

Globus Genomics @ Georgetown

Yuriy Gusev - ICBI, Georgetown University, Washington DC



THE UNIVERSITY OF
CHICAGO



INNOVATION CENTER FOR BIOMEDICAL INFORMATICS

Georgetown | Lombardi
COMPREHENSIVE CANCER CENTER

Innovation Center for Biomedical Informatics

<http://icbi.georgetown.edu/>



Innovation Center for Biomedical Informatics

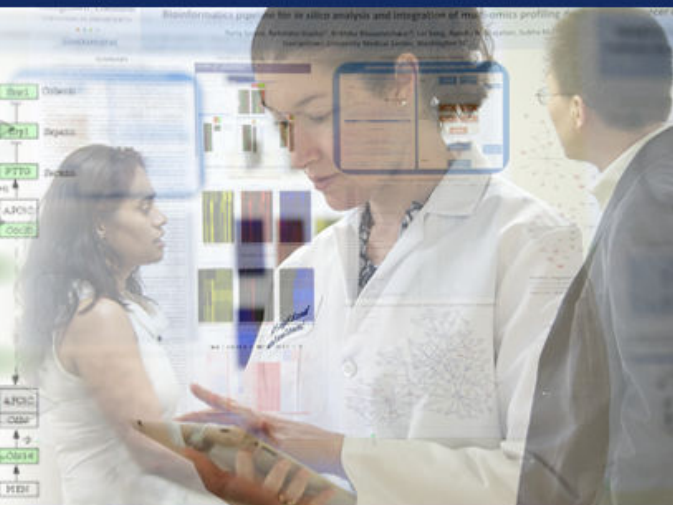


[HOME](#) | [SITEMAP](#) | [CONTACT](#)

[About Us](#) [People](#) [Biomedical Research](#) [Technology](#) [Education & Training](#) [Our Impact](#) [News & Events](#) [Careers](#)

Cutting-edge Research in Biomedical Informatics

The Innovation Center for Biomedical Informatics (ICBI) at GUMC is an academic hub for cutting-edge research in biomedical informatics that will enhance basic and translational research and educate the next generation of scientists and physicians on integrating informatics into biomedical research and clinical practice.



COLLABORATE WITH Us



We welcome you to collaborate with us in the testing and development of a new generation of data management, mining, and analysis tools to help with the "Big Data" challenge facing the genomic scientists today.

SCIENTIST Spotlight



Difei Wang, PhD
Assistant Professor

Dr. Wang joined ICBI in June 2012. His primary research areas of interest include computer aided drug design, pharmacogenomics, next generation sequencing, nucleosome positioning, and nucleosome and

ICBI Blog

Genomes on Cloud 9
posted by Subha Madhavan on 2014-01-12

Keynote Talks at ICBI symposium: Stephen Friend and Eric Hoffman
posted by Laura Sheahan on 2013-10-24

Poster Winners!
posted by Laura Sheahan on 2013-10-24

[>> all blog articles](#)

Innovation Center for Biomedical Informatics: Enabling Translational Genomics Research



Innovation Center for Biomedical Informatics



[HOME](#) | [SITEMAP](#) | [CONTACT](#)

[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](#)

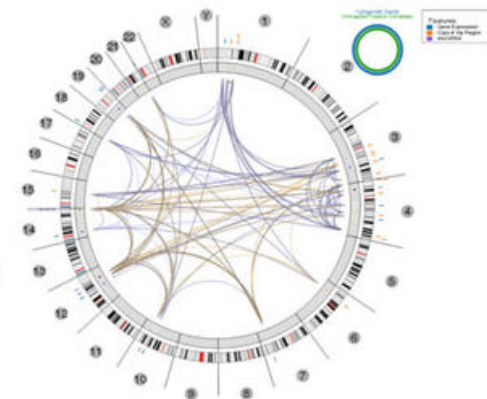
[Oncology Biomarkers & Drug Response](#)

