Genome contamination has minimal impact on the delineation of species

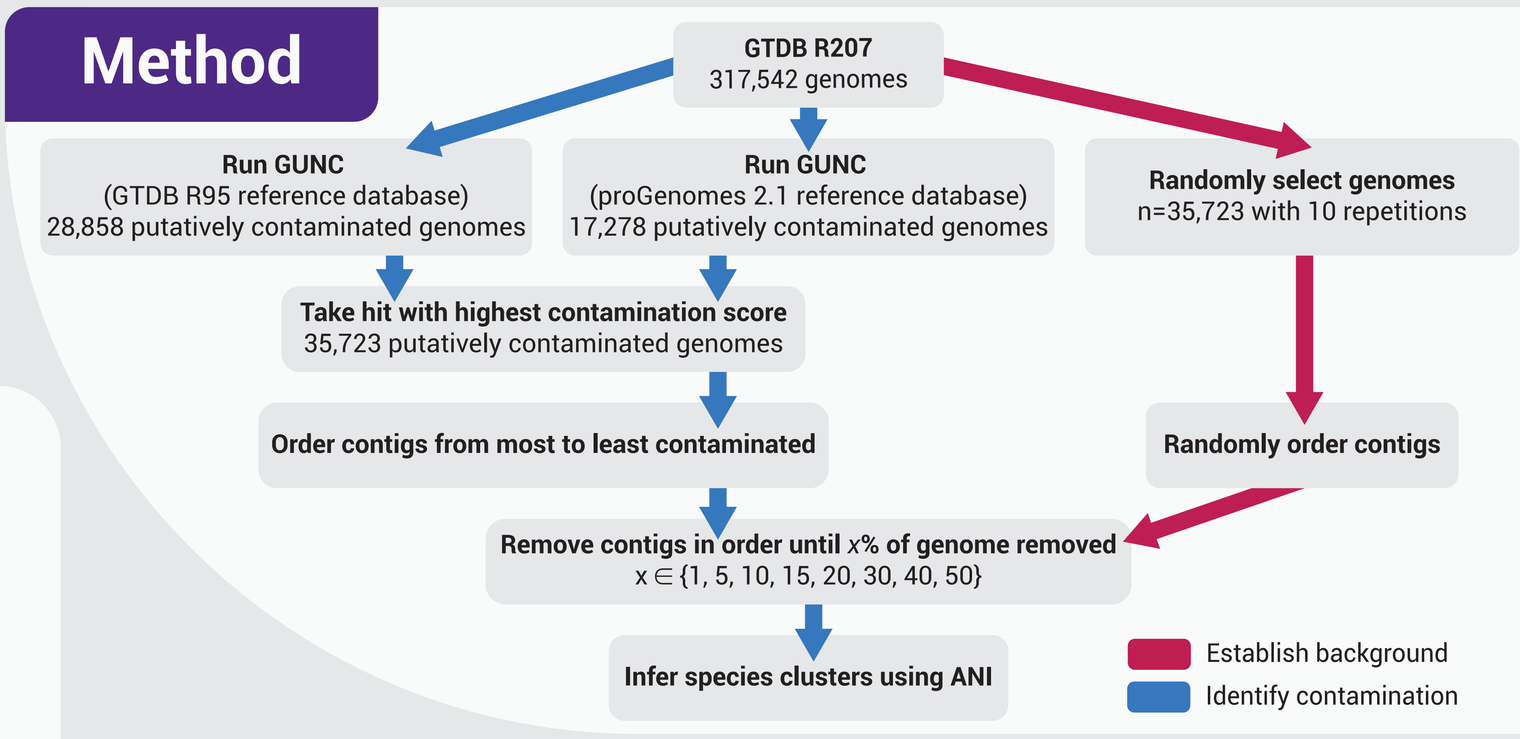
Aaron J. Mussig[1], Pierre-Alain Chaumeil[1], Maria Chuvochina[1], Christian Rinke[1], Donovan H. Parks[1], Philip Hugenholtz[1]

[1] The University of Queensland, School of Chemistry and Molecular Biosciences, Australian Centre for Ecogenomics, St Lucia, QLD, Australia

