

# 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 addition 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:

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



# GDSAP Structure

*Tool  
Development*



*Biomedical and bioinformatics research*

UCSC Genome Bioinformatics



*Publishing*

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

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



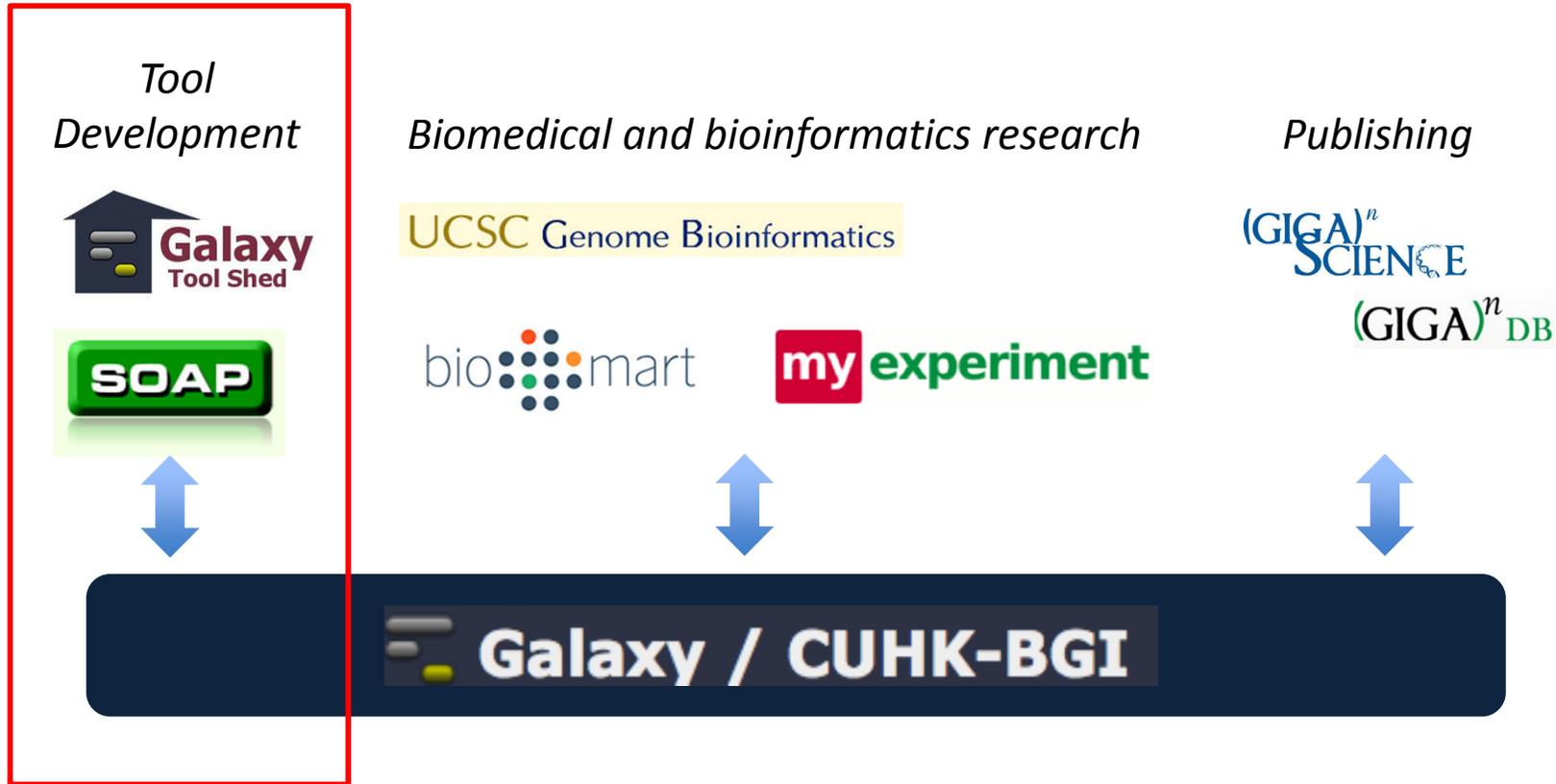
# Galaxy/CUHK-BGI

The screenshot shows a Firefox browser window with several tabs open, including 'CUHK-BGI Innovation Institute...', 'CUHK Communications and P...', 'Events/GCC2012/Abstracts - G...', 'Galaxy', and 'CUHK-BGI Innovation Institute...'. The address bar shows the URL '61.244.113.150/galaxy/'. The main content area displays the Galaxy/CUHK-BGI logo and navigation menu. The left sidebar contains a 'Tools' section with categories like 'CBIIIT TOOLS', 'CUHK-BGI TOOLBOX', 'SOAP Family', 'GALAXY TOOLS', and various tool options. The main content area features the CUHK-BGI logo, the text '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', and a 'History' panel on the right. The 'History' panel shows '0 bytes' and a message: 'Your history is empty. Click 'Get Data' on the left pane to start'. Below the main content, there are sections for 'News' (We are implementing SOAP tools.) and 'Announcement' (Genomic Data Submission and Analytical Platform (GDSAP) is under development.). The footer of the page includes the text '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.', and 'Last Updated: July 20, 2012'.

<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	
✓ <b>SOAP3/GPU</b>	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.
✓ <b>SOAPaligner/soap2</b>	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.
✓ <b>SOAPsplice</b> <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.
✓ <b>SOAPsnp</b>	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.
