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 [27]. Another method that is widely used due to its simplicity and accuracy is transposon mutagenesis along with high-throughput sequencing [7, 13, 14, 19, 23, 25, 26]. 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*. 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

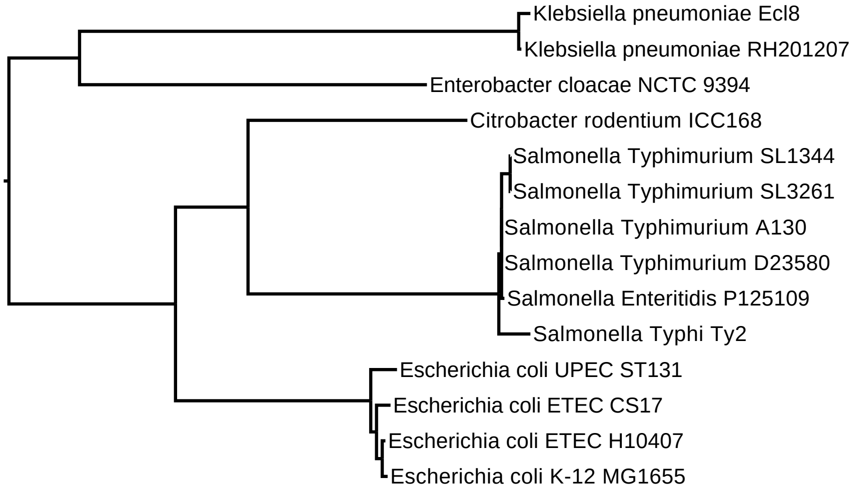
Klebsiella pneumoniae Ecl8
Klebsiella pneumoniae RH201207
Enterobacter cloacae NCTC 9394
Citrobacter rodentium ICC168
Salmonella Typhimurium SL1344
Salmonella Typhimurium SL3261
Salmonella Typhimurium A130
Salmonella Typhimurium D23580
Salmonella Enteritidis P125109
Salmonella Typhi Ty2
Escherichia coli UPEC ST131
Escherichia coli ETEC CS17
Escherichia coli ETEC H10407
Escherichia coli K-12 MG1655

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 [24] on Phylosift [11] amino acid markers.

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 Results

1.1 Are there biases in transposon mutagenesis data?

To evaluate the essentiality of a gene, the number of insertions within that gene was measured. However, if the transposons are biased to specific regions in the genome, it results in false predictions and influences the accuracy of our analysis. Different articles have reported biases in transposon mutagenesis [3, 18, 23]. We performed a detailed study of these biases. The biases that we studied include: origin of replication bias, preferred insertion motif bias, and positional bias within genes.

1.1.1 Origin of replication bias

One possible source of bias is the distance from origin of replication. When the bacteria are under replication during the transposn insertion process, there are more copies of the genes close to the origin of replication than the genes further away. This results in more insertions in the genes near the origin of replication which can influence the accuracy of our predictions. The other factor that we considered was if essential genes are not uniformly distributed in the genome and are clustered near the origin. Rocha and Eduardo [22] have shown that unlike highly expressed genes, essential genes are not enriched near the origin of replication. However, the essential genes are more frequent in the leading strand than the lagging one.

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. 2 shows the results. The figure indicates that the insertion indices decrease when the genes are located further from the origin of replication. To overcome this bias, we normalised our insertion indices by dividing the value of the insertion index by the predicted value by loess for that position and then multiplied this value by the average insertion index.

1.1.2 Preferred insertion motif bias

Another concern is that transposons are biased to certain compositions of nucleotides and high number of insertions in genes reflects the enrichment of the motifs that transposons are inclined to, rather than their essentiality level. For this, we generated a logo from 10 nucleotides flanking the 100 top most frequent insertion sites in each genome. The results in Fig. 3 show a slight bias towards certain combinations of bases.

In addition, we 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. 4. 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 reduces the insertions in A-T rich regions [18]. The other reason is that the genes with low G-C content are enriched in mobile

