DATABASE
The Journal of Biological Databases and Curation

Database, Vol. 2014, Article ID bau018, doi:10.1093/database/bau018

Original article

GigaDB: promoting data dissemination and reproducibility

Tam P. Sneddon[1], Xiao Si Zhe[2], Scott C. Edmunds[2], Peter Li[2], Laurie Goodman[2] and Christopher I. Hunter[2,*]

[1]Department of Genetics, Stanford University, USA, [2]GigaScience team, BGI HK Research Institute, 16 Dai Fu Street, Tai Po Industrial Estate, Hong Kong

***Corresponding author:** Tel: +85 2 36 10 35 32; Email: chris@gigasciencejournal.com

Submitted 13 November 2013; Revised 4 February 2014; Accepted 6 February 2014

Citation details: Sneddon,T.P., Zhe,X.S., Edmunds,S.C., *et al*. GigaDB: promoting data dissemination and reproducibility. *Database* (2014) Vol. 2014: article ID bau018; doi:10.1093/database/bau018.

Often papers are published where the underlying data supporting the research are not made available because of the limitations of making such large data sets publicly and permanently accessible. Even if the raw data are deposited in public archives, the essential analysis intermediaries, scripts or software are frequently not made available, meaning the science is not reproducible. The *GigaScience* journal is attempting to address this issue with the associated data storage and dissemination portal, the *GigaScience* database (GigaDB). Here we present the current version of GigaDB and reveal plans for the next generation of improvements. However, most importantly, we are soliciting responses from you, the users, to ensure that future developments are focused on the data storage and dissemination issues that still need resolving.

Database URL: http://www.gigadb.org

Introduction

The life sciences fields are all embracing the recent advances in technology, whether they are the next-generation sequencing in genetics/genomics (1), high-throughput mass spectrometry in proteomics (2)/metabolomics (3) or the integration of multiple data types in a single analysis (4, 5), with a steady increase in the number of published peer-reviewed articles to match. The accepted ideal of scientific publishing is that researchers provide the data from which they have drawn their conclusions to enable peers and reviewers to retest those data to validate the claims. However, all too often, papers are published where the underlying data are not made available because of factors such as a lack of community repositories or size, meaning there is no opportunity for independent validation of findings. Sometimes the raw data are deposited in archives, but the essential analysis intermediaries, scripts or software are not available, meaning the results of a study are not reproducible by others.

Good science not only finds solutions but also generates more questions. Therefore, the challenge for publishers is not only to provide the resources for scientists to read and validate the results but also to allow them to interrogate the data with their own questions. Despite the majority of high-impact journals setting out editorial policies relating to the public availability and sharing of data, one study showed that <10% of papers deposited full raw data (6).

In addition, with the rapid development of technology comes the *ad hoc* development of many, often untested, software solutions for data analysis. Unfortunately, code availability policies are even more poorly adhered to than data release policies (7). For correct validation of results, the software tools used must also be available and the methods of a paper should include the software versions and parameters used in every step. Some may even argue that hardware can affect analysis, and therefore advocate the creation and use of virtual machines to ensure software always behave consistently. Aiming to incentivize and address this reproducibility gap, here we discuss the current

Database, Vol. 2014, Article ID bau018, doi:10.1093/database/bau018

and future state of the *GigaScience* database (GigaDB), covering the rationale and current data publication ecosystem, the relationship and integration with its companion *GigaScience* journal, the types of data held, the use of standards, citation of data by digital object identifier (DOI) and the planned additions and improvements to the database.

