1 The homologue clustering program

We have developed a program which clusters homologous genes. This program uses jackhmmer from HMMER package for clustering. Figure 1 shows the results of this clustering on all the genes in the 13 species under study. To divide our clusters into ORFan, single copy, and multiple copy groups, we have used the procedure represented in Figure 2.
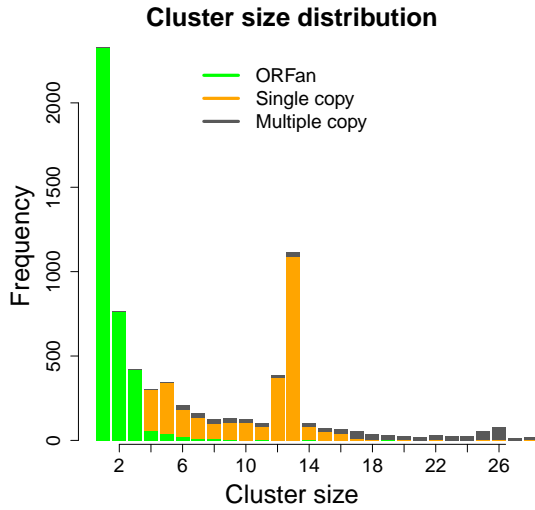

Figure 1: The distribution of the size of the clusters resulted from clustering all homologous genes in the 13 species under study.
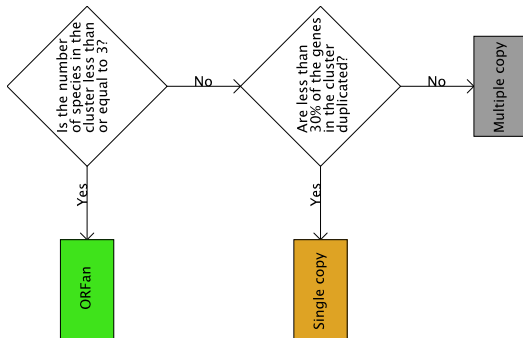

Figure 2: The procedure used to divide clusters into three groups: ORFan, single copy, and multiple copy.

2 Are EnTrI results biased?

Since transposon insertion biases can affect the essentiality level inferred from transposon mutagenesis experiments, the dataset has been tested for three types of biases: the distance from the origin, GC content, and the position of insertions in genes. The distance bias in every individual strain is depicted in Figure 3. 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 4.

