AT homopolymer strings as a motif for self-recognition and repair of genomes in *Salmonella enterica* subspecies I

Jean Guard1,*, Adam R Rivers 2, Justin N Vaughn1, Michael J Rothrock, Jr1, Adelumola Oladeinde1, and Devendra H Shah3

1 Affiliation 1; US Department of Agriculture, US National Poultry Research Center, Athens, GA 30605
2 Affiliation 2; US Department of Agriculture, Genomics and Bioinformatics Research, Gainesville, GA 32608
3 Department of Veterinary Microbiology and Pathology, Washington State University, Pullman, WA 99164
* Correspondence: jean.guard@usda.gov; Tel.: 1+706-546-3446

**Abstract:** Adenine and thymine homopolymer strings of at least 8 nucleotides (AT 8+mers) were characterized in *Salmonella enterica* subspecies I and other Eubacteria. Incidence of the motif differed between Eubacteria but not between *Salmonella enterica* serotypes. Of 481 AT 8+mers loci in serovars Typhimurium, Enteritidis, and Gallinarum, 35 (12.3%) had mutations. We propose that the AT 8+mer motif identifies genomes with optimal gene content and provides self-recognition that facilitates efficient genome repair. A theory that genome regeneration accounts for both serovar diversity and persistence of predominant *Salmonella* serovars associated provides a new framework for investigating root causes of foodborne illness.

**Keywords:** keyword 1; *Salmonella enterica* 2; food safety 3; genome 4; theory 5; single nucleotide polymorphisms 6; recombination 7; serotype

1. Introduction

Approximately 30 of 1500 *Salmonella enterica* subspecies I (*S. enterica*) serovars have been persistent agents of foodborne illness in people for the past several decades [1]. Despite improved biosecurity throughout the food production pipeline, reduction of salmonellosis has plateaued over the past two decades [2]. The inability to reduce salmonellosis indicates new approaches to understanding the biology of this important pathogen are needed. Recently, the most commonly occurring single nucleotide polymorphism (SNP) that caused disruption of a gene in *S. enterica* serovar Enteritidis (Enteritidis) was identified, and it was deletion of a single adenine in a homopolymer string of 8 nucleotides (nt) within the fimbrial gene *sefD* [3]. Mutational analysis, phenotype microarray, and infection experiments in the egg-laying hen indicated that the *sefD* mutation increased organ invasion and mortality in hens, disturbed egg production, enhanced growth of the pathogen to high cell density, and otherwise behaved as a regulator of dimorphism of phenotype [4]. The impact of the discovery was that the performance of a killed vaccine for hens was enhanced by increasing SefD in preparations [5]. The drastic change in biological phenotype imparted by the single base pair deletion suggested that characterization of purine homopolymer strings of adenine, AAAAAAAA, and its pyrimidine base pair (bp) of thymine, TTTTTTTT, in *S. enterica* and other eubacteria should be explored.

A homopolymer of adenine:thymine (AT) with 8 nucleotides or more is abbreviated in this manuscript as an AT 8+mer. It is a DNA motif suggested by conformational studies to bend DNA out of the Z-conformation [6]. Polyadenine regions can impact gene regulation in prokaryotes and can contribute to microsatellite instability in eukaryotes [7-10]. Evidence exists to show that homopolymer nucleotide strings contribute to non-programmed slipped strand replication and the accumulation of errors in DNA [11-13].
Thus, the physicochemical impact of these strings was another reason to catalogue this motif in the genome of *S. enterica*.

To evaluate AT 8+mers in *S. enterica* subspecies, several serovars of *S. enterica* and other bacterial genera were compared for both AT 8+mers and GC 8+mers. *S. enterica* serovars Enteritidis and Typhimurium were analyzed because they are two of the three most common causes of foodborne salmonellosis in the US and abroad. They are from different genomic lineages and have been extensively studied and sequenced. Together they have caused approximately 40% of all foodborne salmonellosis in the US [1]. *S. enterica* Gallinarum was included because it is another poultry-associated pathogen that shares a genomic lineage with Enteritidis. However, its biological impact is different from that of *S. Enteritidis*. It does not cause human salmonellosis; instead, it causes devastating disease in poultry resulting in high morbidity, mortality, and economic loss [14]. Comparing Typhimurium and Enteritidis, which have different genomic lineages yet cause foodborne illness, to Gallinarum, which is genetically related to Enteritidis but has a drastically different epidemiology to both, is a comparative approach used before to link single nucleotide polymorphisms to phenotype [15]. In this study the three genomes were compared to better understand the content of AT 8+mer homopolymer nucleotide strings in *S. enterica*, and the association the motif might have with naturally occurring mutation that disrupts open reading frames of genes.

**Additional background on select *S. enterica* serotypes**

*S. enterica* serovars Enteritidis and Typhimurium differ biologically although both are predominant causes of foodborne illness. One way they differ is in immunological properties of the cell surface. Serovar Typhimurium is a serovar Group B organism, with an antigenic formula of _1,4,[5],12:i:1,2_ [16]. Serovar Enteritidis is a Group D organism, with an antigenic formula of _1,9,12:g,m:-_, thus it is mono-flagellated [16].

