National Center for Genome Analysis Program Year 1 Report
– September 15, 2011 – September 14, 2012

William K. Barnett, Ph.D.
Carrie Ganote
Matthew Vaughn, Ph.D.*
Richard D. LeDuc, Ph.D.
Craig A. Stewart, Ph.D.

Indiana University
PTI Technical Report PTI-TR13-002
4 March 2013

Citation:

* Texas Advanced Computing Center, Austin TX
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 all tough the National Center for Genome Analysis Support is funded by NSF 1062432, the work that we do 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 PTI was supported in part by the Indiana Genomics Initiative. The Indiana Genomics Initiative of Indiana University is supported in part by Lilly Endowment, Inc.
- The PTI was supported in part by the Indiana METACyt Initiative. The Indiana METACyt Initiative of Indiana University is supported in part by Lilly Endowment, Inc.
- This material is based in part upon work supported by the National Science Foundation under Grant No. CNS-0521433.

Further, any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation (NSF).
Table of Contents

1. Executive Summary .......................................................................................................................... 1
2. Introduction ....................................................................................................................................... 2
3. Progress on Program Year 1 milestones .......................................................................................... 3
   3.1. Milestones ................................................................................................................................... 3
4. Accomplishments – software and servers deployed and supported ............................................ 3
   4.1. Metrics of use of software supported by NCGAS ................................................................. 3
   4.2. Technical system architecture ................................................................................................. 8
   4.3. Servers and job statistics ........................................................................................................... 9
   4.4. Utilization breakdown ................................................................................................................ 10
   4.5. Key software deployment /improvement / hardening ............................................................. 11
   4.6. User contacts ............................................................................................................................. 11
   4.7. Short-term consultations ........................................................................................................... 12
   4.8. Long-term consultations ........................................................................................................... 12
   4.9. Project support ......................................................................................................................... 12
5. Training and development activities ............................................................................................... 13
6. Outcomes and work products .......................................................................................................... 14
   6.1. Publications ............................................................................................................................... 14
      6.1.1. Journal papers .................................................................................................................... 14
      6.1.2. Conference papers ............................................................................................................ 14
      6.1.3. Conference proceedings ................................................................................................... 14
      6.1.4. Journal articles ................................................................................................................... 14
      6.1.5. Poster presentations ........................................................................................................... 15
      6.1.6. Presentations ...................................................................................................................... 15
      6.2. Web pages ............................................................................................................................... 16
7. Management ..................................................................................................................................... 16
   7.1. Summary of needs identified with discussion with SAB chair, co-PIs, and key constituents and partners .................................................................................................................. 17
   7.2. Development of a sustainability plan ....................................................................................... 17
8. Science outcomes ............................................................................................................................. 17
   8.1. Science highlights ..................................................................................................................... 18
   8.2. Education, outreach, and training highlight ............................................................................. 24
   8.3. Supported projects .................................................................................................................... 25
      8.3.1. NSF-funded genomics ....................................................................................................... 25
      8.3.2. Non-NSF-funded research projects on the Mason cluster ................................................ 33
Table of Tables

Table 1. Progress towards each of the PY1 milestones proposed for the NCGAS project. ......................... 3
Table 2. Bioinformatics software supported by NCGAS. .................................................................................. 6
Table 3. Technical descriptions of software supported by NCGAS. This table provides reference information about the software installed on Mason in PY1 ........................................................................ 7
Table 4. NCGAS software support across partners. This table shows the level of shared support for bioinformatics software on the various NCGAS partner compute resources in PY1 ....................... 8
Table 5. EOT activities for PY1 for NCGAS. ................................................................................................. 14

Table of Figures

Figure 1. Distribution of the 85 users who submitted jobs to the Mason cluster in PY1. ......................... 9
Figure 2. Number of allocation requests received from REDCap by month for PY1 ......................... 10
Figure 3. Jobs submitted to the Mason Cluster per week, and wall time consumed on the Mason cluster per week over PY1 ........................................................................................................ 11
Figure 4. Map showing the location of NCGAS supported projects. .................................................. 13
Figure 5. A graph showing the improved computational performance of Trinity running on biological data sets of increasing size .................................................................................................................. 19
Figure 6. The Drosophila species of Hawaii can be found in a variety of habitats ranging from upland deserts to this tropical rain forest ........................................................................................................... 20
Figure 7. RNA-sequence analyses pose unique assembly issues that required specialized software. Using Trinity installed on the Mason cluster NCGAS was able to create a list of candidate gene models in copepods. Given the prior knowledge of copepod neuropeptides, the quality of the assembled gene models can be independently verified ........................................................................... 21
Figure 8. Vertical plankton tow in the Gulf of Maine to collect Calanus finmarchicus. Slow tows are performed to collect intact individuals for culturing and laboratory experiments .............................................. 22
Figure 9. Quality control of DNA sequence data. An important first step in any sequencing analysis is determining the overall quality of the sequence reads. NCGAS software allows researchers to test the quality of their practice runs before consuming valuable samples ........................................................................................................ 23
Figure 10. Attendees at the Daphnia Genomics Jamboree use the IU IQ-Wall to collaborate on a research paper ........................................................................................................................................... 24
1. Executive Summary

On September 15, 2011, Indiana University (IU) received three years of support to establish the National Center for Genome Analysis Support (NCGAS). This technical report describes the activities of the first 12 months of NCGAS, during which time NCGAS brought online a large-RAM computational cluster, recruited 25 NSF-funded genomics projects to use the resource, responded to 502 requests for support, and processed 28,523 computational jobs representing a total of 136.83 core years of computing.

NCGAS also laid the framework for creating a truly national-scale center supporting genomics research. By coordinating effort between multiple supercomputing centers, NCGAS is creating a service-oriented computational infrastructure – one that is designed to be approachable by end-users unaccustomed to using traditional supercomputing resources. The benefits of such inter-institutional coordination can be seen from events such as the NCGAS co-hosted Daphnia Genomics Jamboree. At this gathering, dozens of scientists from across the US and Europe spent five days accelerating the completion of the *Daphnia manga* genome. NCGAS-supported staff from both Texas Advanced Computing Center (TACC) and IU gave presentations early in the Jamboree before participants broke into small teams and used NCGAS clusters to perform their analyses. NCGAS also established a Galaxy web portal to allow researchers to use our large-RAM cluster with a familiar web interface, and we worked to increase the computational efficiency of the best-in-class Trinity application for RNA-sequence assembly.
2. Introduction

On September 15, 2011, Indiana University (IU) received a three-year, $1,460,000 grant (award no. 1062432) from the Division of Molecular and Cellular Biosciences of the National Science Foundation (NSF) to establish the National Center for Genome Analysis Support (NCGAS) in partnership with the Texas Advanced Computing Center (TACC). The award established NCGAS as an innovative service center (core facility) that supports the national community of NSF-funded researchers who use genome assembly software, particularly software suitable for assembly of data from next-generation sequencers; large-scale phylogenetic software; and other genome analysis software requiring large amounts of memory. This center is a general source of software support and services provided on the Mason large memory computer cluster at IU, on TACC’s Gordon system, and on San Diego Supercomputer Center’s (SDSC) Dash system. NCGAS provides services such as use of cluster-based genome analysis software, storage of submitted data sets, and a repository of open source genome analysis software. Services particularly support analyses of next-generation sequencer output for de novo assembly, metagenomic projects, and resequencing. NCGAS has established a core of experts and software tools to support research on a variety of nationally funded cyberinfrastructure systems, and has added to the suite of available systems a large memory cluster ideal for this work. By developing a community of investigators and technologists and exploring new modalities of provisioning computational resources, such as "on demand" computing, this project aspires to become a sustainable model for the ongoing, and increasing, need for sequence analysis. The NCGAS website (http://ncgas.org/) provides up-to-date information.

This technical report describes the activities of the first 12 months of NCGAS, from the start date through September 14, 2012. The leadership of NCGAS for this first year was Dr. Craig Stewart, PI of the NSF NCGAS award; Dr. William Barnett (Director of NCGAS), Drs. Michael Lynch, Matthew Hahn, and Geoffrey Fox (Internal Advisory Board (IAB)), and Dr. Richard McCombie (Chair, Science Advisory Board (SAB)). As with any new program, the leadership was tasked with commencing administrative processes that provide support, staffing up, defining and creating services, implementing infrastructures, creating policies and procedures, appointing advisory bodies, establishing feedback mechanisms and metrics, and initiating an outreach campaign – in addition to undertaking the central mission. Our progress towards the central mission of NCGAS is described in this report.

During the course of this first year, NCGAS was not only successful in ‘setting up shop’ and instituting policies, practices, and resources that were tightly aligned with the needs of genomics researchers, but it also supported a significant number of research projects. Indiana University provided, as a resource contribution to NCGAS, the 16-node Mason large memory cluster. Each of Mason’s nodes, outfitted with a 32-core processor and 512 gigabytes (GB) of Random Access Memory (RAM), is architected particularly for memory-intensive applications such as genome assembly. By installing a suite of genome analysis applications on Mason (see section 4) and establishing a low barrier system by which NSF-funded researchers could get access to Mason, its software applications, and bioinformatics support, NCGAS was quickly able to begin providing service to genome scientists while it built its online presence and began outreach effort to the research community. NCGAS also staffed up quickly, immediately transferring Le Shin Wu from IU’s Research Technologies (RT) division of University Information Technology Services (UITS) for computer science support and retaining part-time services from Dr. Thomas Doak, a genomics researcher in IU’s Department of Biology, to provide genomics consulting and outreach. NCGAS also contracted with TACC to support genome projects at their site. In March of PY1, NCGAS hired Dr. Richard LeDuc to provide management leadership and in early PY2 hired Carrie Ganote 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, and made 22 peer-reviewed or invited presentations, and exhibited at two major conferences attended by NSF-funded genomics researchers.
3. Progress on Program Year 1 milestones

3.1. Milestones
NCGAS set out to achieve the following milestones in PY1.

<table>
<thead>
<tr>
<th>Quarter</th>
<th>Task</th>
<th>Completed?</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1 (Oct-Dec)</td>
<td>Establish NCGAS SAB and Year 1 face-to-face</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Initial needs assessment and report</td>
<td>Yes</td>
<td>Completed in discussion with SAB Chair and community leaders</td>
</tr>
<tr>
<td></td>
<td>Establish initial NCGAS web presence</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Q2 (Jan-Mar)</td>
<td>Install software: Velvet, Celera Assembler, SOAPdenovo, ABBySS, Edena,</td>
<td>Yes</td>
<td>Total of 45 software packages installed</td>
</tr>
<tr>
<td></td>
<td>RAXML, and NINJA ready to use on IU Mason, TACC Ranger, and SDSC Gordon.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Storage infrastructure on Data Capacitor and Scholarly Data Archive at IU. Storage allocated and available.</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Year 1 outreach at two events of 100 attendees.</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>NCGAS HUB provides access to software, Knowledge Base documentation, sample sequence collections, schedules and content</td>
<td>Yes</td>
<td>HUB concept replaced with development of Galaxy portals</td>
</tr>
<tr>
<td></td>
<td>Year 1 satisfaction and needs surveys sent out</td>
<td>No</td>
<td>Delayed until PY2</td>
</tr>
<tr>
<td>Q3 (Apr-Jun)</td>
<td>Year 1 survey results report</td>
<td>No</td>
<td>Delayed until PY2</td>
</tr>
<tr>
<td>Q4 (Jul-Sep)</td>
<td>Year 1 SAB teleconference (Stewart, McCombie)</td>
<td>Yes</td>
<td>Stewart, Barnett, Hahn, Lynch attending</td>
</tr>
<tr>
<td></td>
<td>Year 1 Outreach at two events of ≥100 attendees.</td>
<td>Yes</td>
<td>10 outreach events and 22 peer-reviewed presentations or posters</td>
</tr>
<tr>
<td></td>
<td>Code repository ready to accept codes - all above codes.</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Applications evaluated for memory use and performance and repackaged - applications.</td>
<td>Yes</td>
<td>Trinity optimization project with Broad Institute completed.</td>
</tr>
</tbody>
</table>

Table 1. Progress towards each of the PY1 milestones proposed for the NCGAS project.

