MICHIGAN STATE
UNIVERSITY

COLLEGE OF ENGINEERING

Department of Computer Science and Engineering

Michigan State University
3115 Engineering Building
East Lansing, Michigan
48824-1226

(517) 353-3148
FAX: (517) 432-1061

May 19, 2012

Dear Sirs,

We would like the editors to consider our research article entitled "A reference-free algorithm for computational normalization of shotgun sequencing data" for publication in PLoS One.

In this manuscript, we address computational challenges created by the vast genome-scale data sets that are quickly and easily produced by "next-generation" short-read sequencing machines. The specific challenge we address is that of reducing the redundancy and data set size of vast, short-read data sets such as those produced by Illumina.

Our key finding is the development and demonstration of a conceptually very simple computational algorithm for discarding redundant sequence without losing information, called digital normalization. This algorithm takes advantage of the massive redundancy present in shotgun sequencing data sets by selectively subsampling sequencing data based on abundance in the data set. The approach is very computationally convenient: single-pass, streaming, relatively low memory, and fast, as well as potentially parallelizable. Moreover, it converts several sequence analysis approaches into approaches that scale with the size of the source genome, rather than with the size of the data set – a significant achievement in the era of deep short-read sequencing.

In support of our conclusions, we demonstrate that our implementation of this approach loses virtually no information while discarding the majority of errors, and results in "normalized" coverage. Moreover, this approach yields nearly identical results for *de novo* assembly of a variety of different data sets, ranging from microbial genomes to transcriptomes, with a substantial reduction in data size (70-95%) and correspondingly reduced compute requirements. On some single-cell sequencing samples, we also see improvement of assembly contiguity and content after digital normalization.

The only similar work for scaling *de novo* assembly approaches is in the area of error correction (the first publication below), which uses a conceptually different algorithm. Two of the most recent publications in *de novo* sequence assembly that are relevant to this work are listed below; note that in our paper, we use data sets from these two papers to demonstrate the effectiveness of our technique.

Kelley DR, Schatz MC, Salzberg SL. *Quake: quality-aware detection and correction of sequencing errors.* Genome Biol. 2010;11(11):R116. Epub 2010 Nov 29. PubMed PMID: 21114842.

Chitsaz et al, *Efficient de novo assembly of single-cell bacterial genomes from short-read data sets*. Nat Biotechnol. 2011 Sep 18;29(10):915-21. PubMed PMID: 21926975.

Grabherr MG et al, *Full-length transcriptome assembly from RNA-Seq data without a reference genome*. Nat Biotechnol. 2011 May 15;29(7):644-52. doi: 10.1038/nbt.1883. PubMed PMID: 21572440.

Within the field of genomics, our approach dramatically increases the usability of deep sequencing for genome-scale analysis of environmental microbes and non-model organisms. By significantly reducing computational barriers for *de novo* assembly, we not only make existing data sets much easier to analyze but make it possible to analyze much larger, deeper, and more sensitive (e.g. multi-tissue) data sets from transcriptome sequencing. Our technique will also apply to metagenomic studies from the Human Microbiome Project and JGI soil sequencing. Since our approach is implemented as a preprocessing function on data sets, it can be used with any existing assembly pipeline. Our approach is already open source and freely available, with several online tutorials.

More broadly, this approach provides significant leverage to the problem of efficiently using data from the recent sequencing bonanza in biology. While this paper does not directly address biomedical or environmental problems, sequence assembly is important for both; indeed, the need to analyze sequence from evolutionarily or environmentally interesting non-model organisms is our primary motivation for tackling the problem in the first place. By lowering practical barriers to sequence analysis, we hope to further democratize sequence analysis.

The paper also has some potentially interesting consequences for sequence analysis. While we frame the paper primarily in terms of *de novo* sequence assembly, we also introduce an efficient way to look at extremely large sequence data sets as a "streaming" or "online" problem. This concept, in which potentially infinite data streams must be searched for interesting signal, is central to the broader field of "Big Data" analysis in other fields; by introducing a streaming framework for detecting sequence novelty, we believe we can provide leverage on a variety of other bioinformatic problems, including resequencing analysis and heterozygosity resolution. Thus, we believe this work opens up many new avenues for investigation.

Any PLoS One Academic Editors that deal with sequence assembly would be appropriate, such as Jason E Stajich, Jan Aerts, Steven L. Salzberg, or Haixu Tang.

Sincerely,

C. Titus Brown (corresponding/first author)
Assistant Professor
Computer Science and Engineering /
Microbiology and Molecular Genetics
Michigan State University