAS resources is mapping of existing transcript assemblies we have to the newly prepared genome draft v0.6. This will inform quality and completeness of the genome assembly. Next new RNAseq assemblies will be prepared and compared to each other in a quantitative and qualitative manner to distinguish tissue specificities of gene expression and transcript processing. Later in the project we will incorporate small RNA data from additional experiments.

PI: Daniel S. Standage

State: IN
Funding Organization: None, but in areas funded by NSF
Award #: NA
Title: Next-generation innovations in gene and genome annotation
Date initiated: 2013-01-25
Date completed: 2015-01-25

My research interests involve genome informatics in general and gene annotation in particular. In the past our group has worked to provide community resources for comparative plant genomics, and our current collaborations have me taking the lead on assembling and annotating the genome of a non-model social insect species. Our need for HPC resources is driven not only by our support of these efforts, but our desire to innovate in this area. In particular, as the rate of genome sequencing continues to increase, we are working on scalable methods for simultaneously annotating homologous loci across tens, hundreds, or even thousands of related genomes.

PI: Volker Brendel

State: IN
Funding Organization: None, but in areas funded by NSF
Award #: NA
Title: Assembly and analysis of insect and plant genomes
Date initiated: 2013-01-01
Date completed: 2016-12-31

We are assembling insect genomes and transcriptomes using programs like Allpaths-lg and Trinity. We are annotating these and plant genomes using our workflows based on ab initio gene prediction and spliced alignment.

PI: Loren H. Rieseberg

State: IN
Funding Organization: None, but in areas funded by NSF
Award #: NA
Title: Comparative Genomics of Phenotypic variation in the Compositae
Date initiated: 2012-03-29
Date completed: 2017-01-01

Building on the recent advances of the Compositae Genome Project (CGP; http://compgenomics.ucdavis.edu/), this project will develop extensive resources for functional, comparative, and evolutionary genomics in the Compositae. This work, which will integrate genetic, phenotypic, and molecular evolutionary information, will address several major questions in crop and weed science as well as evolutionary biology. The project addresses multiple recommendations of the recent National Research Council report on the National Plant Genome Initiative including understanding processes of domestication and performance in various environments, developing models for accessing germplasm diversity for crop improvement, and providing multidisciplinary computational and wet lab training. The specific aims of this project are to: (i) sequence the gene space of the three most important crops in the Compositae (lettuce, sunflower, and safflower) and Gerbera, a model species for studying plant development, that represent the four major subfamilies within the Compositae; (ii) greatly increase the taxonomic coverage of the CGP's EST database using high throughput sequencing of cDNAs from 25 additional taxa, including six crop species, three weed species, the wild progenitors of ten crops and weeds, representatives of five taxonomically important subfamilies of the Compositae, and an outgroup (the Calyceraceae); (iii) establish the prevalence of copy number variation relative to nucleotide variation and phenotypic diversity using oligonucleotide arrays; (iv) study the effect of whole genome duplications on diversification rates; (v) identify genotypic changes driven by parallel selective pressures across crop and weed lineages; (vi) construct ultra-high density, transcript-based maps using single-feature polymorphisms (SFPs) of lettuce, sunflower and chicory, thereby facilitating detailed comparative analyses of genome evolution; (vii) develop permanent mapping populations (RILs) of key Compositae species to facilitate generation of similar transcript-based maps in other taxa; and (viii) use genetic map-based approaches and candidate gene analyses to dissect the genetic changes underlying multiple phenotypic transitions in the Compositae associated with domestication and the evolution of weediness.

12. Appendix 2. Scientific products

In PY3, the first scientific publications by biologists carrying out research aided and supported by NCGAS. NCGAS staff also published. Additionally, NCGAS staff published a number of papers and made presentations that helped communicate our accomplishments and strategies.

In the bibliographic listings below, NSF-funded PIs and Co-PIs who are or have been clients of NCGAS are listed in bold; students involved in projects receiving NCGAS support are listed in bold and italics. NCGAS staff and NCGAS collaborators at partner sites in XSEDE are listed in italics. In many cases, clients considered the intellectual contributions of NCGAS staff to be sufficiently valuable as to merit inclusion as a co-author.

Peer-reviewed publications

2012

[NCGAS1] Henschel, R., Lieber, M., Wu, L., Nista, P. M., Haas, B. J., LeDuc, R. D. (2012). Trinity RNA-Seq assembler performance optimization. In Proceedings of the 1st Conference of the Extreme Science and Engineering Discovery Environment: Bridging from the eXtreme to the Campus and Beyond, XSEDE '12, New York, NY, USA. ACM. doi:10.1145/2335755.2335842

2013

[NCGAS2] Boscaro, V., Felletti, M., Vannini, C., Ackerman, M. S., Chain, P. S., Malfatti, S., Vergez, L.,M., Shin, M., Doak, T. G., Lynch, M., Petroni, G. (2013). Polynucleobacter necessarius, a model for genome reduction in both free-living and symbiotic bacteria. Proceedings of the National Academy of Sciences (28 October 2013), 201316687+. doi:10.1073/pnas.1316687110

[NCGAS3] Christie, A. E., Roncalli, V., Wu, L.-S. S., Ganote, C. L., Doak, T., and Lenz, P. H. (2013). Peptidergic signaling in calanus finmarchicus (crustacea, copepoda): in silico identification of putative peptide hormones and their receptors using a de novo assembled transcriptome. General and comparative endocrinology, 187:117-135. doi: 10.1016/j.ygcen.2013.03.018

[NCGAS4] Haas, B., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P., Bowden, J., Couger, M., Eccles, D., Li, B., Lieber, M., MacManes, M., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C., Henschel, R., LeDuc, R., Friedman, N., and Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Protocols 8, 1494–1512 (2013) doi:10.1038/nprot.2013.084

[NCGAS5] LeDuc, R., Vaughn, M., Fonner, J.M., Sullivan, M., Williams, J., Blood, P.D., Taylor, J., and Barnett, W. (2013) Perspective: Leveraging the National Cyberinfrastructure for Biomedical Research, Journal of the American Medical Informatics Association. Accepted for their special issue on “big data”. doi:10.1136/amiajnl-2013-002059

[NCGAS6] LeDuc, R., Wu, L.-S., Ganote, C., Doak, T., Blood, P., Vaughn, M., Williams, B. (2013). National Center for Genome Analysis Support Leverages XSEDE to Support Life Science Research. Proceedings of XSEDE 13, San Diego, CA. http://dl.acm.org/citation.cfm?id=2484790

[NCGAS7] Jin, M., Bothfeld, W., Austin, S., Sato, T. K., La Reau, A., Li, H., Foston, M., Gunawan, C., LeDuc, R. D., Quensen, J. F., McGee, M., Uppugundla, N., Higbee, A., Ranatunga, R., Donald, C. W., Bone, G., Ragauskas, A. J., Tiedje, J. M., Noguera, D. R., Dale, B. E., Zhang, Y., and Balan, V. (2013). Effect of storage conditions on the stability and fermentability of enzymatic lignocellulosic hydrolysate. Bioresource technology, 147:212-220. doi: 10.1016/j.biortech.2013.08.018

[NCGAS8] Motamayor, J. C., Mockaitis, K., Schmutz, J., Haiminen, N., Iii, D.L., Cornejo, O., Findley, S. D., Zheng, P., Utro, F., Royaert, S,, Saski, C., Jenkins, J., Podicheti, R., Zhao, M., Scheffler, B. E., Stack, J. C., Feltus, F. A., Mustiga, G. M., Amores, F., Phillips, W., Marelli, J. P., May, G. D., Shapiro, H., Ma, J., Bustamante, C. D., Schnell, R. J., Main, D., Gilbert, D., Parida, L., Kuhn, D. N. The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color. Genome Biol. 2013 Jun 3;14(6):r53. PMID: 23731509

[NCGAS9] Stanton-Geddes, J., Paape, T., Epstein, B., Briskine, R., Yoder, J., Mudge, J., Bharti, A. K., Farmer, A. D., Zhou, P., Denny, R., May, G. D., Erlandson, S., Yakub, M. 2013. Candidate Genes and Genetic Architecture of Symbiotic and Agronomic Traits Revealed by Whole-Genome, Sequence-Based Association Genetics. Medicago truncatula. PLoS ONE 8(5): e65688. doi:10.1371/journal.pone.0065688

[NCGAS10] Swart, E. C., Bracht, J. R. , Magrini, V., Minx, P., Chen, X., Zhou, Y., Khurana, J. S., Goldman, A. D., Nowacki, M., Schotanus, K., Jung, S., Fulton, R. S., Ly, A., McGrath, S., Haub, K., Wiggins, J. L., Storton, D., Matese, J. C., Parsons, L., Chang, W. J., Bowen, M.S., Stover, N. A., Jones, T. A., Eddy, S. R., Herrick, G. A., Doak, T. G., Wilson, R. K., Mardis, E. R., Landweber, L. F. (2013). The Oxytricha trifallax macronuclear genome: a complex eukaryotic genome with 16,000 tiny chromosomes. PLoS Biol. 2013;11(1):e1001473. doi: 10.1371/journal.pbio.1001473. Epub 2013 Jan 29. PubMed

[NCGAS11] Zhang, Q., Doak, T.G., Ye, Y. “Expanding the catalog of cas genes with metagenomes.” Nucleic Acids Res. 2013 Dec 6. [Epub ahead of print] PubMed PMID: 24319142.

