GDSAP- A Galaxy-based platform for large-scale genomics analysis

Tin-Lap, LEE
School of Biomedical Sciences,
CUHK-BGI Innovation Institute of Trans-omics,
The Chinese University of Hong Kong,
Hong Kong SAR, China.

CBIIT

香港中文大學－華大基因研究所
跨組學創新研究院
CUHK-BGI Innovation Institute of Trans-Omics

- Jointly established between The Chinese University of Hong Kong (CUHK) and BGI.
- *"We aim to provide a platform conducive to training of multi-disciplinary talents conversant with the knowledge and application of genomics, proteomics, genetics , computation biology and bioinformatics, by capitalizing on both institutions' expertise and strengths in genomic science."*

Genomic Data Submission and Analytical Platform(GDSAP)

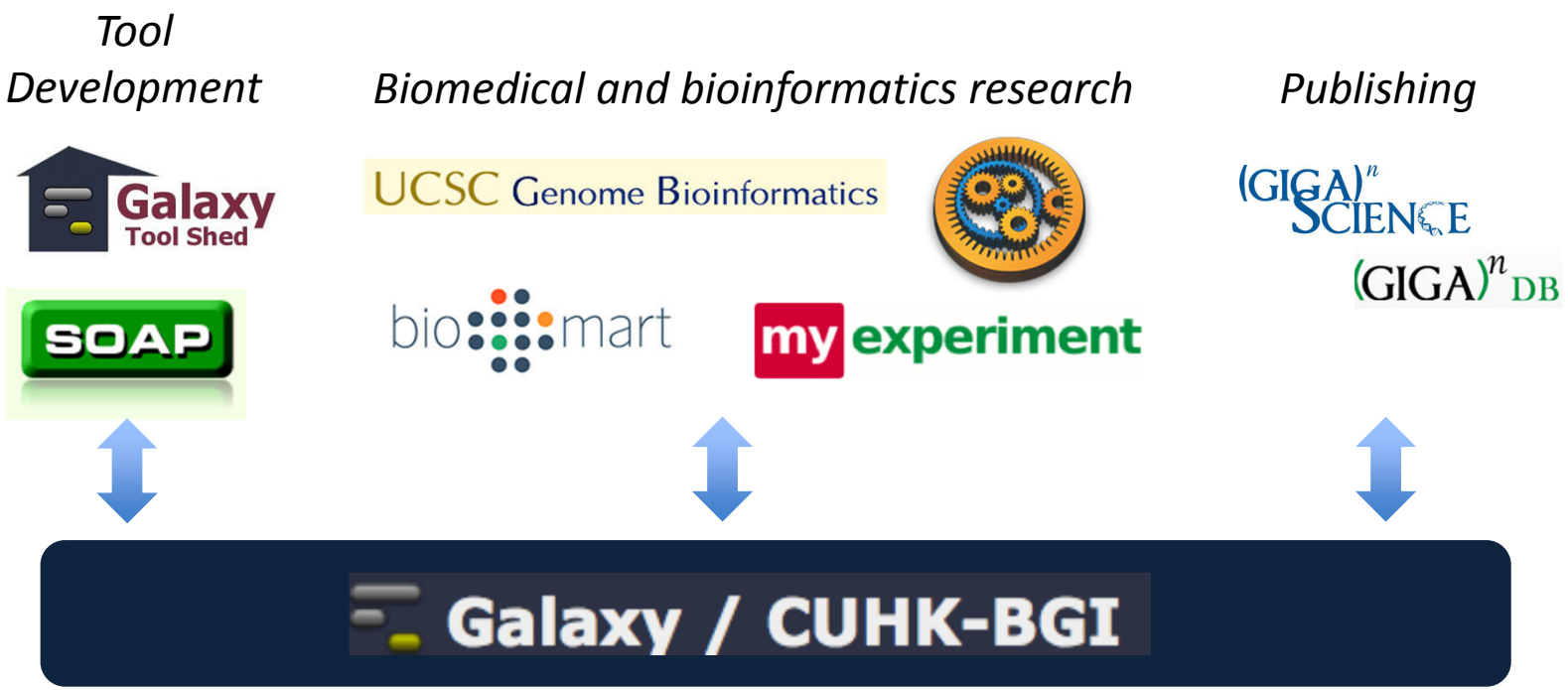
Objectives:

- Provides enhanced functionality in additional to the original Galaxy functions:
  - Customized public instances.
  - Seamless integration with SBS-UCSC genome database mirror and MyExperiemnt workflow environment.
  - Exchange and publish data through *GigaSciences* journal portal.

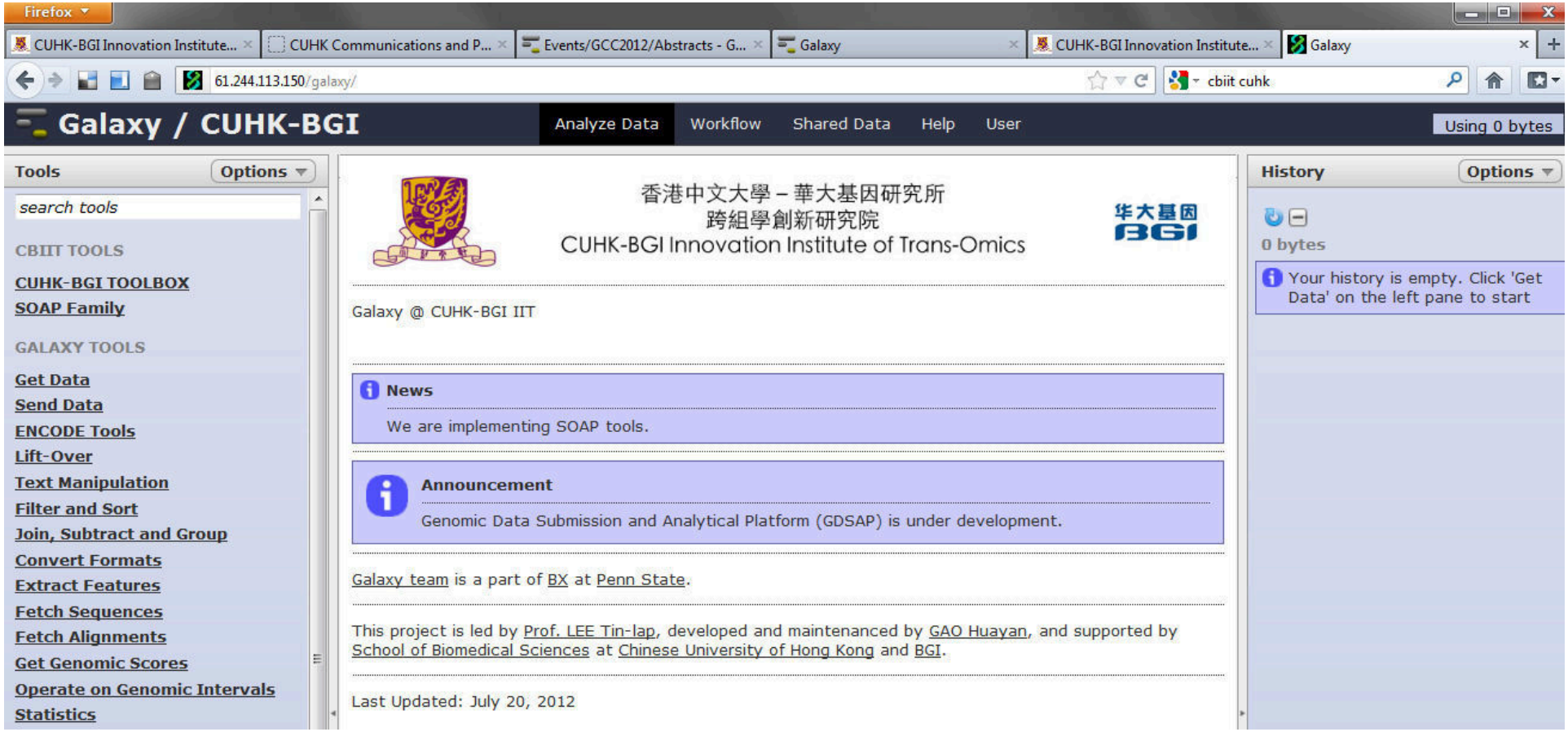
Outcomes:

- Simplies complicated bioinformatics tasks, accelerate data processing and allow flexible analysis.
- Significantly reduce software and hardware costs, encourage research collaboration.