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 5. 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 6 that there is no significant bias in any position.
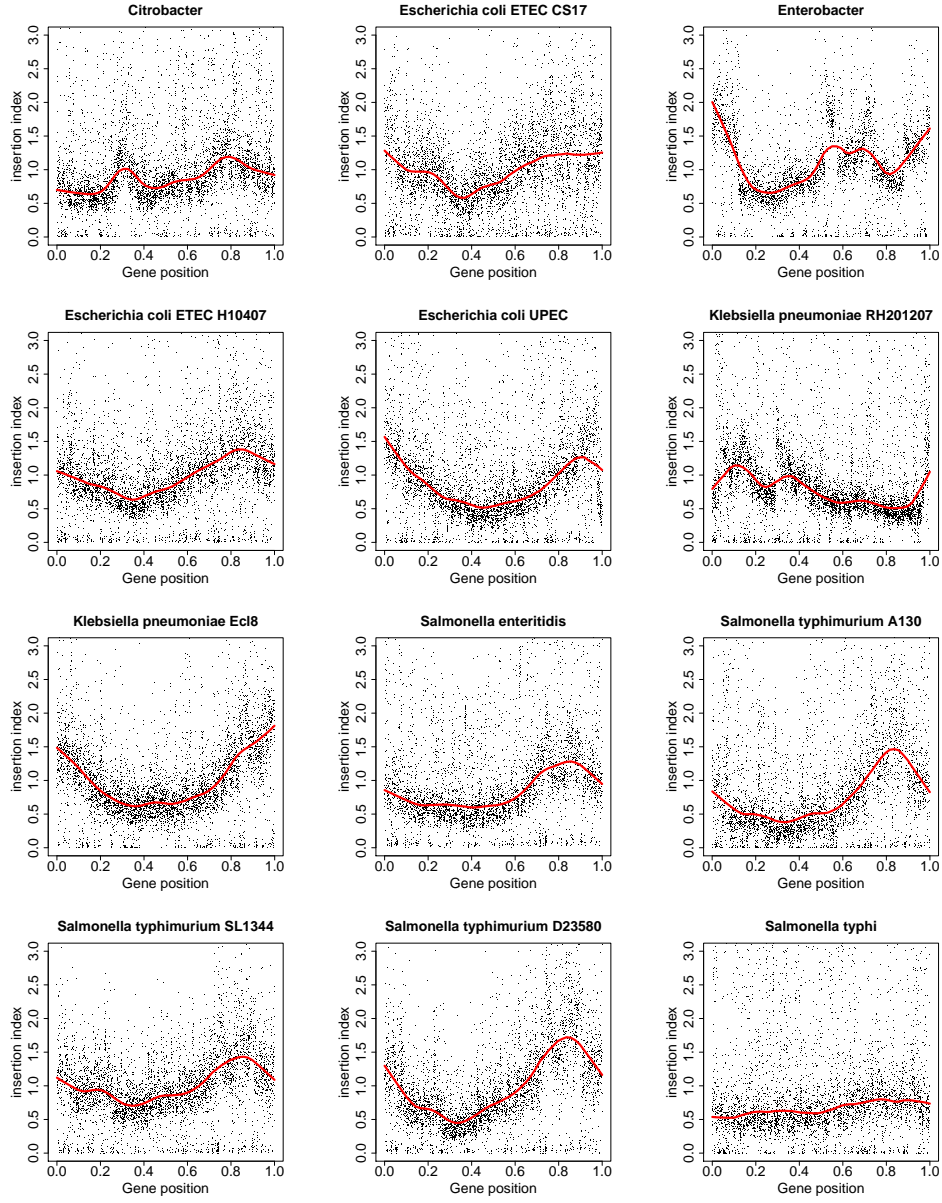
Figure 3: The bias towards the position of the gene for every individual strain. The red curves show the fitted LOESS curves.
