GigaByte – A New Workflow for Rapid Dissemination of Datasets & Tools - Bringing Papers to Life

(GIGA)byte

Publishing at the Speed of Research

eResearch 2021, Wellington, NZ

Nicole Nogoy

Executive Editor,

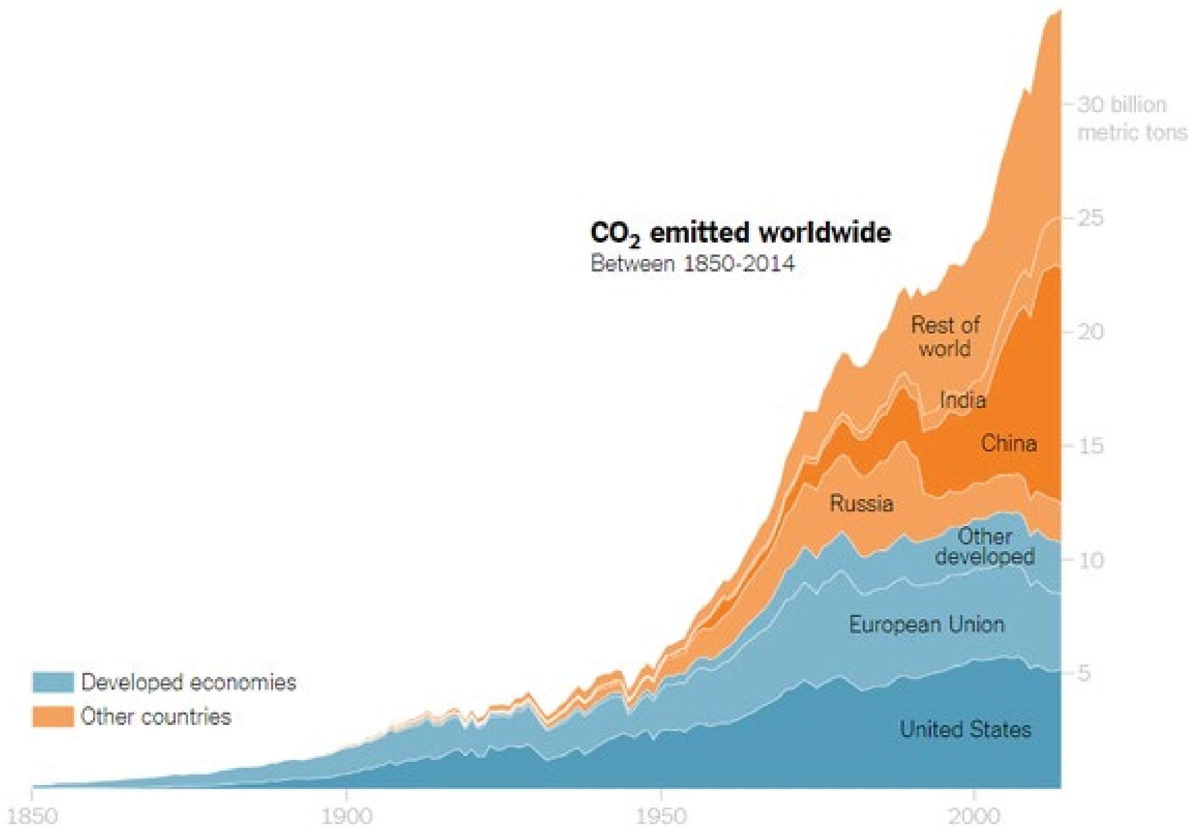
ORCID: 0000-0002-5192-9835

(GIGA)^n SCIENCE PRESS

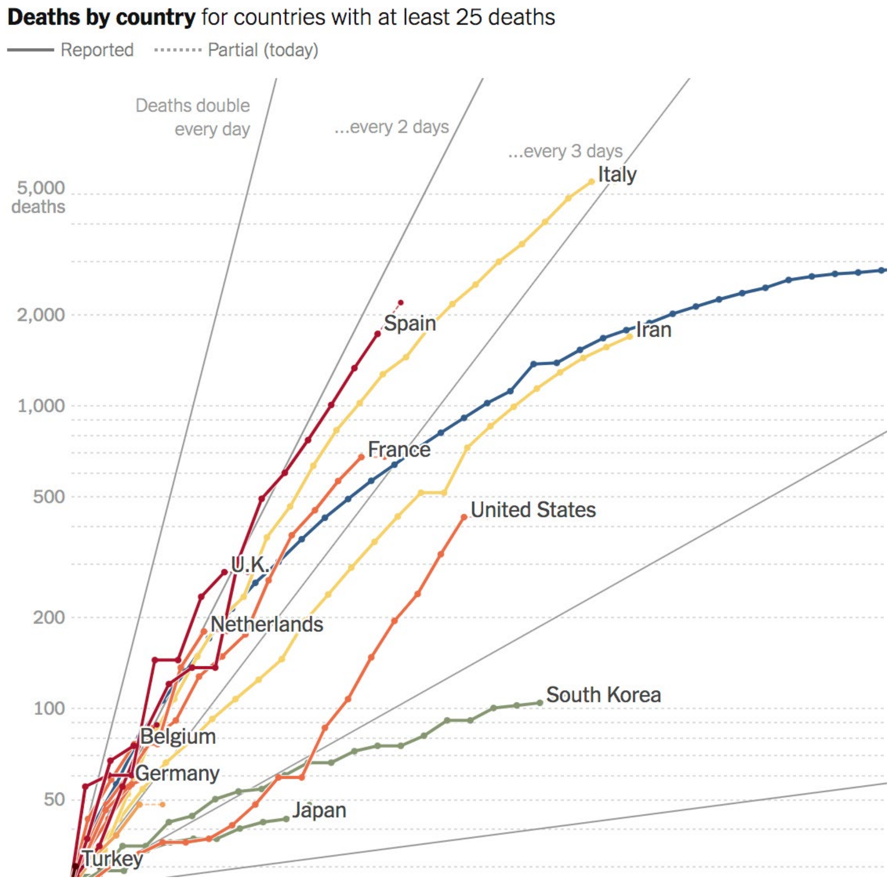
BGI 华大

Urgent Challenges Research Needs to Address

