 doi:10.20944/preprints202106.0384.v1

Article

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

Jean Guard[1,*], Adam R Rivers [2], Justin N Vaughn[1], Michael J Rothrock, Jr[1], Adelumola Oladeinde[1], and Devendra H Shah[3]

[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 *sef*D [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].

 doi:10.20944/preprints202106.0384.v1

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,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2021 doi:10.20944/preprints202106.0384.v1

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 V16.0.0 352) (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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2021 doi:10.20944/preprints202106.0384.v1

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+kmer 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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2021 doi:10.20944/preprints202106.0384.v1

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+kmer 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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2021 doi:10.20944/preprints202106.0384.v1

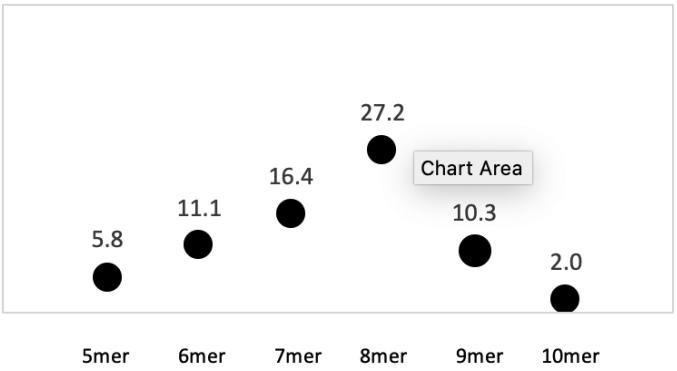

Figure 1. Ratios of AT homopolymers from 5 to 10 nucleotides in *Salmonella enterica* serovar Typhimurium LT2 NC_003197.2. The ratio of AT homopolymer kmers, either adenine or thymine but not mixed, to GC homopolymers was determined using Geneious software as described in text. The range in number of nucleotides per kmer searched was 5 to 10 (see legend label). Results showe 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.

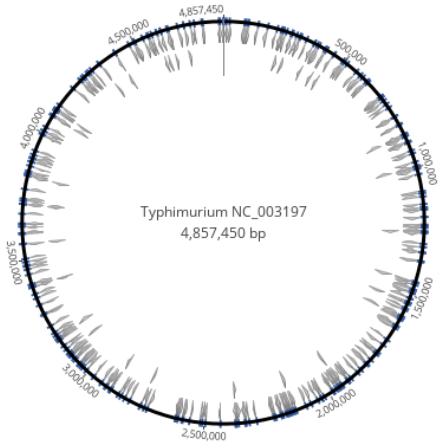

Figure 2. Locations of AT 8+mers in the genome of *Salmonella enterica* serovar Typhimurium LT2 NC_003197.2.

Table 1. Expected versus observed occurrence of homopolmer strings of 8 and more nucleotides in genomes of Eubacteria

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2021 doi:10.20944/preprints202106.0384.v1

Genus species [1]	Other genome information	Number of genomes analyzed		genome size (bp)	Common denominator (nt)[2]	Expected number of 8+kmers	Observed AT 8+mers	Observed GC 8+mers	Observed vs expecterd AT 8+mers	Observed vs expected GC 8+mers
Salmonella enterica	Typhimurium	12	Average	4,890,448	16,299	300.0	332.6	17.2	1.11	0.06
			stdev	50,356	---	3.1	13.0	3.6	---	---
Salmonella enterica	Enteritidis	12	Average	4,686,462	16,299	287.5	323.7	21.5	1.13	0.07
			stdev	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
			stdev	60,270	---	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
			stdev	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
			stdev	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
			stdev	83,305	---	5.1	42.0	2.1	---	---
Shigella sonnei	---	3	Average	4,929,599	16,299	302.4	261.3	11.7	0.86	0.04
			stdev	90,607	---	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
			stdev	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
			stdev	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
			stdev	87,177	---	5.3	7.9	3.1	---	---
Staphylococcus aureus	---	3	Average	2948373	16,299	180.9	108.3	0.0	0.60	0.00
			stdev	114371	---	7.0	10.6	0.0	---	---
Streptococcus pyogenes	---	3	Average	1,895,707	16,299	116.3	263.7	0.3	2.27	0.00
			stdev	42,370	---	2.6	15.4	0.6	---	---
Enterococcus faecalis	---	3	Average	3,090,387	16,299	189.6	649.7	2.0	3.42	0.01
			stdev	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
			stdev	1,349	---	0.1	11.5	0.6	---	---
Bacillus cereus	---	3	Average	5,406,060	16,299	331.7	700.3	13.0	2.11	0.04
			stdev	16615	---	1.0	53.7	6.1	---	---

[2]Genomes included in analysis are listed in supplementary Table S1 with NCBI accession numbers.

