Open (and Big) Data – the next challenge

Beyond dead trees: are publishers the problem or solution?

Scott Edmunds
OASPA Asia, 2nd June 2013
@gigascience
Harnessing Data-Driven Intelligence

Enables:
Using networking power of the internet to tackle problems
Can ask new questions & find hidden patterns & connections
Build on each others efforts quicker & more efficiently
More collaborations across more disciplines
Harness wisdom of the crowds: crowdsourcing, citizen science, crowdfunding

Enabled by:
Removing silos, open licenses, transparency, immediacy
Dead trees not fit for purpose
The problems with publishing

• Scholarly articles are merely advertisement of scholarship. The actual scholarly artefacts, i.e. the data and computational methods, which support the scholarship, remain largely inaccessible — Jon B. Buckheit and David L. Donoho, WaveLab and reproducible research, 1995

• Lack of transparency, lack of credit for anything other than “regular” dead tree publication.

• If there is interest in data, only to monetise & re-silo

• Traditional publishing policies and practices a hindrance
Things holding us back:

• Disincentives to share or communicate:
  – Ingelfinger*! Embargoes, anti preprint & early data release policies
  – Page/method/citation limits

• Disincentives to remix
  – Open source approaches = plagiarism?

• Disincentives to release more quickly/more granularly
  – “Salami Slicing”

• First 2 years of citation data the only currency
  – “Faddism” v long term use or reproducibility. Publication bias.

* T-Shirts available from Graham Steel // http://www.zazzle.co.uk/steelgraham
The consequences: growing replication gap

Out of 18 microarray papers, results from 10 could not be reproduced

**Essay**

*Why Most Published Research Findings Are False*

John P. A. Ioannidis

Consequences: increasing number of retractions

>15X increase in last decade

Strong correlation of “retraction index” with higher impact factor

2. Retracted Science and the Retraction Index [http://iai.asm.org/content/79/10/3855.abstract](http://iai.asm.org/content/79/10/3855.abstract)
Consequences: growing replication gap

1. Ioannidis et al., 2009. Repeatability of published microarray gene expression analyses. *Nature Genetics* 41: 14


More retractions:

>15X increase in last decade

At current % > by 2045 as many papers published as retracted

1. Ioannidis et al., 2009. Repeatability of published microarray gene expression analyses. *Nature Genetics* 41: 14


Global perceptions of Chinese Research

Million RMB rewards for high IF publications = ?

"Faked research is endemic in China"

Focus on quality, not just quantity

China publishes huge amounts of scientific research. Now it must make more of it worth reading, says Changhui Peng.

Science, 29th November 2013: http://www.sciencemag.org/content/342/6162/1035.full
Global perceptions of Chinese Research

Million RMB rewards for high IF publications = ?

“Wide distribution of information is key to scientific progress, yet traditionally, Chinese scientists have not systematically released data or research findings, even after publication."

“There have been widespread complaints from scientists inside and outside China about this lack of transparency.”

“Usually incomplete and unsystematic, [what little supporting data released] are of little value to researchers and there is evidence that this drives down a paper's citation numbers.”

“Faked research is endemic in China”

475, 267 (2011)

Focus on quality, not just quantity

China publishes huge amounts of scientific research. Now it must make more of it worth reading, says Changhui Peng.

---


Science, 29th November 2013: http://www.sciencemag.org/content/342/6162/1035.full

Issues not just in China...

Need:
...to publish protocols BEFORE analysis
...better access to supporting data
...more transparent & accountable review

...to publish replication studies

Nature Rejects Publication of Paper Reporting that STAP Does Not Work

Posted on March 24, 2014
New incentives/credit

Credit where credit is overdue:
“One option would be to provide researchers who release data to public repositories with a means of accreditation.”
“An ability to search the literature for all online papers that used a particular data set would enable appropriate attribution for those who share.”

*Nature Biotechnology* 27, 579 (2009)

- Data
- Software
- Review
- Re-use...

= Credit
GigaSolution: deconstructing the paper
Combines and integrates:

- Open-access journal
- Data Publishing Platform
- Data Analysis Platform

Utilizes big-data infrastructure and expertise from:

- www.gigadb.org
- www.gigasciencejournal.com

华大基因
www.gigadb.org
www.gigasciencejournal.com
Submitter logs in to GigaDB website and uploads Excel submission.

