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. As Figure **??** shows, the insertion index decreases while increasing the distance from the origin. Therefore, there is a bias towards the distance from the origin. The distance bias in every individual strain is depicted in Figure **??**. 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.

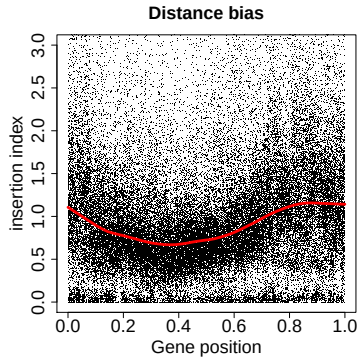

Figure 1: The bias towards the distance from origin (DnaA gene) for all of the species under study. The red curve shows the fitted LOWESS regression curve.

As depicted in Figure **??**, there is no bias towards the GC content of the genes and the regression line is flat. 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 to see if there is any positional bias for any nucleotide. It can be inferred from Figure **??** that there is no bias in any position.

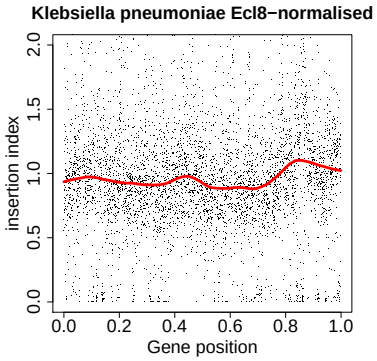

Figure 3: The bias towards GC content. The regression line is depicted by green.

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 **??**.

To test if the same tree can be obtained from the essentiality of genes, we have selected all clusters that contain no more than one gene from each strain and made a binary matrix from the essentiality 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 does not exist in the strain or is not essential, the value is 0. Then, we have generated a distance matrix from these values and plotted a phylogenetic tree. Figure **??** 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

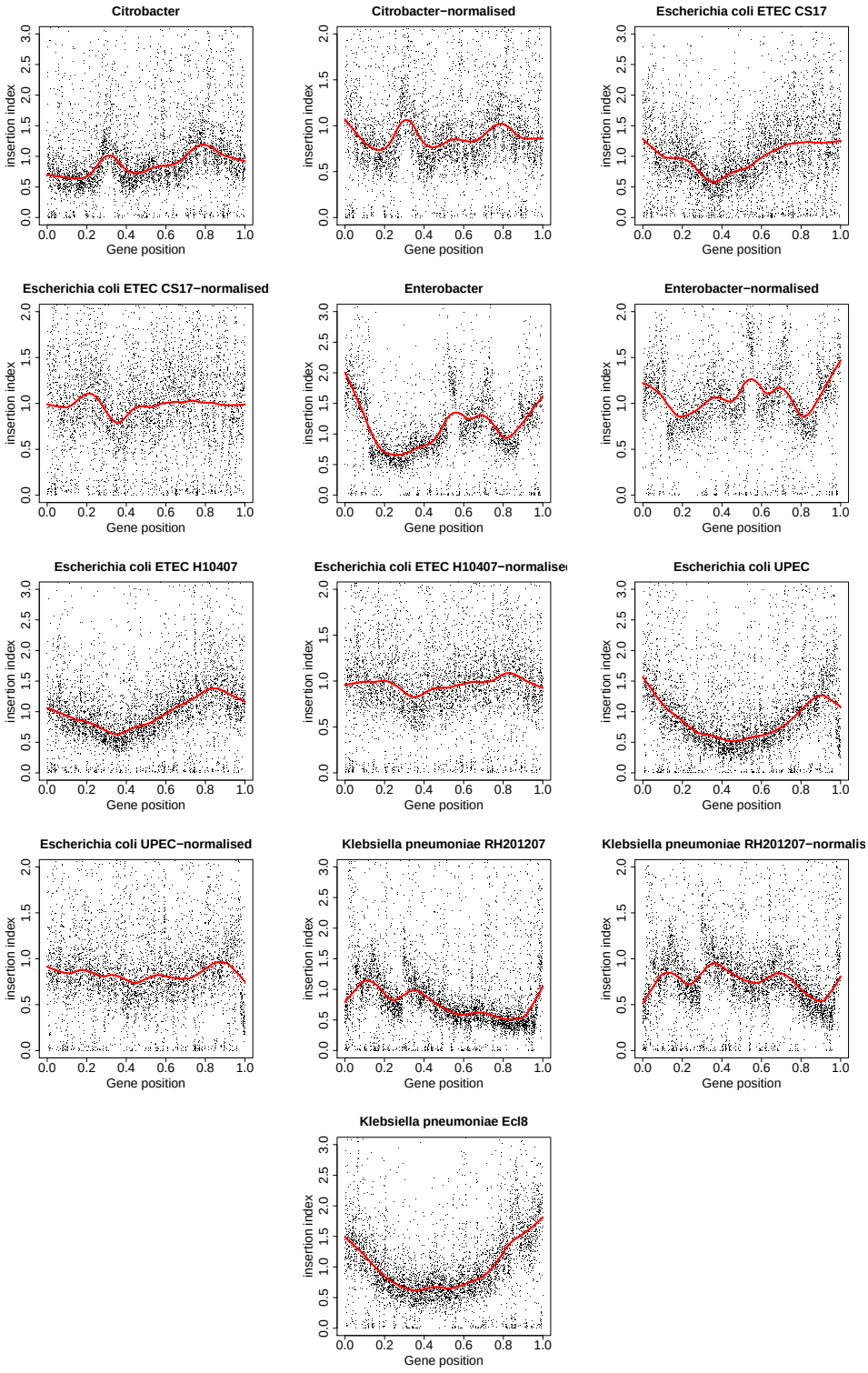

Figure 2: The bias towards the distance from origin for every individual strain. The red curves show the fitted LOWESS curves.