Climate change

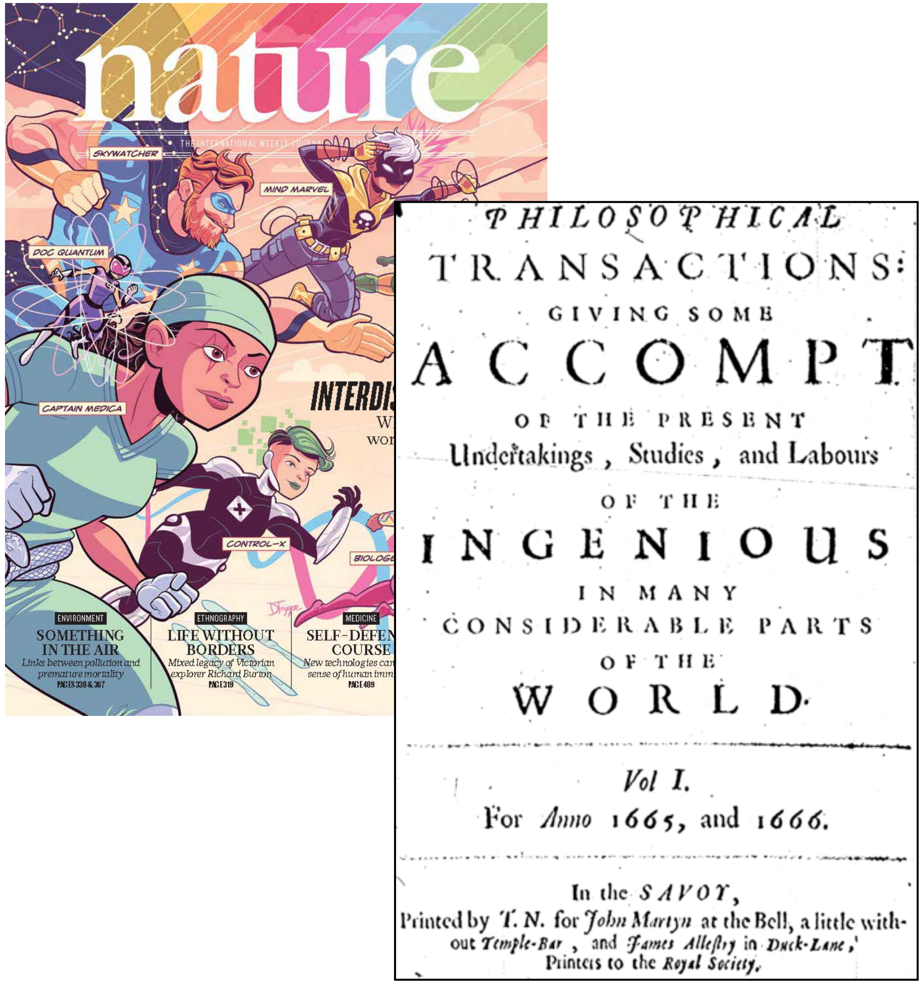


Disease pandemics (e.g., COVID19)



https://www.nytimes.com/2019/02/28/learning/teach-about-climate-change-with-these-24-new-york-times-graphs.html
https://www.nytimes.com/interactive/2020/03/21/upshot/coronavirus-deaths-by-country.html

Urgent Challenges: Publishing to the Rescue?



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*.