[1]The common denominater of 16,299 nucleotides (nt) used to normalize variation in geome size was obtained from *Salmonella enterica* subspecies I serotype Typhimurum LT2 (NC_003197.2) as described in text.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2021 doi:10.20944/preprints202106.0384.v1

Table 2. Categories of genes from 3 serotypes of S. enterica subspecies I that vary in AT 8+mer content

STM gene accession	SEEG gene accession	SEN gene accession	AT 8+mer Sequence variation	Common name of gene	Description of target gene	Gene function
A. Genes of *S.* Typhimurium (STM) that vary from either or both *S.* Gallinarum (SEEG) and *S.* Enteritidis (SEN)						
STM0105	SEEG9184_RS20395	SEN_RS00535	STM 1 bp sub	yabI	All ORFs intact	DedA family protien
STM0212	---	SEN_RS01100	STM 1 bp sub	unnamed	Both ORFs intact	MFS transporter
STM0229	SEEG9184_RS19720	SEN_RS01185	SEEG, SEN 1 bp sub	lpxB	All ORFs intact	Lipid-A-disaccharide synthase
STM0341	SEEG9184_RS23875	SEN_RS01660	SEEG 1 bp del	unnamed	SEEG pseudogene	putative inner membrane protein
STM0342	SEEG9184_RS2540	SEN_RS24170	SEN 1 bp insert	unnamed	All ORFs intact	putative periplasmic protein
STM0555	SEEG9184_RS23515	SEN_RS24190	SEEG, SEN 1 bp sub	unnamed	All pseudogenes	pseudogene in all 3 genomes; frameshift relative to E coli ES3 putative transposase
STM0558	---	SEN_RS02735	SEN 1 bp del	yfdH	Both ORFs intact	putative glycosyltransferase
STM1107_regulon_1,2	---	SEN_RS05020_1,2	(2) STM 1 SNP del	hpaX	na	4-hydroxyphenylacetate permease
STM1130	SEEG9184_RS15695	---	SEEG 1 bp del	unnamed	SEEG pseudogene	N-acetylneuraminic acid mutarotase
STM1240_1,2,3	SEEG9184_RS11205	SEN_RS09405	(1) SEN TTG insert, (3) STM 1 bp del	envF	SEEG pseudogene	lipoprotein with ribosome binding site
STM1520	SEEG9184_RS12690	---	STM 1 bp sub	marR	Both ORFs intact	transcriptional regulator multiple antibiotic resistance protein MAR
STM1550	---	SEN_RS07790	Deletion of 154 bp in SEN	unnamed	SEN pseudogene	type II toxin-antitoxin system mRNA interferase toxin
STM1602	SEEG9184_RS13085	---	SEEG 1 bp del	sifB	SEEG pseudogene	effector protein SifB
STM1635	SEEG9184_RS13270	---	STM 1 bp sub	unnamed	ORF intact	amino acid ABC transporter ATP-binding protein
STM1670	SEEG9184_RS13465	---	SEEG 1 bp del	unnamed	ORF intact	HO protein
STM1698	SEEG9184_RS13605	---	SEEG 1 bp del	steC	SEEG pseudogene	secreted effector kinase SteC
STM1869	SEEG9184_RS14565	---	SEEG 1 bp del	unnamed	SEEG pseudogene	HO protein
STM1939	SEEG9184_RS15200	---	SEEG 1 bp del	unnamed	SEEG pseudogene	putative glucose-6-phosphate dehydrogenase
STM1941	SEEG9184_RS15210	SEN_RS05525	SEEG, SEN 1 bp del	unnamed	SEEG pseudogene	HO protein
STM2047	SEEG9184_RS10150	SEN_RS10650	STM 1 bp sub	pduL	All ORFs intact	propanediol utilization protein
STM2090	SEEG9184_RS10005	SEN_RS10870	SEEG, SEN 1 bp sub	rfbH	All ORFs intact	lipopolysaccharide biosynthesis protein; LPS side chain defect
STM2100	SEEG9184_RS09955	SEN_RS10920	SEEG, SEN 1 bp sub	wcaL	All ORFs intact	colanic acid biosynthesis glycosyltransferase
STM2129	SEEG9184_RS09800	---	SEEG 1 bp del	yegB	SEEG pseudogene	multidrug transporter subunit MdtD
STM2134	SEEG9184_RS09775	---	SEEG 1 bp del	unnamed	ORF intact	HO protein
STM2268_regulon	bp1841830-1842036	bp2360965-2361170	SEN 1 SNP del	micF-regulon	na	Not annotated in SEEG:ribosomal binding site of micF in STM
STM2277	SEEG9184_RS09060	---	SEEG 1 SNP del	ubiG-regulon	na	3-demethylubiquinone-9 3-methyltransferase
STM3022	---	SEN_RS14915	SEN 1 bp sub	unnamed	ORF intact	transporter
STM3166	SEEG9184_RS04980	SEN_RS15645	STM 1 bp del	unnamed	All ORFs intact	putative cation transporter
STM3212	SEEG9184_RS04730	---	STM 1 bp sub	mug	ORF intact	double-stranded uracil-DNA glycosylase
STM3464	SEEG9184_RS01925	---	STM 1 bp del	prkB	ORF intact	putative phosphoribulokinase
STM3638	SEEG9184_RS01040	---	SEEG 1 bp del	lpfC	ORF intact	fimbrial assembly protein; SEEG9184 has 5 bp deletion
