(GIGA)^n SCIENCE

EDITORIAL Open Access

GigaDB: announcing the GigaScience database

Tam P Sneddon, Peter Li and Scott C Edmunds*

Abstract

With the launch of *GigaScience* journal, here we provide insight into the accompanying database *Giga*DB, which allows the integration of manuscript publication with supporting data and tools. Reinforcing and upholding *GigaScience*'s goals to promote open-data and reproducibility of research, *Giga*DB also aims to provide a home, when a suitable public repository does not exist, for the supporting data or tools featured in the journal and beyond.

Background

Internet pioneer Sir Tim Berners-Lee has stated: "Data is a precious thing and will last longer than the systems themselves" [1], and despite the challenges created due to data production in areas such as genomics growing at rates potentially faster than the ability to store and process it, attempts must still be made to capture and safeguard as much of these precious resources as possible. With the goals of *GigaScience* journal to maximize data reuse, dissemination, and transparency, having somewhere to host and curate all of the supporting data and tools surrounding this research is essential, and the *GigaScience* database, *Giga*DB (http://gigadb.org) is key to achieving this.

Main text

As can be seen in *GigaScience's* first issue, a research article on an epigenomics pipeline [2], in addition to having the raw data available in NCBI [SRP005934], also has this and all the supporting data (totaling 84 GB), such as the epigenomics tracks and the tools created for the pipeline[3], hosted in *Giga*DB. This dataset is linked and cited in the paper through a citable DOI (Digital Object Identifier), providing stability, and most importantly, additional discoverability and traceability through its ability to be tracked in the same manner as standard journal citations. Working and partnering with the British Library and DataCite consortium (http://datacite.org), these datasets are searchable and harvestable through their central metadata repository. Outside of the environmental sciences, data citation is still quite a new area, and we have worked closely with our publisher BioMed Central to ensure that citation of data follows DCC and DataCite best practice guidelines. In promoting the open-data movement, data is also released under the most open CC0 waiver, cutting any legal red tape [4], and maximizing its potential re-use. As *Giga*DB uses BGI's extensive computing infrastructure, it has also been populated with datasets produced by BGI, much of it released in a citable form pre-publication.

Releasing data in this novel manner has had a number of successes to date, particularly spurring the crowdsourcing of data from the deadly 2011 *E. coli* 0104:H4 outbreak (also discussed in Mike Schatz's commentary in this launch issue [5]) resulting in what has been termed "open-source genomics" [6]. For more on the background and mechanisms surrounding data citation, please see our recent correspondence [7] in the *BMC Research Notes* Data Sharing, Standardization and Publication series, using the release of the sorghum genome by *Giga*DB and publication in *Genome Biology* last year [8].

*Giga*DB currently comprises over 30 datasets. The largest of these is a hepatocellular carcinoma dataset [9], which consists of 15 Tb of normal and tumor raw data from 88 individuals. Additional data derived and processed from these same individuals, e.g. transcriptome sequence, can also be added to a DOI rapidly after their generation so users can immediately access the data from this ongoing project in a single, permanent place.

The goal of centralizing data and making it reproducible is exemplified by the mouse methylome dataset [3] in which we provide all data necessary to replicate the published results. This includes the raw fastq reads, bam alignment files, the Medusa software package, and the bigwig read-depth files. This and the sorghum study are excellent examples for future data submitters in regards to what can be done to not only comply with but also go beyond minimal journal data policies. Authors not only adhered to

* Correspondence: scott@gigasciencejournal.com
GigaScience, BGI-Hong Kong Co. Ltd, 16 Dai Fu Street, Tai Po Industrial Estate, NT, Hong Kong

our standard journal editorial policies for genomics studies, with raw data deposition in one of the three INSDC databases and assemblies in Genbank, but the sorghum study also deposited additionally processed data to the dbSNP and dbVar databases. The methylome *Giga*DB page also includes data and associated files that do not have equivalent established repositories. The complementary system of releasing data through *Giga*DB and established repositories also has the advantage of making the data available much sooner than the staggered build releases of many of these databases, which can take several months.

The *Giga*DB website is continuing to evolve and the next version will be released later this year. Features in this version will include an extensive search interface allowing users to choose datasets and/or files for download/export by dataset type, file format, sample, species, DOI, external accession etc.

Although most published *Giga*DB datasets are genomic, we can accept any large-scale data including proteomic, environmental, and imaging data. Taking such a broad range of data types makes data interoperability an issue, and we have been working with the ISA-Commons community to see if *Giga*DB can capture study and assay metadata along with relationships between dataset components and take submissions using their ISA-Tab format [10]. We have a nice example in our first issue, with much of the data supporting the epigenomics pipeline paper stored in a more interoperable ISA-compliant manner [3]. Upcoming datasets will include gut metagenomic data and a *Drosophila* genomics workflow dataset. We would like to be as comprehensive as possible, especially in providing a home for data that is not represented in any of the major public databases/repositories, so we encourage you contact us if you have a dataset or tools you would like to submit to *Giga*DB.

Maximising the reuse of published data does not only involve its deposition, along with its metadata, into an open access repository in a standardised format. Results published in scientific articles also have to be reproducible so, for example, comparisons can be made with analyses on new research data [11].

In future editions of *GigaScience*, we will be working with authors to make the computational tools and data processing pipelines described in their papers available and, where possible, executable on an informatics platform. We hope that by making both the data and processes involved in their analysis freely accessible, this novel form of publication will help articles published in our journal to have a much higher impact in the scientific literature, and maximize their reuse within the community.

Competing interests
All authors are employees of *GigaScience* and BGI.