Publishing data

There is a growing awareness of a data reproducibility and access gap, and some funders such as the NSF are starting to mandate data management and sharing plans. There has been a need for infrastructure and mechanisms to facilitate these mandates, as well as calls for better incentives and credit to make the time and effort to do this worthwhile (8). There are a growing number of databases purporting to 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, and it is important to discern where GigaDB fits among these. The ecosystem of biorepositories includes domain-specific databases such as those of the International Nucleotide Sequence Database Consortium (INSDC; GenBank, ENA and DDBJ), as well as the variety of protein and peptide repositories (e.g. PRIDE, AE, GEO), metabolomics (MetaboLights) and imaging repositories (e.g. Morphosource, MorphoBank). There is also now a selection of broad-spectrum databases including GigaDB, Zenodo, FigShare and Dryad. The benefits of the broad-spectrum databases are that they are not restricted by the data types they can hold, meaning 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 high-throughput sequencing, mass-spectrometry and imaging data are not as well served by their file size limitations and charges. GigaDB is able to leverage the tens of petabytes of storage as well as the computational and bioinformatics infrastructure already in place at its host institution, BGI, much cheaper than that of any other publisher, and in a similar manner to Zenodo using the CERN Data Centre.

Although the greater than Moore's Law growth in the amount of sequencing data produced is well publicized, areas of research such as high-throughput phenotyping and functional genomics screens, brain mapping and super-resolution microscopy are also producing increasingly large volumes of data. Furthermore, with an increasing number of integrative 'multi-omics' studies combining genomic, transcriptomic, proteomic and metabolomics data, there is an increasing need for infrastructure to help curate, integrate and present these types of large-scale biological data-oriented studies. The use of cloud-based storage such as Amazon S3, or platforms using this technology, solves many of these issues, but there are cost and stability issues for long-term storage (9), and they are generally not tailored for biological data and metadata.

Journal: GigaScience

GigaScience (10) is an online open-access journal that includes GigaDB as a part of its publishing activities (11). *GigaScience* is co-published in collaboration between the BGI and BioMed Central, the world's largest genomics organization and first commercial open-access publisher, respectively, to cater for biological and biomedical researchers in the era of 'big-data'. The journal's scope covers studies from the entire spectrum of the life sciences that produce and use large-scale data as the center of their work. Studies have shown that citation of work is greatly improved by the availability of associated data (12). Data from *GigaScience* articles are hosted in GigaDB, from where they can be cited to provide a direct link between the study and the data supporting it. The journal also provides a forum for discussions surrounding best practices and issues in handling large-scale data. See http://www.gigasciencejournal.com for additional information about the journal.

Database: GigaDB

GigaDB primarily serves as a repository to host data and software tools associated with articles in *GigaScience*. It also includes a subset of data sets that is not associated with *GigaScience* articles from the funding organization, BGI, enabling them to release their data quicker, and allowing them to be cited and receive credit for its release before publication in journal articles. With support from China National Genebank, a non-profit institute supported by the government of China and operated by BGI that has database infrastructure as a part of its remit, GigaDB has leveraged and taken advantage of the tens of petabytes of storage and computational infrastructure already in place at the BGI to handle and present large-scale biological data sets much cheaper and easier than other similar resources. A good example of the utility of prepublication release of data was from our first data set, the genome of the deadly 2011 *Escherichia coli* outbreak in Germany (13). Its immediate release into the public domain allowed bioinformaticians around the world to perform rapid, publicly available and openly discussed analyses. This provided crucial data in the fight against the pathogen for 2 months before the 'open-source genomic' effort was eventually published in the *New England Journal of Medicine* (14). The rapid release of preliminary data to the research community while enabling their producers to obtain credit

through citation is a growing trend with the BGI, and the wider research community, where the speed of data production far outstrips the ability of researchers to analyze and write up their findings.

GigaDB defines a data set as a group of files (e.g. sequencing data, analyses, imaging files, software programs) that are related to and support an article or study. To maximize their utility to the research community, all data sets in GigaDB are placed under a Creative Commons CC0 waiver. This is increasingly accepted as the most appropriate mechanism for dedicating data to the public domain, as it eliminates legal impediments to integration and reuse of large collections of data, such as attribution stacking (15). Although the data are released without legal restriction, scientific etiquette is built on attribution.

