

Opening peer-review: our new paper on SOAPdenovo2 shows how it works

Scott Edmunds 

Published December 27, 2012

Citation

Edmunds, S. (n.d.). In *GigaBlog*. GigaBlog. <https://doi.org/10.59350/jrf70-d5t78>

Keywords

Open Access, Data Citation, GigaDB, GigaScience, Open Peer Review
Feature Image

Copyright

Copyright © Scott Edmunds 2012. Distributed under the terms of the [Creative Commons Attribution 4.0 International License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

With everyone in a reflective mood as the year comes to a close, one of the big scientific trends of 2012 has obviously been the high profile that open-access and more open methods of carrying out science has received. With the [Elsevier boycott](#), UK [Finch report](#), and launch of a number of innovative new schemes in publishing open-access research and data (including [F1000 Research](#), [eLife](#), [PeerJ](#) and of course [GigaScience](#)), 2012 has been talked of as the year of an "academic spring" that has started to shake up the centuries old, stuffy and closed system of scientific discourse.

On top of changes to the way scientists and readers are demanding they can access and mine the literature and data, new incentives and mechanisms to release and publish data (of which we have [written extensively](#)), the process of peer-review has also come under the spotlight, and there has been a lot of talk on the [deficiencies of this system](#). Many newly launched journals have tried to make the system more transparent, using systems such as post-publication peer-review (e.g. [F1000 Research](#)), pre-print servers (e.g. the increasing acceptance of [arXiv](#) in biology), providing access to anonymized (e.g. [EMBO journals](#)) or partial ([eLife](#)) parts of the peer-review history, or encouraging reviewers to opt-into open peer-review ([PeerJ](#), and [experimented with](#) a little at [PLOS](#)). At [GigaScience](#) we have decided to take this process one step further and ask for [open peer-review as default](#), and as our aims are to promote more open, reproducible and transparent-science we feel it promotes accountability, fairness, and importantly gives credit to reviewers for their hard efforts. A [new publication](#) in the journal today provides a particularly useful example of how this process has worked, so we have decided to highlight it here in [GigaBlog](#), and would welcome feedback and comments on our approach.

What is SOAPdenovo2?

Today we [publish](#) an updated version of BGI's popular SOAPdenovo software application (the [original version](#) having 460 citations according to [googlescholar](#)), a state of the art tool for *de novo* genome assembly. *De novo* assembly – piecing together genomes from sequencing data without the aid of a previously assembled reference, is a particularly computationally intensive and technically challenging task. BGI and their SOAPdenovo tool has been particularly adept in this area, using it to assemble hundreds of new plant and animal species genomes, as well as finding huge amounts of [previously undetected structural changes](#) when applied to individual human genomes. *De novo* genome assembly is an important and competitive area in bioinformatics, and there have been a number of assembly competitions and genome assembler "bake offs" to compare and benchmark the various applications and methods available for carrying this out, the [Assemblathon](#) and [GAGE](#) assembly competitions and evaluations being notable examples of this.

New developments in version 2 of SOAPdenovo have focused on using more efficient algorithms and data structures to reduce the memory requirements, better optimizing and handling of errors and low coverage or heterozygous regions, as well as improved closing of gaps. To demonstrate the improvements and that the application truly is the state-of-the-art for *de novo* assembly of large vertebrate genomes, the authors reassembled BGI's [YH Asian reference genome](#) with the new and original versions of SOAPdenovo, version 2 producing contig sizes 3 times larger, and with nearly two thirds of the maximum memory consumption. Doing

comparisons against other state of the-art assemblers such as ALLPATHS-LG, SOAPdenovo2 outperformed them for many metrics on the Assemblathon and GAGE benchmark datasets tested, really showcasing and demonstrating the potential power and utility of this new application for the bioinformatics community.

Open peer-review, *GigaScience* style

Stating that SOAPdenovo2 can perform better than other state-of-the-art assembly tools is one thing, but to justify and prove this review and testing by independent peers is needed, and the larger and more complicated an application is (particularly an issue for us being a journal that focuses on data heavy research studies), the more challenging this can be. In order to ease, throw light and credit the reviewers in this process *GigaScience* uses a much more transparent, accountable and open peer-review process. Tailoring the process for such data heavy studies our [criteria for publication](#) is based more on relative amount of data created or used, and transparency and availability more than subjective and unpredictable measures such as supposed "impact". For software and methods papers what is being presented obviously has to be an improvement on what is currently available, but for scenarios such as genome assembly where there is obviously no "one-size-fits-all" solution, assessment has to be based on the new method being an improvement in at least one potential application.

