Edmunds *GigaScience* 2014

POSTER Open Access

GigaGalaxy: A GigaSolution for reproducible and sustainable genomic data publication and analysis

Scott Edmunds[1,2], Peter Li[1,2], Huayan Gao[3,4], Ruibang Luo[2,5], Dennis Chan[1], Alex Wong[1], Zhang Yong[2], Tin-Lap Lee[3,4]

Abstract

Today's next generation sequencing (NGS) experiments generate substantially more data and are more broadly applicable to previous high-throughput genomic assays. Despite the plummeting costs of sequencing, downstream data processing and analysis create financial and bioinformatics challenges for many biomedical scientists. It is therefore important to make NGS data interpretation as accessible as data generation. GigaGalaxy (http://galaxy.cbiit.cuhk.edu.hk) represents a NGS data interpretation solution towards the big sequencing data challenge. We have ported the popular Short Oligonucleotide Analysis Package (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 its GigaDB database that links to more than 40 TBs of genomic data. We have begun this effort by re-implementing the data procedures described by Luo *et al*., (*GigaScience* **1**: 18, 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.

Keywords: Galaxy, workflows, reproducible research, genome assembly, next generation sequencing, *GigaScience*

Background

Growing replication gap:

- 10/18 microarray papers cannot be reproduced
- Ioannidis: "Most Published Research Findings Are False"
- >15X increase in retracted papers in last decade
- Lack of incentives to make data/methods available

GigaSolution: deconstructing the paper

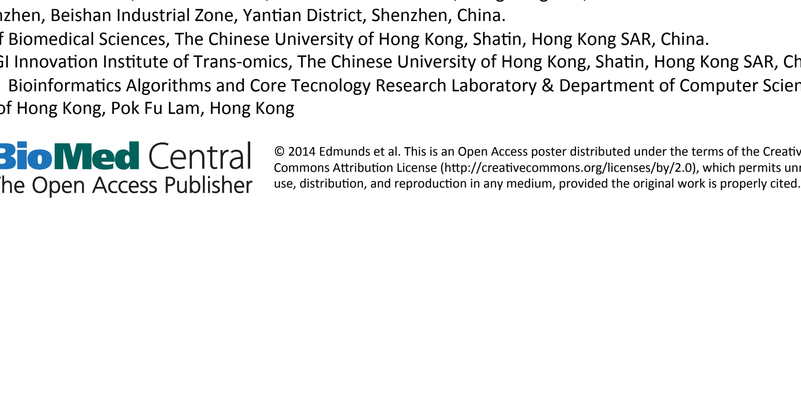
Combine and integrate (via citable DOIs):

Open-access journal
www.gigasciencejournal.com

Data Publishing Platform
gigadb.org

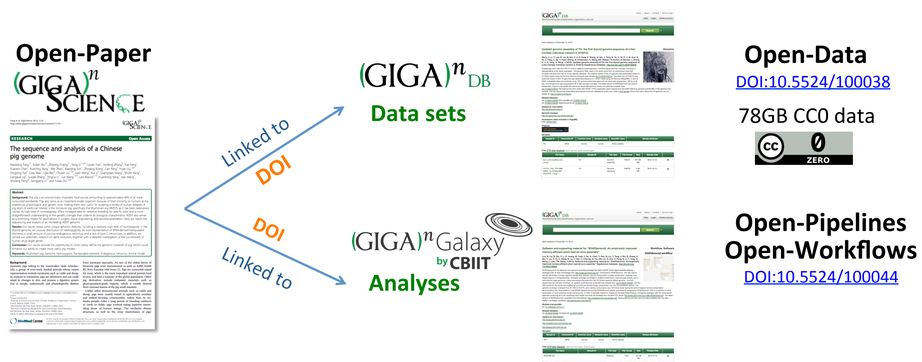
Data Analysis Platform
galaxy.cbiit.cuhk.edu.hk

