How to compare and cluster every known genome in about an hour

Sergey Koren, @sergekoren

Why MinHash?

- Large compression
  - 3 Gbp primate genome
  - 8 kbp vs. 750 MB
  - 5 Tbp of samples
  - 71 MB vs. 1.25 TB

- Fast comparisons
  - Cluster all of RefSeq
  - 46 CPU hours
  - Linear search of RefSeq
  - 1 CPU second

Assembling large genomes with single-molecule sequencing and locality sensitive hashing. Berlin et al. (2015)

Primary overhead is in the sketching, comparisons are instantaneous.

What is a sketch?


Estimating Jaccard with MinHash


Mash distance correlates with ANI

All-pairs comparison of 500 Escherichia genomes
Unsupervised database clustering

RefSeq = 1.5 billion distances in 46 CPU h, sketches <100 MB, linear search in 1s

Whole-genome phylogeny

Each genome = 1,000 values, fasta to phylogeny in <30m on a laptop

Metagenome sample clustering

• 888 HMP and MetaHIT samples (v=10,000, k=21)

4.4 CPU hours (assemblies), 279 CPU hours (reads)

Database search

• Discriminates between B. anthracis and B. cereus
• Bloom filter to remove single-copy k-mers
• Can be used to index/search SRA
• Read mapping on the way (cf. Heng Li's minimap)

POC disease surveillance

Sequencing as a sensor

hint.fm/wind
NOAA NDFD
Sequencing as a sensor

The rise of a digital immune system. Schatz and Phillippy (2012)

Mash preprint on bioRxiv

- Comments welcome

Fast genome and metagenome distance estimation using MinHash. Ondov et al.

- [http://mash.readthedocs.org](http://mash.readthedocs.org)
  - examples and RefSeq database

- Fast distance estimation
  - Database search
  - Rapid species assignment
  - Very large guide trees
  - Sample quality control
  - Metagenome sample clustering

Acknowledgements

- **Mash**
  - Brian Ondov
  - Todd Treangen
  - Adam Phillippy

- **Canu**
  - Adam Phillippy
  - Brian Walenz

- **NHGRI**
  - [GitHub/MarBL](https://github.com/emplbl-genome-informatics/MarBL)

- **Postdocs wanted!**
  - Genome Informatics Section
  - Assembly
  - Structural variation
  - Infectious disease
  - Undiagnosed disease
  - [http://www.genome.gov/27563366](http://www.genome.gov/27563366)