# Data curation as a means to promote reproducibility and discoverability.

**Chris Hunter** 

WCRI 2019, June 5th 2019

Presentation DOI:

# What is (GIGA)<sup>n</sup> SCIENEE

- GigaScience (<u>http://GigaScienceJournal.com</u>) 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



# What is (GIGA)<sup>n</sup><sub>DB</sub>

- 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).



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)<sup>n</sup><sub>DB</sub>
- 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.

### 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	У	First Name, Last Name, Email, Institution/Company, ORCID.
Author list	У	First Name, Last Name, ORCID
Dataset title	У	Manuscript title prefixed with "Supporting data for"
Dataset description	У	Manuscript abstract
Funding information	У	Funding body, program, award ID and awardee
Dataset type	n	Selected from <u>controlled vocabulary</u>
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:PRJNAXXXXX'
Protocols.io link	n	Where authors provide their methods via <u>protocols.io</u> we can embed these in GigaDB datasets, please provide the published widget URL or DOI

http://gigadb.or

## **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 <u>NCBI Taxonomy ID</u> 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 <sup>^</sup>	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 <u>INSDC country list</u>
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.g69.576435, 91.883948
Broad-scale environmental context	R	Please add one or more <u>ENVO terms</u> to describe the broad environment in which sampling occurred e.g. cliff [ENVO:0000087]
Local environmental context	R	Please add one or more <u>ENVO terms</u> to describe the local environment in which sampling occurred as a semicolon separated list, e.g. digestive tract environment [ENVO:01001033]

http://gigadb.org/sile/gi

### **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.

st	Item	Suggested format	Check
	Genome assembly	fasta	
ed	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.or

### Checklists and Guidelines - File metadata

### Our file metadata expectations

Item	Mandatory (y/n)	Description
File name	у	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	у	Short human readable description of the file and its contents
Data type	у	The type of data in the file, selected from a <u>controlled vocabulary</u>
Format	У	Most common formats are automatically assigned by file extension, but can be updated manually if required.
MD5 #value	У	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



## 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 Interoperable \*



### 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/gigas

### 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.

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 Ha



Hans Zauner





Scott Edmunds

Nicole Nogoy

Hongling Zhou





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): <u>http://dx.doi.org/10.5524/100568</u>

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

♠) Chris' GigaDB Page   Admin   ♠) LogOut   ♠) Browse Samples	e.g. Chicken, brain etc	۹	f 🍠	ଷ୍ଟ G+ ୬
(GIGA) <sup>n</sup> <sub>DB</sub> Revolutionizing data dissemination, organization, and use	Home	About ~	Help ~	Terms of use



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. <u>http://dx.doi.org/10.5524/100364</u>



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



Top half of an

Imaging dataset

Contact Submitter

Additional details

Read the peer-reviewed publication(s):

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 Functing 3D Viewer listory

#### 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

### **Bottom half**

### Sample Files Funding 3D Viewer History

#### 3D Models:



### 3D viewer tab



Revolutionizing data dissemination, organization, and use

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

Home

About ~

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



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



\_ @

Terms of use

Help ~

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:



+8 Chris' GigaDB Page Admin +8 LogOut +8 Browse Samples	e.g. Chicken, brain etc	Q f ≇ ø∂ G+ ଲ	
(GIGA) <sup>n</sup> <sub>DB</sub> Revolutionizing data dissemination, organization, and use	Home Abou	t ~ Help ~ Terms of use	Top half Genomic
Supporting data for "Dr.         Dataset type: Genomic         Data released on July 23, 2018         Edwards RJ: Enosi Tuipulotu D;         MR: Sequeira F; Rollins LA; Hold         cane toad, Rhinella marina" Gigas         DOI         10.5524/100483	ft genome assembly of the invasive cane to mos TG: O'Meally D: Richardson MF: Russell TL: Vallinot es EC: Shine R: White PA (2018): Supporting data for "Draf ience Database. <u>http://dx.doi.org/10.5524/100483</u>	oad, Rhinella marina" o M: Carneiro M: Ferrand N: Wilkins t genome assembly of the invasive	dataset
The cane toad ( <i>Rhinella marina</i> formerly <i>Bufo marinus</i> ) is a species native to Central and S rapid adaptation and deleterious impacts on native fauna in invaded regions. However, of availability of a genome would help to close these gaps and accelerate cane toad researce. We report a draft genome assembly for <i>R. marina</i> , the first of its kind for the Bufonidae fit to generate a total of 359.5 Gb of raw sequence data. The final hybrid assembly of 31,392 assembly included full length or partial fragments of 90.6% of tetrapod universal single-of Annotation predicted 25,846 protein coding genes with similarity to known proteins in SV. The <i>R. marina</i> draft genome assembly will be an invaluable resource that can be used to into cane toad evolution and enrich our understanding of their interplay with the ecosyst	uth America that has spread across many regions of the gluspite an iconic status, there are major gaps in our understate. mily. We used a combination of long read PacBio RS II and s scaffolds was 2.55 Gb in length with a scaffold N50 of 168 k py orthologs (n=3950), illustrating that the gene-containing issProt. Repeat sequences were estimated to account for 6 urther probe the biology of this invasive species. Future and em at large.	obe. Cane toads are known for their anding of cane toad genetics. The short read Illumina HiSeq X sequencing db. BUSCO analysis revealed that the pregions have been well-assembled. 3.9% of the assembly. alysis of the genome will provide insights	