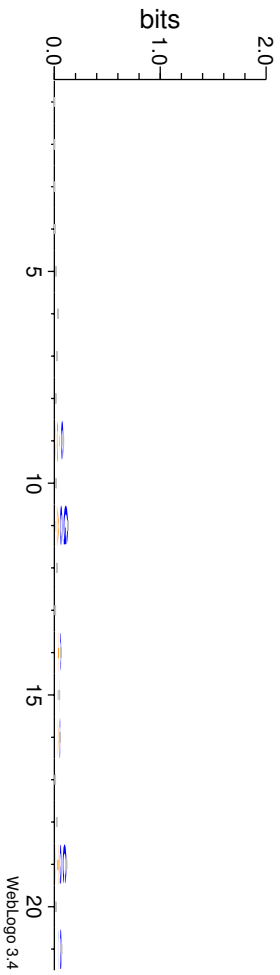

Figure 4: The nucleotides around the insertion sites (the insertion site and 10 nucleotides on each side) are stacked on top of each other and a sequence logo is generated from these sequences using webLogo stand-alone package. The height of the letter stack in each position shows how conserved the bases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

`./speciestree/speciestree.pdf`

Figure 5: The tree is generated using fasttree software

`make-essentiality-tree/essentiality-tree.pdf`

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

between the two strains. The resulted heatmaps can be seen in Figure **??**. The dendograms obtained from heatmaps in this figure are consistent with the species tree to some extent.

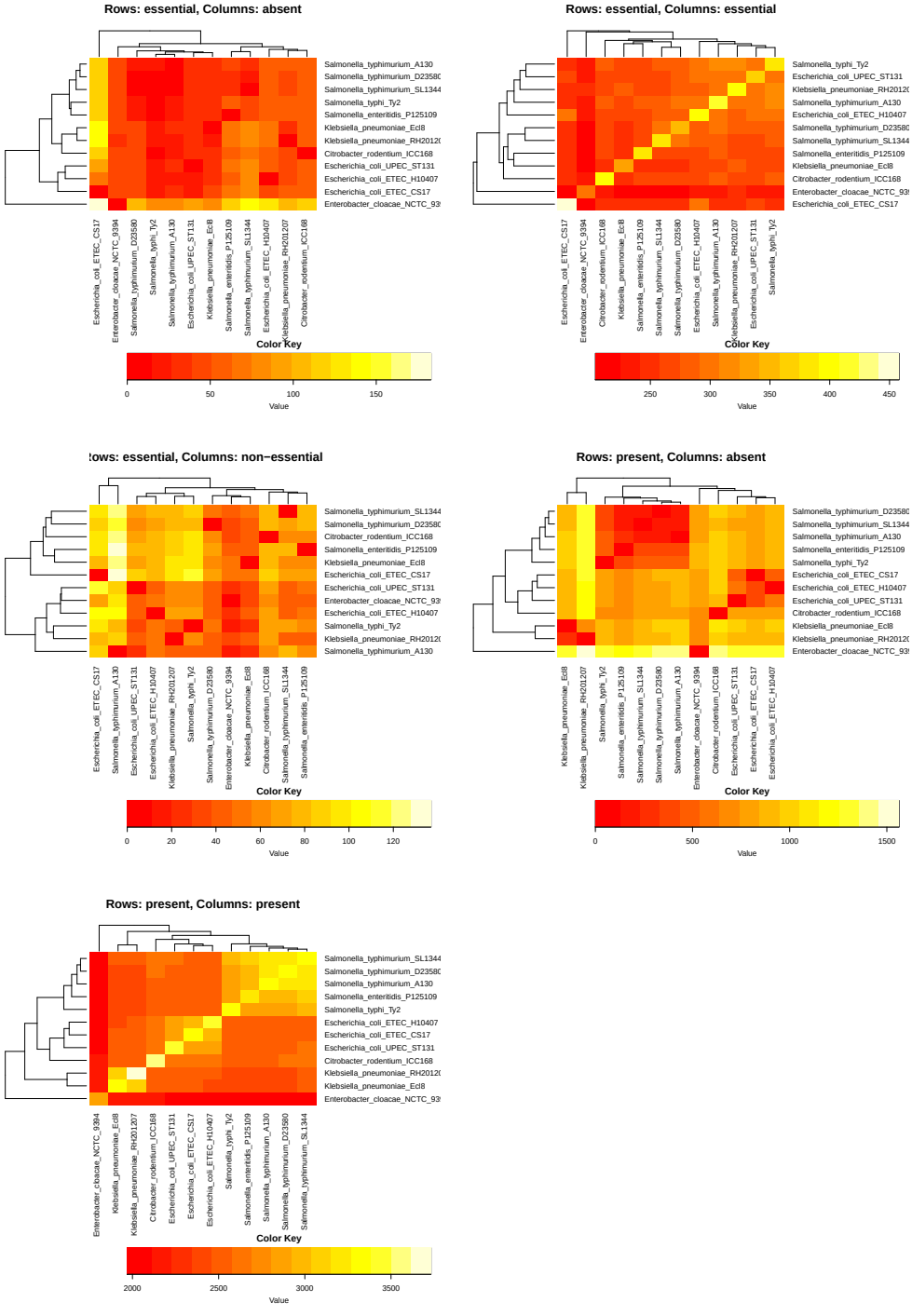

Figure 7: The number of genes that are essential in one strain and absent in the other (upper left), the number of genes that are essential in both strains (upper right), the number of genes that are essential in one strain and present but not essential in the other strain (middle left), the number of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