Author attribution and credit

Through association with DataCite (www.datacite.org), each data set in GigaDB is assigned a DOI that can be used as a standard citation for future use of these data in other articles by the authors and other researchers. All data sets in GigaDB require a title that is specific to the data set, an author list and an abstract that provides information specific to the data included within the set. As much metadata as possible is then provided to DataCite to maximize its discoverability in their repository and in the Thomson-Reuters Data Citation Index (http://thomsonreuters.com/data-citation-index/). Following the Digital Curation Centre's (DCC) best practice formatting and citation guidelines (16), GigaDB data sets are integrated into *GigaScience* through citation in the references, and we have also worked closely with other publishers to ensure that the use of GigaDB data is correctly attributed and cited in their journals as well (17). Citation guidelines are outlined clearly on each data set entry, and functionality to export to reference managers as BibTex files is also included.

In addition to providing mechanisms for citation and reuse, it is important to measure statistics to provide insights into the use of content and further incentivize deposition. DataCite already supplies resolution statistics that are publicly available, and although GigaDB hosts only a small number of data sets at present, the numbers show that GigaDB is receiving over the DataCite average number of hits (http://stats.datacite.org/). GigaDB is keen to help promote and provide alternative metrics for scholarly impact to the journal impact factor, so alternative measures and mechanisms for dissemination and credit are also used, with social media integration and functionality included. The ability to share to Facebook, Twitter and Google+ is provided, and statistics are included at the bottom of every entry. An RSS feed of the latest data sets is also provided.

Data types

The scope covers not only 'omics' type data and the fields of high-throughput biology currently serviced by large public repositories but also the growing range of more difficult-to-access data, such as imaging, neuroscience, ecology, cohort data, systems biology and other new types of large-scale sharable data, as well as software used to analyze large-scale data sets. By archiving software packages and data analyses as executable scripts, for example, all of the pipelines used in the publication of the SOAPdenovo2 genome assembler (18), this incentivizes reproducibility of software papers. By giving DOIs to software, their citeability and longevity is improved while also strengthening the links between publications and code that may not be possible from code repositories.

At the time of writing, GigaDB has issued 64 DOIs to data sets including data from genomic, transcriptomic, metagenomic, epigenetic, proteomic, mass spectrometry, imaging, workflows and software platforms. A total of ~20 TB of data are currently available to download from the GigaDB servers, with the largest single data set comprising ~15 TB (19).

It should be noted that GigaDB will only host data that can be released without restriction, and will not host data that are restricted for any reason, be that ethical or otherwise.

Standards and interoperability

To encourage scientists to submit their data, it is important to make the perceived barrier to data release as low as possible. One way in which GigaDB is doing this already is to provide a multitude of options for submitters, while still ensuring data meets community standards, for example, encourage compliance with the Genome Standard Consortiums (GSC) Minimum Information about any Sequence (MIxS) standard (20). To assist authors in this respect, GigaDB provides a basic spreadsheet template for users to complete and upload in either comma separated value (CSV) or Excel formats. In addition, they are able to accept ISA-Tab (21) formatted files. ISA-Tab is a format used by the BioSharing and ISA Commons communities and is widely used by a growing number of resources and databases such as the Metabolights database (22), Harvard Stem Cell Discovery Engine (23) and Nature's upcoming *Scientific Data* journal (http://www.nature.com/scientificdata/).

So that these different submission formats do not fragment the data, the information is imported from them into a relational database to be stored in a uniform structure for searching and retrieval. In future releases, tools will be

