A journal's experiences of reproducing published data analyses

Peter Li

peter@gigasciencejournal.com

(GIGA)n SCIENCE

Journal and database for large-scale data studies

华大基因 BGI

in conjunction with

BioMed Central
The Open Access Publisher

Editor-in-Chief: Laurie Goodman
Executive Editor: Scott Edmunds
Commissioning Editor: Nicole Nogoy
GigaDB: Chris Hunter, Jesse Xiao
GigaGalaxy: Peter Li

OPEN ACCESS

www.gigasciencejournal.com

(GIGA)n SCIENCE

Search GigaScience for

Home | Articles | Authors | Reviewers | About this journal | My GigaScience

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

Big data in the biological and biomedical sciences

Call for papers

BioData Mining

(GIGA)n SCIENCE

Articles

Editor's picks | Latest | Most viewed

Commentary Open Access
The 3,000 rice genomes project: new opportunities and challenges for future rice research
Li JY, Wang J and Zeigler RS
GigaScience 2014, **3**:8 (28 May 2014)

Data Note Open Access
The 3,000 rice genomes project
The 3,000 rice genomes project
GigaScience 2014, **3**:7 (28 May 2014)

Data Note Open Access
A dataset comprising four micro-computed tomography scans of freshly fixed and museum earthworm specimens
Lenihan J, Kvist S, Fernández R, Giribet G and Ziegler A
GigaScience 2014, **3**:6 (16 May 2014)

Research Open Access
Joint assembly and genetic mapping of the Atlantic horseshoe crab genome reveals ancient whole genome duplication
Nossa CW, Havlak P, Yue JX, Lv J, Vincent KY, Brockmann HJ and Putnam NH
GigaScience 2014, **3**:9 (14 May 2014)

No Publication Fees Until 2015

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

A savings of £1250 (based on 2014 prices)

(GIGA)n DB

GigaDB: The GigaScience Database

The Rice 3000 Genomes Project Data.

Genomic data of the green sea turtle (Chelonia mydas).

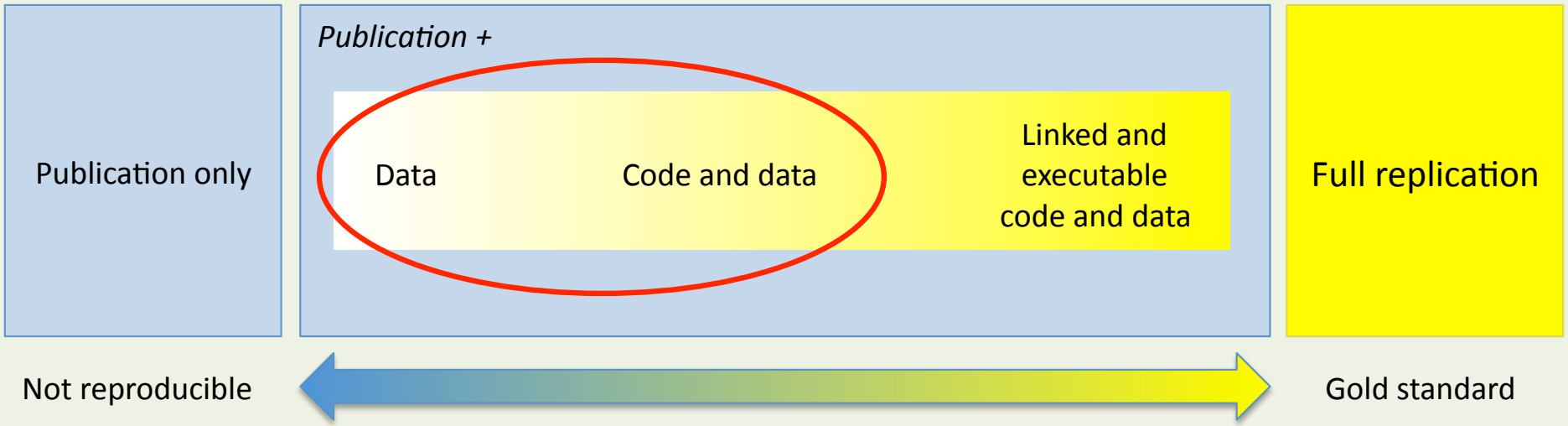
Genomic data of the soft shell turtle (Pelodiscus sinensis).

reproducibility

trust

understanding

Reproducibility spectrum



Adapted from Roger Peng (2011) Reproducible research in computational science. Science 334: 1226-1227.

gigadb.org

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

Home | About | Contact | Terms of use

Help | Login | Create account

GigaDB contains 121 discoverable, trackable, and citable datasets that have been assigned DOIs and are available for public download and use.