Epidemiological patterns for the two predominant pathogens also differ. Enteritidis is an exceptional *Salmonella* pathogen in part because it efficiently contaminates the internal contents of eggs produced by otherwise healthy-appearing hens. It produces a high molecular mass (HMM) O-antigen, which not only protects killing of the pathogen by the host complement system, but also acts as a protective capsule in the hostile environment of the egg [17-19]. Typhimurium is also resistant to complement, but it does not produce HMM O-antigen and, thus, does not survive in the internal contents of eggs to an extent that can be detected by epidemiological surveillance. Both Typhimurium and Enteritidis can contaminate a broad spectrum of other food sources such as the eggshell, the poultry gastrointestinal tract, poultry carcasses, and fresh vegetables. Both serovars can invade organs and survive in macrophages, which contributes to systemic spread during infection [20]. Variation between strains within each serovar occurs but serotype characteristics and general genome organization is maintained [21, 22]. There are serovar-specific patterns in plasmid carriage and fimbrial genes. Comprehensive reviews of the similarities and differences between *Salmonella* serovars are available [23-27].

*S. enterica* serovars Gallinarum and Enteritidis are genetically closely related [28]. Gallinarum’s antigenic formula is _1,9,12:-:-_, which indicates it has the same lipopolysaccharide O-antigen epitopes as Enteritidis; however, it lacks both H1 and H2 flagellin proteins and is thus non-motile. Both Gallinarum and Enteritidis can contaminate the internal contents of eggs; however, Gallinarum has mutations and rearrangements throughout its chromosome that restrict its host range to the avian host, possibly by reducing immunological response to infection and thus facilitating systemic infection [20]. Thus, the most striking differences between the foodborne pathogen and Gallinarum is that the latter makes poultry extremely sick, reduces egg production and causes high mortality. In contrast hens infected with Enteritidis often appear healthy, remain in production,
and thus eggs become contaminated internally and are a source of foodborne illness. The ability of Enteritidis to spread through flocks that appear healthy was one of the contributing factors in its world-wide spread through the layer industry. The differences in the epidemiology, association with food, and virulence characteristics of the three pathogens, all of which occur in the poultry environment, suggested that comparative analysis of S. enterica serovars Typhimurium, Enteritidis, and Gallinarum would help set a baseline for the association between the AT 8+mer motif and naturally occurring mutation of an important food borne pathogen. Other pathogenic Salmonellae and other Eubacteria were also included in analysis.

2. Materials and Methods

2.1 Genomes of Eubacteria analyzed for strings of homopolymers

The database of 1,434 complete genomes of S. enterica subspecies I (taxid:59201), as well as other Eubacteria listed in Table S1, was used as source material as available from the National Center for Biotechnology Information (NCBI) [29]. The last accession date was April 30, 2020. S. enterica serovar Typhimurium LT2 (NC_003197.2) was used as the primary reference sequence to name genes and gene functions, and it was used to order genes [30]. Two other references were S. enterica serovar Enteritidis strain P125109 and S. enterica serovar Gallinarum strain 9184, with respective NCBI accession numbers of NC_011294.1 and CP019035.1 [31, 32]. S. enterica serovars Typhimurium, Typhi, and Enteritidis genomes were over-represented compared to other serovars, and together they comprised 39.4% of all completed genomes available. Only 51.2% of S. enterica subspecies I genomes had a complete adenylate cyclase (cyaA) gene, which is required for virulence as a foodborne pathogen. The other sequences were plasmids, which were not under review in this study. Genome CP018657 is classified as serovar Enteritidis, but all analyses suggest it is serovar Typhimurium; thus, it is excluded from analyses. A broader examination of AT and GC 8+mer homopolymers included Escherichia coli, Proteus mirabilis, Shigella sonnei, Yersinia pseudotuberculosis, Vibrio vulnificus (chromosome I and II), Staphylococcus aureus, Streptococcus pyogenes, Enterococcus faecalis, Bacillus anthracis and Bacillus cereus. Genome databases at NCBI show homopolymer strings, as well as other combinations of low-complexity regions, in lower-case gray font because there is recognition that some sequence strings might be susceptible to alignment error and thus require masking during the alignment process. For the BLAST searches conducted here, each gene was observed for high fidelity of surrounding regions, therefore it is unlikely low complexity impacted observed alignments.

2.2 Incidence and location of homopolymer nucleotide strings

Counting of kmers, locating kmers within genomes, and determining impact on open reading frames within annotated genes was done with Genious Prime 2020.0.3 (Biomatters, Inc., San Diego, California, USA). Homopolymer strings of all 4 nucleotides, ranging from 5 – 20 nucleotides, were catalogued in several S. enterica serovars and other genera (TABLE 1). For S. enterica subspecies I grouped by serovar, at least 12 complete genomes were assessed. For other genera, at least 3 complete genomes were assessed. Averages and standard deviations were calculated. Ttest analysis was used to determine if differences between groups were significant at p < 0.01. Other types of data processing were that the genome of interest was stored in SeqBuilder Pro, Lasergene V16.0.352 (DNASTAR, Madison, Wisconsin, USA) and in Geneious format. Strings of homopolymers of different lengths were entered as windows of text and the genomes were searched. Results were copied into an Excel “.csv” file as Unicode text (Microsoft Excel for Mac, V16.16.20 (200307). The text to column feature, and appropriate delimiters, were used to produce columns of data to calculate distance between nucleotide
strings. The average, standard deviation, and median values between AT 8+mer homopolymers were then calculated.

