Data curation as a means to promote reproducibility and discoverability.

Chris Hunter

Presentation DOI:

What is (GIGA)n SCIENCE

- *GigaScience* (http://GigaScienceJournal.com) is a journal
- Collaboration between OUP and BGI
- Main office in Hong Kong, with editorial staff in China, New Zealand, USA and Europe.
- Fully Open Access
- Completely online (no hard copies)
- Accepts manuscripts from all of the Life Sciences, with a scope of Big Data (can be use of, or generation of)
- Focusing on reproducibility and re-usability when assessing manuscripts

OXFORD UNIVERSITY PRESS

BGI 华大 基因科技造福人类

What is (GIGA)^n DB

- GigaDB is our fully curated and maintained data repository to ensure accessibility of all data required for reproducibility of manuscripts
- We provide a home for the myriad of data types that currently do not already have a stable domain specific repository
- GigaDB is managed using the FAIR Principles for scientific data management and stewardship
- GigaDB datasets are also citable, transforming FAIR into FORCE (FAIR, Open, Research-Object based, Citable Ecosystem).

Sansone,S.A., et al. (2018), DOI:10.1101/215183

Reasons why research might not be reproducible (excluding fraud, errors, mistakes)

- Only summary results shown
- Insufficient or unclear methods
- Access to software/tools used (inc. proprietary)
- Lack of availability of input data (e.g. raw sequences)

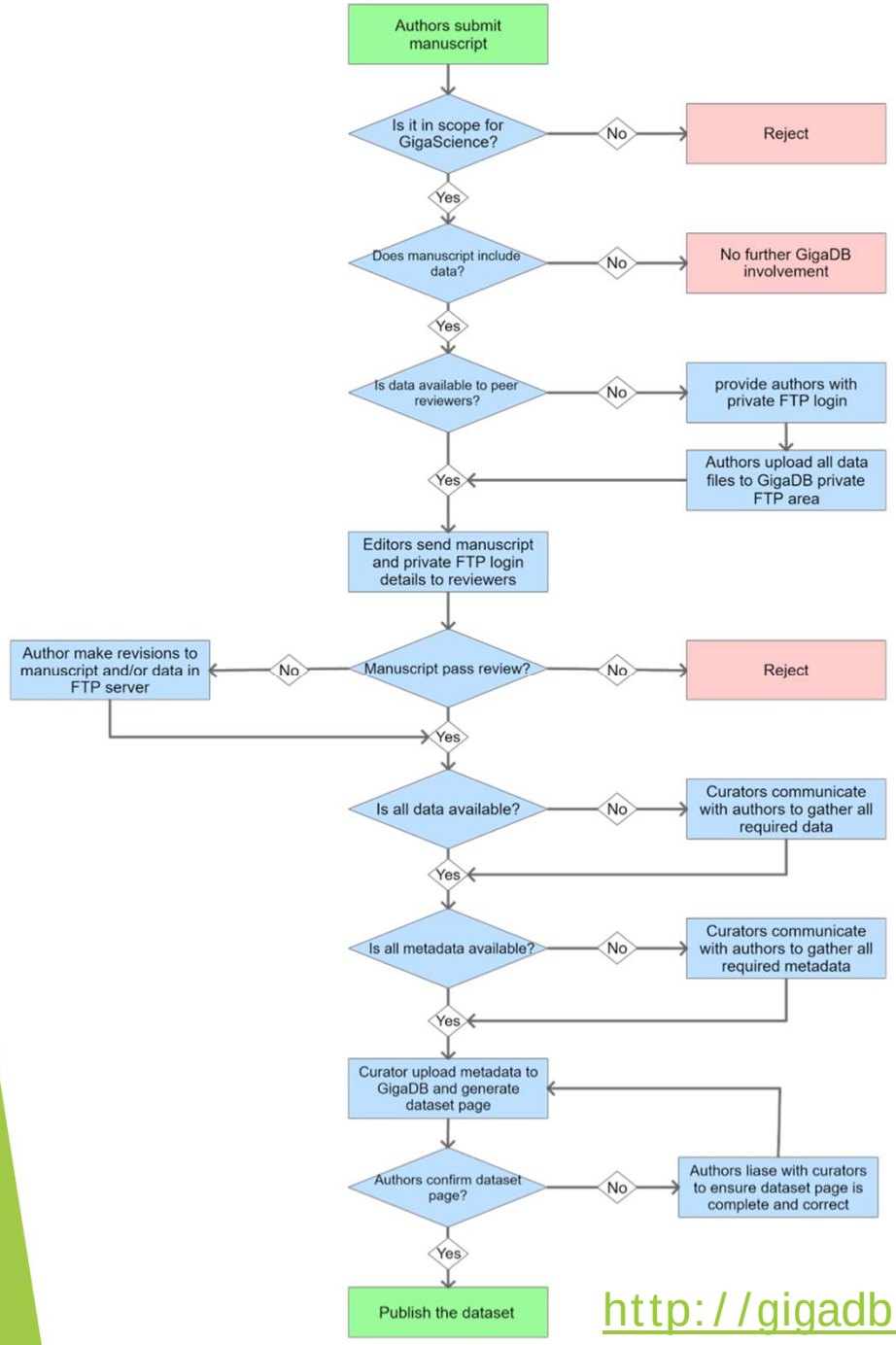
How to combat those issues?

By Data / Transparency review;

- Identify and mandate field specific, stable public repositories
- Provide hosting for things that don't have a stable public repository

(GIGA)^n DB

- Guidelines and checklists of what is expected
- Provide expert guidance to authors



How does the additional work fit into the publishing process?

- Initial data availability check to ensure reviewers have access to everything required
- Provide authors with private FTP drop-box to upload their data
- Curators pull metadata from accepted manuscripts to generate a GigaDB dataset and move data to public FTP area.

http://gigadb.org/site/guide

So what is our transparency review?

