GigaDB and Galaxy: revolutionizing data dissemination, organization and analysis

Peter Li

GigaScience

peter@gigasciencejournal.com

华大基因 BGI

BioMed Central The Open Access Publisher

(GIGA)n SCIENCE

(GIGA)^n SCIENCE

Journal and database for large-scale data

华大基因 BGI

in conjunction with

BioMed Central
The Open Access Publisher

OPEN ACCESS

www.gigasciencejournal.com

GigaScience aims to revolutionize data dissemination, organization, understanding, and use. An online open-access open-data journal, we publish 'big-data' studies from the entire spectrum of life and biomedical sciences. To achieve our goals, the journal has a novel publication format: one that links standard manuscript publication with an extensive database that hosts all associated data and provides data analysis tools and cloud-computing resources.

Our scope covers not just 'omic' type data and the fields of high-throughput biology currently serviced by large public repositories, but also the growing range of more difficult-to-access data, such as imaging, neuroscience, ecology, cohort data, systems biology and other new types of large-scale sharable data.

Editorial Board | Editorial Team | Instructions for authors | FAQ

Articles

Editor's picks | Latest | Most viewed | Most forwarded

The little piggy genome
An important model organism for medical research, the genome of the Wuzhishan inbred miniature pig is presented along with analysis of its unusual genetic features
GigaScience 2012, **1**:16

The Puerto Rican parrot genome
Community funded with an innovative approach to assembly, the critically endangered Puerto Rican Parrot (Amazona vittata) genome is presented here for the first time
GigaScience 2012, **1**:14

Publication fees

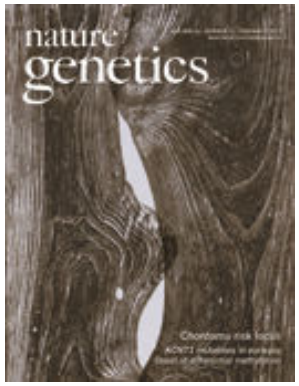
There are currently no article processing charges (APCs) for articles published in *GigaScience* due to generous support from the BGI.

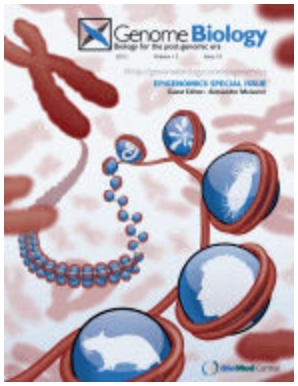
GigaScience is a new integrated database and journal co-published in collaboration between BGI Shenzhen and BioMed Central, to meet the needs of a new generation of biological and biomedical research as it enters the era of "big-data." BGI (formerly known as Beijing Genomics Institute) was founded in 1999 and has since become the largest genomic organization in the world and has a proven track record of innovative, high profile research.

To achieve its goals, *GigaScience* has developed a novel publishing format that integrates manuscript publication with a database that will provide DOI assignment to every dataset. Supporting the open-data movement, we require that all supporting data and source code be publically available in a suitable public repository and/or under a public domain CC0 licence in the BGI

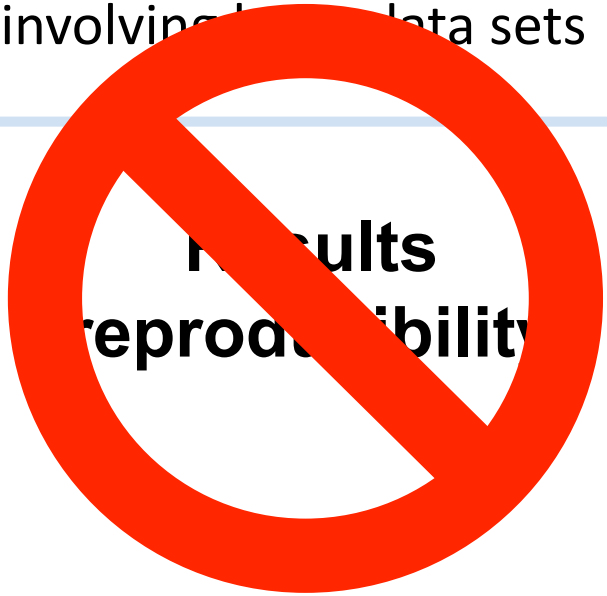
Submit a manuscript
Register
Sign up for article alerts
Contact us
Follow us on Twitter
Support

Why another *omics journal?