STM3674	SEEG9184_RS00845	SEN_RS18190	SEEG, SEN 1 bp sub	lyxK	SEEG pseudogene	carbohydrate kinase
STM4351	---	SEN_RS21445	SEN 1 bp del	unnamed	ORF intact	arginine ABC transporter substrate-binding protein
B. Genes of *S.* Typhimurium (STM) with AT 8+mers but with no homolog to either *S*. Gallinarum or *S.* Enteritidis						
STM0290	---	---	na	unnamed	ORF intact	HO protein
STM0720	---	---	na	unnamed	ORF intact	putative glycosyl transferase
STM0722	---	---	na	unnamed	ORF intact	ABC transporter permease
STM0723	---	---	na	unnamed	ORF intact	similar to E coli ATP-binding component of putrescine transport system; sugar ABC transporter ATP-binding protein
STM0870	---	---	na	unnamed	ORF intact	putative inner membrane protein
STM0929	---	---	na	unnamed	ORF intact	Gifsy-2 prophage excisionase
STM1006	---	---	na	unnamed	ORF intact	Gifsy-2 prophage protein
STM1023	---	---	na	unnamed	ORF intact	Gifsy-2 prophage protein
STM1554	---	---	na	unnamed	ORF intact	putative coiled-coil protein
STM2088	---	---	na	rfbX	ORF intact	O-antigen transferase:LPS side chain defect if mutated; putative transporter
STM2137	---	---	na	unnamed	ORF intact	putative cytoplasmic protein
STM2584	---	---	na	gogB	ORF intact	type III secretion system protein
STM2592	---	---	na	unnamed	ORF intact	similar to phage tail component L
STM2609	---	---	na	unnamed	ORF intact	Gifsy-1 prophage protein
STM2706	---	---	na	unnamed	ORF intact	Fels-2 prophage protein; similar to tail fiber protein in phage P2
STM2754	---	---	na	unnamed	ORF intact	hexulose-6-phosphate synthase
STM2766	---	---	na	unnamed	ORF intact	HO protein
STM2902	---	---	na	unnamed	ORF intact	putative cytoplasmic protein
STM4435	---	---	na	unnamed	ORF intact	inosose isomerase
STM4495	---	---	na	unnamed	ORF intact	type II restriction enzyme methylase subunit
STM4524	---	---	na	hsdS	ORF intact	restriction endonuclease subunit S
C. Genes of *S.* Enteritidis and/or *S.* Gallinarum with similar AT 8+mers but with no homolog in *S.* Typhimurium						
---	SEEG9184_RS10025	SEN_RS10850	same	unnamed	Both ORFs intact	glycosyltransferase family 2 protein
---	SEEG9184_RS10420	SEN_RS10400	same	unnamed	Both ORFs intact	HO protein
---	SEEG9184_RS10430	SEN_RS10390	same	unnamed	Both ORFs intact	SLATT domain-containing protein
---	SEEG9184_RS10595	SEN_RS09970	same	unnamed	Both ORFs intact	HO protein
---	SEEG9184_RS14095	SEN_RS06485	same	unnamed	Both ORFs intact	HO protein
---	SEEG9184_RS14610	---	na	unnamed	ORF intact	HO protein
---	SEEG9184_RS15540	---	na	unnamed	ORF intact	HO protein
---	SEEG9184_RS19390	---	na	unnamed	Both ORFs intact	DUF3987 domain-containing protein
---	SEEG9184_RS19405	---	na	unnamed	Both ORFs intact	HO protein
---	SEEG9184_RS21290	SEN_RS22310	same	unnamed	Both ORFs intact	AAA family ATPase
---	SEEG9184_RS21300	SEN_RS22300	same	unnamed	Both ORFs intact	restriction endonuclease subunit M
---	SEEG9184_RS21305_1,2,3,4	SEN_RS22295_1,2,3,4	same	unnamed	Both ORFs intact	N-6 DNA methylase
---	SEEG9184_RS21500_1,2_regulon	SEN_RS22090_1,2_regulon	(1) SEN 1 SNP sub, (2) same	unnamed	na	AraC family transcriptional regulator
---	SEEG9184_RS21505	SEN_RS22085	SEEG 1 SNP del	sefD	SEEG pseudogene	adhesin
---	SEEG9184_RS21510_1,2,3 (PS)	SEN_RS22080_1,2,3	same	sefC	SEEG pseudogene	outer membrane fimbrial user protein SefC, pseudo in SG, which has an extra A/T to make a 7mer
---	SEEG9184_RS22845	SEN_RS20760	same	unnamed	ORF intact	putative inner membrane protein
D. Genes of *S.* Typhimurium and *S.* Enteritidis with AT 8+mers but no homolog in *S.* Gallinarum						
STM0870	---	SEN_RS04220	same	unnamed	Both ORFs intact	protein transporter: not in SEEG9184; see SG_RS04310
STM1054	---	SEN_RS23040	same	unnamed	SEN pseudogene	Gifsy-2 prophage protein in STM: GC rich region has a deletion in in SEN due to 7bp insert in a guanidine rich fragment, causing a frameshift
STM1666	---	SEN_RS07090	same	unnamed	STM pseudogene	STM has in-frame stop following codon 24; SEN, hypothetical protein
