3. Bioinformatics Results

In this chapter I will present the results from the *in silico* screen for plasmids in the culture collection, the phylogeny and host range of the putative plasmids, and the predicted distribution of plasmid antimicrobial genes. Plasmids serve as a way for bacteria to easily facilitate the sharing of advantageous auxiliary functions between species and even between phyla. Plasmids with mobility genes are a good indication of instance where this sharing is being actively pursued. One example of genes that may be actively shared are the AMR genes. The increased use of antimicrobials from active courses to treat illness, to passive use in deodorants and hand sanitisers has increased the levels of exposure of pathogens to antimicrobials in particular members of the human flora. The gut in particular is proposed as a reservoir of AMR, capable of transmitting these genes to pathogens. Observing the predicted prevalence of AMR containing mobile elements, would allow for the investigation of the scope of this hypothesis, and potentially provide evidence to support it.

3.1 Frequency of Plasmid Detection Across Human Gut Microbiota Phyla

Bacterial whole genome sequences from the Lawley Lab Culture Collection were screened *in silico* using plasmidSPAdes (Antipov D. *et al*., 2016) to identify genomes containing predicted plasmid sequences. PlasmidSPAdes isolates and assembles plasmid contigs from whole genome sequence data based on contig coverage as plasmid coverage often differs from the median chromosomal coverage. This data is presented in Figure 1 as the number of genomes predicted to contain zero, one, two, or more plasmids in each phylum in the culture collection. 653 genomes were scanned; 60% are not predicted to contain any plasmids, and of the 240 genomes predicted to contain plasmid sequences, 70% are predicted to only contain one plasmid. Plasmids are predicted across all phyla in the culture collection phylogeny, with phylum Firmicutes containing the largest number of plasmids (151) and Bacteroidetes the most enriched phylum with a chi squared value of $\chi 2=0.04$.

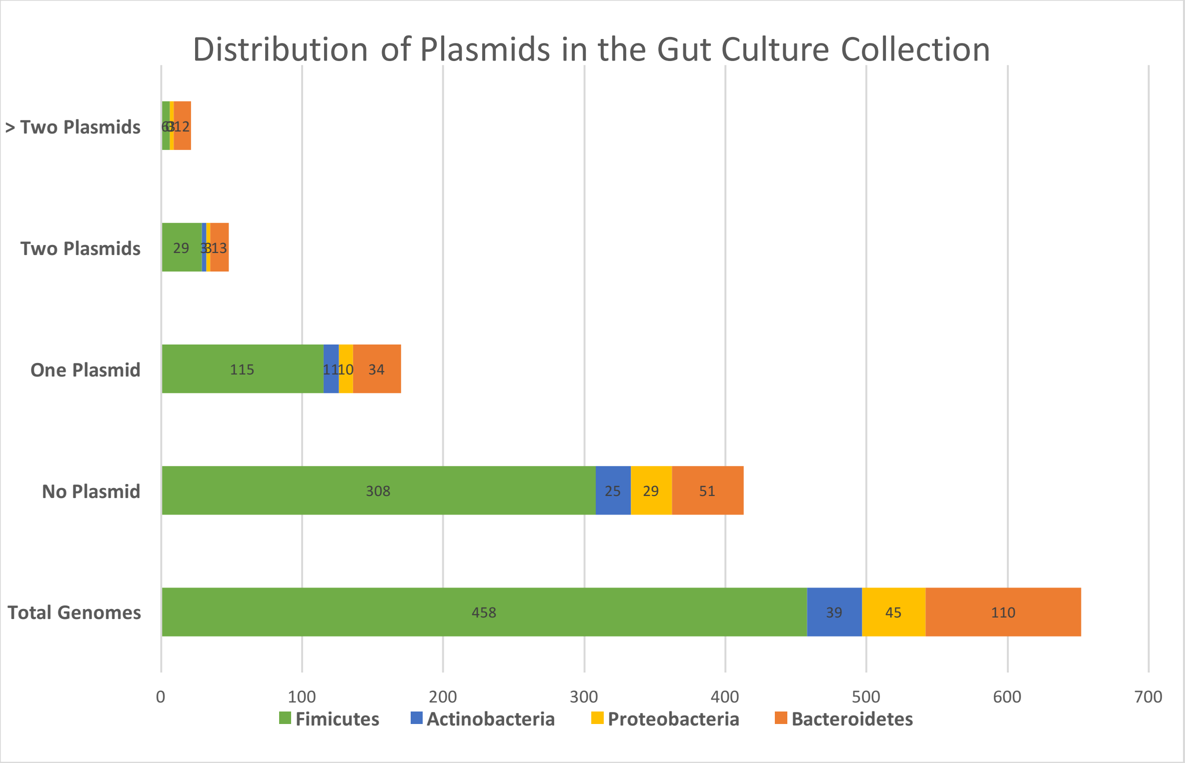

Figure 3.1. *The distribution of predicted plasmids in the 653 genomes of the Lawley Laboratory culture collection*. 40% of the genomes are predicted to harbour plasmid sequences; 70% of those genomes are only predicted to contain one plasmid.