Keywords: <u>cane toad</u> <u>rhinella marina</u> <u>genome sequencing</u> <u>hybrid assembly</u> <u>annotation</u>

Contact Submitter

Additional details

Read the peer-reviewed publication(s):



а

ſ							
Contact Sub	omitter						
Additional	details						Bottom half
Read the pee	r-reviewed pu	blication(s):					
Edwards, R. J., 7(9). <u>doi:10.10</u>	Tuipulotu, D. E 193/gigascience	., Amos, T. G., ( <u>giy095</u>	O'Meally,	, D., Richardson	, M. F., Russell, T	T. L., White, P. A. (2018). Draft genome assembly of the invasive cane toad, Rhinella marina. GigaScience,	
Related data	sets:						
doi:10.5524/1	00483 HasPart	doi:10.5524/1	00374				
Additional in	formation:						
http://www.sl	imsuite.unsw.ee	lu.au/servers	/apollo.ph	he			
Accessions (c	lata generated	as part of th	his study)	):			
BioProject: PR	EB24695						
Sample	Files Fundi	ng Histor	У				
						Table Settings	
Sample	ID ID	omic Cor N	nmon ame	Genbank Name	Scientific Name	Sample Attributes	
SAMEA1045	58286 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).	

DB Page Admin 🛛 🖶 LogOut 📔 🖷 Browse Samples	e.g. Chicken, brain etc	٩		•	6)	G
) $n_{ m DB}$ ; data dissemination, organization, and use	Home	About ~	Help ~	т	erms	ofu



GIG

Supporting data for "Draft genome assembly of the invasive cane toad, Rhinella marina" Dataset type: Genomic

Edwards RI: Enosi Tulpulotu D. Amos TG: O'Meally D. Bichardson MF. Russell TL: Vallinoto M. Carneiro M. Ferrand N. Wilkins MS Sequeira F. Rollina LA. Holmes LC: Shine R. White PA (2018): Supporting data for "Draft genome assembly of the invasive carne toad, Rhineller morino" (gisclenice Database http://dx.odi.org/10.5524/10.043

10.5524/100483

Data released on July 23, 2018

The cance tood (Whitells moting formerly Bigs morinui) is a species native to Central and South America that has spread across many regions of the globe. Cance toads are known for their rapid adaptation and deleterious impacts on native faura in invaded regions. However, despite an iconic status, there are major gaps in our understanding of cance toad genetics. The availability of agenome would help to close these gaps and accelerate cance toad creaters.

We report a draft genome assembly for R. morino, the first of its tild for the Bufonides family. We used a combination of long read PacBio RS II and short read llumina HSeq X sequencing to generate a total of 3955 Gb of raw sequence data. The final hybrid assembly of 31.392 scaffolds was 2.55 Gb in length with a scaffold NSO of 168 ibs. BUSCO analysis revealed that the samebly included full length or partial fragments of 90.6% of tetrapod universal single-copy othologis (n#350). Blustating that the gene-containing regions have been well-assembled. Annotation prediced 53.686 protein coding genes with similarity to known proteins in SwissProt. Repeat sequences were estimated to account for 53.9% of the assembly. The R. morino draft genome assembly libe an invaluable resource that can be used to further prote the biology of this invasive species. Furure analysis of the genome WII provide insight

into cane toad evolution and enrich our understanding of their interplay with the ecosystem at large.

