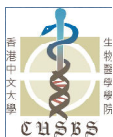
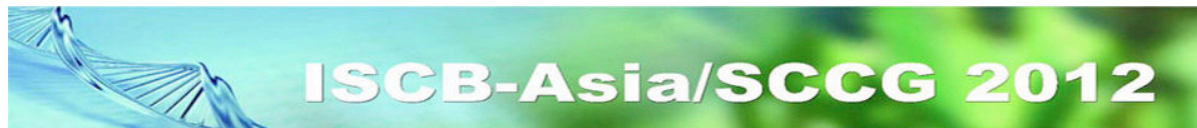


# CBIIT GigaGalaxy – A Galaxy-based Platform for Large-scale Genomics Analysis

Tin-Lap, LEE

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



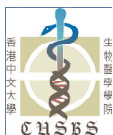
# CBIIT



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.”*



# Big Data Translates into Big Opportunities... and Big Responsibilities



The image shows the cover of the journal *Nature* (Volume 455, Issue 7242, 25 September 2008) with the main headline "BIG DATA" and the sub-headline "SCIENCE IN THE PETABYTE ERA". A yellow dashed box highlights the sub-headline. Below the cover, a snippet of an article titled "Big data: open-source format needed to aid wiki collaboration" by Tin-Lap Lee is visible. The article discusses the challenges of managing large data sets and the need for open-source formats to facilitate collaboration.

**Journal name**  
 • Advance online publication  
 • Current issue  
 • Nature News  
 • Archive  
 • Supplements

*Nature* 455, 461 (25 September 2008) | doi:10.1038/455461c; Published online 24 September 2008

**Big data: open-source format needed to aid wiki collaboration**

Tin-Lap Lee<sup>1</sup>

1. Section on Developmental Genomics, Laboratory of Clinical Genomics, Eunice Kennedy Shriver National Institute of Child Health and Human Development, Bethesda, Maryland 20892, USA  
 Email: [leeti@mail.nih.gov](mailto:leeti@mail.nih.gov)

Wikiomics' (*Nature* 455, 22–25; 2008) points out that the open-source community collaboration offers a smart way to respond to the challenges of managing large data sets. Although lack of recognition and credit may prevent some from participating, this may be only part of the story. Reasons for this can be more complicated.

Managing big data sets is becoming more and more challenging. While open-source formats are convenient, and the GenMAPP pathway database is a good example, the lack of a unified data format for facilitating data exchange between databases can be a killer for someone trying to integrate data. Currently no de facto standard on pathway-data format, which hinders data portability.

# The challenges for biomedical scientists

© Original Artist  
Reproduction rights obtainable from  
[www.CartoonStock.com](http://www.CartoonStock.com)



"This changes everything — I found  
a loophole in the genetic code!"

# The challenges for biomedical scientists





# Data intensive biology *for everyone.*

Galaxy is an open, web-based platform for data intensive biomedical research. Whether on the free public server or your own instance, you can perform, reproduce, and share complete analyses.

## Use Galaxy



Use the free public server

## Get Galaxy



Install locally or in the cloud

## Learn Galaxy



Screencasts, Galaxy 101, ...

## Get Involved



Mailing lists, Tool Shed, wiki

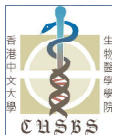
Search all resources

The Galaxy Team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University. The Galaxy Project is supported in part by NHGRI, NSF, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.

# CBIIT GigaGalaxy

## Highlights:

- Provides enhanced functionality in addition to the original Galaxy functions
  - Specialized instances
  - Speed: local servers with SBS-UCSC genome database mirror in Hong Kong
  - Reproducibility: Seamless integration with Taverna/myExperiment workflows
  - Data exchange and publishing: *GigaScience* journal portal/*GigaDB*
  - Customized functions and more.....



# CBIIT GigaGalaxy

## Benefits:

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





galaxy.cbiit.cuhk.edu.hk

**Galaxy / CBIIT** Analyze Data Workflow Shared Data Visualization Help User Using 0 bytes

**Tools**

search tools

**BGI SOAP PACKAGE BETA**

[NGS: Mapping](#)

[NGS: Indel Analysis](#)

[NGS: De Novo Assembly](#)

[NGS: Splice Detection](#)

[NGS: Somatic Mutation Detection](#)

**CBIIT TOOLBOX(UNDER DEVELOPMENT)**

**CUHK-BGI TOOLBOX**

**GALAXY TOOLBOX**

[Get Data](#)

[Send Data](#)

[ENCODE Tools](#)

[Lift-Over](#)

[Text Manipulation](#)

[Filter and Sort](#)

[Join, Subtract and Group](#)

[Convert Formats](#)

[Extract Features](#)

[Fetch Sequences](#)

[Fetch Alignments](#)

[Get Genomic Scores](#)

[Operate on Genomic Intervals](#)

[Statistics](#)

[Wavelet Analysis](#)

 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 

**News**

- **November 22, 2012: An Integration of Taverna Workflows into Galaxy Platform.**  
A project on automatic processing of Taverna Workflows to CBIIT-Galaxy Platform is under development.
- **November 12, 2012: BGI SOAP Package Beta Version is Available now!**  
BGI SOAP Package Beta Version is available for sneak preview. Please send comments or suggestions to [Prof. LEE Tin-lap](#) at CBIIT. Many thanks for the contribution of Peter Li at GigaScience.
- **September 10, 2012: SOAPdenovo tool is ready.**  
The first tool of SOAP package is available now.
- **July 5, 2012: Running Taverna Workflows on Galaxy Platform.**  
Running Taverna Workflows on CBIIT-Galaxy Platform have been successfully tested.
- **January 18, 2012: Announcement**  
Genomic Data Submission and Analytical Platform (GDSAP) is under development with  [GigaScience journal](#).

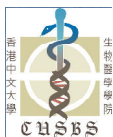
[Galaxy team](#) is a part of [BX](#) at [Penn State](#).

This project is led by [Prof. LEE Tin-lap](#), developed and maintained by [GAO Huayan](#), and supported by [School of Biomedical Sciences](#) at [Chinese University of Hong Kong](#) and [BGI](#).