- Need to disseminate & communicate information openly to global community
- Expensive. Information is held back by paywalls, APCs, and other barriers

- Need to share this quickly and in a trusted form (peer reviewed)
- Laborious, archaic tech, and untransparent processes

- Research data, software and underlying methods and results need to be shared for scrutiny and re-use
- Hard work and little incentive to share

- Needs to be understandable by policy makers, public
- Barriers of language, jargon, and lack of interaction

Attempt to Address This: *GigaScience* + GigaDB (2012)

GigaScience, 9, 2020, 1–11

doi: 10.1093/gigascience/giaa100
Data Note

DATA NOTE

Construction of a chromosome-scale long-read reference genome assembly for potato

Gina M. Pham[1,†], John P. Hamilton[1,†], Joshua C. Wood[1], Joseph T. Burke[1], Hainan Zhao[1], Brieanne Vaillancourt[1], Shujun Ou[2], Jiming Jiang[1,3,4] and C. Robin Buell[1,4,5,*]

[1]Department of Plant Biology, Michigan State University, 612 Wilson Road, East Lansing, MI 48824, USA; [2]Department of Ecology, Evolution, and Organismal Biology, Iowa State University, 2200 Osborne Dr, Ames, IA 50011, USA; [3]Department of Horticulture, Michigan State University, 1066 Bogue St, East Lansing, MI 48824, USA; [4]MSU AgBioResearch, Michigan State University, 446 W. Circle Drive, East Lansing, MI 48824, USA and [5]Plant Resilience Institute, Michigan State University, 612 Wilson Road, East Lansing, MI 48824, USA

[*]**Correspondence address.** C. Robin Buell, Department of Plant Biology, Michigan State University, 612 Wilson Road, East Lansing, MI 48824, USA. E-mail: buell@msu.edu http://orcid.org/0000-0002-6727-4677
[†]These authors contributed equally to this work.