cane toad rhinella n	narina genon	<u>te sequencing</u>	<u>hybrid assembly</u>	annotation	8 🦲	
Additional details	5					
Read the peer-revie	ewed publicat	ion(s):				
Edwards, R. J., Tuipul 7(9). doi:10.1093/gig	lotu, D. E., Amo ascience/giy09	os, T. G., O'Meall 5	y, D., Richardsor	n, M. F., Russell,	T. L.,, White, P. A. (2018). Draft genome assembly of the invasive cane toad, Rhinella marina. GigaS	cienc
Related datasets:						
doi:10.5524/100483	HasPan, Joi 10	5524/100374	>—			
Additional information	tion:					
http://www.slimsuite	unsw.edu.au	servers/apollo.	<u>ehe</u>			
Accessions (data ge	nerated as pa	art of this stud	y):			
Sample Files	Funding	History			O Table Se	ttines
	Taxonomic	Common	Genbank	Scientific		
SAMEA104558286	8386	cane toad	marine toad	Rhinella marina	Sample Actioutes Description:High molecular weight DNA extraction from Cane toad (Rhinelia marina) liver for who genome sequencing Transitions	ole

Displaying 1-1 of 1 Sample(s).

### gaDB dataset

					000000	Contraction of the second s
🕈 Chris' GigaDB Page   Admin   🕂 LogOut   🕫 Browse Samples	e.g. Chicken, brain etc	٩	f ¥	r 💰	G•	2
$(GIGA)^{n}_{DB}$	Home	About Y	Help v	Terms	ofuse	

Revolutionizing data dissemination, organization, and use



Supporting data for "Improving amphibian genomic resources: a multi-tissue reference transcriptome of an iconic invader"

Dataset type: Genomic, Transcriptomic Data released on November 13, 2017

Richardson MF; Sequeira F; Selechnik D; Carneiro M; Vallinoto M; Reid JG; West AJ; Crossland MR; Shine R; Rollins LA (2017): Supporting data for "Improving amphibian genomic resources: a multi-tissue reference transcriptome of an iconic invader" GigaScience Database. <u>http://dx.doi.org/10.5524/100374</u>

DOI 10.5524/100374

Cane toads (Rhinella marina) are an iconic invasive species introduced to four continents and well utilized for studies of rapid evolution in introduced environments. Despite the long introduction history of this species, its profound ecological impacts and its utility for demonstrating evolutionary principles, genetic information is sparse. Here we produce a de novo transcriptome spanning multiple tissues and life stages to enable investigation of the genetic basis of previously identified rapid phenotypic change over the introduced range. Using approximately 1.9 billion reads, from developing tadpoles and 6 adult tissue-specific cDNA libraries, and a transcriptome assembly pipeline encompassing 100 separate de novo assemblies, we constructed 62,202 transcripts, of which we functionally annotated –50 %. Our transcriptome assembly exhibits 90% full-length completeness of the BUSCO (benchmarking universal single-copy orthologs) dataset. Robust assembly metrics and comparisons to several available anuran transcriptomes and genomes indicate that our cane toad assembly is one of the most complete anuran genomic resources available. This comprehensive anuran transcriptome will provide a valuable resource for investigation of genes under selection during invasion in cane toads, but will also greatly expand our general knowledge of anuran genomes, which are underrepresented in the literature. The dataset is publically available in NCBI and GigaDB to serve as a resource for other researchers.

#### Keywords:

de novo assembly bufo marinus cane toad rhinella marina invasive species rna-seg transcriptome anuran amphibian

