Title

Abstract

Many genes have been identified with advances in sequencing technology and genome annotation methods. However, not all of these genes are of the same importance. We have used transposon mutagenesis to investigate gene essentiality in 14 strains of Enterobacteriaceae. We investigated the potential biases of this approach and found an origin of replication bis, no preferred insertion motif bias, a G-C bias in low G-C genes, and positional bias within genes. After correcting for these biases, we investigated the changes in the cohorts of essential genes and compared them to their conservation level. Surprisingly, we found that conserved genes are not necessarily essential, and essential genes are not necessarily conserved. However, on average, essential genes are more likely to be conserved.

1 Introduction

With the advent of sequencing technologies and genome annotation methods many genes have been identified. However, not all of these genes are of the same importance for the growth of an organism. So far, scientists have studied the essential genes in organisms from different domains of life [Luo et al.,]. These studies can give us new insights for developing new antibiotics that target essential genes of pathogenic bacteria [Clatworthy et al., , Peters et al.,] and synthesising minimal genomes [Hutchison et al., b, Hutchison et al., a, Reuß et al.,]. Different methods have been used for studying the essentility of genes in prokaryotes. Baba et al. [Baba et al.,] 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 [Xu et al.,]. Another method that is

widely used due to its simplicity is transposon mutagenesis along with high-throughput sequencing [Gawronski et al., ,van Opijnen et al., ,Langridge et al., ,Christen et al., ,Goodman et al., ,Wetmore et al., ,Rubin et al.,]. 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 [Barquist et al., a]. This can lead to the identification of essential genes.

Now that the essentiality of genes can be evaluated using different methods, it is possible to compare the essentiality data in different organisms and investigate the differentiation of essentiality of their genes. Curtis and Brun [Curtis and Brun,] have studied the essentiality changes in cell cycle genes of three alpha-proteobacteria strains: *Caulobacter crescentus*, *Brevundimonas subvibrioides*, and *Agrobacterium tumefaciens* and concluded that although essential genes responsible for cell functions are conserved, there are many essential genes that are specific to each organism. Freed et al. [Freed et al.,] have investigated the difference between essential genes in *Shigella flexneri* 2a 2457T and *Escherichia coli* K12 BW25113 and shown that some genes have gained essentiality in *Shigella flexneri* while there are no genes that are essential in *Escherichi coli* and not essential in *Shigella flexneri*. Canals et al. [Canals et al.,] have compared the essentiality of genes in *Salmonella* typhimurium and *Salmonella* Typhi and found that the essentiality of genes differs in different organisms. In a similar study, Barquist et al. [Barquist et al., b] have used transposon-directed insertion-site sequencing to study the differentiation of the essentiality of genes and ncRNAs in *Salmonella* serovars Typhi and Typhimurium. Although there are many studies on differentiation of essentiality in different organisms, these studies usually include two or three strains.

Our aim is to study the essentiality of genes in an evolutionary framework in 14 different organisms from Enterobacteriaceae family (Fig. 1). Enterobacteriaceae is a well characterised family of Gram-negative bacteria with a variety of host ranges and pathogenecity [Brenner and Krieg,]. In addition, we added the essential genes of *Escherichia coli* K-12 MG1655 from EcoGene database [Zhou and Rudd,] to our study. In EcoGene the essentiality of genes suggested as essential by Baba et al. [Baba et al.,] has been further studied and only 299 out of 303 genes are marked as essential. We first performed a detailed study of biases than can influence the inference of essentiality. Then, we normalised our data for the biases

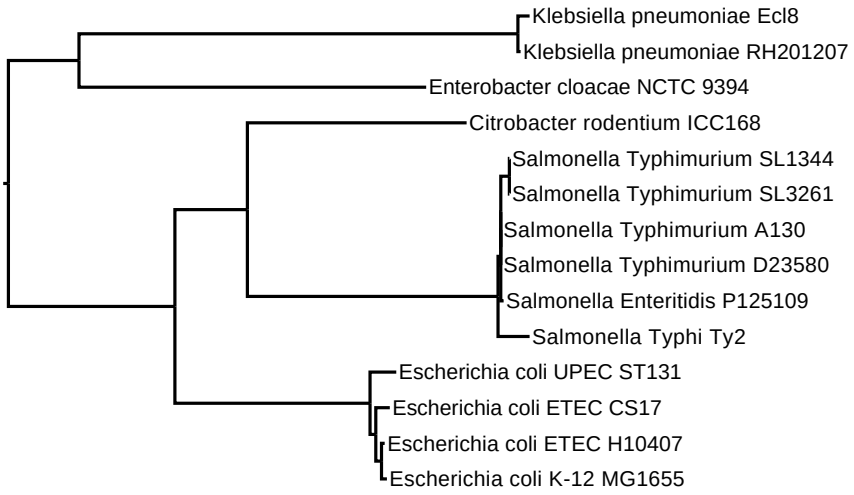

Figure 1. The species tree containing the 13 strains under study and Escherichia *coli* K-12 MG1655 studied in EcoGene [Zhou and Rudd,]. We have generated the tree by running RAxML [Stamatakis,] on Phylosift [Darling et al.,] amino acid markers.

and investigated the essentiality of genes in three classes of genes: genus specific (genes that are present only in one genus), single copy genes (genes with about one instance per genome in all of the genomes that we are studying), and multi-copy genes(genes that are copied multiple times per genome). We have also investigated how essentiality changes in the phylogenetic tree for these organisms.

2 Results and discussion

Throughout time, species can gain or lose genes. We investigated if these gene gain and losses are related to the essentiality of the genes. In this section, we have first described the biases that can affect our study, and then evaluated the essentiality of genes and their conservation and the relationship between these two.

2.1 Are there biases in transposon mutagenesis data?

To evaluate the essentiality of a gene, the number of insertions within that gene was measured as explained in 3.3. 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 [Barquist et al., b, Green et al., , Rubin et al., , Kimura et al.,]. We performed a detailed study of these biases. The biases that we have studied include: origin of replication bias, preferred insertion motif bias, and positional bias within genes.

2.1.1 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 can affect the results is if essential genes are clustered near the origin by nature. Rocha and Eduardo [Rocha,] 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 in Fig. 2. The figure indicates that the insertion indices decrease when the genes are located further from the origin of replication.

2.1.2 Preferred insertion motif bias

Another concern while inferring essentiality from transposon mutagenesis data 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, instead of their essentiality level. For this, we used Weblogo [Crooks et al.,] to generate 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. {Our genomes are G-C rich, so it makes sense to see more G-Cs in this plot. Use BLogo}