2014

[NCGAS12] Hayashi, S., Gesing, S., Quick, R., Teige, S., Ganote, C., Wu, Le-S., Prout, E. (2014). Galaxy based BLAST submission to distributed national high throughput computing resources. Presentation. Presented at the International Symposium on Grids and Clouds (ISGC) 2014 March 23-28. Academia Sinica,Taipei, Taiwan. http://hdl.handle.net/2022/18609

[NCGAS13] Lee, F. J., Rusch, D. B., Stewart, F. J., Mattila, H. R., Newton, I. L. G. (2014). “Saccharide breakdown and fermentation by the honey bee gut microbiome.” Environmental Microbiology. doi: 10.1111/1462-2920.12526.

[NCGAS14] Brown, J. B., Boley, N., Eisman, R., May, G. E., Stoiber, M. H., Duff, M. O., Booth, B. W., Wen, J., Park, S., Suzuki, A. M., Wan, K. H., Yu, C., Zhang, D., Carlson, J. W., Cherbas, L., Eads, B. D., Miller, D., Mockaitis, K., Roberts, J., Davis, C. A., Frise, E., Hammonds, A. S., Olson, S., Shenker, S., Sturgill, D., Samsonova, A. A., Weiszmann, R., Robinson, G., Hernandez, J., Andrews, J., Bickel, P. J., Carninci, P., Cherbas, P., Gingeras, T. R., Hoskins, R. A., Kaufman, T. C., Lai, E. C., Oliver, B., Perrimon, N., Graveley, B. R., and Celniker, S. E. (2014). Diversity and dynamics of the drosophila transcriptome. Nature, advance online publication. doi:10.1038/nature12962

[NCGAS15] Lenz, P. H., Roncalli, V., Hassett, R. P., Wu, L.-S., Cieslak, M. C., Hartline, D. K., and Christie, A. E. (2014). *De Novo* Assembly of a Transcriptome for Calanus finmarchicus (Crustacea,

Copepoda) – The Dominant Zooplankter of the North Atlantic Ocean. PLoS ONE 9(2): e88589. doi:10.1371/journal.pone.0088589

[NCGAS16] Wegrzyn, J.L., Liechty, J.D., Stevens, K.A., Wu, L-S., Loopstra, C. A., Vasquez- Gross, H. A., Dougherty, W. M., Lin, B. Y., Zieve, J. J., Martínez-García, P. J., Holt, C., Yandell, M., Zimin, A. V., Yorke, J. A., Crepeau, M. W., Puiu, D., Salzberg, S. L., de Jong, P. J., Mockaitis, K., Main, D., Langley, C. H., Neale, D. B. (2014). Unique Features of the Loblolly Pine (Pinus taeda L.) Megagenome Revealed Through Sequence Annotation. Genetics, 196(3), 891-909. PMCID: PMC3948814. http://www.genetics.org/content/196/3/891.full.pdf

[NCGAS17] LeDuc, R., Fellers, R. T., Early, B. P., Greer, J. B., Thomas, P. M., and Kelleher, N. L. (2014). The C-Score: A Bayesian Framework to Sharply Improve Proteoform Scoring in High-Throughput Top Down Proteomics. Journal of Proteome Research: 2014 Jul 3;13(7):3231-40. doi: 10.1021/pr401277r.

[NCGAS18] Neale, D., Wegrzyn, J., Stevens, K., Zimin, A., Puiu, D., Crepeau, M., Cardeno, C., Koriabine, M., Morris, A. H., Liechty, J., Garcia, P. M., Gross, H. V., Lin, B., Zieve, J., Dougherty, W., Soriano, S. F., Wu, L. S., Gilbert, D., Marcais, G., Roberts, M., Holt, C., Yandell, M., Davis, J., Smith, K., Dean, J., Lorenz, W., Whetten, R., Sederoff, R., Wheeler, N., McGuire, P., Main, D., Loopstra, C., Mockaitis, K., deJong, P., Yorke, J., Salzberg, S., and Langley, C. (2014). Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biology, 15(3):R59+. doi:10.1186/gb-2014-15-3-r59

Major peer-reviewed publications by unfunded NCGAS partners outside IU

2013

[NCGAS19] Pipes L., Li S., Bozinoski, M., Palermo R., Peng, X., Blood, P., Kelly, S., Weiss, J.M., Thierry-Mieg, J., Thierry-Mieg, D., Zumbo, P., Chen, R., Schroth, G.P., Mason, C.E., Katze, M.G. (2013). The non-human primate reference transcriptome resource (NHPRTR) for comparative functional genomics. Nucleic Acids Res. 41, D1, D906-14.

[NCGAS20] Youssef N. H,. Couger, M. B., Struchtemeyer, C. G., Liggenstoffer, A. S., Prade, R. A., Najar, F. Z., Atiyeh, H. K., Wilkins, M.R., Elshahed, M. S. The genome of the anaerobic fungus Orpinomyces sp. strain C1A reveals the unique evolutionary history of a remarkable plant biomass degrader. Applied and Environmental Microbiology 2013; 79(15):4620–4634.

2014

[NCGAS21] Blood, P.D., Marcus, S., & Schatz, M.C. (2014) Large-scale Sequencing and Assembly of Cereal Genomes Using Blacklight. In Proceedings of the Conference on Extreme Science and Engineering Discovery Environment (XSEDE'14). ACM, New York, NY, USA, DOI: 10.1145/2616498.2616502.

[NCGAS22] Couger, M.B., Pipes, L., Squina, F., Prade, R., Siepel, A., Palermo, R., Katze, M.G., Mason C.E., & Blood, P.D. (2014) Enabling large-scale next-generation sequence assembly with Blacklight. Concurrency and Computation: Practice and Experience, 26, 2157-2166, DOI: 10.1002/cpe.3231.

[NCGAS23] Pipes L. , et al. (2014) "Assessment and improvement of Indian-origin rhesus macaque and Mauritian-origin cynomolgus macaque genome annotations using deep transcriptome sequencing data" Journal of Medical Primatology. (accepted)

[NCGAS24] Ghaffari, N.; Sanchez-Flores, A.; Ryan, D.; Garcia-Orozco, K.D.; Chen, P. L.; Ochoa-Leyva, A.; Lopez-Zavala, A. A.; Carrasco, J. S.; Hong, C.; Brieba, L. G.; Rudino-Pinera, E.; Blood, P. D.; Sawyer, J. A.; Johnson, C. D.; Dindot, S. V.; Sotelo-Mundo, R.R.; Criscitiello, M. F. (2014) Improved transcriptome of the whiteleg shrimp (Litopenaeus vannamei), a dominant crustacean in global seafood mariculture. Nature Scientific Reports (provisionally accepted pending revision).

Publications in press at peer-reviewed journals

[NCGAS25] Chen, X., Bracht, J. R., Goldman, A. D., Dolzhenko, I., Clay, D. M., Swart, E. C., Perlman, D. H., Doak, T. G., Stuart, A., Amemiya, C. T., Sebra, R. P., Landweber, L. F. 2014. The architecture of a scrambled genome reveals massive levels of genomic rearrangement during development. Cell. DOI: 10.1016/j.cell.2014.07.034

[NCGAS26] McGrath, C. L., Gout, J. F., Doak, T. G., Yanagi, A., Lynch, M. (2014). "Insights into Three Whole-Genome Duplications Gleaned from the Paramecium caudatum." Genome Sequence Genetics. In Press.

[NCGAS27] McGrath, C. L., Gout, J. F., Johri, P., Doak, T. G., and Lynch, M. (2014). "Differential retention and divergent resolution of duplicate genes following whole-genome duplication." Genome Research. In Press.

[NCGAS28] Petersen, I. T., Bates, J. E., Dodge, K. A., Lansford, J. E., Pettit, G. S. (2014). "Describing and predicting developmental profiles of externalizing problems from childhood to adulthood." Development and Psychopathology. In press.

