LCA (Least Common Ancestor)

Least-common ancestor (LCA)

- Problem: we need a computationally efficient way to analyze millions of reads with respect to their taxonomic origin.
- Idea: analyze small subsequences instead of complete sequences; combine information from these small subsequences in a second step.
- Try to make the most specific assignment for each subsquence
- Definition: a k-mer is a sequence of length k
  Example: ACGTT consists of the 3 3-mers ACG, CGT, GTT

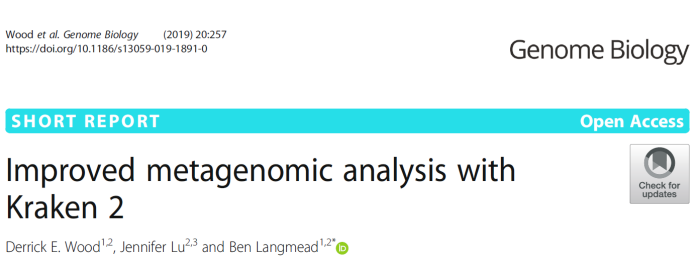

Wood and Salzberg *Genome Biology* 2014, **15**:R46
http://genomebiology.com/2014/15/3/R46

Genome Biology

METHOD Open Access

Kraken: ultrafast metagenomic sequence classification using exact alignments

Derrick E Wood[1,2*] and Steven L Salzberg[2,3]


Wood *et al. Genome Biology* (2019) 20:257
https://doi.org/10.1186/s13059-019-1891-0

Genome Biology

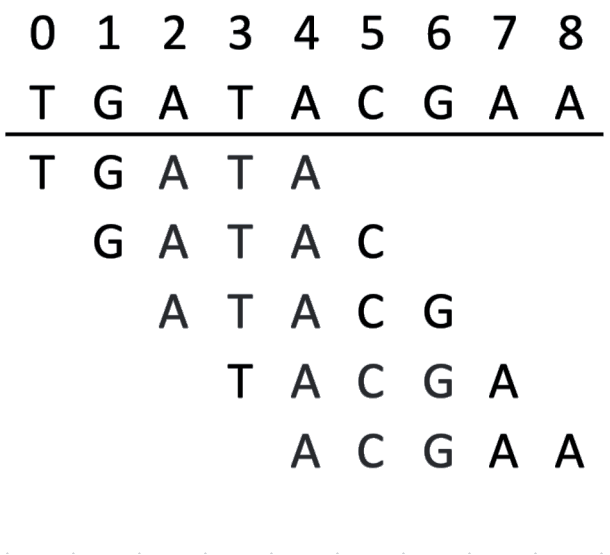
SHORT REPORT Open Access

Improved metagenomic analysis with Kraken 2

Derrick E. Wood[1,2], Jennifer Lu[2,3] and Ben Langmead[1,2*]

