

Methods

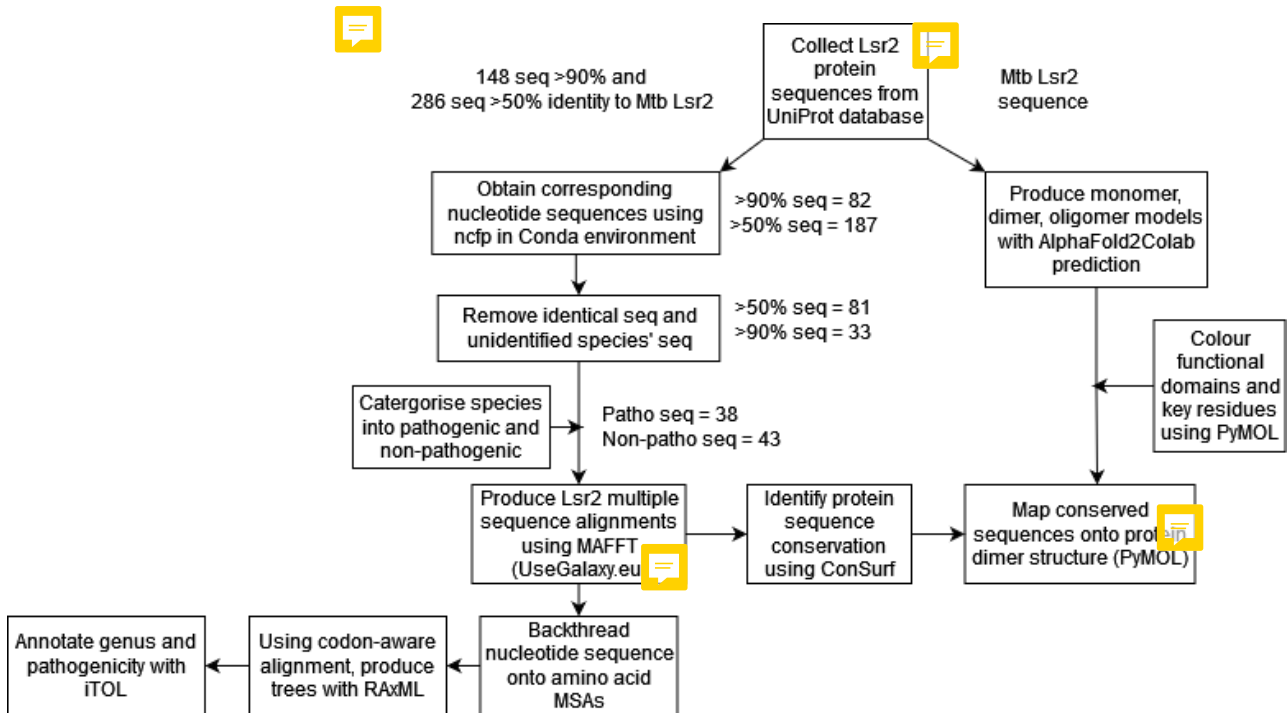


Figure 2: Flowchart showing overview of methods; a combined approach of protein structure modelling with multiple sequence alignments, to map conserved sequence onto protein; and production of phylogenetic trees.

Phylogenetic trees


Sequence data collection


Lsr2 protein sequences were obtained from UniProt in 2 batches and saved in FASTA format (data accessed 26/11/21), 1 batch included sequences with >90% identity to Mtb Lsr2 (148 sequences), and the second batch included sequences with >50% identity to Lsr2 (286 sequences). Sequences shorter than 100 amino acids (protein fragments), were not included.

The corresponding nucleotide sequences for each protein sequence were obtained using ncfp (31). This tool was used instead of back-translating as it finds the corresponding codon for each amino acid in the query sequence using available sequencing reads; this will result in more accurate phylogeny estimations. Not all nucleotide sequences were found for each protein sequence; the corresponding nucleotide sequence were found for 82 of the sequences with >90% identity to Mtb Lsr2, and 187 of the sequences with >50% identity to Mtb Lsr2.


