Int J Digit Libr (2017) 18:99–111
DOI 10.1007/s00799-016-0174-6

Experiences in integrated data and research object publishing using GigaDB

Scott C Edmunds[1] · **Peter Li**[1] · **Christopher I Hunter**[1] · **Si Zhe Xiao**[1] · **Robert L Davidson**[1,2] · **Nicole Nogoy**[1] · **Laurie Goodman**[1]

Received: 29 June 2015 / Revised: 30 April 2016 / Accepted: 10 May 2016 / Published online: 27 May 2016

Abstract In the era of computation and data-driven research, traditional methods of disseminating research are no longer fit-for-purpose. New approaches for disseminating data, methods and results are required to maximize knowledge discovery. The "long tail" of small, unstructured datasets is well catered for by a number of general-purpose repositories, but there has been less support for "big data". Outlined here are our experiences in attempting to tackle the gaps in publishing large-scale, computationally intensive research. *GigaScience* is an open-access, open-data journal aiming to revolutionize large-scale biological data dissemination, organization and re-use. Through use of the data handling infrastructure of the genomics centre BGI, *GigaScience* links standard manuscript publication with an integrated database (GigaDB) that hosts all associated data, and provides additional data analysis tools and computing resources. Furthermore, the supporting workflows and methods are also integrated to make published articles more transparent and open. GigaDB has released many new and previously unpublished datasets and data types, including as urgently needed data to tackle infectious disease outbreaks, cancer and the growing food crisis. Other "executable" research objects, such as workflows, virtual machines and software from several *GigaScience* articles have been archived and shared in reproducible, transparent and usable formats. With data citation producing evidence of, and credit for, its use in the wider research community, *GigaScience* demonstrates a move towards more executable publications. Here data analyses can be reproduced and built upon by users without coding backgrounds or heavy computational infrastructure in a more democratized manner.

Keywords Reproducibility · Open-data · Data publishing · Computational biology · Data citation

✉ Scott C Edmunds
scott@gigasciencejournal.com

[1] GigaScience, BGI-Hong Kong Co, Ltd, 16 Dai Fu Street, Tai Po Industrial Estate, NT, Hong Kong SAR, China

[2] Office for National Statistics, Duffryn, Government Buildings, Cardiff Rd, Newport NP10 8XG, UK

1 Introduction

In a world where zettabytes of electronic information are now produced globally each year [53], quick and easy access to this information is becoming increasingly important in realizing its potential for society and human development. For scientific data in particular, removing silos and opening access to enable new data-driven approaches increase transparency and self-correction, allow for more collaborative and rapid progress, and enable the development of new questions—revealing previously hidden patterns and connections across datasets.

On top of a citation advantage [62], public access to data has had other measurable benefits to individual fields, such as rice research [90]. Further, pressing issues led by the stresses of a growing global population, such as climate change, rapid loss of biodiversity, and public health costs, require urgent and rapid action. Unfortunately, shrinking research budgets in much of the world mean that the access and use of research data that is already being collected need to be maximized as much as possible. There is growing awareness and uptake of open access publishing, with some estimates that up to half the papers currently being published are now free-to-read [83]. Browsing the narrative is only the first step, but key to maximizing the utility of publicly funded research is the ability to access the supporting data and build upon the contents—an area that needs more development.

Funding agencies, such as the US National Science Foundation, have started to mandate data management and share plans for all funded projects, and the NIH is likely to go a step further—investing through their "Big Data to Knowledge" program in the development of a biomedical and healthCAre Data Discovery and Indexing Ecosystem, or "bioCADDIE" (http://biocaddie.ucsd.edu/). The data discovery index enabled through bioCADDIE aims to achieve data what PubMed has achieved for the literature, like Pubmed and PubMed Central, should provide infrastructure and momentum towards mandatory data archiving. In Europe, the Directorate General for Research and Innovation and OpenAIRE (http://www.openaire.eu/) requires grant recipients to make their publications Open Access. They also have an Open Research Data Pilot for selected H2020 projects that mandates a Data Management Plan and deposition of data funded by the grant in a research data repository, such as Zenodo or the upcoming European Open Science Cloud for Research infrastructure. The UK Research Councils (RCUK) are also drafting a Concordat on Open Research Data (http://www.rcuk.ac.uk/research/opendata/) that promotes making research data open and usable, but does not act as a mandate. The lack of mandate at the RCUK level is indicative of the general approach of broad, abstract statements in support of open access and data, but with little hard line necessity. This and the lack of cross-border agreement greatly limits the uptake of such policies in international research.

The other key stakeholders are journals, and while they have supposedly been tightening their policies (culminating in cross publisher schemes, such as the Joint Data Archiving Policy in Ecology Journals [87]), there is still a very long way to go in terms of compliance. Although there are encouraging signs that some open access publishers are starting to address this issue [8]. While a recent survey found 44 out of the top 50 highest impact journals have made policy statements about data sharing, data are available for only a minority of articles [1], and in some cases access to raw data can be as low as 10 % [72]. With increasing dependence on computational methods, code availability policies are unfortunately even more poorly adhered to than data release policies [77].

These deficiencies have led to a 'reproducibility gap', where most studies cannot be reproduced based on limited available information in the published paper. Systematically testing the reproducibility of papers across various research fields, Ioannidis and others determined that the proportion of published research findings are false or exaggerated, and an estimated 85 % of research resources are wasted because of this [33]. For example, an evaluation of a set of microarray studies revealed that only 1 out of 9 could be reproduced in principle [34], and similar results have been seen in preclinical cancer research, where scientific findings could be confirmed in only 11 % of cases [6]. These sorts of figures have been reported in a number of studies, and an analysis of past preclinical research studies indicates that the cumulative prevalence of irreproducible preclinical research exceeds 50 %; in other words, the US$28 B/year spent on preclinical research is not reproducible [24]. With increasing rates of retraction of published work (particularly correlating with supposed 'higher impact' journals [21]), it is important to stem these deficiencies, not only to prevent the waste and even incorrect answers to health and environmental issues that could have deadly consequences, but also to prevent undermining public confidence in science.

Current formats for disseminating scientific information—the static scientific journal article—have fundamentally not changed for centuries and need to be updated for the current, more data-driven and digital age. Especially so considering journal articles are the primary method for judging work achievements and career success. As stated by Buckheit and Donoho, scholarly articles are merely advertisement of scholarship, and the actual scholarly artefacts are the data and computational methods [10]. Further, with particularly interesting and important datasets and tools, curious users outside of traditional academic environments can also be engaged and utilized through citizen science. This approach has already demonstrated novel insights can be gained through widening the user base to include people with different perspectives and a lack of many preconceptions (e.g. Galaxy Zoo [13]). Making scientific and medical information open access has educative potential, and also enables informed decision making by patients, doctors, policy makers and electorates that may not otherwise have access to this information. It also provides material for data miners to extract new knowledge, and enabling open data advocates and developers to build new infrastructure, apps and tools.

1.1 The need for integrated data publishing

One of the key ways to tackle this reproducibility gap and—importantly—to accelerate scientific advances is to make data and code freely and easily available. However, doing this in the real world is far more problematic than it sounds. The current mechanism of simply relying on authors to provide data and materials on request or from their own websites has been clearly shown not to work. The Reproducibility Initiative: Cancer Biology Consortium carried out a study to quantify many of these issues by trying to repeat experiments from 50 highly cited studies published in 2010–2012. Their attempt to obtain data using this approach took, on average, two months to obtain the data for each paper, and in 4 out of 50 cases, the authors had yet to cooperate after a year of chasing [84]. Further, based on an assessment of papers in the ACM conferences and journals, obtaining the code, which is essential for replication and reuse, took two months on average for nearly 44 % of papers [14], and a survey of 200 economics publications found that of the 64 % of the authors that responded, 56 % *would not* share supplementary materials [39].

As an alternative to relying only on authors' time and goodwill, funding agencies can play a role in pushing data and code sharing. However, despite the good intentions of a few forward thinking organizations, such as the NIH and Wellcome Trust, most funders around the world do not enforce this before publication. Journals, therefore, default to being one of the few stakeholders who can make this happen. But, with their current focus more on narrative delivery rather than code and data access, this has also been problematic.