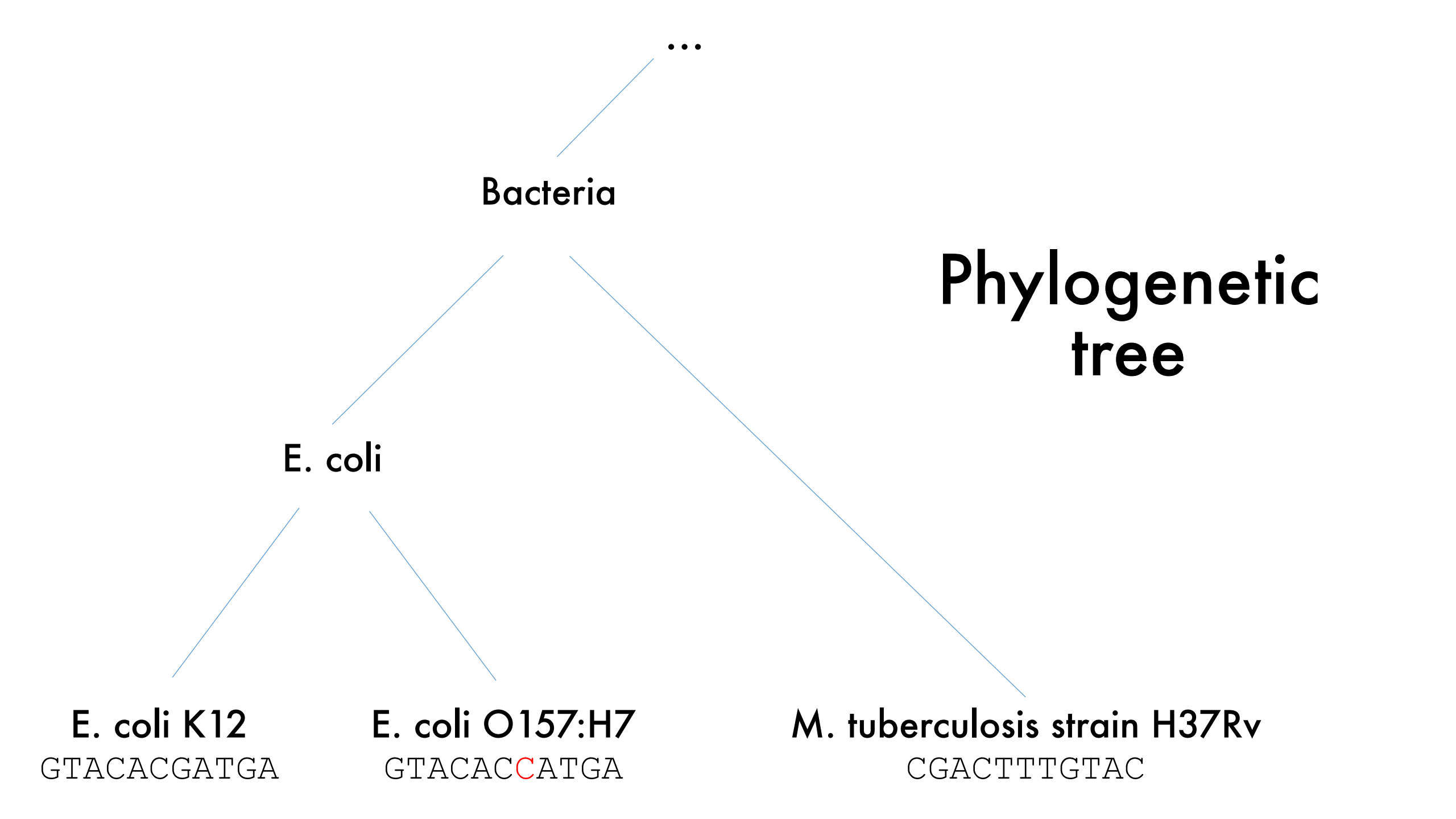
k-mers

- k-mer = Sequence of length k



LCA: 3 example genomes

Species / strain	Reference genome
E. coli K12	GTACACGATGA
E. coli O157:H7	GTACACCATGA
M. tuberculosis strain H37Rv	CGACTTTGTAC



