National Center for Genome Analysis Program Year 2 Report – September 15, 2012 – September 14, 2013

William K. Barnett, Ph.D.
Richard D. LeDuc, Ph.D.
Craig A. Stewart, Ph.D.

Indiana University

PTI Technical Report PTI-TR14-004

25 February 2014

Citation:

Barnett, W.K., LeDuc, R.D., and Stewart, C.A. "National Center for Genome Analysis Program Year 2 Report – September 15, 2012 – September 14, 2013" Indiana University, Bloomington, IN. PTI Technical Report PTI-TR14-004, 2014.

PERVASIVE TECHNOLOGY INSTITUTE
INDIANA UNIVERSITY

RESEARCH TECHNOLOGIES
INDIANA UNIVERSITY
University Information Technology Services
Pervasive Technology Institute

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 although the National Center for Genome Analysis Support is funded by NSF 1062432, 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

National Center for Genome Analysis Program Year 2 Report – September 15, 2012 – September 14, 2013 i

1. Executive Summary iv

2. Introduction ix

3. Consulting and support for biological research in the US x

3.1. Consulting interactions – summary xi

3.2. Projects receiving significant support from NCGAS xii

4. Scientific Products xiv

5. Optimization, hardening, and enhancement of genome analysis software xv

6. Providing support to biologists delivery and assistance in use of supercomputer clusters xvii

6.1. Mason xvii

7. New service for PY2 - Long-term storage of data and Archiving of public data sets xix

8. New service for PY2 - Aid to researchers preparing grant proposals xix

9. Education and outreach xix

9.1. Management and Operations xx

9.2. User Survey xxi

9.3. Sustainability xxii

9.4. Progress on Program Year 2 milestones xxii

10. Appendix 1. Research projects that received extensive consultation and support from NCGAS during PY2. 1

10.1. New NSF-funded Projects 1

10.2. Projects receiving ongoing support during PY2 and initiated in PY1 11

10.3. Research projects supported – not with NSF funding, but in areas that NSF funds 15

11. Appendix 2. Scientific products 17

11.1. Biological research and bioinformatics papers published by NCGAS clients 18

11.1.1. Peer-reviewed journal technical papers – published or in press 18

11.1.2. Peer reviewed journal papers in-review 18

11.1.3. Journal papers in-preparation 19

11.1.4. Poster presentations 19

11.2. Methods-oriented and outreach papers, presentations, and materials by NCGAS staff 19

11.2.1. Peer-reviewed journals 19
11.2.2. Peer-reviewed conference papers 20
11.2.3. Posters 20
11.2.4. Presentations (not peer reviewed) 20
12. Appendix 3: Software Supported by NCGAS 21
12.1. Bioinformatics software supported by NCGAS 21
12.2. Technical descriptions of software supported by NCGAS and provided on the Mason cluster. 25
12.3. NCGAS software support across XSEDE resources 27
13. Appendix 2: NCGAS Education, Outreach, and Training Activities 28
13.1.1. Press Releases 28
13.1.2. Education, outreach, and training events and participants 29

Table of Tables

Table 1. Summary of scientific products created by scientists with the benefit of NCGAS support during PY2. xv
Table 2. Summary of outreach and education activities by NCGAS staff xx
Table 3. Summary of NCGAS user survey results. xxi
Table 4. Comments made in free-text entry sections of NCGAS user survey and NCGAS responses to those comments xxii
Table 5. Accomplishment of NCGAS milestones in PY2 xxiii
Table 6: Bioinformatic Software Supported on the Mason Cluster 21
Table 7: Technical Properties of Bioinformatic Software Supported on the Mason Cluster. 25
Table 8: Bioinformatic Software Supported by Non-Indiana University NCGAS Partners. 27
Table 9. EOT activities for PY2 for NCGAS. 31

Table of Figures

Figure 1. Harriet Alexander (left) and an unidentified helper setting up her incubation experiment on a three-week research cruise on the R/V Kilo Moana in the North Pacific Ocean. xi
Figure 2: Number of Tickets Reported by NCGAS Staff as a Function of the Time Needed to Complete the Ticket. xii
Figure 3: 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. xv
Figure 4: Total Users on the Mason System across PY2 xviii

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 second 12 months of NCGAS.

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

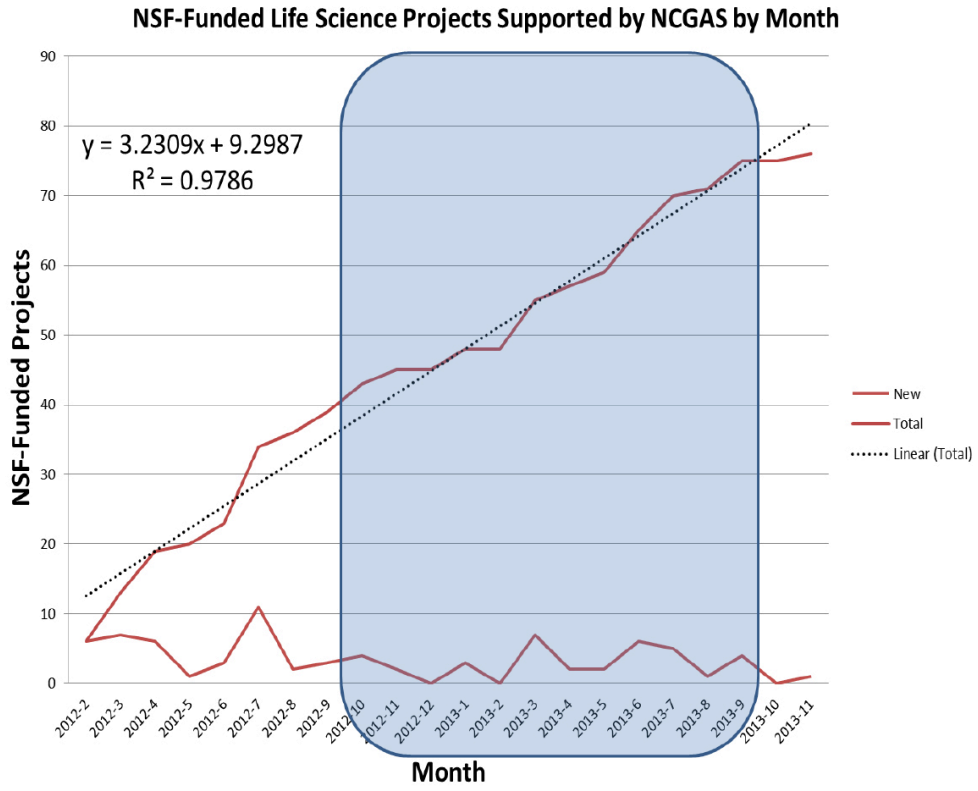
NCGAS addresses critical problems genomics researchers face today. Life sciences research has seen a recent exponential increase in genome sequencing, largely due to the rapidly increase in output of DNA sequencers. Next-generation sequencers generate much more data, straining laboratory and departmental 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.

NCGAS provides services to biologists in the following areas:

- Consulting for biologists undertaking genome analysis
- Optimization, hardening, and enhancement of genome analysis software
- Providing support to biologists delivery and assistance in use of supercomputer clusters for genome analysis, including many supercomputers and supercomputer clusters that are part of XSEDE (the eXtreme Science and Engineering Discovery Environment). 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. Operated by the Texas Advanced Computing Center
    - Blacklight – a shared memory supercomputer run by the Pittsburgh Supercomputing Center, access to a large memory supercomputer cluster to support genome analysis, particular support of genome assembly software.
- Rockhopper, a "cluster on demand" cloud resource where researchers can purchase time on a real supercomputer cluster from Penguin Computing Inc. Long-term storage of data (by default up to 50 TB and with proposals amounts larger than that) and
- Archiving of public data sets, with Dublin Core metadata so that they can be discovered through web searchers, in IU's persistent digital archive (based on DSPACE)
- Aid to researchers preparing grant proposals including providing Letters of Collaboration for your project's NSF grand submission and partnership agreements that commit NCGAS to aiding your research.

In order to disseminate information about the services offered by NCGAS, 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 and education.

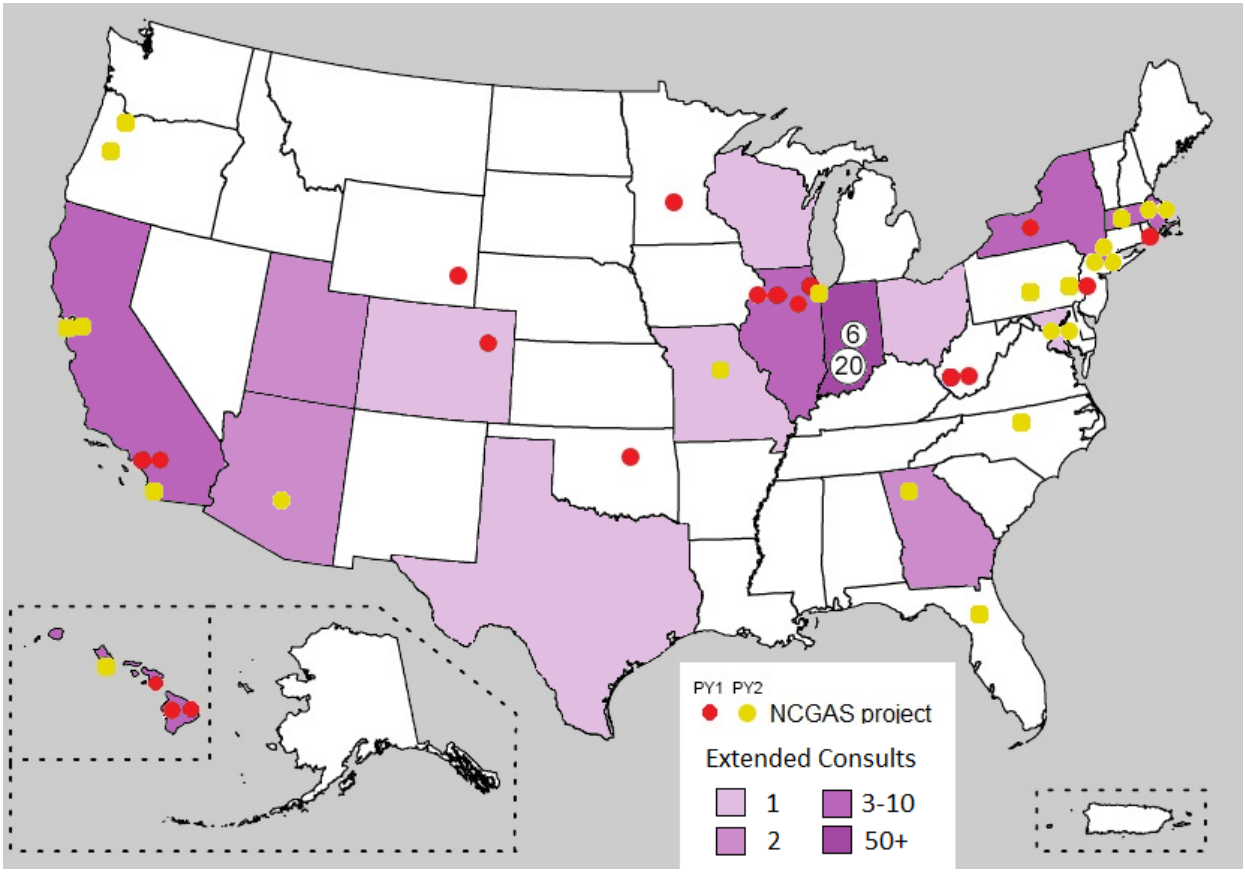
Consulting and support for biological research in the US



Aggregate NCGAS-Supported Projects, per month, showing an addition of just over 3 new projects per month.

The figure above shows the rate at which biologists have requested help with projects over the two years of operation of NCGAS so far. NCGAS received over the life of the project. The surge in requests in early PY1 represents pent-up demand for our services. Since then, allocation requests have arrived at a stable rate of about 3.2 per month. In PY2 NCGAS offered support to 23 new client projects with NSF funding projects, and 9 new projects without NSF funding but in areas potentially fundable by the NSF. The 32 new projects focused on a wide range of topics that included bioinformatics, ecology, entomology, evolution, undergraduate and graduate instruction, mammalogy, microbiology, oceanography, plant biology, and systems biology.

