The National Center for Genome Analysis Support as a Model Virtual Resource for Biologists


William K. Barnett, Ph.D.
National Center for Genome Analysis Support
Summary

- The NGS Big Data Problem
- NCGAS as a National Model
- Current State and Prospects
The Next Generation Sequencing Big Data Problem
Changing genomics data environment

• Sequencing is:
  ➢ Becoming commoditized at large centers,
  ➢ Multiplying at individual labs, and
  ➢ Generating (MUCH) more data

• Analytical capability lacks:
  ➢ Bioinformatics support
  ➢ Computational support
  ➢ Storage support
Sequencing is getting cheaper

Source: http://www.genome.gov/images/content/cost_per_genome.jpg
NGS Data growth is outstripping Storage growth

omicsmaps.com listed 931 Next Generation Sequencers in the US on April 29, 2013
<table>
<thead>
<tr>
<th></th>
<th>Coverage</th>
<th>No. of Reads</th>
<th>Read Length</th>
<th>BAM File Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole Genome</td>
<td>37.7x</td>
<td>975,000,000</td>
<td>115</td>
<td>82 GB</td>
</tr>
<tr>
<td>Whole Genome</td>
<td>38.4x</td>
<td>3,200,000,000</td>
<td>36</td>
<td>138 GB</td>
</tr>
<tr>
<td>Exome</td>
<td>40x</td>
<td>110,000,000</td>
<td>75</td>
<td>5.7 GB</td>
</tr>
</tbody>
</table>

Source: Nancy Spinner Keynote at 2012 Internet2 meeting: [http://events.internet2.edu/2012/fall-mm/agenda.cfm?go=session&id=10002624&event=1149](http://events.internet2.edu/2012/fall-mm/agenda.cfm?go=session&id=10002624&event=1149)
But other disciplines already handle Bigger

One Degree Imager: 1.4 Petabytes per year

Large Hadron Collider: 15 Petabytes per year. The Open Science Grid moves 2 Petabytes per day

The Square Kilometer Array: 2,000 Petabytes per year in 2013
The National Center for Genome Analysis Support as a National Model
1. Bioinformatics consulting for biologists
2. Large memory clusters for assembly
3. Optimized software for better efficiency

- Initially Funded by the National Science Foundation
- Collaboration across multiple institutions
- Open for business at: http://ncgas.org
Research networks are now (much) faster than hard drives.
**NCGAS Service Model**

**Public Cloud**
- Software as a Service
- Platform as a Service
- Infrastructure as a Service

**NEEDS**
- Bioinformatics
- Applications
- Services Layer
- OS Layer
- Hardware Layer
- Network Layer

**NCGAS**
- Expert Consulting
- Galaxy and Tuned Apps
- Parallel Optimized Envs.
- Professional Admins
- Supercomputer Cntrs
- 100 Gbps Internet2

**Aligned with HIPAA for Clinical research at IU**
The Galaxy Instance at Indiana University

using Galaxy! NCGAS is committed to providing support for Indiana University research. Don't hesitate to contact us if you find that you need a tool that is not supported by our current Galaxy or if you have questions or suggestions. We will have the requested tool up and running in two working days time, and failing that, will report the status of the request within that time.

Galaxy is installed and maintained by National Center for Genome Analysis Support NCGAS (NSF Award #1062432) is provided by the Indiana University Mason Compute Cluster

1 by the Indiana University Data Capacitor (NSF Award # 0521433)

2d on the Indiana University Quarry Gateway Hosting

Supported in part by NSF, NHGRI, and the Huck Institutes of the Life Sciences.

The Pervasive Technology Institute at Indiana University

s.org
How does this work at scale?