Carrot and stick approaches, however, are not enough. The hurdles for action need to be lowered, so as to make data FAIR (findable, accessible, interoperable and re-usable) [85]. Data management and curation is an expensive and complicated process that most researchers are simply unable to deal with. Further, data storage does not come cheap: a number of studies on disciplinary research data centres in the UK and Australia, funded by their Research Councils, found the running costs of these data centres to be roughly 1.5 % of the total research expenditure [29].

While Open Access textual content is being worked into policies and mandates from research funders, looking beyond static archived PDFs, a Research Object (RO)-oriented approach to all the products of the research cycle is needed. ROs are semantically rich Linked Data aggregations of resources that provide a layer of structure on top of this textual information, bundling together essential information relating to experiments and investigations. This includes not only the data used, and methods employed to produce and analyse that data, but also links and attributes the people involved in the investigation [3].

There are a growing number of databases that allow scientists to share their findings in more open and accessible ways, as well as a new generation of data journals trying to leverage them. Some areas of biology and chemistry (particularly those working with nucleic acid or X-ray crystal structure data) have been well catered for over many decades with an ecosystem of domain-specific databases, such as those of the International Nucleotide Sequence Database Consortium (INSDC; GenBank, DDBJ and ENA), and Worldwide Protein Data Bank. There is also now a selection of broad-spectrum databases including Zenodo, Dryad, figshare and the DataVerse repositories. These broad-spectrum databases have the benefit of not having data-type restrictions, and researchers can deposit data from the entire set of experiments of a study in a single place. Although these resources cater well for the 'long-tail' of data producers working with tabular data in the megabyte to gigabyte size range, researchers working in more data-intensive areas producing large-scale imaging, high-throughput sequencing and mass spectrometry may not be as well served, due to their file size limitations and charges.

Still, beyond storage, there needs to be more incentive to make this still difficult activity more worthwhile: data and method producers need to be credited doing so [15]. Effectively, wider and more granular methods of credit such as micro- or even nano-publication need to be accepted [60], and new platforms and infrastructure are required to enable this information to be disseminated and shared as easily and openly as possible.

To establish such a mechanism for providing this type of credit, along with data storage and access, code availability, and open use for all of these components, BGI, the world's largest producer of genomics data, and BioMed Central, the world's first commercial open access publisher, built a novel partnership and launched the open access, open data, open code journal *GigaScience*. This aims to provide these elements for biological and biomedical researchers in the era of "big-data [28]. BGI, with extensive computational resources, has a long history of making its research outputs available to the global research community, while BioMedCental has an established open access publishing platform. Bringing these together allowed the creation of a journal that integrates data publishing via the journal's database, GigaDB, a method/workflow/analysis sharing portal based on the Galaxy workflow system ('GigaGalaxy'), and standard narrative publishing. Through this, *GigaScience* aims to finally provide infrastructure and incentives to credit and enable more reproducible research [76]. Here we outline our experiences and approaches in attempt to enable a more fit-for-purpose publishing of large-scale biological and biomedical data-heavy, and computational-intensive research.

2 The *GigaScience* approach

Integrated with the online, open access journal *GigaScience* is a database, GigaDB, that is deployed on an infrastructure provided by BGI Hong Kong, that hosts the data and software tools associated with articles published in *GigaScience* [75]. In a similar manner to Zenodo using the CERN Data Centre, the ability to leverage the tens of petabytes of storage as well as the computational and bioinformatics infrastructure already in place at BGI makes the start-up and ongoing overhead of this data publishing approach more cost effective. On top of the integrated data platform, another feature differentiating *GigaScience* from most other data journals is the in-house and on-hand team of curators, data scientists and workflow management experts to assist and broker the data and method curation, review and dissemination process.

Working with the British Library and the DataCite consortium (http://www.datacite.org), each dataset in GigaDB is assigned a Digital Object Identifier (DOI) that can be used as a standard citation in the reference section for use of these data in articles by the authors and other researchers. Using the DOI system already familiar in publishing, the process of data citation, in which the data themselves are cited and referenced in journal articles as persistently identifiable bibliographic entities, is a way to acknowledge data output [17]. Digital Curation Centres (DCC) best practice guidelines for formatting and citation [2] are carried out, and as much metadata as possible is provided to DataCite to maximize its discoverability in their repository and in the Thomson-Reuters Data Citation Index (http://thomsonreuters.com/data-citation-index/).

GigaDB further removes unenforceable legal barriers by releasing all data under the most open Creative Commons CC0 waiver that is recommended for datasets as it prevents the stacking of attribution requirements in large collections of data [30]. Taking such a broad array of data types, we have also tried to aid data interoperability and integration by taking submissions in the ISA-TAB metadata format [71]. To increase the usability further, we are also working on providing programmatic access to the data by the provision of an application programming interface (API) to GigaDB. Figure 1 illustrates the submission and peer-review workflow of *GigaScience*, GigaDB and GigaGalaxy (Fig. 1). With curation of data and checking of software availability by the in-house team, information about the study and samples is collected and collated, and checked for completeness. This information forms the basis of the DOI, and upon passing review, the data files are then transferred to the GigaDB servers from the submitter. Care is taken to ensure files are in appropriate formats and correctly linked to the relevant metadata. Finally, a DOI is minted and release of the data through the GigaDB website can occur.

3 *GigaScience* experiences

At time of writing, GigaDB has issued 220 DOIs to datasets, the largest (by volume) from "Omics" fields, including sequence-based genomics, transcriptomics, epigenomics, and metagenomics, as well as mass spectrometry-based technologies such as proteomics and metabolomics. A growing number of datasets are from imaging technologies such as MRI, CT and mass spectrometry imaging, as well as other techniques such as electrophysiology and systems biology. Smaller in size, but not in novelty are workflows, virtual machines and software platforms. Roughly 30TB of data is currently available to download from the GigaDB servers, with the largest datasets being sets of agricultural (379 cattle [80] and 3000 rice strains [81], as well as human cancer genomes [37] in the 10–15TB range. Much of the raw sequencing data has been subsequently moved to the subject-specific INSDC databases, but GigaDB hosted them during the initial curation and copying processes, and continues to host intermediate and analysed data for reproducibility purposes. Subscribing to Aspera (http://asperasoft.com) to speed up data transfers demonstrated an up to 30-fold increase in transfer speeds over FTP if users downloaded and installed the free Web browser plug-in. Most datasets are in the 1–10GB range, but many are pushing 100GB, making them impossible to host in other broad-spectrum databases. On top of datasets of relevance to human health (cancer, diabetes, hepatitis B and other human pathogens), many plant genomes of agricultural importance (sorghum, millet, potato, chickpea, cotton, flax, cucumber and wheat) are also available and released without many of the restrictions and material transfer agreements that the agricultural sector often imposes.

3.1 Positive signs in data publishing and citation

GigaScience are keen to promote the use of alternative measures for research assessment other than the impact factor. Google analytics and DataCite resolution statistics show our datasets are receiving over five times the UK DataCite average number of accesses, with some reaching over 1000 resolutions a year, and much higher levels of page views and FTP accesses.