The figure below shows the geographic distribution of groups to which NCGAS offered services in PY1 and PY2. In general, in PY2 NCGAS recruited more projects from further afield than in PY1. Locales of groups to which we offered services included Northern California, Oregon, Pennsylvania, Georgia, and Florida. NCGAS has maintained its constant growth by reaching out of its home state when recruiting new projects, consistent with NCGAS' national focus, and that is resulting in a national base of users that is unquestionably national in scope.



Map showing the location of NCGAS supported projects. Red dots indicate home locations of projects receiving support starting in PY1; yellow dots indicate home locations of projects receiving support starting in PY2. Circled numbers indicate several projects from the same location. Color coding represents the number of extended consultations originating from the state in PY2.

NCGAS categorizes its interactions with clients into short term and extended consultations. A short term consultation is any interaction requiring less than four hours of staff time to resolve the users request; our goal is to resolve any problem of this sort in no more than 3 business days. Extended consultations are any interaction requiring more than four hours to resolve. In 2012 NCGAS staff conducted 392 short term and 143 extended consultations. These consultations represented 314 staff days of interaction. Meanwhile our partners at Texas Advanced Computing Center (TACC) completed 30 support engagements with biologists outside of the University of Texas at Austin.

Scientific outcomes

The primary purpose of NCGAS is to enable biologists to achieve research results that they would not have been able to obtain without NCGAS assistance, or obtain research results faster than would have been possible without NCGAS assistance. The table below summarizes the scientific contributions of NCGAS clients and NCGAS staff during PY2. IT is particularly notable that during PY2 there are now several peer-reviewed technical papers available in print or electronically that were aided by NCGAS services in some fashion.

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	6	3
Peer reviewed journal papers in-review	3	1
Journal papers in-preparation	2	0
Posters (science content, not outreach)	5	0

Optimization, hardening, and enhancement of genome analysis software

The most significant accomplishment in the area of software enhancement and deliver was the creation of a Galaxy portal, accessible to NCGAS users from the NCGAS web page. This permits use of bioinformatics software within a Galaxy workflow to be used on a number of computational resources, including the Mason Cluster, Big Red 2, and the Data Capacitor 2.

In addition, NCGAS staff have managed and updated a total of 52 bioinformatics and genome analysis software packages on a variety of computing resources available to US biologists, including the large memory supercomputer cluster Mason and other supercomputers accessible.

Providing support to biologists delivery and assistance in use of supercomputer clusters

The large-memory Mason cluster is the primary computational resource of NCGAS. Each of the Mason cluster's 16 nodes has a 32-core processor and 512 gigabytes (GB) of Random Access Memory (RAM), and is architected for memory-intensive genome and transcriptome assembly. NCGAS and the Mason cluster grew significantly in 2012. Mason users grew linearly with time growing from 155 to more than 400.

In terms of total number of users and total amount of CPU time delivered, one of the most significant things NCGAS did, through work with its partner the Texas Advanced Computing Center at the University of Texas- Austin, was deploy a broad array of bioinformatics and genome analysis software on the new Stampede supercomputer, including the 27 genome assembly tools listed in Table 11. NCGAS partner TACC added 24 new start-up allocations related to life sciences research – primarily genomics and population genetics - onto their two major XSEDE-allocatable clusters, Stampede and Lonestar.

Long-term storage of data and archiving of public data sets

In response to requests from NCGAS clients, we have added new services during PY2 for storing data including storage of up to 50 terabytes of research data on IU's Scientific Data Archive tape storage system, and services for curation and long-term storage of data sets and final results from genome research in the IUScholarWorks (http://scholarworks.iu.edu) digital repository.

Support for researchers in preparation of grant proposals

Also in response to requests from NCGAS clients, we are now offering a variety of services specifically to support researchers preparing grant proposals to the NSF – particularly writing letters of commitment to provide consulting services, computing resources, and storage of data (this latter in particular in support of preparation of Data Management Plans).

Outreach and education

Staff of the NCGAS communicate vigorously with the science community of the present and with students who make up the source of the science community of tomorrow.

Out focus in outreach to the scientific community we have focused on presentations at conferences, scientific meetings, or workshops. We have made formal presentations at 25 such events with a total attendance of 2,160. There they logged 193 contact hours presenting 15 talks and 5 posters. Overall the 3.7 FTE NCGAS staff attended a scientific meeting, gave a workshop, poster, or talk, or published a paper every two weeks!

Outreach and education activities	
Posters (outreach and about services)	7
Presentations, including outreach-oriented presentations	8
Press releases	3

Outreach and education activities by NCGAS staff

2. Introduction

In September 2011 IU received a three-year, $1,460,000 grant (award no. 1062432) from the National Science Foundation (NSF) Division of Molecular and Cellular Biosciences to establish the National Center for Genome Analysis Support (NCGAS) in partnership with the Texas Advanced Computing Center (TACC). NCGAS is an innovative service center (core facility) that provides software support and services on IU's Mason large-memory computer cluster, TACC's Gordon system, and the Dash system at the San Diego Supercomputer Center (SDSC). It supports NSF-funded researchers who use genome assembly software that assembles data from next-generation sequencers, large-scale phylogenetic software, and other genome analysis software that require large amounts of memory.

NCGAS provides services to biologists in the following areas:

- Consulting for biologists undertaking genome analysis
- Optimization, hardening, and enhancement of genome analysis software
- Providing support to biologists delivery and assistance in use of supercomputer clusters for genome analysis, including many supercomputers and supercomputer clusters that are part of XSEDE (the eXtreme Science and Engineering Discovery Environment). 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. Operated by the Texas Advanced Computing Center
    - Blacklight – a shared memory supercomputer run by the Pittsburgh Supercomputing Center, access to a large memory supercomputer cluster to support genome analysis, particular support of genome assembly software.
    - Rockhopper, a "cluster on demand" cloud resource where researchers can purchase time on a real supercomputer cluster from Penguin Computing Inc.
- Long-term storage of data (by default up to 50 TB and with proposals amounts larger than that) and
- Archiving of public data sets, with Dublin Core metadata so that they can be discovered through web searchers, in IU's persistent digital archive (based on DSPACE)
- Aid to researchers preparing grant proposals including providing Letters of Collaboration for your project's NSF grand submission and partnership agreements that commit NCGAS to aiding your research.

In order to disseminate information about the services offered by NCGAS, 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 and education.

The resources NCGAS provides make it easier for genomics researchers to conduct their science. 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 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, NCGAS provides NSF-funded biologists 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 PY2 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.

In PY 2, NCGAS continued its science support, outreach, and cyberinfrastructure development. The following report details NCGAS efforts in PY2 towards meeting Center goals.

3. Consulting and support for biological research in the US

To understand the role of NCGAS in supporting science projects, consider the problem facing researchers at New York's Lamont-Doherty Earth Observatory (LDEO). Phytoplankton plays a key role in the ocean's ability to absorb atmospheric gases associated with global climate change, so researchers at the LDEO wanted to investigate the nutrient physiology of phytoplankton in a species-specific manner. The growth of phytoplankton, and therefore their ability to sequester atmospheric greenhouse gasses, is usually restricted by the phytoplankton's ability to access nitrogen and phosphorus. Understanding how living phytoplankton populations partition these growth-limiting compounds will allow the researchers to build more accurate models of the ocean's ability to absorb atmospheric carbon.

To answer their questions regarding genes involved in resource partitioning, the researchers needed to assemble RNA sequence data. Without completed genomes on the various phytoplankton species, this was possible only on a large RAM computer such as Mason. Researchers turned to NCGAS, whose bioinformatic and computational resources enabled LDEO faculty and students to examine the bottom-up controls on phytoplankton growth, and to *de novo* assemble their RNA sequence data from each biological sample.

Figure 1. Harriet Alexander (left) and an unidentified helper setting up her incubation experiment on a three-week research cruise on the R/V Kilo Moana in the North Pacific Ocean.

3.1. 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.

In PY2 NCGAS staff reported 398 short-term and 143 extended consultations. Figure 7 shows the breakdown of the time spent on consultations. By far the most common is the short question whose resolution takes less than a half hour. On average, NCGAS fields nearly one per working day. Extended consultations represent well over one third of available staff time, and in PY2 required 314 staff days.

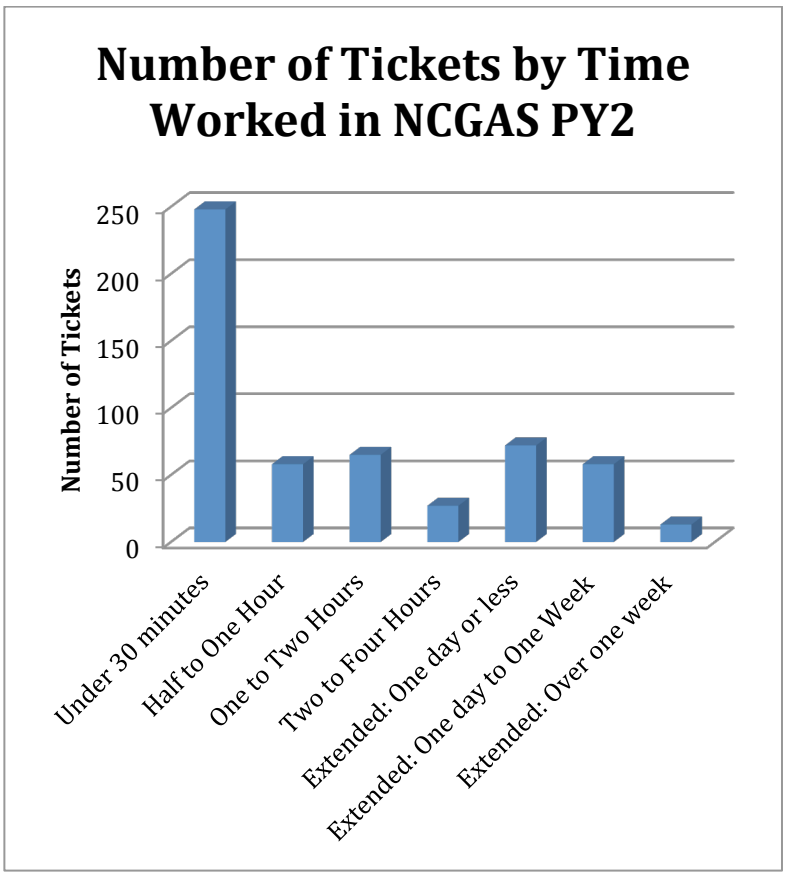


Figure 2: Number of Tickets Reported by NCGAS Staff as a Function of the Time Needed to Complete the Ticket.

3.2. Projects receiving significant support from NCGAS

During PY2, NCGAS added 32 new research projects to the list of supported projects (21 from non-IU institutions). They focused on a wide range of topics that included bioinformatics, ecology, entomology, evolution, undergraduate and graduate instruction, mammalogy, microbiology, oceanography, plant biology, and systems biology. The two maps in Figure 3 show the locations of scientific projects in PY 1 (left) and PY2 (right). In general, in PY2 NCGAS recruited more projects from further afield than in PY1. Locales included Northern California, Oregon, Pensylvania Georgia, and Florida. See **Appendix 1** for a list of all new NCGAS-supported projects for PY2. Meanwhile our partners at TACC completed 30 support engagements with non-UT researchers and onboarded 24 life-science start-up allocations on their major supercomputers Lonestar and Stampede.

Figure 3 shows the rate at which biologists have requested help with projects over the two years of operation of NCGAS so far. NCGAS received over the life of the project. The surge in requests in early PY1 represents pent-up demand for our services. Since then, allocation requests have arrived at a stable rate of about 3.2 per month. In PY2 NCGAS offered support to 23 new client projects with NSF funding projects, and 9 new projects without NSF funding but in areas potentially fundable by the NSF. The 32 new projects focused on a wide range of topics that included bioinformatics, ecology, entomology, evolution, undergraduate and graduate instruction, mammalogy, microbiology, oceanography, plant biology, and systems biology.

