Is essentiality of genes conserved in Enterobacteriaceae?

Fatemeh Ashari-Ghomi [*], **Paul P. Gardner** [*] **and Lars Barquist** [†]

[*]University of Canterbury, and [†]Wurzburg University

Submitted to Proceedings of the National Academy of Sciences of the United States of America

essentiality | conservation | transposon insertion

Introduction

Studying the essentiality of genes helps with identifying the fundamental processes necessary for cell viability [1]. So far, scientists have studied the essential genes in organisms from different domains of life [2]. The results have led to new insights for developing new antibiotics that target essential genes of pathogenic bacteria [3, 4] and synthesising new genomes [5, 6]. Researchers have used different methods for studying the essentility of genes in prokaryotes. Baba et al. [7] 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 [8]. Another method that is widely used due to its simplicity and accuracy is transposon mutagenesis along with high-throughput sequencing [9, 10, 11, 12, 13, 14, 15]. 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 [16]. 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. Barquist et al. [17] 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. [7] are depicted in Fig. 1.

Enterobacteriaceae is a family that includes Gram-negative bacteria with different host ranges and pathogenecity found in soil, water, plants, animals and humans [18]. In humans, various strains from this family can cause diarrhoea, septicaemia, urinary tract infection, meningitis, respiratory disease, and wound and burn infection [18]. Besides, they can infect poultry and livestocks and cause financial losses for farmers [18]. 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.

{A summary of what we have done}

Transposon mutagenesis

Strains. 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 [7]. These strains are all selected from Enterobacteriaceae family.

Transposon insertion workflow. 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 [20] 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.

Essentiality. 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 in all the strains comparable, 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 our genes into three groups: essential genes, non-essential genes, and beneficial losses. We have adapted the pipeline introduced by Barquist et al. [20] to evaluate the essentiality of genes. The insertion index distribution plot has two peaks and a heavy tail Fig. 8. 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.

Are transposons biased towards certain positions?

Different articles have reported biases in transposon mutagenesis [17, 15, 21]. We have performed a thorough study of these biases. 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

Reserved for Publication Footnotes

by the length of the genome. Fig. 4 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.

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. 5. The results show a slight bias towards certain combinations of bases. 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. 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 [21]. The other reason is that the genes with low G-C content are enriched in mobile genetic elements (corrected p-value for the number of repetition of these words in the quartile with lowest G-C and the interquartile genes: pathogenicity island: 4.551893655690 93e-54, prophage: 3.20756138380406e-26, phage: 1.81663788061583e-08, bacteriophage: 0.00285196530662877) compared to the genes with high G-C content and that has caused seeing a different pattern of essentiality in that region.

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 divided our genes into three groups - essential genes, non-essential genes, and beneficial losses - and studied the bias in each group. 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.

Essentiality and conservation

Homolog clustering algorithm (homclust). 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 [19] 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. 2 and its results are illustrated in Fig. 3.

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 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. 3.

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. 8. The figure shows that most of the essential clusters are single copy and most of the beneficial losses are genus specific. {KEGG results}

Evolution of essentiality. 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. The results propose that although conservation of genes follows a tree-like trend, the essentiality does not show a tree-like signal.

- UpSetR results (Fig. 9)
- Stringent
- Dollo law
- Ancestral insertion index

Case study of genes

Core genes.

Accessory genes. ACKNOWLEDGMENTS.

1. Mario Juhas, Leo Eberl, and John I. Glass. Essence of life: essential genes of minimal genomes. 21(10):562–568.
2. Hao Luo, Yan Lin, Feng Gao, Chun-Ting Zhang, and Ren Zhang. DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. 42:D574–D580.
3. Anne E. Clatworthy, Emily Pierson, and Deborah T. Hung. Targeting virulence: a new paradigm for antimicrobial therapy. 3(9):541–548.
4. JasonM. Peters, Alexandre Colavin, Handuo Shi, TomaszL. Czarny, MatthewH. Larson, Spencer Wong, JohnS. Hawkins, CandyH. S. Lu, Byoung-Mo Koo, Elizabeth Marta, AnthonyL. Shiver, EvanH. Whitehead, JonathanS. Weissman, EricD. Brown, LeiS. Qi, KerwynCasey Huang, and CarolA. Gross. A comprehensive, CRISPR-based functional analysis of essential genes in bacteria. 165(6):1493–1506.
5. Clyde A. Hutchison, Scott N. Peterson, Steven R. Gill, Robin T. Cline, Owen White, Claire M. Fraser, Hamilton O. Smith, and J. Craig Venter. Global transposon mutagenesis and a minimal mycoplasma genome. 286(5447):2165–2169.
6. Clyde A. Hutchison, Ray-Yuan Chuang, Vladimir N. Noskov, Nacyra Assad-Garcia, Thomas J. Deerinck, Mark H. Ellisman, John Gill, Krishna Kannan, Bogumil J. Karas, Li Ma, James F. Pelletier, Zhi-Qing Qi, R. Alexander Richter, Elizabeth A. Strychalski, Lijie Sun, Yo Suzuki, Billyana Tsvetanova, Kim S. Wise, Hamilton O. Smith, John I. Glass, Chuck Merryman, Daniel G. Gibson, and J. Craig Venter. Design and synthesis of a minimal bacterial genome. 351(6280):aad6253.
7. Tomoya Baba, Takeshi Ara, Miki Hasegawa, Yuki Takai, Yoshiko Okumura, Miki Baba, Kirill A Datsenko, Masaru Tomita, Barry L Wanner, and Hirotada Mori. Construction

of escherichia coli k-12 in-frame, single-gene knockout mutants: the keio collection. 2:2006.0008.