Fail – submitter is provided error report.

Pass – dataset is uploaded to GigaDB.

Curator contacts submitter with DOI citation and to arrange file transfer (and resolve any other questions/issues).

Curator makes dataset public (can be set as future date if required).

XML is generated and registered with DataCite.

Files
Submitter provides files by ftp or Aspera.

GigaDB
DOI assigned

Curator Review

Excel submission file

Validation checks

Pass – dataset is uploaded to GigaDB.

 rejected – submitter is provided error report.

DOI 10.5524/100003
Genomic data from the crab-eating macaque/cynomolgus monkey (Macaca fascicularis) (2011)

See: http://database.oxfordjournals.org/content/2014/bau018.abstract
• 10-100x faster download than FTP
• Provide curation & integration with other DBs
Beneficiaries of this open data?
Beneficiaries of this open data?

Rice 3K project: 3,000 rice genomes, 13.4TB public data
New Article types v Species Description <2012
Collaborations with Pensoft & PLOS
Cyber-centipede & virtual worms

Rosa Fernandez 1, Sebastian Koeltz, Jennifer Lennihan, Gonzalo Giribet, Alexander Ziegler
1 Museum of Comparative Zoology, Department of Organism and Developmental Biology, Harvard University, Cambridge, Massachusetts, United States of America, 2 Downtown Dallas, Dallas, Texas, United States of America

Abstract

The rapid rise of the number of publications and projects using high-resolution, 3D imaging in protein, nucleic acid, and organellar analyses has led to a number of new techniques that are now furthering our understanding of the evolution of life. In recent years, these technologies have led to the development of new species descriptions, which have been used to determine the species complexes of organisms from living as well as fossilized specimens. In this study, we explore the potential of using non-destructive imaging techniques to reveal morphology and function of organisms from living and fossilized specimens. We also discuss the use of this approach to determine the species complexes of organisms from living as well as fossilized specimens. In addition, we discuss the use of this approach to determine the species complexes of organisms from living as well as fossilized specimens.
Genomics
Barcoding
Imaging
microCT
Video
New & more transparent peer-review: open review

Edmunds GigaScience 2013, 2:1
http://www.gigasciencejournal.com/content/2/1/1

EDITORIAL
Open Access

Peering into peer-review at GigaScience

Scott C Edmunds

Abstract
Fostering and promoting more open and transparent science is one of the goals of GigaScience. One of the ways we have been doing this is by throwing light on the peer-review process and carrying out open peer-review as standard. In this editorial, we provide our rationale for undertaking this policy, give examples of our positive experiences to date, and encourage others to open up the normally opaque publication process.
Reward open & transparent review

End reviewer 3 Downfall parody videos, now!
New & more transparent peer-review: pre-prints

Haldane’s Sieve
Discussing preprints in population and evolutionary genetics
Real-time open-review = paper in arXiv + blogged reviews

Feedback and analysis of the Assemblathon 2 pre-print

There has already been some discussion of the pre-print of the Assemblathon 2 manuscript. Although a pre-print is not the same thing as a peer-reviewed, accepted paper — I don’t want us to get too ahead of ourselves! — I thought it useful to start collecting together some of the online commentaries:

- Homologous blog post 1: highlights a few conclusions from the paper
- Homologous blog post 2: delves into the results, and attempts to estimate some of the costs of genome assembly. Assemblathon co-author Sébastien Boisvert adds some useful comments.
- Haldane’s Sieve post: an invited blog post by lead author Keith Bradnam, that summarizes what the Assemblathons are all about by way of a pizza-themed analogy
- Reevaluating Assembly Evaluations with Feature Response Curves: GAGE and Assemblathons: this is not a blog post, but a recently published paper that evaluates some of the Assemblathon 2 data
- Thoughts on the Assemblathon 2 paper: by C. Titus Brown (a reviewer of the manuscript)
- Homologous blog post 3: reactions to the previous post by C. Titus Brown
- Assemblathon 2 review, round 1, parts thereof: a concise version of C. Titus Brown’s formal manuscript review (minus the specific suggestions)
- On assembly uncertainty (inspired by the Assemblathon 2 debate): blog post by Lex Nederbragt in response to post by C. Titus Brown

www.gigasciencejournal.com/content/2/1/10       http://tmblr.co/ZzXdssfOMJfy
Reward open & transparent review

Real-time open-review = paper in arXiv + blogged reviews

Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species

So far Altmetric has seen 143 tweets from 91 accounts with an upper bound of 90,939 combined followers.

