Globus Genomics @ Georgetown

Yuriy Gusev - ICBI, Georgetown University, Washington DC

Computation Institute

THE UNIVERSITY OF CHICAGO

icbi

INNOVATION CENTER FOR BIOMEDICAL INFORMATICS

Georgetown | Lombardi

COMPREHENSIVE CANCER CENTER

Innovation Center for Biomedical Informatics
http://icbi.georgetown.edu/

Innovation Center for Biomedical Informatics:
Enabling Translational Genomics Research

ICBI — Innovation Center for Biomedical Informatics

Georgetown UNIVERSITY MEDICAL CENTER
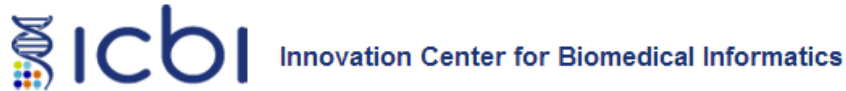
About Us People Biomedical Research Technology Education & Training Our Impact News & Events Careers

biomedical research
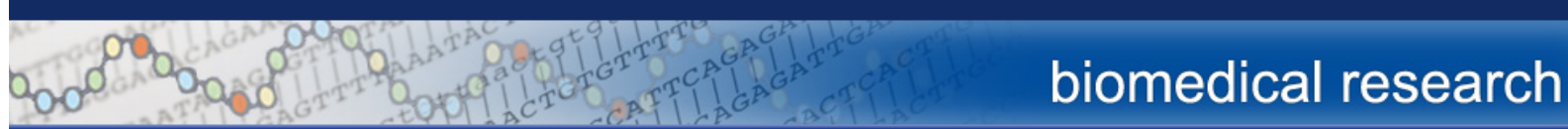
home > Genome Research

search

Biomedical Research

Cancer Research
  Cancer Systems Biology
    U54
  Data Coordination Centers
    Breast and Colon Cancer Family Registries
    Clinical Proteomic Tumor Analysis Consortium
  In Silico Cancer Research
    ISRCE
Drug Discovery
  Drug Informatics Database
  Ligand Family Database
Genome Research
Translational Research
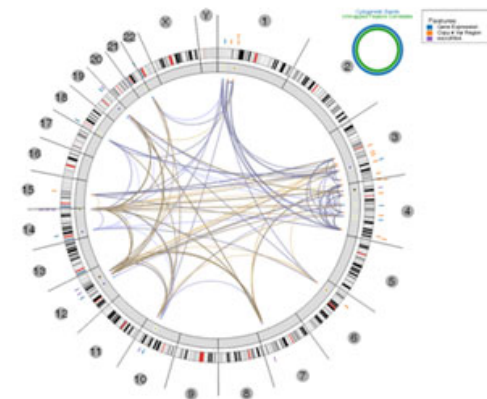 Clinical and Translational Science Award
  G-DOC
  Pharmacogenomics
    Oncology Biomarkers & Drug Response
  Vaccine Adverse Events

GENOME RESEARCH

With the sequencing of the human genome and availability of high power computational methods and various high throughput technologies, biomedical research is poised to undergo revolutionary change. As a result of these developments and emerging technologies, *genome research* continues to evolve as a field that develops novel insights into the genome biology of all organisms, including significant advances in genomic medicine. The field comprises of developments in cutting-edge computational biology and high-throughput methodologies. Faculty and staff members of ICBI engage in genome research on critical projects involving breast, pancreatic and colorectal cancers as well as neurobiology and pediatric diseases. Our efforts in genome research include:

- Determining the function of genes and the elements that regulate genes throughout the genome by utilizing systems biology techniques

- Finding variations in the DNA sequence among people and determining their significance and associations to various genetic and non-genetic diseases. The most common type of genetic variation is known as a single nucleotide polymorphism or SNP (pronounced "snip"). These small differences may help predict a person's risk of particular diseases and response to certain medications (field known as Pharmacogenomics)