8. H. Howard Xu, John D. Trawick, Robert J. Haselbeck, R. Allyn Forsyth, Robert T. Yamamoto, Rich Archer, Joe Patterson, Molly Allen, Jamie M. Froelich, Ian Taylor, Danny Nakaji, Randy Maile, G. C. Kedar, Marshall Pilcher, Vickie Brown-Driver, Melissa McCarthy, Amy Files, David Robbins, Paula King, Susan Sillaots, Cheryl Malone, Carlos S. Zamudio, Terry Roemer, Liangsu Wang, Philip J. Youngman, and Daniel Wall. Staphylococcus aureus TargetArray: comprehensive differential essential gene expression as a mechanistic tool to profile antibacterials. 54(9):3659–3670.
9. Jeffrey D. Gawronski, Sandy M. S. Wong, Georgia Giannoukos, Doyle V. Ward, and Brian 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.
10. Tim van Opijnen, Kip L. Bodi, and Andrew Camilli. Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. 6(10):767–772.
11. Gemma C. Langridge, Minh-Duy Phan, Daniel J. Turner, Timothy T. Perkins, Leopold Parts, Jana Haase, Ian Charles, Duncan J. Maskell, Sarah E. Peters, Gordon Dougan, John Wain, Julian Parkhill, and A. Keith Turner. Simultaneous assay of every salmonella typhi gene using one million transposon mutants. 19(12):2308–2316.
12. Beat Christen, Eduardo Abeliuk, John M Collier, Virginia S Kalogeraki, Ben Passarelli, John A Coller, Michael J Fero, Harley H McAdams, and Lucy Shapiro. The essential genome of a bacterium. 7:528.
13. Andrew L. Goodman, Meng Wu, and Jeffrey I. Gordon. Identifying microbial fitness determinants by insertion sequencing using genome-wide transposon mutant libraries. 6(12):1969–1980.
14. Kelly M. Wetmore, Morgan N. Price, Robert J. Waters, Jacob S. Lamson, Jennifer He, Cindi A. Hoover, Matthew J. Blow, James Bristow, Gareth Butland, Adam P. Arkin, and Adam Deutschbauer. Rapid quantification of mutant fitness in diverse bacteria by sequencing randomly bar-coded transposons. 6(3):e00306–15.
15. Benjamin E. Rubin, Kelly M. Wetmore, Morgan N. Price, Spencer Diamond, Ryan K. Shultzaberger, Laura C. Lowe, Genevieve Curtin, Adam P. Arkin, Adam Deutschbauer, and Susan S. Golden. The essential gene set of a photosynthetic organism. 112(48):E6634–E6643.
16. Lars Barquist, Christine J. Boinett, and Amy K. Cain. Approaches to querying bacterial genomes with transposon-insertion sequencing. 10(7):1161–1169.
17. Lars Barquist, Gemma C. Langridge, Daniel J. Turner, Minh-Duy Phan, A. Keith Turner, Alex Bateman, Julian Parkhill, John Wain, and Paul P. Gardner. A comparison of dense transposon insertion libraries in the salmonella serovars typhi and typhimurium. page gkt148.
18. Don J. Brenner and Noel R. Krieg. Bergey's Manual of Systematic Bacteriology: Volume Two: The Proteobacteria. Springer Science & Business Media.
19. Sean R. Eddy. Accelerated profile HMM searches. 7(10):e1002195.
20. Lars Barquist, Matthew Mayho, Carla Cummins, Amy K. Cain, Christine J. Boinett, Andrew J. Page, Gemma C. Langridge, Michael A. Quail, Jacqueline A. Keane, and Julian Parkhill. The TraDIS toolkit: sequencing and analysis for dense transposon mutant libraries. 32(7):1109–1111.
21. Satoshi Kimura, Troy P. Hubbard, Brigid M. Davis, and Matthew K. Waldor. The nucleoid binding protein h-NS biases genome-wide transposon insertion landscapes. 7(4):e01351–16.
22. Adrian M. Altenhoff, Brigitte Boeckmann, Salvador Capella-Gutierrez, Daniel A. Dalquen, Todd DeLuca, Kristoffer Forslund, Jaime Huerta-Cepas, Benjamin Linard, Ccile Pereira, Leszek P. Pryszcz, Fabian Schreiber, Alan Sousa da Silva, Damian Szklarczyk, Clment-Marie Train, Peer Bork, Odile Lecompte, Christian von Mering, Ioannis Xenarios, Kimmen Sjlander, Lars Juhl Jensen, Maria J. Martin, Matthieu Muffato, Quest for Orthologs Consortium, Toni Gabaldn, Suzanna E. Lewis, Paul D. Thomas, Erik Sonnhammer, and Christophe Dessimoz. Standardized benchmarking in the quest for orthologs. 13(5):425–430.
23. Fabian Schreiber and Erik L. L. Sonnhammer. Hieranoid: hierarchical orthology inference. 425(11):2072–2081.
24. Jaina Mistry, Robert D. Finn, Sean R. Eddy, Alex Bateman, and Marco Punta. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. 41(12):e121–e121.
25. Alexandros Stamatakis. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. 30(9):1312–1313.
26. Aaron E. Darling, Guillaume Jospin, Eric Lowe, Frederick A. Matsen, Holly M. Bik, and Jonathan A. Eisen. PhyloSift: phylogenetic analysis of genomes and metagenomes. 2:e243.