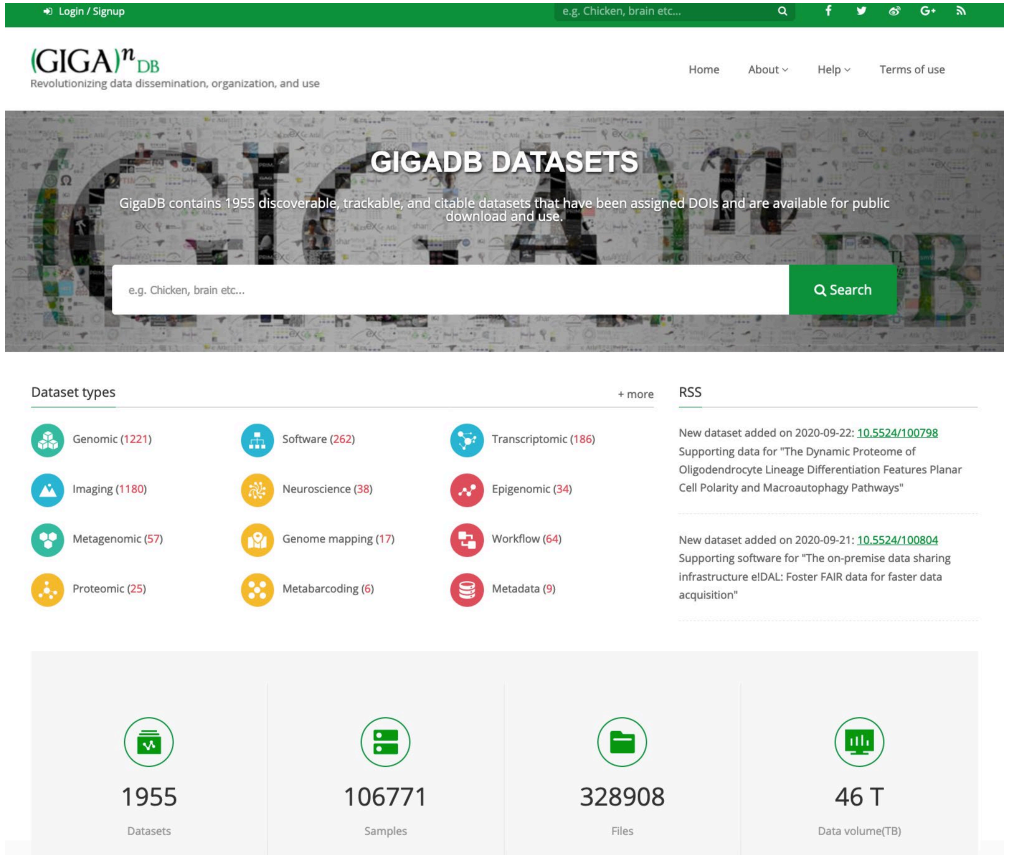
Abstract

Background: Worldwide, the cultivated potato, *Solanum tuberosum* L., is the No. 1 vegetable crop and a critical food security crop. The genome sequence of DM1–3 516 R44, a doubled monoploid clone of *S. tuberosum* Group Phureja, was published in 2011 using a whole-genome shotgun sequencing approach with short-read sequence data. Current advanced sequencing technologies now permit generation of near-complete, high-quality chromosome-scale genome assemblies at minimal cost. **Findings:** Here, we present an updated version of the DM1–3 516 R44 genome sequence (v6.1) using Oxford Nanopore Technologies long reads coupled with proximity-by-ligation scaffolding (Hi-C), yielding a chromosome-scale assembly. The new (v6.1) assembly represents 741.6 Mb of sequence (87.8%) of the estimated 844 Mb genome, of which 741.5 Mb is non-gapped with 731.2 Mb anchored to the 12 chromosomes. Use of Oxford Nanopore Technologies full-length complementary DNA sequencing enabled annotation of 32,917 high-confidence protein-coding genes encoding 44,851 gene models that had a significantly improved representation of conserved orthologs compared with the previous annotation. The new assembly has improved contiguity with a 595-fold increase in N50 contig size, 99% reduction in the number of contigs, a 44-fold increase in N50 scaffold size, and an LTR Assembly Index score of 13.56, placing it in the category of reference genome quality. The improved assembly also permitted annotation of the centromeres via alignment to sequencing reads derived from CENH3 nucleosomes. **Conclusions:** Access to advanced sequencing technologies and improved software permitted generation of a high-quality, long-read, chromosome-scale assembly and improved annotation dataset for the reference genotype of potato that will facilitate research aimed at improving agronomic traits and understanding genome evolution.

Keywords: long-read; chromosome-scale; reference genome; potato

Received: 4 June 2020; **Revised:** 26 August 2020; **Accepted:** 5 September 2020

© The Author(s) 2020. Published by Oxford University Press GigaScience. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.



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

GIGADB DATASETS

GigaDB contains 1955 discoverable, trackable, and citable datasets that have been assigned DOIs and are available for public download and use.

Dataset types

+ more

Genomic (1221)	Software (262)	Transcriptomic (186)
Imaging (1180)	Neuroscience (38)	Epigenomic (34)
Metagenomic (57)	Genome mapping (17)	Workflow (64)
Proteomic (25)	Metabarcoding (6)	Metadata (9)

RSS

New dataset added on 2020-09-22: 10.5524/100798
Supporting data for "The Dynamic Proteome of Oligodendrocyte Lineage Differentiation Features Planar Cell Polarity and Macroautophagy Pathways"

New dataset added on 2020-09-21: 10.5524/100804
Supporting software for "The on-premise data sharing infrastructure e!DAL: Foster FAIR data for faster data acquisition"

1955	106771	328908	46 T
Datasets	Samples	Files	Data volume(TB)