One of the main aims of data citation is to incentivise and credit early release of data, i.e. prior to the publication of the analyses of the data (which can sometimes take years). Here, to promote this activity, GigaDB also includes a subset of datasets from BGI and their collaborators that are not associated with *GigaScience* articles, many of which were released pre-publication. For example, the Emperor and Adelie penguin, and polar bear genomes were released nearly 3 years before they were formally published in research articles. This has enabled, previously impossible, early use of large resources to the scientific community. The polar bear genome [43], for example, has accumulated a number of citations in population genetics [11] and evolutionary biology studies [55] that have benefited from its availability. The polar bear genome analysis was eventually published in May 2014 (nearly three years after its data was publicly released), and is a very encouraging example for authors concerned of "scooping" and negative consequences of early data publication and release. In addition, it highlights the benefits of Data Citation, where despite being used and cited in at least 5 analysis papers, this did not prevent it from a prestigious publication and feature on the cover of *Cell* [44]. Particularly as at that time, Cell Press was the only major biology publisher, to state in a survey carried out by F1000, that they

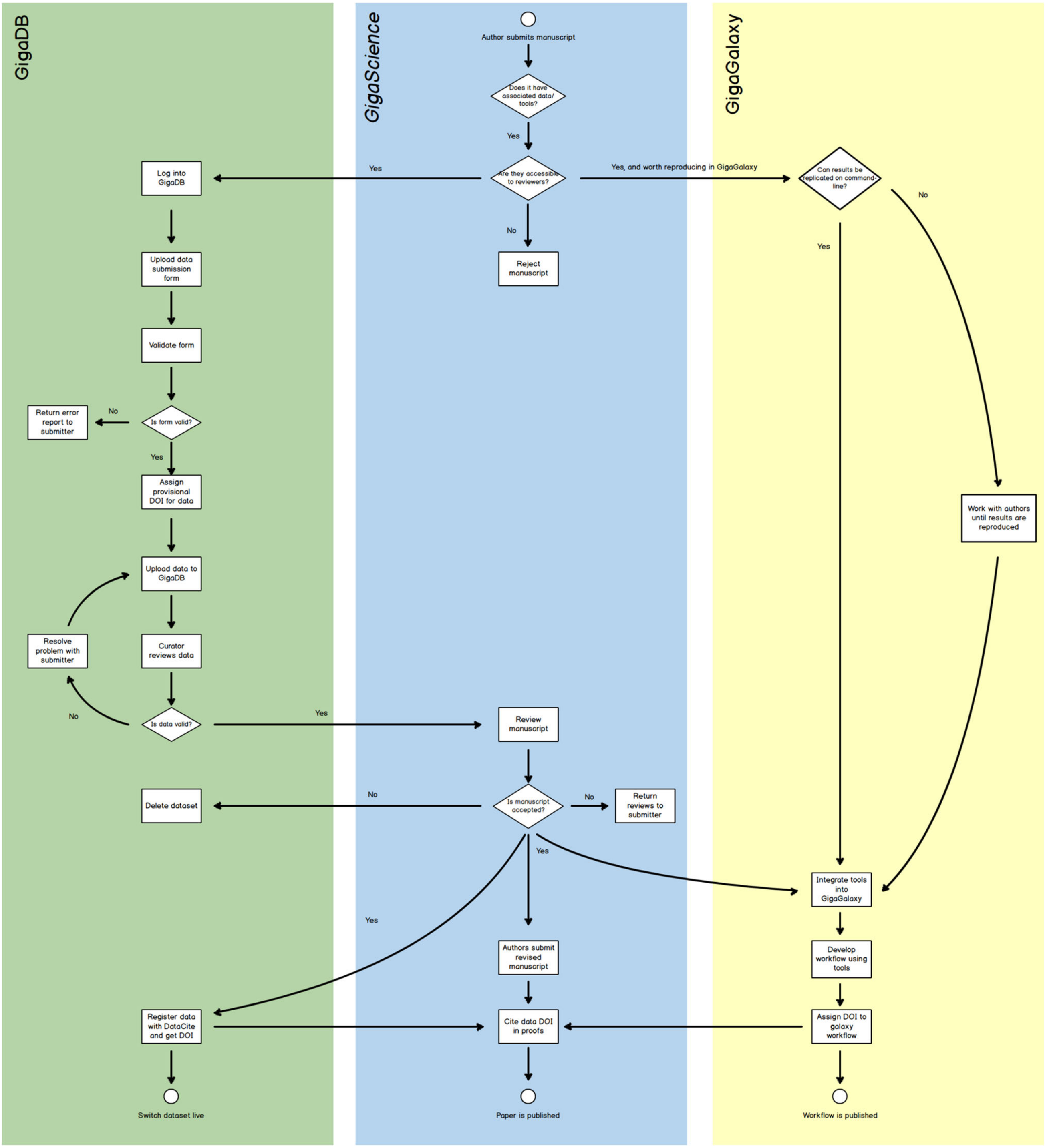
GigaDB
GigaScience
GigaGalaxy
Author submits manuscript
Does it have associated data/ tools?
Yes
Are they accessible to reviewers?
Yes
Yes, and worth reproducing in GigaGalaxy
Can results be replicated on command-line?
No
Log into GigaDB
No
Reject manuscript
Yes
Upload data submission form
Validate form
Return error report to submitter
No
Is form valid?
Yes
Assign provisional DOI for data
Work with authors until results are reproduced
Upload data to GigaDB
Resolve problem with submitter
Curator reviews data
No
Is data valid?
Yes
Review manuscript
Delete dataset
No
Is manuscript accepted?
No
Return reviews to submitter
Yes
Integrate tools into GigaGalaxy
Yes
Authors submit revised manuscript
Develop workflow using tools
Register data with DataCite and get DOI
Cite data DOI in proofs
Assign DOI to galaxy workflow
Switch dataset live
Paper is published
Workflow is published

Fig. 1 *GigaScience* publication pipelines, with curation and review of data and associated pipelines in BGI Hong Kong's ftp servers, and eventual publication in GigaDB. If computational methods and results presented in the paper are suitable to implemented in the Galaxy workflow system (and the results in the paper can also be replicated in review), these are also integrated into the paper via DOIs

would see the publication of data with a DOI as potential prior publication [18].

Most important of all are datasets that assist the fight back against disease and conservation efforts that also have educative potential, and can inspire new more open ways of doing research. What follows are few examples of the most influential and demonstrative datasets we have released, and outline of some of the downstream consequences that the data publication has set in motion.

3.2 The 2011 *E. coli* 0104:H4 outbreak genome: fighting disease outbreaks with "the tweenome"

An excellent example of the utility of pre-publication release of data was from our first dataset and DOI minted—the genome of the deadly *E. coli* 0104: H44 outbreak in Germany that killed over 50 people, infected thousands more, and caused mass panic in Europe in summer of 2011 [43]. Upon receiving DNA from the University Medical Centre Hamburg-Eppendorf, our colleagues at BGI were the first to sequence and release the genome of the pathogen responsible using "Ion Torrent" rapid bench-top sequencing technology. Due to the unusual severity of the outbreak, it was clear that the usual scientific procedure of producing data, analysing it slowly and then releasing it to the public after a potentially long peer-review procedure was inappropriate. By releasing the first genomic data into the public domain within hours of completion of the first round of sequencing, and before it had even finished uploading to the NCBI sequencing repository, the immediate announcement of its availability on twitter, promoted its use. The microbial genomics community around the world immediately took up the challenge to study the organism collaboratively (a process that was dubbed by some bloggers as the first "Tweenome").

Using these data, within the first 24 h of release, researchers from Birmingham University had released their own genome assembly in the same manner, and a group in Spain released analyses from a new annotation platform and set up a collaborative GitHub repository to provide a home to these analyses and data [58]. Within days of the initial release, a potential ancestral strain had been identified by a blogger in the US [51], helping clear Spanish farmers of the blame. Importantly for the treatment side, the many antibiotic resistance genes and pathogenic features were much more quickly and clearly understood. Additionally, releasing the data under a CC0 waiver allowed truly open-source analysis, and the UK Health Protection agency, and other contributors to the GitHub group, followed suit in releasing their work in this way. Within two weeks, two dozen reports were filed in the repository, with contributions spanning North America, Europe, Australia, Hong Kong and Egypt, including files from bloggers without formal biology training [31].

While the authors gained much good feeling and positive coverage from this (despite some inevitable disagreement over credit and what exactly was achieved [82]), they still wanted and, under current research funding assessment systems, required the more traditional form of scientific credit—publication in a prestigious journal. At the time of releasing these data, the consequences of doing so in a citeable form before the publication of associated manuscripts were unclear, especially with varying journal editorial policies regarding pre-publication dissemination of results. Particularly relevant to this is a commonly acknowledged editorial guideline of the *New England Journal of Medicine* outlining limitations on prepublication release of information known as the "Ingelfinger rule" [52]. This policy has made many researchers wary of publicizing preliminary data, as it states that a manuscript may not be considered for publication if its substance has been reported elsewhere, which can include release in press or other non-scientific outlets. In this digital era, however, this policy is looking increasingly out of touch with growing use of social media, blogging and pre-print servers, and with new funder mandates and policies regarding open access to data. It is, therefore, unclear as to how this restriction can be reconciled with various communities' code of practice regarding pre-publication data deposition in public databases.

