Rates of DNA Mutation in Genes and Inte-Gene Regions

Laszlo Csonka, Mark Daniel Ward, Brian French, and Nicole Markley

Department of Biology, Purdue University, West Lafayette, IN

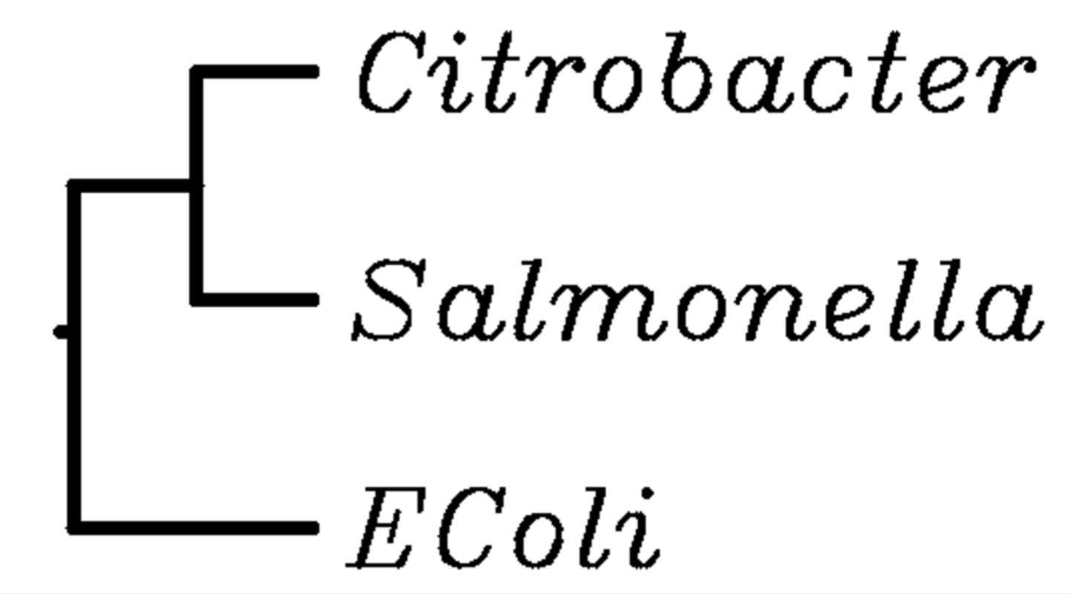
Abstract

The aim of our analysis was to compare the rates of evolution of DNA sequences of genes and intergenic regions. The chromosomes of the related genera of bacteria, Escherichia coli K-12, Salmonella enterica serovar Typhimurium, and Citrobacter koseri, contain extensive regions in which the order of genes of identical. The fact that the gene order is so conserved suggests that both the genes and intergenic spacers were derived by evolution from corresponding sequences in the last common ancestor of the three organisms, and it enabled us to compare the sequence changes in functional genes and in intergenic spacers across the three organisms.

Pairwise comparisons using BLASTN were used to identify regions that had shared genes. ClustalW was then used to identify single base mutations in both genes and inter-gene regions. Mutation frequencies were determined for all paired comparisons. This analysis indicated that the mutation frequencies are significantly higher in intergenic spacers do not have specific sequence-dependent functions, and therefore, they can accumulate mutations more liberally than genes, which are constrained by their function.

Background

Genomes have both useful genetic information coded in genes and presumed useless inter-gene regions termed as gaps. Gene pairs between organisms can be isolated using a tool called **B**asic **L**ocal **A**lignment **S**earch **T**ool, commonly referred to as BLAST. Local alignment can be done with an algorithm called CLUSTALw, and the results from this can be used to isolate base-level changes in sections of DNA.
Salmonella Typhimurium LT2, *Citrobacter Koseri*, and *Escherichia Coli MG1655* were chosen because they are closely related members of *Enterobacteriaceae.*Additionally, the relations of genes can been in the dot plot shown. Barring the region that's inverted, there are large regions of matching genetic information.



The phylogeny tree shows the relative relations of the three bacteria chosen to be analyzed, which is an explanation of why E. Coli and Citrobacter have smaller common regions than Citrobacter and Salmonella.

Motivation

Mutation in genes has a double constrainment of DNA?s built in error-handling as well as maintaining function. Gaps are suspected to have no function need a citation here probably, so they should mutate faster and more randomly. We suspected that choosing pairs of bacteria that were closely related could be used to show rates of mutation in homologous genes and gap regions. By comparing these rates we hoped to show that gaps either are functionless and have random mutations or that they are functional and their mutations have a non-random distribution.