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%,

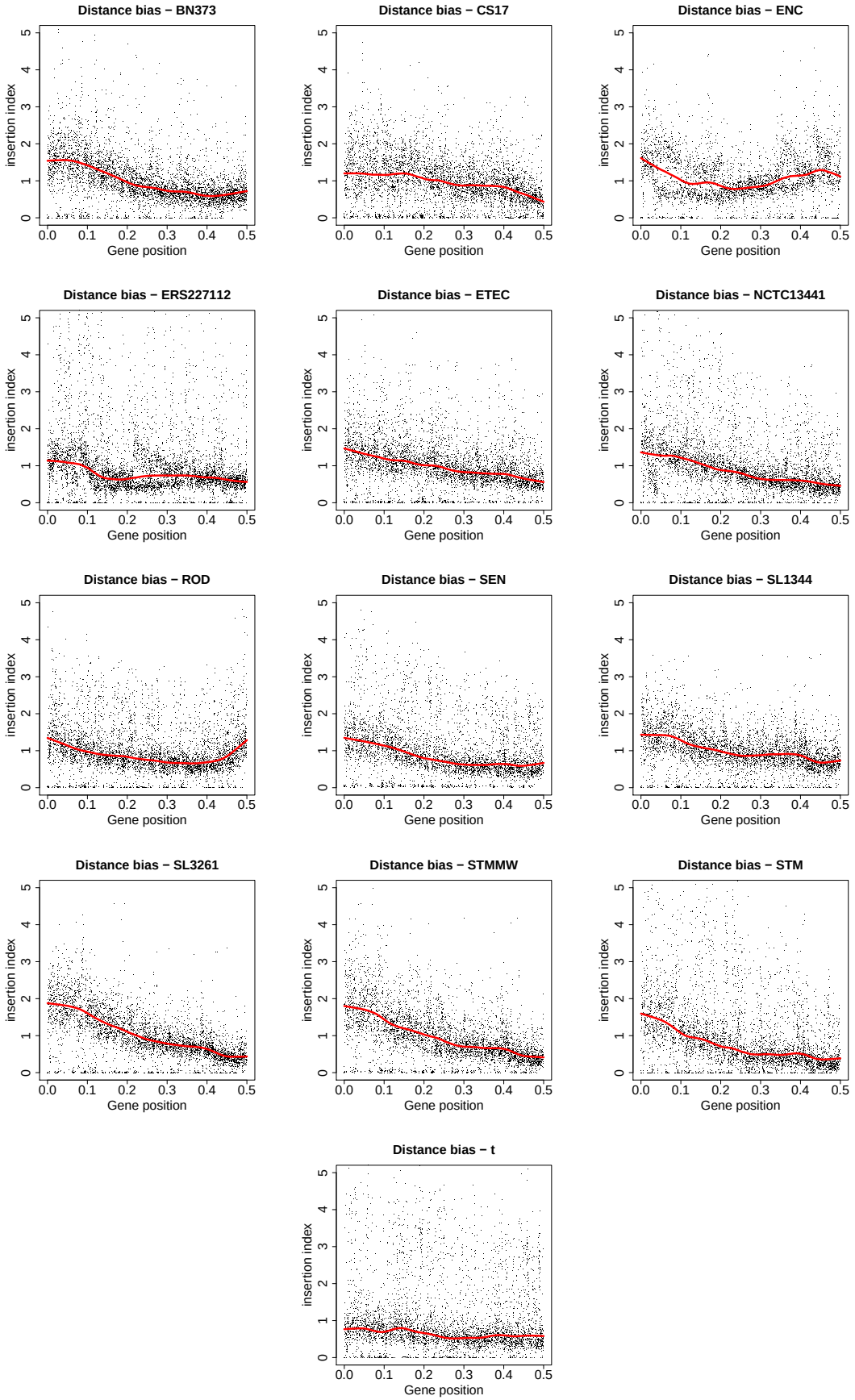

Figure 2. The plots show the distance of the genes from DnaA gene normalised by the lengths of the genomes versus the insertion indices of the genes. The distance from DnaA gene has been calculated in both directions and then the minimum value has been used for

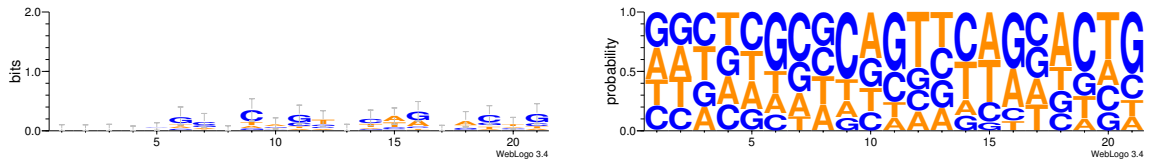

Figure 3. Sequence logo plots generated using sequences from 10 nucleotides flanking the 100 top most frequent insertion sites from each genome. On the left the height of each character corresponds to a bit score for that character (i.e. $2 - \sum f_a \times \log_2 f_a - \frac{1}{\ln 2} \times \frac{3}{2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

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 [Kimura et al.,]. 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 (Fig. 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.

2.1.3 Positional bias within genes

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 percentiles and calculated the mean insertion index for each percentile. Fig. 6 shows almost no bias

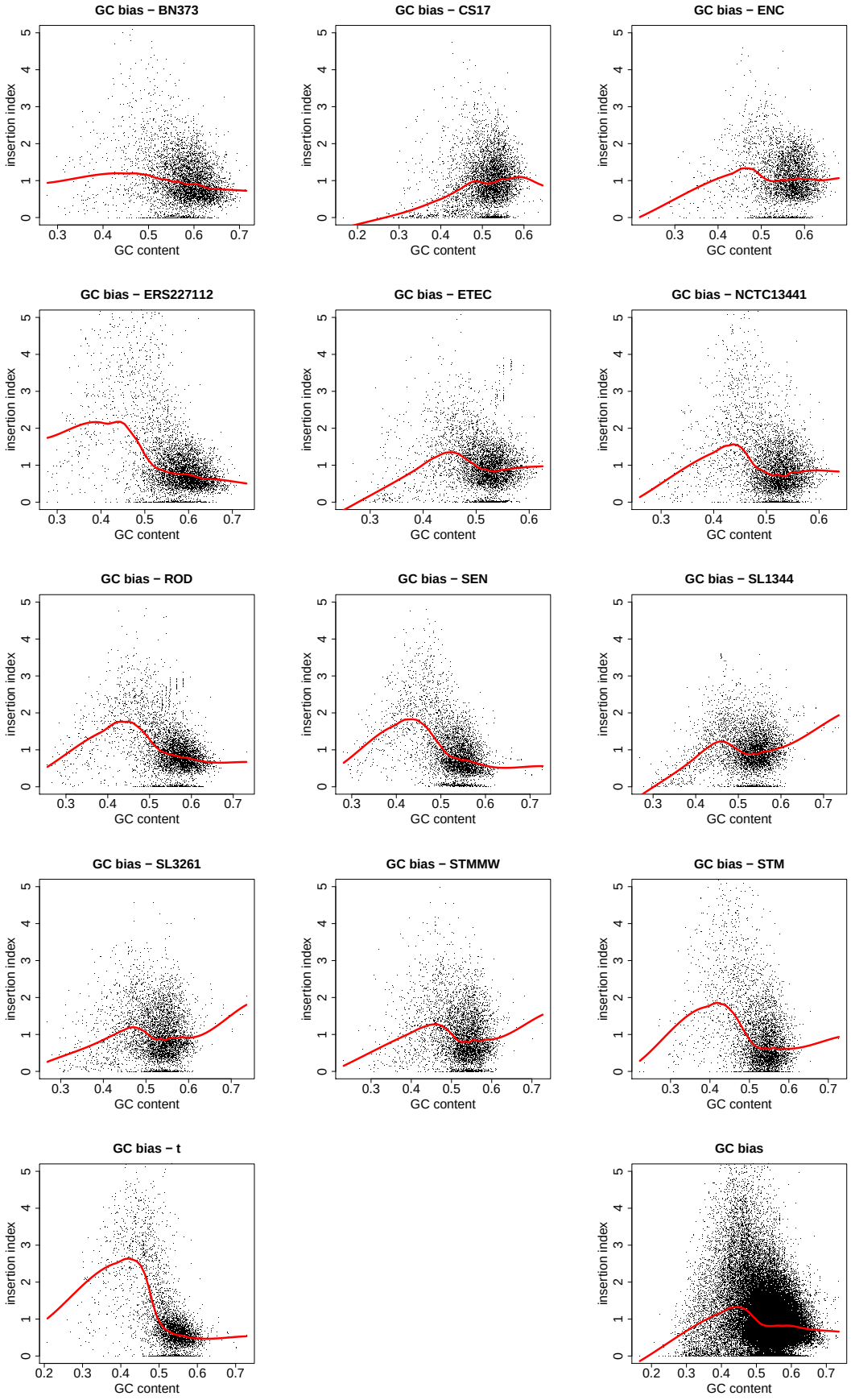

Figure 4. The plots show the ratio of G-C bases in the genes normalised by the lengths of the genes against their insertion indices. The red curves show the loess curve when the smoothness parameter is 0.2.