Database, Vol. 2014, Article ID bau018, doi:10.1093/database/bau018

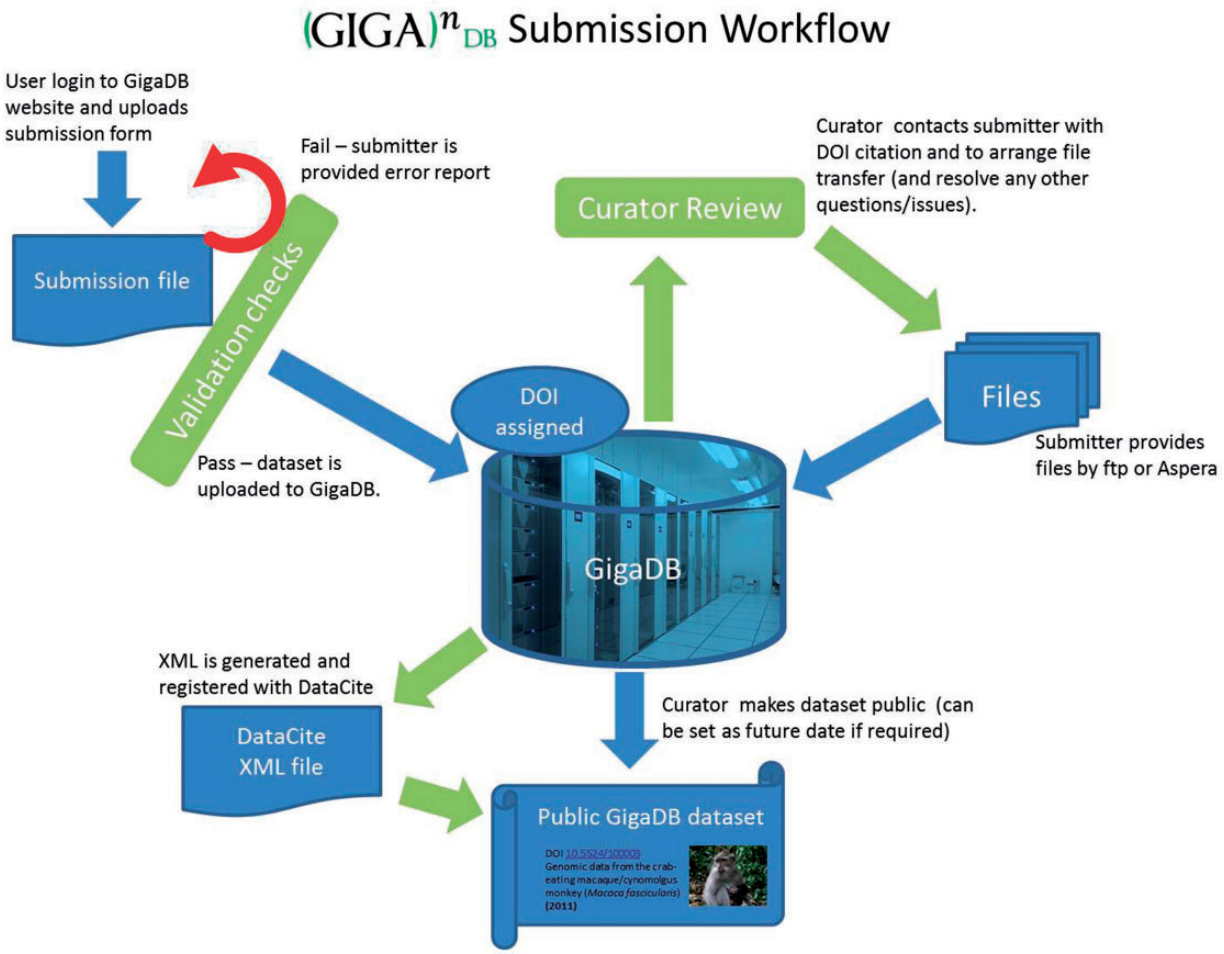

Figure 1. A simplified workflow of data through GigaDB from submission to presentation.

provided to export from the database in a variety of formats via an Application Programming Interface (API) or simple download button on each data set page, thus allowing users to acquire the data set metadata in the format most useful to them.

Figure 1 illustrates the submission workflow, with curation of the basic information before and after the archiving of the data. Initially, information is collected and collated about the study and samples, and checked for completeness. This information forms the basis of the DOI, and the data files (sequence, MS/MS, annotations, images, etc.) are then transferred to the GigaDB servers from the submitter. Care is taken to ensure these files are in appropriate formats and are correctly linked to the relevant metadata. Finally, a DOI is minted through DataCite, and release of the data through the GigaDB Web site can occur.

