

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.”*



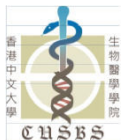
Genomic Data Submission and Analytical Platform(GDSAP)

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



 Galaxy / CUHK-BGI

Galaxy/CUHK-BGI

Firefox

CUHK-BGI Innovation Institute... CUHK Communications and P... Events/GCC2012/Abstracts - G... Galaxy CUHK-BGI Innovation Institute... Galaxy

61.244.113.150/galaxy/ cbiit.cuhk

Galaxy / CUHK-BGI Analyze Data Workflow Shared Data Help User Using 0 bytes

Tools Options

search tools

CBIIT TOOLS

CUHK-BGI TOOLBOX

SOAP Family

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

 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 

Galaxy @ CUHK-BGI IIT

News
We are implementing SOAP tools.

Announcement
Genomic Data Submission and Analytical Platform (GDSAP) is under development.

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.

Last Updated: July 20, 2012

History Options

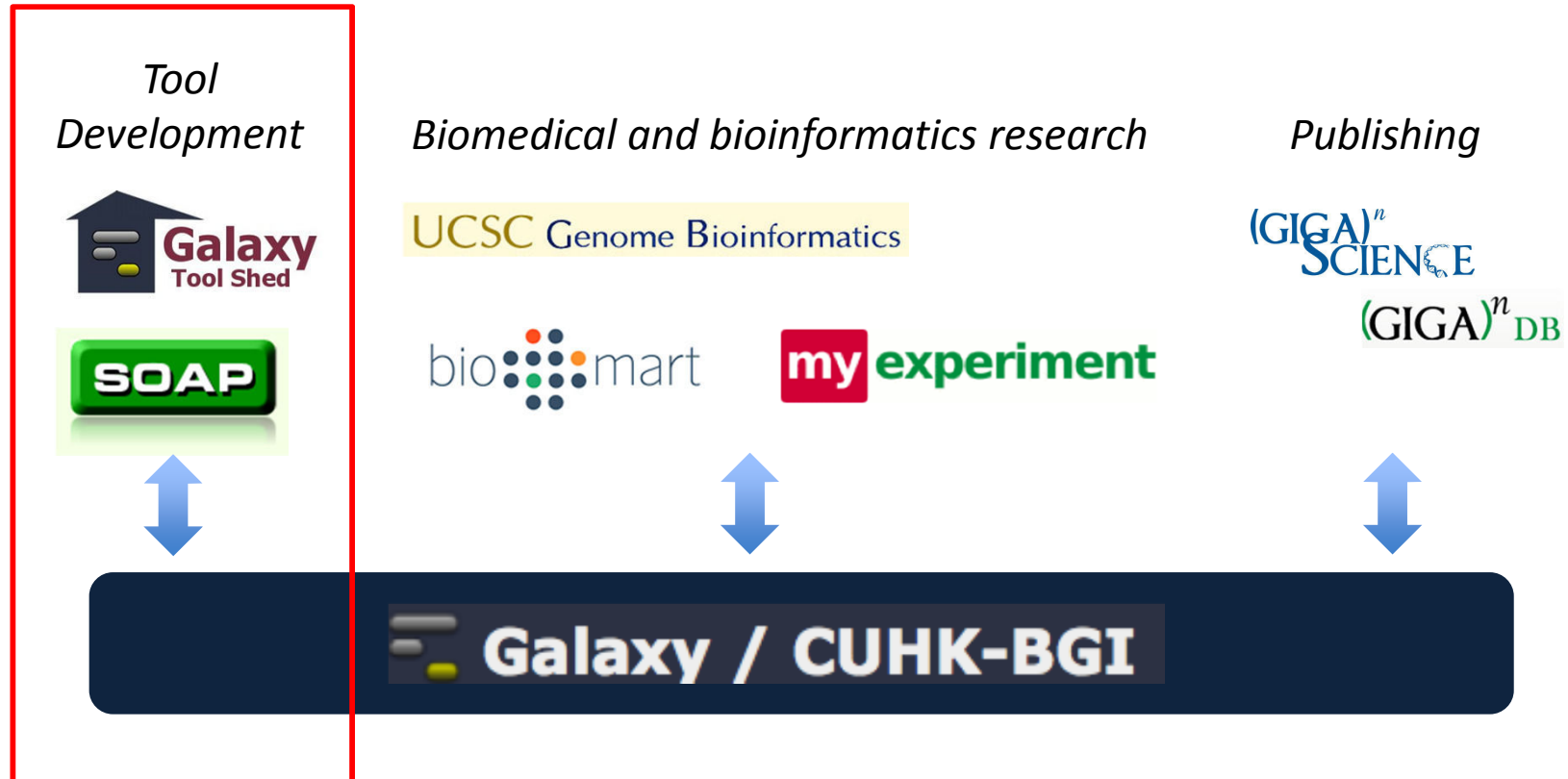
0 bytes

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

<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.

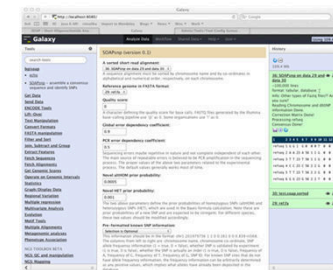
Software	
✓ SOAP3/GPU	SOAP3 is a GPU-based software for aligning short reads with a reference sequence. It can find all alignments with k mismatches, where k is chosen from 0 to 3. When compared with its previous version SOAP2, SOAP3 can be up to tens of times faster.
✓ SOAPaligner/soap2	SOAPaligner/soap2 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

The screenshot shows the Galaxy web interface with the SOAPsnp tool configuration page. 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 sections:

- A sorted short read alignment:** A text input field contains "54: SOAPdenovo_config.. and data 3". Below it, a description states: "A sequence alignment must be sorted by chromosome name and by co-ordinates in alphabetical and numerical order, respectively, on each chromosome."
- Reference genome in FASTA format:** A text input field contains "51: ref.fa".
- Quality score:** A text input field contains "0". Below it, a description states: "A character defining the quality score for base calls. FASTQ files generated by the Illumina base-calling pipeline use '@' as 0. Some organisations use 'I' as 0."