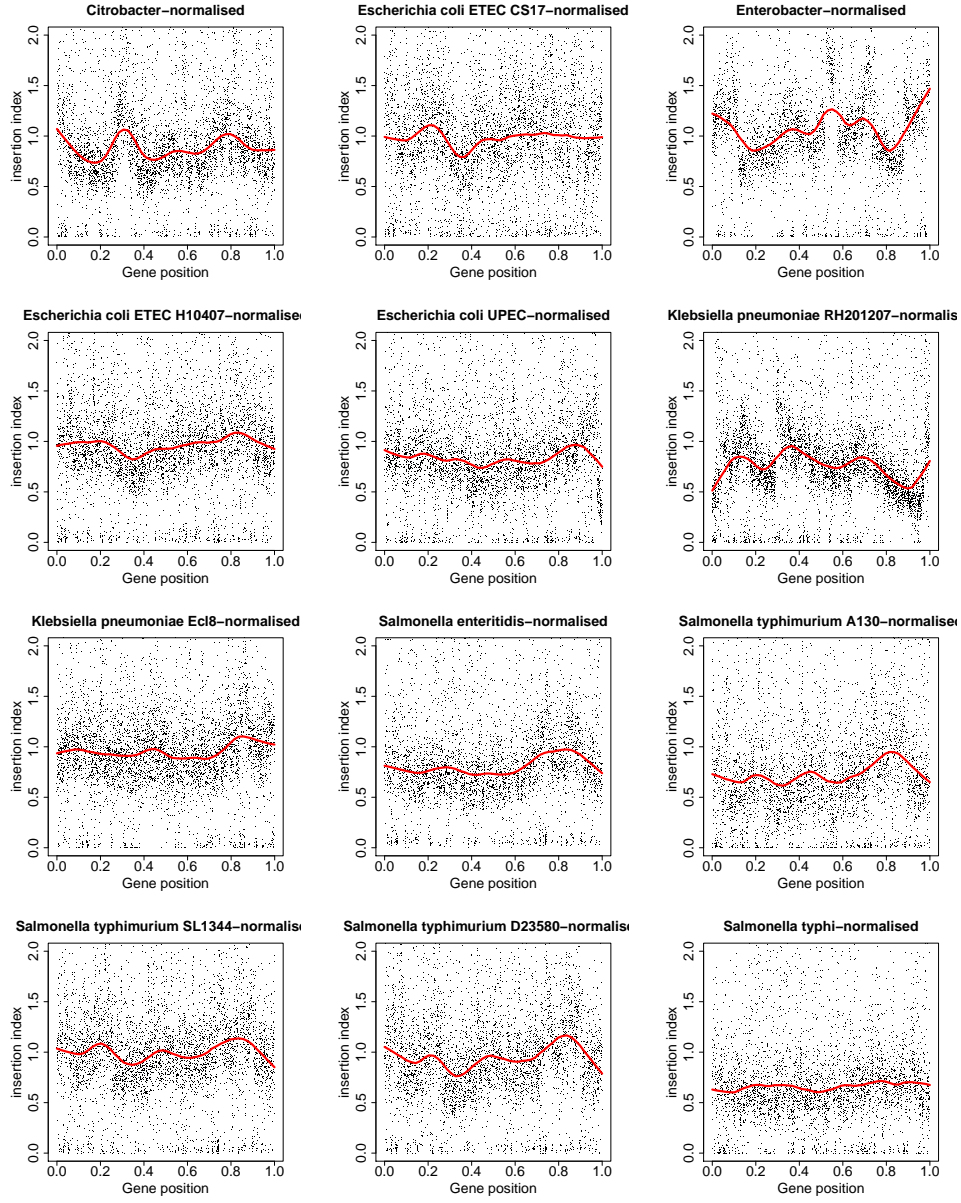
Figure 4: The insertion indices are normalised by the predicted LOESS value and the mean of the insertion indices. The red curves show the fitted LOESS curves.
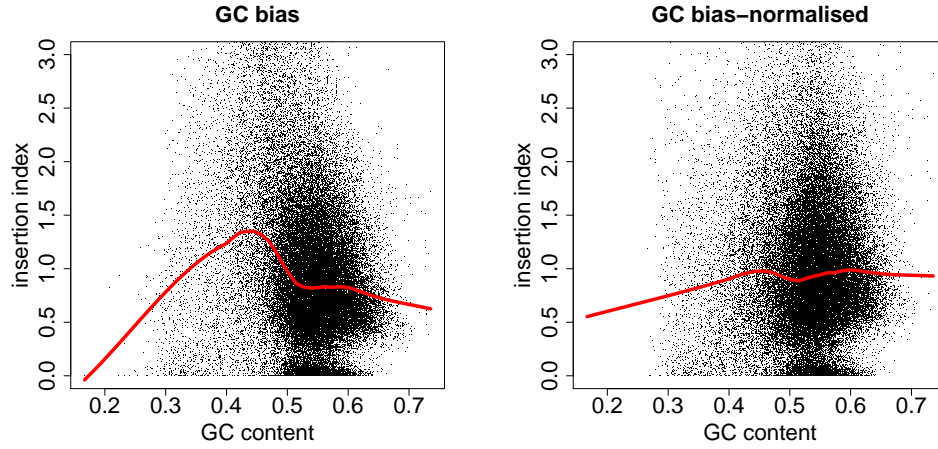
3

