Genome filtering and de-duplication for analysis

To acquire genomes for analysis, 7431 summaries of Acinetobacter baumannii reference assemblies (assembly accessions starting with "GCF\_") were downloaded from NCBI datasets (<https://www.ncbi.nlm.nih.gov/data-hub/genome/>) using the datasets command line interface tool:

$ datasets summary genome --annotated taxon 470 --assembly-source refseq > summary\_all.json

Genomes with contig N50 scores in the bottom 20% were removed from the assemblies, resulting in a total of 5945 genomes. Since there were many closely related genomes, a deduplication process was carried out based on Average Nucleotide Identity (ANI) values estimated using Mash distances generated by Mash version 2.34. Each genome was represented by a compressed min-Hash sketch, and pairwise distances between sketches were used to construct a standard distance matrix.

To perform the deduplication, a custom Python script was developed. The script iterated through each genome in the distance matrix, comparing its distances to other genomes. A chosen threshold Mash distance (tdist) of 0.006 was used. If the pairwise distance between two genomes was less than tdist, the genome with the lower N50 score was discarded. This step was performed recursively until all pairwise distances were greater than tdist, ensuring that only unique genomes remained in the dataset.

10 assemblies were identified as outliers due to their high average Mash distance to all other genomes and were excluded (Fig S1D). Scripts for downloading genomes, creating Mash sketches and distance matrices, and dereplicating genomes and removing outliers are available at<https://github.com/JonWinkelman/genome_deduplication>.

Identification of orthologs in *Acinetobacter baumanii*

We utilized Orthofinder1,2 version 2.5.4 with default settings to determine orthologous relationships between genes in A. baumannii genomes. A total of 233 proteomes from filtered and de-duplicated Acinetobacter genomes were included in the analysis. These genomes include three outgroups, A. baylyi, A. gyllenbergii and A. colistiniresistens that were used to root the species tree.

OrthoFinder computed hierarchical orthologous groups (HOGs) for each internal node in the species tree. These HOGs consist of proteins descended from a single gene in the ancestral species corresponding to the respective internal node. For this particular study, we focused on analyzing HOGs associated with the species tree node representing the last common ancestor of all A. baumannii.

OrthoFinder computes hierarchical orthologous groups (HOGs) for each internal node of the species tree. Each HOG contains all of the proteins that descended from a single gene in ancestral species represented by the internal node. In this analysis we analyzed HOGs for the species tree node representing the last common ancestor of all *A. baumannii.* Orthofinder was run on an AWS EC2 instance with the shell script ./aws\_run\_orthofinder.sh.

Manual identification of orthologs in *Acinetobacter baumanii*

When working with three specific genes, Orthofinder's search for orthologs didn't yield results across all species. In cases where Orthofinder didn't locate an ortholog, we adopted an alternative approach. Specifically, we investigated the presence or absence of neighboring genes. For instance, if an ortholog for the Acinetobacter baumannii 17978-mff gene ACX60\_11495 was detected in a strain, orthologs to all other genes in its operon were found. Conversely, when this ortholog was absent, none of the operon's other genes were found.

In another instance involving the Acinetobacter baumannii 17978-mff gene ACX60\_5080, Orthofinder failed to identify orthologs in two strains. In this case, we observed adjacent orthologs and identified a potential ortholog with an unassigned HOG (Hierarchical Orthologous Group) and more than 95% sequence identity to ACX60\_5080. These findings strongly suggested that this identified potential ortholog was indeed an orthologous gene.

Jupyter notebooks for downloading genomes, creating Mash sketches and distance matrices, dereplicating genomes, removing outliers, processing orthofinder results and creating figures are available at<https://github.com/JonWinkelman/Palmer_baumannii_Ast2_prevalence.git>.

1. Emms, D.M., and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biology *16*, 157. 10.1186/s13059-015-0721-2.

2. Emms, D.M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biology *20*, 238. 10.1186/s13059-019-1832-y.

3. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics *10*, 421. 10.1186/1471-2105-10-421.

4. Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S., and Phillippy, A.M. (2016). Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol *17*, 132. 10.1186/s13059-016-0997-x.

5. Edgar, R.C. (2021). High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny (Bioinformatics) 10.1101/2021.06.20.449169.

6. Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312–1313. 10.1093/bioinformatics/btu033.

7. Le, S.Q., and Gascuel, O. (2008). An Improved General Amino Acid Replacement Matrix. Molecular Biology and Evolution *25*, 1307–1320. 10.1093/molbev/msn067.

8. Letunic, I., and Bork, P. (2007). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics *23*, 127–128. 10.1093/bioinformatics/btl529.