New discoveries from closing *Salmonella* genomes using Pacific Biosciences continuous long reads

Marc W. Allard[1]*, Cary Pirone[1], Tim Muruvanda[1], Maria Hoffmann[3], Ángel A. Soler-García1, Charles Wang[1], Errol Strain[2], Ruth Timme[1], Justin Payne1, Yan Luo[2], Christine E. Keys[1], Chen-Shan Chin[4], Jonas Korlach[4], Steven M. Musser[1], Shaohua Zhao[3], Robert Stones[5], Richard J. Roberts[6], Peter Evans[1], and Eric W. Brown[1]

[1] Office of Regulatory Science, Center for Food Safety and Applied Nutrition, U.S. Food and Drug Administration, College Park, MD, 20742, USA
[2] Office of Food Defense, Communications, and Emergency Response, Center for Food Safety & Applied Nutrition, U.S. Food & Drug Administration, College Park, MD 20740 USA;
[3] Center for Veterinary Medicine, US Food and Drug Administration, Laurel, MD, USA; [4] Pacific Biosciences, Menlo Park, California, USA;
[5] Food and Environment Research Agency, Sand Hutton, York, YO41 1LZ, UK5; 6New England Biolabs, Inc., Ipswich, Massachusetts, USA

Abstract

The newer hierarchical genome assembly process (HGAP) performs de novo assembly using data from a single PacBio® long insert library. To assess the benefits of this method, DNA from several *Salmonella* enterica serovars was isolated from a pure culture. Genome sequencing was performed using Pacific Biosciences RS sequencing technology. The HGAP process enabled us to close sixteen *Salmonella* subsp. *enterica* genomes and their associated mobile elements: The ten serotypes include: *Salmonella enterica* subsp. *enterica* serovar Enteritidis (S. Enteritidis) S. Bareilly, S. Heidelberg, S. Cubana, S. Javiana and S. Typhimurium, S. Newport, S. Montevideo, S. Agona, and S. Tennessee. In addition, we were able to detect novel methyltransferases (MTases) by using the Pacific Biosciences kinetic score distributions showing that each serovar appears to have a novel methylation pattern. For example while all *Salmonella* serovars examined so far have methylase specific activity for 5'-GATC-3'/3'-CTAG-5' and 5'-CAGAG-3'/3'-GTCTC-5' (underlined base indicates a modification), S. Heidelberg is uniquely specific for 5'-ACCANCC-3'/3'-TGGTNGG-5', while S. Typhimurium has uniquely methylase specific for 5'-GATCAG-3'/3'-CTAGTC-5' sites, for the samples examined so far. We believe that this may be due to the unique environments and phages that these serotypes have been exposed to.

Furthermore, our analysis identified and closed a variety of plasmids such as mobilization plasmids, antimicrobial resistance plasmids and IncX plasmids carrying a Type IV secretion system (T4SS). The VirB/D4 T4SS apparatus is important in that it assists with rapid dissemination of antibiotic resistance and virulence determinants. Presently, only limited information exists regarding the genotypic characterization of drug resistance in S. Heidelberg isolates derived from various host species. Here, we characterize two S. Heidelberg outbreak isolates from two different outbreaks. Both isolates contain the IncX plasmid of approximately 35 kb, and carried the genes *vir*B1, *vir*B2, *vir*B3/4, *vir*B5, *vir*B6, *vir*B7, *vir*B8, *vir*B9, *vir*B10, *vir*B11, *vir*D2, and *vir*D4, that are associated with the T4SS. In addition, the outbreak isolate associated with ground turkey carries a 4,473 bp mobilization plasmid and an incompatibility group (Inc) I1 antimicrobial resistance plasmid encoding resistance to gentamicin (*aac*C2), beta-lactam (*bl2b_*tem), streptomycin (*aad*AI) and tetracycline (*tetA, tetR*) while the outbreak isolate associated with chicken breast carries the IncI1 plasmid encoding resistance to gentamicin (*aac*C2), streptomycin (*aad*AI) and sulfisoxazole (*sul*1). Using this new technology we explored the genetic elements present in resistant pathogens which will achieve a better understanding of the evolution of *Salmonella*.