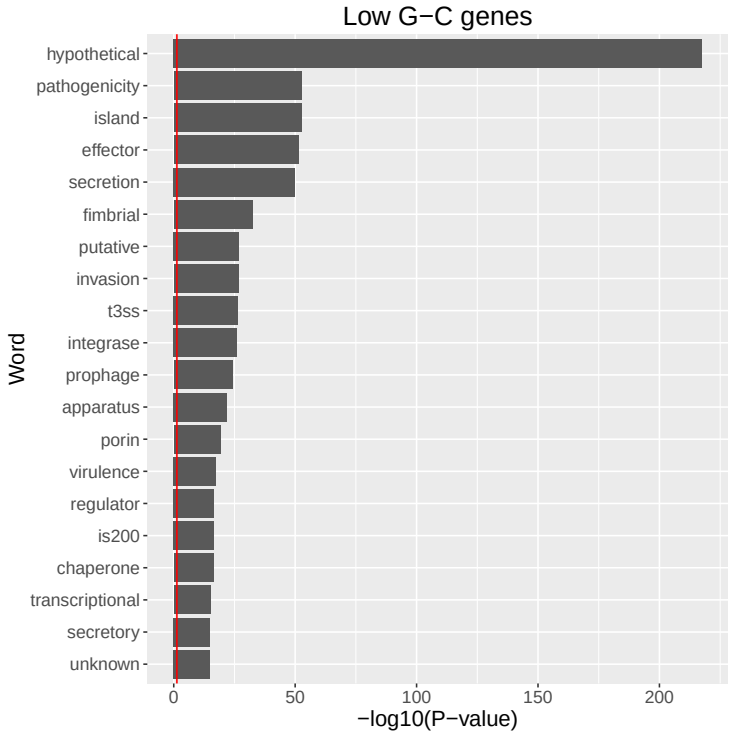

Figure 5. Word enrichment analysis for low G-C genes compared to genes with interquartile G-C level. The red line shows P-value = 0.05. The P-values have been calculated using Fisher's exact test and corrected using Benjamini-Hochberg-Yekutieli.

towards any location when considering all genes together. We also studied the bias in three of the groups defined in Section 3.3: 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 all genes. 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}

2.2 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.

2.2.1 Gene classes

In order to study the relationship between essentiality and conservation, we needed to evaluate the essentiality and conservation of genes. For this, we divided the genes into different levels of essentiality (essential genes, ambiguous, non-essential genes, beneficial losses). We first normalised the biases that exist in the data using the method explained in Section 3.4 and then defined the essentiality level of genes using the method explained in Section 3.3. Moreover, we grouped the genes into different classes of conservation (genus specific, single copy, multi-copy) using the method explaind in Section 3.5.

The results for comparing four levels of essentiality and three classes of conservation are depicted in Fig. 7. The high number of single copy genes in essential level, indicates that there is a set of essential genes in Enterobacteriaceae that are conserved and inclined to keep

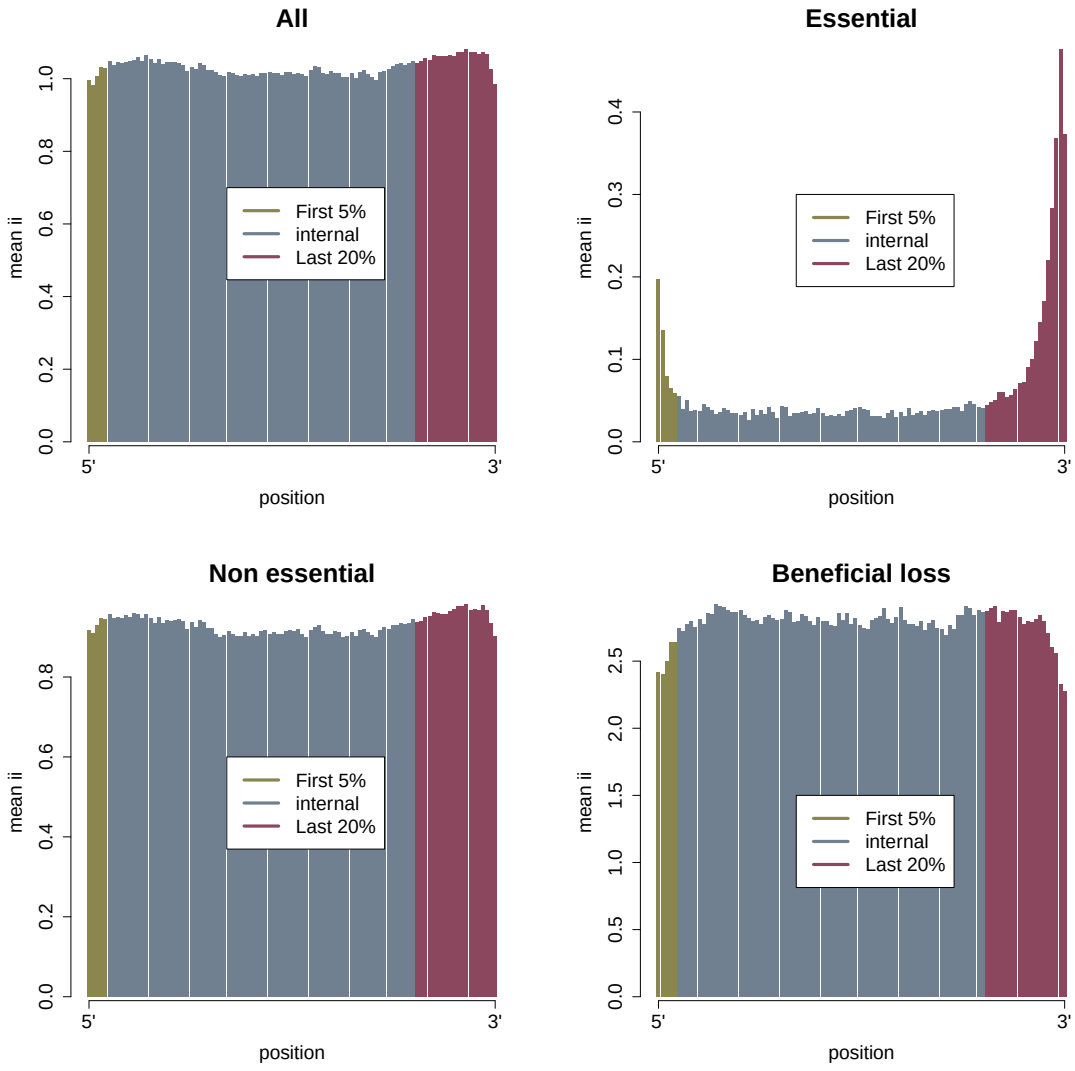

Figure 6. The plots show the average insertion index in the percentiles of all genes (top left), essential genes (top right), non-essential genes (bottom left), and beneficial losses (bottom right). The genes are divided into 3 segments: 5% of the genes on the 5' end, 20% of the genes on the 3' end, and the rest in the middle. These are shown by khaki, slate gray, and violet red respectively.