http://gigasciencejournal.com

http://gigadb.org/

(GIGA)^n SCIENCE

8 years in numbers

765 papers published & 231 Data Notes

5,411 authors from

78 countries

1,575 institutions

DataCite FIND, ACCESS, AND REUSE DATA

(GIGA)^n DB

We've published

46 TB of data

1,914 datasets

327,658 files

Our contents is used in

20+ patents and policy documents

1,000+ news articles and blogs

21,186+ tweets

1 PROSE Award innovation in journal publishing multidisciplinary

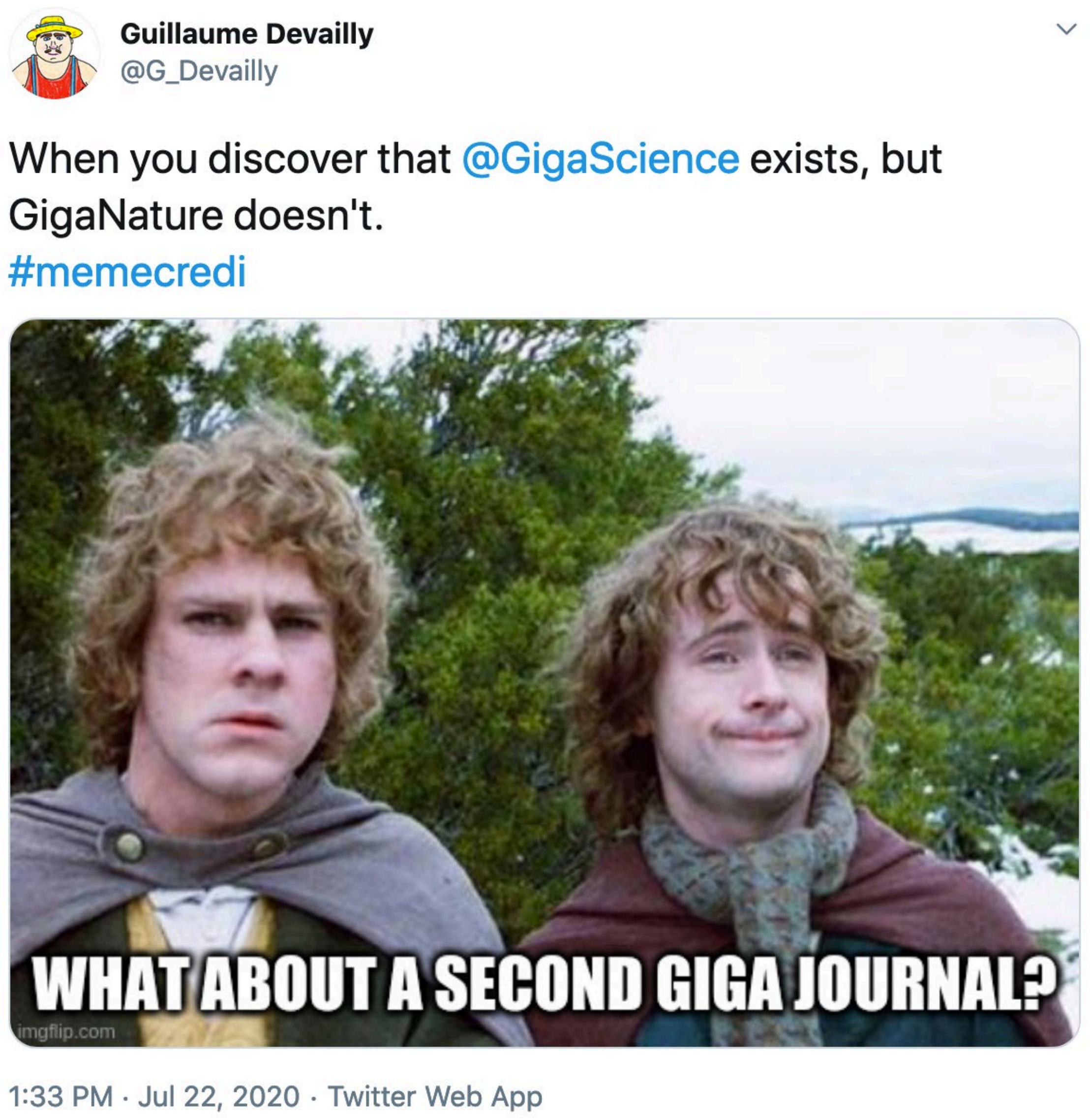
5 editors in 5 locations across 16 time zones

68+ years of editorial experience

3 data curators with 65+ years of experience

Lessons Learned

- Technology is really the bottleneck
- Traditional publishing process is too slow & expensive
- Much too focused on narrative and static “version of record”



Guillaume Devailly
@G_Devailly

When you discover that @GigaScience exists, but GigaNature doesn't.
#memecredi

1:33 PM · Jul 22, 2020 · Twitter Web App

A New Approach

Follow the Software Paradigm?

CODE RELEASE FORK UPDATE REPEAT

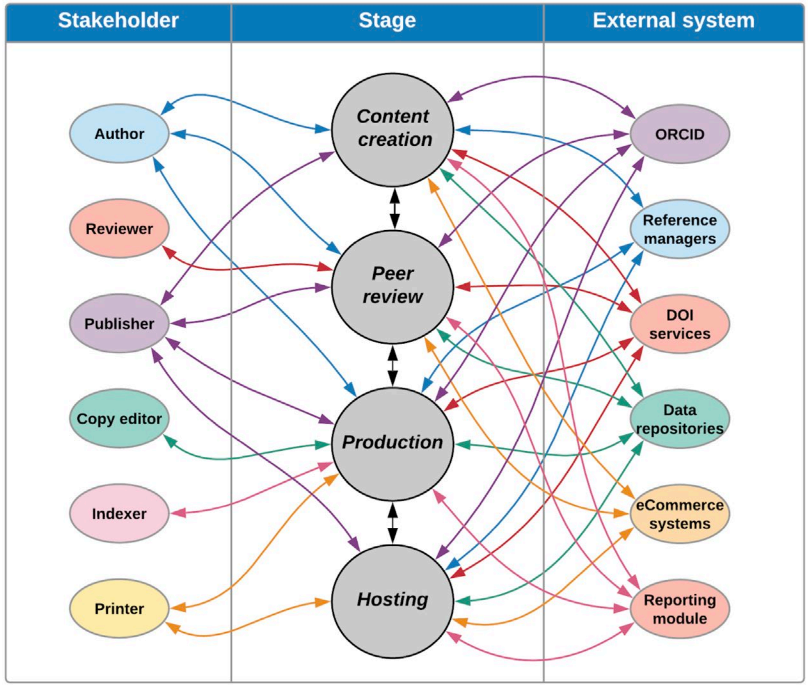
Deconstruct the “Version of Record”?

