
Thomas G. Doak, PI and Manager
National Center for Genome Analysis Support
An outline:

• The science and research NCGAS addresses:
  - as an NSF service (our own grant)
  - beyond NSF (i.e. ITCR) (on others grants)

• What tools and infrastructure XSEDE provides to researchers
  e.g. Jetstream

• Docker (shifter)

• IU and NCGAS again
NCGAS’s primary goals:

- Provide bioinformatics expertise
- Maintain a curated set of applications
- Provide access to HPC resources, especially large-memory clusters = Mason, Bridges
- Build Galaxy instances for our software
- Pursue outreach to biologists
Indiana University
Research Technologies & PTI Service & Cyberinfrastructure Centers
September 2016

NCGAS is embedded in Research Technologies

Associate Dean, RT, and Executive Director, PTI
Craig A. Stewart

Vacancy, Chief Storage Architect
George Turner, Chief Systems Architect

Systems
Matthew Link, Director

High Performance File Systems
Stephen Simmons, Manager
- Thomas Crowe, Nathan Lavender
- Chris Harrison, Ken Rawlings
- Nathan Heald, Shawn Slavin

High Performance Systems
Matt Allen

High Throughput Computing
Robert Quick, Manager
- Alain Deximo
- Chris Pipes
- Kyle Gross
- Elizabeth Prout
- Marina Krenz
- Scott Teige
- Thomas Lee

Research Analytics
Scott Michael, Manager
- Jefferson Davis
- Justin Peters
- Tak Noguchi
- Kevin White

Research Storage
Kristy Kaliback-Rose, Manager
- Danko Antolovic
- Stewart Howard
- Chris Garrison
- Jeff Russ
- Karl Garrison
- Haichuan Yang

Application Virtualization
Stephanie Cox, Manager
- Ellyn Barham-Pruitt
- Steve Schunk
- Sean Mahoney

Research Data Services
M. Esen Tuna, Manager & Technical Advisor
- Nancy Long
- James McCombs
- Charles McClary
- Raminder Singh

Visualization and Analytics
Eric Wernert, Director

Advanced Visualization Lab
Michael Boyles, Manager
- Patrick Beard
- David Reagan
- Ed Dambik
- Jeffrey Rogers
- Chris Eiler
- William Sherman
- Chauncey Freund

Margaret Dolinsky,
Research Scientist

Research Community Tools
Robert Henschel, Director

Associate Dean, RT, and Executive Director, PTI
Craig A. Stewart

Science Community Tools
Robert Henschel, Manager
- Andrew Arenson
- Jim Mullen
- James Dudley
- Radha Surya
- Ryan Long
- Yelena Helen Yezekets
- Abhijeet
- Vacancy

Scientific Applications and Performance Tuning
Raymond Sheppard,
Senior Technical Lead
- H.E. Cicada Dennis
- Junjie Li
- Ben Fulton
- Abhinav Thota
- Arvind Gopu
- Michael Young
- Soichi Hayashi
- Le-Mai Weakley

Research Storage
Kristy Kallback-Rose, Manager
- Danko Antolovic
- Chris Garrison
- Karl Garrison
- Haichuan Yang

Application Virtualization
Stephanie Cox, Manager
- Ellyn Barham-Pruitt
- Sean Mahoney

Vacancy
- Chief Storage Architect
- Chief Systems Architect

Vacancy
- Chief Storage Architect
- Chief Systems Architect

Collaboration and Engagement Support
Thomas Doak, Manager

Advanced Computational Systems
David V. Hancock

Advanced Biomedical IT Core
Richard Merze, Manager
- Andrew Arenson
- Jim Mullen
- James Dudley
- Radha Surya
- Ryan Long
- Yelena Helen Yezekets
- Abhijeet
- Vacancy

National Center for Genome Analysis Support (NCGAS)
Therese Miller, Interim Director
- Raymond Sheppard,
Senior Technical Lead
- H.E. Cicada Dennis
- Junjie Li
- Ben Fulton
- Abhinav Thota
- Arvind Gopu
- Michael Young
- Soichi Hayashi

Digital Humanities Cyberinfrastructure
Tassie Gniady, Coordinator
- Ellie Heald
- Nathan Heald
- Shari Slavin

Digital Humanities Cyberinfrastructure
Tassie Gniady, Coordinator
- Ellie Heald
- Nathan Heald
- Shari Slavin

Research Technologies & PTI Service & Cyberinfrastructure Centers

***Faculty 100% Department of Biology.

*Indicates UITS staff managed by DLP.

