Open (and Big) Data – the next challenge

Beyond dead trees: are publishers the problem or solution?

Scott Edmunds

OASPA Asia, 2nd June 2013

@gigascience

华大基因 BGI

BioMed Central
The Open Access Publisher

Harnessing Data-Driven Intelligence

Enables:

Using networking power of the internet to tackle problems

Can ask new questions & find hidden patterns & connections

Build on each others efforts quicker & more efficiently

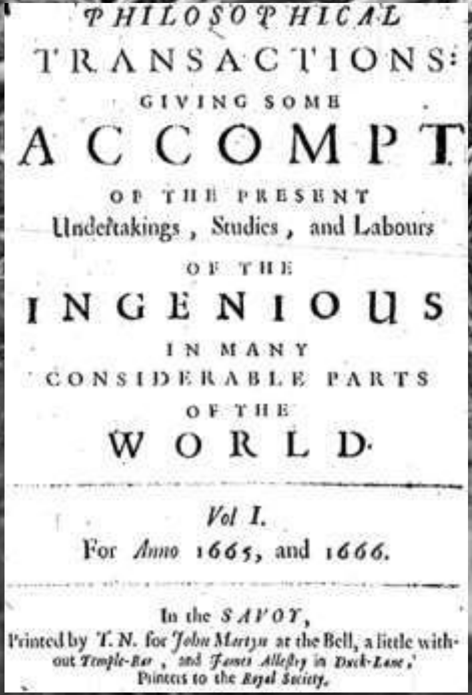
More collaborations across more disciplines

Harness wisdom of the crowds: crowdsourcing, citizen science, crowdfunding

Enabled by:

Removing silos, open licenses, transparency, immediacy

Dead trees not fit for purpose

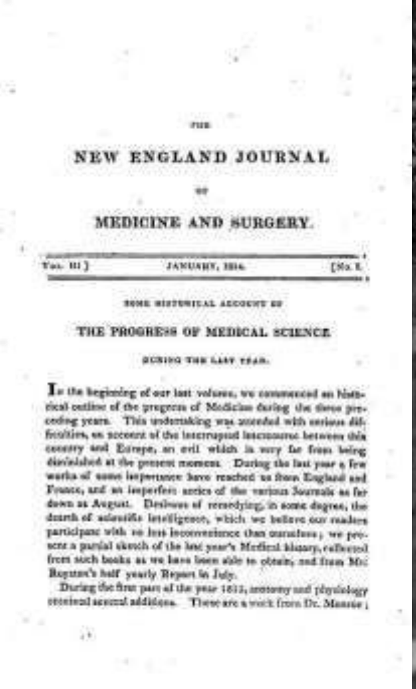


PHILOSOPHICAL
TRANSACTIONS:
GIVING SOME
ACCOMPT
OF THE PRESENT
Undertakings, Studies, and Labours
OF THE
INGENIOUS
IN MANY
CONSIDERABLE PARTS
OF THE
WORLD.

Vol I.
For Anno 1665, and 1666.

In the SAVOY,
Printed by T. N. for John Martyn at the Bell, a little without Temple-Bar, and James Allestry in Duck-Lane,
Printers to the Royal Society.

1665



THE
NEW ENGLAND JOURNAL
OF
MEDICINE AND SURGERY.

Vol. III.] JANUARY, 1814. [No. I.

SOME HISTORICAL ACCOUNT OF
THE PROGRESS OF MEDICAL SCIENCE
DURING THE LAST YEAR.

In the beginning of our last volume, we commenced an historical outline of the progress of Medicine during the three preceding years. This undertaking was attended with serious difficulties, on account of the interrupted intercourse between this country and Europe, an evil which is very far from being diminished at the present moment. During the last year a few works of some importance have reached us from England and France, and an imperfect series of the various Journals as far down as August. Desirous of remedying, in some degree, the dearth of scientific intelligence, which we believe our readers participate with no less inconvenience than ourselves; we present a partial sketch of the last year's Medical history, collected from such books as we have been able to obtain, and from Mr. Royston's half yearly Report in July.

During the first part of the year 1812, anatomy and physiology received several additions. These are a work from Dr. Monro;

1812

A WEEKLY ILLUSTRATED JOURNAL OF SCIENCE

"To the solid ground
Of Nature trusts the mind which builds for aye."—Wordsworth

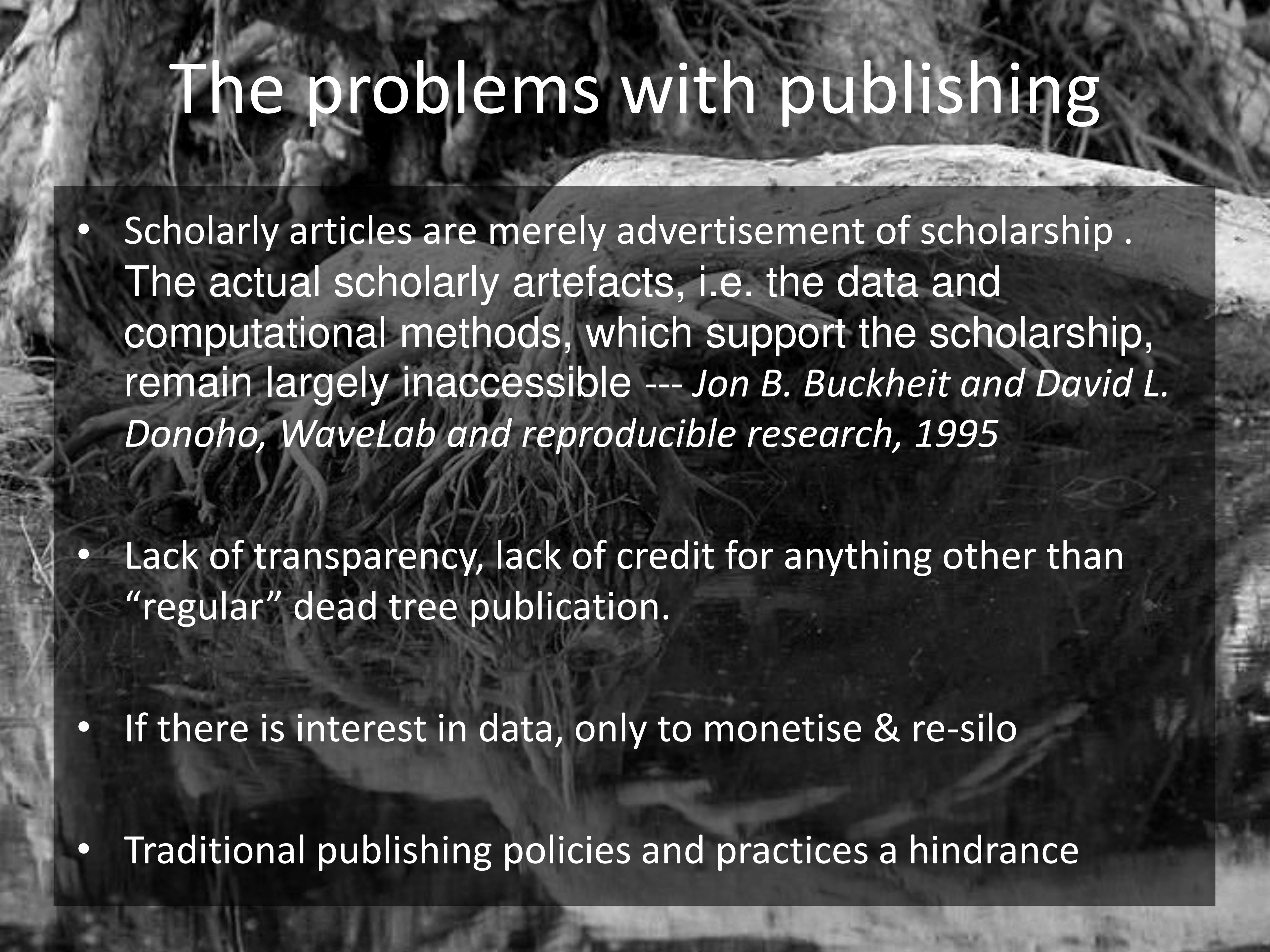
THURSDAY, NOVEMBER 4, 1869

NATURE: APHORISMS BY GOETHE

NATURE! We are surrounded and embraced by her: powerless to separate ourselves from her, and powerless to penetrate beyond her.

[illegible]

1869

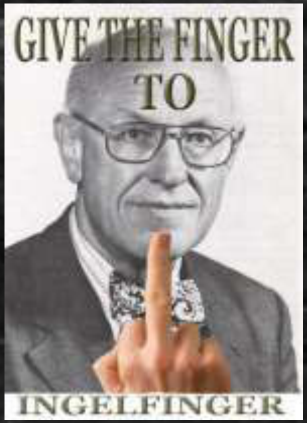


The problems with publishing

- Scholarly articles are merely advertisement of scholarship . The actual scholarly artefacts, i.e. the data and computational methods, which support the scholarship, remain largely inaccessible --- *Jon B. Buckheit and David L. Donoho, WaveLab and reproducible research, 1995*
- Lack of transparency, lack of credit for anything other than “regular” dead tree publication.
- If there is interest in data, only to monetise & re-silo
- Traditional publishing policies and practices a hindrance

Things holding us back:

- Disincentives to share or communicate:
  - Ingelfinger*! Embargoes, anti preprint & early data release policies
  - Page/method/citation limits

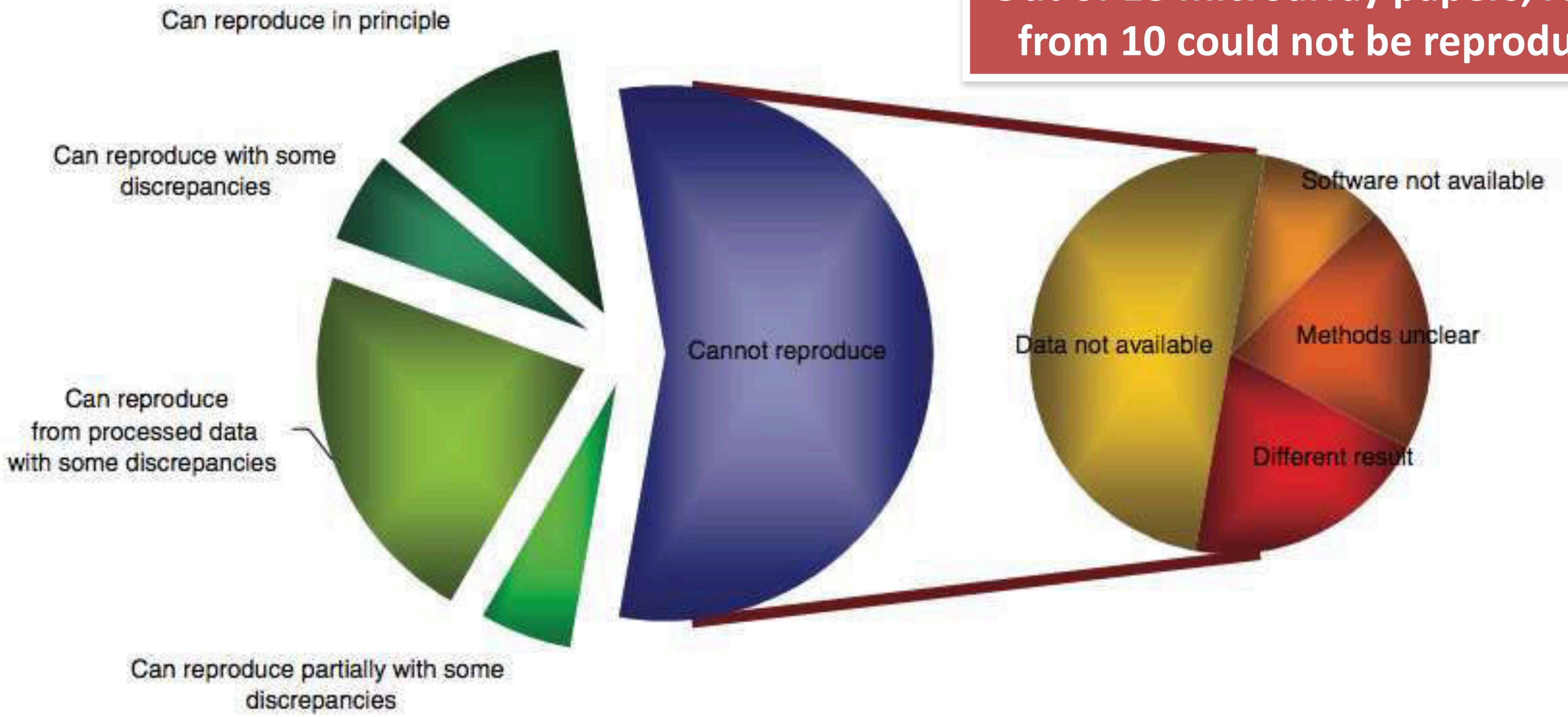


- Disincentives to remix
  - Open source approaches = plagiarism?
- Disincentives to release more quickly/more granularly
  - "Salami Slicing"

- First 2 years of citation data the only currency
  - "Faddism" v long term use or reproducibility. Publication bias.

* T-Shirts available from Graham Steel / http://www.zazzle.co.uk/steelgraham

The consequences: growing replication gap



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

Essay

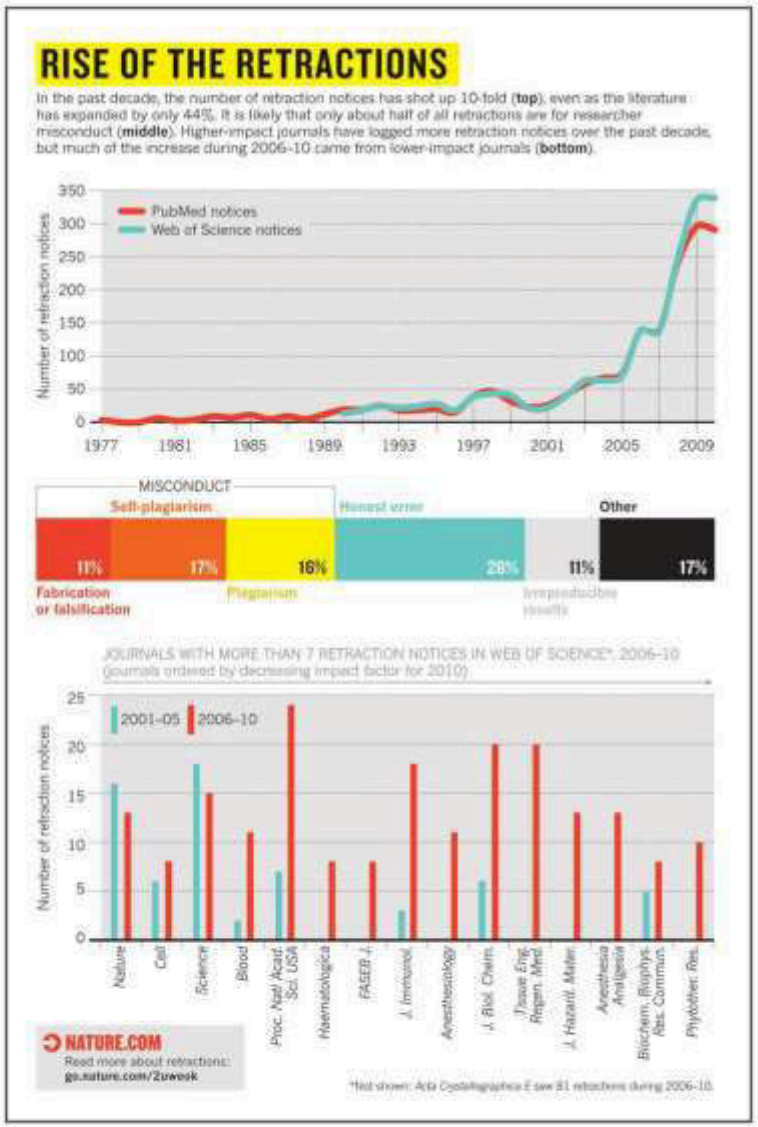
Why Most Published Research Findings Are False

John P. A. Ioannidis

PLOS MEDICINE

1. Ioannidis et al., (2009). Repeatability of published microarray gene expression analyses. *Nature Genetics* 41: 14
2. Ioannidis JPA (2005) Why Most Published Research Findings Are False. *PLoS Med* 2(8)

Consequences: increasing number of retractions



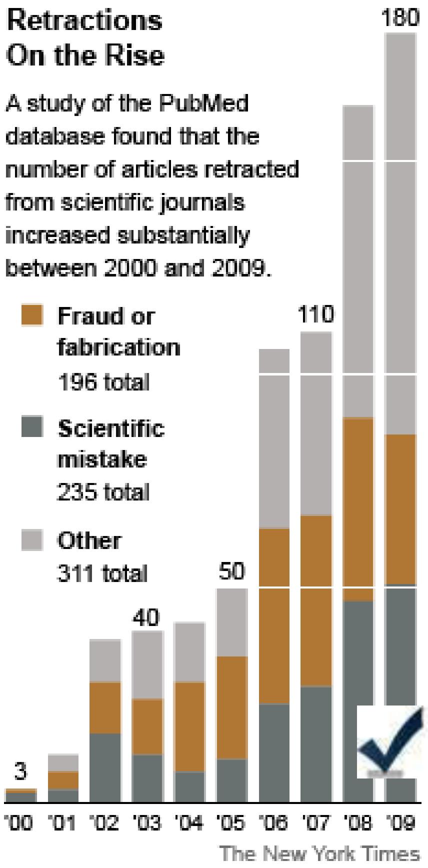
RISE OF THE RETRACTIONS

In the past decade, the number of retraction notices has shot up 10-fold (**top**), even as the literature has expanded by only 44%. It is likely that only about half of all retractions are for researcher misconduct (**middle**). Higher-impact journals have logged more retraction notices over the past decade, but much of the increase during 2006–10 came from lower-impact journals (**bottom**).

*Not shown: Acta Crystallographica E saw 81 retractions during 2006–10.

>15X increase in last decade

Strong correlation of “retraction index” with higher impact factor



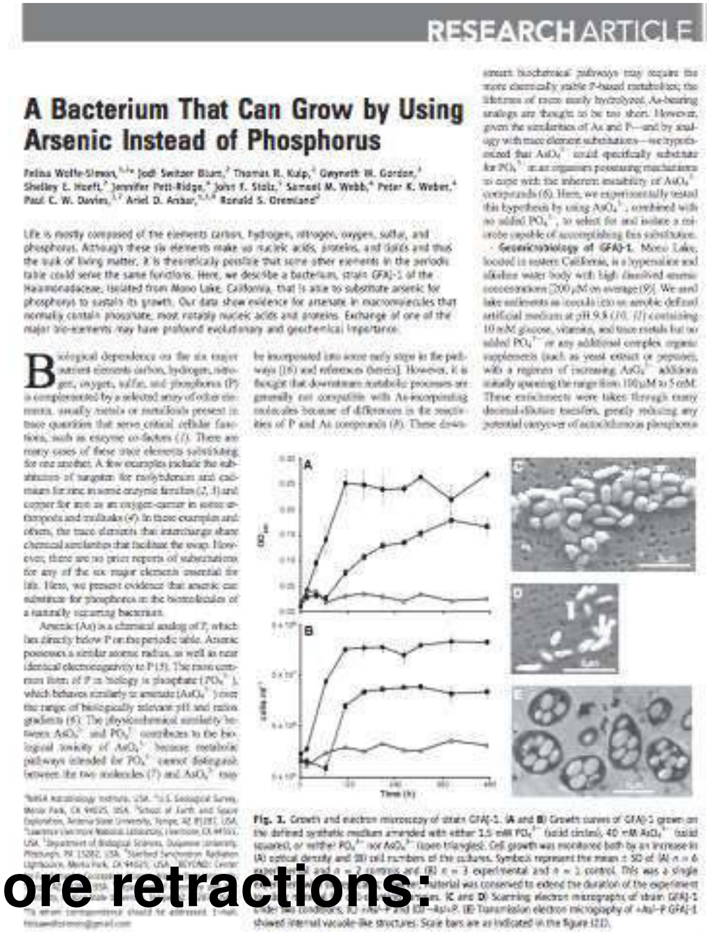
Retractions On the Rise

A study of the PubMed database found that the number of articles retracted from scientific journals increased substantially between 2000 and 2009.

The New York Times

1. **Science publishing: The trouble with retractions** http://www.nature.com/news/2011/111005/full/478026a.html
2. **Retracted Science and the Retraction Index** ˣ http://iai.asm.org/content/79/10/3855.abstract?

Consequences: growing replication gap



RESEARCH ARTICLE

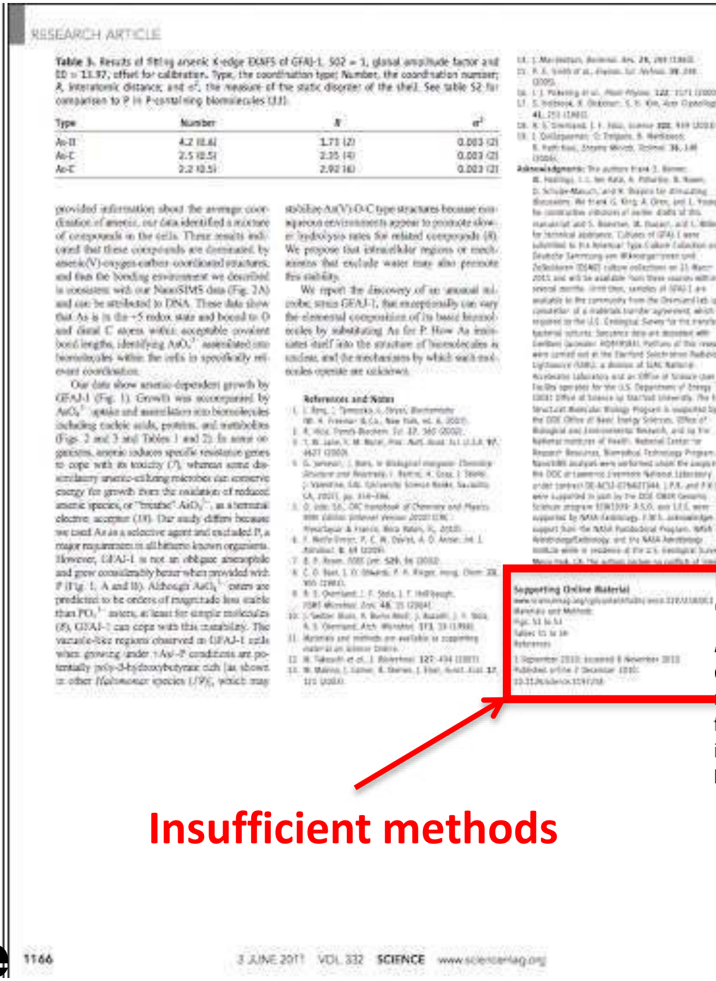
A Bacterium That Can Grow by Using Arsenic Instead of Phosphorus

Felisa Wolfe-Simon, Jodi Switzer Blum, Thomas R. Kulp, Gwyneth W. Gordon, Shelley E. Hoeft, Jennifer Pett-Ridge, John F. Stolz, Samuel M. Webb, Peter K. Weber, Paul C. W. Davies, Ariel D. Anbar, Ronald S. Oremland

Life is mostly composed of the elements carbon, hydrogen, nitrogen, oxygen, sulfur, and phosphorus. Although these six elements make up nucleic acids, proteins, and lipids and thus the bulk of living matter, it is theoretically possible that some other elements in the periodic table could serve the same functions. Here, we describe a bacterium, strain GFAJ-1 of the Halomonadaceae, isolated from Mono Lake, California, that is able to substitute arsenic for phosphorus to sustain its growth. Our data show evidence for arsenate in macromolecules that normally contain phosphate, most notably nucleic acids and proteins. Exchange of one of the major bio-elements may have profound evolutionary and geochemical importance.