✓ <b>SOAPdenovo</b>	SOAPdenovo, a short read de novo assembly tool, is a package for assembling short oligonucleotide into contigs and scaffolds.
✓ <b>SOAPindel</b>	SOAPindel is developed to find the insertion and deletion specially for re-sequence technology.
✓ <b>SOAPsv</b>	SOAPsv is a program for detecting the structural variation .
✓ <b>SOAP v1</b>	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.



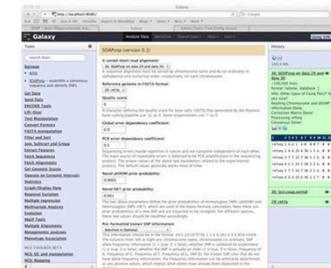
# Galaxy Tool Shed

- 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



# Implement: SOAPsnp

The screenshot displays the Galaxy web interface for the SOAPsnp tool. The browser address bar shows `http://61.244.113.150/galaxy/root`. The page title is "Galaxy / CUHK-BGI". The main content area is titled "SOAPsnp (version 0.1)" and contains the following configuration options:

- A sorted short read alignment:** 54: SOAPdenovo\_config.. and data 3
- Reference genome in FASTA format:** 51: ref.fa
- Quality score:** 0
- Global error dependency coefficient:** 0.9
- PCR error dependency coefficient:** 0.5
- Novel althOM prior probability:** 0.0005
- Novel HET prior probability:** 0.001

The right sidebar shows a history of jobs, including "53: SOAPsnp on data 51 and data 50" and "52: test.mat".

# Implement: SOAPdenovo configuration file

The screenshot displays the Galaxy web interface for the CUHK-BGI instance. The main panel shows the configuration for the 'SOAPdenovo\_config (version 0.1)' tool. The configuration parameters are as follows:

- Maximal read length: 50
- Average insert size: 200
- If sequence needs to be reversed: 0
- In which part(s) the reads are used: 3
- In which order the reads are used while scaffolding: 1
- Fastq file for read 1: 2: illumina\_90\_200\_1.fq
- Fastq file for read 2 always follows fastq file for read 1: 3: illumina\_90\_200\_2.fq
- Fasta file for read 1: Selection is Optional
- Fasta file for read 2 always follows fasta file for read 1: Selection is Optional
- Fastq file for single reads: Selection is Optional
- Fasta file for single reads: Selection is Optional