Already many journals publishing research involving [illegible] data sets

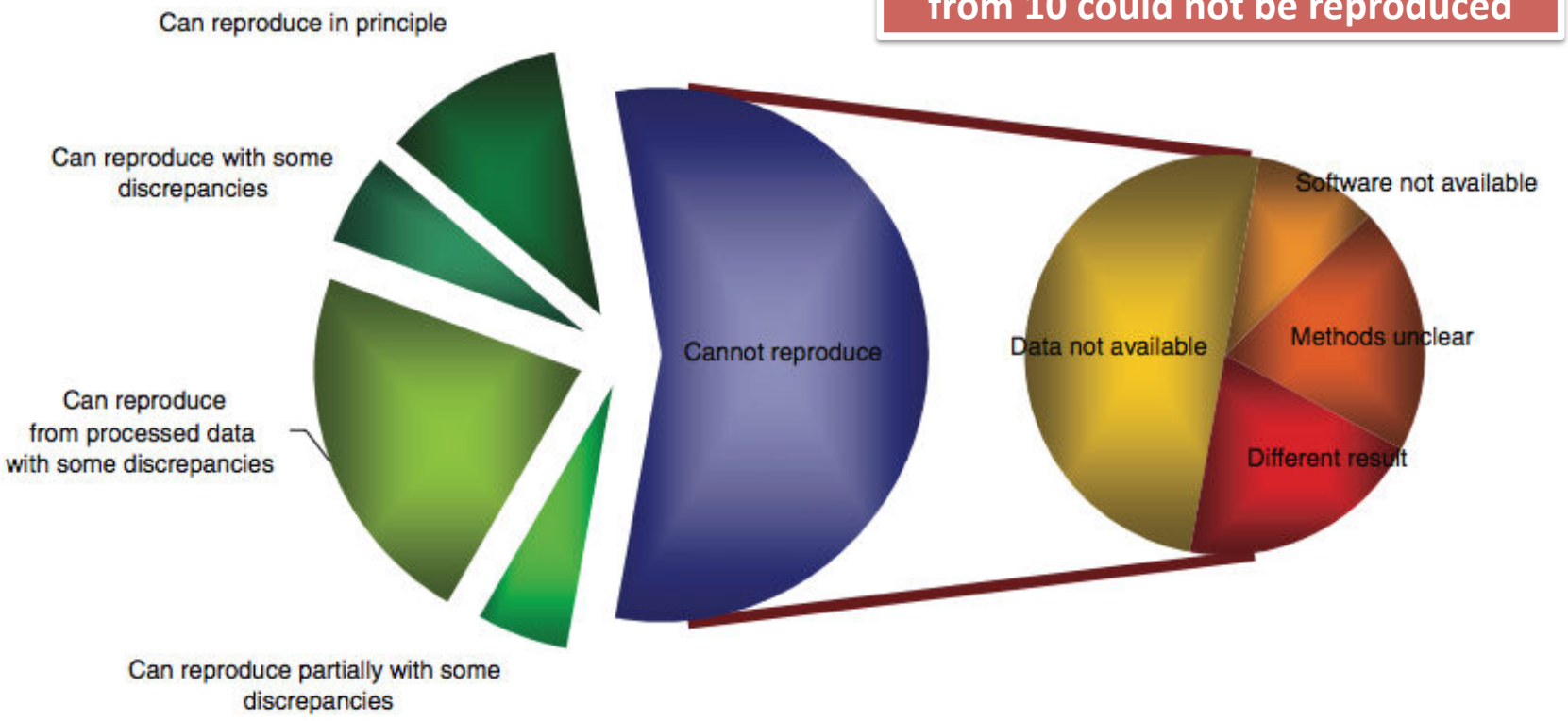


[illegible]ults

[illegible]ibility

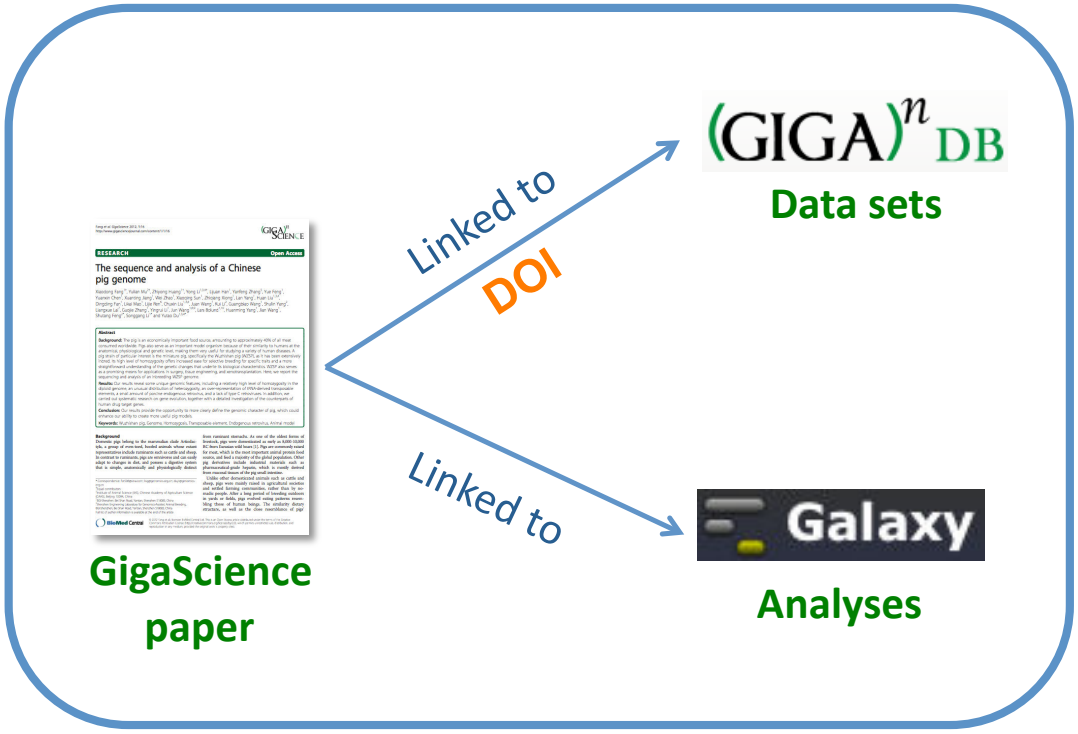
Unrepeatability of scientific results

Out of 18 microarray papers, results from 10 could not be reproduced

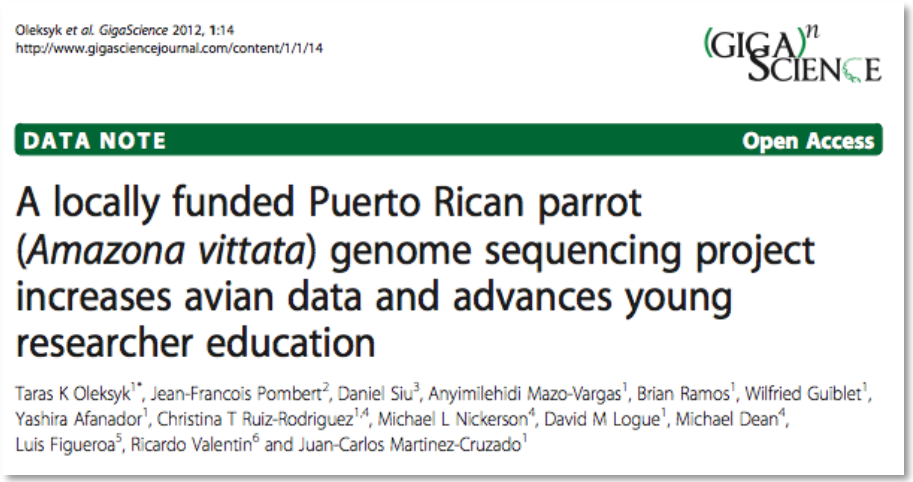


Ioannidis et al., 2009. Repeatability of published microarray gene expression analyses. Nature Genetics 41: 149-155.

How are we supporting data reproducibility?



Community tools for
data reproduction and reuse



Oleksyk *et al. GigaScience* 2012, **1**:14
http://www.gigasciencejournal.com/content/1/1/14

(GIGA)^n SCIENCE

DATA NOTE **Open Access**

A locally funded Puerto Rican parrot (*Amazona vittata*) genome sequencing project increases avian data and advances young researcher education

Taras K Oleksyk[1*], Jean-Francois Pombert[2], Daniel Siu[3], Anyimilehidi Mazo-Vargas[1], Brian Ramos[1], Wilfried Guiblet[1], Yashira Afanador[1], Christina T Ruiz-Rodriguez[1,4], Michael L Nickerson[4], David M Logue[1], Michael Dean[4], Luis Figueroa[5], Ricardo Valentin[6] and Juan-Carlos Martinez-Cruzado[1]

Additional file 15: Table S9. Bioinformatics tools and outputs for scaffold and gene annotation.

Additional file 16: Table S10. An example of annotation output produced by a student in the Genome annotation class using *A. vittata* genome.

Competing interests
Oleksyk TK, Pombert JF, Mazo A, Ramos B, Guiblet W, Afanador Y, Ruiz-Rodriguez CT, Nickerson ML, Logue D, Dean M, Figueroa L, Valentin R, and Martinez-Cruzado JC do not have competing interests. Siu D is employed by Axeq Technologies; the company which carried out the DNA Sequencing.

Authors' contributions
TKO, LF, RV, MD, MLN, DL and JCMC came up with the idea, and designed the experiments. TKO, WG, YA, CTRR and JCMC organized public support and raised the funds. TKO, AMV, BR, YA, CTRR and RV collected, extracted and quantified DNA. DS performed sequencing and assembly by SOAPdenovo. JFP performed assembly by Ray. TKO and WG designed the data browser webpage. TKO, JFP, MLN, DL, MD and JCMC wrote the paper. All authors read and approved the final manuscript.