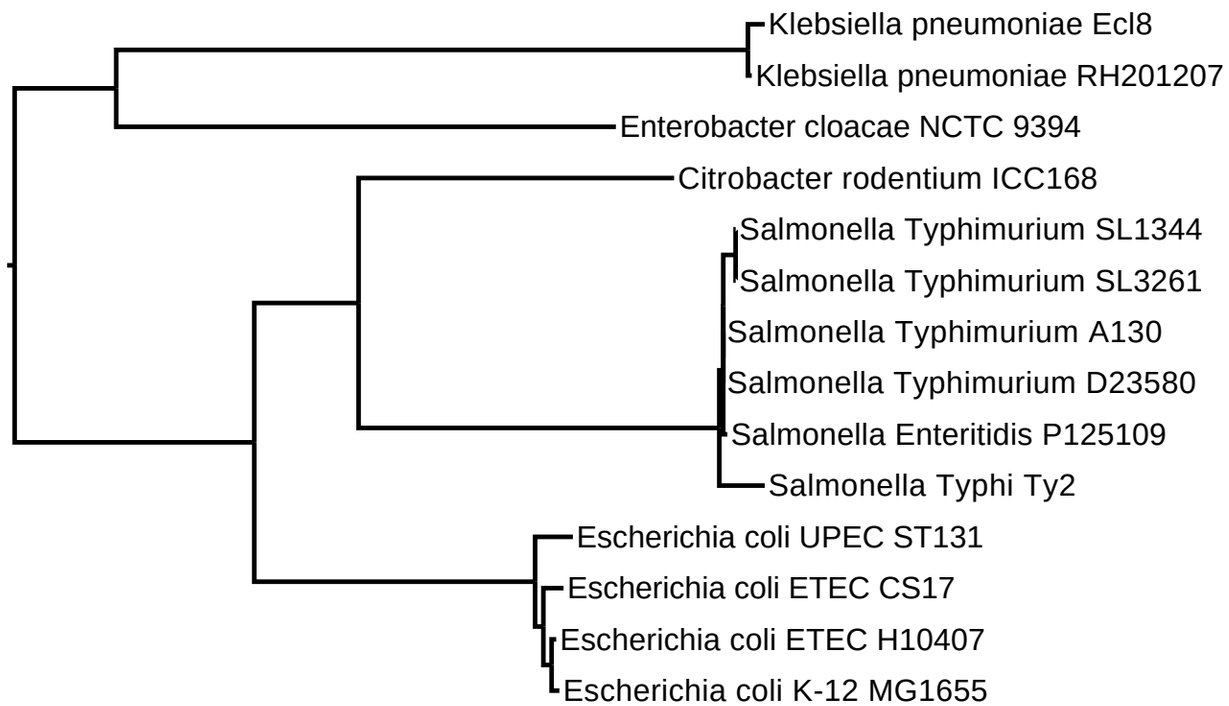

Fig. 1: The species tree containing the 13 strains under study and Escherichia *coli* K-12 MG1655 studied in Keio collection [7]. We have generated the tree by running RAxML [25] on Phylosift [26] amino acid markers.

