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

CONTACT DETAILS:
RDA WG: Data Type Registries (DTR)
NAMES: Scott Edmunds/Laurie Goodman/Susanna-Assunta Sansone
ORGANISATIO: GigaScience/BGI I
EMAIL: scott@gigasciencejournal.com

Objectives:

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.

On-going activities:

We have ported the popular Short Oligonucleotide Analysis Package (http://soap.genomics.org.cn) 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 25 TBs of genomic data.

Results:

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 have also described the experiment using the ISA framework to provide a richer and more interoperable description of the experimental workflows. We hope to revolutionize the publication model with the aim of executable publications, where data analyses can be reproduced and reused.
URL: http://galaxy.cbiit.cuhk.edu.hk and http://gigadb.org/

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

Scott Edmunds1,2, Laurie Goodman1,2, Peter Li1,2, Huayan Gao3,4, Chris Hunter1,2, Si Zhe Zhao1,2, Ruibang Luo2,5, Dennis Chan1, Alex Wong1, Zhang Yong2, Tin-Lap Lee3,4 and the ISA-TAB team6

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) 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 25 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 have also described the experiment using the ISA framework to provide a richer and more interoperable description of the experimental workflows. We hope to revolutionize the publication model with the aim of executable publications, where data analyses can be reproduced and reused.