[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



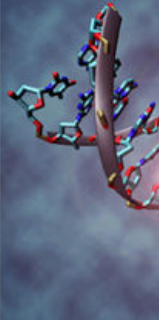
Lombardi Comprehensive Cancer Center
at Georgetown University

[Log In](#)
[register now](#) | [forgot password](#)








Overview
Search
Analyze
My G-DOC

WELCOME











The Georgetown Database of Cancer (G-DOC) is a cutting-edge data integration platform and knowledge discovery system for the oncology and translational research communities. G-DOC users can access public and proprietary clinical and -omics data aggregated from across the Medical Center, along with a comprehensive set of advanced analysis and visualization tools, to generate and test hypotheses across biomedical disciplines.



Data-Type Overview

Data Type	Study Count
PATIENT 	35
microRNA 	8
GENE EXPRESSION 	41
CELL LINE 	9
WGS 	1
METABOLOMICS 	4
COPY NUMBER 	8

Cancer/Study Overview

Disease	Study Count	Patient Count	Biospecimen Count	Available Data Types
BREAST CANCER	27	3653	4566	      
COLON	10	702	1365	  

FINDINGS

NEWS

PUBLICATIONS

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!!!**

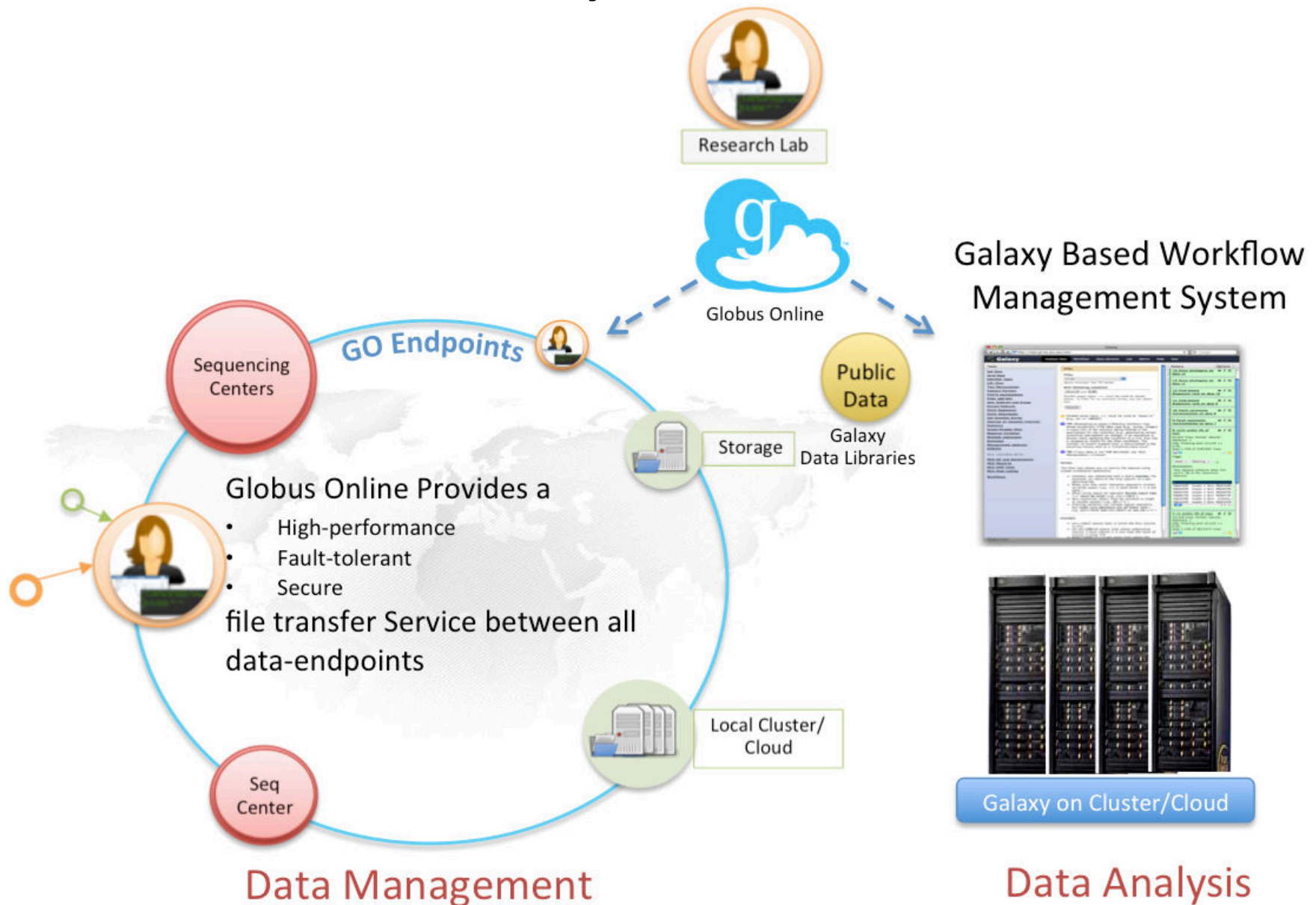
Next Generation Sequencing Data: New challenges for Translational Cancer Research