- Global error dependency coefficient:** A text input field contains "0.9".
- PCR error dependency coefficient:** A text input field contains "0.5". Below it, a description states: "Sequencing errors may be repetitive in nature and not complete independent of each other. The main source of repeatable errors is believed to be PCR amplification in the sequencing process. The proper values of the above two parameters related to the experimental process. The default values generally works most of time."
- Novel althOM prior probability:** A text input field contains "0.0005".
- Novel HET prior probability:** A text input field contains "0.001". Below it, a description states: "The two above parameters define the prior probabilities of homozygous SNPs (althOM) and heterozygous SNPs (HET), which are used in the Bayes formula calculation. Note these are prior probabilities of a new SNP and are expected to be stringent. For different species, these two values should be modified accordingly."
- Pre-formatted known SNP information:** (This section is partially visible at the bottom of the configuration area).

The right sidebar shows the "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! Below this is a table with 12 columns and 6 rows of data.
- 52: test.mat**: 1,845 lines, format: tabular, database: ?. Info: uploaded tabular file. Below this is a table with 3 columns and 5 rows of data.

Implement: SOAPdenovo configuration file

The screenshot displays the Galaxy web interface for CUHK-BGI. The main panel shows the configuration for the SOAPdenovo tool (version 0.1). The configuration includes fields for Maximal read length (50), Average insert size (200), and a checkbox for reversing sequences (unchecked). It also specifies the parts of the reads to be used (3) and the order of reads (1). The Fastq files for read 1 and read 2 are both set to 'Illumina_90_200_1.fq'. The Fasta files for read 1 and read 2 are set to 'Selection is Optional'. The Fastq file for single reads is also set to 'Selection is Optional'. The Fasta file for single reads is set to 'Selection is Optional'.