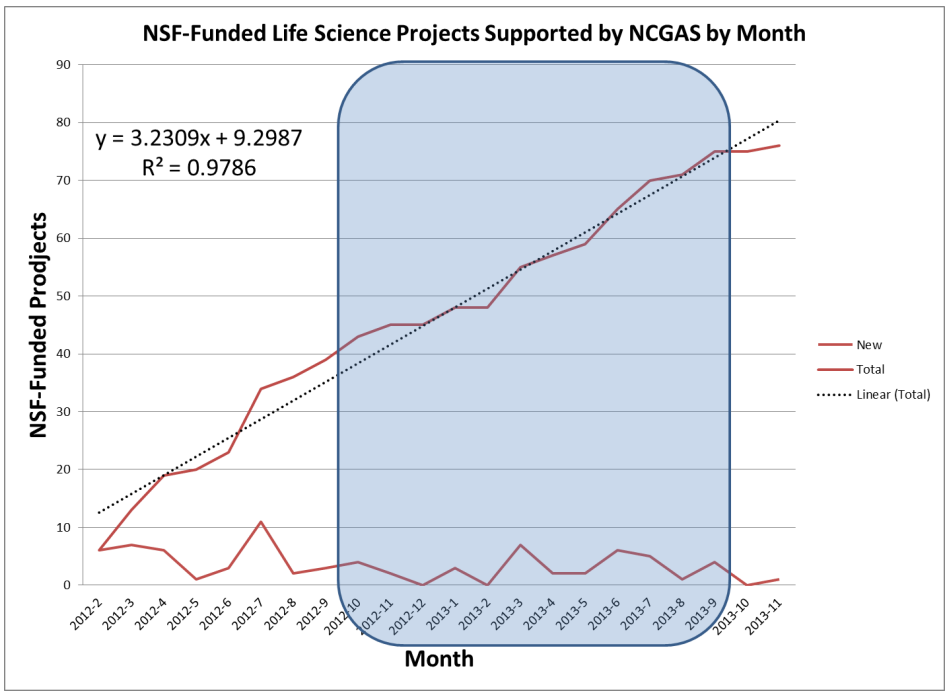


Figure 3. NCGAS-Supported Projects, per month, showing an addition of just over 3 new projects per month.

Figure 4 below shows the geographic distribution of groups to which NCGAS offered services in PY1 and PY2. In general, in PY2 NCGAS recruited more projects from further afield than in PY1. Locales of groups to which we offered services included Northern California, Oregon, Pennsylvania, Georgia, and Florida. NCGAS has maintained its constant growth by reaching out of its home state when recruiting new projects, consistent with NCGAS' national focus, and that is resulting in a national base of users that is unquestionably national in scope.

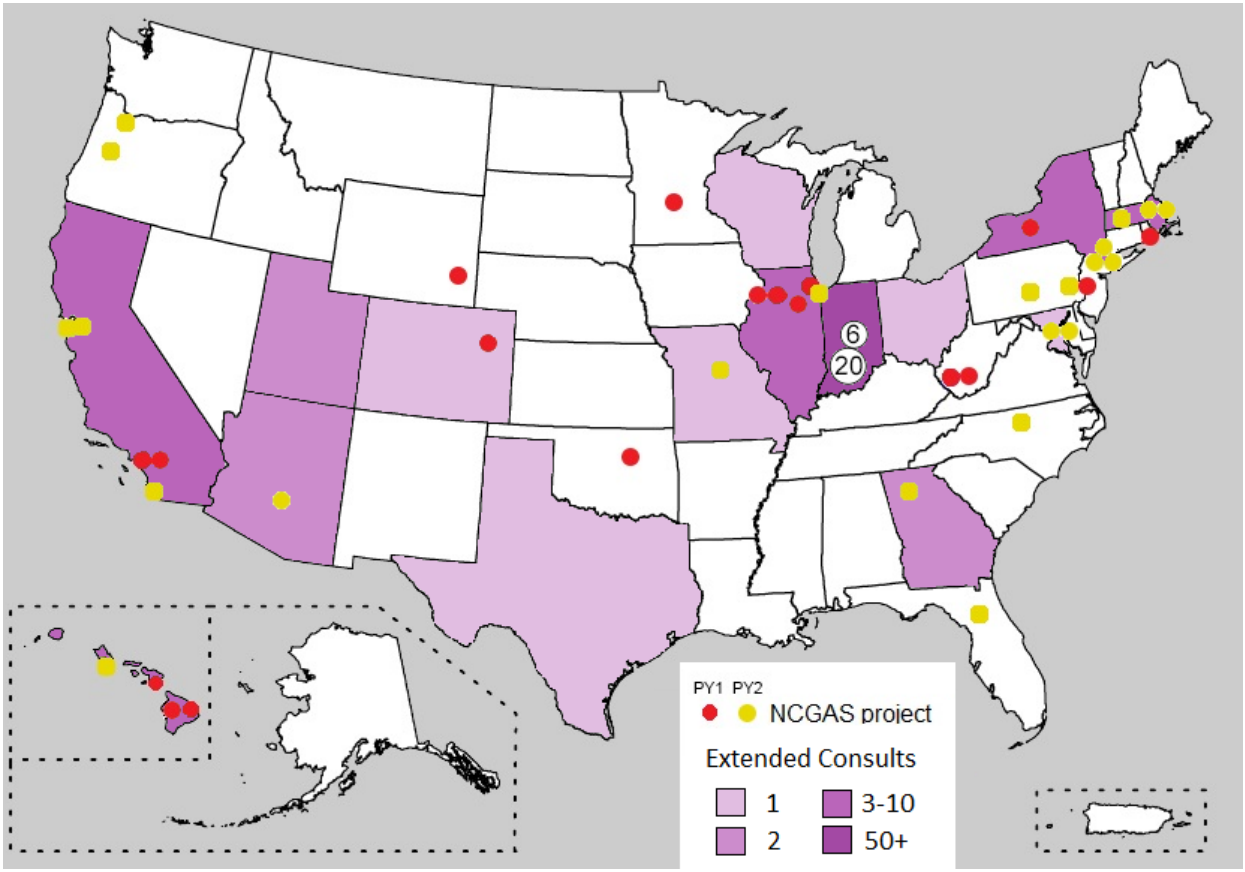


Figure 4. Map showing the location of NCGAS supported projects. Red dots indicate home locations of projects receiving support starting in PY1; blue dots indicate home locations of projects receiving support starting in PY2. Circled numbers indicate several projects from the same location. Color coding represents the number of extended consultations originating from the state in PY2.

NCGAS categorizes its interactions with clients into short term and extended consultations. A short term consultation is any interaction requiring less than four hours of staff time to resolve the users request; our goal is to resolve any problem of this sort in no more than 3 business days. Extended consultations are any interaction requiring more than four hours to resolve. In 2012 NCGAS staff conducted 392 short term and 143 extended consultations. These consultations represented 314 staff days of interaction. Meanwhile our partners at Texas Advanced Computing Center (TACC) completed 30 support engagements with biologists outside of the University of Texas at Austin.

4. Scientific Products

One of the primary purposes of NCGAS is to enable biologists to achieve research results that they would not have been able to obtain without NCGAS assistance, or obtain research results faster than would have been possible without NCGAS assistance. The table below summarizes the scientific contributions of NCGAS clients and NCGAS staff during PY2. IT is particularly notable that during PY2 there are now several peer-reviewed technical papers available in print or electronically that were aided by NCGAS services in some fashion.

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	6	4
Peer reviewed journal papers in-review	4	0
Journal papers in-preparation	3	0
Posters (science content)	5	7
Presentations, including outreach-oriented presentations	0	8

Table 1. Summary of scientific products created by scientists with the benefit of NCGAS support during PY2.

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

5. Optimization, hardening, and enhancement of genome analysis software

The NCGAS mission includes improving the utility of community-developed software and its accessibility to life scientists. In PY2 NCGAS staff at IU initiated three noteworthy projects towards this goal.

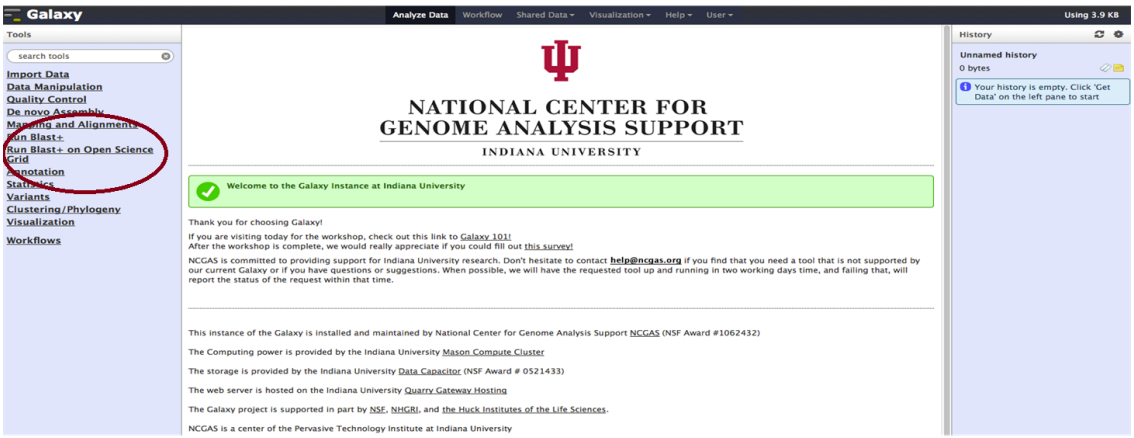


Figure 3: This image shows the web page researchers see when the use the NCGAS instance of the Galaxy web portal to analyze their next-generation DNA or RNA sequence data. The small addition of the option to run BLAST on the Open Science Grid represents a major enhancement in functionality.

The first project involved a seemingly small change to a web page, shown in Figure 5. By adding a link on this website, NCGAS and the NSF-funded Open Science Grid (OSG) made a major step forward in relieving the computational bottleneck facing biologists and medical researchers. The largest computational challenge facing life scientists is comparing new DNA, RNA, and Protein sequences with other known sequences to gain insights into the function of the new sequence. The most commonly used tool for this is a suite of programs known as BLAST. But genomic researchers often need to wait for up to

three weeks for new sequences to be analyzed with BLAST. This is slow and consumes local computer resources. Scientists often use the Galaxy web portal to run smaller BLAST jobs, along with hundreds of other analytic tools, but in the past had to use different, more complex tools for larger jobs. The OSG can run very large computational jobs in parallel, but life scientists found it unapproachable. To rectify this, NCGAS worked with IU's High Throughput Computing group to extend the NCGAS instance of the Galaxy web portal to run large jobs on OSG. Now, life scientists can run BLAST on systems across the nation by automatically breaking apart large BLAST jobs and submitting them to the OSG, This relieves stress on local computational systems, and jobs complete faster.

The second software improvement project for NCGAS in PY2 was the completion of what had initially been a small project started in PY1. In our first year, NCGAS improved the computational efficacy of Trinity, the popular RNA *de novo* sequence assembly program from MIT's Broad Institute. The gains were a fourfold improvement in application runtime and a partnership between NCGAS and the Broad Institute. This turned into a significant improvement in the official, production distribution of the Trinity software. During PY2 NCGAS scientists contributed to a major paper on the use of the Trinity package published in *Nature Protocols*. In the first nine months since its publication, this paper has been cited by 15 peer-reviewed journal articles. Late in PY2 IU/NCGAS received a joint, five-year, $4-million resource improvement grant from the National Cancer Institute. The grant is a partnership with the Broad Institute and Harvard, the lead institution, and the Center for Information Services and High Performance Computing (ZIH) at Technische University in Dresden. The partnership's goal is to advance cancer research by continuing to improve Trinity performance.

Finally, Michael Lynch's laboratory has long studied the interplay between population genetics and genome structure, using an array of organisms and methods: Daphnia, Paramecium, bacteria, wet-lab experiments, and pure computational biology. And now that it is so easy to sequence many genomes for any given creature, it's possible to do true "population genomics". However, while getting the genome sequences is now possible, and getting cheaper every day, the analytic methods to interrogate the data and extract parameters that biologists care about (e.g. population size, past population bottle-necks, pattern of recombination, etc.) are still in active development. Researchers in the Lynch laboratory have been working to develop such methods, using both simulated data and published genome sequences, and are also collecting their own data sets for Daphnia and Paramecium. However, these datasets become quite large, and their analysis requires quite a lot of compute time; NCGAS and IU Research Technologies stepped in to increase the efficiency of some of the Lynch applications, and to act as an enabler for the Lynch lab's successful application for XSEDE resources (~6 Million SUs). This work is now in manuscript form, and will soon be submitted for publication.