Biological dependence on the six major nutrient elements carbon, hydrogen, nitrogen, oxygen, sulfur, and phosphorus (P) is complemented by a selected array of other elements, usually metals or metalloids present in trace quantities that serve critical cellular functions, [illegible]

Fig. 1. Growth and electron microscopy of strain GFAJ-1. [illegible]



RESEARCH ARTICLE

Table 1. Results of fitting arsenic K-edge EXAFS of GFAJ-1. S02 = 1, global amplitude factor and E0 = 11.97, offset for calibration. Type, the coordination type; Number, the coordination number; R, interatomic distance; and σ^2, the measure of the static disorder of the shell. See table S2 for comparison to P in P-containing biomolecules (11).

Type	Number	R	σ^2
As-O	4.2 (0.4)	1.73 (2)	0.003 (2)
As-C	2.5 (0.5)	2.35 (4)	0.003 (2)
As-C	2.2 (0.5)	2.92 (6)	0.003 (2)

provided information about the average coordination of arsenic, our data identified a mixture of compounds in the cells. [illegible]

References and Notes

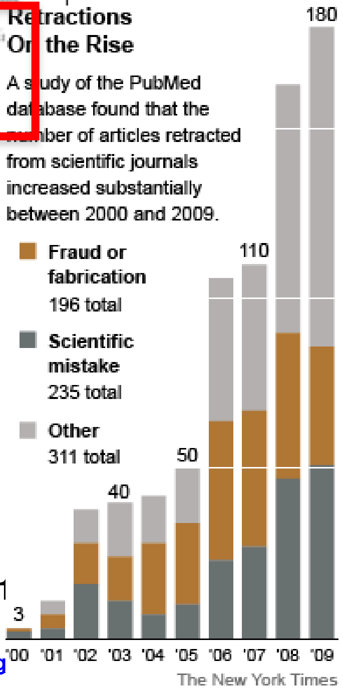
[illegible]

Supporting Online Material
www.sciencemag.org/cgi/content/full/science.1197258/DC1
Materials and Methods
Figs. S1 to S3
Tables S1 to S4
References

1 September 2010; accepted 8 November 2010
Published online 2 December 2010;
10.1126/science.1197258

1166 3 JUNE 2011 VOL 332 SCIENCE www.sciencemag.org

Insufficient methods



Retractions On the Rise

A study of the PubMed database found that the number of articles retracted from scientific journals increased substantially between 2000 and 2009.

▨	Fraud or fabrication 196 total
▨	Scientific mistake 235 total
▨	Other 311 total

The New York Times

More retractions:
>15X increase in last decade
At current % > by 2045 as many papers published as retracted

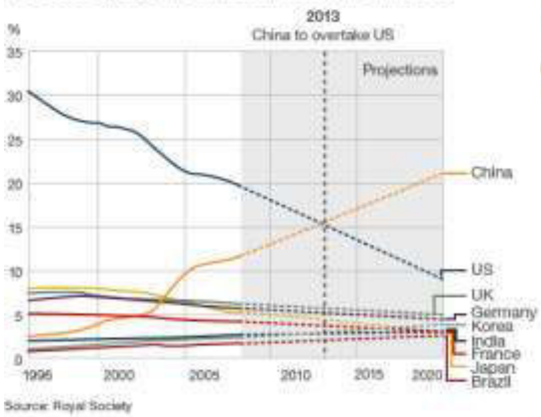
1. Ioannidis et al., 2009. Repeatability of published microarray gene expression analyses. *Nature Genetics* 41
2. Science publishing: The trouble with retractions http://www.nature.com/news/2011/111005/full/478026a.html
3. Bjorn Brembs: Open Access and the looming crisis in science https://theconversation.com/open-access-and-the-looming

Global perceptions of Chinese Research

Million RMB rewards for high IF publications = ?

NewScientist

"Faked research is endemic in China"



Projected growth in citations in scientific literature

Source: Royal Society

China's Publication Bazaar

A Science investigation has uncovered a smorgasbord of questionable practices including paying for author's slots on papers written by other scientists and buying papers from online brokers

Fraud fighter: 'Faked research is endemic in China'

Opinion > Interview p28

Shi-min Fang tells us how risking his life and libel writs to expose scientific misconduct in his native China has just won him the inaugural Maddox prize

nature

International weekly journal of science

475, 267 (2011)

Focus on quality, not just quantity

China publishes huge amounts of scientific research. Now it must mal more of it worth reading, says Changhui Peng.

ALTHOUGH CHINA RANKS SECOND IN TERMS OF **PUBLICATION** OUTPUT, IT RANKS ONLY NINTH IN **CITATION** NUMBERS.

FRAUD

Brawl in Beijing

Critics of Chinese researchers targeted in physical attacks.

New Scientist, 17th Nov 2012: http://www.newscientist.com/article/mg21628910.300-fraud-fighter-faked-research-is-endemic-in-china.html
Nature, 29th September 2010: http://www.nature.com/news/2010/100929/full/467511a.html
Science, 29th November 2013: http://www.sciencemag.org/content/342/6162/1035.full
Nature 20th July 2011: http://www.nature.com/news/2011/110720/full/475267a.html

Global perceptions of Chinese Research

Million RMB rewards for high IF publications = ?

“Wide distribution of information is key to scientific progress, yet traditionally, Chinese scientists have not systematically released data or research findings, even after publication.“

“There have been widespread complaints from scientists inside and outside China about this lack of transparency. ”

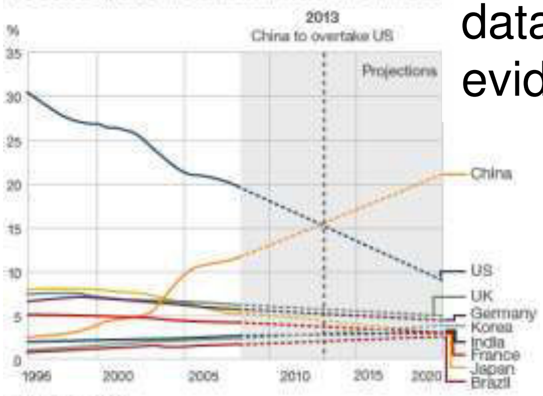
“Usually incomplete and unsystematic, [what little supporting data released] are of little value to researchers and there is evidence that this drives down a paper's citation numbers.”

NewScientist

“Faked research is endemic in China”

China's Publication Bazaar

Projected growth in citations in scientific literatur



Source: Royal Society

nature

International weekly journal of science

475, 267 (2011)

Focus on quality, not just quantity

China publishes huge amounts of scientific research. Now it must mal more of it worth reading, says **Changhui Peng.**

ALTHOUGH CHINA RANKS SECOND IN TERMS OF **PUBLICATION** OUTPUT, IT RANKS ONLY NINTH IN **CITATION** NUMBERS.

FRAUD

Brawl in Beijing

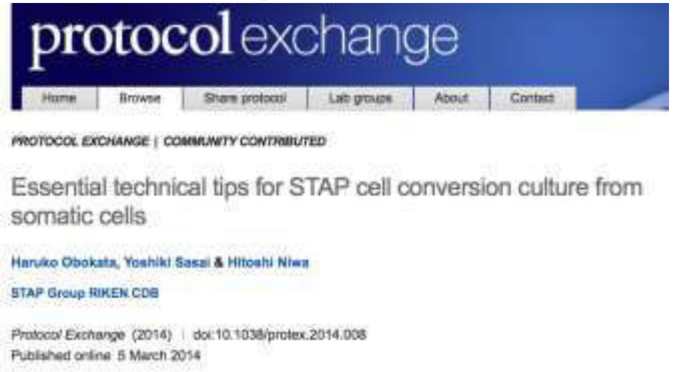
Critics of Chinese researchers targeted in physical attacks.

New Scientist, 17th Nov 2012: http://www.newscientist.com/article/mg21628910.300-fraud-fighter-faked-research-is-endemic-in-china.html
Nature, 29th September 2010: http://www.nature.com/news/2010/100929/full/467511a.html
Science, 29th November 2013: http://www.sciencemag.org/content/342/6162/1035.full
Nature 20th July 2011: http://www.nature.com/news/2011/110720/full/475267a.html

Issues not just in China…

Need:

…to publish protocols BEFORE analysis
…better access to supporting data
…more transparent & accountable review



PROTOCOL EXCHANGE | COMMUNITY CONTRIBUTED

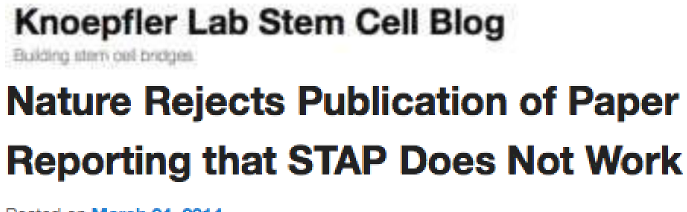
Essential technical tips for STAP cell conversion culture from somatic cells

Haruko Obokata, Yoshiki Sasai & Hitoshi Niwa

STAP Group RIKEN CDB

Protocol Exchange (2014) | doi:10.1038/protex.2014.008
Published online 5 March 2014

…to publish replication studies

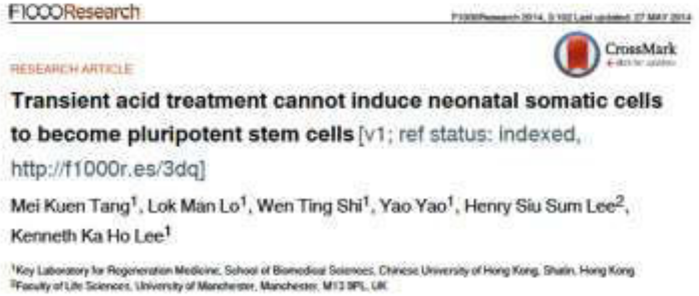


Knoepfler Lab Stem Cell Blog

Building stem cell bridges

Nature Rejects Publication of Paper Reporting that STAP Does Not Work

Posted on March 24, 2014



F1000Research 2014, 3:102 Last updated: 27 MAY 2014

RESEARCH ARTICLE

Transient acid treatment cannot induce neonatal somatic cells to become pluripotent stem cells [v1; ref status: indexed, http://f1000r.es/3dq]

Mei Kuen Tang[1], Lok Man Lo[1], Wen Ting Shi[1], Yao Yao[1], Henry Siu Sum Lee[2], Kenneth Ka Ho Lee[1]

[1] Key Laboratory for Regeneration Medicine, School of Biomedical Sciences, Chinese University of Hong Kong, Shatin, Hong Kong
[2] Faculty of Life Sciences, University of Manchester, Manchester, M13 9PL, UK

New incentives/credit

Credit where credit is overdue:

“One option would be to provide researchers who release data to public repositories with a means of accreditation.”
“An ability to search the literature for all online papers that used a particular data set would enable appropriate attribution for those who share. “
Nature Biotechnology **27**, 579 (2009)

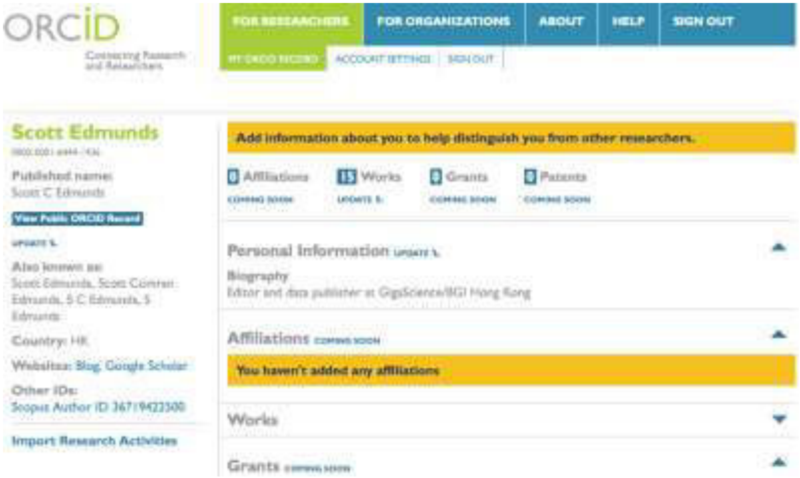
- **Data**
- **Software**
- **Review**
- **Re-use…**

}

= Credit

ODIN
ORCID and DataCite Interoperability Network

DataCite



Scott Edmunds
[illegible]

Published name:
Scott C Edmunds

Also known as:
Scott Edmunds, Scott Conrad Edmunds, S C Edmunds, S Edmunds

Country: HK

Websites: Blog, Google Scholar

Other IDs:
Scopus Author ID 36719422300

Add information about you to help distinguish you from other researchers.

0 Affiliations COMING SOON
15 Works UPDATE
0 Grants COMING SOON
0 Patents COMING SOON

Personal Information

Biography
Editor and data publisher at GigaScience/BGI Hong Kong

Affiliations COMING SOON

You haven't added any affiliations

Works

Grants COMING SOON

GigaSolution: deconstructing the paper

Combines and integrates:

(GIGA)^n SCIENCE — Open-access journal — OPEN ACCESS

(GIGA)^n DB — Data Publishing Platform

(GIGA)^n Galaxy by CBIIT — Data Analysis Platform — open source

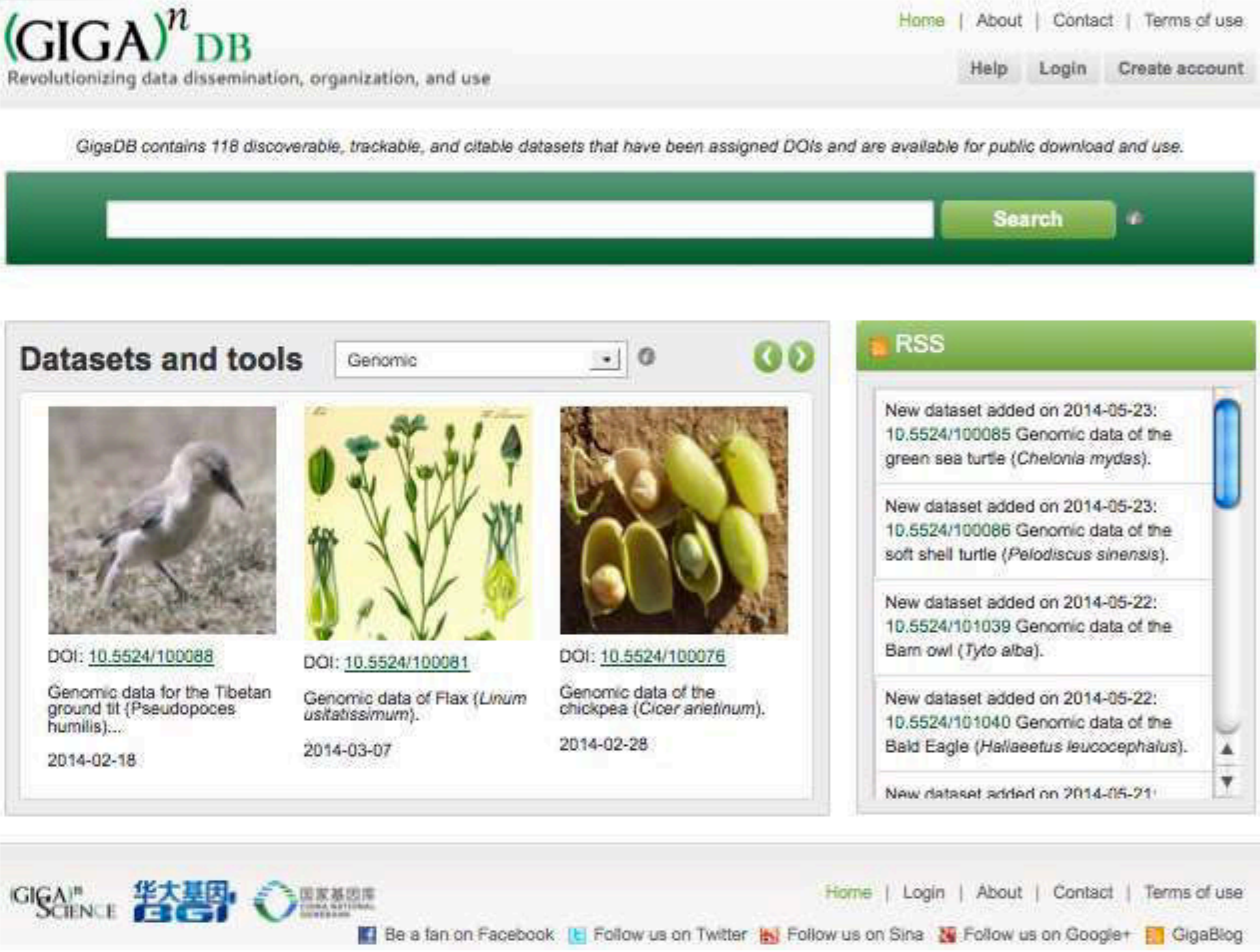
Utilizes big-data infrastructure and expertise from:

华大基因 BGI

www.gigadb.org
www.gigasciencejournal.com

BioMed Central
The Open Access Publisher

Rewarding open data



(GIGA)^n DB

Revolutionizing data dissemination, organization, and use

GigaDB contains 118 discoverable, trackable, and citable datasets that have been assigned DOIs and are available for public download and use.

Datasets and tools

DOI: 10.5524/100088

Genomic data for the Tibetan ground tit (Pseudopoces humilis):...

2014-02-18

DOI: 10.5524/100081

Genomic data of Flax (*Linum usitatissimum*).

2014-03-07

DOI: 10.5524/100076

Genomic data of the chickpea (*Cicer arietinum*).

2014-02-28

RSS

New dataset added on 2014-05-23: 10.5524/100085 Genomic data of the green sea turtle (*Chelonia mydas*).

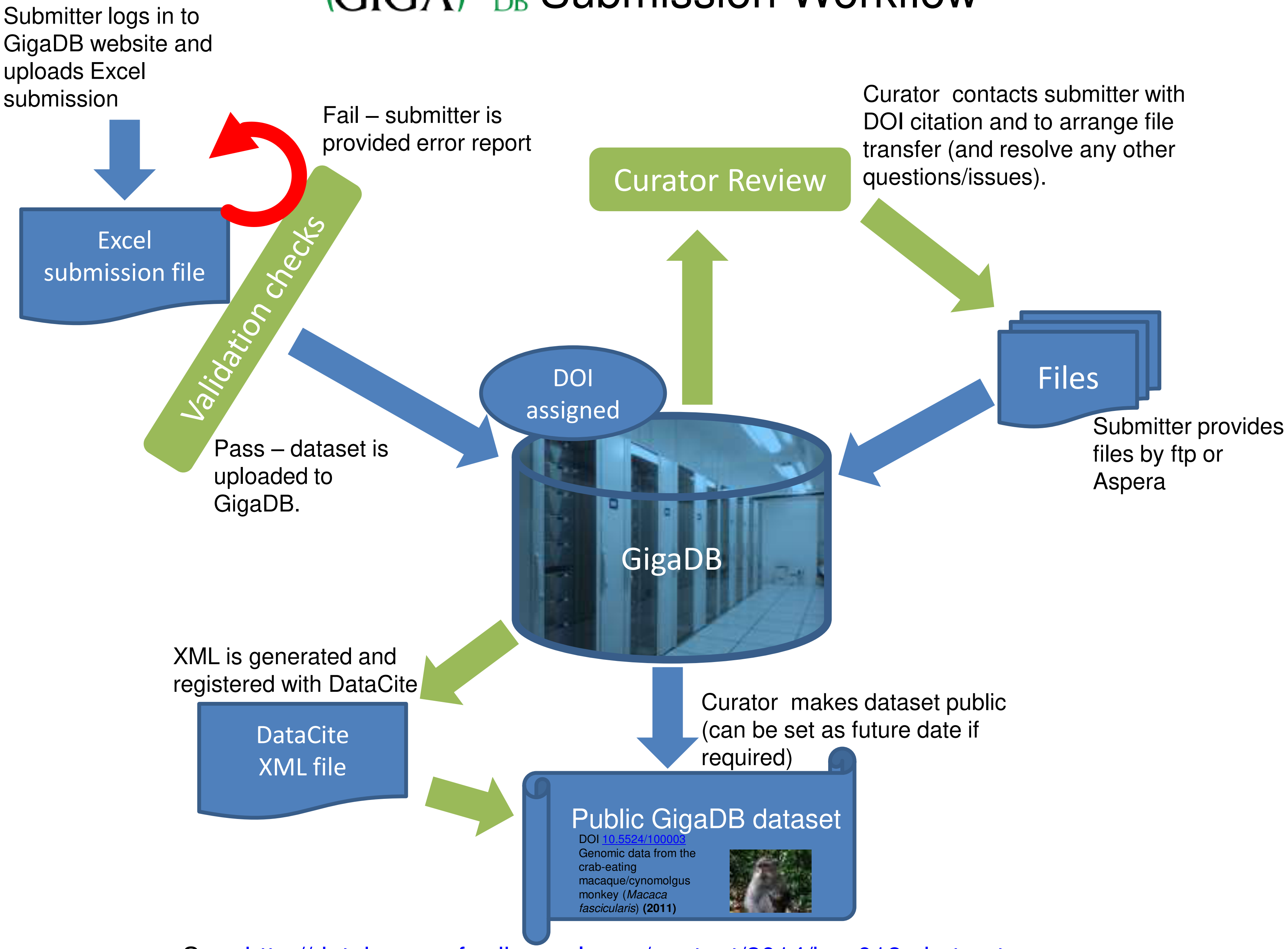
New dataset added on 2014-05-23: 10.5524/100086 Genomic data of the soft shell turtle (*Pelodiscus sinensis*).

New dataset added on 2014-05-22: 10.5524/101039 Genomic data of the Barn owl (*Tyto alba*).

New dataset added on 2014-05-22: 10.5524/101040 Genomic data of the Bald Eagle (*Haliaeetus leucocephalus*).

New dataset added on 2014-05-21:

(GIGA)^{n}DB Submission Workflow



Submitter logs in to GigaDB website and uploads Excel submission

Excel submission file

Fail – submitter is provided error report

Validation checks

Pass – dataset is uploaded to GigaDB.