Sequence processing

The header line of the sequences in FASTA format began with NCBI sequence identifiers which were replaced with species name using the python script (seqid_to_species_name.py) (32), and CSV files containing 2 columns (1 with sequence identifies, and 1 with the corresponding species names). This was done so that the leaves of the tree would be labelled with the species name rather the sequence identifier.


Sequences the same species name were removed, so there is only one representative species in the set. Sequences with unknown species names were also removed. This left 33 >90% identity sequences, and 81 >50% identity sequences. 

The sequences in >50% identity set were also used to create 2 additional sets (pathogenic and non-pathogenic species). The species were classified as pathogenic if there was paper providing evidence of human or animal infection. 

Producing phylogenetic trees and annotation


All tools used for producing the tree are available on use.galaxy.eu (33), with the exception of TrimAL, which is available on Phylemon2. Unless specified, default parameters were used. The nucleotide and protein sequence datasets were uploaded onto Galaxy. MAFFT (34) was used to produce multiple sequence alignments, using the multiple protein sequence FASTA as input. The tool Thread Nucleotides onto a Protein Alignment (Back-Translation) (35) was used to thread the nucleotide sequences onto the multiple sequence alignment. 


The multiple sequence alignments were trimmed using the tool TrimAl (36) with the parameter gappyout, to remove poorly aligned sequences; this tool was run using the webserver Phylemon2 (37).


RAxML (38) used the codon-aware alignment to produce a maximum-likelihood tree and bootstrap analysis. The parameters used are as follows. The model type nucleotide and substitution model GTRGAMMA were used, 500 bootstrap replicates were performed, a random seed for rapid bootstrapping was supplied, and the algorithm Rapid bootstrapping and best ML tree search was used. This process was done twice in total, using the >90% identity sequence alignments and the >50% identify sequence alignments, producing two trees. 


The trees were annotated using iTOL (39), overlaying the phenotype of pathogenic or non-pathogenic onto the tree, and annotating species genus and displaying bootstrap values.

Protein structure modelling


Mtb (strain H37Rv) Lsr2 protein FASTA sequence was obtained from UniProt, (accession = P9WIP7 and date accessed 2/11/21). 

AlphaFold2Colab (40) is a Jupyter Notebook for Google Colaboratory which uses AlphaFold to predict protein structures. The Lsr2 protein sequence was input into AlphaFold2Colab (version 1.2) and run with default settings, producing a predicted structure for Lsr2 monomer. The outputs include PDB files of the protein structure, local Distance Difference Test (IDDT) per position of Lsr2 protein sequence, and a Predicted Aligned Error heatmap. The functional domains of the monomer were grouped and coloured using PyMOL (41). 

To produce the protein dimer structure, the sequence was duplicated, with a colon in the middle, and run again with AlphaFold2Colab (date accessed 02/11/21). Each subunit was differently coloured using PyMOL and the DNA binding residues 'RGR' were shown in stick-view and labelled. 

AlphaFold2Colab was also used to produce a Lsr2 oligomer structure of two bonded Lsr2 dimers. The sequences previously used for the Lsr2 dimer structure were duplicated with another colon interspacing them, and the resulting sequence were used for the oligomer prediction, each subunit of the oligomer structure was then coloured in PyMOL. In order to compare the AlphaFold oligomer prediction with the experimentally derived oligomer dimerization domain structure, another model of the Lsr2 oligomer was produced using the Lsr2 protein dimer structures through duplicating and translating the dimer structure in PyMOL. 

Sequence conservation on protein structure

Multiple sequence alignments were used to map sequence conservation onto the protein dimer structure. 4 sets of amino acid sequences were aligned using MAFFT: the >90% identity sequences, the >50% identity sequences, the pathogenic sequences, and the non-pathogenic sequences. 

The multiple sequence alignments were used as inputs and protein sequence conservation was calculated using ConSurf (42) using Mtb Lsr2 sequence as a reference. This produced an Lsr2 monomer structure with residues coloured based on ConSurf Grade, a measure of evolutionary rate categorised into low (conserved) and high (variable) scores. The conserved and variable residues of the monomer were mapped onto the protein dimer structure. 