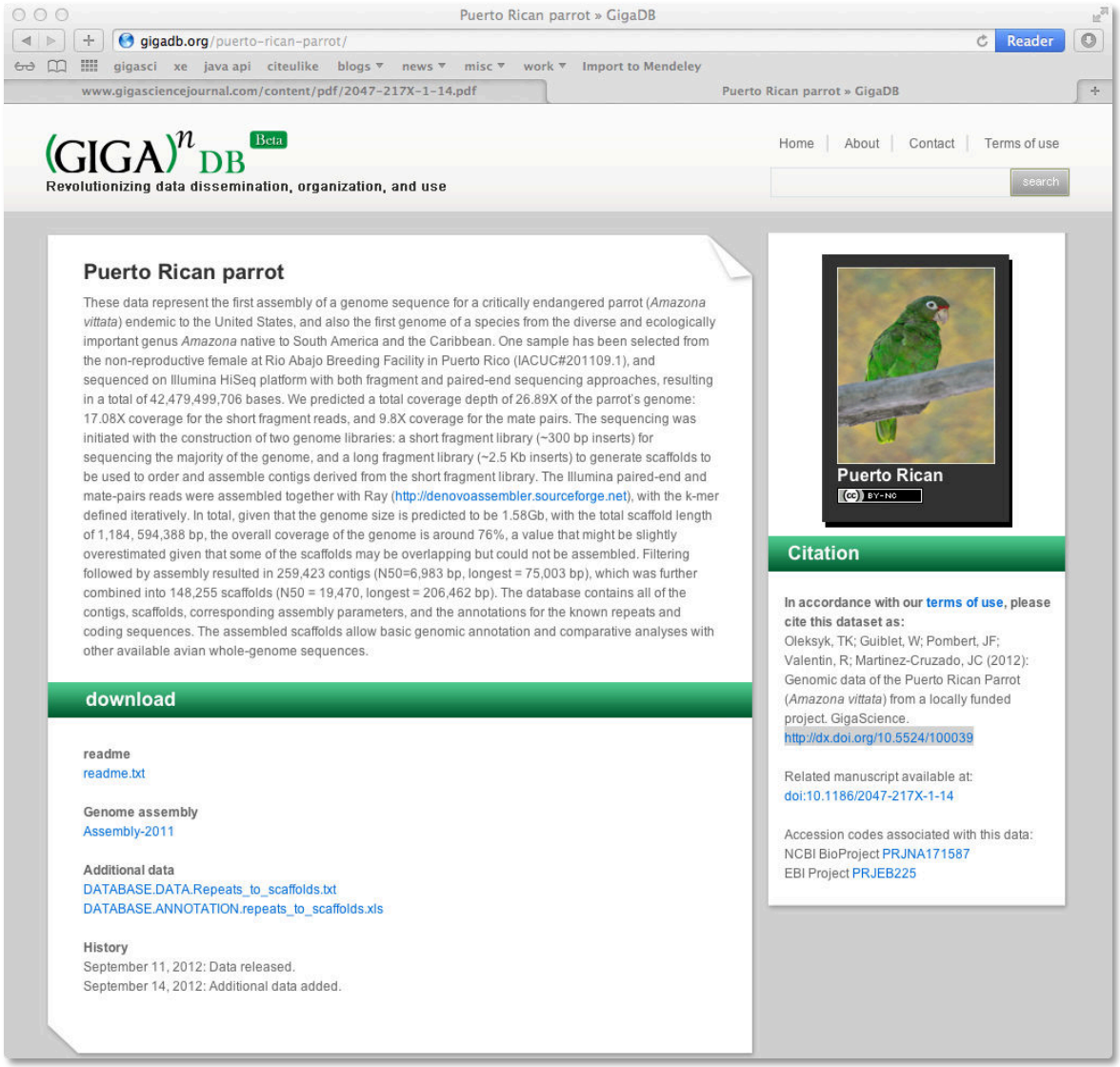
6. Warren WC, Clayton DF, Ellegren H, Arnold AP, Hillier LW, Kunstner A, Searle S, White S, Vilella AJ, Fairley S, *et al*: **The genome of a songbird.** *Nature* 2010, **464**(7289):757–762.
7. Krzywinski MI, Schein JE, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA: **Circos: An information aesthetic for comparative genomics.** *Genome Res* 2009, **19**(9):1639–45.
8. Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, Wang Z, Rasko DA, McCombie WR, Jarvis ED, *et al*: **Hybrid error correction and de novo assembly of single-molecule sequencing reads.** *Nat Biotechnol* 2012, **30**(7):693–700.
9. Oleksyk TK, Guiblet W, Pombert JF, Valentin R, Martinez-Cruzado JC: **Genomic data of the Puerto Rican Parrot (*Amazona vittata*) from a locally funded project.** *GigaScience* 2012. http://dx.doi.org/10.5524/100039.
10. O'Brien SJ: **Genome empowerment for the Puerto Rican parrot – *Amazona vittata*.** *GigaScience* 2012, **1**:13.

doi:10.1186/2047-217X-1-14
Cite this article as: Oleksyk *et al.*: A locally funded Puerto Rican parrot (*Amazona vittata*) genome sequencing project increases avian data and advances young researcher education. *GigaScience* 2012 1:14.

Linking of papers and data by citation of DOIs

Data set DOI

Paper DOI



http://gigadb.org

Search | Advanced Search

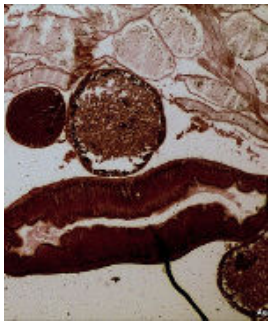
GigaDB is a new database integrated with the GigaScience journal to meet the needs of a new generation of biological and biomedical research as it enters the era of "big-data"… (see more)

Genomic

10.5524/100008

Genomic data from the polar bear (*Ursus maritimus*).

2011-07-06



10.5524/100017

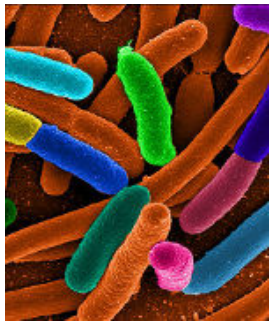
Genomic data from the roundworm *Ascaris suum*.

2011-11-12

10.5524/100003

Genomic data from the crab-eating macaque/cynomolgus monkey (*Macaca fascicularis*).

2011-07-06



10.5524/100001

Genomic data from *Escherichia coli* O104:H4 isolate TY-2482

2011-06-03

10.5524/100005

Genomic data from the Emperor penguin (*Aptenodytes forsteri*).

2011-07-06

(GIGA)^n DB
Revolutionizing data dissemination, organization, and use

Home | About | Contact | Terms of use

Help | Login | Create account

Data released on 2011-11-12

Genomic sequence from an Aboriginal Australian.

Here we present Aboriginal Australian (*Homo sapiens*) genomic sequence obtained from a 100-year-old lock of hair donated by an Aboriginal man from southern Western Australia in the early 20th century. The nuclear genome was sequenced to an overall depth of 6.4-fold. It had a high degree of fragmentation, but no genotype level evidence of sample contamination by DNA from sample handlers of European descent was detected. Over 2 million SNPs were identified after genome mapping and genotyping, approximately 6 percent of which were not previously reported. The mitochondrial genome (mtDNA) was sequenced to an average depth of 338x and found to belong to a new subclade of haplogroup O.