STM2065	---	SEN_RS10740	same	phsA	Both ORFs intact	thiosulfate reductase
STM3023	---	SEN_RS14920	same	yohL/rcnR	Both ORFs intact	transcriptional regulator; Ni(ii)/Co(ii)-binding transcriptional repressor
STM3034	---	SEN_RS14975	same	vapB	Both ORFs intact	toxin-antitoxin system antitoxin VapB
E. Genes of *S.* Typhimurium (STM), *S.* Gallinarum (SEEG), and *S.* Enteritidis with conserved AT 8+mers						
STM0004	SEEG9184_RS20920	SEN_RS00020	same	thrC	All ORFs intact	threonine synthase
STM0022	SEEG9184_RS20830	SEN_RS00110	same	bcfB	All ORFs intact	fimbrial chaperone protein
STM0071	SEEG9184_RS20580	SEN_RS00360	same	caiC	SEEG pseudogene	crotonobetaine/carnitine-CoA ligase
STM0074	SEEG9184_RS20565	SEN_RS00375	same	caiT	All ORFs intact	L-carnitine:gamma-butyrobetaine antiporter
STM0084	SEEG9184_RS20510	SEN_RS00430	same	unnamed	All ORFs intact	phosphatase; sulfatase-like hydrolase/transferase
STM0248	SEEG9184_RS19625	SEN_RS01280	same	yaeD/gmhB	All ORFs intact	D-glycero-beta-D-manno-heptose 1,7-bisphosphate-7-phosphatase
STM0319	SEEG9184_RS19215	SEN_RS01550	same	crl	All ORFs intact	sigma factor-binding protein; curli surface fiber csgA regulation
STM0333	SEEG9184_RS19130	SEN_RS01620	same	unnamed	All ORFs intact	putative LysR family transcriptional regulator
STM0359	SEEG9184_RS23530	SEN_RS01755	same	unnamed	All ORFs intact	HO protein
STM0397_regulon	SEEG9184_RS18805_regulon	SEN_RS01950_regulon	same	phoB	na	phosphate response regulator transcription factor
STM0437	SEEG9184_RS18595	SEN_RS02150	same	unnamed	All ORFs intact	sel1 repeat family protein
STM0497	SEEG9184_RS18275	SEN_RS02445	same	unnamed	All ORFs intact	HO protein
STM0528	SEEG9184_RS18125	SEN_RS02595	same	allD	All ORFs intact	ureidoglycolate dehydrogenase
STM0551	SEEG9184_RS18005	SEN_RS02710	same	unnamed	All ORFs intact	diguanylate cyclase
STM0626_regulon	SEEG9184_RS17645_regulon	SEN_RS03035_regulon	same	dpiB	na	sensor histidine kinase
STM0664_regulon	SEEG9184_RS17450_regulon	SEN_RS03235_regulon	same	gltJ	na	glutamate/aspartate ABC transporter permease
STM0756	SEEG9184_RS17055	SEN_RS03640	same	nadA	All ORFs intact	quinolinate synthase
STM0795	SEEG9184_RS16855	SEN_RS03835	same	bioF	All ORFs intact	8-amino-7-oxononanoate synthase
STM0810	SEEG9184_RS16775	SEN_RS03910	same	unnamed	All ORFs intact	HO protein
STM0827	SEEG9184_RS16695	SEN_RS03995	same	ybiO	All ORFs intact	mechanosensitive channel protein
STM0858	SEEG9184_RS16515	SEN_RS04155	same	unnamed	SEEG pseudogene	electron transfer flavoprotein-ubiquinone oxidoreductase
STM1062	SEEG9184_RS16060	SEN_RS04790	same	uup	All ORFs intact	ABC transporter ATP-binding protein
STM1063	SEEG9184_RS16055	SEN_RS04795	same	pqiA	All ORFs intact	paraquat-inducible protein A
STM1106	SEEG9184_RS15840	SEN_RS05015	same	hpaI	All ORFs intact	2,4-dihydroxyhept-2-ene-1,7-dioic acid aldolase
STM1135	SEEG9184_RS10625	SEN_RS09945	same	serX/ycdW	All ORFs intact	upstream bifunctional glyoxylate/hydroxypyruvate reductase A; ribosome binding site
STM1157	SEEG9184_RS10755	SEN_RS09830	same	yceI	All ORFs intact	HO protein
STM1163_regulon	SEEG9184_RS10790_regulon	SEN_RS09800_regulon	same	pyrC	na	dihydrorotase
STM1169	SEEG9184_RS10820	SEN_RS09770	same	mviM	All ORFs intact	virulence factor MviM; Gfo/Idh/MocA family oxidoreductase
STM1224	SEEG9184_RS11110	SEN_RS09490	same	sifA	All ORFs intact	replication in macrophages; SPI-2 type III secretion system effector SifA
STM1363_regulon	SEEG9184_RS11875	SEN_RS22750	same	rprA	na	regulatory RNA: anti-sense sRNA RprA
STM1392	SEEG9184_RS12025	SEN_RS08580	same	ssrA	All ORFs intact	hybrid sensor histidine kinase/response;two component system sensor kinase
STM1402	SEEG9184_RS12080	SEN_RS08530	same	sseE	All ORFs intact	LcrR family type III secretion system chaperone
STM1409	SEEG9184_RS12115	SEN_RS08495	same	ssaJ	All ORFs intact	EscJ/YscJ/HrcJ family type III secretion inner membrane ring protein
