Supplementary Information

Impact of Related Genomes on MAG binning

By generating a phylogeny of universal single copy genes in our input genomes (Fig. S1) we analysed the relationship between the presence of closely related genomes and the ability of the different MAG-recovery methods to bin chromosomal sequences. Specifically, we regressed phylogenetic distance on this phylogeny with per-bin chromosomal coverage (Fig. S2) and bin purity (Fig. S3). This identified no clear relationship between chromosomal coverage (Fig. S2), or purity (Fig. S3), and the phylogenetic distance to the nearest relative in the metagenome.

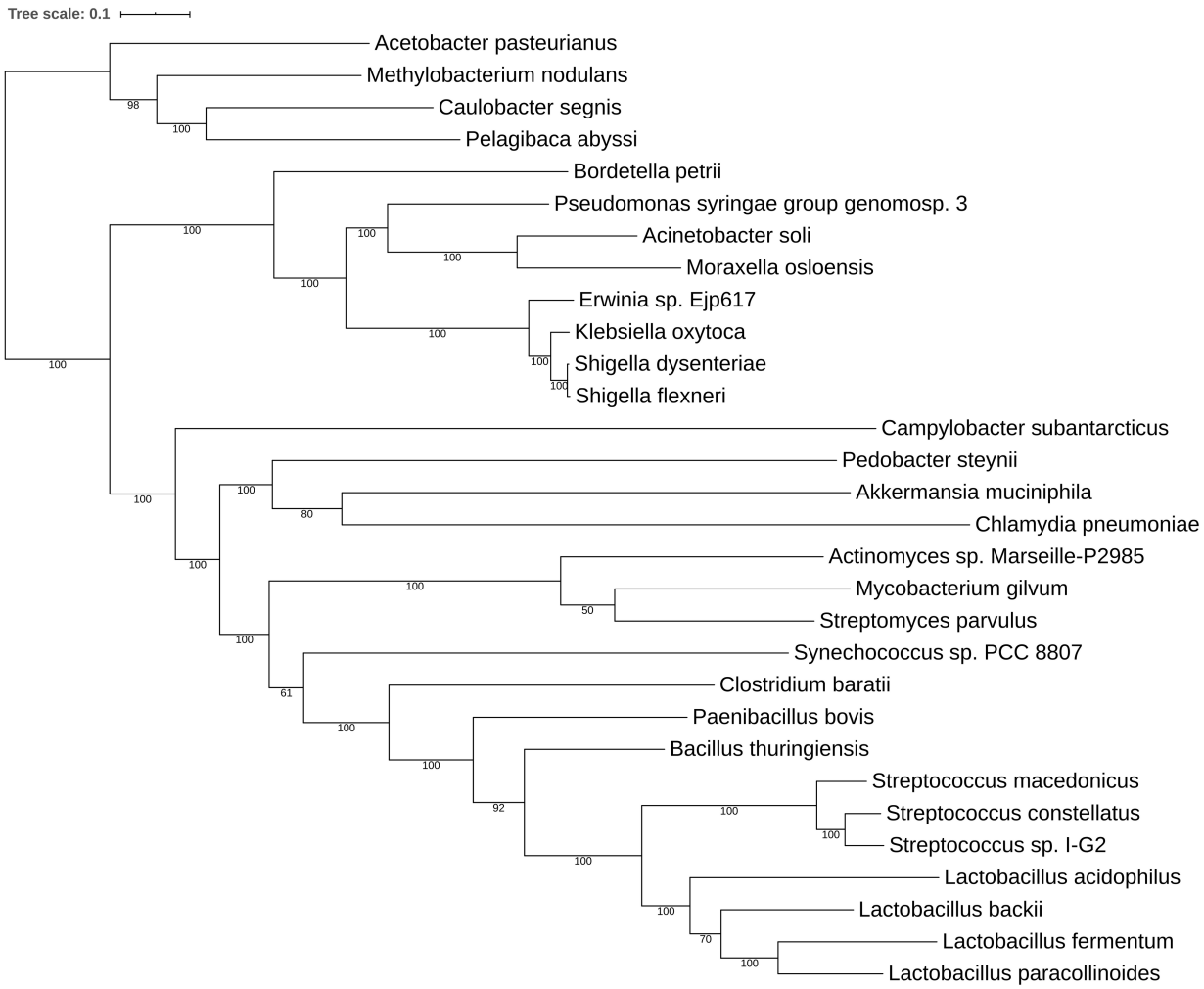


Figure S1: Unrooted universal single-copy gene concatenation maximum-likelihood (IQ-TREE) phylogeny. Percentage of ultrafast-bootstraps (n=1000) supporting each bifurcation are annotated on each node. Phylogeny was visualised using iToL.