4. Accomplishments – software and servers deployed and supported

4.1. Metrics of use of software supported by NCGAS
The following tables detail the bioinformatics software deployed across NCGAS systems. The metrics reported represent important information that contributed to our decisions to support these software packages. The table is drawn from recommendations about software sustainability metrics contained in the report of an NSF-funded workshop (Stewart, C.A., G.T. Almes, D.S. McCaulay and B.C. Wheeler, eds., 2010. Cyberinfrastructure Software Sustainability and Reusability Workshop Final Report. https://scholarworks.iu.edu/dspace/handle/2022/6701). Over time, we expect that NCGAS will contribute code to more NCGAS-supported software. NCGAS during PY1 notably contributed to the optimization of Trinity.
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>abyss</td>
<td>1.2.7</td>
<td>Yes</td>
<td>Yes</td>
<td>Limited agreement for academic use</td>
<td>9</td>
<td>Not yet determined</td>
<td>De novo assembly of DNA for metagenomics, comparative genomics and creation of draft genomes</td>
<td><a href="http://www.bcgsc.ca/platform/bioinfo/software/abyss/releases/1.3.3">http://www.bcgsc.ca/platform/bioinfo/software/abyss/releases/1.3.3</a></td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>1.2.7-openmpi</td>
<td>Yes</td>
<td>Yes</td>
<td>Limited agreement for academic use</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.3.3</td>
<td>No</td>
<td>Yes</td>
<td>Limited agreement for academic use</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.3.3-openmpi</td>
<td>No</td>
<td>Yes</td>
<td>Limited agreement for academic use</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>allpathslg</td>
<td>38445</td>
<td>Yes</td>
<td>Yes</td>
<td>Free to use, change, distribute</td>
<td>9</td>
<td>Not yet determined</td>
<td>Whole-genome shotgun assembly using Illumina long and short insert libraries for greatest accuracy</td>
<td>ftp://ftp.broadinstitute.org/pub/crd/ALLPATHS/Release-LG/latest_source_code/</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>41292</td>
<td>No</td>
<td>Yes</td>
<td>Free to use, change, distribute</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>amos</td>
<td>3.0.0</td>
<td>No</td>
<td>Yes</td>
<td>Artistic License</td>
<td>9</td>
<td>Not yet determined</td>
<td>Assembling, Validating, Comparative, Visualizing, and Scaffolding whole genome sequence data in a pipeline</td>
<td><a href="http://sourceforge.net/projects/amos/files/amos/">http://sourceforge.net/projects/amos/files/amos/</a></td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>arachne</td>
<td>3.2</td>
<td>Yes</td>
<td>Yes</td>
<td>Free to use, change, distribute</td>
<td>9</td>
<td>Not yet determined</td>
<td>Whole genome shotgun assembly of long Sanger reads</td>
<td>ftp://ftp.broadinstitute.org/pub/crd/ARACHNE/</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>bedtools</td>
<td>2.12</td>
<td>Yes</td>
<td>Yes</td>
<td>GNU GPL v2</td>
<td>9</td>
<td>Not yet determined</td>
<td>Discovery of correlated genomic features such as ESTs, polymorphisms, mobile elements, etc.</td>
<td><a href="http://code.google.com/p/bedtools/downloads/list">http://code.google.com/p/bedtools/downloads/list</a>, <a href="http://arm.koji.fedoraproject.org/koji/packageinfo?packageID=10644">http://arm.koji.fedoraproject.org/koji/packageinfo?packageID=10644</a></td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>blat</td>
<td>35</td>
<td>No</td>
<td>Yes</td>
<td>Free for academic, non-profit or personal use, Contact for commercial licensing</td>
<td>9</td>
<td>Not yet determined</td>
<td>Fast alignment of highly similar sequences of DNA/Proteins to find ESTs or to align reads to reference</td>
<td><a href="http://users.soe.ucsc.edu/~kent/src/">http://users.soe.ucsc.edu/~kent/src/</a></td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>bowtie</td>
<td>0.12.7</td>
<td>Yes</td>
<td>Yes</td>
<td>Artistic License</td>
<td>9</td>
<td>Not yet determined</td>
<td>Alignment of short reads to a reference genome in order to approximate coverage, find polymorphisms, and assess assembly quality</td>
<td><a href="http://sourceforge.net/projects/bowtie-bio/files/bowtie/">http://sourceforge.net/projects/bowtie-bio/files/bowtie/</a></td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>2.0.0_b6</td>
<td>No</td>
<td>Yes</td>
<td>Artistic License</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bwa</td>
<td>0.5.9</td>
<td>No</td>
<td>Yes</td>
<td>GNU GPL v3, MIT License</td>
<td>9</td>
<td>Not yet determined</td>
<td>Alignment of long and short reads from a variety of technologies, allows gaps, for approximating coverage, finding polymorphisms, and assessing assembly quality</td>
<td><a href="http://sourceforge.net/projects/bio-bwa/files/">http://sourceforge.net/projects/bio-bwa/files/</a></td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>0.6.1</td>
<td>No</td>
<td>Yes</td>
<td>GNU GPL</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>No</td>
</tr>
<tr>
<td>cd-hit</td>
<td>4.5.6</td>
<td>No</td>
<td>Yes</td>
<td>GNU GPL v2</td>
<td>9</td>
<td>Not yet determined</td>
<td>Clustering program for large sets of protein and DNA to determine relationships between many sequences</td>
<td><a href="https://code.google.com/p/cd-hit/downloads/list">https://code.google.com/p/cd-hit/downloads/list</a></td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>celera</td>
<td>6.1</td>
<td>Yes</td>
<td>No</td>
<td>GNU GPL</td>
<td>9</td>
<td>Not yet determined</td>
<td>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</td>
<td><a href="http://sourceforge.net/projects/wgs-assembler/files/wgs-assembler/">http://sourceforge.net/projects/wgs-assembler/files/wgs-assembler/</a></td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>No</td>
<td>Yes</td>
<td>GNU GPL</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>No</td>
</tr>
<tr>
<td>------------</td>
<td>---------</td>
<td>-------------</td>
<td>--------------</td>
<td>-------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>-------------------------------------------------</td>
<td>--------</td>
<td>---------------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cufflinks</td>
<td>2.0.0_beta</td>
<td>No</td>
<td>Yes</td>
<td>9</td>
<td>Not yet determined Map RNA-Seq reads to reference genomes in order to annotate genes, discover splice variants, and estimate differential expression</td>
<td>Boost License</td>
<td>No</td>
<td><a href="http://cufflinks.cbcb.umd.edu/downloads/">http://cufflinks.cbcb.umd.edu/downloads/</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>edena</td>
<td>2.1.1</td>
<td>Yes</td>
<td>Yes</td>
<td>9</td>
<td>Not yet determined De novo assembly of short reads for smaller genome assembly</td>
<td><a href="http://tophat.cbcb.umd.edu/downloads/">http://tophat.cbcb.umd.edu/downloads/</a></td>
<td>No</td>
<td><a href="http://www.genomic.ch/edena.php">http://www.genomic.ch/edena.php</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>galaxy</td>
<td>1</td>
<td>No</td>
<td>Yes</td>
<td>9</td>
<td>Not yet determined A flexible GUI wrapper for bioinformatics tools allows users to manipulate genomic data and run analyses</td>
<td>mercurial install, see: <a href="http://wiki.galaxyproject.org/Admin/Get">http://wiki.galaxyproject.org/Admin/Get</a> Galaxy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gatk</td>
<td>1.1-33</td>
<td>Yes</td>
<td>Yes</td>
<td>9</td>
<td>Not yet determined Suite of genomics analysis tools with a focus on variant calling and gene finding</td>
<td>New release only: <a href="http://www.broadinstitute.org/gatk/download">http://www.broadinstitute.org/gatk/download</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>genomemapper</td>
<td>0.4.3</td>
<td>Yes</td>
<td>Yes</td>
<td>8</td>
<td>Not yet determined Short read alignment, allows gaps, allows multiple references; used for estimating coverage, finding polymorphisms, variant calling, and quantitative analysis</td>
<td><a href="http://sourceforge.net/projects/genomemapper.html">http://sourceforge.net/projects/genomemapper.html</a></td>
<td>No</td>
<td><a href="http://www.1001genomes.org/software/genomemapper.html">http://www.1001genomes.org/software/genomemapper.html</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gmap</td>
<td>7/20/12</td>
<td>No</td>
<td>Yes</td>
<td>9</td>
<td>Not yet determined Align cDNA to reference to determine gene structure and structural variants</td>
<td><a href="http://research-pub.gene.com/gmap/archive.html">http://research-pub.gene.com/gmap/archive.html</a></td>
<td>No</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mummer</td>
<td>3.22</td>
<td>Yes</td>
<td>Yes</td>
<td>9</td>
<td>Not yet determined Align very large DNA and Protein sequences to reference.</td>
<td><a href="http://sourceforge.net/projects/mummer/files/mummer/">http://sourceforge.net/projects/mummer/files/mummer/</a></td>
<td>No</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ninja</td>
<td>1.2.1</td>
<td>Yes</td>
<td>Yes</td>
<td>9</td>
<td>Not yet determined Infers phylogeny using neighbor-joining tree</td>
<td><a href="http://nimbletwist.com/software/ninja/download.html">http://nimbletwist.com/software/ninja/download.html</a></td>
<td>No</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>novoalign</td>
<td>2.07.13</td>
<td>Yes</td>
<td>Yes</td>
<td>9</td>
<td>Not yet determined Aligns short reads to reference genome for resequencing experiments</td>
<td><a href="http://www.novocraft.com/main/downloadpage.php">http://www.novocraft.com/main/downloadpage.php</a></td>
<td>No</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>picard</td>
<td>1.52</td>
<td>Yes</td>
<td>Yes</td>
<td>9</td>
<td>Not yet determined Provides tools and methods for manipulating sequence alignments for assembly quality assessment, variant calling, and downstream processing.</td>
<td><a href="http://sourceforge.net/projects/picard/files/picard-tools/">http://sourceforge.net/projects/picard/files/picard-tools/</a></td>
<td>No</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>raxml</td>
<td>7.2.6</td>
<td>Yes</td>
<td>Yes</td>
<td>9</td>
<td>Not yet determined Maximum likelihood phylogeny estimation for interpreting relationships between sets of data</td>
<td><a href="http://www.exelixis-lab.org/">http://www.exelixis-lab.org/</a></td>
<td>No</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>samtools</td>
<td>0.1.17</td>
<td>No</td>
<td>Yes</td>
<td>9</td>
<td>Not yet determined Provides tools and methods for manipulating sequence alignments for assembly quality assessment, variant calling, and downstream processing.</td>
<td><a href="http://sourceforge.net/projects/samtools/files/samtools/">http://sourceforge.net/projects/samtools/files/samtools/</a></td>
<td>No</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>shore</td>
<td>0.6.1beta</td>
<td>Yes</td>
<td>Yes</td>
<td>9</td>
<td>Not yet determined Pipeline for mapping short reads to reference genome for finding polymorphisms, variant calling, and quantitative analysis</td>
<td>GNU GPL v3</td>
<td>No</td>
<td><a href="http://sourceforge.net/projects/shore/files/Release">http://sourceforge.net/projects/shore/files/Release</a> 0.6/</td>
<td></td>
<td></td>
</tr>
<tr>
<td>smrt</td>
<td>1.3.1</td>
<td>No</td>
<td>Yes</td>
<td>9</td>
<td>Not yet determined Analysis software specifically designed to support PacBio sequence: barcode handling and the HGAP de novo assembler are included</td>
<td>GNU GPL v3, v3, GNU LGPL, MIT, Apache</td>
<td>No</td>
<td><a href="http://pacificbiosciences.github.com/DevNet/">http://pacificbiosciences.github.com/DevNet/</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>soapdenovo</td>
<td>1.04</td>
<td>No</td>
<td>Yes</td>
<td>9</td>
<td>Not yet determined De novo assembly of short reads for large genomes, creating reference genomes of novel organisms</td>
<td>GNU GPL v3</td>
<td>No</td>
<td>Version 1 not available, see: <a href="http://sourceforge.net/projects/soapdenovo/files/SOAPDenovo2/">http://sourceforge.net/projects/soapdenovo/files/SOAPDenovo2/</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tophat</td>
<td>1.3.3</td>
<td>No</td>
<td>No</td>
<td>9</td>
<td>Not yet determined Alignment for RNA-Seq data against reference for finding splice junctions</td>
<td>Boost License</td>
<td>No</td>
<td><a href="http://tophat.cbcb.umd.edu/downloads/">http://tophat.cbcb.umd.edu/downloads/</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>transabyss</td>
<td>1.3.2</td>
<td>No</td>
<td>Yes</td>
<td>9</td>
<td>Not yet determined Analysis for multiple transcript Abyss assemblies to find splice sites and variants</td>
<td>BCCA Academic License</td>
<td>No</td>
<td><a href="http://www.bcgsc.ca/platform/bioinfo/software/trans-abyss/releases/1.3.2">http://www.bcgsc.ca/platform/bioinfo/software/trans-abyss/releases/1.3.2</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------</td>
<td>---------</td>
<td>-------------------------</td>
<td>----------------------------------</td>
<td>-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>abyss</td>
<td>1.2.7</td>
<td>Requires C++ Boost, sparsehash and Open MPI of Short Reads</td>
<td>Not yet determined</td>
<td>Input: Single or Paired-End Read files in several supported formats; Output: Assembled contigs in Fasta format.</td>
<td>488 Citations of paper doi:10.1101/gr.089532.108.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.2.7-openmp</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.3.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.3.3-openmp</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>allpathslg</td>
<td>38445</td>
<td>Gcc, GMP, Picard, and graphviz</td>
<td>Not yet determined</td>
<td>Input: 100bp Illumina reads from short and long inserts; Output: Assembled contigs graph format</td>
<td>134 Citations of paper doi:10.1073/pnas.1017351108.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>41292</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>amos</td>
<td>3.0.0</td>
<td>Gnu autoconf; subpackages require MUMmer, Boost and QT library</td>
<td>Open source development</td>
<td>Input: AMOS bank; Output: AMOS bank</td>
<td>10 Citations of paper doi:10.1002/0471250953.bi1108s33.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>arachne</td>
<td>3.2</td>
<td>LaTeX, gzip, Xerces-C++ XML Parser</td>
<td>Not yet determined</td>
<td>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.</td>
<td>405 Citations of paper doi:10.1101.gr.208902.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bedtools</td>
<td>2.12</td>
<td>Gcc</td>
<td>Git Repository for Open Source</td>
<td>Input: Sequence format such as BED, GFF, BAM; Output varies by tool but can include text or BED file</td>
<td>281 Citations of paper doi:10.1093/bioinformatics/btg033.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>blat</td>
<td>35</td>
<td>none</td>
<td>Not yet determined</td>
<td>Input: Sequence query and database in fasta, .nib or .2bit format; Output: Alignment in .psl format</td>
<td>3258 Citations of paper doi:10.1101/gr.229202.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bowtie</td>
<td>0.12.7</td>
<td>none</td>
<td>Open source development</td>
<td>Input: set of reads (Fasta, fastq, paired or unpaired, raw, tabular) and an index; Output: list of alignments</td>
<td>2090 Citations of paper doi:10.1186/gb-2009-10-3-r25.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.0.0_b6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bwa</td>
<td>0.5.9</td>
<td></td>
<td>Not yet determined</td>
<td>Input: Query in fastq format, database in fasta format; Output: Alignments in .sai format</td>
<td>1583 Citations of paper doi:10.1093/bioinformatics/btp324.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.6.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cd-hit</td>
<td>4.5.6</td>
<td>Gcc</td>
<td>Not yet determined</td>
<td>Input: Fasta query sequence; Output: cluster file describing members of each cluster in clstr format</td>
<td>107 Citations of paper doi:10.1093/bioinformatics/btg003.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>celera</td>
<td>6.1</td>
<td>Compiler, kmer</td>
<td>Not yet determined</td>
<td>Input: FRG file containing sequence; Output: Assembled contigs in native ASM format, Fasta format. Other outputs for stats and mapping.</td>
<td>739 Citations of paper DOI:10.1126/science.287.5461.2196, which relates to Celera; 19 Citations for doi:10.1093/bioinformatics/btn074.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cufflinks</td>
<td>2.0.0_beta</td>
<td>Gcc, Boost, Samtools, Eigen libraries</td>
<td>Not yet determined</td>
<td>Input: Bam/Sam file; Output: GTF file with transcripts, FPKM tracking files for genes and transcripts</td>
<td>736 Citations for paper doi:10.1058/nbt.1621.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Bioinformatics software supported by NCGAS.