Search

Datasets and tools

All types

DOI: 10.5524/100039

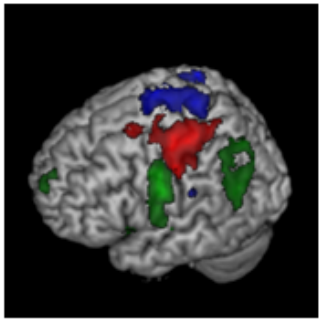
Genomic data of the Puerto Rican Parrot (Amazona vittata) fro...

2012-09-11

DOI: 10.5524/100067

Genomic data from the Insectivorous bat (*Myotis davidii*).

2013-10-31



DOI: 10.5524/100051

A test-retest functional MRI dataset for motor, language and spatial a...

2013-04-08

RSS

New dataset added on 2014-06-18: 10.5524/100049 Metacellulosomics data demonstrating synergism in a soil-derived cellulose-degrading microbial community.

New dataset added on 2014-06-06: 10.5524/100094 Data and software to accompany the paper: Applying compressed sensing to genome-wide association studies.

New dataset added on 2014-05-27: 10.5524/200001 The Rice 3000 Genomes Project Data.

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

(GIGA)n SCIENCE 华大基因 BGI 国家基因库 CHINA NATIONAL GENEBANK

Home | Login | About | Contact | Terms of use

Be a fan on Facebook | Follow us on Twitter | Follow us on Sina | Follow us on Google+ | GigaBlog

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

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.

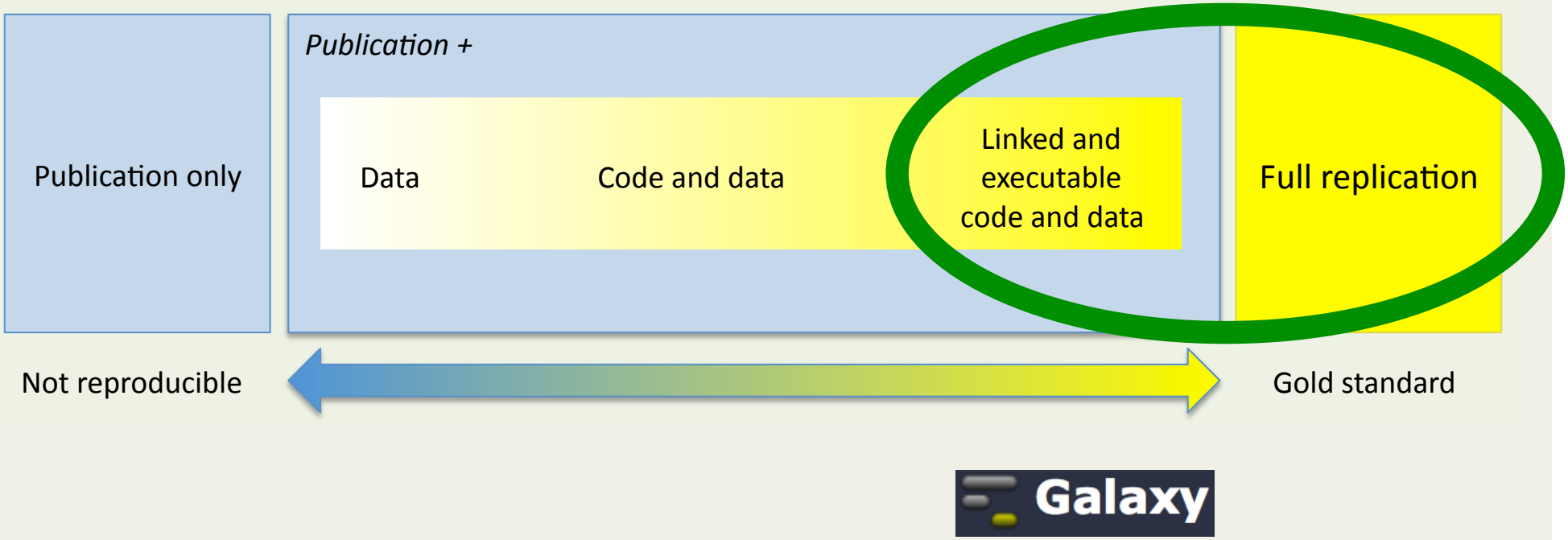
doi:10.1186/2047-217X-1-14
Cite this article as: Oleksyk *et al.*: **A locally funded Puerto Rican parrot (*Amazona vittata*) genome sequencing project increases avian data and advances young researcher education.** *GigaScience* 2012 1:14.

Linking of papers and data by citation of DOIs

Data set DOI

Paper DOI

Reproducibility spectrum



Adapted from Roger Peng (2011) Reproducible research in computational science. Science 334: 1226-1227.