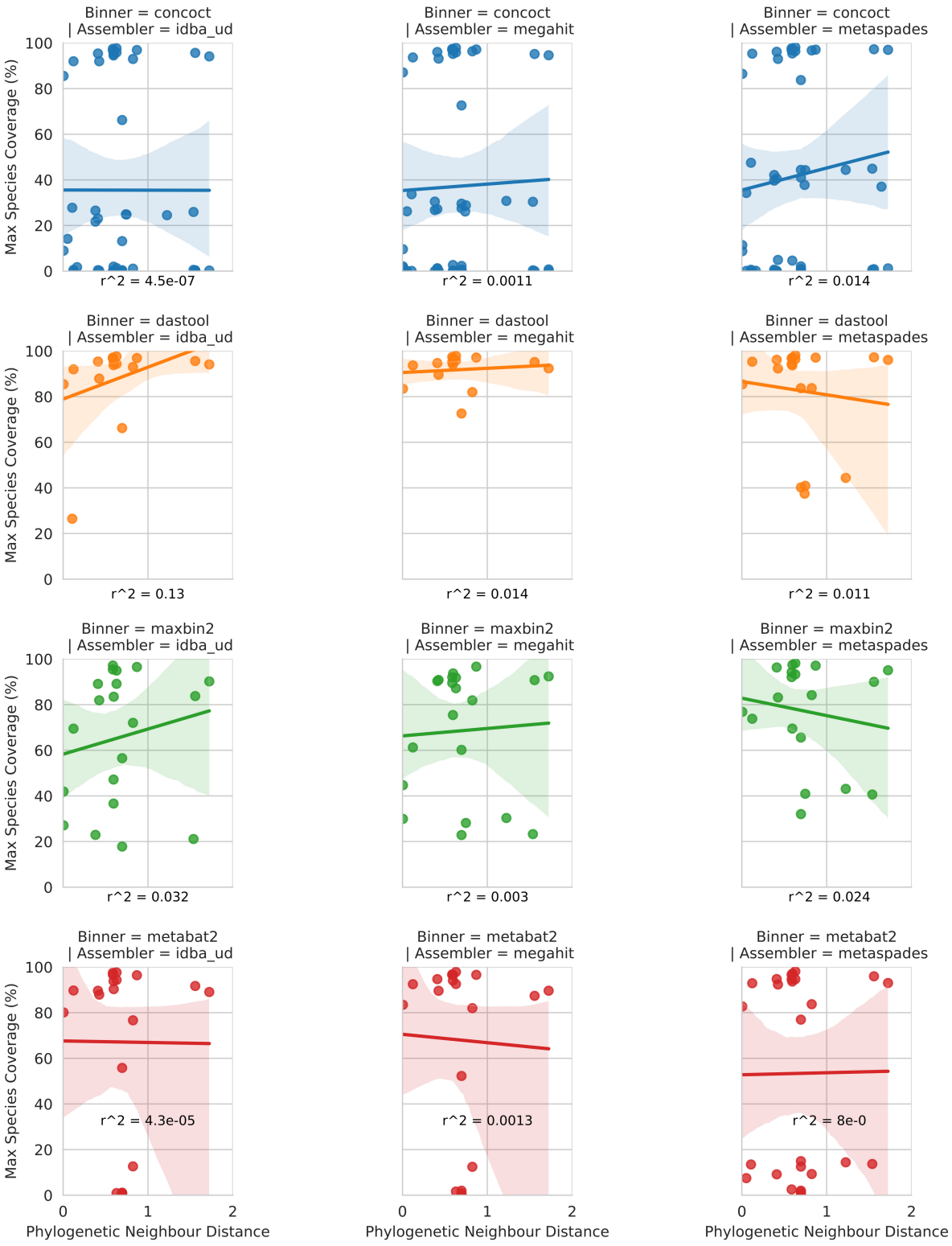


Figure S2: Relationship between phylogenetic distance to closest neighbour input genome on genomic coverage in MAG majority comprised of that taxon. Each dot represents the genomic coverage of a particular genome and the branch distance on an 86-protein concatenated phylogeny between that genome and its nearest neighbour. Rows indicate the binning software and columns the metagenomic assembler. Regression line is a simple linear model fitted in seaborn with R^2 values calculated and annotated on each plot.