This table summarizes the bioinformatics software applications installed and supported on the Mason cluster in NCGAS PY1. In all cases, the software was developed outside of NCGAS and had a production version 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/edswg/reuse/Resources/library/Publications/2008_IGARSS-RRL-Paper.pdf).
<table>
<thead>
<tr>
<th>Software</th>
<th>Version</th>
<th>Language(s)</th>
<th>Dependencies/Platforms</th>
<th>Architecture</th>
<th>Inputs</th>
<th>Outputs</th>
<th>Citations of paper / DOI</th>
</tr>
</thead>
<tbody>
<tr>
<td>edena</td>
<td>2.1.1</td>
<td>none</td>
<td>Not yet determined</td>
<td>Input: Fasta or fastq short reads file; Output: Assembled contigs (fasta format?)</td>
<td>250 Citations of paper doi: 10.1101/gr.072033.107</td>
<td></td>
<td></td>
</tr>
<tr>
<td>galaxy</td>
<td>1</td>
<td>Python, supported tools</td>
<td>Not yet determined</td>
<td>Input and output depend on the tool being used. Web interface or API available.</td>
<td>423 Citations of paper doi:10.1186/gb-2010-11-8-r86</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gatk</td>
<td>1.1-33</td>
<td>Java, R</td>
<td>Not yet determined</td>
<td>Inputs: Fasta, SAM/BAM, ROD, or interval files as per tool; Output: SAM and VCF files</td>
<td>468 Citations of paper doi:10.1101.gr.107524.110</td>
<td></td>
<td></td>
</tr>
<tr>
<td>genomemapper</td>
<td>0.4.3</td>
<td>Gcc</td>
<td>Not yet determined</td>
<td>Inputs: Reference genome in Fasta format, Shore files, Fasta or Fastq data queries; Output: Shore or Bed file with alignments</td>
<td>41 Citations of paper doi:10.1101/gb-2009-10-9-r98</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gmap</td>
<td>7/20/12</td>
<td>Gcc, Perl</td>
<td>Not yet determined</td>
<td>Input: Reference genome in Fasta format, Query in Fasta format; Outputs: Alignment file either compressed or uncompressed</td>
<td>273 Citations of paper doi:10.1093/bioinformatics/btf310</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mummer</td>
<td>3.22</td>
<td>gcc, perl, g++, fig2dev, gnuplot, sfigsed, awk, ar, sh, csh</td>
<td>Open source development</td>
<td>Input: Reference genome in Fasta format, Query in Fasta format; Output: Text output</td>
<td>870 Citations of paper doi:10.1186/gb-2004-5-2-r12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ninja</td>
<td>1.2.1</td>
<td>Java</td>
<td>Not yet determined</td>
<td>Input: Sequence alignment in Fasta format; Output: Phylogenetic tree in Newick or Phylip format</td>
<td>17 Citations of paper DOI: 10.1007/978-3-642-04241-6_31</td>
<td></td>
<td></td>
</tr>
<tr>
<td>novoalign</td>
<td>2.07.13</td>
<td>Bedtools, Samtools, Picard, GATK</td>
<td>Not yet determined</td>
<td>Input: Read files in Fastq or compressed format, Reference genome index created with novoindex; Output: Sam or tabular alignments</td>
<td>No Published paper found for Novoalign; 447 hits for Novoalign by Google Scholar</td>
<td></td>
<td></td>
</tr>
<tr>
<td>picard</td>
<td>1.52</td>
<td>Java</td>
<td>Not yet determined</td>
<td>Input: Sam/Bam or URL, depending on tool; Output: Bam, Bam index, or text file with metrics</td>
<td>No published paper for Picard</td>
<td></td>
<td></td>
</tr>
<tr>
<td>raxml</td>
<td>7.2.6</td>
<td>Gcc</td>
<td>Not yet determined</td>
<td>Input: Phylip file containing alignments to be run; Output: Text files containing tree topologies, logfiles, intermediate files</td>
<td>3089 Citations for paper doi:10.1093/bioinformatics/btl446</td>
<td></td>
<td></td>
</tr>
<tr>
<td>samtools</td>
<td>0.1.17</td>
<td>Gcc</td>
<td>Not yet determined</td>
<td>Input: Sam or Bam file; Output: Varies by tool, Sam, Bam, .vcf, .afs files</td>
<td>1526 Citations of paper doi:10.1093/bioinformatics/bp352.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>shore</td>
<td>0.6.1beta</td>
<td>Gcc, Boost, alignment software (genomemapper, lwa, bowtie), R</td>
<td>Not yet determined</td>
<td>Input: Reference genome in fasta format, reads files in raw format; Output: Statistics, analysis results and quality assessments in a variety of formats</td>
<td>203 Citations of paper doi:10.1101/gr.080200.108</td>
<td></td>
<td></td>
</tr>
<tr>
<td>smrt</td>
<td>1.3.1</td>
<td>Mysql, perl, bash, Java</td>
<td>Not yet determined</td>
<td>Input: Analysis is a GUI with multiple tools; sequence information stored in XML files; Output: XML and HTML results depending on tool</td>
<td>No published paper for SMRT Analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>soapdenovo</td>
<td>1.04</td>
<td>none</td>
<td>Not yet determined</td>
<td>Input: Read files in Fasta, Fastq, and Bam, config file; Output: Contig assemblies and Scaffold assemblies</td>
<td>2 Citations of paper doi:10.1186/2047-217X-1-18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tophat</td>
<td>1.3.3</td>
<td>Gcc, Boost, Samtools</td>
<td>Not yet determined</td>
<td>Input: Reads in Fasta or Fastq format, Bowtie index database; Output: SAM alignment results, BED files with indel and junction results</td>
<td>854 Citations of paper doi:10.1093/bioinformatics/bp120</td>
<td></td>
<td></td>
</tr>
<tr>
<td>transabyss</td>
<td>1.3.2</td>
<td>BWA, Bowtie, Pysam, Samtools, Abyss, Blat, GMAP, Python, Perl, Anchor, xa2multi.pl</td>
<td>Not yet determined</td>
<td>Input: Input file specifying the assemblies to use, Reference genome file and gene annotations; Output: Bam and Bam index files for consensus assembly</td>
<td>130 Citations of paper doi:10.1038/nmeth.1517</td>
<td></td>
<td></td>
</tr>
<tr>
<td>trinityrnaseq</td>
<td>8/20/11</td>
<td>Gcc</td>
<td>Open source development</td>
<td>Input: Fastq files containing reads; Output: Assembled contigs in Fasta format</td>
<td>202 Citations of paper doi: 10.1038/nbt.1883</td>
<td></td>
<td></td>
</tr>
<tr>
<td>velvet</td>
<td>1.1.04</td>
<td>Gcc</td>
<td>Not yet determined</td>
<td>Input: Fasta, fastq, sam, bam, eland, gerald; Output: Assembled contigs in Fasta format, stats file, AMOS file, and graph file.</td>
<td>1461 Citations of paper doi: 10.1101/gr.074492.107</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Technical descriptions of software supported by NCGAS. This table provides reference information about the software installed on Mason in PY1.
<table>
<thead>
<tr>
<th>Software</th>
<th>Gordon (SDSC)</th>
<th>Stampede (TACC)</th>
<th>Lonestar (TACC)</th>
<th>Rockhopper (IU)</th>
</tr>
</thead>
<tbody>
<tr>
<td>abyss</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>allpathslg</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>amos</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>arachne</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bedtools</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>blat</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>bowtie</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>bwa</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>cd-hit</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>celera</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>cufflinks</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>edena</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>galaxy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gatk</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>genomemapper</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gmap</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mummer</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>ninja</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>novoalign</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>picard</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>raxml</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>samtools</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>shore</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>smrt</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>soapdenovo</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>tophat</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>transabyss</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>trinityrnaseq</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>velvet</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>