In accordance with our terms of use, please cite this dataset as:
Related manuscripts:
10.1126/science.1211177 (PubMed: 21940856)

Accessions (data not in GigaDB):
SRA: SRP006388

Samples:

Sample ID	Taxonomic ID	Common name	Genbank name	Scientific name	Sample attributes

Files (FTP site):

DOI	File Name	Sample ID	File Type	File Format	Size	Release Date	Download
10.5524/100010	100726_I198_FC204RWABXX_L3_HUMgspRBBSEW.fq.gz	FILE_CODE	Genome sequence	FASTQ	5.05 GB	2011-11-12	Download
10.5524/100010	100726_I198_FC204RWABXX_L4_HUMgspRBCSEW.fq.gz	FILE_CODE	Genome sequence	FASTQ	5.75 GB	2011-11-12	Download
10.5524/100010	100726_I198_FC204RWABXX_L5_HUMgspRBCSEW.fq.gz	FILE_CODE	Genome sequence	FASTQ	5.42 GB	2011-11-12	Download

Faster download speeds

Aspera data transfer

(GIGA)^n DB BGI Datasets Get DOI®s

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

Roundworm
Schistosoma
Silkworm
Parasitic nematode
Pacific oyster

Human
Asian individual (YH)
- DNA Methylome
- Genome Assembly
- Transcriptome

Cancer (14TB)
Single cell bladder cancer
HBV infected exomes
Ancient DNA
- Saqqaq Eskimo
- Aboriginal Australian

GENOME 10K

Vertebrates
Darwin's Finch
Giant panda Macaque
-Chinese rhesus
-Crab-eating
Mini-Pig
Naked mole rat
Parrot, Puerto Rican
Penguin
- Emperor penguin
- Adelie penguin

Pigeon, domestic
Polar bear
Sheep
Tibetan antelope

Released pre-publication
Paper published in GigaScience

Microbe
E. Coli O104:H4 TY-2482
T2D gut metagenome

Cell-Lines
Chinese Hamster Ovary
Mouse methylomes

PLANTS
Chinese cabbage
Cucumber
Foxtail millet
Pigeonpea
Potato
Sorghum

39 data sets

华大基因 BGI

(GIGA)^n DB

Currently: 39 public datasets

10 citations in references

华大基因 BGI

Humans

Ancient D

- Aborigin

- Saqqaq Eskimo

Asian individual (YH)

nature biotechnology

Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome

Zhiyu Peng, Yanbing Cheng, Bertrand Chin-Ming Tan, Lin Kang, Zhijian Tian, Yuankun Zhu, Wenwei Zhang, Yu Liang, Xueda Hu, Xuemei Tan, Jing Guo, Zirui Dong, Yan Liang, Li Bao & Jun Wang

Affiliations | Contributions | Corresponding author

Nature Biotechnology **30**, 253–260 (2012) | doi:10.1038/nbt.2122

Received 07 December 2011 | Accepted 17 January 2012 | Published online 12 February 2012

40. Tian, Z. *et al*. Transcriptome from a lymphoblastoid cell line taken from the YH Han Chinese individual. *Giga Sci*. http://dx.doi.org/10.5524/100013. (2011).

1000 Genomes
A Deep Catalog of Human Genetic Variation

GENOME 10K.

5k

What about the analyses?

GigaScience paper

Linked to

(GIGA)^n DB

Data sets

Linked to

Galaxy

Analyses

DataCite

doi

biosharing

my experiment

Galaxy

mibbi

comply to common standards

isatab

OBO

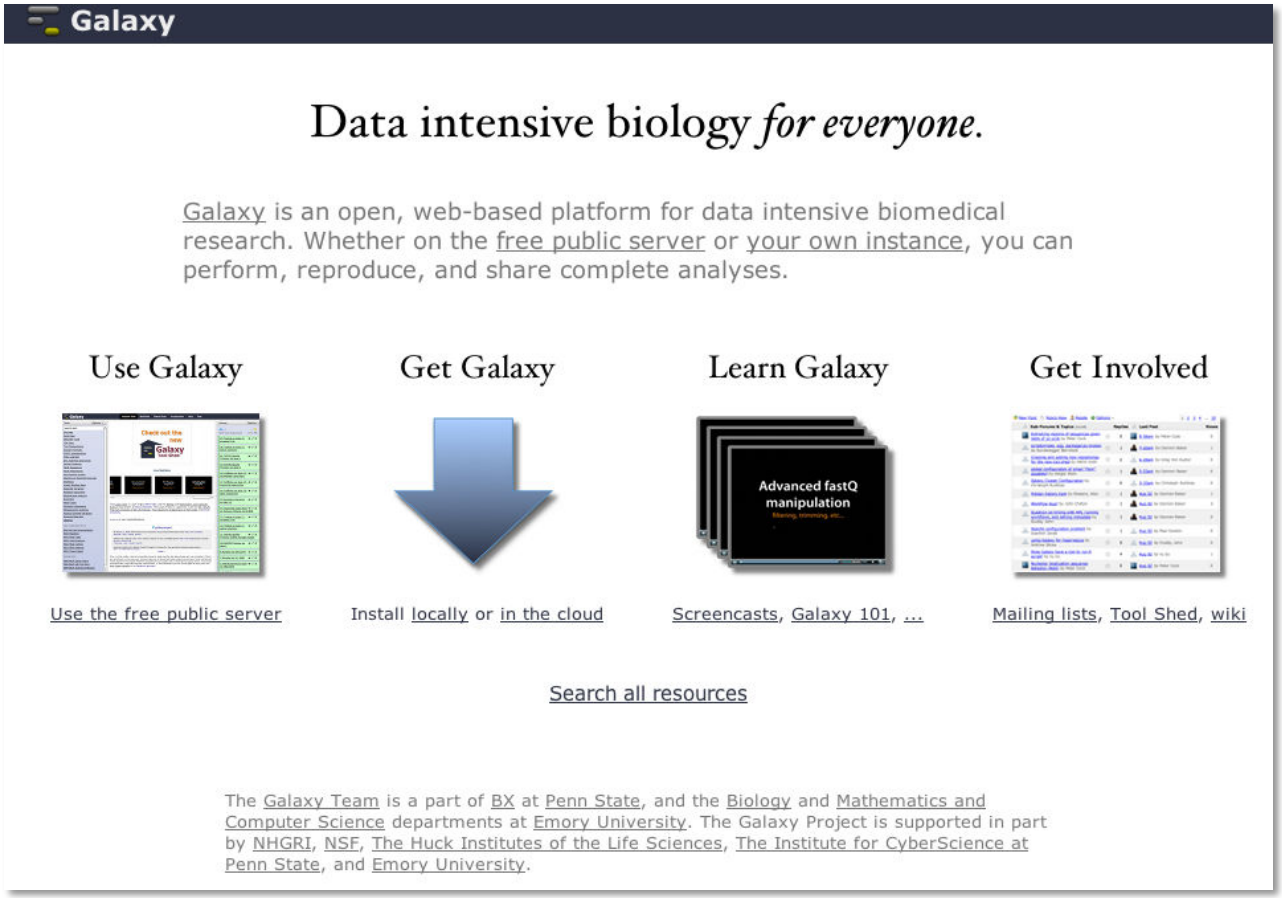
How will we make analyses available for downloading and execution?