The predictions from plasmidSPAdes were used to annotate the culture collection phylogeny, as displayed in Figure 2. Putative plasmids are widely distributed across the phyla in the culture collection, suggesting that there are important functions present on these elements facilitating survival in the gut environment.

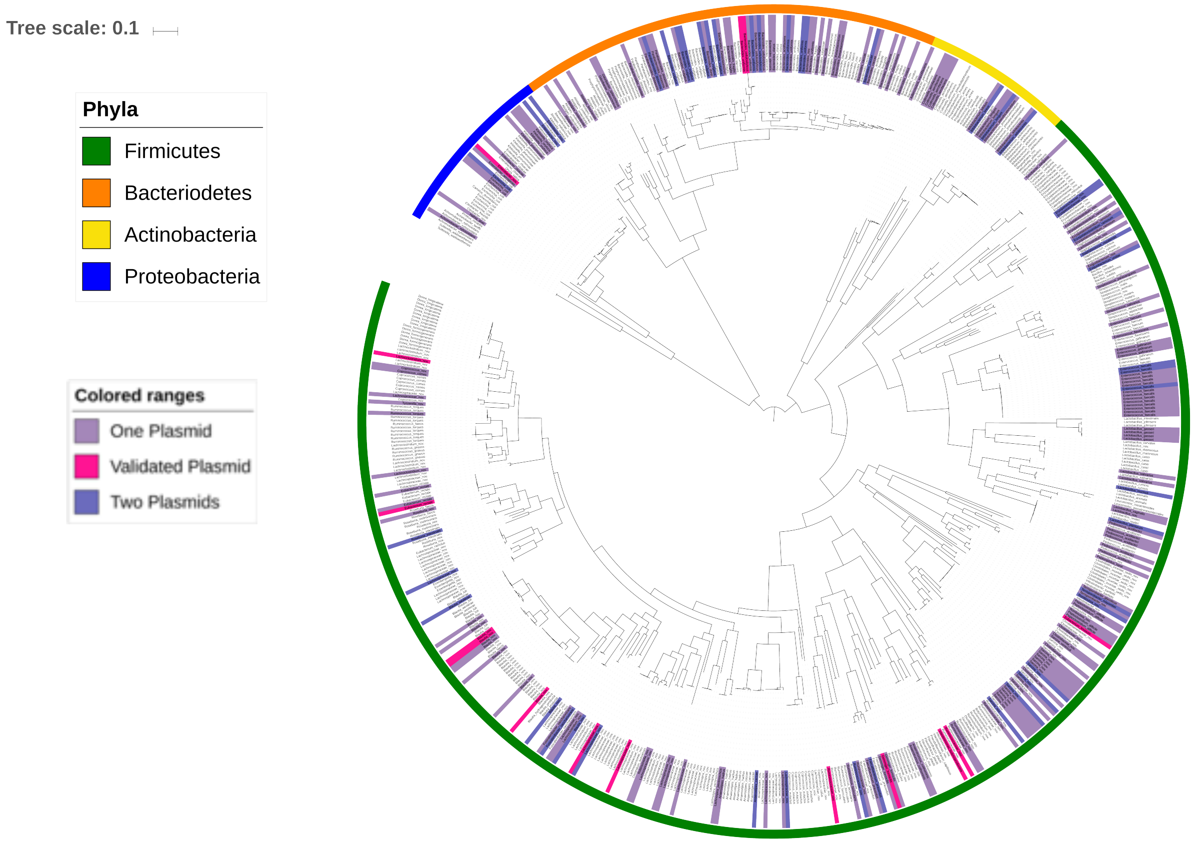

Figure 3.2. *Lawley Lab culture collection phylogeny with plasmid prediction annotations.* 653 bacterial whole genome sequences were scanned for the presence of plasmids. 240 genomes are predicted to contain plasmids: genomes predicted to contain one plasmid are annotated in pale purple, and genomes predicted to contain two plasmids are annotated in deep purple. Genomes where DNA has been experimentally isolated by mini prep and visualised on agarose are annotated in pink.

3.2 Size and Coverage and Distribution of Predicted Plasmids

In an attempt to understand the type of extrachromosomal elements in the culture collection, the plasmid size and coverage information from the plasmidSPAdes assembly graphs were used to plot the distribution of plasmids according to these traits; this data is presented in Figure 3. Plasmids size distribution is predicted to follow a bimodal curve, and plasmid coverage is often higher than that of the chromosome. The predicted plasmids of the culture collection are predominantly small, with 70% falling in the 1-10kb range. The majority of the predicted plasmids are high coverage elements with coverage between two and tenfold higher than the median. This could be an artefact of the prediction methods, or may give insight into the pressures in the gut environment shaping the distribution of plasmids by affecting factors such as carriage costs and the frequency of HGT.