SOAPdenovo_config (version 0.1)

Maximal read length:

Average insert size:

If sequence needs to be reversed:

In which part(s) the reads are used:

In which order the reads are used while scaffolding:

Fastq file for read 1:

Fastq file for read 2 always follows fastq file for read 1:

Fasta file for read 1:

Fasta file for read 2 always follows fasta file for read 1:

Fastq file for single reads:

Fasta file for single reads:

History

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!

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

52: test.mat
1,845 lines
format: tabular, database: ?
Info: uploaded tabular file

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

Implement: SOAPdenovo

The screenshot displays the Galaxy web interface at the URL <http://61.244.113.150/galaxy/root>. The interface includes a top navigation bar with links to 'Analyze Data', 'Workflow', 'Shared Data', 'Admin', 'Help', and 'User'. A status bar at the top right indicates 'Using 395.3 Mb'.

The main content area is titled 'SOAPdenovo (version 1.0.0)'. It contains the following sections:

- Select kmer version:** A dropdown menu set to '127mer Version'.
- SOAPdenovo configuration file:** A dropdown menu set to '54: SOAPdenovo_config.. and data 3'.
- Execute** button.

Below the configuration section, there is a detailed description of the tool:

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 re-sequencing 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.

The right sidebar shows the 'History' section 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 bottom left corner features the CUHKS logo, and the bottom right corner features the CUHK crest.

GDSAP structure

*Bioinformatics
Development*



Biomedical and bioinformatics research

UCSC Genome Bioinformatics



Publishing

(GIGA)ⁿ
SCIENCE

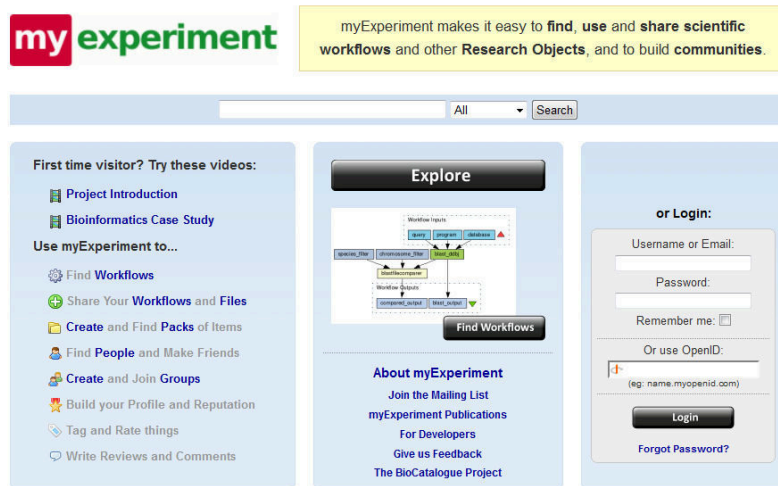
(GIGA)ⁿ
DB



 **Galaxy / CUHK-BGI**



How does it work?



- MyExperiment works as a repository for workflows.

- Taverna workflows.



- New: Galaxy workflows.



- GDSAP 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 \(3\)](#) | [Favourite 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)