DOI assigned

GigaDB

Curator Review

Curator contacts submitter with DOI citation and to arrange file transfer (and resolve any other questions/issues).

Files

Submitter provides files by ftp or Aspera

XML is generated and registered with DataCite

DataCite XML file

Curator makes dataset public (can be set as future date if required)

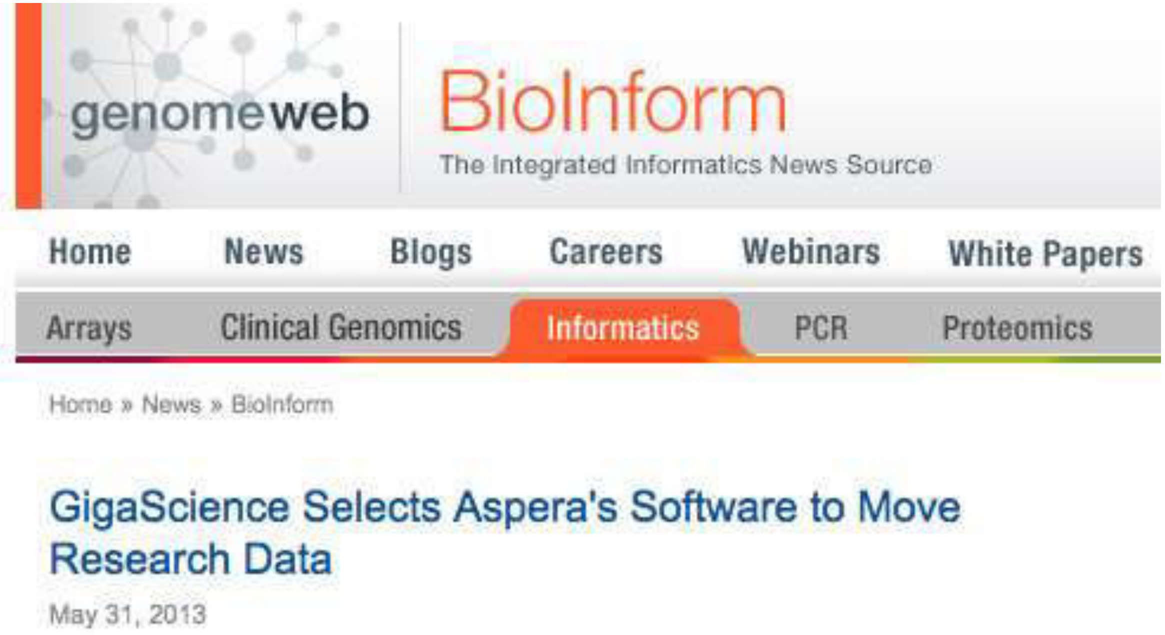
Public GigaDB dataset

DOI 10.5524/100003

Genomic data from the crab-eating macaque/cynomolgus monkey (*Macaca fascicularis*) **(2011)**

See: http://database.oxfordjournals.org/content/2014/bau018.abstract

(GIGA)$^{n}_{\mathrm{DB}}$ aspera



GigaScience Selects Aspera's Software to Move Research Data

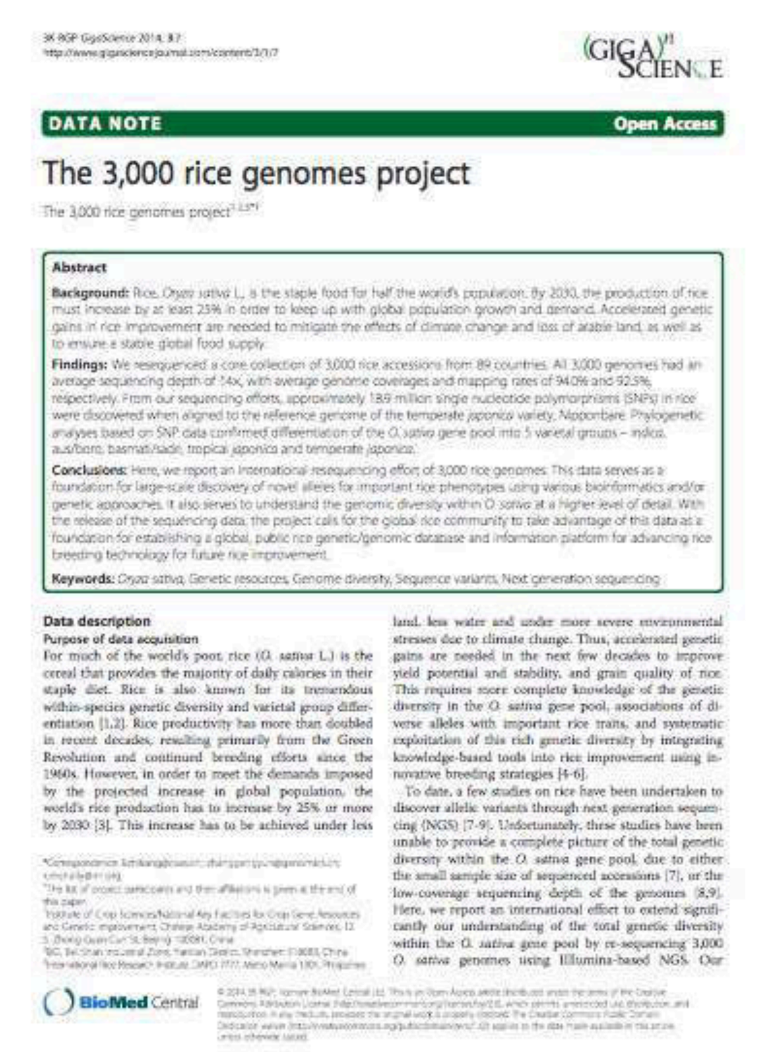
May 31, 2013

- 10-100x faster download than FTP
- Provide curation & integration with other DBs

Beneficiaries of this open data?

Beneficiaries of this open data?

Rice 3K project: 3,000 rice genomes, 13.4TB public data



3K RGP GigaScience 2014, 3:7
http://www.gigasciencejournal.com/content/3/1/7

(GIGA)ⁿ SCIENCE

DATA NOTE | Open Access

The 3,000 rice genomes project

The 3,000 rice genomes project[1,2,3*†]

Abstract

Background: Rice, *Oryza sativa* L., is the staple food for half the world's population. By 2030, the production of rice must increase by at least 25% in order to keep up with global population growth and demand. Accelerated genetic gains in rice improvement are needed to mitigate the effects of climate change and loss of arable land, as well as to ensure a stable global food supply.

Findings: We resequenced a core collection of 3,000 rice accessions from 89 countries. All 3,000 genomes had an average sequencing depth of 14x, with average genome coverages and mapping rates of 94.0% and 92.5%, respectively. From our sequencing efforts, approximately 18.9 million single nucleotide polymorphisms (SNPs) in rice were discovered when aligned to the reference genome of the temperate *japonica* variety, Nipponbare. Phylogenetic analyses based on SNP data confirmed differentiation of the *O. sativa* gene pool into 5 varietal groups – *indica*, *aus/boro*, *basmati/sadri*, tropical *japonica* and temperate *japonica*.

Conclusions: Here, we report an international resequencing effort of 3,000 rice genomes. This data serves as a foundation for large-scale discovery of novel alleles for important rice phenotypes using various bioinformatics and/or genetic approaches. It also serves to understand the genomic diversity within *O. sativa* at a higher level of detail. With the release of the sequencing data, the project calls for the global rice community to take advantage of this data as a foundation for establishing a global, public rice genetic/genomic database and information platform for advancing rice breeding technology for future rice improvement.

Keywords: *Oryza sativa*, Genetic resources, Genome diversity, Sequence variants, Next generation sequencing

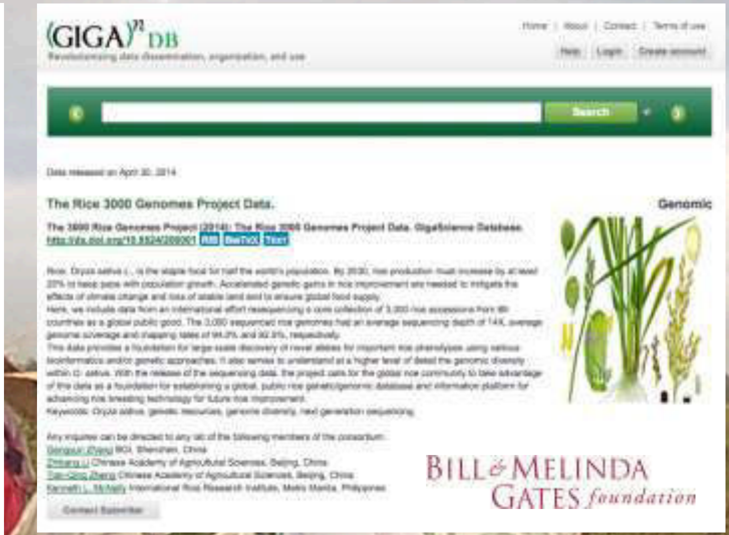
Data description

Purpose of data acquisition

For much of the world's poor, rice (*O. sativa* L.) is the cereal that provides the majority of daily calories in their staple diet. Rice is also known for its tremendous within-species genetic diversity and varietal group differentiation [1,2]. Rice productivity has more than doubled in recent decades, resulting primarily from the Green Revolution and continued breeding efforts since the 1960s. However, in order to meet the demands imposed by the projected increase in global population, the world's rice production has to increase by 25% or more by 2030 [3]. This increase has to be achieved under less land, less water and under more severe environmental stresses due to climate change. Thus, accelerated genetic gains are needed in the next few decades to improve yield potential and stability, and grain quality of rice. This requires more complete knowledge of the genetic diversity in the *O. sativa* gene pool, associations of diverse alleles with important rice traits, and systematic exploitation of this rich genetic diversity by integrating knowledge-based tools into rice improvement using innovative breeding strategies [4-6].

To date, a few studies on rice have been undertaken to discover allelic variants through next generation sequencing (NGS) [7-9]. Unfortunately, these studies have been unable to provide a complete picture of the total genetic diversity within the *O. sativa* gene pool, due to either the small sample size of sequenced accessions [7], or the low-coverage sequencing depth of the genomes [8,9]. Here, we report an international effort to extend significantly our understanding of the total genetic diversity within the *O. sativa* gene pool by re-sequencing 3,000 *O. sativa* genomes using Illumina-based NGS. Our

BioMed Central



(GIGA)ⁿ DB
Revolutionizing data dissemination, organization, and use
Home | About | Contact | Terms of use
Help | Login | Create account
Search
Genomic
BILL & MELINDA GATES foundation
Contact Submitter

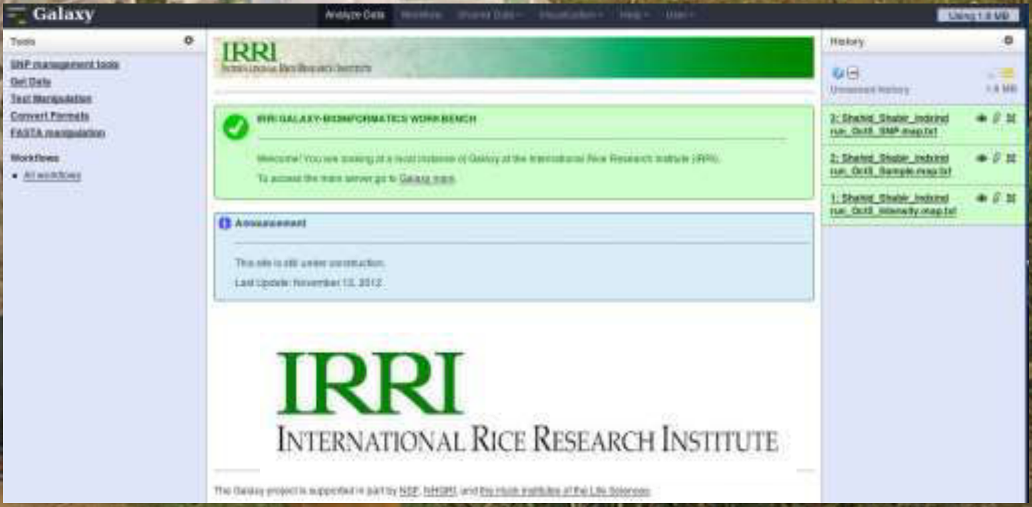
Data released on April 30, 2014

The Rice 3000 Genomes Project Data.

The 3000 Rice Genomes Project (2014): The Rice 3000 Genomes Project Data. GigaScience Database.
http://dx.doi.org/10.5524/200001 RIS BibTeX Text

Rice, Oryza sativa L., is the staple food for half the world's population. By 2030, rice production must increase by at least 25% to keep pace with population growth. Accelerated genetic gains in rice improvement are needed to mitigate the effects of climate change and loss of arable land and to ensure global food supply.
Here, we include data from an international effort resequencing a core collection of 3,000 rice accessions from 89 countries as a global public good. The 3,000 sequenced rice genomes had an average sequencing depth of 14X, average genome coverage and mapping rates of 94.0% and 92.5%, respectively.
This data provides a foundation for large-scale discovery of novel alleles for important rice phenotypes using various bioinformatics and/or genetic approaches. It also serves to understand at a higher level of detail the genomic diversity within O. sativa. With the release of the sequencing data, the project calls for the global rice community to take advantage of this data as a foundation for establishing a global, public rice genetic/genomic database and information platform for advancing rice breeding technology for future rice improvement.
Keywords: Oryza sativa, genetic resources, genome diversity, next generation sequencing

Any inquiries can be directed to any (all) of the following members of the consortium:
Gengyun Zhang BGI, Shenzhen, China
Zhikang Li Chinese Academy of Agricultural Sciences, Beijing, China
Tian-Qing Zheng Chinese Academy of Agricultural Sciences, Beijing, China
Kenneth L. McNally International Rice Research Institute, Metro Manila, Philippines



Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 1.8 MB
Tools
SNP management tools
Get Data
Text Manipulation
Convert Formats
FASTA manipulation
Workflows
All workflows
IRRI
International Rice Research Institute
History
Unnamed history
1.8 MB
3: Shahid_Shabir_indxind run_Oct8_SNP map.txt
2: Shahid_Shabir_indxind run_Oct8_Sample map.txt
1: Shahid_Shabir_indxind run_Oct8_intensity map.txt
IRRI
INTERNATIONAL RICE RESEARCH INSTITUTE

IRRI GALAXY-BIOINFORMATICS WORKBENCH

Welcome! You are looking at a local instance of Galaxy at the International Rice Research Institute (IRRI).

To access the main server go to Galaxy main.

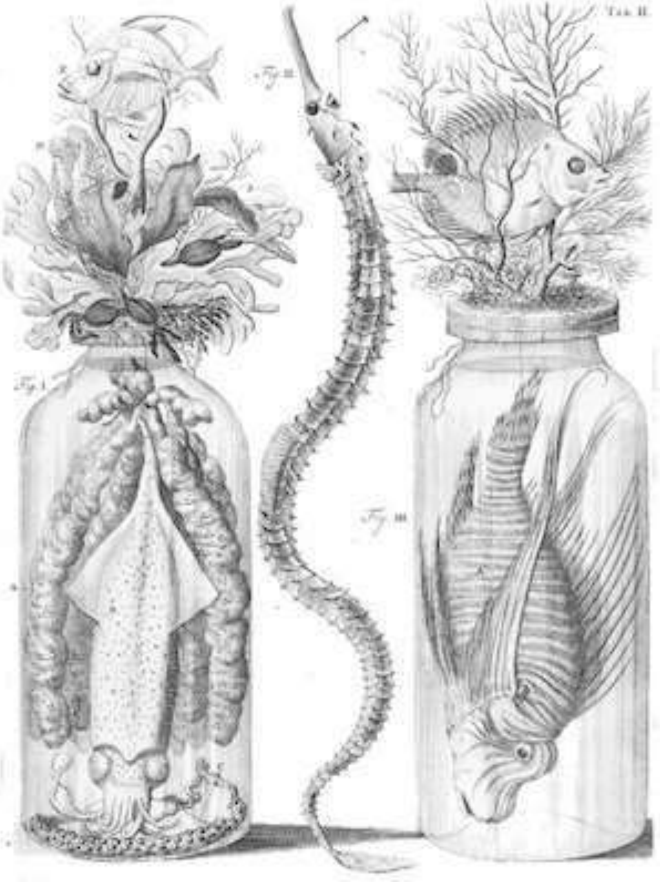
Announcement

This site is still under construction.

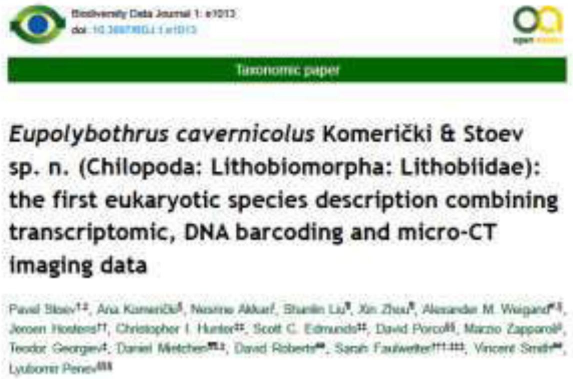
Last Update: November 13, 2012

The Galaxy project is supported in part by NSF, NHGRI, and the Huck Institutes of the Life Sciences.

New Article types v Species Description <2012



Collaborations with Pensoft & PLOS
Cyber-centipedes & virtual worms

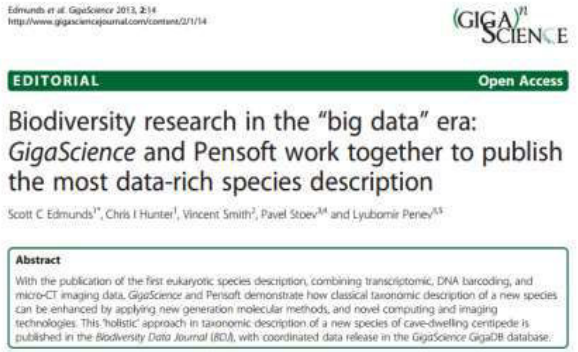


Biodiversity Data Journal 1: e1013
doi: 10.3897/BDJ.1.e1013

Taxonomic paper

Eupolybothrus cavernicolus Komerički & Stoev sp. n. (Chilopoda: Lithobiomorpha: Lithobiidae): the first eukaryotic species description combining transcriptomic, DNA barcoding and micro-CT imaging data

Pavel Stoev[†‡], Ana Komerički[§], Nesrine Akkari[|], Shanlin Liu[¶], Xin Zhou[¶], Alexander M. Weigand[#¤], Jeroen Hostens[††], Christopher I. Hunter[‡‡], Scott C. Edmunds[‡‡], David Porco[§§], Marzio Zapparoli[||], Teodor Georgiev[‡], Daniel Mietchen[¶¶‡], David Roberts[##], Sarah Faulwetter[†††‡‡‡], Vincent Smith[##], Lyubomir Penev[‡§§§]



Edmunds et al. GigaScience 2013, 2:14
http://www.gigasciencejournal.com/content/2/1/14

EDITORIAL Open Access

Biodiversity research in the “big data” era: *GigaScience* and Pensoft work together to publish the most data-rich species description

Scott C Edmunds[1*], Chris I Hunter[1], Vincent Smith[2], Pavel Stoev[3,4] and Lyubomir Penev[4,5]

Abstract

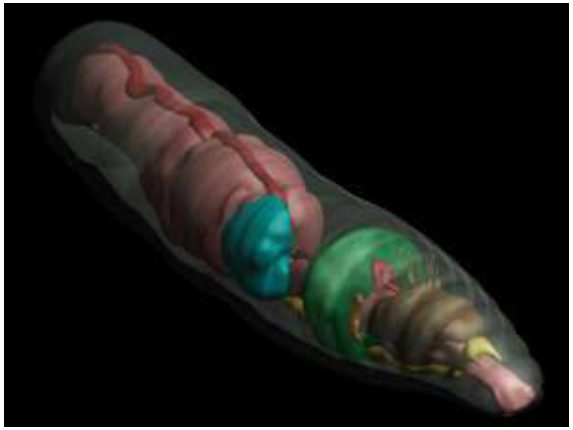
With the publication of the first eukaryotic species description, combining transcriptomic, DNA barcoding, and micro-CT imaging data, *GigaScience* and Pensoft demonstrate how classical taxonomic description of a new species can be enhanced by applying new generation molecular methods, and novel computing and imaging technologies. This 'holistic' approach in taxonomic description of a new species of cave-dwelling centipede is published in the *Biodiversity Data Journal (BDJ)*, with coordinated data release in the *GigaScience* GigaDB database.

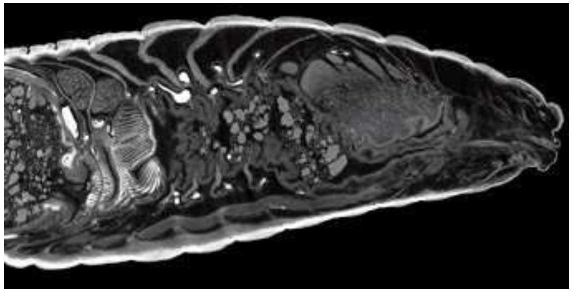
(GIGA)n DB
Home | About | Contact | Terms of use
Help | Login | Create account
Search

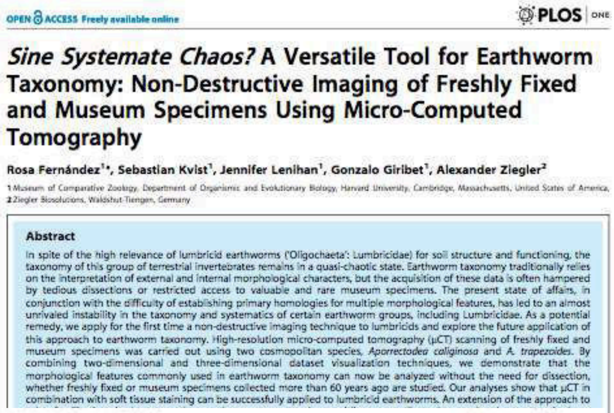
Transcriptomic, DNA barcoding, and micro-CT imaging data from an advanced taxonomic description of a novel centipede species (Eupolybothrus cavernicolus Komerički & Stoev, sp. n.)