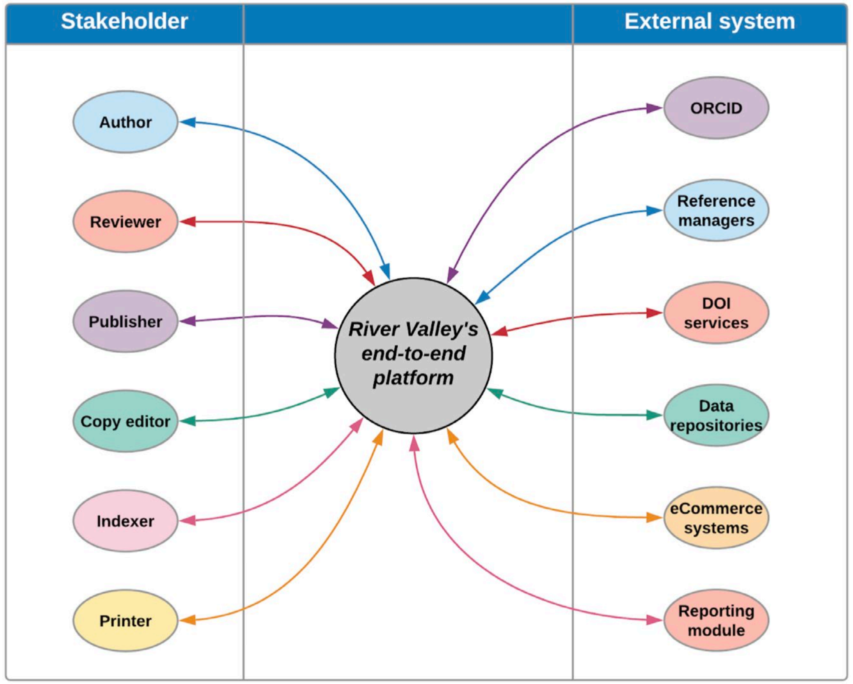
DATA CODE ENTITIES FACTS STABILITY

Move To New XML End-to-End Pipeline

RIVER VALLEY
TECHNOLOGIES

Custom end-to-end workflow makes integrations simpler with one integration point





(GIGA)byte A New Approach With New Tech

Main advantage of workflow is XML from start to end

Several modules acting as one platform: no import/export of files, so fast and accurate

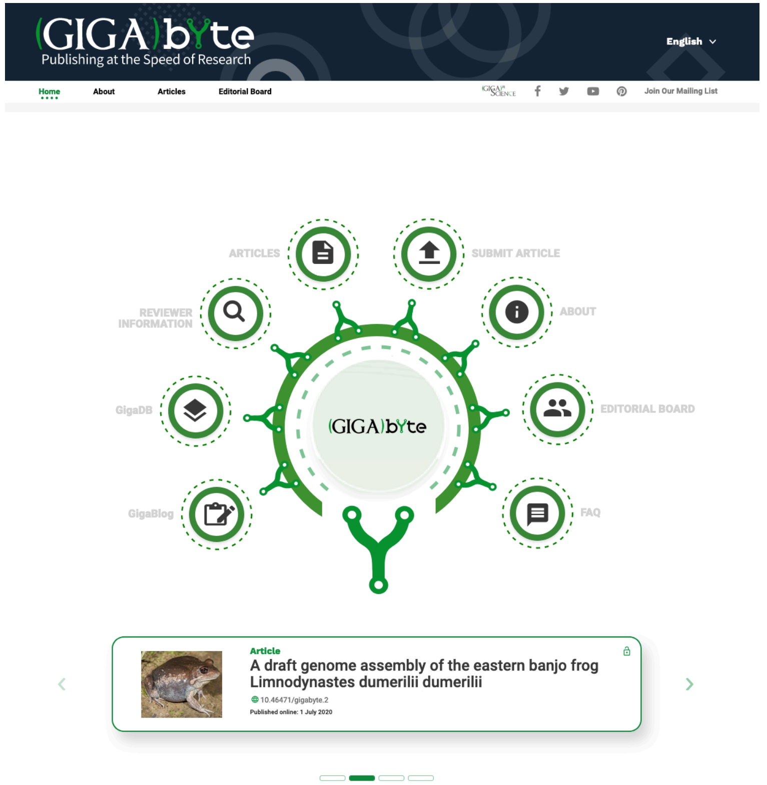
Cutting out production allows huge time & cost saving (currently 4-8 hours NOT 2 months+)

Any number of versions can be published instantly, including typographic quality PDF or updates/forks –minimal effort

Allows instantaneous switch of views – push of a button

Initial focus on forkable products: data + software + updates

Leverage embeddable dynamic content/widgets



Article

A draft genome assembly of the eastern banjo frog Limnodynastes dumerilii dumerilii

