1 Are EnTrI results biased?

Since transposon insertion biases can affect the essentiallity level inferred from transposon mutagenesis experiments, the dataset has been tested for two types of biases: the distance from the origin and GC content. The distance bias in every individual strain is depicted in Figure 1. These plots indicate that the bias is negligible in some strains like Salmonella typhi, while the insertion indices of the genes in other strains need to be normalised by their distances from the origin. We have used the LOESS curve and for each strain, divided the insertion indices by the predicted LOESS value to normalise the insertion indices. As the distribution of new insertion indices will be around 1, we have multiplied the resulted values by the mean of the initial insertion indices to have a distribution around the mean. The results are shown in Figure 2.

We have also checked for GC bias. There was a bias towards the GC content of the genes and we normalised the insertion indices for GC content in the same way as we did for gene position. The results, before and after the normalisation, are plotted in Figure 3. To see if there is any positional bias for any nucleotide, the nucleotides around the insertion sites (the insertion site and 10 nucleotides on each side) are stacked on top of each other and a sequence logo is generated from these sequences. It can be inferred from Figure 4 that there is no significant bias in any position.

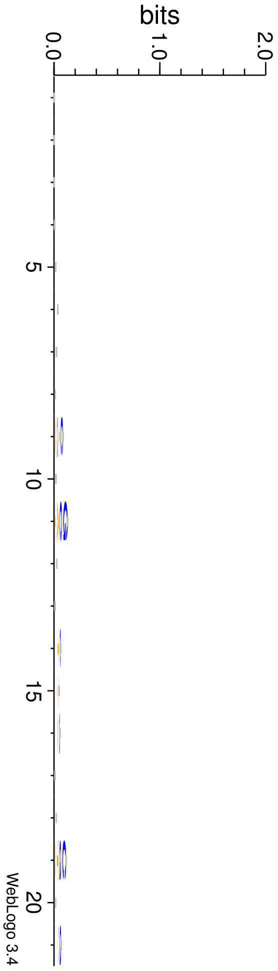

Figure 4: The nucleotides around the insertion sites (the insertion site and 10 nucleotides on each side) are stacked on top of each other and a sequence logo is generated from these sequences using webLogo stand-alone package. The height of the letter stack in each position shows how conserved the bases in that position are.

2 Can we recover phylogenetic information from the essential genes?

To get the evolutionary relationship among all our strains, we collected all clusters with one and only one gene per genome and concatenated all the genes corresponding to every strain. Then aligned them using mafft and generated a phylogenetic tree using fasttree software. The resulted phylogenetic tree is depicted in Figure 5.

To test if the same tree can be obtained from the essentiality of genes, we have selected all clusters that contain exactly one gene from each strain (82 clusters) and made a binary matrix from the es-