Methods

Single long read libraries (~10,000 bp) were sequenced on Pacific Biosciences RS Sequencer. De novo assembly of sequences was performed by the hierarchical genome assembly process (HGAP) into a single circular chromosome and associated plasmids. Sequences were annotated using the NCBI Prokaryotic Genomes Automatic Annotation Pipeline (PGAAP).

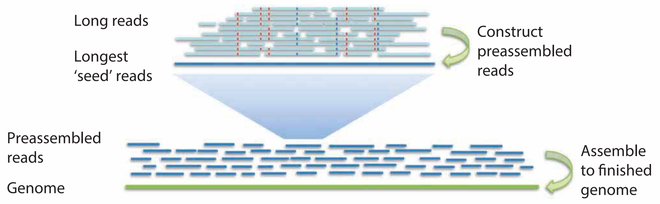


Figure 1. The HGAP assembly process. Diagram from Chin et al., 2013

Methylated bases were detected based on their kinetic signatures and motifs were identified using SMRT Analysis Software. Kinetic signatures were assigned based on IPD (time between successive pulses) and ID (interpulse duration).

MTase genes were identified in genome sequences via homology searches using in-house software and REBASE. Motif specificities were assigned to each putative homologue based on comparison of sequence alignments to genes of known specificities.

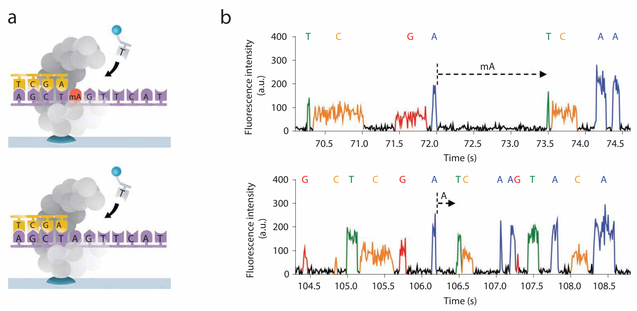


Figure 2. Detection of methylated bases during SMRT sequencing a. The addition of a fluorescence-tagged base by a polymerase b. Fluorescence signals over time. Note the lengthened duration between successive signals after the polymerase passes a methylated base (upper panel) vs. a non-methylatebase (lower panel). Diagram taken from Flusberg et al., 2010.

Phenotypic patterns were determined using the GEN III plates (Biolog, Inc., Hayward, CA) which contain various growth compounds (see Table 2). S. Enteritidis (SE) isolates were grown in BUG agar, then suspended in inoculated fluid A and transferred to the GEN plate. Optical density was adjusted to 90% transmittance. Growth kinetic was monitored for 36 hours incubated at 35 degrees.

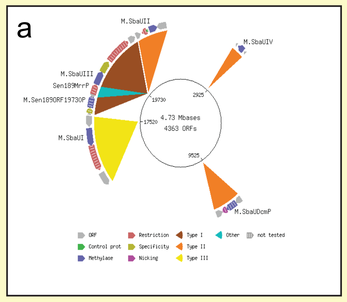
Results

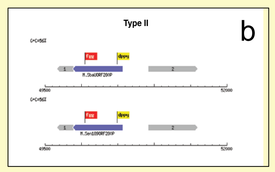
We identified 12 methylated motifs across 6 *Salmonella* serovars. Methylase specificity is unique to each serovar, and also varies within the serovars S. Heidelberg and S. Typhimurium.