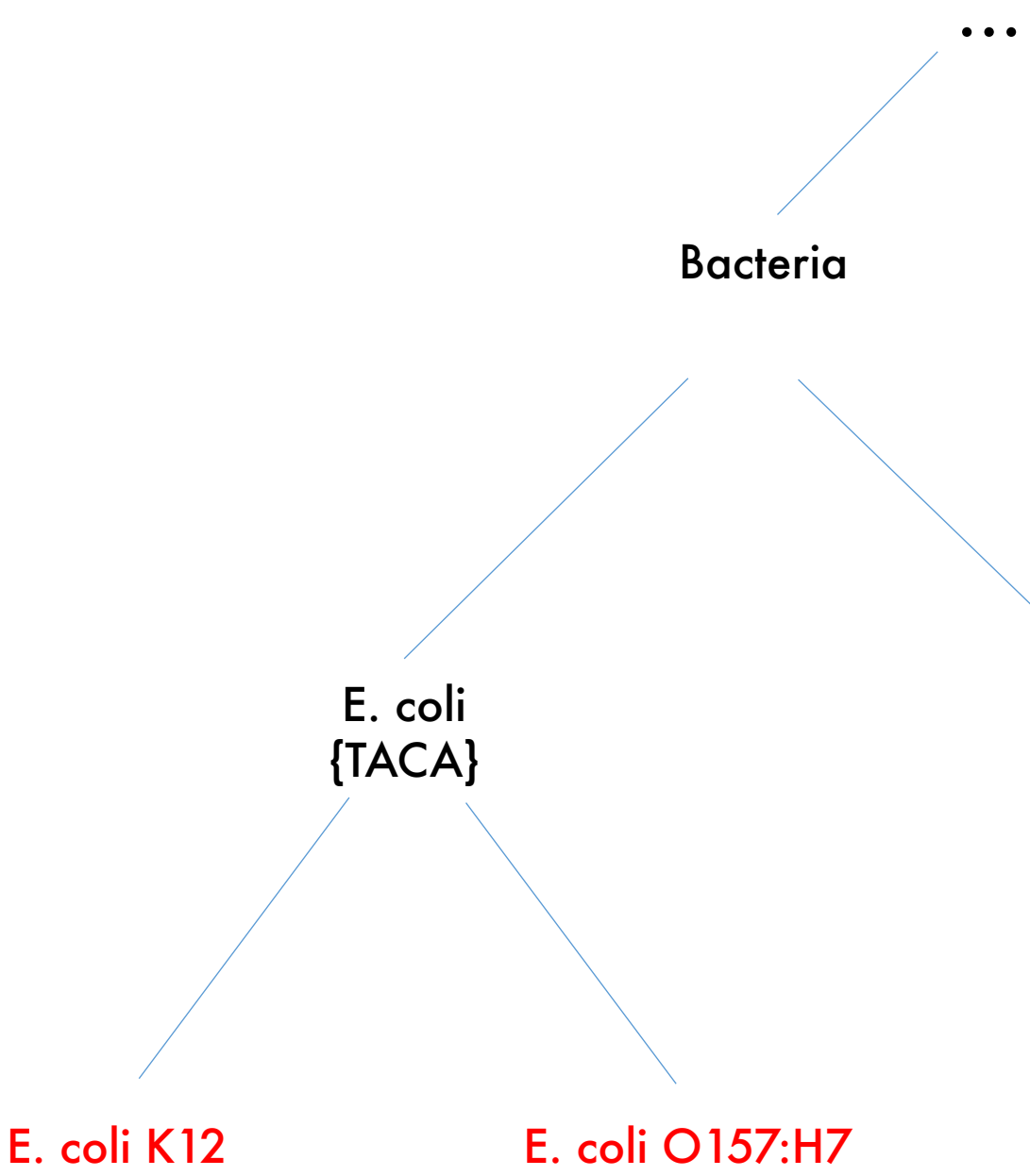
Phylogenetic tree

LCA: Building the database

Split the input genomes into k-mers (e.g. 4-mere) and determine the phylogenetic placement of each k-mer.