**History**

0 bytes

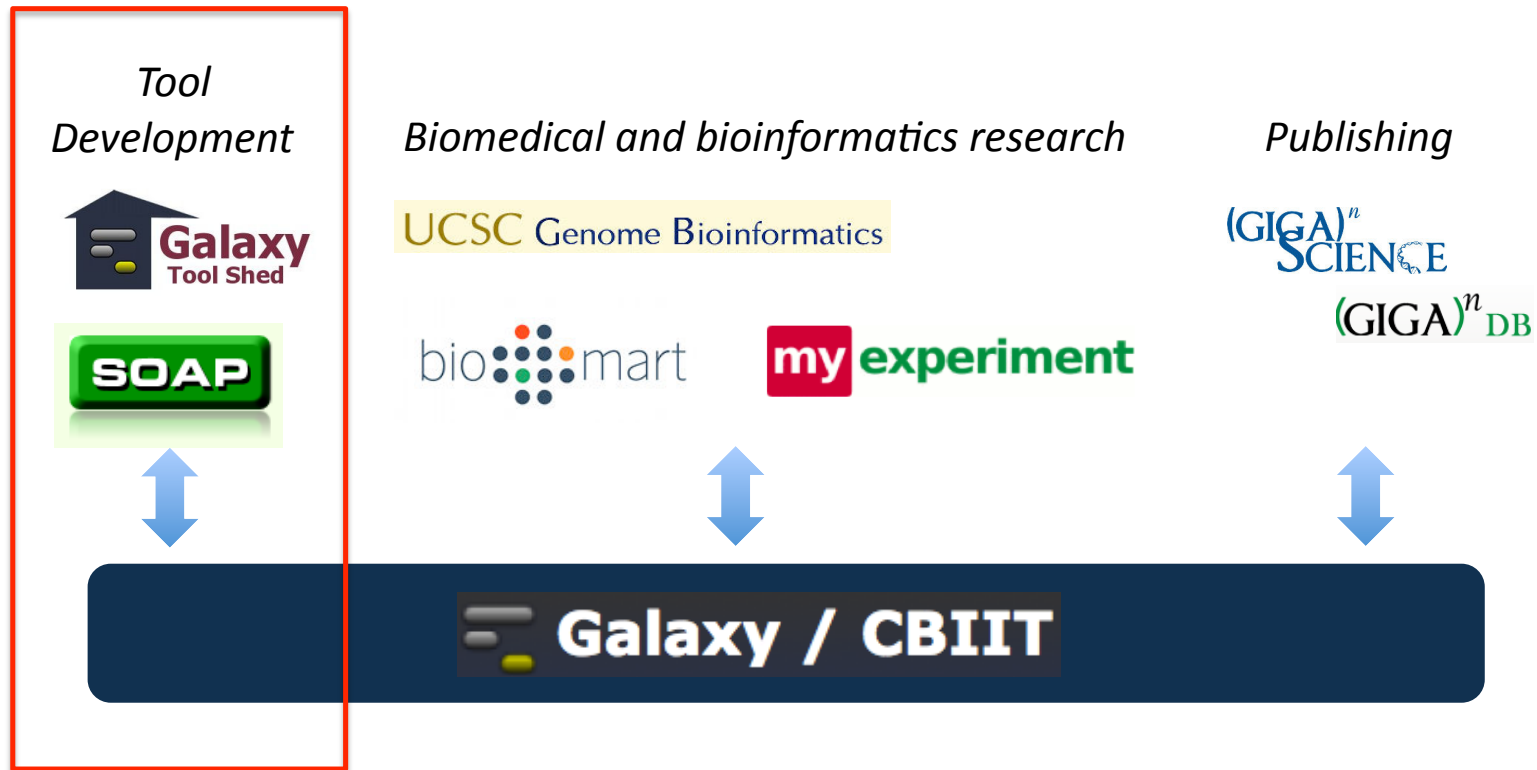
**i** Your history is empty. Click 'Get Data' on the left pane to start



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



# CBIIT GigaGalaxy 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** <sup>NEW</sup>

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

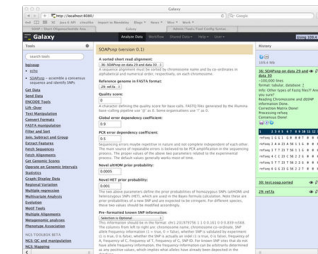
- Enables sharing of Galaxy tools across Galaxy servers around the world.
- SOAP package tools configured for use in Galaxy.
  - SOAPsnp/SOAPdenovo



Command line call

Python wrapper

Tool XML config file



# NGS mapping: SOAP1

**Galaxy / CBIIT**

Analyze DataWorkflowShared DataVisualizationHelpUser

Tools

search tools

**BGI SOAP PACKAGE BETA**

**NGS: Mapping**

- [soap1](#) – short oligonucleotide alignment tool
- [soap2](#) – improved version of soap1

**NGS: Indel Analysis**

**NGS: De Novo Assembly**

**NGS: Splice Detection**

**NGS: Somatic Mutation Detection**

**CBIIT TOOLBOX(UNDER DEVELOPMENT)**

**CUHK-BGI TOOLBOX**

**GALAXY TOOLBOX**

[Get Data](#)

[Send Data](#)

[ENCODE Tools](#)

[Lift-Over](#)

[Text Manipulation](#)

[Filter and Sort](#)

[Join, Subtract and Group](#)

[Convert Formats](#)

[Extract Features](#)

**soap1 (version 0.1)**

Select a reference sequence:

What type of mapping do you want to perform?:

Single

FASTA file:

SOAP settings to use:

Default

Default settings is suitable for most mapping needs. If you want full control, use Full parameter list

Execute

**What it does**

SOAP performs efficient gapped and ungapped alignment of short oligonucleotides onto reference sequences. The program is designed to handle short reads generated by parallel sequencing using the new generation Illumina-Solexa sequencing technology. SOAP is compatible with numerous applications, including single-read or pair-end resequencing, small RNA discovery and mRNA tag sequence mapping. SOAP supports multi-threaded parallel computing, and has a batch mode for query multiple data sets.

**Single-end sequencing**

SOAP will allow a certain number of mismatches or one continuous gap when aligning a read onto a reference sequence. The best hit of each read which has the minimal number of mismatches or the smallest gap is reported. For multiple equal best hits, the user can instruct SOAP to report all hits, a random one, or disregard



# NGS mapping: SOAP2

**Galaxy / CBIIT**

Analyze DataWorkflowShared DataVisualizationHelpUser

Tools

search tools

BGI SOAP PACKAGE BETA

NGS: Mapping

- [soap1](#) – short oligonucleotide alignment tool
- [soap2](#) – improved version of soap1

NGS: Indel Analysis

NGS: De Novo Assembly

NGS: Splice Detection

NGS: Somatic Mutation Detection

CBIIT TOOLBOX(UNDER DEVELOPMENT)

CUHK-BGI TOOLBOX

GALAXY TOOLBOX

Get Data

Send Data

ENCODE Tools

Lift-Over

Text Manipulation

Filter and Sort

Join, Subtract and Group

Convert Formats

Extract Features

Fetch Sequences

Fetch Alignments

Get Genomic Scores

Operate on Genomic Intervals

soap2 (version 0.2)

Select a reference genome from your history or use a built-in index:

Use built-in index

Select a reference genome:

hs\_chr10

What type of mapping do you want to perform?:

Single

FASTA file:

SOAP settings to use:

Default

Default settings is suitable for most mapping needs. If you want full control, use Full parameter list

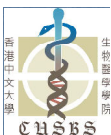
Execute

What it does

SOAP2 (also known as SOAPaligner) is a member of the SOAP (Short Oligonucleotide Analysis Package). This second version of the SOAP tool for short oligonucleotide alignment features fast, accurate alignment for huge amounts of short reads generated by the Illumina/Solexa Genome Analyzer.

