Interpretation of all results

The phylogenetic trees constructed from the concatenated alignments of homologous genes and proteins (see Figures 1 and 7) illustrate clustering patterns among E. coli genomes, but do not accurately represent their evolutionary history.

Individual gene and protein trees (see Figures 2-6 and 8-12) highlight gene-specific and protein-specific evolutionary trajectories, capturing variations not evident in the concatenated trees.

Homologous sequences were identified using stringent BLAST cut-offs of e-value < 1e-5 and identity ≥ 96% but < 100%, ensuring high-confidence matches while avoiding identical sequences.

Attempts to create trees from concatenated alignments of all protein and gene homologs failed because the sequences were of different lengths. To address this, all sequences across the homologous sets were aligned together in a single step, which resolved the length issue but introduced potential biases due to combining unrelated sequences. The concatenation approach amplifies shared evolutionary signals but makes it difficult to see the differences between individual genes and proteins, which limits the trees' biological interpretability.

The individual amino acid trees (Figures 8-12) show shorter branch lengths compared to the nucleotide trees (Figures 2-6), reflecting stronger functional constraints on protein evolution.

Individual homologous trees reveal finer-scale evolutionary dynamics, which are masked in concatenated alignments, emphasising the need for a genome-wide or pan-genome approach.

Overall, while the trees provide useful clustering insights, a more careful selection of homologous sets (like different cut offs) and improved concatenation strategies would enhance phylogenetic accuracy and biological relevance.