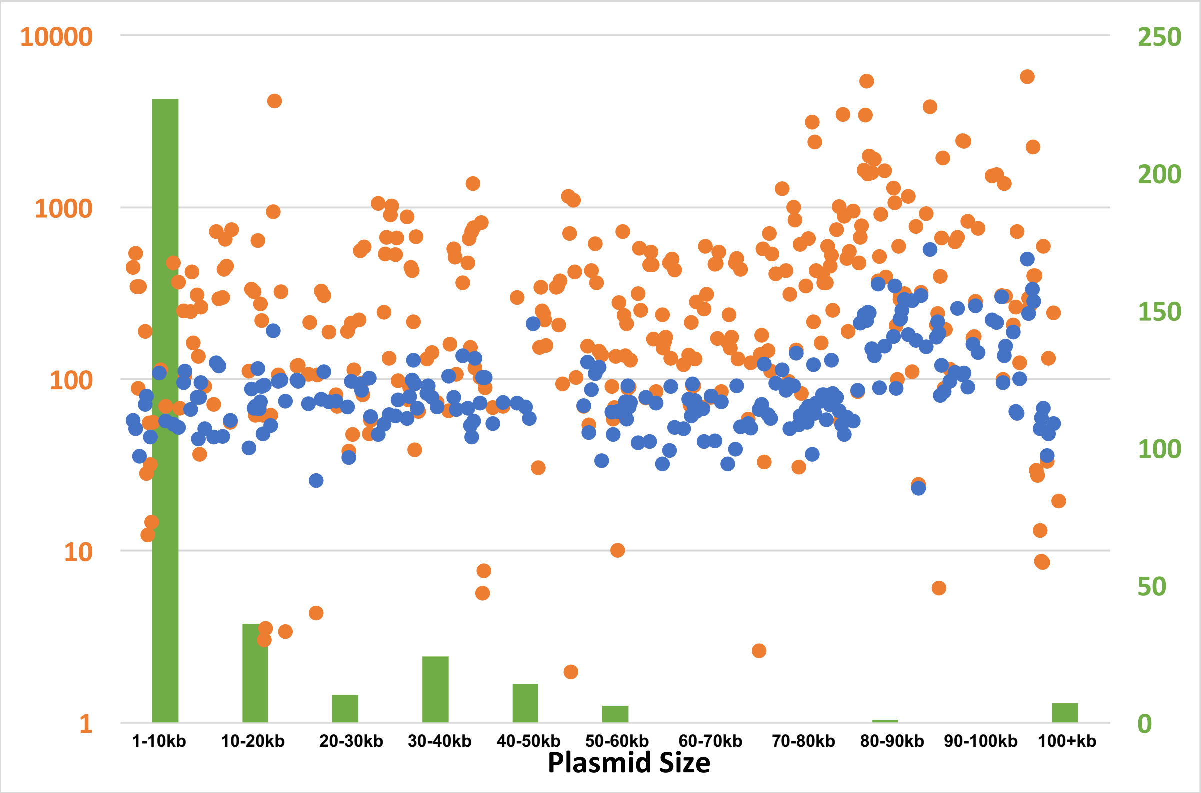

Figure 3.3. *Detected plasmid coverage compared to chromosomal median coverage and plasmid distribution by size*. Triaxial graph displaying predicted plasmid distribution by size and coverage. The green dataset plotted on the right-hand axis displays the plasmid counts for each size range. The blue and orange datasets plotted on the left-hand side show the median chromosomal coverage and plasmid coverage on a logarithmic scale to display the fold difference between them. The majority of plasmids are small, 1-10kb, and high coverage, 2-10 fold above median.

3.3 Plasmid Classification and Phylogeny

As discussed in Chapter 1 plasmids are classified either by function or by the inc grouping system, this can be done *in silico* through replicon typing. Replicon typing uses a computational alignment to associate plasmids to one of the known incompatibility groups based on the sequence of the plasmid's *rep* gene. Attempts to assign a replicon type to the predicted plasmid sequences using PlasmidFinder (Carattoli A. *et al.,* 2014) were unsuccessful, yielding only 19 hits against the 240 genomes predicted to contain plasmids. An alternative approach involved aligning the plasmid sequences with BLASTx against all the rep sequences on the pfam database for domain PF01051. This domain is common to both RepB and RepA and was the Rep protein domain that was the most frequently identified during plasmid annotation. All the sequences did yield hits to *rep* genes; however, the top hits were to uncharacterised plasmids. Therefore, the plasmids were not able to be provisionally typed by association.

Subsequently, annotated *rep* genes were used to infer a phylogeny of the culture collection genomes. Two trees were built, one using *repE* (Figure 4) as the most abundant rep gene in the culture collection, and the other using *repA* (Figure 5), which is more commonly considered a conserved or standard rep gene. This would allow me to observe clusters of related plasmids in the absence of replicon typing. The *repE* gene was most frequently annotated in the culture collection. An alignment of the culture collection genomes to the *repE* sequence from one of the putative plasmids was used to build the phylogeny in Figure 4. This phylogeny shows clusters of related *rep* genes across the different phyla.