Chart includes appointed staff only. See http://uits.iu.edu/scripts/ose.cgi?ltxt.help.
Indiana University
Research Technologies & PTI Service & Cyberinfrastructure Centers

Associate Dean, RT, and Executive Director, PTI
Craig A. Stewart

Vacancy, Chief Storage Architect
George Turner, Chief Systems Architect

High Performance Systems
Matthew Link, Director

High Performance File Systems
Stephen Simms, Manager

Thomas Crowe  Nathan Lavender
Chris Garrison  Ken Rawlings
Nathan Heald  Shawn Slavin

High Throughput Computing
Robert Quick, Manager

Alain Deximo  Chris Pipes
Kyle Gross  Elizabeth Prout
Marina Krenz  Scott Teige
Thomas Lee

Research Storage
Kristy Kaliback-Rose, Manager

Danko Antolovic  Steward Howard
Chris Garrison  Jeff Russ
Karl Garrison  Haichuan Yang

Application Virtualization
Stephanie Cox, Manager

Ellyn Barham-Pruitt  Steve Schunk
Sean Mahoney

Visualization and Analytics
Eric Wernert, Director

Advanced Visualization Lab
Michael Boyles, Manager

Patrick Beard  David Reagan
Ed Dambik  Jeffery Rogers
Chris Eller  William Sherman
Chauncey Frend

Margaret Dolinsky, Research Scientist

Research Analytics
Scott Michael, Manager

Jefferson Davis  Justin Peters
Tak Noguchi  Kevin Wilhite

Advanced Biomedical IT Core
Richard Merze, Manager

Andrew Arenson  Jim Mullen
James Dudley  Radha Surya
Ryan Long  Yelena Helen Yezertes
Abhijeet  Vacancy
Malapure

Scientific Applications Performance Tuning
Raymond Sheppard, Senior Technical Lead

H.E. Cicada Dennis  Junjie Li
Ben Fulton  Abhinav Thota
Aydind Gopu  Michael Young
Shoji Hayashi  Le-Mai Weakley

Science Community Tools
Robert Henschel, Director

Associate Dean, RT, and Executive Director, PTI
Craig A. Stewart

Science Gateways Group
Marlon Pierce, Manager

Erona Abeyesinghe  Sudhakar
Yu “Marie” Ma  Panidighantam
Suresh Marru  Jun Wang
Supon Nakandala  Vacancy

Science Gateway Cyberinfrastructure
Tassie Gniady, Coordinator

Digital Humanities Cyberinfrastructure
Craig Stewart, Principal Investigator
Michael Lynch, Co-Principal Investigator
Matthew Hahn, Co-Principal Investigator
Scott Daniel Michaels, Co-Principal Investigator

National Center for Genome Analysis Support (NCGAS)
Therese Miller, Interim Director

Thom Doak, Manager

Vacancy

Collaboration and Engagement Support
Vacancy, Manager

Advanced Computational Systems
David Y. Hancock

High Performance Systems
Matt Allen

Collin Gayde  Mike Lowe
Jeff Gronek  Joseph Rinkovksy
Bret Hammond  Jenett Tillotson
Peggy Lindenlaub

Jetstream Cyberinfrastructure

Education Outreach Training
Robert Ping, Manager

Jeremy Fischer, Sr Technical Advisor
Winona Snapp-Childs, Lead Cyberinfrastructure Analyst
Julie Wernert, Information Manager

Campus Bridging and Research Infrastructure
Richard Knepper, Manager

Eric Coulter  Aaron Wells
Barbara Hallock  Sarah Williams

Community Engagement and Interoperability
Therese Miller, Senior Interoperability

CICADAS

Indiana University

NCGAS is embedded in Research Technologies

September 2016

Chart includes appointed staff only. See http://uits.iu.edu/scripts/ose.cgi?ltxt.help.
*Indicates UITS staff managed by DLP.

***Faculty 100% Department of Biology.
Supporting NCGAS Genomics Research at PSC

Philip D. Blood, Ph.D.
Senior Computational Scientist
Pittsburgh Supercomputing Center
**Bridges:**

- 2×16 TB of cache-coherent shared memory, 4096 cores
- Ideal for genome sequence assembly
- High bandwidth, low latency interprocessor communication

*Bridges* leverages its large memory for *interactivity* and to seamlessly support applications through *virtualization*, *gateways*, familiar and *productive programming* environments, and *data-driven* workflows.
From our recent NSF survey:

Types of Research (Multi-Pick):

- RNA-Seq approaches? 26%
- Microbiome sequencing? 14%
- Whole genome structural analysis, such as nucleosome mapping or high-resolution DNase sensitivity? 16%
- Gene function analysis using parallel transposon insertion sequencing? 10%
- Protein-DNA interactions with ChIP-seq? 9%
- Methylation studies with whole-genome bisulfate sequencing? 6%
- Gene expression profiles with microarrays? 6%
- Proteomics and Mass Spec? 3%
- Genome annotation and ortholog discovery? 2%
- Other 8%
From our recent NSF survey:

Would any of the follow services be helpful to you (Multi-Pick):

- Bioinformaticians on-call: 12%
- Large RAM infrastructure: 21%
- Fast CPU nodes: 23%
- Installed curated genomics applications: 13%
- Installation of published bioinformatic applications: 11%
- Mounting of applications as web-accessible Galaxy tools: 20%
- Rapid access to reference data for analysis: 20%
Galaxy Anatomy and Physiology

Tool bar – contains the available steps to apply to data

Focus pane – shows options, parameters, and output for current item.