10.46471/gigabyte.2

Published online: 1 July 2020

https://gigabytejournal.com/

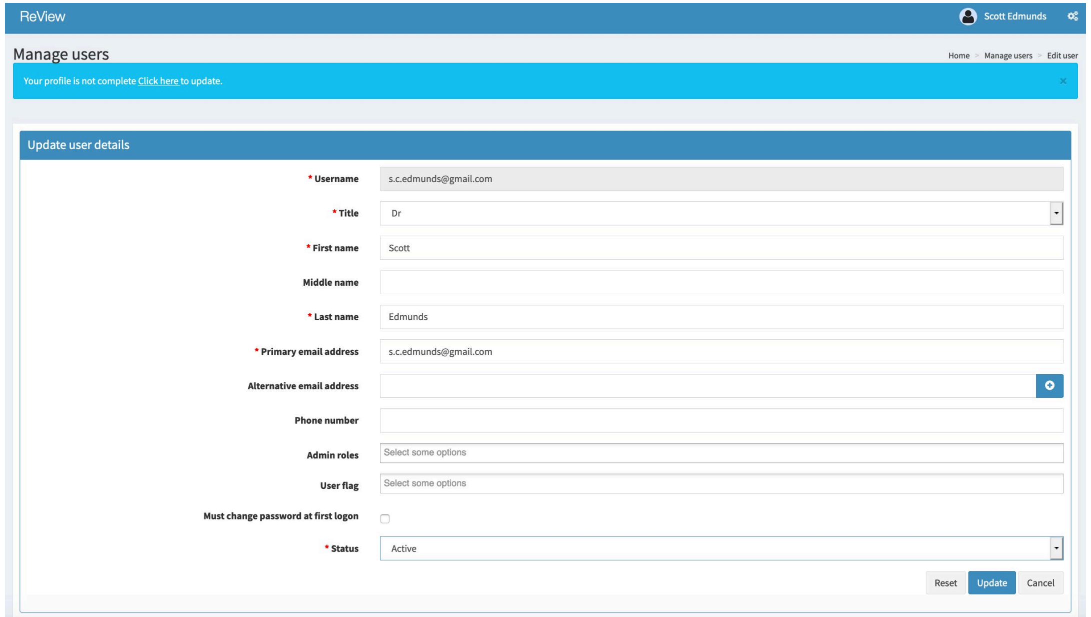
RIVER VALLEY TECHNOLOGIES

(GIGA)^n SCIENCE PRESS

Thinking About Users: Authors, Reviewers, Readers

Reconfigured for short, easy to write & review data & software papers

Streamlined questionnaire-based review



ReView

Scott Edmunds

Manage users

Home > Manage users > Edit user

Your profile is not complete Click here to update. ×

Update user details

* Username: s.c.edmunds@gmail.com

* Title: Dr

* First name: Scott

Middle name:

* Last name: Edmunds

* Primary email address: s.c.edmunds@gmail.com

Alternative email address:

Phone number:

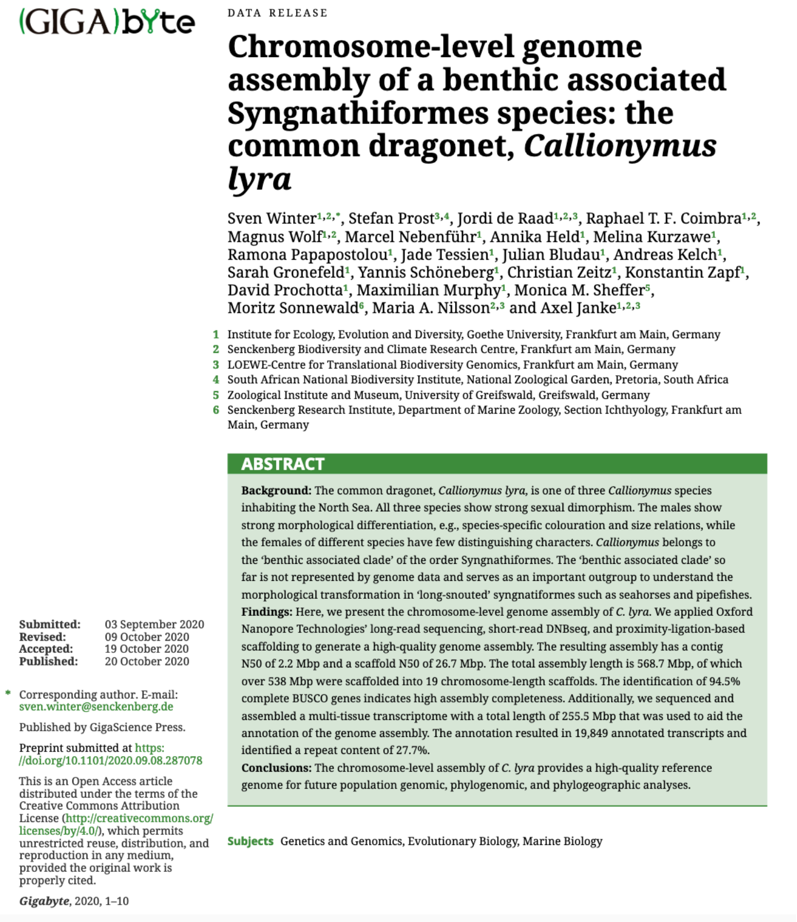
Admin roles: Select some options

User flag: Select some options

Must change password at first logon ☐

* Status: Active

Export as PDF, XML, HTML… “on the fly”