STM1432_1,2	SEEG9184_RS12225_1,2	SEN_RS08380_1,2	same	ydhO	All ORFs intact	putative cell wall-associated hydrolase;C40 family peptidase
STM1477	SEEG9184_RS12450	SEN_RS08155	same	ydgI	All ORFs intact	putative amino acid transporter/permease
STM1484	SEEG9184_RS12495	SEN_RS08120	same	unnamed	All ORFs intact	serine protease
STM1560	SEEG9184_RS12850	SEN_RS07750	same	treZ/treY	SEEG treY pseudogene	malto-oligosyltrehalose trehalohydrolase
STM1630	SEEG9184_RS13245	SEN_RS07380	same	unnamed	All ORFs intact	HO protein
STM1716	SEEG9184_RS13730	SEN_RS06820	same	sohB	All ORFs intact	protease
STM1798	SEEG9184_RS14165	SEN_RS06415	same	ycgR	All ORFs intact	flagellar brake protein
STM1813	SEEG9184_RS14230	SEN_RS06340	same	ycgL	All ORFs intact	putative cytoplasmic protein
STM1854	SEEG9184_RS23435	SEN_RS06135	same	unnamed	All ORFs intact	HO protein; membrane protein
STM1862	SEEG9184_RS14495	SEN_RS06085	same	pagO	All ORFs intact	EamA/DMT family transporter (PAGO protein)
STM1887	SEEG9184_RS14940	SEN_RS05795	same	yebK	All ORFs intact	transcriptional regulator of HexR;MurR/RpiR family transcriptional regulator
STM1889	SEEG9184_RS14950	SEN_RS05785	same	lpxM	All ORFs intact	lauroyl-Kdo(2)-lipid IV(A) myristoyltransferase
STM1940	SEEG9184_RS15205	SEN_RS05530	same	unnamed	All ORFs intact	putative cell wall-associated hydrolase
STM1963	SEEG9184_RS15315	SEN_RS05415	same	amyA	All ORFs intact	cytoplasmic alpha-amylase
STM2020	SEEG9184_RS23790	SEN_RS10515	same	cbiO	SEEG pseudogene	cobalt transport atp-binding protein CbiO: B12 synthesis associated?
STM2082	SEEG9184_RS10050	SEN_RS10825	same	rfbP	All ORFs intact	UDP-phosphate galactose phosphotransferase; bifunctional enzyme; LPS side chain defect
STM2086	SEEG9184_RS10030	SEN_RS10845	same	rfbU	All ORFs intact	mannosyl transferase
STM2093	SEEG9184_RS09990	SEN_RS10885	same	rfbI	All ORFs intact	CDP-6-deoxy-delta-3,4-glucoseen reductase; LPS
STM2097	SEEG9184_RS09970	SEN_RS10905	same	rfbB	All ORFs intact	dTDP-glucose 4,6-dehydratase
STM2112	SEEG9184_RS09895	SEN_RS10980	same	wcaD	All ORFs intact	colanic acid polymerase WcaD
STM2113	SEEG9184_RS09890	SEN_RS10985	same	wcaC	All ORFs intact	colanic acid biosynthesis glycosyltransferase
STM2118	SEEG9184_RS09865	SEN_RS11010	same	wza	All ORFs intact	polysaccharide exporter
STM2118_regulon	not annotated_regulon	SEN_RS11010_regulon	same	unnamed	na	see STM2118
STM2241	SEEG9184_RS09260	SEN_RS11570	same	sspH2	SEEG pseudogene	E3 ubiquitin--protein ligase; induced by the SPI-2 regulatory ssrA/B
STM2245	SEEG9184_RS09225	SEN_RS11595	same	unnamed	All ORFs intact	putative outer membrane protein
STM2273	SEEG9184_RS09075	SEN_RS11730	same	unnamed	All ORFs intact	MR-MLE family protein: starvation sensing protein?
STM2274	SEEG9184_RS09070	SEN_RS11735	same	unnamed	SEEG pseudogene	MFS transporter
STM2328	SEEG9184_RS08795	SEN_RS12005	same	nuoA	All ORFs intact	NADH-quinone oxidoreductase subunit A
STM2386	SEEG9184_RS08495	SEN_RS12305	same	smrB	All ORFs intact	endonuclease
STM2397	SEEG9184_RS08410	SEN_RS12380	same	pgtB	All ORFs intact	sensor histidine kinase
STM2403	SEEG9184_RS08370	SEN_RS12420	same	glk	All ORFs intact	glucokinase
STM2449	SEEG9184_RS08135	SEN_RS12660	same	unnamed	All ORFs intact	putative acetyltransferase
STM2475	SEEG9184_RS08005	SEN_RS12780	same	unnamed	All ORFs intact	HO protein
STM2490	SEEG9184_RS07925	SEN_RS12855	same	gcvR	All ORFs intact	transcriptional repressor of gcv operon
STM2494	SEEG9184_RS07895	SEN_RS12885	same	unnamed	All ORFs intact	HO protein
STM2498	SEEG9184_RS07875	SEN_RS12905	same	upp	All ORFs intact	uracil phosphoribosyltransferase
STM2691	SEEG9184_RS07115	SEN_RS13585	same	unnamed	SEEG pseudogene	type I secretion system permease/ATPase: TolC family OMP
STM2782	SEEG9184_RS07005	SEN_RS13665	same	mig-14	All ORFs intact	antimicrobial resistance protein Mig-14
STM2819	SEEG9184_RS06805	SEN_RS13870	same	yqaA/yqaB	SEEG yqaB pseudogene	HO protein
STM2862	SEEG9184_RS06575	SEN_RS14085	same	sitB	All ORFs intact	manganese/iron transporter ATP-binding protein
