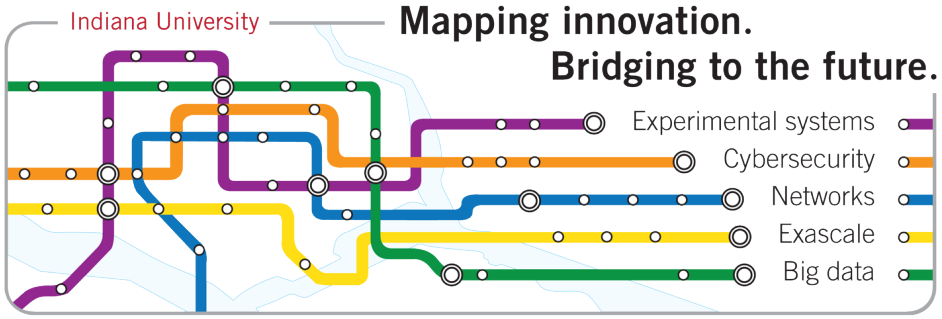
Biology on a National Scale Parallel File System

Richard LeDuc
Manager
National Center for Genome Analysis Support

November 13, 2012



Summary

NCGAS and its mission
NCGAS cyberinfrastructure
The role of the Data Capacitor
Scaling genomics analysis

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
- Storage support

NATIONAL CENTER FOR GENOME ANALYSIS SUPPORT

INDIANA UNIVERSITY

Funded by National Science Foundation

1. Large memory clusters for assembly
2. Bioinformatics consulting for biologists
3. Optimized software for better efficiency

Collaboration across multiple institutions

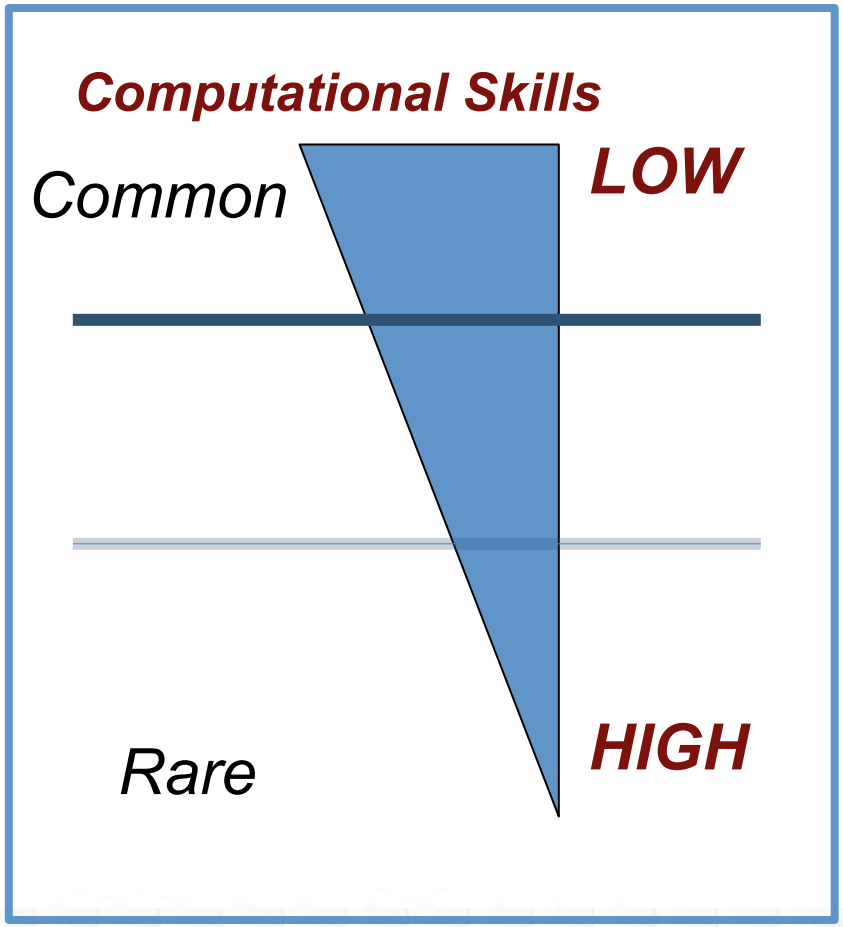
Open for business at: http://ncgas.org

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)