The right-hand side of the interface shows a 'History' panel with two entries:

- 53: SOAPsnp on data 51 and data 50**  
~100,000 lines  
format: tabular, database: ?  
Info: Other types of Fastq files?  
Are you sure?  
Reading Chromosome and dbSNP information Done.  
Correction Matrix Done!  
Processing refseq  
Consensus Done!
- 52: test.mat**  
1,845 lines  
format: tabular, database: ?  
Info: uploaded tabular file

The left-hand side of the interface shows a 'Tools' panel with various categories of tools, including CBIIT TOOLS, CUHK-BGI TOOLBOX, SOAP Family, GALAXY TOOLS, and various utility tools like Get Data, Send Data, ENCODE Tools, etc.

# Implement: SOAPdenovo

The screenshot shows the Galaxy web interface for the SOAPdenovo tool. The browser address bar displays `http://61.244.113.150/galaxy/root`. The page header includes the Galaxy logo, navigation tabs (Analyze Data, Workflow, Shared Data, Admin, Help, User), and a memory usage indicator (Using 395.3 Mb).

**Tools** (Options):

- CBIIT TOOLS
- CUHK-BGI TOOLBOX
- SOAP Family
  - msort msort
  - SOAPSnp - assemble a consensus sequence and identify SNPs
  - SOAPdenovo\_config - SOAPdenovo config file
  - SOAPdenovo SOAPdenovo
- GALAXY TOOLS
  - 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
  - Graph/Display Data
  - Regional Variation

**SOAPdenovo (version 1.0.0)**

Select kmer version:  
127mer Version

Support large kmer up to 127 to utilize long reads. Three version are provided. Please notice that, with longer kmer, the quantity of nodes would decrease significantly, thus the memory consumption is usually smaller than double with shifted version.

SOAPdenovo configuration file:  
54: SOAPdenovo\_config.. and data 3

Execute

**What it does** SOAPSnp is a member of the SOAP (Short Oligonucleotide Analysis Package). Despite its name, the program is a resequencing utility that can assemble consensus sequence for the genome of a newly sequenced individual based on the alignment of the raw sequencing reads on the known reference. The SNPs can then be identified on the consensus sequence through the comparison with the reference. In the first Asian genome resequencing project, evaluation of SOAPSnp result on Illumina HapMap 1M BeadChip Duo genotyping sites shows great accuracy. Over 99% of the genotyping sites are covered at over 99.9% consistency. Further PCR plus Sanger sequencing of the inconsistent SNP sites confirmed majority of the SOAPSnp results.

SOAPSnp uses a method based on Bayes'theorem (the reverse probability model) to call consensus genotype by carefully considering the data quality, alignment, and recurring experimental errors. All these kinds of information was integrated into a single quality score for each base in PHRED scale to measure the accuracy of consensus calling. Currently, it supports the alignment format of SOAPaligner.

**History** (Options):

- 53: SOAPSnp on data 51 and data 50  
~100,000 lines  
format: tabular, database: ?  
Info: Other types of Fastq files??  
Are you sure?  
Reading Chromosome and dbSNP information Done.  
Correction Matrix Done!  
Processing refseq  
Consensus Done!
- 52: test.mat  
1,845 lines  
format: tabular, database: ?  
Info: uploaded tabular file

Table 1 (from History 53):

1	2	3	4	5	6	7	8	9	10	11	12
refseq 1	G	G	1	G	0	0	0	0	T	0	0
refseq 2	A	A	23	A	56	1	1	G	0	0	!
refseq 3	T	T	23	T	56	1	1	G	0	0	!
refseq 4	C	C	23	C	56	2	2	G	0	0	!
refseq 5	T	T	23	T	56	2	2	G	0	0	!
refseq 6	G	G	23	G	56	2	2	T	0	0	!