Publications in review at peer-reviewed journals

[NCGAS29] Kuhn, D. N., Dillon, N. L., Innes, D. J., Wu, L. S., Mockaitis, K. Development of Single Nucleotide Polymorphism (SNP) Markers from the Mango (Mangifera indica) Transcriptome for Mapping and Estimation of Genetic Diversity. 2014

[NCGAS30] Sanchez-Puerta, M. V., Zubko, M., Palmer, J. D. (2014). "The mitochondrial genome of a Solanaceae cybrid plant is highly chimeric and retains a single form of most genes." Plant Cell.

[NCGAS31] Uy, K., LeDuc, R., Ganote, C., Price, D. (2014). Physiological effects of heat stress on Hawaiian picture-wing Drosophila: genome-wide expression patterns and stress-related traits. CONPHYS 2014-052.

Technical reports (not peer-reviewed)

2012

[NCGAS32] Barnett, W. K., Shankar, G., Hancock, D. Y., Allen, M., Seiffert, K., Boyles, M., Rogers, J. L., Wernert, E., Link, M. R., Stewart, C. A. (2012). 2012 Annual Report - Advanced Biomedical Information Technology Core. http://hdl.handle.net/2022/15229

[NCGAS33] Bolte, J., Wernert, J., Seiffert, K., Simms, S. C., Link, M. R., Pierce, M., Marru, S., Hancock, D. Y., Miller, T., Stewart, C. A. (2012). Indiana University Pervasive Technology Institute – Research Technologies: XSEDE Service Provider and XSEDE subcontract report (PY1: 1 July 2011 to 30 June 2012). http://hdl.handle.net/2022/14702

[NCGAS34] Miller, T., Ping, R. J., Plale, B., Stewart, C. A. (2012). 2012 annual report on training, education, and outreach activities of the Indiana University Pervasive Technology Institute and affiliated organizations. http://hdl.handle.net/2022/17577

[NCGAS35] Stewart, C. A., Miller, T., Blood, P., Tillotson, J., Froelich, W., DeStefano, L., Rivera, L. (2012). XSEDE 12 Conference Final Report. Indiana University. http://hdl.handle.net/2022/18493

2013

[NCGAS36] Barnett, W. K., Ganote, C., Vaughn, M., LeDuc, R. D., Stewart, C. A. (2013). National Center for Genome Analysis Program Year 1 Report – September 15, 2011 – September 14, 2012. http://hdl.handle.net/2022/15340

[NCGAS37] Stewart, C. A., Barnett, W. K., Link, M. R., Shankar, G., Miller, T., Michael, S., Henschel, R., Boyles, M. J., Wernert, E., Quick, R. (2013). Services and support for IU School of Medicine and Clinical Affairs Schools by the UITS/PTI Advanced Biomedical Information Technology Core and Research Technologies Division in FY 2013 - Extended Version. http://hdl.handle.net/2022/17216

2014

[NCGAS38] Barnett. W. K., LeDuc, R. D., Stewart, C. A. National Center for Genome Analysis Program Year 2 Report – September 15, 2012 – September 14, 2013. (2014). Indiana University. http://hdl.handle.net/2022/17387

[NCGAS39] Barnett. W. K., Stewart, C. A. National Center for Genome Analysis Program Year 3 Report – September 15, 2013 – September 14. (2014). Indiana University. http://hdl.handle.net/2022/18513

[NCGAS40] Ping, R. J., Miller, T., Plale, B., Stewart, C. A. 2013 annual report on training, education, and outreach activities of the Indiana University Pervasive Technology Institute and affiliated organizations. (2014). Indiana University. http://hdl.handle.net/2022/17581

Presentations

2011

[NCGAS41] Stewart, C. A. 2011. XSEDE Campus Bridging. (Presentation) CASC (Coalition for Academic Scientific Computation) Meeting. Arlington, VA, 7 Sept 2011. http://hdl.handle.net/2022/13608

[NCGAS42] Stewart, C. A., Henschel, R., Barnett, W. K., Doak, T. 2011. Experiences with a large-memory HP cluster - performance on benchmarks and genome codes. (Presentation) HP- CAST 17 - HP

Consortium for Advanced Scientific and Technical Computing World-Wide User Group Meeting. Seattle, WA, 12 Nov 2011. http://hdl.handle.net/2022/13879

[NCGAS43] Stewart, C. A., Link, M. R., Turner, G., Barnett, W. K. 2011. Penguin Computing and Indiana University partner for 'above campus' and campus bridging services to the community. (Presentation) IEEE/ACM SC11 Conference. Seattle, WA, 14-17 Nov 2011. http://hdl.handle.net/2022/13880

[NCGAS44] Stewart, C. A., Wheeler, B. C. 2011. Cyberinfrastructure Begins at Home. (Presentation) IBM Multi- customer Briefings, IEEE/ACM SC11 Conference. Seattle, WA, 15 Nov 2011). http://hdl.handle.net/2022/13888

2012

[NCGAS45] Barnett, W. K. (2012). A Nation-Wide Area Networked File System for Very Large Scientific Data. (2012). (Presentation) Bio-IT World (Boston, MA, Apr 25 2012). Available from: http://hdl.handle.net/2022/14538

[NCGAS46] Barnett, W.K. (2012). High Performance Data Management and Computational Architectures for Genomics Research at National and International Scales. (Presentation) Bio-IT World Asia (Singapore, 7 June 2012). Available from: http://hdl.handle.net/2022/14540

[NCGAS47] Jacobs, M., Stewart, C.A. 2012. Penguin Computing / IU Partnership HPC "cluster as a service" and Cloud Services. (Presentation) Coalition for Academic Scientific Computation (Arlington, VA, 29 Feb). http://hdl.handle.net/2022/14441

[NCGAS48] LeDuc, R., Barnett, W. K. (2012). The National Center for Genome Analysis Support and Galaxy. Galaxy Community Conference (Chicago, IL, Jul 2012). http://hdl.handle.net/2022/14619

[NCGAS49] LeDuc, R. Careers in Computing and Science. Clark State University, Dayton OH, 2012-09-27. https://scholarworks.iu.edu/dspace/handle/2022/14745

[NCGAS50] LeDuc, R. D. (2012). Genomics, Transcriptomics, and Proteomics: Engaging Biologists. (Presentation). eScience (Oct 8 2012, Chicago IL). http://hdl.handle.net/2022/14746

[NCGAS51] Stewart, C.A. (2012). Campus Bridging. (Presentation) XSEDE (eXtreme Environment for Science and Engineering Discovery) Advisory Board Meeting (Chicago, IL, 23 Apr). Available from: http://hdl.handle.net/2022/14443

[NCGAS52] Stewart, C.A. (2012). Cyberinfrastructure Begins at Home. (Presentation) Rutgers University (New Brunswick, NJ, 20 Feb). http://hdl.handle.net/2022/14442

[NCGAS53] Stewart, C.A., Marru, S. Knepper, R., Hancock, D.Y., Wernert, J., Aikman, C., Bolte, J., Brown, P., Miller, T. M. 2012. Indiana University collected XSEDE update. (Presentation) XSEDE (eXtreme Environment for Science and Engineering Discovery) Quarterly Meeting (Austin, TX, 6-7 Mar). http://hdl.handle.net/2022/14444

2013

[NCGAS54] Barnett, W. K., LeDuc, R. D. (2013). Next Generation Cyberinfrastructures for Next Generation Sequencing and Genome Science. (Presentation). The AAMC 2013 Information Technology in Academic Medicine Conference (Jun 5-7 2013 Vancouver, BC). http://hdl.handle.net/2022/16668

[NCGAS55] Ganote, C. L., Doak, T. (2013). Intro to Bioinformatics - Assembling a Transcriptome (Presentation). Clark State student visit and workshop (June 12-13 2013 Bloomington Indiana). http://hdl.handle.net/2022/16682

[NCGAS56] Ganote, C., Doak, T. (2013). Intro to Using Galaxy for Bioinformatics. (Presentation). IU Galaxy for Bioinformatics Workshop, Indiana University (Sept 17 2013, Bloomington Indiana). http://hdl.handle.net/2022/17204

[NCGAS57] LeDuc, R., Barnett, W. K. (2013). National Center for Genome Analysis Support Leverages XSEDE to Support Life Science Research, XSEDE 13, 2013-07-23 San Diego, CA. https://scholarworks.iu.edu/dspace/handle/2022/16402

[NCGAS58] LeDuc, R. D., Barnett, W. K. (2013). The National Center for Genome Analysis and Support. (Presentation). Plant and Animal Genome Conference 2013 (Jan 12-16 2013, San Diego, CA). http://hdl.handle.net/2022/15283