Question

What degree of similarity is there between gene and gap mutation rates, and can this be used to determine if there is any functional constrainment in gaps?

Hypothesis

We hypothesize that there is no preservation of order shown in gap regions, so they should be substantially more random than genes.

Illustrations

Figure 1

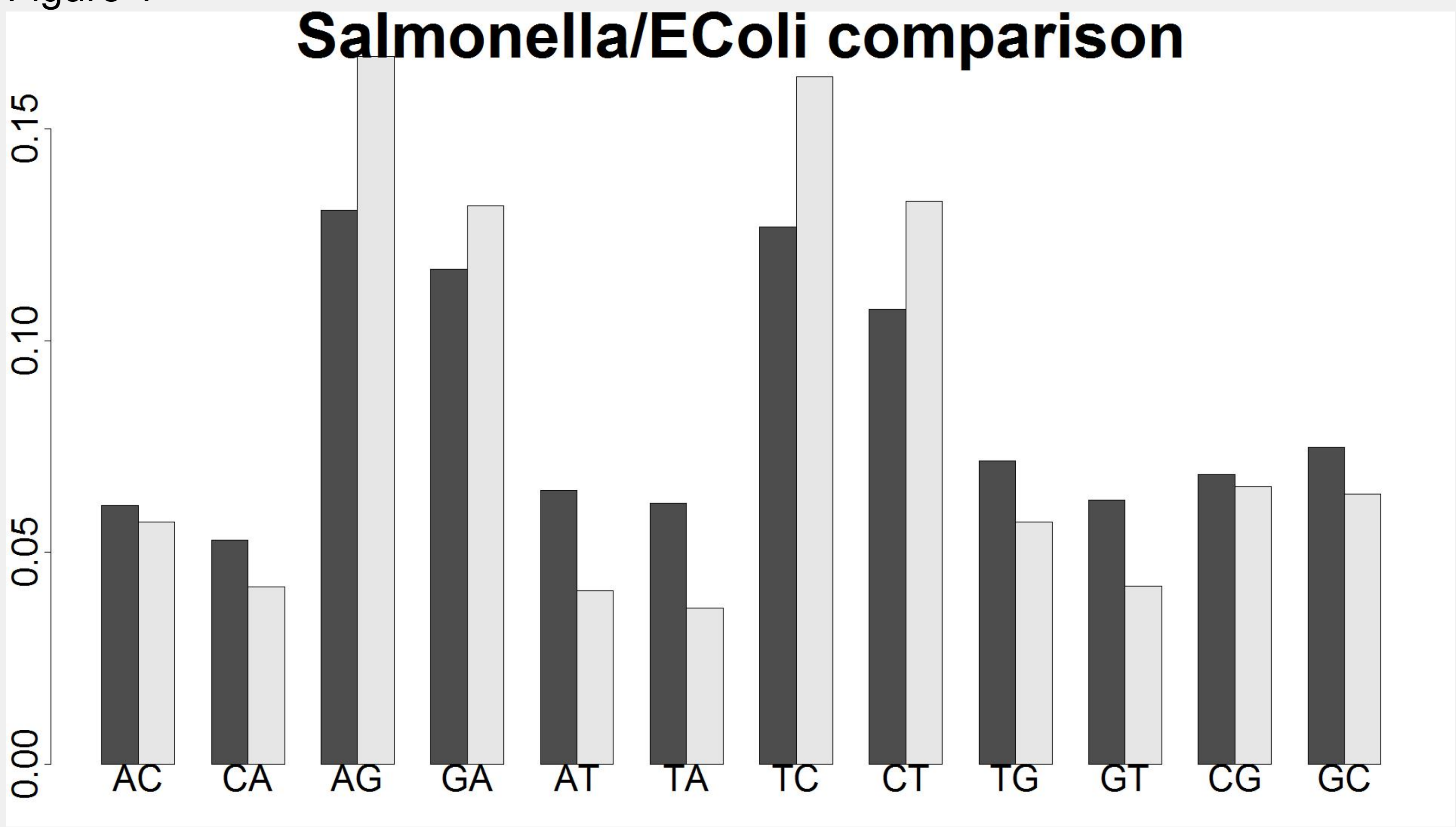


Figure 2

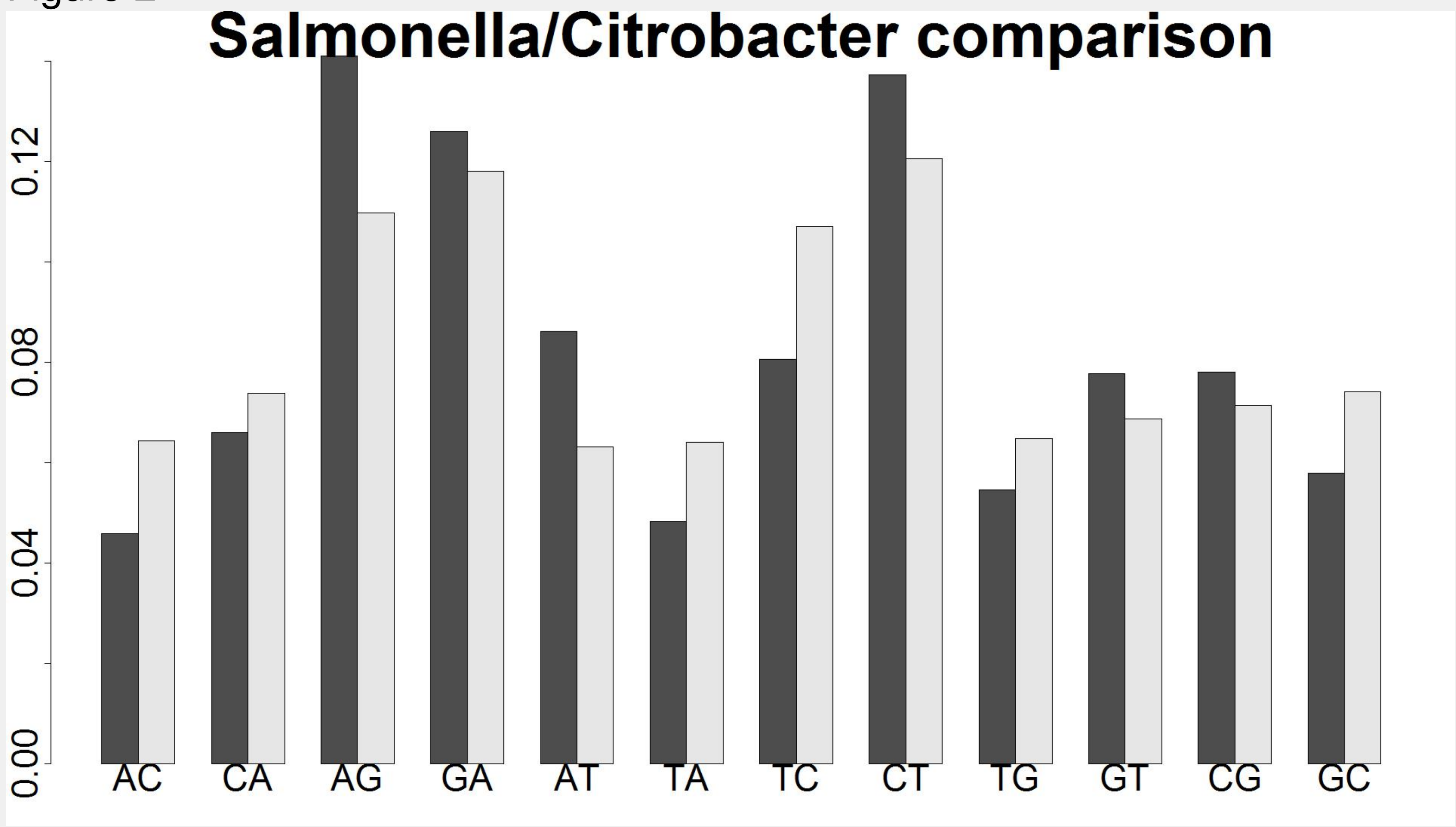
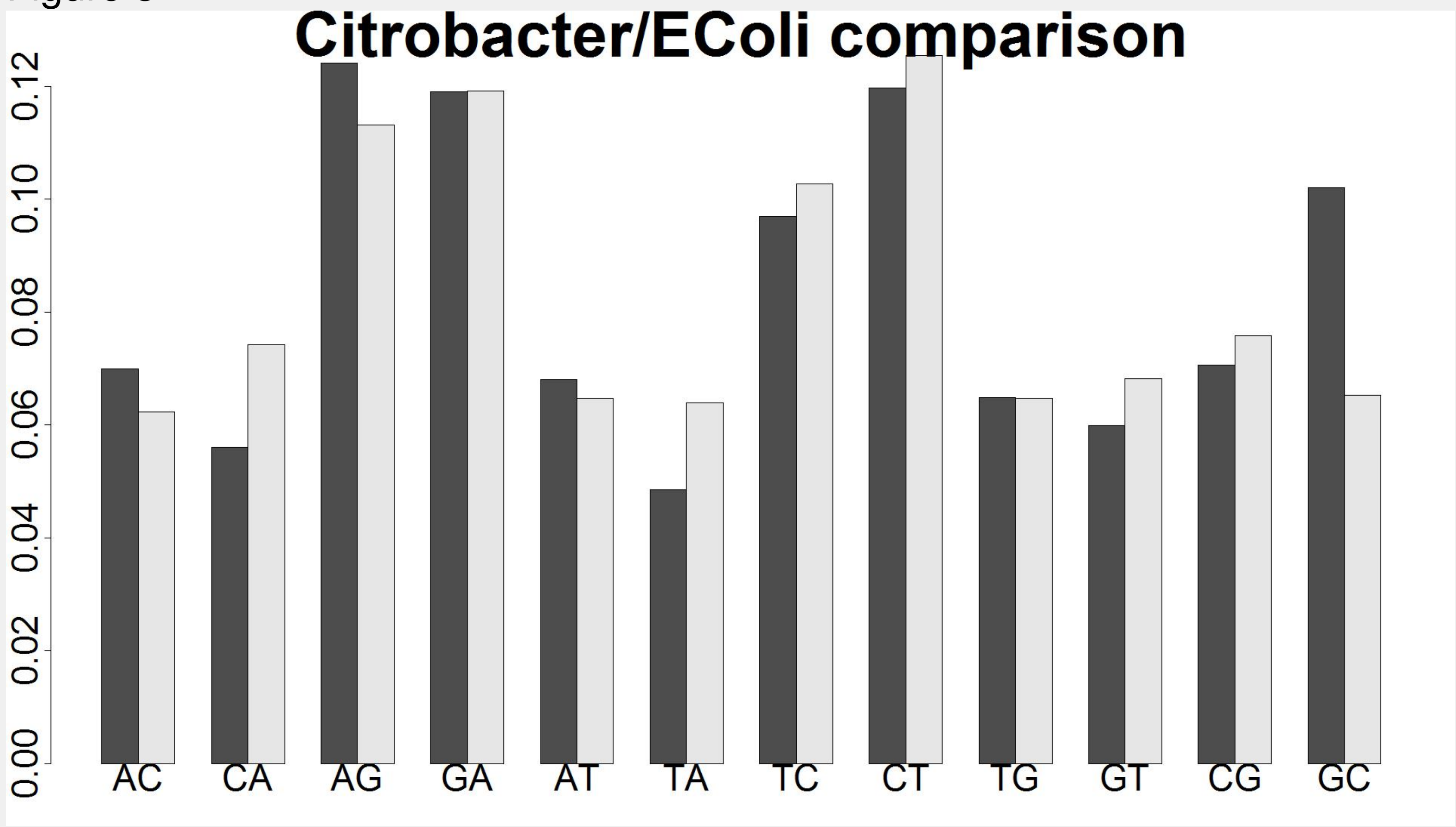