2.3 Determination of a common denominator for comparison of genomes from bacteria of different genera

S. enterica subspecies I serovar Typhimurium LT2 was the reference genome used to produce a common denominator to normalize genomes of different sizes. Every AT 8+mer for Typhimurium LT2 was tabulated and classified as intergenic, intragenic, or regulatory using Genious Prime 2020.0.3 (Table S2). The same software was used to generate a map of AT kmers within the circular chromosome. Another approach used to establish a baseline incidence of AT 8+mers occurring in genes was to generate a list of random numbers using the 4,600 predicted genes of the reference genome. Two hundred random numbers were generated between 1–4600 corresponding to numbered genes, a FASTA file was compiled, and the number of AT 8+mers within the randomly generated sets was determined.

2.4 Comparison of AT 8+mers in 3 S. enterica serovars that vary in epidemiological parameter

A FASTA file was generated from the list of genes and regulatory regions having AT 8+mers from Typhimurium LT2. The reference genome FASTA file was used for BLAST searches against the other two serovars for detailed analysis, namely S. enterica Enteritidis P125109 and Gallinarum 9184. Each genome was sequentially processed for AT 8+mers as it appeared on either strand in either direction, using Geneious Prime 2020.0.3 functions. Differences occurring within AT 8+mers for the 3 genomes were tabulated. Other manipulations of genes used data available at NCBI or were further analyzed with Lasergene V16.0.0 (352) (DNASTAR, Madison, Wisconsin, USA).

3. Results

3.1 The AT 8+mer motif in Eubacteria is specific to Genus and species

Table 1 lists all genera evaluated for the motif. First, S. enterica subspecies I serovars were collated to include 12 different complete genomes for serovars Typhimurium, Enteritidis, and Typhi. A fourth S. enterica group included 12 foodborne Salmonellae of mixed serovars associated with poultry and/or foodborne disease, only strains with complete genomes were analyzed because gaps associated with draft genomes would impact results. Table 1 also lists results from analysis of 3 strains each from a variety of Eubacteria genera; in addition, the outlier group for S. enterica was 12 strains of Escherichia coli (E. coli). Values greater than 1 indicated that more than the expected number of motifs were observed in comparison to S. enterica after normalizing for the size of the genome, and less than 1 indicates fewer motifs were observed than expected.

Results of comparisons between Eubacteria were as follows: 1) AT 8+mers in S. enterica groups were significantly more frequent than what was observed for E. coli (p < 0.005); 2) The range in results was a minimum of 90.0 AT 8+mers for Vibrio vulnificus cII to a maximum of 712.7 for Proteus mirabilis; 3) Standard deviations between strains in each Genera ranged from 2.3 for Yersinia pseudotuberculosis to 84.1 for Enterococcus faecalis; 4) All the genera examined, including S. enterica and E. coli, had a relative paucity of GC 8+mers as compared to AT 8+mers; thus, it appears there is a bias for Eubacteria maintaining AT 8+mers in genomes, or inversely, selecting against GC 8+mers; 5) Each genus appeared distinctly different from others; thus, conservation of AT 8+mers appears to be species specific; 6) Vibrio vulnificus had 180 and 90 AT 8+mers in chromosomes cI and cII respectively; thus, AT 8+mer content might be a chromosomal characteristic that maintains the organization of chromosomes.
Genomes varied widely in size across the Eubacteria, and a common denominator was needed to normalize data. To produce a common denominator, the reference genome of serovar Typhimurium LT2 was mapped for the location of all AT 8+mers. On average the motif occurred every 16,634nt (Table S2). The AT 8+mers appeared to be dispersed throughout the entire genome of serovar Typhimurium LT2 (Figure 2). The range of AT 8+mer distance was 11 to 117,141nt, and the median was 11,578nt (Table S2). Distances of 52,048nt or greater between motifs were over 3 standard deviations and were thus possibly deficient in AT 8+mers. Of 13 putatively deficient regions, the 4 longest regions were assessed for phage genes, pseudo genes, insertion elements, transposases, ribosome binding sites and regulons. The 4 regions were located between nucleotides i) 1368633-1444823 (76,198nt), ii) 2612956-2730097 (117,148nt), iii) 4124625-4209022 (84,404nt), and iv) 4342879-4418289 (75,418). At this time, no feature could be found that differentiated AT 8+mer deficient regions from regions with shorter distances between AT 8+mers.

3.2 The AT 8+mer motif in *Salmonella enterica* is not specific to serotype

The genome of reference strain *S. enterica* serovar Typhimurium LT2 is 52.2% GC. When data were expressed as ratios of AT:GC homopolymer strings, the AT 8mer homopolymers (e.g. AAAAAAAA and TTTTTTTT) were much more prevalent than GC 8mers in the reference genome (Figure 1). In total there were 294 AT 8mers and 11 GC 8mers in the reference serovar, which is a ratio of 27 AT 8mers to every GC 8mer. AT strings longer than 8bp were less frequently observed (Figure 1). To account for every AT kmer of at least 8 nucleotides, the longer motifs were added to 8mers in further analyses; thus, the term AT 8+mer is applied throughout to describe the motif. As was referenced in the introduction, the length of the homopolymer impacts the physicochemical bending properties of DNA and thus we wanted to account for every kmer of 8 nucleotides or more.