During peer review we host all of the supporting information and data (totaling 78GB in this case) and our curators work and make all of it available to the peer-reviewers from our ftp servers. In this case we worked with three groups of expert reviewers (8 independent experts in total) who thoroughly tested the software against various tools and datasets provided to ensure the claims made by the authors were correct. On top of providing all of the test data and scripts and tools that support the paper, to aid the process the authors also provide detailed pipelines with the tools and configured packages including commands and necessary utilities to reproduce the different tests carried out in the paper.

Whilst used in a number of medical journals, almost unprecedentedly in biology we ask as default all of the reviewers to carry out open peer-review, and in this case all 8 of them consented and signed their names to the reports that are now [available to view](#) from the pre-publication history section associated with our published articles. To see how this looks you can follow the history of the SOAPdenovo2 paper [here](#).

While we do have the option for reviewers to opt-out and anonymize their reports if they have concerns about this process, it is encouraging that for all of the papers we have reviewed so far none have asked to do this. A number of new journals are starting to encourage reviewers to sign their reports, but this is the default option for *GigaScience*, with the option of opting out if the referees have reasons to remain anonymous. We also give the reviewers the option of making confidential comments to the editors (particularly on ethical and policy issues), but so far the quality of the reports has generally been very constructive, and [previous studies](#) on open peer-review have also found that quality and courteousness of reviews were increased, with little if any negative effects. By making the process more open and transparent competing interests and biases are reduced, and reviewers are able to take credit for the hard efforts they have put into the review process, and even declare and include it in their CV if they wish as we

would like to put the content of accepted papers reviews under a [CC-BY license](#). The benefits of this increased transparency to readers are also useful, as they do not have to take it on trust that published manuscripts were reviewed by qualified reviewers, and for educational purposes they can see good examples of how peer review operates.

Promoting reproducibility, *GigaScience* style

On top of boosting transparency and reproducibility of peer-review of data-heavy studies, *GigaScience* also carries this over to the publication process, and this paper is also an excellent example of this goal. On top of SOAPdenovo2 meeting our requirements of being open source and having its code in a repository ([sourceforge](#)), the authors also provide detailed pipelines with the tools and configured packages including commands and necessary utilities to reproduce the different tests carried out in the paper. With 78GB of test data and 30MB of tools and scripts being much larger than any other journal is able to handle, we have made all of these available from our [GigaDB database](#) as separate citable DOIs. Taking this a step further, on top of being able to be downloaded by ftp and our [Aspera](#) license (allowing up to 10-100X faster access), the software and analyses are also currently being integrated into our Galaxy-workflow system based data platform.

While we have previously published software articles and pipeline studies combining reference datasets and tools before (see our [methyloome pipeline paper](#) with 84GB of [supporting information](#)), this is the first paper that we have given separate DOIs to the [tools](#) and [data](#). The logic for doing this is that both can now be credited to potentially different groups of authors, the data and methods/analyses may be used and cited independently of each other, and each can be tracked and credited to each author via DOIs listed in their [ORCID](#) account. We feel that it is important to credit method as well as data production, and while [DataCite](#) currently recognizes "Software" as a resource type, we are encouraging and working with them to add "Workflow" to their list of handled objects.

Work is ongoing on the workflow and data platform side, but we are currently in the process of reviewing a number of other software articles (called [Technical Note](#) in *GigaScience*), and if you have similar studies you are interested in having reviewed in a more transparent and constructive manner please contact us at editorial@gigasciencejournal or submit it through our [online submission system](#). As process is still currently evolving and being fine-tuned we would welcome any feedback via this blog, [twitter](#) or email. We would like to thank the team of reviewers of [this](#) and our other manuscripts so far for their hard efforts, as well as the authors for being so helpful in making their work and data available in such a reproducible manner. Many journals are tentatively starting to experiment going down a partially more open route, but from our positive experiences so far we would encourage them and others to be more bold and go all of the way.

Further Reading

1. Luo R et al., SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler *GigaScience* 2012, **1**:18
2. van Rooyen S et al., Effect of open peer review on quality of reviews and on reviewers'

recommendations: a randomised trial. *BMJ* 1999, **318**:23-7

3. Walsh E et al., Open peer review: a randomised controlled trial. *Br J Psychiatry* 2000, **176**:47-51.

4. 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>

5. 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>

UPDATE 24th Jan 2013: we have [produced an editorial](#) on our peer-review policies based on this blog and the feedback we received on it. Also check out the great work the Homolog_us blog has done [testing and studying SOAPdenovo2](#) making the source-code even more transparent via this wiki: <http://homolog.us/wiki1/index.php?title=SOAPdenovo2>

The post [Opening peer-review: our new paper on SOAPdenovo2 shows how it works](#) appeared first on [GigaBlog](#).