Is the essentiality of genes conserved in Enterobacteriaceae?
Is gene essentiality conserved?
Are essential genes conserved?

Abstract

Introduction

Studying the essentiality of genes helps with identifying the fundamental processes necessary for cell viability [17]. So far, scientists have studied the essential genes in organisms from different domains of life [20]. The results have led to new insights for developing new antibiotics that target essential genes of pathogenic bacteria [8, 21] and synthesising new genomes [15, 16]. Researchers have used different methods for studying the essentility of genes in prokaryotes. Baba et al. [1] have made a library of single gene deletions using phage lambda Red recombination system to screen essential genes while another group have used antisense RNA knockdowns for this purpose [26]. Another method that is widely used due to its simplicity and accuracy is transposon mutagenesis along with high-throughput sequencing [7, 13, 14, 19, 22, 24, 25]. In this method, pools of single insertion mutants are constructed using transposon mutagenesis and the effect of each mutation on the survival of mutants is evaluated by sequencing the survivors [2]. This can lead to the identification of essential genes.

Although the essentiality of genes has been studied in a variety of organisms, there is still room to study the evolutionary conservation of essentiality. Curtis and Brun [10] have studied the essentiality changes in cell cycle genes of three alpha-proteobacteria strains: *Caulobacter crescentus*, *Brevundimonas subvibrioides*, and *Agrobacterium tumefaciens*.

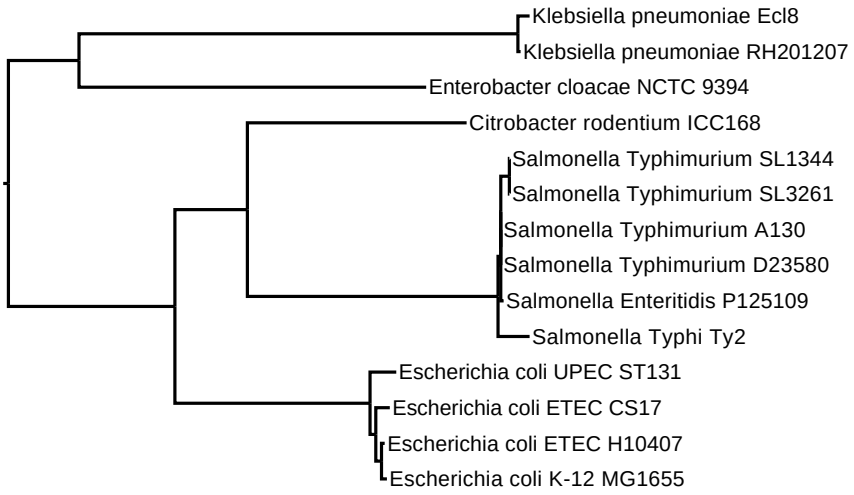

Figure 1. The species tree containing the 13 strains under study and Escherichia *coli* K-12 MG1655 studied in Keio collection [1]. We have generated the tree by running RAxML [23] on Phylosift [11] amino acid markers.

Canals et al. [6] have compared the essentiality of genes in *Salmonella* typhimurium and *Salmonella* Typhi. In a similar study, Barquist et al. [3] have used transposon-directed insertion-site sequencing to study the differentiation of the essentiality of genes in *Salmonella* serovars Typhi and Typhimurium which has led to divergence in their pathogenecity and host ranges. We extend this research by studying 13 bacterial strains from Enterobacteriaceae. These strains and *Escherichia coli* K-12 MG1655 studied by Baba et al. [1] are depicted in Fig. 1.

{A summary of what we have done}

1 Transposon mutagenesis

overview paragraph.

We have studied 2 *Klebsiella* strains, an *Enterobacter* strain, a *Citrobacter* strain, 6 *Salmonella* strains, and 3 *Escherichia* strains and compared the essentiality of genes in these strains and *Escherichia coli* K-12 MG1655 from another study [1]. These strains are all selected from Enterobacteriaceae family. Enterobacteriaceae is a family that includes Gram-negative bacteria with different host ranges and pathogenecity found in soil, water,

plants, animals and humans [5]. In humans, various strains from this family can cause diarrhoea, septicaemia, urinary tract infection, meningitis, respiratory disease, and wound and burn infection [5]. Besides, they can infect poultry and livestocks and cause financial losses for farmers [5]. Here, we perform a transposon-directed insertion-site sequencing experiment to study the conservation of essentiality of genes in strains from 5 different species in this family.

We have used Tn5 transposon to generate single inserted mutants and placed our mutants in a selective media for Tn5. We have picked the mutants and pooled them and used the splinkerette adapter and primers designed in [4] for PCR enrichment. Then we have sequenced the fragments and mapped them back to the genome to figure out the number of insertions that have been tolerated in each position of the genome. The number of insertions in a gene can imply the degree of essentiality for that gene.

We have used a value called insertion index to evaluate the essentiality of a gene. This value is calculated by summing up the number of transposon insertion sites observed in a gene. Since the lengths of the genes are different, we have normalised the insertion index by dividing it by gene length. Our experiment has been performed on different strains and the library density is different in each experiment. Therefore, in order to make the insertion indices comparable in all the strains, we have normalised our insertion indices by the ratio between the number of insertions in the whole genome and the length of the genome. We have not studied genes shorter than 100 base-pairs as they might not be targeted by any transposon due to their shortness.

We have divided the genes into three groups: essential genes, ambiguous, non-essential genes, and beneficial losses. We have adapted the pipeline introduced by Barquist et al. [4] to evaluate the essentiality of genes. The insertion index distribution plot has two peaks and a heavy tail as shown in Fig. 2. The first peak shows the genes with no or just a few insertions which are considered as essential genes. We have fitted an exponential distribution to the first peak and a gamma distribution to the second one. Then, we have calculated the log odds ratio for belonging to each of these distributions for each gene. The region that has log odds value between -2 and 2 is called the ambiguous region, the genes belonging to the first peak are essential and the rest of the genes are not essential. Among genes that are not

essential, any gene for which the value of the cumulative distribution function for the gamma distribution is greater than or equal to 0.99 is considered as a beneficial loss and the other genes are non-essential genes.

2 Are there biases in transposon mutagenesis data?

OVERVIEW PARAGRAPH:

Different articles have reported biases in transposon mutagenesis [3, 18, 22]. We have performed a detailed study of these biases. **LIST TYPES OF BIAS HERE.**

ORIGIN OF REPLICATION BIAS: To study the bias towards the position of the genes, we plotted the insertion index for each gene versus the distance of the gene from the origin of replication normalised by the length of the genome. Fig. 3 shows the results. The red line is a loess curve that has been fitted to the data when the smoothness parameter equals 0.2. The figure indicates that the insertion indices decrease when the genes are located further from the origin of replication. A possible explanation for this phenomenon is that the bacteria were under replication while being infected with transposons. Therefore, the number of gene copies close to the origin of replication was greater and more insertions have occurred in these genes. To overcome this bias, we have normalised our insertion indices by dividing the value of the insertion index by the predicted value by loess for that position and then multiplying this value by the average insertion index.

mention possible confounding factor for this study. Essential genes may (are???) non-uniformly distributed in the genome (I presume they are clustered near the origin). This would be expected to lower the insertion index near the origin, however, the number of non-essential genes is $\approx 10\times$ the number of essential genes, therefore the signature of biased numbers of insertions near the origin remains.

PREFERRED INSERTION MOTIF BIAS: We have tested whether our transposons are biased towards certain motifs. For this, we have generated a logo from 10 nucleotides flanking the 100 top most frequent insertion sites in each genome. The results are depicted in Fig. 4. The results show a slight bias towards certain combinations of bases.