GDSAP Structure

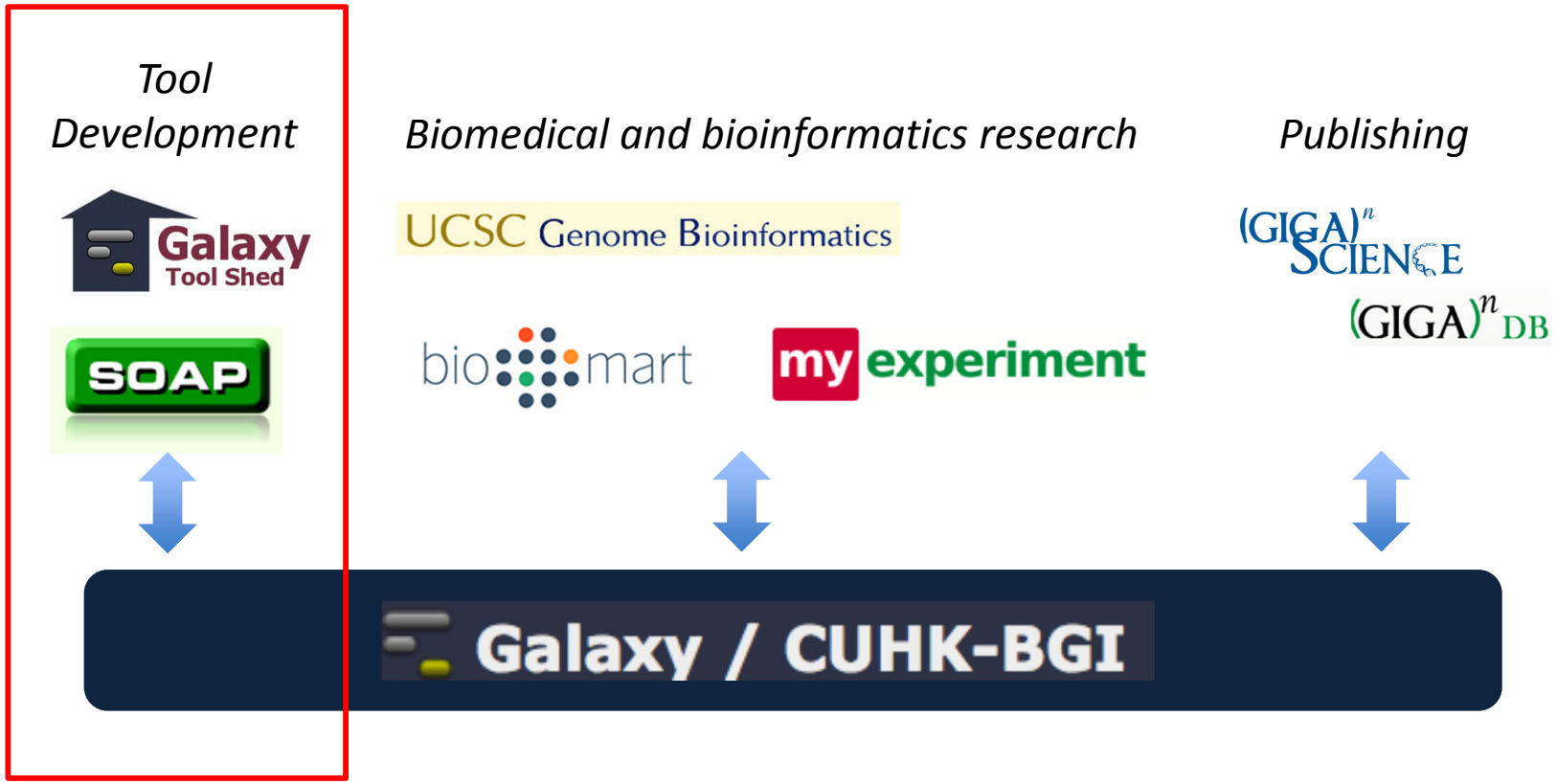


Galaxy/CUHK-BGI



http://www.cuhk.edu.hk/cbiit/galaxy.html

GDSAP Structure



What is SOAP?

- **SOAP** - a tool package that provides full solution to NGS data analysis by BGI.

Software	
SOAP3/GPU	SOAP3 is a GPU-based software for aligning short reads with a reference sequence. It can find all alignments with k mismatches, where k is chosen from 0 to 3. When compared with its previous version SOAP2, SOAP3 can be up to tens of times faster.
SOAPaligner/soap2	SOAPaligner/soap2 is a program for faster and efficient alignment for short oligonucleotide onto reference sequences. SOAPaligner/soap2 is compatible with numerous applications, including single-read or pair-end resequencing.
SOAPsplice NEW	SOAPsplice is designed to use RNA-Seq reads for genome-wide ab initio detection of splice junction sites and identification of alternative splicing (AS) events.
SOAPsnp	SOAPsnp is an accurate consensus sequence builder based on soap1 and SOAPaligner/soap2's alignment output. It calculates a quality score for each consensus base, which can be used for any latter process to call SNPs.
SOAPdenovo	SOAPdenovo, a short read de novo assembly tool, is a package for assembling short oligonucleotide into contigs and scaffolds.
SOAPindel	SOAPindel is developed to find the insertion and deletion specially for re-sequence technology.
SOAPsv	SOAPsv is a program for detecting the structural variation .
SOAP v1	SOAP v1 is available all the same.

Why SOAP?

- Galaxy has been using SAMtools for consensus sequence calling, but the recent upgrade has left this part out, which is very limited to some biologists.

- SOAPsnp is the only other method that can call full consensus sequences besides SAMtools.

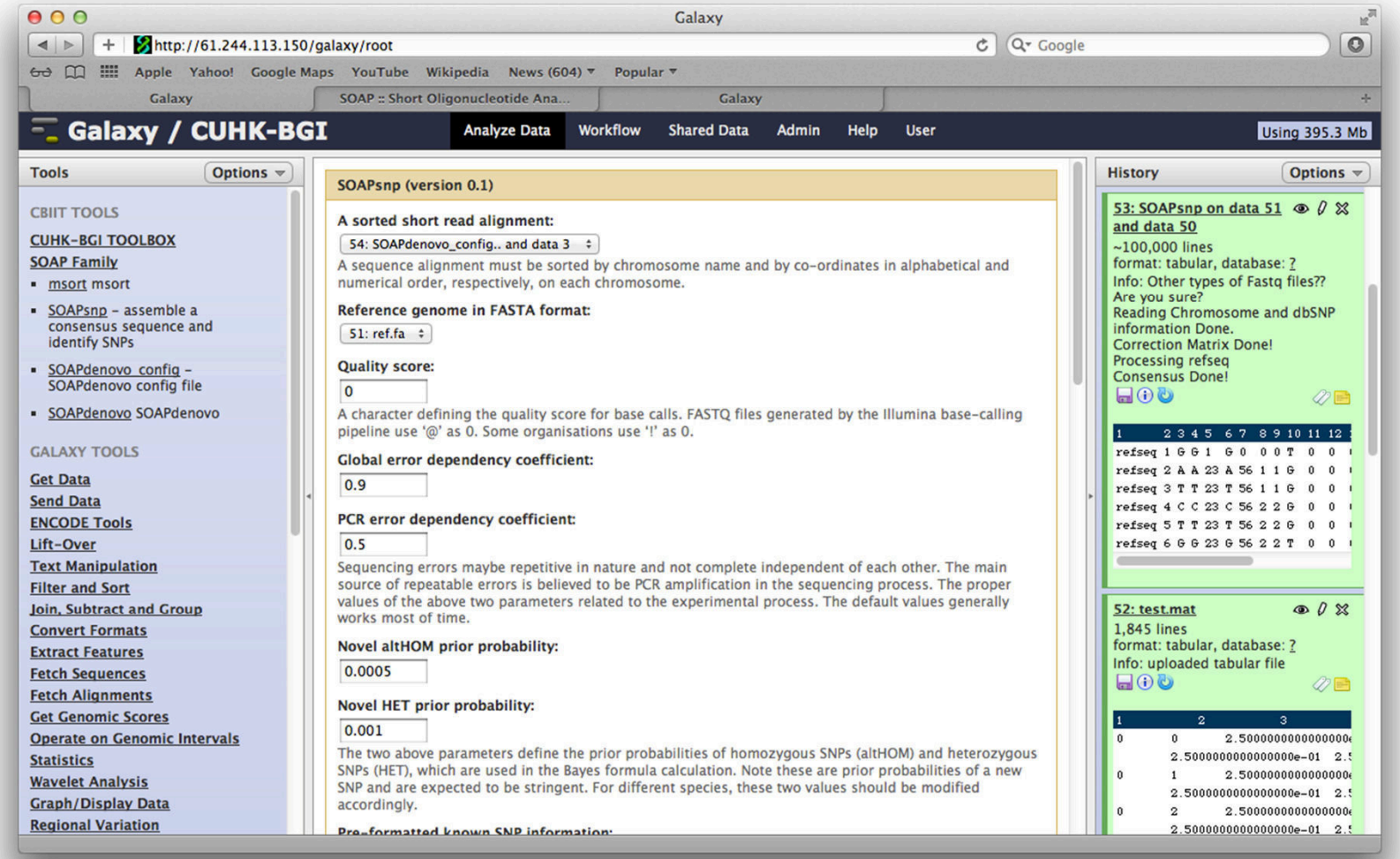
- The main galaxy site supports none of the SOAP tools, including SOAPsnp.