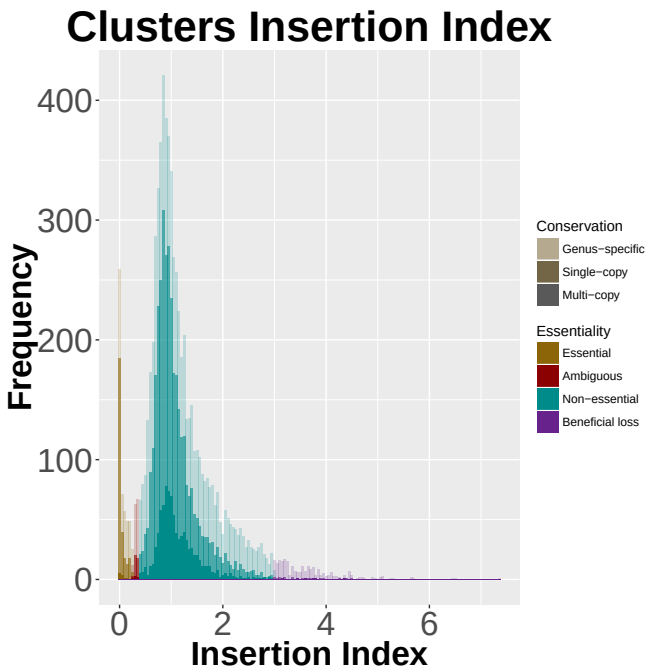

Figure 7. The genes have been clustered into orthologous groups using Hieranoid and paralogous groups using Jackhmmer and divided into 3 groups: genus specific, single copy, and multi-copy genes. Then, the essentiality of the clusters has been defined using the insertion indices of the genes in the clusters. The figure shows that most of the essential genes are in single copy group, while most of the beneficial losses are genus-specific.

their essentiality. However, the relatively high number of essential genes in genus specific class implies that each genus has a set of essential genes that makes it distinct. The figure also shows that beneficial losses are over represented in genus specific class. Therefore, beneficial losses are mostly recent genes that the organism tends to lose in the long run. Besides, most of the multi-copy clusters are non-essential and there are only a few multi-copy clusters that are essential. This can be explained by the redundancy that genes can keep even after $\sim$ 100 million years [Dean et al.,]. In the presence of two redundant variations of one gene, if we knock out one copy by transposon mutagenesis, the other copy compensates and the organism can still survive.

To study which functions are enriched in each class of essentiality, we used the word enrichment method explained in Section 3.6. Fig. 8 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 which is explained in Section 3.6. The results (Fig. 9) 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 [Kanehisa and Goto,], the pathway enrichment analysis does not show the enrichment of mobile genetic elements in beneficial-losses.

2.2.2 The evolution of essentiality

In this section, we compared the number of genes that were essential and conserved at each level of the phylogenetic tree. We were interested to see if the trend of the essentiality variation between different organisms is in agreement with the phylogenetic tree. In other words, we have tested if organisms close together have more essential genes in common than organisms that have separated earlier in the phylogenetic tree.

We needed to cluster sets of orthologous genes in the bacteria that we were studying. Homologous clusters introduced in 3.5 were not useful for this purpose as sets of paralogous genes with different essentiality levels can make essentiality inference ambiguous. Plenty of methods are proposed for this purpose. Altenhoff et al. have compared 15 of these methods [Altenhoff et al.,] and shown that Hieranoid [Schreiber and Sonnhammer,] is among three methods that keep a balance between precision and recall. We used Hieranoid for clustering orthologous genes. Hieranoid need a species tree for clustering genes. We used the method explained in Section 3.2 to generate this tree.

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. The genes that are conserved in all bacteria being studied are called core genes and the genes that are conserved and essential are called core

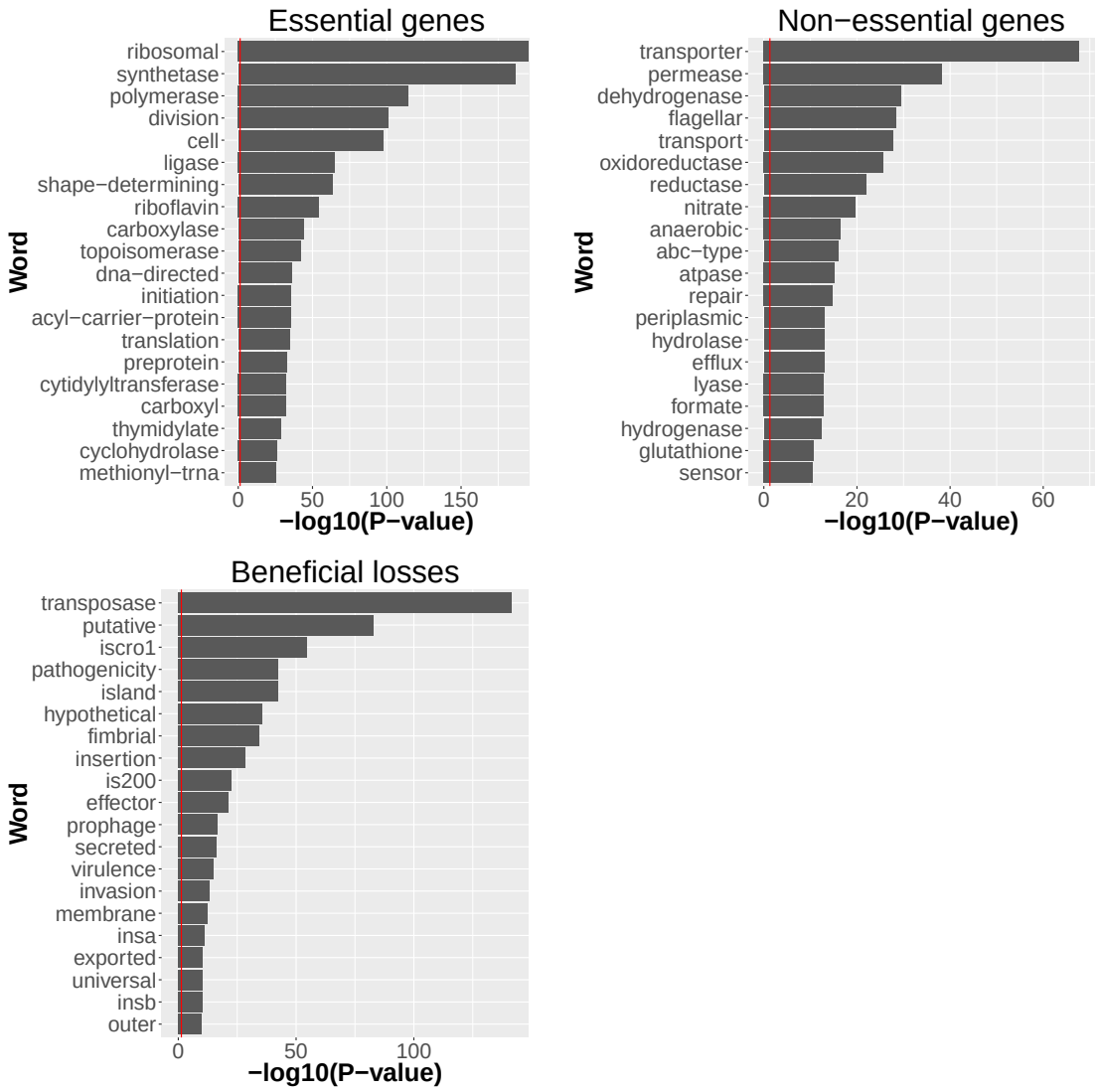

Figure 8. Word enrichment analysis for essential genes, non-essential genes, and beneficial losses compared to other genes. The red line shows P-value = 0.05. The P-values have been calculated using Fisher's exact test and then corrected using Benjamini-Hochberg-Yekutieli procedure.

