Reviewer’s report

Title: SOAPdenovo2: Empirically improved memory-efficient short-read de novo assembler

Version: 1 Date: 22 August 2012

Reviewer: Aleksey Zimin

Reviewer’s report:

This paper describes improvements to the original SOAPdenovo genome assembly software by BGI. The improvements are rather minor and I am not convinced that they will result in significant improvements when the assembler is used on real-life data sets.

Based on my comments listed below, the paper in its current shape requires much more than a major revision. The authors need to present more and different kinds of results to merit the publication, as well as clarify the presentation. The paper needs to be rewritten and resubmitted. Therefore I recommend that the paper is rejected at this time.

Major compulsory revisions.

1. Authors compare the performance of the SOAPdenovo2 to SOAPdenovo1 and Allpaths-LG on faux data set from Assemblathon 1. Faux data usually is much easier to assemble than the real data and the evaluation presented in the subsequent works (e.g. GAGE assembly competition by Salzberg et al., 2011) painted completely different picture. The comparisons of assemblers using faux data have no practical value for determining real-life usability and performance in de novo genome assembly projects. Papers that present genome assembly software must demonstrate the performance of the software on real-life data sets for which finished sequence exists. For example one can use data sets from the GAGE project (Salzberg et al., 2011). The data sets are available at http://gage.cbcb.umd.edu. Alternatively, one can download data for mouse B6 genome available at SRA. Then authors can create an assembly, compare it to the finished sequence and clearly comment on the contiguity, connectivity and correctness of the assembly. For example one can split the assembly at the locations of all misassemblies and compare the resulting N50 size to the N50 size of the original assembly for both contigs and scaffolds. A comparison to another major software package (such as Allpaths-LG) on the same data set would be a big plus.

Therefore I ask that authors demonstrate the performance of the SOAPdenovo2 on a real mammalian (one chromosome is sufficient) genome data set and compare the resulting assembly to the finished sequence to evaluate the performance of the assembler.

2. The performance of SOAPdenovo2 was evaluated against SOAPdenovo1 using YH chomorome data. However, the data used for the two assemblies was
different, the new assembly with SOAPdenovo2 has more data in it. Yes, the assembly is better, but would SOAPdenovo1 generate the same improvements with the additional data?

Minor essential revisions
1. This paper is written in a style of more of a technical report than a structured scientific publication. Material is presented in haphazard fashion. I would call this paper an initial draft that would require more work to be worthy of a scientific publication. I suggest that the authors divide their paper into more common Introduction -- Results -- Methods -- Discussion sections, or alike.

2. Authors should clarify what improvements were made to the GapCloser module described on page 4. Are the improvements described implemented in the main assembler code of in the GapCloser? It is not clear from the text.

Level of interest: An article of limited interest

Quality of written English: Needs some language corrections before being published

Statistical review: No, the manuscript does not need to be seen by a statistician.

Declaration of competing interests:
No competing interests.