strain	S.Bareilly	S. Heidelberg			S.Javiana	S.Cubana	S.Typhimurium				S.Enteriditis			
Methylase specificity	189	318	2069	2064	1992	2050	1921	1158	3511	2049	6	45	111	158
5'-GATC-3'/3'-CTAG-5'														
5'-CAGAG-3'/3'-GTCTC-5'														
5'-ATGCAT-3'/3'-TACGTA-5'														
5'-CAGCTG-3'/3'-GTCGAC-5'														
5'-GATCAG-3'/3'-CTAGTC-5'														
5'-ACCANCC-3'/3'-TGGTNGG-5'														
5'-CCGAN5GTC-3'/3'-GGCTN5CAG-5'														
5'-GAGN6RTAYG-3'/3'-CTCN6YATRC-5'														
5'-GN2TAYN5RTGG-3'/3'-CN2ATRN5YACC-5'														
5'-RAACN5TGA-3'/3'-YTTGN5ACT-5'														
5'-GGAN6ATTA-3'/3'-CCTN6TAAT-5'														
5'-GGYAN6TCG3'/3'-CCRTN6AGC-5'														

Partial Modification
Complete Modification

Table 1. Methylation Motifs detected during SMRT Sequencing of 6 *Salmonella* serovars. Numbers below serovar name represent different strains.





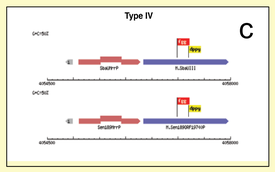


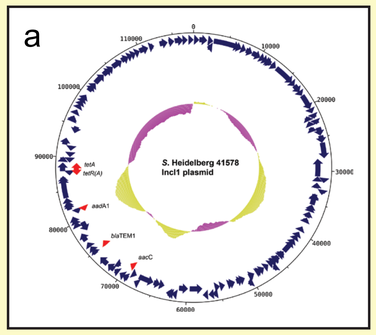
Figure 3. MTases identified in *Salmonella* Bareilly (S. Bareilly).
a. Overview of chromosomal MTases
b. A type II MTase from a plasmid. fgg = the generic sequence motif found in the SAM binding site, dppy = the generic sequence motif in the catalytic site
c. A novel type IV MTase from the chromosome

MTase sequences can be downloaded from REBASE http://rebase.neb.com/rebase/rebase.html

CFSAN Number	6	45	111	158
Clade	9	3	6	1
Isolate type	Environmental	Environmental	Environmental	Clinical
Methylation site	**Profile**			
5'-GATC-3'/3'-CTAG-5'				
5'-CAGAG-3'/3'-GTCTC-5'				
5'-ATGCAT-3'/ 3'-TACGTA-5'				
5'-GGYAN6TCG3'/ 3'-CCRTN6AGC-5'				
5'-CCAGGAAY-3'/ 3'-GGTCCTTR-5'				
Compound or Growth Condition	**Metabolic Profile**			
Dextrin	3	2	2	1
D-Maltose	3	2	2	1
D-Trehalose	3	2	2	1
pH 6	4	3	3	2
pH 5	4	3	3	2
D-Raffinose	3	2	2	1
N-Acetyl-D-Glucosamine	3	2	2	1
N-Acetyl-D-Mannosamine	3	2	2	1
8% NaCl	3	2	2	1
D-Mannose	3	2	2	1
D-Galactose	3	2	2	1
L-Fucose	3	2	2	1
1% Sodium Lactate	4	3	3	2
D-Sorbitol	3	2	2	1
D-Aspartic Acid	3	2	2	1
Troleandomycin	4	3	3	2
Rifamycin SV	4	3	3	2
Glycyl-L-Proline	3	2	2	1
L-Alanine	3	2	2	1
L-Glutamice Acid	3	2	2	1
Lincomycin	4	3	3	2
Nlaproff 4	5	3	3	2
Glucoronamide	3	2	2	1
Vancomycin	4	3	3	2
D-Malic Acid	2	1	1	0
Nalidixic Acid	4	1	1	0
Tween 40	3	2	2	1
α-Hydroxy Butyric Acid	3	2	2	1
α-Keto-Butyric Acid	3	2	2	1
Propionic Acid	4	2	2	1
Sodium Bromate	3	2	2	1
Phage	P125109	RE-2010	RE-2010	RE-2010
Plasmid	PSEEE-0956_35	PSEEE-0956_35 PSEE-3072_19	PSEEE-0956_35	PSEEE-0956_35 PSEE-1729_15

Table 2. Methylation, phenotypic, and phage profiles of selected S. Enteritidis isolated from egg-related outbreaks. Numbers represent amount of overall growth. Red represents complete methylation and blue represents partial methylation. Note how phenotypic metabolic profiles correspond to methylation pattern for one variant C methyl site (CCAGGAAY) among four S. Enteritidis isolates examined. Molecular biology experiments are ongoing to confirm preliminary work. These S. Enteritidis shell egg isolates were initially published in Allard et al. 2013a. Phages and plasmids also vary with strain, and may affect methylation.