Therefore, from a publishing perspective, the release of these data in a citeable format was a useful test case of how new and faster methods of communication and data dissemination can complement and work alongside more traditional systems of scientific communication and credit. Fortunately, the open source analysis was eventually published in the *New England Journal of Medicine*, ironically the journal responsible for the Ingelfinger rule [69]. It was positive to see that maximizing the use of the data by putting it into the public domain did not trump scientific etiquette and convention that allowed those producing the data to be attributed and get credit.

While these data and approach aided the development of rapid diagnostic methods [23] and targeted bactericidal agents to kill the pathogen [73], the project's potentially biggest legacy may be as an example of open-science, data-citation, and use of CC0 data. After releasing the data under a CC0 waiver, this allowed truly open source analysis. Furthermore, a team at the sequencing company Pacific-Biosystems quickly followed this style of sharing by releasing their data openly and speedily without wasting time on legal wrangling [12]. The lessons from this have subsequently been used to influence UK and EU science policy, with the Royal Society in the UK using the *E. coli* crowdsourcing as an example of "the power of intelligently open data", and highlighting it on the cover of their influential "Science as an Open Enterprise" report [70].

3.3 Promoting data: bird genomes lead the way

The first Data Note we published was a first-pass genome assembly covering 76 % of the genome of the rare Puerto Rican Parrot, crowdfunded with money raised from private donations, art and fashion shows in Puerto Rico [57]. The social and traditional media attention that the authors attracted from this paper enabled the authors to receive even more sponsorship, and using this has funded an improved higher coverage version of the genome, as well as comparative genomics studies [63]. The bird genomics community has embraced rapid data publishing, and following the Puerto Rican Parrot example, a number of our first datasets released were bird genomes. These datasets eventually became part of the Avian Phylogenomics Project, utilizing the genomics of modern birds to unravel how they emerged and evolved after the mass extinction that wiped out their dinosaur ancestors 66 million years previously. The decision by this community to release these data, in some cases up to 3 years prior to the main consortium publications—in 2011 for the Adelie and Emperor Penguins [91,92], the Pigeon (eventually published in 2013) [74], and Darwin Finch (released in 2012, and as yet unpublished) [59]—was a positive step demonstrating the benefits of early data release. Along with the extra early release of these first species, the remaining 42 avian genomes for this interdisciplinary, international project were released seven or eight months before the project was published. Releasing one bird genome per day on Twitter and Facebook doubled the traffic to these datasets on GigaDB over the month, and generated many retweets and positive comments from other avian researchers. In addition to releasing the data, *GigaScience* also published two Data Notes alongside the over thirty other consortium papers in *Science* and various BMC Journals. The two Data Notes presented a more narrative-style way to publish data, where the authors described the details of data production and access for all of the comparative genomics data [91] and phylogenomics data [35] from the bird species that supports all of these studies. On top of the 4TB of raw data in the SRA repository, we hosted 150GB of data from all of the assemblies in GigaDB, and many other datasets that do not have subject-specific repositories, such as the optical maps for the Budgie and Ostrich [25,95], including the thousands of files used in the phylogenomic work. Two DOIs in GigaDB [36,92] collect together all of the individual bird genome DOIs, and also provide a link to a compressed single archive file for those who wish to retrieve the complete set.

3.4 Real-time research: rapid advances in new technology through data access

Another place where fast and open dissemination of large-scale data can promote scientific advance is in areas of new technology that undergo quick changes and cannot wait to be disseminated via standard publishing venues. An example is the release of data from the Oxford Nanopore MinION™ portable single-molecule sequencer in 2014 on GigaDB. Being a new and rapidly improving technology with regular chemistry changes and updates, there was much demand for access to test data, but few platforms or repositories ready and able to handle the volumes and un-standardized formats of data produced by it.

While the first publication presenting on the data was quite negative about the quality, due to the difficulties in sharing it, there was a lack of supporting evidence provided [54]. Other groups claimed to have had more success, but there was a need and much demand to share these early data to resolve these arguments. While groups were able to share individual reads via figshare [45], the raw datasets were 10–100× larger than the size restrictions set by this platform. Working with authors at Birmingham University, we helped them release the first reference bacterial genome dataset sequenced on the MinION™ in GigaDB on September 10th 2014 [64]; and after peer review, published the Data Note article describing it just over five weeks later [65]. Being 125GB in size, this was a challenging amount of data to transfer around the world, and our curators worked with the EBI to enable their pipelines to take the raw data. But, this only became possible several weeks after the data were released.

Being the first MinION™-generated genome in the public domain, it was immediately acquired and used as test data for tools [46] and teaching materials [56]. Further, being able to rapidly review and disseminate this type of data, *GigaScience* also published the first MinION™ clinical amplicon sequencing paper and data in March 2015 [38]. The clinical applications of this tool continue to grow, with the Birmingham group recently demonstrating the utility of MinION™ sequencing via involvement in the Ebola crisis in the field in West Africa [67]. The “real-time” nature of this type of technology demonstrates that publishing needs to become more real-time to keep up.

3.5 Tools for taxonomy 2.0: sea urchins, earthworms and multi-data type species descriptions of the “cyber centipede”

The rate of species extinction is lending increasing urgency to the description of new species, but in this supposedly networked era, the process of cataloguing the rich tapestry of life is little changed since the time of Linnaeus. Fortunately, this process is being dragged into the twenty-first century, as the procedure of describing animal species finally entered the electronic era in 2012 with the acceptance of electronic taxonomy publication and registration with ZooBank, the official registry of the ICZN [32]. Concerned with growing disappearance rates, some scientists have encouraged moving to a so-called ‘turbo taxonomy’ approach, where rapid species description is needed to manage conservation [68].

A demonstrative collaboration between *GigaScience* and Pensoft Publishers has pushed the boundaries of opening up by the digital era further still, presenting an innovative approach to describing new species by creating a new kind of ‘specimen’, the ‘cybertype’ [78]. This consists of detailed and three-dimensional (3D) computer images of a specimen that can be downloaded anywhere in the world and a swathe of data types to suit modern biology, including its gene catalogue (transcriptome), DNA barcodes, and video of the live animal, in addition to the traditional morphological description. This approach has been illustrated by the description of a newly discovered cave centipede species from a remote karst region of Croatia—the 'cyber centipede' *Eupolybothrus cavernicolus*, with all of the data hosted and curated and integrated using ISA-TAB metadata in the GigaDB database [79].

This digital representation of an exemplar type specimen shows there is the potential for new forms of collections that can be openly accessed and used without the physical constraint of loaning specimens or visiting museum collections. It also means digital specimens can also be viewed alive and in three dimensions. While this new species subterranean lifestyle may protect it from some of the threats on the surface, this new type of species description also provides an example of how much previously uncharacterized information, including animal behaviour, internal structure, physiology and genetic makeup, can potentially be preserved for future generations [17].

While museum specimens can degrade, this "cybertype" specimen has the potential to be a digital message in a bottle for future generations that may not have access to the species. This publication encouraged further submissions and publications from this community, such as 141 magnetic resonance imaging scans of 98 extant sea urchin species [96], three high-resolution microCT scans of brooding brittle stars [40], as well as a coordinated publication with journal, *PLOS One* [22], publishing the nearly 40GB of microCT data supporting a paper describing, in high resolution and 3D, the morphological features commonly used in earthworm taxonomy [41]. Despite some of the folders being close to 10GB in size, the data reviewers were able to retrieve each of those in as little as half an hour using our high-speed Aspera internet connection.

3.6 Executable data: publishing software, workflows and other Research Objects

The growth in "big data" has led to scientists doing more computation, but the nature of the work has exposed limitations in our ability to evaluate published findings. One barrier is the lack of an integrated infrastructure for distributing reproducible research to others [61]. To tackle this, in addition to data, we are also hosting the materials and methods used in the data analyses reported in papers published in *GigaScience* in our repository, GigaDB. Publishing Technical Notes describing software and pipelines, all associated code is released under OSI (Open Source Initiative)-compliant licenses to allow software to be freely used, modified, and shared.