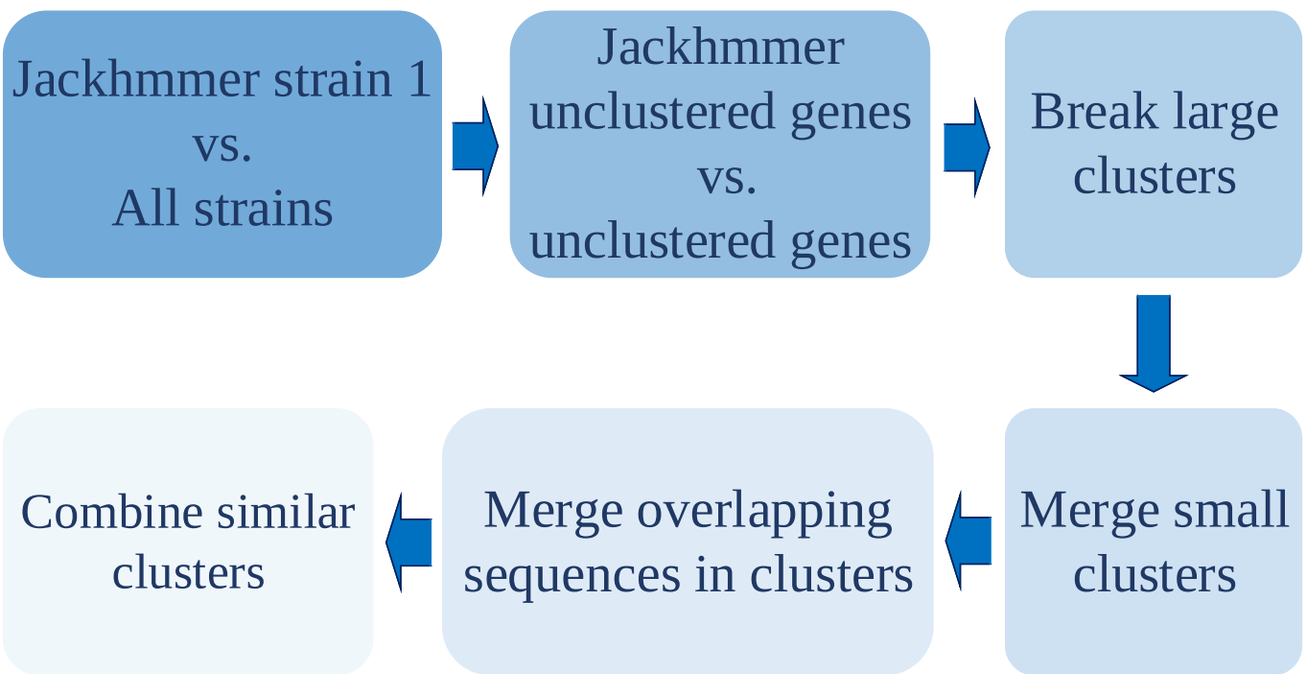

Fig. 2: The steps of our proposed algorithm for clustering homologous genes.

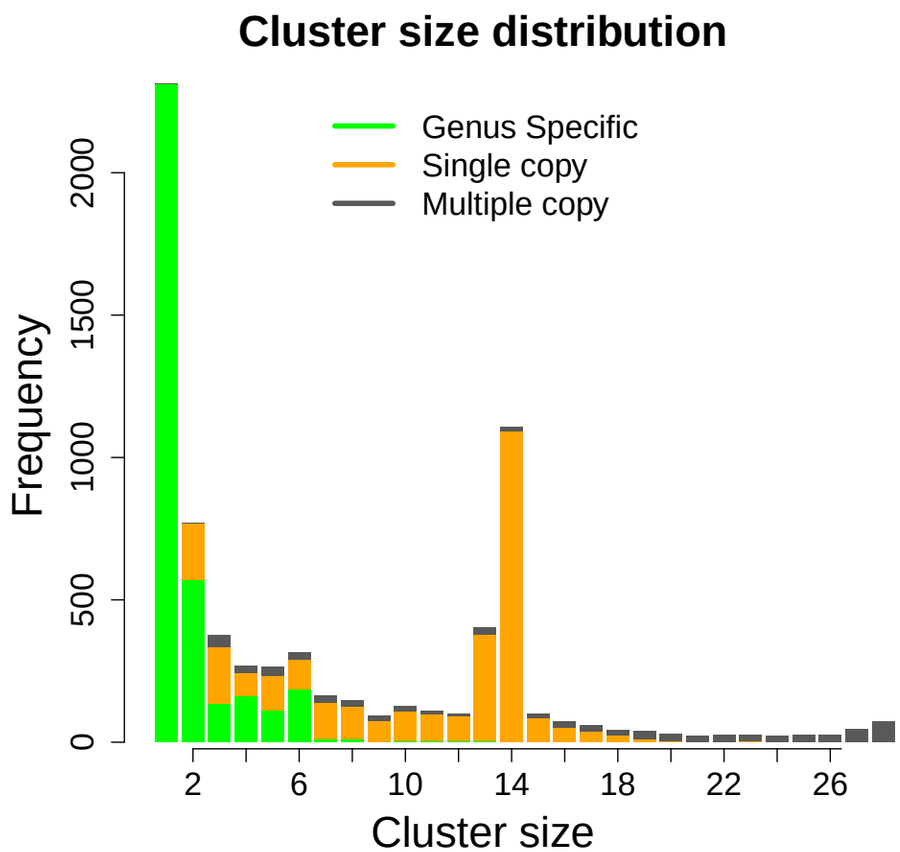

Fig. 3: Size distribution for all clusters of homologous genes. Genus specific genes are genes that are present only in one genus, single copy genes are present in more than one genus and more than 70% of them are not duplicated, and multi-copy genes are present in more than one genus and less than 70% of them are not duplicated.