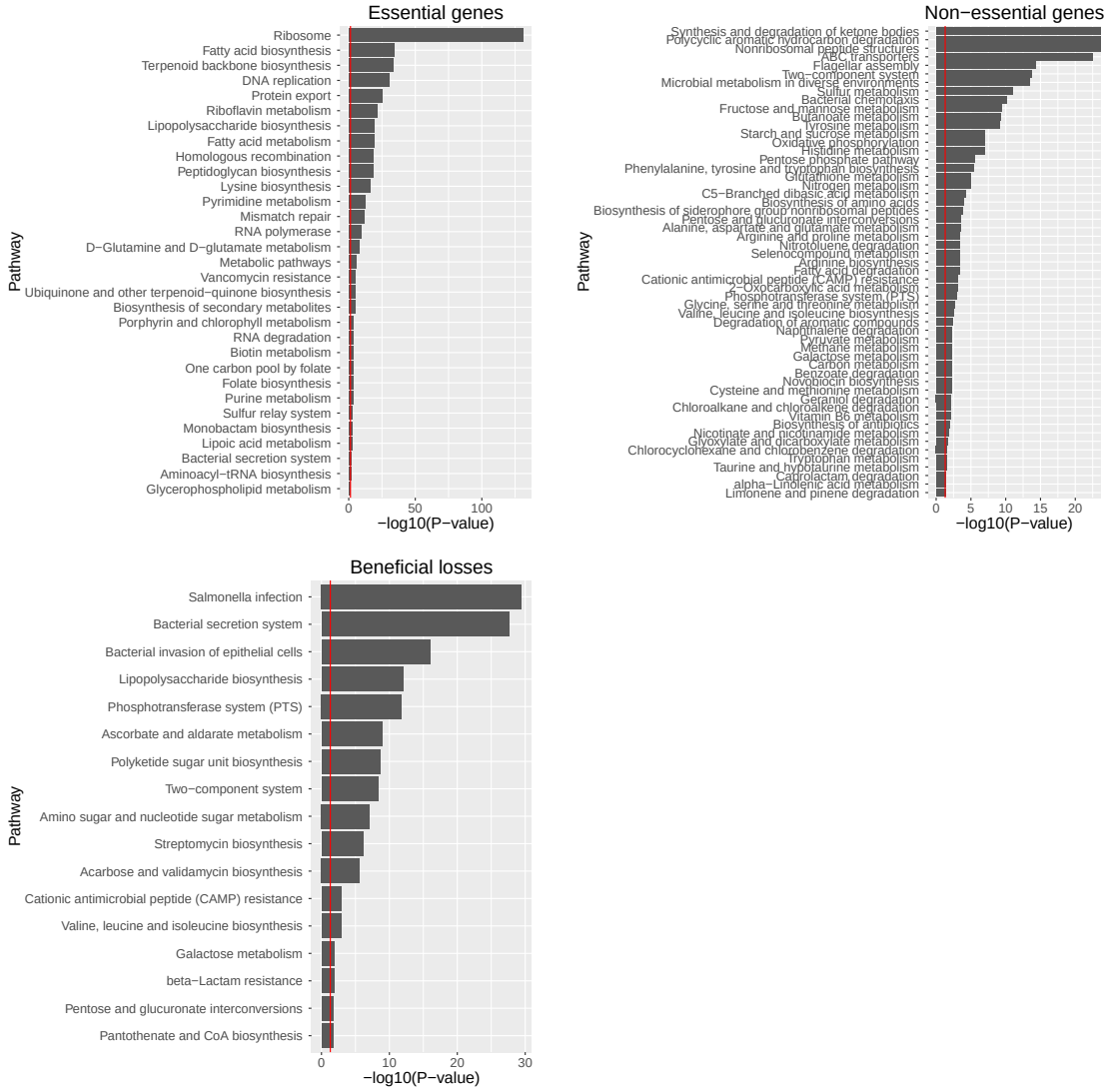

Figure 9. Pathway enrichment analysis for beneficial losses, essential genes, and non-essential genes compared to other genes. The red line shows P-value = 0.05. The P-values are calculated using hypergeometric test and then corrected using Benjamini-Hochberg-Yekutieli procedure.

essential genes. We counted the number of genes that were core between every combination of bacteria, and the number of core essential genes between those combinations. We used UpSetR package [Conway and Gehlenborg,] in R to visualise the results in Fig. 10. As shown in the figures, among 1908 genes that are core between all the bacteria under study, only 184 are core essential. We looked at subsets of the genes that were core 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 core 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 is due to the small number of essential genes. Each bacterium has about 300 to 500 essential genes among which 184 is core essential between all bacteria. A portion of the remaining essential genes are specific to each bacterium and many are shared between all bacteria except one. The number of remaining essential genes is so low that causes the essentiality trend not to be tree-like. Furthermore, some of the predicted essential genes might be artefacts of transposon mutagenesis method.

To further study the trend of essentiality changes, 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 inferring core genes and core essential genes. These methods are explained in Section 3.7. The numbers predicted using these three methods are shown in Fig. 11. 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 conflicting 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 explained in Section 3.6 on the 184 genes in intersection method and 89 genes that are core essential using ancestral insertion index and not core essential using the

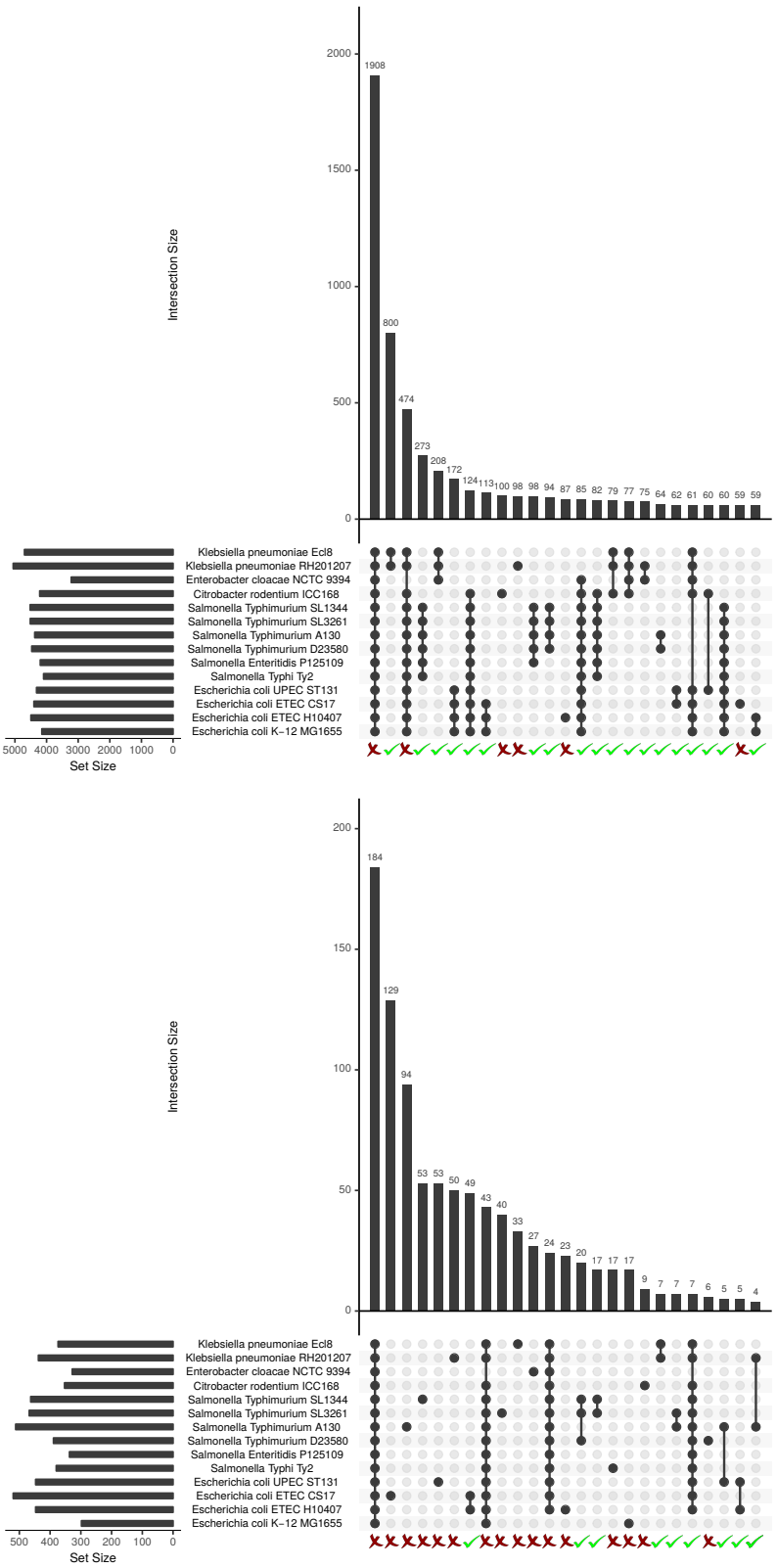

Figure 10. The first figure shows the number of core genes between each group of species and the second figure shows the number of core essential genes. The bars show the number of genes that are core between the strains marked with black circles. The tick marks show phylogenetically informative columns and the cross marks show non informative columns.