Results from analysis of AT 8+mers between *S. enterica* serotypes were: i) The incidence of AT 8+mers in the reference genome for serovar Typhimurium LT2 was the lowest of the 12 strains in the group, which suggests that using the serovar as a reference would not over-estimate the incidence of AT 8+mers for *S. enterica* or other genera; ii) The range of AT 8+mers per *S. enterica* grouping in Table 1 was from 315.6 to 332.6, and the average was 322.2 +/- 12.83 AT 8+mers; iii) The standard deviations for AT 8+mers in serovar Typhimurium and in the group of mixed serovars were, respectively, 13.0 and 13.9; iv) Serovars Enteritidis and Typhi, with respective standard deviations of 10.5 and 5.9, appeared more clonal than Typhimurium, which agrees with current knowledge; v) the foodborne serovars, namely Typhimurium, Enteritidis, and the group of mixed serovars, had a more variable motif content than host restricted Typhi. Overall, the *S. enterica* serovar groups were not significantly different from each other. There were not enough completed genomes of the host-restricted serovar Gallinarum to include it in analysis.

3.3 The AT 8+mer motif in poultry-associated serovars of *Salmonella*

Table 2 lists all genes and regulatory regions with at least one AT 8+mer in serovars Typhimurium, Enteritidis, and/or Gallinarum. Genes were listed in the order in which they appeared in the reference genome for Typhimurium LT2 (NC_003197.2). Some genes in serovars Enteritidis and/or Gallinarum did not have homologs in the Typhimurium reference strain, and vice versa. Six categories of genes were listed, and a total of 175 genes and 13 regulons were included. The number of pseudogenes found with the motif for Typhimurium, Gallinarum, and Enteritidis were 3, 22, and 5, respectively, and each genome had a total of 40, 287, and 96 pseudogenes each. Overall, 7.5%, 8.4%, and 5.2% of genes of pseudogenes had the motif, respectively. In total 30 pseudogenes out of 481 loci (6.2%) were identified as having the motif. For the 188 total genes and regulatory sites listed, 4.2% of genes had AT 8+mers for a *S. enterica* genome with an average of 4517 genes.

3.4. Figures, Tables and Schemes
Figure 1. Ratios of AT homopolymers from 5 to 10 nucleotides in *Salmonella enterica* serovar Typhimurium LT2 NC_003197.2. The ratio of AT homopolymer kmers, either adenine or thymine but not mixed, to GC homopolymers was determined using Geneious software as described in text. The range in number of nucleotides per kmer searched was 5 to 10 (see legend label). Results show that a nucleotide motif of 8 was the most common encountered, and that approximately 27 AT homopolymers were found for every 1 GC AT homopolymer in the reference sequence of *S. enterica* LT2 NC_003197.2.

Figure 2. Locations of AT 8+mers in the genome of *Salmonella enterica* serovar Typhimurium LT2 NC_003197.2.

Table 1. Expected versus observed occurrence of homopolymer strings of 8 and more nucleotides in genomes of Eubacteria
| Genus species  | Other genomic information | Number of genomes analyzed | Genome size (bp) | Common denominator (nt) | Expected number of SNVs | Observed AT % | Observed GC % | Observed AT vs expected AT % | Observed GC vs expected GC % |
|---------------|---------------------------|---------------------------|------------------|------------------------|-------------------------|--------------|--------------|-----------------------------|-----------------------------|
| Salmonella enterica | Typhimurium | 12 | Average | 4,890,448 | 16,299 | 300.0 | 332.6 | 17.2 | 1.11 | 0.06 |
| | | | stddev | 50,356 | | 3.1 | 13.0 | 3.6 | --- | --- |
| Salmonella enterica | Enteritidis | 12 | Average | 4,666,462 | 16,299 | 287.5 | 323.7 | 21.5 | 1.13 | 0.07 |
| | | | stddev | 20,384 | | 1.3 | 10.5 | 4.2 | --- | --- |
| Salmonella enterica | Typhi | 12 | Average | 4,770,414 | 16,299 | 292.7 | 316.9 | 29.5 | 1.08 | 0.10 |
| | | | stddev | 60,279 | | 3.7 | 5.9 | 3.2 | --- | --- |
| Salmonella enterica | mixed | 12 | Average | 4,713,701 | 16,299 | 289.2 | 315.6 | 17.2 | 1.09 | 0.06 |
| | | | stddev | 80,652 | | 4.9 | 13.9 | 4.4 | --- | --- |
| Escherichia coli | | | 12 | Average | 5,087,133 | 16,299 | 312.1 | 281.9 | 18.3 | 0.90 | 0.06 |
| | | | stddev | 262,098 | | 16.1 | 30.2 | 6.5 | --- | --- |
| Proteus mirabilis | | | 3 | Average | 4,124,431 | 16,299 | 253.0 | 712.7 | 15.7 | 2.82 | 0.06 |
| | | | stddev | 83,305 | | 5.1 | 42.0 | 2.1 | --- | --- |
| Shigella sonnei | | | 3 | Average | 4,929,559 | 16,299 | 302.4 | 261.3 | 11.7 | 0.86 | 0.04 |
| | | | stddev | 50,607 | | 5.6 | 8.4 | 3.5 | --- | --- |
| Yersinia pseudotuberculosis | | | 3 | Average | 4,802,245 | 16,299 | 294.6 | 429.3 | 120.3 | 1.46 | 0.41 |
| | | | stddev | 118,706 | | 7.3 | 2.3 | 3.8 | --- | --- |
| Vibrio vulnificus | chromosome I | 3 | Average | 3,330,104 | 16,299 | 204.3 | 180.0 | 10.3 | 0.88 | 0.05 |
| | | | stddev | 79,423 | | 4.9 | 14.0 | 4.9 | --- | --- |
| Vibrio vulnificus | chromosome II | 3 | Average | 1,756,668 | 16,299 | 107.8 | 90.0 | 3.3 | 0.83 | 0.03 |
| | | | stddev | 87,177 | | 5.3 | 7.9 | 3.1 | --- | --- |
| Staphylococcus aureus | | | 3 | Average | 2,948,373 | 16,299 | 180.9 | 108.3 | 0.0 | 0.60 | 0.00 |
| | | | stddev | 114,871 | | 7.0 | 10.6 | 0.0 | --- | --- |
| Streptococcus pyogenes | | | 3 | Average | 1,855,707 | 16,299 | 116.3 | 263.7 | 0.3 | 2.27 | 0.00 |
| | | | stddev | 42,370 | | 2.6 | 15.4 | 0.6 | --- | --- |
| Enterococcus faecalis | | | 3 | Average | 3,050,287 | 16,299 | 189.6 | 649.7 | 2.0 | 3.42 | 0.01 |
| | | | stddev | 117,259 | | 7.2 | 84.1 | 3.5 | --- | --- |
| Bacillus anthracis | | | 3 | Average | 5,228,732 | 16,299 | 320.8 | 432.0 | 1.3 | 1.35 | 0.00 |
| | | | stddev | 1,349 | | 6.1 | 11.5 | 0.6 | --- | --- |
| Bacillus cereus | | | 3 | Average | 5,406,060 | 16,299 | 315.7 | 700.3 | 13.0 | 2.11 | 0.04 |
| | | | stddev | 166,15 | | 1.0 | 53.7 | 6.1 | --- | --- |