Table 2 (from History 52):

1	2	3
0	0	2.5000000000000000e-01
0	1	2.5000000000000000e-01
0	2	2.5000000000000000e-01
0	3	2.5000000000000000e-01

# GDSAP structure

*Bioinformatics  
Development*



*Biomedical and bioinformatics research*

UCSC Genome Bioinformatics



*Publishing*

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

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



# my experiment How does it work?

The screenshot shows the myExperiment website interface. At the top left is the myExperiment logo. A yellow banner contains the text: "myExperiment makes it easy to find, use and share scientific workflows and other Research Objects, and to build communities." Below this is a search bar with a dropdown menu set to "All" and a "Search" button. The main content area is divided into three columns. The left column, titled "First time visitor? Try these videos:", lists "Project Introduction" and "Bioinformatics Case Study". Below this, "Use myExperiment to..." lists several actions: "Find Workflows", "Share Your Workflows and Files", "Create and Find Packs of Items", "Find People and Make Friends", "Create and Join Groups", "Build your Profile and Reputation", "Tag and Rate things", and "Write Reviews and Comments". The middle column, titled "Explore", features a workflow diagram with a "Find Workflows" button and a section for "About myExperiment" with links to "Join the Mailing List", "myExperiment Publications", "For Developers", "Give us Feedback", and "The BioCatalogue Project". The right column, titled "or Login:", contains a login form with fields for "Username or Email:", "Password:", and "Remember me: [checkbox]". It also includes an "Or use OpenID:" field with a plus sign icon and a "Login" button, along with a "Forgot Password?" link.

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



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)

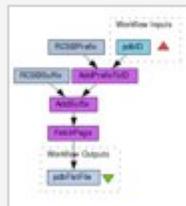
Original Uploader

Credits: Tomoinn

License: Creative Commons Attribution 3.0 Unported License



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

CBIIIT 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

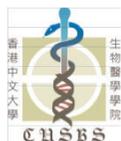
Execute

#### 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



香港中文大學  
生物醫學學院  
CUHKS



# Galaxy workflow

The screenshot shows the myExperiment website interface. The browser address bar displays the URL: [http://www.myexperiment.org/galaxy?galaxy\\_url=https://main.g2.bx.psu.edu/](http://www.myexperiment.org/galaxy?galaxy_url=https://main.g2.bx.psu.edu/). The page title is "myExperiment - Workflows". The navigation menu includes "Home", "Users", "Groups", "Workflows", "Files", "Packs", "Services", and "Topics". The "Workflows" section is active, showing a search for "Basic RNA-Seq Analysis - Differential Expression (Functional Genomics Workshop 2012) (v1)".

**Search filter terms:** [Search box]

**Sort by:** Rank

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

**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)

**Workflow Details:**

- Galaxy** **Basic RNA-Seq Analysis - Differential Expression (Functional Genomics Workshop 2012) (v1)** [View] [Download (v1)]
- Original Uploader:** David De Roure
- Created:** 16/07/12 @ 21:20:44 | **Last updated:** 16/07/12 @ 21:27:56
- License:** No license
- Description:** 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



