Automated Single Cell Data Decontamination Pipeline

Kristin Tennessen[1*], Amrita Pati[1]

[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.

Automated Single Cell Data Decontamination Pipeline

Kristin Tennessen, Amrita Pati

Integration & Validation: Tanja Woyke, Scott Clingenpeel, Natalia Ivanova, Patrick Schwientek, Stephan Trong, James Han, Nikos Kyrpides

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.[1,2] 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.[4] The process includes blasting ribosomal RNA sequences and protein coding genes, as well as visual analyzation of k-mer frequency plots and GC content. These tools are available through the IMG website.[5]

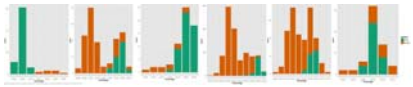
An existing software tool, DeconSeq[3], automatically removes sequence contaminants. However, not only do all sources of contamination need to be known before runtime, the databases for the contaminants need to be selected as input to DeconSeq.

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.

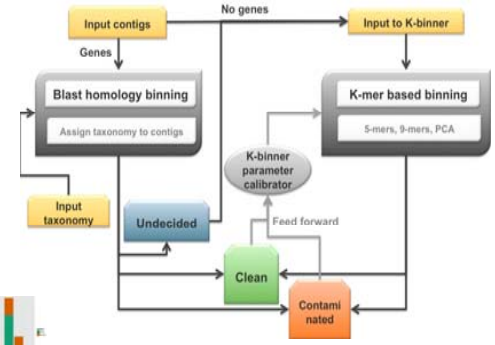
MATERIALS & METHODS

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).

	Endophyte	MDM
Number of datasets	129	201
Median number of contigs per dataset	95	57
Median contig length (nts)	5,728	4,023
Median GC%	59	40
Median contamination %	14	20

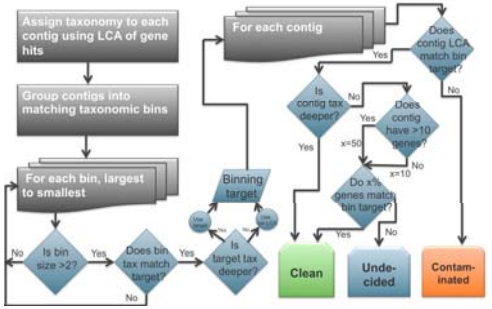


Pipeline Schematic



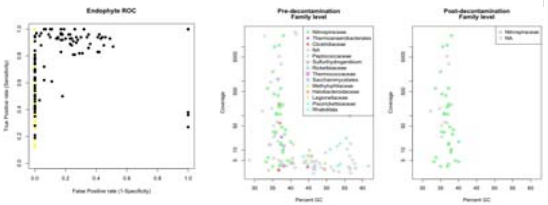
Blast Homology Binning

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 arget 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.



		Endophyte		MDM	
		contig	base	contig	base
Sensitivity	median	0.68	0.82	0.64	0.93
	mean	0.65	0.75	0.66	0.90
Specificity	median	1.00	1.00	0.91	0.67
	mean	0.95	0.93	0.84	0.65

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 9-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

CONCLUSIONS

The Automated Single Cell Data Decontamination pipeline is a valuable tool for preventing the dissemination of contaminated data into public databases, avoiding wasted hours of misleading analysis, and thwarting the publication of erroneous conclusions due to contamination of single cell datasets. The fully automated nature of the pipeline relieves expert scientists of hours of manual screening and produces a reliable, clean dataset.

REFERENCES

[1] Rinke C, Schwientek P, Sczyrba A, Ivanova NN,Anderson IJ, Cheng J-F, Darling A, Malfatti S, SwanBK, Gies EA et al. 2013.Insights into the phylogeny and coding potential of microbial dark matter. Nature 499:431-437

[2] Rex R Malmstrom, Sébastien Rodrigue, Katherine H Huang, Libusha Kelly, Suzanne E Kern, Anne Thompson, Sara Roggensack, Paul M Berube, Matthew R Henn and Sallie W Chisholm. Ecology of uncultured Prochlorococcus clades revealed through single-cell genomics and biogeographic analysis. The ISME Journal (2013) **7,** 184–198.

[3] Schmieder R and Edwards R: Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS ONE* 2011, 6:e17288.

[4] Clingenpeel, Scott. JGI Microbial Single Cell Program Single Cell Data Decontamination. http://img.jgi.doe.gov/w/doc/SingleCellDataDecontamination.pdf

[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 US Department of Energy Joint Genome Institute is supported by the Office of Science of the US Department of Energy under Contract No. DE-AC02-05CH11231

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.