Figure S3: Relationship between phylogenetic distance to closest neighbour input genome on bin purity. Each dot shows the number of other input genomes detectable in a given MAG bin in relation to the branch distance on an 86-protein concatenated phylogeny between the majority genome in that bin

and its nearest neighbour. McFadden's pseudo-R^2 calculated from fitted Poisson logistic regression models are annotated on each plot.

Recovery of Specific Gene Content

We explored the ability of different approaches to find open reading frames (ORFs) within MAGs. Overall, the total number of predicted ORFs in MAGs followed a similar trend (Fig. S4) as the chromosomal coverage and purity (Fig. 2). Of the four binning tools, CONCOCT performed the worst, finding <30% of the number of ORFs in our reference genomes used to construct the synthetic data. MetaBAT2 performed second worst at ~80%. DASTool recovered a similar number to our reference and Maxbin2 detected 7-46% more genes. The Assembler method did not significantly impact the number of genes predicted with the exception of Maxbin2, in which IDBA_UD was the closest to reference and metaSPAdes predicted 46% more ORFs. Given that there is reason to suspect that there are some issues with the ORF calling in the MAGs. i.e., some tools produced more predicted ORFs than reference, it could be the case that some of these sequences are present in the assemblies (with errors/gaps), but are not being identified as ORFs, or are broken into multiple ORFs, leading to issues downstream labeling them correctly as AMR/VF genes. Regardless of different tools producing a different number of ORFs, the recovery of AMR/VF is pretty consistent regardless of how many ORFs are predicted.

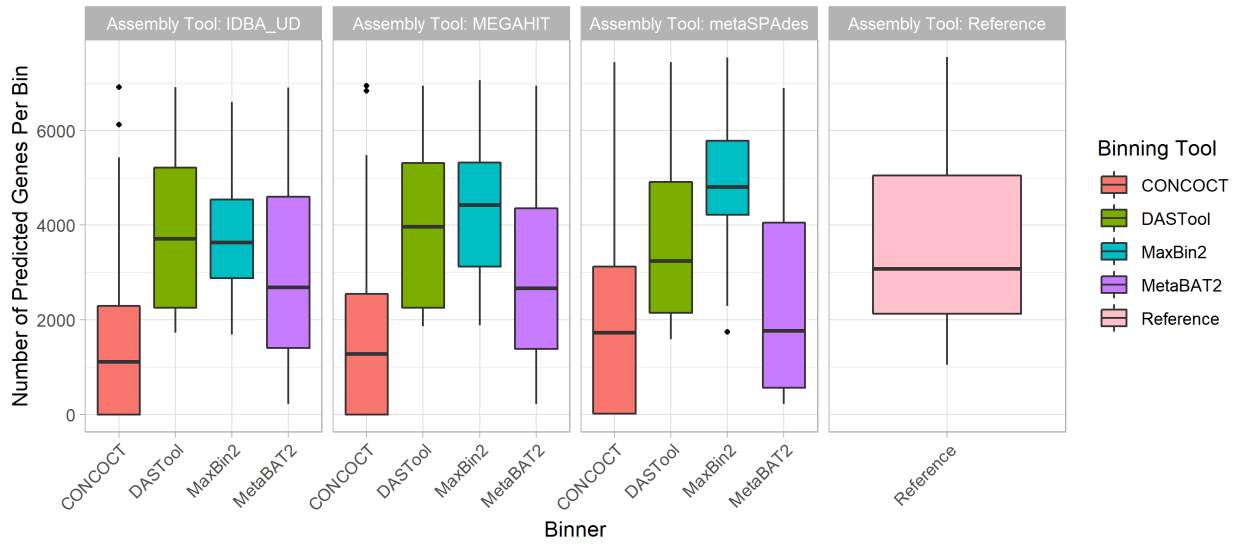


Figure S4: Predicted Gene Content. The total number of open reading frames (ORF) predicted followed the same trend as chromosomal coverage and purity. The assemblers (colored bars) did not contribute to variability in the number of ORFs detected. Of the 4 binners, CONCOCT recovered <30% of our reference genome ORFs. DASTool and MetaBAT2 predicted a similar number as our reference genomes.

Comparisons of Rates of Loss

Combining the performance metrics for Figs. 3, 4, 5, and 6 to compare the rates of loss of different components emphasises some of the observed patterns (see Fig. S5). This highlights that genomic components (GIs and plasmids) and plasmids in particular are lost at a higher rate than individual gene types during MAG recovery.