Compared to version 1 of SOAP, SOAP2 is one order of magnitude faster so that, for example, it requires only 2 minutes to align one million single-end reads onto the human reference genome. Another improvement of SOAP2 is that it now supports a wide range of read lengths.

SOAP2 made improvements in time and space efficiency by a re-implementing the basic data structures and algorithms used in SOAP1. The core algorithms and the indexing data structures (2way-BWT) were developed by T.W. Lam, Alan Tam, Simon Wong, Edward Wu and S.M. Yiu of the Algorithms Research group at the Department of Computer Science, the University of Hong Kong.





# SOAPsnp

**Galaxy / CBIIT**

Analyze DataWorkflowShared DataVisualizationHelpUser

Tools

search tools

BGI SOAP PACKAGE BETA

**NGS: Mapping**

- [soap1](#) – short oligonucleotide alignment tool
- [soap2](#) – improved version of soap1

**NGS: Indel Analysis**

**NGS: De Novo Assembly**

**NGS: Splice Detection**

**NGS: Somatic Mutation Detection**

CBIIT TOOLBOX(UNDER DEVELOPMENT)

**CUHK-BGI TOOLBOX**

GALAXY TOOLBOX

**Get Data**

**Send Data**

**ENCODE Tools**

**Lift-Over**

**Text Manipulation**

**Filter and Sort**

**Join, Subtract and Group**

**Convert Formats**

**Extract Features**

**Fetch Sequences**

**Fetch Alignments**

**Get Genomic Scores**

**Operate on Genomic Intervals**

soap2 (version 0.2)

Select a reference genome from your history or use a built-in index:

Use built-in index

Select a reference genome:

hs\_chr10

What type of mapping do you want to perform?:

Single

FASTA file:

SOAP settings to use:

Default

Default settings is suitable for most mapping needs. If you want full control, use Full parameter list

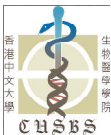
Execute

**What it does**

SOAP2 (also known as SOAPaligner) is a member of the SOAP (Short Oligonucleotide Analysis Package). This second version of the SOAP tool for short oligonucleotide alignment features fast, accurate alignment for huge amounts of short reads generated by the Illumina/Solexa Genome Analyzer.

Compared to version 1 of SOAP, SOAP2 is one order of magnitude faster so that, for example, it requires only 2 minutes to align one million single-end reads onto the human reference genome. Another improvement of SOAP2 is that it now supports a wide range of read lengths.

SOAP2 made improvements in time and space efficiency by a re-implementing the basic data structures and algorithms used in SOAP1. The core algorithms and the indexing data structures (2way-BWT) were developed by T.W. Lam, Alan Tam, Simon Wong, Edward Wu and S.M. Yiu of the Algorithms Research group at the Department of Computer Science, the University of Hong Kong.



# SOAPpopindel

**Galaxy / CBIIT**

Analyze DataWorkflowShared DataVisualizationHelpUser

Tools

search tools

**BGI SOAP PACKAGE BETA**

**NGS: Mapping**

- [soap1](#) – short oligonucleotide alignment tool
- [soap2](#) – improved version of soap1

**NGS: Indel Analysis**

- [msort](#) – sort tabular data by multiple fields
- [SOAPsnp](#) – assemble a consensus sequence and identify SNPs
- [SOAPpopindel](#) – detection of splice junctions

**NGS: De Novo Assembly**

**DE NOVO ASSEMBLY TOOLS**

- [SOAPdenovo](#) – perform de novo genome assembly
- [SOAPdenovo2](#) – perform de novo genome assembly with optimized k-mer determination
- [SOAPdenovo-trans](#) – perform de novo transcriptome assembly
- [SoapCoverage](#) – calculate the coverage and depth of target sequences out of soap mapping result

SOAPpopindel (version 0.1)

Depth file:

See documentation below for information on this format

Ploidy:

2

Execute

**What it does**

SOAP-PopIndel is a novel probabilistic framework for fast and sensitive indel genotyping at the population level. This tool was created by Haojing Shao and Hanjiudai Yin at BGI.

**Tool parameters**

Depth file:

The depth file is a temporary data format used for storing supporting read information. It could be generated by your raw data, realignment or other. The format for a depth file is as follows:

First line:  
Chromosome Position Non-reference\_allele 1 (Non-reference\_allele 2) (Non-reference\_allele 3)

Second line:  
Number of supporting reference read counts for sample 1, sample 2,sample N

Third line:  
Number of supporting first non-reference allele read counts for sample 1, sample 2, sample N

# NGS De Novo Assembly: SOAPdenovo

**Galaxy / CBIIT** Analyze Data Workflow Shared Data Visualization Help User

**Tools**

search tools

**BGI SOAP PACKAGE BETA**

**NGS: Mapping**

- [soap1](#) – short oligonucleotide alignment tool
- [soap2](#) – improved version of soap1

**NGS: Indel Analysis**

- [msort](#) – sort tabular data by multiple fields
- [SOAPsnp](#) – assemble a consensus sequence and identify SNPs
- [SOAPpopindel](#) – detection of splice junctions

**NGS: De Novo Assembly**

**DE NOVO ASSEMBLY TOOLS**

- [SOAPdenovo](#) – perform de novo genome assembly
- [SOAPdenovo2](#) – perform de novo genome assembly with optimized k-mer determination
- [SOAPdenovo-trans](#) – perform de novo transcriptome assembly
- [SoapCoverage](#) – calculate the coverage and depth of target sequences out of soap mapping result

**SOAPdenovo (version 0.1)**

**Maximum read length:**

**library**

**libraries 1**

**Average insert size:**

**Reverse sequence?:**

**Which operations should the reads be used for?:**

**Which order are the reads used while scaffolding:**

**What type of data are you using?:**

**What type of data are you using?:**

**Forward FASTQ file:**

# NGS De Novo Assembly: SOAPdenovo2

**Galaxy / CBIIT**

Analyze DataWorkflowShared DataVisualizationHelpUser

Tools

**NGS: Indel Analysis**

- [msort](#) – sort tabular data by multiple fields
- [SOAPsnp](#) – assemble a consensus sequence and identify SNPs
- [SOAPpopindel](#) – detection of splice junctions

**NGS: De Novo Assembly**

DE NOVO ASSEMBLY TOOLS

- [SOAPdenovo](#) – perform de novo genome assembly
- [SOAPdenovo2](#) – perform de novo genome assembly with optimized k-mer determination
- [SOAPdenovo-trans](#) – perform de novo transcriptome assembly
- [SoapCoverage](#) – calculate the coverage and depth of target sequences out of soap mapping result
- [GapCloser](#) – close the gaps using the abundant pair relationships of short reads

SOAPDENOV2 MODULES

- [config](#)
- [pregraph](#)
- [contig](#)
- [map](#)
- [scaff](#)

**SOAPdenovo2 (version 0.1)**

Select a configuration file from history or create a new one?:

Select configuration file from history:

SOAP settings to use:

Default settings is suitable for most mapping needs. If you want full control, use Full parameter list

**Execute**

**What it does**  
SOAPdenovo is a novel short-read assembly method that can build a de novo draft assembly for the human-sized genomes. The program is specially designed to assemble Illumina GA short reads. It creates new opportunities for building reference sequences and carrying out accurate analyses of unexplored genomes in a cost effective way.