Table 4. NCGAS software support across partners. This table shows the level of shared support for bioinformatics software on the various NCGAS partner compute resources in PY1.

TACC graduate student Manoj Dhanpal has led a yearlong effort to explore the suitability of over a dozen community bioinformatics codes for acceleration on the traditional multicore and the new many-core Xeon Phi chips. He has compiled several benchmarks that will guide our acceleration efforts for these applications. Along the way he has reported or solved substantial issues with the Intel development toolkit, including lack of SSE intrinsics which are widely used in bioinformatics programs.

4.2. Technical system architecture

Indiana University provided, as a resource contribution to NCGAS, the 16-node Mason large memory cluster computer. Each of Mason’s nodes, outfitted with a 32-core processor and 512 GB of RAM connected to the 1 petabyte (PB) Data Capacitor storage system, is architected particularly for memory-intensive applications such as genome assembly. By installing a suite of genome analysis applications on Mason (Table 2 and Table 3) and establishing a low barrier system by which NSF-funded researchers could get access to Mason, its software applications, and bioinformatics support, NCGAS was quickly able to begin providing service to genome scientists while it built its online presence and began outreach effort to the research community. TACC provided services on their infrastructure systems, most notably The Ranger system. Ranger consisted of 3,936 16-way SMP compute nodes providing 15,744 AMD Opteron processors for a total of 62,976 compute cores, 123 terabytes (TB) of total memory, and 1.7 PB
of raw global disk space. At the writing of this report, the Ranger system has been decommissioned and replaced with the Stampede system. Additional capacity is available on the SDSC Gordon System. Gordon consists of 1024 compute nodes and 64 I/O nodes. Each compute node contains two 8-core 2.6 GHz Intel EM64T Xeon E5 (Sandy Bridge) processors and 64 GB of DDR3-1333 memory. The I/O nodes each contain two 6-core 2.67 GHz Intel X5650 (Westmere) processors, 48 GB of DDR3-1333 memory, and sixteen 300 GB Intel 710 solid-state drives. It is attached to a 4 PB file system.

### 4.3. Servers and job statistics

The status of 85 users who submitted jobs on Mason in PY1 is broken out in Figure 1. Collectively, these users submitted 28,523 separate jobs using 136.83 core years of computation. This is approximately 26.7% of the total core years available in a year on Mason. This utilization figure is necessarily low, however, because throughout PY1 a single user concurrently owned all 32 cores on a node independent of how many cores the user’s job actually requested. This configuration was required in order to allow large memory jobs to fully utilize the 504 GB of free RAM available on each node. Given the unique management challenges of large RAM clusters, it is impossible to know precisely the overall utilization of the system, but we can put bounds on it. The 26.7% utilization figure represents a minimum bound. If we assume each job requesting less than 32 cores ran on its own node, Mason would have been 107% occupied in PY1. Of course, jobs requiring less than a full node are usually stacked with other small jobs from the same user, which accounts for this figure being unrealistically high. Mason was no less than 26.7% utilized in PY1, and it could have been nearly fully utilized. Improvements are being made constantly to increase the availability of the system and to shorten queue times for submitted jobs.

![Figure 1. Distribution of the 85 users who submitted jobs to the Mason cluster in PY1.](image)

**Non IU NSF-Funded** are those people associated with NSF genomics projects that are not from Indiana University. **IU NSF-Funded** are students, staff or faculty of Indiana University who are associated with an NSF-funded genomics project. **IU Non-NSF Funded** are non-UITS staff or faculty who are not associated with NSF funding. **IU Students** are Indiana University students who are not associated with an NSF-funded genomics project or employed by UITS to support the Mason system. Lastly, **IU Staff** are primarily UITS staff associated with technical aspects of the Mason cluster.

**NSF-Funded Users:** The 41 NSF-funded users who submitted jobs on Mason in PY1 represent 23 separate NSF award numbers. A description of each of the supported NSF funded projects is given in Section 8.3.1.
**Recruitment of External NSF-Funded Projects:** Representatives of NCGAS presented posters, gave talks, and exhibited at nine national and international conferences that potential users were expected to attend. The figures below show the recruitment of users by month across PY1 by type. Notice first that the data gathering started in January of 2012, and then focus on the non-IU NSF funded recruitment. *Non-IU NSF* refers to non-Indiana University NSF-funded life science projects. *IU-NSF* refers to IU faculty initiated requests that were associated with active NSF life science related awards. Lastly, *IU-Non-NSF* represents Indiana University users taking advantage of the 25% of Mason reserved for any and all IU large RAM computational tasks. The spike in Non-IU NSF-funded projects that occurred in June 2012 is the result of NCGAS exhibiting at the First Joint Congress on Evolutionary Biology in Ottawa Canada. The steady low level of recruitment is attributed to Dr. Tom Doak attending a number of smaller highly focused meetings, and word of mouth.

![Graph showing recruitment by month](image)

**Figure 2. Number of allocation requests received from REDCap by month for PY1.** Cumulative requests are shown to the right. The peak in requests in June corresponds to NCGAS having a table at the First Joint Congress on Evolutionary Biology to disseminate information about NCGAS services.

### 4.4. Utilization breakdown

The figure below shows two views of the overall activity on the Mason Cluster in PY1. Although the specifics vary depending on whether you look at the number of jobs submitted or the wall time used, the same pattern holds. In the first few months Mason’s use was dominated by pent-up need for large RAM computation at IU, as well as a significant amount of tuning done by UITS systems administrators – who make up the largest component of the IU-non NSF users. Later in the year there are bursts of computation associated with non-IU NSF life science projects. Users submitting jobs were classified as Non-IU NSF-funded, IU NSF-Funded, IU Non-NSF funded, and IU Students.
Figure 3. Jobs submitted to the Mason Cluster per week, and wall time consumed on the Mason cluster per week over PY1.

4.5. **Key software deployment /improvement / hardening**

The NCGAS mission goes beyond simply installing software on our supported systems. An additional component is improving the utility of community-developed software and its accessibility to life scientists. In PY1 NCGAS staff initiated two noteworthy projects towards this goal.

- **Launched two Galaxy web portals for genomics analysis.** The Galaxy web portal creates a user-friendly interface to pre-defined computational resources with a selection of features that are of particular interest to life scientists. A Galaxy portal is easily used by students and life science faculty. By concealing the technical complexities of using supercomputers behind an intuitive and well-documented graphical user interface, Galaxy lowers the amount of non-biological skills researchers need in order to analyze their experimental results. NCGAS installed a Galaxy server targeting Mason in support of IU students and faculty, and a second for NSF-funded genomics projects. Efforts are underway in PY2 to expand this second instance so that users can launch jobs from this portal targeting all NCGAS-supported computational resources at all partner institutions.

- **Optimized Trinity in partnership with the Broad Institute.** The open source *de novo* RNA sequence assembly application, Trinity, is a best-in-class tool that had required excessive computational resources. By partnering with the Broad Institute, an independent non-profit research institute originally founded as a partnership between Harvard University and the Massachusetts Institute of Technology, NCGAS has contributed to improving the computational performance of Trinity such that it is now the fastest *de novo* assembler for RNA sequencing data. This effort is detailed in Section 8.1 of this report.

Additionally, TACC undergraduate Eric Dawson has contributed to testing and use case development of the TACC Agave API, which will play a significant role in providing powerful computational resources to life sciences researchers performing genomic analyses.

4.6. **User contacts**

The interaction between NCGAS staff and the user community is classified into three categories: short-term consultations, long-term 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. Long-term consultations require more than four hours of effort and can be either technical or scientific. The former usually revolve around complex technical issues that exceed the reasonable understanding of a domain scientist. These include requests to install software packages with complex dependencies or to troubleshoot error messages received from failed jobs. It is not uncommon
for these consultations to require NCGAS staff to interact with the user, the systems administrators, and 
the software community that developed the failing software. Scientific long-term consultations are 
primarily the domain of Drs. Doak and LeDuc. As a genomic scientist and biometrician respectively, they 
can advise domain scientists on appropriate methodologies for processing and analyzing data. Lastly there 
are supported programs. This category is used to describe separate research initiatives or individual grant 
recipients who have asked for allocations on NCGAS resources. In practice, in PY1 supported projects 
have required one of two levels of support. The first are those projects which had staff knowledgeable in 
the use of high performance computer clusters, and that simply needed access to the resources, and some 
modest amount of short-term consultations to analyze their data. The second class of projects are those 
that had domain scientists and clearly defined research goals that required next-generation sequencing 
data analysis, but that lacked both the computational resources and the skills to perform the analysis – and 
frequently they were unclear on the best course of analysis given the sequence data to which they already 
had access.

In PY1 NCGAS staff reported 480 short-term consultations, 22 long-term consultations, 25 NSF-funded 
supported programs, and 14 non-NSF-funded supported programs. Effort-tracking systems have been 
implemented in PY2 to allow more detailed analysis of the time spent on the various categories of 
support, and to provide statistical analysis of NCGAS staff effort.

At TACC, support staff are alerted to the presence of NCGAS as a resource for genomics researchers and 
will be offered the opportunity to copy in NCGAS consultants on TACC support tickets.

4.7. **Short-term consultations**

Overall NCGAS staff reported responding to 480 short-term tickets. The majority of these requests were 
completed the same day they arrived, but those requests requiring the coordinated action of NCGAS and 
IU Research Technologies staff, such as technical questions regarding the installation of software or 
determining the cause of job failure on the Mason cluster, were sent to the RT tracking system. Of the 89 
short-term consultations completed in PY1 using the RT tracking system, 25% were completed in less 
than 2.1 days, and half in less than 10.0 days. A small number of long-standing issues pulled the mean 
completion time up to 33.0 days.

