Charcoal: filtering contamination in metagenome-assembled genome bins

Taylor Reiter
Some Institute of Technology
tereiter@ucdavis.edu

Lots of Other People
Another University
cat@example.com

C. Titus Brown
Some Institute of Technology
ctbrown@ucdavis.edu
*

2020-05-22

Abstract

This is the abstract.

It consists of two paragraphs.

Keywords: metagenome-assembled genome bin; contamination; scaled MinHash

1 Introduction

Metagenomic sequencing has revolutionized our understanding of microorganisms. Advances in high-throughput sequencing combined with scalable analysis methods have generated hundreds of thousands of draft genomes, termed metagenome-assembled genomes (MAGs), from culturable and unculturable organisms from diverse environments. These

*Corresponding author; Email: ctbrown@ucdavis.edu

techniques have revealed extreme diversity in microbial life and metabolism, revealing the Candidate Phyla Radiation and other really cool things (CITE: BANFIELD LAB STUFF). Recently, large-scale analysis efforts have demonstrated that *de novo* metagenome analysis techniques like assembly and binning capture a large amount of diversity in some metagenomes (CITE: PASSOLI, NAYFACH, ALMEIDA). These efforts have led to a rapid expansion in draft genomes in public repositories like GenBank (CITATION). Increased observation of draft genomes across the tree of life better enables researches to contextualize new sequencing data and the roles that microorgansisms play in nutrient cycling, disease, etc.

These efforts rely on our ability to correctly assembly and bin metagenomic sequencing data, and to detect contamination when assembly and binning fail. The quality of MAGs has traditionally been assessed by the presence of single copy marker genes specific to an inferred lineage (CITE: CHECKM). While single copy marker excel in the estimation of completeness, other methods may better measure contamination as all contiguous sequences binned into a MAG do not contain marker genes. Long k-mers capture relatedness between organisms, where a k=31 captures species-level similarity (CITE: METAPALETTE). K-mers offer an alternative metric to identify contamination, especially in sequences lacking marker genes. Here we describe Charcoal, an automated method for filtering contaminant contiguous sequences from MAGs. We show... We show... We envisage that charcoal will complement marker gene-based approaches for contamination estimation, removing problematic sequences before they are further analyzed or propagated in public databases.

Here are two sample references: Feynman and Vernon Jr. [1963; Dirac, 1953]. Bibliography will appear at the end of the document.

2 Materials and methods

2.1 Overview

Charcoal identifies and removes contamination in metagenome-assembled genomes using k-mer based methods. Charcoal is developed around the tool sourmash, which computes and compares scaled MinHash signatures for nucleotide sequences based on k-mers (CITE: F1000, JOSS). A k-mer is a nucleotide sequence of length k. When k is long enough, k-mer are generally are specific to a taxonomic lineage (CITE: METAPELLETE, FASTANI). Scaled signatures (MATH ASSOCIATED WITH GARUNTEES ON K-MERS OBSERVED), thus XXX and reducing computational XXX.

Sourmash represents k-mers as hashes.

To identify contamination, charcoal first assigns a majority lineage to the input genome. Charcoal uses sourmash `gather` to identify genome sequences in a database

("reference genomes") that are present in the input genome. The hashes from these reference genomes are assigned a lineage up to the user-specified rank according to the lowest common ancestor (LCA). Using this information, charcoal determines the majority rank by counting the number of times each rank is observed in the hashes of the input genome. Charcoal then defines the fraction of identifiable hashes in the input genome, and the fraction of identifiable hashes that are anchored to the majority lineage. If fewer than 10% of hashes are identifiable, charcoal requests that the user provide a lineage to enable decontamination. While we refer to the most represented rank as the "majority lineage", the winning lineage can represent a majority or a plurality of hashes. Charcoal requests that the user provide a lineage when the majority lineage contains fewer than 20% of identified hashes.

Using the majority lineage, charcoal tests each contig in the input genome for contamination. Charcoal currently removes contamination for three reasons: 1) First, charcoal compares the contig against the database of reference genomes created during majority lineage assignment and determines the best match. If the best match has a different lineage up to the configured rank than that of the majority lineage, the contig is removed as a contaminant. 2) Second, charcoal computes the LCA for all hashes on the contig. If the most common LCA is above the configured rank, the contig is removed as a contaminant. 3) Third, charcoal uses the computed LCA from the previous step and determines whether the lineage differs from the majority lineage at the configured rank. If the lineages differ, the contig is removed as a contaminant.

Optionally, the user can specify a lineage (e.g. `d__Eukaryota`), and charcoal will remove contigs that have a lineage different from the user-specified one. This allows charcoal to remove contigs that are in a database from a genome which is not related to anything in a database. Charcoal reports whether the provided linage agrees with k-mer classification at or above the genus level.

Charcoal outputs clean and contaminant contigs as separate fastq files. Charcoal reports the majority lineage, the fraction of identified hashes and the fraction of identified hashes that match the majority lineage, the number of clean, contaminant, and missed contigs, and the reason each contig was removed. Additionally, charcoal reports the breakdown of known genomes in the clean contigs as estimated by sourmash `gather`, as well as size and lineage of the primary `gather` result.

Importantly, charcoal will not remove a contig if it is unidentifiable. While these contigs could still be contaminants, charcoal assumes contigs are clean for which it has no information. Therefore, charcoal will fail to detect contamination for very short contigs which contain no selected k-mers, as well as contigs with novel DNA content.

Identifying all lineages in the genome with sourmash gather is the most compute intensive step in the decontamination process.

2.2 Availability

Charcoal is written in python3 and can be installed via conda or pip. It depends on sourmash and snakemake. The source code is available at github.com/dib-lab/charcoal.