Galaxy Tool Shed

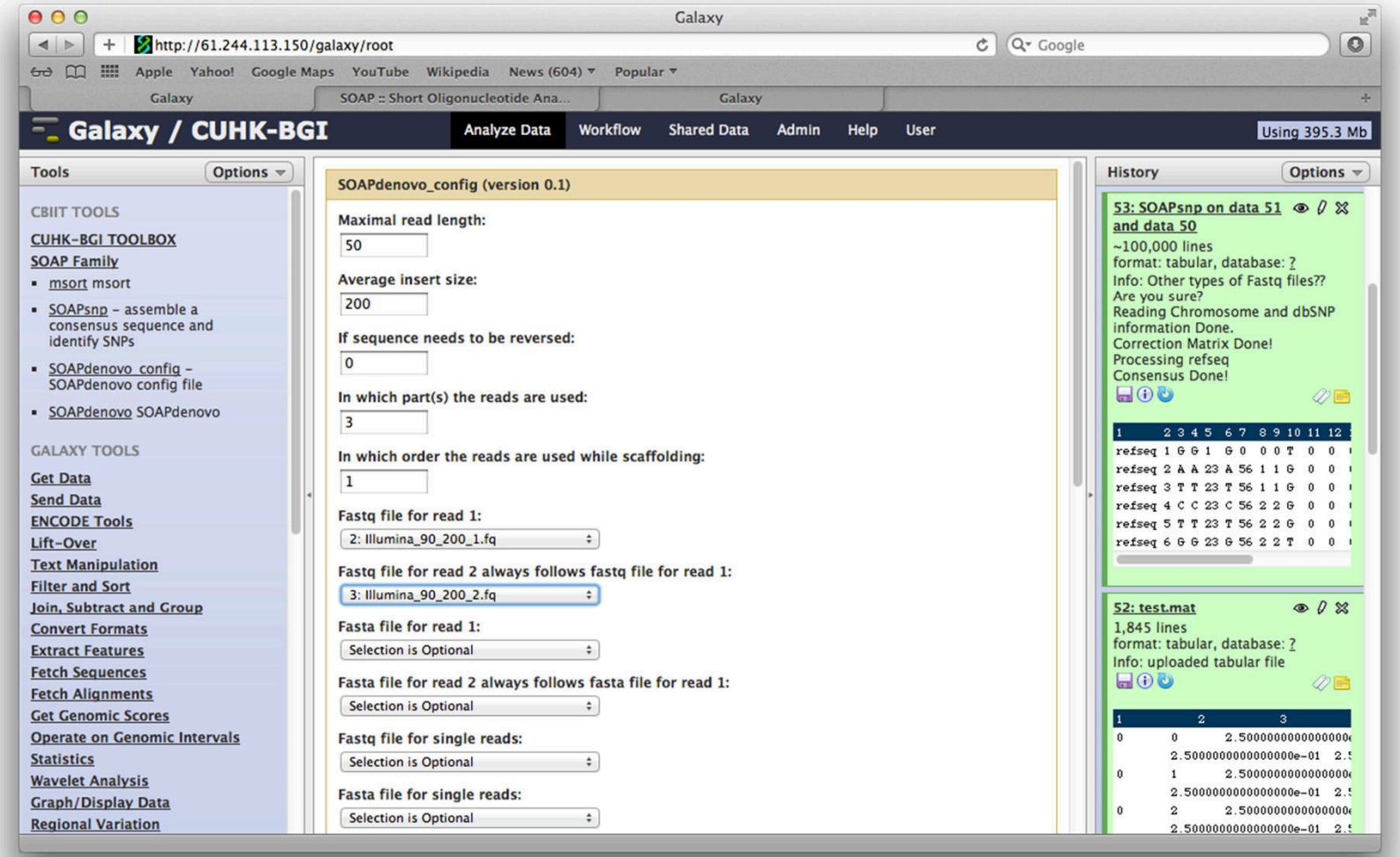
- Enables sharing of Galaxy tools across Galaxy servers around the world.

- SOAP package tools configured for use in Galaxy.
  - SOAPsnp/SOAPdenovo

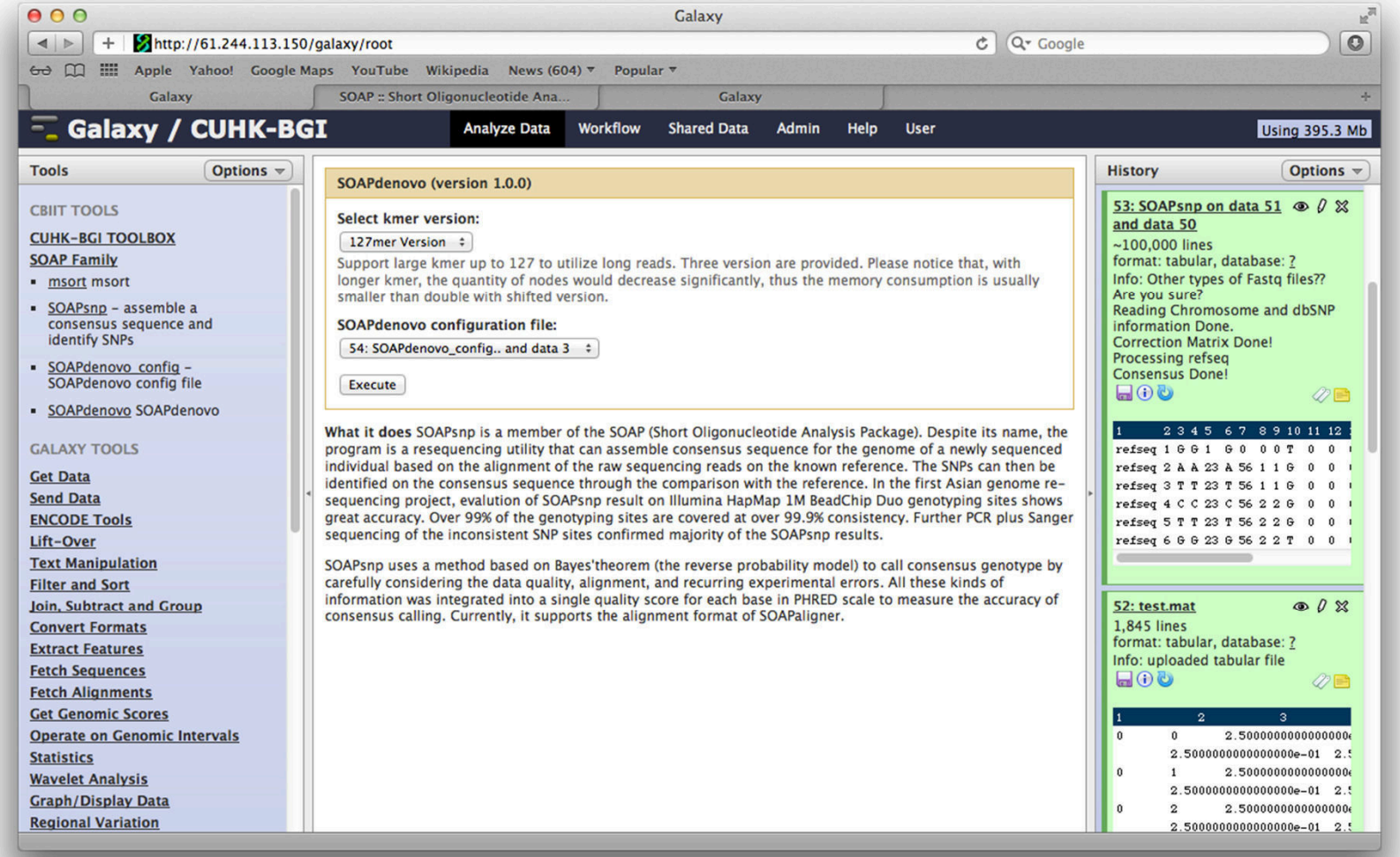
Implement: SOAPsnp



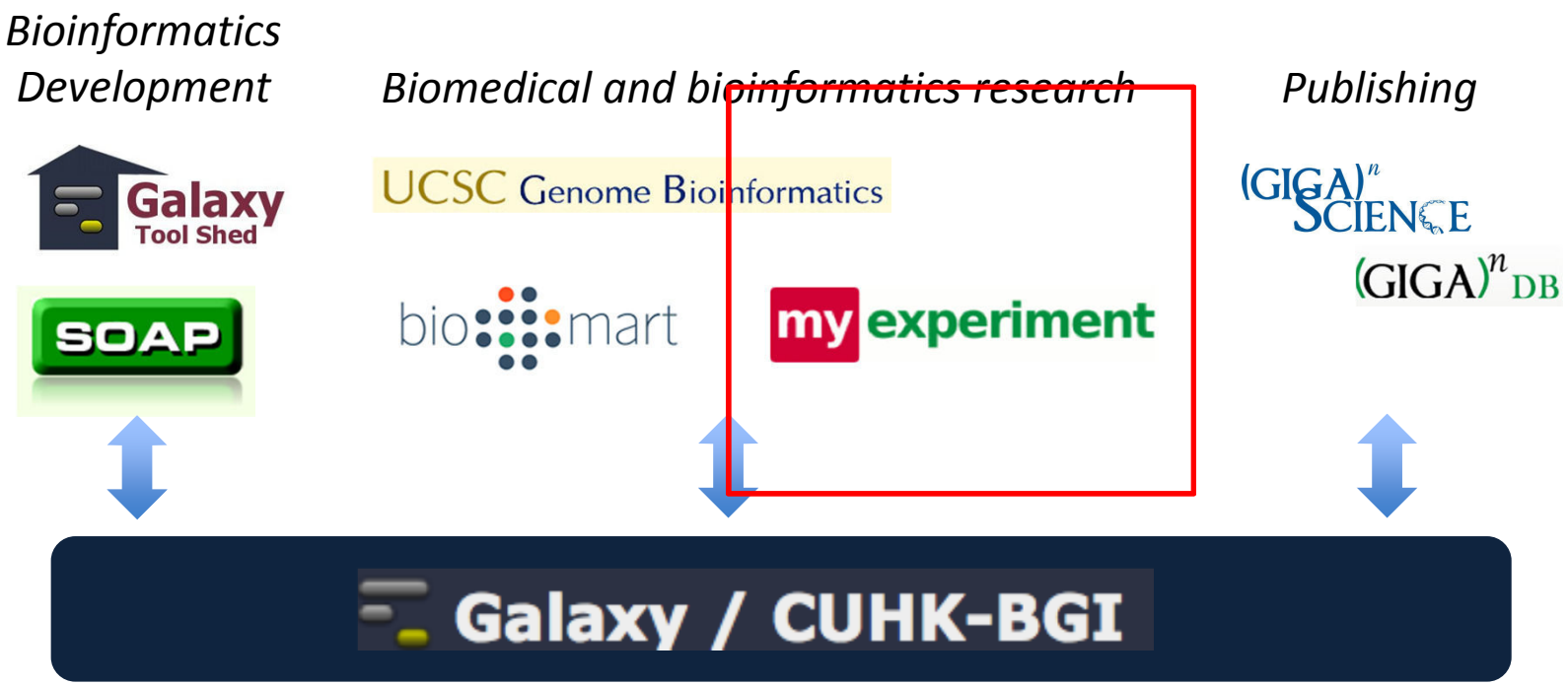
Implement: SOAPdenovo configuration file