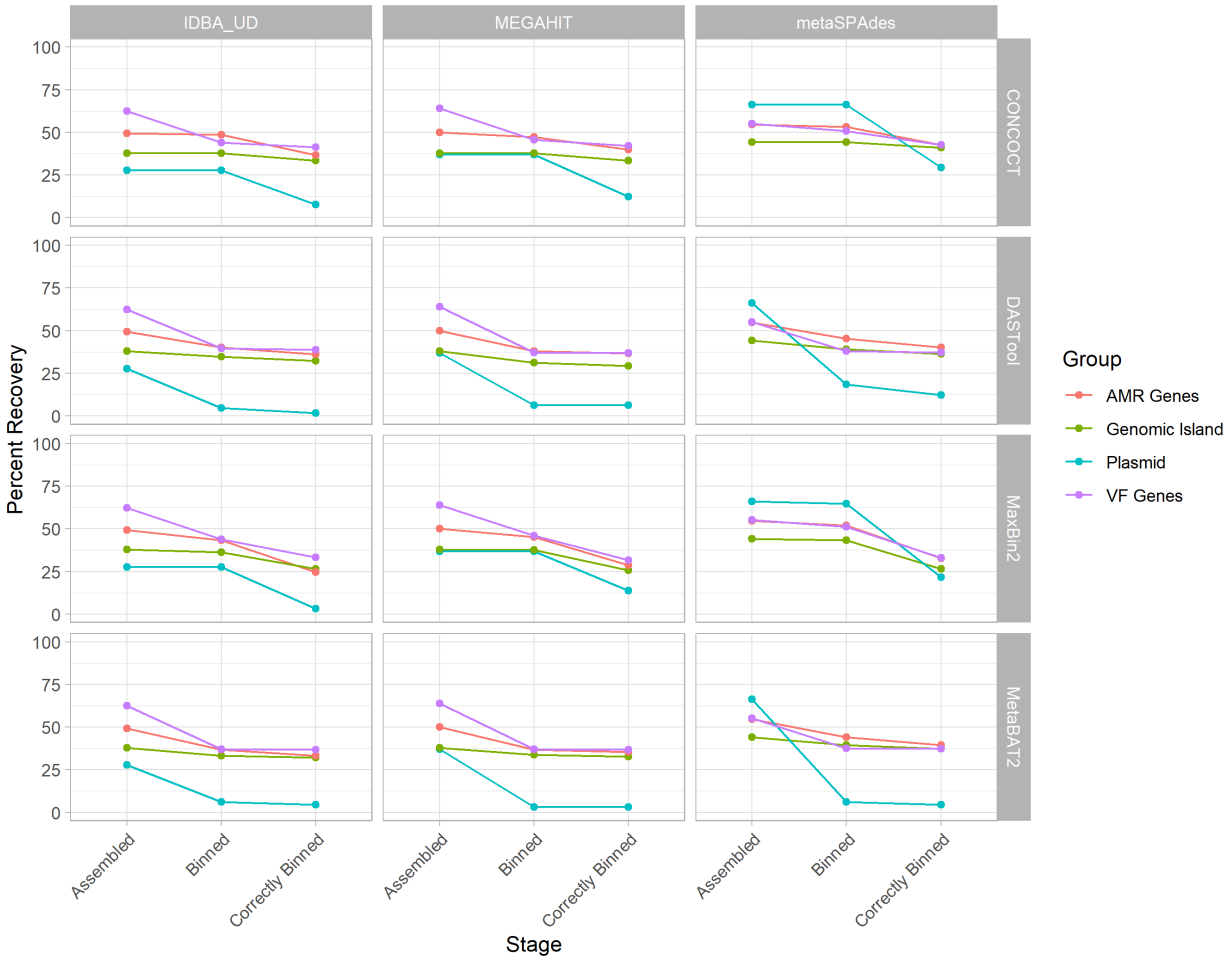
