(GIGA)^n SCIENCE

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 20 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
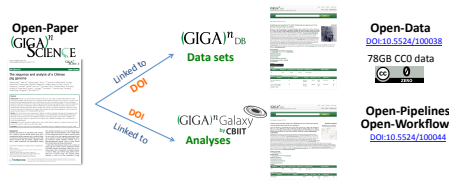Combine and integrate (via citable DOIs):

(GIGA)^n SCIENCE — Open-access journal
www.gigasciencejournal.com — OPEN ACCESS

(GIGA)^n DB — Data Publishing Platform
gigadb.org — OD

(GIGA)^n Galaxy by CBIIT — Data Analysis Platform
galaxy.cbiit.cuhk.edu.hk — open source

GigaGalaxy: screenshot

Example: SOAPdenovo2

Linking papers to data and analyses

Open-Paper **(GIGA)^n SCIENCE** — Linked to DOI — **(GIGA)^n DB** Data sets — Linked to DOI — **(GIGA)^n Galaxy by CBIIT** Analyses

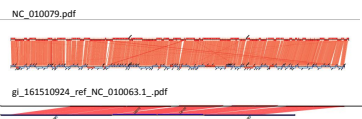Open-Data DOI:10.5524/100038 — 78GB CC0 data

Open-Pipelines Open-Workflows DOI:10.5524/100044

Implement paper pipelines in GigaGalaxy

Visualization of results:

e.g. GAGE metrics and CONTIGuator 2 outputs:

NC_010079.pdf

gi_161510924_ref_NC_010063.1_.pdf

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. Luo R et al., SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler *GigaScience* 2012, 1:18
5. 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.
6. 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
7. Galardini et al.:CONTIGuator: a bacterial genomes finishing tool for structural insights on draft genomes. *Source Code for Biology and Medicine* 2011 6:11.

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

Submit your next manuscript containing large-scale data and workflows to GigaScience and take full advantage of:

- No space constraints, and unlimited data and workflow hosting in GigaDB and GigaGalaxy
- Article processing charges for all submissions in 2013 covered by BGI
- Open access, open data and highly visible work freely available for distribution
- Inclusion in PubMed and Google Scholar

(GIGA)^n SCIENCE