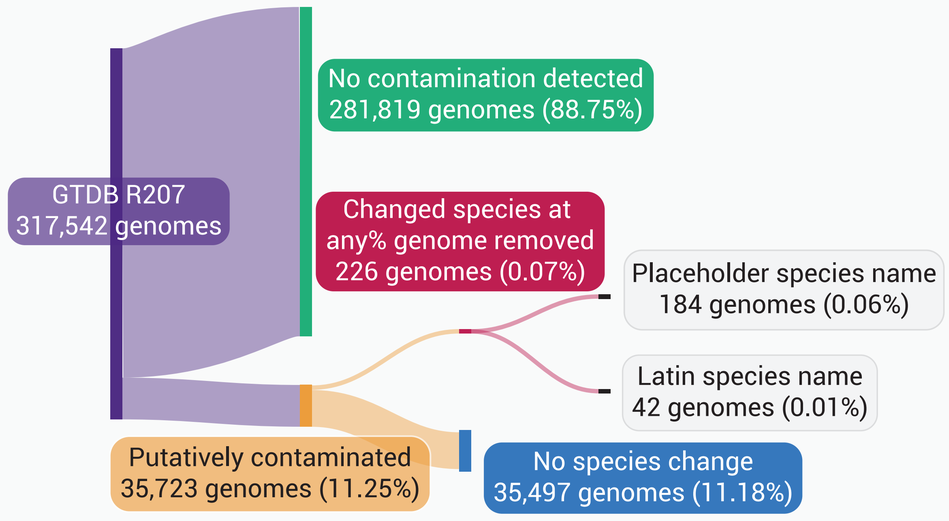
Contamination is a known issue, but to what extent does this impact taxonomy?

We assess the impact of putative contamination on species delineated using average nucleotide identity (ANI).

Method

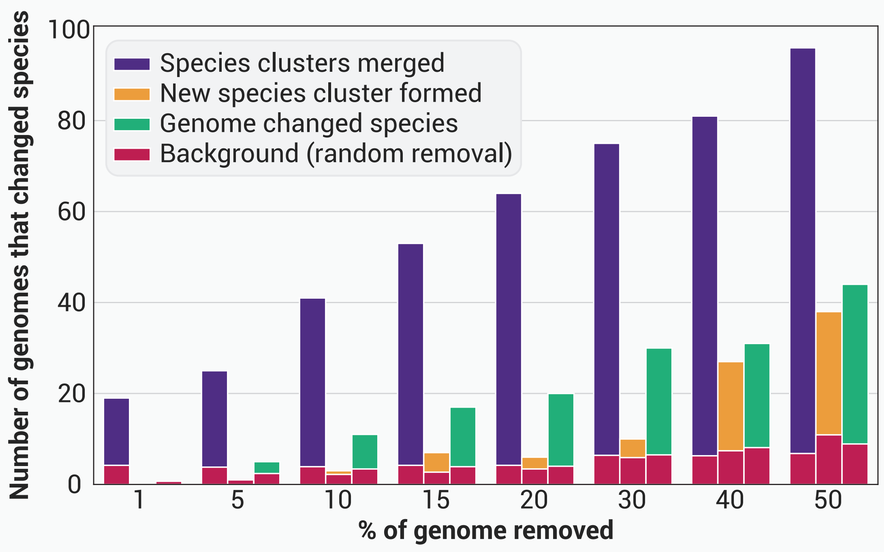


Species clustering is robust to contamination



- 35,723 of 317,542 (11.25%) GTDB R207 genomes were putatively contaminated.
- Identifying contamination using a reference database has limitations:
  i) Taxonomically novel genomes are insufficiently represented.
  ii) Order that genomes appear in the GUNC reference database impacts results.