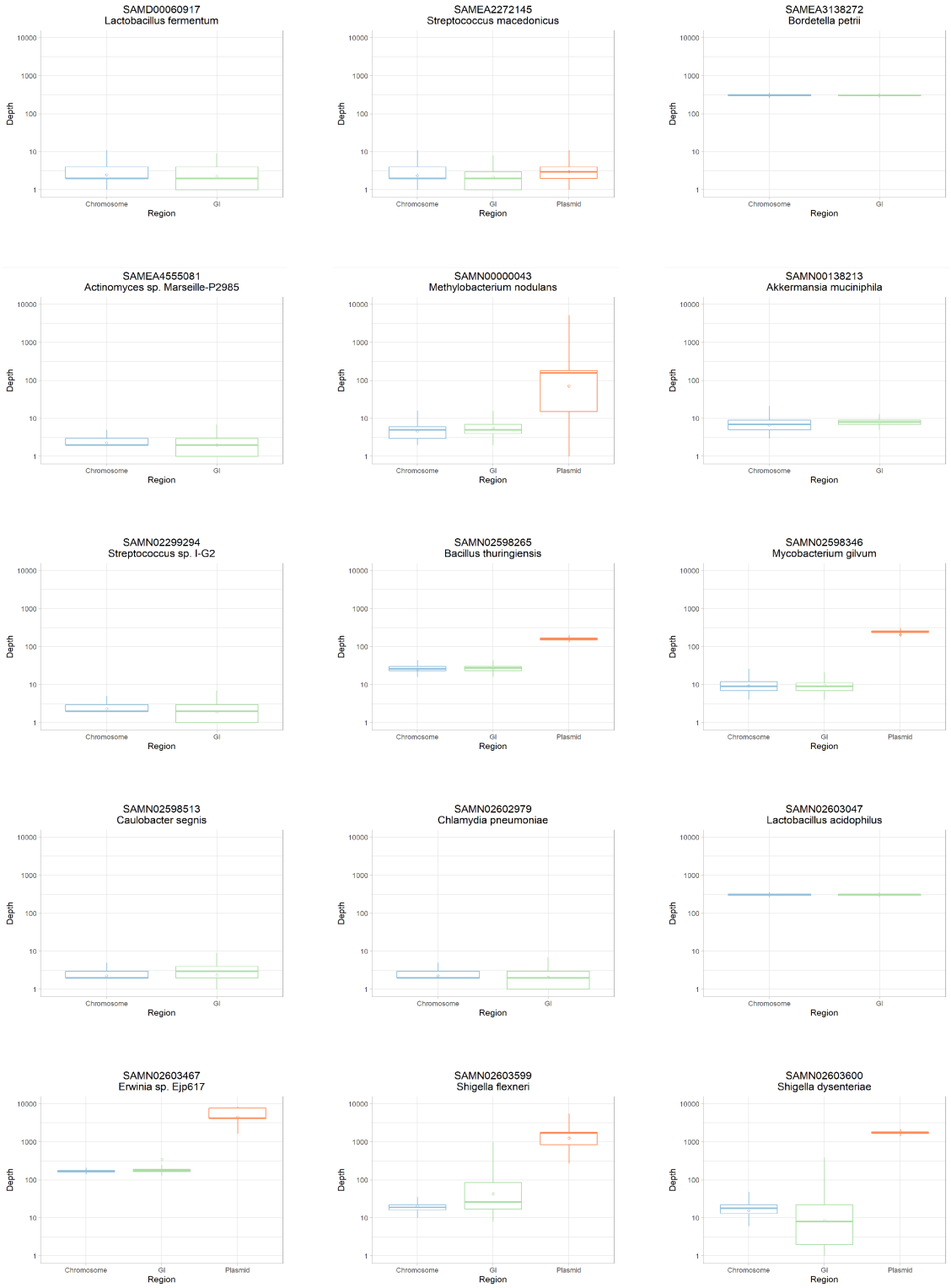