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

Workflow Type
Taverna 1

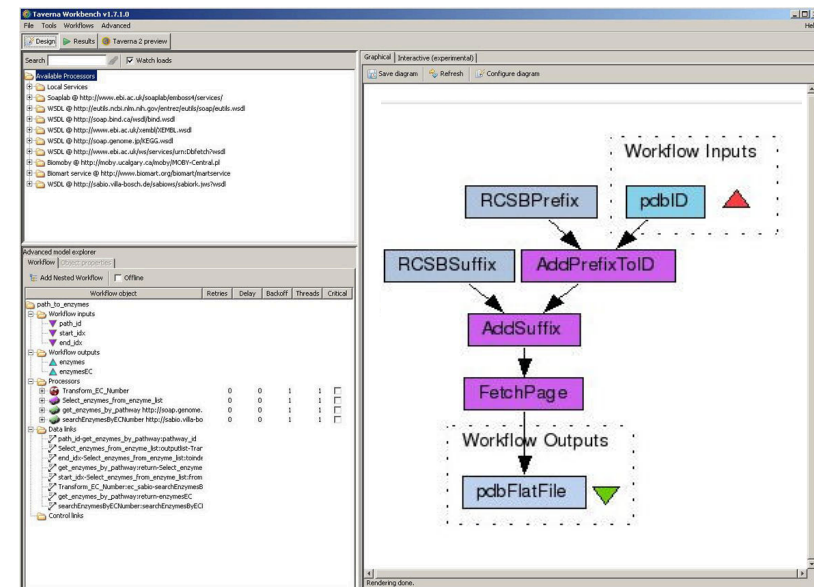
Original Uploader
 Alan Williams

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



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-BGI

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

The screenshot displays the myExperiment - Workflows web interface. The browser address bar shows the URL: http://www.myexperiment.org/galaxy?galaxy_url=https://main.g2.bx.psu.edu/. The page features a navigation bar with tabs: Home, Users, Groups, Workflows (selected), Files, Packs, Services, and Topics. A search bar is located below the navigation bar.

The main content area is titled "Workflows" and shows a list of 9 results. The left sidebar contains filter options:

- Search filter terms: [Search box]
- Sort by: Rank
- 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)

The main workflow listed is "Basic RNA-Seq Analysis - Differential Expression (Functional Genomics Workshop 2012) (v1)". It is created by David De Roure. The workflow details include:

- Created: 16/07/12 @ 21:20:44 | Last updated: 16/07/12 @ 21:27:56
- License: No license
- Original Uploader: David De Roure
- Rating: 0.0 / 5 (0 ratings) | Versions: 1 | Reviews: 0 | Comments: 0

The right sidebar contains user information for "Xiaoxia..." and links to My Profile, My Messages, My Memberships, My History, and My News. Below this is a "My Stuff" section showing 0 Friends, 0 Groups, and 2 Workflows. The "Workflows" section lists GetCities and GetCities2. The "My Favourites" section is also visible.

Import (1)

The image displays two overlapping browser windows. The background window shows the Galaxy website interface with a workflow titled "Basic RNA-Seq Analysis - Differential Expression (Functional Genomics Workshop 2012) (David De Roure)". The foreground window shows the Galaxy workflow editor for an imported RNA-Seq workflow. The workflow canvas contains several tools connected by lines, including "FASTQ Groomer", "Tophat for Illumina", "flagstat", "BAM File to Convert", "Cufflinks", and "Map with BWA for Illumina". The left sidebar lists various tool categories like "Get Data", "Send Data", "ENCODE Tools", etc. The right sidebar shows the "Details" panel for the selected tool, including its description and citation information.

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/Pack

Import
Import this workflow into myExperiment

Workflow Components

Inputs (3)

Steps (4)

Outputs (11)

Citations (0) Version History

Tools

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

FASTQ Groomer
File to groom
output_file (fastqsanger, fastqcssanger, fastqsolexa, fastqillumina)

FASTQ Groomer
File to groom
output_file (fastqsanger, fastqcssanger, fastqsolexa, fastqillumina)

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)

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)

flagstat
BAM File to Convert
output1 (txt)

Map with BWA for Illumina
Forward FASTQ file
Reverse FASTQ file
output (sam)

Cufflinks
SAM or BAM file of aligned reads
Global model (for use in Trac)
genes_expression (tabular)
transcripts_expression (tabular)
assembled_isoforms (gtf)
total_map_mass (txt)