**System requirements**  
SOAPdenovo2 is designed for assembling large plant and animal genomes, although it also works well on bacteria and fungal genomes. The code runs on 64-bit/32-bit Linux/MAC OSX systems with a minimum of 5G physical memory. Approximately 150 GB of memory is required to process large genomes such as those of humans.

**Notes**  
This Galaxy tool is a wrapping of SOAPdenovo version 2.04. This version supports large k-mers of up to 127 in size to process long reads. Changes which have been made to SOAPdenovo-1 include:  

1. Merging of the 63mer and 127mer versions.
2. A new module named "sparse-pregraph" has been written which can reduce considerable computational consumption.
3. The "Multi-kmer" method has been introduced in the "contig" step to allow the utilization of the advantages of small and large k-mers.



# CBIIT GigaGalaxy structure

*Bioinformatics  
Development*



*Biomedical and bioinformatics research*

UCSC Genome Bioinformatics

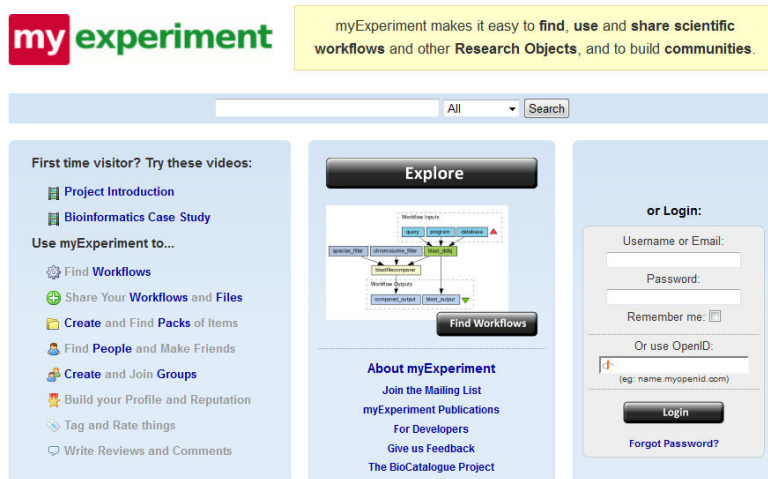


*Publishing*





# How does it work?



<http://www.myexperiment.org>

- **myExperiment** -a repository for workflows.

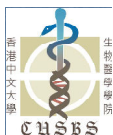
➤ Taverna workflows.



➤ New: Galaxy workflows.



- **CBIIT GigaGalaxy** integration





# Taverna workflow

**my experiment** [About](#) | [Mailing List](#) | [Publications](#) [Log in](#) | [Register](#)

[Home](#) [Users](#) [Groups](#) [Workflows](#) [Files](#) [Packs](#) [Services](#) [To](#)

Home > Workflows > Fetch PDB flatfile from RCSB server

### Workflow Entry: Fetch PDB flatfile from RCSB server

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

[License](#) | [Credits](#) (1) | [Attributions](#) (0) | [Tags](#) (8) | [Featured in Packs](#) (1) | [Ratings](#) (1) | [Attributed By](#) (3) | [Favoured By](#) (0) | [Citations](#) (0) | [Version History](#) | [Reviews](#) (0) | [Comments](#) (0)

**Version 1 (of 1)**

Version created on: 05/03/08 @ 14:13:24 by: Alan Williams | [Revision comments](#)

Last edited on: 31/03/08 @ 16:01:41 by: Alan Williams

**Title:** Fetch PDB flatfile from RCSB server

**Type:** Taverna 1

**Preview**

(Click on the image to get the full size)

```
graph TD; subgraph Inputs; RCSBPrefix; pdbID; end; RCSBPrefix --> RCSBSuffix; RCSBSuffix --> AddPrefixToID; pdbID --> AddPrefixToID; AddPrefixToID --> AddSuffix; AddSuffix --> FetchPage; FetchPage --> Output; subgraph Outputs; Output[pdbFlatFile]; end;
```

[Download Scalable Diagram \(SVG\)](#)

**Workflow Type**  
Taverna 1

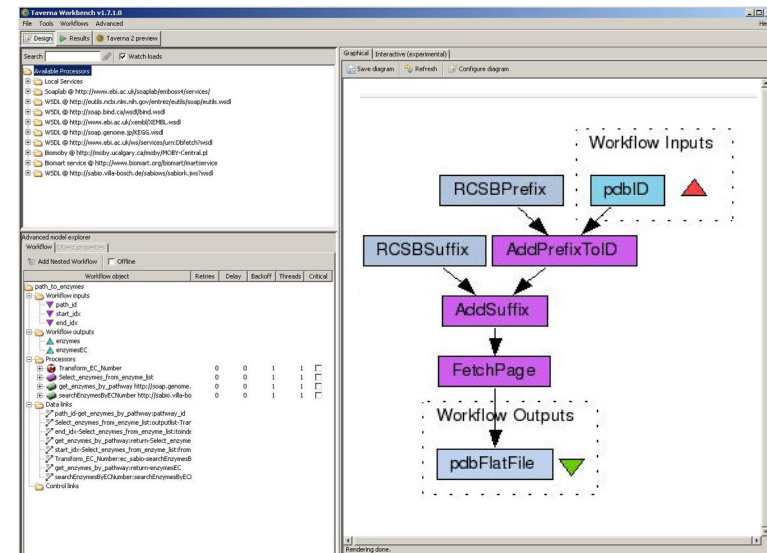
**Original Uploader**  
  
Alan Williams

**License**  
All versions of this workflow are licensed under:

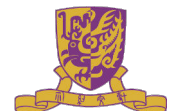
**Credits** (1)  
(People/Groups)  
Tomoinn

**Attributions** (0)  
(Workflows/Files)  
None


**Tags** (8)  
[Original Uploader tags](#)  
bioinformatics | example | mygrid | pdb | protein | protein structure | rcsb | taverna