Figure S5: Comparison of rates of loss for different genomic components and gene types across assemblers and binning tools. Each line represents a different component as indicated by the legend with assemblers indicated by row and binning tool by column. This shows that regardless of approach genomic components (GIs and plasmids) are lost at a higher rate than individual VF or AMR genes.

Detailed Simulated Read Depth Analysis

Depth of Simulated Reads By Species (A)



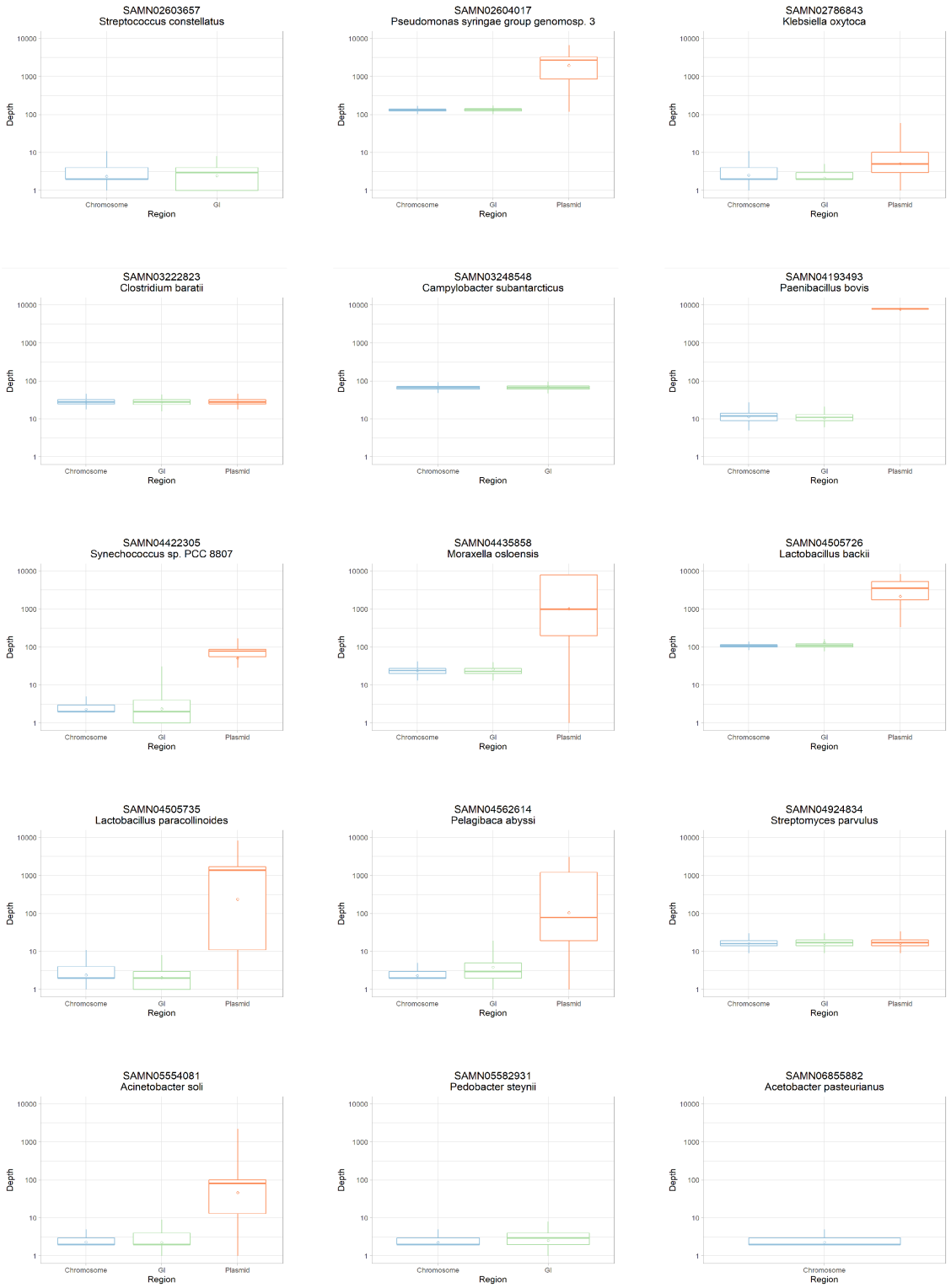
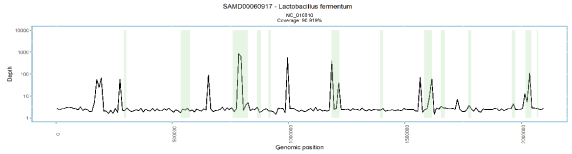
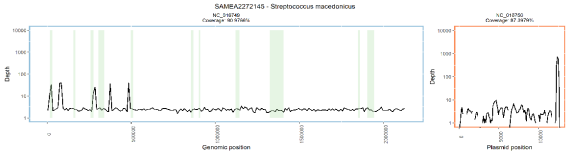
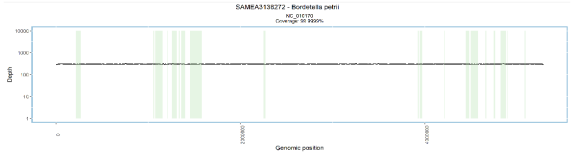


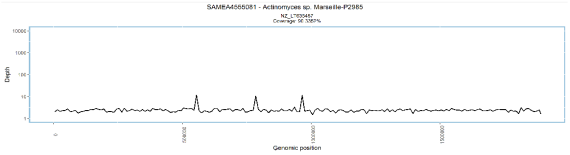
Figure S6: Average Read Depth Per Genome. Across all of the reference species (facet), the read depth of plasmids (orange) is considerably higher relative to chromosomes (blue), likely due to the copy number regime randomly assigned. GIs (green) exhibited relatively lower read depth compared to chromosomes. The variability in read depth is notably higher in and around GIs and plasmids.