1 Genomes included in analysis are listed in supplementary Table S1 with NCBI accession numbers.

2 The common denominator of 16,299 nucleotides (nt) used to normalize variation in genome size was obtained from Salmonella enterica subspecies I serotype Typhimurium LT2 (NC_003397.2) as described in text.
Table 2. Categories of genes from 3 serotypes of S. enterica subspecies I that vary in AT 8+mer content

| Serotype | Category | Function | AT 8+mer Content | Sequence Analysis |
|----------|----------|----------|------------------|------------------|
| Typhimurium | - | - | - | - |
| Paratyphi | - | - | - | - |
| Enteritidis | - | - | - | - |
4. Discussion

The AT 8+mer motif was located in genes and regulatory regions that impact phenotype, growth potential, virulence and metabolism of Salmonella enterica subspecies I. In addition, there is biological evidence that AT 8+mers influence evolution at the scale of the single nucleotide. For example, A and T homopolymers impact transcription termination in Archea [33]. The canine herpesvirus thymidine kinase gene has mutational hotspots at stretches of 8 adenines [34]. T7 bacteriophage RNA polymerases undergo transcription slippage at A and T homopolymers [35]. As mentioned previously for S. enterica serovar Enteritidis, a mutational hotspot in 1 of 8 adenines increased virulence [3].

While there is reason to suspect AT 8+mers as mutational hotspots, the conundrum exists that there must be a mechanism for repair of accumulating mutations. Otherwise, evolution of any one serotype of S. enterica would be unidirectional towards extinction. There are several examples of Salmonella serotypes, e.g. Typhimurium, Enteritidis, Newport, Infantis and Heidelberg, that continuously circulate over decades; however, the majority of serotypes cause illness inconsistently, rarely, or never [1, 16]. For this reason, we theorize there is another function for AT 8+mers. It is proposed that AT 8+mers align sections of genomes during replication, DNA acquisition, and DNA repair processes, thus maintaining a general organization of the S. enterica genome. This function would result in repair of mutations occurring between stretches of wildtype AT 8+mers during the replication/repair process and/or during acquisition of new DNA by homologous recombination [36, 37]. It would also account for an inherent mechanism of self-recognition, which would facilitate preferential, but not exclusive, DNA exchange within a Genus species. The pan-genome of S. enterica subspecies I has a mosaic structure between serotypes, with frequent inversions, deletions, and insertions occurring between serotypes; however, the chromosomal arrangement of many Salmonella lineages is comparatively stable [25, 32, 38, 39]. AT 8+mers being important to the processes of DNA replication, repair and acquisition by repair mechanisms and homologous recombination would account for i) the stability of some serotypes with conserved genome features that are persistent, e.g. serovar Typhimurium [1], ii) the occasional emergence of a new serotype that happens to undergo clonal expansion in an environment favorable for growth, e.g. serovar Tennessee in peanut butter [40, 41], iii) the rare emergence of a hybrid strain following a major recombination event that results in rapid proliferation of a serotype with new biological properties, e.g. serovar Enteritidis and its ability to contaminate and survive in the internal contents of eggs [42], and iv) the periodic emergence and disappearance of serotypes that are not optimized for the survival in the environment in which they are generated.

S. enterica serovars with similar AT 8+mer content would thus be expected to maintain the ability to form Holliday structures at least within subspecies I. In contrast, the two chromosomes of Vibrio vulnificus could be inhibited from recombination in part because the AT 8+mer content differs substantially. E. coli and S. enterica are natural exchangers, and an area of future research is to evaluate if some, but not all, AT 8+mer content in chromosomal segments of different Genus species align to facilitate the formation of Holliday structures that are an integral part of homologous recombination [43-45].