On top of archiving snapshots of code and scripts in our GigaDB servers and to allow more dynamic source code management, we also have a journal GitHub page for tools that are not in a code repository (see many examples in http://github.com/gigascience). In addition, we have developed a data reproducibility platform based on the popular Galaxy workflow system to host histories and workflows and communicate computational analyses in an interactive manner [26]. GigaGalaxy is a project prototyping the use of Galaxy to enable computational experiments to be documented and published with all computational outputs directly connected, allowing readers to inspect intermediate data and analysis steps, as well as reproduce some or all of the experiment, and modify and re-use methods. Specific analyses of data from selected papers are re-implemented as Galaxy workflows in GigaGalaxy using the procedure shown in Fig. 1, in which all of this technical reproducibility work is done in-house. Making data analyses available using the popular Galaxy platform democratises the use of many complicated computational tools. Users do not need knowledge of computer programming nor do they need to learn the implementation details of any single tool, and can run much of it off our computational resources. It also enables more visual and easy-to-understand representations of methods, an example being the test cases from our publication demonstrating a suite of Galaxy tools to study genome diversity [4]. We provide further documentation in GigaGalaxy on data analyses as well as diagrams generated by cytoscape.js to visualize how input datasets, workflows and histories are related to each example analysis. Plus for the sake of additional citability and reproducibility, the Galaxy XML files are also hosted in GigaDB [4]. Moreover, there are implemented workflows from other papers, including a case study in reproducibility from our SOAPdenovo2 paper [49] that managed to exactly recreate all of the benchmarking results listed in the paper [27].

With open source software environments such as R and Python continuing to grow in popularity, there are a number of reporting tools being integrated into it, such as Knitr and Jupyter. These tools enhance reproducible research and automated report generation by supporting execution embedded within various document formats. One example was a paper publishing a huge cache of electrophysiology data resources important for studying visual development [19]. On top of the 1GB of supporting data and code being available from GigaDB, the authors also produced the paper in a dynamic manner, creating it using R and the Knitr library. Following the reproducible research paradigm, this allows readers to see and use the code that generated each figure and table and know exactly how the results were calculated, adding confidence in the research output and allowing others to easily build upon previous work [20].

Another approach in disseminating more accessible and dynamic research outputs is through the use of virtual machines, giving reviewers and future users the ability to reproduce the experiments described in a paper, without the need to install complex, version-sensitive and interdependent prerequisite software components. With a number of submitters following this approach we have reviewed and published a number of virtual machines, one example being a paper publishing novel MRI tools and data [88]. Publishing and packaging the test data alongside tools, scripts and software required to run the experiments, this is available to download from GigaDB as a "virtual hard disk" that will specifically allow researchers to directly run the experiments themselves and to add their own annotations to the data set [89]. A related, but more lightweight approach is to use containers, such as Docker, applying virtualisation techniques to encapsulate analysis workflows to make them independent from the host it is executed on. *GigaScience* recently published its first example of a containerized analysis workflow that can be executed virtually anywhere using a Docker container of metagenomics data and pipelines [9].

Submitting code, workflows and these more dynamic research objects can be complicated, but being based at a research institute and having curators, data scientists and workflow management experts in-house enables us to help authors curate these resources if they are unable to. Leveraging the functionality that many of these open source platforms (e.g. GitHub and DockerHub) provide for cloning their contents as an archive makes it a relatively trivial task for us to take these snapshots. Referees are asked, and in most cases do carry out thorough data reviews, validation and reproducibility checks, although if they are unable to do this sufficiently rigorously, our in-house team often steps in as an additional reproducibility referee. These additional overheads are covered through combined Open Access Article and Data Publishing charges, as well as support from external funders, such as China National Genebank (a non-profit institute supported by the government of China). Being able to take advantage of the tens of petabytes of storage and computational infrastructure already in place at BGI and China National Genebank keeps the overheads low enough to provide these value-added services.

3.7 Lessons learned in reproducible data publishing

In the three and half years since the formal launch of the journal, we have published an extremely diverse range of data types, and our publishing pipelines and practices have evolved significantly in this time. Looking back at the successes and difficulties over this period, there are a number of lessons that are worth sharing. Being a journal focussing on "big data", the challenges of data volumes are the most obvious one. While the "long tail" of small, unstructured datasets is easy to handle by ourselves and others, our chosen niche focussing on the small proportion of data producers and fields generating the bulk of global research data volumes has been challenging. While demonstrating it is possible to publish a 10TB+ dataset such as the 3000 rice genomes [81], it subsequently took DNAnexus a month to download these data from our FTP servers. In the subsequent year after publication, the processed data and analyses carried out on the DNAnexus platform has now taken the total data volumes to 120TB. This has been made publically available in the cloud as an Amazon Web Services (AWS) Public Dataset, but if we were going to host this in our GigaDB server it would take one user at least one year to download (https://aws.amazon.com/public-data-sets/3000-rice-genome/), given the speed DNAnexus received the data from us. Similarly, a number of the terascale datasets we have presented have had to be shipped to us on hard disks. This method being increasingly impractical if we wanted to send the hard disks on in the same manner to downstream users. Even datasets in the 100GB range have been challenging to get hold of from less-connected corners of the world, where a microCT imaging dataset from South Africa took one month to be copied to our servers due to bandwidth problems and regular power cuts at their university, requiring the process to be restarted a number of times [40]. Popular datasets require greater bandwidth, and the *E. coli* nanopore dataset mentioned above had to be mirrored in AWS S3 for the first month to cope with the huge short-term demand [64].

On top of data volumes, reproducibility has been the other major challenge and learning experience. To carry out a case study in reproducibility we used what we hoped was one of our most scrutinized papers, the bioinformatics tool SOAPdenovo2 [49]. We subjected the publication to a number of data models including ISA-TAB, Research Object, and Nanopublications and despite managing to exactly recreate all of the results listed in the paper, it identified a small number of inaccuracies in the interpretation and discussion [27]. Due to these deficiencies being uncovered, the authors produced a correction article to officially communicate the amendment to their initial report [50]. The open, detailed and transparent approach to reviewing data and methods used by *GigaScience* has also uncovered methodological problems in companion papers published in other journals. For example, in reviewing a Data Note presenting metabolomics and lipidomics data [47] discrepancies in the previously reported methods came to light. This leads to an Erratum being published in the *J Proteome Res* explaining care should be taken when interpreting some of the results [48]. Availability and scrutiny of code is just as important as data, and our MinION™ sequenced reference bacterial genome Data Note [65] had to be corrected after an error in a script had been reported due to a fix supplied by an anonymous online contributor on Github [66]. While correcting and keeping the scientific record accurate, these approaches are expensive in time and money, and cost–benefit decisions need to be made on how much effort should expended to maintain this level of reproducibility. Notable among the examples discussed in this paper, it took about half a man-month worth of resources to reproduce the results reported in our SOAPdenovo2 paper, and cost around $1000 of AWS credits to replicate the result of our "Dockerised" metagenomics data and pipeline paper. These costs need to be balanced against the US$28 B/year wasted just on irreproducible preclinical research [24], and will be much cheaper and more effective if reproducible practices like version control and workflow management are carried out by the authors at the time the experiments are carried out rather than retrospectively by the journal. This will also make the review process much easier, and take less time of the external reviewers or in-house team at the journal. The investments and moves towards distributed computing infrastructure should also be made in distributing training in reproducible research practices. Schemes, such as Software and Data Carpentry [86] are essential to take much of the load off the peer review and publication process in ensuring the accuracy and reproducibility of the literature, and the *GigaScience* team has participated in a number of workshops, hackathons and "bring your own data parties" for these very reasons.

4 Conclusions

The papers and datasets presented provide examples of how novel mechanisms of dissemination can aid and speed up important areas of research, such as disease outbreaks, biodiversity and conservation research. Whilst trying to host all of the supporting materials and methods used in *GigaScience* papers in our GigaDB database, it is often the case that this is still not enough information to understand how the results of a scientific study were produced. A more comprehensive solution is required for users to reproduce and reuse the computational procedures described in scientific publications. This deeper and more hands-on scrutiny in the publication and review process is likely to identify more methodological inconsistencies in presented work, and as we have seen from some of our own published work initially at least, there may be increase in Errata and Corrections as a result. Rather than be seen in a negative light as airing out the "dirty laundry", journals and authors should see this as an essential part of the scientific process, and be proud to be part of the self-correcting nature of the research cycle.

