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

Scott Edmunds1,2, Peter Li1,2, Huayan Gao1,4, Ruibang Luo2,5, Dennis Chan1, Alex Wong1, Zhang Yong2, Tin-Lap Lee1,4

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 (SOAP) (http://soap.genomics.org.cn) as well as supporting tools such as Contiguator 2 (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 20TBs 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):

GigaDB: Open-data journal
GigaGalaxy: Data Publishing Platform
GigaSolution: Data Analysis Platform
GigaGalaxy Database

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 JPA. (2005) Contradicted and initially strong findings are the most likely to be spuriously significant. PLoS Medicine 2: e124.
7. Open-access journal, open data and highly visible work freely available for distribution

Acknowledgments:

This work was supported by the Hong Kong Research Grants Council (HKU 759110) and by the National Natural Science Foundation of China (61003162, 61103200, 61100005, and 61100003).

Financial support from:

Financial support from the Research Grants Council of Hong Kong and the Chinese Academy of Sciences.