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

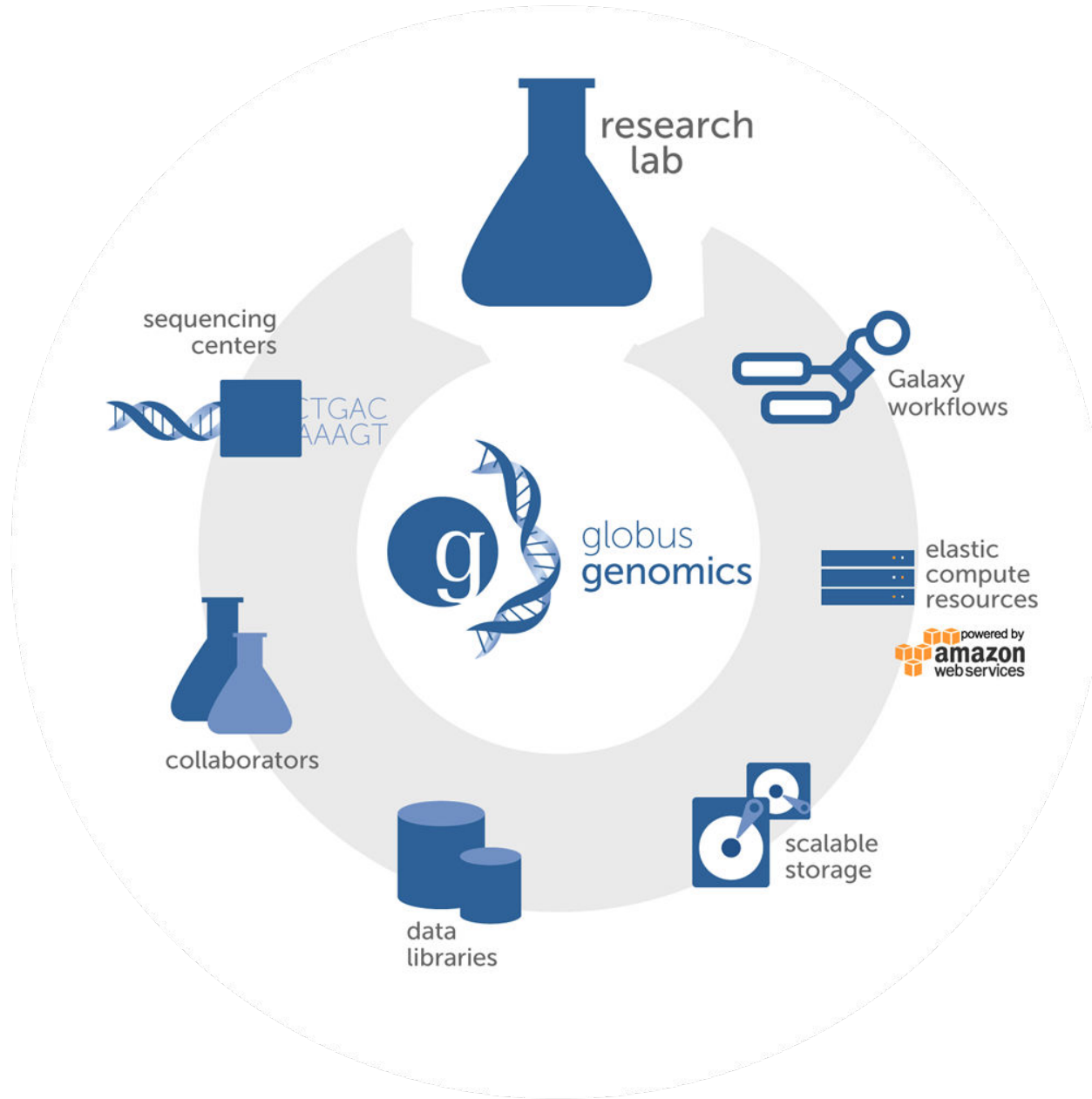
Globus + Galaxy = Globus Genomics



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

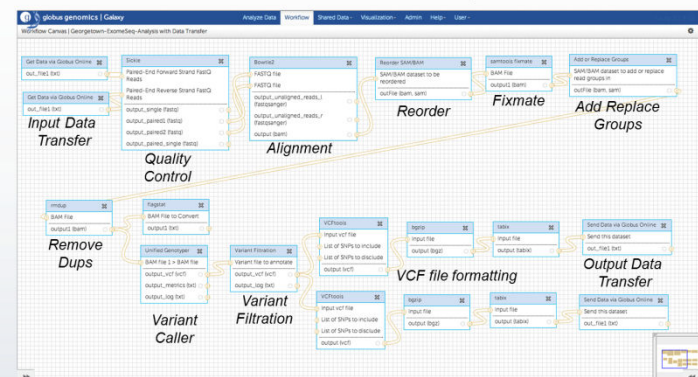
❖ Whole-genome and Exome Seq Workflow

- 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.

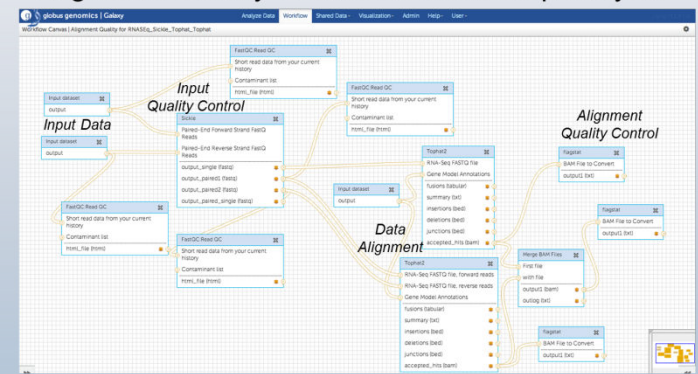
❖ RNA-Seq Workflow

- 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.