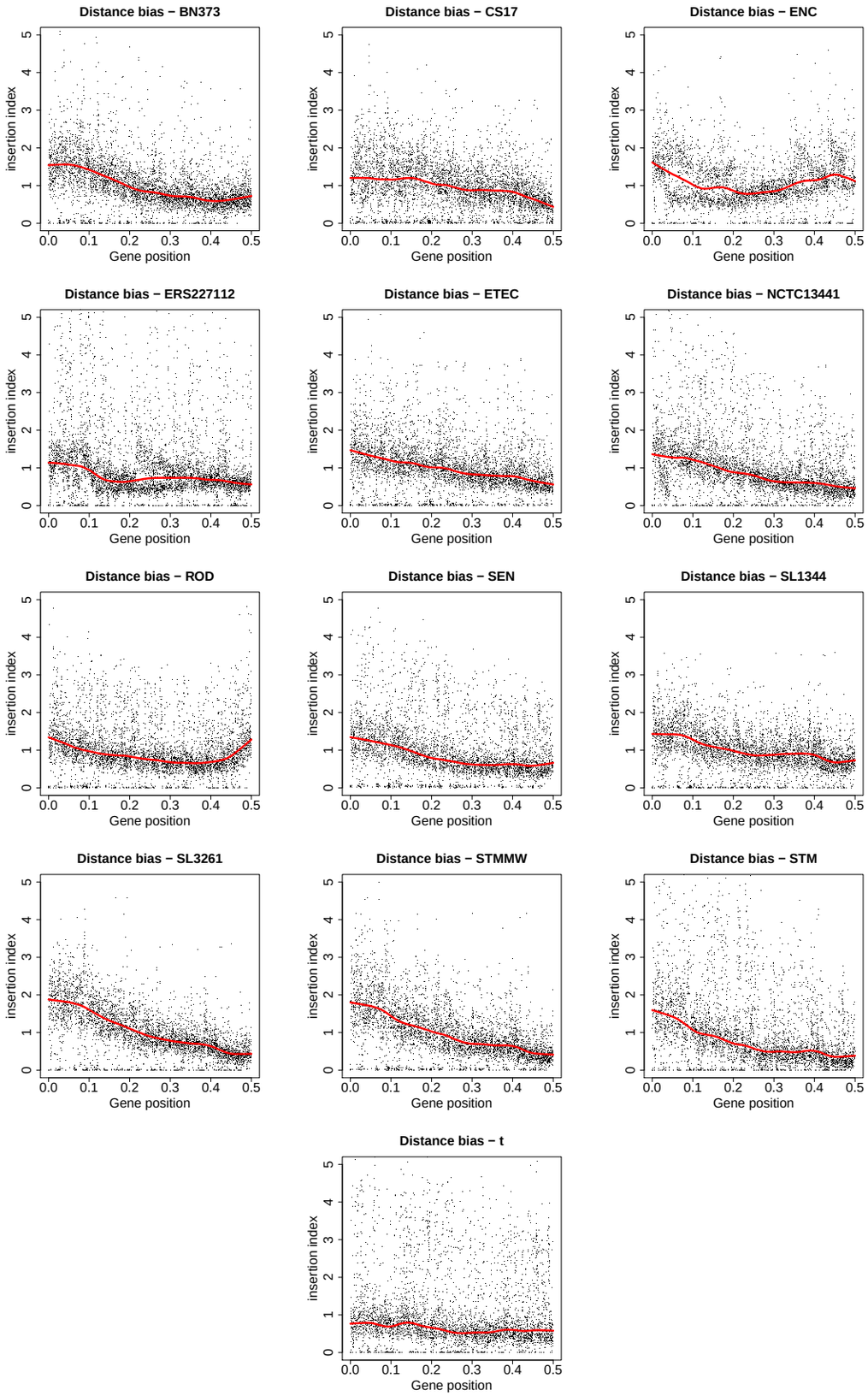

Fig. 4: The plots show the position of the genes within the genome (normalised by the lengths of the genomes) versus the insertion indices of the genes

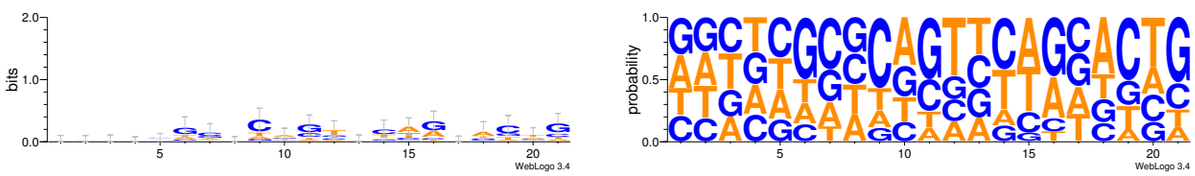

Fig. 5: We have generated the logos from 10 nucleotides flanking the 100 top most frequent insertion sites.