[NCGAS59] LeDuc, R. (2013). Systems Biology Data Analysis.. Computational Approaches to Analyzing Microarray Data. Madison WI, July 14 2013. http://www.btci.org/courses/intermediate/caamd/2012/caamd12.html. Madison, WI

[NCGAS60] LeDuc, R. (2013). Leveraging the National Cyberinfrastructure for Top Down Mass Spectrometry. Consortium of Top Down Proteomics (Lightning talk), Concurrent with ASMS, 2013-06-08. Minneapolis, MN. https://scholarworks.iu.edu/dspace/handle/2022/16679

[NCGAS61] LeDuc, R. (2013). Using Prior Knowledge to Improve Scoring in High-Throughput Top-Down Proteomics Experiments. American Society of Mass Spectrometrists annual meeting, 2013-06-08, Minneapolis, MN. https://scholarworks.iu.edu/dspace/handle/2022/16680

[NCGAS62] LeDuc, R. (2013). Statistical Consideration for Identification and Quantification in Top-Down Proteomics. American Society for Mass Spectrometry, Sanibel Conference, St Pete Beach FL, 2013-01-27. https://scholarworks.iu.edu/dspace/handle/2022/15284

2014

[NCGAS63] Ganote, C., Hayashi, S. (2014). Galaxy Deployment on Heterogenous Hardware. (Presentation). Galaxy Community Conference 2014 (Jun 30 – Jul 2, 2014, Baltimore, MD). http://hdl.handle.net/2022/18500

[NCGAS64] Ganote, C., Wu, L., Doak, T. (2014). Moving Large Data to Galaxy. (Presentation). IU Bioinformatics Clinic, Indiana University (July 14-18 2014, Bloomington Indiana). http://hdl.handle.net/2022/18501

[NCGAS65] Ganote, C., Wu, L., Doak, T. (2014). Galaxy for Data Provenance. (Presentation). IU Bioinformatics Clinic, Indiana University (July 14-18 2014, Bloomington Indiana). http://hdl.handle.net/2022/18502

[NCGAS66] Ganote, C., Wu, L., Doak, T. (2014). RNA-Seq Demo on Galaxy. (Presentation). IU Bioinformatics Clinic, Indiana University (July 14-18 2014, Bloomington Indiana). http://hdl.handle.net/2022/18503

[NCGAS67] Haas, B. (2014). High-Performance De novo Transcript Reconstruction Leveraging Distributed Memory and Massive Parallelization. ftp://ftp.broad.mit.edu/pub/users/bhaas/TrinityBeta/BioIT_Trinity_bhaas_2014.pdf

[NCGAS68] Kuhn, D. (2014). Development of SNP markers from the mango (Mangifera indica L.) transcriptome for mapping and estimation of genetic diversity. Seminar by D. Kuhn (USDA-ARS) Plant and Animal Genome XXII Conference, San Diego, CA, USA, January 11, 2014 https://pag.confex.com/pag/xxii/webprogram/Paper11012.html

[NCGAS69] Langley, C. H. (2014). Loblolly Pine Genome v1.0. Plant Seminar. Plant and Animal Genome XXII Conference, San Diego, CA, USA, January 11, 2014 https://pag.confex.com/pag/xxii/webprogram/Paper10691.html

[NCGAS70] Mockaitis. (2014). K. Gene expression in conifers revealed on a comprehensive scale: Sequencing, assembly and classification of the loblolly pine transcriptome. Seminar by K. Mockaitis (Indiana University Dept. of Biology), Plant and Animal Genome XXII Conference, San Diego, CA, USA, January 11, 2014 https://pag.confex.com/pag/xxii/webprogram/Paper10694.html

[NCGAS71] Stewart, C.A. (2014). High Performance Computing serving Life Science Research Needs. Presentation. Presented at Universitaet zu Koeln, Koeln, Germany, 3 July 2014. http://hdl.handle.net/2022/18507

[NCGAS72] Stewart, C.A. (2014). Serving national scientific communities - genome analysis as an example. ZKI-Frühjahrstagung, Berlin, Deutschland. 25 March 2014. http://http://hdl.handle.net/2022/17383

[NCGAS73] Stewart, C.A. (2014). Cyberinfrastructure as a strategic university asset - for Hessian HPC Competence Center Leaders. Presentation. Presented at Technische Universitaet Darmstadt, Darmstadt, Germany, 30 June 2014. http://hdl.handle.net/2022/18476

[NCGAS74] Stewart, C.A. (2014). Information technology support for your local university community - Presentation for IT staff of TU-Darmstadt. Presentation. Presented at Technische Universitaet Darmstadt, Darmstadt, Germany, 30 June 2014. http://hdl.handle.net/2022/18475

Workshops

2014
[NCGAS75]. IU Bioinformatics Clinic. Held at Indiana University Bloomington, Bloomington, IN, July 14-18, 2014. http://ittraining.iu.edu/training/browse.aspx?workshop=BIOCL#workshop609

Posters

2012

[NCGAS76] Doak, T., LeDuc, R., Wu, L.-S., Stewart, C.A., Henschel, R., Barnett, W.K. (2012). The National Center for Genome Analysis Support (Poster). (Presentation) International Conference on Genomics in the Americas. Philadelphia, PA, 27 Sept 2012. http://hdl.handle.net/2022/14769.

[NCGAS77] Doak, T., LeDuc, R., Wu, L.-S., Stewart, C.A., Henschel, R., Barnett, W.K. (2012). The National Center for Genome Analysis Support (Poster). (Presentation) International Society of Protistologists -- North American Section. North Carolina Central University, Durham, NC, 22 Sept 2012. http://hdl.handle.net/2022/14769

[NCGAS78] Doak, T. G., Wu, L.-S., Stewart, C.A., Henschel, R., Barnett, W.K. (2012). National Center for Genome Analysis Support (Poster). (Presentation) Pacific Symposium on Biocomputing Kona, HI, Jan 5 2012. http://hdl.handle.net/2022/14539

[NCGAS79] Doak, T. G., Wu, L.-S., Stewart, C.A., Henschel, R., Barnett, W.K. (2012). National Center for Genome Analysis Support (Poster). (Presentation) Plant and Animal Genome Conference San Diego CA, Jan 16, 2012. http://hdl.handle.net/2022/14539

[NCGAS80] Doak, T., LeDuc, R., Wu, L.-S., Stewart, C.A., Henschel, R., Barnett, W.K. (2012). National Center for Genome Analysis Support. Poster presented at the New Directions for Investigating Biodiversity of Ciliates workshop, National Evolutionary Synthesis Center Durham, NC. September 19-22, 2012. http://hdl.handle.net/2022/14769

2013
[NCGAS81] Hallock, B. L. (2013). Cyberinfrastructure Resources for Bioinformatics Research. (Poster). Bio-IT World Expo (April 2013, Boston, MA). http://hdl.handle.net/2022/16601

Data products

2012
[NCGAS82] Trinity software: NCGAS contributed several software enhancements that are included in the standard distribution of software available from the definitive Trinity sourceforge repository: RNA-Seq De novo Assembly Using Trinity. First version of the Trinity software to contain these enhancements was version trinityrnaseq_r2012-06-08, and these enhancements remain in every release since that version. Available from: http://trinityrnaseq.sourceforge.net

2014
[NCGAS83] Mlrho software version 2.1. NCGAS contributed several software enhancements that are included in the standard distribution of software available from the definitive mlrho sourceforge repository CIPaGES/mlrho. (2014). https://github.com/CIPaGES/mlrho

[NCGAS84] Cannon, J.R., Cammarata, M. B., Robotham, S. A., Cotham, V.C., Shaw, J.B., Fellers, R. T., Early, B. P., Thomas, P. M., Kelleher, N. L., Brodbelt, J.S. (2014). Ultraviolet Photodissociation for Characterization of Whole Proteins on a Chromatographic Time Scale. http://hdl.handle.net/2022/17316

[NCGAS85] Durbin, K., Fellers, R., Ioanna, N., Kelleher, N., Compton, P. (2014). Autopilot: An Online Data Acquisition Control System for the Enhanced High-throughput Characterization of Intact Proteins. http://hdl.handle.net/2022/17234

[NCGAS86] Li, Y., Compton, P. D., Tran, J. C., Ntai, I., Kelleher, N. L. (2014). Optimizing capillary electrophoresis for top-down proteomics of 30-80 kDa proteins. (Wiley VCH (Proteomics), http://hdl.handle.net/2022/17235

Press Releases

2012

[NCGAS87] IU partnership results in faster Trinity RNA sequencing software. http://newsinfo.iu.edu/news/page/normal/22885.html

2013

[NCGAS88] IU center partners on $4M grant to advance cancer research. http://itnews.iu.edu/articles/2013/iu-center-partners-on-4m-grant-to-advance-cancer-research.php