Genomic, Transcriptomic, Imaging

[illegible]







OPEN ACCESS Freely available online

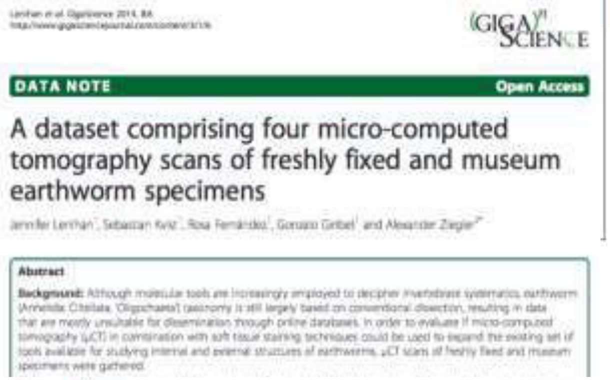
Sine Systemate Chaos? A Versatile Tool for Earthworm Taxonomy: Non-Destructive Imaging of Freshly Fixed and Museum Specimens Using Micro-Computed Tomography

Rosa Fernández[1*], Sebastian Kvist[1], Jennifer Lenihan[1], Gonzalo Giribet[1], Alexander Ziegler[2]

1 Museum of Comparative Zoology, Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts, United States of America, 2 Ziegler Biosolutions, Waldshut-Tiengen, Germany

Abstract

In spite of the high relevance of lumbricid earthworms ('Oligochaeta': Lumbricidae) for soil structure and functioning, the taxonomy of this group of terrestrial invertebrates remains in a quasi-chaotic state. Earthworm taxonomy traditionally relies on the interpretation of external and internal morphological characters, but the acquisition of these data is often hampered by tedious dissections or restricted access to valuable and rare museum specimens. The present state of affairs, in conjunction with the difficulty of establishing primary homologies for multiple morphological features, has led to an almost unrivaled instability in the taxonomy and systematics of certain earthworm groups, including Lumbricidae. As a potential remedy, we apply for the first time a non-destructive imaging technique to lumbricids and explore the future application of this approach to earthworm taxonomy. High-resolution micro-computed tomography (µCT) scanning of freshly fixed and museum specimens was carried out using two cosmopolitan species, *Aporrectodea caliginosa* and *A. trapezoides*. By combining two-dimensional and three-dimensional dataset visualization techniques, we demonstrate that the morphological features commonly used in earthworm taxonomy can now be analyzed without the need for dissection, whether freshly fixed or museum specimens collected more than 60 years ago are studied. Our analyses show that µCT in combination with soft tissue staining can be successfully applied to lumbricid earthworms. An extension of the approach to



Lenihan et al. GigaScience 2014, 3:6
http://www.gigasciencejournal.com/content/3/1/6

DATA NOTE Open Access

A dataset comprising four micro-computed tomography scans of freshly fixed and museum earthworm specimens

Jennifer Lenihan[1], Sebastian Kvist[1], Rosa Fernández[1], Gonzalo Giribet[1] and Alexander Ziegler[2*]

Abstract

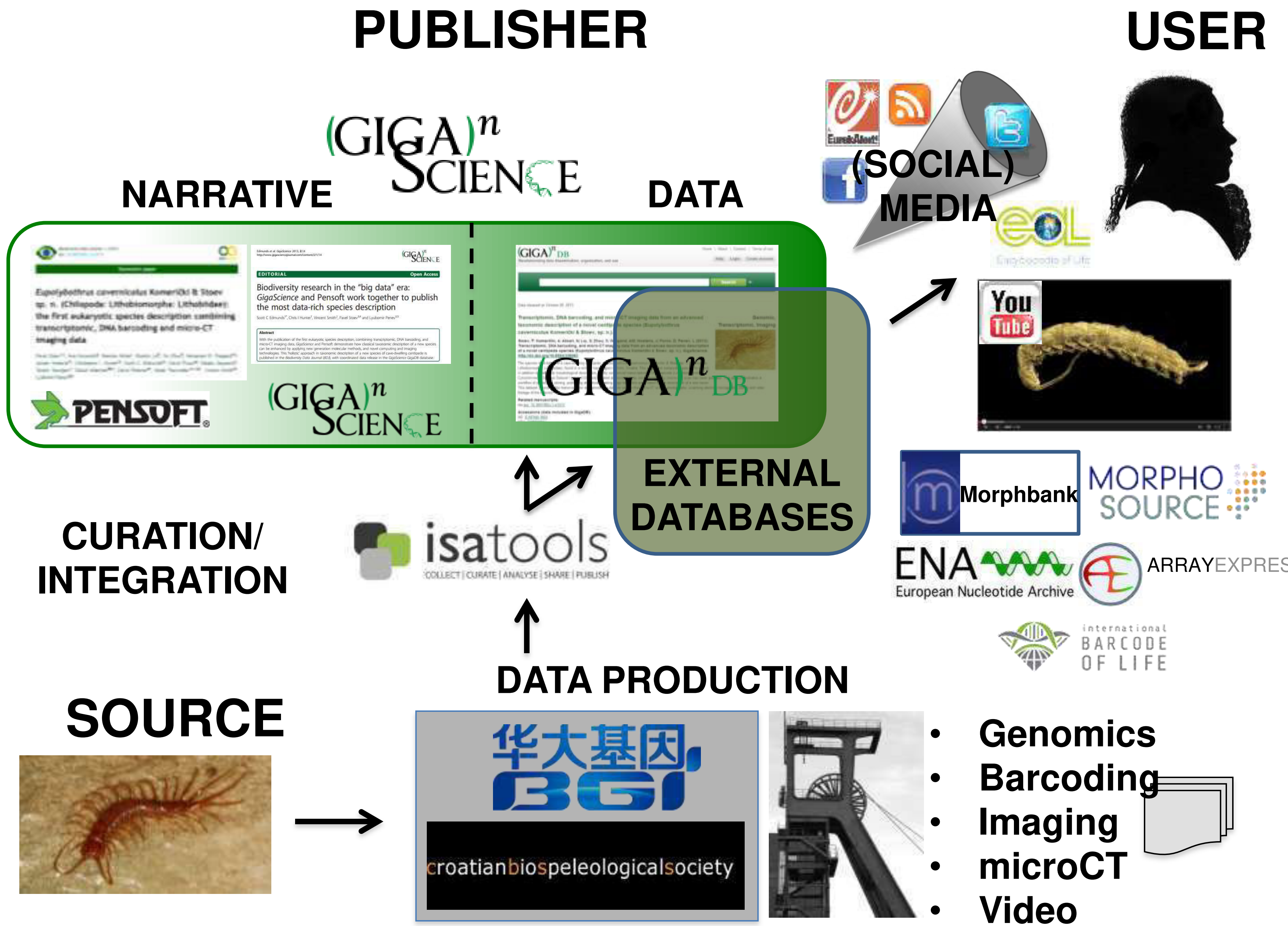
Background: Although molecular tools are increasingly employed to decipher invertebrate systematics, earthworm (Annelida: Clitellata: 'Oligochaeta') taxonomy is still largely based on conventional dissection, resulting in data that are mostly unsuitable for dissemination through online databases. In order to evaluate if micro-computed tomography (µCT) in combination with soft tissue staining techniques could be used to expand the existing set of tools available for studying internal and external structures of earthworms, µCT scans of freshly fixed and museum specimens were gathered.

(GIGA)n DB
Home | About | Contact | Terms of use
Help | Login | Create account
Search

MicroCT scans of freshly fixed and museum earthworm specimens

Imaging

[illegible]


PUBLISHER
USER
(GIGA)^n SCIENCE
NARRATIVE
DATA
(SOCIAL) MEDIA
EurekAlert!
eol Encyclopedia of Life
You Tube
PENSOFT
(GIGA)^n SCIENCE
(GIGA)^n DB
EXTERNAL DATABASES
Morphbank
MORPHO SOURCE
ENA European Nucleotide Archive
ARRAYEXPRES
international BARCODE OF LIFE
CURATION/ INTEGRATION
isatools
COLLECT | CURATE | ANALYSE | SHARE | PUBLISH
DATA PRODUCTION
SOURCE
华大基因 BGI
croatianbiospeleologicalsociety
Genomics
Barcoding
Imaging
microCT
Video
Eupolybothrus cavernicolus Komerički & Stoev sp. n. (Chilopoda: Lithobiomorpha: Lithobiidae): the first eukaryotic species description combining transcriptomic, DNA barcoding and micro-CT imaging data
Transcriptomic, DNA barcoding, and micro-CT imaging data from an advanced taxonomic description of a novel centipede species (Eupolybothrus cavernicolus Komerički & Stoev, sp. n.)
Genomic, Transcriptomic, Imaging

Edmunds et al. GigaScience 2013, 2:14
http://www.gigasciencejournal.com/content/2/1/14

EDITORIAL Open Access

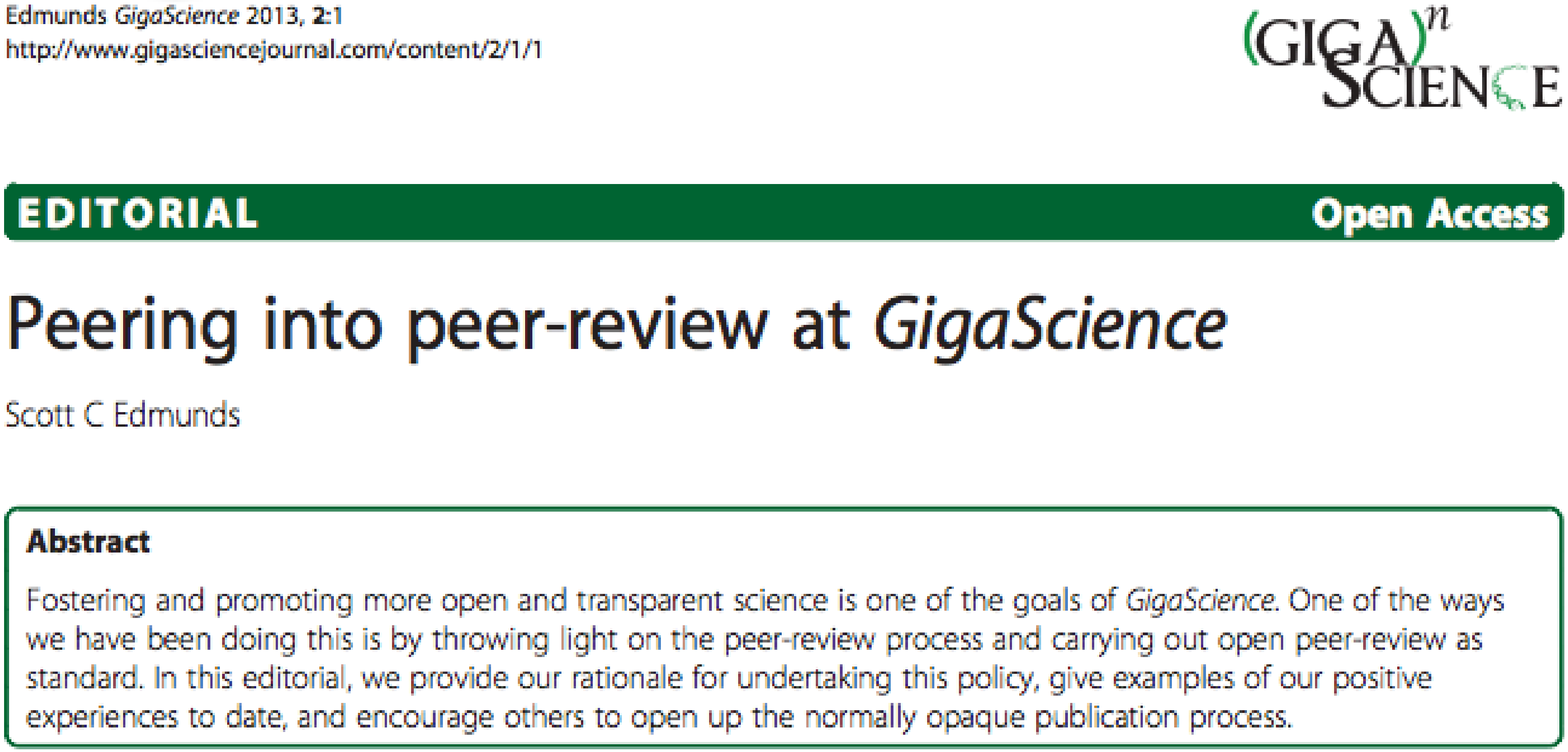
Biodiversity research in the "big data" era: *GigaScience* and Pensoft work together to publish the most data-rich species description

Scott C Edmunds[1*], Chris I Hunter[1], Vincent Smith[2], Pavel Stoev[3,4] and Lyubomir Penev[3,5]

Abstract

With the publication of the first eukaryotic species description, combining transcriptomic, DNA barcoding, and micro-CT imaging data, *GigaScience* and Pensoft demonstrate how classical taxonomic description of a new species can be enhanced by applying new generation molecular methods, and novel computing and imaging technologies. This 'holistic' approach in taxonomic description of a new species of cave-dwelling centipede is published in the *Biodiversity Data Journal* (BDJ), with coordinated data release in the *GigaScience* GigaDB database.

New & more transparent peer-review: open review



Edmunds *GigaScience* 2013, **2**:1
http://www.gigasciencejournal.com/content/2/1/1

(GIGA)^n SCIENCE

EDITORIAL Open Access

Peering into peer-review at *GigaScience*

Scott C Edmunds

Abstract

Fostering and promoting more open and transparent science is one of the goals of *GigaScience*. One of the ways we have been doing this is by throwing light on the peer-review process and carrying out open peer-review as standard. In this editorial, we provide our rationale for undertaking this policy, give examples of our positive experiences to date, and encourage others to open up the normally opaque publication process.

BioData Mining

BMC Series Medical Journals

BMJ

Reward open & transparent review

End reviewer 3 Downfall parody videos, now!

New & more transparent peer-review: pre-prints

Cornell University Library

arXiv.org

PeerJ PrePrints

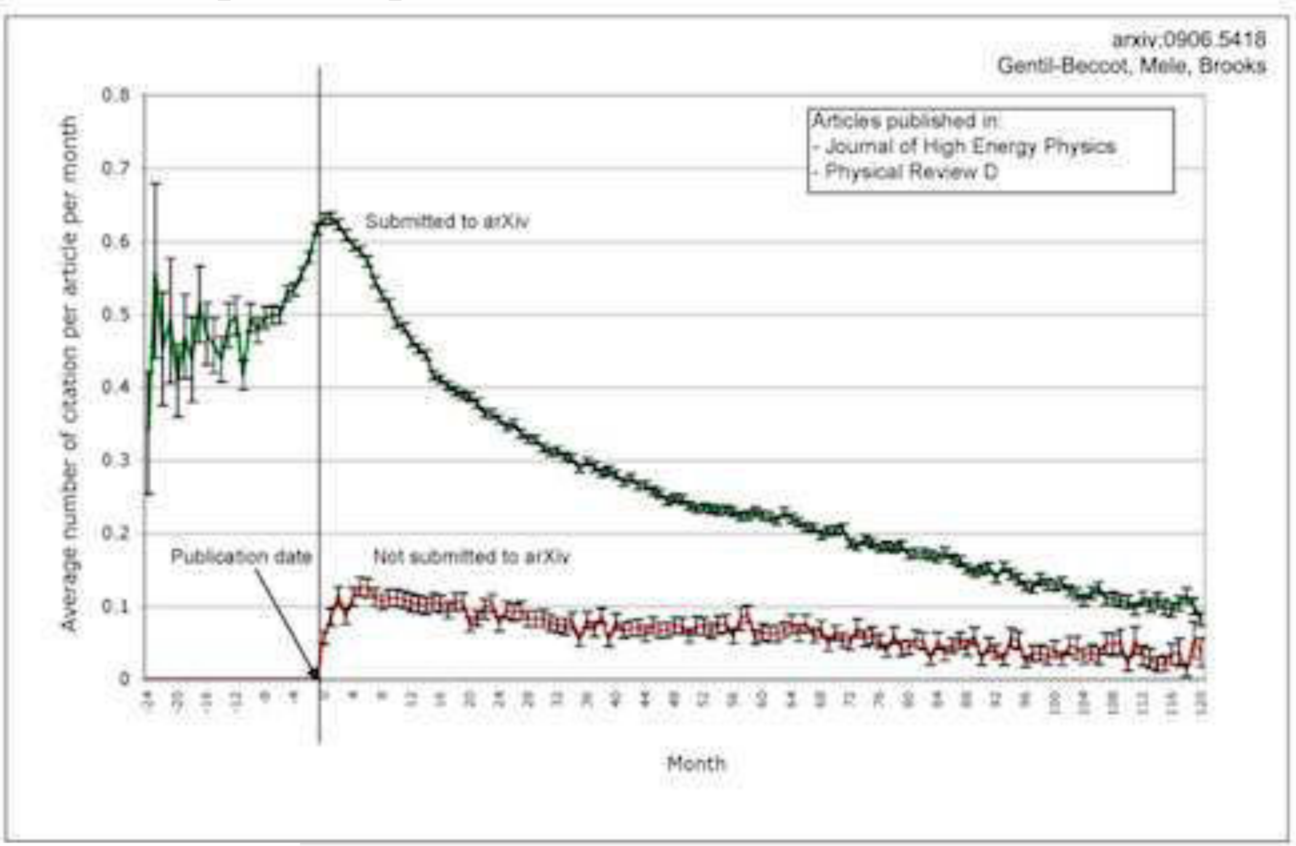
bioRxiv beta

THE PREPRINT SERVER FOR BIOLOGY

natureprecedings

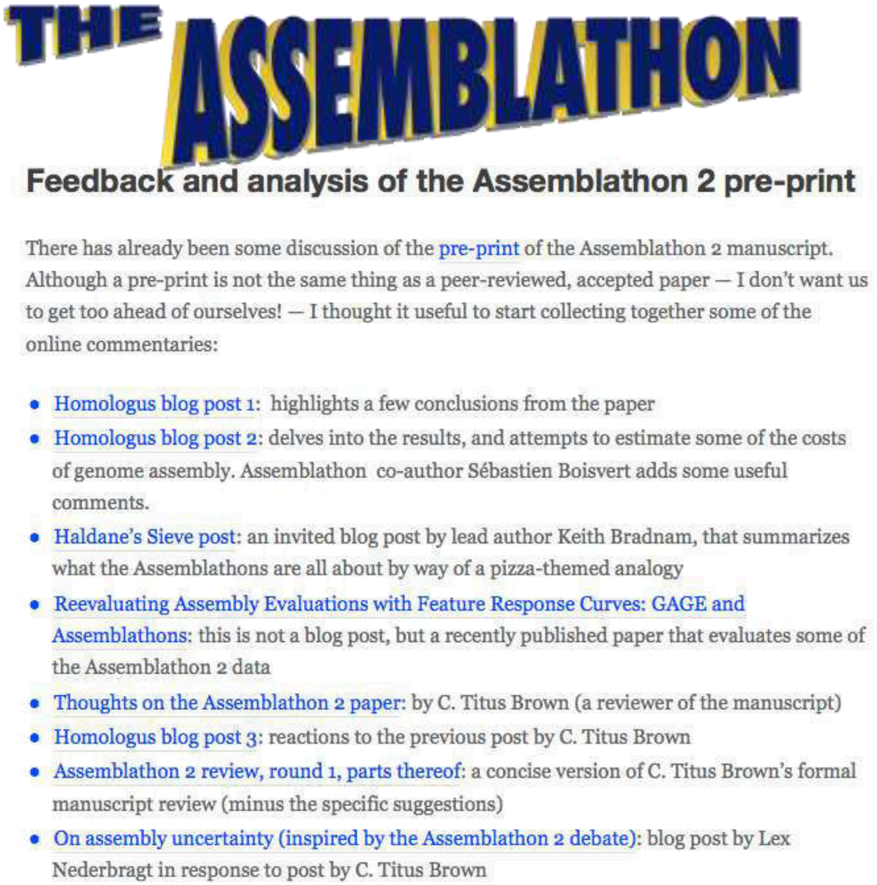
Haldane's Sieve

Discussing preprints in population and evolutionary genetics


arxiv:0906.5418
Gentil-Beccot, Mele, Brooks

Reward open & transparent review

Real-time open-review = paper in arXiv + blogged reviews



Feedback and analysis of the Assemblathon 2 pre-print

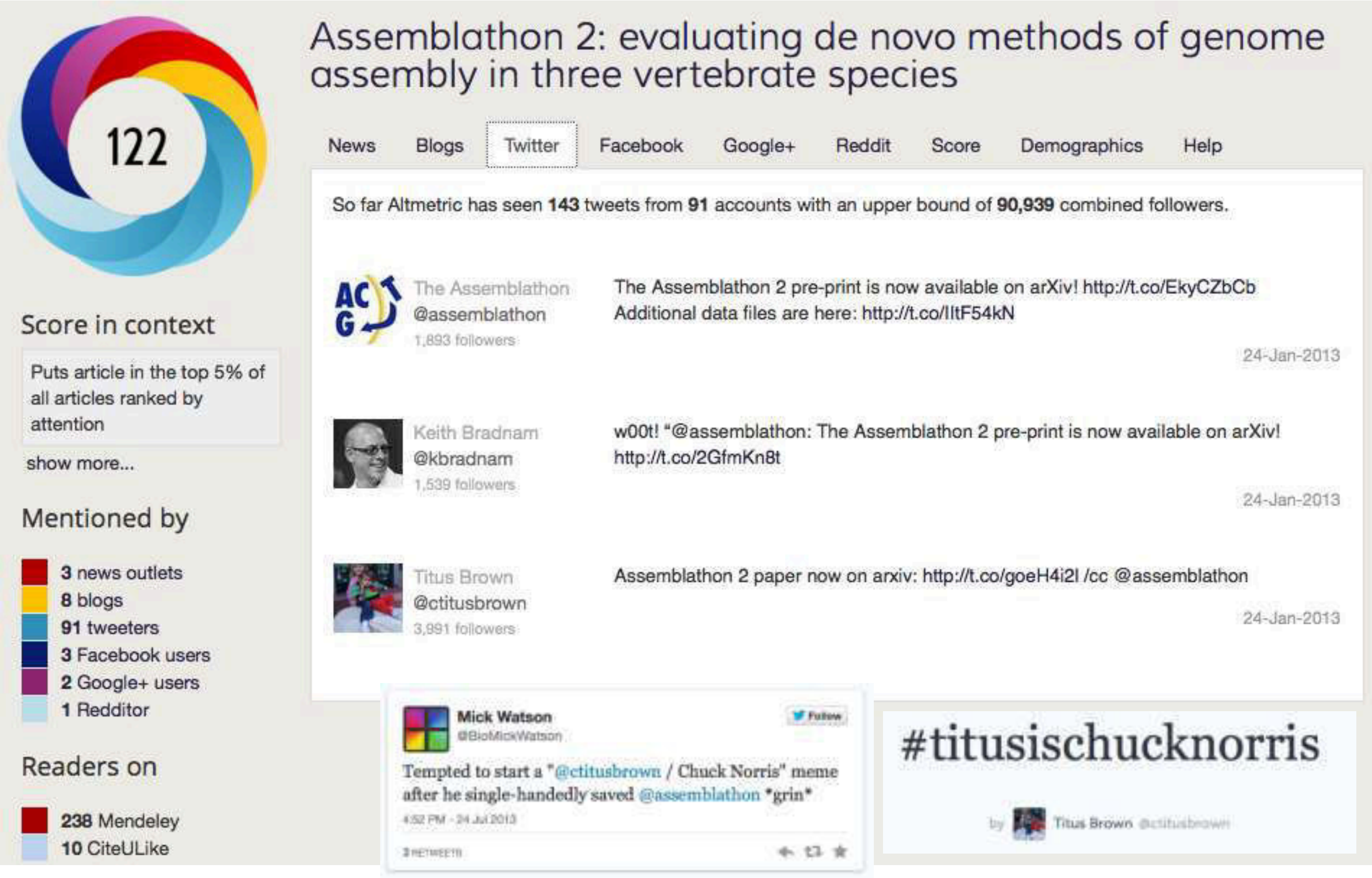
There has already been some discussion of the pre-print of the Assemblathon 2 manuscript. Although a pre-print is not the same thing as a peer-reviewed, accepted paper — I don't want us to get too ahead of ourselves! — I thought it useful to start collecting together some of the online commentaries:

- Homologus blog post 1: highlights a few conclusions from the paper
- Homologus blog post 2: delves into the results, and attempts to estimate some of the costs of genome assembly. Assemblathon co-author Sébastien Boisvert adds some useful comments.
- Haldane's Sieve post: an invited blog post by lead author Keith Bradnam, that summarizes what the Assemblathons are all about by way of a pizza-themed analogy
- Reevaluating Assembly Evaluations with Feature Response Curves: GAGE and Assemblathons: this is not a blog post, but a recently published paper that evaluates some of the Assemblathon 2 data
- Thoughts on the Assemblathon 2 paper: by C. Titus Brown (a reviewer of the manuscript)
- Homologus blog post 3: reactions to the previous post by C. Titus Brown
- Assemblathon 2 review, round 1, parts thereof: a concise version of C. Titus Brown's formal manuscript review (minus the specific suggestions)
- On assembly uncertainty (inspired by the Assemblathon 2 debate): blog post by Lex Nederbragt in response to post by C. Titus Brown

www.gigasciencejournal.com/content/2/1/10 http://tmblr.co/ZzXdssfOMJfy

Reward open & transparent review

Real-time open-review = paper in arXiv + blogged reviews



122

Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species

So far Altmetric has seen **143** tweets from **91** accounts with an upper bound of **90,939** combined followers.