Figure 3



Procedure

Genomes were run through BLAST to identify matching regions for each pair comparison. BLAST matches were presumed to be homologous genes. Two matching regions with a non-matching region between them led the non-matching region to be labeled as a gap. Any gap longer than 500 bases was presumed to contain a gene that wasn't present in the compared genome and was therefore discarded.
Matched regions (both genes and gaps) were run through CLUSTALw and any base mismatch preceded and followed by at least 3 matching bases was presumed to be a mutation. Mutation rates were then tabulated, plotted, and chi-squared tested on mutation counts.

Conclusion

Looking at the three figures off to the left, we have depictions of the three pairwise comparisons. The plots show the relative frequency of mutations in genes vs gaps. Gaps are shown in black, genes are in white. What immediately stands out is that the A to G/G to A and T to C/C to T mutation is clearly more frequent. Likely, this is because A and G are the two purine bases and T and C are the two pyrimidines.
Running statistical tests across the comparison, the gene and gap mutations are clearly from different distributions. Comparing the distribution of gene and gap mutations for Salmonella and E Coli with insertions and deletions left in, a chi-squared test returns a probability that they came from the same distribution of less than 2.2x10≐16. When insertions and deletions are ignored, the likelihood that gene and gap mutations are from the same distribution rises to 1.766x10≐13. Similar numbers arise from chi-squared testing the other two pairwise comparisons.
This is conclusive evidence that genes and gaps have differing distributions for mutation frequencies. From this, we can conclude that there is some mechanism limiting mutations in genes. As mentioned in the background section, this is likely due to genes having to maintain function despite mutation, while gaps can mutate freely.