- Developing and applying genome-based computational strategies for the early detection, diagnosis, and treatment of disease

- Understanding the 3-dimensional structures of proteins and identify their functions in the context of genome variations as related to drug response or disease causation
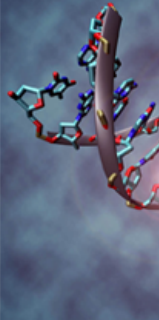
GDOC: Our Web Platform for Translational Cancer Research
gdoc.georgetown.edu

Next Generation Sequencing Revolution – Clinical Applications for Personalized Medicine

NGS projects at ICBI

- Typical Translational Projects:
  - Mutli-Omics Profiling of Colorectal Cancer: : **40 samples**
    Whole Exome Seq

  - Early Detection of Alzheimer's: **156 samples**
    Whole Genome Seq; RNA Seq; microRNAseq

  - Pre-Term Birth: **2000 samples**
    Whole Genome Seq; RNA Seq; microRNAseq

  Computational Solutions: customized pipelines at AWS
  Problem: not suitable for large scale service/production

Next Generation Sequencing (NGS):
New Challenges and Opportunities

- **Translational Research Projects:**

  ~20 to ~ 200 samples (5Gb-50Gb/sample)

- Data management: scalability, accessibility, cost

- Data Transfer, Processing and Analysis:

  – **Example: Whole Exome seq - 40 samples CRC (3 disk drives)**

- Some of the existing tools do not scale well:

  – **Examples: TopHat, Cufflinks (>1 month to process ~80 samples)**

- Paradigm shift: "Data to Tools" vs "Tools to Data"

- Data re-usage - *in silico* research

Solution?: Cloud computing with Genomics computational environment – Potential candidate: Galaxy CloudMan

But: scalability problems starting from <10 samples;

Also: does not solve the data transfer and management problems.

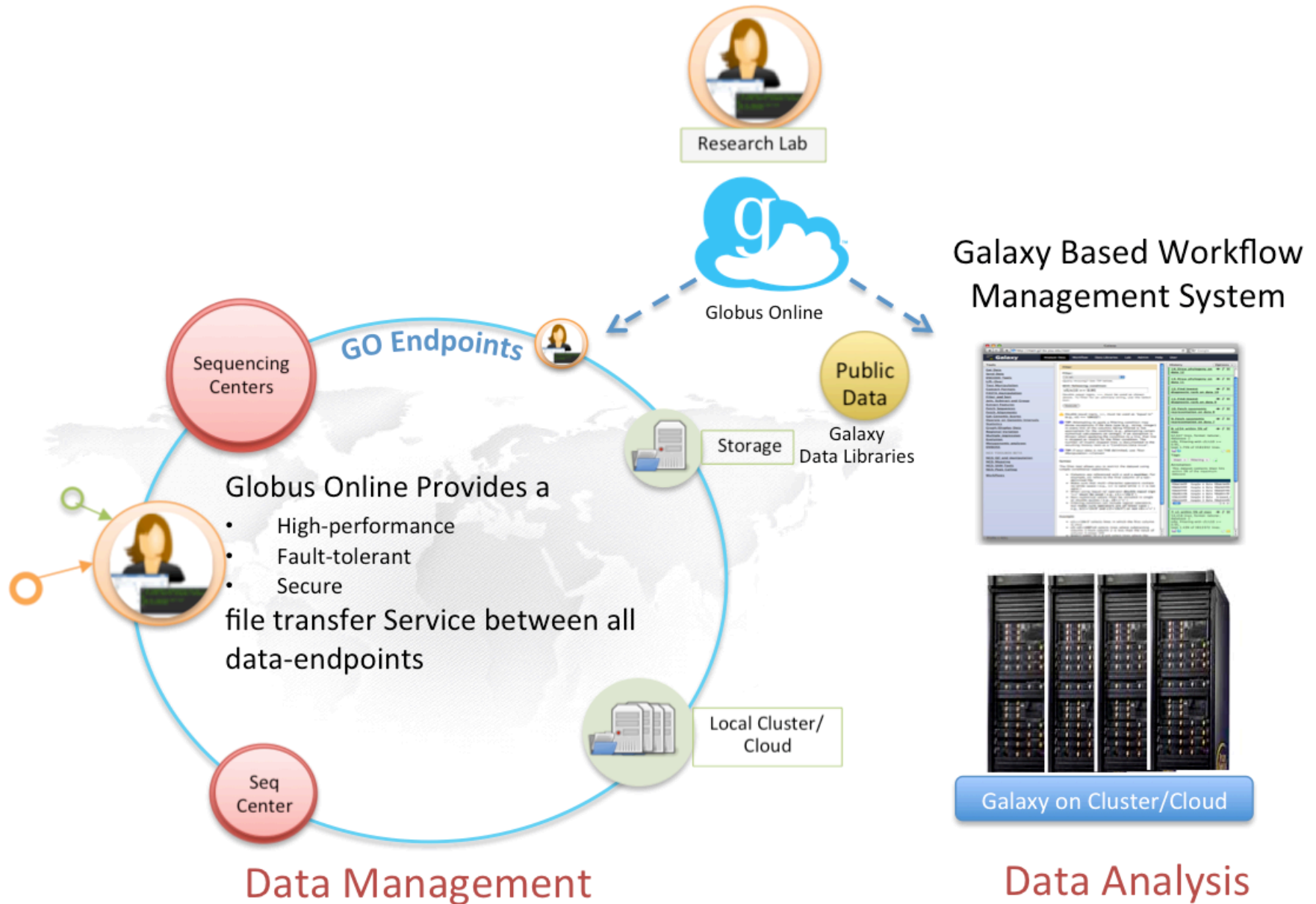Several Commercial Solutions: tested 5 – **No Scalability!!!**