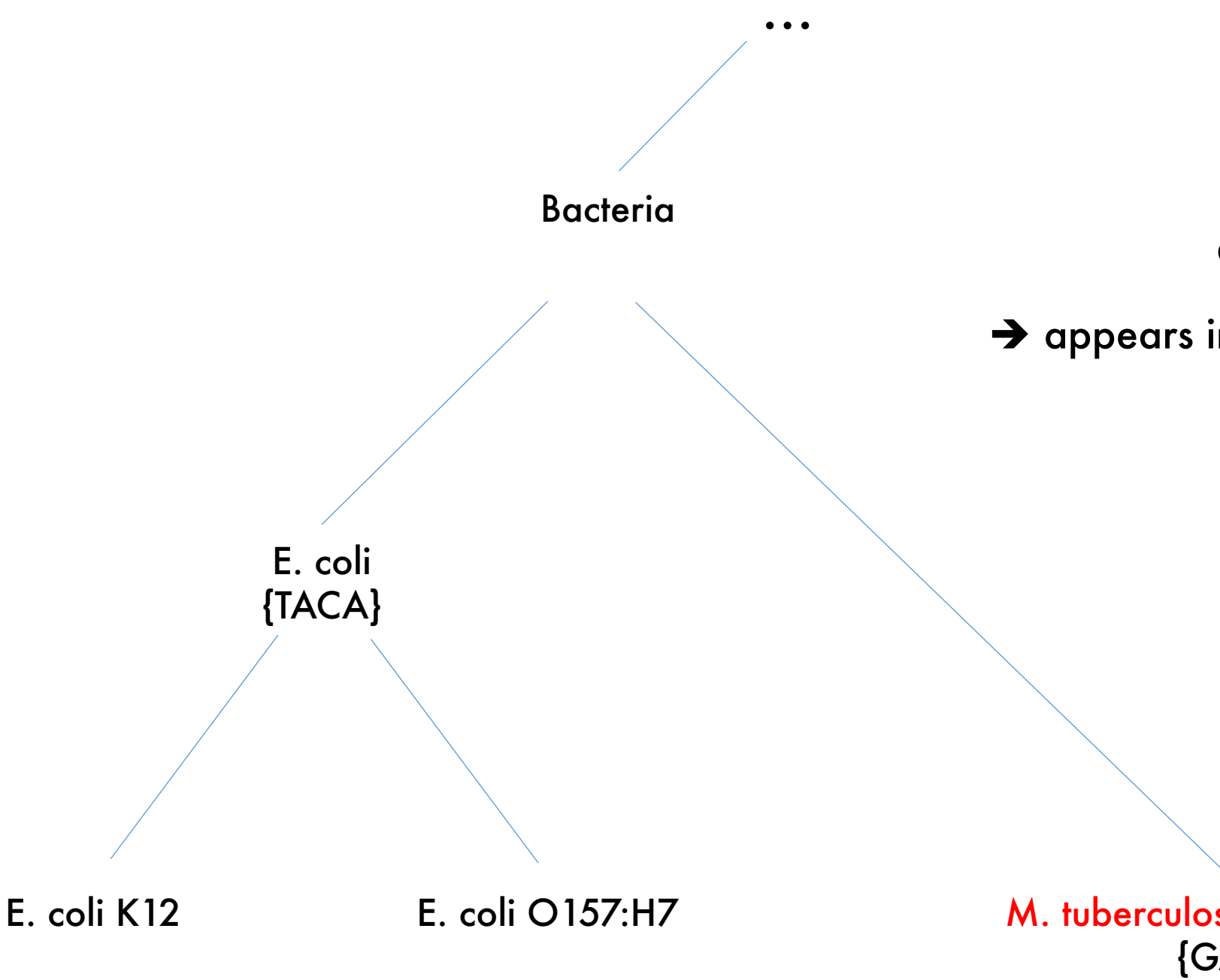
(I.e. determine the lowest node in the phylogenetic tree that fulfils the condition that all genomes that contain the k-mer are below the chosen node).

Species / strain	Reference genome 4-mers
E. coli K12	{GTAC, TACA, ACAC, CACG, ACGA, CGAT, GATG, ATGA}
E. coli O157:H7	{GTAC, TACA, ACAC, CACC, ACCA, CCAT, CATG, ATGA}
M. tuberculosis strain H37Rv	{CGAC, GACT, ACTT, CTTT, TTTG, TTGT, TGTA, GTAC}