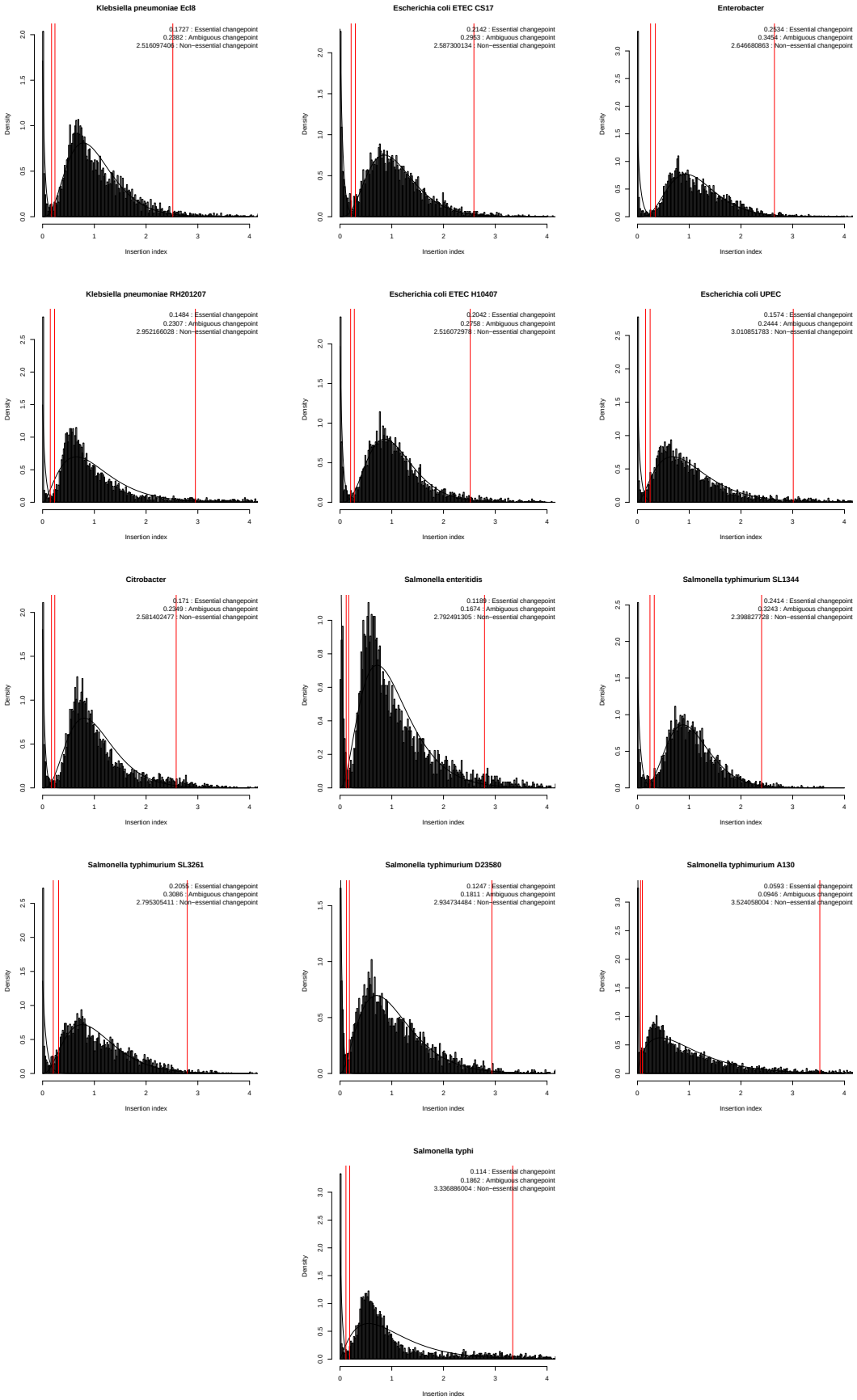

Figure 2. Plots show the insertion index distribution for each genome. The plots are divided into 4 regions using red lines. These regions from left to right are: essential, ambiguous, non-essential, and beneficial loss.

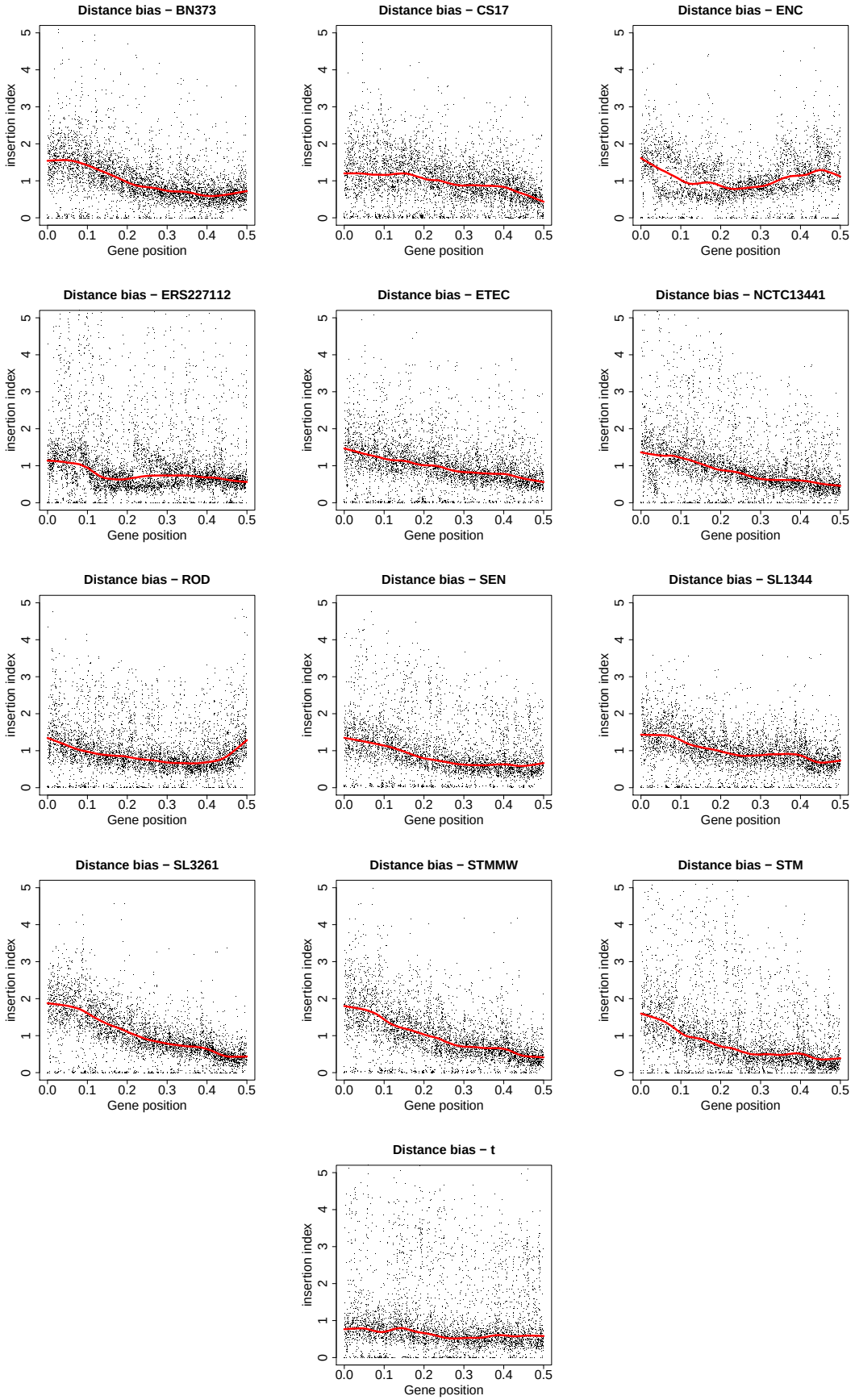

Figure 3. The plots show the distance of the genes from DnaA gene normalised by the lengths of the genomes versus the insertion indices of the genes. The distance from DnaA gene has been calculated in both directions and then the minimum value has been used for