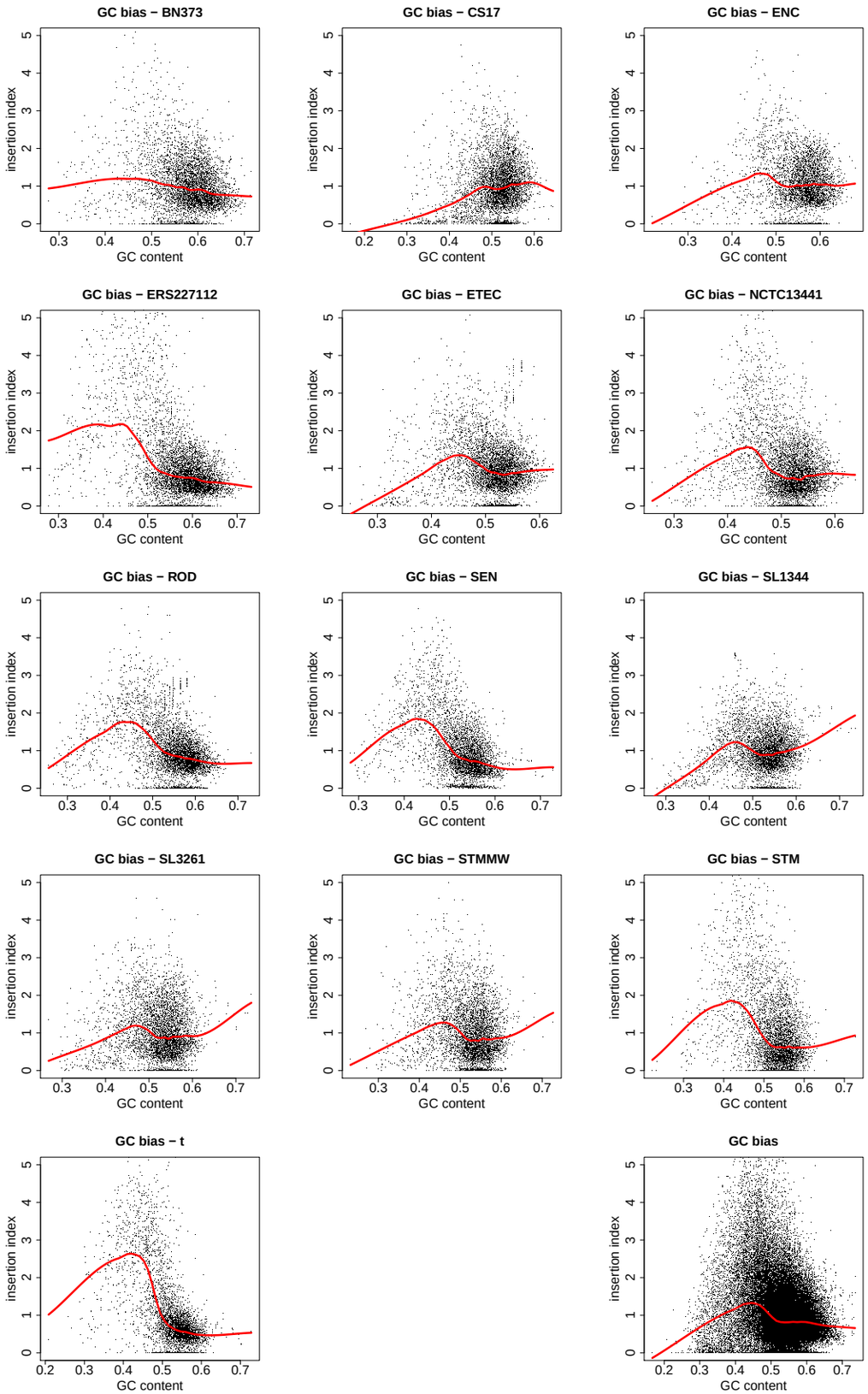

Fig. 6: The plots show the G-C contents of the genes (normalised by the lengths of the genes) against their insertion indices