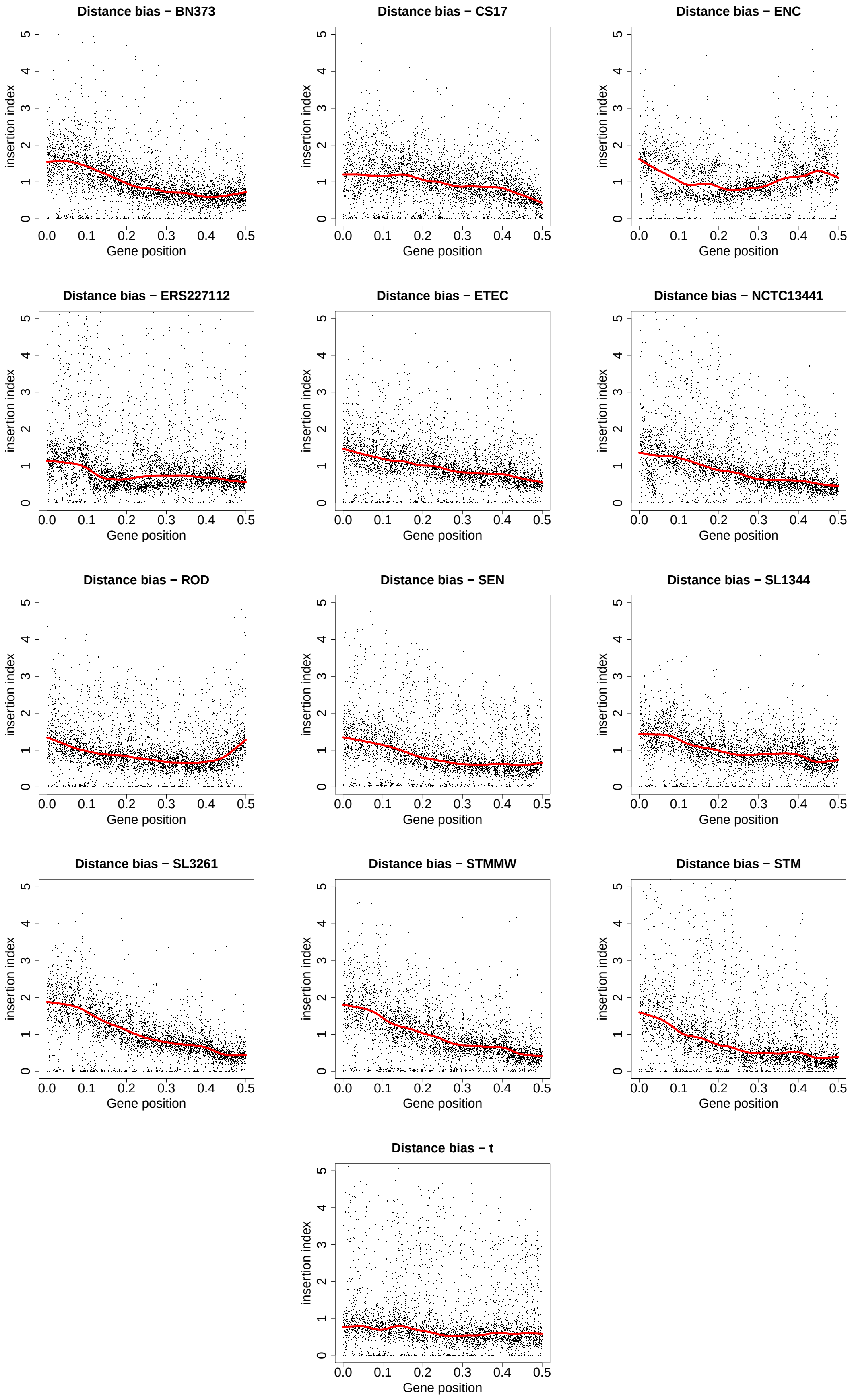

Figure 2. The plots show the distance of the genes from DnaA gene normalised by the lengths of the genomes versus the insertion indices of the genes. The distance from DnaA gene has been calculated in both directions and then the minimum value has been used for

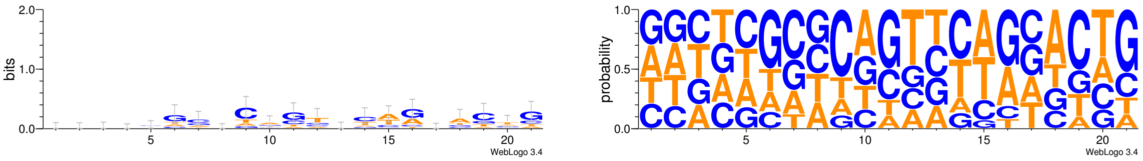

Figure 3. Sequence logo plots generated using sequences from 10 nucleotides flanking the 100 top most frequent insertion sites from each genome. On the left the height of each character corresponds to a bit score for that character (i.e. $2 - \sum f_a \times \log_2 f_a - \frac{1}{\ln 2} \times \frac{3}{2 \times n}$, 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.

genetic elements compared to the genes with average G-C content (5) 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? It is not mentioned in the paper.

The other question that we tried to answer was whether insertions are tolerated in some regions in a gene? For example, can essential genes tolerate insertions at their 3' end without losing their funtionality? To address this question, we divided every gene into 100 bins and calculated the mean insertion index for each bin. Fig. 6 shows almost no bias towards any location. We also studied the bias in each of 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. 6). 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

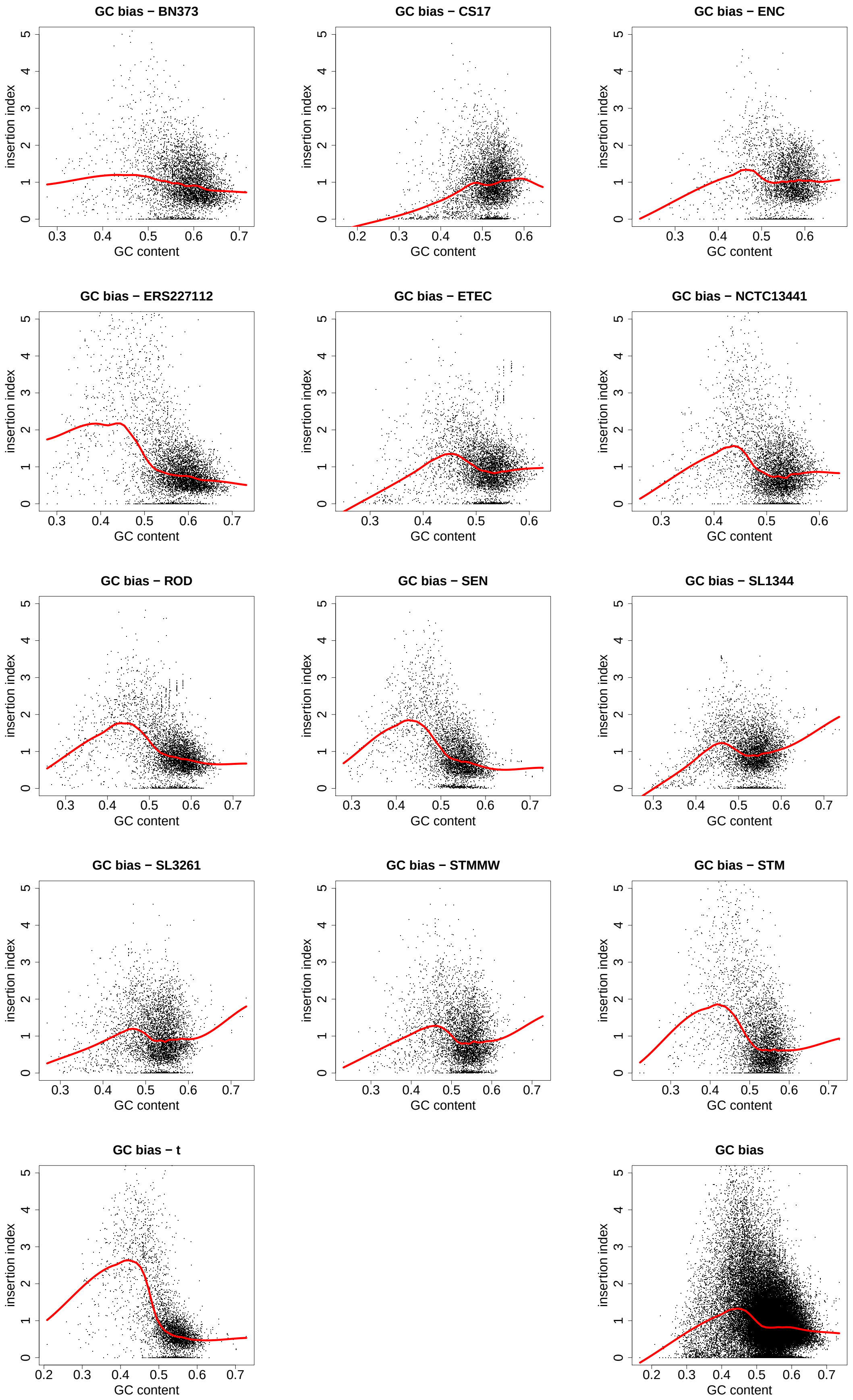

Figure 4. The plots show the ratio of G-C bases in the genes normalised by the lengths of the genes against their insertion indices. The red curves show the loess curve when the smoothness parameter is 0.2.