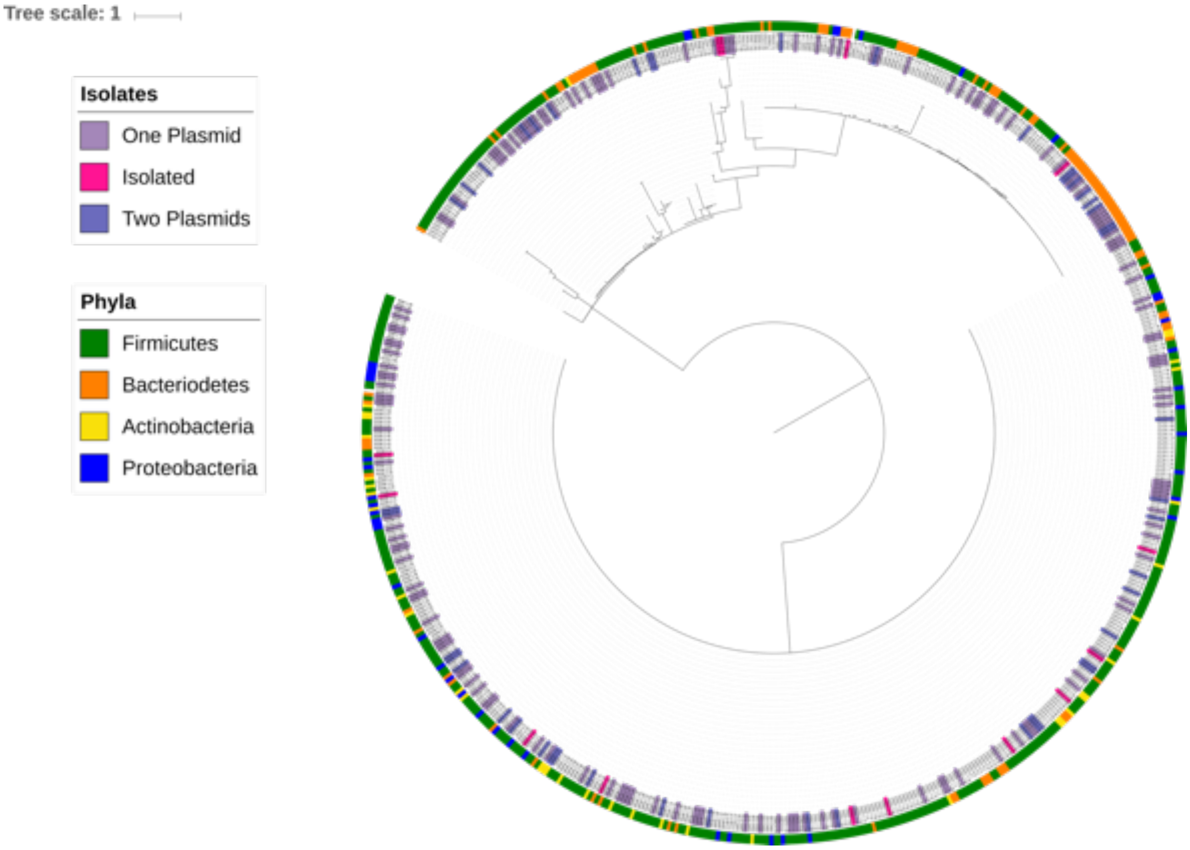

Figure 3.4. *Phylogenetic tree constructed using the repE gene*. A phylogeny of the 653 genomes in the culture collection was built from an alignment of the *repE* gene. Plasmids from all 4 phyla are clustered across the top third of the tree, displaying relatedness between the *rep* genes in those genomes. The shallow branch along the bottom third of the tree, may be due to an absence of the *repE* gene in these samples, of the presence of a very pervasive gene. Genomes predicted to contain one plasmid are annotated in pale purple, and genomes predicted to contain two plasmids are annotated in deep purple. Genomes where DNA has been experimentally isolated by mini prep and visualised on agarose are annotated in pink.

The *repA* gene is significantly less annotated and possibly less represented in the culture collection, however it is more commonly recognised as a conserved or core plasmid gene. The *repA* gene sequences were aligned across the culture collection and the resulting phylogeny presented in Figure 5; once again demonstrating clustering of the *rep* sequences across phyla.