Can the results in a *GigaScience* paper be replicated using Galaxy?

Pilot project

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*]

Replicate

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/.
*ALLPATHS-LG was using the latest version 42807.

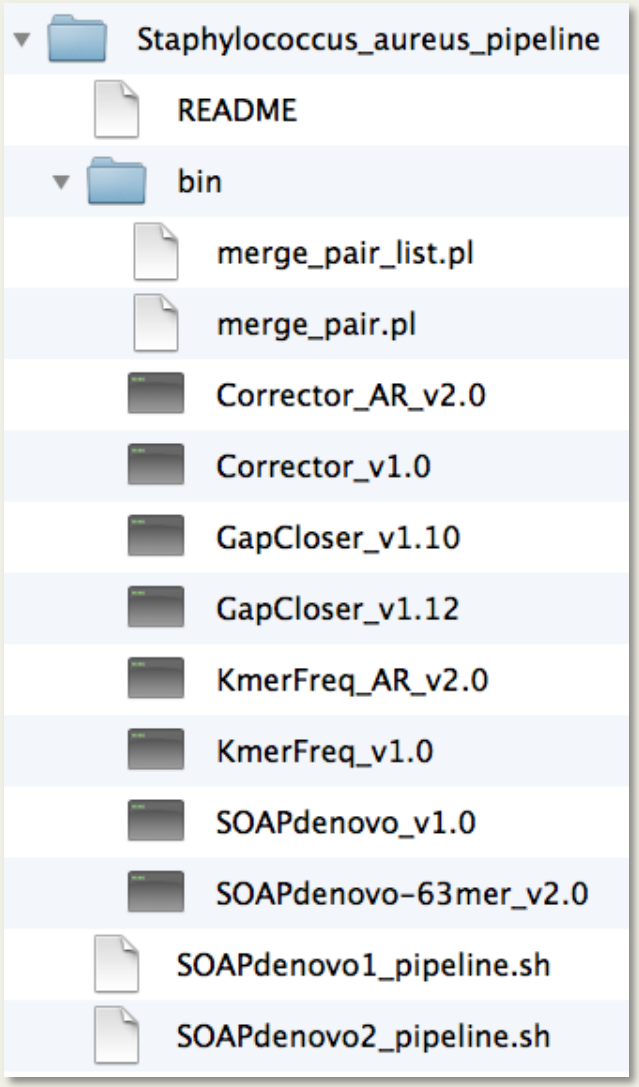
Tools

Files (FTP site) (Aspera): Aspera user name: gigadb , password: gigadb **Table Settings**

File Name	Sample ID	File Type	File Format	Size	Release Date	
README.pdf		Readme	PDF	231.99 KB	2012-12-13	
Assemblathon1_pipeline.tgz		Software	UNKNOWN	10.02 MB	2012-12-13	
Bombus_impatiens_pipeline.tgz		Software	UNKNOWN	4.89 MB	2012-12-13	
Rhodobacter_sphaeroides_pipeline.tgz		Software	UNKNOWN	4.89 MB	2012-12-13	
Staphylococcus_aureus_pipeline.tgz		Software	UNKNOWN	4.33 MB	2012-12-13	
YH_pipeline.tgz	YH	Software	UNKNOWN	7 MB	2012-12-13	
readme.txt		Readme	TEXT	0.29 KB	2012-12-13	