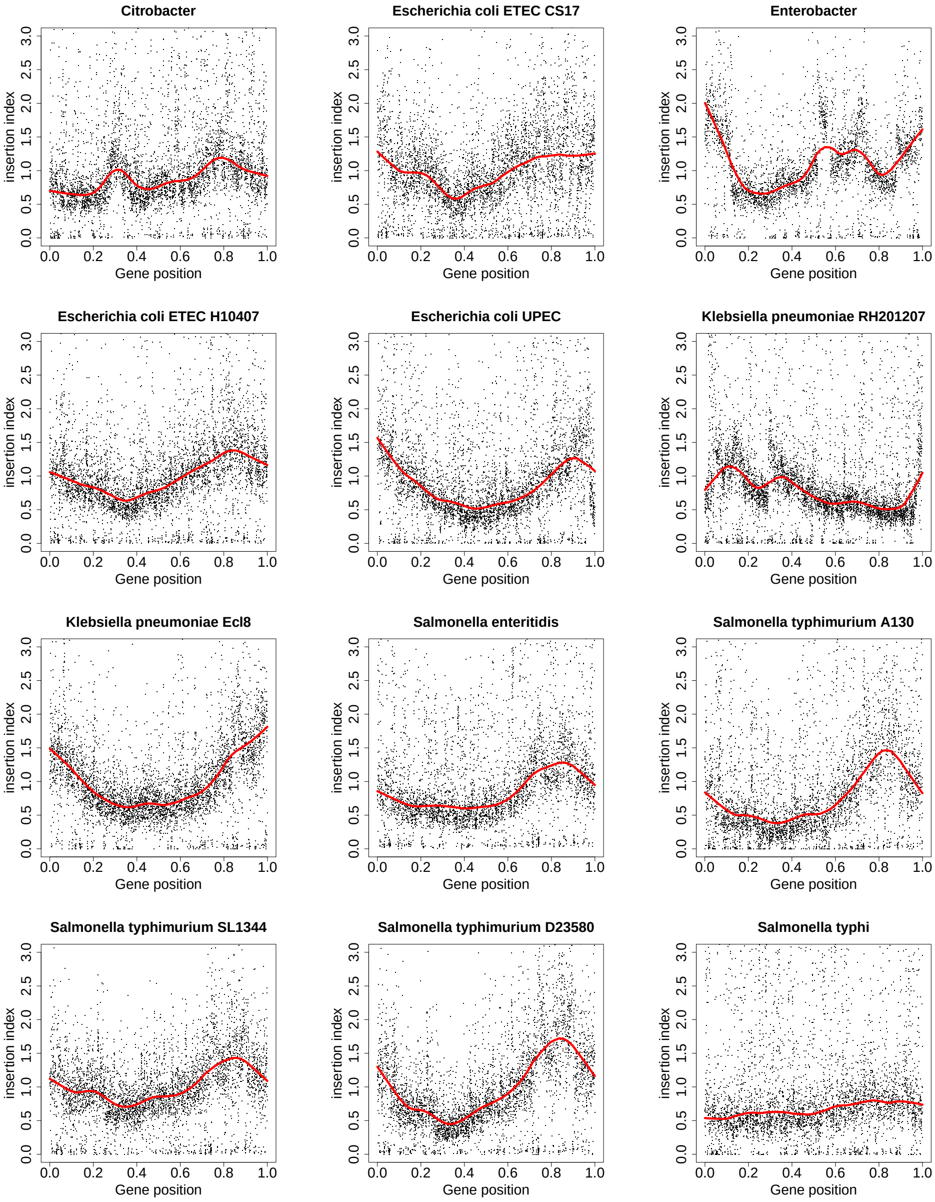

Figure 1: The bias towards the position of the gene for every individual strain. The red curves show the fitted LOESS curves.