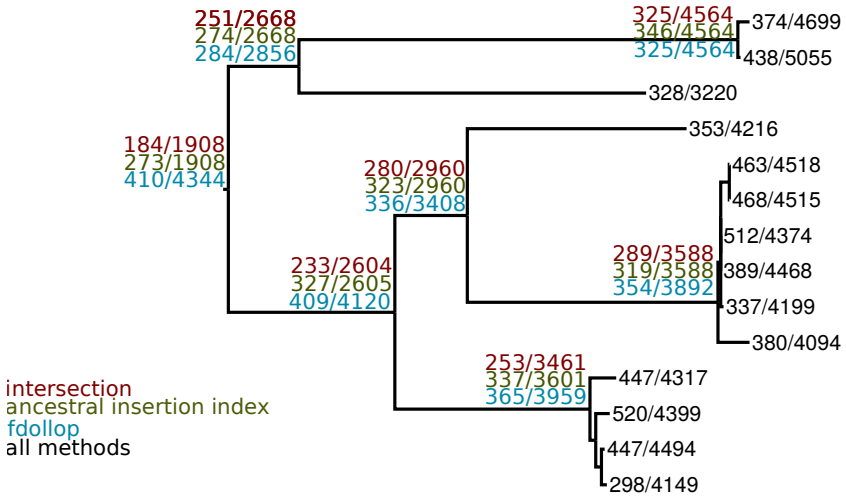

Figure 11. The tree shows the species tree in Fig.1 annotated by the number of core essential genes and core genes at each node. We used three methods to define core essential genes and core genes. The numbers at the leaves are the same using all these three methods. At the internal nodes, red shows the numbers using the intersection method, green shows the ancestral insertion index method, and turquoise shows fdollop method.

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.12.

3 Materials and Methods

3.1 Transposon mutagenesis

We studied 2 *Klebsiella* strains, an *Enterobacter* strain, a *Citrobacter* strain, 6 *Salmonella* strains, and 3 *Escherichia* strains (Fig. 1) and compared the essentiality of genes in these strains and *Escherichia coli* K-12 MG1655 from another study [Baba et al.,]. These strains are all selected from Enterobacteriaceae family and a transposon mutagenesis study have been performed on them. We generated single inserted mutants using Tn5 transposon and placed the mutants in a selective media for Tn5. Then, we picked the mutants and pooled them and performed PCR enrichment using the method described in [Barquist et al., c]. We sequenced the fragments and mapped them back to the genome to figure out the number of

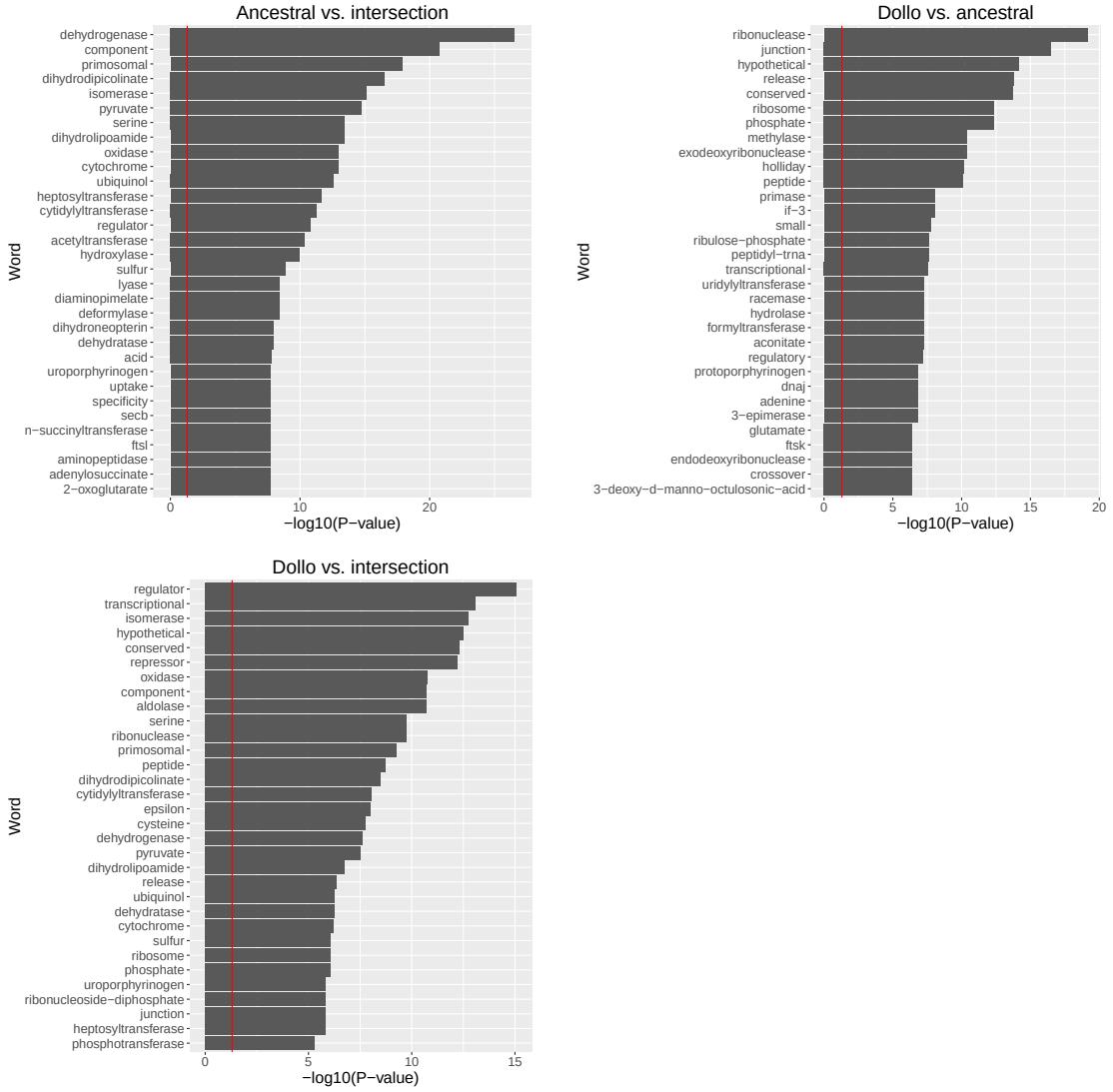

Figure 12. The figure shows the difference between core essential genes in intersection and ancestral insertion index methods, ancestral insertion index and fdollop methods, and intersection and fdollop methods. The red line shows P-value = 0.05. The P-values have been calculated using Fisher's exact 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.

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.

3.2 Generating the species tree

We first ran search and align commands from PhyloSift package [Darling et al.,] to select gene markers for generating a phylogenetic tree and aligning them. Then, we concatenated the protein alignments for all 14 genomes and ran RaxML [Stamatakis,] with "PROTGAMMALG4M" amino acid substitution model, "a" algorithm and 100 alternative runs on distinct starting trees.

3.3 Essentiality levels

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, the insertion indices were then normalised by dividing them 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 the insertion indices by the ratio between the number of insertions in the whole genome and the length of the genome.