4.8. **Long-term consultations**

NCGAS enters into long-term consulting projects with NSF-funded researchers who need assistance with 
analysis of next-generation sequence data, using the Mason cluster, or both. A long-term consultation is 
any contact taking more than four hours of staff time to complete. NCGAS staff reported 22 long-term 
consultations in PY1.

4.9. **Project support**

NCGAS provides support to projects with NSF-funded researchers who need assistance with analysis of 
next-generation sequence data. Frequently NCGAS support consists of providing access to the Mason 
cluster, providing assistance with a series of short-term tickets, and then standing back and allowing the 
researchers to proceed. In several cases the researchers request much greater support including one or 
more long-term consultations to provide bioinformatics guidance as well as NCGAS staff analyzing data 
for projects lacking committed bioinformaticians. There are 25 NSF-funded projects supported by 
NCGAS and another 14 non-NSF funded. Figure 4 provides a map of the United States showing the 
distribution of these projects. In PY1 half the total projects came from Indiana, while the remainder came 
from 12 different states.
Figure 4. Map showing the location of NCGAS supported projects.

5. Training and development activities

Education, training, and outreach (EOT) are an important component of the NCGAS mission. In PY1 NCGAS staff gave 10 EOT presentations which tended to focus on increasing awareness of the national cyberinfrastructure and how it can benefit life-scientists. These presentations are in addition to our scientific efforts detailed in section 6.

<table>
<thead>
<tr>
<th>Type</th>
<th>Title</th>
<th>Location</th>
<th>Date</th>
<th>Hours</th>
<th>Number of Participants</th>
<th>Number of individuals from traditionally underserved groups (TUGs)*</th>
<th>Method†</th>
<th>Funding Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Presentation on NCGAS Services</td>
<td>National Center for Genome Analysis Support</td>
<td>Plant and Animal Genome Conference</td>
<td>16 Jan ’12</td>
<td>2</td>
<td>20</td>
<td>9</td>
<td>S</td>
<td>NCGAS</td>
</tr>
<tr>
<td>Presentation on NCGAS Services</td>
<td>National Center for Genome Analysis Support</td>
<td>Pacific Symposium on Biocomputing</td>
<td>5 Jan ’12</td>
<td>1</td>
<td>10</td>
<td>3</td>
<td>S</td>
<td>NCGAS</td>
</tr>
<tr>
<td>Presentation on cyberinfrastructure resources</td>
<td>Cyberinfrastructure Begins at Home</td>
<td>Rutgers University New Brunswick NJ</td>
<td>20 Feb ’12</td>
<td>2</td>
<td>22</td>
<td>4</td>
<td>S</td>
<td>NCGAS</td>
</tr>
<tr>
<td>Presentation on HPC and Cloud Services</td>
<td>POD/IU Partnership “Cluster as a service” and Cloud Services</td>
<td>Coalition for Academic Scientific Computation</td>
<td>29 Feb ’12</td>
<td>1</td>
<td>18</td>
<td>3</td>
<td>S</td>
<td>NCGAS</td>
</tr>
<tr>
<td>Training for IU Biology Faculty on NCGAS services</td>
<td>Bioinformatics Support for Experimental Biologists</td>
<td>Jordan Hall, Indiana University</td>
<td>26 Mar ’12</td>
<td>1</td>
<td>30</td>
<td>15</td>
<td>S</td>
<td>NCGAS</td>
</tr>
</tbody>
</table>
Presentation on NCGAS Services
National Center for Genome Analysis Support
Midwest Protozoology Society Meeting
21 Apr, ‘12
1
20
10 S NCGAS

Presentation on NCGAS national infrastructure
National Infrastructure for High Speed Data
Bio-IT World
25 Apr ‘12
1
100
0 S NCGAS

Presentation on technology services for research
Data Management? I’m a Biologist
Research Computing Day, U. of Florida
25 Apr ‘12
1
50
20 S NCGAS

Training on RNA sequencing software
Designing RNA-Seq Experiments
Daphnia Genomics Jamboree, Indiana University
22 May ‘12
3
22
14 S NCGAS

Presentation on NCGAS national infrastructure
HPC Data Management on National and International Scale
Bio-IT World Asia, Singapore
7 Jun ‘12
2
40
0 S NCGAS

Totals
10 Events
15
332
78

Table 5. EOT activities for PY1 for NCGAS.
*Traditionally underrepresented groups as defined by the NSF.
† All events were conducted (S)ynchronously, e.g., in front of a live audience.

6. Outcomes and work products
NCGAS had a productive first year. In PY1 NCGAS staff presented 21 posters of scientific presentations, and co-authored on peer-reviewed paper.

6.1. Publications
Because NCGAS was in its first year, and the typical period for biology research to generate publications is approximately 18 – 24 months, there are not yet any scientific publications resulting from NCGAS assistance with data analysis. However, NCGAS staff made a number of presentations and publications to both scientific and cyberinfrastructure audiences that helped communicate its accomplishments and strategies.

6.1.1. Journal papers
N/A

6.1.2. Conference papers

6.1.3. Conference proceedings
N/A

6.1.4. Journal articles
N/A
6.1.5. Poster presentations


6.1.6. Presentations


6.2. Web pages
The main NCGAS web page at: http://ncgas.org was developed during this period. It has nine informational and service sections:

- About NCGAS: http://ncgas.org/about.php
- Services: http://ncgas.org/services.php
- News, Events & Training: http://ncgas.org/events.php
- Request Allocation: http://ncgas.org/request.php
- Software: http://ncgas.org/software.php
- Partners: http://ncgas.org/partners.php
- Staff: http://ncgas.org/who.php
- Contact Us: http://ncgas.org/contact.php

7. Management
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 NCGAS 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. In providing NCGAS as a service and Mason as a cyberinfrastructure resource, NCGAS is supported by the following staff: Dr. Tom Doak (scientist), Dr. Le-Shin Wu (computer scientist), Rich Knepper (software management), Dr. Robert Henschel (software optimization), Dr. Scott Michael (software optimization and management), Bret Hammond (systems Administration), Matthew Allen (systems administration), and Carrie Ganote (bioinformatics support). Service resources at TACC are provided by Matt Vaughn, their lead bioinformatician, and by technical support services funded by the NCGAS grant. Additionally, an internal advisory board comprised of NCGAS co-PIs at Indiana University (Drs. Michael Lynch, Matthew Hahn, and Geoffrey Fox) provides guidance for NCGAS operations.
7.1. **Summary of needs identified with discussion with SAB chair, co-PIs, and key constituents and partners**

The critical and overarching needs identified in discussions with SAB chair (Dr. Richard McCombie of Cold Spring Harbor Labs), co-PIs, and key constituents and partners for NCGAS activities going forward are as follows:

- Establish a higher profile nationally through web presence, more displays / posters at national conferences, and through electronic/social media
- Establish clearer, stronger, and more visible partnerships with major national cyberinfrastructure resource providers such as XSEDE and the supercomputer centers that are part of XSEDE
- Continue to focus on long-term support of important scientific research – recognizing that benefits will become visible in PYs 2 and 3 – but find more and better ways to report incremental activities and assistance to the scientific community
- Select a set of conferences at which NCGAS becomes a regular and well-recognized participant
- Develop our national services so that next year’s map of users makes the main location of NCGAS much less obvious

7.2. **Development of a sustainability plan**

NCGAS is pursuing a multi-pronged strategy toward long-term sustainability, which includes the following key components:

- The National Science Foundation has recently changed the Advances in Biological Informatics program, which initially funded NCGAS. This change created a new category of grant awards, called sustaining awards, in which the NSF pays half of the total operating cost of a biological service center and the hosting university pays the other half. It seems feasible to imagine partial funding from IU and partial funding from the NSF, through a second "sustaining" award as an element of the long-term sustainability of NCGAS.
- Researchers funded by the National Institutes of Health have expressed interest in using NCGAS services. The NSF award that created NCGAS specified that NCGAS is intended to serve NSF-funded researchers. NIH-funded researchers have a tradition of being willing (and able) to pay for supporting services such as bioinformatics. As a second component of our multi-pronged sustainability strategy, we are investigating offering services to NIH-funded researchers on a "cost-recovery plus" basis and searching for NIH funding to serve NIH researchers to complement NSF funding supporting NSF researchers. Partnerships with organizations such as the Broad Institute are a key element of executing this strategy.
- NCGAS is partnering with other national scale centers to increase its national exposure. We have formed a partnership with the Cyberinfrastructure for Phylogenetic Research (CIPRES) at the San Diego Supercomputing Center. CIPRES provides computational support for phylogenetic research by providing vetted software solutions for calculating evolutionary distances and building phylogenetic trees. By simplifying data transfer between NCGAS and CIPRES, comparative genomics research can be conducted using specialized resources for genome assembly at NCGAS followed by phylogenetics at CIPRES.

8. **Science outcomes**

With a focus on long-term support for biological research and bioinformatics projects, PY1 was focused more on starting projects than finishing them. Here we report four science highlights, and one education, outreach, and training highlight. In section 8.3, we list the projects and scientific research we are supporting during PY1.
8.1. Science highlights

Project title: Optimization of Trinity RNA sequence analysis toolkit
PI or project lead: Robert Henschel, Indiana University
Funding agency: NSF
Award number(s): MCB 1062432

Outcome: NCGAS and IU Research Technologies’ Scientific Applications and Performance Tuning (SciAPT) group developed a first major deliverable in our efforts to support the open source bioinformatics software community. At XSEDE12 Robert Henschel presented a paper detailing our work to improve the computational performance of the Trinity application. Trinity is a de novo sequence assembly tool optimized for RNA-sequence analysis – arguably the most complex next-generation sequencing (NGS) computation currently undertaken in the life sciences. RNA-sequencing accounts for the majority of all next-generation sequence data collected worldwide. With over 60% of all NGS studies being RNA-sequence analysis, Trinity is the RNA-sequencing application of choice for researchers at NCGAS.

Impact / benefits: By greatly improving the performance of best-in-class bioinformatics software, less computer infrastructure is needed to meet the growing demand. Here software that is critical for a large and growing number of life scientists and which requires large-memory clusters was optimized such that four times the number of samples can use analyzed with the existing computational resources.

Explanation: Trinity has very specific computation that is needed to gain biological insights from the sequences. It was originally developed at the Broad Institute – the premier genomic research and gene sequencing organization in the US. Trinity was recognized as giving the most biologically correct assemblies – but unfortunately, at that time it took over an order of magnitude more time than competing applications to compute these superior results. While the open source community worked to generally improve the performance of Trinity, NCGAS partnered with the Broad Institute to optimize the application to running on high performance supercomputers and clusters. By the time we were done, Trinity ran faster on Mason than competing applications that gave biologically inferior results. Trinity was over four times faster than when we started. All the improvements NCGAS made have been returned to the open source community, and are currently known to be used by at least two other supercomputing centers. As a result of this project, fewer researchers are being forced to use inferior RNA-sequencing assemblies simply because they were unable to access sufficient computational resources to use Trinity. This will result in better and more rapid genomic science across the globe.

This project increased the performance of the most popular genomics analysis tool in current practice and in turn it allows IU, and every other supercomputing center, to accomplish more genome science with the same computational resources. Realistically, shortening the runtime of an individual RNA-sequencing assembly by a few hours does little to shorten the time to completion of the overall studies, which often take months to years from conception to publication. Instead, the reduced runtime reduces the total number of large RAM systems that need to be in place nationally to meet the entire aggregate national need. The broader impacts of this project were that it established the NCGAS and SciAPT as the national authority for optimizing Trinity, and for high performance genomics code optimization in general.