Contact Submitter

CICA	ge Admin	a LogOut	e Browse Si	amples	e.g. Chicken, brain etc Q 🛉 🌶 🛷 G+ 🔈
volutionizing data	DB dissemination	, organization, a	ind use		Home About - Help - Terms of use
a the		D	Support Dataset typ Data releas Edwards R <u>MR Seque</u> cane toad, DOI 10	ing data for we Genomic wed on July 23, 21 J: <u>Enosi Tuipule</u> ira E: <u>Rollins L4</u> Rhinella marina <sup>*</sup>	r "Draft genome assembly of the invasive cane toad, <i>Rhinella marina</i> " ors stu D: Amos T&-O'Meally D: Richardson ME: Russell TL- Yallinoto M: Carneiro M: Ferrand N: Wilkins Nolmes EC: Shine B: White PA (2018): Supporting data for "Draft genome assembly of the invasive "GigaScience Database. <u>http://dx.doi.org/10.5524/100483</u>
The cane toad (Rhine apid adaptation and wailability of a geno Ve report a draft gero o generate a total of ssembly included fr unnotation predicted the <i>R. marina</i> draft g nto cane toad evolu	Illa marina form d deleterious in me would help nome assembi f 359.5 Gb of ra all length or pa d 25.846 protei tenome assemi tion and enrich	nerly Bufo marin mpacts on native to close these ( y for <i>R. marina</i> , aw sequence da w sequence da in coding genes bly will be an im n our understan	etus) is a species r e fauna in invade gaps and acceler the first of its kir ta. The final hyb of 90.6% of tetra with similarity tr valuable resource ding of their inte	native to Centra ed regions. How rate cane toad r nd for the Bufor rid assembly of pod universal s to known proteir te that can be us erplay with the e	I and South America that has spread across many regions of the globe. Cane toads are known for their ever, despite an iconic status, there are major gaps in our understanding of cane toad genetics. The essarch. lidee family. We used a combination of long read PacBio RS II and short read IIlumina HSeq X sequencing 31.392 scaffolds was 2.55 Gb in length with a scaffold NS0 of 168 kb, BUSCO analysis revealed that the lingle-copy orthologic (m3950), illustrating that the gene-containing regions have been well-assembled. It is "NuistPort Repar sequences were estimated to account for 63 % of the assembly. sed to further probe the biology of this invasive species. Future analysis of the genome will provide insights consistem at large.
(eywords: <u>ane toad</u> <u>rhinella n</u> Contact Submitter	narina genom	ne sequencing	<u>hybrid assembly</u>	annotation	🛛 🖲 🔇
teach the peer revie idwards, R. J. Tulpout 199. doi:10.102 Pisig telated datas ts: didditional inf rmai tstp://www.sti-pute taccessions (d. ta ge sioProject: PRI B245 Samp Files	exect publicat lotu, D. E., Amd assience/giv09 HasPart doi:10 tion: .unsw.edu.au/ merated as pa 195 Funding	ion(s): ss, T. G., O'Meall <u>5</u> <u>1.5524/100374</u> <u>iservers'apollo</u> ; ret of this study History	y, D., Richardsor ehe y):	n, M. F., Russell,	T. L, White, P. A. (2018). Draft genome assembly of the invasive cane tood. Rhinella marina. GgaScience.
	Taxonomia	Common	Gaphaci	Scientife	O Table Settings
Sample ID	ID	Name	Name	Name	Sample Attributes
	8386	cane toad	marine toad	Rhinella marina	Description-High molecular weight DNA extraction from Cane toad (Rhinelia marina) liver for whole genome sequencing the second

Funding History Sample Files (FTP site) Table Settings Release File Name Description Size Date canetoad.v2.2.busco.stdout STDOUT for BUSCO v2.5 short run on genome (text) 2018-07-16 7.71 KB . canetoad.v2.2.busco.tgz Gzipped tarball of BUSCO v2.5 short run (directory tarball) 124.52 2018-07-16 + MB canetoad.v2.2.fasta.gz Gzipped genome assembly fasta (fasta) 662.54 2018-07-16 . MB canetoad.v2.2.highquality.aln.tgz Gzipped directory containing MAFFT alignments of predicted orthologues for high quality predicted proteins 40.62 MB 2018-07-16 + (fasta) 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 64.69 MB 2018-07-16 + Newick) canetoad.v2.2.maker.gff3.gz Gzipped MAKER gene annotations (GFF3) 2018-07-16 12.65 MB + 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).