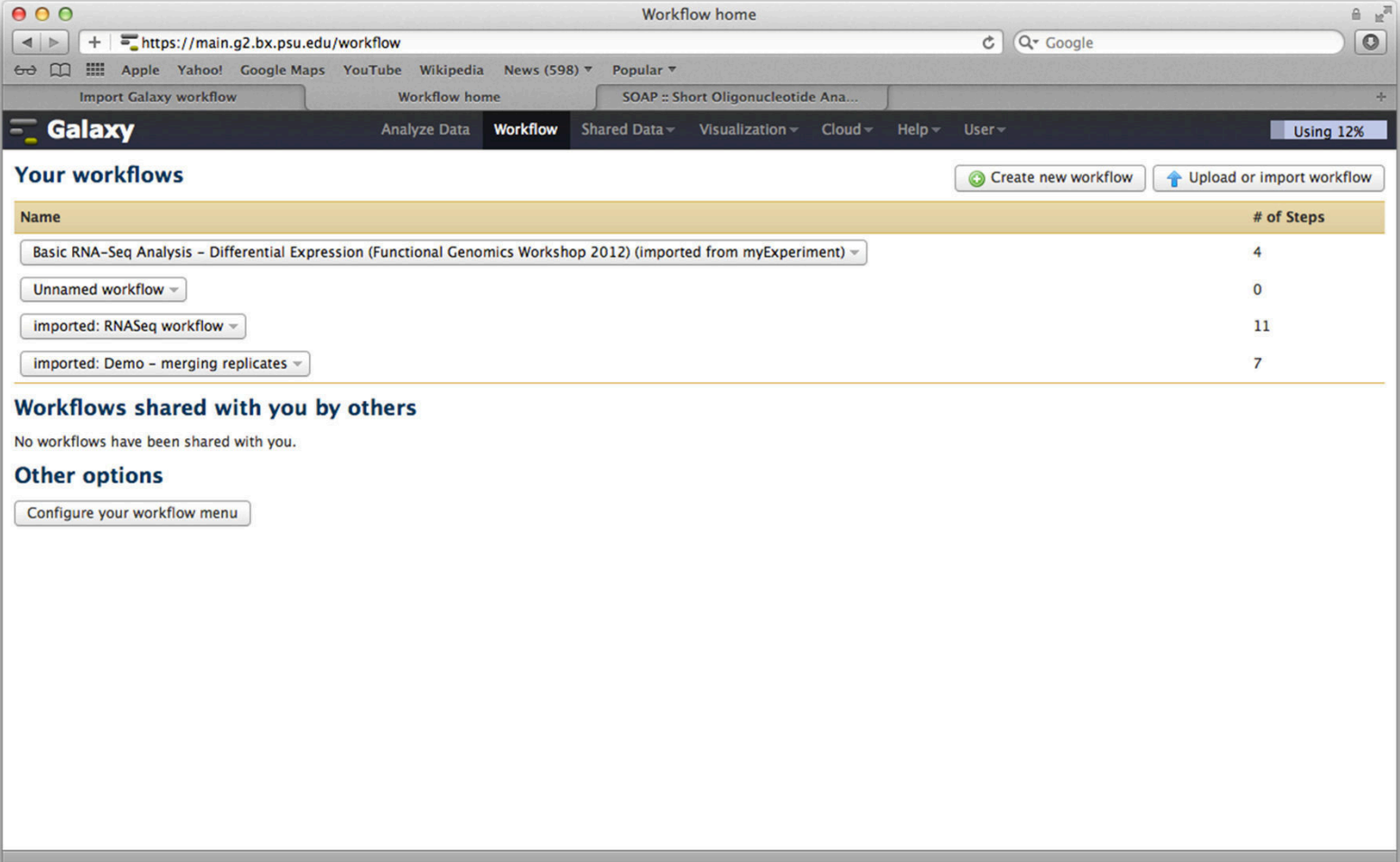
Details
output1 : Create
Add actions to this step: actions are applied when this workflow step completes.