Figure 5: The bias towards GC content (left) and the data after normalising GC bias using LOESS curve and mean (right). The LOESS curve is shown in red.
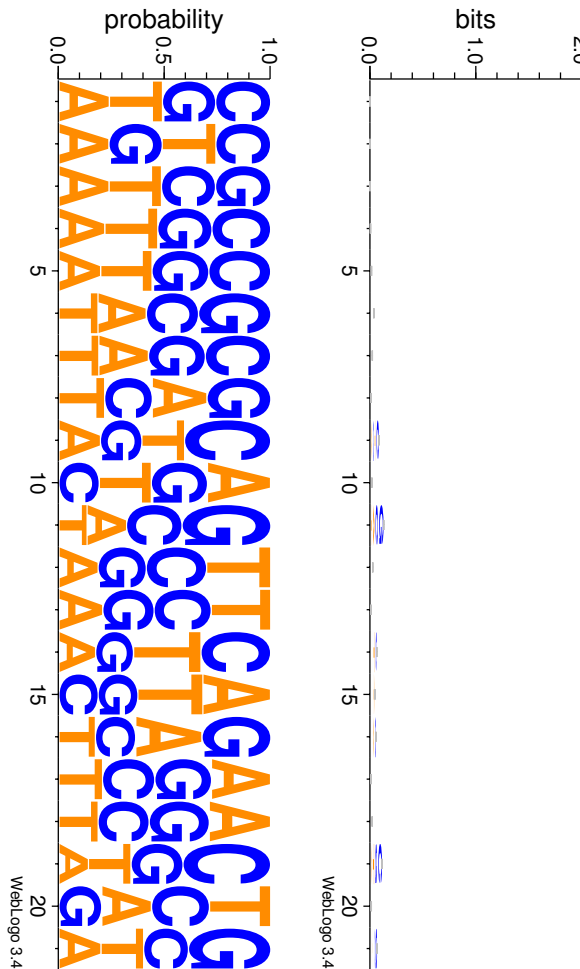