- Specialist BioCurators read the methods and results sections to ensure that;
  - All the methods are clear and explicit
  - All the tools/software/scripts are available somewhere and cited appropriately
  - All the input data are available in public stable repositories
  - All the results are available in machine readable formats (i.e. not just PDF)
- To assist authors, reviewers and biocurators we provide guidelines and checklists to help ensure everyone knows whats expected

Checklists and Guidelines - Dataset

- We provide a checklist of expected metadata to accompany each dataset

Item	Imported directly from manuscript (y/n)	Description
Submitting author	y	First Name, Last Name, Email, Institution/Company, ORCID.
Author list	y	First Name, Last Name, ORCID
Dataset title	y	Manuscript title prefixed with "Supporting data for"
Dataset description	y	Manuscript abstract
Funding information	y	Funding body, program, award ID and awardee
Dataset type	n	Selected from controlled vocabulary
Keywords	n	Please list upto 5 keywords, separated by semicolons. All keywords are converted to lowercase.
Additional information links	n	Any URLs to FTP servers or webpages associated with your dataset as semicolon separated lists
Thumbnail image	n	An appropriate image to represent the dataset. Title, Credit, Source and License (CC0 or public domain only) details will be required.
External accessions	n	If any data that you wish to publish in GigaDB has been submitted to to an external resource such as EBI or NCBI, please provide the accession(s) as a semicolon separated list in the format 'SRA:SRPXXXXXX' ; BioProject:PRJNAXXXXXX'
Protocols.io link	n	Where authors provide their methods via protocols.io we can embed these in GigaDB datasets, please provide the published widget URL or DOI

http://gigadb.org/site/guide

Checklists and Guidelines - Samples

- We provide a checklist of expected metadata to accompany each sample
- We do push authors to provide more metadata where possible.
- GSC MIxS standards are expected for sequenced samples

Attribute	Requirement *	Description
Sample name^	R	Use an alphanumeric string to uniquely identify each sample used in your study, you may use BioSample IDs if you have them.
Species tax ID	R	Please enter the NCBI Taxonomy ID for the species used in your study. NB this is mandatory for any sequenced samples.
Species name^	R	Please enter the bionomial (Genus species) name for the species of this sample
Description^	R	Human readable description of sample, it should be unique within a dataset i.e. no two samples are identical so the description should reflect that.
Geographic location (country and/or sea,region)	R	The geographical origin of the sample as defined by the country or sea name followed by specific region name. Country or sea names should be chosen from the INSDC country list
Geographic location (latitude and longitude)	R	The geographical origin of the sample as defined by latitude and longitude. The values should be reported in decimal degrees and on WGS84 system e.g. -69.576435, 91.883948
Broad-scale environmental context	R	Please add one or more ENVO terms to describe the broad environment in which sampling occurred e.g. cliff [ENVO:00000087]
Local environmental context	R	Please add one or more ENVO terms to describe the local environment in which sampling occurred as a semicolon separated list, e.g. digestive tract environment [ENVO:01001033]

http://gigadb.org/site/guide

Checklists and Guidelines - Files

- We provide a checklist of files that we would expect to be submitted along with a dataset.
- This example lists the files that we would anticipate for a genomic and/or transcriptomic dataset.

Item	Suggested format	Check
Genome assembly	fasta	
Coding gene annotations	GFF	
Coding gene nucleotide sequences	fasta	
Coding gene translated sequences	protein fasta	
Repeats/transposable elements/ncRNAs /other annotations	GFF	
Gene family alignments (multi-fasta)	multi-fasta	
Phylogenetic tree files (newick)	newick	
BUSCO output file(s) (text)	text	
SNP annotations (VCF)	VCF	
Any perl/python scripts created for analysis process	py, pl, etc	
readme.txt including all file names with a brief description of each	text	

In addition these might be included for Transcriptomic datasets;

Item	Suggested format	Check
De novo transcriptome assembly	fasta	
Aligned reads	bam	
Expression levels	fpkm table	

http://gigadb.org/site/guidegenome

Checklists and Guidelines – File metadata

- Our file metadata expectations

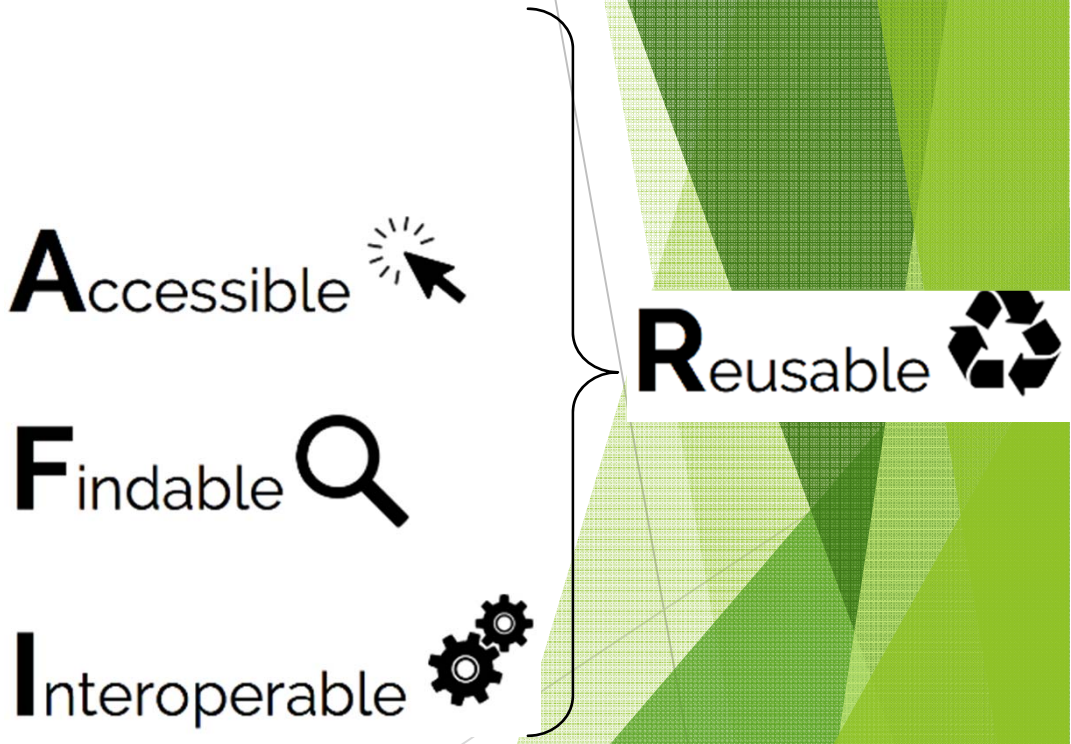
Item	Mandatory (y/n)	Description
File name	y	The exact name of the file including relative file path. Ideally it should be unique within the dataset. Filenames should only include the following characters a-z,A-Z,0-9,_,-,+,. Filenames should not include spaces, we recommend using the underscore (_) in place of spaces.
Description	y	Short human readable description of the file and its contents
Data type	y	The type of data in the file, selected from a controlled vocabulary
Format	y	Most common formats are automatically assigned by file extension, but can be updated manually if required.
MD5 #value	y	These are calculated automatically on our server and added to the database on submitters behalf.
File-Sample association	n	If the sample is derived from a particular sample (in GigaDB) an explicit link can be made between sample(s) and file(s) by adding the Sample ID to the file attributes.
Additional attributes	n	If files have metadata that should be included with them they can be added as attributes, the most common example is Licenses

http://gigadb.org/site/guide

Summary:

Through GigaDB we;

- Reduce fraud (by making it easier to see)
- Spot errors before publication
- More organised data
- Credit for data producers
- Greater visibility of data
- Increased exposure of article
- Allows interactive data/results/research



What the future may hold

- A new submission wizard to assist authors to provide the data required for transparency and reproducibility.
- Integration with GigaOMERO to enable more interactivity with the ever growing number of imaging datasets.
- File Uploader tool – better user interface, resumable uploads, integrated with submission wizard.
- All GigaDB website development is done in GitHub,
  - https://github.com/gigascience/gigadb-website
  - Feel free to suggest new features there, or even contribute to the code!

https://github.com/gigascience/gigadb-website

Acknowledgments

- Thank you for your attention
- BGI – main funders
- All *GigaScience* authors; past, present and future!
- Hypothes.is , protocols.io, SketchFab, GitHub & all the other opensource collaborators that work with us towards better science and sharing.