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)



The screenshot shows the Galaxy web interface for workflow management. The browser address bar displays `https://main.g2.bx.psu.edu/workflow`. The page title is "Workflow home". The navigation bar includes "Import Galaxy workflow", "Workflow home", and "SOAP :: Short Oligonucleotide Ana...". The main header features the "Galaxy" logo and a navigation menu with "Analyze Data", "Workflow", "Shared Data", "Visualization", "Cloud", "Help", and "User". A status bar indicates "Using 12%".

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

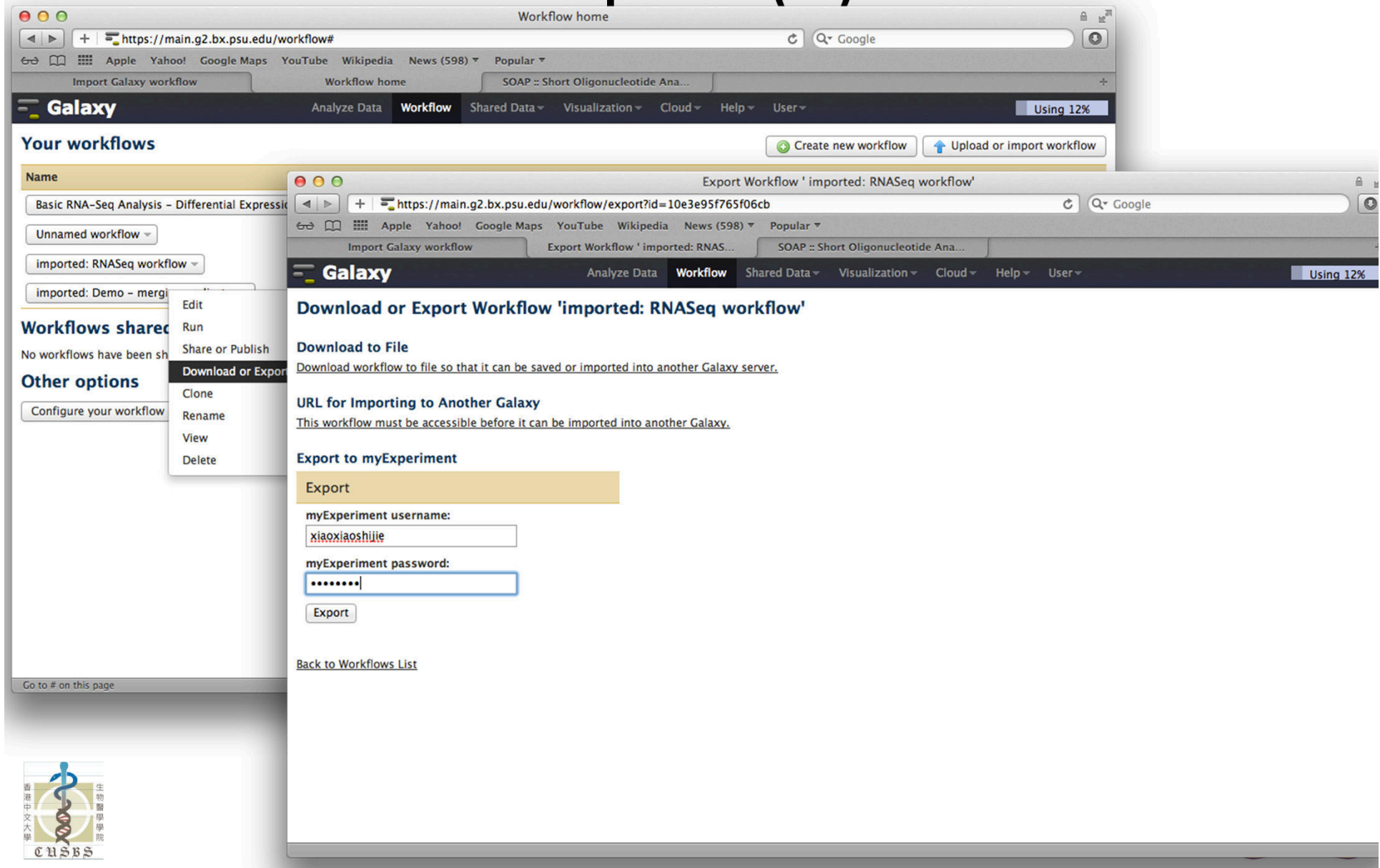
Workflows shared with you by others

No workflows have been shared with you.

