The National Center for Genome Analysis Support and Galaxy

William K. Barnett, Ph.D. (Director)
Richard LeDuc, Ph.D. (Manager)
National Center for Genome Analysis Support

Galaxy Community Conference July 27, 2012
Summary

• NCGAS and its mission
• NCGAS cyberinfrastructure
• The 100 Gigabit demonstration
• Scaling genomics analysis
• Trinity optimization
Changing genomics analytical needs

• Next Gen sequencers are generating more data and getting cheaper

• Sequencing is:
  ➢ Becoming commoditized at large centers and
  ➢ Multiplying at individual labs

• Analytical capacity has not kept up
  ➢ Bioinformatics support
  ➢ Computational support (thousand points solution)
  ➢ Storage support
NCGAS widens the analytical bottleneck

• Funded by National Science Foundation
• Large memory clusters for assembly
• Bioinformatics consulting for biologists
• Optimized software for better efficiency
• Open for business at: http://ncgas.org
Making it easier for Biologists

- Galaxy interface provides a “user friendly” window to NCGAS resources
- Supports many bioinformatics tools
- Available for both research and instruction.
NCGAS Cyberinfrastructure at IU

- Mason large memory cluster (512 GB/node)
- Quarry cluster (16 GB/node)
- Data Capacitor (1 PB at 20 Gbps throughput)
- Research File System (RFS) for data storage
- Research Database Cluster for managing data sets.
- All interconnected with a high speed internal network (40 Gbps)
GALAXY.IU.EDU Model

Individual labs can get duplicate boxes – provided they support it themselves.

Virtual box hosting
Galaxy.IU.edu

The host for each tool is configured to meet IU needs

UITS/NCGAS establishes tools, hardens them, and moves them into production.

Quarry

Mason

RFS

Data Capacitor

A custom Galaxy tool can be made to import data from the RFS to the DC.

Policies on the DC guarantee that untouched data is removed with time.

National Center for Genome Analysis Support: http://ncgas.org

GCC. July 27, 2012
NCGAS Sandbox Demo at SC 11

- **STEP 1:** data preprocessing, to evaluate and improve the quality of the input sequence
- **STEP 2:** sequence alignment to a known reference genome
- **STEP 3:** SNP detection to scan the alignment result for new polymorphisms

National Center for Genome Analysis Support: [http://ncgas.org](http://ncgas.org)
Two Options for Computation and Storage

- NCGAS Mason
  - Free for NSF users

- IU POD
  - 12 cents per core hour

- Data Capacitor
  - NO data storage charges

- Amazon EC2
  - 20 cents per core hour

- Amazon Cloud Storage
  - $80 – 120 per TB per month
This Architecture Scales!

- Commodity Internet (1Gbps but highly variable)
- Internet2 (100Gbps)
- DDR3 SDRAM (51.2 Gbps, 6.4GBps)
- IU Data Capacitor (20 Gbps throughput)
- NLR to Sequencing Centers (10Gbps/link)
- Ultra SCSI 160 Disk (1.2 Gbps, 160 MBps)
- Commodity Internet (1Gbps but highly variable)
How would this work at scale?

1. Biologists use Galaxy to execute workflows
2. Sequence data mounted via Lustre WAN or automatically transferred using Internet2
3. Data Capacitor flows data into Mason or other computational clusters
4. Data Capacitor mounts or mirrors reference data from NCBI or other sources
5. Results delivered through web interfaces and to visualization or other science tools

National Center for Genome Analysis Support:  http://ncgas.org
Performance Improvements

Inchworm, GraphFromFastA, ReadsToTranscripts, QuantifyGraph

Core Utilization

I/O MB/s

Runtime (hours)

RESEARCH TECHNOLOGIES
INDIANA UNIVERSITY

ZIH
Center for Information Technology Services & High Performance Computing

NATIONAL CENTER FOR GENOME ANALYSIS SUPPORT
INDIANA UNIVERSITY

Richard LeDuc
GCC 2012
7/27/2012
Final Results

- Original (Mason)
- General opt. (Mason)
- Optimized (Mason)
- Optimized (Blacklight)

Runtime (hours)

Data set size (Gbp)
Trinity Results

- Significantly reduced runtime, while maintaining correctness of results
- Results are published
- Source code is commit to official SourceForge repository
- Continued support for HPC optimization for Trinity
- Brian Haas at Broad is developing Trinity workflows for Galaxy
In Sum…

• NG Sequencing is creating a analytical problem that cannot be solved at sequencing centers

• NCGAS can provide a global scale infrastructure to better serve the needs of biologists who cannot become bioinformaticians to accomplish their research.

• Trinity is no longer a resource hog
Thank You

Questions?

Bill Barnett (barnettw@iu.edu)
Rich LeDuc (rleduc@iu.edu)