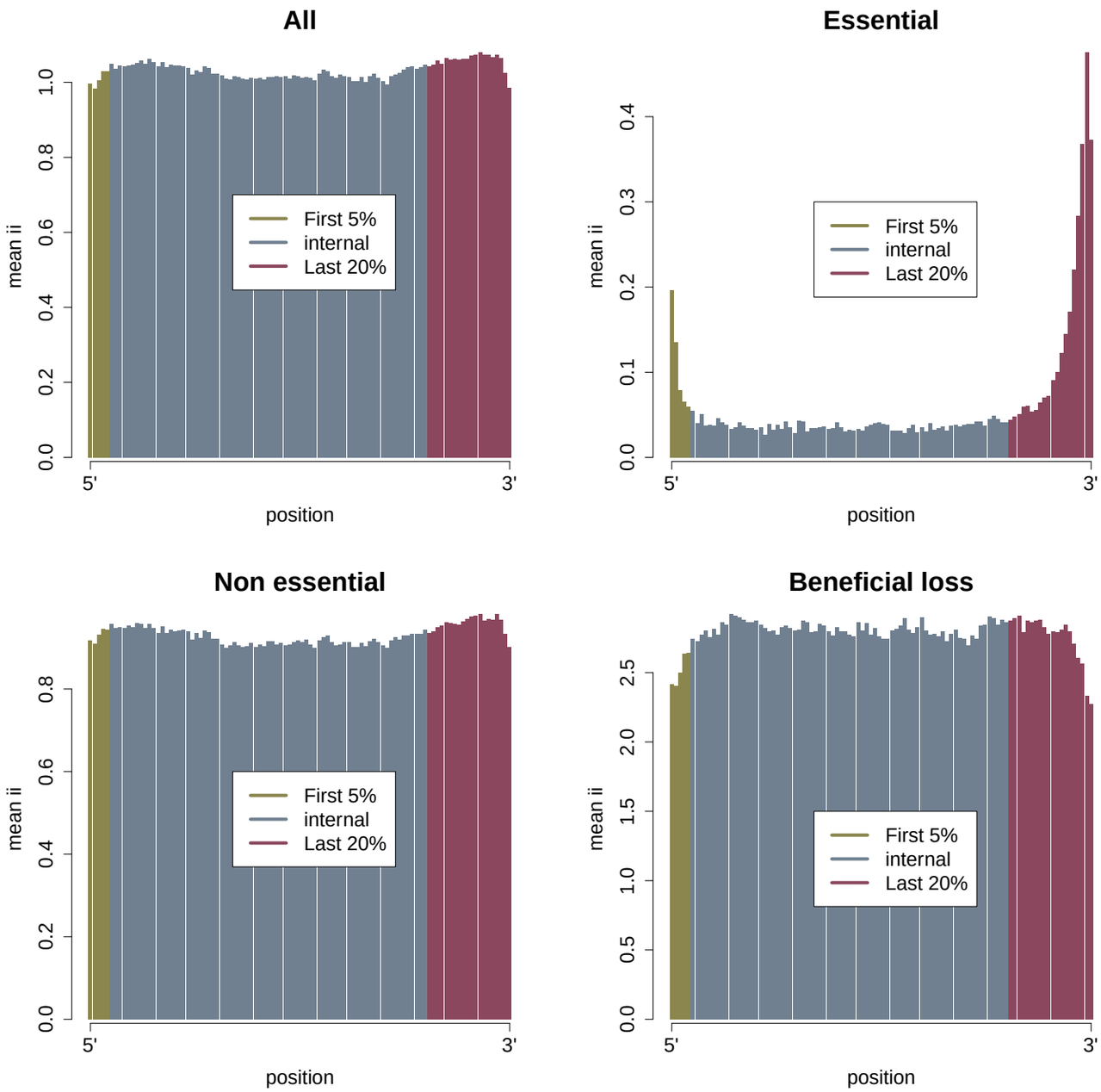

Fig. 7: We have divided our genes into 3 segments: 5% of the genes on the 5' end, 20% of the genes on the 3' end, and the rest in the middle.