Other options

[Configure your workflow menu](#)

Export (1)



The image shows a two-step process for exporting a workflow from the Galaxy platform. The background window displays the 'Your workflows' page with a list of workflows. A context menu is open over the 'imported: RNASeq workflow' entry, showing options like 'Edit', 'Run', 'Share or Publish', and 'Download or Export'. The foreground window shows the 'Export Workflow' page, which provides instructions and a form for exporting the workflow to a file or to a myExperiment account.

Workflow home
Workflow home

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

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'
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

香港中文大學
生物醫學學院
CUHKS

Export (2)

The image displays two overlapping browser windows illustrating the export of a Galaxy workflow to myExperiment.

Top Window (Galaxy):

- URL: https://main.g2.bx.psu.edu/workflow/export_to_myexp?id=10e3e95f765f06cb
- Message: Workflow 'imported: RNASeq workflow' successfully exported to myExperiment. [Click here to view the workflow on myExperiment](#). Return to [workflow list](#).

Bottom Window (myExperiment):

- 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
- Workflow Entry: RNASeq workflow
- Created at: 18/07/12 @ 02:32:15
- Version 1 (of 1): Version created on: 18/07/12 @ 02:32:15 by: Xiaoxiaoshijie
- Workflow Type: Galaxy
- Original Uploader: Xiaoxia...
- License: All versions of this Workflow are not licensed.
- Preview: (Click on the image to get the full size)

Right Sidebar (myExperiment):

- New/Upload: Workflow [dropdown] GO
- User Profile: Xiaoxia... (My Profile, My Messages, My Memberships, My History, My News)
- My Stuff: 0 Friends | 0 Groups | 3 Workflows
- Workflows: GetCities, GetCities2, RNASeq workflow

Bottom Left Logo:

香港中文大學 生物醫學學院 CUSBS

GDSAP structure

*Bioinformatics
Development*



Biomedical and bioinformatics research

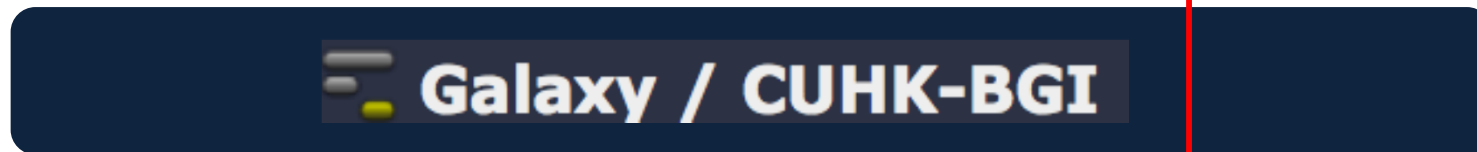
UCSC Genome Bioinformatics



Publishing

(GIGA)ⁿ
SCIENCE

(GIGA)ⁿ DB



(GIGA)ⁿ SCIENCE

Now taking submissions...

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



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 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.gigasiencejournal.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.

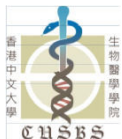
DNA sequences determined for research purposes have been routinely archived since 1982, when the EMBL Data Library was founded. This was closely followed by the formation of GenBank first at the US Department of Energy and then transferred to NIH, and in 1987 by the DNA Databank of Japan. These three 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



The screenshot shows the GigaDB website homepage. At the top left is the logo **(GIGA)ⁿ DB** with a green "Beta" badge and the tagline "Revolutionizing data dissemination, organization, and use". To the right are navigation links: Home | About | Contact | Terms of use. The main section has a blue background with the text "GigaDB" in large black font. Below it is a search bar with the placeholder "SEARCH by Species, DOI, Data Type" and a "GO" button. A paragraph states: "GigaDB contains discoverable, trackable, and citable data that have been assigned DOIs and are available for public download and use." The footer contains the logos for **(GIGA)ⁿ SCIENCE** and **华大基因 BGI**, followed by the same navigation links and social media icons for Facebook, Twitter, Sina, and GigaBlog.