Next Generation Sequencing Data:
New challenges for Translational Cancer Research

6 KEY ISSUES FOR NCI CANCER CLOUDS

The biomedical research community has identified six issues that need to be addressed in the NCI Cloud Initiative:

1. Data access
2. Computing capacity
3. Data interoperability
4. Training
5. Usability
6. Governance

Globus + Galaxy = Globus Genomics

Research Lab

Globus Online

Galaxy Based Workflow
Management System

GO Endpoints

Sequencing
Centers

Public
Data

Storage

Galaxy
Data Libraries

Globus Online Provides a
- High-performance
- Fault-tolerant
- Secure
file transfer Service between all
data-endpoints

Local Cluster/
Cloud

Seq
Center

Galaxy on Cluster/Cloud

Data Management

Data Analysis

Pilot Project: Georgetown ICBI and Globus Genomics Team at CI/U Chicago

- **Pilot Project Aims:**
  1. To manage and move NGS data on a cloud
  2. To bring Tools to Data
  3. To make it scalable for translational projects

NGS data management workflows
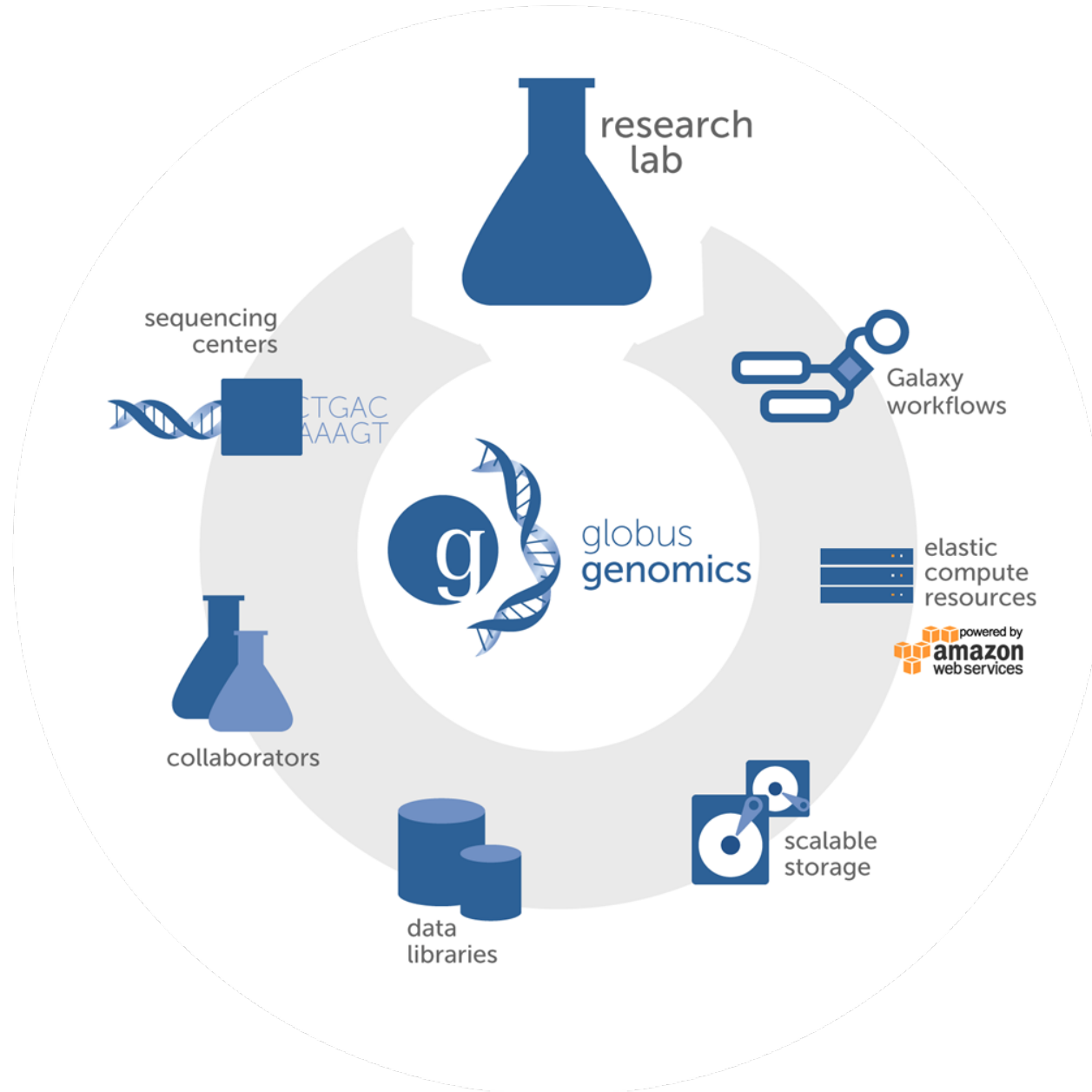
Main NGS workflows in use at ICBI:

ICBI Workflows

❖ Whole-genome and Exome Seq Workflow

➢ In collaboration, the Globus Genomics and ICBI teams, tested and benchmarked the analytical workflows

➢ Workflows include data transfer from data source to analysis platform using Globus Online.

❖ RNA-Seq Workflow

➢ Performed quality control for alignment of Transcriptome data.

➢ Includes multiple read filtering tools (Fastx toolkit, native Galaxy filtering tools) to achieve optimal alignment statistics.

➢ Includes comparing performance of Tophat2 and RSEM alignment.