The Assemblathon
@assemblathon
1,893 followers

The Assemblathon 2 pre-print is now available on arXiv! http://t.co/EkyCZbCb Additional data files are here: http://t.co/lItF54kN

24-Jan-2013

Keith Bradnam
@kbradnam
1,539 followers

w00t! "@assemblathon: The Assemblathon 2 pre-print is now available on arXiv! http://t.co/2GfmKn8t

24-Jan-2013

Titus Brown
@ctitusbrown
3,991 followers

Assemblathon 2 paper now on arxiv: http://t.co/goeH4i2l /cc @assemblathon

24-Jan-2013

Score in context

Puts article in the top 5% of all articles ranked by attention

Mentioned by

- 🟥 **3** news outlets
- 🟨 **8** blogs
- 🟦 **91** tweeters
- 🟦 **3** Facebook users
- 🟪 **2** Google+ users
- ⬜ **1** Redditor

Readers on

- 🟥 **238** Mendeley
- ⬜ **10** CiteULike

Mick Watson
@BioMickWatson

Tempted to start a "@ctitusbrown / Chuck Norris" meme after he single-handedly saved @assemblathon *grin*

4:52 PM - 24 Jul 2013

3 RETWEETS

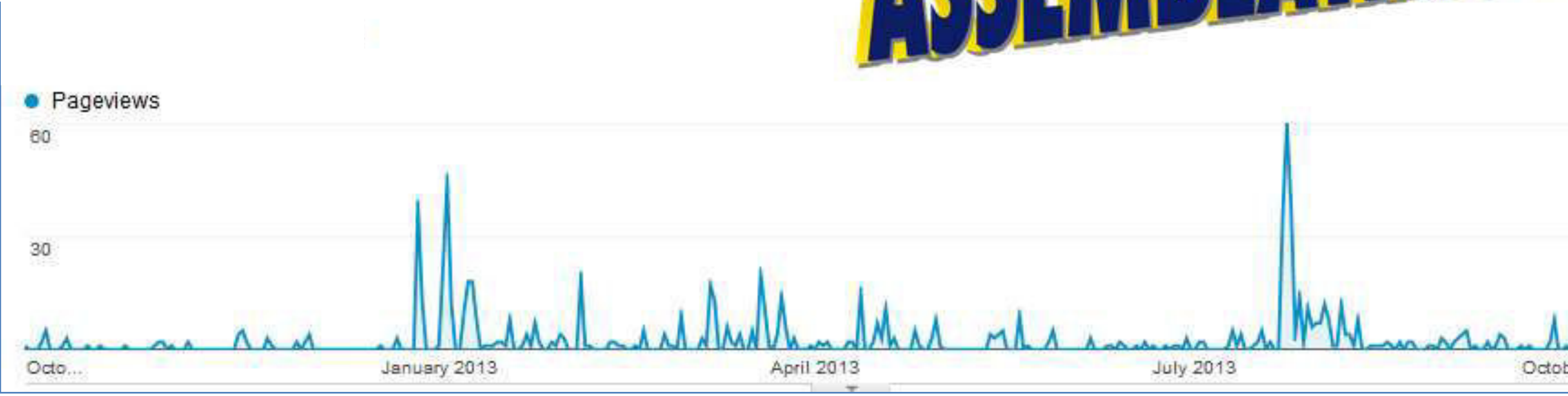
#titusischucknorris

by Titus Brown @ctitusbrown

Readers are interested in open review

SOAPdenovo2
Large-size genome de-novo assembler

THE ASSEMBLATHON



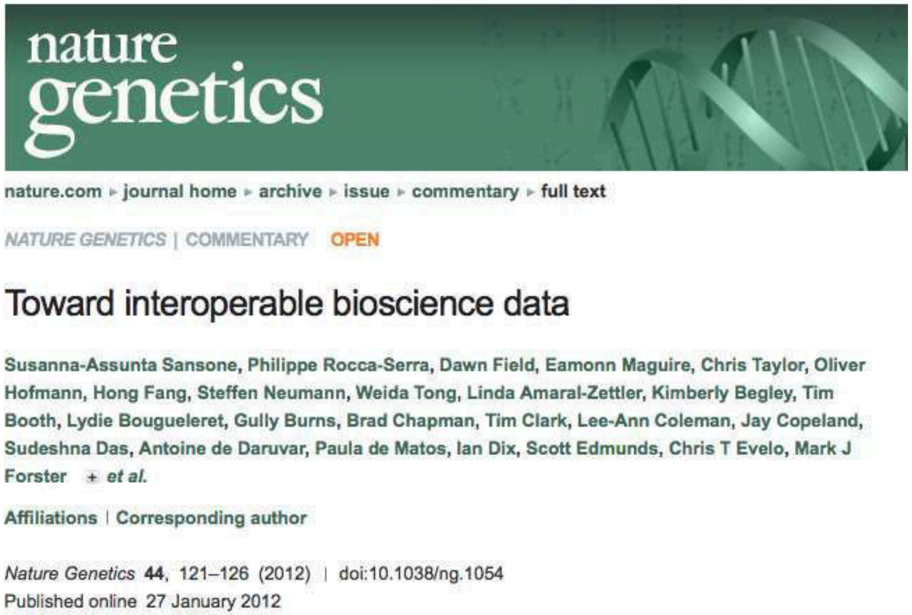
Next step to link to ORCID iD

casrai
Connecting Research

Reward better handling of metadata…

Novel tools/formats for data interoperability/handling.

isainfrastructure



nature genetics

nature.com ▸ journal home ▸ archive ▸ issue ▸ commentary ▸ full text

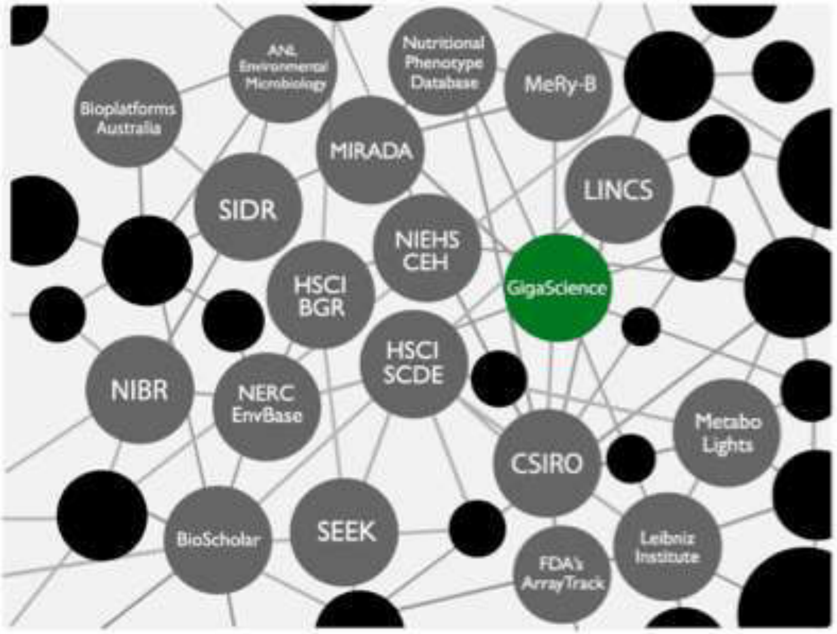
NATURE GENETICS | COMMENTARY OPEN

Toward interoperable bioscience data

Susanna-Assunta Sansone, Philippe Rocca-Serra, Dawn Field, Eamonn Maguire, Chris Taylor, Oliver Hofmann, Hong Fang, Steffen Neumann, Weida Tong, Linda Amaral-Zettler, Kimberly Begley, Tim Booth, Lydie Bougueleret, Gully Burns, Brad Chapman, Tim Clark, Lee-Ann Coleman, Jay Copeland, Sudeshna Das, Antoine de Daruvar, Paula de Matos, Ian Dix, Scott Edmunds, Chris T Evelo, Mark J Forster + *et al.*

Affiliations | Corresponding author

Nature Genetics **44**, 121–126 (2012) | doi:10.1038/ng.1054
Published online 27 January 2012



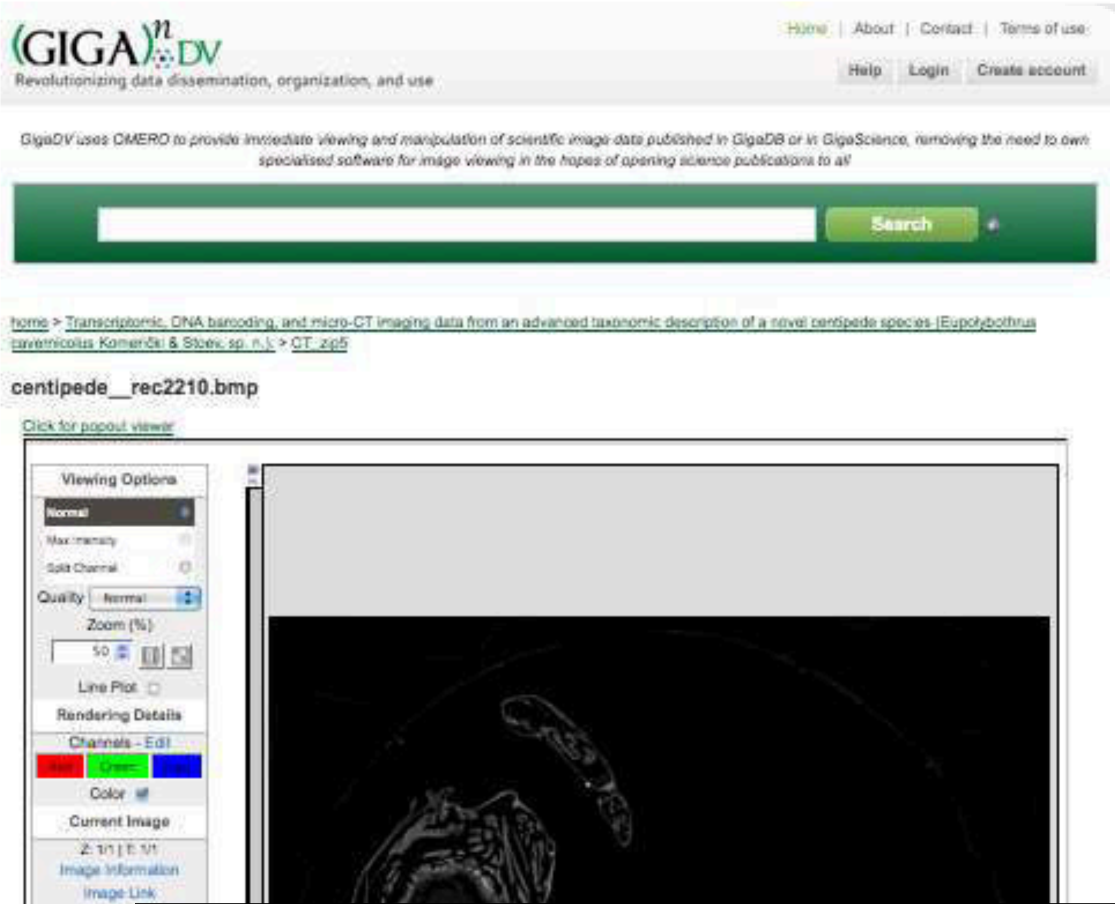
isacommons
isacommons.org

isacreator configurator

npg
nature publishing group
SCIENTIFIC DATA

SCIENTIFIC DATA

Rewarding and aiding reproducibility



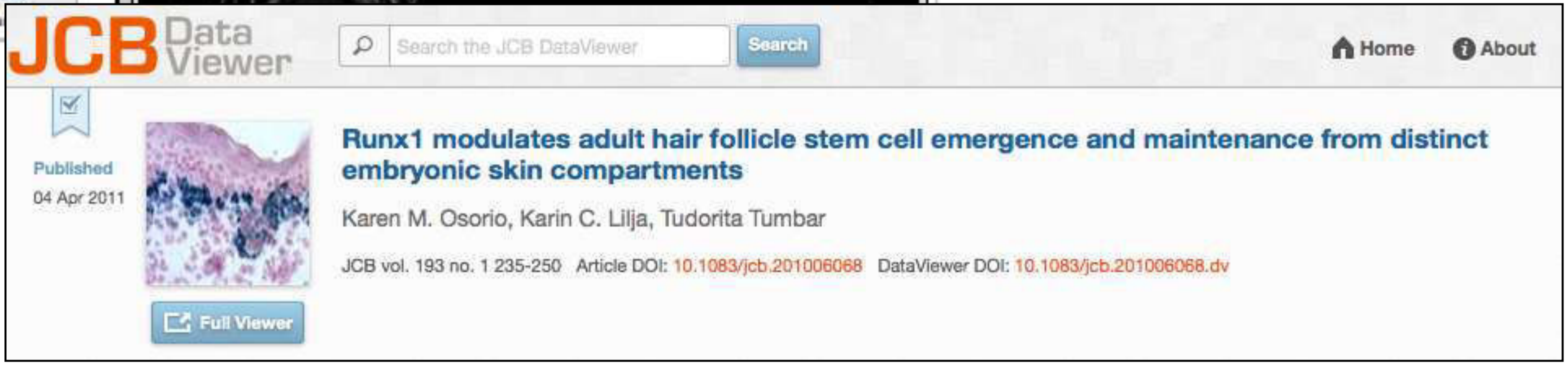
GigaDV uses OMERO to provide immediate viewing and manipulation of scientific image data published in GigaDB or in GigaScience, removing the need to own specialised software for image viewing in the hopes of opening science publications to all

home > Transcriptomic, DNA barcoding, and micro-CT imaging data from an advanced taxonomic description of a novel centipede species (Eupolybothrus cavernicolus Komerički & Stoev, sp. n.); > CT_zip5

centipede__rec2210.bmp

OME

OMERO: providing access to imaging data…



Published
04 Apr 2011

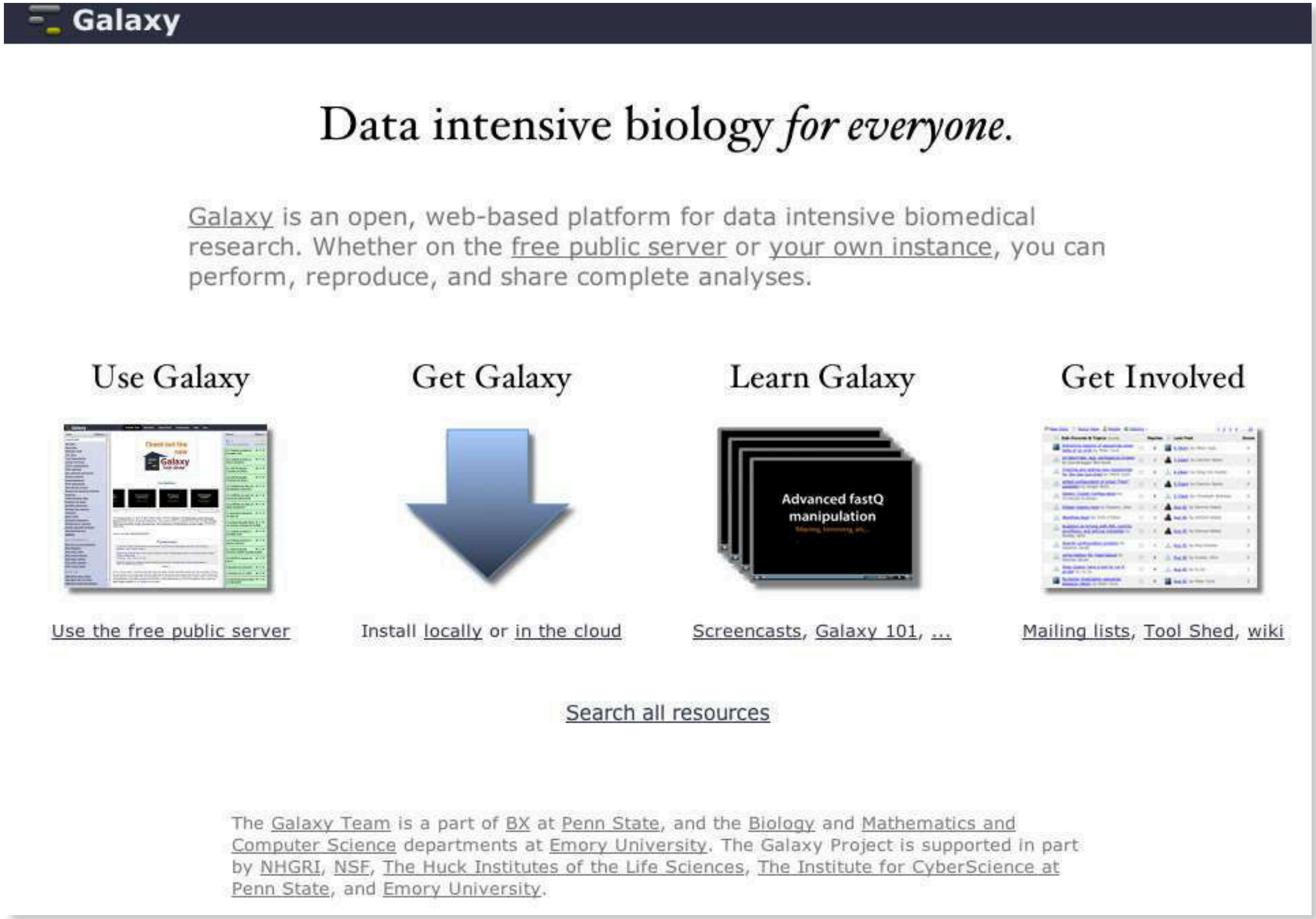
Runx1 modulates adult hair follicle stem cell emergence and maintenance from distinct embryonic skin compartments

Karen M. Osorio, Karin C. Lilja, Tudorita Tumbar

JCB vol. 193 no. 1 235-250 Article DOI: 10.1083/jcb.201006068 DataViewer DOI: 10.1083/jcb.201006068.dv

Rewarding and aiding reproducibility

Implement workflows in a community-accepted format



Galaxy

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	Get Galaxy	Learn Galaxy	Get Involved
Use the free public server	Install locally or in the cloud	Screencasts, Galaxy 101, ...	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.