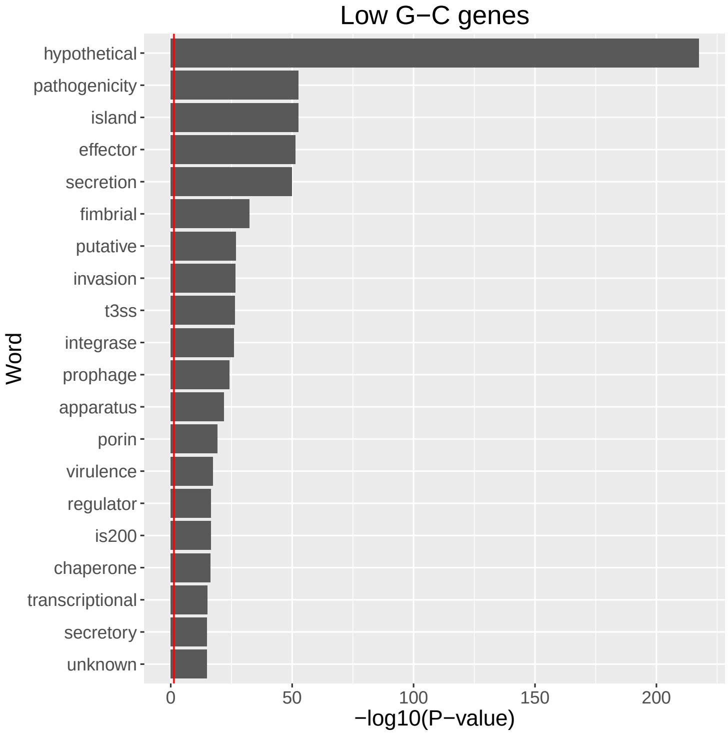

Figure 5. Word enrichment analysis for low G-C genes compared to genes with interquartile G-C level. The red line shows P-value = 0.05. The P-values have been calculated using Fisher's exact test and corrected using Benjamini-Hochberg-Yekutieli.

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 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 distance from the origin of replication bias and bias towards the position of insertion within genes is depicted in Fig. 7.

1.2 Essentiality and conservation

Essential genes are needed for the growth of organisms. Because of that, one might think that essential genes should not be lost in a short period of time throughout evolution, unless they are no longer needed in new organisms or they are replaced by new pathways. Therefore, it is expected that most of the essential genes are conserved in different organisms from the same family. We have tested this idea by comparing the essentiality and conservation of genes in Enterobacteriaceae family.

To study whether each gene in our 13 organisms is conserved we 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 running Jackhmmer with more permissive parameters. Finally, the program merges overlapping sequences in each cluster and combines similar clusters. The program is summarised in Fig. 8 and the distribution of cluster lengths after clustering the genes of 13 strains under study is plotted in Fig. 9.

1.2.1 Gene classes

We needed to evaluate essentiality and conservation of genes to study the relationship between this two. So, we divided our genes into different levels of essentiality using the

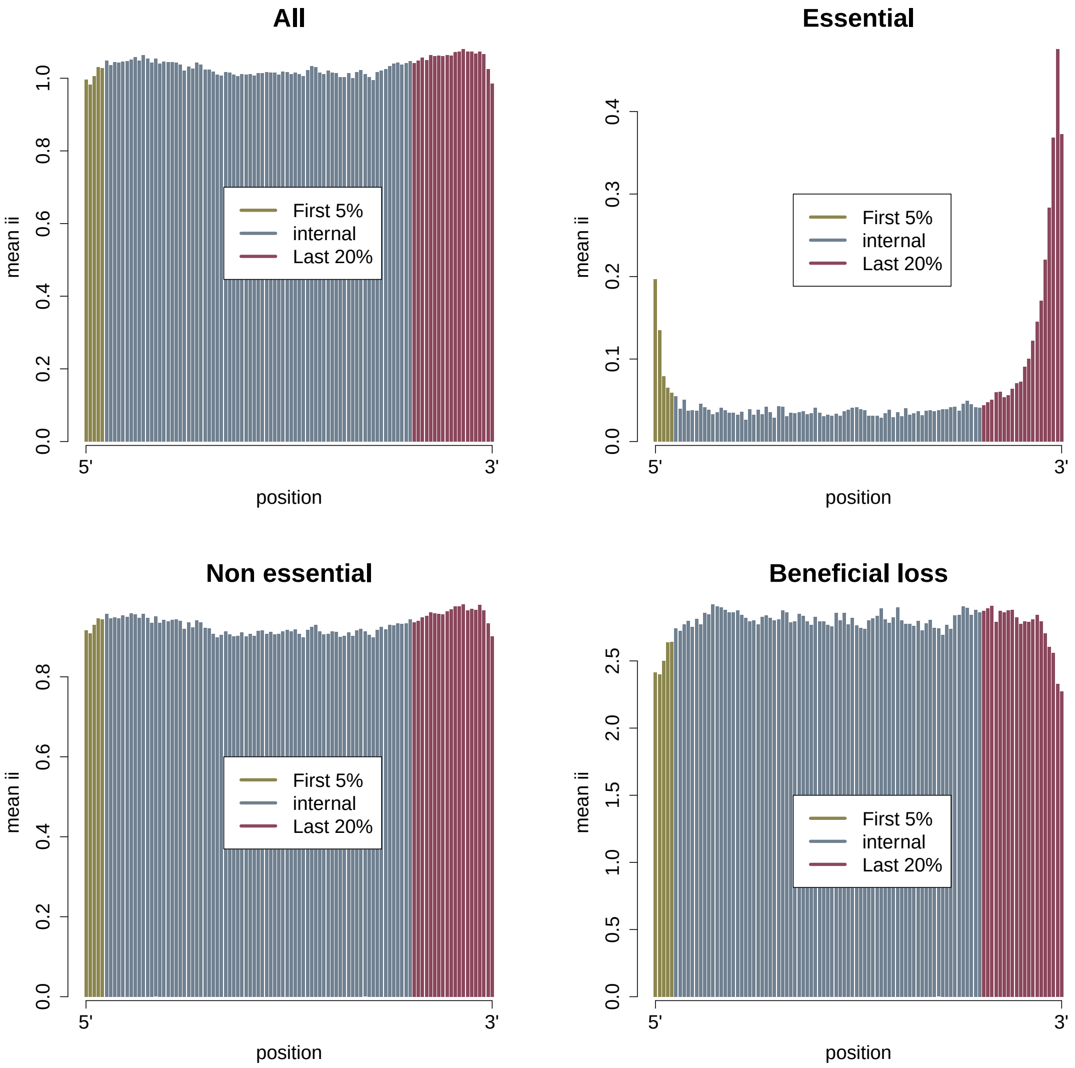

Figure 6. The plots show the average insertion index in 100 percentiles in all genes (top left), essential genes (top right), non-essential genes (bottom left), and beneficial losses (bottom right). Each bin shows the average insertion index for 1% of the genes. The genes have been divided into 3 segments: 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.