BGI 华大
基因科技造福人类

Why not join our mailing list by registering for a user account on GigaDB.org
We send out a quarterly newsletter to our mailing list with news of exciting datasets released, new developments in *Giga*DB and information on upcoming conferences that we will be attending.

The Giga Team

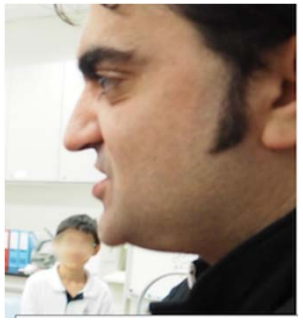
Laurie Goodman

Chris Armit

Mary Ann Tuli

Jesse Xiao

Hans Zauner



Scott Edmunds

Nicole Nogoy

Hongling Zhou

Peter Li

Qi Chen

Join our mailing list by registering for a user account on GigaDB.org

We send out a quarterly newsletter to our mailing list with news of exciting datasets released, new developments in *Giga*DB and information on upcoming conferences that we will be attending.

Anatomy of a GigaDB dataset

Using example datasets:

Supporting data for "Chromosome-scale genome assembly of kiwifruit Actinidia eriantha with single-molecule sequencing and chromatin interaction mapping"

Tang *et al*. (2019): http://dx.doi.org/10.5524/100568

And

Supporting data for "A micro X-ray computed tomography dataset of South African hermit crabs (Crustacea: Decapoda: Anomura: Paguroidea)"

Landschoff *et al.* (2018): http://dx.doi.org/10.5524/100364

Supporting data for "A micro X-ray computed tomography dataset of South African hermit crabs (Crustacea: Decapoda: Anomura: Paguroidea)"

Dataset type: Imaging

Data released on March 06, 2018

Landschoff J; Du Plessis A; Griffiths CL (2018): Supporting data for "A micro X-ray computed tomography dataset of South African hermit crabs (Crustacea: Decapoda: Anomura: Paguroidea)" GigaScience Database. http://dx.doi.org/10.5524/100364

DOI 10.5524/100364

Along with the conventional deposition of physical types at natural history museums, the deposition of three-dimensional (3D) image data has been proposed for rare and valuable museum specimens, such as irreplaceable type material. Micro computed tomography (µCT) scan data of five hermit crab species from South Africa, two of rare specimens, three of holotypic specimens, and two of selected paratypes, depicted main identification characters of calcified body parts. However, low image contrasts, especially in larger (>50 mm total length) specimens did not allow sufficient 3D reconstructions of weakly calcified or fine characters, such as soft tissue of the pleon, mouthparts, gills, or of the setation. Reconstructions of soft tissue were sometimes possible, depending on individual sample and scanning characteristics. The raw data of seven scans are publicly available for download from the GigaDB repository.

Keywords:
microct µct nanoct 3d cybertype e-type diogenidae paguridae parapaguridae taxonomy deep sea species

Contact Submitter

Additional details

Read the peer-reviewed publication(s):

Top half of an Imaging dataset

Bottom half

Additional details

Read the peer-reviewed publication(s):

LANDSCHOFF, J., & RAHAYU, D. L. (2018). A new species of the hermit crab genus Diogenes (Crustacea: Decapoda: Diogenidae) from the coast of KwaZulu-Natal, South Africa. Zootaxa, 4379(2), 268. doi:10.11646/zootaxa.4379.2.7

Additional information:

https://www.thingiverse.com/thing:3242383

https://www.thingiverse.com/thing:3242300

https://www.thingiverse.com/thing:3241997

https://www.thingiverse.com/thing:3241871

https://www.thingiverse.com/thing:3242503

https://www.thingiverse.com/thing:3242746

https://www.thingiverse.com/thing:3242926

Sample | Files | Funding | 3D Viewer | History

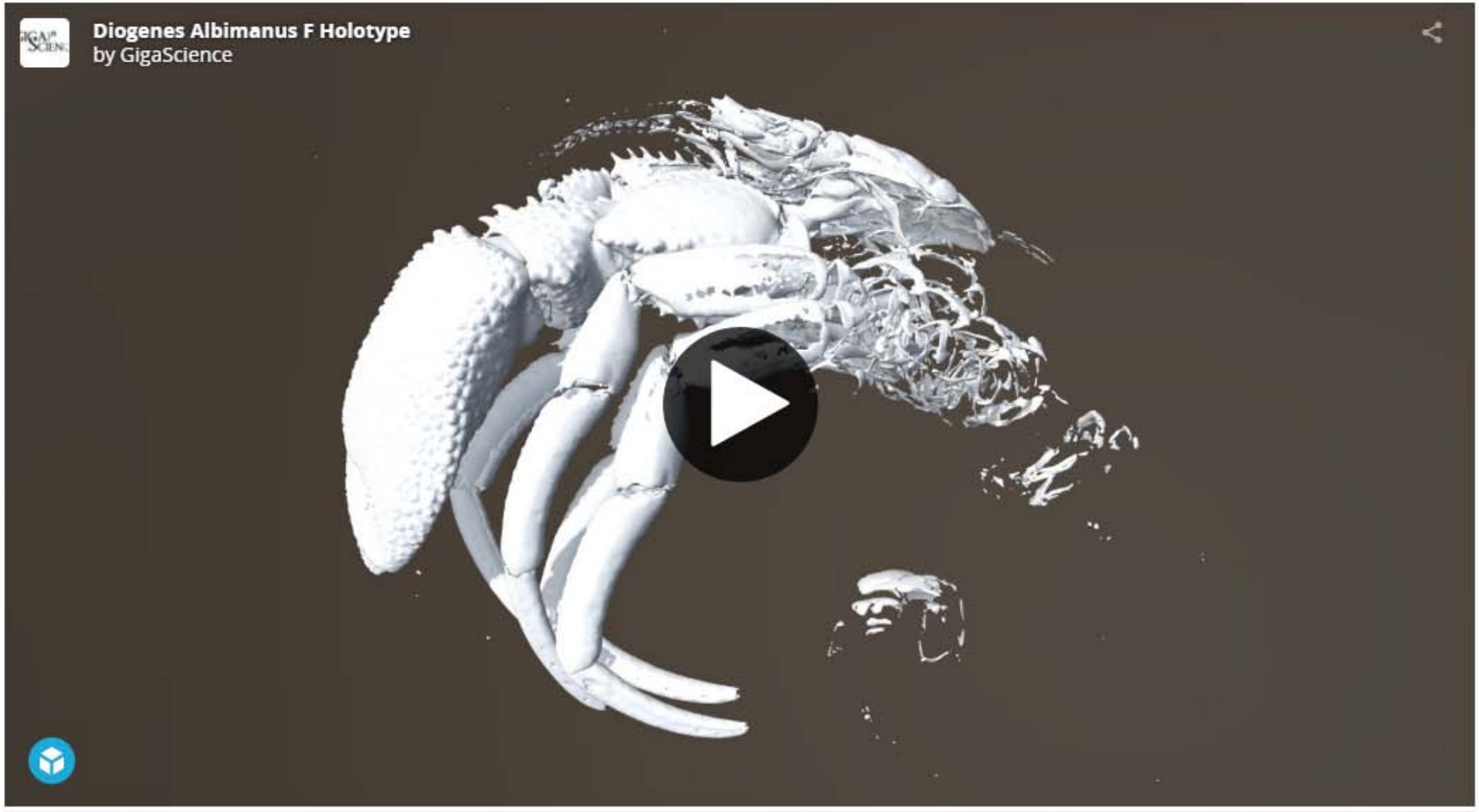
Table Settings

Sample ID	Taxonomic ID	Common Name	Genbank Name	Scientific Name	Sample Attributes
Cancellus_macrothrix	-1	None assigned	None assigned	None assigned	Description:diogenid hermit crab, specimen with bo... Specimen voucher:SAMC MB-A066204 Species-a:Cancellus macrothrix ... +
Diogenes_albimanus_f_holotype	1127697			Diogenes albimanus	Description:diogenid hermit crab, ovigerous female... Specimen voucher:SAMC MB-A066353 Sex:female ... +
Goreopagurus_poorei_m	1377936			Goreopagurus poorei	Description:pagurid hermit crab Specimen voucher:USNM 1292090 Sex:male ... +
Pagurus_species_f_paratype	6746	hermit crabs		Pagurus	Description:pagurid hermit crab, paratype Specimen voucher:SAMC MB-A066770 Sex:female ... +
Pagurus_species_m_holotype	6746	hermit crabs		Pagurus	Description:pagurid hermit crab, holotype

3D viewer tab

Sample | Files | Funding | 3D Viewer | History

3D Models:



Supporting data for "Chromosome-scale genome assembly of kiwifruit Actinidia eriantha with single-molecule sequencing and chromatin interaction mapping"

Dataset type: Genomic, Transcriptomic

Data released on February 19, 2019

Tang W; Sun X; Yue J; Tang X; Jiao C; Yang Y; Niu X; Miao M; Zhang D; Huang S; Shi W; Li M; Fang C; Fei Z; Liu Y (2019): Supporting data for "Chromosome-scale genome assembly of kiwifruit Actinidia eriantha with single-molecule sequencing and chromatin interaction mapping" GigaScience Database. http://dx.doi.org/10.5524/100568

DOI 10.5524/100568

Kiwifruit (Actinidia spp.) is a dioecious plant with fruits containing abundant vitamin C and minerals. A handful of kiwifruit species have been domesticated, among which the A. eriantha is increasingly favored in breeding due to its superior commercial traits. Recently, elite cultivars from A. eriantha have been successfully selected and further studies on their biology and breeding potential require genomic information which is currently unavailable. Here, we provide a high quality A. eriantha genome as well as its gene annotation. Availability of these data will facilitate the breeding program in the future.

Keywords:
kiwifruit actinidia eriantha genome assembly single molecular sequencing hi-c

Contact Submitter

Additional details

Contact Submitter

Additional details

Read the peer-reviewed publication(s):

Tang, W., Sun, X., Yue, J., Tang, X., Jiao, C., Yang, Y., ... Liu, Y. (2019). Chromosome-scale genome assembly of kiwifruit Actinidia eriantha with single-molecule sequencing and chromatin interaction mapping. GigaScience, 8(4). doi:10.1093/gigascience/giz027

Additional information:

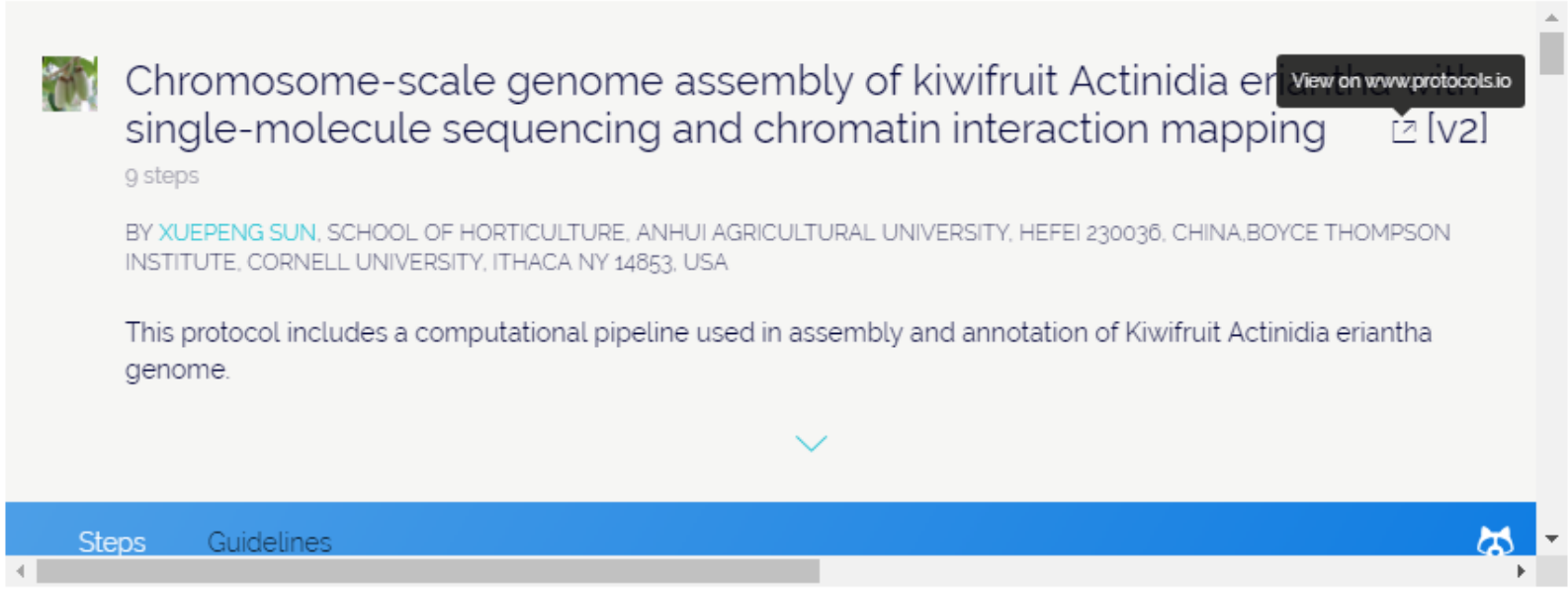
http://bdg.hfut.edu.cn/kir/

Accessions (data generated as part of this study):

BioProject: PRJNA480681

Sample | Files | Funding | Protocols.io | History

Protocols.io:



Chris' GigaDB Page | Admin | LogOut | Browse Samples

e.g. Chicken, brain etc...

(GIGA)^{n}DB

Revolutionizing data dissemination, organization, and use

Home About Help Terms of use

Supporting data for "Draft genome assembly of the invasive cane toad, *Rhinella marina*"

Dataset type: Genomic

Data released on July 23, 2018

Edwards RJ; Enosi Tuipulotu D; Amos TG; O'Meally D; Richardson MF; Russell TL; Vallinoto M; Carneiro M; Ferrand N; Wilkins MR; Sequeira F; Rollins LA; Holmes EC; Shine R; White PA (2018): Supporting data for "Draft genome assembly of the invasive cane toad, *Rhinella marina*" GigaScience Database. http://dx.doi.org/10.5524/100483

DOI 10.5524/100483

The cane toad (*Rhinella marina* formerly *Bufo marinus*) is a species native to Central and South America that has spread across many regions of the globe. Cane toads are known for their rapid adaptation and deleterious impacts on native fauna in invaded regions. However, despite an iconic status, there are major gaps in our understanding of cane toad genetics. The availability of a genome would help to close these gaps and accelerate cane toad research.

We report a draft genome assembly for *R. marina*, the first of its kind for the Bufonidae family. We used a combination of long read PacBio RS II and short read Illumina HiSeq X sequencing to generate a total of 359.5 Gb of raw sequence data. The final hybrid assembly of 31,392 scaffolds was 2.55 Gb in length with a scaffold N50 of 168 kb. BUSCO analysis revealed that the assembly included full length or partial fragments of 90.6% of tetrapod universal single-copy orthologs (n=3950), illustrating that the gene-containing regions have been well-assembled. Annotation predicted 25,846 protein coding genes with similarity to known proteins in SwissProt. Repeat sequences were estimated to account for 63.9% of the assembly.

The *R. marina* draft genome assembly will be an invaluable resource that can be used to further probe the biology of this invasive species. Future analysis of the genome will provide insights into cane toad evolution and enrich our understanding of their interplay with the ecosystem at large.

Keywords:
cane toad rhinella marina genome sequencing hybrid assembly annotation

Contact Submitter

Additional details

Read the peer-reviewed publication(s):

Top half of a Genomic dataset

Contact Submitter

Additional details

Read the peer-reviewed publication(s):

Edwards, R. J., Tuipulotu, D. E., Amos, T. G., O'Meally, D., Richardson, M. F., Russell, T. L., ... White, P. A. (2018). Draft genome assembly of the invasive cane toad, Rhinella marina. GigaScience, 7(9). doi:10.1093/gigascience/giy095

Related datasets:

doi:10.5524/100483 HasPart doi:10.5524/100374

Additional information:

http://www.slimsuite.unsw.edu.au/servers/apollo.php

Accessions (data generated as part of this study):

