

The GigaSolution to data publication, reuse and integration.

Christopher I Hunter, Peter Li, Xiao Si Zhe, Robert Davidson, Laurie Goodman & Scott C Edmunds

Affiliation: GigaScience, BGI-HK Research Institute, 16 Dai Fu Street, Tai Po Industrial Estate, Hong Kong SAR, China.

Correspondence to chris@gigasciencejournal.com

To meet the needs of a new generation of biological and biomedical research in the era of “big-data”, BGI and BioMed Central have formed a unique partnership to publish the journal *GigaScience*. *GigaScience* is a novel publishing platform that combines the open-access article publishing expertise of BMC with the bioinformatics expertise and extensive computational storage space at BGI. The journal’s affiliated database, *GigaDB* (Figure 1), serves as a repository that hosts the data and tools associated with *GigaScience* publications. It also provides a rapid data release mechanism for datasets that are not associated with *GigaScience* articles that have not previously been published elsewhere by giving each the dataset a DOI, making them citable in a standard (and countable) manner in the reference section of papers that use these data.

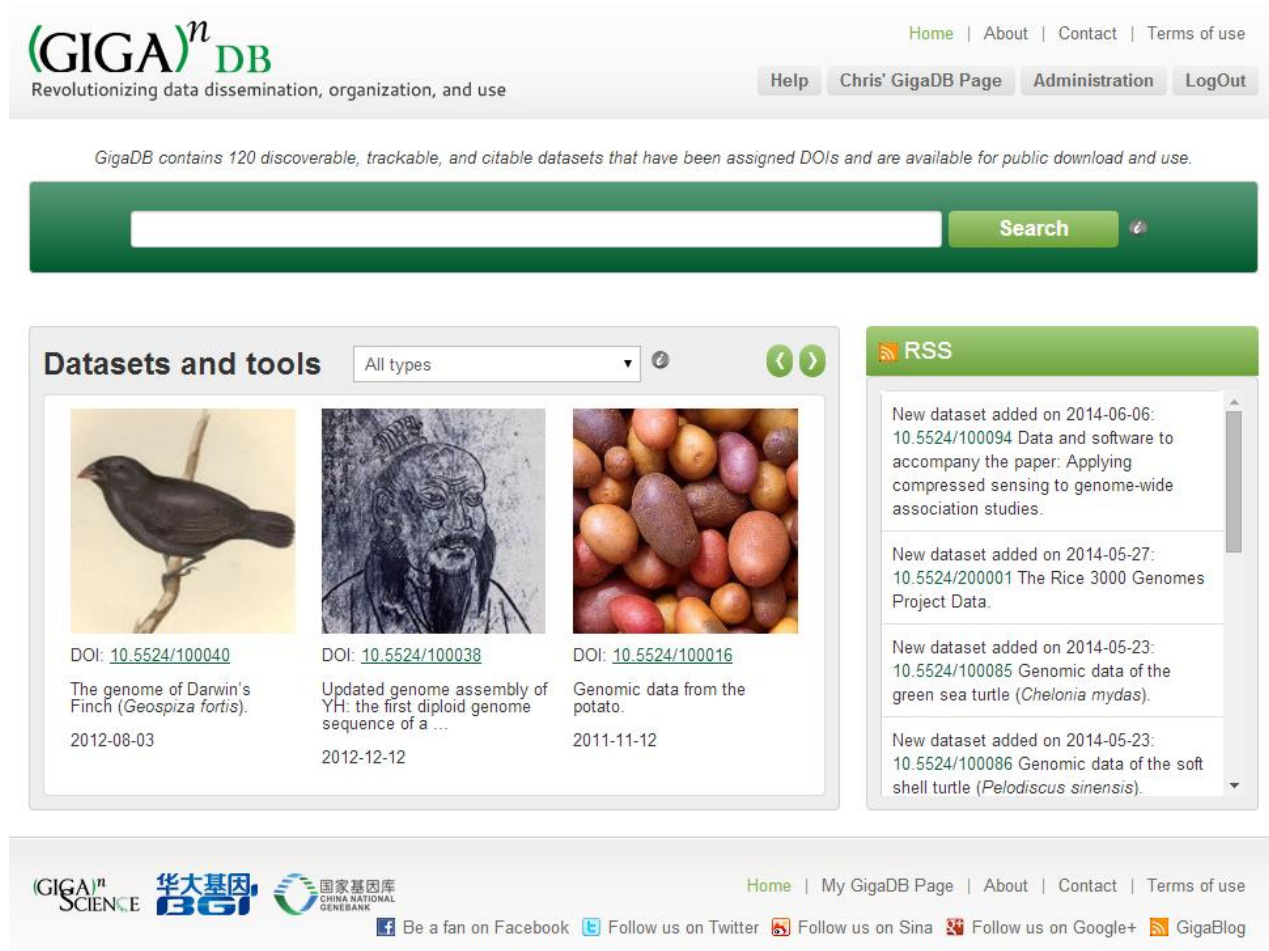


Figure 1. The home page of GigaDB website

In its first 18 months, over 100 datasets (>20 TB in size) have been made available in *GigaDB* — all under a CC0 Waiver (the most open sharing waiver available). These datasets include the first ~50 bird genomes from an avian phylogenomics study (Figure 2)(Zhang *et al.*, 2014) some of these datasets were made available before they were published in scientific journals. *GigaDB* will also host data from the Rice 3000 genomes project (The 3000 rice genomes project *et al.*, 2014), the 10,000 genome project (G10K), 1000 plant transcriptomes project, as well other smaller scale genome projects, and numerous non-genomic datasets (e.g. imaging, proteomic, metabolomics, etc.), some of which currently have no formal community data repository.



Data released on May 16, 2014

The avian phylogenomic project data.

Zhang, G; Li,B; Li,C; Gilbert,MTP; Jarvis,E; The Avian Genome Consortium; Wang,J. (2014); The avian phylogenomic project data. GigaScience Database. <http://dx.doi.org/10.5524/101000> [RIS](#) [BioTeX](#) [Text](#)