[NCGAS89] NCGAS Makes HPC a Mainstay Tool for Biologists. http://www.hpcwire.com/2013/08/22/ncgas_makes_hpc_a_mainstay_tool_for_biologists/

[NCGAS90] Genome analysis support center announces new initiatives to foster discovery. http://ncgas.org/news/press-release-011413.php

2014

[NCGAS91] NCGAS supports IU biologists who received $6.2 million to advance research on bacterial evolution. http://news.indiana.edu/releases/iu/2014/05/army-research-office-grant.shtml

[NCGAS92 IU genome center sheds light on North Atlantic fishery decline. http://news.indiana.edu/releases/iu/2014/05/army-research-office-grant.shtml

Web publications (IU Knowledge base documents available directly or through a searchable interface at kb.iu.edu)

[NCGAS93] Indiana University. (2014). Software supported by NCGAS. Indiana University Knowledge Base. https://kb.iu.edu/d/bdko

[NCGAS94] Indiana University. (2014). What is the National Center for Genome Analysis Support. (NCGAS)? Indiana University Knowledge Base. https://kb.iu.edu/d/bbhg

[NCGAS95] Indiana University. (2014). At IU, what is the policy about installing software on Mason?. Indiana University Knowledge Base. https://kb.iu.edu/d/bbzs.

[NCGAS96] Indiana University. (2014). At IU, what software is available on the research computing systems? Indiana University Knowledge Base. https://kb.iu.edu/d/bade.

[NCGAS7] Indiana University. (2014). Using PAUP* 4.0 on Quarry at IU. Indiana University Knowledge Base. https://kb.iu.edu/d/aybk

[NCGAS98] Indiana University. (2014). At IU, what research computing services are available? Indiana University Knowledge Base. https://kb.iu.edu/d/anrf

[NCGAS99] Indiana University. (2014). Research and high-performance computing. Indiana University Knowledge Base. https://kb.iu.edu/d/apes

[NCGAS100] Indiana University. (2014). Mason at Indiana University. Indiana University Knowledge Base. https://kb.iu.edu/d/bbhh

[NCGAS101] Indiana University. (2014). Policies regarding UITS research systems. Indiana University Knowledge Base. https://kb.iu.edu/d/avkh

[NCGAS102] Indiana University. (2014). Glossary. Indiana University Knowledge Base. https://kb.iu.edu/d/glos

[NCGAS103] Indiana University. (2014). What is SOAPdenovo, and where is it installed on XSEDE? Indiana University Knowledge Base. https://kb.iu.edu/d/batn

[NCGAS104] Indiana University. (2014). At IU, what supercomputer systems are available for academic research? Indiana University Knowledge Base. https://kb.iu.edu/d/alde

[NCGAS105] Indiana University. (2014). Indiana University. (2014). What is Velvet, and where is installed on XSEDE? Indiana University Knowledge Base. https://kb.iu.edu/d/bato

[NCGAS106] Indiana University. (2014). What kind of research is being done in the domain sciences using XSEDE digital services? Indiana University Knowledge Base. https://kb.iu.edu/d/avbu

[NCGAS107] Indiana University. (2014). At IU, how do I use NAMD on Big Red II, Quarry, or Mason? Indiana University Knowledge Base. https://kb.iu.edu/d/bdkr

[NCGAS108] Indiana University. (2014). What is GPG, and how do I use it to encrypt files on Quarry and Mason at IU? Indiana University Knowledge Base. https://kb.iu.edu/d/awio

[NCGAS109] Indiana University. (2014). How do I monitor memory and CPU usage on Quarry, Mason, or Big Red II? Indiana University Knowledge Base. https://kb.iu.edu/d/bedc

[NCGAS110] Indiana University. (2014). For batch jobs on Big Red II, Quarry, or Mason at IU, how do I specify the required parallel file systems?

[NCGAS111] Indiana University Knowledge Base. https://kb.iu.edu/d/baht

[NCGAS112] Indiana University. (2014). At IU, on Big Red II, Quarry, or Mason, how do I change my login shell or passphrase? Indiana University Knowledge Base. https://kb.iu.edu/d/avmj

[NCGAS113] Indiana University. (2014). How do I use the IU Cyberinfrastructure Gateway to monitor batch jobs on Big Red II, Quarry, and Mason? Indiana University Knowledge Base. https://kb.iu.edu/d/bdsj

[NCGAS114] Indiana University. (2014). On Big Red II, Mason, Quarry, and Rockhopper at IU, how do I use Modules to manage my software environment? Indiana University Knowledge Base. https://kb.iu.edu/d/bcwy

[NCGAS115] Indiana University. (2014). On Big Red II, Quarry, or Mason at IU, why is my job sitting in the queue, and when will it run? Indiana University Knowledge Base. https://kb.iu.edu/d/awgw

13. Appendix 3: Software supported by NCGAS

13.1 Bioinformatics software supported by NCGAS

Table 9 summarizes the bioinformatics software applications installed and supported on the Mason cluster in NCGAS PY3. In all cases, the software was developed outside of NCGAS and a production version was released by the development community. In most cases, software was added to Mason in response to user requests. The "Readiness Reuse Level" refers to the levels described in Marshall and Downs (http://earthdata.nasa.gov/sites/default/files/esdswg/reuse/Resources/library/Publications/2008_IGARSS-RRL-Paper.pdf).

Table 9. Bioinformatic software supported on the Mason cluster.

Software	Version	Installed and supported by NCGAS as of Sept 20, 2012	Installed and supported by NCGAS as of Sept 20, 2013	Installed and supported by NCGAS as of Aug 31, 2014	License Terms	Current Reusability Readiness Level	Software Development Plan	Current and Anticipated Uses	Source	NCGAS optimizations incorporated into dist	Comes with test suite
abyss	1.3.3	Yes	Yes	Yes	Limited agreement for academic use	9	Not yet determined	De novo assembly of DNA for metagenomics, comparative genomics and creation of draft genomes	http://www.bcgsc.ca/platform/bioinfo/software/abyss/releases/1.3.3	No	No
	1.3.3-openmpi	Yes	Yes	Yes	Limited agreement for academic use	9					
	1.3.6	No	Yes	Yes	Limited agreement for academic use	9					
	1.3.6-openmpi	No	Yes	Yes	Limited agreement for academic use	9					
	1.5.1-openmpi	No	No	Yes	Limited agreement for academic use	9				No	Yes
	1.5.2-openmpi	No	No	Yes	Limited agreement for academic use	9				No	Yes
allpathslg	41292	Yes	Yes	Yes	Free to use, change, distribute	9	Not yet determined	Whole-genome shotgun assembly using Illumina long and short insert libraries for greatest accuracy	ftp://ftp.broadinstitute.org/pub/crd/ALLPATHS/Release-LG/latest_source_code/	No	Yes
	43460	No	Yes	Yes		9					
	45684	No	Yes	Yes		9					
	48894	No	No	Yes		9					

amos	3.0.0	Yes	Yes	Yes	Artistic License	9	Not yet determined	Assembling, Validating, Comparative, Visualizing, and Scaffolding whole genome sequence data in a pipeline	http://sourceforge.net/projects/amos/files/amos/	No	Yes
arachne	3.2	Yes	Yes	Yes	Free to use, change, distribute	9	Not yet determined	Whole genome shotgun assembly of long Sanger reads	ftp://ftp.broadinstitute.org/pub/crd/ARACHNE/	No	Yes
bedtools	2.12	Yes	Yes	Yes	GNU GPL v2	9	Not yet determined	Discovery of correlated genomic features such as ESTs, polymorphisms, mobile elements, etc.	http://code.google.com/p/bedtools/downloads/list, http://arm.koji.fedoraproject.org/koji/packageinfo?packageID=10644	No	Yes
	2.20.1	No	No	Yes	GNU GPL v2	9	Not yet determined	Discovery of correlated genomic features such as ESTs, polymorphisms, mobile elements, etc.	http://code.google.com/p/bedtools/downloads/list, http://arm.koji.fedoraproject.org/koji/packageinfo?packageID=10644	No	Yes
bio3d	1.1-4	No	Yes	Yes	GNU GPL v2	9	Not yet determined	R package containing utilities for the analysis of protein structure, sequence and trajectory data.	https://bitbucket.org/Grantlab/bio3d/	No	No
bioconductor	2.10	No	Yes	Yes	GPL-2 + file LICENSE	9	Not yet determined	R package for the analysis and comprehension of high-throughput genomic data.	http://www.bioconductor.org/install/	No	No
blat	35	Yes	Yes	Yes	Free for academic, non-profit or personal use. Contact for commercial licensing	9	Not yet determined	Fast alignment of highly similar sequences of DNA/Proteins to find ESTs or to align reads to reference	http://users.soe.ucsc.edu/~kent/src/	No	No
bowtie	0.12.8	No	Yes	Yes	Artistic License	9	Not yet determined	Alignment of short reads to a reference genome in order to approximate coverage, find polymorphisms, and assess assembly quality	http://sourceforge.net/projects/bowtie-bio/files/bowtie/	No	No
	2.1.0	No	Yes	Yes	GNU GPL v3	9			http://sourceforge.net/projects/bowtie-bio/files/bowtie2/	No	No
	2.2.3	No	No	Yes	GNU GPL v3	9			http://sourceforge.net/projects/bowtie-bio/files/bowtie2/	No	No
bwa	0.6.2	No	Yes	Yes	GNU GPL v3, MIT License	9	Not yet determined	Alignment of long and short reads from a variety of technologies, allows gaps, for approximating coverage, finding polymorphisms, and assessing assembly quality	http://sourceforge.net/projects/bio-bwa/files/	No	No
	0.7.2	No	Yes	Yes	GNU GPL v3, MIT License	9	Not yet determined		http://sourceforge.net/projects/bio-bwa/files/	No	No
	0.7.6a	No	No	Yes	GNU GPL v3, MIT License	9	Not yet determined		http://sourceforge.net/projects/bio-bwa/files/	No	No
	0.7.10	No	No	Yes	GNU GPL v3, MIT License	9	Not yet determined		http://sourceforge.net/projects/bio-bwa/files/	No	No