STM2878	SEEG9184_RS06495	SEN_RS14165	same	sptP	All ORFs intact	pathogenicity island 1 effector protein StpP
STM2886	SEEG9184_RS06455	SEN_RS14205	same	sicA	All ORFs intact	CesD/SycD/LcrH family type III secretion system chaperone: surface presentation of antigens
STM2897	SEEG9184_RS06400	SEN_RS14260	same	invE	All ORFs intact	type III secretion system gatekeeper InvE
STM2898	SEEG9184_RS06395	SEN_RS14265	same	invG	All ORFs intact	type III secretion system outer membrane ring protein InvG
STM2932	SEEG9184_RS06225	SEN_RS14430	same	ygbE	All ORFs intact	putative inner membrane protein
STM2945	SEEG9184_RS06150	SEN_RS14495	same	sopD	All ORFs intact	SPI-1 type III secretion system effector
STM2966	SEEG9184_RS05995	SEN_RS23740	same	csrB	All ORFs intact	regulatory RNA in LT2
STM3005	SEEG9184_RS05790	SEN_RS14830	same	mutH	All ORFs intact	DNA mismatch repair endonuclease MutH
STM3118	SEEG9184_RS05225	SEN_RS15405	same	unnamed	All ORFs intact	putative acetyl-CoA hydrolase
STM3121	SEEG9184_RS05210	SEN_RS15420	same	unnamed	All ORFs intact	putative LysR family transcriptional regulator
STM3185	SEEG9184_RS04880	SEN_RS15745	same	yqiE	All ORFs intact	ADP-ribose diphosphatase
STM3233	SEEG9184_RS04620	SEN_RS15990	same	yqjG	All ORFs intact	glutathione-dependent reductase
STM3274	SEEG9184_RS04430	SEN_RS16165	same	yhbU	All ORFs intact	putative protease
STM3311	SEEG9184_RS04240	SEN_RS16355	same	yrbD/mlaD	All ORFs intact	outer membrane lipid asymmetry maintenance protein MlaD
STM3328	SEEG9184_RS04150	SEN_RS16440	same	arcB	All ORFs intact	two-component sensor histidine kinase; senses redox conditions
STM3480	SEEG9184_RS01850	SEN_RS17205	same	yhfL	All ORFs intact	DUF4223 domain-containing protein
STM3484	SEEG9184_RS01830	SEN_RS17225	same	dam	All ORFs intact	DNA adenine methylase
STM3515	SEEG9184_RS01665	SEN_RS17380	same	malT	All ORFs intact	transcriptional regulator
STM3658	SEEG9184_RS00930	SEN_RS18105	same	yiaH	SEEG pseudogene	acetyltransferase
STM3691	SEEG9184_RS00750	SEN_RS18275	same	sadA	All ORFs intact	trimeric autotransporter adhesin; SEEG has 42 bp deletion, no AT 8+mer involved
STM3697_regulon	SEEG9184_RS00715_regulon	SEN_RS18305_regulon	same	unnamed	na	RBS for mandelate racemase/muconate lactonizing protein
STM3794	SEEG9184_RS00245	SEN_RS18745	same	tsx, nupG E. coli	All ORFs intact	DeoR family transcriptional regulator
STM3806	bp37560-37846	bp3881833-3882117	same	unnamed	All pseudogenes	not annotated in SEN or SEEG; pseudogene in STM; frameshift
STM3825	SEEG9184_RS00075	SEN_RS18905	same	torT	All ORFs intact	TMAO reductase system periplasmic protein
STM3914	SEEG9184_RS03405	SEN_RS19330	same	rhlB	All ORFs intact	ATP-dependent RNA helicase RhlB
STM4039	SEEG9184_RS23265	SEN_RS24515	same	unnamed	SEN, SEEG pseudogenes	HO protein
STM4076	SEEG9184_RS02975	SEN_RS20130	same	ydeZ/lsrD	All ORFs intact	putative sugar transport protein
STM4086	SEEG9184_RS03025	SEN_RS20180	same	glpK	All ORFs intact	glycerol kinase
STM4196	SEEG9184_RS22850	SEN_RS20755	same	unnamed	All ORFs intact	HO protein
STM4260	SEEG9184_RS22640	SEN_RS20975	same	unnamed	All ORFs intact	predicted cation efflux pump; HLYD secretioncation transporter
STM4266-regulon	SEEG9184_RS22610_regulon	SEN_RS21005_regulon	same	soxR	na	redox-sensitive transcriptional activator SoxR: contains iron-sulfur center for redox-sensing (MerR family)
STM4314	SEEG9184_RS22360	SEN_RS21255	same	unnamed	All ORFs intact	putative luxR family bacterial regulatory
STM4401	SEEG9184_RS21915	SEN_RS21705	same	ytfG	All ORFs intact	NAD(P)-dependent oxidoreductase
STM4485	SEEG9184_RS21575	SEN_RS22030	same	idnK	All ORFs intact	gluconate kinase
STM4585	SEEG9184_RS21020	SEN_RS22580	same	gpmB	All ORFs intact	phosphoglycerate mutase
STM4585_regulon	SEEG9184_RS21020_regulon	SEN_RS22580_regulon	same	gpmB	na	phosphoglycerate mutase regulon
STM4600	SEEG9184_RS20940	SEN_RS22655	same	lasT	All ORFs intact	tRNA/rRNA methyltransferase