e.g. TACA

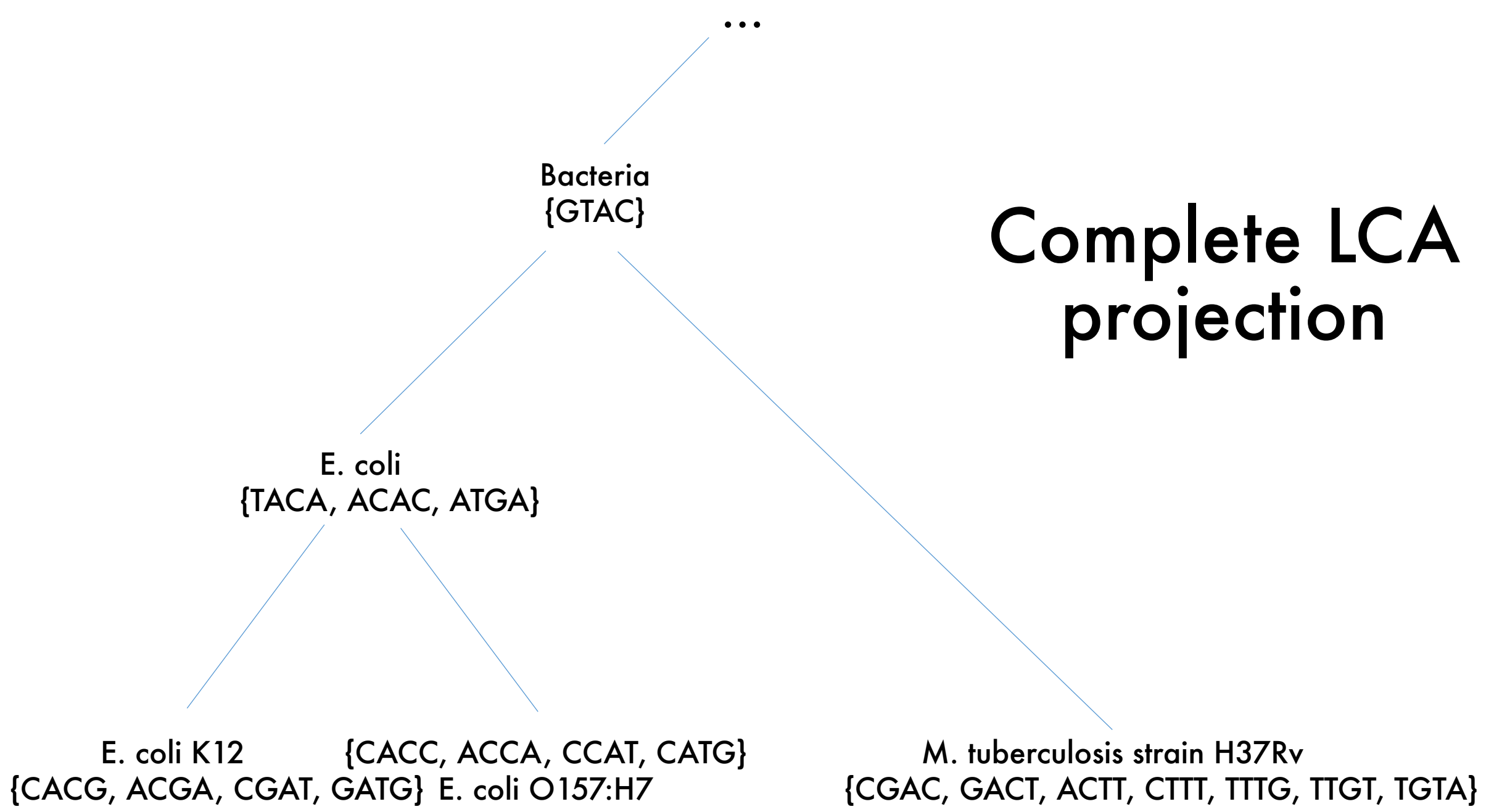
➔ appears in both E. coli