cd-hit	4.5.6	Yes	Yes	Yes	GNU GPL v2	9	Not yet determined	Clustering program for large sets of protein and DNA to determine relationships between many sequences	https://code.google.com/p/cdhit/downloads/list	No	No
celera	6.1	No	No	No	GNU GPL	9	Not yet determined	De novo assembly of whole-genome shotgun reads from a variety of sequencers using paired end reads at least 64bp long, for assembly of novel organisms and to incorporate multiple sources for greater accuracy	http://sourceforge.net/projects/wgs-assembler/files/wgs-assembler/	No	No
	7	Yes	Yes	Yes	GNU GPL	9	Not yet determined			No	No
cufflinks	2.0.2	No	Yes	Yes	Boost License	9	Not yet determined	Map RNA-Seq reads to reference genomes in order to annotate genes, discover splice variants, and estimate differential expression	http://cufflinks.cbcb.umd.edu/downloads/	No	Yes
	2.1.1	No	Yes	Yes	Boost License					No	Yes
cutadapt	1.2.1	No	Yes	Yes	MIT License (MIT)	9	Not yet determined	A tool of removing adapter sequences from high-throughput sequencing data.	http://code.google.com/p/cutadapt/downloads/list	No	No
cytoscape	2.8.3	No	Yes	Yes	GNU LGPL	9	Not yet determined	An open source software platform for visualizing molecular interaction networks and biological pathways	http://www.cytoscape.org/download.html	No	Yes
edena	2.1.1	Yes	Yes	Yes	Free to private and educational use, License included with software	9	Not yet determined	De novo assembly of short reads for smaller genome assembly	http://www.genomic.ch/edena.php	No	Yes
fastqc	0.10.1	No	Yes	Yes	GNU GPL v3 or later	9	Not yet determined	A quality control tool for high throughput sequence data.	http://www.bioinformatics.babraham.ac.uk/projects/download.html - fastqc	No	No
	0.11.2	No	No	Yes	GNU GPL v3 or later	9	Not yet determined		http://www.bioinformatics.babraham.ac.uk/projects/download.html - fastqc	No	No
galaxy	1	Yes	Yes	Yes	http://opensource.org/licenses/AFL-3.0	9	Not yet determined	A flexible GUI wrapper for bioinformatics tools allows users to manipulate genomic data and run analyses	mercurial install, see: http://wiki.galaxyproject.org/Admin/GetGalaxy	No	Yes
gatk	1.1-33	Yes	Yes	Yes	Free to use, change, distribute	9	Not yet determined	Suite of genomics analysis tools with a focus on variant calling and gene finding	New release only: http://www.broadinstitute.org/gatk/download	No	Yes
genomemapper	0.4.3	Yes	Yes	Yes	GNU GPL v3	8	Not yet determined	Short read alignment, allows gaps, allows multiple references; used for estimating coverage, finding polymorphisms, variant calling, and quantitative analysis	http://1001genomes.org/software/genomemapper.html	No	No
gmap	2012-11-09	No	Yes	Yes	Free to use and modify for own purpose; Copyright (c) 2005-2011 Genentech, Inc.	9	Not yet determined	Align cDNA to reference to determine gene structure and structural variants	http://research-pub.gene.com/gmap/archive.html	No	Yes
	2014-02-20	No	No	Yes		9	Not yet determined		http://research-pub.gene.com/gmap/archive.html	No	Yes
	2014-05-15	No	No	Yes		9	Not yet determined		http://research-pub.gene.com/gmap/archive.html	No	Yes

hmmer	3.0	No	No	Yes		9	Not yet determined	A tool for searching sequence databases for homologs of protein sequences, and for making protein sequence alignments	http://hmmer.janelia.org/software	No	No
khmer	1.0	No	No	Yes	Free to use, change, distribute	9	Not yet determined	A fast protocol for digitally normalizing large sets of short-read sequence data	https://github.com/ged-lab/khmer	No	No
macs	1.4.2	No	No	Yes	Artistic License	9	Not yet determined	Model-based Analysis of ChIP-Seq (MACS) on short reads sequencers such as Genome Analyzer.	http://liulab.dfci.harvard.edu/MACS/	No	No
maker	2.27-beta	No	No	Yes	Artistic License 2.0/ GNU GPL	9	Not yet determined	A genome annotation pipeline	http://www.yandell-lab.org/software/maker.html	No	No
metamos	1.1	No	No	Yes	Free to use	9	Not yet determined	an integrated assembly and analysis pipeline for metagenomic data	https://github.com/marbl/metAMOS	No	No
mlRho	1.7	No	No	Yes	GNU GPL v2	9	Not yet determined	Software for Estimating the Population Mutation and Recombination Rates from Shotgun-Sequenced Diploid Genomes	http://guanine.evolbio.mpg.de/mlRho/	Yes	No
	2.8	No	No	Yes		9	Not yet determined				
mothur	1.31	No	No	Yes	GNU GPL v3	9	Not yet determined	A tool for analyzing 16S rRNA gene sequences and can be easily used to analyze data generated by Sanger, PacBio, IonTorrent, 454 and Illumina	http://www.mothur.org/wiki/Download_mothur	No	No
	1.32	No	No	Yes		9	Not yet determined				
mummer	3.22	Yes	Yes	Yes	Artistic License	9	Not yet determined	Align very large DNA and Protein sequences to reference.	http://sourceforge.net/projects/mummer/files/mummer/	No	Yes
ninja	1.2.1	Yes	Yes	Yes	GNU LGLP	9	Not yet determined	Infers phylogeny using neighbor-joining tree	http://nimbletwist.com/software/ninja/download.html	No	No
namd	2.9	No	Yes	Yes	Free to research and educational use	9	Not yet determined	A parallel molecular dynamics code designed for high-performance simulation of large biomolecular systems	http://www.ks.uiuc.edu/Development/Download/download.cgi?PackageName=NAMD	No	No
novoalign	2.07.13	Yes	Yes	Yes	Free to use for non-profit projects and organizations	9	Not yet determined	Aligns short reads to reference genome for resequencing experiments	http://www.novocraft.com/main/downloadpage.php	No	No
	3.00.02	No	No	Yes		9	Not yet determined		http://www.novocraft.com/main/downloadpage.php	No	No
oases	0.2.08	No	Yes	Yes	GNU GPL v3	9	Not yet determined	De novo transcriptome assembler for very short reads	http://www.ebi.ac.uk/~zerbino/oases/	No	No
picard	1.52	Yes	Yes	Yes	Apache License v2, MIT License	9	Not yet determined	Provides tools and methods for manipulating sequence alignments for assembly quality assessment, variant calling, and downstream processing.	http://sourceforge.net/projects/picard/files/picard-tools/	No	No
raxml	7.2.6	Yes	Yes	Yes	GNU GPL v2	9	Not yet determined	Maximum likelihood phylogeny estimation for interpreting relationships between sets of data	http://www.exelixis-lab.org/	No	No
	7.2.8	Yes	Yes	Yes							