- The Assemblathon @assemblathon 1,893 followers
  The Assemblathon 2 pre-print is now available on arXiv! http://t.co/ExyCZbC8
  Additional data files are here: http://t.co/IltF54iN
  24-Jan-2013

- Keith Bradnam @kbradnam 1,539 followers
  w00t! *@assemblathon: The Assemblathon 2 pre-print is now available on arXiv!
  http://t.co/2GfmKnt6t
  24-Jan-2013

- Titus Brown @ctitusbrown 3,891 followers
  Assemblathon 2 paper now on arxiv: http://t.co/goeH4I2l/ @assemblathon
  24-Jan-2013

Readers on
- 238 Mendeley
- 10 CiteULike

#titusischucknorris
Readers are interested in open review

Next step to link to ORCID
Reward better handling of metadata…

Novel tools/formats for data interoperability/handling.

Toward interoperable bioscience data


Affiliations | Corresponding author

Nature Genetics 44, 121-126 (2012) | doi:10.1038/ng.1054
Published online 27 January 2012
Rewarding and aiding reproducibility

OMERO: providing access to imaging data...

Runx1 modulates adult hair follicle stem cell emergence and maintenance from distinct embryonic skin compartments

Karen M. Osorio, Karin C. Lilja, Tudorita Tumbar

JCB vol. 193 no. 1 235-250  Article DOI: 10.1083/jcb.201006088  DataViewer DOI: 10.1083/jcb.201006088.dv
Rewarding and aiding reproducibility

Implement workflows in a community-accepted format

http://galaxyproject.org
GigaGalaxy is supported by

GigaGalaxy is Gigascience's Galaxy-based platform for supporting the reproducibility of data analyses. We will be using GigaGalaxy to provide computational tools and workflows that further document and/or reproduce the data analyses reported in papers published in our journal. In addition, we will be making the next-generation sequencing tools developed by BGI available from this platform. GigaGalaxy is a joint project with Prof. Tin Loi Lai and Xiuyan Gao at the CUHK-BGI Innovation Institute of Trans-Omics. We are supported by BGI and by The China National Genebank.
GigaScience example 1C: Measuring the nucleotide diversity of aye-aye populations

Data released on December 30, 2013

GigaGalaxy workflows and histories from “Galaxy tools to study genome diversity”

Related manuscripts:
doi:10.1186/2041-211X-2-17

Additional Information:
http://galaxy.cbiit.gov/run/gigascience/g/obedoyanima2013
SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler

Ruibang Luo, Binghang Liu, Yinlong Xie, Zhenyu Li, Weihua Huang, Jiaying Yuan, Guangzhu He, Yanxiang Chen, Qi Pan, Yunjie Liu, Jingbo Tang, Guangxiong Wu, Hao Zhang, Yujian Shi, Yong Liu, Chang Yu, Bo Wang, Yao Lu, Changlei Han, David W Cheung, Siu-Ming Yiu, Shaoliang Peng, Zhu Xiaojian, Guangming Liu, Xiangke Liao, Yingrui Li, Huanming Yang, Jian Wang, Tak-Wah Lam, and Jun Wang

Abstract

**Background:** There is a rapidly increasing amount of de novo genome assembly using next-generation sequencing (NGS) short reads; however, several big challenges remain to be overcome in order for this to be efficient and accurate. SOAPdenovo has been successfully applied to assemble many published genomes, but it still needs improvement in continuity, accuracy and coverage, especially in repeat regions.

**Findings:** To overcome these challenges, we have developed its successor, SOAPdenovo2, which has the advantage of a new algorithm design that reduces memory consumption in graph construction, resolves more repeat regions in contig assembly, increases coverage and length in scaffold construction, improves gap closing, and optimizes for large genome.