Current publishing practices, which have barely changed since the seventeenth century, are particularly poorly evolved to handle this; so beyond data storage, we are looking to produce more dynamic and executable research objects. To this end, we have been developing a data reproducibility platform based on the popular Galaxy workflow system, and the first examples of this have already been published. While there is an archival version of the code hosted in GigaDB, and dynamic version linked through code repositories, such as our GitHub page, being able to visualize and execute parts of the pipelines and workflows allows a totally different perspective, allowing reviewers and future users to 'kick the wheels' and 'get under the hood' of computational methods and analyses. We are starting also to work with virtual machines and docker containers to ensure the software presented always behaves consistently. This will democratize science further, allowing users without coding experience or access to high-performance computing infrastructure to access and utilize the increasingly complicated and data-driven research that they have funded. As research data volumes continue to grow near exponentially, anecdotally demonstrated here from our publishing of growing numbers of datasets in the Giga- and Tera-scale, it is becoming increasingly technically challenging to publish and disseminate large-scale data to potential users. Alternative strategies are required, and taking the lead from industry-standard big data processing approaches used in cloud computing, we and others need to move from being "data publishers" to "compute publishers". The proof-of-concept examples presented here in publishing virtual machines and Docker containers already demonstrate the feasibility of this, and the next stage is to make this scalable and standardized. The approaches of groups, such as the Bioboxes community [7] to create standardized interfaces that make scientific software interchangeable, show a potential path towards doing this, and *GigaScience* is focussing its efforts over the coming years to be at the forefront of these efforts.

Acknowledgements The authors would like to thank BGI and China National Genebank for supporting *GigaScience* with seed funding and data, and the BBSRC and NERC for collaboration funding.

References

1. Alsheikh-Ali, A.A., Qureshi, W., Al-Mallah, M.H., Ioannidis, J.P.A.: Public availability of published research data in high-impact journals. PloS One **6**(9), e24357. doi:10.1371/journal.pone.0024357 (2011)
2. Ball, A., Duke, M.: How to cite datasets and link to publications. DCC How-to Guides. Digital Curation Centre, Edinburgh (2015). http://www.dcc.ac.uk/resources/how-guides/cite-datasets
3. Bechhofer, S., Buchan, I., De Roure, D., Missier, P., Ainsworth, J., et al.: Why linked data is not enough for scientists. Future Gener. Comput. Syst. **29**(2), 599–611 (2013)

4. Bedoya-Reina, O.C., Ratan, A., Burhans, R., Kim, H.L., Giardine, B., Riemer, C., et al.: Galaxy tools to study genome diversity. GigaSci. **2**(1), 17. doi:10.1186/2047-217X-2-17 (2013)
5. Bedoya-Reina, O., Ratan, A., Burhans, R., Kim, H., Giardine, B., Riemer, C., Miller, W.: GigaGalaxy workflows and histories from "Galaxy tools to study genome diversity". GigaScience Database. doi:10.5524/100069 (2013)
6. Begley, C.G., Ellis, L.M.: Drug development: raise standards for preclinical cancer research. Nature, **483**(7391), 531–533. doi:10.1038/483531a (2012)
7. Belmann, P., Dröge, J., Bremges, A., McHardy, A.C., Sczyrba, A., Barton, M.D.: Bioboxes: standardised containers for interchangeable bioinformatics software. Gigascience **15**(4), 47 (2015)
8. Bloom, T., Ganley, E., Winker, M.: Data access for the open access literature: PLOS's data policy. PLoS Med. **11**(2), e1001607. doi:10.1371/journal.pmed.1001607 (2014)
9. Bremges, A., Maus, I., Belmann, P., Eikmeyer, F., Winkler, A., Albersmeier, A., Pühler, A., Schlüter, A., Sczyrba, A.: Deeply sequenced metagenome and metatranscriptome of a biogas-producing microbial community from an agricultural production-scale biogas plant. Gigascience **4**, 33 (2015)
10. Buckheit, J.B., Donoho, D.L.: WaveLab and reproducible research. In Antoniadis, A., Oppenheim, G. (eds.) Wavelets and statistics, (pp. 55–81). Springer, New York (1995) (retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.53.6201)
11. Cahill, J.A., Green, R.E., Fulton, T.L., Stiller, M., Jay, F., Ovsyanikov, N., et al.: Genomic evidence for island population conversion resolves conflicting theories of polar bear evolution. PLoS Genet. **9**(3), e1003345. doi:10.1371/journal.pgen.1003345 (2013)
12. Check Hayden, E.: Open-data project aims to ease the way for genomic research. Nature (2012). doi:10.1038/nature.2012.10507
13. Clery, D.: Galaxy evolution. Galaxy zoo volunteers share pain and glory of research. Science. **333**(6039), 173–175 (2011)
14. Collberg, C., Proebsting, T., Warren, A.M.: Repeatability and Benefaction in Computer Systems Research. University of Arizona TR 14-04 (2015). http://reproducibility.cs.arizona.edu/v2/RepeatabilityTR.pdf
15. Credit where credit is overdue: Nat. Biotech. **27**(7), 579. doi:10.1038/nbt0709-579 (2009)
16. Edmunds, S.C.: The Latest Weapon in Publishing Data: the Polar Bear. GigaBlog (2014). http://blogs.biomedcentral.com/gigablog/2014/05/14/the-latest-weapon-in-publishing-data-the-polar-bear/
17. Edmunds, S.C., Hunter, C.I., Smith, V., Stoev, P., Penev, L.: Biodiversity research in the "big data" era: GigaScience and Pensoft work together to publish the most data-rich species description. GigaScience **2**(1), 14. doi:10.1186/2047-217X-2-14 (2013)
18. Edmunds, S., Pollard, T.: Adventures in data citation: sorghum genome data exemplifies the new gold standard. BMC Res. Notes **5**(1), 223 (2012). doi:10.1186/1756-0500-5-223
19. Eglen, S., Weeks, M., Jessop, M., Simonotto, J., Jackson, T., Sernagor, E.: A data repository and analysis framework for spontaneous neural activity recordings in developing retina. GigaScience **3**(1), 3. doi:10.1186/2047-217X-3-3 (2014)
20. Eglen, S., Weeks, M., Jessop, M., Simonotto, J., Jackson, T., Sernagor, E.: Supporting material for "A data repository and analysis framework for spontaneous neural activity recordings in developing retina". GigaScience Database (2014). doi:10.5524/100089
21. Fang, F.C., Casadevall, A.: Retracted science and the retraction index. Infect. Immun. **79**(10), 3855–3859 (2011). doi:10.1128/IAI.05661-11
22. Fernández, R., Kvist, S., Lenihan, J., Giribet, G., Ziegler, A.: Sine systemate chaos? A versatile tool for earthworm taxonomy: non-destructive imaging of freshly fixed and museum specimens using micro-computed tomography. PloS One **9**(5), e96617 (2014). doi:10.1371/journal.pone.0096617
23. Francis, O.E., Bendall, M., Manimaran, S., Hong, C., Clement, N.L., Castro-Nallar, E., et al.: Pathoscope: species identification and strain attribution with unassembled sequencing data. Genome Res. **23**(10), 1721–1729 (2013). doi:10.1101/gr.150151.112
24. Freedman, L.P., Cockburn, I.M., Simcoe, T.S.: The economics of reproducibility in preclinical research. PLOS Biol. **13**(6), e1002165 (2015). doi:10.1371/journal.pbio.1002165
25. Ganapathy, G., Howard, J.T., Koren, S., Phillippy, A., Zhou, S., Schwartz, D., Schatz, M., Aboukhalil, R., Ward, J.M., Li, J., Li, B., Fedrigo, O., Bukovnik, L., Wang, T., Wray, G., Rasolonjatovo, I., Winer, R., Knight, J.R., Warren, W., Zhang, G., Jarvis, E.D.: De novo high-coverage sequencing and annotated assemblies of the budgerigar genome GigaScience Database (2013). doi:10.5524/100059
26. Goecks, J., Nekrutenko, A., Taylor, J.: Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol **11**(8), R86 (2010). doi:10.1186/gb-2010-11-8-r86
27. González-Beltrán, A., Li, P., Zhao, J., Avila-Garcia, M.S., Roos, M., Thompson, M., et al.: From peer-reviewed to peer-reproduced in scholarly publishing: the complementary roles of data models and workflows in bioinformatics. PLoS One **10**(7), e0127612 (2015). doi:10.1371/journal.pone.0127612
28. Goodman, L., Edmunds, S.C., Basford, A.T.: Large and Linked in Scientific Publishing. GigaScience **1**(1), 1 (2012). doi:10.1186/2047-217X-1-1
29. Houghton, J., Gruen, N.: Open Research Data. Report to the A ustralian National Data Service (ANDS) (2014). http://apo.org.au/files/Resource/open-research-data-report.pdf
30. Hrynaszkiewicz, I., Cockerill, M.J.: Open by default: a proposed copyright license and waiver agreement for open access research and data in peer-reviewed journals. BMC Res. Notes **5**(1), 494 (2012). doi:10.1186/1756-0500-5-494
31. Huang, B.: Reverse Engineering Superbugs « Bunnie's blog (2011). http://www.bunniestudios.com/blog/?p=1676
32. ICZN: Amendment of Articles 8, 9, 10, 21 and 78 of the International Code of Zoological Nomenclature to expand and refine methods of publication. ZooKeys **219**, 1–10 (2012). doi:10.3897/zookeys.219.3944
33. Ioannidis, J.P.A.: How to Make More Published Research True. PLoS Med. **11**(10), e1001747 (2014). doi:10.1371/journal.pmed.1001747
34. Ioannidis, J.P.A., Allison, D.B., Ball, C.A., Coulibaly, I., Cui, X., Culhane, A.C., et al.: Repeatability of published microarray gene expression analyses. Nat. Genet. **41**(2), 149–55 (2009). doi:10.1038/ng.295
35. Jarvis, E.D., Mirarab, S., Aberer, A.J., Li, B., Houde, P., Li, C., et al.: Phylogenomic analyses data of the avian phylogenomics project. GigaScience **4**(1), 4 (2015). doi:10.1186/s13742-014-0038-1
36. Jarvis, E., Mirarab, S., Aberer, A., Houde, P., Li, C., Ho, S., Zhang, G., et al.: Phylogenomic analyses data of the avian phylogenomics project. GigaScience Database (2014). doi:10.5524/101041
37. Kan, Z., Zheng, H., Liu, X., Li, S., Barber, T., Gong, Z., et al.: Hepatocellular carcinoma genomic data from the Asian Cancer Research Group. GigaScience (2012). doi:10.5524/100034
38. Kilianski, A., Haas, J.L., Corriveau, E.J., Liem, A.T., Willis, K.L., Kadavy, D.R., et al.: Bacterial and viral identification and differentiation by amplicon sequencing on the MinION nanopore sequencer. GigaScience **4**(1), 12 (2015). doi:10.1186/s13742-015-0051-z
39. Krawczyk, M., Reuben, E.: (Un)Available upon Request: Field Experiment on Researchers' Willingness to Share Supplementary Materials. Account. Res. (2012). doi:10.1080/08989621.2012.678688

