Dear Editors,

We would like the editors to consider our manuscript entitled "Illumina Sequencing Artifacts Revealed by Connectivity Analysis of Metagenomic Datasets" for publication in *PLOS ONE*.

In this manuscript, we demonstrate the presence of sequencing artifacts in metagenonomic datasets from the human gut, rumen, and soil habitats. These sequences within an assembly graph are very highly connected and are minimally assembled into contigs. We demonstrate this by showing that the removal of these sequences prior to assembly results in similar assembly content, and importantly, allows for the partitioning of disconnected portions of a metagenomic assembly graph. As the computational requirements of assembly are directly related to the number and diversity of sequences within a dataset, the ability to remove artificially highly-connective sequences and subdivide the graph into partitions are critical steps towards enabling the scaling of de novo metagenomic assembly for very diverse environmental datasets

This work extends the previous literature of known sequencing artifacts in previous publications to characterize sequences which cause position specific biases in metagenomic reads.

> Gomez-Alvarez V, Teal TK, Schmidt TM (2009) Systematic artifacts in metagenomes from complex microbial communities. The ISME Journal.
>
> Keegan KP, Trimble WL, Wilkening J, Wilke A, Harrison T, et al. (2012) A platform-independent method for detecting errors in metagenomic sequencing data: Drisee. PLoS Comput Biol.
>
> Niu B, Fu L, Sun S, Li W (2010) Artificial and natural duplicates in pyrosequencing reads of metagenomic data. BMC Bioinformatics.

De novo metagenomic assembly is a developing research area, the following published metagenome-specific assemblers leverage the variable coverage of communities:

> Namiki T, Hachiya T, Tanaka H, Sakakibara Y (2012) MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. Nucleic Acids Research.
>
> Peng Y, Leung HCM, Yiu SM, Chin FYL (2011) Meta-IDBA: a de novo assembler for metagenomic data. Bioinformatics.

The presence of sequencing biases which we have identified in these metagenomic datasets would appear as artificially high coverage regions and result in much larger assembly requirements and potentially erroneous assemblies. We show that the removal

of highly connective sequences from several metagenomes results in comparable assemblies for these assemblers and others.

To review this research article, we suggest:

PLOS ONE Editors:
Michael Watson, The Roslin Institute, mick.watson@roslin.ed.ac.uk
Bas E. Dutilh, Radboud University Medical Center, dutilh@cmbi.ru.nl
Haixu Tang, Indiana University, hatang@indiana.edu

External Reviewers:
Mihai Pop, Associate Professor, U. Maryland, mpop@umiac.umd.edu
Michael Schatz, Cold Spring Harbor Laboratory, mschatz@cshl.edu
Francis Chin, University of Hong Kong, chin@cs.hku.hk

Sincerely,

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