Figure 5 shows the importance of software enhancements created by NCGAS running on IU’s Mason computing cluster and the Pittsburgh Computing Center’s Blacklight cluster. Comparing the optimized Mason (blue squares) to the Original Mason (red circles) shows the nearly fourfold improvement in runtime that was achieved by NCGAS and RT work on this widely used software. Comparing the Original Mason (red circles) to Optimized Blacklight (black diamonds) shows that our performance improvements generalize across very different types of supercomputer clusters.
Figure 5. A graph showing the improved computational performance of Trinity running on biological data sets of increasing size.
Project title: Evolutionary genetics of fruit flies in Hawaii
PI or project lead: Donald Price, University of Hawaii-Hilo
Funding agency: NSF
Award number(s): 0833211

Outcome: NCGAS annotated 14,000 microarray probes developed for a different *Drosophila* species.

Impact / benefits: By re-annotating the array, it allowed the commercially available microarray to be used to study the role of elevation in causing changes in fly morphology of Hawaiian fruit flies.

Explanation: Professor Donald Price of University of Hawaii-Hilo is attempting to understand the role of environment on gene expression in fruit fly species of the Hawaiian Islands. *Drosophila* have evolved on the Hawaiian islands to yield hundreds of distinct species with different behaviors, morphologies, and developmental programs, and Dr. Price is using these species for comparative developmental studies. Starting with EST data from Hawaiian species Dr. Price produced microarrays representing 14,000 genes, but these genes were unannotated. NCGAS bioinformaticians were able to annotate the functions of these genes by aligning them to sequences of known function in reference species. With this annotation, Dr. Price can use them for systems-level studies comparing the differential expression of genes across environmental conditions. Of particular interest to Dr. Price is the role of elevation in causing changes in fly morphology.

Figure 6. The *Drosophila* species of Hawaii can be found in a variety of habitats ranging from upland deserts to this tropical rain forest.
Project title: RNA sequencing of marine copepods
PI or project lead: Petra H. Lenz at Pacific Biosciences Research Center, University of Hawaii
Funding agency: NSF
Award number(s): OCE 1040597

Outcome: NCGAS co-assembled 92 GB of RNA-sequence data to create a reference assembly.

Impact / benefits: Developing a reference assembly for marine copepods will allow researchers to conduct differential gene expression studies in these organisms. This will give insights on how environmental perturbations influence survival strategies of animal populations.

Explanation: Professor Petra H. Lenz at Pacific Biosciences Research Center, University of Hawaii has an ongoing project on the development and neuroecology of zooplankton sensory systems, especially in marine copepods. Since the genomes are too large to practically sequence, an RNAseq approach has been taken. First, the RNA seq reads from a number of developmental stages are co-assembled with Trinity, generating 92 GB of output data. This reference assembly can then be used – in place of a genomic reference – to quantify the reads from the individual stages in order to give differential expression of genes in different developmental stages. Using this approach, Professor Lenz has started specific studies of the interplay of environment and genetics on the jumping behavior of copepods. This will give insights on how environmental perturbations influence survival strategies of animal populations.

![Contig Length in Assembly](image)

Figure 7. RNA-sequence analyses pose unique assembly issues that required specialized software. Using Trinity installed on the Mason cluster NCGAS was able to create a list of candidate gene models in copepods. Given the prior knowledge of copepod neuropeptides, the quality of the assembled gene models can be independently verified.
Figure 8. Vertical plankton tow in the Gulf of Maine to collect Calanus finmarchicus. Slow tows are performed to collect intact individuals for culturing and laboratory experiments.
Project title: Support of Sequencing Projects at Marshall University
PI or project lead: Wendy Trzyna, Gary Shultz
Funding agency: NSF
Award number(s): EPS1003907

Outcome: NCGAS assisted two faculty from Marshall University in West Virginia. We assisted in the initial quality assessment and reference assembly of the ameba Acanthamoeba, and assisted in the initial metagenomic assembly of Ohio River water samples.

Impact / benefits: In providing support to researchers at EPSCoR-supported institutions, NCGAS demonstrates how national-scale resources can benefit institutions without their own internal supercomputing resources.

Explanation:
Amoeba RNA sequencing – studying salt tolerance and pathogenicity
Professor Wendy Trzyna of Marshall University studies the ubiquitous amoeba Acanthamoeba. While generally a benign soil inhabitant, Acanthamoeba is also an opportunistic pathogen. Dr. Trzyna is interested in the role of salt concentration on gene expression in different Acanthamoeba isolates – one determinant of their distribution. We are currently working with Dr. Trzyna to assemble both genomic and RNAseq data (30 GB of sequence data) for three different strains of Acanthamoeba with different salt tolerances, in preparation for studies in the differential expression of genes in this organism as well as for comparative genomic studies.

Microbial community diversity
Professor Gary Shultz of Marshall University is attempting to determine the diversity, stability, and resilience of the microbial community and its genetic potential across spatial and temporal scales of the ecosystems of the Ohio River. This will allow researchers to understand the relationships between microbial diversity, the environment they inhabit, and the metabolic processes they perform. NCGAS resources are helpful because the PI has little access to bioinformatics expertise at his home institution. NCGAS personal have partnered with Dr. Shultz and IU researchers who specialize in metagenomic analysis to initiate analysis of his initial data sets, comparing samples taken for the Ohio River during summer and winter months.

Figure 9. Quality control of DNA sequence data. An important first step in any sequencing analysis is determining the overall quality of the sequence reads. NCGAS software allows researchers to test the quality of their practice runs before consuming valuable samples.
8.2. **Education, outreach, and training highlight**

We have one significant education, outreach, and training (EOT) highlight to report in PY1, as follows:

- **Project title:** Daphnia Genomics Jamboree,  
- **PI or project lead:** John Colbourne, Ph.D., Center for Genomics and Bioinformatics, Indiana University  
- **Funding agency:** Indiana University Center for Genomics and Bioinformatics and NSF  
- **Award number(s):** NSF 1062432

**Outcome:** *Daphnia manga* is an important freshwater indicator species. This meeting allowed scientists to work together to prepare an initial draft of the *D. manga* genome which is needed before the environmental researchers can start taking advantage of powerful next-generation sequencing methodologies in their field work.

**Impact / benefits:** *Daphnia manga* has long been used as an indicator species for determining the health of bodies of freshwater. The draft genome prepared at this meeting is an important first step to allow field biologists to use powerful tools such as RNA-sequencing to measure the genetic changes associated with environmental perturbations in natural *Daphnia* populations.

**Explanation:** NCGAS co-sponsored the International *Daphnia* Genome Analysis Jamboree held in the Cyberinfrastructure Building in Bloomington for one week in May 2012. During this time 22 researchers from North America and Europe worked to accelerate the genomic annotation of the important ecological indicator organism *Daphnia manga*.

![Figure 10. Attendees at the Daphnia Genomics Jamboree use the IU IQ-Wall to collaborate on a research paper.](image-url)
Support projects

The primary goal of NCGAS is to provide computational support to existing genomics and life science projects. In this section we briefly outline each project receiving some form of support in PY1. Section 9.3.1 lists the 25 NSF funded projects that were supported in PY1 in chronological order of their request for support. This is followed by a listing of the remaining projects supported in PY1.

8.3.1. NSF-funded genomics

We received 25 requests for genomics support during Project Year 1. These requests are listed below in chronological order.

Date: 2/3/12 13:09  
Institution: Boyce Thompson Institute for Plant Research  
Grant: 820612  
Title: Petunia genome project  
Our NSF funded project is for the tomato genome sequence. We are collaborating with an independent Petunia sequencing project, with the goal to integrate the results in a comparative genomics approach on the SGN website (http://solgenomics.net) with the tomato data.

Date: 2/4/12 15:55  
Institution: Northern Illinois University  
Grant: 820612  
Title: Tomato Chromosome 1 and 10 Sequencing, Coordination and Bioinformatics for the International Solanaceae Genome Initiative  
The Sol Genomics Network & The Petunia Platform Group are currently attempting to assemble genomic sequences for Petunia inflata and Petunia axillaris, the progenitor species to the commercial garden petunia. 2X-PE 100 nt illumina reads (65-70X coverage) have been produced for both genomes, along with RNA-Seq data from 6 tissue samples of P. inflata. Current assembly efforts have used SOAPdenovo with different k-mer values. We wish to use SOAPdenovo, ABySS and/or ALLPATHS software to increase the quality of the assembly. Current computer resources at Cornell and Northern Illinois University are limited for running multiple assemblies, so we are seeking NCGAS resources to complete that task. Further 3 kb and 8 kb PE GS FLX+ sequences may be added to the current datasets to improve assembly. Once assembled our intention is to carry out genome annotation & comparison using the MAKER program running on Cornell and NIU servers. This is an international project involving molecular biologists and geneticists from 16 Institutions in 10 countries, including 4 U.S. universities (NIU, Cornell, U Florida, Michigan State). Sequence assembly and public data release will be coordinated through the Sol Genomics Network, a NSF-supported center at the Boyce Thompson Institute for Plant Research at Cornell University.

Date: 2/12/12 15:51  
Institution: Princeton University  
Grant: 900544  
Title: Assembly and analysis of the scrambled germline genome of Oxytricha trifallax.  
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.

Date: 3/5/12 12:51  
Institution: Cornell University  
Grant: 703908  
Title: Gramene: A Platform for Comparative Plant Genomics

Gramene is a curated data resource for comparative genome analysis in grasses and other plants. The database integrates information about genomic sequence, genes, proteins, biochemical pathways, maps and markers, QTL, germplasm, and genetic and phenotypic diversity. Currently we are developing an enrichment algorithm, and we need to do analysis on large-scale diversity data sets.

Date: 3/13/12 10:49  
Institution: Indiana University  
Grant: NSF EF-0328516-A006; NSF EF-0827411  
Title: Paramnecium and Daphnia genomes

The lab of Michael Lynch studies a variety of aspects of genome architecture and evolution. While some of this work deals with specific gene families, much of it requires analysis of entire genomes. In addition to analyzing published genomes, we are actively generating genomic sequences of our own. The two largest projects are now population genomics of Daphnia, and an NSF-funded study comparing the genomic structures of a number of Paramecium species. For this project, we are sequencing the genomes of up to 20 paramecium species, using mixed next generation sequencing methods, and generating paired RNAseq data. The genomes need to be de novo assembled, and the RNAseq data mapped to the finished genomes. The immediate goal of the Paramecium project is to study the evolutionary fates of duplicate genes after whole-genome duplication. RNAseq for each genome is currently being used to aid in genome annotations and will later be used to detect expression differences between paralogs.

Date: 3/13/12 23:44  
Institution: Indiana University  
Grant: NSF 1027529  
Title: The Geraniaceae genomes project: Accelerated and coordinated evolution across the three plant genomes

Plant cells contain genomes in three distinct compartments, the mitochondrion, nucleus and plastid. Over time, thousands of genes transferred among these genomes and now there is extensive communication among the compartments and considerable conservation and stability of the genomes. The plant family Geraniaceae represents an important exception to this pattern because its mitochondrial and plastid genomes have experienced remarkably accelerated rates of change in gene content, gene order, and rates of nucleotide substitutions. The cause of these accelerated rates is unknown, but may be directed by genes encoded in the nucleus. This project investigates the basis for this accelerated evolution with the aim of understanding how different genomes within a cell can influence one another and co-evolve over time. The project will sequence, from 30 members of the Geraniaceae, the DNA in mitochondrial and plastid genomes and the genes expressed in the nucleus. These data will be analyzed to elucidate the mechanisms of inter-compartmental crosstalk and co-evolution in plant cells. The goals of this project are to determine the extent of genomic upheaval in the mitochondrial and plastid genomes of the Geraniaceae and to identify the correlated changes in the nuclear genome that have driven this instability. The large and complex data sets generated in this project - 60 organelle genomes (some of them many Mb in size and full of repetitive DNA) and complete nuclear transcriptomes from 30 plants - require powerful, high-
speed computing systems as available through NCGAS for efficient genome and gene assembly and analysis.