Implement: SOAPdenovo



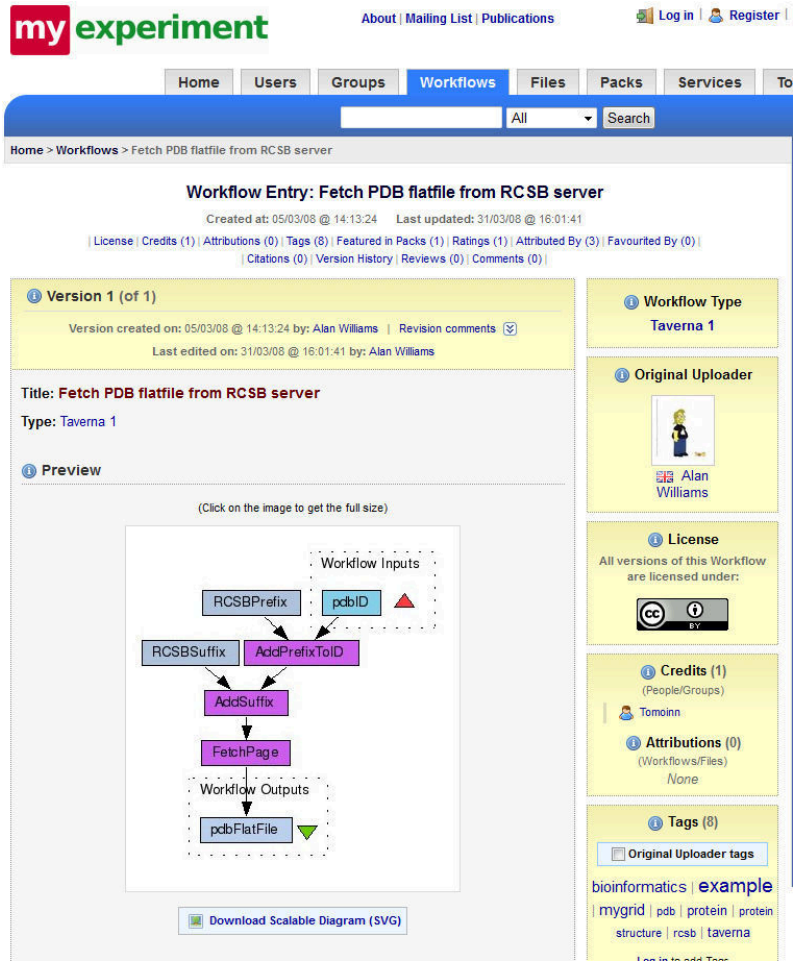
GDSAP structure

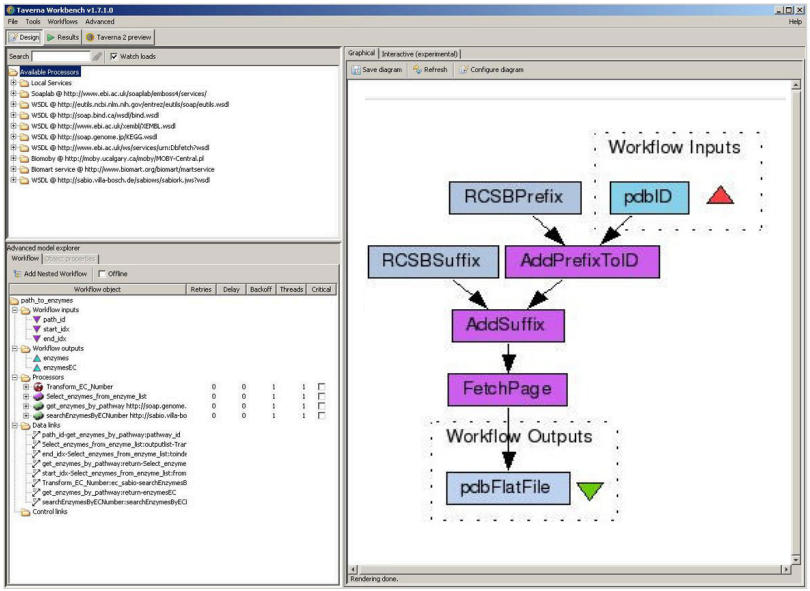


How does it work?

- MyExperiment works as a repository for workflows.
- Taverna workflows.
- New: Galaxy workflows.
- GDSAP integration

Taverna workflow





Taverna 1

Fetch PDB flatfile from RCSB server (v1)

View

Download (v1)

Created: 05/03/08 @ 14:13:24 | **Last updated:** 31/03/08 @ 16:01:41

Credits: Tomoinn

License: Creative Commons Attribution 3.0 Unported License

Original Uploader

Alan Williams

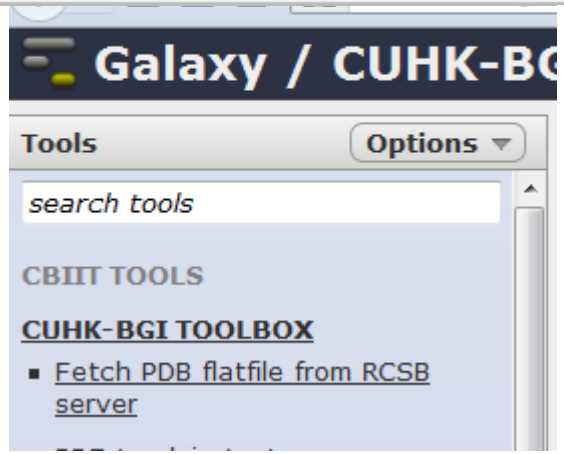
Given an identifier such as '1crn' fetches the PDB format flatfile and returns the corresponding 3D image of the protein.

Rating: 3.0 / 5 (1 rating) | **Versions:** 1 | **Reviews:** 0 | **Comments:** 0 | **Citations:** 0

Viewed: 296 times | **Downloaded:** 112 times

Tags (8):

bioinformatics | example | mygrid | pdb | protein | protein structure | rcsb | taverna



Fetch PDB flatfile from RCSB server (version 1.0.0)

Select source for pdbID:

Type manually

Enter pdbID:

1crn

Would you also like the raw results as a zip file:

No

Execute

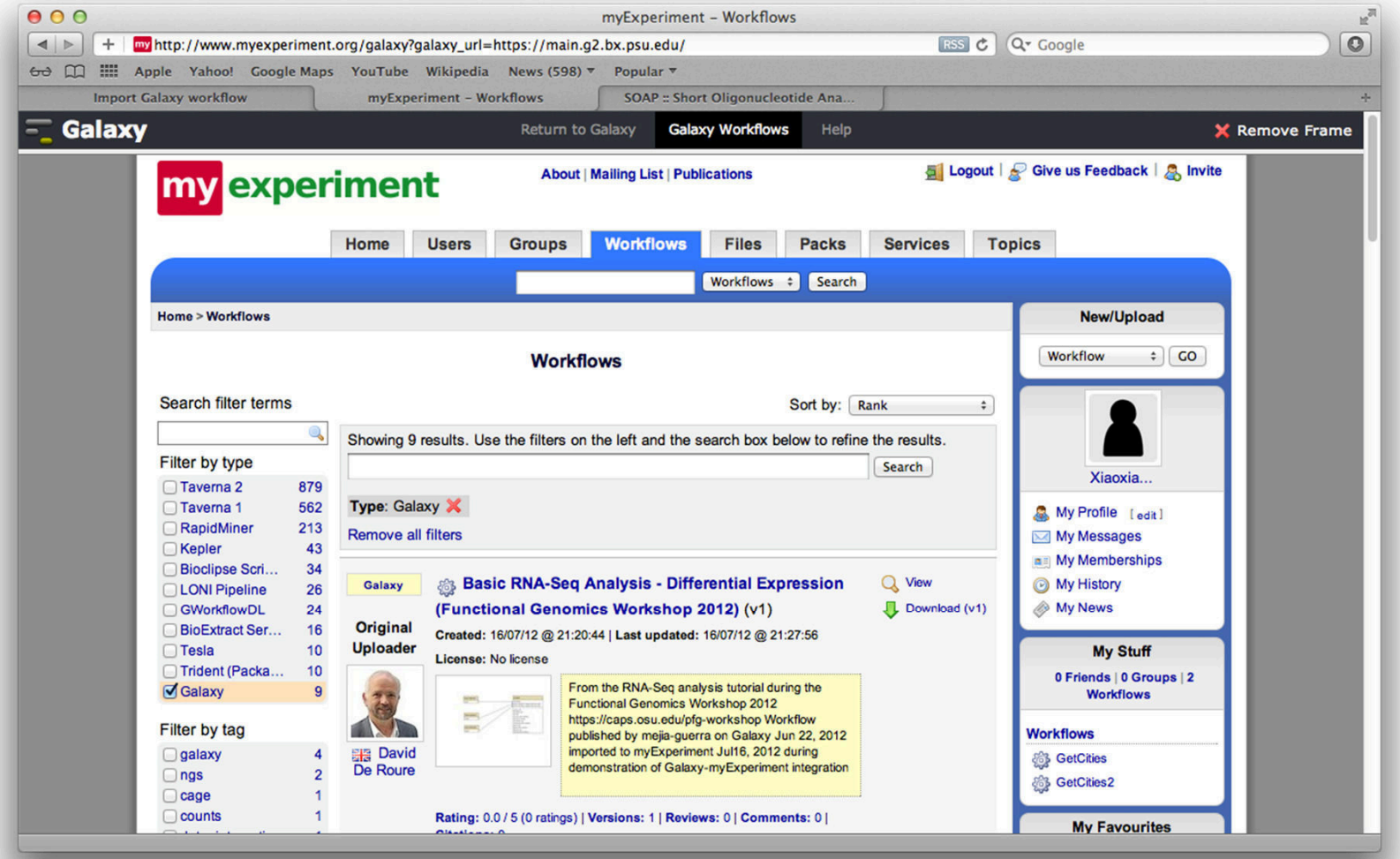
What it does

Given an identifier such as '1crn' fetches the PDB format flatfile and returns the corresponding 3D image of the protein.