<http://www.taverna.org.uk/>






**Taverna 1**  **Fetch PDB flatfile from RCSB server (v1)** [View](#)


**Created:** 05/03/08 @ 14:13:24 | **Last updated:** 31/03/08 @ 16:01:41 [Download \(v1\)](#)

**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

**Galaxy / CUHK-B**

**Tools** **Options**

search tools

**CBIT TOOLS**

**CUHK-BGI TOOLBOX**

- Fetch PDB flatfile from RCSB server

**Fetch PDB flatfile from RCSB server (version 1.0.0)**

Select source for pdbID:  
 Type manually

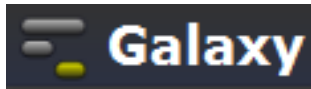
Enter pdbID:  
 1crn

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

myExperiment - Workflows

http://www.myexperiment.org/galaxy?galaxy\_url=https://main.g2.bx.psu.edu/

Import Galaxy workflow myExperiment - Workflows 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

Home > Workflows

Workflows

Search filter terms Sort by: Rank

Showing 9 results. Use the filters on the left and the search box below to refine the results.

Type: Galaxy X Remove all filters

Filter by type

- ☐ Taverna 2 879
- ☐ Taverna 1 562
- ☐ RapidMiner 213
- ☐ Kepler 43
- ☐ Bioclipse Scri... 34
- ☐ LONI Pipeline 26
- ☐ GWorkflowDL 24
- ☐ BioExtract Ser... 16
- ☐ Tesla 10
- ☐ Trident (Packa... 10
- ☒ Galaxy 9

Filter by tag

- ☐ galaxy 4
- ☐ ngs 2
- ☐ cage 1
- ☐ counts 1

Original Uploader

David De Roure

Basic RNA-Seq Analysis - Differential Expression (Functional Genomics Workshop 2012) (v1)

Created: 16/07/12 @ 21:20:44 | Last updated: 16/07/12 @ 21:27:56

License: No license

From the RNA-Seq analysis tutorial during the Functional Genomics Workshop 2012  
https://caps.osu.edu/pfg-workshop Workflow published by mejia-guerra on Galaxy Jun 22, 2012 imported to myExperiment Jul16, 2012 during demonstration of Galaxy-myExperiment integration

Rating: 0.0 / 5 (0 ratings) | Versions: 1 | Reviews: 0 | Comments: 0

New/Upload Workflow GO

Xiaoxia...

My Profile [edit]  
My Messages  
My Memberships  
My History  
My News

My Stuff  
0 Friends | 0 Groups | 2 Workflows

Workflows  
GetCities  
GetCities2

My Favourites



# Import (1)

The image shows two overlapping browser windows from the Galaxy project management system.

The top window displays the workflow's history and download options. The URL is <http://www.myexperiment.org/workflows/3028.html>. The workflow is titled "Basic RNA-Seq Analysis - Differential Expression (Functional Genomics Workshop 2012) (David De Roure) [Galaxy Workflow]". It was published by mejia-guerra on Galaxy Jun 22, 2012 and imported to myExperiment Jul 16, 2012. The workflow is attributed to (0) and has no associated workflows or files. The left sidebar shows the "Download" section with a "Download Workflow File/Pack" button. The "Import" section has a button to "Import this workflow into myExperiment". The "Workflow Components" section shows the workflow has 3 inputs, 4 steps, and 11 outputs. The bottom of the sidebar shows "Citations (0)" and "Version History".

The bottom window shows the workflow canvas for the "imported: RNASeq workflow". The canvas displays a complex pipeline of tools connected by lines. The tools include:

- FASTQ Groomer (multiple instances)
- File to groom
- RNA-Seq FASTQ file
- Tophat for Illumina
- flagstat
- BAM File to Convert
- Cufflinks
- Map with Bowtie for Illumina
- Map with BWA for Illumina

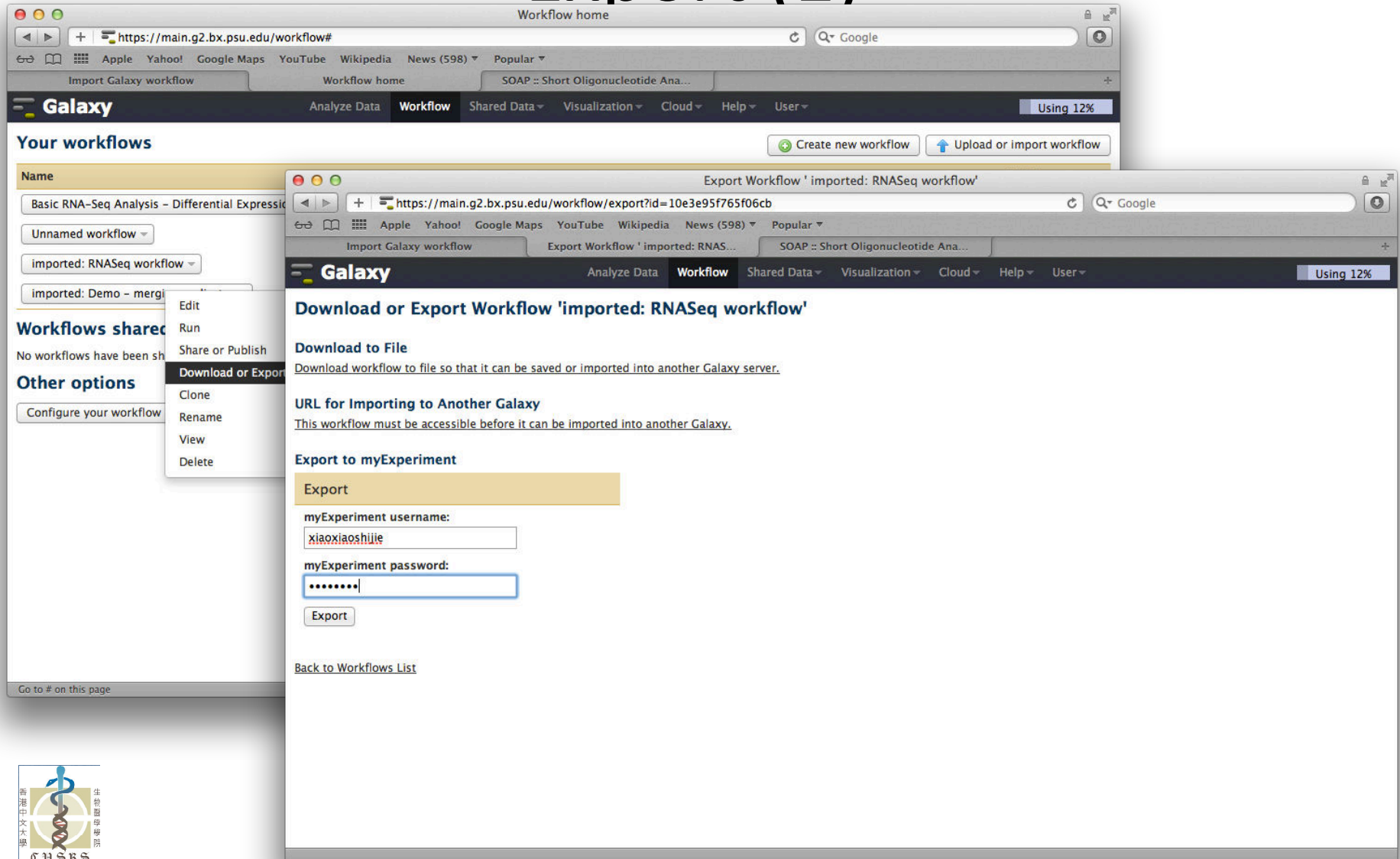
The right sidebar shows the "Details" section for the selected tool, "Cufflinks". It includes a "Create" button, a description of the tool, and a citation: "For the underlying tool, please cite Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, C., Abecasis, G., Durbin, R., 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)