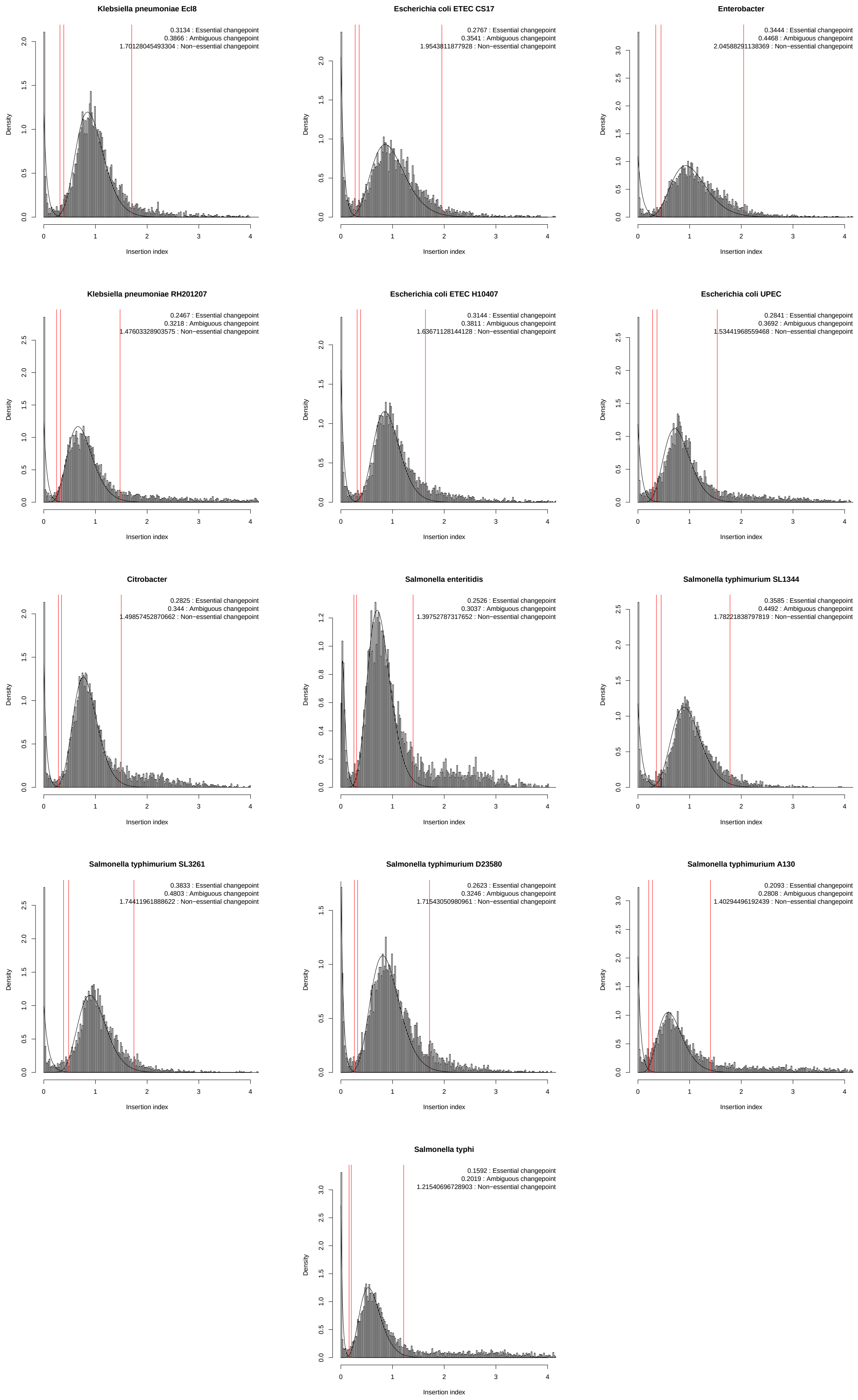

Figure 7. Plots show the insertion index distribution for each genome after correcting for distance from the origin of replication bias and bias towards the position of insertion within genes. The plots are divided into 4 regions using red lines. These regions from left to right

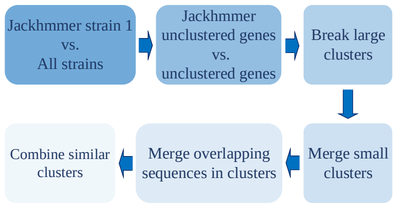

Figure 8. The steps of our proposed algorithm for clustering homologous genes.

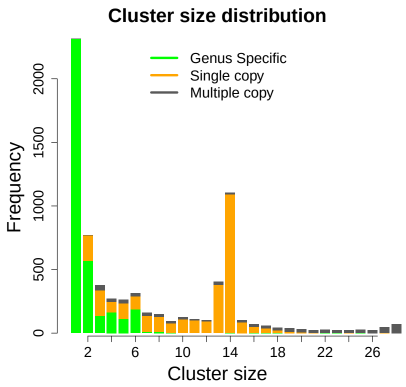

Figure 9. Size distribution for all clusters of homologous genes. Genus specific genes are genes that are present only in one genus, single copy genes are present in more than one genus and more than 70% of them are not duplicated, and multi-copy genes are present in more than one genus and less than 70% of them are not duplicated.

method explained in Section 2.2. Moreover, we used our proposed algorithm to define different levels of conservation.

We divided the clusters of homologous genes into three groups based on their conservation. Genus specific clusters contain genes that are present 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. 9. The results for comparing these three groups and the four levels of essentiality (essential, ambiguous, non-essential, and beneficial losses) are depicted in Fig. 10. The figure shows that most of the essential clusters are single copy and most of the beneficial losses are genus specific. Besides, most of the multi-copy genes are non-essential.

To study which functions are enriched in each class of essentiality, we gathered the“note” section for each gene from their embl files. Then, we counted the repeat number of each

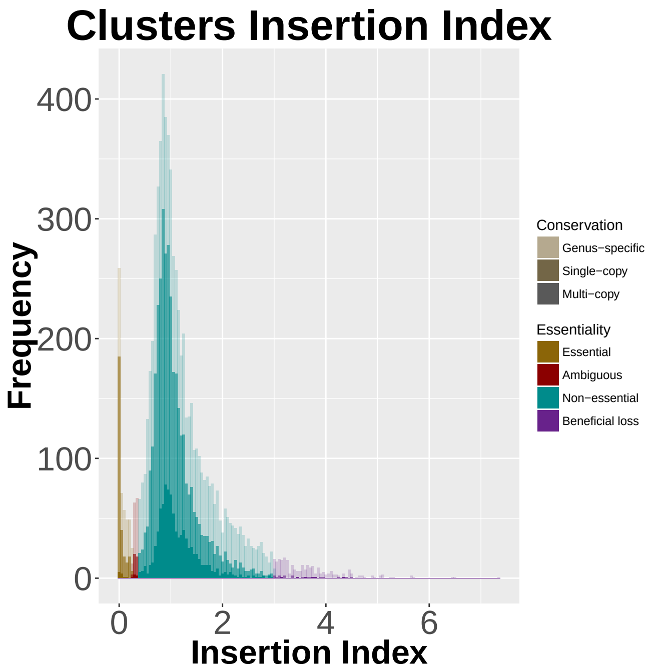

Figure 10. The genes have been clustered into orthologous groups using Hieranoid and paralogous groups using Jackhmmer and divided into 3 groups: genus specific, single copy, and multi-copy genes. Then, the essentiality of the clusters has been defined using the insertion indices of the genes in the clusters. The figure shows that most of the essential genes are in single copy group, while most of the beneficial losses are genus-specific.