BioProject: PRJEB24695

Sample | Files | Funding | History

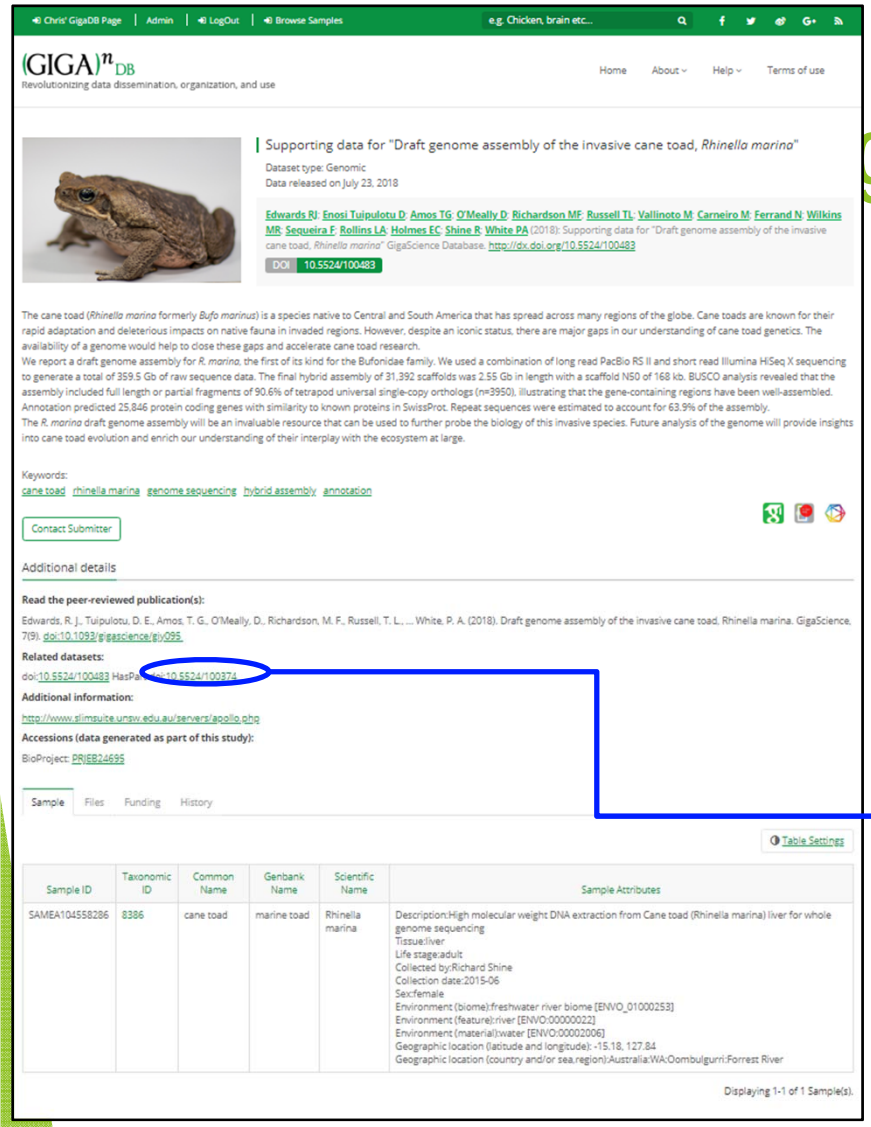
Table Settings

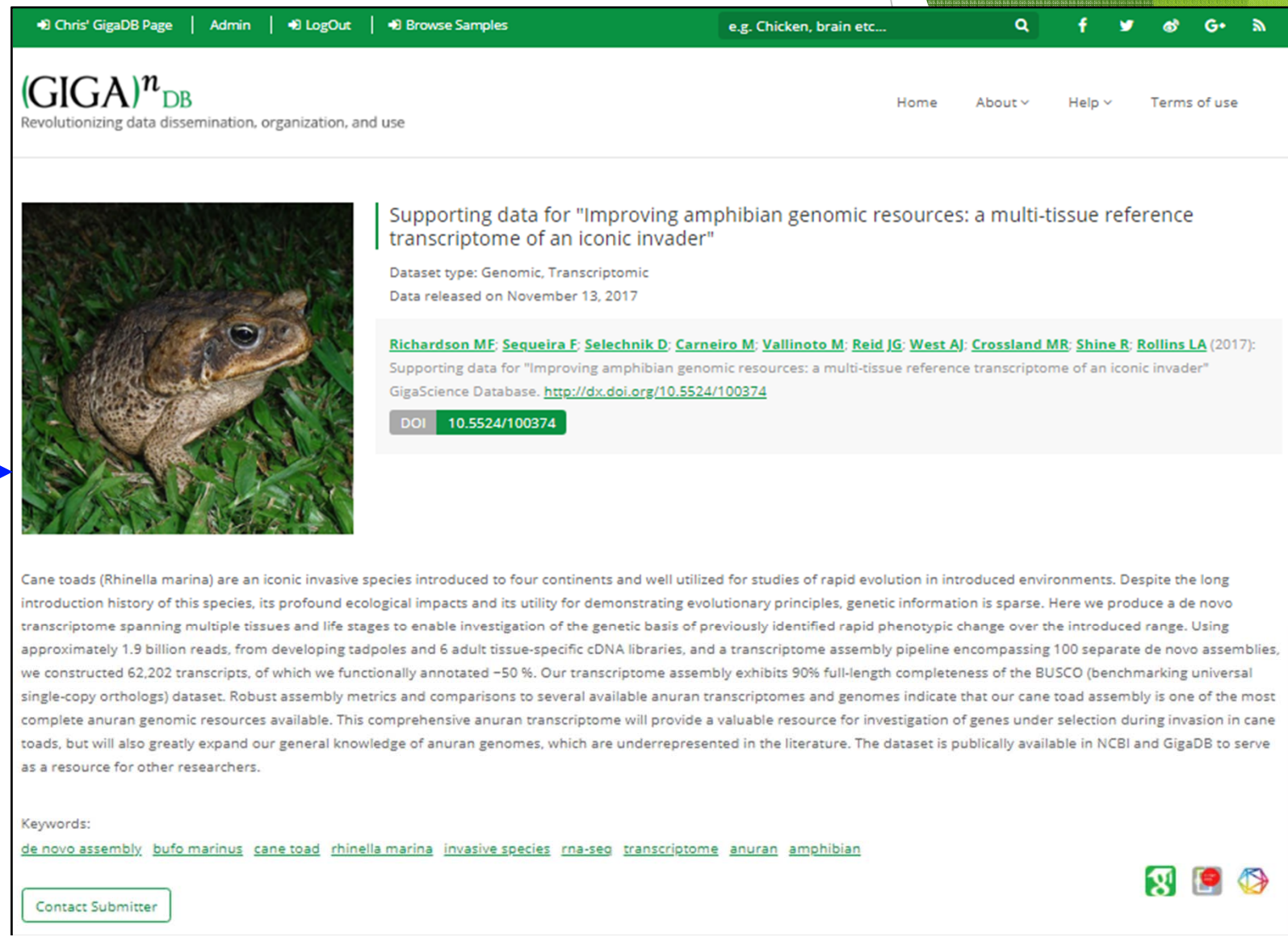
Sample ID	Taxonomic ID	Common Name	Genbank Name	Scientific Name	Sample Attributes
SAMEA104558286	8386	cane toad	marine toad	Rhinella marina	Description:High molecular weight DNA extraction from Cane toad (Rhinella marina) liver for whole genome sequencing Tissue:liver Life stage:adult Collected by:Richard Shine Collection date:2015-06 Sex:female Environment (biome):freshwater river biome [ENVO_01000253] Environment (feature):river [ENVO:00000022] Environment (material):water [ENVO:00002006] Geographic location (latitude and longitude): -15.18, 127.84 Geographic location (country and/or sea,region):Australia:WA:Oombulgurri:Forrest River