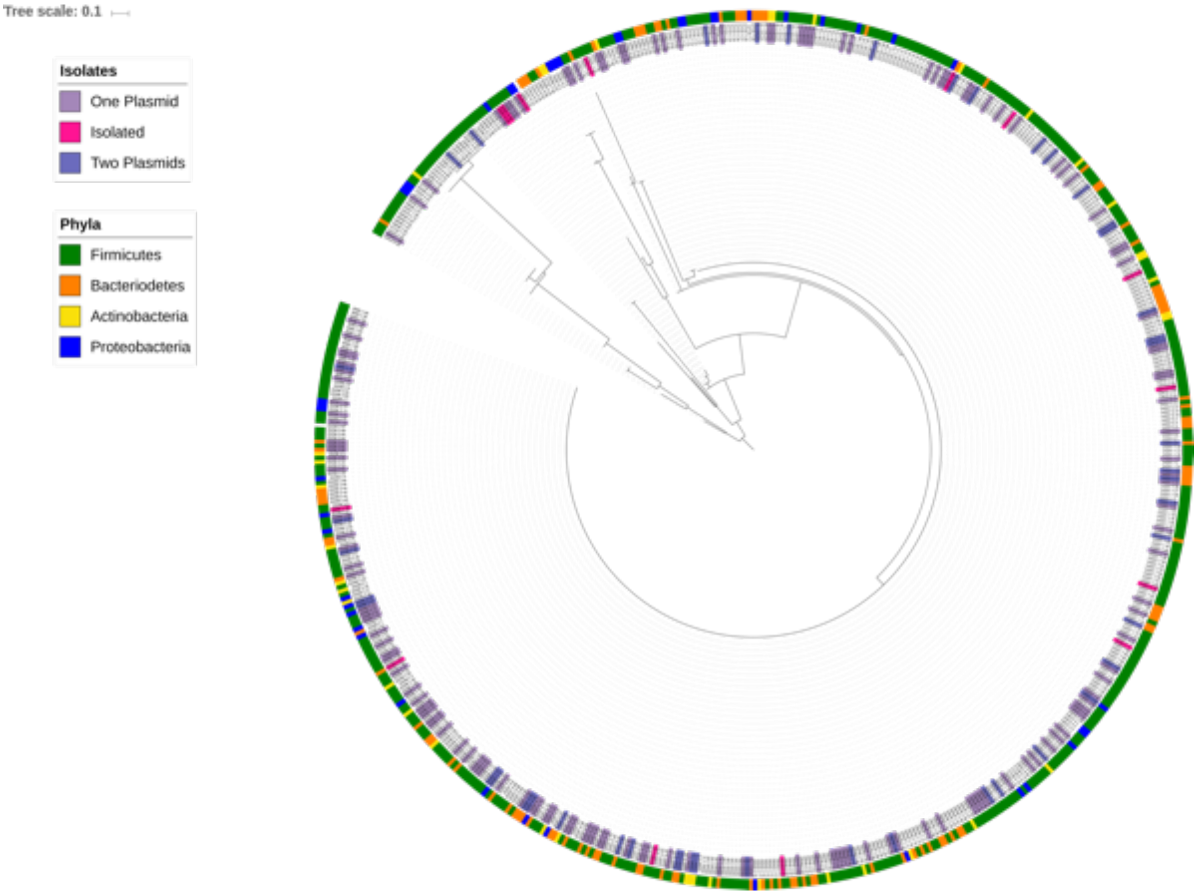

Figure 3.5. *Phylogenetic tree constructed using the repA gene.* A phylogeny of the 653 genomes in the culture collection was built from an alignment of the *repA* gene. Plasmids from all 4 phyla are clustered in a small segment of the tree, displaying relatedness between the *rep* genes in those genomes. The shallow branch is broader in this tree; and as the *repA* gene is not as abundantly annotated in the culture collection this may be due to an absence of the *repA* gene in these samples, or once again the presence of a very pervasive gene. Genomes predicted to contain one plasmid are annotated in pale purple, and genomes predicted to contain two plasmids are annotated in deep purple. Genomes where DNA has been experimentally isolated by mini prep and visualised on agarose are annotated in pink.

The clustering of the nodes along the top third of the tree in Figure 4 does not match the phyla-based clustering of the species tree; this incongruence suggests the sharing of plasmids across broad phylogenetic distances. These interactions were annotated on the culture collection tree to visualise the connections between genomes, producing the chord diagram in Figure 6. The chords (pink) show links between Firmicutes and Bacteroidetes, Firmicutes and Proteobacteria, and Bacteroidetes and Actinobacteria, indicating the possible presence of broad host range plasmids.