**Date:** 4/5/12 20:27  
**Institution:** Indiana University  
**Grant:** DBI-0845494  
**Title:** CAREER: Computational and statistical genomics of gene families  
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.

**Date:** 4/9/12 9:53  
**Institution:** Oklahoma State University  
**Grant:** 924401  
**Title:** Collaborative Research: Organism-environment interactions - impact of cultural eutrophication on Daphnia tracked by genomics, physiology and resurrection ecology  
Use microarray and other technologies to discover genes and variation in their transcription linked to the phenotypic plasticity and adaptation to lake-eutrophication.

**Date:** 4/12/12 11:59  
**Institution:** Indiana University Bloomington  
**Grant:** DBI-0845685  
**Title:** Computational Protein Function Annotation for Metagenomics  
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.

**Date:** 6/12/12 11:07  
**Institution:** Marshall University  
**Grant:** 1003907  
**Title:** The Metagenome of the Ohio River  
The goal of the proposed project is to determine the diversity, stability, and resilience of the microbial community and its genetic potential across spatial and temporal scales of the ecosystems within the Ohio River. These data will allow researchers to understand the relationships between microbial diversity, the environment they inhabit, and the metabolic processes they perform. The overall hypothesis underlying this study is that every ecosystem has an identifiable microbial community that remains relatively stable over time, but that responds to changes in the ecosystem. This response may be at the taxonomic or the genetic level. Next-generation sequencing methods will be used to determine the microbial diversity and genetic potential of the microbial community within the ecosystems that make up the Ohio River. These data are vital to understanding the stability and resilience of microbial communities across relatively steep gradients of time and across ecosystem boundaries and will demonstrate how the microbial communities in different ecosystems react to natural and anthropogenic disturbances. The anticipated outcome of this project is that taxonomic and genetic data will be collected and analyzed to show relationships between microbial species, genes, and environmental parameters as well as provide baselines for the construction of models for the prediction of ecosystem response to perturbation. NCGAS resources are imperative for the completion of this project as the PI has little access to bioinformatics expertise at his home institution.
**Title: Comparative Genomics of Acanthamoeba**

The Acanthamoeba genus comprises a diverse group of protists found ubiquitously in soil and aquatic environments. They are important ecologically because of their abundance, and also because some strains are opportunistic pathogens capable of causing human infections. Hundreds of strains have been described, many displaying distinct physiological properties, though the molecular basis for these differences is largely unknown and morphological similarity makes it difficult to assign isolates to species. Commonly used genotyping methods, based on ribosomal gene sequences, help to establish phylogenetic relationships among, but do not consistently correlate with any particular phenotypic characteristic. Acanthamoeba undergoes encystment when environmental stress is encountered. Encystment is a cellular differentiation event resulting in a dormant form of the cell encased in a highly resistant cellulose structure. The coordinated intracellular networks controlling this process and the genes involved have not yet been characterized. The overall goals of this project are twofold: 1) Generate whole genome sequences for a diverse array of Acanthamoeba isolates, and using comparative genomics define the core and accessory genes that comprise the 'pan-genome' of this eukaryotic microbe, and 2) Generate and analyze transcriptomes for diverse strains of Acanthamoeba growing in/responding to various environmental conditions and as they undergo encystment, in order to link activated or repressed genes to metabolic processes or phenotypes related to differentiation. Acanthamoeba is well suited for these studies because encystment is a clearly identifiable phenotypic change, and the conditions for stimulating this event have been defined in this laboratory. We have already characterized a diverse panel of environmental isolates of Acanthamoeba, at the phenotypic level, which will be used throughout these studies.

**Title: Genomic outcomes of repeated hybrid speciation**

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.

**Title: Phylogeny, biogeography, and diversification in Pedicularis (Orobanchaceae)**

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.

**Date:** 7/11/12 23:30  
**Institution:** University of Delaware  
**Grant:** 1129816  
**Title:** IRES: Research in Industrial Projects for Students (RIPS) - Hong Kong

We are working under industrial mentors at BGI-Shenzhen for our RIPS-2012 project. We have been given two very large data sets (13GB, dna-seq reads, fastq format) and one reference genome for Daphnia pulex. Our project goal is to identify copy number variant genes or regions across two Daphnia populations (one adapted to cadmium and one not) using dna-seq data (this task is commonly completed using micro-arrays instead). We first need access to aligners and possibly snp detectors, and later in the project, will need to run our own code (probably on Bioperl) for copy number variation detection. We request to use this server until November 30th, 2012 to continue this research at our home institutions if necessary and to further refine our results to present at conferences.

**Date:** 7/12/12 13:26  
**Institution:** Colorado State University  
**Grant:** 5336590  
**Title:** Genome evolution in plethodontid salamanders: molecular and modeling-based analyses of genomic gigantism

This project is examining genome evolution in salamanders, which have the largest tetrapod genomes (15 - 120 Gb). We are using low coverage genome shotgun data to assess the repetitive landscapes of these species. Because we are using low-coverage data, it is critical to us that we evaluate the strength of the inferences we're able to make from our data. Thus, we are doing a lot of subsampling analyses from fully sequenced genomes to generate distributions of datasets equivalent to ours, and we are then assessing our ability to repetitive landscape identify patterns in the data. We are also doing a lot of de novo repeat analyses on full and subsampled datasets. We are really time-limited by computational power at this point. Colorado State University's clusters aren't able to accommodate our needs in a way that allows our research to proceed in a timely manner.

**Date:** 7/15/12 15:07  
**Institution:** College of Charleston  
**Grant:** DEB 1132229  
**Title:** Collaborative Research: Jaws and Backbone: Chondrichthyan Phylogeny and a Spine for the Vertebrate Tree of Life

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

Date: 7/17/12 12:02
Institution: UCLA
Grant: DMS-0931852
Title: Institute for Pure and Applied Mathematics

Institute for Pure and Applied Mathematics (IPAM) is an NSF math institute through DMS. IPAM supports RIPS-Hong Kong, a summer research experience for undergraduate math students held in Hong Kong in partnership with HKUST. One team is working on a project sponsored by BGI entitled “The genomic basis for metal adaptation in Daphnia” and will need to use the Mason Cluster to analyze large data sets (13GB each).

Date: 7/17/12 12:20
Institution: UCLA
Grant: DMS-0931852
Title: Institute for Pure and Applied Mathematics

The Institute for Pure and Applied Mathematics (IPAM) is an NSF math institute through DMS. IPAM sponsors RIPS-Hong Kong, a summer research experience for undergraduate math students held in Hong Kong with HKUST. One team is working on a project sponsored by BGI entitled ‘The genomic basis for metal adaptation in Daphnia’ and will require use of the Mason Server to analyze large data sets.

Date: 7/23/12 11:36
Institution: University of Chicago
Grant: NSF GRFP 2010-2013
Title: The genomic evolution of behavior in mutualistic acacia-ants

Three monophyletic species groups in the ant genus Pseudomyrmex are obligate mutualists of the plant genera Acacia, Triplaris, and Tachigali. These ants nest in and feed from their host plants and, in exchange, aggressively protect their hosts against herbivores, pathogenic fungi, and encroaching plants. However, most species of Pseudomyrmex are non-mutualistic and nest in hollow twigs that they only timidly protect. This difference in behavior is consistent and pronounced. Mutualistic behavior evolved convergently at least three times and the sister group to each clade of mutualists is non-mutualistic. Therefore, the evolutionary history of this group is incredibly useful for studying the evolution of mutualism and behavior generally. Our project focuses on genomic evolution in the group, with a first pass at selection pressures in various gene families but with future plans of comparing gene expression levels. We have Illumina sequence data for five of the seven species to be included in this study and will analyze all but the first of these genomes using reference-based assembly, cutting computation costs. While several ant genomes have now been published, none are closely related enough to our target group to make reference assembly possible meaning that we must begin with de novo genome assembly of at least one Pseudomyrmex species. We have data appropriate for ALLPATHS-LG assembly but our current computational resources are insufficient for ALLPATHS assembly. Therefore, our primary purpose on the Mason NCGAS system is ALLPATHS genome assembly.
Title: EAGER: Application of transcriptomics to investigate organism-environment relationships in marine zooplankton

This proposal is to develop transcriptomics approaches to investigate gene regulation as a function of environmental cycles and in response to experimental manipulation. The focus is on a model planktonic crustacean, Calanus finmarchicus. This calanoid copepod, is highly abundant in the North Atlantic, with populations extending from the Gulf of Maine and Labrador Sea to the North Sea.

Title: A Medicago truncatula HapMap as a Platform for Exploring the Genetics of Legume Symbioses

Legumes, the third largest family of flowering plants, are notable for their ability to form symbiotic relationships with rhizobia bacteria. This symbiosis leads to massive amounts of biological nitrogen worldwide, providing a major source of organic fertilizer and vegetable protein for humans and animals. Medicago truncatula is a widely studied model species for legume genomics and one important question focuses on the identity of naturally occurring genes that control variation in symbiosis in legumes. This project will use association mapping techniques to create a Medicago 'HapMap'. In brief, 384 diverse genetic lines obtained from collaborators at INRA-Montpellier, Ecole National Superieur Agronomique de Toulouse (ENSAT) and the Noble Foundation will be resequenced using next generation sequencing technology for sequence polymorphisms (SNPs) between the different Medicago lines. SNP discovery through genome resequencing is possible because a reference sequence for the gene-rich euchromatin of Medicago has already been created through previous NSF funding. The massive database of SNPs between Medicago lines enables the prediction of genome segments with shared ancestry (haplotypes), which can then be associated statistically with trait variation in symbiosis. Because of the exceedingly high level of SNP density, association mapping can approach the resolution of a single gene. Association mapping with whole genome SNP data requires substantial computational resources. NCGAS resources will allow us to perform permutations of our analyses to more fully explore our results.

Title: Oceanography research in Public Interest

The Kuruma prawn, Marsupenaeus japonicus, is an economically and nutritionally important species of the Penaeidae family of decapod crustaceans. To date, the sequencing of its whole genome is unavailable as a non-model organism. Transcriptomic information is also scarce for this species. In this study, we performed de novo transcriptome sequencing to produce the first comprehensive expressed sequence tag (EST) dataset for M. japonicus using high-throughput sequencing technologies, but analyse these data need High Performance Computing Systems.
Plant cells contain genomes in three distinct compartments, the mitochondrion, nucleus and plastid. Over time, thousands of genes transferred among these genomes and now there is extensive communication among the compartments and considerable conservation and stability of the genomes. The plant family Geraniaceae represents an important exception to this pattern because its mitochondrial and plastid genomes have experienced remarkably accelerated rates of change in gene content, gene order, and rates of nucleotide substitutions. The cause of these accelerated rates is unknown, but may be directed by genes encoded in the nucleus. This project investigates the basis for this accelerated evolution with the aim of understanding how different genomes within a cell can influence one another and co-evolve over time. The project will sequence, from 30 members of the Geraniaceae, the DNA in mitochondrial and plastid genomes and the genes expressed in the nucleus. These data will be analyzed to elucidate the mechanisms of inter-compartmental crosstalk and co-evolution in plant cells. The goals of this project are to determine the extent of genomic upheaval in the mitochondrial and plastid genomes of the Geraniaceae and to identify the correlated changes in the nuclear genome that have driven this instability. The large and complex data sets generated in this project - 60 organelle genomes (some of them many Mb in size and full of repetitive DNA) and complete nuclear transcriptomes from 30 plants - require powerful, high-speed computing systems as available through NCGAS for efficient genome and gene assembly and analysis.