History – shows steps previously taken to manipulate input data sets

National Center for Genome Analysis Support: [http://ncgas.org](http://ncgas.org)
We’re reaching most states

National Center for Genome Analysis Support Users 2016
Representing approximately 105 institutions
MDIBL hosted Environmental Genomics in Salisbury Cove, Maine. Over one week, 9 students and 9 post doc/faculty from 14 universities learned to design RNA-seq experiments, create Daphnia RNA-seq libraries, manage and analyze the data, and present their results.

NCGAS partnered with the workshop for the first time in 2016 and provided:

- Three reserved Karst nodes, totaling
  - 48 processors
  - 48 GBs of RAM
- On site training and consultation on cluster use, bioinformatics, and statistics. Immediate coordination with IU sysadmin.
- Rapid processing of ~250Gb of sequence data, when data arrival was delayed

As a result of partnership with NCGAS, the workshop was able to include 4.5x more data throughput, enabling more realistic and complex experimental designs.
We ask that you acknowledge our grant in any published work that uses our resources. Collaborations and authorship are requested for intellectual contributions.

THE FACTS
- 16-nodes, 500GB RAM
- 10TB project space
- Bioinformatics software
- Galaxy instance
- 50TB archive space/user
IU students and faculty have access to Galaxy @ IU using their Indiana University credentials. Even with access to the resources available at IU, the interface between users and computer clusters can be daunting. We provide support to IU affiliates through Galaxy to accomplish their bioinformatics analyses without the need for a degree in computer science.

Click here to enter Galaxy @ IU

Trinity CTAT Galaxy, hosted by Indiana University and the Broad Institute, is a free-to-use public interface for Trinity users

Click here to enter Trinity_CTAT Galaxy

GenePattern is a freely available computational biology open-source software package developed at the Broad Institute of MIT and Harvard, for the analysis of genomic data. NCGAS now hosts a free GenePattern server, with increased computational resources.

Click here to enter IU’s Public GenePattern server

Galaxy Main, hosted by Penn State University and Emory University, is a free-to-use public service that includes hundreds of tools and a server with 250 GB of storage space per user. Our own Galaxy instances are based off of this technology.

Click here to enter Galaxy Main
Welcome to the Trinity Galaxy Instance

Thank you for choosing Galaxy!
Get started with some help moving files into Galaxy. Feel free to visit our FAQ page for additional information.
We are committed to helping you succeed with your research. Don't hesitate to contact help@ncgas.org if you need help or if you have questions or suggestions.

This Galaxy instance is running on hardware that is scheduled to be unavailable on the first Tuesday of every month for maintenance. Jobs that are started before this time will resume after maintenance.

The Trinity project is supported by the National Cancer Institute of the National Institutes of Health under award number U24CA180922.

This instance of the Galaxy is installed and maintained by National Center for Genome Analysis Support NCGAS (NSF Award #1062432)
The Trinity Grant punched Karst “condo” nodes

Science DMZ
(demilitarized zone)

Karst Cluster

T1
Dedicated to Trinity:
- 3 IBM/Lenovo NeXtScale nx360 M4 server nodes

T2
- Each node has:
  - Up to 3.4GHz clock
  - 16 cores
  - hyperthreading
  - 512GB RAM

T3
- Red Hat EL6
- Torque + MOAB

10 gigE
National Center for Genome Analysis Support and Trinity Galaxy Users 2016

Representing approximately 233 institutions
National Center for Genome Analysis Support and Trinity Galaxy Users 2016

Representing approximately 545 institutions in 49 countries

Users by State
Galaxy Job CPU Hours by Input File Size

Data from 2014-04-03 to 2016-09-02

<table>
<thead>
<tr>
<th>Input File Size</th>
<th>0-1h</th>
<th>1-5h</th>
<th>5-10h</th>
<th>10-30h</th>
<th>30-50h</th>
<th>&gt;50h</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;40GB</td>
<td>7</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>46</td>
</tr>
<tr>
<td>30-40GB</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>18</td>
</tr>
<tr>
<td>20-30GB</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>28</td>
</tr>
<tr>
<td>10-20GB</td>
<td>12</td>
<td>5</td>
<td>7</td>
<td>8</td>
<td>5</td>
<td>39</td>
</tr>
<tr>
<td>5-10GB</td>
<td>19</td>
<td>21</td>
<td>2</td>
<td>10</td>
<td>13</td>
<td>85</td>
</tr>
<tr>
<td>0-5GB</td>
<td>137</td>
<td>60</td>
<td>38</td>
<td>246</td>
<td>181</td>
<td>102</td>
</tr>
</tbody>
</table>
Trinity metrics: Just for fun