Because of the size of files being moved around, GigaDB makes use of the Aspera software (http://asperasoft.com) to speed up the transfers. In our experience, up to a 30-fold increase in transfer speeds over FTP has been observed, which means that 2 GB of data can be moved in <1 minute instead of 30 minutes on a 1-MBps network connection. A free Web browser plug-in from the developers of Aspera has to be downloaded and installed for users to transfer data to and from GigaDB. However, we are actively looking at alternative open-source methods of high-speed data transfer to keep with the ethos of open data.

The 'cyber-centipede'

A good example of the utility of GigaDB can be seen from recent collaboration between *GigaScience*, *China National GeneBank*, *BGI-Shenzhen* and *Pensoft Publishers*, which resulted in the publication of the first eukaryotic species description combining transcriptomic, DNA barcoding, morphology and X-ray microtomography imaging data. This 'holistic' approach in taxonomic description of a new species of cave-dwelling centipede was published in the Biodiversity Data Journal (24), with coordinated data release in GigaDB (25). It demonstrates how one can use the Internet to move accessibility of typed species out of museum collections and species description out of the

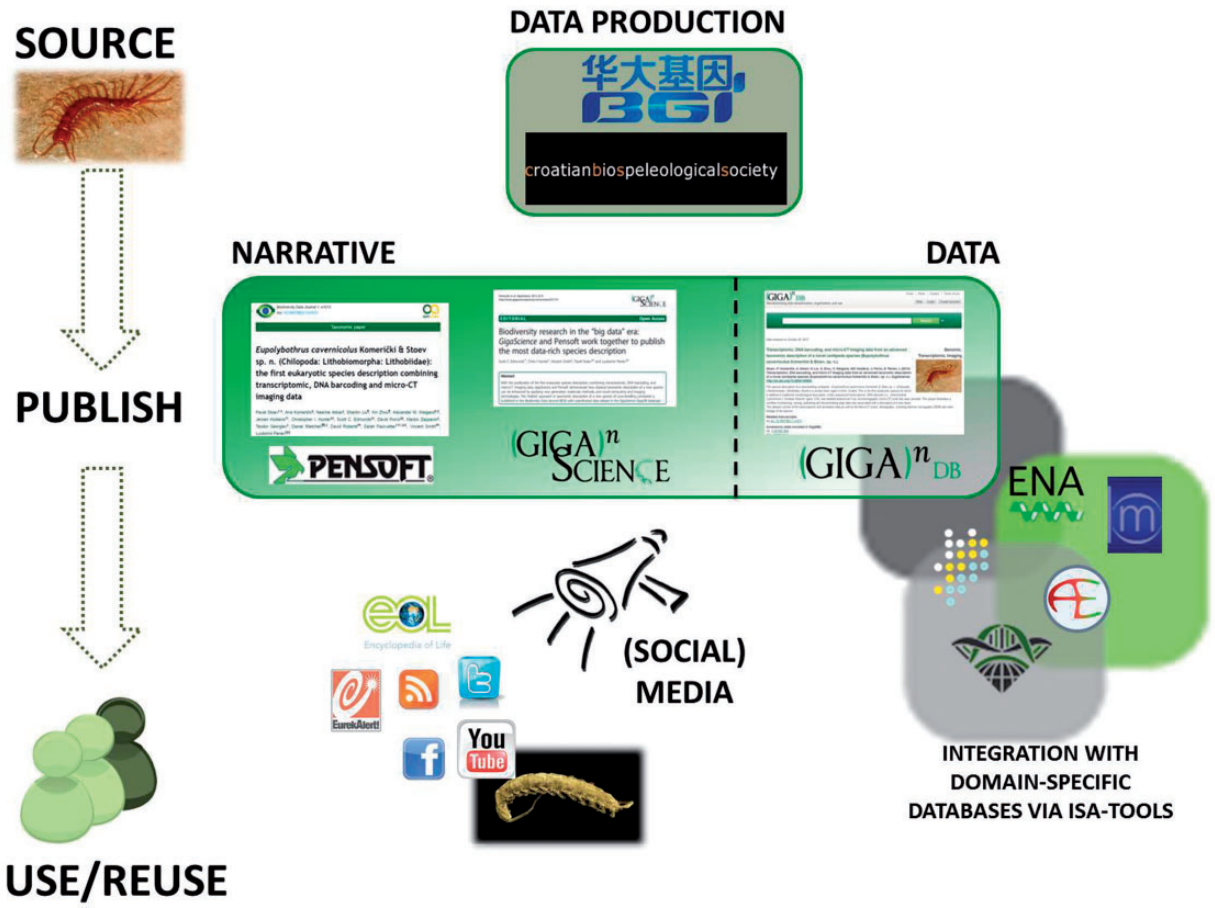

Figure 2. The integrated approach to data dissemination and attribution, using the example of the cyber-centipede data-rich species description.

print era. This attempt at a data-rich 'cybertype' mixed traditional morphological description, a transcriptomic profile, a 3D X-ray view and a movie of the living specimen to document important traits of its behavior. Figure 2 shows the extent and variety of data and metadata associated with this publication, all of which are collected together in, or linked to from, GigaDB. It also illustrates how multiple different media can be successfully integrated using the power of the ISA-tab metadata format. This study demonstrates how classical taxonomic description of a new species can be enhanced by applying new-generation molecular methods and novel computing and imaging technologies.

Searching for data sets

The recently released new GigaDB Web site includes an improved full-text search facility powered by the open-source software package, Sphinx (http://sphinxsearch.com/docs/current.html), which allows users to easily find data sets. Registered users can also save searches and get automatic e-mail notifications about newly released data sets matching their criteria.