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

Making it easier for Biologists

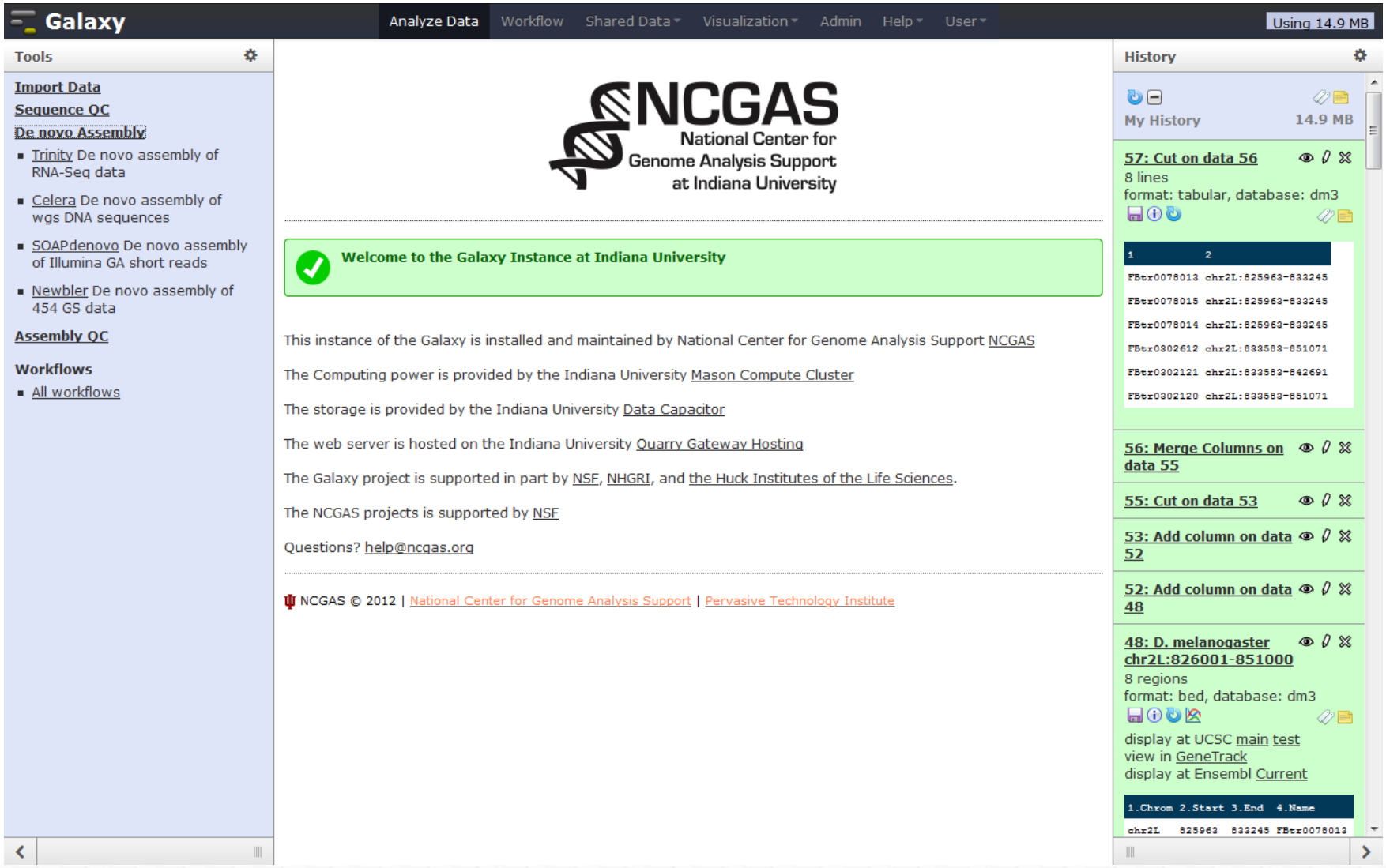


Galaxy

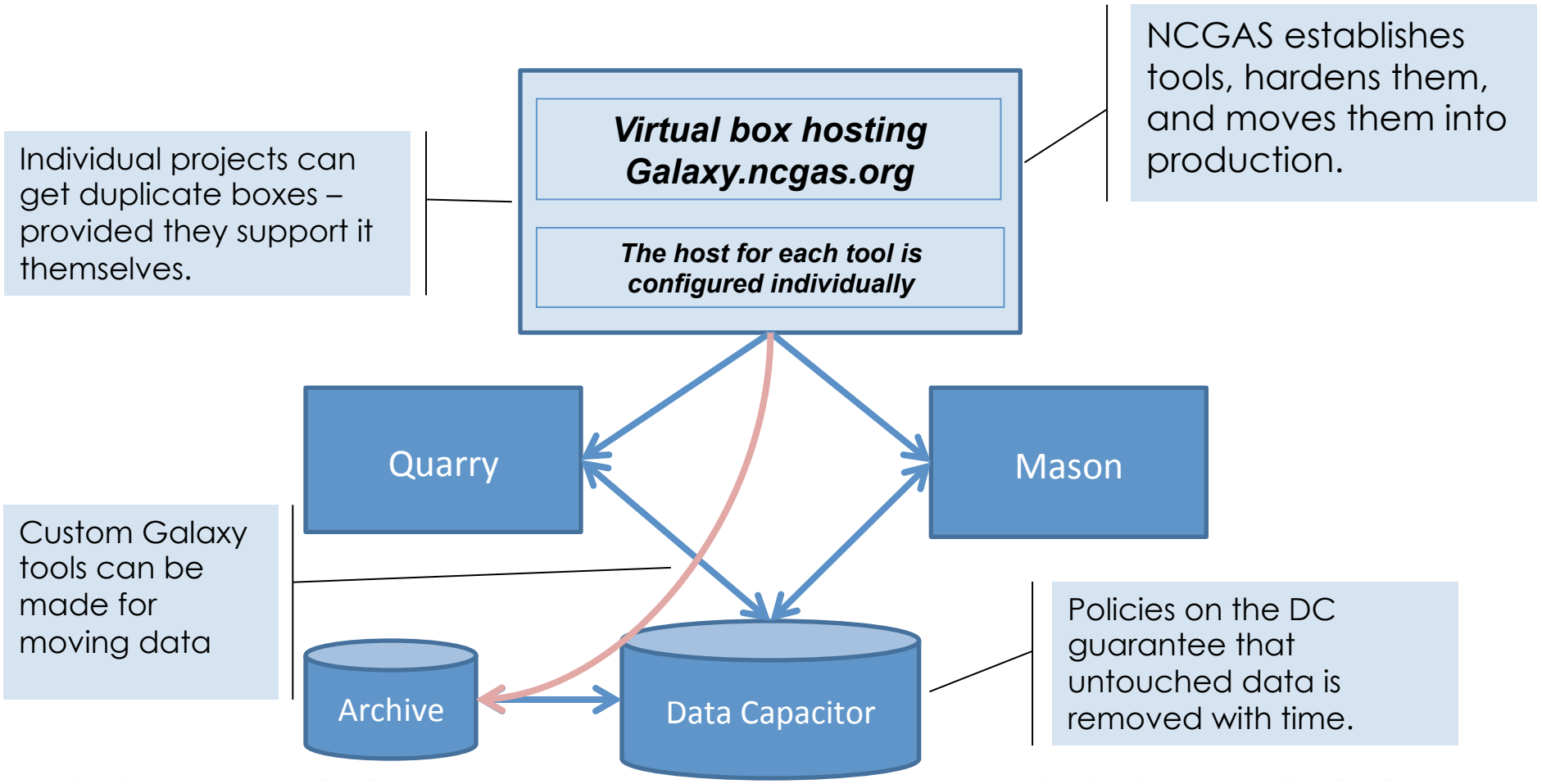
Web interface to NCGAS resources

Supports many bioinformatics tools

Available for both research and instruction.


Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 14.9 MB
Tools
Import Data
Sequence QC
De novo Assembly
Trinity De novo assembly of RNA-Seq data
Celera De novo assembly of wgs DNA sequences
SOAPdenovo De novo assembly of Illumina GA short reads
Newbler De novo assembly of 454 GS data
Assembly QC
Workflows
All workflows
NCGAS
National Center for Genome Analysis Support at Indiana University
Welcome to the Galaxy Instance at Indiana University
This instance of the Galaxy is installed and maintained by National Center for Genome Analysis Support NCGAS
The Computing power is provided by the Indiana University Mason Compute Cluster
The storage is provided by the Indiana University Data Capacitor
The web server is hosted on the Indiana University Quarry Gateway Hosting
The Galaxy project is supported in part by NSF, NHGRI, and the Huck Institutes of the Life Sciences.
The NCGAS projects is supported by NSF
Questions? help@ncgas.org
NCGAS © 2012 | National Center for Genome Analysis Support | Pervasive Technology Institute
History
My History
14.9 MB
57: Cut on data 56
8 lines
format: tabular, database: dm3
1 2
FBtr0078013 chr2L:825963-833245
FBtr0078015 chr2L:825963-833245
FBtr0078014 chr2L:825963-833245
FBtr0302612 chr2L:833583-851071
FBtr0302121 chr2L:833583-842691
FBtr0302120 chr2L:833583-851071
56: Merge Columns on data 55
55: Cut on data 53
53: Add column on data 52
52: Add column on data 48
48: D. melanogaster chr2L:826001-851000
8 regions
format: bed, database: dm3
display at UCSC main test
view in GeneTrack
display at Ensembl Current
1.Chrom 2.Start 3.End 4.Name
chr2L 825963 833245 FBtr0078013