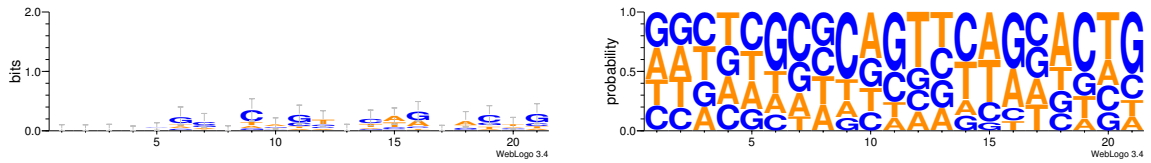

Figure 4. We have generated sequence logo plots using sequences from 10 nucleotides flanking the 100 top most frequent insertion sites from each genome. On the left the height of each character corresponds to a bit score for that character (i.e. $2 - \sum f_a \times \log_2 f_a - \frac{1}{\ln 2} \times \frac{3}{2 \times n}$, where f_a is the relative frequency of base a and n is the number of sequences). To put it in simple words, the height of the set of characters shows how biased that position is and the height of each character shows the amount of bias towards that character. On the right the height of each character shows the relative frequency of that character.

In addition, we have investigated if the G-C content of genes can change the number of insertions by plotting the number of G-C bases in a gene normalised by the length of the gene versus insertion index Fig. 5. The red lines show the loess curve when the smoothness parameter is 0.2. As the figure shows, when G-C content is less than 40%, the insertion index is low, however when it is higher than 50%, the insertion index is almost constant. A possible reason for this phenomena is the association of A-T rich sequences and histone-like nucleotide structuring (H-NS) proteins, which causes a reduction in the insertions in A-T rich regions [18]. The other reason is that the genes with low G-C content are enriched in mobile genetic elements compared to the genes with average G-C content (6) and this has caused seeing a different pattern of essentiality in that region.

- model H-NS binding sites? CGWTWHWww Lang et al (2007)
- seems unlikely – show bulk of genes are around 50% G+C (add box-whisker plots to scatter diagrams?)
- check Freed, Silander paper – the missing piece of genome, was this low G+C?

The other bias that we have considered is the bias towards certain locations in a gene. We have divided every gene into 100 bins and calculated the mean insertion index for each bin. Fig. 7 shows almost no bias towards any location. We have also studied the bias in each of

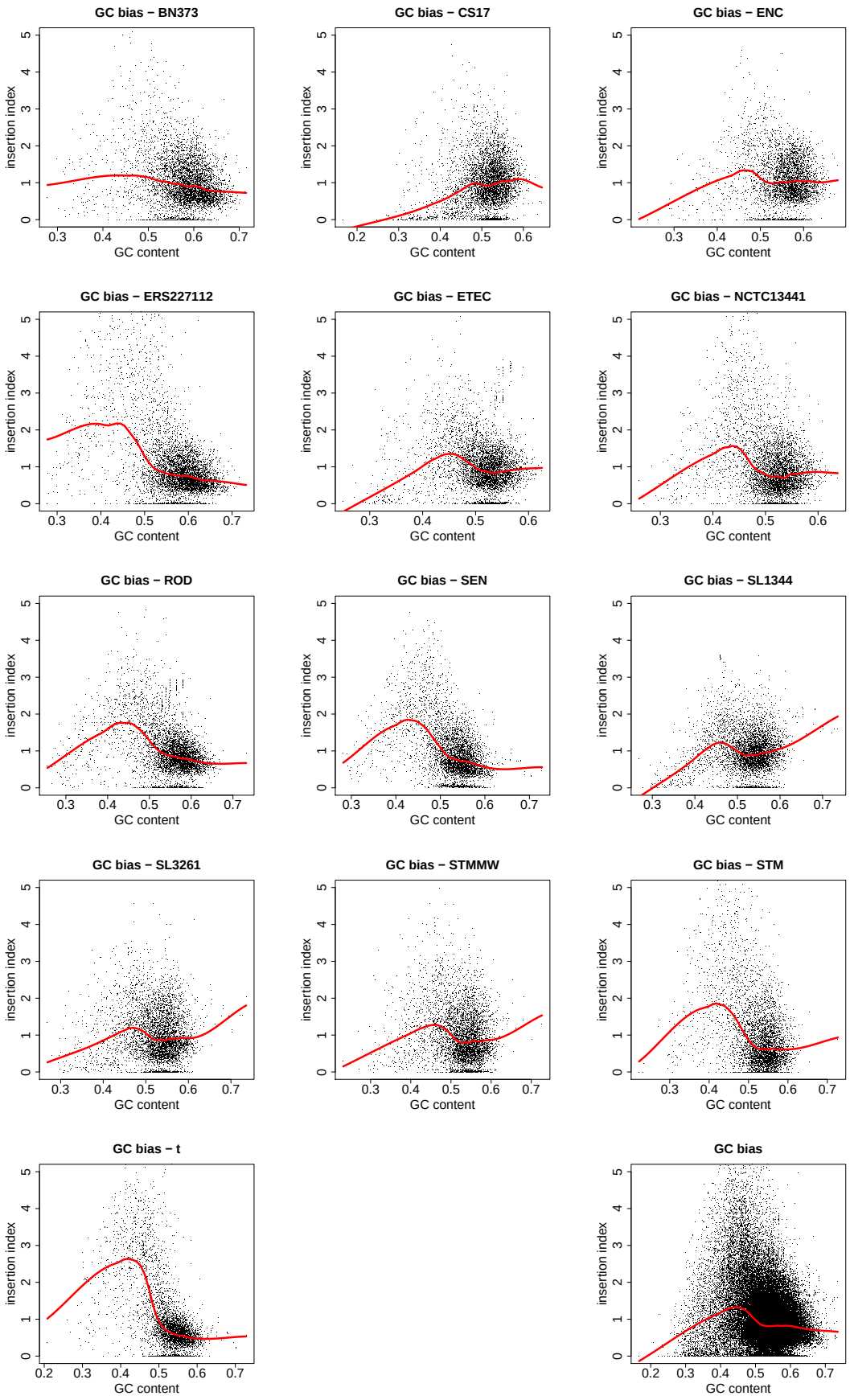

Figure 5. The plots show the ratio of G-C bases in the genes normalised by the lengths of the genes against their insertion indices. The red curves show the loess curve when the smoothness parameter is 0.2.