Open source

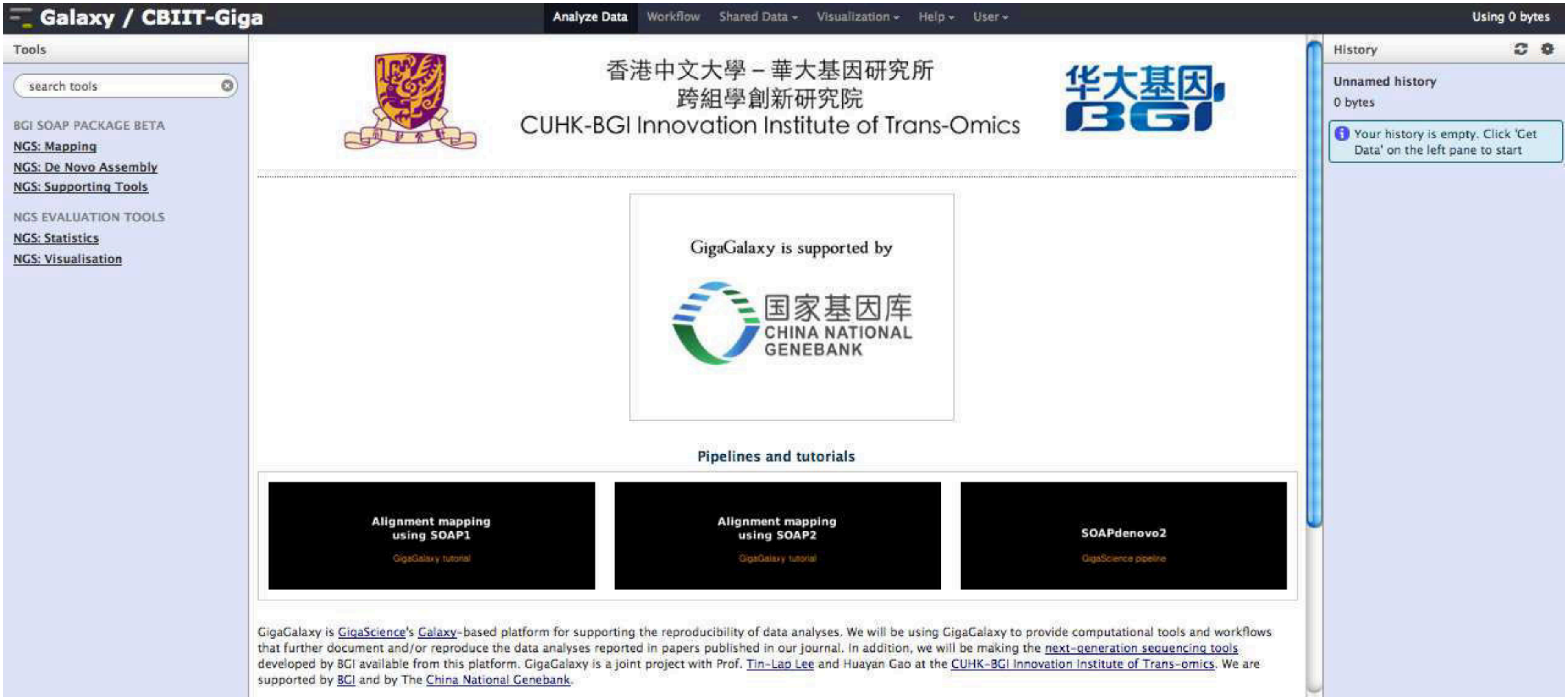
Over 36,000 main Galaxy server users

Over 1,000 papers citing Galaxy use

Over 55 Galaxy servers deployed

http://galaxyproject.org

(GIGA)^{n}Galaxy
by 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

华大基因 BGI

GigaGalaxy is supported by

国家基因库
CHINA NATIONAL GENEBANK

Pipelines and tutorials

Alignment mapping using SOAP1
GigaGalaxy tutorial

Alignment mapping using SOAP2
GigaGalaxy tutorial

SOAPdenovo2
GigaScience pipeline

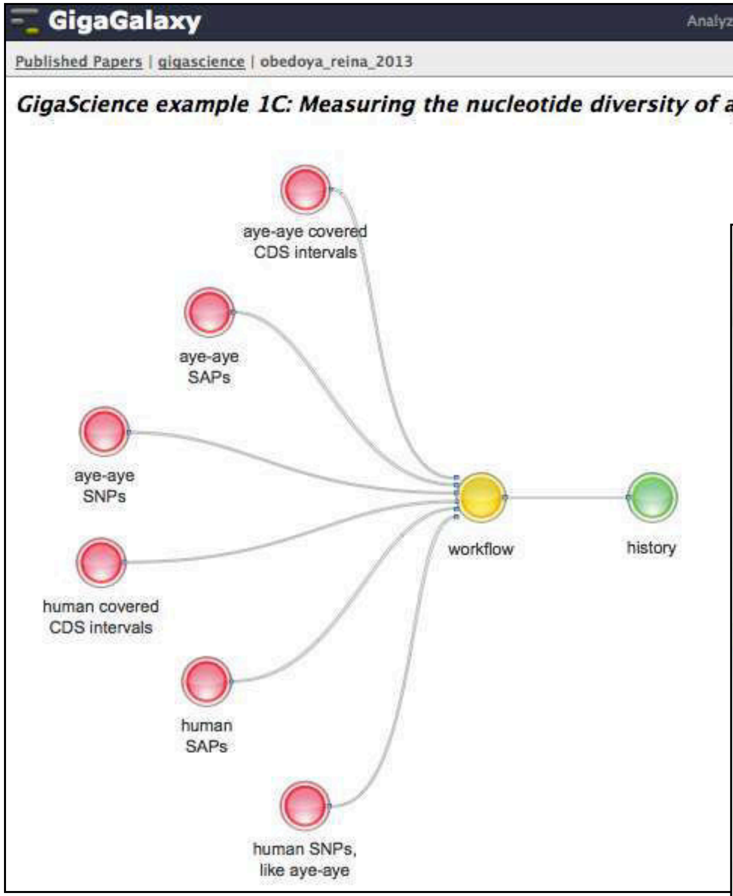
GigaGalaxy is GigaScience's Galaxy-based platform for supporting the reproducibility of data analyses. We will be using GigaGalaxy to provide computational tools and workflows that further document and/or reproduce the data analyses reported in papers published in our journal. In addition, we will be making the next-generation sequencing tools developed by BGI available from this platform. GigaGalaxy is a joint project with Prof. Tin-Lap Lee and Huayan Gao at the CUHK-BGI Innovation Institute of Trans-omics. We are supported by BGI and by The China National Genebank.

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

galaxy.cbiit.cuhk.edu.hk

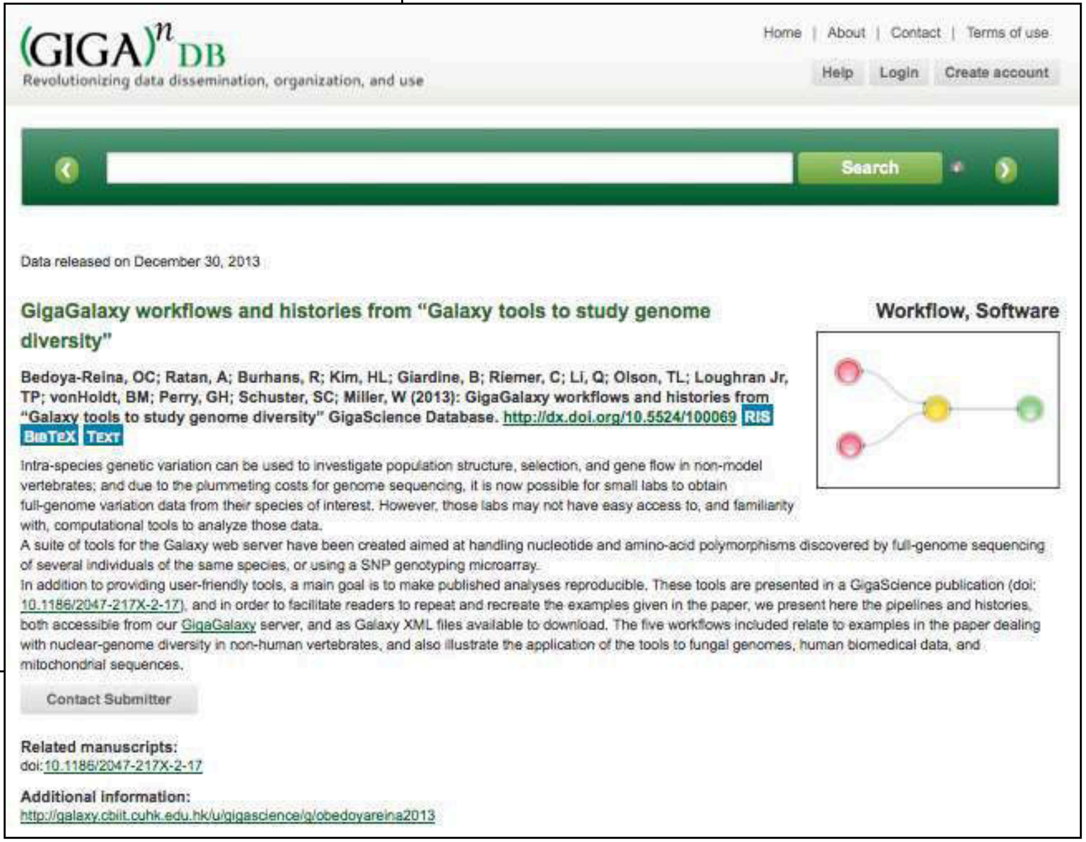
(GIGA)^n Galaxy by CBIIT

Visualizations & DOIs for workflows



Published Papers | gigascience | obedoya_reina_2013

GigaScience example 1C: Measuring the nucleotide diversity of aye-aye populations



Revolutionizing data dissemination, organization, and use

Data released on December 30, 2013

GigaGalaxy workflows and histories from "Galaxy tools to study genome diversity"

Workflow, Software

Bedoya-Reina, OC; Ratan, A; Burhans, R; Kim, HL; Giardine, B; Riemer, C; Li, Q; Olson, TL; Loughran Jr, TP; vonHoldt, BM; Perry, GH; Schuster, SC; Miller, W (2013): GigaGalaxy workflows and histories from "Galaxy tools to study genome diversity" GigaScience Database. http://dx.doi.org/10.5524/100069 RIS BibTeX Text

Intra-species genetic variation can be used to investigate population structure, selection, and gene flow in non-model vertebrates; and due to the plummeting costs for genome sequencing, it is now possible for small labs to obtain full-genome variation data from their species of interest. However, those labs may not have easy access to, and familiarity with, computational tools to analyze those data.

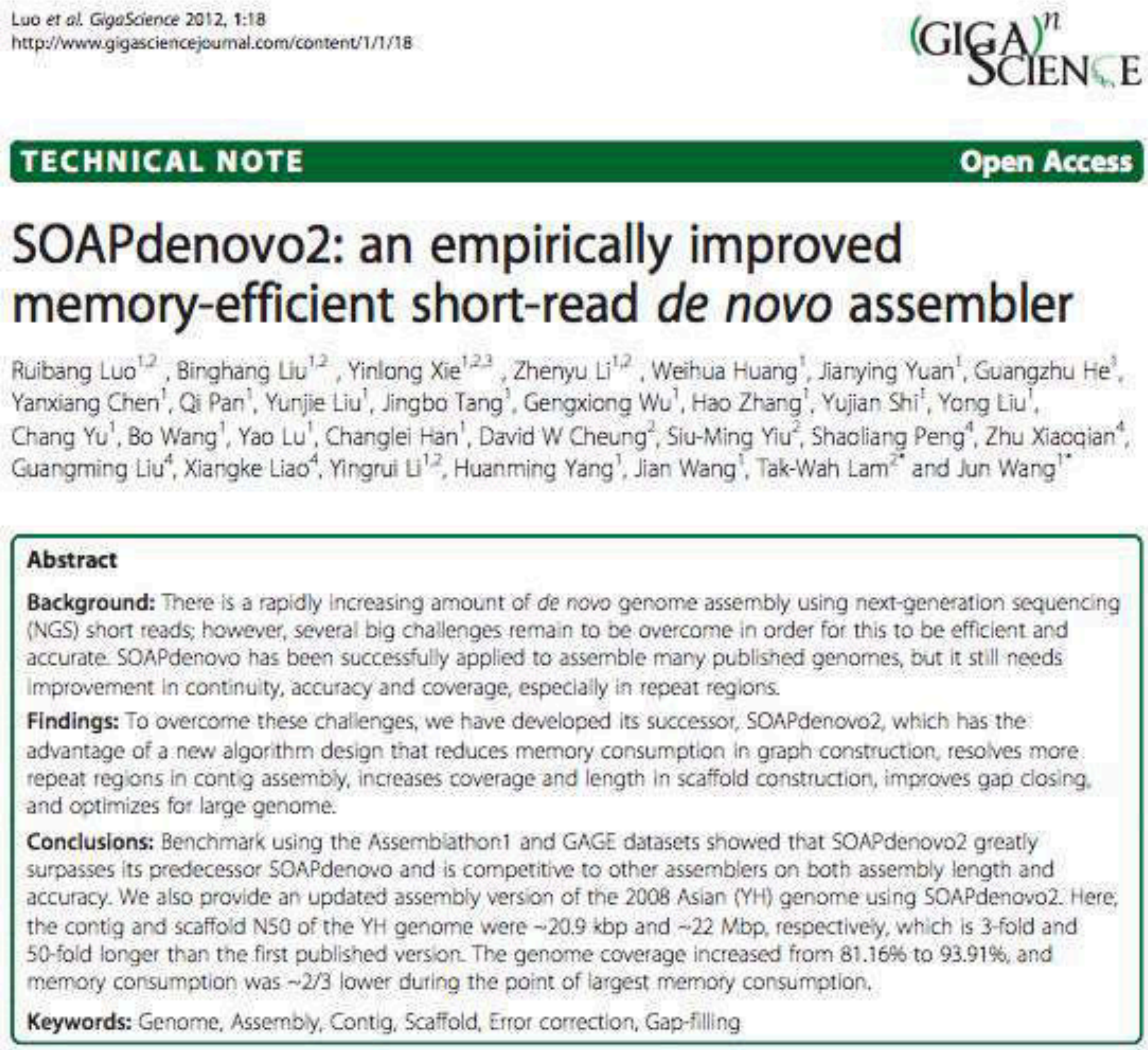
A suite of tools for the Galaxy web server have been created aimed at handling nucleotide and amino-acid polymorphisms discovered by full-genome sequencing of several individuals of the same species, or using a SNP genotyping microarray.

In addition to providing user-friendly tools, a main goal is to make published analyses reproducible. These tools are presented in a GigaScience publication (doi: 10.1186/2047-217X-2-17), and in order to facilitate readers to repeat and recreate the examples given in the paper, we present here the pipelines and histories, both accessible from our GigaGalaxy server, and as Galaxy XML files available to download. The five workflows included relate to examples in the paper dealing with nuclear-genome diversity in non-human vertebrates, and also illustrate the application of the tools to fungal genomes, human biomedical data, and mitochondrial sequences.

Contact Submitter

Related manuscripts:
doi:10.1186/2047-217X-2-17

Additional information:
http://galaxy.cbiit.cuhk.edu.hk/u/gigascience/g/obedoyareina2013



Luo *et al. GigaScience* 2012, **1**:18
http://www.gigasciencejournal.com/content/1/1/18

(GIGA)^n SCIENCE

TECHNICAL NOTE **Open Access**

SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler

Ruibang Luo[1,2], Binghang Liu[1,2], Yinlong Xie[1,2,3], Zhenyu Li[1,2], Weihua Huang[1], Jianying Yuan[1], Guangzhu He[1], Yanxiang Chen[1], Qi Pan[1], Yunjie Liu[1], Jingbo Tang[1], Gengxiong Wu[1], Hao Zhang[1], Yujian Shi[1], Yong Liu[1], Chang Yu[1], Bo Wang[1], Yao Lu[1], Changlei Han[1], David W Cheung[2], Siu-Ming Yiu[2], Shaoliang Peng[4], Zhu Xiaoqian[4], Guangming Liu[4], Xiangke Liao[4], Yingrui Li[1,2], Huanming Yang[1], Jian Wang[1], Tak-Wah Lam[2*] and Jun Wang[1*]

Abstract

Background: There is a rapidly increasing amount of *de novo* genome assembly using next-generation sequencing (NGS) short reads; however, several big challenges remain to be overcome in order for this to be efficient and accurate. SOAPdenovo has been successfully applied to assemble many published genomes, but it still needs improvement in continuity, accuracy and coverage, especially in repeat regions.

Findings: To overcome these challenges, we have developed its successor, SOAPdenovo2, which has the advantage of a new algorithm design that reduces memory consumption in graph construction, resolves more repeat regions in contig assembly, increases coverage and length in scaffold construction, improves gap closing, and optimizes for large genome.

Conclusions: Benchmark using the Assemblathon1 and GAGE datasets showed that SOAPdenovo2 greatly surpasses its predecessor SOAPdenovo and is competitive to other assemblers on both assembly length and accuracy. We also provide an updated assembly version of the 2008 Asian (YH) genome using SOAPdenovo2. Here, the contig and scaffold N50 of the YH genome were ~20.9 kbp and ~22 Mbp, respectively, which is 3-fold and 50-fold longer than the first published version. The genome coverage increased from 81.16% to 93.91%, and memory consumption was ~2/3 lower during the point of largest memory consumption.

Keywords: Genome, Assembly, Contig, Scaffold, Error correction, Gap-filling

How are we supporting data reproducibility?

Open-Paper

(GIGA)^n SCIENCE

[DOI:10.1186/2047-217X-1-18](DOI:10.1186/2047-217X-1-18)

Highly accessed >23,000 accesses

Linked to DOI

(GIGA)^n DB

Data sets

Open-Data

[DOI:10.5524/100038](DOI:10.5524/100038)

CC 0 ZERO

78GB CC0 data

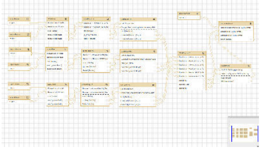
DOI Linked to

(GIGA)^n Galaxy by CBIIT

Analyses

Open-Pipelines

Open-Workflows

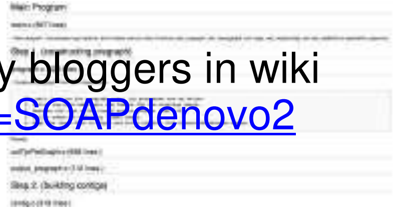


[DOI:10.5524/100044](DOI:10.5524/100044)

Open-Review

7 reviewers tested data in ftp server & named reports published

Enabled code to being picked apart by bloggers in wiki http://homolog.us/wiki/index.php?title=SOAPdenovo2



Open-Code

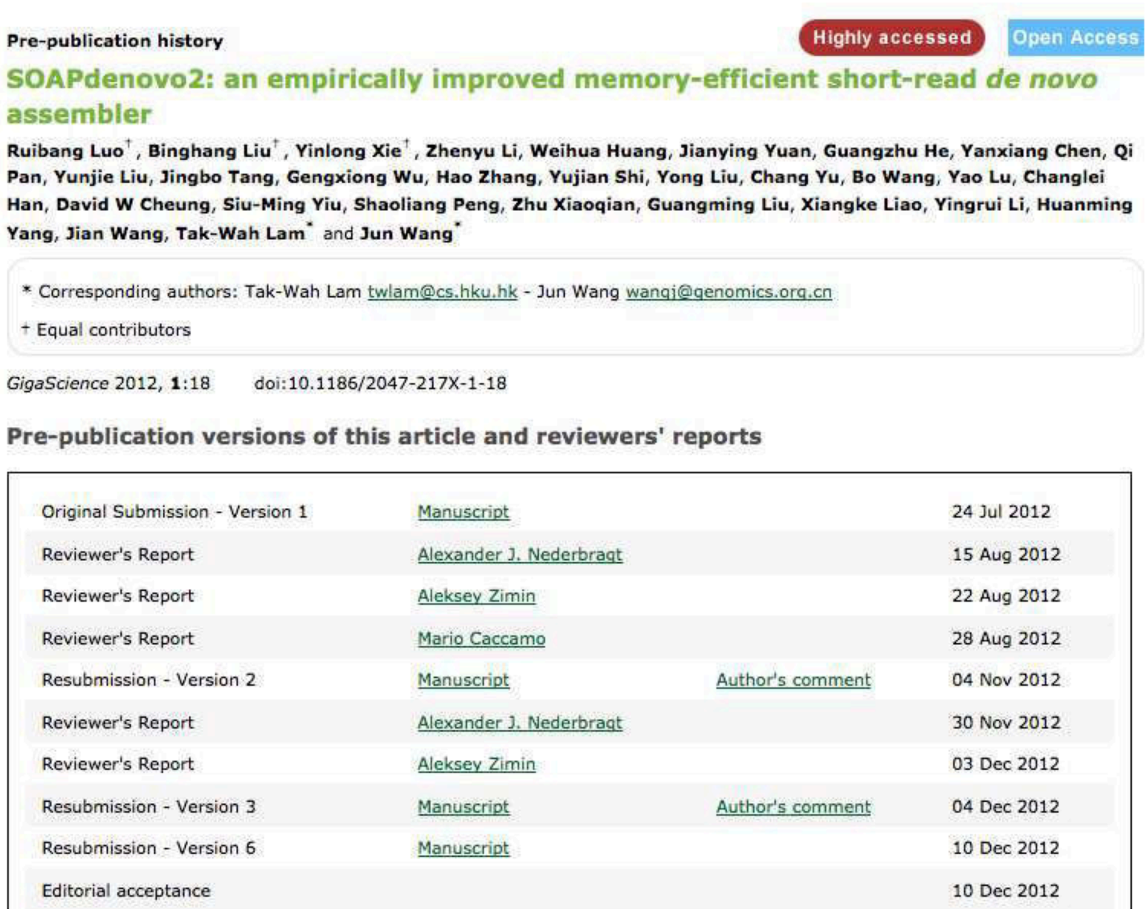
source forge open source

>20,000 downloads

Code in sourceforge under GPLv3: http://soapdenovo2.sourceforge.net/

Reward open & transparent review

7 referees downloaded & tested data, then signed reports



Pre-publication history

Highly accessed Open Access

SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler

Ruibang Luo†, Binghang Liu†, Yinlong Xie†, Zhenyu Li, Weihua Huang, Jianying Yuan, Guangzhu He, Yanxiang Chen, Qi Pan, Yunjie Liu, Jingbo Tang, Gengxiong Wu, Hao Zhang, Yujian Shi, Yong Liu, Chang Yu, Bo Wang, Yao Lu, Changlei Han, David W Cheung, Siu-Ming Yiu, Shaoliang Peng, Zhu Xiaoqian, Guangming Liu, Xiangke Liao, Yingrui Li, Huanming Yang, Jian Wang, Tak-Wah Lam* and **Jun Wang***

* Corresponding authors: Tak-Wah Lam twlam@cs.hku.hk - Jun Wang wangj@genomics.org.cn

† Equal contributors

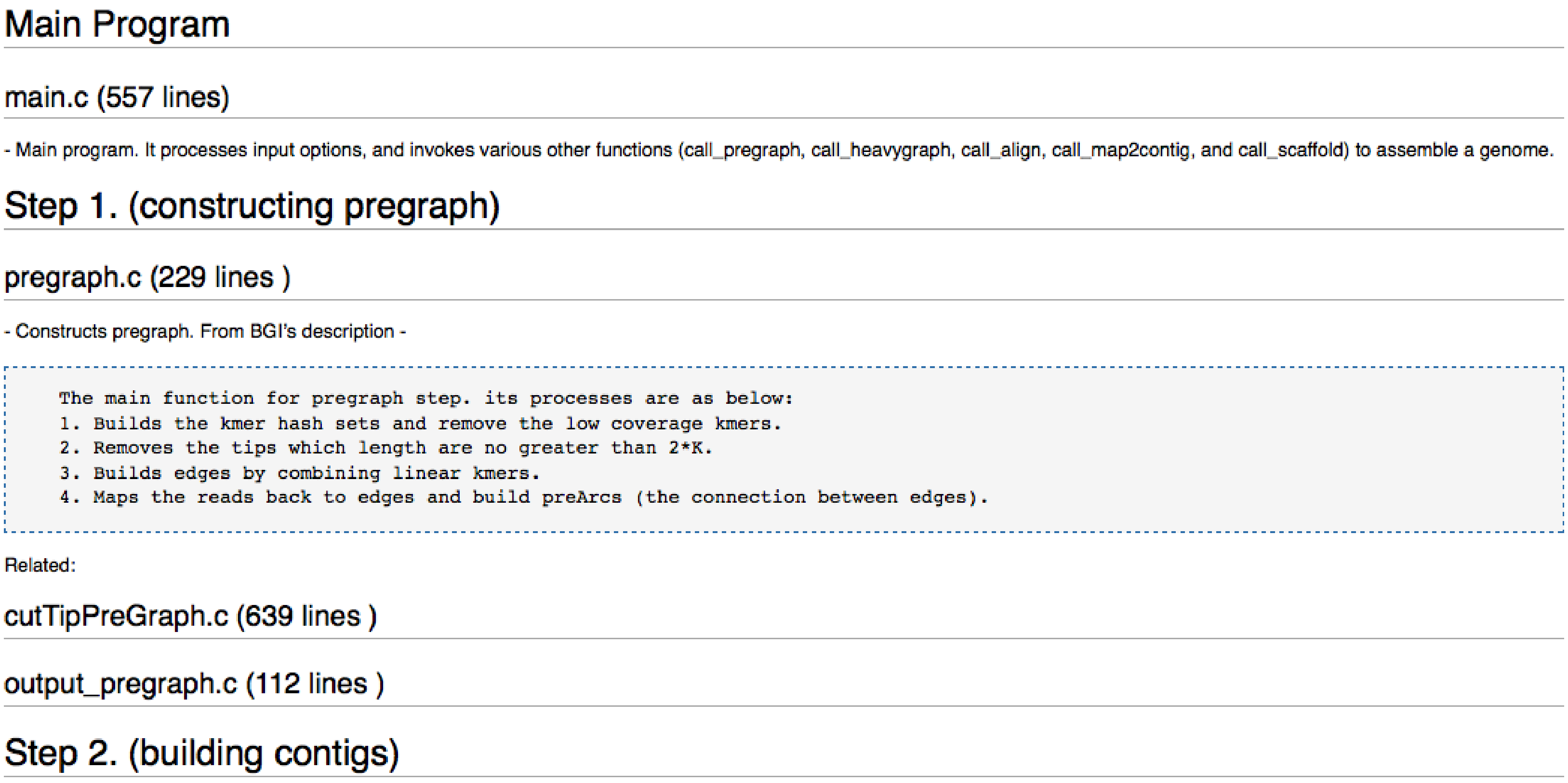
GigaScience 2012, **1**:18 doi:10.1186/2047-217X-1-18

Pre-publication versions of this article and reviewers' reports

Original Submission - Version 1	Manuscript		24 Jul 2012
Reviewer's Report	Alexander J. Nederbragt		15 Aug 2012
Reviewer's Report	Aleksey Zimin		22 Aug 2012
Reviewer's Report	Mario Caccamo		28 Aug 2012
Resubmission - Version 2	Manuscript	Author's comment	04 Nov 2012
Reviewer's Report	Alexander J. Nederbragt		30 Nov 2012
Reviewer's Report	Aleksey Zimin		03 Dec 2012
Resubmission - Version 3	Manuscript	Author's comment	04 Dec 2012
Resubmission - Version 6	Manuscript		10 Dec 2012
Editorial acceptance			10 Dec 2012

Reward open & transparent review

Post publication: bloggers pull apart code/reviews in blogs + wiki:



Main Program

main.c (557 lines)

- Main program. It processes input options, and invokes various other functions (call_pregraph, call_heavygraph, call_align, call_map2contig, and call_scaffold) to assemble a genome.