www.gigaDB.org





37 Datasets with DOI[®]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	10.5524/1000010 - Genomic sequence from an Aboriginal Australian	Biosample: 259765 - Aboriginal Australian human	SNPs	vcf	AusAboriginal.hg19.var.filtered.snps.sampled.vcf.gz	<input checked="" type="checkbox"/>
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	<input checked="" type="checkbox"/>

Link to GigaDB landing pages

Link to sample if applicable

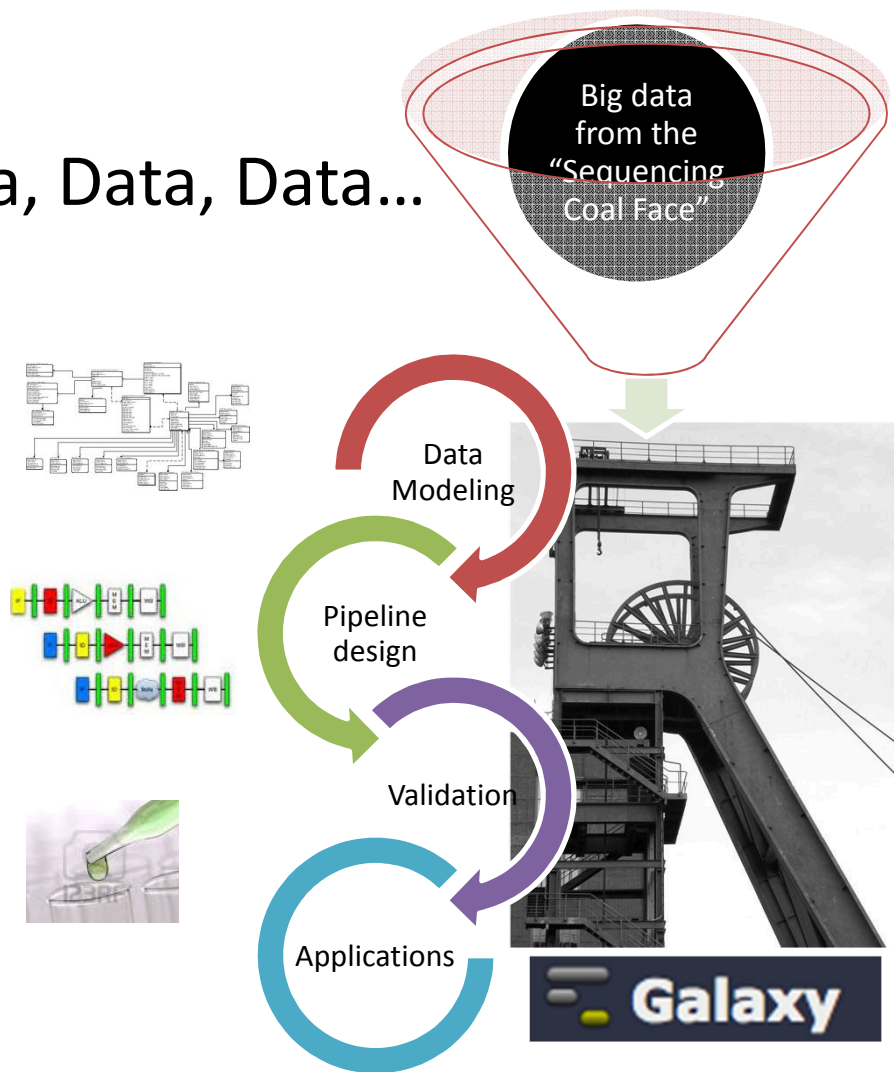
Download to Galaxy

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