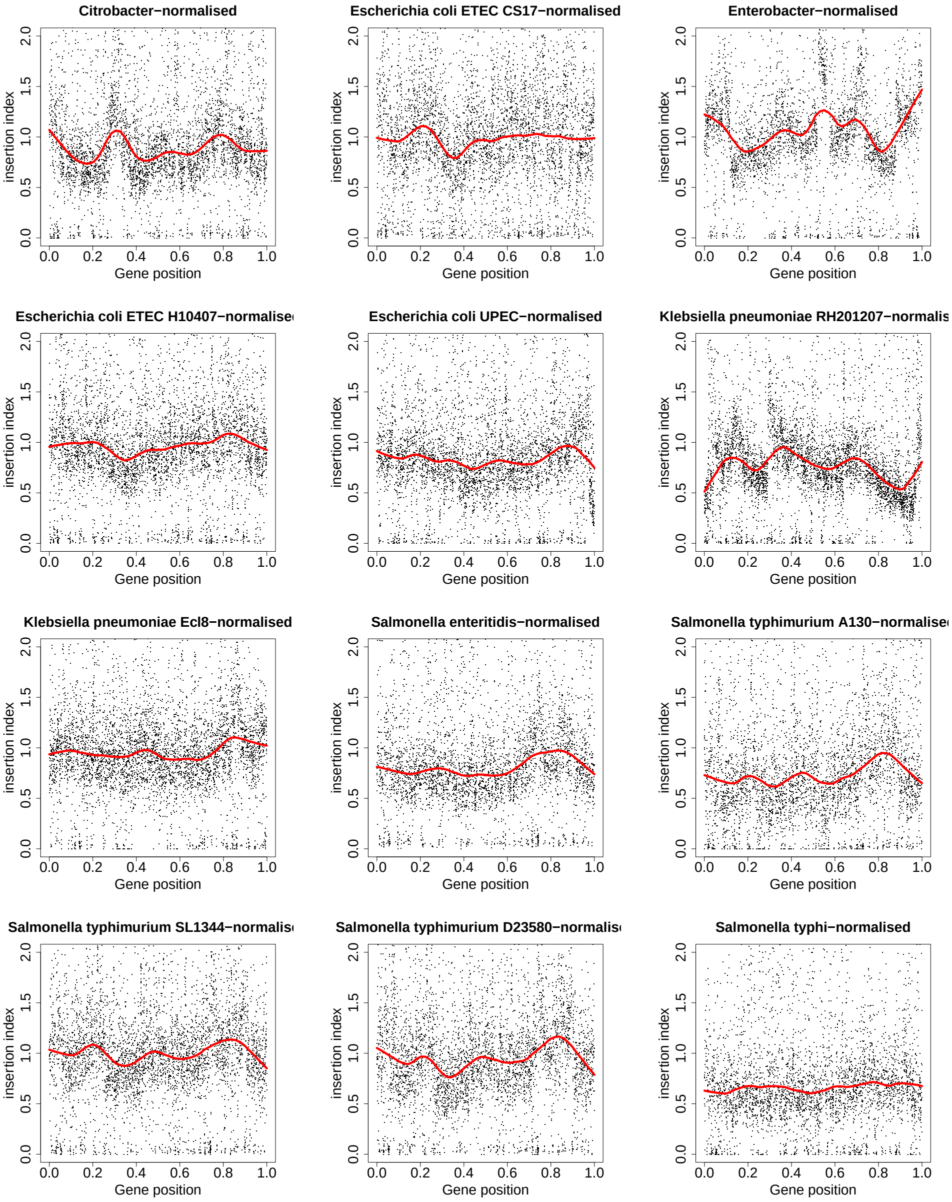

Figure 2: The insertion indices are normalised by the predicted LOESS value and the mean of the insertion indices. The red curves show the fitted LOESS curves.

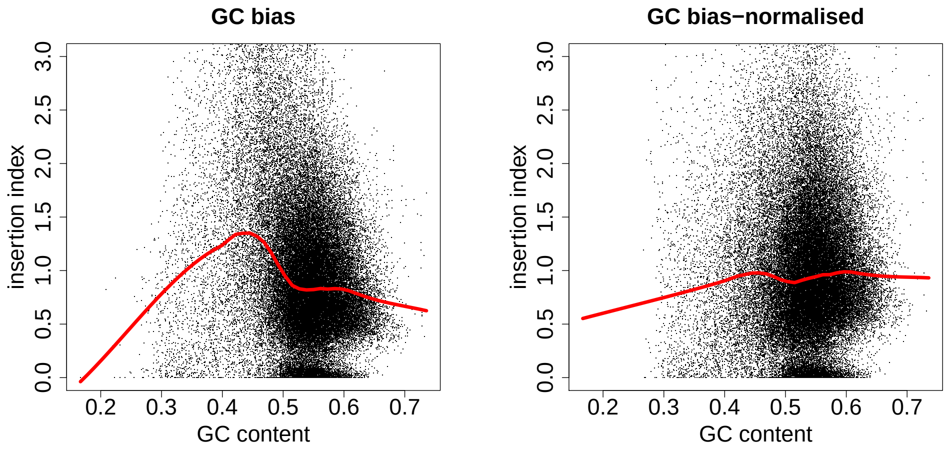

Figure 3: The bias towards GC content. The LOESS curve is shown in red.