Figure 6: The nucleotides around the insertion sites (the insertion site and 10 nucleotides on each side) are stacked on top of each other and a sequence logo is generated from these sequences using webLogo stand-alone package. The height of the letter stack in each position shows how conserved the bases in that position are.

The bias towards the position of insertions within genes has been studied since the genes can tolerate more insertions at the 3' end of the gene. We have divided our genes into 3 parts: first 60 nucleotides, last 60 nucleotides, and the rest and plotted the normalised number of insertions in each of these groups. Figure 7 summarises the results.

3 Gene essentiality

After normalising the insertion indices, we have used these values to determine the essentiality of genes. The insertion index distribution for each single strain is pictured in Figure 8. If the insertion index is less than 0.2, the corresponding gene is regarded as essential, if it is between 0.2 and 2, the gene is non-essential, and otherwise, it is considered as a beneficial loss.

Figure 9 shows the insertion index distribution for all clusters. We have divided the clusters into ORFan, single copy, and multiple copy groups and studied the essentiality in these groups.
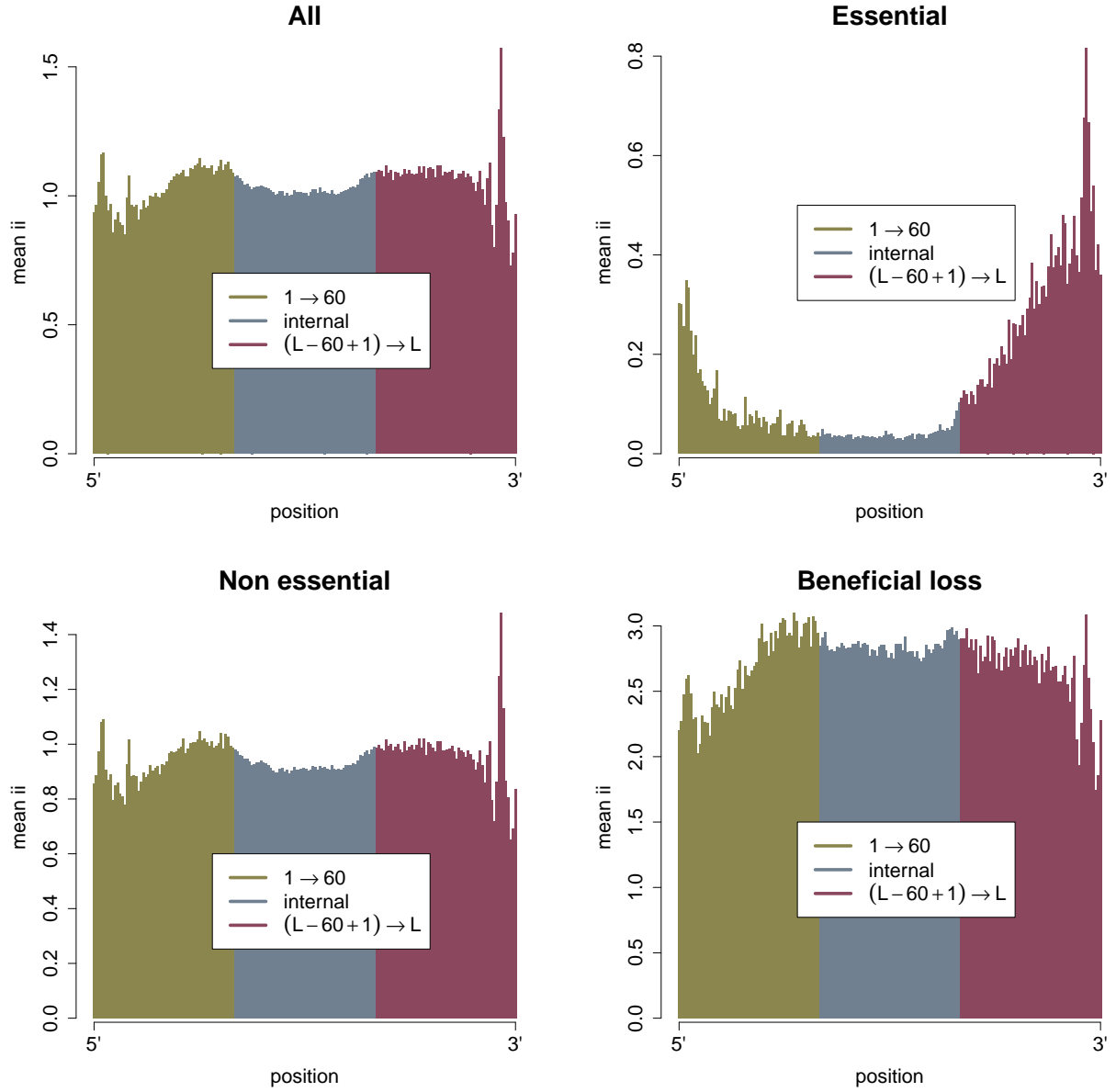
4