Acknowledgements
The authors would like to thank Laurie Goodman and Alexandra Basford of the *GigaScience* team for support and feedback on the *Giga*DB project, Tin-Lap Lee and Huayan Gao at the Chinese University of Hong Kong for work on the data platform, and Qiong Luo, Senhong Wang and Yan Zhao at the Hong Kong University of Science and Technology for assistance with building the database. We would also like to think our collaborators at the British Library, DataCite, and ISA-Commons for their support and advice.

Author' contributions
All authors have been working on *Giga*DB and have contributed to this editorial. All authors read and approved the final manuscript.

Received: 1 June 2012 Accepted: 12 July 2012
Published: 12 July 2012

References

1. *Isn't it semantic.* http://www.ecs.soton.ac.uk/about/berners-lee.php.
2. Wilson GA, Dhami P, Feber A, Cortázar D, Suzuki Y, Schulz R, Schär P, Beck S: **Resources for methylome analysis suitable for gene knockout studies of potential epigenome modifiers.** *GigaScience* 2012, **1**:3.
3. Wilson G, Dharmi P, Saito Y, Cortázar D, Kunz C, Schär P, Beck S: **Resources for the MeDUSA (Methylated DNA Utility for Sequence Analysis) MeDIP-seq computational analysis pipeline for the identification of differentially methylated regions, and associated methylome data from 18 wild-type and mutant mouse ES, NP and MEF cells.** *GigaScience* 2012, http://dx.doi.org/10.5524/100035.
4. Hayden EC: **Open-data project aims to ease the way for genomic research.** *Nature* 2012, http://dx.doi.org/10.1038/nature.2012.10507.
5. Schatz MC, Phillippy AM: **The rise of a digital immune system.** *GigaScience* 2012, **1**:4.
6. Rohde H, Qin J, Cui Y, Li D, Nicholas ME, Loman J, Hentschke M, Chen W, Pu F, Peng Y, Li J, Xi F, Li S, Li Y, Zhang Z, Yang X, Zhao M, Wang P, Guan Y, Cen Z, Zhao X, Christner M, Kobbe R, Loos S, Oh J, Yang L, Danchin A, Gao GF, Song Y, Li Y, Yang H, Wang J, Xu J, Pallen MJ, Wang J, Aepfelbacher M, Yang R: **the *E. coli* O104:H4 Genome Analysis Crowd-Sourcing Consortium: Open-Source Genomic Analysis of Shiga-Toxin–Producing *E. coli* O104:H4.** *N Engl J Med* 2011, **365**:718–724.
7. Edmunds SC, Pollard TJ, Hole B, Basford A: **Adventures in data citation: sorghum genome data exemplifies the new gold standard.** *BMC Res Notes* 2012, **5**:223.
8. Zheng L-Y, Guo X-S, He B, Sun L-J, Peng Y, Dong S-S, Liu T-F, Jiang S, Ramachandran S, Liu C-M, Jing H-C: **Genome-wide patterns of genetic variation in sweet and grain sorghum (*Sorghum bicolor*).** *Genome Biol* 2011, **12**:R114.
9. Kan Z, Zheng H, Liu X, Li S, Barber TD, Gong Z, Gao H, Hao K, Willard MD, Xu J, Hauptschein R, Rejto PA, Fernandez J, Wang G, Zhang Q, Wang B, Chen R, Wang J, Lee NP, Lee WH, Ariyaratne PN, Tennakoon C, Mulawadi FH, Wong KF, Liu AM, Chan KL, Hu Y, Chou WC, Buser C, Zhou W, Lin Z, Peng Z, Yi K, Chen S, Li L, Fan X, Yang J, Ye R, Ju J, Wang K, Estrella H, Deng S, Wulur IH, Liu J, Ehsani ME, Zhang C, Loboda A, Sung WK, Aggarwal A, Poon RT, Fan ST, Wang J, Hardwick J, Reinhard C, Dai H, Li Y, Luk JM, Mao M, The Asian Cancer Research Group: **Hepatocellular carcinoma genomic data from the Asia Cancer Research Group.** *GigaScience.* 2012, http://dx.doi.org/10.5524/100034.
10. Sansone SA, Rocca-Serra P, Field D, Maguire E, Taylor C, Hofmann O, Fang H, Neumann S, Tong W, Amaral-Zettler L, Begley K, Booth T, Bougueleret L, Burns G, Chapman B, Clark T, Coleman LA, Copeland J, Das S, de Daruvar A, de Matos P, Dix I, Edmunds S, Evelo CT, Forster MJ, Gaudet P, Gilbert J, Goble C, Griffin JL, Jacob D, Kleinjans J, Harland L, Haug K, Hermjakob H, Ho Sui SJ, Laederach A, Liang S, Marshall S, McGrath A, Merrill E, Reilly D, Roux M, Shamu CE, Shang CA, Steinbeck C, Trefethen A, Williams-Jones B, Wolstencroft K, Xenarios I, Hide W: **Toward interoperable bioscience data.** *Nat Genet* 2012, **44**(2):121–126.
11. Ioannidis JP, Allison DB, Ball CA, Coulibaly I, Cui X, Culhane AC, Falchi M, Furlanello C, Game L, Jurman G, Mangion J, Mehta T, Nitzberg M, Page GP, Petretto E, van Noort V: **Repeatability of published microarray gene expression analyses.** *Nature Gen.* 2009, **41**(2):149–155.

doi:10.1186/2047-217X-1-11
Cite this article as: Sneddon *et al.*: GigaDB: announcing the GigaScience database. *GigaScience* 2012 1:11.