Based on insertion indices, the genes were divided into four groups: essential genes, ambiguous, non-essential genes, and beneficial losses. We utilised the pipeline introduced by Barquist et al. [Barquist et al., c] to evaluate the essentiality of genes. The insertion index distribution plot has two peaks and a heavy tail as shown in Fig. 13. A loess curve was fitted to this distribution to find where the two peaks separate from each other. 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.

3.4 Bias correction

We observed a distance from origin of replication bias and also a positional bias within genes. Hence, these biases needed to be corrected before inferring the essentiality of genes from their insertion indices. We did not include genes shorter than 100 base-pairs in our study as they might not be targeted by any transposon due to their shortness.

To overcome the distance from origin of replication bias, we divided the value of insertion index for a gene in a specific position by the predicted value by loess for that position. This value was then multiplied by the average insertion index. To overcome the positional bias within genes, we calculated the insertion index for genes by ignoring 5% from the 5' end and 20% from the 3' end of the genes. 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. 14.

3.5 Conservation classes

To study whether each gene in the 14 organisms is conserved we proposed a program that clusters homologous proteins. This program uses Jackhmmer from HMMER package [Eddy,] 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

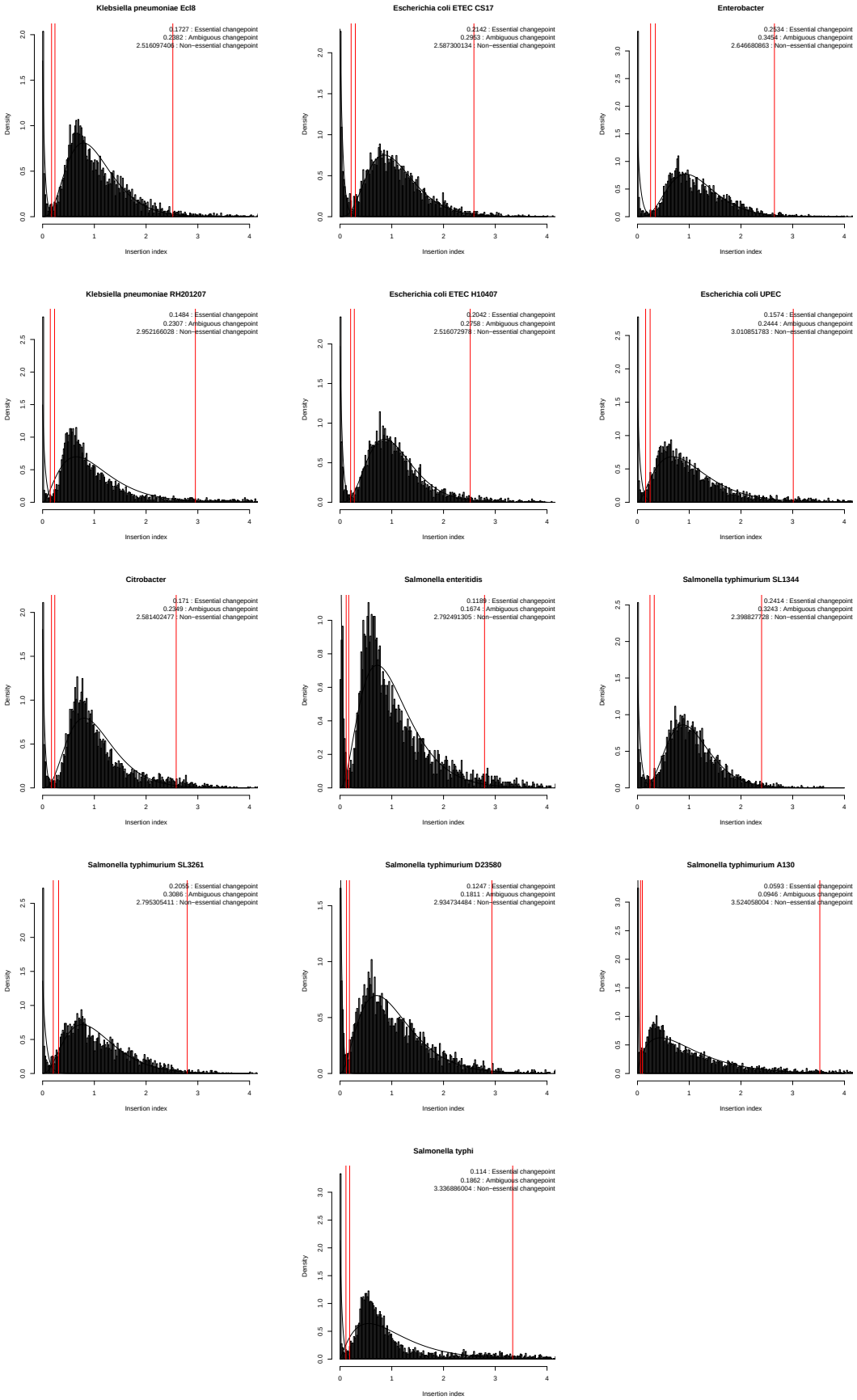

Figure 13. Plots show the insertion index distribution for each genome. The plots are divided into 4 regions using red lines. These regions from left to right are: essential, ambiguous, non-essential, and beneficial loss.