Long read technology has allowed us to sequence and close mobilization plasmids, antimicrobial resistance plasmids, and virulence plasmids.



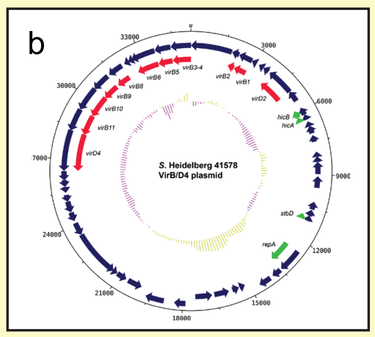


Figure 4. Plasmids from S. Heidelberg
a. IncI1 antimicrobial resistance plasmid: The coding sequences (CDS) are shown in dark blue and resistance genes are shown in red.
b. VirB/D4 virulence plasmid: The coding sequences (CDS) are shown in dark blue, genes from the Type IV secretion system are shown in red, genes responsible for plasmid stability and replication are shown in green.

Significance

Most techniques for methylation profiling are limited to the detection of 5-methylcytosine (5mC). SMRT sequencing can detect various methylated bases, but can best detect 6-methyladenine (6mA), the most common methylated base in bacteria. Several recent studies have used this technique to document methylation patterns and explore their biological significance in bacteria (Fang et al., 2013; Murray et al., 2013). More studies are needed and we intend to sequence hundreds of additional enteric pathogens over the coming year with an emphasis on *Salmonella*.

Here, we identify the methylation motifs and associated MTases in 14 strains from 6 serovars of *Salmonella*. Some of these motifs are common, including CAG^{m6}AG, C^{m6}AGCTG, ATGC^{m6}AT, and G^{m6}ATC. However, some motifs and their associated MTases are novel and merit further study. For example, we found a Type I MTase in S. Bareilly which forms CCG*m6*ANNNNNGTC. This is the first enzyme with this specificity to be characterized, but four identical genes are present in other Salmonella genomes. We also found a Type IIG combination MTase and REase that forms GATC^{m6}AG which has not been previously characterized. The function of these and other MTases in *Salmonella* are unknown, but will be interesting to explore. It is well established that MTases form part of restriction modification (RM) systems that protect bacterial cells from foreign DNA, and recent studies also demonstrate that methylation can regulate chromosome replication, cell cycle control, gene expression, and pathogenicity (Davis et al., 2013, Fang et al., 2013).

We also show that methylase specificity varies both within and between serotypes of *Salmonella* (Tables 1 and 2). The cause of variation is unknown, but our preliminary data (Table 2) suggests methylation may correlate with phenotypic profile. Methylation may also vary due to environmental effects, or with the introduction of novel MTases via phages and/or plasmids.

We closed 16 strains of *Salmonella* and their associated mobile elements. Virulence and resistance functions are often encoded in extrachromosomal plasmids. Thus, sequencing of these plasmids provides evidence of molecular level changes which influence virulence, resistance, and pathogenicity which is critical to prevent future outbreaks of pathogenic foodborne bacteria. When examined in a phylogenetic context, these data also broaden our understanding of *Salmonella* evolution.

These collaborations have effectively integrated CFSAN genomics into the National Antimicrobial Resistance Monitoring System (NARMS). Long read sequencing technology will assist us to investigate the evolution of antimicrobial resistance in zoonotic pathogens and the resistance associated with mobile genetic elements. This information is necessary to inform FDA's policy and regulatory decision making on antimicrobial use in animals and the process of how foodborne pathogens are entering into the food supply. Bioinformatics analysis and experimental result will help show the genes responsible for resistance, and we expect to discover novel genes and novel linkages that have not been described previously. Many more such discoveries will come from these *Salmonella* isolates such as the new antimicrobial gene discovered in *Campylobacter coli* (Chen et al., 2013).