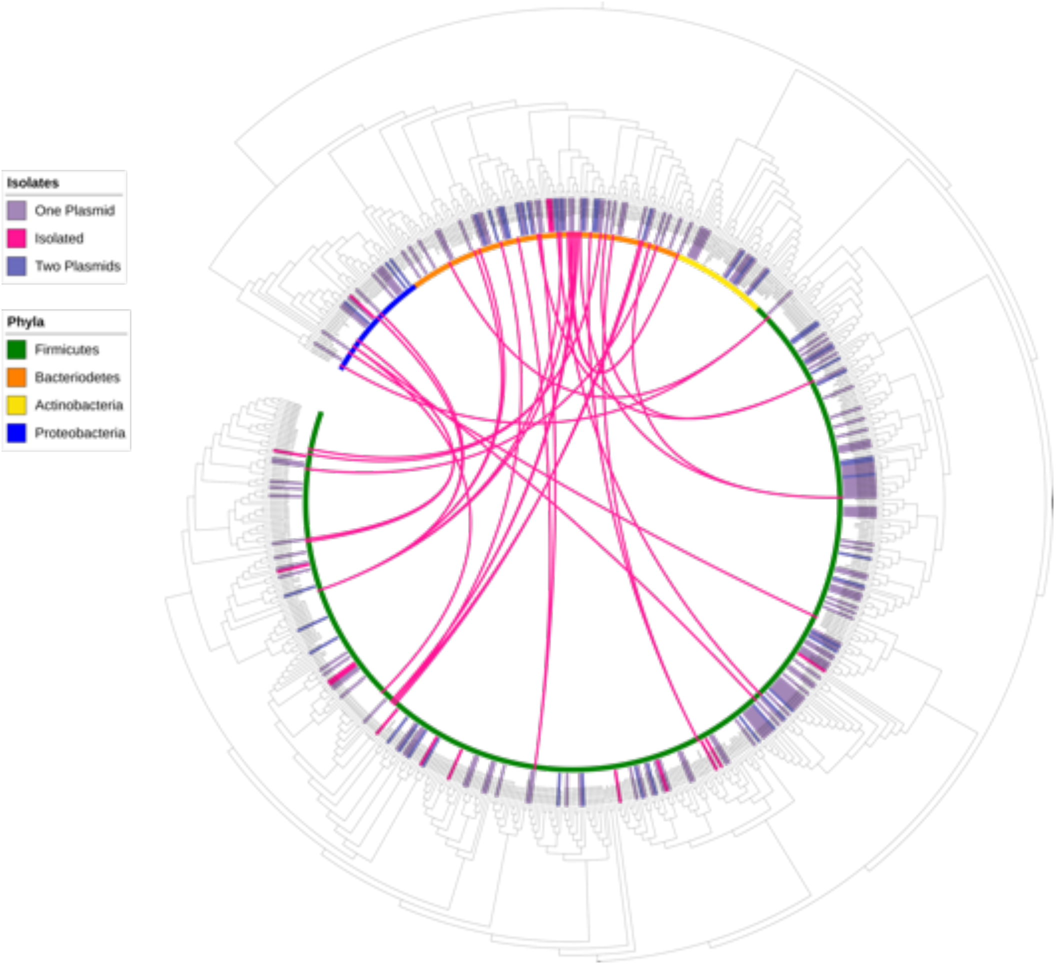

Figure 3.6. *Chord diagram displaying links between genomes containing closely related repE genes.* The chords (pink) display links between isolates with closely related *repE* genes, indicating potential instances of plasmid sharing between isolates of differing phyla and the presence of broad host range plasmids. Genomes predicted to contain one plasmid are annotated in pale purple, and genomes predicted to contain two plasmids are annotated in deep purple. Genomes where DNA has been experimentally isolated by mini prep and visualised on agarose are annotated in pink.

As with the *repE* phylogeny the clustering of the nodes in the *repA* tree does not match the phyla-based clustering of the species tree. These interactions were used to create a second chord diagram on the culture collection tree (Figure 7). In this instance, the chords show interactions between Firmicutes and Actinobacteria, Firmicutes and Proteobacteria, and Proteobacteria and Actinobacteria. A *Klebsiella* node is shown to have several connections to several *Blautia* species, including a species containing a predicted small resistance element, and one containing a mid-size conjugation plasmid. Both of these genetic elements were experimentally isolated, and these strains may be good candidates for demonstrating horizontal gene transfer *in vitro.*