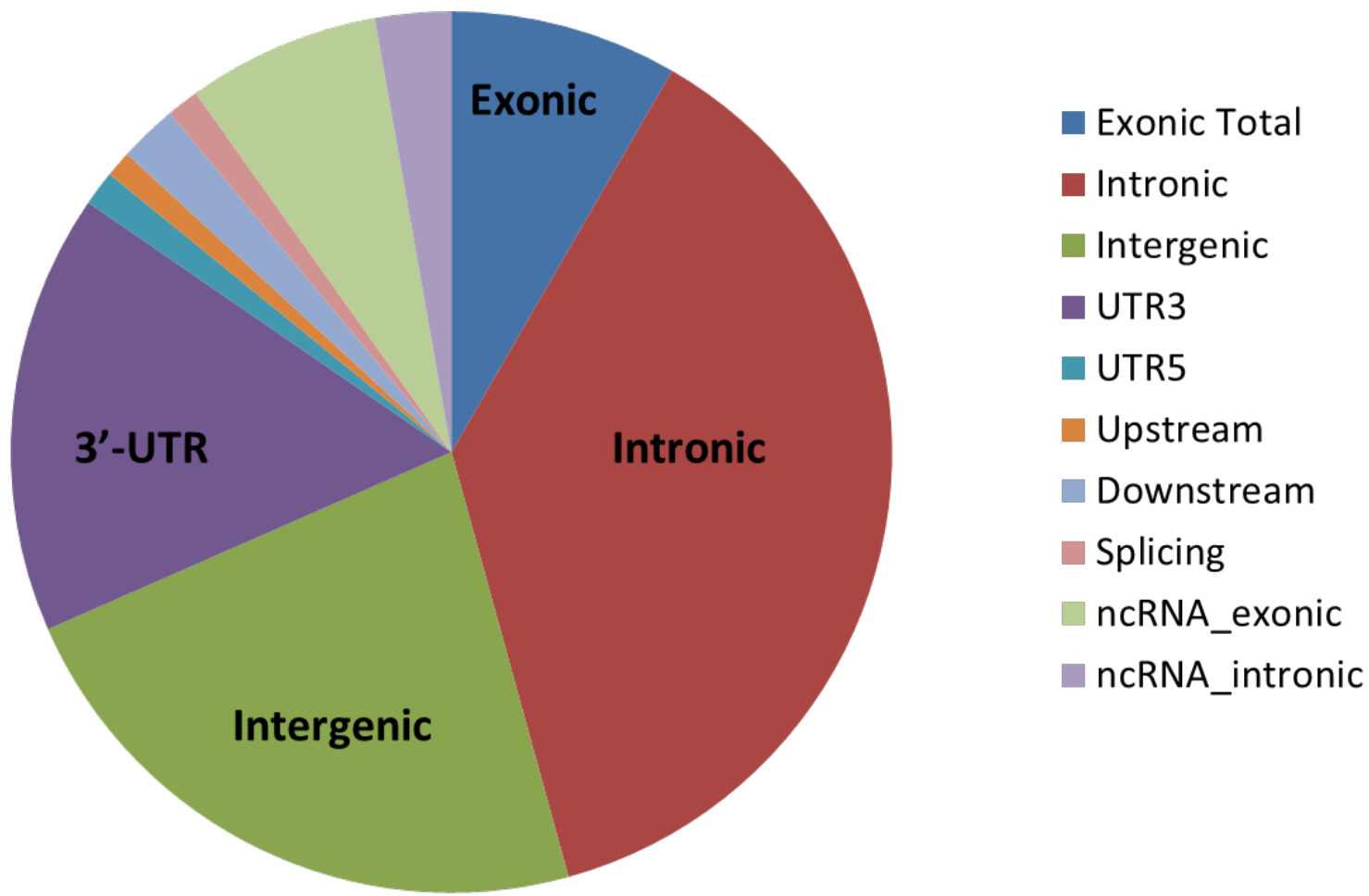
Whole-Genome and Exome Analysis Workflows



Alignment Quality Control for RNA-Seq Analysis



RNA-seq: Variants % Genomic Location



Implementation of ICBI workflows at Globus Genomics Instance

globus genomics | Galaxy

[Analyze Data](#)
[Workflow](#)
[Shared Data](#)
[Visualization](#)
[Admin](#)
[Help](#)
[User](#)

Using 5.7 TB

Tools

[Tool Installer](#)