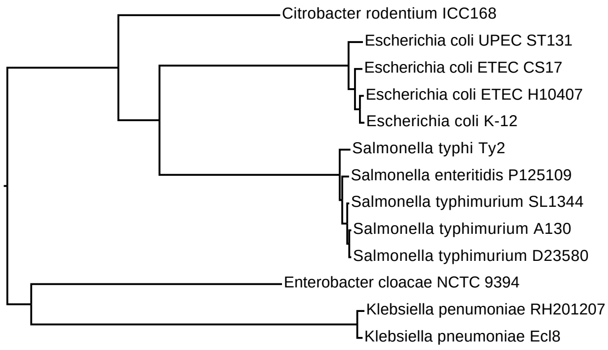
Citrobacter rodentium ICC168
Escherichia coli UPEC ST131
Escherichia coli ETEC CS17
Escherichia coli ETEC H10407
Escherichia coli K-12
Salmonella typhi Ty2
Salmonella enteritidis P125109
Salmonella typhimurium SL1344
Salmonella typhimurium A130
Salmonella typhimurium D23580
Enterobacter cloacae NCTC 9394
Klebsiella penumoniae RH201207
Klebsiella pneumoniae Ecl8

Figure 5: The tree is generated from all the clusters with one and only one gene per strain.

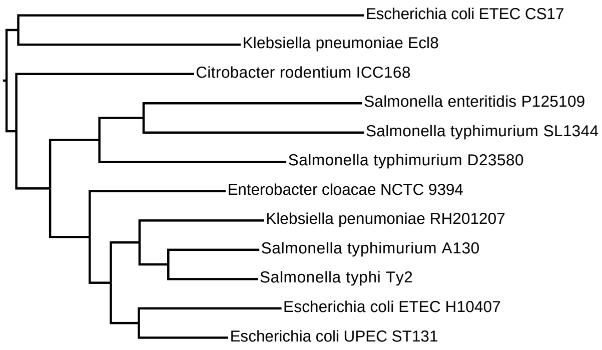
Escherichia coli ETEC CS17
Klebsiella pneumoniae Ecl8
Citrobacter rodentium ICC168
Salmonella enteritidis P125109
Salmonella typhimurium SL1344
Salmonella typhimurium D23580
Enterobacter cloacae NCTC 9394
Klebsiella penumoniae RH201207
Salmonella typhimurium A130
Salmonella typhi Ty2
Escherichia coli ETEC H10407
Escherichia coli UPEC ST131

Figure 6: The phylogenetic tree is generated using "phylip neighbor" software.

sentiality of the genes in these clusters. If a gene is essential in a strain, the corresponding value in the matrix is 1 and if the gene is not essential, the value is 0. Then, we have generated a distance matrix from these values using Bray-Curtis distance and plotted a phylogenetic tree. Figure 6 indicates that the resulting tree does not maintain the phylogenetic information of the species under study.

We have also compared every pair of strains and calculated the number of genes that are essential in one strain and absent in the other, the number of genes that are essential in both strains, the number of genes that are essential in one strain and present but not essential in the other strain, the number of genes that are present in one strain and absent in the other, and the number of genes that are shared between the two strains. The resulted heatmaps can be seen in Figure 7. The dendograms obtained from heatmaps in this figure are consistent with the species tree to some extent.