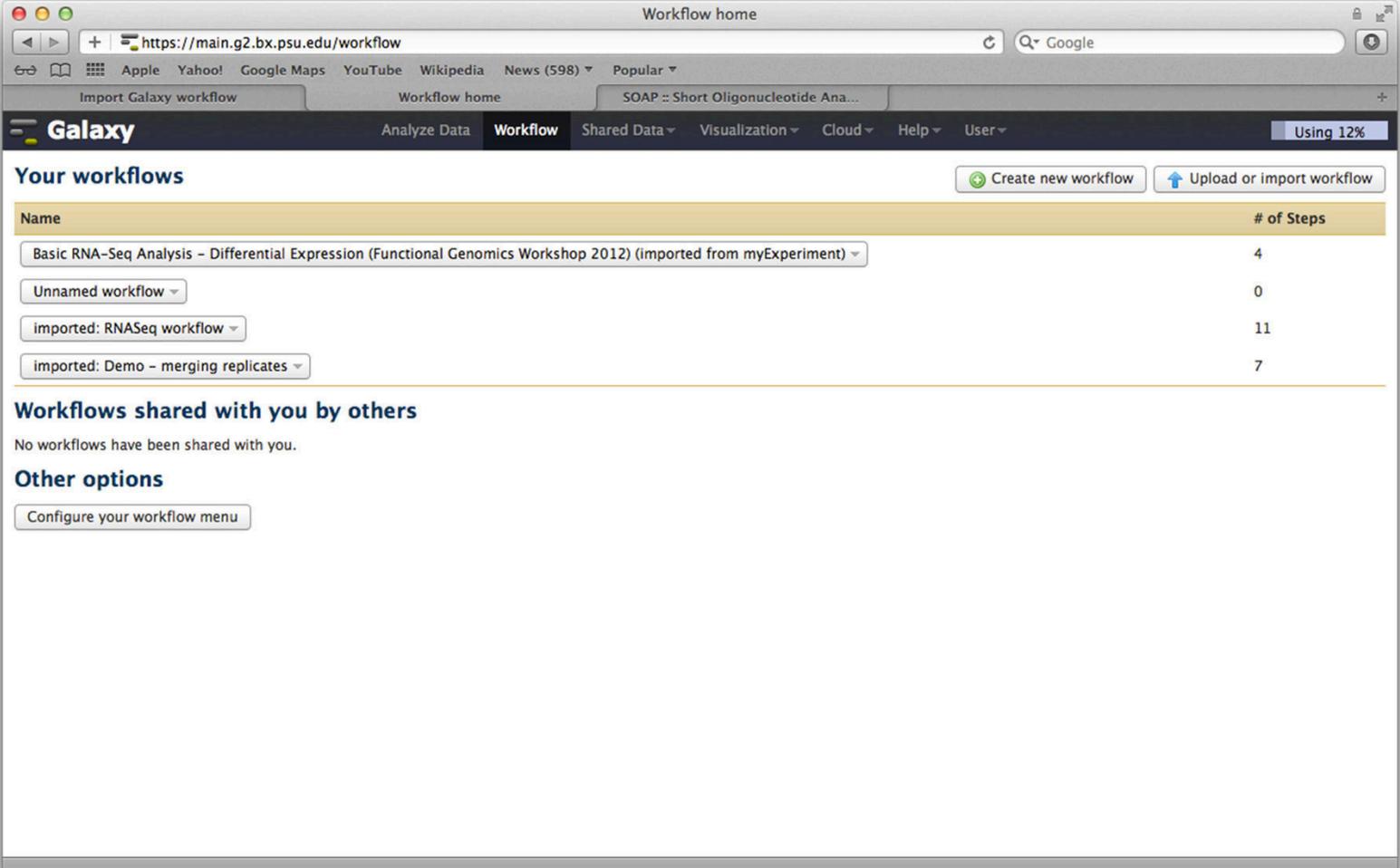
# Import (1)

The image displays two overlapping browser windows from the Galaxy platform. The background window shows a workflow page for 'Basic RNA-Seq Analysis' published by mejia-guerra. The foreground window shows the 'Workflow Canvas' for an imported RNA-Seq workflow. The canvas contains several tool steps connected by data lines:

- Two 'FASTQ Groomer' tools, each receiving 'File to groom' input and producing 'output\_file' (fastqsanger, fastqcssanger, fastqsolexa, fastqillumina).
- A 'Map with Bowtie for Illumina' tool receiving 'Forward FASTQ file' and 'Reverse FASTQ file' inputs, producing 'output (sam)', 'output\_suppressed\_reads\_l (fastq)', 'output\_suppressed\_reads\_r (fastq)', 'output\_unmapped\_reads\_l (fastq)', and 'output\_unmapped\_reads\_r (fastq)'. Its output is connected to a 'Map with BWA for Illumina' tool.
- A 'Tophat for Illumina' tool receiving 'RNA-Seq FASTQ file' and 'accepted\_hits (bam)' inputs, producing 'insertions (bed)', 'deletions (bed)', 'junctions (bed)', and 'accepted\_hits (bam)'. Its output is connected to a 'flagstat' tool.
- A 'flagstat' tool receiving 'BAM File to Convert' input and producing 'output1 (txt)'. Its output is connected to another 'flagstat' tool.
- A 'Cufflinks' tool receiving 'SAM or BAM file of aligned reads' and 'Global model' inputs, producing 'genes\_expression (tabular)', 'transcripts\_expression (tabular)', 'assembled\_isoforms (gtf)', and 'total\_map\_mass (txt)'. Its output is connected to a 'BAM File to Convert' tool.
- A 'BAM File to Convert' tool receiving 'output1 (txt)' and producing 'output1 (txt)'.

The interface includes a 'Tools' sidebar on the left with categories like 'Get Data', 'Send Data', 'ENCODE Tools', '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', and 'Evolution'. The right sidebar shows 'Details' for the selected tool, including 'What it does', 'Citation', and 'Add actions to this step'.

# Import (2)



The screenshot shows the Galaxy workflow management interface. The browser address bar displays `https://main.g2.bx.psu.edu/workflow`. The page title is "Workflow home". The navigation menu includes "Analyze Data", "Workflow", "Shared Data", "Visualization", "Cloud", "Help", and "User". A "Using 12%" indicator is visible in the top right.