In addition to the keyword search, one can further filter the results on a number of fields to narrow the results to those most relevant. For example, if a search is performed for the keyword 'cancer', six data sets are currently returned containing 7852 files. These data sets can be further filtered to find human cancer data sets by using the common name 'human' to reduce the returned results to two data sets containing 5000 files. If registered and logged-in, users can save the search and filter criteria to their own GigaDB user area and be automatically notified of any new data sets matching those criteria.

Future directions

GigaDB is currently operational and attracting an increasing number of users, but we are keen to ensure that we increase our usefulness to the research community by the addition of multiple new features over the coming months. Here, we highlight some of those features.

Database, Vol. 2014, Article ID bau018, doi:10.1093/database/bau018

Although we try to host the materials and methods used in the data analyses, which are reported in *GigaScience* papers in GigaDB, it is often the case that this does not provide sufficient information to fully understand how the results of a scientific study were produced. A more comprehensive solution is required for our users to reproduce and reuse the computational procedures described in *GigaScience*. To this end, we are currently developing a data reproducibility platform, whereby data analyses are provided or re-implemented as Galaxy workflows (26). The first examples of this have already been published with testing of the *de novo* genome assembly tool, SOAPdenovo2 (27), and a population genomics toolkit (28). In the future, we hope to enable this Galaxy platform to directly access data stored in GigaDB through its API, which is currently in development. This API will also allow other computational tools to directly access the information held in our database.

In addition to the currently available methods of submission, we are developing an online wizard to allow our users an interactive online-guided submission process, which we hope will further encourage deposition of all available metadata by reducing the burden and suggesting the inclusion of data that may not have been considered relevant by the submitter.

Also in development is the implementation of a citation tracker to enable us to supply direct links to citations of GigaDB data sets, which will allow authors to see who, where and when their data sets are being reused. On top of the statistics that DataCite provides, we currently use Google Analytics internally to monitor traffic on the GigaDB Web site, which we hope will enable us to tailor future developments to areas of the Web site that lose traffic, but these data could also provide alternative statistics about the popularity of data sets.