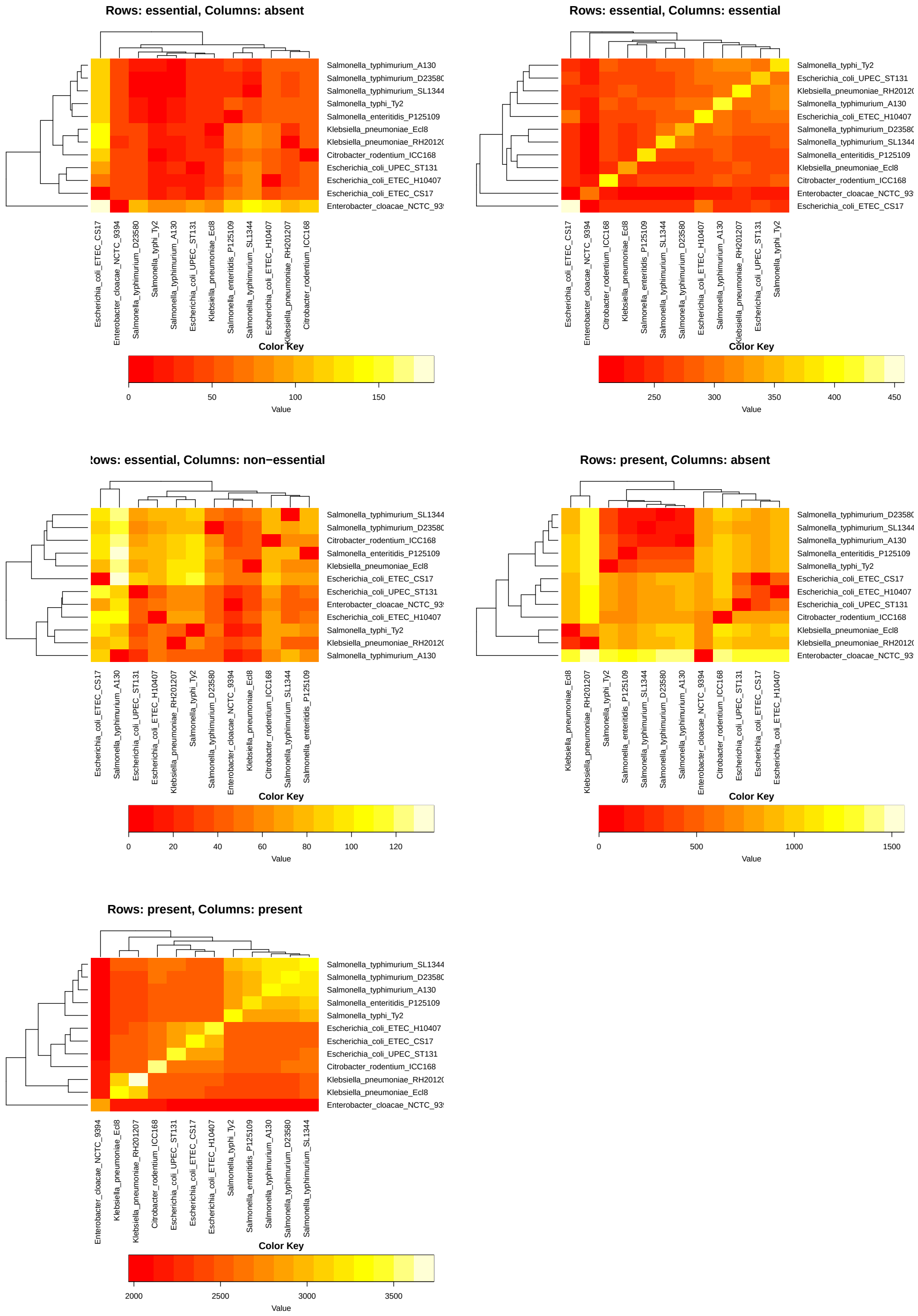

Figure 7: The number of genes that are essential in one strain and absent in the other (upper left), the number of genes that are essential in both strains (upper right), the number of genes that are essential in one strain and present but not essential in the other strain (middle left), the number of genes that are present in one strain and absent in the other (middle right), and the number of genes that are shared between the two strains (lower left).