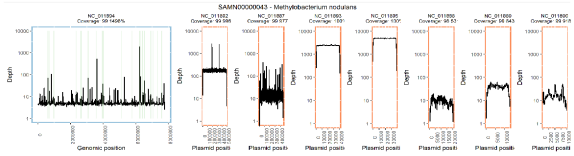
Per Base Depth of Simulated Reads By Species (A)

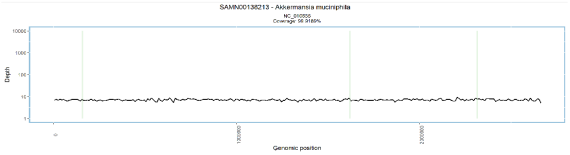


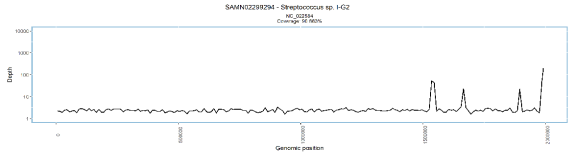


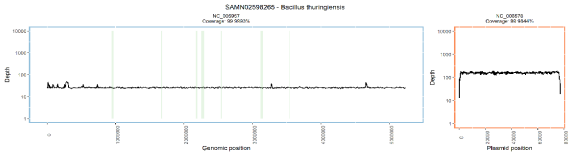


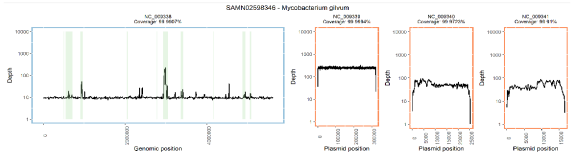


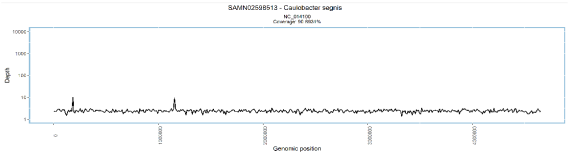


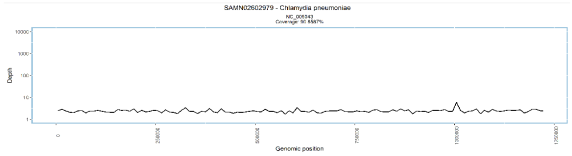


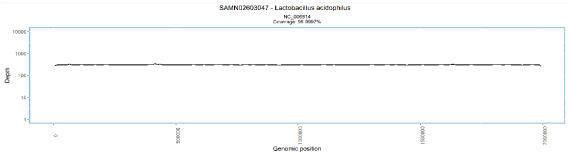


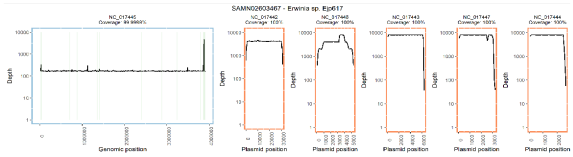


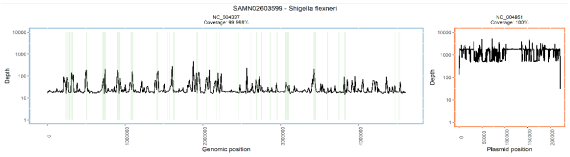


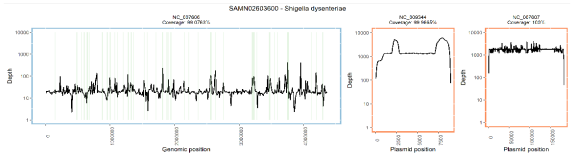


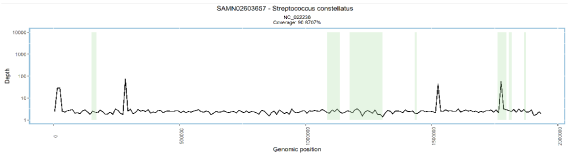












Per Base Depth of Simulated Reads By Species (B)