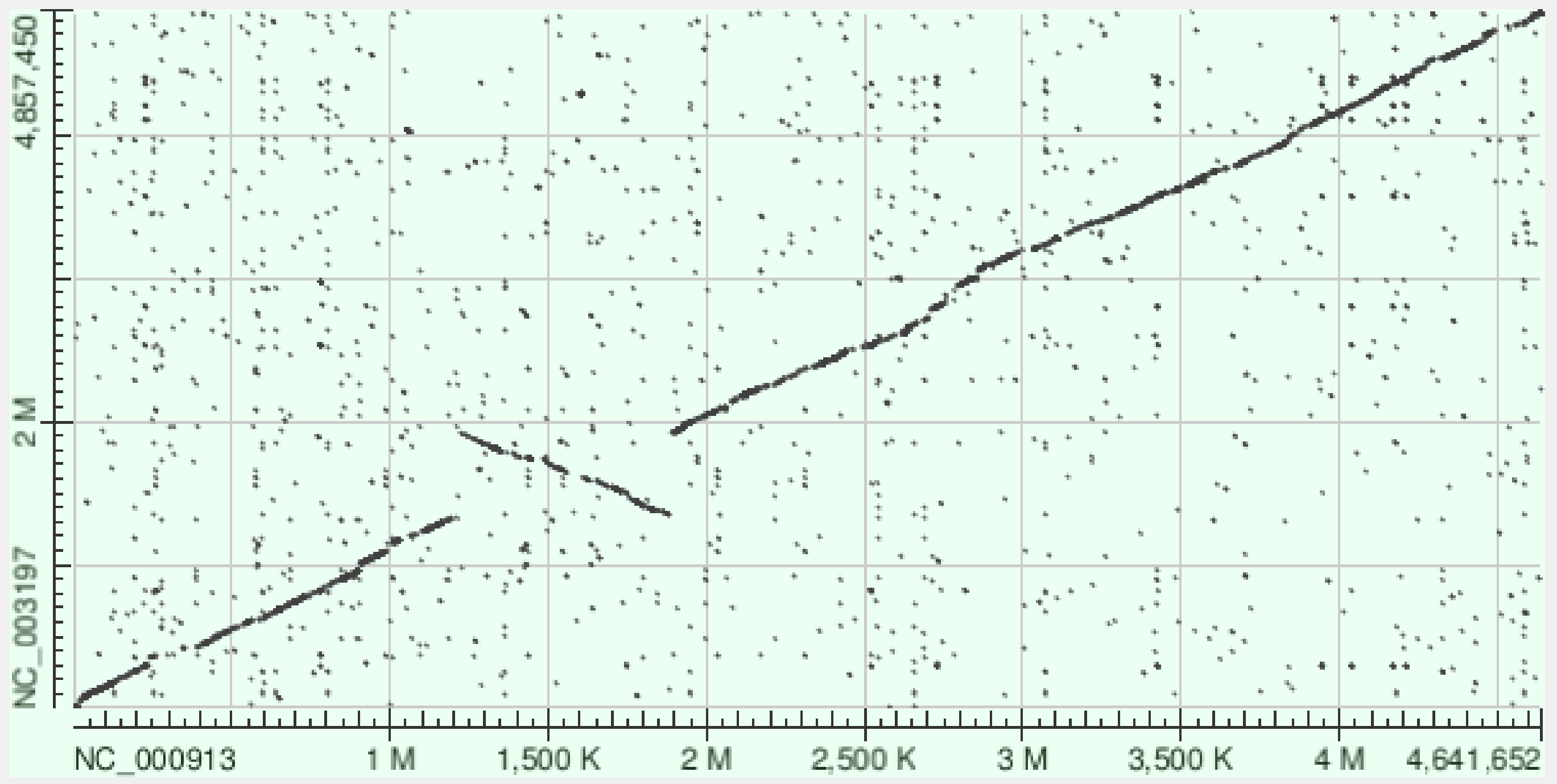
Acknowledgments

This research is supported by NSF grant DMS #1246818.

Citations

Thompson, J. D., Gibson, T. J. and Higgins, D. G. 2002. Multiple Sequence Alignment Using ClustalW and ClustalX. Current Protocols in Bioinformatics. 00:2.3:2.3.1?2.3.22.
Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, David J. Lipman, Basic local alignment search tool, Journal of Molecular Biology, Volume 215, Issue 3, 1990, Pages 403-410, ISSN 0022-2836, http://dx.doi.org/10.1016/S0022-2836(05)80360-2. (http://www.sciencedirect.com/science/article/pii/S0022283605803602) BLAST accessed through https://blast.ncbi.nlm.nih.gov/Blast.cgi

Dot Plot Example


2 M
NC_003197
NC_000913
1 M
1,500 K
2 M
2,500 K
3 M
3,500 K
4 M