[1] Results are from comparing 3 genomes, namely S. Typhimurium NC_003197.2 (STM), S. Enteritidis NC_011294.1 (SEN), and S. Gallinarum CP019035.1 (SEEG).

[2] Genes of STM that vary in AT 8+mer sequences as compared to SEEG and SEN are listed in both sections A and B.

[3] Abbreviations: na, not applicable; ORF, open reading frame.

 doi:10.20944/preprints202106.0384.v1

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 1 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 transfactor 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: PMCPMC1069516.
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: PMCPMC5574713.
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, Haesebrouck F, Van Immerseel F. The *Salmonella* Enteritidis lipopolysaccharide biosynthesis gene rfbH is required for survival in egg albumen. Zoonoses 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, Henzler DJ, Guard-Petter J. Lipopolysaccharide O-chain microheterogeneity of *Salmonella* serotypes Enteritidis and Typhimurium. Environ Microbiol. 2001;3(5):332-42. Epub 2001/06/26. doi: 10.1046/j.1462-2920.2001.00200.x. PubMed PMID: 11422320.
20. Huang K, Fresno AH, Skov S, Olsen JE. Dynamics and Outcome of Macrophage Interaction Between *Salmonella* Gallinarum, *Salmonella* Typhimurium, and *Salmonella* Dublin and Macrophages From Chicken and Cattle. Front Cell Infect Microbiol. 2019;9:420. Epub 2020/01/31. doi: 10.3389/fcimb.2019.00420. PubMed PMID: 31998655; PubMed Central PMCID: PMCPMC6966237.
21. Foley SL, Nayak R, Hanning IB, Johnson TJ, Han J, Ricke SC. Population dynamics of *Salmonella enterica* serotypes in commercial egg and poultry production. Appl Environ Microbiol. 2011;77(13):4273-9. Epub 2011/05/17. doi: 10.1128/AEM.00598-11. PubMed PMID: 21571882; PubMed Central PMCID: PMCPMC3127710.
22. Branchu P, Bawn M, Kingsley RA. Genome Variation and Molecular Epidemiology of *Salmonella enterica* Serovar Typhimurium Pathovariants. Infect Immun. 2018;86(8). Epub 2018/05/23. doi: 10.1128/IAI.00079-18. PubMed PMID: 29784861; PubMed Central PMCID: PMCPMC6056856.
23. McMillan EA, Gupta SK, Williams LE, Jove T, Hiott LM, Woodley TA, et al. Antimicrobial Resistance Genes, Cassettes, and Plasmids Present in *Salmonella enterica* Associated With United States Food Animals. Front Microbiol. 2019;10:832. Epub 2019/05/07. doi: 10.3389/fmicb.2019.00832. PubMed PMID: 31057528; PubMed Central PMCID: PMCPMC6479191.
24. Desai PT, Porwollik S, Long F, Cheng P, Wollam A, Bhonagiri-Palsikar V, et al. Evolutionary Genomics of *Salmonella enterica* Subspecies. mBio. 2013;4(2). Epub 2013/03/07. doi: 10.1128/mBio.00579-12. PubMed PMID: 23462113; PubMed Central PMCID: PMCPMC3604774.
25. Achtman M, Hale J, Murphy RA, Boyd EF, Porwollik S. Population structures in the SARA and SARB reference collections of *Salmonella enterica* according to MLST, MLEE and microarray hybridization. Infect Genet Evol. 2013;16:314-25. Epub 2013/03/20. doi: 10.1016/j.meegid.2013.03.003. PubMed PMID: 23507027.
26. Turcotte C, Woodward MJ. Cloning, DNA nucleotide sequence and distribution of the gene encoding the SEF14 fimbrial antigen of *Salmonella* enteritidis. J Gen Microbiol. 1993;139(7):1477-85. Epub 1993/07/01. doi: 10.1099/00221287-139-7-1477. PubMed PMID: 8371111.
27. Clouthier SC, Collinson SK, Kay WW. Unique fimbriae-like structures encoded by sefD of the SEF14 fimbrial gene cluster of *Salmonella* enteritidis. Mol Microbiol. 1994;12(6):893-901. Epub 1994/06/01. doi: 10.1111/j.1365-2958.1994.tb01077.x. PubMed PMID: 7934897.
28. Matthews TD, Schmieder R, Silva GG, Busch J, Cassman N, Dutilh BE, et al. Genomic Comparison of the Closely-Related *Salmonella* enterica Serovars Enteritidis, Dublin and Gallinarum. PLoS One. 2015;10(6):e0126883. Epub 2015/06/04. doi: 10.1371/journal.pone.0126883. PubMed PMID: 26039056; PubMed Central PMCID: PMCPMC4454671.
29. Sayers EW, Agarwala R, Bolton EE, Brister JR, Canese K, Clark K, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2019;47(D1):D23-D8. Epub 2018/11/06. doi: 10.1093/nar/gky1069. PubMed PMID: 30395293; PubMed Central PMCID: PMCPMC6323993.
30. McClelland M, Sanderson KE, Spieth J, Clifton SW, Latreille P, Courtney L, et al. Complete genome sequence of *Salmonella enterica* serovar Typhimurium LT2. Nature. 2001;413(6858):852-6. Epub 2001/10/26. doi: 10.1038/35101614. PubMed PMID: 11677609.
31. Thomson NR, Clayton DJ, Windhorst D, Vernikos G, Davidson S, Churcher C, et al. Comparative genome analysis of *Salmonella* Enteritidis PT4 and *Salmonella* Gallinarum 287/91 provides insights into evolutionary and host adaptation pathways. Genome Res. 2008;18(10):1624-37. Epub 2008/06/28. doi: 10.1101/gr.077404.108. PubMed PMID: 18583645; PubMed Central PMCID: PMCPMC2556274.