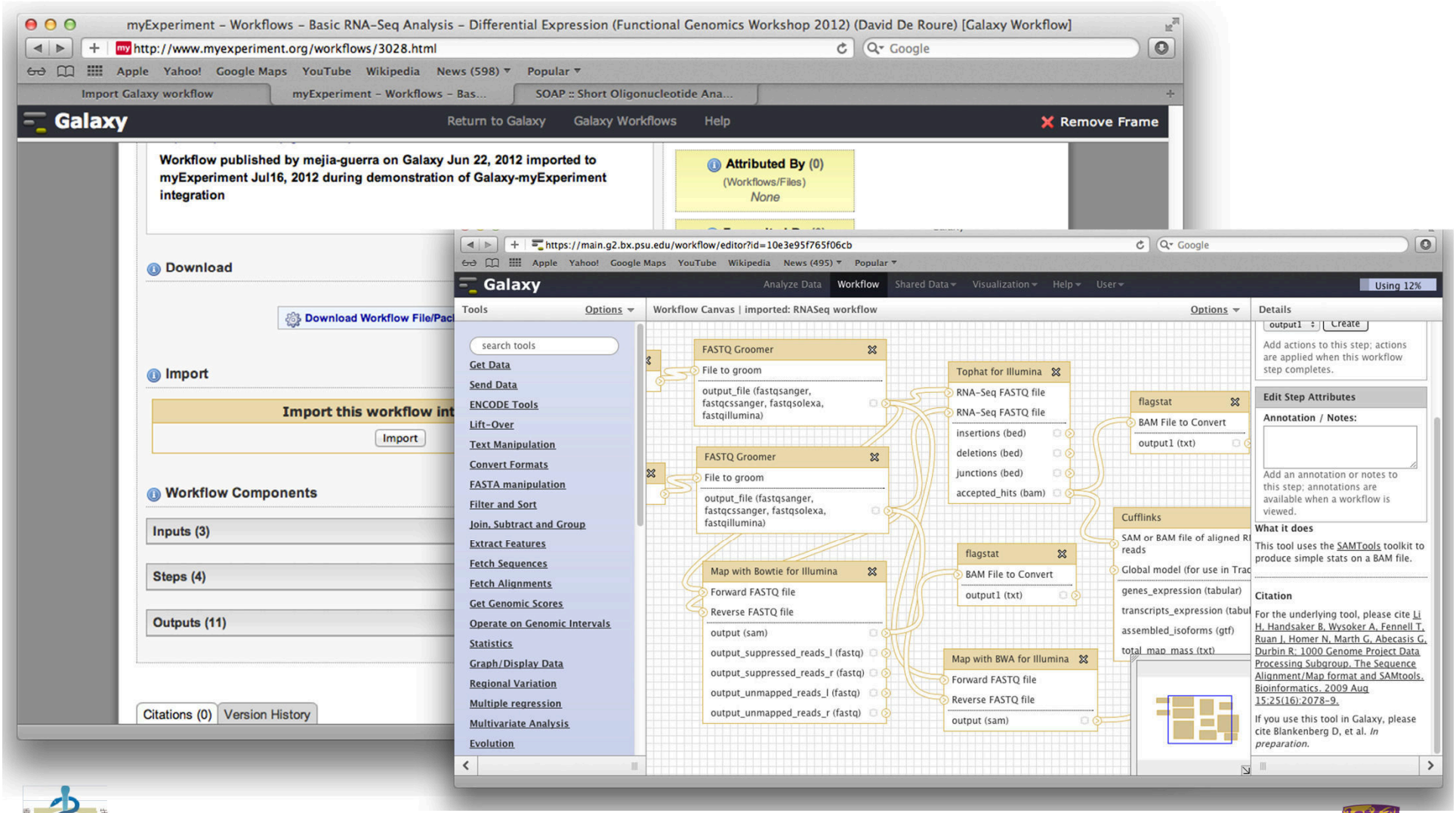
Inputs

pdbID PDB identifier such as '1crn' Examples include:
1crn

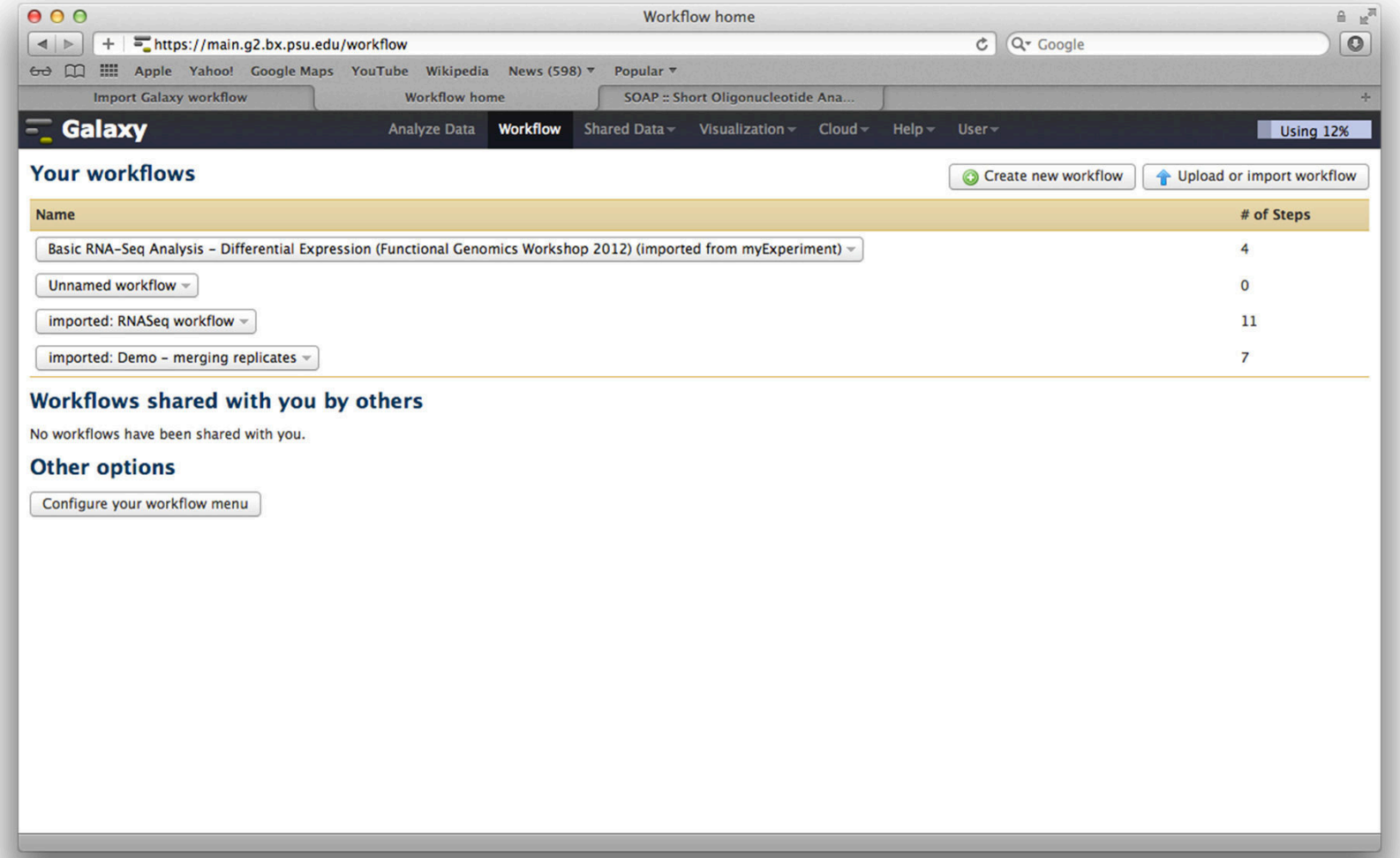
Galaxy workflow



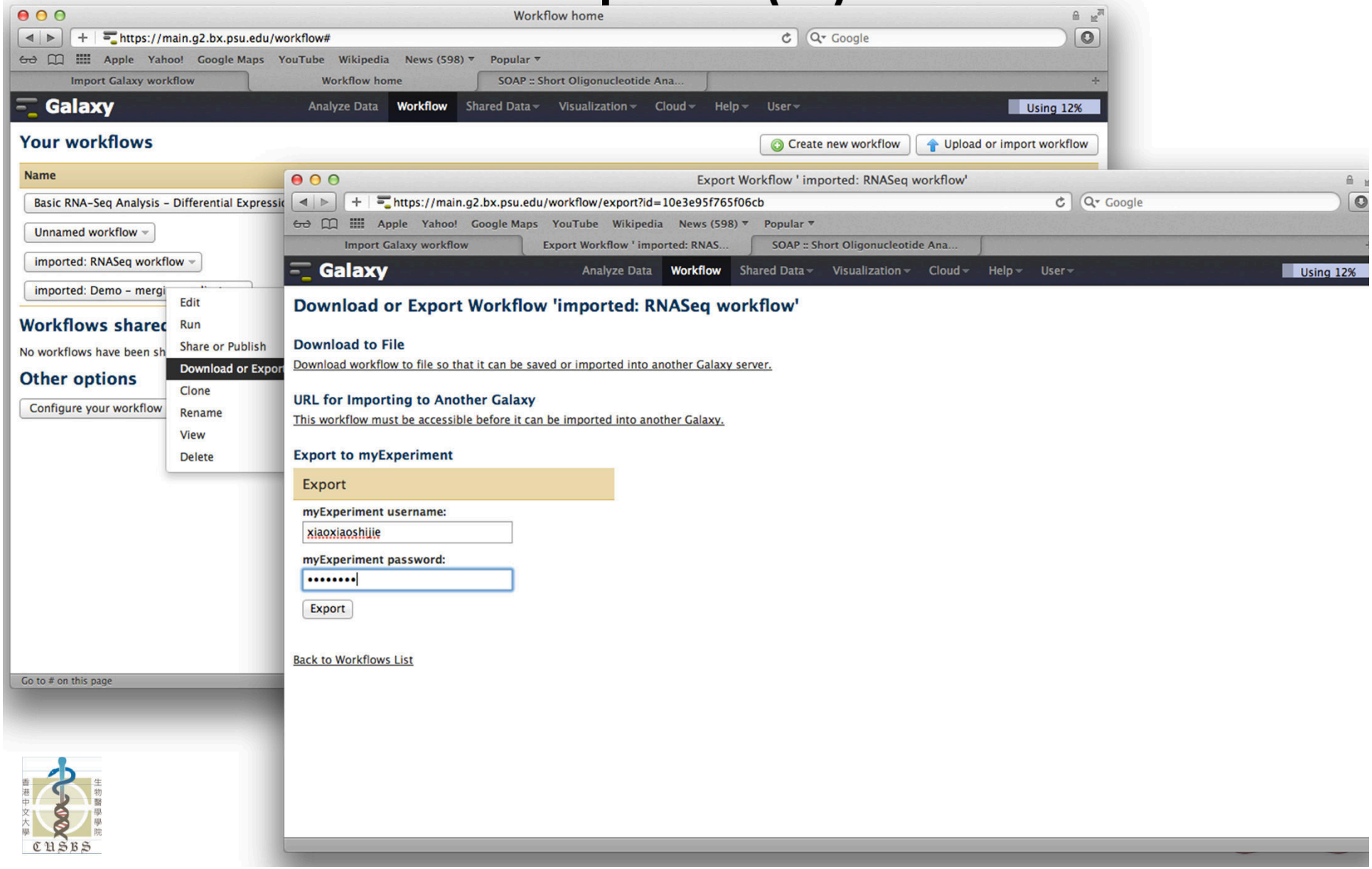
Import (1)


myExperiment – Workflows – Basic RNA-Seq Analysis – Differential Expression (Functional Genomics Workshop 2012) (David De Roure) [Galaxy Workflow]
http://www.myexperiment.org/workflows/3028.html
Galaxy
Return to Galaxy
Galaxy Workflows
Help
Remove Frame
Workflow published by mejia-guerra on Galaxy Jun 22, 2012 imported to myExperiment Jul16, 2012 during demonstration of Galaxy-myExperiment integration
Attributed By (0)
(Workflows/Files)
None
Download
Download Workflow File/Pac
Import
Import this workflow int
Import
Workflow Components
Inputs (3)
Steps (4)
Outputs (11)
Citations (0)
Version History
https://main.g2.bx.psu.edu/workflow/editor?id=10e3e95f765f06cb
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 12%
Tools
Options
search tools
Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Workflow Canvas | imported: RNASeq workflow
Options
FASTQ Groomer
File to groom
output_file (fastqsanger, fastqcssanger, fastqsolexa, fastqillumina)
Tophat for Illumina
RNA-Seq FASTQ file
RNA-Seq FASTQ file
insertions (bed)
deletions (bed)
junctions (bed)
accepted_hits (bam)
flagstat
BAM File to Convert
output1 (txt)
Cufflinks
SAM or BAM file of aligned reads
Global model (for use in Trac
genes_expression (tabular)
transcripts_expression (tabul
assembled_isoforms (gtf)
total_map_mass (txt)
Map with Bowtie for Illumina
Forward FASTQ file
Reverse FASTQ file
output (sam)
output_suppressed_reads_l (fastq)
output_suppressed_reads_r (fastq)
output_unmapped_reads_l (fastq)
output_unmapped_reads_r (fastq)
Map with BWA for Illumina
Forward FASTQ file
Reverse FASTQ file
output (sam)
Details
output1
Create
Add actions to this step; actions are applied when this workflow step completes.
Edit Step Attributes
Annotation / Notes:
Add an annotation or notes to this step; annotations are available when a workflow is viewed.
What it does
This tool uses the SAMTools toolkit to produce simple stats on a BAM file.
Citation
For the underlying tool, please cite Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078-9.
If you use this tool in Galaxy, please cite Blankenberg D, et al. In preparation.

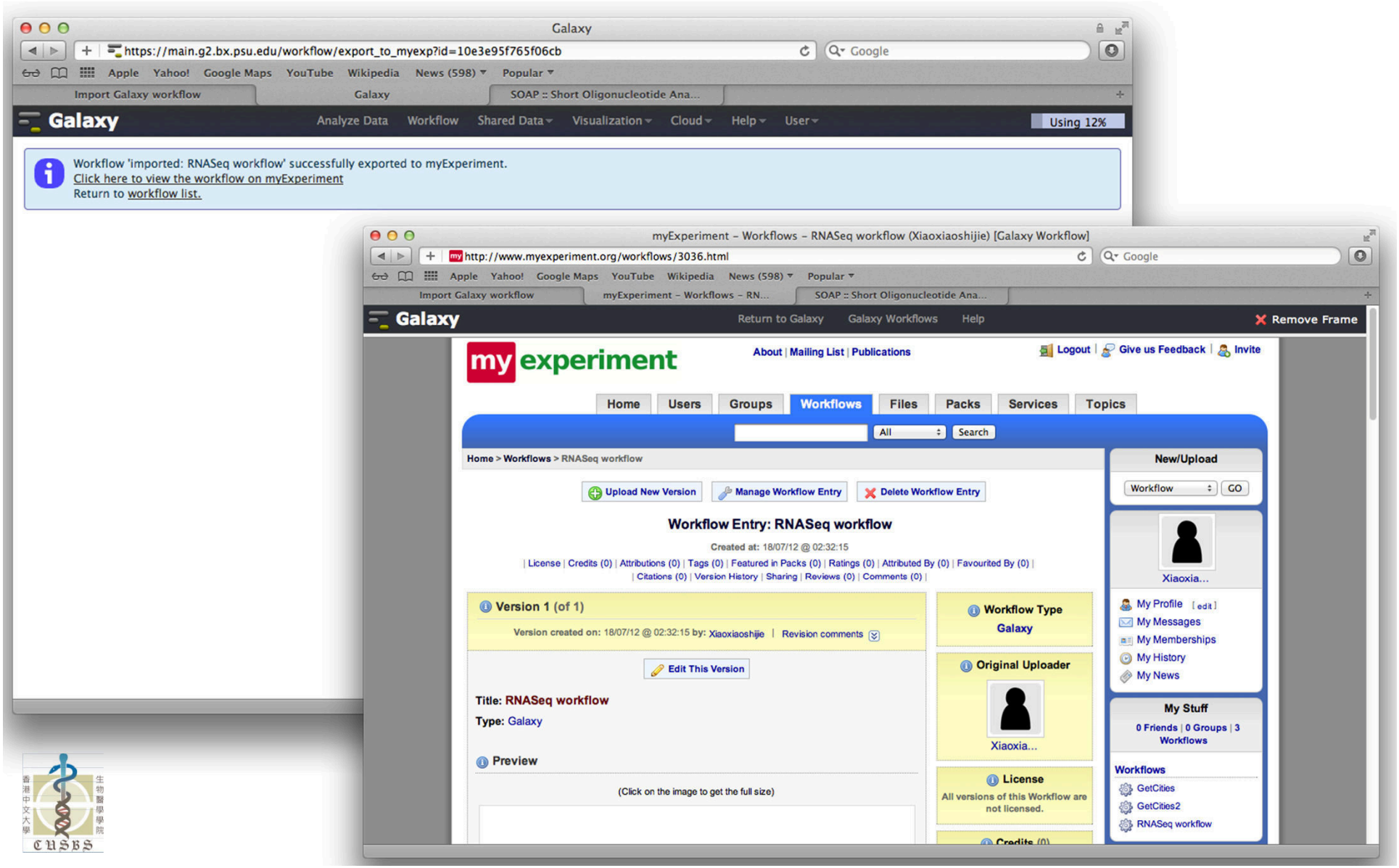
Import (2)