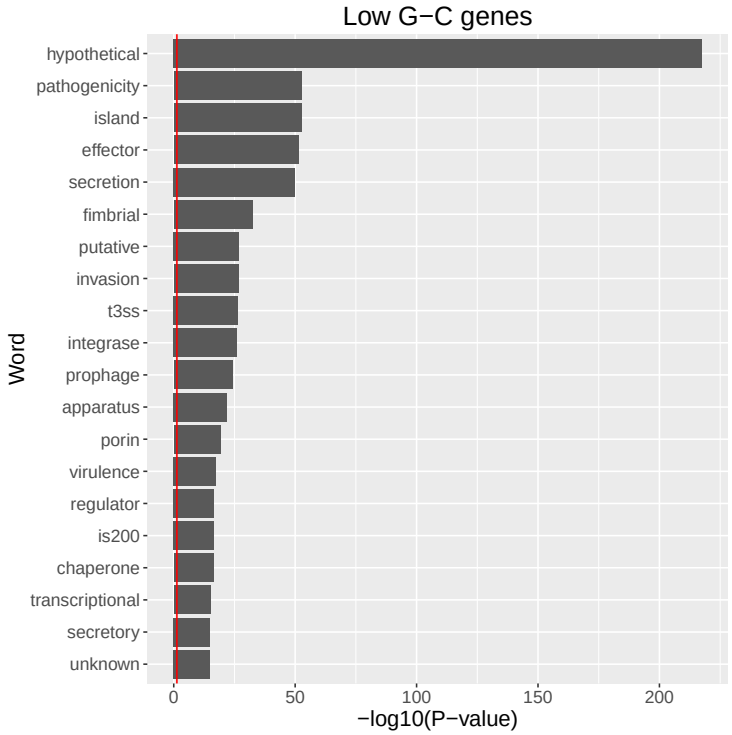

Figure 6. Word enrichment analysis for low G-C genes compared to genes with interquartile G-C level. The red line shows P-value = 0.05. The P-values have been calculated using Fisher's exact test and corrected using Benjamini-Hochberg-Yekutieli.

the groups: essential genes, non-essential genes, and beneficial losses. The results imply that the number of insertions in the internal region of the essential genes is outnumbered by the number of insertions in the 5' and 3' ends while it is the opposite in beneficial losses. The case for the non-essential genes is similar to the average (Fig. 7). High number of insertions at the 3' end of essential genes implies that the functional part of the genes are located before the insertions. On the other hand, high number of insertions at the 5' end of the essential genes indicates there might be alternative start codons in the 5' end or it might be because of alignment errors. {To be tested} We have calculated the insertion index for genes by ignoring 5% from the 5' end and 20% from the 3' end of the genes to overcome these biases. The insertion index distribution for each genome after correcting for all distance from the origin of replication bias and bias towards the position of insertion within genes is depicted in Fig. 8.

3 Essentiality and conservation

To study whether each gene in our 13 organisms is conserved we have proposed a program that clusters homologous proteins. This program uses Jackhmmer from HMMER package [12] to compare protein sequences. It first compares a set of query proteins against all given proteins and clusters homologous proteins using Jackhmmer. Then it selects all sequences that were not selected in the first step and compares them together and clusters those protein sequences. In the next step, it breaks down large clusters by using Jackhmmer with more stringent parameters within the clusters and also merges clusters which have a single member by relaxing Jackhmmer parameters. Finally, the program merges overlapping sequences in each cluster and combines similar clusters. The program is summarised in Fig. 9 and the distribution of cluster lengths after clustering the genes of 13 strains under study is plotted in Fig. 10.

3.1 Gene classes

To study the conservation of genes, we have used homclust and divided the clusters of homologous genes into three groups. Genus specific clusters contain genes that are present

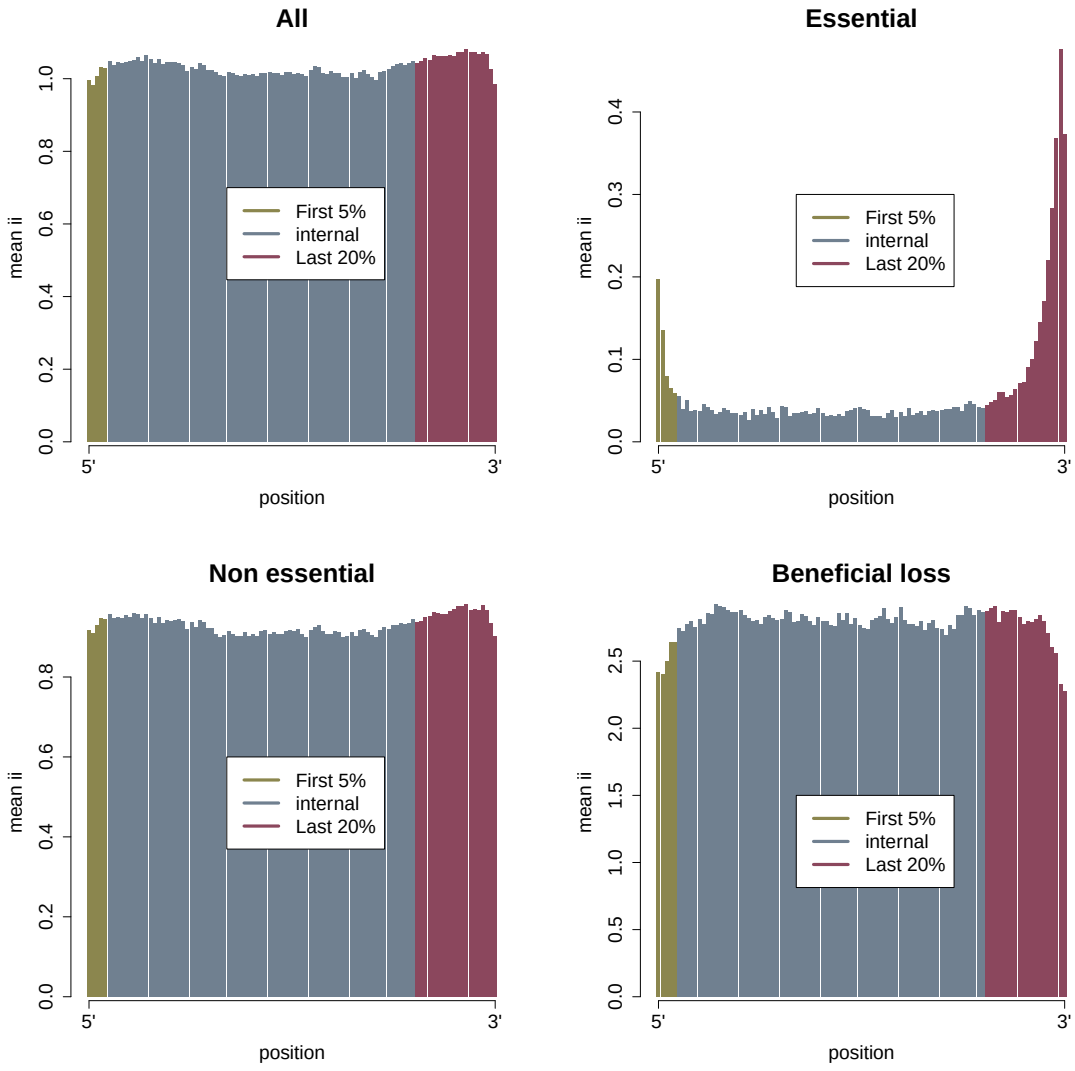

Figure 7. The plots show the average insertion index in over the genes in all genes (top left), essential genes (top right), non-essential genes (bottom left), and beneficial losses (bottom right). Each bin shows the average insertion index for 1% of the genes. The genes have been divided into 3 segments: 5% of the genes on the 5' end, 20% of the genes on the 3' end, and the rest in the middle. These are shown by khaki, slate gray, and violet red respectively.