### Your workflows

Buttons: [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

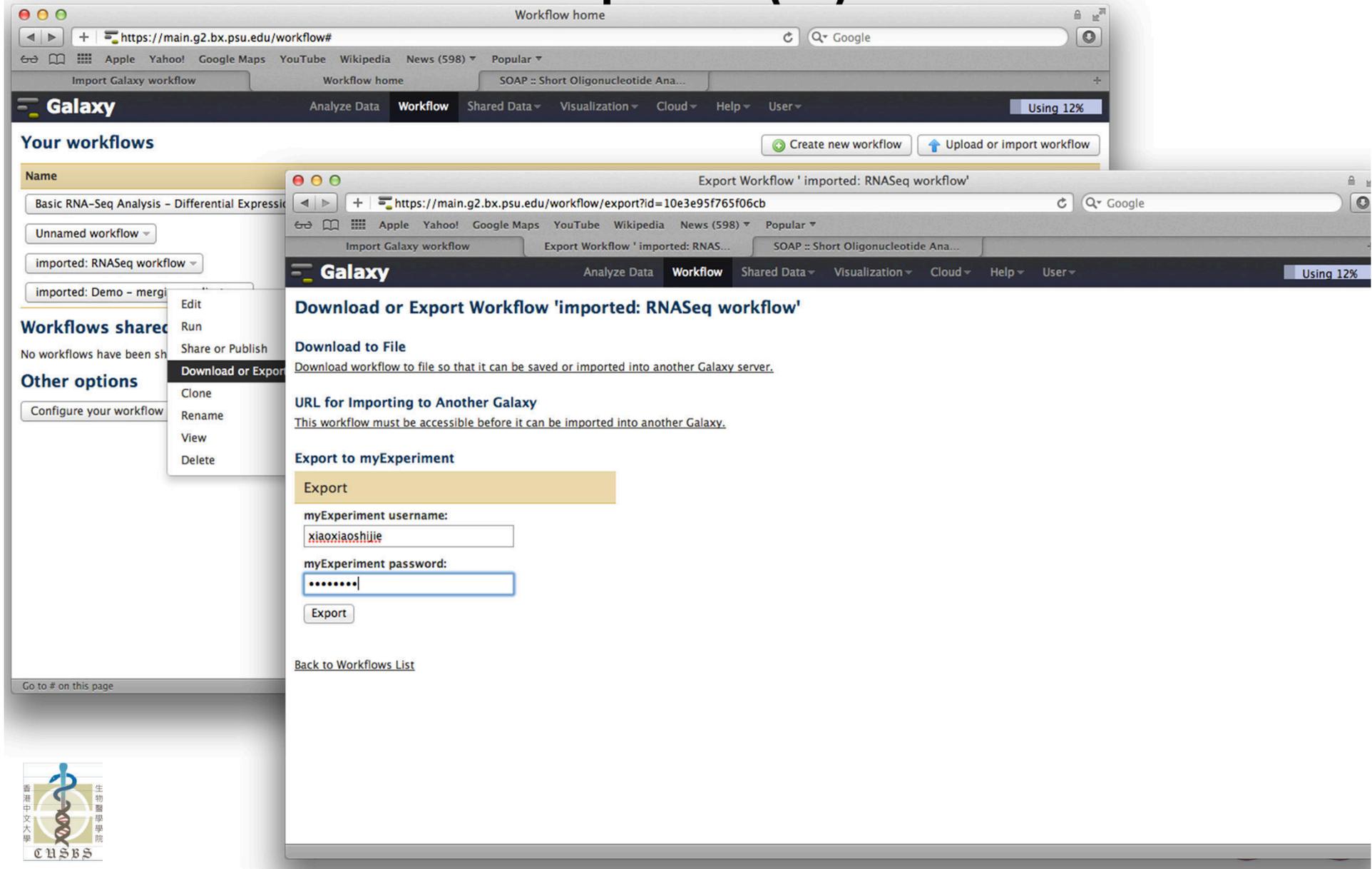
### Workflows shared with you by others

No workflows have been shared with you.

### Other options

[Configure your workflow menu](#)

# Export (1)



The image shows two overlapping browser windows from the Galaxy web interface. The background window is titled "Workflow home" and shows the "Your workflows" section. A workflow named "imported: RNASeq workflow" is selected, and a context menu is open over it with "Download or Export" highlighted. The foreground window is titled "Export Workflow 'imported: RNASeq workflow'" and shows the export options. The "Export" option is selected, leading to a form for "myExperiment" with fields for username and password, and an "Export" button.

**Workflow home**

Workflow home | SOAP :: Short Oligonucleotide Ana... | Using 12%

### Your workflows

Create new workflow | Upload or import workflow

Name
Basic RNA-Seq Analysis - Differential Expressi...
Unnamed workflow
imported: RNASeq workflow
imported: Demo - mergi

Workflows shared: No workflows have been shared

Other options: Configure your workflow

- Edit
- Run
- Share or Publish
- Download or Export**
- Clone
- Rename
- View
- Delete

**Export Workflow 'imported: RNASeq workflow'**

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](#)

Go to # on this page



CUHK  
香港中文大學  
生物醫學學院

# Export (2)

The image shows two overlapping browser windows. The top window is the Galaxy interface, displaying a notification: "Workflow 'imported: RNASeq workflow' successfully exported to myExperiment. Click here to view the workflow on myExperiment. Return to workflow list." The bottom window is the myExperiment website, showing the details for a workflow entry titled "RNASeq workflow".

**Galaxy Window:**  
URL: [https://main.g2.bx.psu.edu/workflow/export\\_to\\_myexp?id=10e3e95f765f06cb](https://main.g2.bx.psu.edu/workflow/export_to_myexp?id=10e3e95f765f06cb)  
Notification: Workflow 'imported: RNASeq workflow' successfully exported to myExperiment. Click here to view the workflow on myExperiment. Return to workflow list.