Bioinformatics data analyses as workflows

Example workflow: Investigate the evolutionary relationships between proteins

Query → UniProt → Protein sequences → TCoffee Multiple Sequence Alignment → Multiple sequence alignment

Implement GigaScience workflows in a community-accepted format



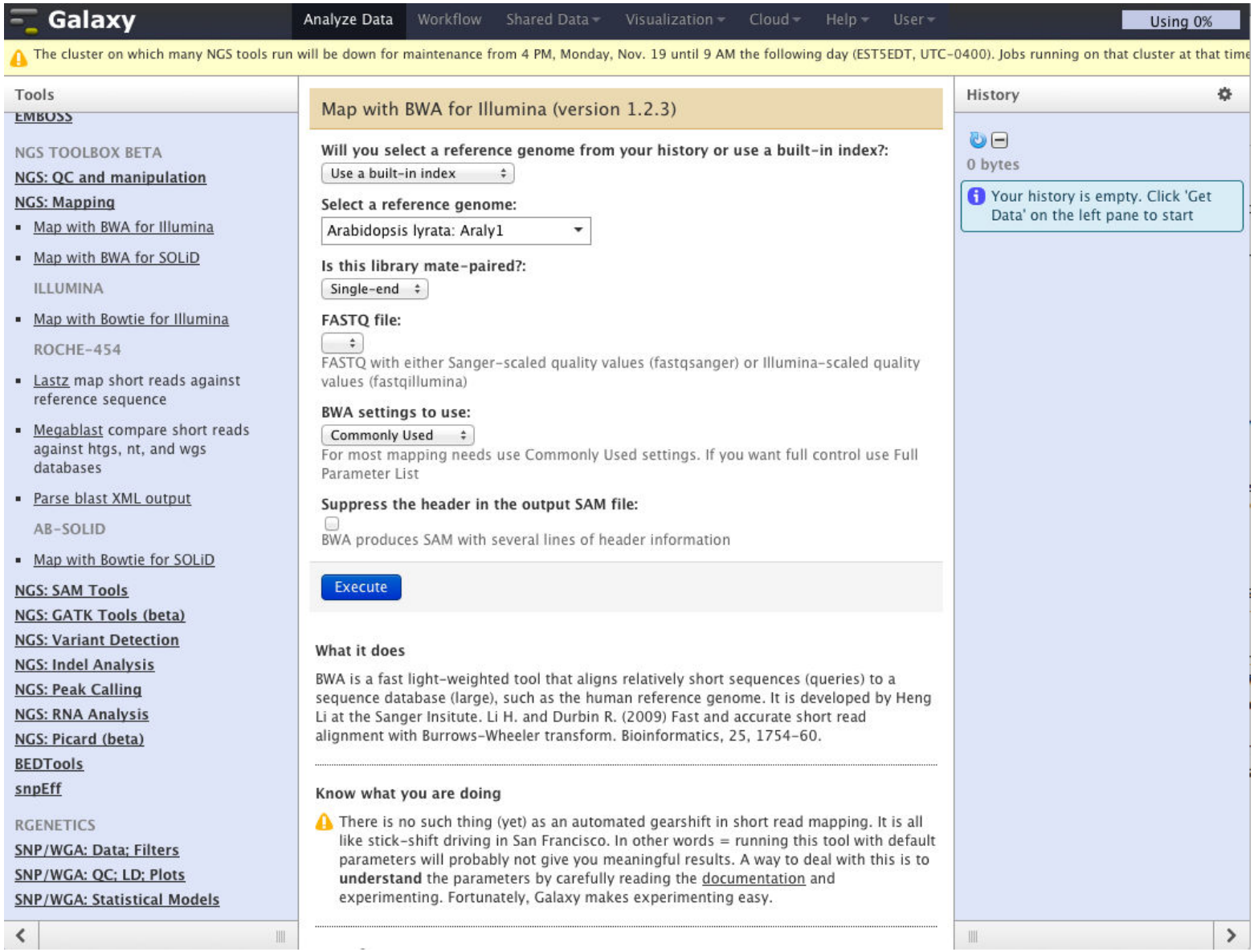
Open source

Over 20,000 main Galaxy server users

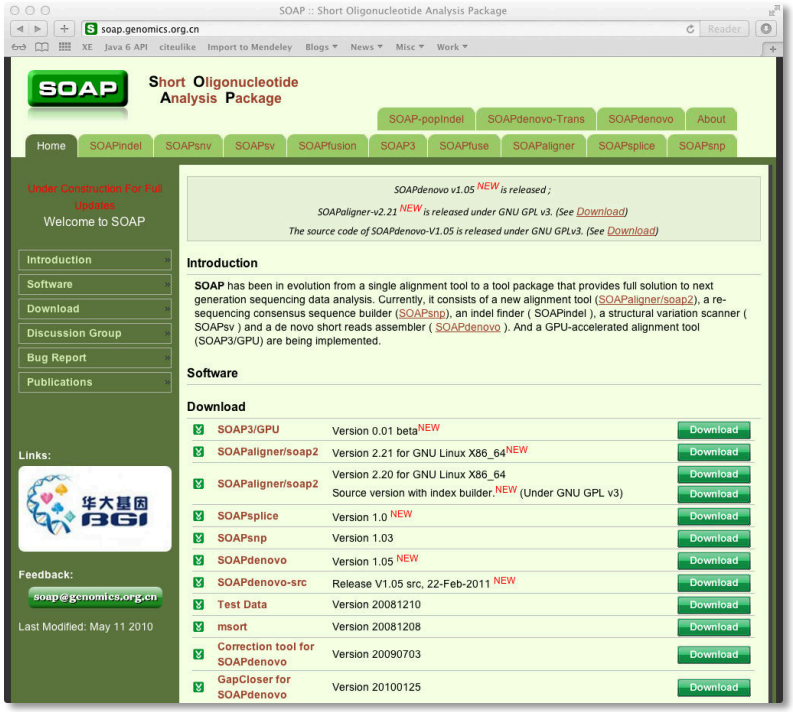
Over 500 papers citing Galaxy use

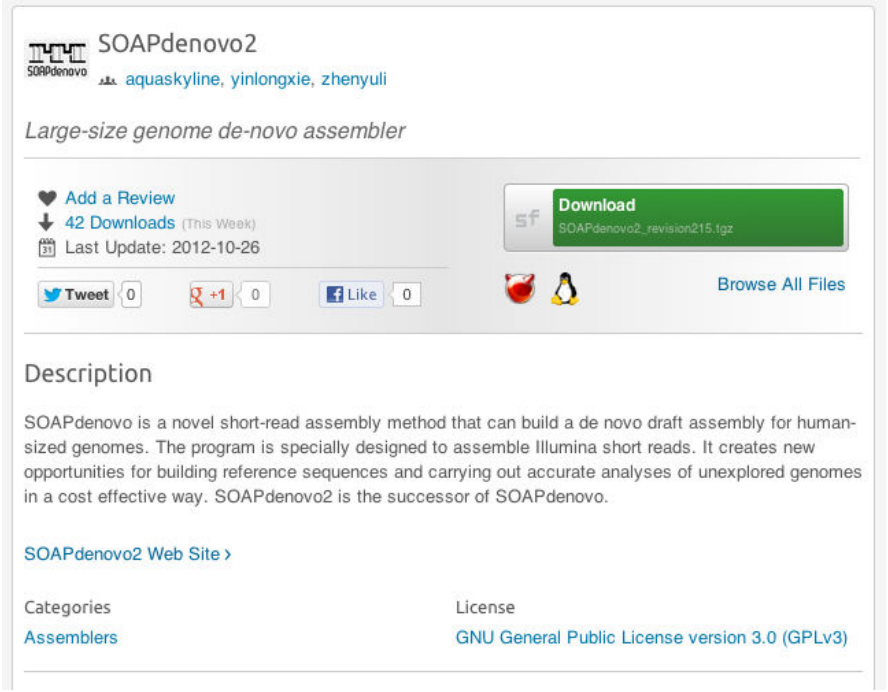
Over 55 Galaxy servers deployed

http://galaxyproject.org


Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Cloud
Help
User
Using 0%
The cluster on which many NGS tools run will be down for maintenance from 4 PM, Monday, Nov. 19 until 9 AM the following day (EST5EDT, UTC-0400). Jobs running on that cluster at that time
Tools
EMBOSS
NGS TOOLBOX BETA
NGS: QC and manipulation
NGS: Mapping
Map with BWA for Illumina
Map with BWA for SOLiD
ILLUMINA
Map with Bowtie for Illumina
ROCHE-454
Lastz map short reads against reference sequence
Megablast compare short reads against htgs, nt, and wgs databases
Parse blast XML output
AB-SOLID
Map with Bowtie for SOLiD
NGS: SAM Tools
NGS: GATK Tools (beta)
NGS: Variant Detection
NGS: Indel Analysis
NGS: Peak Calling
NGS: RNA Analysis
NGS: Picard (beta)
BEDTools
snpEff
RGENETICS
SNP/WGA: Data; Filters
SNP/WGA: QC; LD; Plots
SNP/WGA: Statistical Models
Map with BWA for Illumina (version 1.2.3)