strains

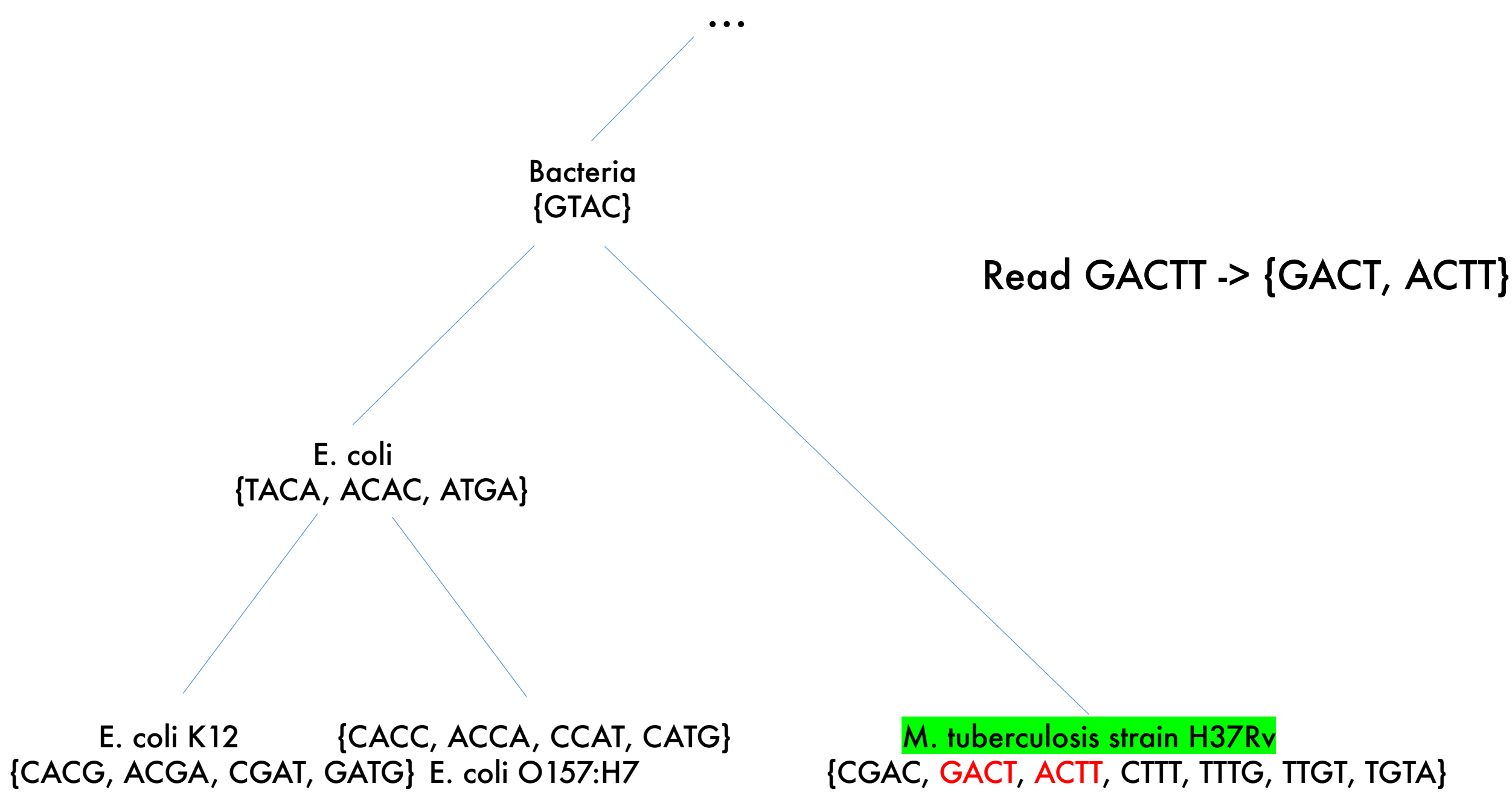


e.g. GACT