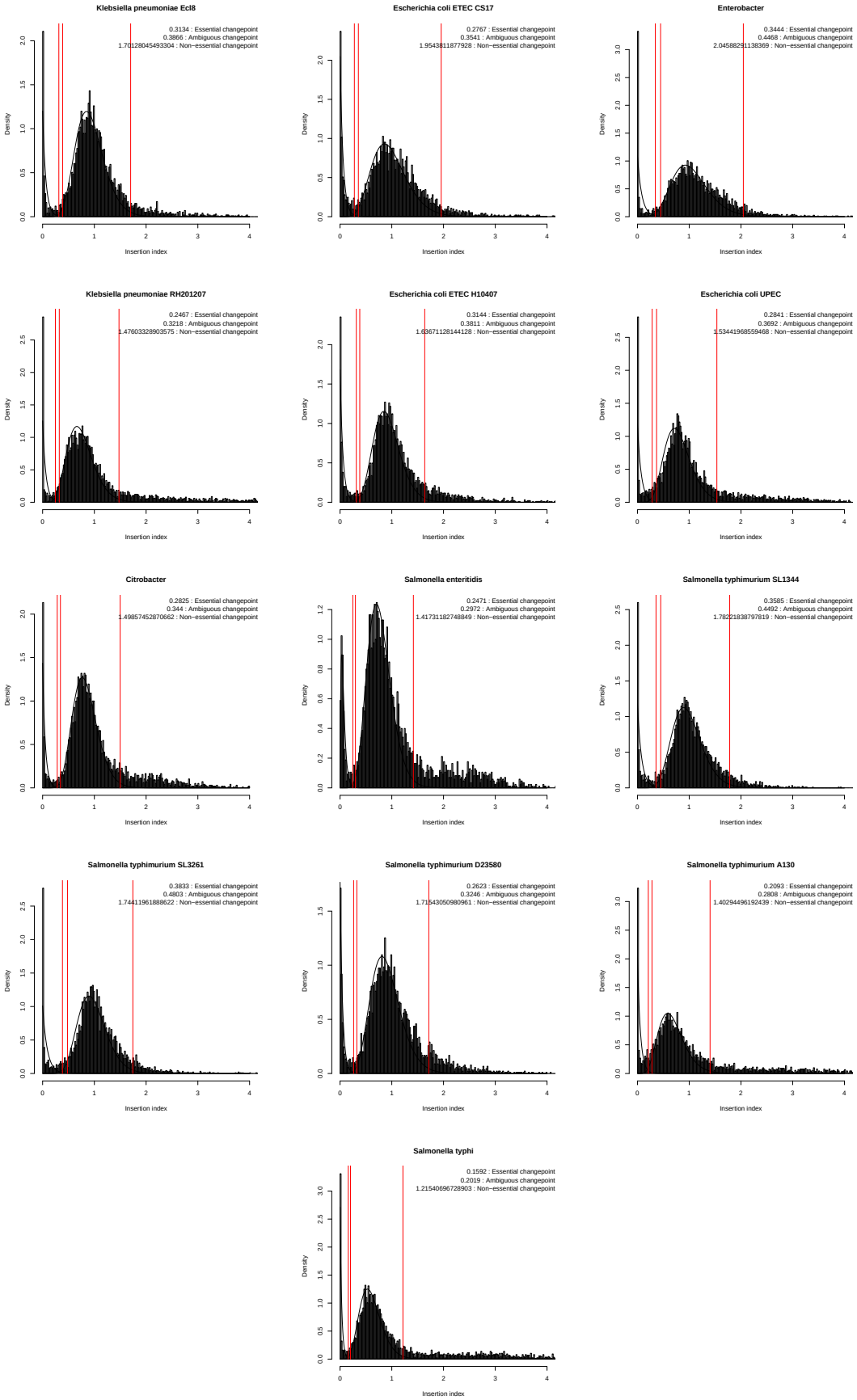

Figure 14. Plots show the insertion index distribution for each genome after correcting for distance from the origin of replication bias and bias towards the position of insertion within genes. The plots are divided into 4 regions using red lines. These regions from left to right

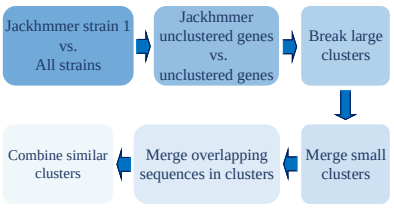

Figure 15. The steps of our proposed algorithm for clustering homologous genes.

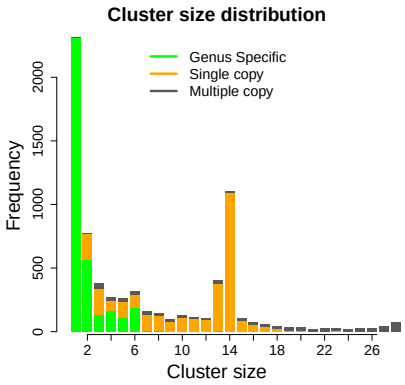

Figure 16. Size distribution for all clusters of homologous genes. Genus specific genes are genes that are present only in one genus, single copy genes are present in more than one genus and more than 70% of them are not duplicated, and multi-copy genes are present in more than one genus and less than 70% of them are not duplicated.

summarised in Fig. 15 and the distribution of cluster lengths after clustering the genes of 14 strains under study is plotted in Fig. 16.

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. 16.

3.6 Word and pathway enrichment analysis

To study the enrichment of words in a group of genes compared to a second group, We gathered the descriptions of genes from their embl files. Then, we counted the repeat number of each word for the genes in each group 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 were then corrected using Benjamini-Hochberg-Yekutieli procedure.

For pathway enrichment analysis, we downloaded pathway datasets for strains that were available in KEGG database [Kanehisa and Goto,]. 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.

3.7 Defining core and core essential genes

We used three different methods for defining core and core essential genes: intersection, ancestral insertion index, and Dollo law. These three methods are explained in what follows.

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 3.3 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.

4 Conclusion

In this paper, we studied the relationship between the essentiality and conservation of genes in 14 bacteria from Enterobacteriaceae family. We first studied the biases that can affect our results and showed that transposon insertions are more abundant near the origin of replication. In addition, there is a slight preferred insertion motif bias and a G-C bias in A-T rich genes. Moreover we showed that transposon insertions are more abundant near the ends of essential genes compared to the internal region, while it is the opposite in beneficial losses.

After correcting biases, we studied the essentiality versus conservation by dividing the genes into three classes of conservation: genus specific, single copy, and multi-copy, dividing them into four levels of essentiality: essential, ambiguous, non-essential, and benefical losses. We found that essential genes are mostly single copy, however there is a considerable number of genus specific genes in essential level which help to distinguish between genera. Furthermore, beneficial losses are mostly genus specific which means they are mostly new genes. The other finding was that multi-copy genes are mostly non-essential which can happen due to redundancy.

We also found that the pattern of essentiality changes is not in agreement with the phylogenetic tree due to the small number of essential genes. Also the ratio between core essential genes and core genes can be either constant or increasing depending on the method that we use for defining core and core essential genes. {WHAT?}

Overall, we have found that conserved genes are not necessarily essential, and essential genes are not necessarily conserved. However, on average, essential genes are more likely to be conserved.

Acknowledgments

References

Altenhoff et al., . Altenhoff, A. M., Boeckmann, B., Capella-Gutierrez, S., Dalquen, D. A., DeLuca, T., Forslund, K., Huerta-Cepas, J., Linard, B., Pereira, C., Pryszcz, L. P.,

Schreiber, F., da Silva, A. S., Szklarczyk, D., Train, C.-M., Bork, P., Lecompte, O., von Mering, C., Xenarios, I., Sjölander, K., Jensen, L. J., Martin, M. J., Muffato, M., Quest for Orthologs Consortium, Gabaldón, T., Lewis, S. E., Thomas, P. D., Sonnhammer, E., and Dessimoz, C. Standardized benchmarking in the quest for orthologs. 13(5):425–430.

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

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

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

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

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

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

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

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

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

Crooks et al., . Crooks, G. E., Hon, G., Chandonia, J.-M., and Brenner, S. E. WebLogo: a sequence logo generator. 14(6):1188–1190.

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

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

Dean et al., . Dean, E. J., Davis, J. C., Davis, R. W., and Petrov, D. A. Pervasive and persistent redundancy among duplicated genes in yeast. 4(7):e1000113.

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

Freed et al., . Freed, N. E., Bumann, D., and Silander, O. K. Combining shigella tn-seq data with gold-standard e. coli gene deletion data suggests rare transitions between essential and non-essential gene functionality. 16(1):203.

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

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

Green et al., . Green, B., Bouchier, C., Fairhead, C., Craig, N. L., and Cormack, B. P. Insertion site preference of mu, tn5, and tn7 transposons. 3:3.

Hutchison et al., a. Hutchison, C. A., Chuang, R.-Y., Noskov, V. N., Assad-Garcia, N., Deerinck, T. J., Ellisman, M. H., Gill, J., Kannan, K., Karas, B. J., Ma, L., Pelletier, J. F., Qi, Z.-Q., Richter, R. A., Strychalski, E. A., Sun, L., Suzuki, Y., Tsvetanova, B.,

Wise, K. S., Smith, H. O., Glass, J. I., Merryman, C., Gibson, D. G., and Venter, J. C. Design and synthesis of a minimal bacterial genome. 351(6280):aad6253.

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

Kanehisa and Goto, . Kanehisa, M. and Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. 28(1):27–30.

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

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

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

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

Reuß et al., . Reuß, D. R., Commichau, F. M., Gundlach, J., Zhu, B., and Stülke, J. The blueprint of a minimal cell: MiniBacillus. 80(4):955–987.

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

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

Schreiber and Sonnhammer, . Schreiber, F. and Sonnhammer, E. L. L. Hieranoid: hierarchical orthology inference. 425(11):2072–2081.

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

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

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

Xu et al., . Xu, H. H., Trawick, J. D., Haselbeck, R. J., Forsyth, R. A., Yamamoto, R. T., Archer, R., Patterson, J., Allen, M., Froelich, J. M., Taylor, I., Nakaji, D., Maile, R., Kedar, G. C., Pilcher, M., Brown-Driver, V., McCarthy, M., Files, A., Robbins, D., King, P., Sillaots, S., Malone, C., Zamudio, C. S., Roemer, T., Wang, L., Youngman, P. J., and Wall, D. Staphylococcus aureus TargetArray: comprehensive differential essential gene expression as a mechanistic tool to profile antibacterials. 54(9):3659–3670.

Zhou and Rudd, . Zhou, J. and Rudd, K. E. EcoGene 3.0. 41:D613–D624.