Whole-Genome and Exome Analysis Workflows
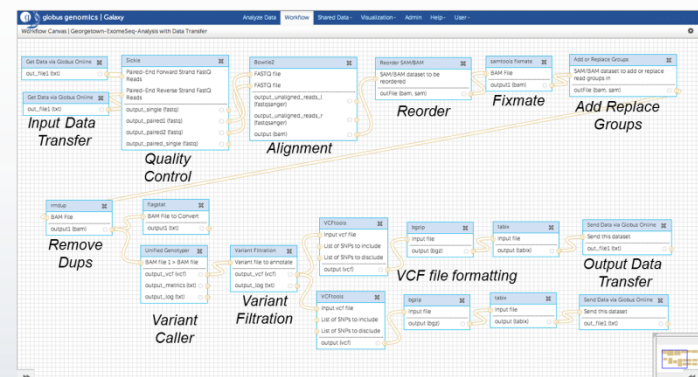
Alignment Quality Control for RNA-Seq Analysis
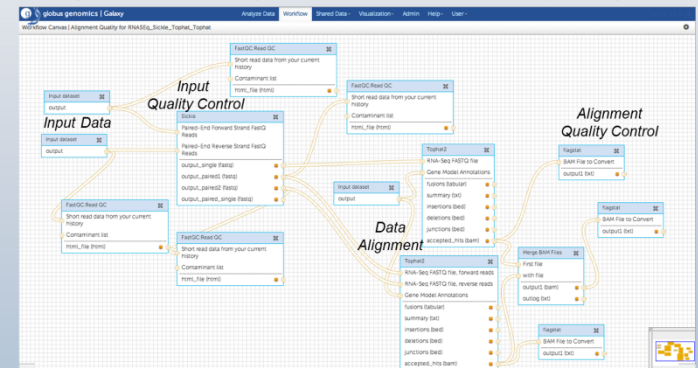
RNA-seq: Variants % Genomic Location

Legend:
- Exonic Total
- Intronic
- Intergenic
- UTR3
- UTR5
- Upstream
- Downstream
- Splicing
- ncRNA_exonic
- ncRNA_intronic
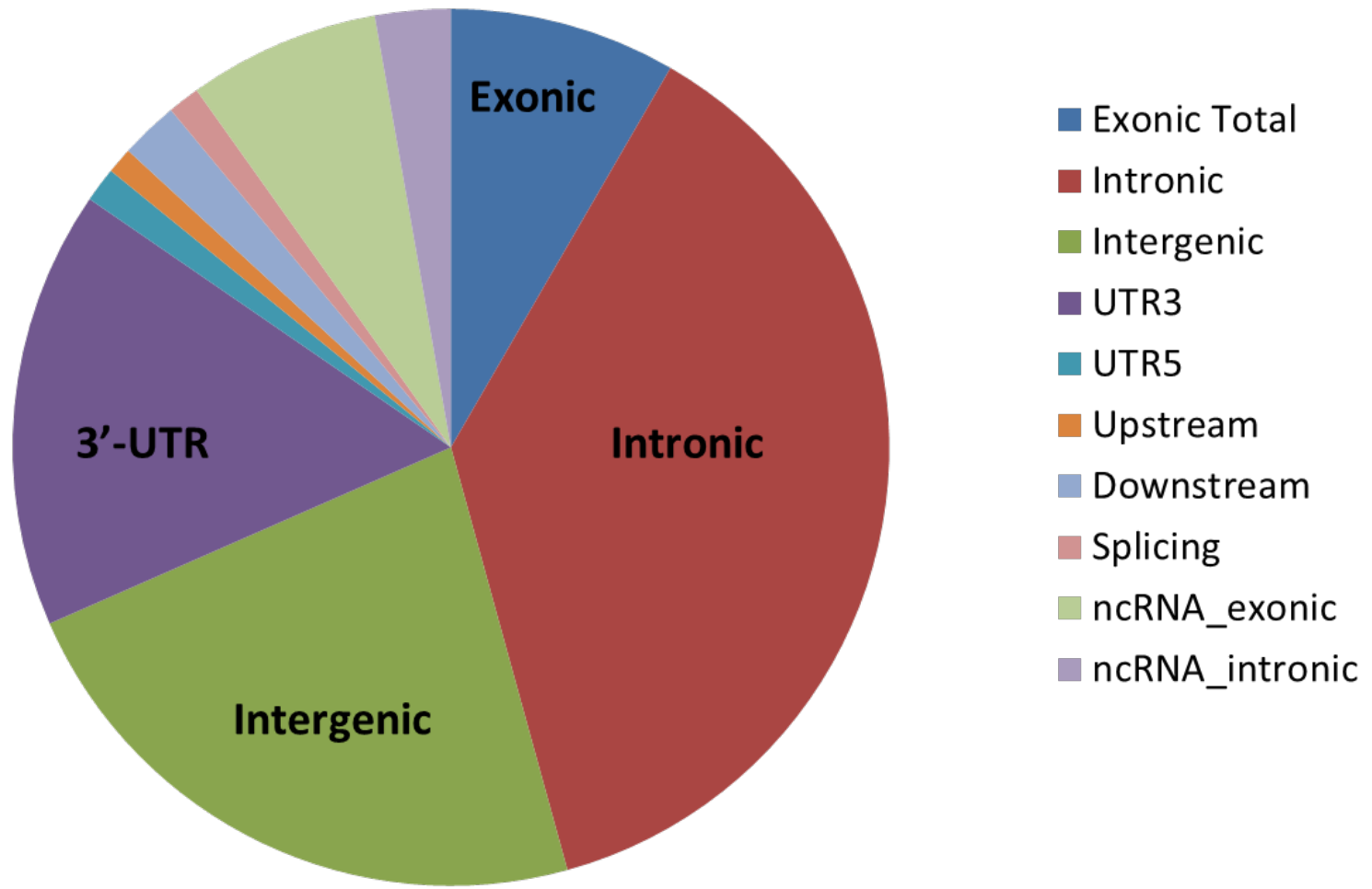
Implementation of ICBI workflows at Globus Genomics Instance

