Geography Shapes the Population Genomics of *Salmonella enterica* Dublin

Gavin J. Fenske[1], Anil Thachil[2], Patrick L. McDonough[2], Amy Glaser[2], and Joy Scaria[1,*]

[1]Department of Veterinary and Biomedical Sciences, South Dakota State University

[2]Department of Population Medicine and Diagnostic Sciences, Cornell University

*Corresponding author: E-mail: joy.scaria@sdstate.edu.

Accepted: July 16, 2019

Abstract

Salmonella enterica serotype Dublin (*S.* Dublin) is a bovine-adapted serotype that can cause serious systemic infections in humans. Despite the increasing prevalence of human infections and the negative impact on agricultural processes, little is known about the population structure of the serotype. To this end, we compiled a manually curated data set comprising of 880 *S.* Dublin genomes. Core genome phylogeny and ancestral state reconstruction revealed that region-specific clades dominate the global population structure of *S.* Dublin. Strains of *S.* Dublin in the UK are genomically distinct from US, Brazilian, and African strains. The geographical partitioning impacts the composition of the core genome as well as the ancillary genome. Antibiotic resistance genes are almost exclusively found in US genomes and are mediated by an IncA/C2 plasmid. Phage content and the *S.* Dublin virulence plasmid were strongly conserved in the serotype. Comparison of *S.* Dublin to a closely related serotype, *S. enterica serotype* Enteritidis, revealed that *S.* Dublin contains 82 serotype specific genes that are not found in *S. Enteritidis*. Said genes encode metabolic functions involved in the uptake and catabolism of carbohydrates and virulence genes associated with type VI secretion systems and fimbria assembly respectively.

Key words: *Salmonella* Dublin, genomic epidemiology, antimicrobial resistance, *Salmonella enteritidis*, pangenome, phylogeography.

Introduction

Salmonella enterica serotype Dublin (*S.* Dublin) is a host-adapted serotype of *S. enterica* that is primarily associated with cattle. In contrast to many enteric diseases affecting cattle that are presented primarily with diarrhea in young calves, *S.* Dublin infection can manifest as both enteric and systemic forms in older calves (Wallis et al. 1995). When cattle ingest sufficient infectious dose of *S.* Dublin, typically $>10^6$ CFU's (Nielsen 2013), *S.* Dublin could colonize the gut of the animal. After colonization, *S.* Dublin invades enteric cells in the ileum and jejunum and subsequently traverses to the mesenteric lymph nodes ultimately causing systemic infection (Pullinger et al. 2007). It has been shown that a virulence plasmid carried by *S.* Dublin is partly responsible for the systemic phase of the infection; removal of the plasmid or the Salmonella plasmid virulence (spv) genes carried upon the plasmid attenuates systemic infections (Wallis et al. 1995; Libby et al. 1997). Analogous to Salmonella Typhi infections in humans, *S.* Dublin is known to establish a carrier state in susceptible cattle. Carrier animals harbor the bacteria in internal organs and lymph areas and sporadically shed *S.* Dublin through feces and milk (Nielsen et al. 2004). Such carriers tend to help to maintain *S.* Dublin infection rates in local dairy herds and cases of human infections after drinking raw milk contaminated with the pathogen have been documented (Small and Sharp 1979; Werner et al. 1979; Maguire et al. 1992). The duration and severity of shedding is highly variable between animals. Some animals may begin shedding *S.* Dublin in feces as soon as 12–48 h after infection (Nielsen 2013). Shedding has been detected up to six months after the initial discovery that an animal is a carrier (Nielsen et al. 2004).

Current genomics of *S.* Dublin has primarily focused upon the identification of antimicrobial resistance (AMR) homologues and mobile genetic elements such as prophages and plasmids (Langridge et al. 2015; Matthews et al. 2015; Carroll et al. 2017). *S.* Dublin genome diversification appears to be driven by horizontal gene transfer and genome degradation resulting in pseudogenes (Langridge et al. 2015). However, many of these studies focused on a smaller set of *S.* Dublin, typically <30 genomes, and focused on comparisons to closely related serotypes. The population structure of *S.* Dublin has yet to be resolved, especially regarding isolates from outside of the US. Because of the importance of the pathogen in animal agriculture and human health, establishing the population structure and pangenome of the serotype would provide valuable insight into the evolution of *S.* Dublin. Phylogeographical clustering is evident in the population of *S.* Dublin and impacts the composition of the core and ancillary genomes.