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 calculate P-values. The P-values are then corrected using Benjamini-Hochberg-Yekutieli procedure. Fig. 11 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 membrane associated 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 also conducted a pathway enrichment analysis for these three groups. For this, we 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 merged these databases and used the hypergeometric test to find which pathways were enriched in each essentiality class. Finally, we corrected the P-values using Benjamini-Hochberg-Yekutieli. The results (Fig. 12) suggest similar results to the enrichment analysis that we had done on the description of the genes. However, as mobile genetic elements are not stored in KEGG database, the pathway enrichment analysis does not show the enrichment of mobile genetic elements in beneficial-losses.

1.2.2 The evolution of essentiality

If essential genes are more likely to be conserved than non-essential genes, we expect to see species that are closer together have more essential genes in common than other species. Moreover, we expect the ratio of core essential genes to core genes to increase as we go up in the phylogenetic tree. We have put these two ideas to test in this chapter.

In order to test if the essentiality of genes follows a tree-like trend, we compared the number of genes that were conserved in different bacteria in our study and the number of genes that were essential in these bacteria. For this, we counted the number of genes that

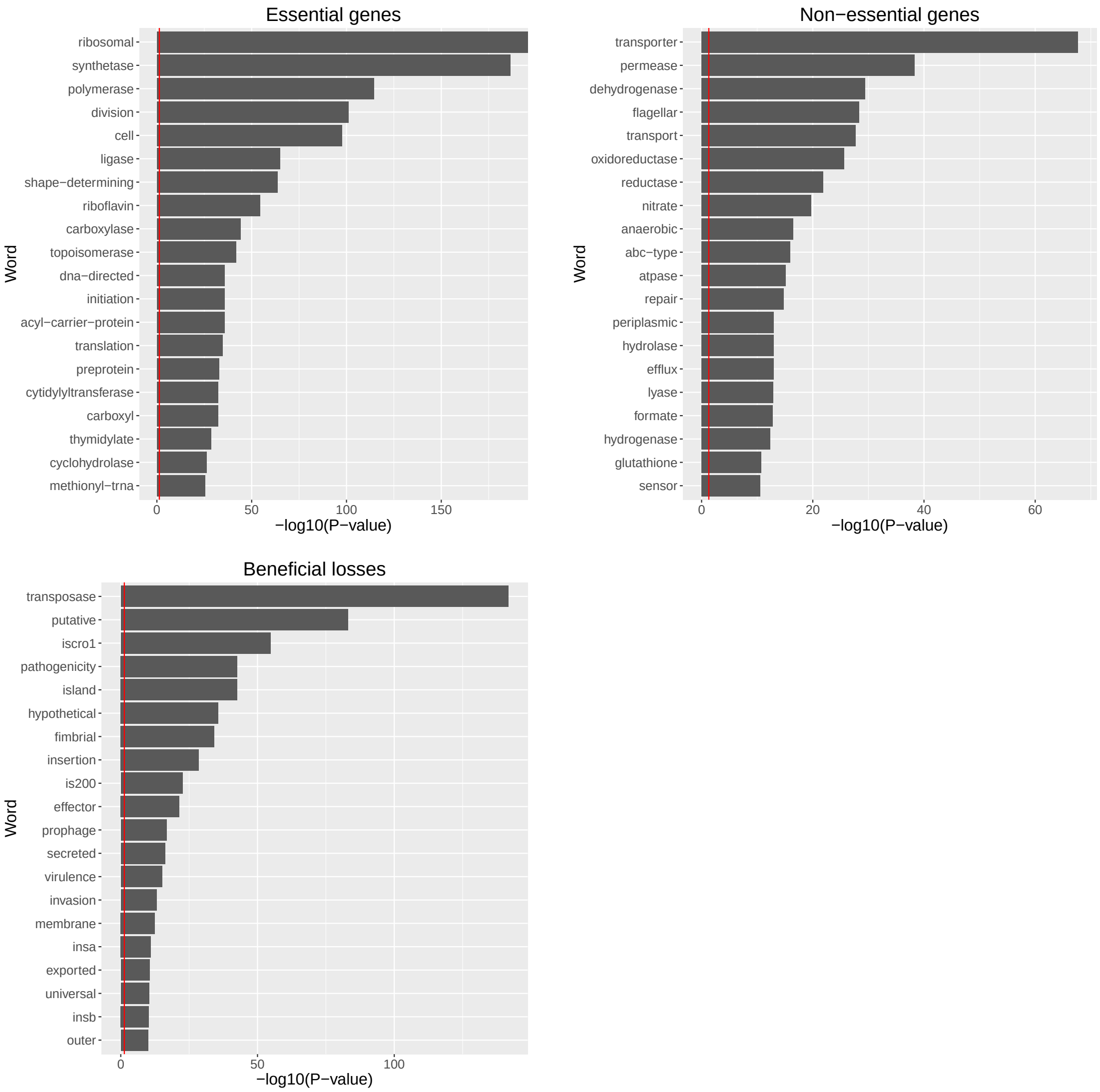

Figure 11. Word enrichment analysis for essential genes, non-essential genes, and beneficial losses compared to other genes. The red line shows P-value = 0.05. The P-values have been calculated using Fisher's exact test and then corrected using Benjamini-Hochberg-Yekutieli procedure.