5. Conclusion

In summary, we suggest that AT 8+mers are a motif in the genome of Eubacteria that facilitates DNA replication, repair, and exchange while also maintaining speciation. In regards to Salmonella enterica subspecies I, the motif is proposed to contribute to the emergence of serotypes, and at the same time, maintain some genomes with optimized gene content that are highly successful as foodborne pathogens [46-49]. Future research on the AT 8+mer contribution to genome organization, fidelity of replication, and ability to restore mutated gene content will require proof of concept experimentation. Biological experimentation at an applied level will focus on finding environmental niches within food production systems that facilitate genomic exchange and repair mechanisms. Application for improving food safety will involve determining effective interventions.
Analyzing the impact of genetic repair on the safety of the food supply may require methods with detection limits that are orders of magnitude lower than those used to currently detect contaminating bacteria. This is because a successful recombinant may at first be a rare cell type [50, 51]. Further analysis into the impact of AT 8+mers on the ability of *S. enterica* to survive and persist in environments associated with foodborne illness is thus warranted.

**Supplementary Materials:** The following are available online at www.mdpi.com/xxx/s1, Table S1: List of bacterial genomes analyzed for AT 8+mer homopolymers, Table S2: Location and classification of all AT8+mers in Typhimurium LT2.

**Author Contributions:** For research articles with several authors, a short paragraph specifying their individual contributions must be provided. The following statements should be used “Conceptualization, Jean Guard and Adam Rivers; methodology, all authors; validation, Jean Guard, Adam Rivers, and Justin Vaughn; formal analysis, Adam Rivers; investigation, Jean Guard; resources, Jean Guard.; data curation, Jean Guard; writing—original draft preparation, Jean Guard.; writing—review and editing, all authors.; visualization, all authors.; supervision, Jean Guard; project administration, Jean Guard and Adam Rivers.; funding acquisition, Jean Guard and Adam Rivers. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by The U.S. Department of Agriculture, Agricultural Research Service project plan number 6040-32000-012-00-D and by The National Institute of Food and Agriculture, Agriculture and Food Research Initiative Grant Number 2019-67021-29924.

**Data Availability Statement:** The database analyzed for this project can be found at the National Center for Biotechnology Institute (NCBI) at https://www.ncbi.nlm.nih.gov.

**Conflicts of Interest:** The authors declare no conflict of interest.

**References**

1. CDC. An Atlas of *Salmonella* in the United States, 1968-2011: Laboratory-based Enteric Disease Surveillance. Atlanta, Georgia: Centers for Disease Control and Prevention (CDC), 2013.

2. Tack DM, Ray L, Griffin PM, Cieslak PR, Dunn J, Rissman T, et al. Preliminary Incidence and Trends of Infections with Pathogens Transmitted Commonly Through Food - Foodborne Diseases Active Surveillance Network, 10 U.S. Sites, 2016-2019. MMWR Morb Mortal Wkly Rep. 2020;69(17):509-14. Epub 2020/05/01. doi: 10.15585/mmwr.mm6917a1. PubMed PMID: 32352955.

3. Guard J, Cao G, Luo Y, Baugher JD, Davison S, Yao K, et al. Genome sequence analysis of 91 *Salmonella* Enteritidis isolates from mice caught on poultry farms in the mid 1990s. Genomics. 2020;112(1):528-44. Epub 2019/04/12. doi: 10.1016/j.ygeno.2019.04.005. PubMed PMID: 30974149.

4. Morales CA, Guard J, Sanchez-Ingunza R, Shah DH, Harrison M. Virulence and metabolic characteristics of *Salmonella enterica* serovar enteritidis strains with different sefD variants in hens. Appl Environ Microbiol. 2012;78(18):6405-12. Epub 2012/06/26. doi: 10.1128/AEM.00852-12. PubMed PMID: 22729535; PubMed Central PMCID: PMCPMC3426706.

5. Sanchez-Ingunza R, Guard J, Morales CA, Icard AH. Reduction of *Salmonella* Enteritidis in the spleens of hens by bacterins that vary in fimbrial protein SefD. Foodborne Pathog Dis. 2015;12(10):836-43. Epub 2015/07/29. doi: 10.1089/fpd.2015.1971. PubMed PMID: 26218804; PubMed Central PMCID: PMCPMC4601671.

6. Reich Z, Friedman P, Levin-Zaidman S, Minsky A. Effects of adenine tracts on the B-Z transition. Fine tuning of DNA conformational transition processes. J Biol Chem. 1993;268(11):8261-6. Epub 1993/04/15. PubMed PMID: 8463336.

7. Hines ER, Kolek OI, Jones MD, Serey SH, Sirjani NB, Kiela PR, et al. 1,25-dihydroxyvitamin D3 down-regulation of PHEX gene expression is mediated by apparent repression of a 110 kDa transactor that binds to a polyadenine element in the promoter. J Biol Chem. 2004;279(45):46406-14. Epub 2004/09/01. doi: 10.1074/jbc.M404278200. PubMed PMID: 15337762.

8. Lindemose S, Nielsen PE, Mollegaard NE. Polyamines preferentially interact with bent adenine tracts in double-stranded DNA. Nucleic Acids Res. 2005;33(6):1790-803. Epub 2005/03/25. doi: 10.1093/nar/gki319. PubMed PMID: 15788751; PubMed Central PMCID: PMCPMC4601671.