The screenshot shows the Galaxy web interface for workflow management. The browser address bar displays `https://main.g2.bx.psu.edu/workflow`. The interface includes a navigation bar with tabs for 'Import Galaxy workflow', 'Workflow home', and 'SOAP :: Short Oligonucleotide Ana...'. Below this is a header with the 'Galaxy' logo and navigation links: 'Analyze Data', 'Workflow' (selected), 'Shared Data', 'Visualization', 'Cloud', 'Help', and 'User'. A status bar indicates 'Using 12%'. The main content area is titled 'Your workflows' and contains a table of existing workflows. To the right of the table are buttons for 'Create new workflow' and 'Upload or import workflow'. Below the table, there is a section 'Workflows shared with you by others' stating 'No workflows have been shared with you.', followed by 'Other options' with a button 'Configure your workflow menu'.

| 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          |

# Export (1)



The image shows two overlapping browser windows from the Galaxy project management tool. The background window, titled 'Workflow home', displays a list of workflows under 'Your workflows'. A workflow named 'imported: RNASeq workflow' is selected, and a context menu is open with the 'Download or Export' option highlighted. The foreground window, titled 'Export Workflow 'imported: RNASeq workflow'', shows the export options. Under 'Download or Export Workflow 'imported: RNASeq workflow'', there are three main sections: 'Download to File' (with a link to download the workflow to a file), 'URL for Importing to Another Galaxy' (with a link to the workflow's URL), and 'Export to myExperiment'. The 'Export to myExperiment' section is expanded, showing a form with fields for 'myExperiment username' (filled with 'xiaoxiaoshijie') and 'myExperiment password' (filled with '\*\*\*\*\*'). An 'Export' button is visible below the password field. A 'Back to Workflows List' link is at the bottom of the foreground window. In the bottom left corner, there is a logo for CUSSS (CUHK University Science and Safety Service) featuring a caduceus and the text '香港中文大學' and 'CUSSS'.

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 Expression

Unnamed workflow

imported: RNASeq workflow

imported: Demo - merge

Edit

Run

Share or Publish

Download or Export

Clone

Rename

View

Delete

Workflows shared

No workflows have been shared

Other options

Configure your workflow

Go to # on this page

Export Workflow 'imported: RNASeq workflow'

https://main.g2.bx.psu.edu/workflow/export?id=10e3e95f765f06cb

Galaxy

Analyze Data Workflow Shared Data Visualization Cloud Help User

Using 12%

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:

\*\*\*\*\*

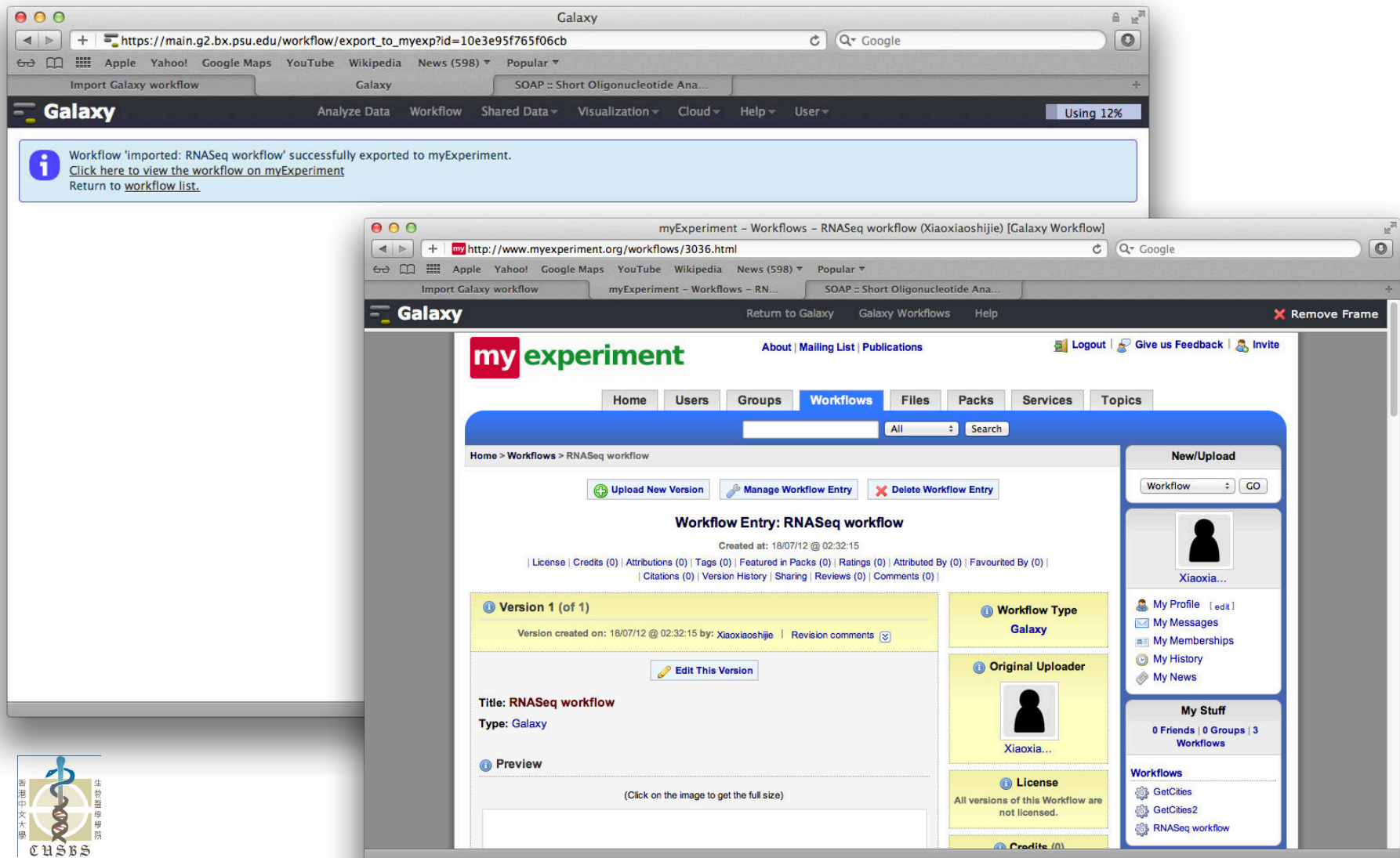
Export

Back to Workflows List

CUSSS



# Export (2)



The image shows two overlapping browser windows. The background window is the Galaxy web interface at [https://main.g2.bx.psu.edu/workflow/export\\_to\\_myexp?id=10e3e95f765f06cb](https://main.g2.bx.psu.edu/workflow/export_to_myexp?id=10e3e95f765f06cb). It displays a success message: "Workflow 'imported: RNASeq workflow' successfully exported to myExperiment. Click here to view the workflow on myExperiment. Return to workflow list." The foreground window is the myExperiment interface at <http://www.myexperiment.org/workflows/3036.html>. It shows the details of the "RNASeq workflow" (Version 1 of 1), created on 18/07/12 at 02:32:15 by Xiaoxiaoshijie. The workflow is titled "RNASeq workflow" and is of type "Galaxy". The interface includes navigation tabs (Home, Users, Groups, Workflows, Files, Packs, Services, Topics), a search bar, and a sidebar with user information (Xiaoxiaoshijie) and workflow details (Workflow Type: Galaxy, Original Uploader: Xiaoxiaoshijie, License: All versions of this Workflow are not licensed).

Galaxy

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]

my experiment

Home | Users | Groups | Workflows | Files | Packs | Services | Topics

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: Xiaoxiaoshijie

License: All versions of this Workflow are not licensed.

Credits (0)

New/Upload

Workflow GO

Xiaoxiaoshijie

My Profile [edit] | My Messages | My Memberships | My History | My News

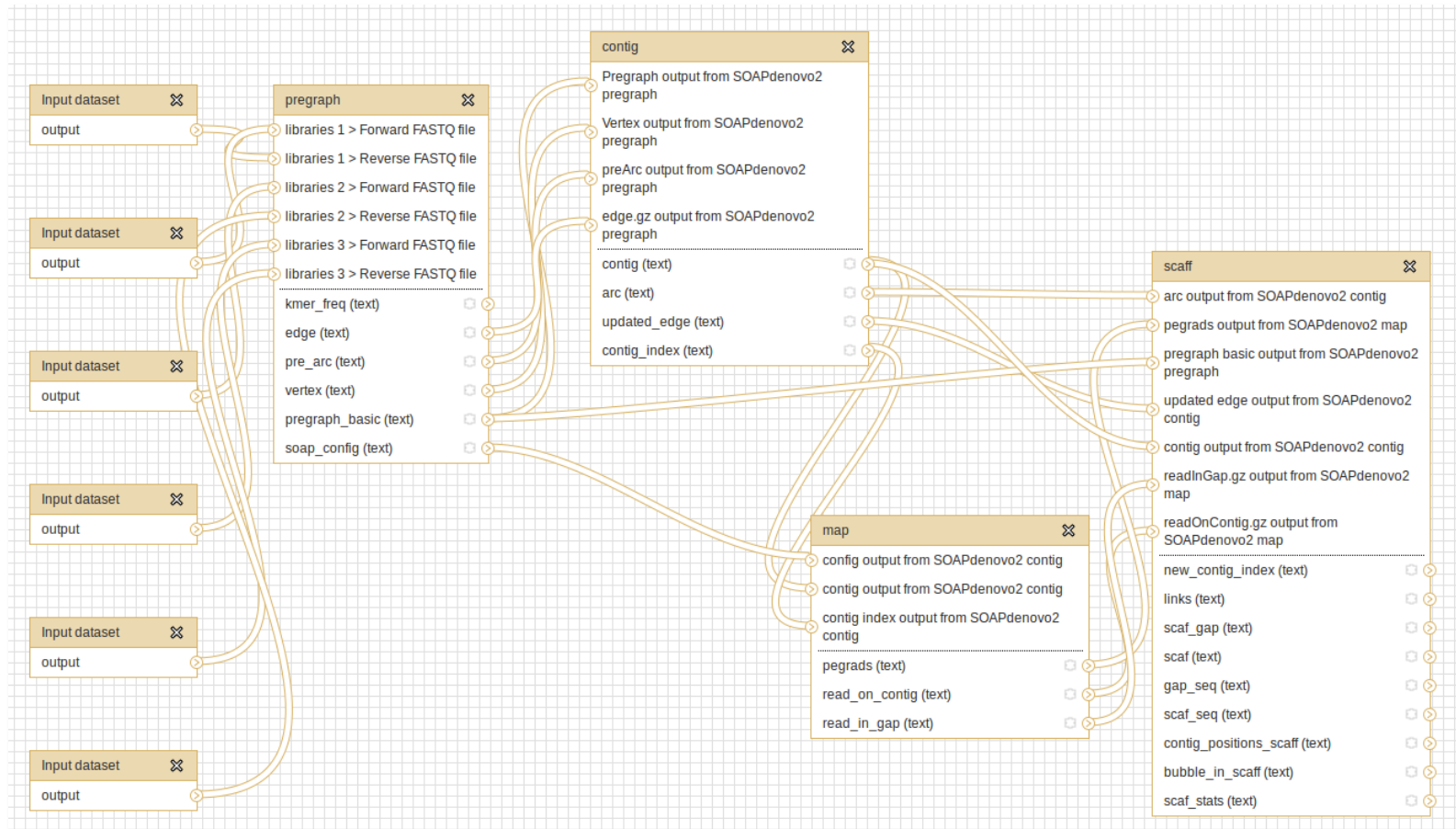
My Stuff

0 Friends | 0 Groups | 3 Workflows

Workflows

GetCities | GetCities2 | RNASeq workflow

# SOAPdenovo2 Galaxy workflow





# CBIIT GigaGalaxy structure

*Bioinformatics  
Development*



*Biomedical and bioinformatics research*

UCSC Genome Bioinformatics



*Publishing*



# (GIGA)<sup>n</sup> SCIENCE

Now launched...