Aside from deploying the NCGAS software stack on Lonestar, and the majority on Stampede, TACC continues to have an active role in NCGAS-related software improvement and hardening. With a particular interest in the Intel Xeon Phi coprocessor, TACC proved guidance and test cases to the laboratory of Henry Tufo, who is working to accelerate BLAST+ on this architecture. Additionally, TACC staff worked to develop an accelerated R 3.02 package that is able to take advantage of both the latest MPI stack for parallel operations, the Intel Math Kernel Library, and is able to automatically offload some linear algebra operations to the Xeon Phi. TACC staff are also actively researching use of system-level checkpointing and containers to support longer run times needed by many computational biologists.

In NCGAS PY2, TACC also began hosting the Galaxy Main portal using their cloud environment. The Galaxy project consists of two major components, the first is the actual software used to generate a Galaxy web portal, and the second is an instance of that portal referred to as Galaxy Main. This portal currently has over 30,000 registered users and is one of the singularly most used web applications in bioinformatics. TACC is actively researching how to scale the Galaxy Main portal out to the rest of the TACC systems, with an objective of being able to support Galaxy jobs running across the XSEDE network once and for all.

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

Providing support to biologists delivery and assistance in use of supercomputer clusters for genome analysis, including many supercomputers and supercomputer clusters that are part of XSEDE (the eXtreme Science and Engineering Discovery Environment). 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. Operated by the Texas Advanced Computing Center
- Blacklight – a shared memory supercomputer run by the Pittsburgh Supercomputing Center, Access to a large memory supercomputer cluster to support genome analysis, particular support of genome assembly software.
- Rockhopper, a "cluster on demand" cloud resource where researchers can purchase time on a real supercomputer cluster from Penguin Computing Inc.

6.1. Mason

The large-memory Mason cluster is the primary computational resource of NCGAS. Each of the Mason cluster's 16 nodes has a 32-core processor and 512 gigabytes (GB) of Random Access Memory (RAM), and is architected for memory-intensive genome and transcriptome assembly. NCGAS and the Mason cluster grew significantly in 2012. Mason users grew linearly with time growing from 155 to more than 400.

The number of Mason users has grown in a remarkably linear fashion throughout PY2 (see Figure below). No clear distinction is made between students and researchers – a natural result of the boutique nature of the Mason cluster. When students use a system like Mason, they usually do so to gain skills in genome and transcriptome assembly and to analyze research data gathered for their own projects. Regardless of the difficulty of accounting for students and researchers separately, clearly Mason usage has exceeded expectations. Across PY2, we added over 20 users per month and more than doubled our number of registered users. We started PY2 with 155 total users, and ended with well over 400.

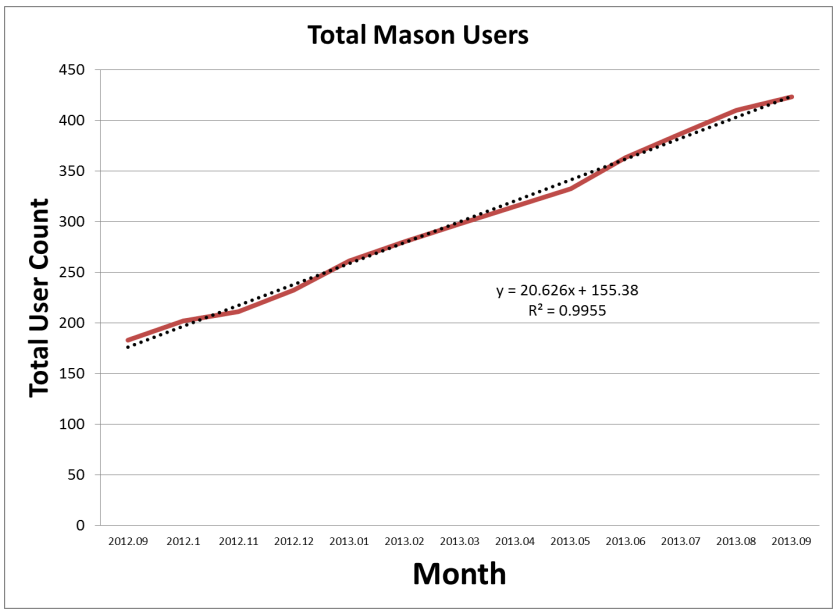


Figure 5: Total Users on the Mason System across PY2

The combination of the structure of many genome analysis software and the architecture of the Mason cluster complicates understanding of its utilization using traditional supercomputer center metrics such as CPU-hours or core-hours. Mason has 16 nodes, each with 32 cores. Most genome assembly codes – the software that most needs large memory clusters – are single threaded, using just a single core on a CPU. Such codes may need to access large amounts of memory, but codes like this don't rack up accumulated use of core hours the same way as some other codes. Furthermore, to ensure good security of individual research activities, systems policy was that a given user had sole use of any given node at any given time. This rule was useful in supporting complex genomic workflows where certain steps are highly parallel and others are serial. On average, over PY2, each node had 1.007 core minutes per CPU minute. Overall, then, the system was well utilized, but the single-user-per-node policy left many cores idle. When a new multi-user-per-node queue was added around week 40, the number of used core seconds skyrocketed, even without a noticeable jump in the number of jobs submitted. It is, therefore, best to consider Mason usage in terms of the number of user jobs submitted. The number of jobs on the system varies greatly over time, with most of the variability being due to student jobs. Although less than a third of the total were submitted by non-NSF funded researchers. NSF-funded users submitted 33,902 jobs to the Mason cluster in PY2.

Other XSEDE systems

In terms of total number of users and total amount of CPU time delivered, one of the most significant things NCGAS did, through work with its partner the Texas Advanced Computing Center at the University of Texas- Austin, was deploy a broad array of bioinformatics and genome analysis software on the new Stampede supercomputer including the 27 applications listed in Table 11. NCGAS partner TACC added 24 new start-up allocations related to life sciences research – primarily genomics and population genetics - onto their two major XSEDE-allocatable clusters, Stampede and Lonestar.

One of the key services offered by NCGAS to the national community is the maintenance of software on supercomputers and supercomputer clusters that are coordinated through XSEDE –the NSF-funded eXtreme Science and Engineering Discovery Environment.

Appendix 3 provides a set of 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 resources

7. New service for PY2 - Long-term storage of data and Archiving of public data sets

In response to requests from NCGAS clients, we have added new services during PY2 for storing data including:

- Storage of up to 50 terabytes of research data on IU's Scientific Data Archive tape storage system. These data are stored in duplicate - one copy in Indianapolis and another in Bloomington. Storage is committed for at least three years.
- Services for curation and long-term storage of data sets and final results from genome research in the IUScholarWorks (http://scholarworks.iu.edu) digital repository. Storage of such final data products is unlimited; this system is IU's official data archive. IUScholarWorks data can be made public immediately or be held for a future date determined by the researcher.

8. New service for PY2 - Aid to researchers preparing grant proposals

Also in response to requests from NCGAS clients, we are now offering a variety of services specifically to support researchers preparing grant proposals to the NSF – particularly writing letters of commitment to provide consulting services, computing resources, and storage of data (this latter in particular in support of preparation of Data Management Plans).

During PY2, NCGAS wrote four letters of support including two for national centers, and two supporting individual research projects.

9. Education and outreach

Staff of the NCGAS communicate vigorously with the science community of the present and with students who make up the source of the science community of tomorrow.

Outreach to the biology research community: To attract NSF-funded genomics and life-science projects, NCGAS exhibits and speaks at events that attract potential users. These include major life science conferences such as Evolution, the annual meeting of the Society for Molecular Biology and Evolution, and meetings that draw traditionally underserved groups, such as the annual meeting of the Hispanic Association of Colleges and Universities.

Our academic papers fall into two categories. First are papers for a broad academic audience, such as our recent perspectives piece in the Journal of the American Medical Informatics Association encouraging medical informatics use of the national cyberinfrastructure, and our contribution to the Trinity tutorial in Nature Protocols. Second are papers involving NCGAS staff intellectual contributions to individual research projects.

Public presentations include general announcements of our services, techniques for leveraging the national CI, and specific scientific contributions. Examples of the former include Dr. Barnett's Silver

Medallion-winning poster, "The National Center for Genome Analysis Support: Providing Free or Low-Cost Bioinformatics Support at Scale" (2013 AMIA Translational Sciences Summit, San Francisco), and Dr. LeDuc's talk "Optimizing the National Cyberinfrastructure for Lower Bioinformatic Costs: Making the Most of Resources for Publicly Funded Research," (RNA-Seq 2013 Summit, Boston, MA). The latter includes talks and posters given at the International Society of Protistologists 3rd North American Section Meeting in North Carolina.

Support for Students: As well as supporting individual student research, in PY2 NCGAS sponsored or was involved in five meetings designed to educate students. These ranged from small workshops introducing the Galaxy web portal, to invited presentations on domain knowledge unique to NCGAS scientists. A total of 182 made use of the IU Mason cluster during PY2, engaging in research or research education activities. Of these, 4 were undergraduates and 60 were graduate students. We offered a total of 7 training sessions, amounting to 30 contact hours to students. A total of 155 students attended these training sessions and classes.

Outreach to Underserved Populations and in EPSCOR states:

In PY2 NCGAS sought to reach researchers in minority communities. We exhibited and/or participated in two conferences and meetings that target underserved communities: the American Indian Higher Education Commission (AIHEC) AIHEC 40th Anniversary Conference, and the Hispanic Association of Colleges and Universities. We met with a total of 93 individuals at these conferences to inform them of our services, but none of these led to ongoing consulting interactions with NCGAS, although it did lead to NCGAS participation at two workshops in PY3.

NCGAS has also focused on support to EPSCOR states. The mission of the NSF Experimental Program to Stimulate Competitive Research (EPSCoR) is to "strengthen research and education in science and engineering throughout the United States and to avoid undue concentration of such research and education." NCGAS leverages IU's strength in national-scale computer networking to support researchers at institutions in EPSCoR states that may have difficulty accessing large RAM computers. By the end of PY2 we had supported seven projects from EPSCoR listed states.

Appendix 4 lists the 25 events where NCGAS logged 193 contact hours and whose audiences totaled up to 1,708. Of these participants a total of 658 represented members of traditionally underserved groups as defined by the NSF.

Outreach and education activities	
Posters (outreach and about services)	7
Press releases	3
Presentations, including outreach-oriented presentations	8
Total attendees at events	1708
Total attendees at events who were members of traditionally underserved groups as defined by NSF	658
Total contact hours in presentations and EOT events	193

Table 2. Summary of outreach and education activities by NCGAS staff

9.1. Management and Operations

As principal investigator of the NSF award that established NCGAS, Dr. Craig Stewart provides overall leadership. Dr. William Barnett directs NCGAS, and Dr. Richard LeDuc manages its operations. For

internal coordination, Drs. Barnett and LeDuc oversee a monthly "kitchen cabinet" that reviews status on research projects, coordinates the development and execution of cyberinfrastructure, and resolves process and other tactical issues. The following staff support NCGAS as a service and Mason as a cyberinfrastructure resource: Dr. Tom Doak (scientist), Dr. Le-Shin Wu (computer scientist), Rich Knepper (software management), Robert Henschel (software optimization), Dr. Scott Michael (software optimization and management), Bret Hammond (systems administration), Matthew Allen (systems administration), and Carrie Ganote (bioinformatics support). TACC service resources are provided by their lead bioinformatician Matt Vaughn, and by technical support services funded by the NCGAS grant. An internal advisory board comprising NCGAS co-PIs at IU (Drs. Michael Lynch, Matthew Hahn, and Geoffrey Fox) provides guidance for NCGAS operations.

9.2. *User Survey*

NCGAS conducted its first User Survey in mid-PY2, covering the first 18 months of operation. We sent surveys to a total of 52 people who had used our services since September 2011. We received a total of 22 responses, for a response rate of 42.3%.

The services used by survey respondents are shown in Table 7:

Service	% utilization
Consulting (short-term and extended)	18%??
Project support	23%
Mason use	95%
Storage	36%
Help with grant preparation	0%

Table 3. Summary of NCGAS user survey results.

Positive comments in free text sections of the survey included compliments on services in general, satisfaction with the availability of large memory machines; software, and the speed at which requested software is installed; and the level of support, including scientific and bioinformatics consulting.