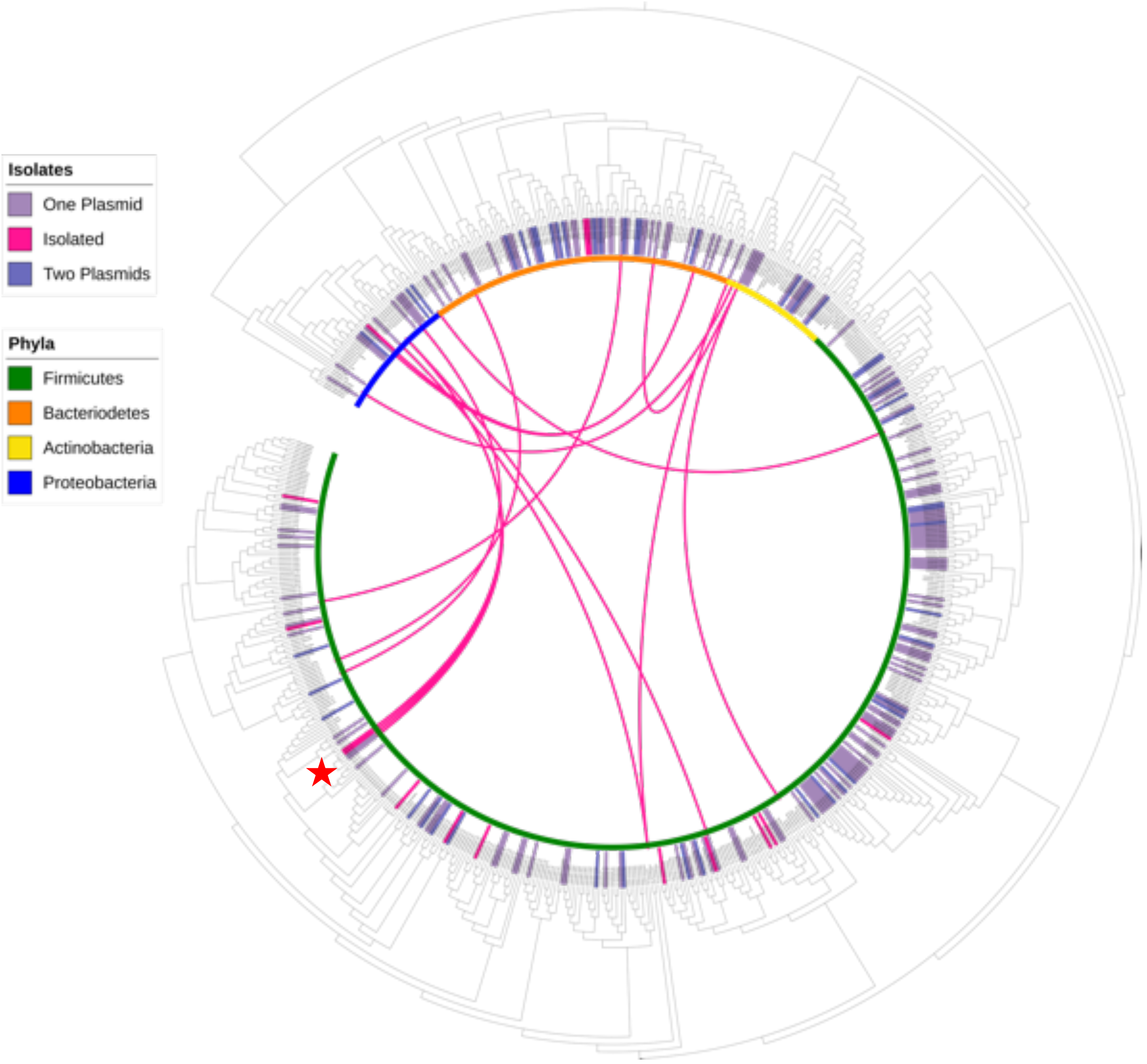

Figure 3.7. *Chord diagram displaying links between genomes containing closely related repA genes*. The chords (pink) indicate potential instances of plasmid sharing between isolates of differing phyla and the presence of broad host range plasmids. The *Blautia* species marked by a star are connected to a strain of *Klebsiella pneumonia*. The plasmids contained in the *Blautia* include a small resistance element, and a mid-range element with conjugation sequences, both of which have been visualised experimentally. Genomes predicted to contain one plasmid are annotated in pale purple, and genomes predicted to contain two plasmids are annotated in deep purple. Genomes where DNA has been experimentally isolated by mini prep and visualised on agarose are annotated in pink.

3.4 AMR Gene Distribution

The other method for classifying plasmids is based on function, and a function that readily leads itself to validation is AMR. In addition, once validated, the AMR phenotypes can be used as selectable markers when testing plasmid mobility. Examining the abundance of AMR on gut bacterial plasmids also begins to address the hypothesis of the gut microbiome as an AMR reservoir and location for HGT. ARIBA (Antibiotic Resistance Identification by Assembly, Hunt M. *et al*, 2017, bioRxiv) was used to predict the prevalence of AMR genes in the culture collection. AMR genes were predicted in the culture collection genomes from raw sequence reads by implementing ARIBA with the CARD and SRST2_ARGannot databases. AMR genes were also identified within the predicted plasmid assemblies using BLAST with the CARD, SRST2_ARGannot (SA), and ResFinder (RF) databases. (Zankari E. *et al.*, 2012; McArthur A.G. *et al,* 2013; Inouye M. *et al.,* 2014)

The incidence of AMR gene prediction was then compared between plasmid containing genomes and non-plasmid containing genomes, displayed as percentages in Figure 8. The result demonstrated almost equal populations of resistant and non-resistant bacteria: 53% resistant to 47% non-resistant with CARD, and 51% resistant to 49% non-resistant with SA. Scanning the genomes predicted to contain plasmids with the same databases demonstrated an increase in the proportion of reported AMR genes in the genomes predicated to contain plasmids 60% resistant to 40% non-resistant with CARD, and 55% resistant to 45% non-resistant with SA. The increase of AMR predictions in plasmid containing genomes compared to all genomes is significant when using the CARD predictions but not the SA predictions, with chi squared values of χ^2=0.035 and χ^2=0.36 respectively.