The evolutionary relationship of modern birds is one of the most challenging questions in systematic biology and has been debated for centuries. We proposed to rebuild the avian phylogenetic tree by using whole genome data, thus we have collected genomes of 48 bird species, representing 36 orders of bird class. The chicken, zebra finch, and turkey genomes, which were sequenced in Sanger method, were collected from public domain. Another three genomes, pigeon, peregrine falcon, and duck, have been published during the development of this project. The data posted here include the full genome assemblies of 42 bird species, the repeat and gene annotation produced by our new pipeline, 8295 1:1 syntenic orthologous genes, and the whole genome alignment data for all bird species. The detailed information of the published genomes can be accessed from their own publications. The 42 genomes first released here were sequenced and assembled with NGS technology in whole genome shotgun strategy. Using an homology-based method, we annotated 13000~17000 protein-coding genes in each avian genome.

So far as we know, the avian phylogenomic project is the biggest comparative genomics project to date. The unprecedented genomic data presented here will contribute to the downstream analyses in many fields, including phylogenetics, comparative genomics, neurology, development biology, etc. Below are listed the links to all the individual species data used in this study. In addition, the entire dataset has been compressed into a single archive file for those who wish to retrieve the complete set.

Adelle Penguin - *Pygoscelis adeliae* - PYGAD - [doi:10.5524/100006](http://dx.doi.org/10.5524/100006)
 American Crow - *Corvus brachyrhynchos* - CORBR - [doi:10.5524/101008](http://dx.doi.org/10.5524/101008)
 American Flamingo - *Phoenicopterus ruber ruber* - PHORU - [doi:10.5524/101035](http://dx.doi.org/10.5524/101035)
 Anna's Hummingbird - *Calypte anna* - CALAN - [doi:10.5524/101004](http://dx.doi.org/10.5524/101004)
 Bald Eagle - *Haliaeetus leucocephalus* - HALLE - [doi:10.5524/101040](http://dx.doi.org/10.5524/101040)
 Barn Owl - *Tyto alba* - TYTAL - [doi:10.5524/101039](http://dx.doi.org/10.5524/101039)
 Bar-tailed Trogon - *Apaloderma vittatum* - APAVI - [doi:10.5524/101016](http://dx.doi.org/10.5524/101016)
 Brown Mesite - *Mesitornis unicolor* - MESUN - [doi:10.5524/101030](http://dx.doi.org/10.5524/101030)
 Budgerigar - *Melopsittacus undulatus* - MELUN - [doi:10.5524/100059](http://dx.doi.org/10.5524/100059)
 Carmine Bee-eater - *Merops nubicus* - MERNU - [doi:10.5524/101029](http://dx.doi.org/10.5524/101029)