There were several suggestions for improvements and changes in services in free text comments as well. Comments and our responses to them are presented in the following table

User comment	NCGAS response
Timing and alerts. Respondents requested longer run times, shorter wait times, more transparent queuing, longer wall times, and an alert system for when runs fail.	NCGAS allows requested wall times of up to 10 days, and will extend this upon request from a user. Additional queues were added in PY2 to shorten wait times.
Node use. Several users responded that a single job per node is inefficient, even with a single-processor job.	A shared node queue was added in PY2 to allow single core jobs from multiple users to share a node.
Communications. Fourteen percent reported some level of communications difficulties with the system.	Anyone experiencing difficulty connecting to an NCGAS resource should contact help@ncgas.org immediately. We will work promptly to identify and correct communications difficulties.

Data storage and transfer. Concerns included the permanence of and limitations on amount of data storage	In PY2 NCGAS added archival space on the SDA for permanent storage of analysis results, and access to IU ScholarWorks for archiving public results and datasets..
Lack of software for error-checking submission scripts, and e-mail alerts for job status or job failures. Likewise, using software. Some respondents wanted more introductory material, example scripts, etc.	NCGAS staff are happy to work individually with users experiencing difficulty with submitting and monitoring jobs on the Mason system. Please contact help@ncgas.org with specific questions.
Accessing information about NCGAS. Respondents wanted a list of applications on the web pages, and email updates on software and upgrades.	As we already provide these, we viewed this comment as a communication failure on our part, and restructured and reformatted our web pages.

Table 4. Comments made in free-text entry sections of NCGAS user survey and NCGAS responses to those comments

9.3. *Sustainability*

NCGAS is pursuing long-term sustainability based on economies of scale. Most research projects using NGS sequencing and requiring large RAM compute resources do not require a full-time bioinformatician, but only a fraction of a year of support. It is unusual to employ professional staff for just a few months. It is easier for these projects to subcontract assembly and large RAM compute needs to a center such as NCGAS. We received two new awards in PY2 based on this approach, listed below.

Date: 09/01/2013

Institution: Northwestern University

Grant: Subcontract from NIH: CPTAC Grant to SAIC SAIC-Frederick

Title: Combined Top Down and Bottom Up Proteomics of CompRef Cancer Samples

NCGAS will provide biostatistical and bioinformatic computation for proteomics quantification. This subcontract will fund 0.15 FTE of an analyst's time.

Date: 10/1/2013

Institution: Broad Institute

Grant: NIH 1U24CA180922-01, Regev, Aviv

Title: Trinity: Transcriptome Assembly for Genetic and Functional Analysis of Cancer

This is a five-year subcontract to support the Broad Institute's development of the Trinity software package. In PY4 NCGAS will develop a public-facing web server for launching Trinity jobs on human cancer samples.

9.4. *Progress on Program Year 2 milestones*

Progress on milestones for NCGAS are shown in Table 5.

NCGAS Progress on Year 2 Grant Milestones

Quarter	Task	Notes
Q1 (Oct-Dec)	Assess usage (250 student and 25 researcher accounts) and consulting (100 short-term and 10 long-term consults) Barnett	NCGAS completed 398 short-term consults and 143 extended consults in PY2. We ran jobs from 64 students and 117 staff and faculty.
Q2 (Jan-Mar)	Year 2 satisfaction and needs surveys sent out Doak	One survey done over the first year and a half.
	Year 2 fall outreach at 2 events (100 attendees) Doak	NCGAS had 193 contact hours at 25 outreach events attended by 2,160 people.
	All installed applications at latest version, KB documents up to date, made easy to install if necessary & in repository Wu, Repasky, System Programmer	See Appendix 3 for complete software list
Q3 (Apr-Jun)	Year 2 survey results report Doak	Completed
	Year 2 spring outreach at 2 events (100 attendees) Doak	Completed
Q4 (Jul-Sep)	Assess usage (250 student and 25 researcher accounts) and consulting (100 short-term and 10 long-term consults) Barnett	423 total user accounts

Table 5. Accomplishment of NCGAS milestones in PY2.

10. Appendix 1. Research projects that received extensive consultation and support from NCGAS during PY2.

The primary goal of NCGAS is to provide computational support to existing genomics and life science projects. This section outlines each project supported in PY2: first, the 23 NSF-funded projects in order of their request for support, and then the remaining nine projects.

10.1. New NSF-funded Projects

PI: Jacob Freimer
State: CA
Funding Organization: NSF
Award #: 1000122262
Title: Mapping mRNA and RNA binding protein interactions
Date initiated: 7/30/13
Date completed: 7/30/15

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: Jason Stajich
State: CA
Funding Organization: NSF
Award #: 1027542
Title: Genome-wide Impact of mPing Transposition on Rice Phenotypic Diversity
Date initiated: 11/1/12
Date completed: 6/1/14

Reference guided assembly of several strains of rice, which have varying degrees of transposable element amplification to identify complete genomes. One of these strains are parents in a cross to generate RILs and RIL genotyping by resequencing will be performed to associate phenotypes with TE migrations.

PI: Jeffrey Boore
State: CA
Funding Organization: NSF
Award #: W000466595
Title: The Genomic Consequences of Asexuality
Date initiated: 7/11/13
Date completed: 12/31/14

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
State: CA
Funding Organization: NSF
Award #: 1155030
Title: Collaborative Research: Ecological genomics of stress response in an intertidal copepod
Date initiated: 7/1/13
Date completed: 4/1/15

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: C. Eduardo Vallejos
State: FL
Funding Organization: NSF
Award #: 0923975 and 0920145
Title: Development of a Gene-Based Ecophysiology Model, and Genetic analysis of root traits associated with domestication of the common bean (Phaseolus vulgaris L.)
Date initiated: 5/10/13
Date completed: 7/30/14

Both projects aim to identify QTLs controlling some aspect of growth and development. These include time-to-events, and growth rates or allometric relationships. Recombinant inbred families (n=180) have been phenotyped at five sites (from North Dakota to Colombia; 1st project) using weekly destructive samplings, or daily root scans (2nd project). QTL analyses using high-resolution linkage maps (genotyping-by sequencing) have allowed us to identify several major QTLs. Aligning the linkage maps to the genome sequence will likely reveal the identity of candidate genes for the QTLs. These alignments will be facilitated by the fact that the linkage maps were constructed with SNPs located in 100 base reads. We have obtained Illumina HiSeq data (98X genome coverage) and would like to perform an assembly using Velvet. The sequencing data occupies 170 GB (data cleaned and trimmed), and contains 600 million 100-base reads with a minimum score of 33.

PI: Christopher Beck
State: GA
Funding Organization: NSF
Award #: 815135
Title: Developing a bean beetle curriculum development network
Date initiated: 4/25/13
Date completed: 3/1/14

The goal of this project is to develop a network of faculty who are using the bean beetle, Callosobruchus maculatus, as a model system for teaching inquiry-based lab. As a part of this project, we are assembling a partial genome sequence and transcriptomes for use in teaching. NCGAS resources will be used for genome and transcriptome assembly and gene prediction

PI: Mahdi Belcaid
State: HA
Funding Organization: NSF
Award #: 1260169
Title: Multispecies connectivity: Comparative analysis of marine connectivity and its drivers for the coral reefs of Hawaii
Date initiated: 6/20/13
Date completed: 6/20/15

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: Neil Kelleher
State: IL
Funding Organization: NSF
Award #: DMS-0800631
Title: Statistical Approaches to Integration of Mass Spectral and Genomic Data of Yeast Histone Modifications
Date initiated: 12/1/12
Date completed: 11/30/13

New statistical and analytical methods will be developed to study the regulatory role of histone modifications in Saccharomyces cerevisiae. Gene activities in eukaryotic cells are concertedly regulated by transcription factors and chromatin structure. The basic repeating unit of chromatin is the nucleosome, an octamer containing two copies each of four core histone proteins. While nucleosome occupancy in promoter regions typically occludes transcription factor binding, thereby repressing global gene expression, the role of histone modification is more complex. Histone tails can be modified in various ways, including acetylation, methylation, phosphorylation, and ubiquitination. Even the regulatory role of histone acetylation, the best characterized modification to date, is still not fully understood. Mass spectral and genome-wide microarray data from Saccharomyces cerevisiae offer new opportunities to evaluate the regulatory effects of histone modifications. The investigators will develop statistical methods for identifying target genes of histone modifications and associated DNA sequence features of histone modifications. They will also develop computational and statistical methods for predicting histone modifications and their interactions.

PI: J. Andrew DeWoody
State: IN
Funding Organization: NSF
Award #: DGE-1333468
Title: An evaluation of MHC based mate choice in captive koala (Phascolarctos cinereus)
Date initiated: 7/12/13
Date completed: 8/31/15

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: Melissa Pespeni
State: IN
Funding Organization: NSF
Award #: DBI-1103716
Title: Horned beetle transcriptome assembly, Onthophagus Sagittarius
Date initiated:
Date completed: 3/1/13

Genome-wide tests for selection between closely related yet phenotypically diverged species promise to reveal the genes and functional groups of proteins underlying phenotypic evolution. Among horned dung beetles, Onthophagus sagittarius and O. taurus differ radically in the shape and physical location of horns and sexual dimorphism, yet are only 5 million years diverged. O. sagittarius differs in three major ways: 1) reverse sexual dimorphism where females have more prominent horns than males, 2) male horns are in a novel location of the head, and 3) there is no male dimorphism. Recently a transcriptome sequence was completed for O. taurus using 454 sequencing of normalized cDNA libraries (Choi et al. 2010 BMC Genomics). Here we aim to generate a transcriptome sequence for O. sagittarius using an RNA-seq approach on the Illumina platform. To discover fixed and polymorphic SNPs between these species, we will map short reads from O. sagittarius transcriptome to the 40,000 contigs of the O. taurus 454 transcriptome. SNP data within and among species will be used to identify genes with signatures of positive and negative selection (e.g. DN/DS, McDonald-Kreitman test). Another transcriptome (454) of a closely related species may be available soon, O. nigriventris, which would allow branch-specific analyses. I will need NCGAS resources for assistance with assembling the transcriptome of O. sagittarius.

PI: Jeffrey Blanchard
State: MA
Funding Organization: NSF
Award #: 1237491
Title: Microbial Forest Soil Community Dynamics
Date initiated: 4/4/13
Date completed: 4/4/15

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: John R. Finnerty
State: MA
Funding Organization: NSF
Award #: 924749
Title: LiT: Rel Homology Domain Signal Transduction Pathways in the Sea Anemone Nematostella vectensis
Date initiated: 2/1/13
Date completed: 1/31/14

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: Sean Patrick Mullen
State: MA
Funding Organization: NSF
Award #: 1020136
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: Bastian Bentlage
State: MD
Funding Organization: NSF
Award #: 1046075
Title: Can evolutionary history predict how changes in biodiversity impact the productivity of ecosystems?
Date initiated: 3/21/13
Date completed: 8/31/14

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: Endymion D. Cooper
State: MD
Funding Organization: NSF
Award #: DEB-1036506
Title: Collaborative Research: Assembling the Green Algal Tree of Life (GRAToL)
Date initiated: 6/1/13
Date completed: 12/1/15

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
State: MO
Funding Organization: NSF
Award #: MCB 1127059
Title: GEPR: Deciphering Mechanisms of Prezygotic Reproductive Isolation in Solanum
Date initiated: 8/9/13
Date completed: 8/9/15

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
State: NC
Funding Organization: NSF
Award #: DEB
Title: Analysis of Genomic Data
Date initiated: 7/25/13
Date completed: 7/25/15

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: Sonya Dyhrman
State: NY
Funding Organization: NSF
Award #: CCF-424599
Title: Center for Microbial Oceanography: Research and Education
Date initiated: 3/14/13
Date completed: 3/14/14

Life on Earth most likely originated as microbes in the sea. Today microbes inhabit and sustain all of Earth's biotopes and comprise the dominant form of life in marine environments. They catalyze key biogeochemical transformations of nutrients and trace elements that maintain productivity of the oceans, form and consume greenhouse gases, and are a critical component of marine food webs, in short making the planet habitable. Recent advances in molecular ecology, genomics, remote sensing, and ecological modeling now make it possible to synthesize a more integrated and predictive understanding of the biology, ecology, and biogeochemical roles of the microbes that dominate Earth's largest biome: the global ocean. The goal of this Center for Microbial Oceanography: Research and Education (C-MORE) is to fuse these otherwise separate disciplines into a targeted inquiry into the microbial ecology of the oceans, through an integrated Center. The primary distinguishing feature of the Center is that its work involves all scales of biological organization, from genomes to biomes. Here we are examining how phytoplankton differentially partition their nitrogen and phosphorus niche space using comparative metatranscriptomics. Ultimately, these data will identify bottom-up controls on phytoplankton growth by investigating the nutrient physiology of phytoplankton in a species-specific way.