http://gigadb.org/dataset/100044

Tools and data



Genome Assembly Gold-Standard Evaluations

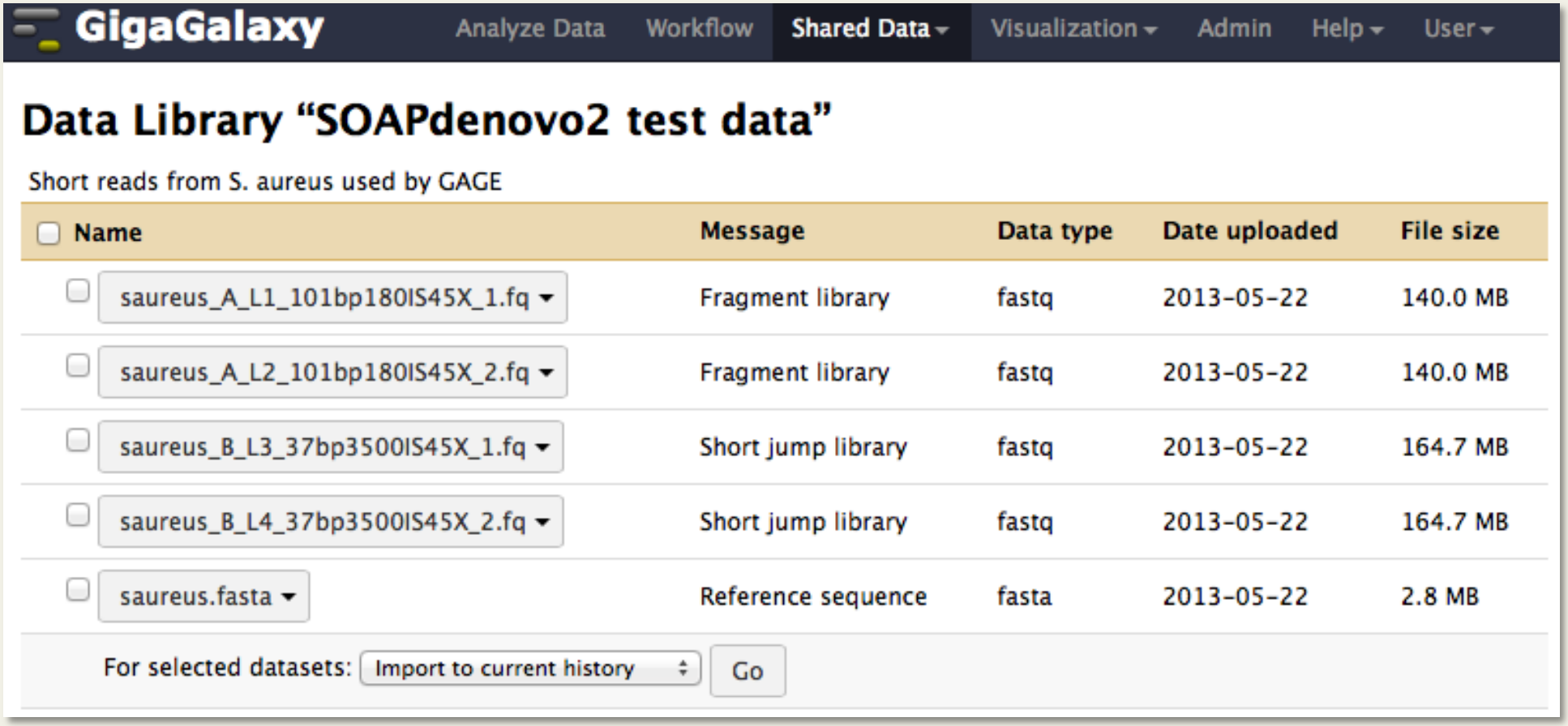
http://gage.cbcb.umd.edu/data/index.html

***Staphylococcus aureus*: Data download page**

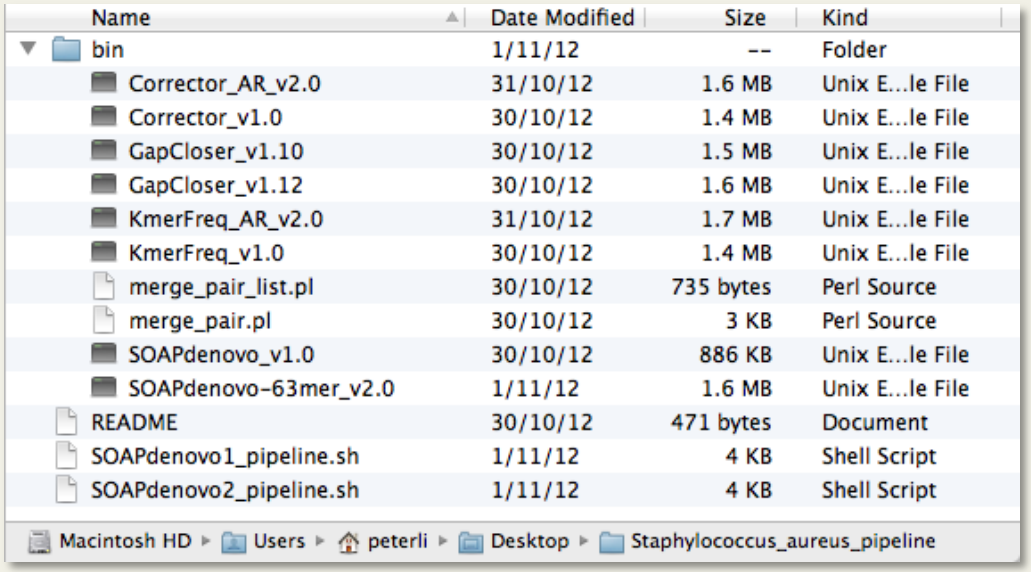
- Library 1: Fragment
  - Avg Read length: 101bp
  - Insert length: 180bp
  - # of reads: 1,294,104
  - Fastq read file 1
  - Fastq read file 2

- Library 2: Short jump library
  - Avg Read length: 37bp
  - Insert length: 3500bp
  - # of reads: 3,494,070
  - Fastq read file 1
  - Fastq read file 2