40. Landschoff, J., Du Plessis, A., Griffiths, C.L.: A dataset describing brooding in three species of South African brittle stars, comprising seven high-resolution, micro X-ray computed tomography scans. Gigascience **4**, 52 (2015). doi:10.1186/s13742-015-0093-2
41. Lenihan, J., Kvist, S., Fernández, R., Giribet, G., Ziegler, A.: A dataset comprising four micro-computed tomography scans of freshly fixed and museum earthworm specimens. GigaScience **3**(1), 6 (2014). doi:10.1186/2047-217X-3-6
42. Li, B., Zhang, G., Willersleve, E., Wang, J., Wang, J.: Genomic data from the polar bear (Ursus maritimus). GigaScience (2011). doi:10.5524/100008
43. Li, D., Xi, F., Zhao, M., Chen, W., Cao, S., Xu, R.,. Consortium, T. E. coli O. T.-2482 isolate genome sequencing: Genomic data from Escherichia coli O104:H4 isolate TY-2482. BGI Shenzhen (2011). doi:10.5524/100001
44. Liu, S., Lorenzen, E.D., Fumagalli, M., Li, B., Harris, K., Xiong, Z., et al.: Population genomics reveal recent speciation and rapid evolutionary adaptation in polar bears. Cell **157**(4), 785–794 (2014). doi:10.1016/j.cell.2014.03.054
45. Loman, N., Quick, J., Calus, S.: A P. aeruginosa serotype-defining single read from our first Oxford Nanopore run (2014). doi:10.6084/m9.figshare.1052996
46. Loose, M.: minoTour (2014). http://minotour.nottingham.ac.uk/
47. Luan, H., Meng, N., Liu, P., Fu, J., Chen, X., Rao, W., Jiang, H., Xu, X., Cai, Z., Wang, J.: Non-targeted metabolomics and lipidomics LC–MS data from maternal plasma of 180 healthy pregnant women. Gigascience **4**, 16 (2015a). doi:10.1186/s13742-015-0054-9
48. Luan, H., Meng, N., Liu, P., Feng, Q., Lin, S., Fu, J., Davidson, R., Chen, X., Rao, W., Chen, F., Jiang, H., Xu, X., Cai, Z., Wang, J.: Correction to "Pregnancy-Induced Metabolic Phenotype Variations in Maternal Plasma". J. Proteome Res. **14**(7), 3005 (2015). doi:10.1021/acs.jproteome.5b00430
49. Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., et al.: SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience **1**(1), 18 (2012). doi:10.1186/2047-217X-1-18
50. Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., Wang, J.: Erratum: SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience **4**, 30 (2015). doi:10.1186/s13742-015-0069-2
51. Mike the Mad Biologist: I Don't Think the German Outbreak E. coli Strain Is Novel: Something Very Similar Was Isolated Ten Years Ago.... (2011). http://mikethemadbiologist.com/2011/06/03/i_dont_think_the_german_e_coli/
52. Marshall, E.: Franz Ingelfinger's Legacy Shaped Biology Publishing. Science **282**(5390), 861 (1998). doi:10.1126/science.282.5390.861
53. Mervis, J.: Agencies rally to tackle big data. Science **336**(6077), 22–22 (2012). doi:10.1126/science.336.6077.22
54. Mikheyev, A.S., Tin, M.M.Y.: A first look at the Oxford Nanopore MinION sequencer. Mol. Ecol. Res. **14**(6) (2014). doi:10.1111/1755-0998.12324
55. Morgan, C.C., Foster, P.G., Webb, A.E., Pisani, D., McInerney, J.O., O'Connell, M.J.: Heterogeneous models place the root of the placental mammal phylogeny. Mol. Biol. Evol. **30**(9), 2145–2156 (2013). doi:10.1093/molbev/mst117
56. Nederbragt, A.J.: INF-BIO5121/9121 fall 2014 de novo assembly. GitHub (2014). http://github.com/lexnederbragt/INF-BIOx121_fall2014_de_novo_assembly
57. Oleksyk, T.K., Pombert, J.-F., Siu, D., Mazo-Vargas, A., Ramos, B., Guiblet, W., et al.: A locally funded Puerto Rican parrot (Amazona vittata) genome sequencing project increases avian data and advances young researcher education. GigaScience **1**(1), 14 (2012). doi:10.1186/2047-217X-1-14
58. Pareja-Tobes, P., Manrique, M., Pareja-Tobes, E., Pareja, E., Tobes, R.: BG7: a new approach for bacterial genome annotation designed for next generation sequencing data. PloS One **7**(11), e49239 (2012). doi:10.1371/journal.pone.0049239
59. Parker, P., Li, B., Li, H., Wang, J.: The genome of Darwin's Finch (Geospiza fortis). GigaScience (2012). doi:10.5524/100040
60. Patrinos, G.P., Al Aama, J., Al Aqeel, A., Al-Mulla, F., Borg, J., Devereux, A., et al.: Recommendations for genetic variation data capture in developing countries to ensure a comprehensive worldwide data collection. Human Mutat. **32**(1), 2–9 (2011). doi:10.1002/humu.21397
61. Peng, R.D.: Reproducible research in computational science. Sci (N.Y.) **334**(6060), 1226–1227 (2011). doi:10.1126/science.1213847
62. Piwowar, H.A., Vision, T.J.: Data reuse and the open data citation advantage. PeerJ **1**, e175 (2013). doi:10.7717/peerj.175
63. Proffitt, A.: The People's Parrot: the First Community-Sponsored Genome. *Bio-IT World* (2011). http://www.bio-itworld.com/2012/09/28/peoples-parrot-first-community-sponsored-genome.html
64. Quick, J., Loman, N.: Bacterial whole-genome read data from the Oxford Nanopore Technologies MinION^{TM} nanopore sequencer. GigaScience Database (2014). doi:10.5524/100102
65. Quick, J., Quinlan, A.R., Loman, N.J.: A reference bacterial genome dataset generated on the MinION(TM) portable single-molecule nanopore sequencer. GigaScience **3**(1), 22 (2014). doi:10.1186/2047-217X-3-22
66. Quick, J., Quinlan, A.R., Loman, N.J.: Erratum: a reference bacterial genome dataset generated on the MinION(TM) portable single-molecule nanopore sequencer. Gigascience **4**, 6 (2015). doi:10.1186/s13742-015-0043-z
67. Quick, J., Loman, N.J., Duraffour, S., Simpson, J.T., Severi, E., Cowley, L., Carroll M.W.: Real-time, portable genome sequencing for Ebola surveillance. Nature. **530**(7589), 228–232 (2016). doi:10.1038/nature16996
68. Riedel, A., Sagata, K., Suhardjono, Y.R., Tänzler, R., Balke, M.: Integrative taxonomy on the fast track—towards more sustainability in biodiversity research. Frontiers Zool. **10**(1), 15 (2013). doi:10.1186/1742-9994-10-15
69. Rohde, H., Qin, J., Cui, Y., Li, D., Loman, N.J., Hentschke, M., et al.: Open-Source Genomic Analysis of Shiga-Toxin–Producing *E. coli* O104:H4. N. Engl. J. Med. **365**, 718–724 (2011). doi:10.1056/NEJMoa1107643
70. Royal Society: Science as an open enterprise (2012). http://royalsociety.org/policy/projects/science-public-enterprise/report/
71. Sansone, S.A., Rocca-Serra, P., Field, D., Maguire, E., Taylor, C., Hofmann, O., et al.: Toward interoperable bioscience data. Nat. Genet. **44**(2), 121–126 (2012). doi:10.1038/ng.1054
72. Savage, C.J., Vickers, A.J.: Empirical study of data sharing by authors publishing in PLoS journals. PloS One **4**(9), e7078 (2009). doi:10.1371/journal.pone.0007078
73. Scholl, D., Gebhart, D., Williams, S.R., Bates, A., Mandrell, R.: Genome sequence of *E. coli* O104:H4 leads to rapid development of a targeted antimicrobial agent against this emerging pathogen. PLoS One **7**(3), e33637 (2012). doi:10.1371/journal.pone.0033637
74. Shapiro, M.D., Kronenberg, Z., Li, C., Domyan, E.T., Pan, H., Campbell, M., et al.: Genomic diversity and evolution of the head crest in the rock pigeon. Sci. (N.Y.) **339**(6123), 1063–1067 (2013). doi:10.1126/science.1230422
75. Sneddon, T.P., Li, P., Edmunds, S.C.: GigaDB: announcing the GigaScience database. GigaScience **1**(1), 11 (2012). doi:10.1186/2047-217X-1-11
76. Sneddon, T.P., Si Zhe, X., Edmunds, S.C., Li, P., Goodman, L., Hunter, C.I.: GigaDB: promoting data dissemination and reproducibility. Database **2014**(0), bau018–bau018 (2014). doi:10.1093/database/bau018