GigaGalaxy: screenshot

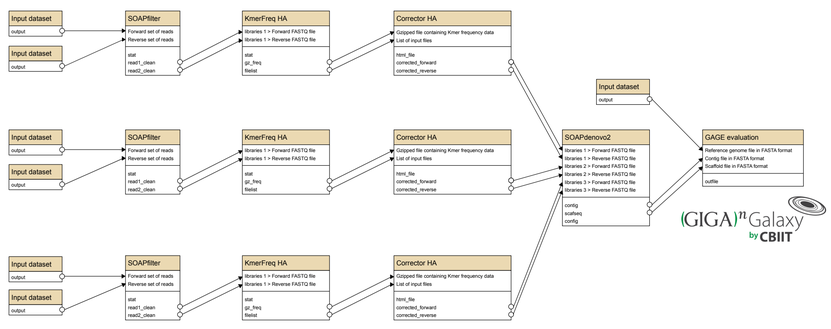


Example: SOAPdenovo2

Linking papers to data and analyses

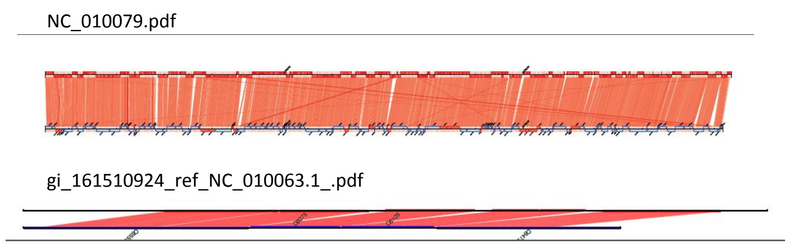


Implement paper pipelines in GigaGalaxy



Visualization of results:

e.g. GAGE metrics and CONTIGuator 2 outputs:



References

1. Ioannidis et al., Repeatability of published microarray gene expression analyses. *Nature Genetics* 2009 41: 14
2. Science publishing: The trouble with retractions *Nature* 2011 478, 26-28
3. Ioannidis J. Why Most Published Research Findings Are False. *PLoS Med* 2005 2(8): e124.
4. Sneddon et al., GigaDB: promoting data dissemination and reproducibility. *Database* 2014: bau018
5. Luo et al., SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler *GigaScience* 2012, 1:18
6. Wang, J; et al., (2012): Updated genome assembly of YH: the first diploid genome sequence of a Han Chinese individual (version 2, 07/2012). GigaScience Database. http://dx.doi.org/10.5524/100038.
7. Luo, R; et al., (2012): Software and supporting material for "SOAPdenovo2: An empirically improved memory-efficient short read de novo assembly". GigaScience Database. http://dx.doi.org/10.5524/100044
8. Galardini et al.:CONTIGuator: a bacterial genomes finishing tool for structural insights on draft genomes. *Source Code for Biology and Medicine* 2011 6:11.
9. GigaGalaxy code in GitHub under ASL v3.0 open source license https://github.com/gigascience

Acknowledgements

Thanks to:

Laurie Goodman, Chris Hunter, Xiao Si Zhe, Rob Davidson, Tam Sneddon (*GigaScience*), Shaoguang Liang (BGI-SZ), Qiong Luo, Senghong Wang, Yan Zhou (HKUST), Mark Viant (Birmingham Uni), Marco Galardini (Unifi)

Financial support from:

Correspondence: scott@gigasciencejournal.com
1. BGI HK Research Institute, 16 Dai Fu Street, Tai Po Industrial Estate, Hong Kong SAR, China.
2. BGI-Shenzhen, Beishan Industrial Zone, Yantian District, Shenzhen, China.
3. School of Biomedical Sciences, The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China.
4. CUHK-BGI Innovation Institute of Trans-omics, The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China.
5. HKU-BGI Bioinformatics Algorithms and Core Tecnology Research Laboratory & Department of Computer Science, University of Hong Kong, Pok Fu Lam, Hong Kong

doi:10.6084/m9.figshare.713512
Cite this poster as: GigaGalaxy: A GigaSolution for reproducible and sustainable genomic data publication and analysis. Scott C. Edmunds, Peter Li, Huayan Gao, Ruibang Luo, Dennis Chan, Alex Wong, Zhang Yong, Tin-Lap Lee. figshare http://dx.doi.org/10.6084/m9.figshare.713512