9. Jung BH, Beck SE, Cabral J, Chau E, Cabrera BL, Fiorino A, et al. Activin type 2 receptor restoration in MSI-H colon cancer suppresses growth and enhances migration with activin. Gastroenterology. 2007;132(2):633-44. Epub 2007/01/30. doi: 10.1053/j.gastro.2006.11.018. PubMed PMID: 17258738; PubMed Central PMCID: PMCPMC4154562.

10. Agnoli K, Haldipurkar SS, Tang Y, Butt AT, Thomas MS. Distinct Modes of Promoter Recognition by Two Iron Starvation sigma Factors with Overlapping Promoter Specificities. J Bacteriol. 2019;201(3). Epub 2018/11/21. doi: 10.1128/JB.00507-18. PubMed PMID: 30455278; PubMed Central PMCID: PMCPMC6349086.

11. Roberts JD, Nguyen D, Kunkel TA. Frameshift fidelity during replication of double-stranded DNA in HeLa cell extracts. Biochemistry. 1993;32(15):4083-9. Epub 1993/04/20. doi: 10.1021/bi00066a033. PubMed PMID: 8385995.
12. Traverse CC, Ochman H. Genome-Wide Spectra of Transcription Insertions and Deletions Reveal That Slippage Depends on RNA:DNA Hybrid Complementarity. mBio. 2017;8(4). Epub 2017/08/31. doi: 10.1128/mBio.01230-17. PubMed PMID: 28851848; PubMed Central PMCID: PMCPMC5574173.

13. Gragg H, Harfe BD, Jinks-Robertson S. Base composition of mononucleotide runs affects DNA polymerase slippage and removal of frameshift intermediates by mismatch repair in Saccharomyces cerevisiae. Mol Cell Biol. 2002;22(24):8756–62. Epub 2002/11/26. doi: 10.1128/mcb.22.24.8756-8762.2002. PubMed PMID: 12446792; PubMed Central PMCID: PMCPMC139878.

14. Wigley P. Salmonella enterica serovar Gallinarum: addressing fundamental questions in bacteriology sixty years on from the 9R vaccine. Avian Pathol. 2017;46(2):119-24. Epub 2016/10/30. doi: 10.1080/03079457.2016.1240866. PubMed PMID: 27791403.

15. Guard J, Morales CA, Fedorka-Cray P, Gast RK. Single nucleotide polymorphisms that differentiate two subpopulations of Salmonella enteritidis within phage type. BMC Res Notes. 2011;4:369. Epub 2011/09/29. doi: 10.1186/1756-0500-4-369. PubMed PMID: 21942987; PubMed Central PMCID: PMCPMC3220660.

16. Grimont P, Weill F-X. Antigenic formulae of the Salmonella serovars. 9th Edition. Paris, France: WHO Collaborating Centre for Reference and Research on Salmonella. Paris, France: World Health Organization, 2007.

17. Guard-Bouldin J, Gast RK, Humphrey TJ, Henzler DJ, Morales C, Coles K. Subpopulation characteristics of egg-contaminating Salmonella enterica serovar Enteritidis as defined by the lipopolysaccharide O-chain. Appl Environ Microbiol. 2004;70(5):2756-63. Epub 2004/05/07. doi: 10.1128/aem.70.5.2756-2763.2004. PubMed PMID: 15128529; PubMed Central PMCID: PMCPMC404386.

18. Gantois I, Ducatelle R, Pasmans F, Haesebrock F, Van Immerseel F. The Salmonella Enteritidis lipopolysaccharide biosynthesis gene rfbH is required for survival in egg albumen. Zoosones Public Health. 2009;56(3):145-9. Epub 2008/11/08. doi: 10.1111/j.1863-2378.2008.01195.x. PubMed PMID: 18990194.

19. Parker CT, Liebana E, Hengler DJ, Guard-Jones CA, Mitchell EJL, Clouthier SC, Collinson SK, Kay WW. Unique fimbriae according to MLST, MLEE and microarray hybridization. Infect Genet Evol. 2013;16:314-23. Epub 2013/05/21. doi: 10.1016/j.meegid.2013.04.005. PubMed PMID: 23785719; PubMed Central PMCID: PMCPMC3769191.

20. Droop MR, Kay WW, Clouthier SC. Selection of Salmonella enterica serovar Enteritidis during passage through the chicken egg. J Gen Microbiol. 1996;142(3):735-42. Epub 1996/03/15. doi: 10.1128/9781. PubMed PMID: 8591161; PubMed Central PMCID: PMCPMC3051077.

21. Clouthier SC, Collinson SK, Kay WW. Unique fimbriae-like structures encoded by selD of the SEF14 fimbrial gene cluster of Salmonella enterica. Mol Microbiol. 1994;12(6):893-901. Epub 1994/06/01. doi: 10.1111/j.1365-2958.1994.tb01077.x. PubMed PMID: 7934897.
Med PMID: 32333808.

Richards AK, Hopkins BA, Shariat NW. Conserved CRISPR arrays in Salmonella. Front Microbiol. 2020;11(6):e1007261. Epub 2020/04/06. doi: 10.3389/fmicb.2018.010344. PubMed PMID: 30474557; PubMed Central PMCID: PMCPMC6260861.