The incidence of AMR gene prediction in the culture collection was then compared to the prediction of AMR genes on the putative plasmid sequences. The plasmid assemblies were scanned using BLAST (blastn), against three databases at a similarity level of 70%. AMR genes were more frequently predicted to be on putative plasmid sequences than in the culture collection with 68% resistant to 32% non-resistant. The increase in predicted AMR genes on the putative plasmids sequences is significant when compared with the prediction on the whole genome sequences of the culture collection with both databases- chi squared values of χ^2=0.002(CARD) and χ^2=0.009(SA).

The correlation between frequency of mobility genes and AMR genes was also investigated. Plasmid annotations were searched for *mob* and *tra* genes and 90 of the 240 genomes predicted to contain plasmids also contain plasmid mobility elements. The proportion of AMR reported in this set of plasmids was compared to the amount of AMR reported in the culture collection. There is an increase in predicted AMR on the putative plasmid sequences with mobile elements with 75% resistant and 25% non-resistant. This result is significant when compared to the predictions made using the SA database χ^2=0.023, and falls just short of significance with the CARD database χ^2=0.062. These AMR predictions were validated on the species with experimentally isolated plasmids, and these results are discussed in the next chapter.

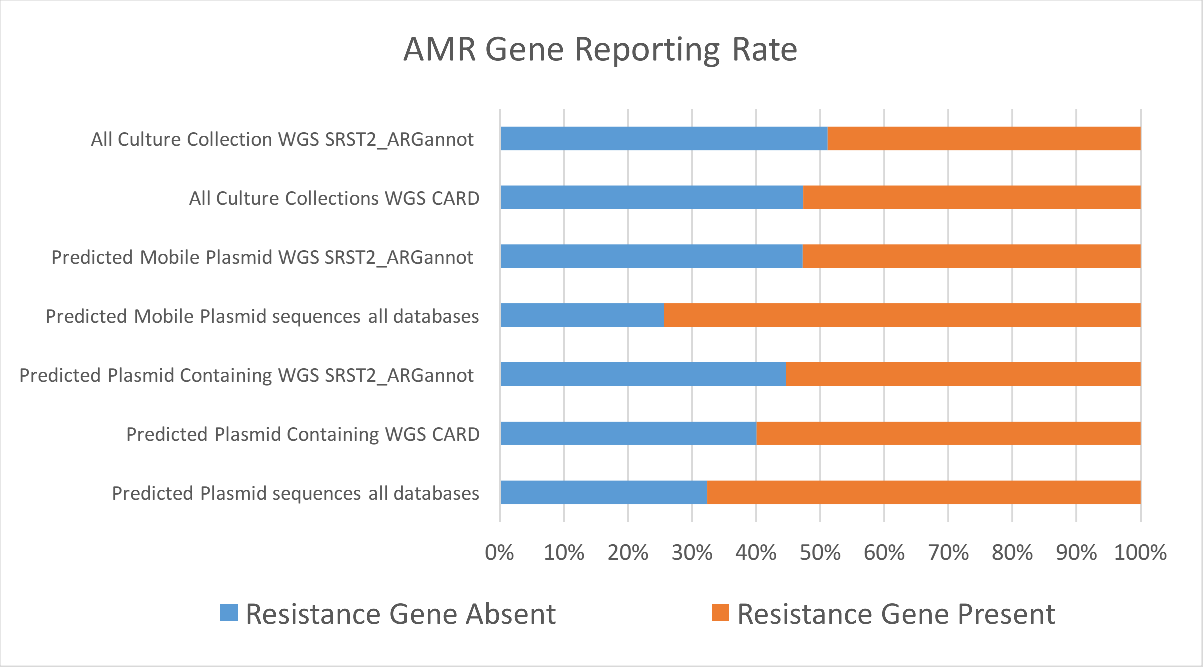

Figure 3.8. Percentage bar graph displaying the proportions of predicted resistance in the whole culture collection, predicted plasmid containing species, and on predicted plasmid sequences. Predictions were completed using the CARD and SRST2_ARGannot databases for the whole genome predictions, and using the CARD, SRST2_ARGannot, and ResFinder databases on the predicted plasmid sequences. There is a significant increase in the reporting rate of AMR genes in plasmid containing species compared to the rate in the culture collection when using the CARD database but not with the SRST2_ARGannot database. There is a significant increase in the reporting rate on the predicted plasmid sequences when compared to the reporting rate in the whole culture collection. When comparing the reporting rate from genomes with potential mobile elements, there is only a significant increase in the reporting rate from the predicted sequences themselves and not the whole genome sequence.