Top Ten Galaxy Users by Job Submissions

Data from 2014-04-03 to 2016-09-02

- luenlee_yahoo_com
- jfmclaughlin_alaska_edu
- e_schuster_ucl_ac_uk
- kevin_winker_alaska_edu
- alballou_ncsu_edu
- gwilymh_sfu_ca
- chrltsd_mcmaster_ca
- fspaulding_alaska_edu
- lorena_bioinf_uni_leipzig_de
- Lukacdm_njms_rutgers_edu

Job Count
Use GenePattern

The GenePattern team and collaborating organizations maintain several servers that can be used without installing any software. All that is needed is to register (Note that each server must be registered for separately). More information about the servers is provided below.

GenePattern @ Broad
https://genepattern.broadinstitute.org/gp

The Broad Institute hosts a publicly available GenePattern server.

- The job purge for the public Broad-hosted server is set to 7 days.
- There is a 30 GB quota on jobs and uploaded data.
- Most of the modules and pipelines available from the Broad Institute (see the Modules page of the GenePattern web site) are available on the Broad-hosted server. Several modules are available only on the Broad-hosted server because they require customized server configuration. (If you are interested in these modules, contact us for more information). A small number of modules are not available on the Broad-hosted server because they run only on the Windows platform; the Broad-hosted server runs under Unix.

GenePattern @ Indiana University
http://gp.indiana.edu/gp/

The GenePattern team in collaboration with Indiana University’s (IU) National Center for Genome Analysis Support (NCGAS) hosts a public server on IU’s high performance computing system. This server has more capacity to better accommodate next generation sequencing analysis and other compute intensive analyses.

- This resource is only available for academic and non-profit users.
- The job purge for the IU server is currently set to 30 days, as opposed to 7 days on the Broad server, so your jobs will remain on the server longer.
- There is currently no quota on the IU server, with regard to jobs or data upload. This may change in the future.
- The server is unavailable on the first Tuesday of every month for planned maintenance. We will communicate any other planned downtime as far in advance as possible.

@Pittsburgh (PSC) too
A Roadmap to the Research Information Superhighway: over 200 supercomputer centers are interconnected across a series of high speed physical networks. The resources in these centers are shared across organizations such as XSEDE and OSG. Specialized centers use XSEDE and OSG to support specialized user communities.
Anyone can use XSEDE

Total Research Funding Supported by XSEDE in Program Years 1-4

$1.68 billion in research supported by XSEDE in PY1-PY4 (July 2011-June 2015)

Research funding only. XSEDE leverages and integrates additional infrastructure, some funded by NSF (e.g. Track 2 systems) and some not (e.g. Internet2).
Research Computing on Cloudy Platforms
Jetstream: A national research and educational cloud

J. Michael Lowe (jomlowe@iu.edu)  George Turner (turnerg@iu.edu)
Jetstream System Engineer  Chief Systems Architect
IU High Performance Systems  IU Research Technologies

Operating Innovative Networks Workshop, Indiana University - Bloomington, 12-July-2016

XSEDE allocations: PIs may have support from any funding agency or funding source.
What is Jetstream?

• User-friendly, widely accessible cloud environment

• User-selectable library of preconfigured virtual machines; no need for system administration skills.

• NSF’s first production cloud facility supporting all areas of science and engineering within NSF’s scope

• Enable discoveries across disciplines such as biology, atmospheric science, economics, network science, observational astronomy, and social sciences.
Containers: Docker and Shifter

Shifter is built on docker, uses the images, etc. It replaces the run management (our end) to work with clusters and Cray. The repo retrieval is only slightly different (goes through a filter first), but can pull images made in docker and from dockerhub.

Basically, shifter is like docker bubble wrap for HPC systems, and people familiar with docker would have little to learn to run shifter: just little nuances.

Sheri Sanders

Shifter is "just another docker engine"... meaning it will run docker images... much the same way as the "original docker engine" runs docker images. Users can thus use all the existing docker containers from docker hub. The big picture: there is no issue using Shifter, even if all your work is in "the docker world".

Robert Henschel
1) The NCGAS Model
   1) How to work with us:
      i. We can host a server.
      ii. We can provide a user interface, such as Galaxy
      iii. We can “serve a community” as we do now for NSF researchers.
      iv. We’re good at genomics/bioinformatics; image analyses would probably involve an additional RT group.

Summarizing:
Thank You

Questions?
Tom Doak (tdoak@iu.edu)
help@ncgas.org