Modifying RNAseq workflows on a fly
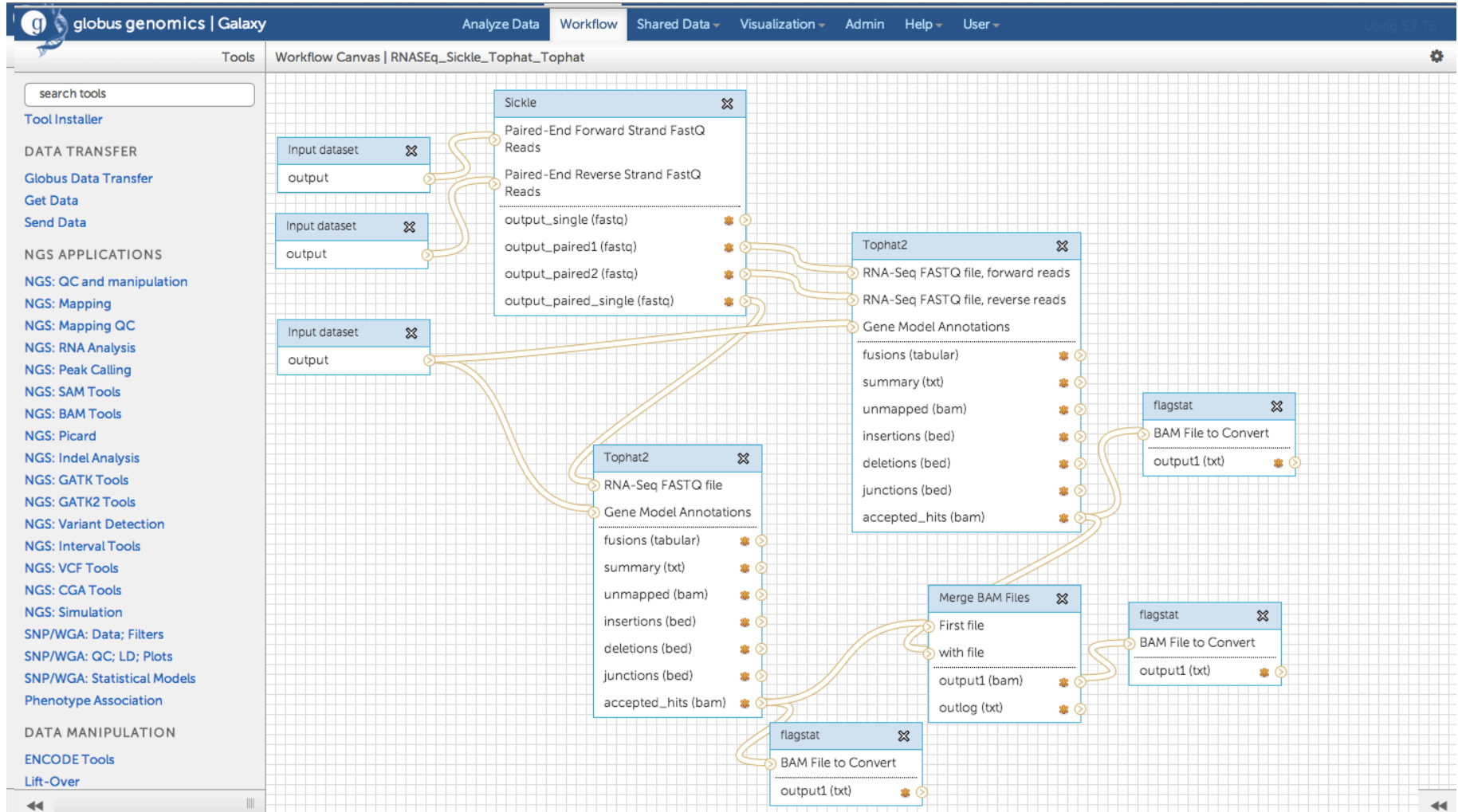
Batch Processing Example:
RNAseq from TCGA Ovarian Cancer Study: 21 samples
Traceable, Reproducible, Manageable