PI: Sonya Dyhrman
State: NY
Funding Organization: NSF
Award #: OCE-1316036
Title: Collaborative Research: Chemical and biological characterization of phosphonate and polyphosphate dynamics in marine phytoplankton (MMETSP)
Date initiated: 3/14/13
Date completed: 3/14/14

Although phosphorus (P) is recognized as an essential nutrient that influences both marine primary production and community structure, difficulties in the characterization of specific P compounds in the dissolved organic phosphorus pool (DOP) have greatly hindered our understanding of P biogeochemistry. Microbial taxa produce a range of DOP compounds in response to physio-chemical conditions, yet counterintuitively, measurements of DOP bond classes suggest that DOP remains surprisingly invariant with depth and region in the ocean. Given the importance of DOP as a phosphorus source to microorganisms and the climatic and ecosystem implications of DOP utilization, understanding this surprising observation remains a fundamental challenge in marine P biogeochemistry. To address this challenge we are tracking phosphorus composition in different algal taxa grown under a range of environmental conditions. Complimenting this work are surveys of gene expression in transcriptomes of different algal taxa from these treatments. When combined these data will provide a better mechanistic understanding of DOP production in the ocean.

PI: Sonya Dyhrman
State: NY
Funding Organization: NSF
Award #: OCE-0925284
Title: Quantification of Trichodesmium spp. vertical and horizontal abundance patterns and nitrogen fixation in the western north Atlantic (Trichodesmium metatranscriptomes)
Date initiated: 3/14/13
Date completed: 3/14/14

The diazotroph Trichodesmium spp. constitutes a major pathway of nitrogen flow into marine planktonic ecosystems, but estimates of its impact on global nitrogen budgets vary widely. Sampling is made difficult by the fragility of the organism with the consequence that Trichodesmium spp. are difficult to manipulate in both field and laboratory experiments. A recent transatlantic survey revealed unexpectedly high abundance of Trichodesmium spp. at depth, suggesting the vertical distribution of the organism within the euphotic zone may be more uniform than previously thought. Application of a simple bio-optical model of productivity to the observed profile of abundance suggests the depth-integrated nitrogen fixation rate could be three to five times higher than that based on the canonical profile of exponential decrease in abundance with depth. This raises a key question: is there a similar vertical distribution in waters further to the south, where Trichodesmium spp. are an order of magnitude more abundant overall? If so, are the deep populations actively fixing nitrogen? If so, the implications for the global nitrogen budget would be substantial. To answer these questions, we have conducted two cruises to survey the waters of the southern Sargasso Sea, where Trichodesmium spp. are commonly found in high abundance. Rate measurements are being compiled to assess whether or not the deep populations are actively fixing nitrogen and complimented with metatranscriptome analyses to independently identify patterns in nitrogen fixation and examine potential controlling factors on Trichodesmium growth and its activities.

PI: Alice Barkan
State: OR
Funding Organization: NSF
Award #: MCB-1243641;IOS-0922560
Title: Deciphering the Code for RNA Recognition by PPR Proteins; and Macromolecular Networks Underlying Chloroplast Biogenesis
Date initiated: 9/11/13
Date completed: 9/11/15

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
State: OR
Funding Organization: NSF
Award #: 1150203
Title: Upon Which Selection Can Act: Quantifying How Mutation and Environment Generate Genotypic & Phenotypic Variation in an Emerging Ecological & Evolutionary Genomic Model
Date initiated: 7/4/13
Date completed: 7/1/17

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: Tim Yen
State: PA
Funding Organization: NSF
Award #: CA169706
Title: Chemosensitization of Pancreatic Cancer Cells by Curcumin and Vitamin D Receptor
Date initiated:
Date completed: 1/24/13

We have RNAseq data needing to be analyzed aiming to identify critical target genes of VDR that promote survival in response to drugs.

PI: Vincent P Buonaccorsi
State: PA
Funding Organization: NSF
Award #: 1248096
Title: RCN-UBE - GCAT-SEEK: The Genome Consortium for Active Undergraduate Research and Teaching Using Next-Generation Sequencing
Date initiated: 6/17/13
Date completed: 1/31/15

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.

10.2. Projects receiving ongoing support during PY2 and initiated in PY1

PI: Marcus Breese
State: IN
Funding Organization: None, but in areas funded by NSF
Award #: NA
Title: Next gen-sequencing analysis
Date initiated: 3/12/12
Date completed: 3/12/14

I want to examine the usefulness of an extremely high-memory system in mapping of next generation sequencing data as compared to current lower-memory clusters.

PI: Matthew Hahn
State: IN
Funding Organization: None, but in areas funded by NSF
Award #: NA
Title: Yeast Multinucleotide Mutational Events/CAFE error modeling and testing
Date initiated: 7/11/12
Date completed: 7/11/13

I am currently working on two projects which require me to run software which takes days on my personal computer and would like to speed this up. One project involves intense analysis of the yeast genome, comparing up to four strains of the organism against each other and analyzing for local mutational events. The other project requires me to run hundreds of simulations using the Cafe software which estimates gene family evolution rate. Each simulation can take from 10 minutes to an hour, so running hundreds of these takes quite some time. Hopefully by utilizing the Mason server I can speed up my work and get results faster.

PI: Matthew Hahn
State: IN
Funding Organization: NSF
Award #: DBI-0845494
Title: CAREER: Computational and statistical genomics of gene families
Date initiated: 4/5/12
Date completed: 4/5/14

Identification of the number of genes in a gene family is critically dependent on accurate genome assembly. NCGAS will allow us to explore the effects of assembly on this number.

PI: Melissa Toups
State: IN
Funding Organization: None, but in areas funded by NSF
Award #: NA
Title: Genomic consequences of sex-chromosome evolution in Aedes aegypti
Date initiated: 3/12/12
Date completed: 8/1/13

In some species the non-recombining region of the sex chromosome includes only a small portion of the chromosome (homomorphic sex chromosomes), whereas in other species this region encompasses the entirety of the sex chromosomes (heteromorphic sex chromosomes). In the accepted model of sex-chromosome evolution, the non-recombining region progressively expanded from only the portion near the sex-determining locus to nearly the full extent of the sex chromosomes. However, why this progression from homomorphic sex chromosomes occurs in some species and not other remains a puzzling phenomenon. Investigating this phenomenon in Aedes aegypti, which has been inferred to have had homomorphic sex chromosomes for at least the last 100-150 million years [6], may provide some insight. In order to gain insight into the forces that maintain homomorphic sex chromosomes, we will assemble the genome and examine a variety of processes hypothesized to be associated specifically with heteromorphic sex-chromosome systems in the homomorphic sex-chromosome system of Ae. aegypti. The Aedes aegypti genome is comprised of over 1500 scaffolds, with only ~250 scaffolds mapped onto chromosomes. In order to determine the complete genetic content of the individual chromosomes in Ae. aegypti, I will create a high resolution linkage map using restriction-site associated DNA (RAD) tags in an F6 recombinant population. I have developed ~1500 markers. I am currently using Rqtl to construct the linkage map. However, the computers in the Hahn lab do not have enough memory for me to perform the analyses. Using the NCGAS will allow me to perform these analyses.

PI: Yuzhen Ye
State: IN
Funding Organization: NSF
Award #: DBI-0845685
Title: Computational Protein Function Annotation for Metagenomics
Date initiated: 4/14/12
Date completed: 4/14/14

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
State: IL
Funding Organization: NSF
Award #: 1119098
Title: Phylogeny, biogeography, and diversification in Pedicularis (Orobanchaceae)
Date initiated: 7/11/12
Date completed: 7/11/14

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: Laura Landweber
State: NJ
Funding Organization: NSF
Award #: 900544
Title: Assembly and analysis of the scrambled germline genome of Oxytricha trifallax.
Date initiated: 2/20/12
Date completed: 2/20/14

The unicellular ciliate Oxytricha trifallax possesses two types of nuclei: a transcriptionally active somatic macronucleus and a germline micronucleus involved in sexual conjugation. During development of the soma from the germline, Oxytricha accomplishes 95% genome reduction by eliminating a large number of noncoding sequences that interrupt gene segments and rearranging the remaining DNA fragments by inversions or permutations to assemble functional genes. While the sequencing of Oxytricha's somatic genome has reached final draft stage, much less is known about the germline genome. The Landweber lab is sequencing the germline genome in order to better understand processes of genome rearrangement such as internal deletion, unscrambling and chromosome fragmentation. The genome is estimated to be ~1Gb large and contains a large number of transposons and repetitive elements, which pose big challenges to the assembly. We are currently generating hundreds of millions of short reads using a combination of shotgun Illumina sequencing and fosmid sequencing. We plan to assemble the germline genome with short read assembly tools SOAPdenovo and ABYSS. The NCGAS computing resources will greatly facilitate the assembly and downstream analysis of the genome.

PI: Thomas M. Williams
State: OH
Funding Organization: NSF
Award #: 1146373
Title: Collaborative Research: The structure, function, and evolution of a regulatory network controlling sexually dimorphic fruit fly development
Date initiated: 1/1/12
Date completed: 1/1/14

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. To carry out these gene expression analyses we seek the support of NCGAS.

PI: Punidan Jeyasingh
State: OK
Funding Organization: NSF
Award #: 924401
Title: Collaborative Research: Organism-environment interactions - impact of cultural eutrophication on Daphnia tracked by genomics, physiology and resurrection ecology
Date initiated: 4/9/12
Date completed: 07/13/13

Use microarray and other technologies to discover genes and variation in their transcription linked to the phenotypic plasticity and adaptation to lake-eutrophication.

PI: Alex Buerkle
State: WY
Funding Organization: NSF
Award #: 1050149
Title: Genomic outcomes of repeated hybrid speciation
Date initiated: 7/15/12
Date completed: 7/15/14

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.

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

PI: Kenro Kusumi
State: AZ
Funding Organization: None, but in areas funded by NSF
Award #: NA
Title: Decoding the Evolution of Tail Regeneration in Anole Lizards
Date initiated: 10/18/12
Date completed: 08/31/13

Lizards are the vertebrates most closely related to humans that have the regenerative capacity to form extensive new musculoskeletal and spinal cord tissue (reviewed in Alibardi, 2010). Lizards, including the first reptile sequenced to date, Anolis carolinensis, have evolved the ability to autotomize, or self-amputate, their tails in response to a threat, and then to regenerate hyaline cartilage, de novo muscle groups, and spinal cord in a new tail. One of the most spectacular examples of lizard diversity and adaptive radiation in vertebrates are the over 375 species of anoles that inhabit the islands and mainland surrounding the Caribbean basin. Within the Anolis genus, there are substantial differences in the ability of the tail to autotomize and regenerate. We will be building on our RNA-Seq transcriptomic analysis of the regenerating tail in A. carolinensis with a comparison of genomic and transcriptomic data in multiple Anolis species in order to identify coding and noncoding sequence regulating differences in regenerative capacity. This analysis will require de novo genome assembly of other Anolis species, including gene annotation, as well as transcriptomic analysis of RNA-Seq data in the regenerating tail, including identification of specific sites of sequence divergence between species as well as between individuals of a species. The A. carolinesis genome is estimated to be approximately 1.78 gigabase pairs (Gb), and we expect other Anolis species to be of similar size. Due to the size of this genome, the analysis, particularly the de novo genome assembly, will be memory intensive and as such we request time on the Mason high-memory cluster.

PI: Alexander Christou
State: IN
Funding Organization: None, but in areas funded by NSF
Award #: NA
Title: I529 Classwork
Date initiated: 3/24/13
Date completed: 5/5/13

NCGAS resources for the purpose of classwork (I529 with professor Ye - group class project). This project served 17 students.

PI: Christine Picard
State: IN
Funding Organization: None, but in areas funded by NSF
Award #: NA
Title: De novo genome assembly of Phormia regina
Date initiated: 6/19/13
Date completed: 6/19/15

We have obtained ~100Gb of genomic sequence data (paired-end Illumina reads) for which we are assembling the genome de novo (no reference genome). We have a preliminary assembly done using CLC Genomics Workbench on a local machine, but we'd like to get some additional assemblies for comparison. We'd like to use Abyss and SOAPdenovo (and perhaps a 3rd option). We have also collected an additional ~30Gb of RAD-seq data on the same species and would like to use Stacks (open source) software for the analysis of this data in order to do some SNP discovery.