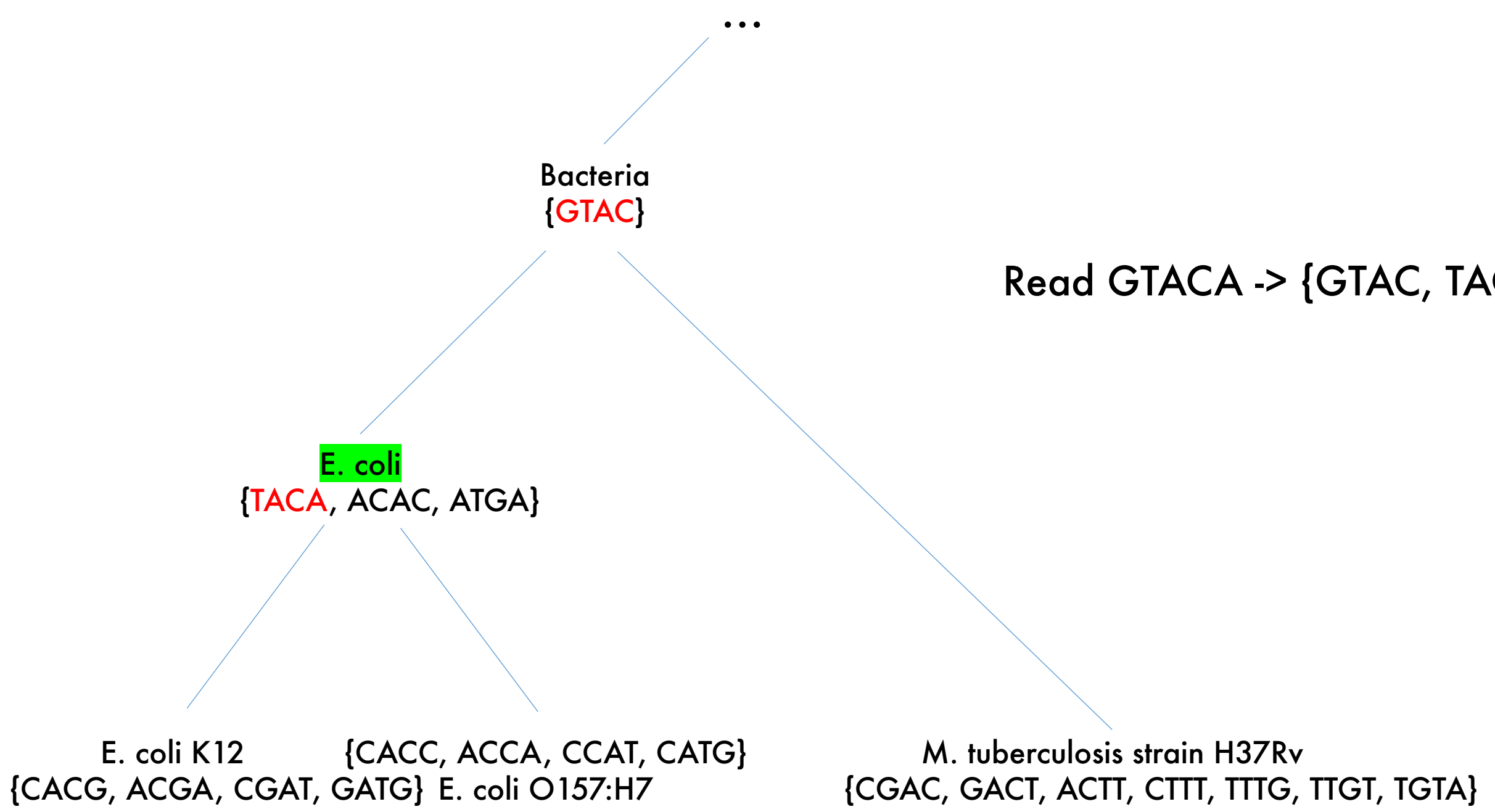
➔ appears in M. tuberculosis strain H37Rv.