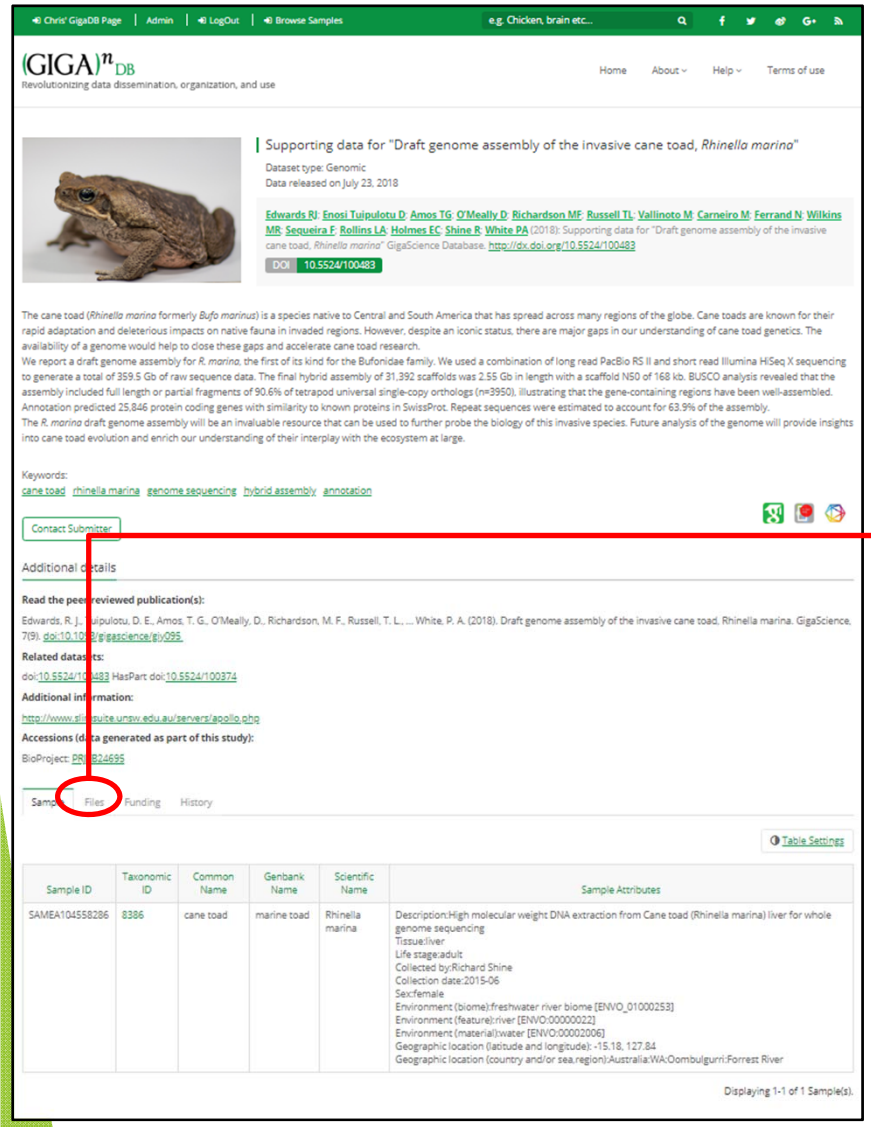
Displaying 1-1 of 1 Sample(s).

Bottom half

gaDB dataset







Sample Files Funding History

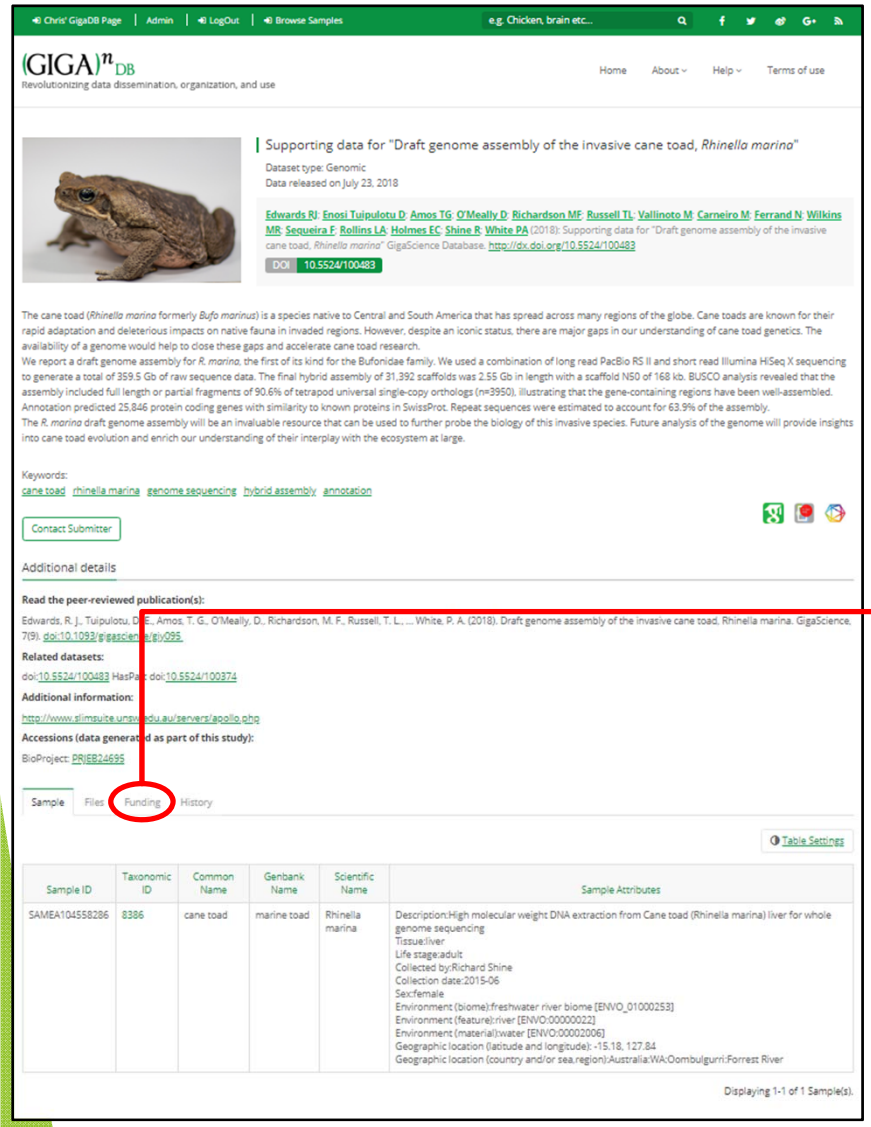
(FTP site)

Table Settings

File Name	Description	Size	Release Date	
canetoad.v2.2.busco.stdout	STDOUT for BUSCO v2.5 short run on genome (text)	7.71 KB	2018-07-16	
canetoad.v2.2.busco.tgz	Gzipped tarball of BUSCO v2.5 short run (directory tarball)	124.52 MB	2018-07-16	
canetoad.v2.2.fasta.gz	Gzipped genome assembly fasta (fasta)	662.54 MB	2018-07-16	
canetoad.v2.2.highquality.aln.tgz	Gzipped directory containing MAFFT alignments of predicted orthologues for high quality predicted proteins (fasta)	40.62 MB	2018-07-16	
canetoad.v2.2.highquality.faa.gz	Gzipped high quality subset of predicted proteins (protein fasta)	2.57 MB	2018-07-16	
canetoad.v2.2.highquality.fna.gz	Gzipped high quality subset of predicted transcripts (fasta)	4.97 MB	2018-07-16	
canetoad.v2.2.highquality.iqtree.tgz	Gzipped directory containing IQ-TREE maximum likelihood trees for high quality predicted proteins (text and Newick)	64.69 MB	2018-07-16	
canetoad.v2.2.maker.gff3.gz	Gzipped MAKER gene annotations (GFF3)	12.65 MB	2018-07-16	
canetoad.v2.2.md5sum	MD5 checksum values for above file	0.69 KB	2018-07-16	
canetoad.v2.2.proteins.faa.gz	Gzipped predicted coding gene translated sequences (protein fasta)	12.88 MB	2018-07-16	