rsem	1.2.5	No	No	Yes								
sam2counts	1.0	No	Yes	Yes	GNU GPL	9	Not yet determined	Count number of mapped reads per reference in SAM files (often for RNA-Seq experiments)	https://github.com/vsbuffalo/sam2counts	No	No
samtools	0.1.18	Yes	Yes	Yes	BSD License, MIT License	9	Not yet determined	Provides tools and methods for manipulating sequence alignments for assembly quality assessment, variant calling, and downstream processing.	http://sourceforge.net/projects/samtools/files/samtools/	No	No
	0.1.19	No	Yes	Yes							
scythe	0.992-beta	No	No	Yes	MIT License	9	Not yet determined	A Naive Bayesian approach to classify contaminant substrings in sequence reads.	https://github.com/ucdavis-bioinformatics/scythe	No	No
shore	0.6.1beta	Yes	Yes	Yes	GNU GPL v3	9	Not yet determined	Pipeline for mapping short reads to reference genome for finding polymorphisms, variant calling, and quantitative analysis	http://sourceforge.net/projects/shore/files/Release_0.6/	No	No
smrt	1.3.1	Yes	Yes	Yes	GNU GPL v2, v3, GNU LGPL, MIT, Apache	9	Not yet determined	Analysis software specifically designed to support PacBio sequence: barcode handling and the HGAP de novo assembler are included	http://pacificbiosciences.github.com/DevNet/	No	Yes
	2.0.1	No	No	Yes		9	Not yet determined		http://pacificbiosciences.github.com/DevNet/	No	Yes
soapdenovo	1.04	Yes	Yes	Yes	GNU GPL v3	9	Not yet determined	De novo assembly of short reads for large genomes, creating reference genomes of novel organisms	Version 1 not available, see: http://sourceforge.net/projects/soapdenovo2/files/SOAPdenovo2/	No	No
	1.05	Yes	Yes	Yes							
	R240	No	No	Yes							
sra-toolkit	2.1.15	No	Yes	Yes	GPL v2 or greater	9	Not yet determined	A package of tools used to work with the Sequence Read Archive (SRA)	http://eutils.ncbi.nih.gov/Traces/sra/?view=software	No	No
	2.3.5-2	No	Yes	Yes		9	Not yet determined		http://eutils.ncbi.nih.gov/Traces/sra/?view=software	No	No
stacks	1.06	No	No	Yes								
tophat	2.0.5	No	Yes	Yes	Boost License	9	Not yet determined	Alignment for RNA-Seq data against reference for finding splice junctions	http://tophat.cbcb.umd.edu/downloads/	No	Yes
	2.0.7	No	No	Yes		9	Not yet determined		http://tophat.cbcb.umd.edu/downloads/	No	Yes
transabyss	1.3.2	Yes	Yes	Yes	BCCA Academic License	9	Not yet determined	Analysis for multiple transcript Abyss assemblies to find splice sites and variants	http://www.bcgsc.ca/platform/bioinfo/software/trans-abyss/releases/1.3.2	No	Yes
trinityrnaseq	10/05/12	No	Yes	Yes	BSD License	9	Open Source Development	De novo assembly of RNA-Seq data for differential expression and gene finding of novel organisms	http://sourceforge.net/projects/trinityrnaseq/files/	Yes	Yes
	02/05/13	No	Yes	Yes							
	04/13/2014	No	No	Yes							
	07/17/2014	No	No	Yes							
velvet	1.2.03-k111-openmp	Yes	Yes	Yes	GNU GPL v2	9	Not yet determined	De novo assembly of short reads with paired ends for smaller genome assembly of novel organisms	https://github.com/dzerbino/velvet	No	Yes
	1.2.08-k111-openmp	Yes	Yes	Yes							

	1.2.10-k111-openmp	No	No	Yes							

13.2 Technical descriptions of software supported by NCGAS and provided on the Mason cluster.

Table 10 provides reference information about the software installed on Mason in PY3.

Table 10. Technical properties of bioinformatic software supported on the Mason cluster.

Software	Version	Software Dependencies?	Software Development Methodology	Software Functionality	Resulting Publications
abyss	1.3.3 1.3.3-openmpi 1.3.6 1.3.6-openmpi 1.5.1-openmpi 1.5.2-openmpi	Requires C++ Boost, sparsehash and Open MPI of Short Reads	Not yet determined	Input: Single or Paired-End Read files in several supported formats; Output: Assembled contigs in Fasta format.	
allpathslg	41292 43460 45684 48894	Gcc, GMP, Picard, and graphviz	Not yet determined	Input: 100bp Illumina reads from short and long inserts; Output: Assembled contigs graph format	
amos	3.0.0	Gnu autoconf; subpackages require MUMmer, Boost and QT library	Open source development	Input: AMOS bank; Output: AMOS bank	
arachne	3.2	LaTeX, gzip, Xerces-C++ XML Parser	Not yet determined	Input: reads in Fasta format, quality scores, an xml ancillary tree, config file, and genome size file; Output: assembled bases in Fasta, assembled qualities, logs, reads, links, unplaced.	
bedtools	2.12 2.20.1	Gcc	Git Repository for Open Source	Input: Sequence format such as BED, GFF, BAM; Output varies by tool but can include text or BED file	
bio3d	1.1-4	R	Not yet determined	Input: reads in PDB, Fasta, AMBER Binary netCDF, CRD, PQR files; Output: reads in PDB, Fasta, AMBER Binary netCDF, CRD, PQR, PCS, NMA files	
bioconductor	2.10	R	Not yet determined	Input: Sequence format such as. Fasta BED, GFF, BAM; Output varies	
blat	35	none	Not yet determined	Input: Sequence query and database in Fasta, .nib or .2bit format; Output: Alignment in .psl format	
bowtie	0.12.8 2.1.0 2.2.3	Gcc	Open source development	Input: set of reads (Fasta, Fastq, paired or unpaired, raw, tabular) and an index; Output: list of alignments	
bwa	0.6.2 0.7.2 0.7.6a 0.7.10	Gcc	Not yet determined	Input: Query in Fastq format, database in Fasta format; Output: Alignments in .sai format	
cd-hit	4.5.6	Gcc	Not yet determined	Input: Fasta query sequence; Output: cluster file describing members of each cluster in clstr format	

celera	6.1 7	Gcc, kmer	Not yet determined	Input: FRG file containing sequence; Output: Assembled contigs in native ASM format, Fasta format. Other outputs for stats and mapping.	
cufflinks	2.0.2 2.1.1	Gcc, Boost, Samtools, Eigen libraries	Not yet determined	Input: Bam/Sam file; Output: GTF file with transcripts, FPKM tracking files for genes and transcripts	
cutadapt	1.2.1	Gcc python	Not yet determined	Input: Fastq file; Output: Fastq file	
cytoscape	2.8.3	none	Not yet determined	Input: networks and attributes in text format; Output: image or text files	
edena	2.1.1	none	Not yet determined	Input: Fasta or Fastq short reads file; Output: Assembled contigs (Fasta format?)	
Fastqc	0.10.1 0.11.2	Java	Not yet determined	Input: Fastq, Bam/Sam files; Output: Fastq, Bam/Sam files	
galaxy	1	Python, supported tools	Not yet determined	Input and output depend on the tool being used. Web interface or API available.	
gatk	1.1-33	Java, R	Not yet determined	Inputs: Fasta, Sam/Bam, ROD, or interval files as per tool; Output: SAM and VCF files	
genomemapper	0.4.3	Gcc	Not yet determined	Inputs: Reference genome in Fasta format, Shore files, Fasta or Fastq data queries; Output: Shore or Bed file with alignments	
gmap	2012-11-09 2014-02-20 2014-05-15	Gcc, Perl	Not yet determined	Input: Reference genome in Fasta format, Query in Fasta format; Outputs: Alignment file either compressed or uncompressed	
hmmer	3.0	Gcc	Open source development	Input: Sequence alignment in Stockholm format, profile HMM file, database sequence in Fasta format; Outputs: Text output	
khmer	1.0	python	Git Repository for Open Source	Input: Fasta or Fastq files; Output: Text output	
macs	1.4.2	python	Git Repository for Open Source	Input: ELAND, Bed, ELANDMULTI, ELANDEXPORT, ELANDMULTIPET, Bam/Sam; Output: Bed and Text output	
maker	2.27-beta	Phrap/cross_match, RMBLAST, RepeatMasker, SNAP, Exonerate, Augustus, perl	Not yet determined	Input: genome, est , and protein in Fasta format; Outputs: gene annotations in GFF3 format	
metamos	1.1	Perl, python, R, Gcc, curl, wget	Git Repository for Open Source	Input: Fasta, Fastq, or SFF files: Output: Assemblies in Fasta format and test output	
mlrho	1.7 2.8	Gcc, Gsl	Not yet determined	Input: Assembly in Fasta format; Output: Text output	
mothur	1.31 1.32	Gcc	Open source development	Input: Sequence in Fastq format; Output: Sequence in Fastq format, text output	
mummer	3.22	gcc, perl, g++, fig2dev, gnuplot, sfig sed, awk, ar, sh, csh	Open source development	Input: Reference genome in Fasta format, Query in Fasta format; Output: Text output	
ninja	1.2.1	Java	Not yet determined	Input: Sequence alignment in Fasta format; Output: Phylogenetic tree in Newick or Phylip format	
namd	2.9	Gcccd	Not yet determined	Input: PDB, PSF, configuration, and force field parameter files; Output: binary and text files	
novoalign	2.07.13 3.00.02	Bedtools, Samtools, Picard, GATK	Not yet determined	Input: Read files in Fastq or compressed format, Reference genome index created with novoindex; Output: Sam or tabular alignments	
oases	0.2.08	velvet	Not yet determined	Input: Read files in Fastq or compressed format; Output: Transcript assemblies	
picard	1.52	Java, Picard	Not yet determined	Input: Sam/Bam or URL, depending on tool; Output: Bam, Bam index, or text file with metrics	
raxml	7.2.6	Gcc	Not yet	Input: Phylip file containing alignments to be	