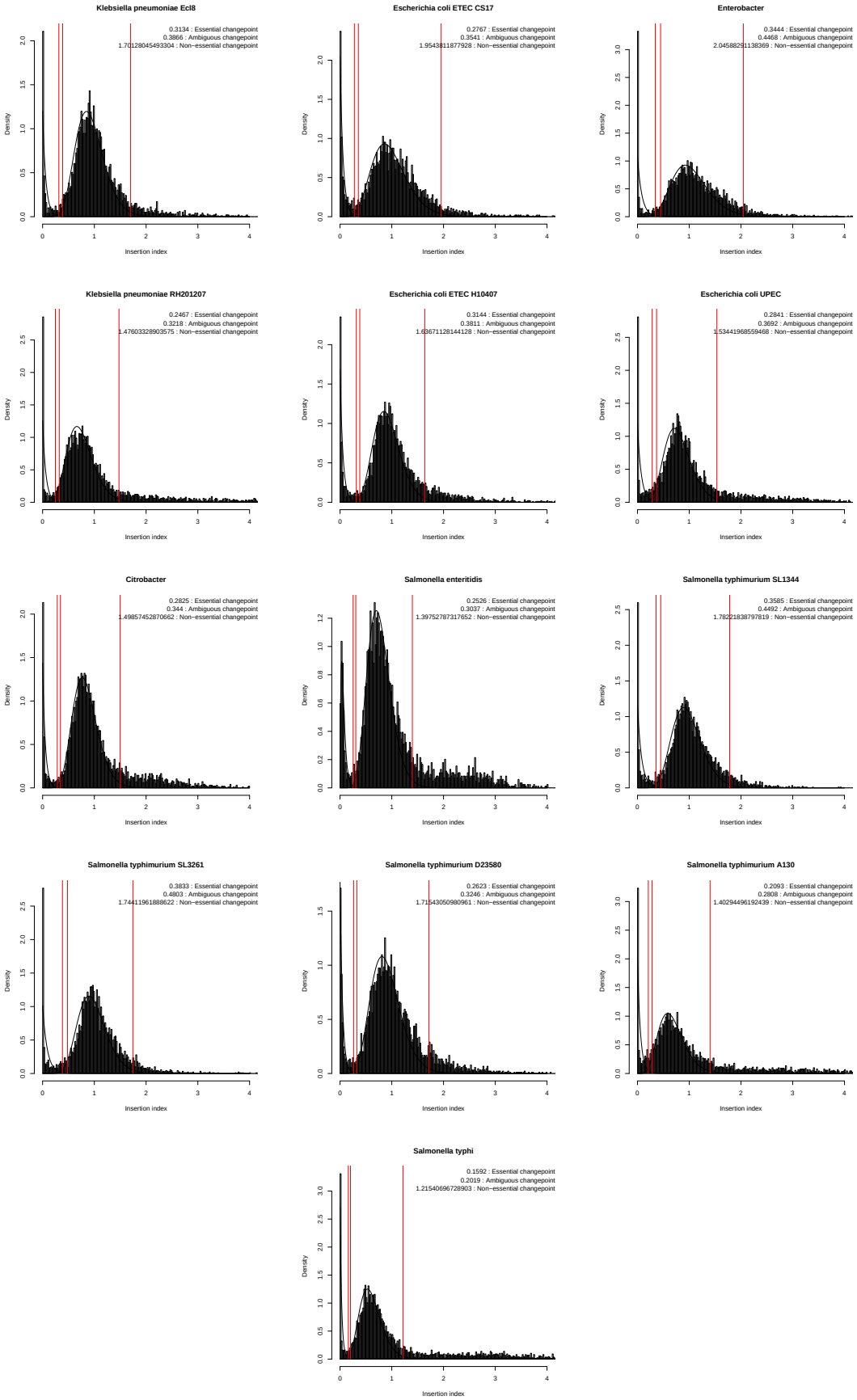

Figure 8. Plots show the insertion index distribution for each genome after correcting for distance from the origin of replication bias and bias towards the position of insertion within genes. The plots are divided into 4 regions using red lines. These regions from left to right

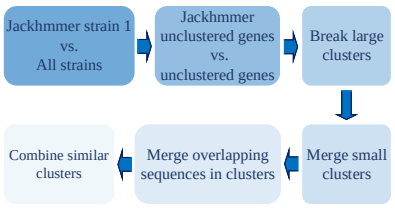

Figure 9. The steps of our proposed algorithm for clustering homologous genes.

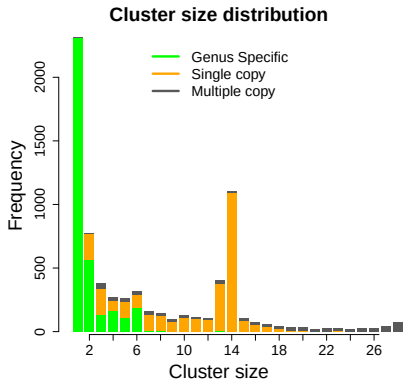

Figure 10. Size distribution for all clusters of homologous genes. Genus specific genes are genes that are present only in one genus, single copy genes are present in more than one genus and more than 70% of them are not duplicated, and multi-copy genes are present in more than one genus and less than 70% of them are not duplicated.

only in one genus, the genes in single copy clusters are present in more than one genus and more than 70% of them are not duplicated, and the genes in multi-copy clusters are present in more than one genus and less than 70% of them are not duplicated. These three groups are depicted in Fig. 10.

We have divided the clusters of orthologous genes into 4 groups based on their essentiality (essential, ambiguous, non-essential, and beneficial loss) and 3 groups based on their conservation (genus specific, single copy, and multi-copy). The results are depicted in Fig. 11. The figure shows that most of the essential clusters are single copy and most of the beneficial losses are genus specific.

To study which functions are enriched in each class of essentiality, we have gathered the"note" section for each gene from their embl files. Then, we have counted the repeat number of each word for the genes in each class and the genes that do not belong to that class and the number of all other words in these two groups and used a Fisher's exact test to

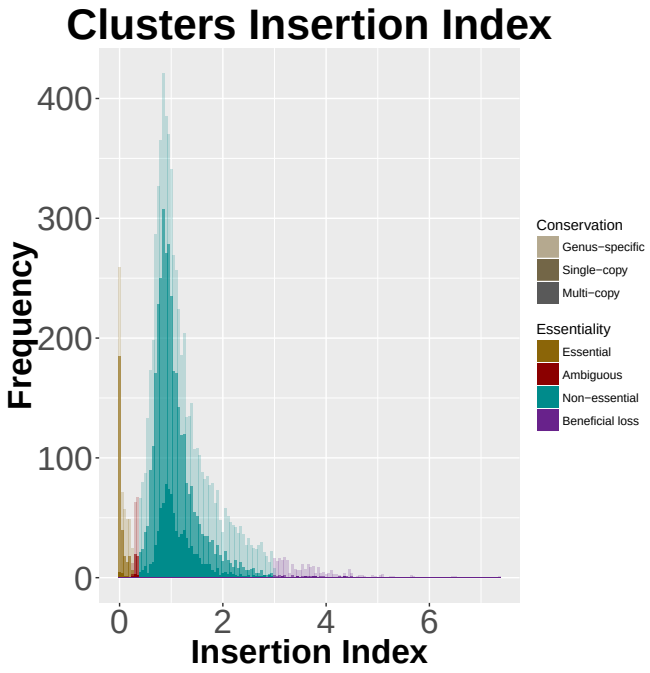

Figure 11. The genes have been clustered into orthologous groups using Hieranoid and paralogous groups using Jackhmmer and divided into 3 groups: genus specific, single copy, and multi-copy genes. Then, the essentiality of the clusters has been defined using the insertion indices of the genes in the clusters. The figure shows that most of the essential genes are in single copy group, while most of the beneficial losses are genus-specific.

calculate P-values. The P-values are then corrected using Benjamini-Hochberg-Yekutieli procedure. Fig. 12 shows the top 20 enriched words for each essentiality class. The results show an enrichment of the genes related to replication, transcription, translation, division, and rod shape determining proteins in essential class. The non-essential genes are mostly transport proteins, flagellar proteins, ATPase, and DNA repair proteins. Beneficial losses are enriched in transposase enzymes, putative and hypothetical proteins, and mobile elements. Beneficial losses also contain many fimbrial proteins which probably has occurred because these proteins are not needed in a rich lab medium {TRUE?}.