Figure 7: We have divided our genes into 3 parts: first 60 nucleotides (position 1-60), last 60 nucleotides (position 121-180), and the rest (position 61-120) and plotted the normalised number of insertion sites in each of these groups.
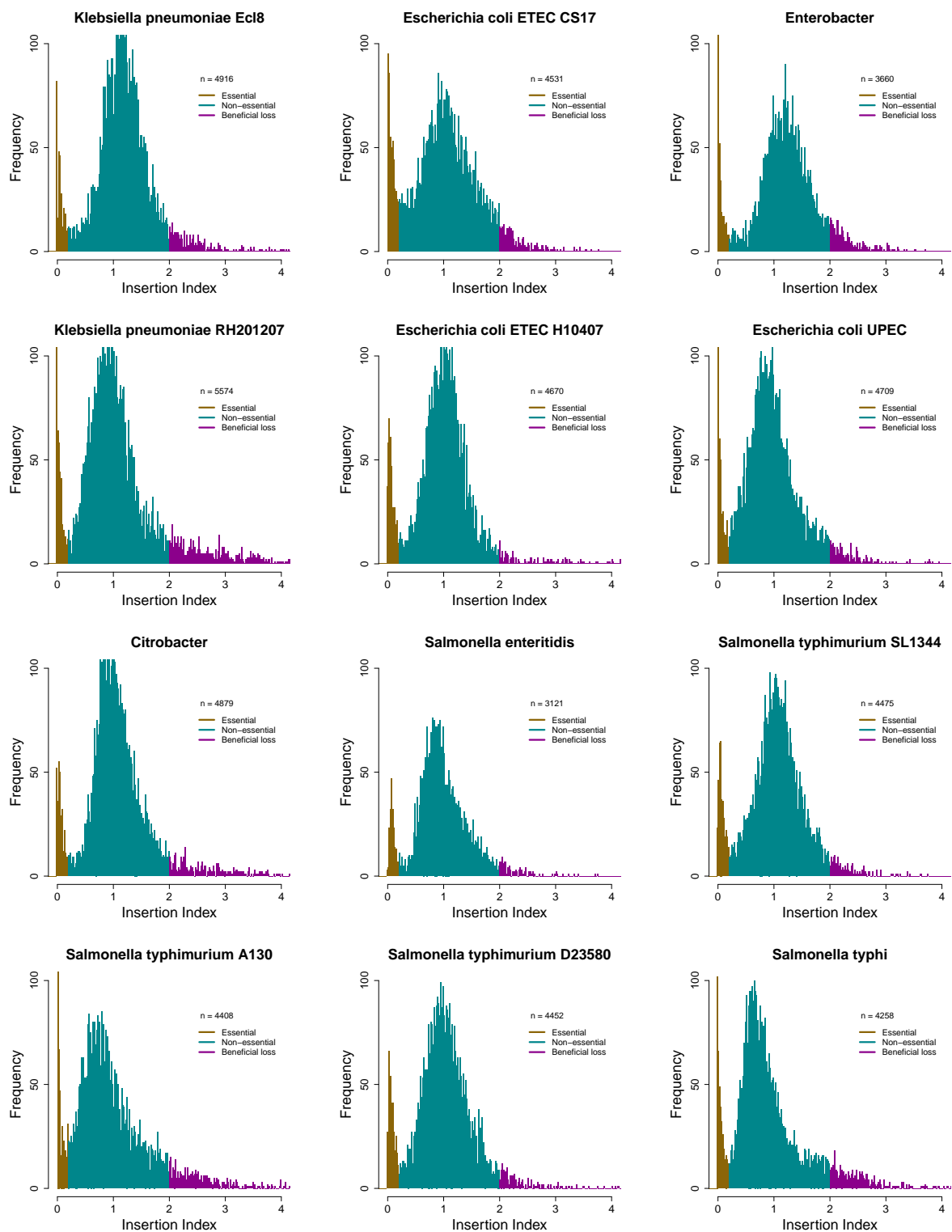
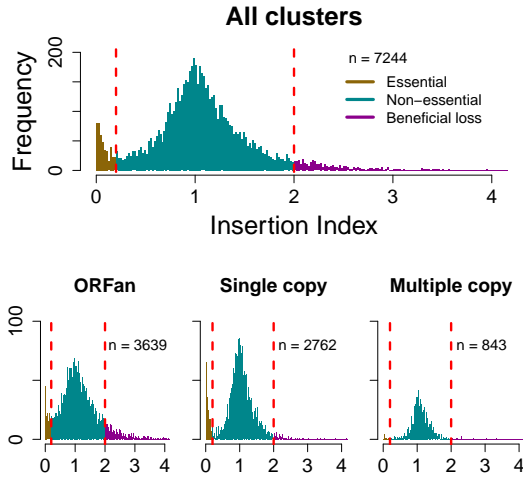Figure 8: Insertion index distribution for every single strain.

Figure 9: Essentiality of clusters in all clusters, OR-Fan, single copy clusters, and multiple copy clusters.
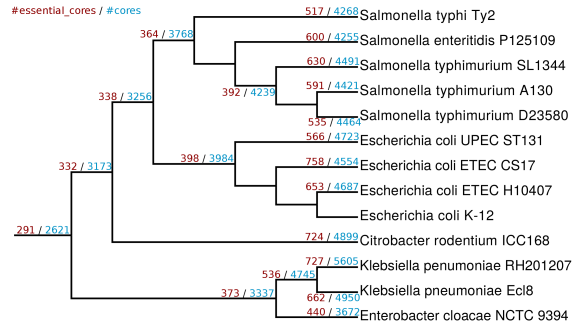

Figure 10: The tree is generated from all genes that are copied only once per genome using an approximately-maximum-likelihood method called fasttree.

