Automated Single Cell Data Decontamination Pipeline

Kristin Tennessen¹*, Amrita Pati¹

¹ LBNL - Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, USA

*To whom correspondence should be addressed: Email: KTennessen@lbl.gov

March 21, 2014

ACKNOWLEDGMENTS:

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. Integration & Validation: Tanja Woyke, Scott Clingenpeel, Natalia Ivanova, Patrick Schwientek, Stephan Trong, James Han, Nikos Kyrpides.

DISCLAIMER:

LBNL: This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California.
**INTRODUCTION**
Recent technological advancements in single-cell genomics have encouraged the classification and functional assessment of microorganisms from a wide span of the biosphere’s phylogeny. Environmental processes of interest to the DOE, such as bioremediation and carbon cycling, can be elucidated through the genomic lens of these unculturable microbes.

However, contamination can occur at various stages of the single-cell sequencing process. Contaminated data can lead to wasted time and effort on meaningless analyses, inaccurate or erroneous conclusions, and pollution of public databases.

A fully automated decontamination tool is necessary to prevent these instances and increase the throughput of the single-cell sequencing process.

**BACKGROUND**
Screening single-cell datasets for contaminants is currently a very manually-intensive procedure. The processing time for one highly-trained scientists to decontaminate one single cell dataset is several hours.

The manual single cell decontamination procedure contains both homology and feature-based screening procedures, the consensus of which can be used to classify a sequence as contaminated or clean. The process includes blasting ribosomal RNA sequences and protein coding genes, as well as visual analysis of k-mer frequency plots and GC content. These tools are available through the IMG website.

An existing software tool, DeconSeq, automatically removes sequence contaminants. However, not only do all sources of contamination need to be known before runtime, the databases for the contaminants need to be selected as input to DeconSeq.

The Single Cell Data Decontamination Pipeline is a fully-automated software tool which classifies unscreened contigs from single cell datasets through a combination of homology and feature-based methodologies using the organism’s nucleotide sequences and known NCBI taxonomy. The software is freely available to download and install, and can be run on any system.

**MATERIALS & METHODS**
The Single Cell Data Decontamination Pipeline (SCDDP) was developed from analysis of 330 manually screened single cell datasets. These datasets can be broken into two groups: Endophyte (129 datasets) and Microbial Dark Matter (201 datasets).

<table>
<thead>
<tr>
<th>Number of datasets</th>
<th>129</th>
<th>201</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median number of contigs per dataset</td>
<td>95</td>
<td>57</td>
</tr>
<tr>
<td>Median contig length (nt)</td>
<td>1,728</td>
<td>4,023</td>
</tr>
<tr>
<td>Median GC%</td>
<td>58</td>
<td>46</td>
</tr>
<tr>
<td>Median contamination %</td>
<td>14</td>
<td>22</td>
</tr>
</tbody>
</table>

The pipeline classifies unscreened contigs through a combination of homology and feature-based methodologies. The input to the pipeline is the known NCBI taxonomy of the target organism, and the fasta file of the nucleotide contig sequences. The algorithm uses the known NCBI taxonomy of the dataset to classify contigs as Clean, Contaminated, or Undecided. Then, the Undecided contigs are classified as Clean or Contaminated using a k-mer based binning method.

**RESULTS**
Results of automated screening of single cell datasets with the Single Cell Data Decontamination Pipeline vary depending on whether the known NCBI taxonomy can be used to classify contigs with blast homology binning. The pipeline was calibrated for a large specificity rate in order to produce a very clean dataset.

Most of the Endophyte datasets have known taxonomy deeper than domain, and thus are screened using both the blast homology and 5-mer binning tools. The MDM datasets have no known taxonomy deeper than domain, thus only 5-mer binning is used to classify the sequences.

The average complete running time for the pipeline is 12 minutes per 1.5 megabase of sequence data, using 16 cores.

**REFERENCES**
5 Markowitz VM et al. (2014) IMG 4 version of the integrated microbial genomes comparative analysis system, Nucl. Acids Res. 42, D560-D567.

**ACKNOWLEDGMENTS**
The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.