• Order GigdbB Rage           Admin           • Order           Admin           • Order           Admin           Admin         Admin           Admin         Admin           Admin <th></th> <th></th> <th></th>						
The cane toad (Rhinello morino formerly Bufo morinuz) 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 dose these gaps and accelerate cane toad research. We record a drift genome assembly for <i>R morino</i> , the first of its kind for the Buforing et amily. We used a combination of long read PacBio RS II and short read Illumina HSeo X sequencine.	Sample Files Funding History	ng History				
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 NSO of 168 kb. BUSCO analysis revealed that the assembly included full length or partial for otherapid approxed of strappod universal single-copy orthology inv3500, illustrating that the gene-containing regions have been well-assembled. Annotation predicted 25.364 protein coding genes with similarity to known proteins in Singles/Dc. Reparts equences were estimated to account for 6.379 of the assemble.	Funding body	Awardee	Award ID			
The R. marino 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.	Australian Research Council	L A Rollins	DE150101393			
Keywords: cane toad inlinella marinal genome asspuencing hybrid assembly annotation	Australian Research Council	R Shine	FL120100074			
Contact Submitter	Australian Research Council	E C Holmes	FL170100022			
Additional details Read the peer-reviewed publication(s):	Australian Research Council	M R Wilkins	LE150100031			
Edwards, R. J., Tulpulosu, 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:10.1039/gigascient.algo.095.	Fundação para a Ciência e a Tecnologia	M Carneiro	IF/00283/2014/CP1256/CT0012			
Related datasets: doi:10.5524100458_HasPa + doi:10.55241100374 Additional (Incomparing)	Brazilian National Council for Scientific and Technological Development	M Vallinoto	302892/2016-8			
http://www.slimpube.unnew.edu.au/servers/apollo.ghp Accessions (data generati d as part of this study):	Bioplatforms Australia	E C Holmes				
BioProject <u>PR(B22695</u>	Australian Research Council	R Shine	DP160102991			
Sample Files Funding History	Fundação para a Ciência e a Tecnologia	F Sequeira	UID/BIA/50027/2013			
Samnia ID D. Nama Nama Camula Errikuras	Fundo Europeu De Desenvolvimento Regional	F Sequeira	COMPETE, POCI-01-0145-FEDER-006821			
SAMEA104558286 8386 cane toad marine toad Rhinella genome sequencing molecular weight DNA extraction from Cane toad (Rhinella marina) liver for whole genome sequencing						
Tissuetiver Life stage adult Collected by/Rhard Shine Collected by/Rhard Shine Collected by/Rhard Shine Collected by/Rhard Shine Collection date: 2015-06 Serfemale Environment (feature):river (ENVO_0000023) Environment (feature):river (ENVO_00000020) Environment (materialiswater (ENVO_00000000) Geographic location (duction and longtrude): -151.8, 127.84 Geographic location (country and/or sea.region):Australia:WA:Oombulguri:Forrest River Displaying 1-1 of 1 Sample(s).						

Comments DECRA Fellowship Australian Laureate Fellowship Australian Laureate Fellowship