Materials and Methods

Genome Sequencing and Comparative Data Set

We sequenced 43 *S.* Dublin clinical isolates that were collected by the Animal Disease Research Laboratory (ADRDL, South Dakota State University) and 33 isolates collected by the Animal Health Diagnostic Center (AHDC, Cornell University). Strains were grown aerobically in Luria Bertani broth at 37 °C for 12 h. DNA was isolated from resultant pellets using the DNeasy Blood and Tissue Kit (Qiagen, Hilden, Germany). Paired-end sequencing was conducted using the Illumina MiSeq platform and 250 base paired V2 chemistry. For the comparative genome analysis and the construction of global population structure, 1,020 publicly available *S.* Dublin genome data were downloaded from the NCBI Sequence Read Archive (SRA). Raw sequence data as well as the metadata tables were downloaded and manually parsed to include samples that contained a positive *S.* Dublin serotype that were sequenced using the Illumina platform. The prefetch utility (SRA toolkit) was used to download the SRA files which were written into paired-end fastq files with the fastq-dump tool (SRA toolkit).

Genome Assembly and Validation

Paired-end reads were assembled into contigs using Shovill (Torsten Seemann) given the following parameters: Minimum contig length 200, depth reduction 100× and an estimated genome size of 4.8 Mbp. The Shovill pipeline is as follows: Read depth reduction per sample to ~100× of the estimated genome size; read sets below the 100× threshold are not affected by the reduction. After reduction, reads are conservatively error corrected with Lighter (Song et al. 2014). Spades (v3.12.0) (Bankevich et al. 2012) was used to generate the assembly using default parameters. After assembly, small indels and assembly errors are corrected using Pilon (Walker et al. 2014). Genome assemblies were passed to the software assembly-stats (https://github.com/sanger-pathogens/assembly-stats; last accessed September, 2018) to gauge basic assembly properties such as contig number, N50, and genome length. Samples were eliminated from the data set if the assemblies were fragmented, defined here as a contig number >300 or an N50 <25,000 bp.

Serotype Prediction

Genomes that passed assembly validation were submitted to serotype validation. The program SISTR (Yoshida et al. 2016) was download and run locally to validate the serotypes using both cgMLST, SISTR method, and Mash (Ondov et al. 2016). A positive Dublin serotype was defined as a confirmed Dublin prediction from both the Mash and cgMLST identification methods.

Pangenome Reconstruction

Curated genome assemblies were annotated using the software Prokka (Seemann 2014). A manually annotated reference *S. enterica* Typhimurium (ASM694v2) genbank file was downloaded (ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/Salmonella_enterica/reference/GCF_000006945.2_ASM694v2; last accessed September, 2018) and formatted to a Prokka database file. Said reference database was used to augment the existing Prokka databases and facilitated consistent nomenclature of core Salmonella genes. Resultant general feature files 3 (.gff) from Prokka were used as the input to the program Roary (Page et al. 2015). PRANK (Loytynoja 2014) was used within Roary to conduct the alignment of core genes.

Phylogeny Reconstruction

Two distinct methods were used to generate phylogenomic trees. In the first method, the core gene alignment file from Roary was passed to the software SNP-Sites (Page et al. 2016) using the flags –cb to discard gaps and include monomorphic sites. The final alignment file is 2,431,413 base pairs longs. Model-test NG (https://github.com/ddarriba/modeltest; last accessed November, 2018) was used to define a substitution model for phylogeny and was run to optimize a model for RAxML. Generalized Time Reversible (GTR) + G4 was the best scoring model and used to generate the maximum-likelihood tree. The interactive Tree of Life (iTOL) (Letunic and Bork 2016) was used to visualize phylogenomic trees. kSNP3 (Gardner et al. 2015) was the second method employed to generate a reference independent SNP phylogeny based upon kmers, rather than gene polymorphisms. The program was run to generate a fasta matrix based upon kmers found in 99% of genomes. Said matrix was passed to Model-test NG same as above and a maximum-likelihood tree was generated using RAxML (Stamatakis 2014) with the GTR model.