GigaDB is supporting the ORCID registry (http://orcid.org/), and, where possible, we collect the ORCID identifier of any author(s) of data sets, allowing the linking of data sets by author to each other and to any external resources linked to those authors by ORCID identifiers. We hope to be able to extend the use of ORCID to allow users to log into GigaDB using their ORCID details. On top of external funding, we have received to set up this platform, some of the data hosting costs will have to be worked into the article processing charges of the journal with a view to our long term sustainability. With our low overheads, this will be a one-off payment that should be less than the fees and subscriptions that cloud storage and many other data publishing platforms and journals charge.

Conclusions

Maximizing the reuse of published data does not only involve its deposition, along with its metadata, into an open-access repository in a standardized format. Results published in scientific articles also have to be reproducible so, for example, comparisons can be made with analyses on new research data. GigaDB is positioning itself to service the research community in these respects with the provision of its open-access database with citable DOIs.

Since its launch in 2011, GigaDB has been providing a platform for data dissemination and publication, but we wish to solicit open discussions from others, to ensure that future developments are focusing on the issues that need resolving. We are always keen to listen to our users and the scientific community to enable us to provide the tools and services that are required by researchers (database@gigasciencejournal.com).

Acknowledgments

We thank Cogini for their professional Web development services. We also thank Sen Hong Wang and Yan Zhou of the Qiong Luo lab at HKUST, Shaoguang Liang, Alexandra Basford, Dennis Chan and Alex Wong of the BGI-HK research institute, who have provided support to the early development of the GigaDB services. Everyone's combined efforts are available in our GitHub server, and any future assistance and attention that our code can receive from the community would be much appreciated; https://github.com/gigascience.

Funding

The GigaScience Database was set up with funding from the BGI HK Research Institute, and continues to be supported in part by them, together with funding from the China National Gene Bank (CNGB). Funding for open access charge: BGI HK Research Institute.

Conflict of interest. None declared.

References

1. Liu,L., Li,Y., Li,S. *et al*. (2012) Comparison of next-generation sequencing systems. *J. Biomed. Biotechnol.*, **2012**, 251364. doi: 10.1155/2012/251364.
2. Wasinger,V.C., Zeng,M. and Yau,Y. (2013) Current status and advances in quantitative proteomic mass spectrometry. *Int. J. Proteomics*, **2013**, 180605. doi: 10.1155/2013/180605.
3. Putri,S.P., Yamamoto,S., Tsugawa,H. and Fukusaki,E. (2013) Current metabolomics: technological advances. *J. Biosci. Bioeng.*, **116**, 9–16. doi: 10.1016/j.jbiosc.2013.01.004.
4. Gonzalez,A., King,A., Robeson,M.S. III *et al*. (2012) Characterizing microbial communities through space and time. *Curr. Opin. Biotechnol.*, **23**, 431–436. doi: 10.1016/j.copbio.2011.11.017.
5. Haider,S. and Pal,R. (2013) Integrated analysis of transcriptomic and proteomic data. *Curr. Genomics*, **14**, 91–110.