GALAXY.NCGAS.ORG Model



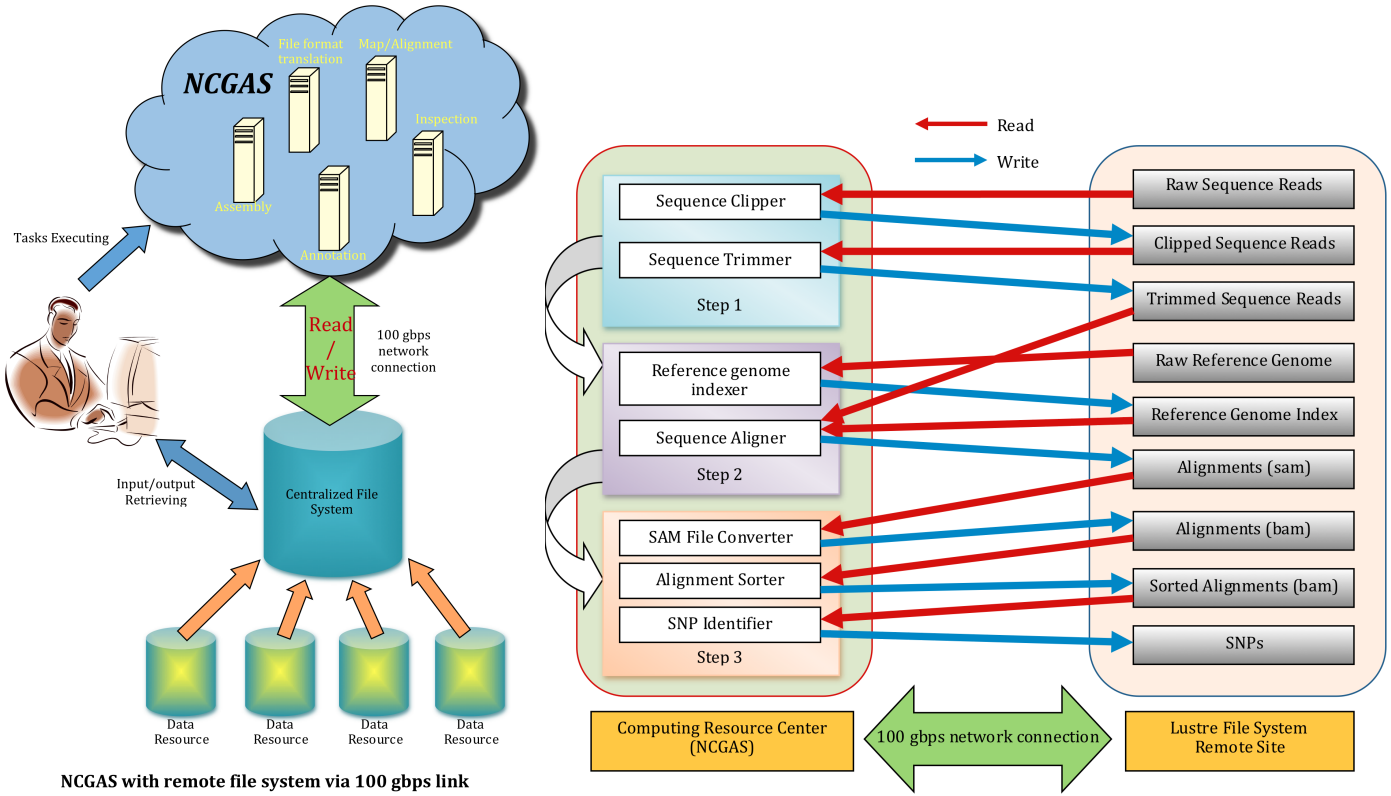
Individual projects can get duplicate boxes – provided they support it themselves.