Workflow home
https://main.g2.bx.psu.edu/workflow
Google
Apple Yahoo! Google Maps YouTube Wikipedia News (598) Popular
Import Galaxy workflow
Workflow home
SOAP :: Short Oligonucleotide Ana...
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Cloud
Help
User
Using 12%
Your workflows
Create new workflow
Upload or import workflow
Name
# of Steps
Basic RNA-Seq Analysis - Differential Expression (Functional Genomics Workshop 2012) (imported from myExperiment)
4
Unnamed workflow
0
imported: RNASeq workflow
11
imported: Demo - merging replicates
7
Workflows shared with you by others
No workflows have been shared with you.
Other options
Configure your workflow menu

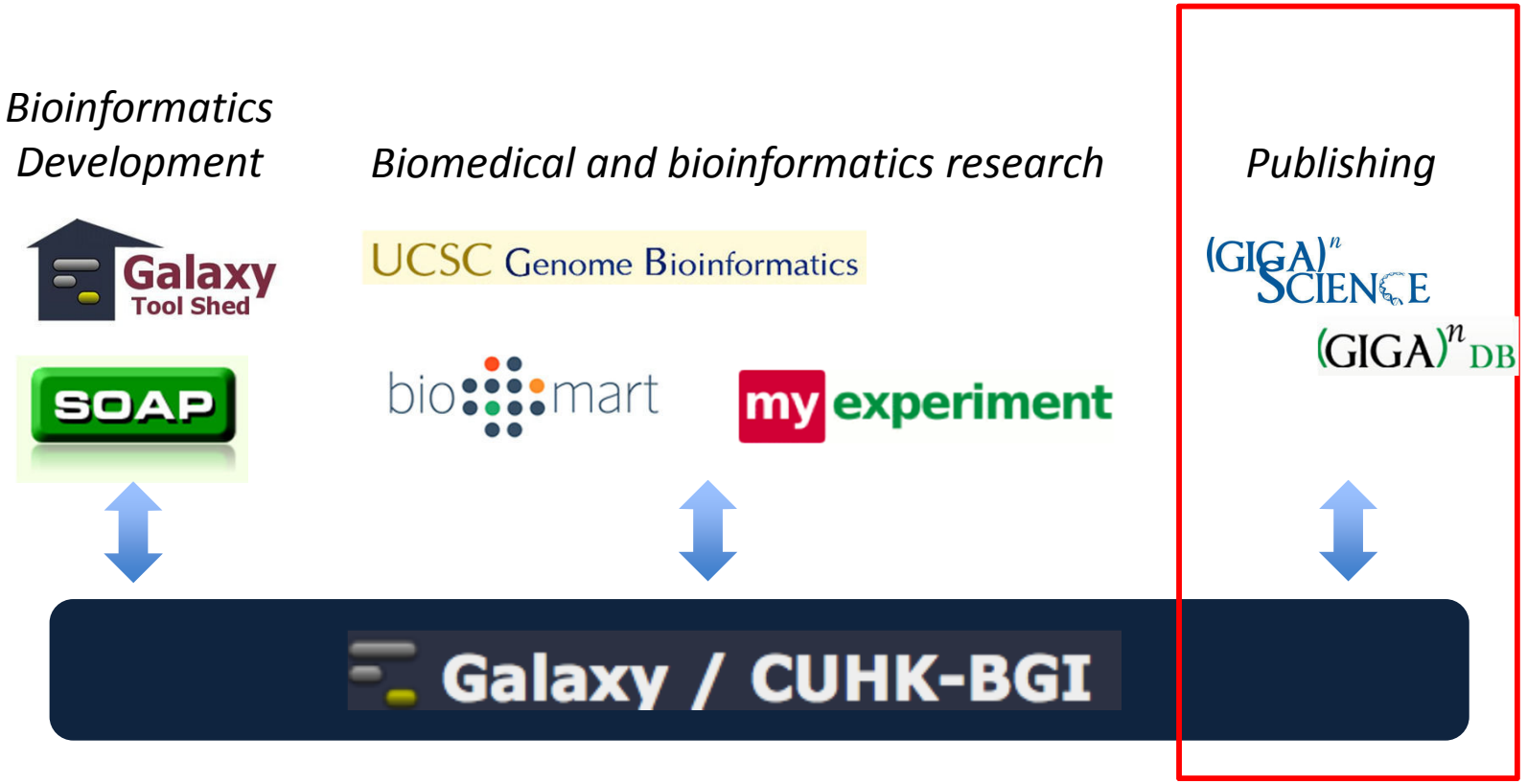
Export (1)


Workflow home
https://main.g2.bx.psu.edu/workflow#
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Cloud
Help
User
Using 12%
Your workflows
Create new workflow
Upload or import workflow
Name
Basic RNA-Seq Analysis - Differential Expressi
Unnamed workflow
imported: RNASeq workflow
Edit
Run
Share or Publish
Download or Export
Clone
Rename
View
Delete
Workflows shared
No workflows have been sh
Other options
Configure your workflow
Export Workflow ' imported: RNASeq workflow'
https://main.g2.bx.psu.edu/workflow/export?id=10e3e95f765f06cb
Download or Export Workflow 'imported: RNASeq workflow'
Download to File
Download workflow to file so that it can be saved or imported into another Galaxy server.
URL for Importing to Another Galaxy
This workflow must be accessible before it can be imported into another Galaxy.
Export to myExperiment
Export
myExperiment username:
xiaoxiaoshijie
myExperiment password:
Export
Back to Workflows List

Export (2)


Galaxy
https://main.g2.bx.psu.edu/workflow/export_to_myexp?id=10e3e95f765f06cb
Google
Apple Yahoo! Google Maps YouTube Wikipedia News (598) Popular
Import Galaxy workflow
Galaxy
SOAP :: Short Oligonucleotide Ana...
Galaxy
Analyze Data Workflow Shared Data Visualization Cloud Help User
Using 12%
Workflow 'imported: RNASeq workflow' successfully exported to myExperiment.
Click here to view the workflow on myExperiment
Return to workflow list.
myExperiment – Workflows – RNASeq workflow (Xiaoxiaoshijie) [Galaxy Workflow]
http://www.myexperiment.org/workflows/3036.html
Google
Apple Yahoo! Google Maps YouTube Wikipedia News (598) Popular
Import Galaxy workflow
myExperiment – Workflows – RN...
SOAP :: Short Oligonucleotide Ana...
Galaxy
Return to Galaxy
Galaxy Workflows
Help
Remove Frame
my experiment
About | Mailing List | Publications
Logout | Give us Feedback | Invite
Home
Users
Groups
Workflows
Files
Packs
Services
Topics
All
Search
Home > Workflows > RNASeq workflow
Upload New Version
Manage Workflow Entry
Delete Workflow Entry
Workflow Entry: RNASeq workflow
Created at: 18/07/12 @ 02:32:15
| License | Credits (0) | Attributions (0) | Tags (0) | Featured in Packs (0) | Ratings (0) | Attributed By (0) | Favourited By (0) |
| Citations (0) | Version History | Sharing | Reviews (0) | Comments (0) |
Version 1 (of 1)
Version created on: 18/07/12 @ 02:32:15 by: Xiaoxiaoshijie | Revision comments
Edit This Version
Title: RNASeq workflow
Type: Galaxy
Preview
(Click on the image to get the full size)
Workflow Type
Galaxy
Original Uploader
Xiaoxia...
License
All versions of this Workflow are not licensed.
New/Upload
Workflow
GO
Xiaoxia...
My Profile [edit]
My Messages
My Memberships
My History
My News
My Stuff
0 Friends | 0 Groups | 3 Workflows
Workflows
GetCities
GetCities2
RNASeq workflow

GDSAP structure



Now taking submissions…

(GIGA)^n SCIENCE

Large-Scale Data Journal/Database

华大基因 BGI

In conjunction with:

BioMed Central
The Open Access Publisher

Editor-in-Chief: Laurie Goodman, PhD
Editor: Scott Edmunds, PhD
Assistant Editor: Alexandra Basford, PhD

OPEN ACCESS

www.gigasciencejournal.com

GigaScience is go…

Volume 1 (2012) - July 2012

Editorial Open Access
GigaDB: announcing the GigaScience database
Tam P Sneddon, Peter Li, Scott C Edmunds
GigaScience 2012, **1**:11 (12 July 2012)
Abstract | Full text | PDF