PI: 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: 1/25/13
Date completed: 1/25/15

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 by our support of these efforts and 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: Irene Newton
State: IN
Funding Organization: None, but in areas funded by NSF
Award #: NA
Title: Wolbachia transcriptome assembly
Date initiated: 3/7/2013
Date completed: 9/30/2013

We would like to determine the transcriptional profile of three different strains of an intracellular bacterium called Wolbachia. Each strain causes slightly different phenotypes upon infection and two, although closely related, cause extremely different phenotypes. Differences in gene expression will inform downstream genetic and genomic analyses for our lab.

PI: Matthew Segar
State: IN
Funding Organization: None, but in areas funded by NSF
Award #: NA
Title: An n-gram based probabilistic method for de novo sequence assembly
Date initiated: 10/15/12
Date completed: 12/31/12

For next-generation sequencers, reconstructing an organism's genome from millions of reads is a computationally expensive task. Our algorithm solves this problem by organizing and indexing the reads using n-grams, which are short, fixed-length DNA sequences of length n. These n-grams are used to efficiently locate putative read joins, thereby eliminating the need to perform an exhaustive search over all possible read pairs. We have developed a novel n-gram method for the assembly of genomes from next-generation sequencers. Specifically, a probabilistic, iterative approach was utilized to determine the most likely reads to join through development of a new metric that models the probability of any two arbitrary reads being joined together. I would like to use NCGAS resources to compare our method to other publicly available software packages. I have presented this work as ISMB 2012 and am in the process of writing a paper.

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: 1/1/13
Date completed: 1/1/15

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: Milan Radovich
State: IN
Funding Organization: None, but in areas funded by NIH
Award #: NA
Title: Comprehensive genomic analyses of next-generation sequencing data from The Cancer Genome Atlas
Date initiated: 6/17/13
Date completed: 12/31/13

The Cancer Genome Atlas is a NCI-led initiative to apply a plethora of genome-wide technologies to 25,000 tumors representing the most common cancers seen in the US. Our group is interested in leveraging this vast genomic resource coupled with clinical annotation to understand the transcriptional and genomic dysregulation that underlies these malignancies. Specifically, our group plans to analyze the entire cohort of TCGA DNA & RNA-sequencing data to understand differential 3'UTR usage and its association with clinical outcome and microRNA interactions; to identify and determine the prevalence of somatically mutated non-coding RNAs; to search for novel forms of somatic mutations; and to understand the role of hypoxia in transcriptome regulation. Because of the size of the data and the need for scalability to complete these analyses across thousands of tumor samples, we are requesting assistance from NCGAS to enable timely download of the data from TCGA, and to provide high-capacity data storage and computing resources for data processing. We anticipate several high-impact publications and funding opportunities from this work, but more importantly, we plan to use this data to inform our drug and biomarker development to translate these results back to the cancer clinic.

11. Appendix 2. Scientific products

In PY2, 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 metric inclusion as a co-author.

11.1. Biological research and bioinformatics papers published by NCGAS clients

11.1.1. Peer-reviewed journal technical papers – published or in press

1) Mingjie J., 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., Richmond, K., Donald, P., Keating, D., Bone, G., Tiedje, J.M., Noguera, D.R., Dale, B.E., Landick, R., Zhang, Y. and Balan, V. (2013) Effect of storage conditions on the stability and fermentability of lignocellulosic hydrolysate. Bioresource Technology, Accepted
2) 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) The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color Nature Protocols 8, 1494–1512 (2013) doi:10.1038/nprot.2013.084
3) Christie, A.E., ***Roncalli, V.***, *Wu, L.-S., Ganote, C.L., Doak, T.G.*, **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.
4) Motamayor JC, **Mockaitis K**, Schmutz J, Haiminen N, Iii DL, Cornejo O, Findley SD, Zheng P, Utro F, Royaert S, Saski C, Jenkins J, Podicheti R, Zhao M, Scheffler BE, Stack JC, Feltus FA, Mustiga GM, Amores F, Phillips W, Marelli JP, May GD, Shapiro H, Ma J, Bustamante CD, Schnell RJ, Main D, **Gilbert D**, Parida L, Kuhn DN. (2013) 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. [Epub ahead of print] PMID: 23731509
5) Swart EC, ***Bracht JR***, Magrini V, Minx P, ***Chen X***, Zhou Y, ***Khurana JS, Goldman AD***, Nowacki M, Schotanus K, Jung S, Fulton RS, Ly A, McGrath S, Haub K, Wiggins JL, Storton D, Matese JC, Parsons L, Chang WJ, Bowen MS, Stover NA, Jones TA, Eddy SR, Herrick GA, *Doak TG*, Wilson RK, Mardis ER, **Landweber LF**. 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 PMID: 23382650; PubMed Central PMCID: PMC3558436.
6) ***Zhang Q***, *Doak TG*, **Ye Y**. Expanding the catalog of cas genes with metagenomes. Nucleic Acids Res. 2013 Dec 6. [Epub ahead of print] PubMed PMID: 24319142.
7) **Stanton-Geddes J**, Paape T, Epstein B, Briskine R, Yoder J, et al. (2013) Candidate Genes and Genetic Architecture of Symbiotic and Agronomic Traits Revealed by Whole-Genome, Sequence-Based Association Genetics in Medicago truncatula. PLoS ONE 8(5): e65688. doi:10.1371/journal.pone.0065688

11.1.2. Peer reviewed journal papers in-review

1) **Denton, J.F**., J. Lugo Martinez, **A. Tucker**, ***D.R. Schrider***, W.C. Warren, and **M.W. Hahn** (submitted) Extensive error in the number of genes inferred from draft genome assemblies. PLoS Computational Biology.
2) ***Ioanna Ntai***, Kyunggon Kim, Ryan T. Fellers, Owen S. Skinner, Archer D. Smith IV, Bryan P. Early, John P. Savaryn, *Richard D. LeDuc*, Paul M. Thomas, and **Neil L. Kelleher** (Under Review) Applying Label-Free Quantitation to Top Down Proteomics. Analytic Chemistry
3) ***Karen Lizel GoChua Uy***, *Richard LeDuc, Carrie Ganote*, and **Donald K Price** (Under Review) Physiological effects of heat stress on Hawaiian picture-wing Drosophila: genome-wide expression patterns and stress-related traits. The Journal of Experimental Biology
4) ***Amy S. Biddle, Kelly N. Haas*** and **Jeffrey L. Blanchard**. (In review) Early establishment and persistence of microcosm community diversity through three years of serial transfers. In review.

11.1.3. Journal papers in-preparation

1) **Kusum**i and others (In Preparation) Sequencing of two divergent anole lizard species for comparative analysis of genomic divergence in adaptive radiation. Targeting Genome Biology.
2) ***Chen, X***., ***Bracht, J.R., Goldman, A.D***., Dolzhenko, E., Swart, E.C., ***Clay, D.M***., Perlman, D.H., *Doak, T.G.*, Stuart, A., Amemiya, C.T., and **Landweber, L.F.** (In preperation) The architecture of a scrambled genome reveals massive levels of genomic rearrangement during development.
3) Kristen M. DeAngelis, Grace Pold, Begum Topcuoglu, Linda van Diepen, Rebecca Varney, **Jeffrey Blanchard**, Jerry Melillo, Serita Frey . (In preparation) Measures of diversity suggest that a small subset of species respond to long-term warming.

11.1.4. Poster presentations

1) ***McGrath, Casey L., Jean-Francois Gout, Parul Johri***, *Thomas G. Doak*, **Michael Lynch** Consequences of whole genome duplication in Paramecium sps. Talk presented at the International Society of Protistologists 3rd North American Section Meeting , North Carolina Central University NC, 22-23 September, 2012.
2) ***McGrath, Casey L., Jean-Francois Gout, Parul Johri,*** *Thomas G. Doak*, **Michael Lynch**. 2013. Consequences of whole genome duplication in Paramecium. Talk presented at the International Society of Protistologists 3rd North American Section Meeting, North Carolina Central University NC, 22-23 September, 2012.
3) ***McGrath, Casey L., Jean-Francois Gout, Parul Johri***, *Thomas G. Doak*, **Michael Lynch** Consequences of whole genome duplication in Paramecium sps. Talk presented at the Midwest Protozoology Society Meeting, Bradley University, Peoria IL, April 27, 2013
4) ***McGrath, Casey L., Jean-Francois Gout, Parul Johri***, *Thomas G. Doak*, **Michael Lynch** Consequences of whole genome duplication in Paramecium sps. Talk presented at the 2013 GDRE IRSES Paramecium Genome Dynamics and Evolution, Tallinn Estonia, 12 - 16 May 2013.
5) ***McGrath, Casey L., Jean-Francois Gout, Parul Johri***, *Thomas G. Doak*, **Michael Lynch** Consequences of whole genome duplication in Paramecium sps. Talk presented at the 2013 FASEB Ciliate Molecular Biology meeting, Steamboat Springs, CO., July 7 - 12, 2013

11.2. Methods-oriented and outreach papers, presentations, and materials by NCGAS staff

11.2.1. Peer-reviewed journals

1) *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". Published on line 8/20/2013: doi:10.1136/amiajnl-2013-002059.
2) **Hahn, M.W**., ***S.V. Zhang***, and L.C. Moyle (in press) Sequencing, assembling, and correcting draft genomes using recombinant populations. G3.
3) *Richard D. LeDuc*, ***Ryan T. Fellers***, Bryan P. Early, ***Joseph B. Greer***, Paul M. Thomas, and **Neil L. Kelleher** (Accepted pending Revision) The C-Score: A Bayesian Framework to Sharply Improve Proteoform Scoring in High-Throughput Top Down Proteomics. Journal of Proteomics Research

11.2.2. Peer-reviewed conference papers

1) *LeDuc, R., Wu, L.-S., Ganote, C., Doak, T., Blood, P., Vaughn, M., and Williams, B*. (2013) National Center for Genome Analysis Support Leverages XSEDE to Support Life Science Research. Proceedings of XSEDE 13, San Diego CA. 7/22/2013.

11.2.3. Posters

1) *Barnett, William K. and Richard D. LeDuc.* "The National Center for Genome Analysis Support: Providing Free or Low-Cost Bioinformatics Support at Scale", Poster presented at the 2013 AMIA Translational Sciences Summit, San Francisco, CA. March 19, 2013. - Note that this poster was awarded a Silver Medallion.
2) *Doak, Thomas G., Richard LeDuc, Le-Shin Wu, Craig A. Stewart, Robert Henschel, William K. Barnett.* 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.
3) *Doak, Thomas G., Richard LeDuc, Le-Shin Wu, Craig A. Stewart, Robert Henschel, William K. Barnett.* National Center for Genome Analysis Support. Poster presented at the 1st International Conference on Genomics, The Children's Hospital of Philadelphia, Philadelphia PE, Sept. 27-28, 2012.
4) *Doak, Thomas G., Richard LeDuc, Le-Shin Wu, Craig A. Stewart, Robert Henschel, William K. Barnett*. National Center for Genome Analysis Support. Poster presented at the Midwest Protozoology Society Meeting, Saturday, Bradley University, Peoria IL. April 27, 2013.
5) *Doak, Thomas G., Richard LeDuc, Le-Shin Wu, Craig A. Stewart, Robert Henschel, William K. Barnett.* National Center for Genome Analysis Support. Poster presented at the 2013 FASEB Ciliate Molecular Biology meeting, Steamboat Springs, CO., July 7 - 12, 2013
6) *Doak, Thomas G., Richard LeDuc, Le-Shin Wu, Craig A. Stewart, Robert Henschel, William K. Barnett.* National Center for Genome Analysis Support. Poster presented at the 2013 GDRE IRSES Paramecium Genome Dynamics and Evolution, Tallinn Estonia, 12 - 16 May 2013.
7) *Doak, Thomas G.* Why NCGAS, and what is NCGAS. Talk presented at the 2013, NCGAS Summer interns workshop, Indiana University, Bloomington IN.