**myExperiment Window:**  
URL: <http://www.myexperiment.org/workflows/3036.html>  
Page Title: myExperiment - Workflows - RNASeq workflow (Xiaoxiaoshijie) [Galaxy Workflow]  
Navigation: Home, Users, Groups, **Workflows**, Files, Packs, Services, Topics  
Search: All [Search]  
Breadcrumbs: Home > Workflows > RNASeq workflow  
Actions: Upload New Version, Manage Workflow Entry, Delete Workflow Entry  
Workflow Entry: RNASeq workflow  
Created at: 18/07/12 @ 02:32:15  
Metadata: 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.  
Credits: (0)  
Right Sidebar:  
New/Upload: Workflow [GO]  
User: Xiaoxia...  
My Profile [edit], My Messages, My Memberships, My History, My News  
My Stuff: 0 Friends | 0 Groups | 3 Workflows  
Workflows: GetCites, GetCites2, RNASeq workflow



# GDSAP structure

*Bioinformatics  
Development*



*Biomedical and bioinformatics research*

UCSC Genome Bioinformatics



*Publishing*

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

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



(GIGA)<sup>n</sup> Now taking submissions...  
SCIENCE

Large-Scale Data  
Journal/Database



In conjunction with:



**Editor-in-Chief:** Laurie Goodman, PhD

**Editor:** Scott Edmunds, PhD

**Assistant Editor:** Alexandra Basford, PhD



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



# GigaScience is go...

**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, Jal 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 Dhani, 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](#)

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



## 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

(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)



# 37 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

### - Saqqaq Eskimo

### Aboriginal Australian



## 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	<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

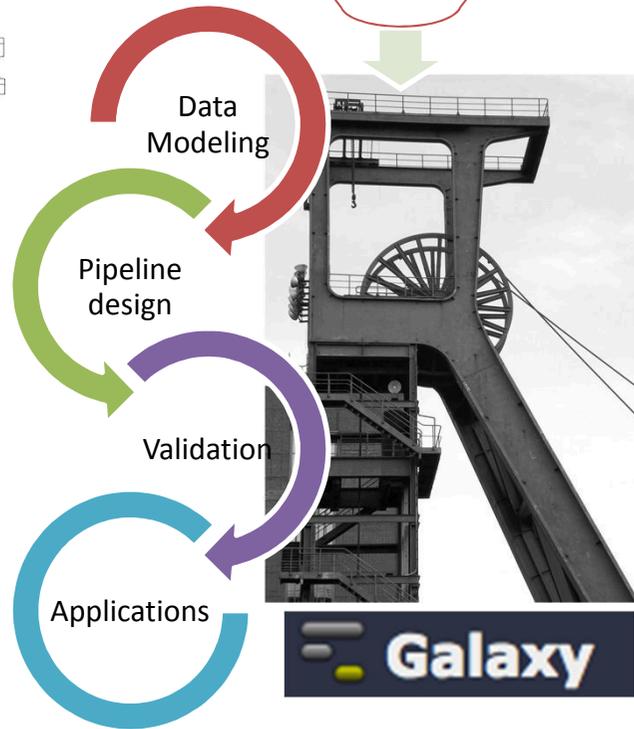
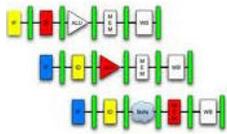
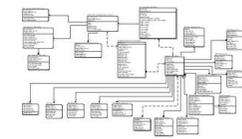
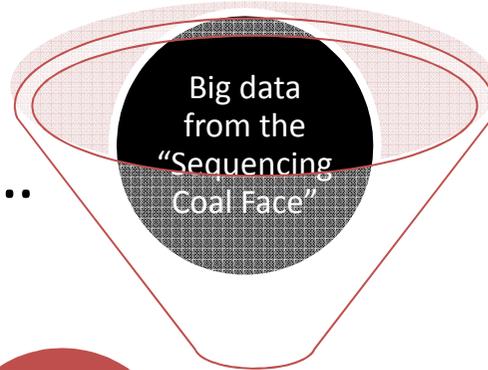
Download to Galaxy

Link to our documentation on file formats



# 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