Summary of Results of Pilot Project:

Achievements

- Completed setup of Globus Online endpoints and validated data transfer capabilities
- Wrapped additional tools and validated execution of Whole Genome, Exome, and RNA-Seq pipelines utilizing Globus Genomics
- Ran all three targeted pipelines at scale against large data sets demonstrating significant speed-up of execution compared to serial approaches
- Optimized the Globus Genomics environment in AWS to efficiently handle burst requirements through elastic provisioning / de-provisioning of compute capacity
- Gathered performance and quality data associated with running all three pipelines at scale on the optimized Globus Genomics instance
- **Jointly prepared and presented several posters: ICBI Symposium 2013; NIH translational Genomics Symposium etc.**
- **Developed Platform to share & learn bioinformatics best practices and technical expertise**

In Progress: **Large batch processing benchmarks -real-life case studies**

- RNAseq: 21 samples from TCGA Ovarian Cancer Study
  - Much wider range of input file sizes: 5 to 25 Gb
  - Total time: upload and execution of RSEM pipeline:

    18 hrs (as long as it takes for the largest single sample)

- Exome seq: 78 samples from Lung Cancer Study (EBI)
  - Much larger file sizes –from 2 to 13 Gb
  - Bowtie2 pipeline: 10 hours

    (as long as it takes for the largest single sample)

Future Plans:

- Transition to Production Instance of GG:
  - Working with Genomics Core at Lombardi Cancer center to establish NGS bioinformatics services for Lombardi researchers

- Additional Pipelines Development - focusing on RNAseq: ncRNAseq; viral RNA seq etc.

- We are interested in:
  - NGS data publishing;
  - Adopting Globus Genomics for Education/Training:
    (Massive Data Initiative at Georgetown)

Acknowledgements:

- CI/Argonne/U Chicago Team:
  - Ian Foster, Utpal Davé, Ravi Madduri, Dinanath Sulakhe, Alex Rodriguez, Lukasz Lacinski
- ICBI/Georgetown Team:
  - Subha Madhavan, Michael Harris, Yuriy Gusev

    Krithika Bhuvaneshwar , Lei Song, Robinder Gauba

THANKS!