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

Custom Galaxy tools can be made for moving data

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

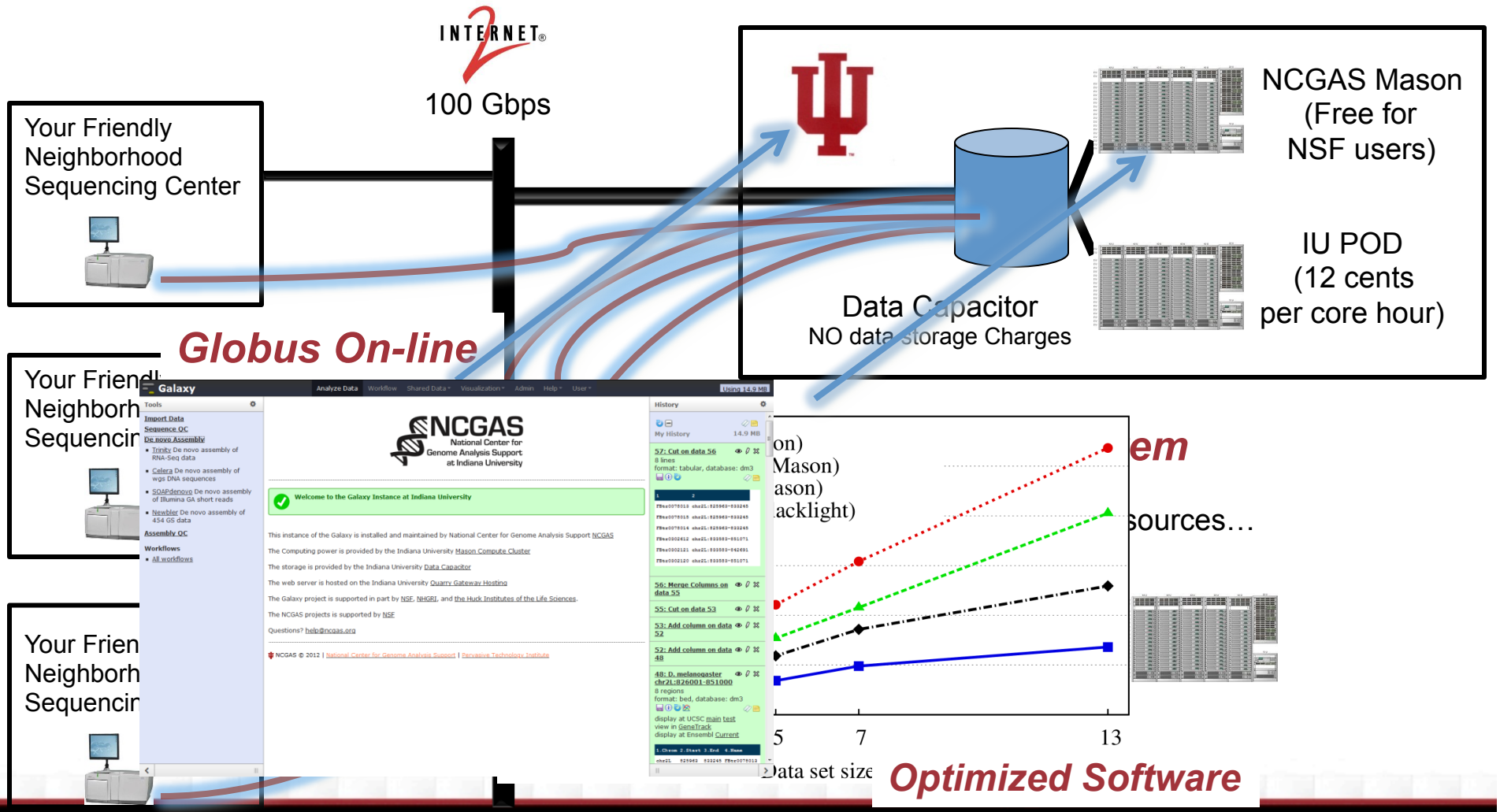
NCGAS Sandbox Demo at SC 11



NCGAS with remote file system via 100 gbps link

- *STEP 1: **data pre-processing**, 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*

Moving Forward



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

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.

The Data Capacitor allows NCGAS to create a web portal for “compute in place” analysis of genomic data across widely distributed resources.

Thank You

Questions?

Bill Barnett (barnettw@iu.edu)
Rich LeDuc (rleduc@iu.edu)
Le-Shin Wu (lewu@iu.edu)
Carrie Ganote (cganote@iu.edu)

NATIONAL CENTER FOR GENOME ANALYSIS SUPPORT
INDIANA UNIVERSITY

Acknowledgements & disclaimer

This material is based upon work supported by the National Science Foundation under Grants No. ABI-1062432
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

Please cite as: LeDuc, R.D., The National Center for Genome Analysis Support, presented at LOCATION OF TALK