11.2.4. Presentations (not peer reviewed)

1) *Richard LeDuc and Bill Barnett,* National Center for Genome Analysis Support Leverages XSEDE to Support Life Science Research, XESED 13, 2013-07-23 San Diego, CA. https://scholarworks.iu.edu/dspace/handle/2022/16402
2) *Richard LeDuc*, Optimizing the National Cyberinfrastructure for Lower Bioinformatic Costs: Making the Most of Resources for Publicly Funded Research. RNA-Seq 2013, 2013-06-20. Boston, MA
3) *Richard LeDuc*, Systems Biology Data Analysis. July; 2013. Computational Approaches to Analyzing Microarray Data (http://www.btci.org/courses/intermediate/caamd/2012/caamd12.html). Madison, WI
4) *Richard LeDuc*, 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
5) *Richard LeDuc,* Using Prior Knowledge to Improve Scoring in High-Throughput Top-Down Proteomics Experiments. American Society of Mass Spectrometerists annual meeting, 2013-06-08, Minneapolis, MN. https://scholarworks.iu.edu/dspace/handle/2022/16680

6) *Richard LeDuc,* 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
7) *Richard LeDuc*, Careers in Computing and Science. Clark State University, Dayton OH, 2012-09-27. https://scholarworks.iu.edu/dspace/handle/2022/14745
8) *Bill Barnett*, The National Center for Genome Analysis Support as a Model Virtual Resource for Biologists, Internet2 Network Infrastructure for the Life Sciences Focused Technical Workshop. Berkeley, CA, July 17, 2013

12. Appendix 3: Software Supported by NCGAS

12.1. Bioinformatics software supported by NCGAS

This table summarizes the bioinformatics software applications installed and supported on the Mason cluster in NCGAS PY2. In all cases, the software was developed outside of NCGAS and a production version was released by its 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 6: 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?	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	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	Limited agreement for academic use	9					
	1.3.6	No	Yes	Limited agreement for academic use	9					

	1.3.6-openmpi	No	Yes	Limited agreement for academic use	9					
allpathslg	41292	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		9					
	45684	No	Yes		9					
amos	3.0.0	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	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	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	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	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	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	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	GNU GPL v3	9			http://sourceforge.net/projects/bowtie-bio/files/bowtie2/	No	No
bwa	0.6.2	No	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	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	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	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	GNU GPL	9	Not yet determined			No	No
cufflinks	2.0.2	No	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	Boost License					No	Yes
cutadapt	1.2.1	No	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	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	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	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
galaxy	1	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/Get Galaxy	No	Yes
gatk	1.1-33	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	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	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
mummer	3.22	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	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	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	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
oases	0.2.08	No	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	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	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							
sam2counts	1.0	No	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	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							
shore	0.6.1beta	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	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
soapdenovo	1.04	Yes	Yes	GNU	9	Not yet	De novo assembly of short reads	Version 1 not available,	No	No

	1.05	Yes	Yes	GPL v3		determined	for large genomes, creating reference genomes of novel organisms	see: http://sourceforge.net/projects/soapdenovo2/files/SOAPdenovo2/		
sra-toolkit	2.1.15	No	Yes	GPL v 2 or gr eater	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
tophat	2.0.5	No	Yes	Boost Licens e	9	Not yet determined	Alignment for RNA-Seq data against reference for finding splice junctions	http://tophat.cbcb.umd.edu/downloads/	No	Yes
transabyss	1.3.2	Yes	Yes	BCCA Acade mic Licens e	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	BSD Licens e	9	Open Source Developmen t	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							
velvet	1.2.03-k111	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
	1.2.03-k111-openmp	Yes	Yes							

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

This table provides reference information about the software installed on Mason in PY2

Table 7: Technical Properties of Bioinformatic Software Supported on the Mason Cluster.

Software	Version	Software Dependencies?	Software Development Methodology	Software Functionality
abyss	1.3.3 1.3.3-openmpi 1.3.6 1.3.6-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	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	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
blat	35	none	Not yet determined	Input: Sequence query and database in Fasta, .nib or .2bit format; Output: Alignment in .psl format

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
bowtie	0.12.8 2.1.0	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	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.
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
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
edena	2.1.1	none	Not yet determined	Input: Fasta or Fastq short reads file; Output: Assembled contigs (Fasta format?)
Fastqc	0.10.1	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	11/09/12	Gcc, Perl	Not yet determined	Input: Reference genome in Fasta format, Query in Fasta format; Outputs: Alignment file either compressed or uncompressed
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	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 7.2.8	Gcc	Not yet determined	Input: Phylip file containing alignments to be run; Output: Text files containing tree topologies, logfiles, intermediate files
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
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	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	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	Gcc	Not yet determined	Input: SRA file, Output: Fasta/Fastq file
tophat	2.0.5	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	10/05/12 02/05/13	Gcc, Samtools, Bowtie	Open source development	Input: Fastq files containing reads; Output: Assembled contigs in Fasta format
velvet	1.2.03-k111 1.2.03-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.

12.3. NCGAS software support across XSEDE resources.

This table 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 8: Bioinformatic Software Supported by Non-Indiana University NCGAS Partners.

Software	Stampede (TACC)	Lonestar (TACC)	Rockhopper (IU)	Blacklight (PSC)
abyss		X	X	X
allpathslg	X	X	X	X
amos	X	X		
AFNI				
arachne	X			
bedtools	X	X		X
bio3d				
bioconductor				
bitseq				
blat	X	X		X
bowtie	X	X	X	X
bwa	X	X	X	X
cafe				
cd-hit		X		
celera			X	
cufflinks	X	X		X
cytoscape				
edena	X	X		
egglib				
euler				
fastqc	X	X		X
fsl				

galaxy				
gatk	X	X		
genomemapper	X	X		
gmap	X	X		
hmmer	X	X	X	X
maker		X		
metamos				
mlRho				
mothur				
mummer	X	X		
namd	X	X	X	X
ngsutils				
ninja	X	X		
novoalign		X		
oases	X	X		X
picard	X	X		X
raxml	X	X		
rsem				X
samtools	X	X		X
scythe	X	X		
shore		X		
smrt				
soapdenovo	X	X	X	X
spm				
sra-toolkit	X	X		
stacks		X		
tophat	X	X		X
transabyss				X
trinityrnaseq	X	X		X
velvet	X	X		X

13. Appendix 2: NCGAS Education, Outreach, and Training Activities

13.1.1. Press Releases

1) March 2013: Article for Inside XSEDE. NCGAS and XSEDE join forces for life sciences research. The National Center for Genome Analysis Support (NCGAS) and the Extreme Science and Engineering Discovery Environment (XSEDE) recently teamed up to showcase their growing support for life sciences research. Positioning themselves as bioinformatics resources for genome science, the two groups hosted a joint booth at the annual Plant and Animal Genome (PAG) Conference, which took place Jan. 12-16 in San Diego.
2) Jan. 14, 2013: Genome analysis support center announces new initiatives to foster discovery. Now entering its second year of operation, the National Center for Genome Analysis Support (NCGAS) provides software, expert consultation and computational resources to help life science researchers analyze genome data. NCGAS is expanding its reach by adding tools, services and partners to help biological research communities make important new scientific discoveries. Distribution: Sent to these targeted publications: Chronicle of Higher Education, the Herald-

Times, Inside Indiana Business, GenomeWeb.com, Bio-IT World, Supercomputing Online, HPCwire, Indianapolis Star, Scientific Computing World
Placements: HPCwire, Supercomputing Online (link no longer active)

3) July 23, 2013: Genome analysis center adds data storage, curation to its services. The National Center for Genome Analysis and Support (NCGAS) has announced new services to help biological researchers store and share their work, speeding scientific discovery and breakthroughs in many fields of study. NCGAS directors made the announcement at the eXtreme Science and Engineering Discovery Environment (XSEDE) 2013 annual conference in San Diego.
Distribution: Sent to 1,161 media outlets
Placements: Supercomputing Online

13.1.2. Education, outreach, and training events and participants

Type	Title	Location	Date	Hours	Number of Participants	Number of individuals from traditionally underserved groups (TUGs)*	Method†	Funding Sources
Tutorial / workshop	Careers in Computing Science Clark State	Springfield, OH	09/27/12	1	22	4		NSF Award #1062432
Tutorial / workshop	RNA-Seq Workshop at the World Genome Data Analysis Summit	San Francisco, CA	11/26/12	3	8	3		NSF Award #1062432
Conference display (walk-up booth)	NCGAS exhibits at PAG XX	San Diego, CA	Jan-13	32	200	35		NSF Award #1062432
Conference talk/presentation/panel	Statistical Consideration for Identification and Quantification in Top-Down Proteomics	St Pete Beach, FL	2013-01-27	1	127	20		Invited
Conference talk/presentation/panel	Internet2 Annual Meeting	Arlington, VA	April 22, 2013	1	8	0		NSF Award #1062432
Academic meeting	Midwest Protozoology Society Meeting	Bradley University Peoria, Illinois	April 27, 2013	8	50	5		*NSF EF-0328516-A006*
Academic meeting	GDRE IRSES "Paramecium Genome Dynamics and Evolution"	Tallinn Estonia	12 - 16 May 2013	8	50	0		*Self*

Conference talk/presentation/panel	Next Generation Cyberinfrastructures for Next Generation Sequencing and Genome Science	Vancouver, CA	6/5/2012	1	60	3		*NSF Award #1062432*
Conference talk/presentation/panel	Leveraging the National Cyberinfrastructure for Top Down Mass Spectrometry	Minneapolis, MN	6/8/2013	1	120	3		*NSF Award #1062432*
Tutorial / workshop	Clark State CC visits NCGAS	Bloomington, IN	6/14/2013	5	11	2		*NSF Award #1062432*
Conference talk/presentation/panel	Optimizing the National Cyberinfrastructure for Lower Bioinformatic Costs: Making the Most of Resources for Publicly Funded Research	Boston, MA	6/20/2013	1	130	2		NSF Award #1062432
Conference display (walk-up booth)	NCGAS exhibit at Evolution 2013	Snowbird, UT	June 21-25	20	50	25		NSF Award #1062432
Conference talk/presentation/panel	FASEB Ciliate Molecular Biology	Steamboat Springs, Colorado	July 7 - 12, 2013	1	200	5		*NSF Award #1062432*
Conference display (walk-up booth)	NCGAS exhibit at SMBE 2013	Chicago, IL	July 7-11	20	20	5		NSF Award #1062432
Conference talk/presentation/panel	The National Center for Genome Analysis Support as a Model Virtual Resource for Biologists	Internet2 Network Infrastructure for the Life Sciences Focused Technical Workshop. Berkeley, CA	July 17, 2013	1	50	0		*NSF Award #1062432*
Conference talk/presentation/panel	XSEDE 13 BOF	San Diego, CA	July 21, 2013	1	35	12		NSF Award #1062432
Tutorial / workshop	Systems Biology Data Analysis	Computational Approaches to Analyzing Microarray Data BioPharmaceutical Technology Center Institute Madison WI	July 24-28	4	17	6		*Invited*

Conference display (walk-up booth)	American Indian Higher Education Commission (AIHEC) AIHEC 40th Anniversary Conference	Santa Fe, New Mexico	August 7-10, 2013	32	48	400		*NSF Award #1062432*
Tutorial / workshop	New Directions for Investigating Biodiversity of Ciliates	National Evolutionary Synthesis Center and North Carolina Central University Durham, NC	September 19-22, 2012	12	30	3		*NSF Award #1062432*
Academic meeting	International Society of Protistologists 3rd North American Section Meeting	North Carolina Central University	22-23 September, 2012	8	50	10		*NSF Award #1062432*
Conference display (walk-up booth)	Cluster 13	Indianapolis, IN	9/24 and 9/25	8	60	15		
Academic meeting	1st International Conference on Genomics	The Children's Hospital of Philadelphia, Philadelphia PE	Sept. 27-28, 2012	8	200	20		*NSF Award #1062432*
Tutorial / workshop	IU Galaxy for Bioinformatics	Bloomington, IN	09/30/2013, 10/01/13	2	6	0		*NSF Award #1062432*
Tutorial / workshop	Galaxy Workshop at IU Bloomington	Bloomington, IN	10/19/12	3	111	35		*Base*
Conference display (walk-up booth)	Hispanic Association of Colleges and Universities/Chicago, IL	Chicago, IL	Oct. 26-28	11	45	45		*NSF Award #1062432*
Totals	**25 Events**			193	1,708	658		

Table 9. EOT activities for PY2 for NCGAS.

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

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