Will you select a reference genome from your history or use a built-in index?:
Use a built-in index
Select a reference genome:
Arabidopsis lyrata: Araly1
Is this library mate-paired?:
Single-end
FASTQ file:
FASTQ with either Sanger-scaled quality values (fastqsanger) or Illumina-scaled quality values (fastqillumina)
BWA settings to use:
Commonly Used
For most mapping needs use Commonly Used settings. If you want full control use Full Parameter List
Suppress the header in the output SAM file:
BWA produces SAM with several lines of header information
Execute
What it does
BWA is a fast light-weighted tool that aligns relatively short sequences (queries) to a sequence database (large), such as the human reference genome. It is developed by Heng Li at the Sanger Insitute. Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25, 1754-60.
Know what you are doing
There is no such thing (yet) as an automated gearshift in short read mapping. It is all like stick-shift driving in San Francisco. In other words = running this tool with default parameters will probably not give you meaningful results. A way to deal with this is to understand the parameters by carefully reading the documentation and experimenting. Fortunately, Galaxy makes experimenting easy.
History
0 bytes
Your history is empty. Click 'Get Data' on the left pane to start
Tool list
Tool parameterisation
Results panel

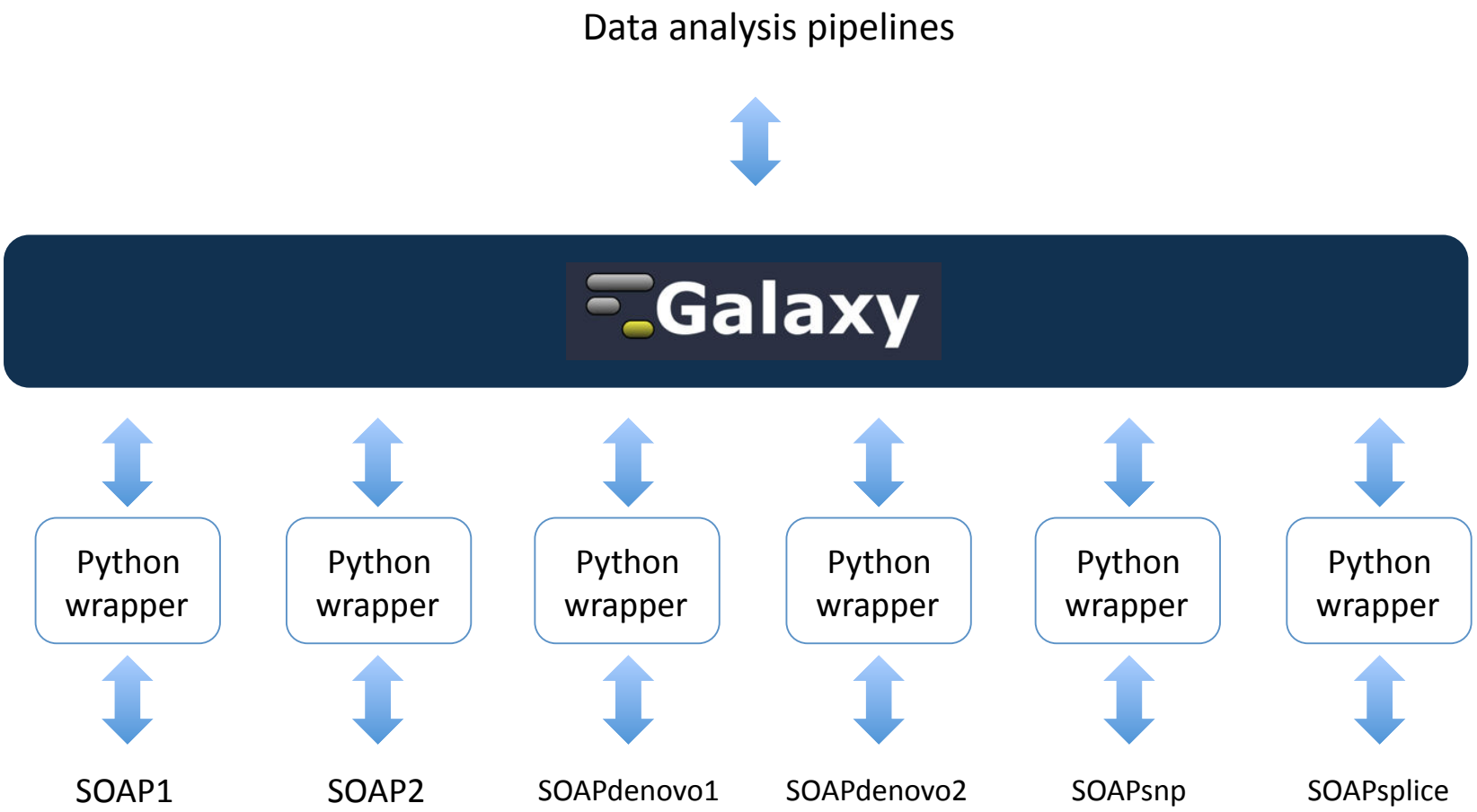
Pilot project - Integrate BGI SOAP package into Galaxy