Step 1. (constructing pregraph)

pregraph.c (229 lines)

- Constructs pregraph. From BGI's description -

```
The main function for pregraph step. its processes are as below:
1. Builds the kmer hash sets and remove the low coverage kmers.
2. Removes the tips which length are no greater than 2*K.
3. Builds edges by combining linear kmers.
4. Maps the reads back to edges and build preArcs (the connection between edges).
```

Related:

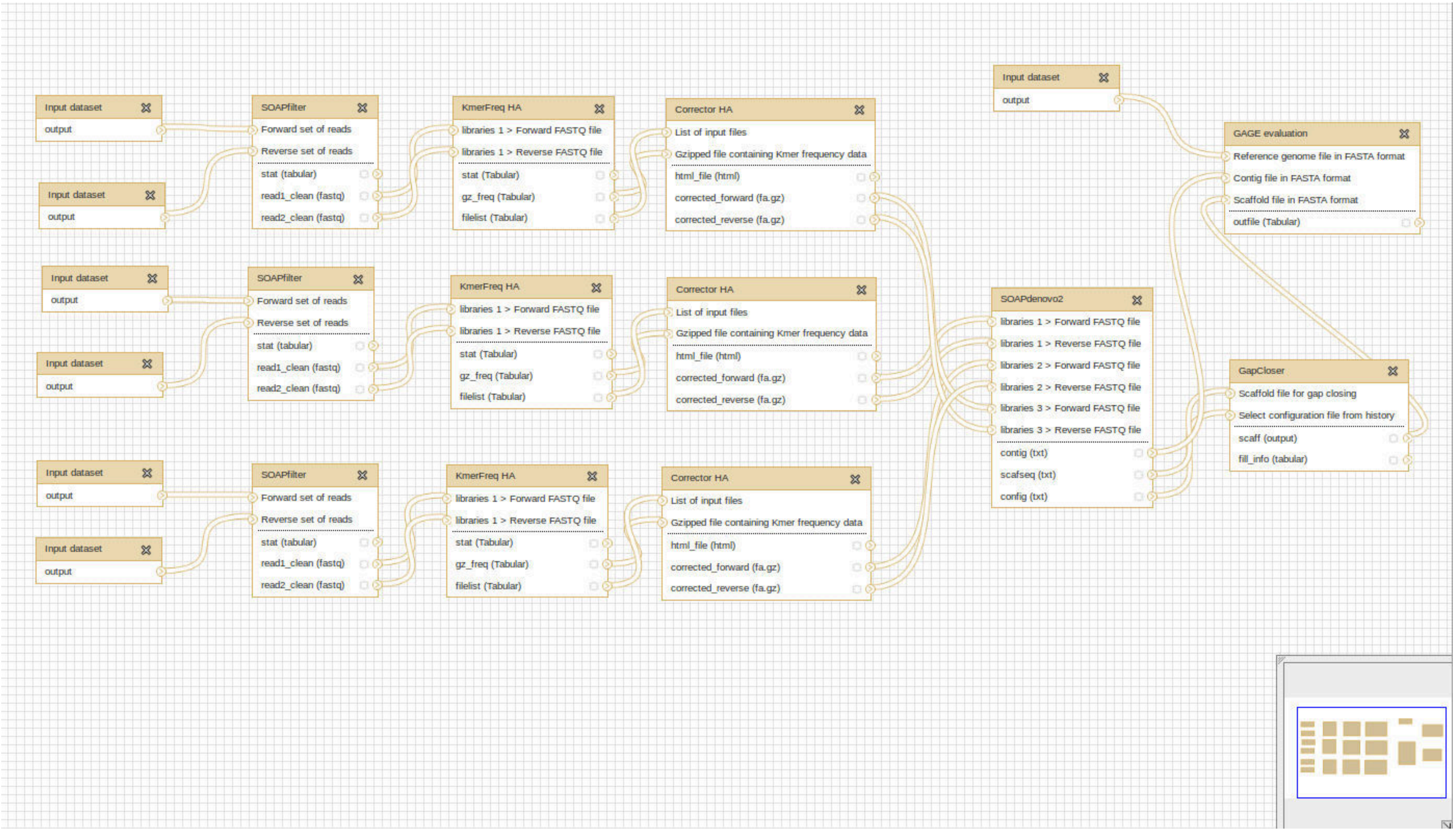
cutTipPreGraph.c (639 lines)

output_pregraph.c (112 lines)

Step 2. (building contigs)

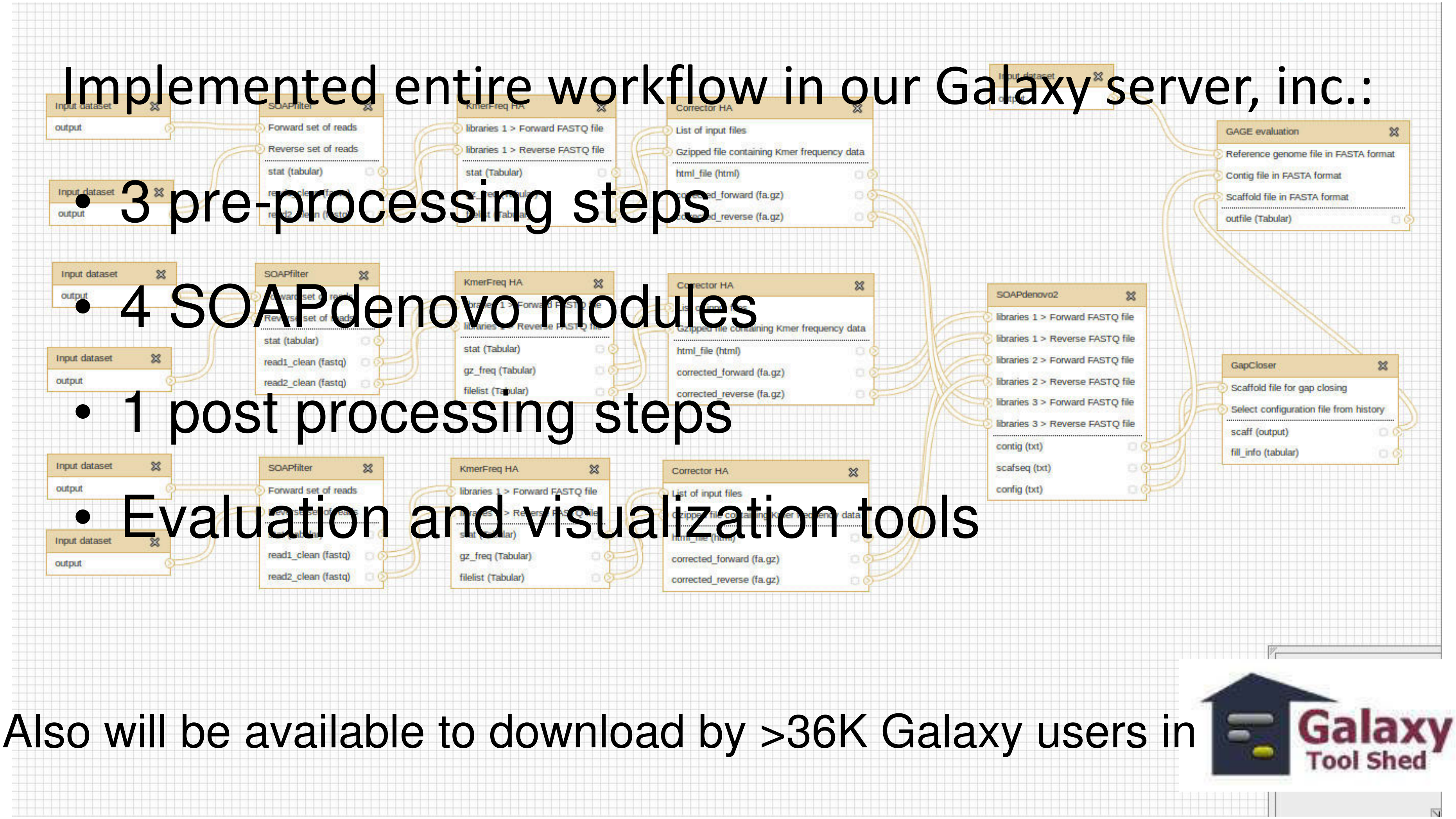
SOAPdenov2 wiki: http://homolog.us/wiki1/index.php?title=SOAPdenovo2
Homologus blogs: http://www.homolog.us/blogs/category/soapdenovo/

SOAPdenovo2 workflows implemented in (GIGA)^{n}Galaxy by CBIIT


Input dataset
output
Input dataset
output
Input dataset
output
Input dataset
output
Input dataset
output
Input dataset
output
SOAPfilter
Forward set of reads
Reverse set of reads
stat (tabular)
read1_clean (fastq)
read2_clean (fastq)
KmerFreq HA
libraries 1 > Forward FASTQ file
libraries 1 > Reverse FASTQ file
stat (Tabular)
gz_freq (Tabular)
filelist (Tabular)
Corrector HA
List of input files
Gzipped file containing Kmer frequency data
html_file (html)
corrected_forward (fa.gz)
corrected_reverse (fa.gz)
SOAPdenovo2
libraries 1 > Forward FASTQ file
libraries 1 > Reverse FASTQ file
libraries 2 > Forward FASTQ file
libraries 2 > Reverse FASTQ file
libraries 3 > Forward FASTQ file
libraries 3 > Reverse FASTQ file
contig (txt)
scafseq (txt)
config (txt)
Input dataset
output
GAGE evaluation
Reference genome file in FASTA format
Contig file in FASTA format
Scaffold file in FASTA format
outfile (Tabular)
GapCloser
Scaffold file for gap closing
Select configuration file from history
scaff (output)
fill_info (tabular)

galaxy.cbiit.cuhk.edu.hk

SOAPdenovo2 workflows implemented in (GIGA)^{n}Galaxy by CBIIT


Input dataset
output
SOAPfilter
Forward set of reads
Reverse set of reads
stat (tabular)
read1_clean (fastq)
read2_clean (fastq)
KmerFreq HA
libraries 1 > Forward FASTQ file
libraries 1 > Reverse FASTQ file
stat (Tabular)
gz_freq (Tabular)
filelist (Tabular)
Corrector HA
List of input files
Gzipped file containing Kmer frequency data
html_file (html)
corrected_forward (fa.gz)
corrected_reverse (fa.gz)
SOAPdenovo2
libraries 2 > Forward FASTQ file
libraries 2 > Reverse FASTQ file
libraries 3 > Forward FASTQ file
libraries 3 > Reverse FASTQ file
contig (txt)
scafseq (txt)
config (txt)
GAGE evaluation
Reference genome file in FASTA format
Contig file in FASTA format
Scaffold file in FASTA format
outfile (Tabular)
GapCloser
Scaffold file for gap closing
Select configuration file from history
scaff (output)
fill_info (tabular)

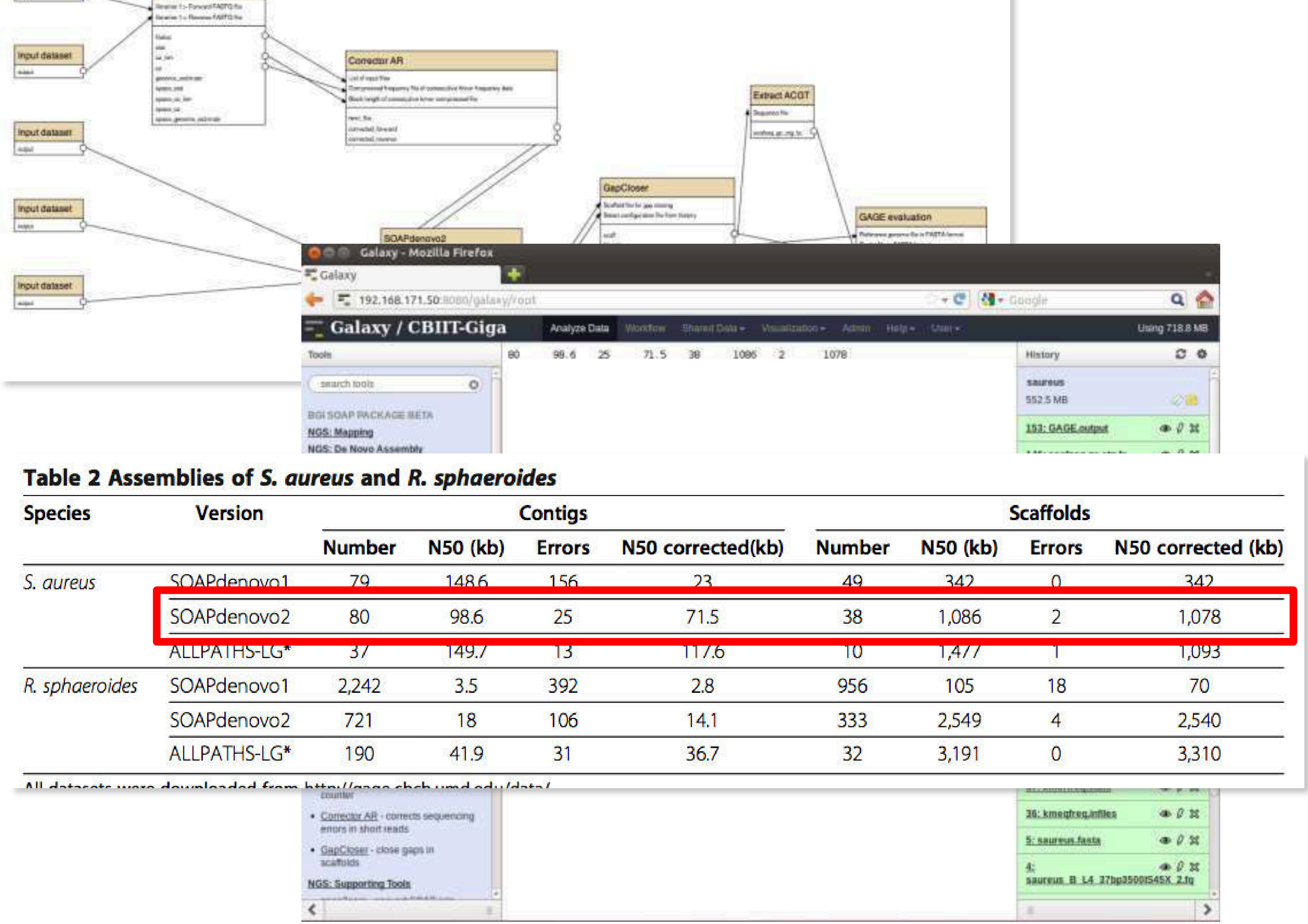
Implemented entire workflow in our Galaxy server, inc.:

- 3 pre-processing steps
- 4 SOAPdenovo modules
- 1 post processing steps
- Evaluation and visualization tools

Also will be available to download by >36K Galaxy users in Galaxy Tool Shed

galaxy.cbiit.cuhk.edu.hk

SOAPdenovo2 *S. aureus* pipeline



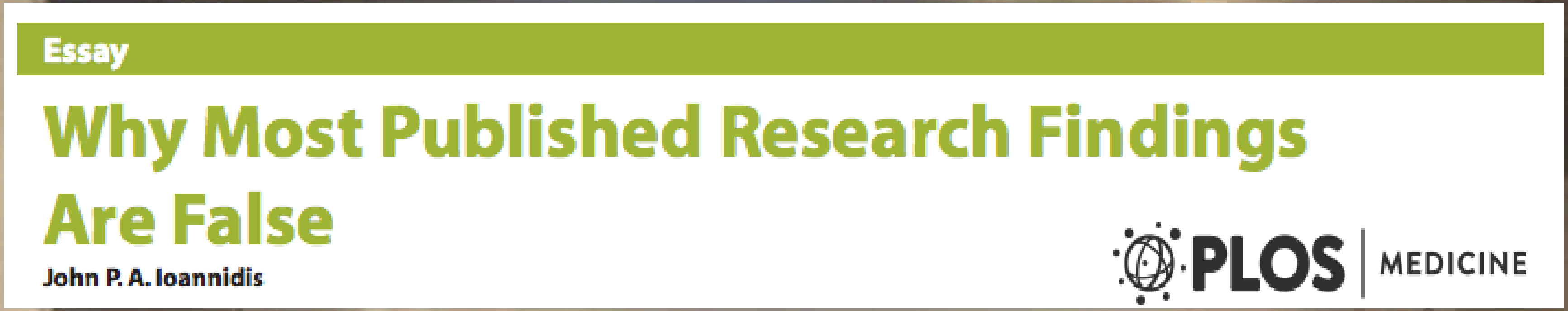
```
80   98.6   25   71.5   38   1086   2   1078
```

Table 2 Assemblies of *S. aureus* and *R. sphaeroides*

Species	Version	Contigs				Scaffolds			
		Number	N50 (kb)	Errors	N50 corrected(kb)	Number	N50 (kb)	Errors	N50 corrected (kb)
S. aureus	SOAPdenovo1	79	148.6	156	23	49	342	0	342
	SOAPdenovo2	80	98.6	25	71.5	38	1,086	2	1,078
	ALLPATHS-LG*	37	149.7	13	117.6	10	1,477	1	1,093
R. sphaeroides	SOAPdenovo1	2,242	3.5	392	2.8	956	105	18	70
	SOAPdenovo2	721	18	106	14.1	333	2,549	4	2,540
	ALLPATHS-LG*	190	41.9	31	36.7	32	3,191	0	3,310

All datasets were downloaded from http://gage.cbcb.umd.edu/data/

Taking a microscope to peer review



Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

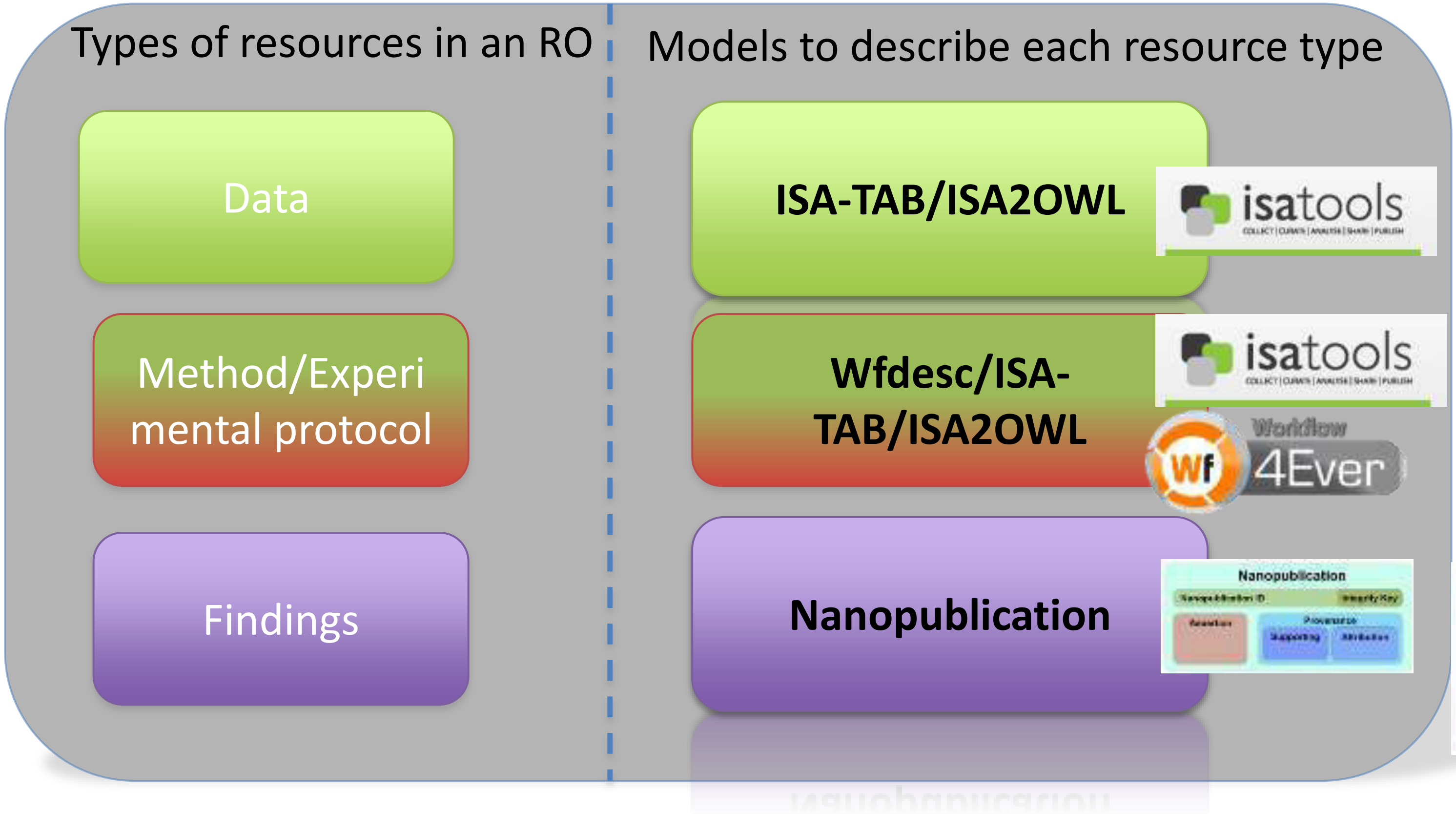
PLOS | MEDICINE

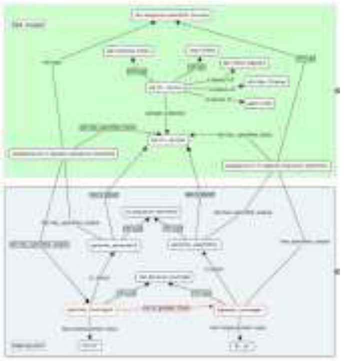
The SOAPdenovo2 Case study

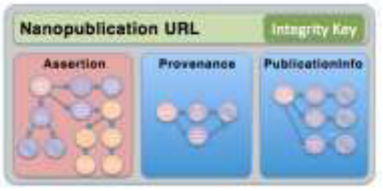
Subject to and test with 3 models:

isacommons researchobject

nanopub.org
People, Tools and Knowhow for Datapublishing



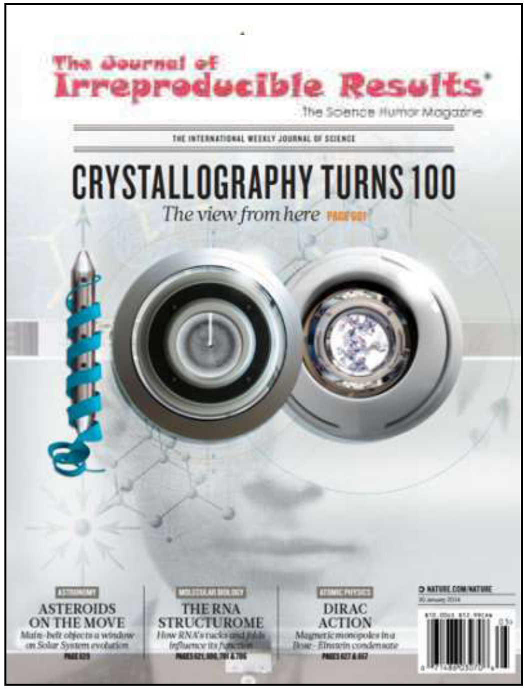




Lessons learned:

- Most published research findings are false. Or at least have errors.
- On a semantic level (via nanopublications) discovered 4 minor errors in text (interpretation not data)
- Is possible to push button(s) & recreate a result from a paper
- Reproducibility is COSTLY. How much are you willing to spend?
- Much easier to do this before rather than after publication

“Regular” Journal



The Journal of Irreproducible Results®

The Science Humor Magazine

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE

CRYSTALLOGRAPHY TURNS 100

The view from here PAGE 601

ASTRONOMY

ASTEROIDS ON THE MOVE

Main-belt objects a window on Solar System evolution

PAGE 629

MOLECULAR BIOLOGY

THE RNA STRUCTUROME

How RNA's tracks and folds influence its function

PAGES 621, 666, 701 & 706

ATOMIC PHYSICS

DIRAC ACTION

Magnetic monopoles in a Bose–Einstein condensate

PAGES 627 & 657

NATURE.COM/NATURE

30 January 2014

“Conscientious” Online Journal

Wilson et al. GigaScience 2012, 1:3
http://www.gigasciencejournal.com/content/1/1/3

GIGA SCIENCE

RESEARCH Open Access

Resources for methylome analysis suitable for gene knockout studies of potential epigenome modifiers

Gareth A Wilson[1*], Pawandeep Dhami[1], Andrew Feber[1], Daniel Cortázar[2], Yuka Suzuki[1], Reiner Schulz[3], Primo Schär[2] and Stephan Beck[1*]

Abstract

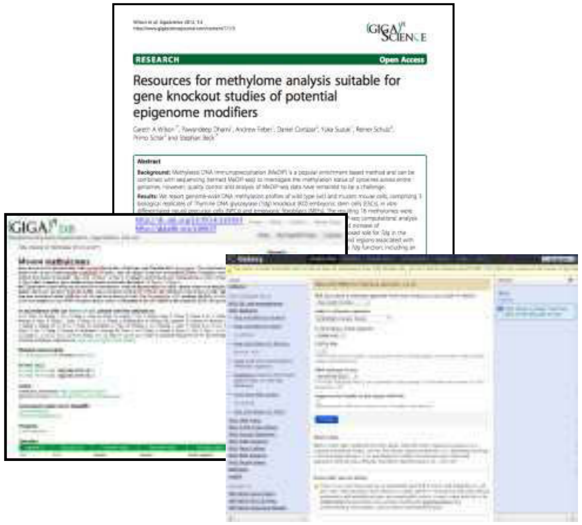
Background: Methylated DNA immunoprecipitation (MeDIP) is a popular enrichment based method and can be combined with sequencing (termed MeDIP-seq) to interrogate the methylation status of cytosines across entire genomes. However, quality control and analysis of MeDIP-seq data have remained to be a challenge.

Results: We report genome-wide DNA methylation profiles of wild type (wt) and mutant mouse cells, comprising 3 biological replicates of Thymine DNA glycosylase (Tdg) knockout (KO) embryonic stem cells (ESCs), in vitro differentiated neural precursor cells (NPCs) and embryonic fibroblasts (MEFs). The resulting 18 methylomes were analysed with MeDUSA (Methylated DNA Utility for Sequence Analysis), a novel MeDIP-seq computational analysis pipeline for the identification of differentially methylated regions (DMRs). The observed increase of hypermethylation in MEF promoter-associated CpG islands supports a previously proposed role for Tdg in the protection of regulatory regions from epigenetic silencing. Further analysis of genes and regions associated with the DMRs by gene ontology, pathway, and ChIP analyses revealed further insights into Tdg function, including an association of TDG with low-methylated distal regulatory regions.

Conclusions: We demonstrate that MeDUSA is able to detect both large-scale changes between cells from different stages of differentiation and also small but significant changes between the methylomes of cells that only differ in the KO of a single gene. These changes were validated utilising publicly available datasets and confirm TDG's function in the protection of regulatory regions from epigenetic silencing.

Keywords: Methylome, MeDIP-seq, Epigenetics, Epigenomics, DNA methylation, Computational pipeline, MeDUSA

“Deconstructed” Journal



Wilson et al. GigaScience 2012, 1:3
http://www.gigasciencejournal.com/content/1/1/3

GIGA SCIENCE

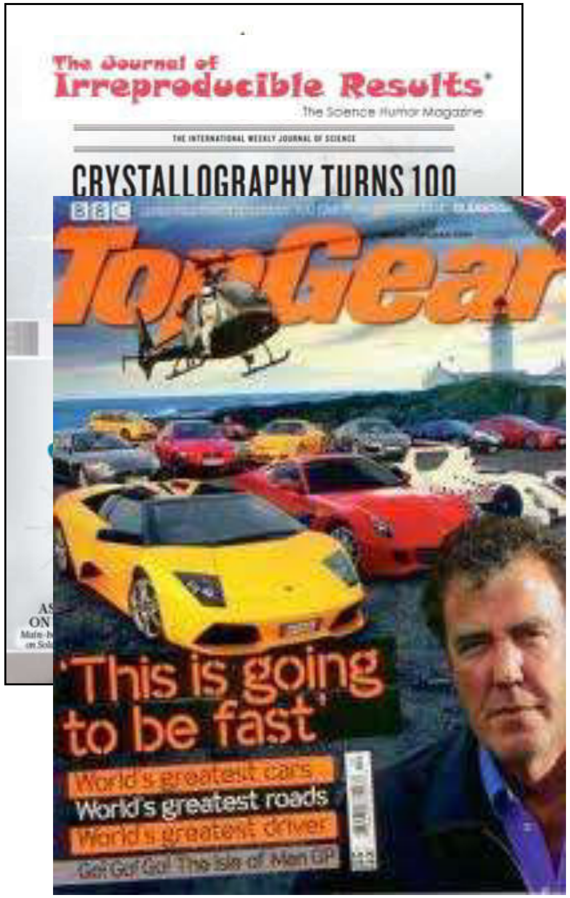
RESEARCH Open Access

Resources for methylome analysis suitable for gene knockout studies of potential epigenome modifiers

Gareth A Wilson[1*], Pawandeep Dhami[1], Andrew Feber[1], Daniel Cortázar[2], Yuka Suzuki[1], Reiner Schulz[3], Primo Schär[2] and Stephan Beck[1*]

Mouse methylomes

“Regular” Journal


The Journal of Irreproducible Results
The Science Humor Magazine
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
CRYSTALLOGRAPHY TURNS 100
BBC
TopGear
'This is going to be fast'
World's greatest cars
World's greatest roads
World's greatest driver
Go! Go! Go! The Isle of Man GP

“Conscientious” Online Journal

Wilson et al. GigaScience 2012, 1:3
http://www.gigasciencejournal.com/content/1/1/3

GIGA SCIENCE

RESEARCH Open Access

Resources for methylome analysis suitable for gene knockout studies of potential epigenome modifiers

Gareth A Wilson1*, Pawandeep Dhami1, Andrew Feber1, Daniel Cortázar2, Yusuke Suzuki2, Reiner Schulz3, Primo Schär2 and Stephan Beck1*

Abstract

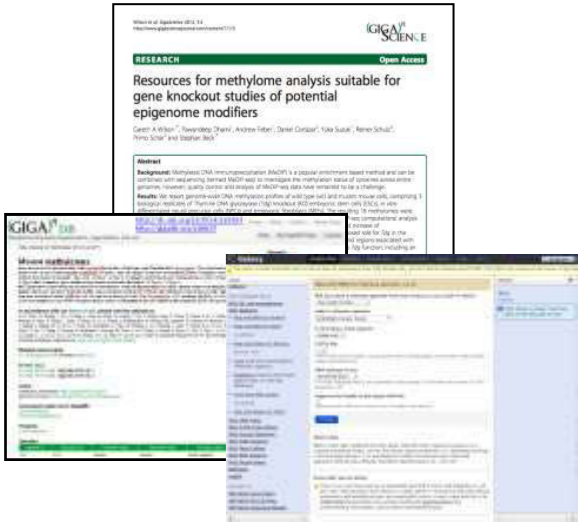
Background: Methylated DNA immunoprecipitation (MeDIP) is a popular enrichment based method and can be combined with sequencing (termed MeDIP-seq) to interrogate the methylation status of cytosines across entire genomes. However, quality control and analysis of MeDIP-seq data have remained to be a challenge.

Results: We report genome-wide DNA methylation profiles of wild type (wt) and mutant mouse cells, comprising 3 biological replicates of Thymine DNA glycosylase (Tdg) knockout (KO) embryonic stem cells (ESCs), in vitro differentiated neural precursor cells (NPCs) and embryonic fibroblasts (MEFs). The resulting 18 methylomes were analysed with MeDUSA (Methylated DNA Utility for Sequence Analysis), a novel MeDIP-seq computational analysis pipeline for the identification of differentially methylated regions (DMRs). The observed increase of hypermethylation in MEF promoter-associated CpG islands supports a previously proposed role for Tdg in the protection of regulatory regions from epigenetic silencing. Further analysis of genes and regions associated with the DMRs by gene ontology, pathway, and ChIP analyses revealed further insights into Tdg function, including an association of TDG with low-methylated distal regulatory regions.

Conclusions: We demonstrate that MeDUSA is able to detect both large-scale changes between cells from different stages of differentiation and also small but significant changes between the methylomes of cells that only differ in the KO of a single gene. These changes were validated utilising publicly available datasets and confirm TDG's function in the protection of regulatory regions from epigenetic silencing.

Keywords: Methylome, MeDIP-seq, Epigenetics, Epigenomics, DNA methylation, Computational pipeline, MeDUSA

“Deconstructed” Journal



GIGA SCIENCE

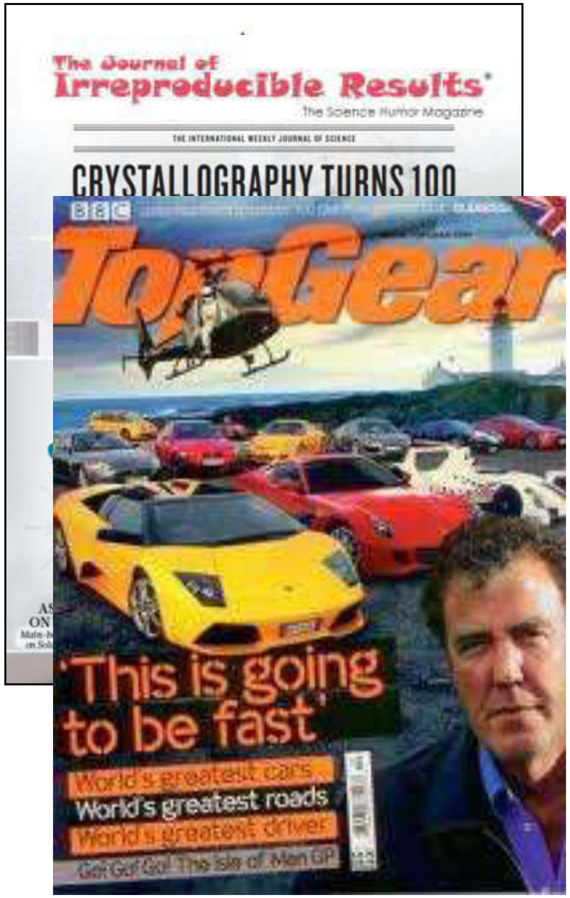
RESEARCH Open Access

Resources for methylome analysis suitable for gene knockout studies of potential epigenome modifiers

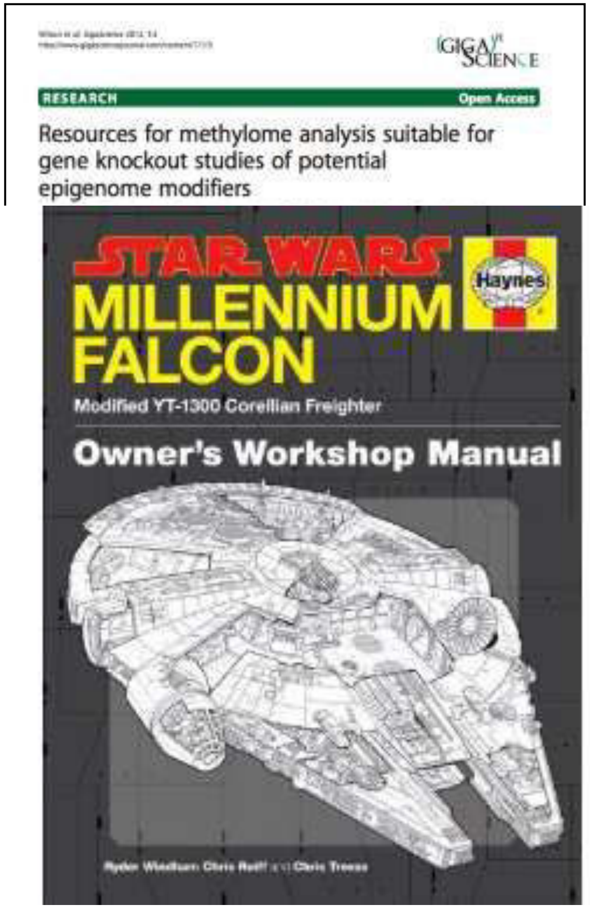
GIGA DB

Mouse methylomes

“Regular” Journal


The Journal of Irreproducible Results
The Science Humor Magazine
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
CRYSTALLOGRAPHY TURNS 100
BBC
TopGear
'This is going to be fast'
World's greatest cars
World's greatest roads
World's greatest driver
Go! Go! Go! The Isle of Man GP

“Conscientious” Online Journal

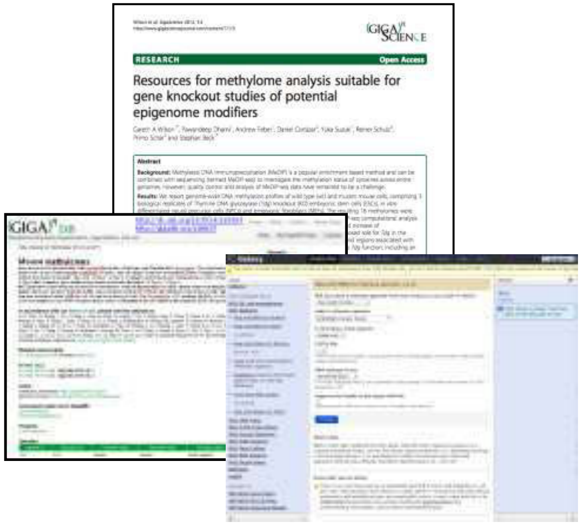

GIGA SCIENCE

RESEARCH Open Access

Resources for methylome analysis suitable for gene knockout studies of potential epigenome modifiers

STAR WARS
MILLENNIUM FALCON
Modified YT-1300 Corellian Freighter
Owner's Workshop Manual
Haynes
Ryder Windham Chris Reiff and Chris Trevas

“Deconstructed” Journal


GIGA SCIENCE

RESEARCH Open Access

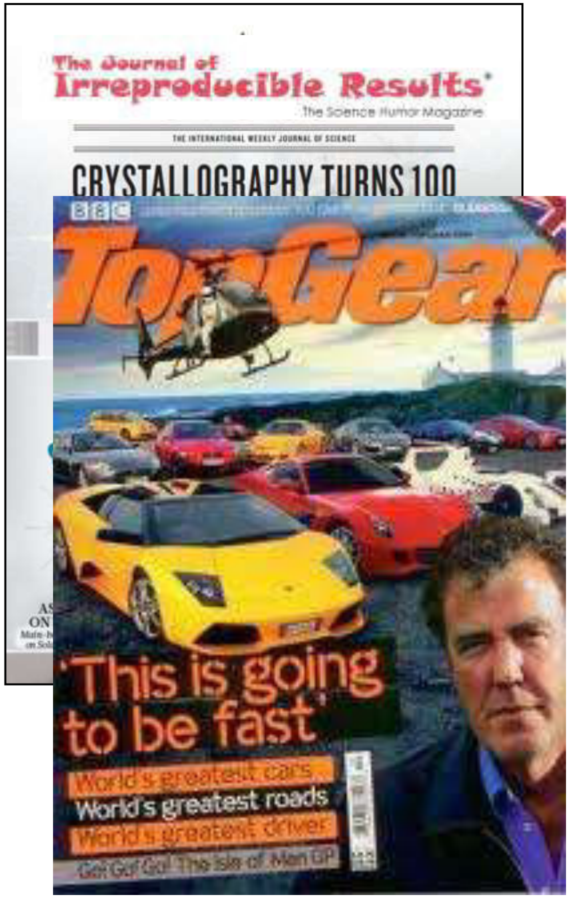
Resources for methylome analysis suitable for gene knockout studies of potential epigenome modifiers

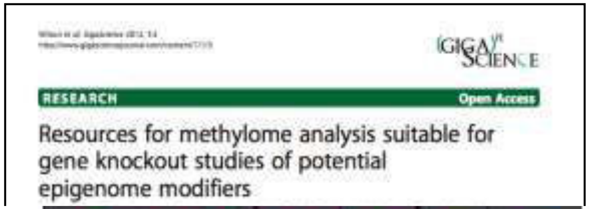
Abstract

“Regular” Journal

“Conscientious” Online Journal

“Deconstructed” Journal


The Journal of Irreproducible Results
The Science Humor Magazine
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
CRYSTALLOGRAPHY TURNS 100
BBC
TopGear
'This is going to be fast'
World's greatest cars
World's greatest roads
World's greatest driver
Go! Go! Go! The Isle of Man GP


GIGA SCIENCE

RESEARCH Open Access

Resources for methylome analysis suitable for gene knockout studies of potential epigenome modifiers

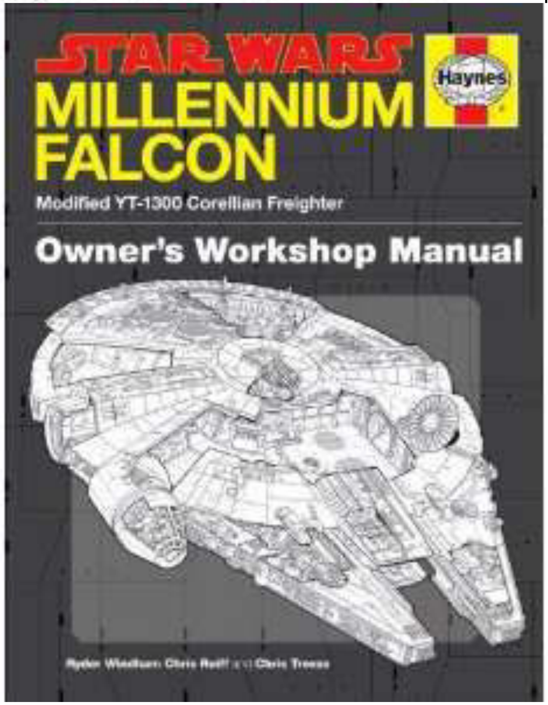

STAR WARS
MILLENNIUM FALCON
Haynes
Modified YT-1300 Corellian Freighter
Owner's Workshop Manual
Ryder Windham Chris Reiff and Chris Trevas

Image Source: http://commons.wikimedia.org/wiki/File:System-Mechanic-California.jpg

Help us make it happen!

Give us data, papers & pipelines*

Contact us:

scott@gigasciencejournal.com
editorial@gigasciencejournal.com
database@gigasciencejournal.com

* APC's currently generously covered by BGI until 2015

华大基因 BGI   www.gigasciencejournal.com   BioMed Central The Open Access Publisher

Thanks to:

(GIGA)^n SCIENCE team:

Peter Li
Huayan Gao
Chris Hunter
Jesse Si Zhe
Nicole Nogoy
Laurie Goodman
Amye Kenall (BMC)

Our collaborators:

Ruibang Luo (BGI/HKU)
Shaoguang Liang (BGI-SZ)
Tin-Lap Lee (CUHK)
Qiong Luo (HKUST)
Senghong Wang (HKUST)
Yan Zhou (HKUST)

Case study:

Marco Roos (LUMC)
Mark Thompson (LUMC)
Jun Zhao (Lancaster)
Susanna Sansone (Oxford)
Philippe Rocca-Serra (Oxford)
Alejandra Gonzalez-Beltran (Oxford)

DataCite isacommons researchobject nanopub.org People, Tools and Knowhow for Datapublishing

Funding from:

华大基因 BGI 国家基因库 CHINA NATIONAL GENEBANK CBIIT NATURAL ENVIRONMENT RESEARCH COUNCIL

@gigascience
facebook.com/GigaScience
blogs.biomedcentral.com/gigablog/

www.gigadb.org
galaxy.cbiit.cuhk.edu.hk
www.gigasciencejournal.com

(GIGA)^n SCIENCE

BioMed Central
The Open Access Publisher