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 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 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 using "phylip neighbour" software which uses a neighbour joining method. Figure 11 indicates that the resulting tree does not maintain the phylogenetic information of the species under study.

4 Can we recover phylogenetic information from the essential genes?
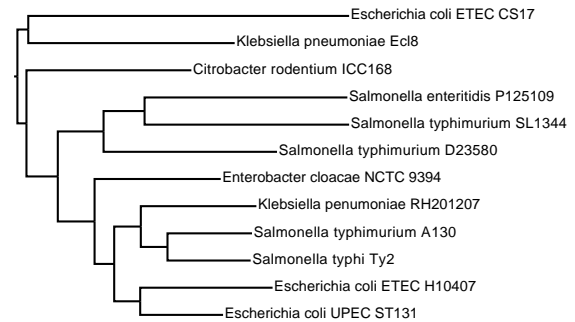
Figure 11: The tree is generated from all the genes that are copied only once per genome. We have made a binary matrix from the essentiality of these genes. Then, we have generated a distance matrix from these values using Bray-Curtis distance and plotted a phylogenetic tree using "phylip neighbor" software which uses a neighbour joining method.

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 10.

7

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 12.

5 Are conserved genes more likely to be essential?

To find out whether conserved genes are more likely to be essential or not we have counted the number of essential, sometimes essential and never essential genes in both core and accessory genomes. The results are summarised in Table 1. Core genes are the ones that have at least one copy per genome, otherwise, the genes are called accessory. The p-values show that the essentiality of genes is not independent of their conservation.
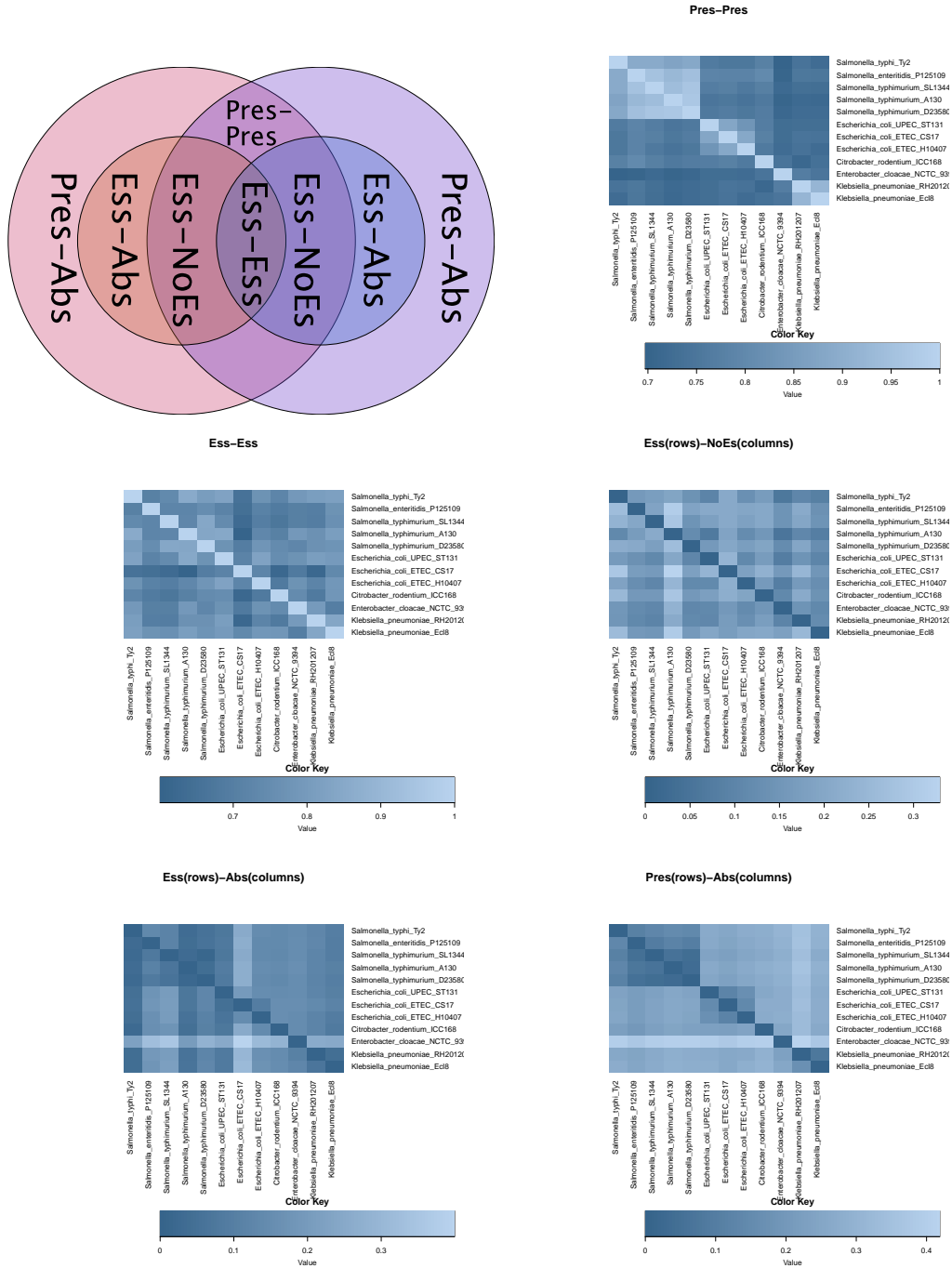
Figure 12: The heat maps shows the normalised number of genes in each group. The genes can be present in both strains (Pres-Pres in the Venn diagram and the upper right heat map), essential in both strains (Ess-Ess in the Venn diagram and the middle left heat map), essential in one species and not essential in the other (Ess-NoEs in the Venn diagram and the middle right heat map), essential in one species and absent in the other species (Ess-Abs in the Venn diagram and the bottom left heat map), and present in one species and absent in the other (pres-Abs in the Venn diagram and the bottom right heat map).