DATA TRANSFER
[Globus Data Transfer](#)
[Get Data](#)
[Send Data](#)

NGS APPLICATIONS
[NGS: QC and manipulation](#)
[NGS: Mapping](#)
[NGS: Mapping QC](#)
[NGS: RNA Analysis](#)
[NGS: Peak Calling](#)
[NGS: SAM Tools](#)
[NGS: BAM Tools](#)
[NGS: Picard](#)
[NGS: Indel Analysis](#)
[NGS: GATK Tools](#)
[NGS: GATK2 Tools](#)
[NGS: Variant Detection](#)
[NGS: Interval Tools](#)
[NGS: VCF Tools](#)
[NGS: CGA Tools](#)
[NGS: Simulation](#)
[SNP/WGA: Data; Filters](#)
[SNP/WGA: QC; LD; Plots](#)
[SNP/WGA: Statistical Models](#)
[Phenotype Association](#)

DATA MANIPULATION
[ENCODE Tools](#)
[Lift-Over](#)

system status

GET STARTED

[Workflow for Illumina RNA-seq »](#)

Provide information on differential gene expression between NGS samples including alleles and spliced transcripts. This analysis is for paired-end sequences. Includes QC, mapping to hg19 and expression of genes.

[Workflow for Illumina Exome-seq »](#)

This analysis is an efficient strategy to selectively sequence the coding regions of the genome. The goal of this approach is to identify the functional variations in the exome regions. Analysis for paired-end sequences. Includes QC, mapping to hg19 and variants list.

[Workflow for Illumina ChIP-seq »](#)

ChIP-seq combines chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing to identify the binding sites of DNA-associated proteins. It can be used to map global binding sites precisely for any protein of interest. Analysis includes QC, mapping to hg19 and identification of peaks.

Your History

2014Apr3.WES~ERR166339~Gerogetown-ExomeSeq-Analysis_input_output~Thu_Apr_03_2014_-1:40:38_PM
98.0 GB

27: tabix on data 24

26: tabix on data 22

25: snps

24: bgzip on data 21

23: indels

22: bgzip on data 20

21: VCFtools on data 18

20: VCFtools on data 18

19: Variant Filtration on data 14 (log)

18: Variant Filtration on data 14 (Variant File)