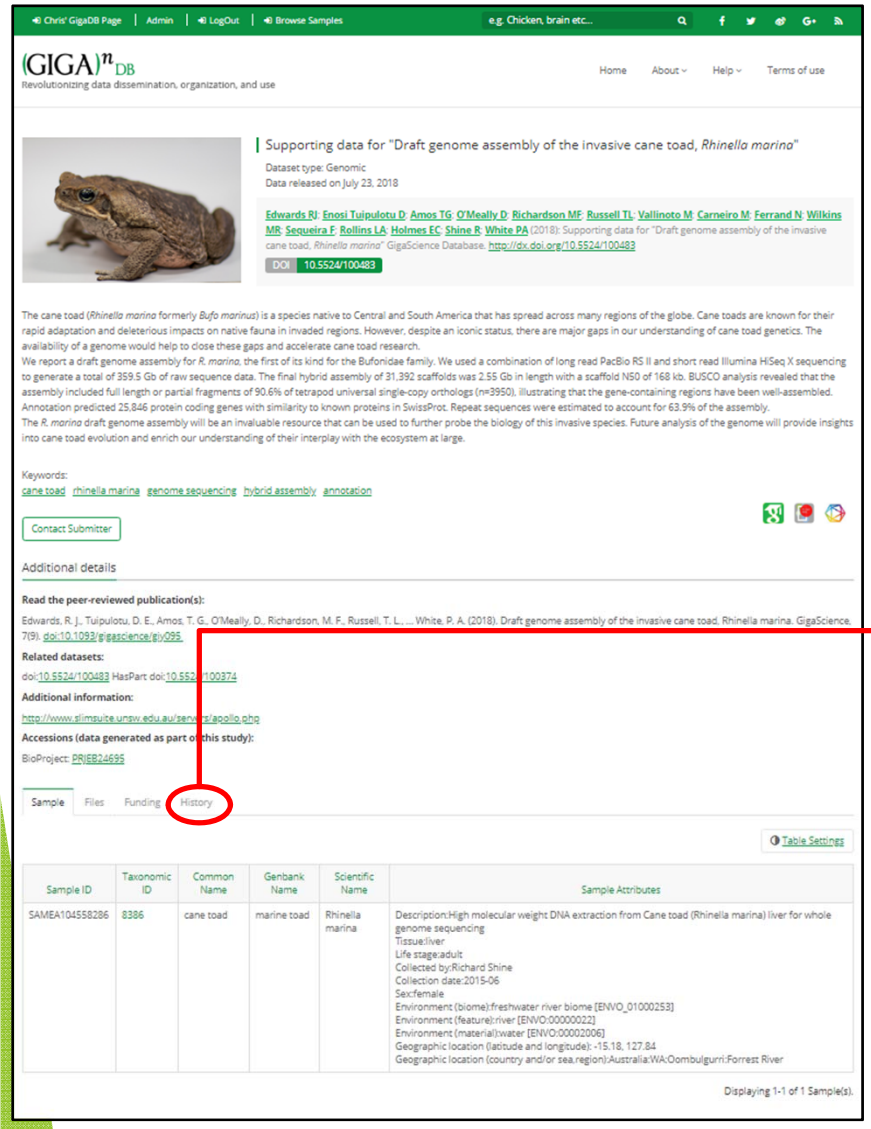
1 2 Next >

Displaying 1-10 of 13 File(s).



Sample Files Funding History

Funding body	Awardee	Award ID	Comments
Australian Research Council	L A Rollins	DE150101393	DECRA Fellowship
Australian Research Council	R Shine	FL120100074	Australian Laureate Fellowship
Australian Research Council	E C Holmes	FL170100022	Australian Laureate Fellowship
Australian Research Council	M R Wilkins	LE150100031	
Fundação para a Ciência e a Tecnologia	M Carneiro	IF/00283/2014/CP1256/CT0012	FCT investigator grant
Brazilian National Council for Scientific and Technological Development	M Vallinoto	302892/2016-8	CNPq fellowship
Bioplatforms Australia	E C Holmes		
Australian Research Council	R Shine	DP160102991	
Fundação para a Ciência e a Tecnologia	F Sequeira	UID/BIA/50027/2013	
Fundo Europeu De Desenvolvimento Regional	F Sequeira	COMPETE, POCI-01-0145-FEDER-006821	



Sample Files Funding **History**

Date	Action
July 23, 2018	Dataset publish
July 23, 2018	Title updated from : Supporting data for "Draft genome assembly of the invasive cane toad, *Rhinella marina*"
July 23, 2018	Title updated from : Supporting data for "Draft genome assembly of the invasive cane toad, Rhinella marina"
September 19, 2018	File readme.txt updated
August 22, 2018	Manuscript Link added : 10.1093/gigascience/giy095

Abstract

O-090

Data curation as a means to promote reproducibility and discoverability

CI. Hunter GigaScience, BGI, Hong Kong

The GigaScience DataBase (GigaDB) is the fundamental infrastructure that enables GigaScience to move beyond the traditional static and descriptive journal article; if a GigaScience article has data associated with it, we curate the metadata and host the files in GigaDB.

While traditional research journals may use a variety of databases to host data for figures associated with their manuscripts GigaDB hosts all the underlying data (where appropriate) to ensure complete transparency and reproducibility of the work.

GigaDB is more than just a file server for supplemental files as all GigaDB datasets are curated by GigaDB staff to ensure the data, metadata and links to associated data are correct, complete and sufficient for purpose, in line with the FAIR (Findable, Accessible, Interoperable and Reusable) principles.

The GigaDB curation process affords a secondary check of the manuscript after peer review, with an emphasis on the identification of underlying and/or intermediary files and data that are required for reproducibility. This does not relieve the reviewers of this responsibility but it can help prevent publication of manuscripts that are missing important data, and help ensure that GigaScience remains true to the Open Access ethos and FAIR principles.

This additional stage can add a small amount of time to the publication process, but this is often mitigated by starting the curation of the dataset prior to formal acceptance while minor revisions are still being made to the manuscript.

On top of organising the data on behalf of the authors, GigaDB tries to educate on best practice for data sharing and organisation during the submission process, and in workshops around the world. We hope that this will increase awareness of data sharing regardless of where authors choose to publish their manuscripts in the future.

The addition of any metadata increases the discoverability of data and GigaDB ensures these details are as discoverable as possible by inclusion of extensive metadata with compliance to various external standards (e.g Schema.org and DataCite) as well as providing an API for programmatic querying.

I hope you met GigaScience team members at posters:

PM-159

The shareability of Hong Kong research experiment

Scott Edmunds

PT-177

Tools for improving transparency of published articles

Si Zhe (Jesse) Xiao