**Conclusions:** Benchmark using the Assembiathon1 and GAGE datasets showed that SOAPdenovo2 greatly surpasses its predecessor SOAPdenovo and is competitive to other assemblers on both assembly length and accuracy. We also provide an updated assembly version of the 2008 Asian (YH) genome using SOAPdenovo2. Here, the contig and scaffold N50 of the YH genome were ~20.9 kbp and ~22 Mbp, respectively, which is 3-fold and 50-fold longer than the first published version. The genome coverage increased from 81.16% to 93.91%, and memory consumption was ~2/3 lower during the point of largest memory consumption.

**Keywords:** Genome, Assembly, Contig, Scaffold, Error correction, Gap-filling
How are we supporting data reproducibility?

Open-Paper

DOI: 10.1186/2047-217X-1-18

>23,000 accesses

(GIGA)_n Data sets

Linked to DOI

Open-Data

DOI: 10.5524/100038

78GB CC0 data

Open-Pipelines

Open-Workflows

DOI: 10.5524/100044

Open-Review

7 reviewers tested data in ftp server & named reports published

Enabled code to being picked apart by bloggers in wiki


Open-Code

Code in sourceforge under GPLv3:

http://soapdenovo2.sourceforge.net/

>20,000 downloads
**Reward open & transparent review**

7 referees downloaded & tested data, then signed reports

---

<table>
<thead>
<tr>
<th>Pre-publication history</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler</td>
</tr>
<tr>
<td>Ruibang Luo¹, Binghang Liu¹, Yinlong Xie¹, Zhenyu Li, Weihua Huang, Jianying Yuan, Guangzhao He, Yanxiang Chen, Qi Pan, Yunjie Liu, Jingbo Tang, Gangxiong Wu, Hao Zhang, Yujian Shi, Yong Liu, Chang Yu, Bo Wang, Yao Lu, Changlei Han, David W Cheung, Siu-Ming Yiu, Shaoliang Peng, Zhu Xiaqian, Guangming Liu, Xiangke Liao, Yingrui Li, Huanming Yang, Jian Wang, Tak-Wah Lam* and Jun Wang*</td>
</tr>
</tbody>
</table>

* Corresponding authors: Tak-Wah Lam twlam@cs.hku.hk - Jun Wang wancj@cenomics.org.cn

† Equal contributors


Pre-publication versions of this article and reviewers’ reports

<table>
<thead>
<tr>
<th>Original Submission - Version 1</th>
<th>Manuscript</th>
<th>24 Jul 2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reviewer’s Report</td>
<td>Aleksey Zimin</td>
<td>22 Aug 2012</td>
</tr>
<tr>
<td>Reviewer’s Report</td>
<td>Marie Caccamo</td>
<td>28 Aug 2012</td>
</tr>
<tr>
<td>Resubmission - Version 2</td>
<td>Manuscript</td>
<td>Author’s comment</td>
</tr>
<tr>
<td>Reviewer’s Report</td>
<td>Aleksey Zimin</td>
<td>03 Dec 2012</td>
</tr>
<tr>
<td>Resubmission - Version 3</td>
<td>Manuscript</td>
<td>Author’s comment</td>
</tr>
<tr>
<td>Resubmission - Version 6</td>
<td>Manuscript</td>
<td>10 Dec 2012</td>
</tr>
<tr>
<td>Editorial acceptance</td>
<td></td>
<td>10 Dec 2012</td>
</tr>
</tbody>
</table>
Reward open & transparent review

Post publication: bloggers pull apart code/reviews in blogs + wiki:

Main Program

main.c (557 lines)

- Main program. It processes input options, and invokes various other functions (call_prograph, call_heavygraph, call_align, call_map2contig, and call_scaffold) to assemble a genome.