	7.2.8		determined	run; Output: Text files containing tree topologies, logfiles, intermediate files	
rsem	1.2.5	Gcc, perl, R	Not yet determined	Input: Reference sequence in Fasta or GTF format, sequence alignment in Sam/Bam format; Output: Text output, wiggle plot files, Fasta, Fastq files	
sam2counts	1.0	Python, cython, pysam	Git Repository for Open Source	Input: Sequence alignment in Sam format, Output: Text output	
samtools	0.1.18 0.1.19	Gcc	Not yet determined	Input: Sam or Bam file; Output: Varies by tool, Sam, Bam, .vcf, .afs files	
scythe	0.992-beta	Gcc	Git Repository for Open Source	Input: Sequence in Fasta format; Output: Trimmed sequence in Fasta format	
shore	0.6.1beta	Gcc, Boost, alignment software (genomemapper, bwa, bowtie), R	Not yet determined	Input: Reference genome in Fasta format, reads files in raw format; Output: Statistics, analysis results and quality assessments in a variety of formats	
smrt	1.3.1 2.0.1	Mysql, perl, bash, Java	Not yet determined	Input: Analysis is a GUI with multiple tools; sequence information stored in XML files; Output: XML and HTML results depending on tool	
soapdenovo	1.04 1.05 R240	none	Not yet determined	Input: Read files in Fasta, Fastq, and Bam, config file; Output: Contig assemblies and Scaffold assemblies	
sra-toolkit	2.1.15 2.3.5-2	Gcc	Not yet determined	Input: SRA file, Output: Fasta/Fastq file	
stacks	1.06	Gcc, Perl	Not yet determined	Input: Sequence in Fastq format; Output: Text output	
tophat	2.0.5 2.0.7	Gcc, Boost, Samtools	Not yet determined	Input: Reads in Fasta or Fastq format, Bowtie index database; Output: SAM alignment results, BED files with indel and junction results	
transabyss	1.3.2	BWA, Bowtie, Pysam, Samtools, Abyss, Blat, GMAP, Python, Perl, Anchor, xa2multi.pl	Not yet determined	Input: Input file specifying the assemblies to use, Reference genome file and gene annotations; Output: Bam and Bam index files for consensus assembly	
trinityrnaseq	10/05/2012 02/05/2013 04/13/2014 07/17/2014	Gcc, Samtools, Bowtie	Open source development	Input: Fastq files containing reads; Output: Assembled contigs in Fasta format	
velvet	1.2.03-k111-openmp 1.2.08-k111-openmp 1.2.100-k111-openmp	Gcc	Not yet determined	Input: Fasta, Fastq, Sam, Bam, eland, gerald; Output: Assembled contigs in Fasta format, stats file, AMOS file, and graph file.	

13.3 NCGAS software support across XSEDE resources.

Table 11 shows the level of shared support for bioinformatics software on the various NCGAS partner compute resources in PY2 that are coordinated and allocated through XSEDE.

Table 11. Bioinformatic software supported by non-Indiana University NCGAS partners.

Software	Mason	Quarry	Big Red 2	Blacklight	Stampede	Lonestar
abyss	X			X		X
afni	X					
allpathslg	X			X	X	X
amos	X				X	X
arachne	X				X	
bedtools	X	X	X	X	X	X
bio3d	X					
bioconductor	X	X				
bioperl	X				X	X
biopython	X					
bitseq	X					
blat	X	X		X	X	X
bowtie	X			X	X	X
bowtie2	X		X	X	X	X
bwa	X			X	X	X
cafe	X					
cdhit	X					X
celera	X					
cufflinks	X	X		X	X	X
cutadapt	X					
cytoscape	X					
edena	X				X	X
egglib	X					
euler	X					
fastqc	X			X	X	X
fsl	X	X				
galaxy	X					
gatk	X			X	X	X
genomemapper	X				X	X
gmap	X				X	X
hmmer	X	X		X	X	X
khmer	X					
macs	X	X	X			
maker	X					X
metamos	X					
mlrho	X					
mothur	X	X				
mothur_mpi	X	X				
mummer	X				X	X

namd	X			X	X	X
ngsutils	X	X	X			
ninja	X				X	X
novoalign	X				X	X
oases	X			X	X	X
picard	X				X	X
raxml	X	X			X	X
rsem	X			X		
sam2counts	X					
samtools	X	X	X	X	X	X
scythe	X				X	X
shore	X					X
smrt	X					
soapdenovo	X			X		
soapdenovo2	X			X	X	X
spm	X	X				
sratoolkit	X				X	X
stacks	X				X	X
tophat	X					X
tophat2	X			X	X	X
transabyss	X			X		
trinityrnaseq	X			X	X	X
velvet	X			X	X	X
velvetoptimiser	X					

14. Appendix 4: NCGAS education, outreach, and training activities

14.1. Press Releases

- NCGAS supports IU biologists who received $6.2 million to advance research on bacterial evolution. http://news.indiana.edu/releases/iu/2014/05/army-research-office-grant.shtml

- IU genome center sheds light on North Atlantic fishery decline. http://news.indiana.edu/releases/iu/2014/05/army-research-office-grant.shtml

14.2. Education, outreach, and training events and participants

Table 12. EOT activities for PY3 for NCGAS.

Type	Title	Location	Date	Hours	Number of Participants	Number of individuals from traditionally underserved groups (TUGs)*	Method†	Funding Sources
Conference talk/presentation/ panel	Internet2	Denver CO	4/7/14	1	100	0		
Tutorial / workshop	Joint School of Nanoscience and Nanoengineering	Greensboro NC	3/7/14	6	50	53		48-124-71
Tutorial / workshop	Society of Research Administrators (SRA) North Carolina Chapter Meeting	Pinehurst NC	3/3/14	24	50	53		48-124-71
Conference talk/presentation/ panel	NSF IGERT Training Grants Deep Genomics Symposium	Tuscon AZ	4/3/14	1	120	23		4012456
Conference talk/presentation/ panel	Midwest Protozoology meeting	St. Louis MO	4/12/14	8	40	26		48-124-71
Conference talk/presentation/ panel	SACNAS National Conference	San Antonio TX	10/2/13	32	50	71		48-124-71
Conference talk/presentation/ panel	Supercomputing 2013	Denver CO	11/17/13	32	12	6		48-124-71
Conference talk/presentation/ panel	Plant and Animal Genome Conference	San Diego CA	4/25/14	24	20	12		48-124-71

Academic meeting	IUPUI Medical School Career Week	Indianapolis IN	4/24/14	1	15	19		
Tutorial / workshop	IUS Campus Visit	New Albany IN	6/12/14	4	5	1		
Tutorial / workshop	NEON	Boulder CO	7/13/14	24	25	13		trip paid by NEON
Academic meeting	Genomics in July - CIB	Bloomington IN	7/25/14	1	18	9		NSF #1062432
Tutorial / workshop	Genomics in July - CIB	Bloomington IN	7/22/14	21	12	7		NSF #1062432
Tutorial / workshop	IU Bioinformatics Clinic 2014	Bloomington IN	7/16/14	8	4	3		
Conference talk/presentation/ panel	Galaxy Community Conference 2014	Baltimore MD	7/1/14	1	150	43		NSF #1062432
Tutorial / workshop	Galaxy Community Conference 2014	Baltimore MD	6/28/14	16	20	4		NSF #1062432
Totals	**16 Events**			204	691	343		

***Traditionally underrepresented groups as defined by the NSF.**

† **All events were conducted synchronously, e.g., in front of a live audience.**