32. Allard MW, Luo Y, Strain E, Pettengill J, Timme R, Wang C, et al. On the evolutionary history, population genetics and diversity among isolates of *Salmonella* Enteritidis PFGE pattern JEGX01.0004. PLoS One. 2013;8(1):e55254. Epub 2013/02/06. doi: 10.1371/journal.pone.0055254. PubMed PMID: 23383127; PubMed Central PMCID: PMCPMC3559427.
33. Santangelo TJ, Cubonova L, Skinner KM, Reeve JN. Archaeal intrinsic transcription termination in vivo. J Bacteriol. 2009;191(22):7102-8. Epub 2009/09/15. doi: 10.1128/JB.00982-09. PubMed PMID: 19749050; PubMed Central PMCID: PMCPMC2772485.
34. Yamada S, Matsumoto Y, Takashima Y, Otsuka H. Mutation hot spots in the canine herpesvirus thymidine kinase gene. Virus Genes. 2005;31(1):107-11. Epub 2005/06/21. doi: 10.1007/s11262-005-2206-y. PubMed PMID: 15965615.
35. Koscielniak D, Wons E, Wilkowska K, Sektas M. Non-programmed transcriptional frameshifting is common and highly RNA polymerase type-dependent. Microb Cell Fact. 2018;17(1):184. Epub 2018/11/27. doi: 10.1186/s12934-018-1034-4. PubMed PMID: 30474557; PubMed Central PMCID: PMCPMC6260861.
36. 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;108(6):697-710. Epub 2018/04/01. doi: 10.1111/mmi.13959. PubMed PMID: 29603442.
37. 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: PMCPMC7854157.
38. Achtman M, Zhou Z, Alikhan NF, Tyne W, Parkhill J, Cormican M, et al. Genomic diversity of *Salmonella enterica* -The UoWUCC 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.
39. 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: PMCPMC5886390.
40. Dong HJ, Cho S, Boxrud D, Rankin S, Downe F, Lovchik J, et al. Single-nucleotide polymorphism typing 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: PMCPMC5412032.
41. Wilson MR, Brown E, Keys C, Strain E, Luo Y, Muruvanda T, et al. Whole Genome DNA Sequence Analysis of *Salmonella* subspecies enterica serotype Tennessee obtained from related peanut butter foodborne outbreaks. PLoS One. 2016;11(6):e0146929. Epub 2016/06/04. doi: 10.1371/journal.pone.0146929. PubMed PMID: 27258142; PubMed Central PMCID: PMCPMC4892500.
42. 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.
43. 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.
44. 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.
45. 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.
46. Zhou Z, McCann A, Litrup 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.
47. 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.
48. 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.
49. 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: 29312269; PubMed Central PMCID: PMCPMC5744181.
50. 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.
51. 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.