Comprehensive evaluation of genomic redundancy and taxonomic coherence in large biological databases

This manuscript (permalink) was automatically generated from dib-lab/2022-paper-genomic-tax-redundancy@1f44b3c on November 19, 2022.

Authors

- **John Doe**
  XXXX-XXXX-XXXX-XXXX · johndoe · johndoe
  Department of Something, University of Whatever · Funded by Grant XXXXXXXX

- **Jane Roe** ✉
  XXXX-XXXX-XXXX-XXXX · janeroe
  Department of Something, University of Whatever; Department of Whatever, University of Something

✉ — Correspondence possible via GitHub Issues or email to Jane Roe <jane.roe@whatever.edu>.

Abstract

A central challenge in bacterial genomics and taxonomy is that genomic databases are increasingly large and contain redundant content. This impacts the accuracy of taxonomic profilers and metagenome analysis tools that rely on these databases. Here, we probe the practical and theoretical limits of genomic identification and taxonomic classification by exploring *unicity distance* and *Shannon entropy*. Unicity distance provides an estimate of how many k-mers are required to precisely identify a single reference genome from within a database, while Shannon entropy of k-mers describes the informativeness of a k-mer for taxonomic classification. We show that approximately 30% of genomes in GTDB rs207 have infinite unicity and that 99% of k-mers can resolve taxonomy at the species, genus, or family level. We conclude that unicity distance and Shannon entropy provide simple metrics for evaluating genomic redundancy and taxonomic coherence of large genomic reference databases.

Introduction

Introduction goes here.

Results

Many k-mers are genome specific

Shannon entropy of k-mers can be used to measure taxonomic informativeness

We measured the species distribution in GTDB rs207 for 21.2 million hashes, representing 21.2 billion 31-mers, and calculated the Shannon entropy of species for each hash (equationXX). Per Table XX, 92.8% of hashes uniquely identify a specific family.

Taxonomic level	# perfectly informative hashes	cumulative % total
species	21,150,287	92.8%
genus	1,262,281	XXX%
family	170,249	YYY%

(Make point that nucleotide k-mers are not necessarily specific beyond family ref protein paper.)

Many k-mers with non-zero entropy come from a few specific genomes

Explore taxonomic incoherence and database contamination.

Unicity distance can be used to estimate genomic redundancy

We next ask, how many genomes can be distinguished from each other using a combinatorial collection of k-mers? To do this, we estimate the *unicity distance* of each genome in the database, where the unicity distance is defined as the smallest set of hashes capable of uniquely identifying a genome. (k=31, scaled=1000)

Table YY shows that approximately 29.2% of the genomes in GTDB rs207 cannot be distinguished uniquely by any combination of k-mers with these parameters.

Unicity distance	Number of genomes	Percent of genomes
1	48,630	15.3%
infinite	92,564	29.2%

Discussion

Discussion goes here.

Methods

Methods go here.

References