We have also conducted a pathway enrichment analysis for these three groups. For this, we have downloaded pathway datasets for strains that were available in KEGG database. This includes pathways for Cirobacter rodentium ICC168, Salmonella Enteritidis P125109, Enterobacter cloacae NCTC 9394, Salmonella Typhimurium D23580, Escherichia coli ETEC H10407, Salmonella Typhimurium SL1344, Escherichia coli K-12 MG1655, and Salmonella Typhi Ty2. Then we have merged these databases and used a hypergeometric test to find which pathways are enriched in each essentiality class. Finally, we have corrected the P-values using Benjamini-Hochberg-Yekutieli. The results are depicted in 13.

3.2 The evolution of essentiality

In order to test if the essentiality of genes follows a tree-like trend, we have compared the number of genes that are conserved in different strains in our study and the number of genes that are essential in these strains. For this, we have counted the number of genes that are core between every combination of our strains, and the number of genes that are core essential between those combinations. We have used UpSetR package [9] in R to visualise the results in Fig. 15. As shown in the figures, among 1908 genes that are core between all the strains under study, only 184 are core essential. The results propose that although conservation of genes follows a tree-like trend (800 genes are core in Klebsiellas, 273 genes are core in Salmonellas, 208 genes are core in Klebsiellas and Enterobacter, ...), the essentiality does not show a tree-like signal and most of the large sets of core essential genes belong to only one strain. We believe this happens due to the small number of essential genes.

To test if essential genes are more likely to be conserved, we looked at different levels in

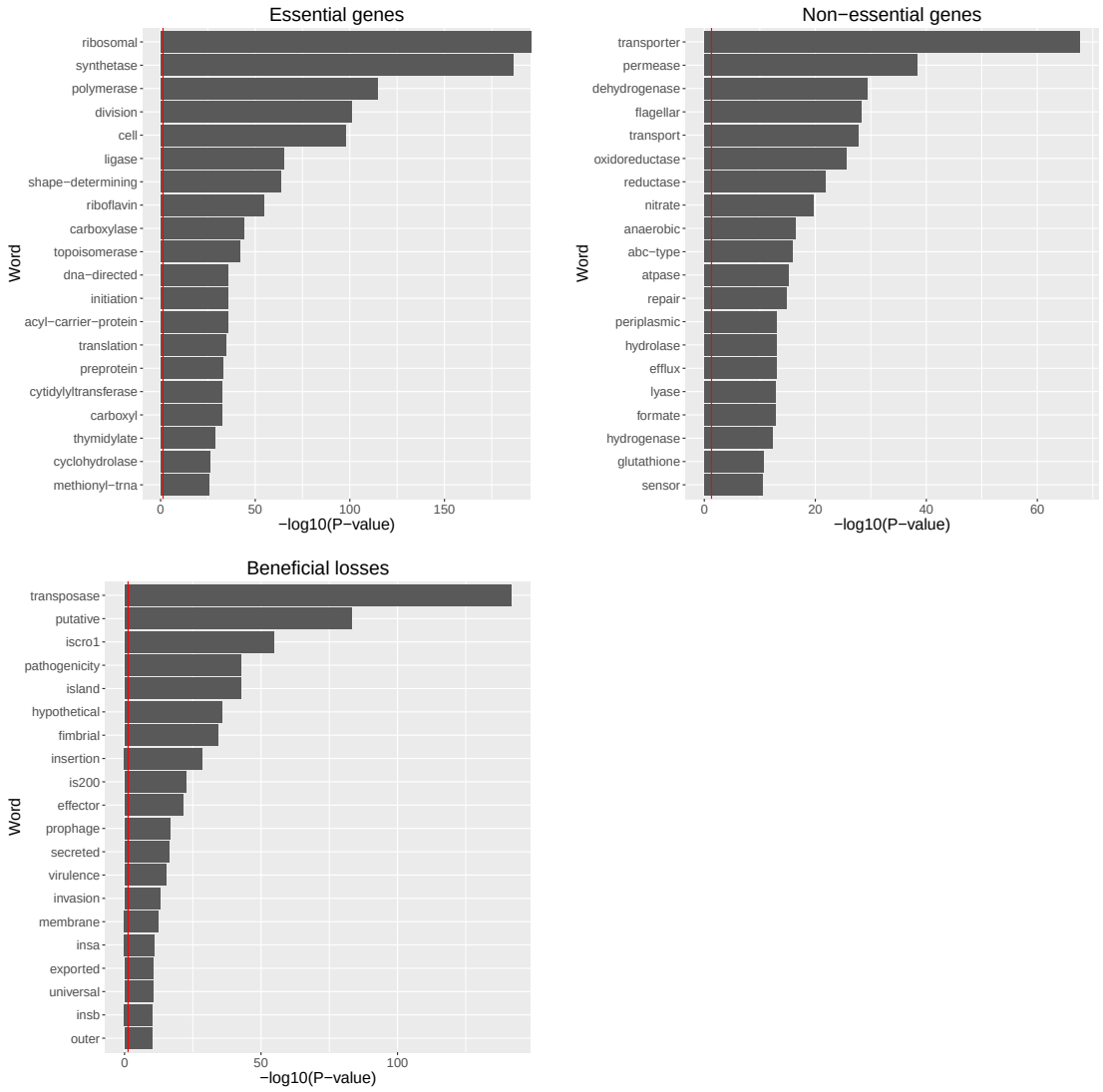

Figure 12. Word enrichment analysis for essential genes, non-essential genes, and beneficial losses compared to other genes. The red line shows P-value = 0.05. The P-values have been calculated using Fisher's exact test and then corrected using Benjamini-Hochberg-Yekutieli procedure.