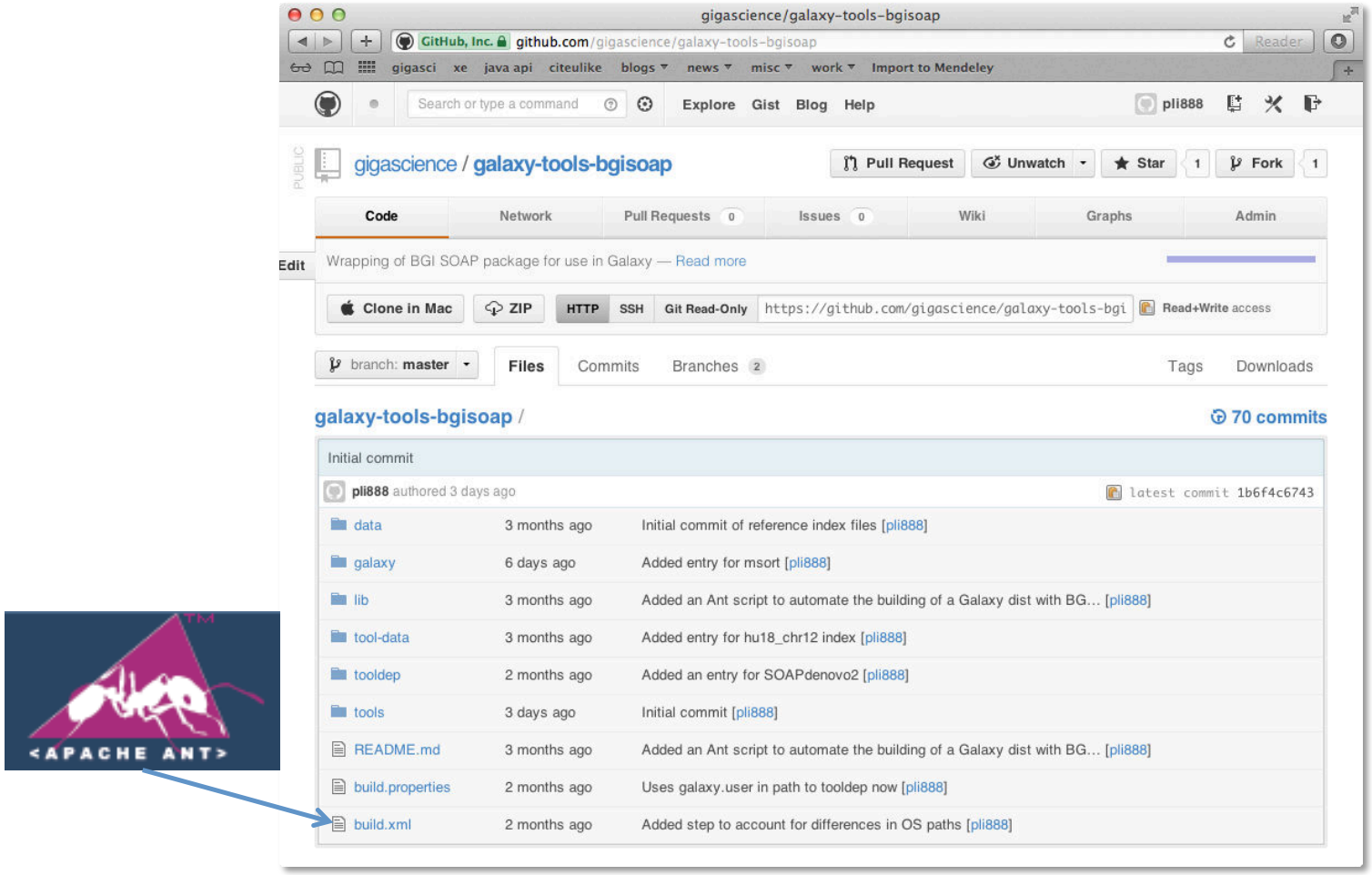


Enable SOAP tools to be used from within Galaxy workflows

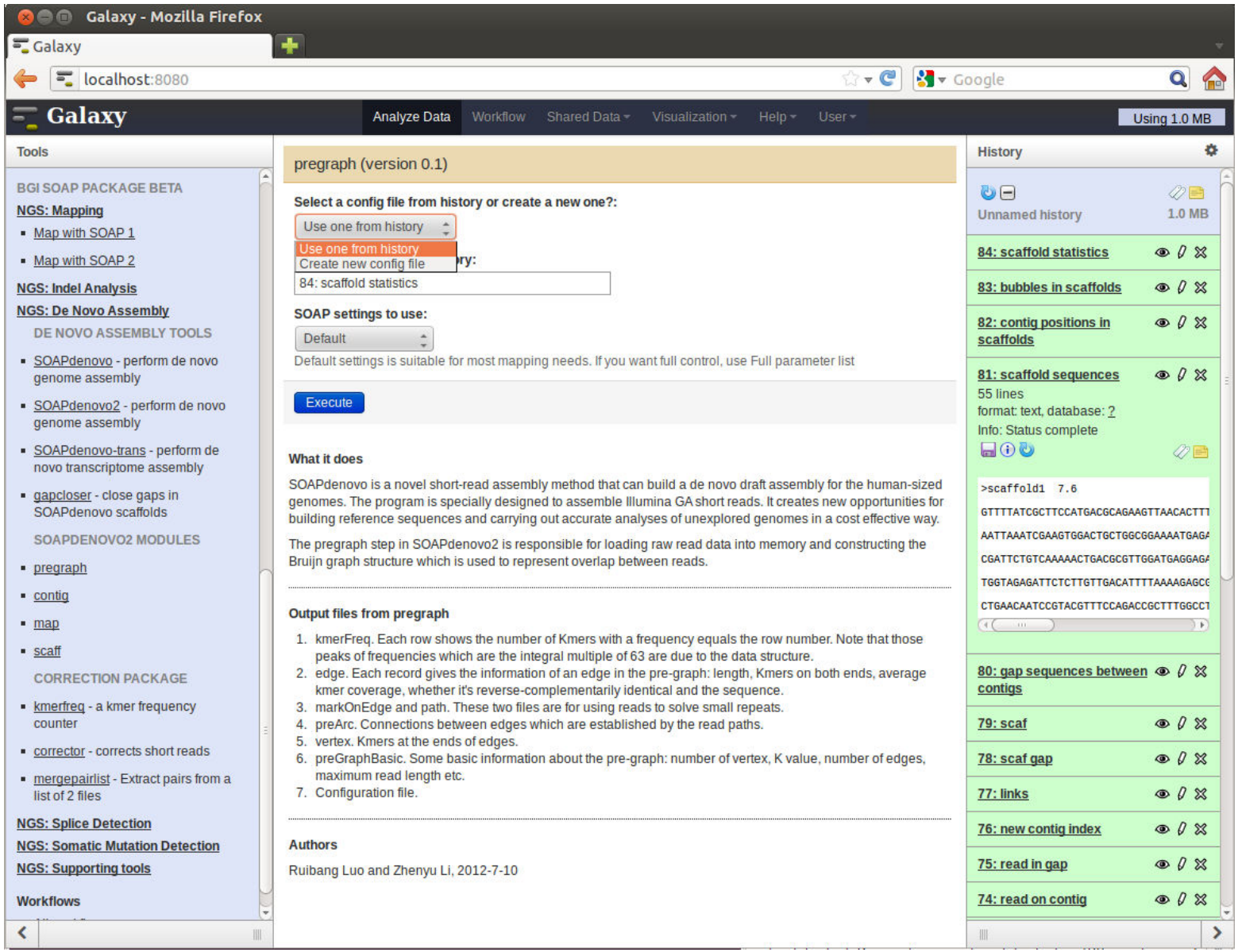
Integrate BGI SOAP package into Galaxy



GitHub open code repository



https://github.com/gigascience

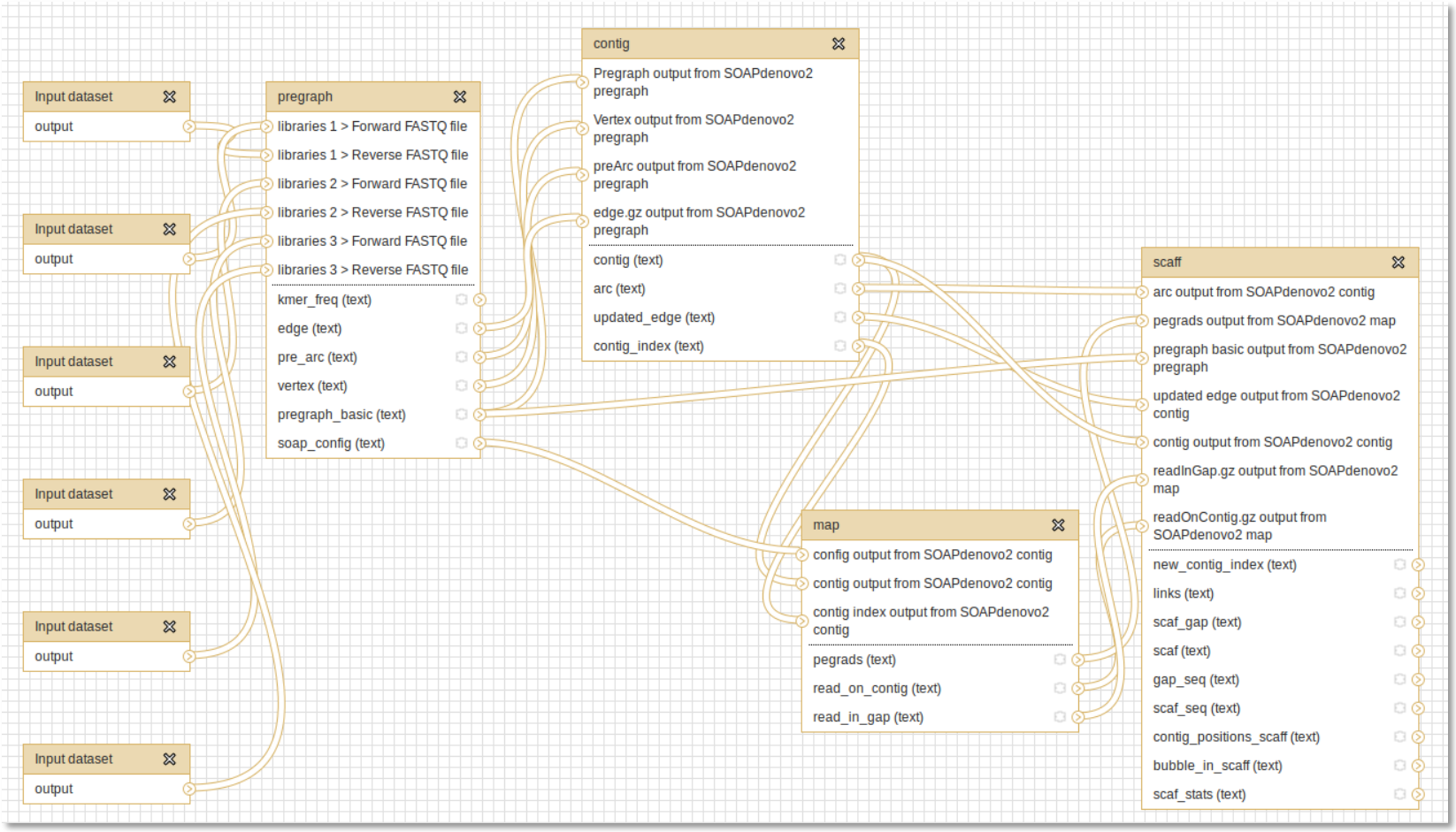


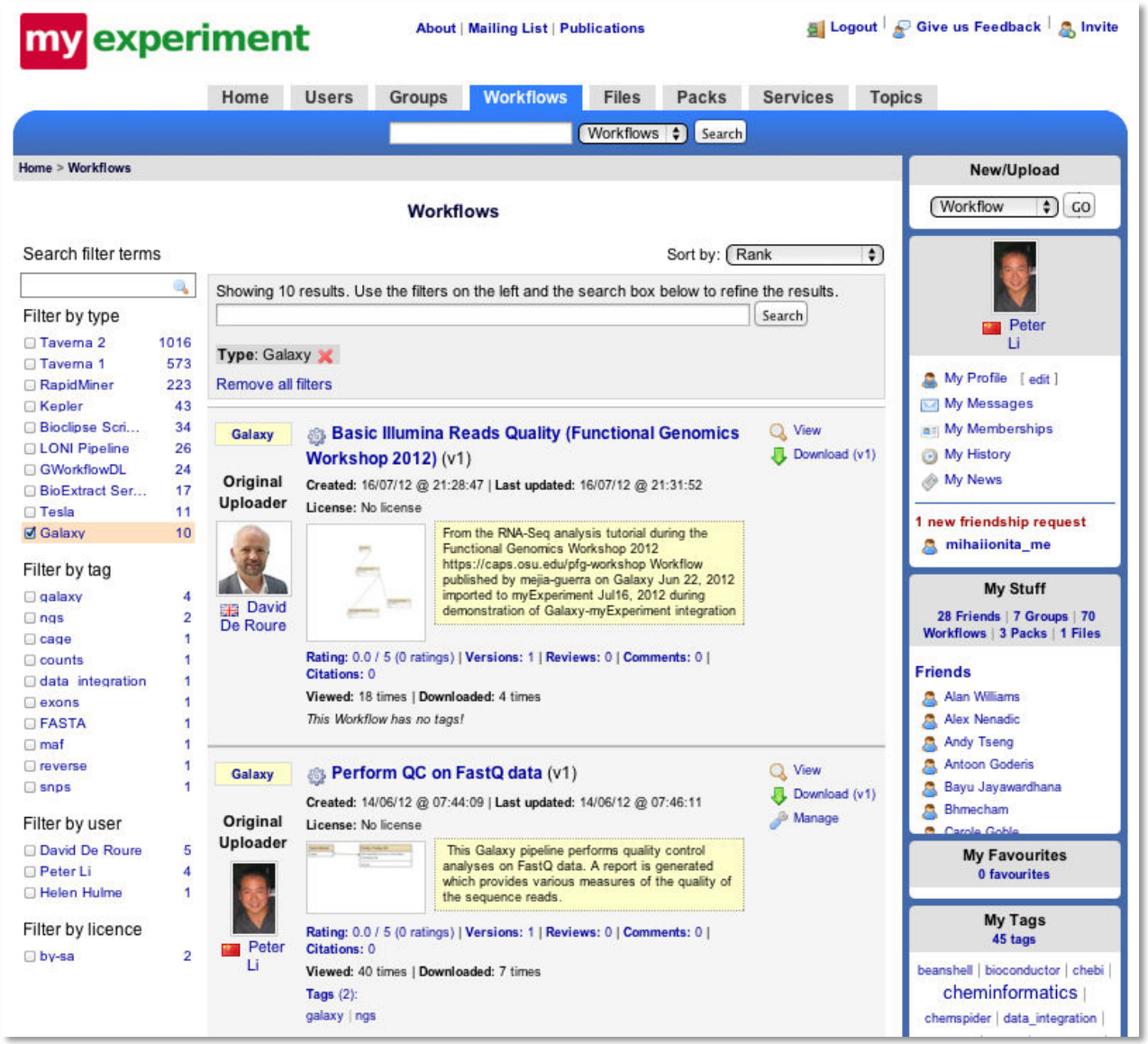
Tool list

Tool parameterisation

Results panel

SOAPdenovo2 Galaxy workflow





http://www.myexperiment.org

Why publish in GigaScience?

Benefit

- Data hosted in GigaDB
- Allocation of DOIs to data
- Metadata in isa-tab format
- Galaxy tool integration
- Use of tools in Galaxy workflows

Added value

- No need to use own servers
- Citable data
- Aids reuse of data
- Supports reuse of tools
- Improves documentation
- Shows how tool can be used with other bioinf. software

Thanks to:

- Tin-Lap Lee and Huayan Gao - CUHK
- Tam, Jesse, Scott, Nicole & Laurie - GigaScience

peter@gigasciencejournal.com

华大基因 BGI

BioMed Central
The Open Access Publisher

(GIGA)^n SCIENCE