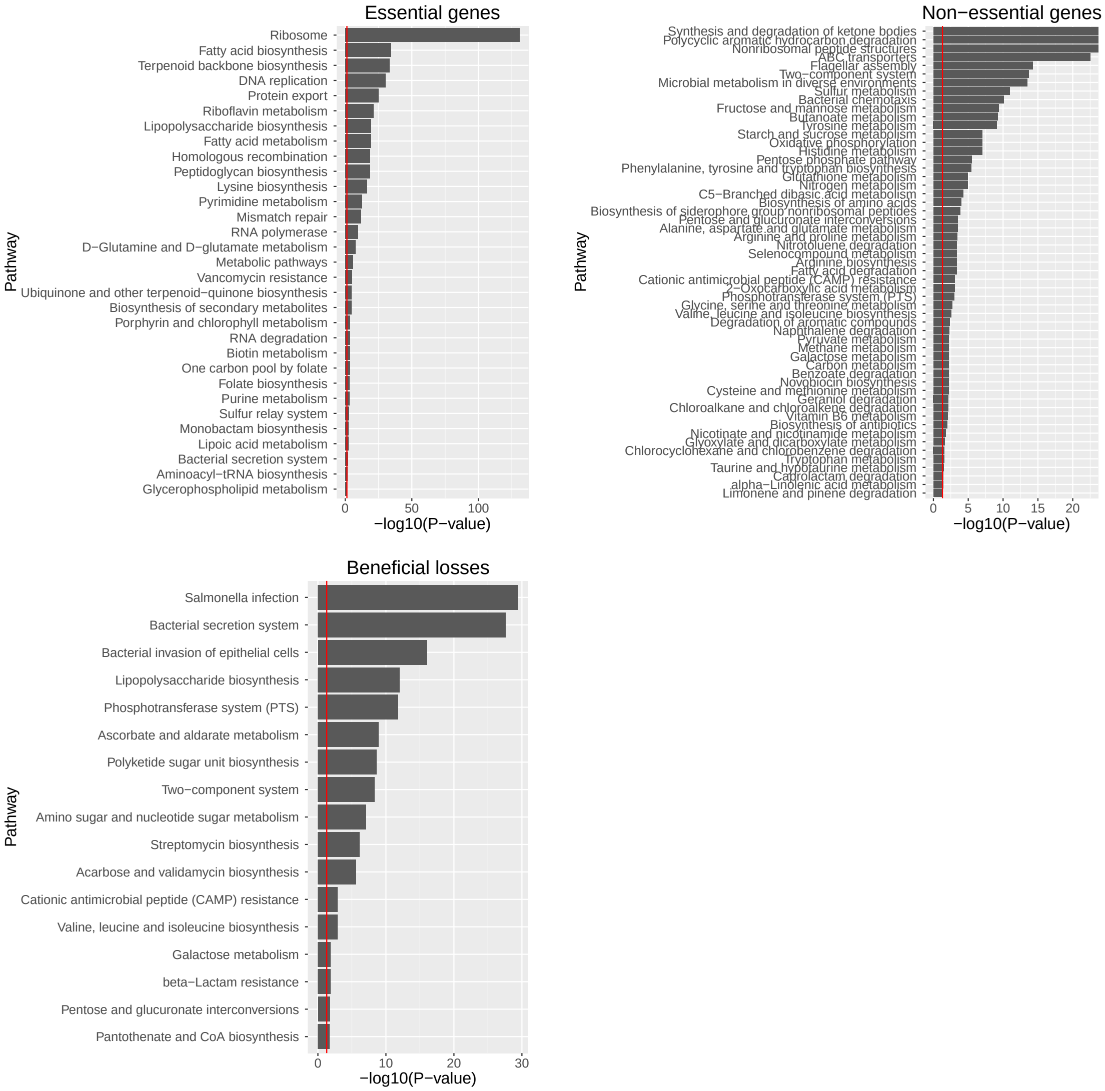

Figure 12. Pathway enrichment analysis for beneficial losses, essential genes, and non-essential genes compared to other genes. The red line shows P-value = 0.05. The P-values are calculated using hypergeometric test and then corrected using Benjamini-Hochberg-Yekutieli procedure.

were core between every combination of our bacteria, and the number of genes that were core essential between those combinations. We used UpSetR package [9] in R to visualise the results in Fig. 13. As shown in the figures, among 1908 genes that are core between all the strains under study, only 184 are core essential. We looked at subsets of the genes that were essential (core) in every combination of our bacterial strains to see whether they were phylogenetically informative or not. A phylogenetically informative subset is a subset that is essential (core) in two or more bacteria but not in all bacteria. We have marked the phylogenetically informative sets of genes with ticks and the uninformative ones with crosses. The results propose that although conservation of genes follows a tree-like trend with many phylogenetically informative sets of genes with high cardinality, the essentiality does not show a tree-like signal and most of the large sets of core essential genes belong to only one bacteria. We believe this happens due to the small number of essential genes.

Furthermore, we looked at different levels in the species tree and calculated the ratio of the number of core essential genes to the number of core genes in each level. We used three different methods for this. The first method was intersecting over core genes and core essential genes, so, genes are core in a node if and only if they are core in all the descendants of that node and are core essential if and only if they are core essential in all the descendants of that node.

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 averaged over the insertion indices of the pair of closest children of the ancestral node. We repeated this and averaged the averages until we reached the ancestral node. Then, we plotted the insertion indices and fitted an exponential and a gamma distribution to the plot as described in Section 2.2 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 (gain of essentiality) is highly improbable, so it tries to have up to one occurrence of gain of genes (gain of essentiality) and minimise the number of times that a gene (the essentiality of a gene) 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. 14.