6. Alsheikh-Ali,A.A., Qureshi,W., Al-Mallah,M.H. and Ioannidis,J.P.A. (2011) Public availability of published research data in high-impact journals. *PLoS One*, **6**, e24357. doi: 10.1371/journal.pone.0024357.
7. Stodden,V., Guo,P. and Ma,Z. (2013) Toward reproducible computational research: an empirical analysis of data and code policy adoption by journals. *PLoS One*, **8**, e67111. doi: 10.1371/journal.pone.0067111.
8. Editorial, (2009) Credit where credit is overdue. *Nat. Biotechnol.*, **27**, 579. http://dx.doi.org/10.1038/nbt0709-579.
9. Rosenthal,D.S.H. and Vargas,D.L. (2013) Distributed digital preservation in the cloud. *Int. J. Digit. Curation*, **8**, 107–119. doi: 10.2218/ijdc.v8i1.248.
10. Goodman,L., Edmunds,S.C. and Basford,A.T. (2012) Large and linked in scientific publishing. *Gigascience*, 1, 1. doi: 10.1186/2047-217X-1-1.
11. Sneddon,T.P., Li,P. and Edmunds,S.C. (2012) GigaDB: announcing the GigaScience database. *Gigascience*, **1**, 11. doi: 10.1186/2047-217X-1-11.
12. Piwowar,H.A., Day,R.S. and Fridsma,D.B. (2007) Sharing detailed research data is associated with increased citation rate. *PloS One*, **2**, e308. doi: 10.1371/journal.pone.0000308.
13. Li,D., Xi,F., Zhao,M. *et al*. (2011) Genomic data from *Escherichia coli* O104:H4 isolate TY-2482. *Gigascience*, **365**, 718–724. http://dx.doi.org/10.5524/100001.
14. Rohde,H., Qin,J., Cui,Y. *et al*. (2011) Open-source genomic analysis of Shiga-toxin–producing *E. coli* O104:H4. *N. Engl. J. Med*, **365**, 718–724. doi: 10.1056/NEJMoa1107643.
15. Hrynaszkiewicz,I., Busch,S. and Cockerill,M.J. (2013) Licensing the future: report on BioMed Central's public consultation on open data in peer-reviewed journals. *BMC Res. Notes*, **6**, 318. doi: 10.1186/1756-0500-6-318.
16. Ball,A. and Duke,M. (2012) *How to Cite Datasets and Link to Publications. DCC How-to Guides*. Digital Curation Centre, Edinburgh. http://www.dcc.ac.uk/resources/how-guides.
17. Edmunds,S.C., Pollard,T.J., Hole,B. and Basford,A.T. (2012) Adventures in data citation: sorghum genome data exemplifies the new gold standard. *BMC Res. Notes*, **2**, 223. doi: 10.1186/1756-0500-5-223.
18. Luo,R., Liu,B., Xie,Y. *et al*. (2012) SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *Gigascience*, **1**, 18. doi: 10.1186/2047-217X-1-18.
19. Kan,Z., Zheng,H., Liu,X. *et al*. (2012) Hepatocellular carcinoma genomic data from the Asian Cancer Research Group. *Gigascience*, **44**, 765–769. http://dx.doi.org/10.5524/100034.
20. Yilmaz,P., Kottmann,R., Field,D. *et al*. (2011) Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat. Biotechnol.*, **29**, 415–420. doi: 10.1038/nbt.1823.
21. Sansone,S.A., Rocca-Serra,P., Field,D. *et al*. (2012) Toward interoperable bioscience data. *Nat. Genet.*, **44**, 121–126. doi: 10.1038/ng.1054.
22. Haug,K., Salek,R.M., Conesa,P. *et al*. (2013) MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res.*, **41**, D781–D786. doi: 10.1093/nar/gks1004.
23. HoSui,S.J., Begley,K., Reilly,D. *et al*. (2011) The stem cell discovery engine: an integrated repository and analysis system for cancer stem cell comparisons. *Nucleic Acids Res.*, **40**, D984–D991.
24. Stoev,P., Komerički,A., Akkari,N. *et al*. (2013) *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. doi: 10.3897/BDJ.
25. Stoev,P., Komerički,A., Akkari,N. *et al*. (2013) Transcriptomic, DNA barcoding, and micro-CT imaging data from an advanced taxonomic description of a novel centipede species (Eupolybothrus cavernicolus Komerički & Stoev, spn.). *Gigascience*, **1**, e1013. http://dx.doi.org/10.5524/100063.
26. Goecks,J., Nekrutenko,A., Taylor,J. and the Galaxy Team. (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, 11, R86. doi: 10.1186/gb-2010-11-8-r86.
27. Luo,R., Liu,B., Xie,Y. *et al*. (2012) Software and supporting material for 'SOAPdenovo2: an empirically improved memory-efficient short read de novo assembly'. *Gigascience Database*, **1**, 18. http://dx.doi.org/10.5524/100044.
28. Bedoya-Reina,O., Ratan,A., Burhans,R. *et al*. (2013) Galaxy tools to study genome diversity. *Gigascience*, **2**, 17. http://dx.doi.org/10.1186/2047-217X-2-17.