The grass family (Poaceae) is an economically and ecologically important, species-rich radiation of monocots, which comprises over 11,000 species. Despite many evolutionary studies within the family using both molecular and structural data, a number of phylogenetic issues remain unresolved. The lack of a well-resolved phylogeny for the grasses impedes progress in understanding the morphological evolution and ecological diversification of the family. High throughput sequencing techniques now make the collection of genome-scale datasets feasible. We propose to sequence complete chloroplast genomes (plastomes) and ten low copy nuclear loci from 100 species representing major clades and subclades of Poaceae to complement sequences in over 50 other species. Our immediate need for NCGAS resources is for de novo assembly of complete plastomes from Illumina-generated sequence files. Sequences will be analyzed to: 1) perform comprehensive phylogenetic analyses of grasses to obtain the most fully resolved phylogeny possible; 2) explore the tempo and mode of evolution of the plastid and nuclear regions to better understand their phylogenetic utility; 3) use this phylogeny to address unresolved issues of the timing of evolutionary divergences and 4) develop a better understanding of some factors responsible for the ecological diversification of the grasses.

The identification of the genetic basis of adaptation is a major objective of evolutionary biology that has broad implications for ecology and conservation biology. Recent advances in the population genomics
and landscape genetics disciplines have greatly facilitated the identification of adaptive loci and allowed for the integration of environmental variables using a landscape genomics approach. This research merges spatially referenced genetic, morphological, and environmental data from a model system of Anolis lizards in the Caribbean using a Geographic Information Systems (GIS) framework. Anolis lizards are an ideal group to investigate the genetic mechanism of adaptation because they have diverged into over 400 species and are a classic example of adaptive radiation and speciation. The GIS-based analytical approach created will serve as a model framework for additional investigations of Anolis and other taxa. Our research objectives are to: 1) identify adaptive loci in A. marmoratus complex using population genomic methods and identify their chromosomal position based on the annotated Green Anole (A. carolinensis) genome map; 2) determine if a similar set of adaptive loci are identified in the different A. marmoratus subspecies and if the adaptive loci correlate to morphological features and environmental variables; and 3) model the spatial distribution of adaptive and neutral loci on Guadeloupe to determine how the environment affects neutral and adaptive gene flow. The National Center for Genome Analysis Support facility would be used to analyze our genomic and environmental data.

8.3.2. Non-NSF-funded research projects on the Mason cluster

We supported 14 additional requests on the Mason cluster during Project Year 1. These requests are listed below in chronological order.

**Date:** 2/17/12 10:27  
**Institution:** IUPUI Chemistry department  
**Grant:** None (IU users)  
**Title:** Quantum Isotropic Periodic Sum method for long range interactions.

We are implementing a new method to deal with electrostatic interactions based on the original Isotropic Periodic Sum method already applied to classical molecular dynamics simulations. The idea of the method is to treat the image charges of the simulation cell in a mean field approximation. Our goal is to implement the IPS method when the Quantum interactions are not neglected. The method should be faster than the Ewald Summation method where the image contribution is computed in the k-space. Short test simulations demonstrate a CPU time twice lower for the IPS method.

**Date:** 2/17/12 10:36  
**Institution:** IUPUI Department of Medical & Molecular Genetics  
**Grant:** None (IU users)  
**Title:** The detection and validation of novel drug-drug interactions that change the risks in myopathy and myocardial infarction in two independent EMR databases

National Health Report Statistics showed that there are an estimated 34.4 million hospital admissions per year, excluding newborns. Drug-drug interactions (DDIs) are a major cause of morbidity and mortality, responsible for nearly 3% of all hospital admissions [1] and 4.8% of admissions in the elderly [2]. DDIs are also a common form of medical error, representing 3% to 5% of all inpatient medication errors [3]. With increasing rates of polypharmacy [4], the incidence of DDIs will likely increase in the coming years. Current DDI research investigates different aspects of drug interactions. In vitro pharmacology experiments use intact cells (e.g. hepatocytes), microsomal protein fractions, or recombinant systems to investigate drug interaction mechanisms [5]. Pharmaco-epidemiology (in populo) uses a population based approach and large electronic medical record (EMR) databases to investigate the contribution of a DDI to drug efficacy and adverse drug reactions (ADRs) [6]. Very recent bioinformatics approaches, especially biomedical literature text mining based (BTM), can detect novel DDI signals from either the published literature or large clinical databases [7]. Although these approaches are complementary, they are usually conducted independently. Thus, DDIs represent an ideal model for translational research that will benefit enormously from interdisciplinary collaboration. It calls for an innovative DDI research approach, which
can investigate a large number of DDIs, and study their pharmacological mechanisms and clinical significance.

Date: 2/20/12 11:12  
Institution: School of Informatics and Computing  
Grant: None (IU users)  
Title: Human Pose Estimation from 2D Static Images  
This project is about detecting human pose from 2D images. As a CS PhD student, I work in the area of computer vision, and in this project, we hope to develop new models and new algorithms to let the machine to automatically parse the image data, and localize human body parts in the image. To this end, we need a lot of computation resources. The server machine we have right now is 24-core, and is not sufficient enough for us to run our matlab code. We hope to make it faster (like ~3X times) than what we achieve now, in order to catch up our project deadline (early April).

Date: 3/7/12 0:02  
Institution: IU  
Grant: None (IU User)  
Title: Next gen-sequencing analysis  
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.

Date: 3/9/12 13:24  
Institution: Indiana University  
Grant: None (IU)  
Title: Genomic consequences of sex-chromosome evolution in Aedes aegypti  
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.

Date: 3/26/12 14:22  
Institution: Indiana University - Bloomington  
Grant: None (IU Faculty)  
Title: Host determinants of Alphavirus replication  
Alphaviruses are pathogens of humans and livestock with worldwide distribution. Their impact on public health globally makes understanding their interaction with the host and development of interventions a high priority. Most alphavirus are transmitted obligatorily by a mosquito vector in which a lifelong
persistent infection is established. This pattern of infection in arthropods contrasts with the pathogenic infection established in vertebrate hosts. This contrast in infection pattern is mirrored in cultured cells; infection of vertebrate cells results in an acute cytolytic infection, whereas infection of mosquito cells results in a persistent, non-cytolytic infection. The molecular interactions between virus and host that result in this host specific manifestation of persistence are not understood. This project employs a proteomic approach for the identification and quantification of cellular proteins altered during an alphavirus infection of mosquito cells. The basic concept for the proposed research is straightforward; using quantitative LC-MS/MS analyses of subcellular fractions from infected and uninfected mosquito cells proteins altered in abundance as a consequence of infection will be identified. Analyses of mosquito cells at different times post-infection will allow the identification of proteins that change in quantity or location over the course of infection as the transition from acute to persistent infection occurs. Access to NCGAS resources and expertise will allow accurate and appropriate analysis of LC-MS/MS data facilitating the identification of cellular targets for further analysis as to their role in virus infection and the establishment of viral persistence.

Date: 3/26/12 23:21
Institution: Indiana University
Grant: None (IU user)
Title: Comparative Genomics of Phenotypic variation in the Compositae

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

Date: 4/3/12 15:06
Institution: Indiana University
Grant: None (IU)
Title: Human Microbiome Project

We are analyzing next generation sequencing obtained from hundreds of human microbiome samples. I am currently working on discovery of repetitive elements in samples.
Title: Parallelization of heterogeneous workloads for Imaging Genomic Browser

With collaborators in the IU Medical School, we are applying our next-generation parallel programming libraries to a recent application in genome analysis, described here: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3065788/. The application enables a user to explore correlations between genotypes and brain structure. It presents a challenging target for parallelization: first, workloads are dynamic, driven by a user manipulating a GUI; second, workloads include both 3D image processing and genome analysis components, the former of which is a good candidate for GPU execution. Our software framework balances parallelism between CPUs and GPUs on multiple nodes, and thus the ideal platform for evaluation of our techniques is a cluster with both GPUs and a high number of CPU cores per node (so as to simultaneously test scaling of multi-threading, distribution, and CPU/GPU partitioning). For this reason we are interested in using the new Delta cluster.

Title: The Geraniaceae genomes project: Accelerated and coordinated evolution across the three plant genomes

Plant cells contain genomes in three distinct compartments, the mitochondrion, nucleus and plastid. Over time, thousands of genes transferred among these genomes and now there is extensive communication among the compartments and considerable conservation and stability of the genomes. The plant family Geraniaceae represents an important exception to this pattern because its mitochondrial and plastid genomes have experienced remarkably accelerated rates of change in gene content, gene order, and rates of nucleotide substitutions. The cause of these accelerated rates is unknown, but may be directed by genes encoded in the nucleus. This project investigates the basis for this accelerated evolution with the aim of understanding how different genomes within a cell can influence one another and co-evolve over time. The project will sequence, from 30 members of the Geraniaceae, the DNA in mitochondrial and plastid genomes and the genes expressed in the nucleus. These data will be analyzed to elucidate the mechanisms of inter-compartmental crosstalk and co-evolution in plant cells. The goals of this project are to determine the extent of genomic upheaval in the mitochondrial and plastid genomes of the Geraniaceae and to identify the correlated changes in the nuclear genome that have driven this instability. The large and complex data sets generated in this project - 60 organelle genomes (some of them many Mb in size and full of repetitive DNA) and complete nuclear transcriptomes from 30 plants - require powerful, high-speed computing systems as available through NCGAS for efficient genome and gene assembly and analysis.

Title: Exploiting comparative genomics and metagenomics approaches for function prediction and understanding functional diversity.

This project will use the currently available vast amount of metagenomic data from different sources to 1) develop new methods for function prediction 2) develop rapid annotation systems for newly sequenced metagenomes (for public access) 3) understand diversity in the metagenomes for establishing criteria to identify new metagenomes to sequence.
Date: 6/12/12 22:32  
Institution: Indiana University  
Grant: None (IU faculty)  
Title: Assembly and analyses of conifer transcriptomes for USDA, NIFA, AFRI Sustainable Bioenergy Challenge Area: National Loblolly Pine Genome Sequencing  
This collaboration (UC Davis, UMd, TAMU, CHORI, WSU and IU, Mockaitis) is sequencing and assembling the genomes of Pinus taeda (loblolly pine) as well as three additional conifers. Mockaitis is responsible for broad transcriptome generation for annotation and functional genomics. Most data are Illumina strand-specific paired sequence reads. These will be subjected to assembly and comparative analyses in progressive experiments. The first need for NCGAS resources is mapping of existing transcript assemblies we have to the newly prepared genome draft v0.6. This will inform quality and completeness of the genome assembly. Next new RNAseq assemblies will be prepared and compared to each other in a quantitative and qualitative manner to distinguish tissue specificities of gene expression and transcript processing. Later in the project we will incorporate small RNA data from additional experiments.

Date: 7/9/12 11:53  
Institution: Indiana University  
Grant: None (IU)  
Title: Bioinformatics Research Project(s)  
One of my current projects is concerned with evaluating the peculiarities of gene predictors, and the effect of low quality assembly on gene counts and the 'resolution' of exons. To answer this question I have a range of predictions on a reference genome and a simulated data set, and need to compare the exon positions (208k x 248k) which is not feasible on a single computer. Another project is related to genome assembly of simulated reads, and will require significant processing and memory resources.

Date: 7/9/12 11:56  
Institution: Indiana University  
Grant: None (IU)  
Title: Yeast Multinucleotide Mutational Events/CAFE error modeling and testing  
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.