In collaboration with the 100k pathogen genome project we have released several closed *Salmonella* genomes. http://www.ncbi.nlm.nih.gov/bioproject/186441 We have also published the genome of S. Javiana (Allard et al., 2013b).

To read more about FDA's *Salmonella* genomics efforts see below:
Genometrakr network where we have recently released >500 unpublished draft genomes for food safety into the SRA database.
http://www.ncbi.nlm.nih.gov/bioproject/183844
We also recently opened a public access url describing FDA's whole genome sequencing program.
http://www.fda.gov/Food/FoodScienceResearch/WholeGenomeSequencingProgramWGS/default.htm

References

- Allard, M.W., Luo,Y., Strain, E., Pettengill, J., Timme, R., Wang, C., Li, C., Keys, C.E., Zheng, J., Stones, R., Wilson, M., K., Musser, S.E., and E. W. Brown. 2013. On the Evolutionary History, Population Genetics and Diversity among Isolates of *Salmonella* Enteritidis PFGE Pattern JEGX01.0004. PLOS ONE 8(1): e55254
- Allard, M.W., Muruvanda, T., Strain, E., Timme, R., Luo, Y., Wang, C., Keys, C.E., Payne, J., Cooper, T., Luong, K., Song, Y., Chin, C.-S., Korlach, J., Roberts, R.J., Evans, P., Musser, S.M., and E.W. Brown. 2013. Fully Assembled Genome Sequence for *Salmonella enterica* subsp. *enterica* Serovar Javiana CFSAN001992. Genome Announcements 1(2): e00081-13.
- Chen, Y., Mukherjee, S.,Hoffmann, M., Kotewicz, M.L., Young, S., Abbott, J., Luo, Y., Davidson, M.K., Allard, M., McDermott, P., and S. Zhao. 2013. Whole Genome Sequencing of a Gentamicin-Resistant Campylobacter coli isolated from United States Retail Meats Reveals Novel Plasmid Mediated Aminoglycoside Resistance 3 Genes. Antimicrob. Agents Chemother. Aug 19. [Epub ahead of print]
- Davis, B., Chao, M.C., and M. K. Waldor. 2013. Entering the era of bacterial epigenomics with single molecule real time DNA sequencing. Current Opinion in Microbiology 16: 192-198.
- Fang, G., Munera, D., Friedman, D., Mandlik, A., Chao, M., Banerjee, O., Feng, Z., Losic, B., Mahajan, M., Jabado, O., Deikus, G., Clark, T., Luong, K., Murray, I., Davis, B., Keren-Paz, A., Chess, A., Roberts, R., Korlach, J., Turner, S., Kumar, V., Waldor, M., and E. Schadt. 2013. Genome-wide mapping of methylated adenine residues in pathogenic Escherichia coli using single-molecule real-time sequencing. Nature Biotechnology 30 (12): 1232-1242.
- Hoffman, M., Luo, Y., Lafon, P.C., Timme, R., Allard, M.W., McDermott, P.F., Brown, E.W., and S. Zhaoa. 2013. Genome Sequences of *Salmonella enterica* Serovar Heidelberg Isolates Isolated in the United States from a Multistate Outbreak of Human *Salmonella* Infections. Genome Annoucements 1(1): 00004-12
- Murray, I., Clark, T., Morgan, R., Boitano, M., Anton, B., Luong, K., Fomenkov, A., Turner, S., Korlach, J., and R. Roberts. 2012. The methylomes of six bacteria. Nucleic Acids Research 40 (22): 11450–11462

Acknowledgements: We thank Pat McDermott and Shaohua Zhao. for access to NARMS isolates. Please direct any queries to our strain curator Dwayne Roberson, at Dwayne.Roberson@fda.hhs.gov