2.3 Datasets and benchmarking

3 Results

Generate a figure.

You can reference this figure as follows: Fig. 1.

Generate a table.

	ID	code
1	1	a
2	2	b
3	3	c

Table 1: This is the table caption

You can reference this table as follows: Table 1.

4 Discussion

You can cross-reference sections and subsections as follows: Section 2 and Section **??**.

5 THINGS TO ADD

- fastani is an accepted way to do average nucleotide identity calculate, and it relies on k-mers.
- from checkm paper: > Bias in genome quality estimates: Quality estimates based on individual marker genes or collocated marker sets exhibit a bias resulting in completeness being overestimated and contamination being underestimated. This bias is the result of marker genes residing on foreign DNA that are otherwise absent in a genome being mistakenly interpreted as an indication of increased completeness as opposed to contamination.
- we don't deal with strain heterogeneity, as this occurs below the species-level aggregation in the LCA

6 formatting notes

A numbered list:

1) First point
2) Second point
    - Subpoint

A bullet list:

- First point
- Second point

7 Questions for titus:

1. What does REASON 3 actually mean?
2. in `class ContigsDecontaminator`, how does the object transit from REASON 1 checks to REASON 2 & 3 checks? What is a class? How is it being used here?
3. are hashes assigned their LCA lineage? so if it's shared between all strains of a species, it gets the species lineage? and if it's shared in two phyla of bacteria, it gets the "bacteria" lineage?

8 Results outline:

- charcoal estimates low contamination in non-representative GTDB genomes
- charcoal assigns correct taxonomy to all non-representative GTDB genomes
- charcoal vs. checkm
    - charcoal vs. checkm: mgnify, tara
    - checkm on charcoal clean
    - prokka on charcoal dirty
- charcoal vs. refineM and magpurify
- verfication of contamination/contam case studies
    - case studies
    - user-provided lineages
    - cDBG/spacegraphcats?
    - user-specified lineages

Acknowledgements

We thank everyone involved.

Including funding sources.

References

P.A.M. Dirac. The lorentz transformation and absolute time. *Physica*, 19(1-–12):888–896, 1953. doi: 10.1016/S0031-8914(53)80099-6.

R.P Feynman and F.L Vernon Jr. The theory of a general quantum system interacting with a linear dissipative system. *Annals of Physics*, 24:118–173, 1963. doi: 10.1016/0003-4916(63)90068-X.

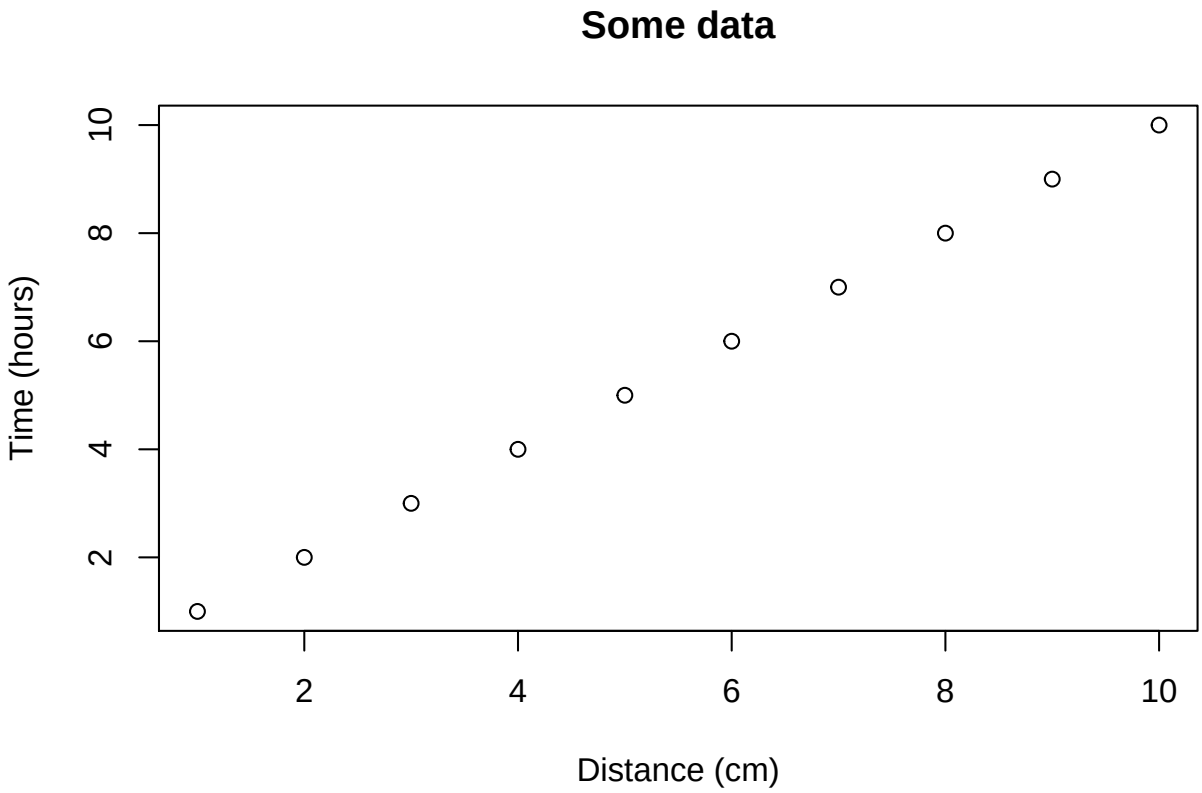

Figure 1: This is the first figure.