## LARGE-SCALE DATA JOURNAL/DATABASE



IN CONJUNCTION WITH:



**EDITOR-IN-CHIEF: LAURIE GOODMAN, PHD**

**EDITOR: SCOTT EDMUNDS, PHD**

**COMMISSIONING EDITOR: NICOLE NOGOY, PHD**



[www.gigasciencejournal.com](http://www.gigasciencejournal.com)



**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, Mariys 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, Folkner 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

# GigaScience is go...

Cochrane et al. *GigaScience* Preview  
<http://www.gigascejournal.com>

(GIGA)<sup>n</sup>  
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 of 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



The screenshot shows the GigaDB website homepage. At the top left is the logo for (GIGA)<sup>n</sup> DB Beta, with the tagline "Revolutionizing data dissemination, organization, and use". To the right of the logo are navigation links: Home, About, Contact, and Terms of use. The main content area has a blue background with a large "GigaDB" title. Below the title is a search bar with the placeholder text "SEARCH by Species, DOI, Data Type" and a "GO" button. Under the search bar, a paragraph states: "GigaDB contains discoverable, trackable, and citable data that have been assigned DOIs and are available for public download and use." At the bottom of the page, there are logos for (GIGA)<sup>n</sup> SCIENCE and BGI (华大基因). To the right of these logos are social media links: "Be a fan on Facebook", "Follow us on Twitter", "Follow us on Sina", and "GigaBlog".

(GIGA)<sup>n</sup> DB Beta  
Revolutionizing data dissemination, organization, and use

Home | About | Contact | Terms of use

# GigaDB

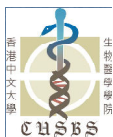
SEARCH by Species, DOI, Data Type **GO**

GigaDB contains discoverable, trackable, and citable data that have been assigned DOIs and are available for public download and use.

(GIGA)<sup>n</sup> SCIENCE 华大基因 BGI

Home | About | Contact | Terms of use

[Be a fan on Facebook](#) [Follow us on Twitter](#) [Follow us on Sina](#) [GigaBlog](#)



[www.gigaDB.org](http://www.gigaDB.org)







# 40 Datasets with DOI<sup>®</sup>s

## Invertebrate

### Ant

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

### Roundworm

### Schistosoma

### Silkworm



## Human

### Asian individual (YH) v1+v2

- DNA Methylome
- Genome Assembly
- **Transcriptome**

### Cancer (14TB)

### Hep B infected exomes

### Single Cell Bladder Cancer

### Ancient DNA

### Sagqaq Eskimo

### Aboriginal Australian

CUSS



## Vertebrates

### Giant panda

### Macaque

- Chinese rhesus
- Crab-eating

### Mini-Pig

### Naked mole rat

### Parrot

### 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



Sequencing of Life

# GigaDB v2 export to CBIIT GigaGalaxy

## Results

| Species | Dataset type | Dataset  | Sample  | File type | File format   | File name   | Include in download                 |
|---------|--------------|--|---|-----------|---|---|-------------------------------------|
| Human   | Genomic      | <a href="#">10.5524/1000010</a> - Genomic sequence from an Aboriginal Australian | Biosample: <a href="#">259765</a> - Aboriginal Australian human | SNPs      | vcf  | AusAboriginal.hg19.var.filtered.snps.sampled.vcf.gz | <input checked="" type="checkbox"/> |
| Human   | Genomic      | <a href="#">10.5524/1000010</a> - Genomic sequence from an Aboriginal Australian | Biosample: <a href="#">259765</a> - Aboriginal Australian human | SNPs      | vcf  | AusAboriginal.hg19.var.filtered.snps.vcf.gz         | <input checked="" type="checkbox"/> |

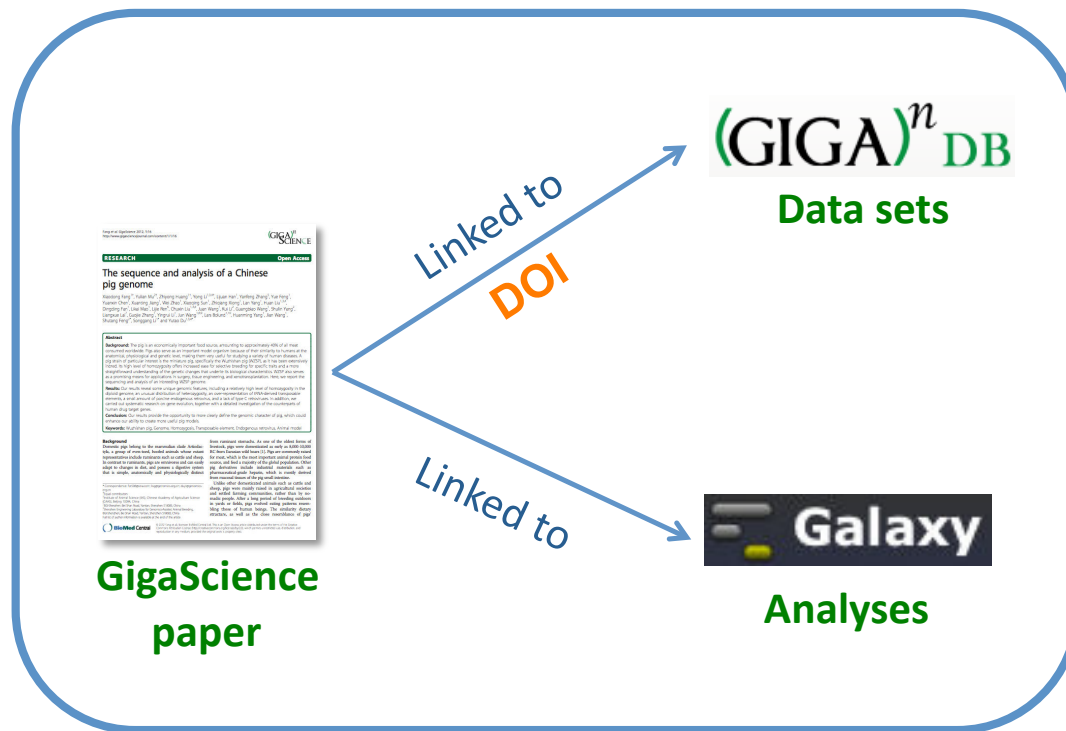
Link to GigaDB landing pages

Link to sample if applicable

Download to Galaxy

Link to our documentation on file formats

# How are we supporting data reproducibility?



Community tools for  
data reproduction and reuse





# CBIIT GigaGalaxy

# Data, Data, Data...



(GIGA)<sup>n</sup>



Tin-Lap Lee, CUHK



# Acknowledgements

- **Lee Lab (CUHK)**

- Huayan Gao



- **GigaScience**

- Scott Edmunds
  - Peter Li
  - Tam Sneddon



- **myExperiment**

- Finn Bacall
  - Dave De Roure



- **NBIC**

- Kostas Karasavvas



- **BGI-Hong Kong**

- Dennis Chan
  - Edmond Leung

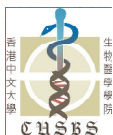
- BGI-Shenzhen**

- Ruiqiang Li
    - Ruibang Luo
    - Haofu Wu
    - SOAP team members



- **Galaxy team**

- Nate Coraor



Thank you