17: Send: ERR166339_flagstat.txt

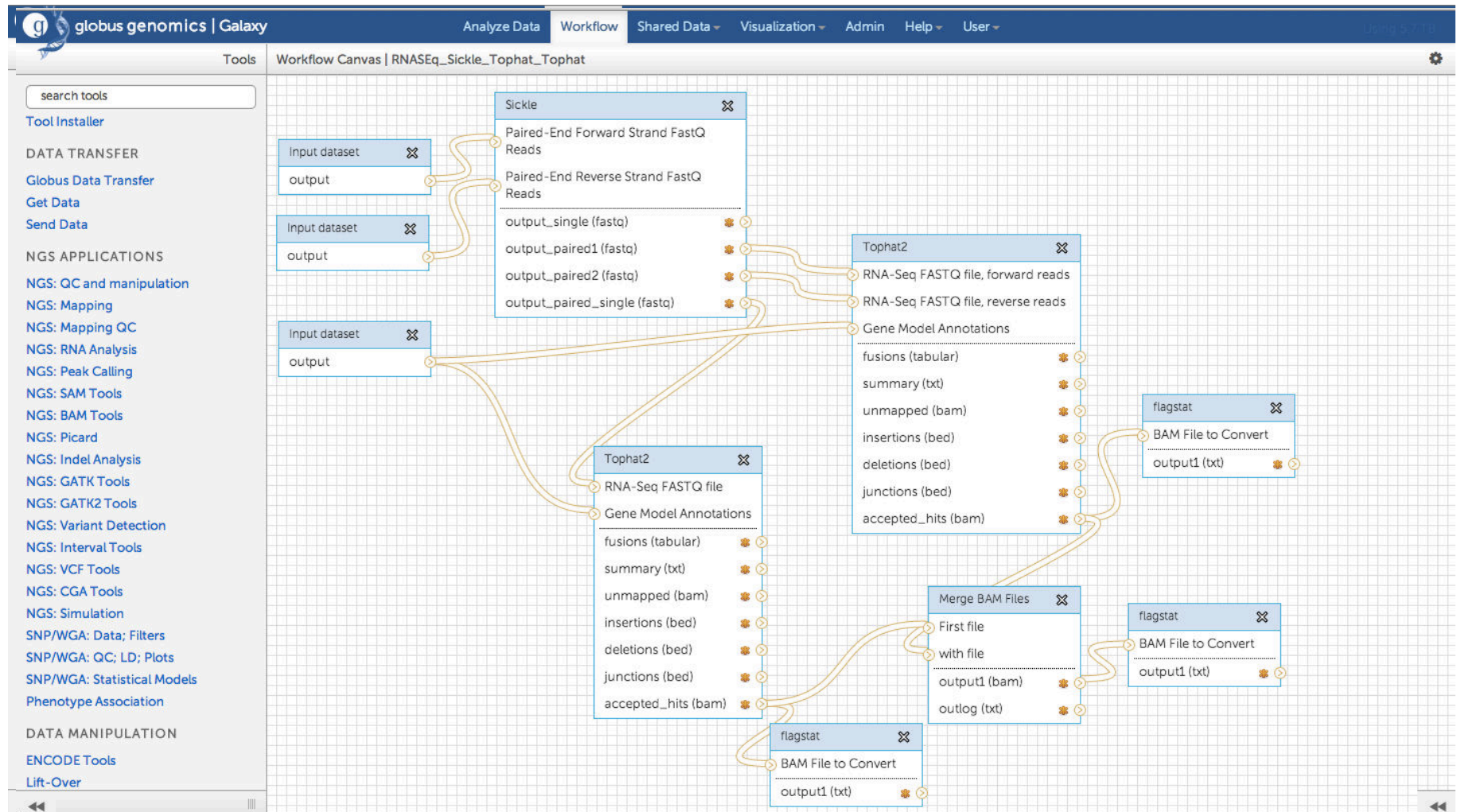
16: Unified Genotyper on data 12 (log)

15: Unified Genotyper on data 12 (metrics)

14: Unified Genotyper on data 12 (VCF)

13: flagstat on data 12

Modifying RNAseq workflows on a fly



Batch Processing Example:

RNAseq from TCGA Ovarian Cancer Study: 21 samples

Traceable, Reproducible, Manageable

globus genomics | Galaxy

Analyze Data Workflow Shared Data Visualization Admin Help User Using 5.7 TB

Tools

search tools

Tool Installer

DATA TRANSFER

Globus Data Transfer

Get Data

Send Data

NGS APPLICATIONS

NGS: QC and manipulation

NGS: Mapping

NGS: Mapping QC

NGS: RNA Analysis

NGS: Peak Calling

NGS: SAM Tools

NGS: BAM Tools

NGS: Picard

NGS: Indel Analysis

NGS: GATK Tools

NGS: GATK2 Tools

NGS: Variant Detection

NGS: Interval Tools

NGS: VCF Tools

NGS: CGA Tools

NGS: Simulation

SNP/WGA: Data; Filters

SNP/WGA: QC; LD; Plots

SNP/WGA: Statistical Models

Phenotype Association

DATA MANIPULATION

ENCODE Tools

Lift-Over

Workflow Name	Status	Tags	Size	Date
2014Feb25TCGAOV~C09BFACXX_2_ACAAC~RNASeq.workflow~Tue_Feb_25_2014_-l:47:30_AM	33	2	0 Tags	138.4 GB
2014Feb25TCGAOV~D0DWTACXX_8_ATACGG~RNASeq.workflow~Tue_Feb_25_2014_-l:06:05_AM	31	0 Tags	100.9 GB	Feb 25, 2014
2014Feb25TCGAOV~D0DWTACXX_8_ATAATT~RNASeq.workflow~Tue_Feb_25_2014_-l:05:08_AM	29	2	0 Tags	159.7 GB
2014Feb25TCGAOV~D0DWTACXX_7_AGATAG~RNASeq.workflow~Tue_Feb_25_2014_-l:04:08_AM	31	0 Tags	67.1 GB	Feb 25, 2014
2014Feb25TCGAOV~D0DWTACXX_6_ACCGGC~RNASeq.workflow~Tue_Feb_25_2014_-l:03:06_AM	31	0 Tags	77.7 GB	Feb 25, 2014
2014Feb25TCGAOV~D0DWTACXX_6_ACCCAG~RNASeq.workflow~Tue_Feb_25_2014_-l:02:06_AM	31	0 Tags	210.1 GB	Feb 25, 2014
2014Feb25TCGAOV~D0DWTACXX_5_AAGGAC~RNASeq.workflow~Tue_Feb_25_2014_-l:01:03_AM	31	0 Tags	86.7 GB	Feb 25, 2014
2014Feb25TCGAOV~D0DWTACXX_4_AAGCGA~RNASeq.workflow~Tue_Feb_25_2014_-l:59:56_AM	31	0 Tags	85.1 GB	Feb 25, 2014
2014Feb25TCGAOV~D0DWTACXX_4_AAGACT~RNASeq.workflow~Tue_Feb_25_2014_-l:58:36_AM	29	2	0 Tags	175.5 GB
2014Feb25TCGAOV~D0DWTACXX_3_AAAGCA~RNASeq.workflow~Tue_Feb_25_2014_-l:57:29_AM	31	0 Tags	81.3 GB	Feb 25, 2014
2014Feb25TCGAOV~D0DWTACXX_3_AAACAT~RNASeq.workflow~Tue_Feb_25_2014_-l:56:28_AM	29	2	0 Tags	164.3 GB
2014Feb25TCGAOV~D0DWEACXX_4_ATCTAT~RNASeq.workflow~Tue_Feb_25_2014_-l:55:25_AM	30	1	0 Tags	141.0 GB
2014Feb25TCGAOV~D0DWEACXX_4_ATCTCTA~RNASeq.workflow~Tue_Feb_25_2014_-l:54:27_AM	31	0 Tags	137.9 GB	Feb 25, 2014
2014Feb25TCGAOV~D0DWEACXX_1_GGCTAC~RNASeq.workflow~Tue_Feb_25_2014_-l:53:25_AM	31	0 Tags	68.0 GB	Feb 25, 2014
2014Feb25TCGAOV~D0DWEACXX_1_CTTGTA~RNASeq.workflow~Tue_Feb_25_2014_-l:52:25_AM	31	0 Tags	227.6 GB	Feb 25, 2014

georgetown.globusgenomics.org/history/list?sharing=All&sort=-update_time&f-name=All&f-tags=All&f-deleted=False&operation=Switch&use_panels=False&id=cab7881291755468

Your History

2014Feb25TCGAOV~C09BFACXX_2_ACAAC~RNASeq.workflow~Tue_Feb_25_2014_-l:47:30_AM

138.4 GB

35: Send: C09BFACXX_2_ACAAC_TxBamStat.txt

34: Send: C09BFACXX_2_ACAAC_GenBamStat.txt

33: tabix on data 31

32: tabix on data 30

31: bgzip on data 29

30: bgzip on data 28

29: removeindels

28: keepindels

27: send.vcf

26: Variant Filtration on data 22 (log)

25: Variant Filtration on data 22 (Variant File)

24: Unified Genotyper on data 20 (log)

23: Unified Genotyper on data 20 (metrics)

22: Unified Genotyper on data 20 (VCF)

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!