Data in GigaGalaxy

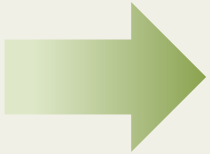


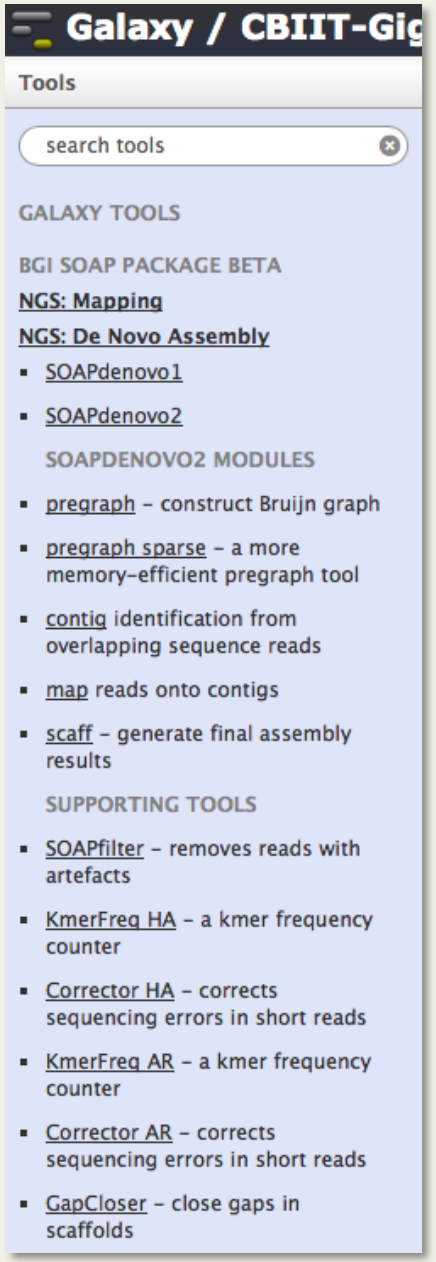
Integration of SOAPdenovo2 into GigaGalaxy

