# Materials and Methods

**Dataset Description**

Three research datasets were identified that consisted of M. bovis sequence data that was also accompanied by reservoir of infection metadata. These datasets came from studies in the United States of America (Salvador et al., 2019), the United Kingdom (Crispell et al., 2019), and New Zealand (Crispell et al., 2017). In total, 782 M. bovis pair read sequencing isolates were selected for further analysis and downloaded from NCBI Sequence Read Archive (SRA) using sratoolkit (\*\*\*).

**Processing, Assembly, and Annotation**

To ensure that all the isolates that were downloaded from SRA were of good quality for downstream steps of the analysis, read trimming was utilized to improve the quality of isolate FASTQ files. Trimmomatic (\*\*\*) was used to cut sequence reads whose average PHRED score was below the 20 threshold on a 4 bp sliding window. Reads that were below 36 bp in length were also expunged from the analysis. The trimmed reads were used to construct a genome assembly using SPAdes with the ‘careful’ option specified (\*\*\*). Statistics associated with the assembled scaffolds was calculated using QUAST, and further statistics were computed when comparing these assemblies with (\*\*\*). Visual inspection of the distributions of assembly statistics such as N50, Genome length, Genome Fraction, and GC content suggested that there were poor quality assemblies within the data as outliers. Therefore, a filtering scheme was devised that allowed for good quality genomes to be retained while clearly poor quality genomes were excluded. Annotation of the quality assessed genomes was then done by using prokka (\*\*\*) with the newly updated M. bovis genome annotation file (Crispell, 2019).