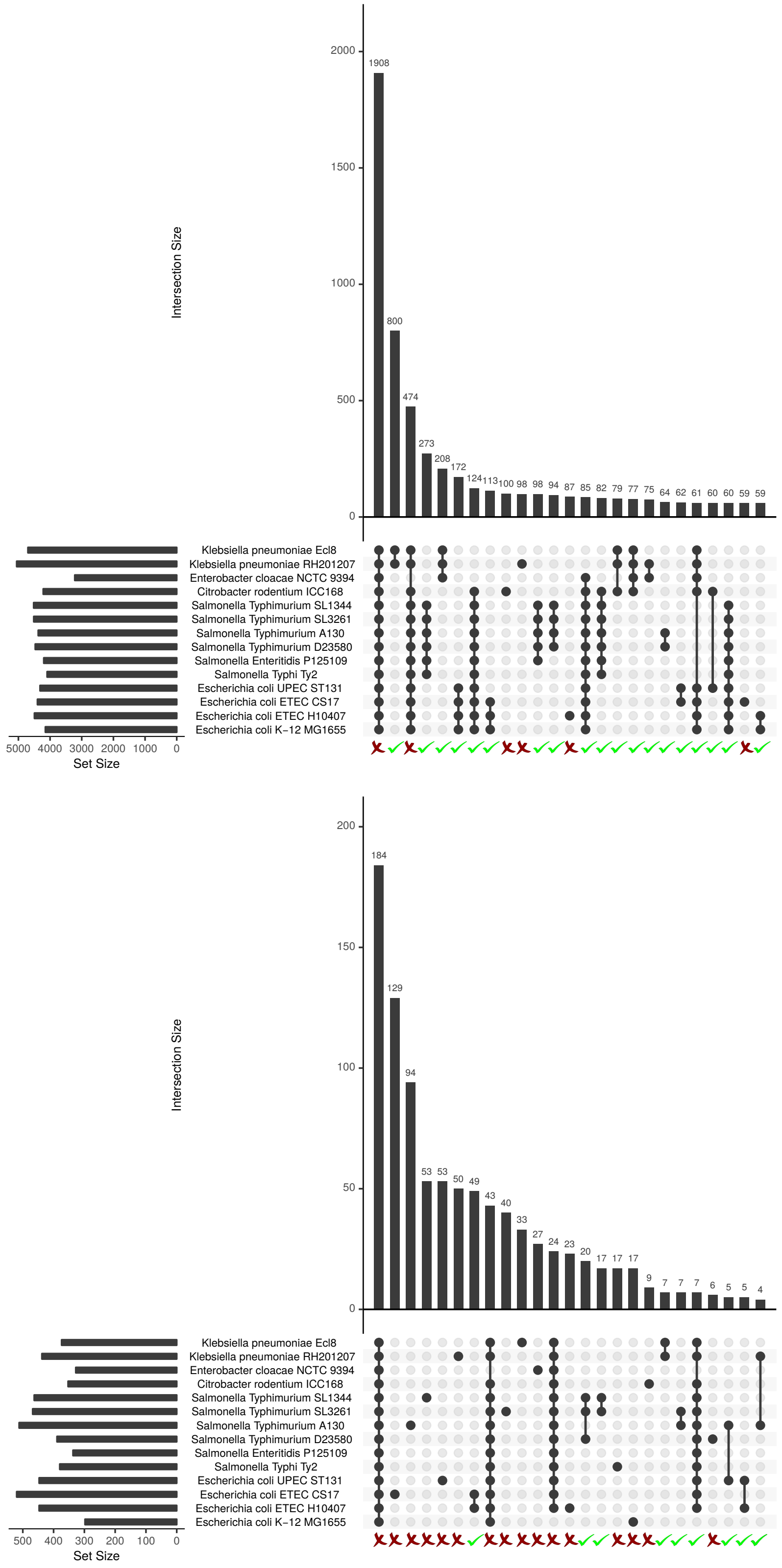

Figure 13. The first figure shows the number of core genes between each group of species and the second figure shows the number of core essential genes. The bars show the number of genes that are core between the strains marked with black circles. The tick marks show phylogenetically informative columns and the cross marks show non-informative columns.

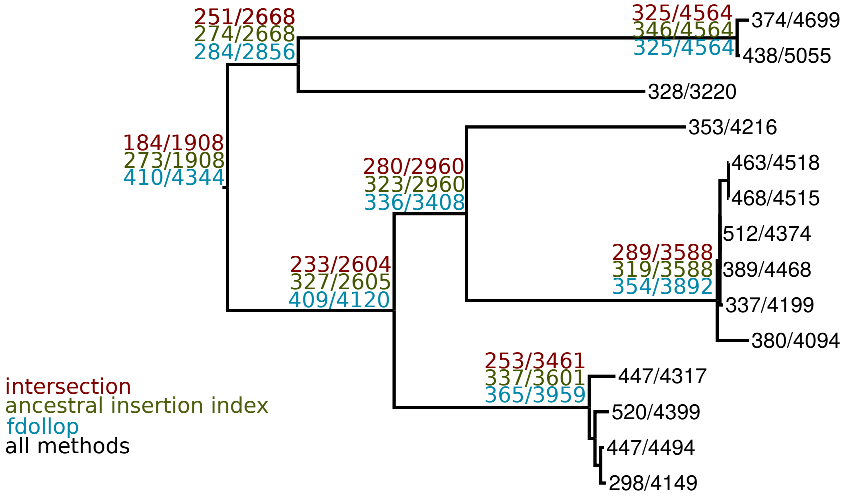

Figure 14. The tree shows the species tree in Fig.1 annotated by the number of core essential genes and core genes at each node. We used three methods to define core essential genes and core genes. The numbers at the leaves are the same using all these three methods. At the internal nodes, red shows the numbers using the intersection method, green shows the ancestral insertion index method, and turquoise shows fdollop method.

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.

As these three methods lead to different results, we compared the differences between the genes found in these three methods. For this, we compared the set of core essential genes resulted from intersection and ancestral insertion index methods. Then, we performed word enrichment analysis that was explained before on the 184 genes in intersection method and 89 genes that are core essential using ancestral insertion index and not core essential using the intersection method. Moreover, the intersection and fdollop methods and also ancestral insertion index and fdollop method were compared using the same procedure. The results are depicted in Fig.15.