Brandis G, Cao S, Hughes D. Co-evolution with recombination affects the stability of mobile genetic element insertions within gene families of Salmonella. Mol Microbiol. 2018(6):697-710. Epub 2018/04/01. doi: 10.1111/mmi.13959. PubMed PMID: 29603442.

Brandis G, Cao S, Hughes D. Measuring Homologous Recombination Rates between Chromosomal Locations in Salmonella. Bio Protoc. 2019;9(3):e3159. Epub 2019/02/05. doi: 10.21769/BioProtoc.3159. PubMed PMID: 33654967; PubMed Central PMCID: PMCPMC645157.

Achtman M, Zhou Z, Alikhan NF, Tyne W, Parkhill J, Cormican M, et al. Genomic diversity of Salmonella enterica - The UoWUC 10K genomes project. Wellcome Open Res. 2020;5:223. Epub 2021/02/24. doi: 10.12688/wellcomeopenres.16291.2. PubMed PMID: 33614977; PubMed Central PMCID: PMCPMC7869069.

Alikhan NF, Zhou Z, Sergeant MJ, Achtman M. A genomic overview of the population structure of Salmonella. PLoS Genet. 2018;14(4):e1007261. Epub 2018/04/06. doi: 10.1371/journal.pgen.1007261. PubMed PMID: 29621240; PubMed Central PMCID: PMCPMC6886390.

Dong HJ, Cho S, Boxrud D, Rankin S, Downe F, Lovchik J, et al. Single-nucleotide polymorphism analysis for molecular subtyping of Salmonella Tennessee isolates associated with the 2007 nationwide peanut butter outbreak in the United States. Gut Pathog. 2017;9:25. Epub 2017/05/05. doi: 10.1186/s13099-017-0176-y. PubMed PMID: 28469710; PubMed Central PMCID: PMCPMC4892500.

Wilson MR, Brown E, Keys C, Strain E, Luo Y, Muruvanda T, et al. Whole Genome DNA Sequence Analysis of Salmonella enterica serovar Newport. J Bacteriol. 2010;192(24):6465-76. Epub 2010/10/12. doi: 10.1128/jb.00969-10. PubMed PMID: 20935094; PubMed Central PMCID: PMCPMC3008538.

J. G, D. S, C. M, D. C. Evolutionary trends associated with niche specialization as modeled by whole genome analysis of egg-contaminating Salmonella enterica serovar Enteritidis. In: Porwollik S, editor. Salmonella: From Genome to Function. San Diego: Caister Academic Press; 2011. p. 91 - 106.

Galitski T, Roth JR. Pathways for homologous recombination between chromosomal direct repeats in Salmonella typhimurium. Genetics. 1997;146(3):751-67. Epub 1997/07/01. PubMed PMID: 9215885; PubMed Central PMCID: PMCPMC1208049.

Sanderson KE, Liu SL. Chromosomal rearrangements in enteric bacteria. Electrophoresis. 1998;19(4):569-72. Epub 1998/05/20. doi: 10.1002/elps.1150190417. PubMed PMID: 9588803.

Kuzminov A. Homologous Recombination-Experimental Systems, Analysis, and Significance. EcoSal Plus. 2011;4(2). Epub 2011/12/01. doi: 10.1128/ecosalplus.7.2.6. PubMed PMID: 26442506; PubMed Central PMCID: PMCPMC4190071.

Zhou Z, McCann A, Littrip E, Murphy R, Cormican M, Fanning S, et al. Neutral genomic microevolution of a recently emerged pathogen, Salmonella enterica serovar Agona. PLoS Genet. 2013;9(4):e1003471. Epub 2013/05/03. doi: 10.1371/journal.pgen.1003471. PubMed PMID: 23637636; PubMed Central PMCID: PMCPMC3630104.

Angal V, Harbottle H, Mazzoni CJ, Helmuth R, Guerra B, Didelot X, et al. Evolution and population structure of Salmonella enterica serovar Newport. J Bacteriol. 2010;192(24):6465-76. Epub 2010/10/12. doi: 10.1128/jb.00969-10. PubMed PMID: 20935094; PubMed Central PMCID: PMCPMC3008538.

ark CJ, Andam CP. Distinct but Intertwined Evolutionary Histories of Multiple Salmonella enterica Subspecies. mSystems. 2020;5(1). Epub 2020/01/16. doi: 10.1128/mSystems.00515-19. PubMed PMID: 31937675; PubMed Central PMCID: PMCPMC6967386.

Liu Y, Zhang DF, Zhou X, Xu L, Zhang L, Shi X. Comprehensive Analysis Reveals Two Distinct Evolution Patterns of Salmonella Flagellin Gene Clusters. Front Microbiol. 2017;8:2604. Epub 2018/01/10. doi: 10.3389/fmicb.2017.02604. PubMed PMID: 29332269; PubMed Central PMCID: PMCPMC5744181.

Richards AK, Hopkins BA, Shariat NW. Conserved CRISPR arrays in Salmonella enterica serovar Infantis can serve as qPCR targets to detect Infantis in mixed serovar populations. Lett Appl Microbiol. 2020;71(2):138-45. Epub 2020/04/26. doi: 10.1111/lam.13296. PubMed PMID: 32333808.

Shariat N, Dudley E. CRISPR Typing of Salmonella isolates. Methods Mol Biol. 2021;2182:39-44. Epub 2020/09/08. doi: 10.1007/978-1-0716-0791-6_5. PubMed PMID: 32894485.