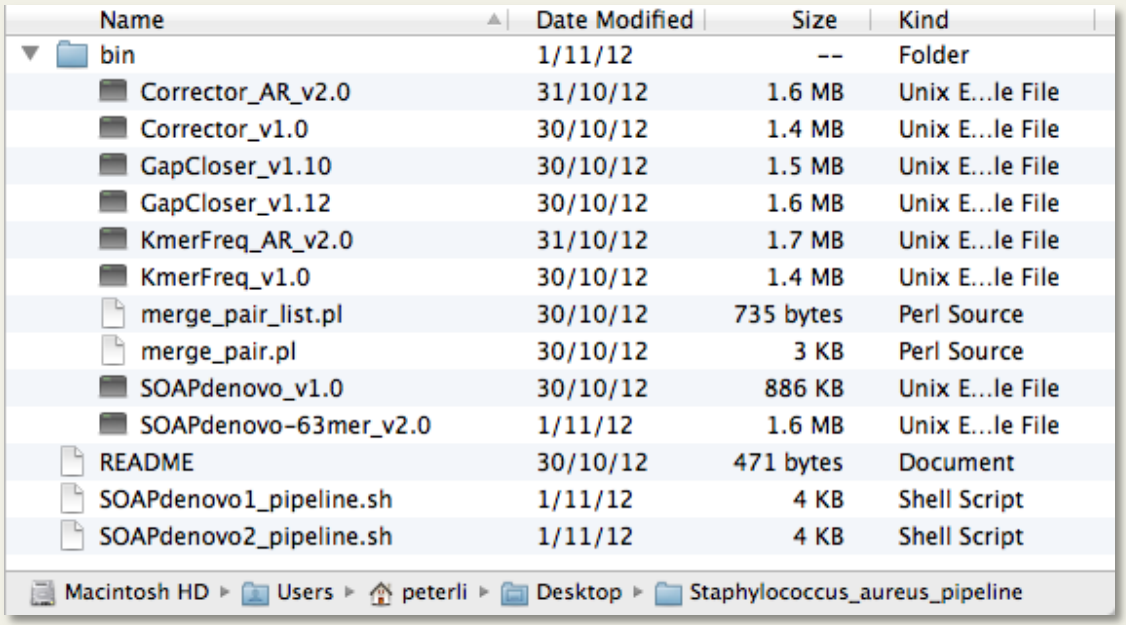


Name	Date Modified	Size	Kind
bin	1/11/12	--	Folder
Corrector_AR_v2.0	31/10/12	1.6 MB	Unix E...le File
Corrector_v1.0	30/10/12	1.4 MB	Unix E...le File
GapCloser_v1.10	30/10/12	1.5 MB	Unix E...le File
GapCloser_v1.12	30/10/12	1.6 MB	Unix E...le File
KmerFreq_AR_v2.0	31/10/12	1.7 MB	Unix E...le File
KmerFreq_v1.0	30/10/12	1.4 MB	Unix E...le File
merge_pair_list.pl	30/10/12	735 bytes	Perl Source
merge_pair.pl	30/10/12	3 KB	Perl Source
SOAPdenovo_v1.0	30/10/12	886 KB	Unix E...le File
SOAPdenovo-63mer_v2.0	1/11/12	1.6 MB	Unix E...le File
README	30/10/12	471 bytes	Document
SOAPdenovo1_pipeline.sh	1/11/12	4 KB	Shell Script
SOAPdenovo2_pipeline.sh	1/11/12	4 KB	Shell Script

Macintosh HD ▸ Users ▸ peterli ▸ Desktop ▸ Staphylococcus_aureus_pipeline







Name	Date Modified	Size	Kind
bin	1/11/12	--	Folder
Corrector_AR_v2.0	31/10/12	1.6 MB	Unix E...le File
Corrector_v1.0	30/10/12	1.4 MB	Unix E...le File
GapCloser_v1.10	30/10/12	1.5 MB	Unix E...le File
GapCloser_v1.12	30/10/12	1.6 MB	Unix E...le File
KmerFreq_AR_v2.0	31/10/12	1.7 MB	Unix E...le File
KmerFreq_v1.0	30/10/12	1.4 MB	Unix E...le File
merge_pair_list.pl	30/10/12	735 bytes	Perl Source
merge_pair.pl	30/10/12	3 KB	Perl Source
SOAPdenovo_v1.0	30/10/12	886 KB	Unix E...le File
SOAPdenovo-63mer_v2.0	1/11/12	1.6 MB	Unix E...le File
README	30/10/12	471 bytes	Document
SOAPdenovo1_pipeline.sh	1/11/12	4 KB	Shell Script
SOAPdenovo2_pipeline.sh	1/11/12	4 KB	Shell Script