(GIGA)bYte

DATA RELEASE

Chromosome-level genome assembly of a benthic associated Syngnathiformes species: the common dragonet, *Callionymus lyra*

Sven Winter[1,2,*], Stefan Prost[3,4], Jordi de Raad[1,2,3], Raphael T. F. Coimbra[1,2], Magnus Wolf[1,2], Marcel Nebenführ[1], Annika Held[1], Melina Kurzawe[1], Ramona Papapostolou[1], Jade Tessien[1], Julian Bludau[1], Andreas Kelch[1], Sarah Gronefeld[1], Yannis Schöneberg[1], Christian Zeitz[1], Konstantin Zapf[1], David Prochotta[1], Maximilian Murphy[1], Monica M. Sheffer[5], Moritz Sonnewald[6], Maria A. Nilsson[2,3] and Axel Janke[1,2,3]

1 Institute for Ecology, Evolution and Diversity, Goethe University, Frankfurt am Main, Germany
2 Senckenberg Biodiversity and Climate Research Centre, Frankfurt am Main, Germany
3 LOEWE-Centre for Translational Biodiversity Genomics, Frankfurt am Main, Germany
4 South African National Biodiversity Institute, National Zoological Garden, Pretoria, South Africa
5 Zoological Institute and Museum, University of Greifswald, Greifswald, Germany
6 Senckenberg Research Institute, Department of Marine Zoology, Section Ichthyology, Frankfurt am Main, Germany

ABSTRACT

Background: The common dragonet, *Callionymus lyra,* is one of three *Callionymus* species inhabiting the North Sea. All three species show strong sexual dimorphism. The males show strong morphological differentiation, e.g., species-specific colouration and size relations, while the females of different species have few distinguishing characters. *Callionymus* belongs to the ‘benthic associated clade’ of the order Syngnathiformes. The ‘benthic associated clade’ so far is not represented by genome data and serves as an important outgroup to understand the morphological transformation in ‘long-snouted’ syngnatiformes such as seahorses and pipefishes.

Findings: Here, we present the chromosome-level genome assembly of *C. lyra*. We applied Oxford Nanopore Technologies’ long-read sequencing, short-read DNBseq, and proximity-ligation-based scaffolding to generate a high-quality genome assembly. The resulting assembly has a contig N50 of 2.2 Mbp and a scaffold N50 of 26.7 Mbp. The total assembly length is 568.7 Mbp, of which over 538 Mbp were scaffolded into 19 chromosome-length scaffolds. The identification of 94.5% complete BUSCO genes indicates high assembly completeness. Additionally, we sequenced and assembled a multi-tissue transcriptome with a total length of 255.5 Mbp that was used to aid the annotation of the genome assembly. The annotation resulted in 19,849 annotated transcripts and identified a repeat content of 27.7%.

Conclusions: The chromosome-level assembly of *C. lyra* provides a high-quality reference genome for future population genomic, phylogenomic, and phylogeographic analyses.

Subjects Genetics and Genomics, Evolutionary Biology, Marine Biology

Submitted: 03 September 2020
Revised: 09 October 2020
Accepted: 19 October 2020
Published: 20 October 2020

* Corresponding author. E-mail: sven.winter@senckenberg.de

Published by GigaScience Press.

Preprint submitted at https://doi.org/10.1101/2020.09.08.287078

Gigabyte, 2020, 1–10

https://gigabytejournal.com/

Focusing Beyond Version of Record: Allows Different Views

Interactive Features – Increases Understanding and Trust

Thanks to:

BGI 华大

(GIGA)ⁿ SCIENCE PRESS

RIVER VALLEY TECHNOLOGIES

Laurie Goodman, *Publisher*
Scott Edmunds, *Editor-in-Chief*
Hans Zauner, *Assistant Editor*
Hongling Zhao, *Assistant Editor*
Peter Li, Head of *IT*
Chris Hunter, *Lead BioCurator*
Chris Armit, *Data Scientist*
Mary Ann Tulli, *Data Editor*
Rija Ménagé, Senior Software Engineer
Ken Cho, Systems Programmer Analyst
Chen Qi, *Shenzhen Office*

Follow us:

@GigaByteJournal

facebook.com/GigaByteJournal

http://gigasciencejournal.com/blog/

editorial@gigabytejournal.com

Submit now, free APCs till 28th Feb 2021

https://gigabytejournal.com/