Complete LCA projection

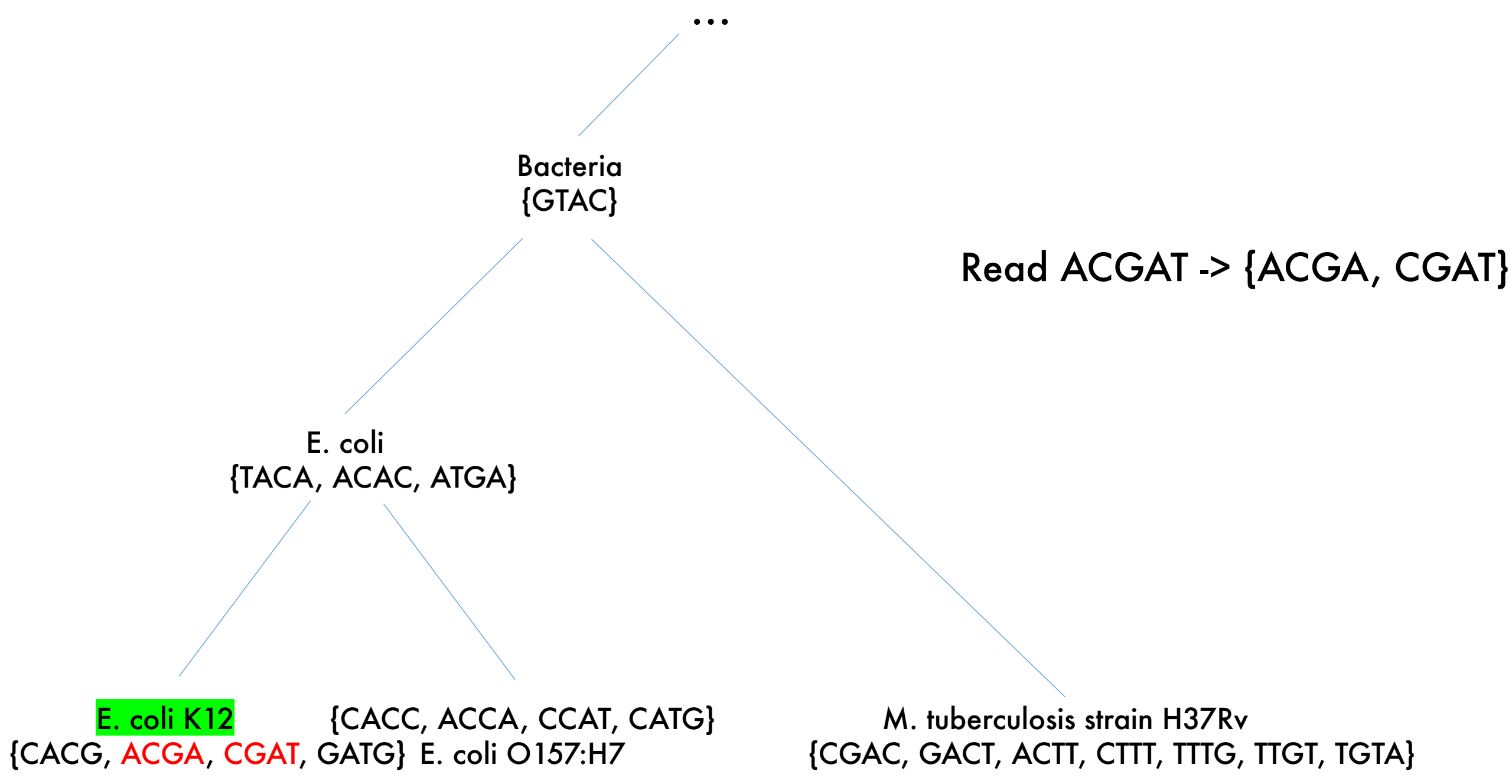




Read GACTT -> {GACT, ACTT}



Read GTACA -> {GTAC, TACA}



Read ACGAT -> {ACGA, CGAT}

Task [15 minutes]

Species / strain	**Reference genome**
Homo sapiens	`CACGACGTACG`
Pan troglodytes (chimp)	`CATGACGTCCG`
E. coli O157:H7	`GTACACCATGA`
M. tuberculosis strain H37Rv	`CGACTTTGTAC`

- What is the LCA of these species?
- Build an LCA classification tree (6-mers)
- Where do the LCA hits of the k-mers in the read `ACGACGTC` localize in the tree?
- Where would we assign the read?
- Does this sequence exist in any of our reference genomes?
- Which conclusions would we draw from that?