Step 1. (constructing pregraph)

pregraph.c (229 lines)

- Constructs pregraph. From BGI's description -

```
The main function for pregraph step. Its processes are as below:
1. Builds the kmer hash sets and remove the low coverage kmers.
2. Removes the tips which length are no greater than 2*K.
4. Maps the reads back to edges and build preArcs (the connection between edges).
```

Related:

cutTipPreGraph.c (639 lines)

output_pregraph.c (112 lines)

Step 2. (building contigs)


Homologus blogs: http://www.homolog.us/blogs/category/soapdenovo/
SOAPdenovo2 workflows implemented in galaxy.cbiit.cuhk.edu.hk
SOAPdenovo2 workflows implemented in galaxy.cbiit.cuhk.edu.hk

Implemented entire workflow in our Galaxy server, inc.:

- 3 pre-processing steps
- 4 SOAPdenovo modules
- 1 post processing steps
- Evaluation and visualization tools

Also will be available to download by >36K Galaxy users in galaxy.cbiit.cuhk.edu.hk
Table 2 Assemblies of *S. aureus* and *R. sphaeroides*

<table>
<thead>
<tr>
<th>Species</th>
<th>Version</th>
<th>Contigs</th>
<th></th>
<th>Scaffolds</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Number</td>
<td>N50 (kb)</td>
<td>Errors</td>
<td>N50 corrected (kb)</td>
</tr>
<tr>
<td><em>S. aureus</em></td>
<td>SOAPdenovo1</td>
<td>79</td>
<td>148.6</td>
<td>156</td>
<td>73</td>
</tr>
<tr>
<td></td>
<td>SOAPdenovo2</td>
<td>80</td>
<td>98.6</td>
<td>25</td>
<td>71.5</td>
</tr>
<tr>
<td></td>
<td>ALLPATHS-LG*</td>
<td>37</td>
<td>149.7</td>
<td>13</td>
<td>117.6</td>
</tr>
<tr>
<td><em>R. sphaeroides</em></td>
<td>SOAPdenovo1</td>
<td>2,242</td>
<td>3.5</td>
<td>392</td>
<td>2.8</td>
</tr>
<tr>
<td></td>
<td>SOAPdenovo2</td>
<td>721</td>
<td>18</td>
<td>106</td>
<td>14.1</td>
</tr>
<tr>
<td></td>
<td>ALLPATHS-LG*</td>
<td>190</td>
<td>41.9</td>
<td>31</td>
<td>36.7</td>
</tr>
</tbody>
</table>

All datasets were downloaded from [http://sra.ebi.ac.uk/sra/?read=1&logid=46](http://sra.ebi.ac.uk/sra/?read=1&logid=46).
Taking a microscope to peer review

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

PLOS Medicine
The SOAPdenovo2 Case study
Subject to and test with 3 models:

Types of resources in an RO
- Data
- Method/Experimental protocol
- Findings

Models to describe each resource type
- ISA-TAB/ISA2OWL
- Wfdesc/ISA-TAB/ISA2OWL
- Nanopublication
Lessons learned:

• Most published research findings are false. Or at least have errors.

• On a semantic level (via nanopublications) discovered 4 minor errors in text (interpretation not data)

• Is possible to push button(s) & recreate a result from a paper

• Reproducibility is COSTLY. How much are you willing to spend?

• Much easier to do this before rather than after publication
"Regular" Journal

"Conscientious" Online Journal

"Deconstructed" Journal
“Regular” Journal

“Conscientious” Online Journal

“Deconstructed” Journal
“Regular” Journal

“Conscientious” Online Journal

“Deconstructed” Journal
Help us make it happen!

Give us data, papers & pipelines*

Contact us:

scott@gigasciencejournal.com
editorial@gigasciencejournal.com
database@gigasciencejournal.com

* APC’s currently generously covered by BGI until 2015

www.gigasciencejournal.com
Thanks to:

team:
Peter Li
Huayan Gao
Chris Hunter
Jesse Si Zhe
Nicole Nogoy
Laurie Goodman
Amye Kenall (BMC)

Our collaborators:
Ruibang Luo (BGI/HKU)
Shaoguang Liang (BGI-SZ)
Tin-Lap Lee (CUHK)
Qiong Luo (HKUST)
Senghong Wang (HKUST)
Yan Zhou (HKUST)

Case study:
Marco Roos (LUMC)
Mark Thompson (LUMC)
Jun Zhao (Lancaster)
Susanna Sansone (Oxford)
Philippe Rocca-Serra (Oxford)
Alejandra Gonzalez-Beltran (Oxford)

Funding from: CBIIT

Our collaborators:

Case study:

www.gigadb.org
galaxy.cbiit.cuhk.edu.hk
www.gigasciencejournal.com