77. Stodden, V., Guo, P., Ma, Z.: Toward reproducible computational research: an empirical analysis of data and code policy adoption by Journals. PloS One **8**(6), e67111 (2013). doi:10.1371/journal.pone.0067111
78. Stoev, P., Komerički, A., Akkari, N., Liu, S., Zhou, X., Weigand, A.M., et al.: Eupolybothrus cavernicolus Komerički & Stoev sp. n. (Chilopoda: Lithobiomorpha: Lithobiidae): the first eukaryotic species description combining transcriptomic, DNA barcoding and micro-CT imaging data. Biodivers. Data J. **1**, e1013 (2013). doi:10.3897/BDJ.1.e1013
79. Stoev, P., Komerički, A., Akkari, N., Liu, S., Zhou, X., Weigand, A., Penev, L.: Transcriptomic, DNA barcoding, and micro-CT imaging data from an advanced taxonomic description of a novel centipede species (Eupolybothrus cavernicolus Komerički Stoev, sp n). GigaScience (2013). doi:10.5524/100063
80. Stothard, P., Liao, X., Arantes, A.S., Pauw, M.D., Coros, C., Plastow, G.S., Sargolzaei, M., Crowley, J.J., Basarab, J.A., Schenkel, F., Moore, S., Miller, S.P.: Bovine whole-genome sequence alignments from the Canadian Cattle Genome Project GigaScience Database (2015). doi:10.5524/100157
81. The 3000 Rice Genomes Project: The Rice 3000 Genomes Project Data. GigaScience Database (2014). doi:10.5524/200001
82. Turner, M.: *E. coli* outbreak strain in genome race. Nature (2011). doi:10.1038/news.2011.430
83. Van Noorden, R.: Half of 2011 papers now free to read. Nature **500**(7463), 386–7 (2013). doi:10.1038/500386a
84. Van Noorden, R.: Sluggish data sharing hampers reproducibility effort. Nature (2015). doi:10.1038/nature.2015.17694
85. Wilkinson, M., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, Arie, B. et al.: The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data. **3**, 160018 (2016). doi:10.1038/sdata.2016.18
86. Wilson G.: Software Carpentry: lessons learned [version 2; referees: 3 approved]. F1000Research **3**, 62 (2016). doi:10.12688/f1000research.3-62.v2
87. Whitlock, M.C.: Data archiving in ecology and evolution: best practices. Trends Ecol. Evol. **26**(2), 61–65 (2011). http://www.sciencedirect.com/science/article/pii/S0169534710002697
88. Wollny, G., Kellman, P.: Free breathing myocardial perfusion data sets for performance analysis of motion compensation algorithms. GigaScience **3**(1), 23 (2014). doi:10.1186/2047-217X-3-23
89. Wollny, G., Kellman, P.: Supporting material for: "Free breathingly acquired myocardial perfusion data sets for performance analysis of motion compensation algorithms". GigaScience Database (2014). doi:10.5524/100106
90. Yang, H.: Support the Manchester Manifesto?: a case study of the free sharing of human genome data. Prometheus **29**(3), 337–341 (2011). doi:10.1080/08109028.2011.631275
91. Zhang, G., Lambert, D.M; Wang, J.: Genomic data from Adelie penguin (Pygoscelis adeliae). GigaScience (2011a). doi:10.5524/100006
92. Zhang, G., Lambert, D.M, Wang, J.: Genomic data from the Emperor penguin (Aptenodytes forsteri). GigaScience (2011b). doi:10.5524/100005
93. Zhang, G., Li, B., Li, C., Gilbert, M.T.P., Jarvis, E.D., Wang, J.: Comparative genomic data of the Avian Phylogenomics Project. GigaScience **3**(1), 26 (2014a). doi:10.1186/2047-217X-3-26
94. Zhang, G., Li, B., Li, C., Gilbert, M., Jarvis, E., & The Avian Genome Consortium Wang, J: The avian phylogenomic project data. GigaScience Database (2014b). doi:10.5524/101000
95. Zhang, G; Li, B; Li, C; Gilbert, M, P; Ryder, O; Jarvis, E, D; The Avian Genome Consortium,; Wang, J: Genomic data of the Ostrich (Struthio camelus australis). GigaScience Database (2014c). doi:10.5524/101013
96. Ziegler, A., Faber, C., Mueller, S., Nagelmann, N., Schröder, L.: A dataset comprising 141 magnetic resonance imaging scans of 98 extant sea urchin species. GigaScience **3**(1), 21 (2014). doi:10.1186/2047-217X-3-21