1. Researchers use Galaxy to transfer files and run jobs
   - Files transferred to Data Capacitor over Internet2
   - Data Capacitor mounted on several Clusters
   - Data Capacitor mounts reference data from NCBI

2. Workflows execute on best system for that analysis

3. Results delivered through web interfaces and to visualization or other science tools
Rates

For NSF Funded Researchers at IU or XSEDE Allocations:

- Consulting: $0 – currently 2.5 FTE bioinformaticians
- Data Transfer: $0
- Data Storage: $0 – contemplating 25 TB allocation/project
- Computation: $0

Other Federally funded Researchers running on the POD

- Consulting: $60/hour for short consults
- Data Transfer: $20 per transfer
- Data Storage: $.10/GB/month
- Computation: $.09/core hour (128GB/node cluster)
Current State and Prospects
What have we done so far?

1. Partnering on 37 research projects, managing 73 TB
2. IU infrastructure aligned with HIPAA
3. XSEDE Tier 2 Service Provider
4. Optimized Trinity with Broad Institute
5. LustreWAN node at U. Hawaii at Hilo

Next:
1. LustreWAN node at NCBI (July 2013)
2. More Science
3. More LustreWAN nodes at sequencing sites
NCGAS Partners

Funding From:

- NCGAS Partners
- Pervasive Technology Institute
- INDIANA UNIVERSITY
- XSEDE
  Extreme Science and Engineering Discovery Environment
- TACC
  Texas Advanced Computing Center
- Pittsburgh Supercomputing Center
- CIPRES
  Cyber Infrastructure for Phylogenetic Research
- SDSC
  San Diego Supercomputer Center
- NSF
- NIH
  National Institutes of Health
Questions?

Bill Barnett (barnettw@iu.edu)

The NCGAS Team at IU:
Rich LeDuc (rleduc@iu.edu)
Le-Shin Wu (lewu@iu.edu)
Thomas Doak (tdoak@indiana.edu)
Carrie Ganote (cganote@iu.edu)
**Acknowledgements and Disclaimers**

- This material is based upon work supported by the National Science Foundation under Grant No. ABI-1062432, Craig Stewart, PI. William Barnett, Matthew Hahn, and Michael Lynch, co-PIs.

- This work was supported in part by the Lilly Endowment, Inc. and the Indiana University Pervasive Technology Institute

- Any opinions presented here are those of the presenter(s) and do not necessarily represent the opinions of the National Science Foundation or any other funding agencies
License Terms


Items indicated with a © are under copyright and used here with permission. Such items may not be reused without permission from the holder of copyright except where license terms noted on a slide permit reuse.

Except where otherwise noted, contents of this presentation are copyright 2011 by the Trustees of Indiana University.

This document is released under the Creative Commons Attribution 3.0 Unported license (http://creativecommons.org/licenses/by/3.0/). This license includes the following terms: You are free to share – to copy, distribute and transmit the work and to remix – to adapt the work under the following conditions: attribution – you must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work). For any reuse or distribution, you must make clear to others the license terms of this work.
JIM WILLIAMS
DIRECTOR, INTERNATIONAL NETWORKING
WILLIAM@INDIANA.EDU
US DOMESTIC NETWORKING

- Very strong high-performance backbone (100G national footprint) both from ESnet and Internet2
- Performance to your institution and your lab may vary due to local considerations
- Internet2 and ESnet have network related scientist support functions and services
US AND INTERNATIONAL NETWORKING

- NSF funded IRNC program provides 10G connectivity to Europe, Asia and South/Latin America
- Within those regions national networks provide varying degrees of connectivity
- IRNC investigators anxious to assist researchers with their connectivity issues
GÉANT and sister networks enabling user collaboration across the globe

April 2013
FUTURES

- Domestic networking in many countries is 100G. International networking moving in that direction.
- ANA-100G supplies 100G TA connectivity between NYC and AMS.
- ESnet interested in extending the 100G ESnet network to Europe (EEX)
SUPPORT IS ALWAYS THE ISSUE

- To contact IRNC investigators: http://www.irnclinks.net/
- To contact Internet2: http://www.internet2.edu/science/
  http://www.internet2.edu/health/
- To contact ESnet: http://es.net/