Macintosh HD ▸ Users ▸ peterli ▸ Desktop ▸ Staphylococcus_aureus_pipeline

Downloaded pipeline is missing two tools for reproducibility

Required pipeline

Short reads

KmerFreq_AR

Corrector_AR

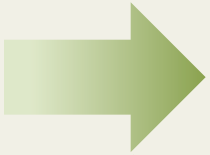
SOAPdenovo2

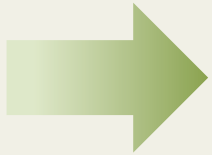
GapCloser

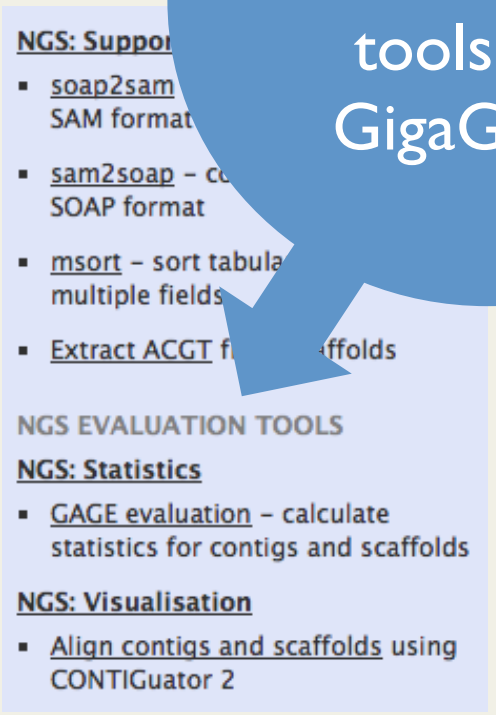
ExtractACGT

GAGE eval

Table 2 N50 & corrected N50 scores







Need to add two extra tools into GigaGalaxy

SOAPdenovo2 *S. aureus* pipeline

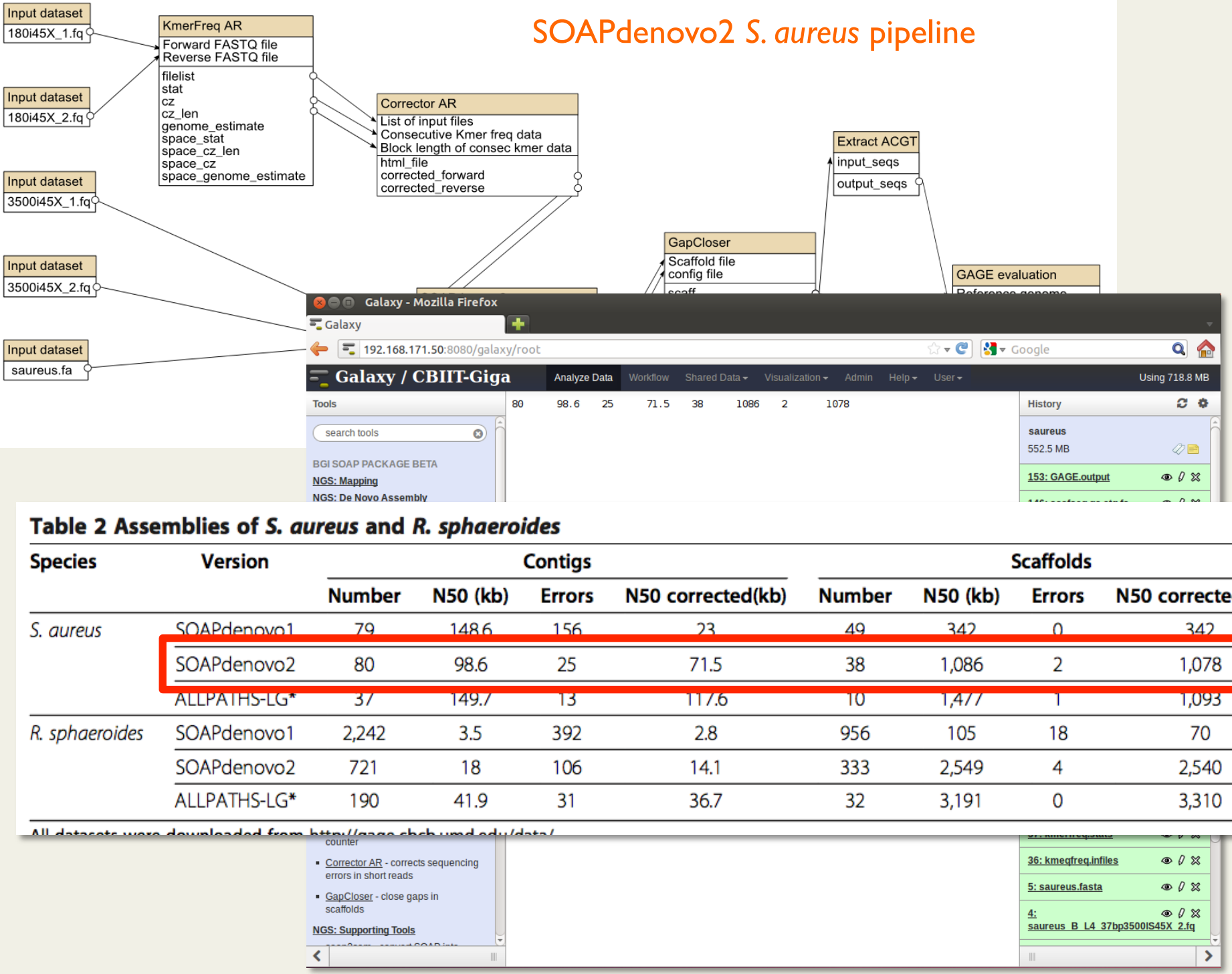


Table 2 Assemblies of *S. aureus* and *R. sphaeroides*

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/

Published and Galaxy-reproduced statistics of genome assemblies of S. aureus and R. sphaeroides

Published

Species	Tool	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	1086	2	1078
	ALL-PATHS-LG	37	149.7	13	117.6	10	1477	1	1093
R. sphaeroides	SOAPdenovo1	2242	3.5	392	2.8	956	105	18	70
	SOAPdenovo2	721	18	106	14.1	333	2549	4	2540
	ALL-PATHS-LG	190	41.9	31	36.7	32	3191	0	3310

Reproduced

Species	Tool	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	1086	2	1078
	ALL-PATHS-LG	37	149.7	13	119.0	11	1477	1	1093
R. sphaeroides	SOAPdenovo1	2241	3.5	400	2.8	956	106	24	68
	SOAPdenovo2	721	18	106	14.1	333	2549	4	2540
	ALL-PATHS-LG	190	41.9	30	36.7	32	3191	0	0