Species	Core				Accessory				Fisher's test
Name	Ess	SEss	NEss	Sum	Ess	SEss	NEss	Sum	p-value
Salmonella typhimurium A130	266	665	1621	2552	53	327	1528	1908	$< 2.2e - 16$
Salmonella typhimurium D23580	267	670	1624	2561	54	339	1550	1943	$< 2.2e - 16$
Salmonella typhimurium SL1344	266	674	1630	2570	63	350	1546	1959	$< 2.2e - 16$
Salmonella enteritidis P125109	266	679	1615	2560	61	311	1015	1387	$= 2.669e - 14$
Salmonella typhi Ty2	265	628	1564	2457	56	294	1505	1855	$< 2.2e - 16$
Escherichia coli UPEC ST131	299	685	1635	2619	62	323	1767	2152	$< 2.2e - 16$
Escherichia coli ETEC CS17	280	656	1603	2539	107	356	1586	2049	$< 2.2e - 16$
Escherichia coli ETEC H10407	284	656	1623	2563	89	329	1743	2161	$< 2.2e - 16$
Citrobacter rodentium ICC168	304	685	1630	2619	96	356	1868	2320	$< 2.2e - 16$
Enterobacter cloacae NCTC 9394	280	623	1526	2429	34	129	1128	1291	$< 2.2e - 16$
Klebsiella penumoniae RH201207	357	864	1780	3001	69	279	2314	2662	$< 2.2e - 16$
Klebsiella pneumoniae Ecl8	357	840	1770	2967	61	225	1751	2037	$< 2.2e - 16$

Table 1: The number of essential (Ess), sometimes essential (SEss) and never essential (NEss) genes in both core and accessory genomes. Core genes are the ones that have at least one copy per genome, otherwise, the genes are called accessory.