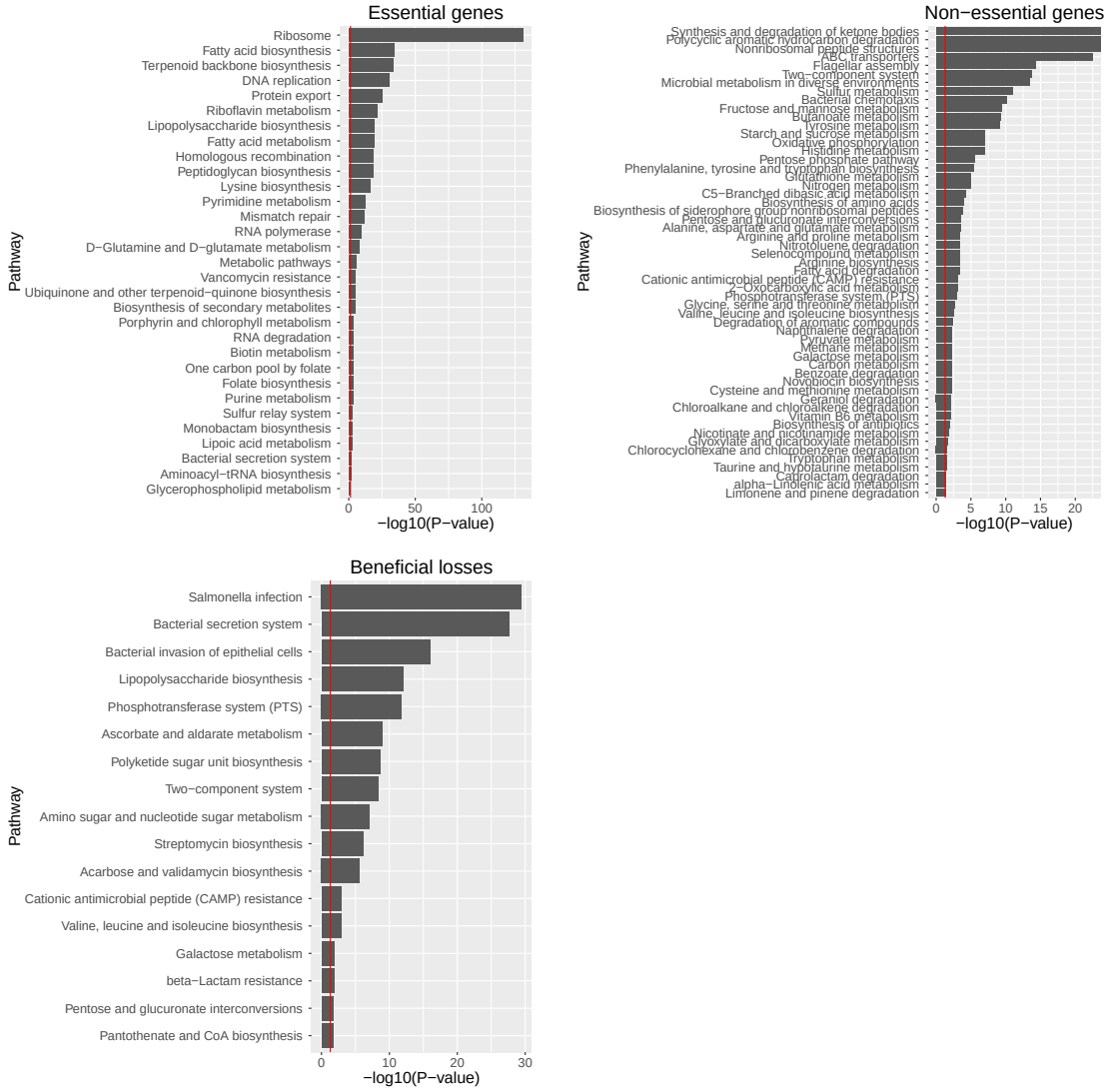

Figure 13. Pathway enrichment analysis for beneficial losses, essential genes, and non-essential genes compared to other genes. The red line shows P-value = 0.05. The P-values have been calculated using hypergeometric test and then corrected using Benjamini-Hochberg-Yekutieli procedure.

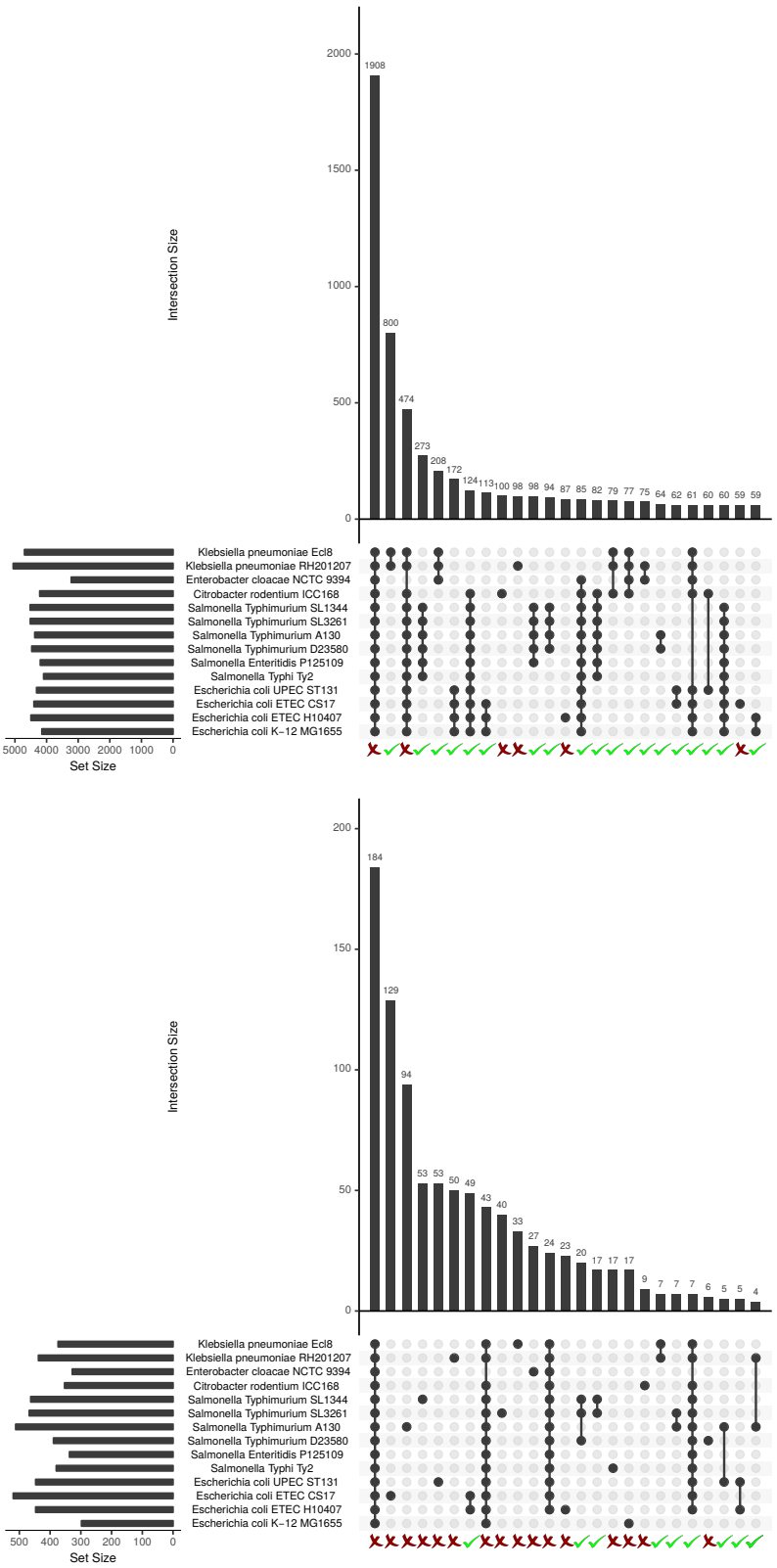

Figure 14. The first figure shows the number of core genes between each group of species and the second figure shows the number of core essential genes. The bars show the number of genes that are core between the strains marked with black circles. The tick marks show phylogenetically informative columns and the cross marks show non informative columns.

the species tree and calculated the ratio of the number of core essential genes to the number of core genes in each level. For the essential genes to be more likely to be conserved, this ratio should increase as we go up in the tree. We have used three different methods for this. The first method is intersecting over core genes and core essential genes, so, genes are core in a node if and only if they are present in all the strains in its child nodes and are core essential if and only if they are essential in all the strains in its child nodes.

The second method which is called ancestral insertion index uses intersection for core genes but a different definition for core essential genes. In this method, we have averaged over the insertion indices of the pair of closest children of the node that we are calculating the core genes for. We have repeated this and averaged the averages until we reached the node under study. Then, we have plotted the insertion indices and fitted an exponential and a gamma distribution to the plot and found the essential genes at that level.

The third method is using Dollo law to define core genes and core essential genes. This method, assumes that the gain of genes (essentiality) is highly improbable, so it tries to have up to one occurrence of gain of genes (essentiality) and minimise the number of times that the genes (the essentiality) has been lost. Using this method, we can predict which genes were present in the common ancestor of our strains and which genes were essential in it. The numbers predicted using these three methods are shown in Fig. 16. As this figure shows, the ratio between core essential and core genes is almost constant using the intersection and dollo method; however, this ratio increases as we go higher in the tree using the ancestral insertion index method.

References

1. T. Baba, T. Ara, M. Hasegawa, Y. Takai, Y. Okumura, M. Baba, K. A. Datsenko, M. Tomita, B. L. Wanner, and H. Mori. Construction of escherichia coli k-12 in-frame, single-gene knockout mutants: the keio collection. 2:2006.0008.

2. L. Barquist, C. J. Boinett, and A. K. Cain. Approaches to querying bacterial genomes with transposon-insertion sequencing. 10(7):1161–1169.