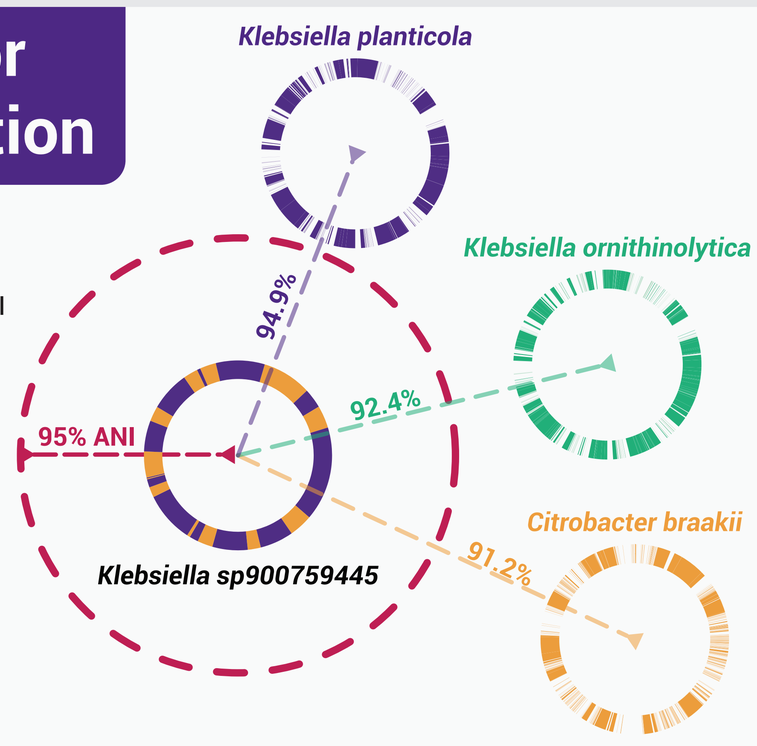
Up to 226 genomes (0.07%) change species assignment when removing putatively contaminated contigs



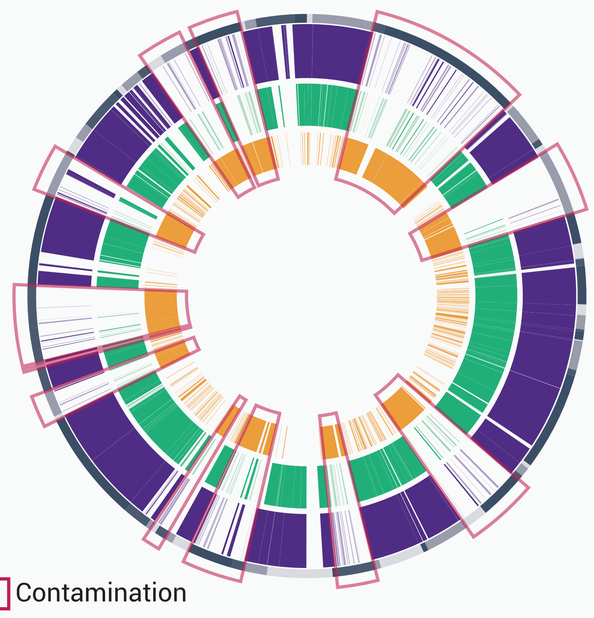
- Even with removing half of the genome, there were only 226 genome that changed species.
- 184 of the genome changes were in nomenclaturely unimportant species (i.e. placeholder species).
- The 226 genomes belonged to 206 unique species clusters from 134 unique genera.