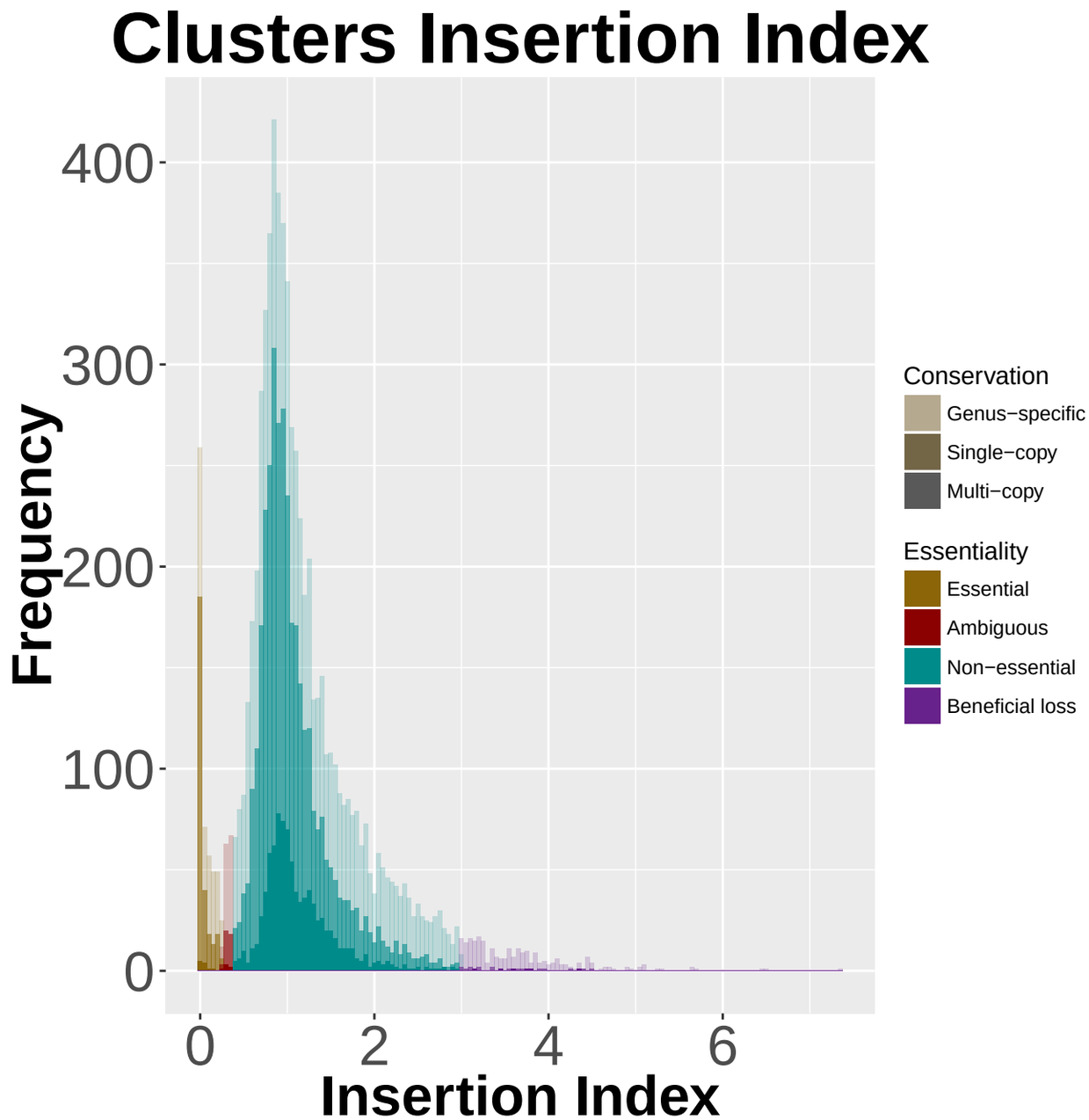

Fig. 8: The genes have been clustered into orthologous groups using Hieranoid and paralogous groups using Jackhmmer and divided into 3 groups: genus specific, single copy, and multi-copy genes. Then, the essentiality of the clusters has been defined using the insertion indices of the genes in the clusters. The figure shows that most of the essential genes are in single copy group, while most of the beneficial losses are genus-specific.

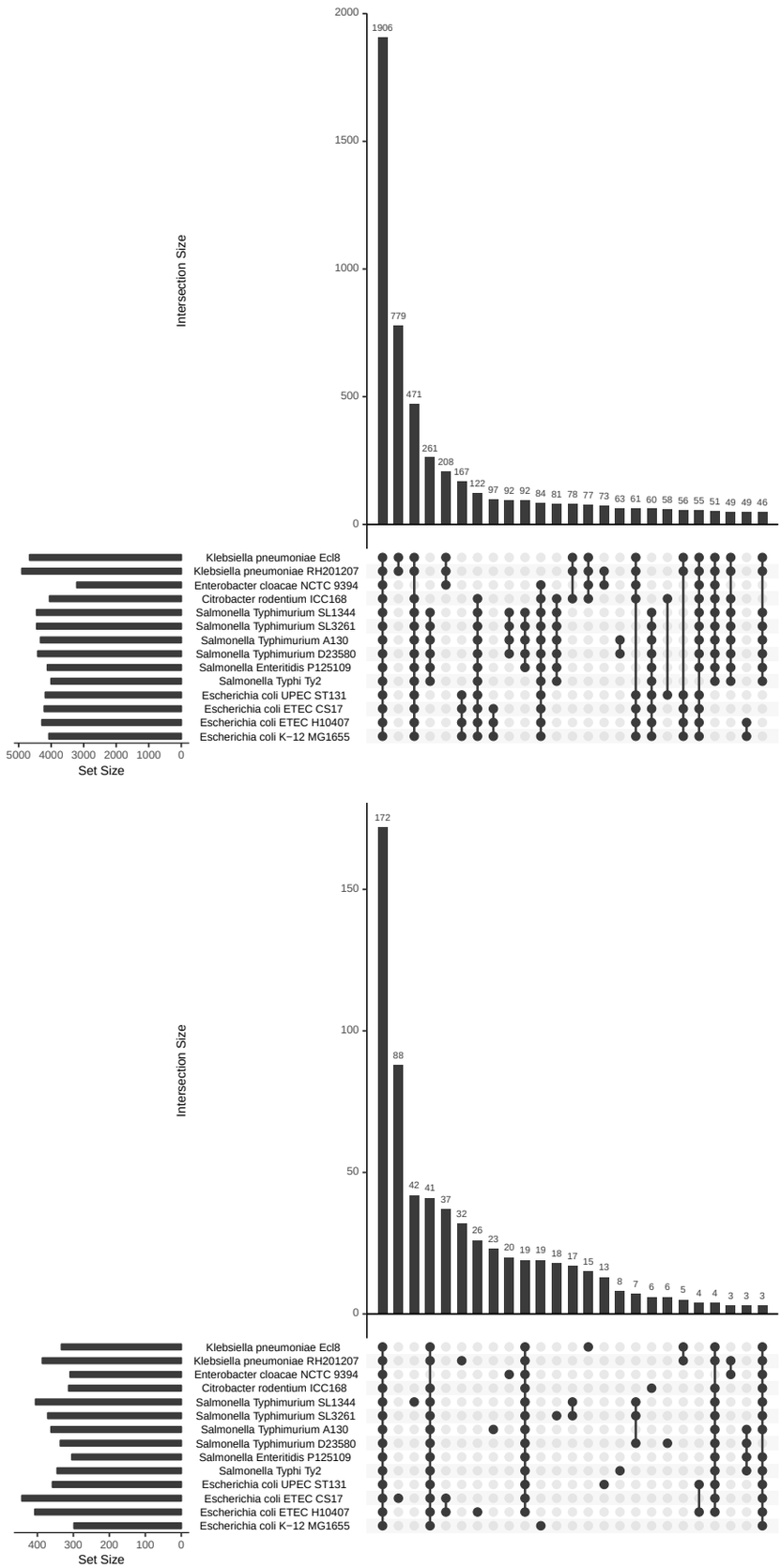

Fig. 9: The first figure shows the number of core genes between each group of species and the second figure shows the number of core essential genes.