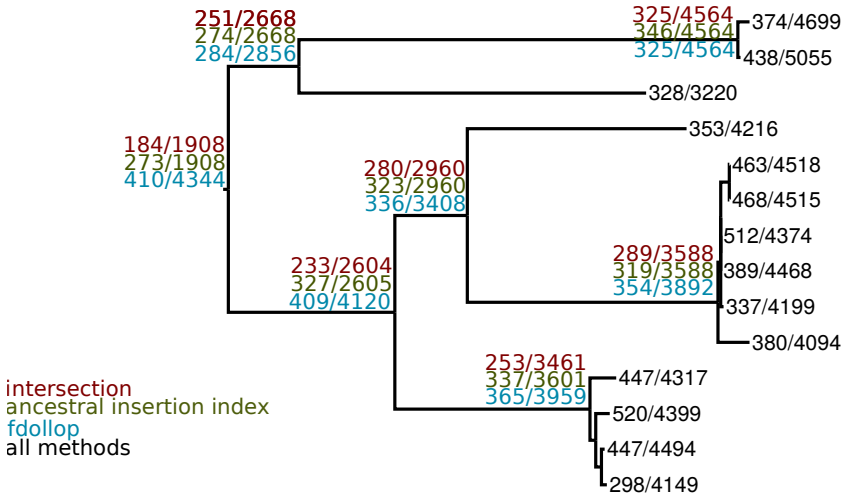

Figure 15. The tree shows the species tree in Fig.1 annotated by the number of core essential genes and core genes at each node. We have used three methods for defining core essential genes and core genes. The numbers at the leaves are the same using all these three methods. At the internal nodes, red shows the numbers using the intersection method, green shows the ancestral insertion index method, and turquoise shows fdollop method.

3. L. Barquist, G. C. Langridge, D. J. Turner, M.-D. Phan, A. K. Turner, A. Bateman, J. Parkhill, J. Wain, and P. P. Gardner. A comparison of dense transposon insertion libraries in the salmonella serovars typhi and typhimurium. page gkt148.

4. L. Barquist, M. Mayho, C. Cummins, A. K. Cain, C. J. Boinett, A. J. Page, G. C. Langridge, M. A. Quail, J. A. Keane, and J. Parkhill. The TraDIS toolkit: sequencing and analysis for dense transposon mutant libraries. 32(7):1109–1111.

5. D. J. Brenner and N. R. Krieg. *Bergey's Manual® of Systematic Bacteriology: Volume Two: The Proteobacteria*. Springer Science & Business Media.

6. R. Canals, X.-Q. Xia, C. Fronick, S. W. Clifton, B. M. Ahmer, H. L. Andrews-Polymenis, S. Porwollik, and M. McClelland. High-throughput comparison of gene fitness among related bacteria. 13:212.

7. B. Christen, E. Abeliuk, J. M. Collier, V. S. Kalogeraki, B. Passarelli, J. A. Coller, M. J. Fero, H. H. McAdams, and L. Shapiro. The essential genome of a bacterium. 7:528.

8. A. E. Clatworthy, E. Pierson, and D. T. Hung. Targeting virulence: a new paradigm for antimicrobial therapy. 3(9):541–548.

9. J. Conway and N. Gehlenborg. UpSetR: A more scalable alternative to venn and euler diagrams for visualizing intersecting sets.

10. P. D. Curtis and Y. V. Brun. Identification of essential alphaproteobacterial genes reveals operational variability in conserved developmental and cell cycle systems. 93(4):713–735.

11. A. E. Darling, G. Jospin, E. Lowe, F. A. Matsen, H. M. Bik, and J. A. Eisen. PhyloSift: phylogenetic analysis of genomes and metagenomes. 2:e243.

12. S. R. Eddy. Accelerated profile HMM searches. 7(10):e1002195.

13. J. D. Gawronski, S. M. S. Wong, G. Giannoukos, D. V. Ward, and B. J. Akerley. Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for haemophilus genes required in the lung. 106(38):16422–16427.

14. A. L. Goodman, M. Wu, and J. I. Gordon. Identifying microbial fitness determinants by insertion sequencing using genome-wide transposon mutant libraries. 6(12):1969–1980.

15. C. A. Hutchison, R.-Y. Chuang, V. N. Noskov, N. Assad-Garcia, T. J. Deerinck, M. H. Ellisman, J. Gill, K. Kannan, B. J. Karas, L. Ma, J. F. Pelletier, Z.-Q. Qi, R. A. Richter, E. A. Strychalski, L. Sun, Y. Suzuki, B. Tsvetanova, K. S. Wise, H. O. Smith, J. I. Glass, C. Merryman, D. G. Gibson, and J. C. Venter. Design and synthesis of a minimal bacterial genome. 351(6280):aad6253.

16. C. A. Hutchison, S. N. Peterson, S. R. Gill, R. T. Cline, O. White, C. M. Fraser, H. O. Smith, and J. C. Venter. Global transposon mutagenesis and a minimal mycoplasma genome. 286(5447):2165–2169.

17. M. Juhas, L. Eberl, and J. I. Glass. Essence of life: essential genes of minimal genomes. 21(10):562–568.

18. S. Kimura, T. P. Hubbard, B. M. Davis, and M. K. Waldor. The nucleoid binding protein h-NS biases genome-wide transposon insertion landscapes. 7(4):e01351–16.

19. G. C. Langridge, M.-D. Phan, D. J. Turner, T. T. Perkins, L. Parts, J. Haase, I. Charles, D. J. Maskell, S. E. Peters, G. Dougan, J. Wain, J. Parkhill, and A. K. Turner. Simultaneous assay of every salmonella typhi gene using one million transposon mutants. 19(12):2308–2316.

20. H. Luo, Y. Lin, F. Gao, C.-T. Zhang, and R. Zhang. DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. 42:D574–D580.

21. J. Peters, A. Colavin, H. Shi, T. Czarny, M. Larson, S. Wong, J. Hawkins, C. S. Lu, B.-M. Koo, E. Marta, A. Shiver, E. Whitehead, J. Weissman, E. Brown, L. Qi, K. Huang, and C. Gross. A comprehensive, CRISPR-based functional analysis of essential genes in bacteria. 165(6):1493–1506.

22. B. E. Rubin, K. M. Wetmore, M. N. Price, S. Diamond, R. K. Shultzaberger, L. C. Lowe, G. Curtin, A. P. Arkin, A. Deutschbauer, and S. S. Golden. The essential gene set of a photosynthetic organism. 112(48):E6634–E6643.

23. A. Stamatakis. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. 30(9):1312–1313.

24. T. van Opijnen, K. L. Bodi, and A. Camilli. Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. 6(10):767–772.

25. K. M. Wetmore, M. N. Price, R. J. Waters, J. S. Lamson, J. He, C. A. Hoover, M. J. Blow, J. Bristow, G. Butland, A. P. Arkin, and A. Deutschbauer. Rapid quantification of mutant fitness in diverse bacteria by sequencing randomly bar-coded transposons. 6(3):e00306–15.

26. H. H. Xu, J. D. Trawick, R. J. Haselbeck, R. A. Forsyth, R. T. Yamamoto, R. Archer, J. Patterson, M. Allen, J. M. Froelich, I. Taylor, D. Nakaji, R. Maile, G. C. Kedar,

M. Pilcher, V. Brown-Driver, M. McCarthy, A. Files, D. Robbins, P. King, S. Sillaots, C. Malone, C. S. Zamudio, T. Roemer, L. Wang, P. J. Youngman, and D. Wall. Staphylococcus aureus TargetArray: comprehensive differential essential gene expression as a mechanistic tool to profile antibacterials. 54(9):3659–3670.