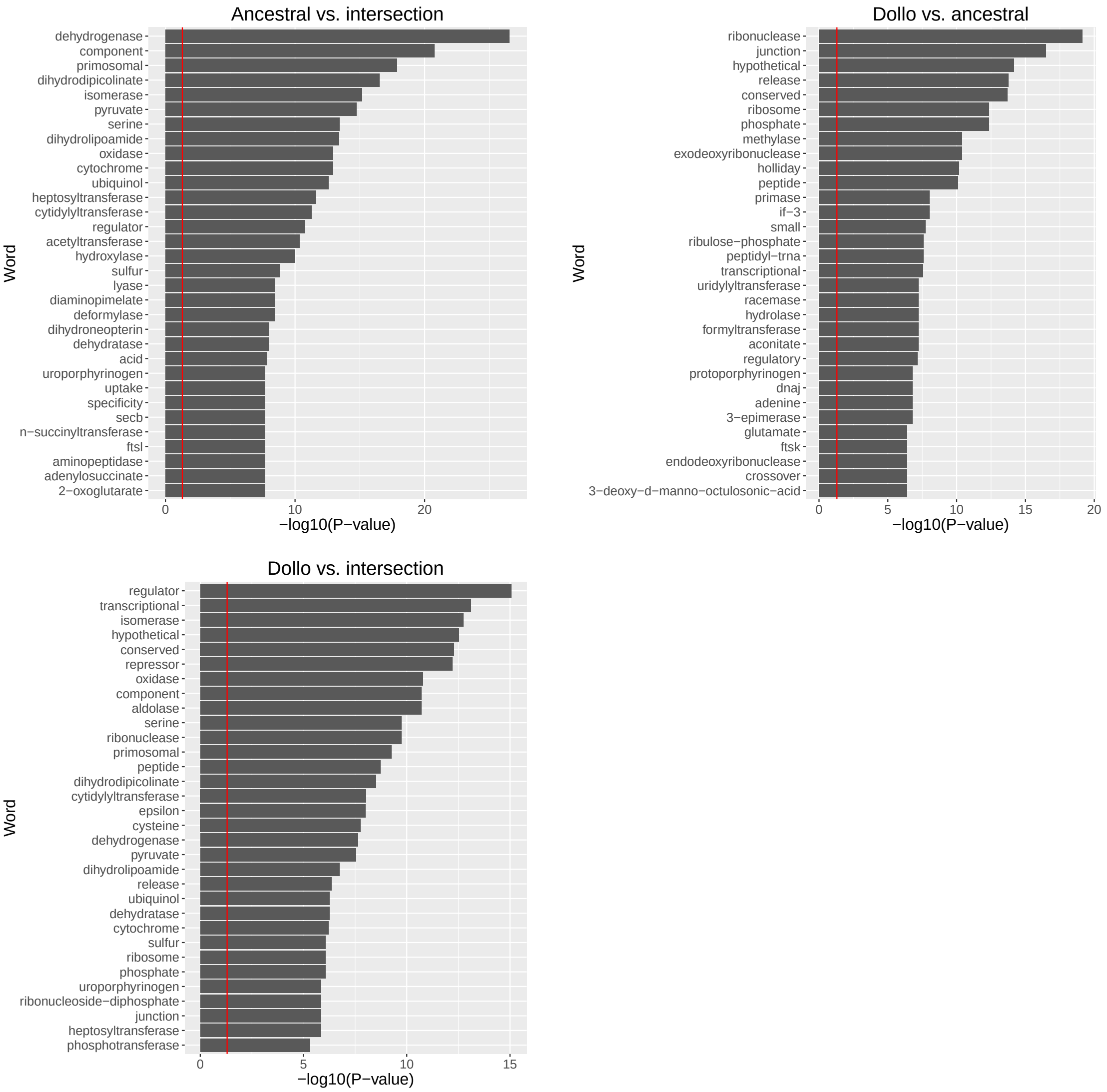

Figure 15. The figure shows the difference between core essential genes in intersection and ancestral insertion index methods, ancestral insertion index and fdollop methods, and intersection and fdollop methods. The red line shows P-value = 0.05. The P-values have been calculated using Fisher's exact test and then corrected using Benjamini-Hochberg-Yekutieli procedure. The top left figure shows words that are enriched in core essential genes found using ancestral insertion index method but not enriched in core essential genes found using intersection method. The top right figure shows words that are enriched in core essential genes found using fdollop method but not enriched in core essential genes found using ancestral insertion index method. The bottom left figure shows words that are enriched in core essential genes found using fdollop method but not enriched in core essential genes found using intersection method.

2 Materials and Methods

2.1 Transposon mutagenesis

Throughout time, species can gain or lose genes. We have investigated if these gene gain and losses are related to the essentiality of the genes themselves. For this, we have evaluated the essentiality of genes and their conservation and the relationship between these two. 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 well characterised family of Gram-negative bacteria with a variety of host ranges and pathogenecity [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 used transposon mutagenesis which is a frequently used method for the study of essentiality. We generated single inserted mutants using Tn5 transposon and placed our mutants in a selective media for Tn5. Then, we picked the mutants and pooled them and performed PCR enrichment using the method described in [4]. We 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 implies the degree of essentiality for that gene.

2.2 Essentiality

The more transposon insertions we observe in a gene after sequencing the genomes, the less essential the gene is. In order to quantify the essentiality of genes, we used a measure named insertion index which is proportional to the number of insertions in a gene.

To calculate the insertion index for each gene, we summed up the number of transposon insertion sites observed in that gene. Since the lengths of the genes are different, we normalised the insertion index by dividing it by gene length. Our experiment is 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 normalised our insertion indices by the ratio between the number of insertions in the whole genome and the length of the genome. We did not genes shorter than 100 base-pairs as they might not be targeted by any transposon due to their shortness.

Based on insertion indices, we divided the genes into four groups: essential genes, ambiguous, non-essential genes, and beneficial losses. We utilised 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. 16. The first peak shows the genes with no or just a few insertions which are considered as essential genes. We fitted an exponential distribution to the first peak and a gamma distribution to the second one. Then, we 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.

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.

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.

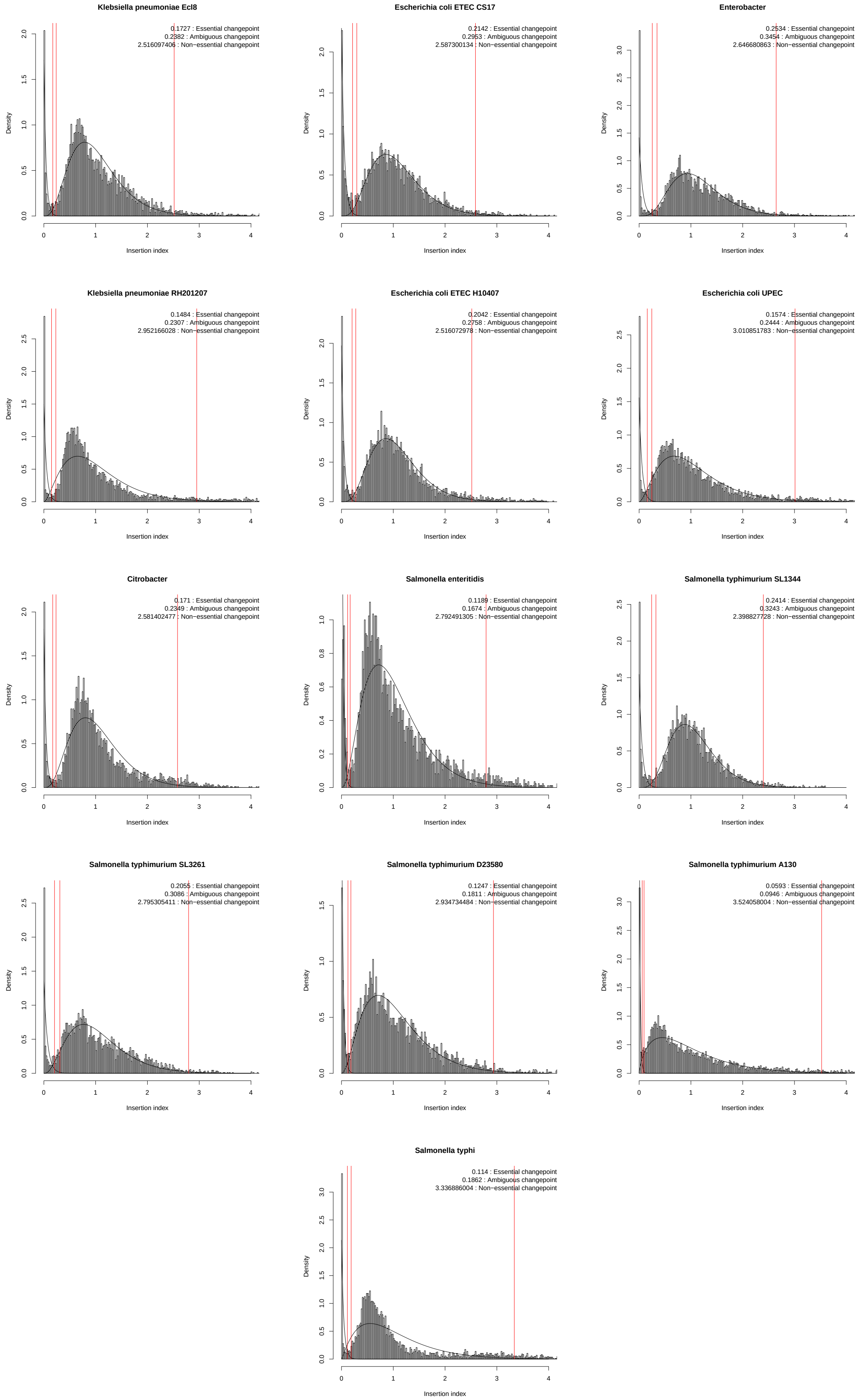

Figure 16. Plots show the insertion index distribution for each genome. The plots are divided into 4 regions using red lines. These regions from left to right are: essential, ambiguous, non-essential, and beneficial loss.

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. E. P. C. Rocha. The replication-related organization of bacterial genomes. 150(6):1609–1627.

23. 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.

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

25. 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.

26. 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.

27. 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.