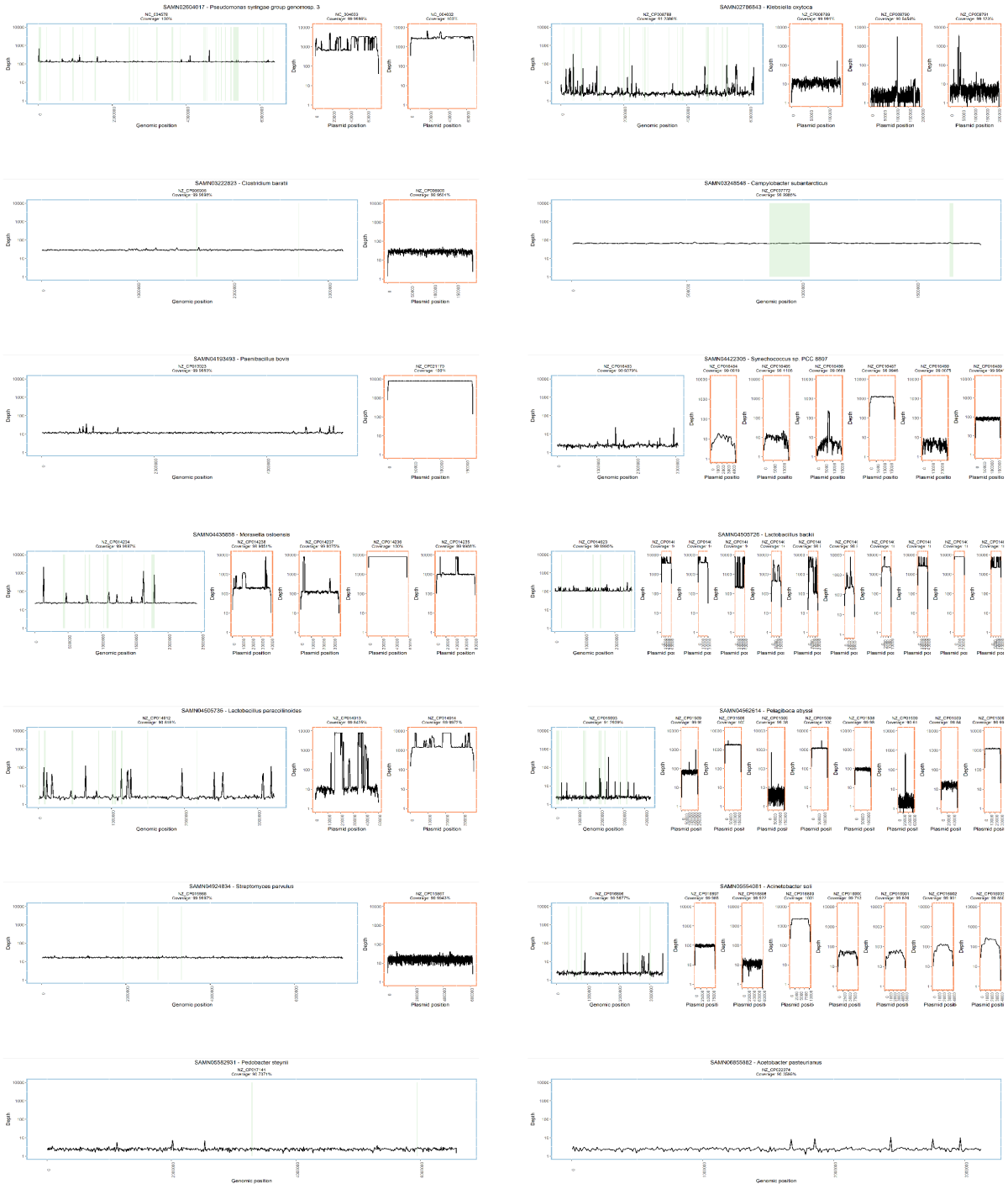


Figure S7: Per Base Read Depth Per Species. The per base (x-axis) read depth (y-axis) of each species is plotted individually. Overall, the read depth of chromosomes (blue boxes) is much lower than the

read depth of plasmids (orange boxes). GIs within the chromosome are highlighted in green. At a per base level, we see a much lower read depth at the beginning and the end of each replicon as well as a higher variability in read depth for GIs and plasmids.