Published Workflows

search name, annotation, owner, and tag

Advanced Search

Name ↓	Annotation	Owner	Community Rating	Community Tags	Last Updated
obedoya_reina_2013: Example 1A: aye-aye-populations		gigascience			Jan 23, 2014
obedoya_reina_2013: Example 1B: aye-aye FST		gigascience			Jan 23, 2014
obedoya_reina_2013: Example 1C: aye-aye diversity		gigascience			Jan 23, 2014
obedoya_reina_2013: Example 2: chicken		gigascience			Jan 23, 2014
obedoya_reina_2013: Example 3: canids		gigascience			Jan 24, 2014
obedoya_reina_2013: Example 4: ABT		gigascience			Jan 23, 2014
rluo_2012: Example 1: soapdenovo2 saureus		gigascience			Mar 26, 2014
rluo_2012: Example 2: soapdenovo2 rsphaeroides		gigascience			Jan 30, 2014
rluo_2012: Example 3: soapdenovo1 saureus		gigascience			Mar 26, 2014
rluo_2012: Example 4: soapdenovo1 rsphaeroides		gigascience			Mar 26, 2014
rluo_2012: Example 5: ALLPATHS-LG saureus		gigascience			Feb 10, 2014
rluo_2012: Example 6: ALLPATHS-LG rsphaeroides		gigascience			Feb 10, 2014

GigaGalaxy | Analyze Data | Workflow | Shared Data | Visualization | Admin | Help | User | Using 105.3 GB

Published Pages | gigascience | SOAPdenovo2 S. aureus

Pipeline: Luo et al., (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* 1:18.

A new version of SOAPdenovo was published in *GigaScience* late in 2012. In this paper, Luo and colleagues described improvements made to their *de novo* genome assembler which was tested on a number of short read data sets including the YH Asian genome. Through our new GigaGalaxy platform, we have made available Galaxy workflows that replicates the data analysis procedures used by Luo et al., (2012) to assemble genomes from short read data sets obtained from *S. aureus* and *R. sphaeroides*. Specifically, these two workflows replicate the metrics shown in Table 2 of the GigaScience paper:

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

Species	Version	Contigs					
		Number	N50 (kb)	Errors	N50 corrected(kb)	Number	N5
S. aureus	SOAPdenovo1	79	148.6	156	23	49	
	SOAPdenovo2	80	98.6	25	71.5	38	1
	ALLPATHS-LG*	37	149.7	13	117.6	10	1
R. sphaeroides	SOAPdenovo1	2,242	3.5	392	2.8	956	
	SOAPdenovo2	721	18	106	14.1	333	2
	ALLPATHS-LG*	190	41.9	31	36.7	32	3

Workflow

The pipeline that reproduces the above *S. aureus* results requires the following steps:

About this Page

Author

gigascience

Related Pages

All published pages
Published pages by gigascience

Rating

Community
(0 ratings, 0.0 average)

Yours

Tags

Community: none

Yours:

http://galaxy.cbiit.cuhk.edu.hk/u/gigascience/p/soapdenovo2-s-aureus

Observations

- Complete scientific reproduction is difficult
  - Time and effort required
- Requires help from authors
- Do we need education and training in scientific reproducibility?


增值機
Add Value Machine
以現金增值 To add value by bank note
1
2
3
機號E2

F1000Research

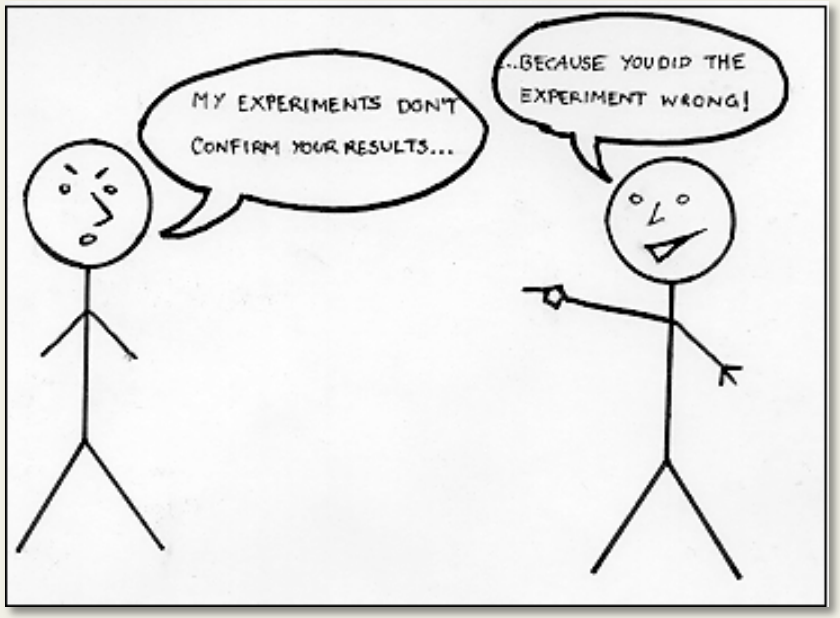
F1000Research 2014, 3:102 Last updated: 27 MAY 2014

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



http://www.cf.ac.uk/socsi/contactsandpeople/harrycollins/image-36548-web.gif

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 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