FCT investigator grant

	8- 1 minut		Browse Si	imples	eg	Chicken, brain etc		٩	f ¥	🛷 G• 🔊
GIGA) <sup>n</sup> volutionizing data	DB dissemination	, organization, a	and use				Home	About ~	Help ~	Terms of use
			Support Dataset typ Data releas <u>Edwards R</u> <u>MR: Seque</u> cane toad, I DOI 10	ing data for e: Genomic ed on July 23, 2 I: <u>Enosi Tuipulu</u> I: <u>Enosi Tuipulu</u> Rhinella marina' 15524/100483	r "Draft genome ass 018 3: <u>Holmes EC: Shine R: Whi</u> 3: <u>Holmes EC: Shine R: Whi</u> GigaScience Database. <u>htt</u>	embly of the i : <u>Richardson ME;</u> te PA (2018): Supp ://dx.doi.org/10.5!	nvasive o Russell TL; orting data f 524/100483	ane toad, <u>Vallinoto M</u> : or "Draft gen	Rhinella n Carneiro M: E	n <b>arina"</b> Ferrand N: Wilkins y of the invasive
e cane toad (Rhine pid adaptation and valiability of a geno e report a draft ge sembly included fr nnotation predicter en <i>R. marina</i> draft g to cane toad evolu sywords: <u>ine toad</u> <u>rhinella</u> n	Ila marina form d deleterious in me would help nome assembl f 359.5 Gb of ri all length or pa d 45.846 protei d 25.846 protei d 25.846 protei tion and enrich	nerly Bufo marin mpacts on natiw to close these i y for <i>R. marina</i> , aw sequence da ritial fragments - in coding genes bly will be an in n our understan	hus) is a species n e fauna in invade gaps and acceler the first of its kir ta. The final hyb of 90.6% of tetra with similarity of vith similarity of vith similarity of vith similarity of vith similarity of vith similarity of hybrid assembly	native to Centra ed regions. How rate cane toad r nd for the Bufor rid assembly of pod universal s o known proteir e that can be u erplay with the o	I and South America that hi evere, despite an iconic stats seearch, iidae family. We used a com 31.392 staffolds was 2.55 ( staffolds was 2.55 ( staffolds family). The state staffold and state seet set of surther probe the bio acosystem at large.	s spread across m s, there are major bination of long re bin length with a 0). Illustrating that ences were estima logy of this invasiv	any regions gaps in our ad PacBio R scaffold N50 the gene-co ted to accou e species. Fu	of the globe. I understandin 5 II and short of 168 kb. Bl ntaining regio int for 63.9% ture analysis	Cane toads ar g of cane toai read Illumina ISCO analysis ons have beer of the assemb of the genom	re known for their d genetics. The HISeq X sequencing r revealed that the n well-assembled. by. te will provide insights
Contact Submitter										8 🖪 🔇
dditional detail	5									
dditional detail: aad the peer-revie dwards, R. J., Tuipui 9). <u>doi:10.1093/gig</u> elated datasets: bi:10.5524/100483 dditional informa to://www.slimsulte	s wwed publicat lotu, D. E., Amo ascience/giy09 HasPart doi: <u>10</u> tion: kunsw.edu.au/	ion(s): ss, T. G., O'Meall 5. 1.552. (100374 (serv: 's/apollo.)	y, D., Richardsor	n, M. F., Russell,	T. L White P. A. (2018). [	iraft genome asser	nbly of the i	nvasive cane	toad, Rhinella	marina. GigaScience,
dditional details ead the peer-revid twards, R. J., Tuipul 9). <u>doi:10.1092/gig</u> elated datasets: bi: <u>10.5524/100483</u> dditional informa to://www.slimsuks ccessions (data ge oProject: <u>PRIEB246</u>	s ewed publicat iotu, D. E., Amo ascience/giy09 HasPart doi: <u>10</u> HasPart doi: <u>10</u> eursw.edu.au/ enerated as pa 195	ion(s): ss. T. G. O'Meall S	y, D., Richardsor phg y):	n, M. F., Russell,	T. L White, P. A. (2018). I	haft genome asse	nbly of the i	nvasive cane	toad, Rhinella	marina. GigaScience,
dditional detail: ad the peer-revi wards, R. J., Tuipul 9. dditol.1099/gig alated datasets: bit0.5524/100439 ditional informa- ter//www.slimsuist ccessions (data ge oProject: PRIEB246 Sample Files	s weed publicat iotu, D. E. Amo ascience/gi)03 HasPart doi:10 tion: kunsw.edu.au/ merated as pa 195 Funding	ion(s): S. T. G. O'Meall S. S.552, 1100374 iserv s'apollo, i art of this study History	y, D., Richardsor ohg y):	n, M. F., Russell,	T. L White, P. A. (2018). I	braft genome asser	mbly of the i	nvasive cane i	coad. Rhinella	marina. GigaScience,
dditional detail: ad the peer-revie wards, R. J., Tuipul 9), <u>doi:10.1032/gip</u> elated datasets: bi:10.5524/100483 dditional informa to://www.slimsuks dditional informa to://ww	s weed publicat iotu, D. E., Ame ascience/gi)099 HasPart doi:10 tion: .unsw.edu.au/ merated as pa 195 Funding	ion(s): ss, T. G., O'Meal 5. 1.552 1100374 1.552 1200374 starvy sciaoolio, htt of this study History	y, D. Richardson <u>oh</u> g y):	, M. F. Russell,	T. L White, P. A. (2018). 8	kraft genome asse	mbly of the i	nvasive cane i	coad, Rhinella	marina GigaScience.
dditional details aad the peer-revie wards. R. J. Tuipul 91. doi:10.1032/pg 91. doi:10.1032/pg abed dataset: sh:10.5524/100483 dditional informa ro//www.imsuks cossions(data ge opProject: PRIE8246 Sample ID	s wwed publicat toru, D. E. Arm ascience/gi/09 HasPart doi:10 tion: wunsw.edu.au/ merrated as pa 195 Funding Taxonomic ID	ion(s): sr. T. G. O'Meall <u>5</u> <u>1552</u> <u>100374</u> <u>1552</u> <u>100374</u> <u>1552</u> <u>100374</u> <u>1552</u> <u>100374</u> <u>1552</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>100374</u> <u>1003755555555555555555555555555555555555</u>	y, D., Richardson ehg yk: Genbank Name	, M. F., Russell, Scientific Name	T. L White, P. A. (2016). D	vaft genome asse	mbly of the i	nvasive cane i	Rhinella	Marina, GigaScience,

 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
 Title updated from : Supporting data for "Draft genome assembly of the invasive cane toad, Rhinella marina"

 September 19, 2018
 Title updated from : Supporting data for "Draft genome assembly of the invasive cane toad, Rhinella marina"

 August 22, 2018
 Manuscript Link added : 10.1093/gigascience/giy095



### Abstract

### O-090

Data curation as a means to promote reproducibility and discoverability

Cl. 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