Commentary Open Access
On the evolving portfolio of community-standards and data sharing policies: turning challenges into new opportunities
Susanna-Assunta Sansone, Philippe Rocca-Serra
GigaScience 2012, **1**:10 (12 July 2012)
Abstract | Full text | PDF

Commentary Open Access
Data sharing and publishing in the field of neuroimaging
Janis L Breeze, Jean-Baptiste Poline, David N Kennedy
GigaScience 2012, **1**:9 (12 July 2012)
Abstract | Full text | PDF

Review Open Access
Tissue sampling methods and standards for vertebrate genomics
Pamela BY Wong, Edward O Wiley, Warren E Johnson, Oliver A Ryder, Stephen J O'Brien, David Haussler, Klaus-Peter Koepfli, Marlys L Houck, Polina Perelman, Gabriela Mastromonaco, Andrew C Bentley, Byrappa Venkatesh, Ya-ping Zhang, Robert W Murphy, G10KCOS
GigaScience 2012, **1**:8 (12 July 2012)
Abstract | Full text | PDF

Technical Note Open Access
The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome
Daniel McDonald, Jose C Clemente, Justin Kuczynski, Jai Rideout, Jesse Stombaugh, Doug Wendel, Andreas Wilke, Susan Huse, John Hufnagle, Folker Meyer, Rob Knight, J Caporaso
GigaScience 2012, **1**:7 (12 July 2012)
Abstract | Full text | PDF

Commentary Open Access
Badomics words and the power and peril of the ome-meme
Jonathan A Eisen
GigaScience 2012, **1**:6 (12 July 2012)
Abstract | Full text | PDF | 1 comment

Review Open Access
A call for an international network of genomic observatories (GOs)
Neil Davies, Chris Meyer, Jack A Gilbert, Linda Amaral-Zettler, John Deck, Mesude Bicak, Philippe Rocca-Serra, Susanna Assunta-Sansone, Kathy Willis, Dawn Field
GigaScience 2012, **1**:5 (12 July 2012)
Abstract | Full text | PDF

Commentary Open Access
The rise of a digital immune system
Michael C Schatz, Adam M Phillippy
GigaScience 2012, **1**:4 (12 July 2012)
Abstract | Full text | PDF | Editor's summary

Research Open Access
Resources for methylome analysis suitable for gene knockout studies of potential epigenome modifiers
Gareth A Wilson, Pawandeep Dhami, Andrew Feber, Daniel Cortázar, Yuka Suzuki, Reiner Schulz, Primo Schär, Stephan Beck
GigaScience 2012, **1**:3 (12 July 2012)
Abstract | Full text | PDF | Editor's summary

Review Open Access
The future of DNA sequence archiving
Guy Cochrane, Charles E Cook, Ewan Birney
GigaScience 2012, **1**:2 (12 July 2012)
Abstract | Full text | PDF | Editor's summary

Editorial Open Access
Large and linked in scientific publishing
Laurie Goodman, Scott C Edmunds, Alexandra T Basford
GigaScience 2012, **1**:1 (12 July 2012)
Abstract | Full text | PDF | Editor's summary

(GIGA)n SCIENCE

REVIEW Open Access

The future of DNA sequence archiving

Guy Cochrane*, Charles E Cook and Ewan Birney

Abstract

Archives operating under the International Nucleotide Sequence Database Collaboration currently preserve all submitted sequences equally, but rapid increases in the rate of global sequence production will soon require differentiated treatment of DNA sequences submitted for archiving. Here, we provide a background on the establishment and operation of public data repositories and present the issues the community faces given the current overwhelming increase in data output. We also propose a way forward through the use a graded system in which the ease of reproduction of a sequencing-based experiment and the relative availability of a sample for resequencing be used as a means to define the level of lossy compression to the stored data.

Keywords DNA, sequence, archive, compression, storage, image

The vast majority of living organisms utilise nucleic acid as their primary store of genetic information. The technology to sequence DNA routinely was developed in the 1970s, but advances over time have since reduced cost and increased output. As the cost of sequencing has laboratory techniques in which DNA and RNA can be cut, ligated, interconverted and replicated *in vitro*. Coupled with the decreasing cost of sequencing, DNA has become a convenient readout for a variety of molecular biology assays. This started with the development of EST and cDNA technologies, was followed by high-throughput genome sequencing and then progressed through routine large-scale transcriptome sequencing, and finally to yet more intensive processes such as RNA-seq, Chip-seq and DNaseI-seq. We have even witnessed the development of DNA sequencing-based methods with no direct biological role, such as the mathematical exploration of a combinatoric space and the development of unique synthetic tags for property tracking.

DNA sequences determined for research purposes have been routinely archived since 1982, when the EMBL Data Library was founded. This was closely followed by the formation of GenBank first at the US Department of Energy and then transferred to NIH, and in 1987 by the DNA Databank of Japan. These three centres joined to form a tripartite collaboration, the INSDC, to archive and provide access to all DNA sequences generated by publicly funded research [3]. This data archiving project has gone through many changes in its 30-year history, responding both to advances in sequencing technology and to changes in the use of DNA sequence information.

Data Publishing

www.gigaDB.org

37 Datasets with DOI®s

Invertebrate
Ant
- Florida carpenter ant
- Jerdon's jumping ant
- Leaf-cutter ant

Roundworm
Schistosoma
Silkworm

1000 Genomes
A Deep Catalog of Human Genetic Variation

Human
Asian individual (YH) v1+v2
- DNA Methylome
- Genome Assembly
- **Transcriptome**

Cancer (14TB)
Hep B infected exomes
Single Cell Bladder Cancer
Ancient DNA
- Saqqaq Eskimo
- Aboriginal Australian

GENOME 10K

Vertebrates
Giant panda
Macaque
- Chinese rhesus
- Crab-eating

Mini-Pig
Naked mole rat
Penguin
- Emperor penguin
- Adelie penguin
Pigeon, domestic
Polar bear
Sheep
Tibetan antelope

Microbes
***E. Coli* O104:H4 TY-2482**

Cell-Line
Chinese Hamster Ovary
Mouse Methylomes

Released pre-publication
Non-BGI
Paper in GigaScience

Plants
Chinese cabbage
Cucumber
Foxtail millet
Pigeonpea
Potato
Sorghum

Coming soon…
Microbiome data
Parrot

GDSAP: Genomic Data Submission and Analytical platform

GigaDB v2 export to GDSAP

Results

Species	Dataset type	Dataset	Sample	File type	File format	File name	Include in download
Human	Genomic	10.5524/1000010 - Genomic sequence from an Aboriginal Australian	Biosample: 259765 - Aboriginal Australian human	SNPs	vcf	AusAboriginal.hg19.var.filtered.snps.sampled.vcf.gz	☑
Human	Genomic	10.5524/1000010 - Genomic sequence from an Aboriginal Australian	Biosample: 259765 - Aboriginal Australian human	SNPs	vcf	AusAboriginal.hg19.var.filtered.snps.vcf.gz	☑

Link to GigaDB landing pages

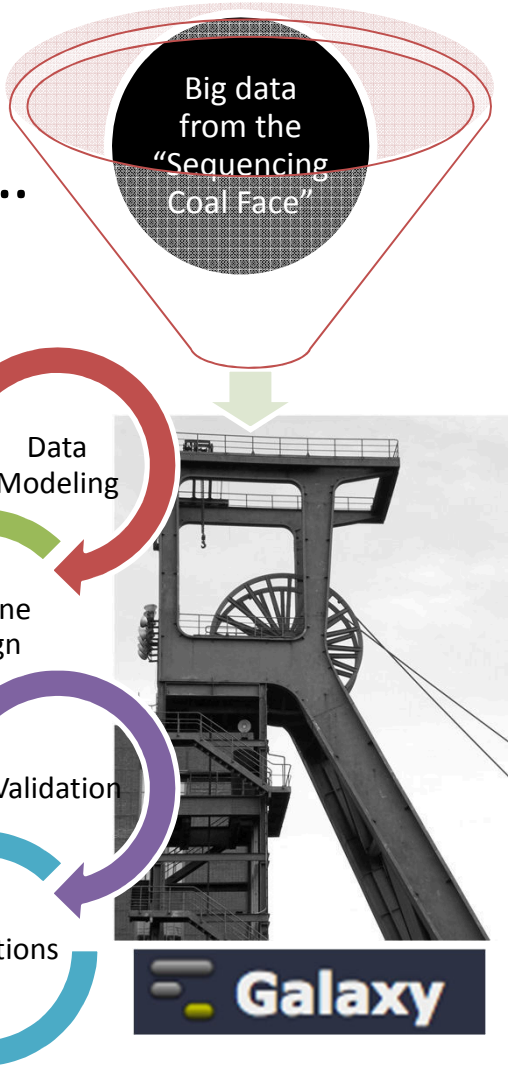
Link to sample if applicable

Download to Galaxy

Link to our documentation on file formats

GDSAP: Genomic Data Submission and Analytical platform

Data, Data, Data…



Tin-Lap Lee, CUHK

Acknowledgements

- **Lee Lab (CUHK)**
  - Huayan Gao

- **GigaScience**
  - Scott Edmunds
  - Peter Li
  - Tam Sneddon

- **BGI-Hong Kong**
  - Dennis Chan
  - Edmond Leung

- **Galaxy team**
  - Nate Coraor

- **myExperiment**
  - Finn Bacall
  - Dave De Roure

- **NBIC**
  - Kostas Karasavvas

Thank you