The perfect storm for species reclassification

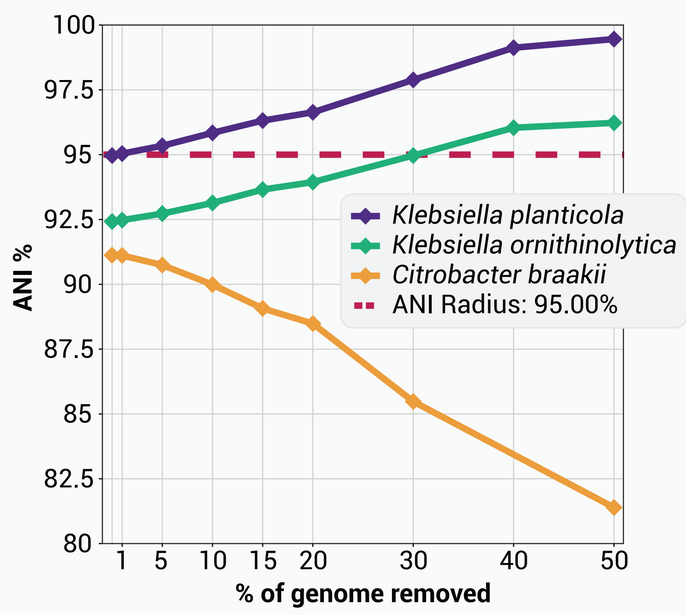
- Chimeric assemblies decrease interspecies ANI to closely related species.
- Removal of foreign contigs increases ANI to neighbouring species clusters.
- If interspecies ANI similarity is extremely close, a species change is more likely.
- Placeholder species are disproportionately affected.



Whole genome alignment shows *C. braakii* contamination



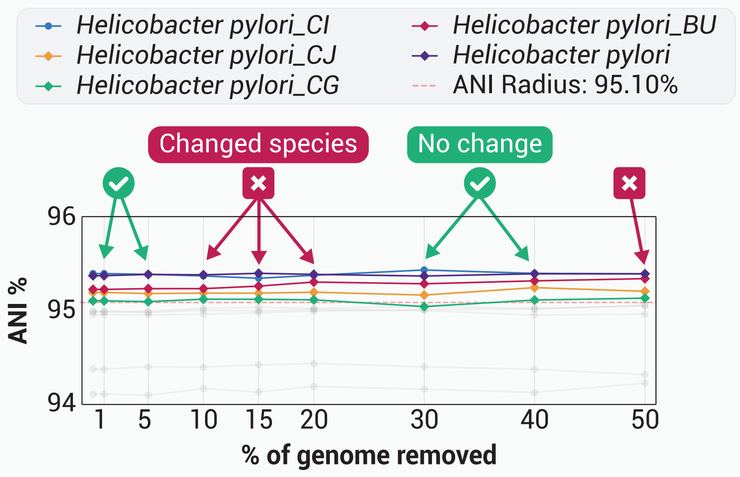
Removing contamination merges *K. sp900759445* and *K. planticola*



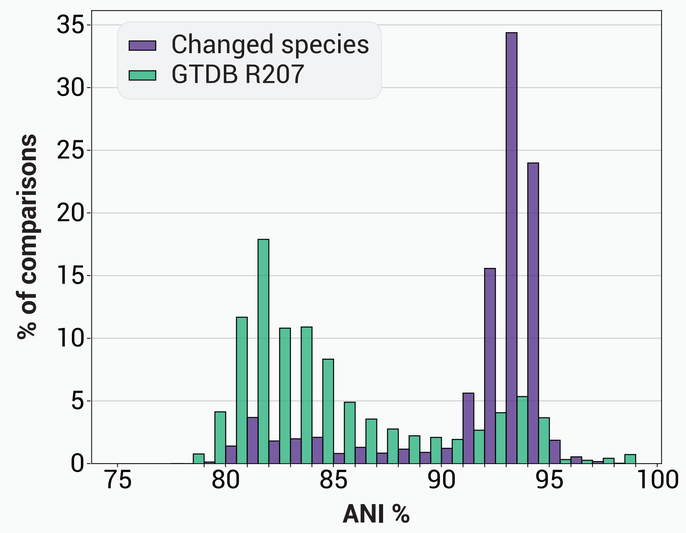
Closely related genomes blur interspecies boundaries

- 76.75% of the interspecies ANI comparisons were >92% ANI for the 226 genomes that changed species, compared to 16.58% for all genomes in GTDB R207.
- Significantly more non-representative genomes fall within multiple species ANI radii.
- Genomes are more likely to have a species reassignment if they are close to multiple representatives.

Species assignments can be on a knifes edge



Interspecies ANI is higher for genomes that changed species



Australian Centre for Ecogenomics

a.mussig@uq.edu.au
@aaronmussig
0000-0002-9988-0866

The University Of Queensland
AUSTRALIA
CREATE CHANGE