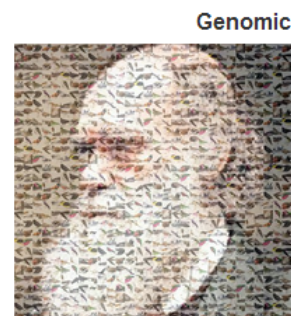


Figure 2. <http://dx.doi.org/10.5524/101000> The avian phylogenomics project dataset stub in GigaDB.

Through our association with DataCite, each dataset in *GigaDB* is assigned a DOI that can be used as a standard citation in the reference section of a paper, improving access and use of these data in articles by the authors and other researchers.

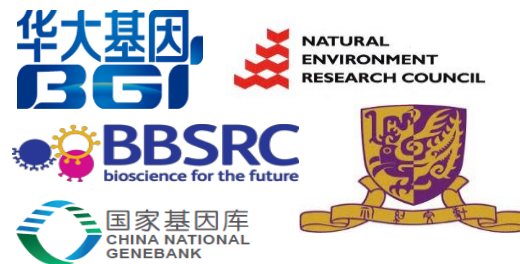
In order to make NGS data interpretation as accessible as data generation, we have implemented "GigaGalaxy" (<http://galaxy.cbiit.cuhk.edu.hk>). We have ported the popular Short Oligonucleotide Analysis Package (SOAP <http://soap.genomics.org.cn>) as well as supporting tools such as Contiguator2 (<http://contiguator.sourceforge.net>) into the Galaxy framework, to provide seamless

NGS mapping, de novo assembly, NGS data format conversion and sequence alignment visualization. Our vision is to create an open publication, review and analysis environment by integrating GigaGalaxy into the publication platform at *GigaScience* and together with GigaDB, to help integrate data and analyses used in publications. We have begun this effort by re-implementing the data procedures described by Luo *et al.*, (Luo *et al.*, 2012) as Galaxy workflows so that they can be shared in a manner which can be visualized and executed in GigaGalaxy. We hope to revolutionize the publication model with the aim of executable publications, where data analyses can be reproduced and reused.

Acknowledgments

We would like to thank the following for their financial support

BGI Research, Shenzhen, China
China National GeneBank, China
Chinese University of Hong Kong, Hong Kong
BBSRC, UK
Natural Environment Research Council, UK



In addition to the authors, we would like acknowledge the works carried out by the following people; Alexandra Basford, Huayan Gao (GigaScience), Yong Zhang (BGI; China National Genebank) Shaoguang Liang (BGI-SZ), Alex Wong, Dennis Chan (BGI-HK), Tin-Lap Lee (CUHK), Qiong Luo, Senghong Wang, Yan Zhou (HKUST) and Mark Viant (Birmingham Uni).

References

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, et al. (2012): Soapdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* 1(1):18. doi: [10.1186/2047-217X-1-18](https://doi.org/10.1186/2047-217X-1-18)

The 3000 rice genomes project. (2014): The 3,000 Rice Genomes Project. *GigaScience* 3:7. doi: [10.1186/2047-217X-3-7](https://doi.org/10.1186/2047-217X-3-7)

Zhang, G; Li,B; Li,C; Gilbert,MTP; Jarvis,E; The Avian Genome Consortium; Wang,J. (2014): The avian phylogenomic project data. *GigaScience Database*. doi: [10.5524/101000](https://doi.org/10.5524/101000)