BEAST2 Phylogeny

The core gene alignment file from Roary was passed to SNP-Sites given the flags -cb to generate an alignment for BEAST2 (Bouckaert et al. 2014). BEAUti was used to generate the xml using a strict molecular clock and a constant coalescent population model. GTR+G4 was used for the nucleotide substitution model. The chain length was set to 10,000,000 sampling every 1,000 trees for the log file with a seed value of 33.

Antibiotic Resistance Homolog Prediction

Antibiotic resistance gene homologs were predicted in the genomes using the software package Abricate (Seemann 2018). The NCBI Bacterial Antimicrobial Resistance Reference Gene Database was used as a reference. Positive hits are defined here as homologs with >90% sequence identity and >60% of target coverage.

Plasmid Replicon Identification

Analogues to AMR homologue detection. Abricate was used to BLAST genomes assemblies against the PlasmidFinder (Carattoli et al. 2014) database for identification of plasmid replicons. Positive hits were defined as a sequence identity > 90% and at least 60% query coverage.

Identification of Prophage Regions

The web service PHASTER was used to identify prophage regions in the genomes (13, 14). As the number of genomes in the query was high, a bash script was used to submit genome assemblies via the API provided. Resulting text files from PHASTER were downloaded, concatenated, and parsed using R (R Core Team 2018). Prophage regions were considered if they were marked as "intact."

Data Analysis

Logistic PCA was conducted using the R package logisticPCA (17). Binary, 0 and 1, presence–absence matrices were prepared from the Roary output with the rows corresponding to genes and the columns corresponding to genomes. Genes with a prevalence >99% or <5% were removed to aid in computational speed and reduced confounding effects of misannotation. All other statistical analysis and plotting were conducted using R using the packages: ggplot2(Wickham 2009), ggtree (Yu et al. 2017), and ComplexHeatmap (Gu et al. 2016).

Results

S. Dublin Global Population Structure

To begin the investigation on *S.* Dublin, 74 isolates of *S.* Dublin were sequenced using the Illumina MiSeq platform. For comparative analysis, 1,020 publicly available *S.* Dublin genomes were downloaded from NCBI Sequence Read Archive (NCBI SRA). Genomes were assembled and subjected to a two-step validation process. The first validation step was to assess genome assembly quality; genome assemblies with >300 contigs or a N50 values <25,000 base pairs were discarded from the analysis set. The second validation step was serotype verification using the program SISTR (Yoshida et al. 2016). Genomes were retained in the data set if the serotype prediction using core genome multilocus sequence typing (cgMLST) and MASH (Ondov et al. 2016) agreed on a prediction of *S.* Dublin. After assembly and validation, a high-quality data set of 880 genomes were used for further analysis. Table 1 describes the full data set and the full metadata for the genome data set is provided in supplementary table 1, Supplementary Material online. On the basis of their origin, genomes are grouped into four major geographical regions: Africa (4%), Brazil (13%), the United Kingdom (20%), and the United States (62%). In terms of host species of isolation, human genomes are the largest constituent representing nearly 38% of genomes followed by bovine isolates (30%), food isolates (24%), and isolates from various sources or without metadata were classified as other (8%). Clear sampling bias is evident in the publicly available genome data set as more than half of the genomes retrieved are of US origin. Additionally, nearly 88% (231/262) of bovine isolates originate from the US.

Core genome variation was first investigated using core gene ($n = 4{,}098$) polymorphism trees (figs. 1 and 2, supplementary fig. 1, Supplementary Material online). Core genome phylogeny revealed strong geographical demarcation between genomes (fig. 1). Five major clades are seen in the phylogeny and correspond to the major geographical regions in the study: Africa, Brazil (two clades), the UK, and the US (fig. 2). The geographical clustering is conserved using both core gene (fig. 1), core kmer (supplementary fig. 1, Supplementary Material online), and ancestral state reconstruction through BEAST2 (Bouckaert et al. 2014; fig. 2). Figure 1 is rooted to a reference *S.* Enteritidis (AM933172). *S.* Enteritidis was chosen as an outgroup of *S.* Dublin based upon the serotype phylogeny provided by SISTR (https://lfz.corefacility.ca/sistr-app/; last accessed September, 2018). Some host

Table 1
Metadata Characteristics of 880 *S.* Dublin Comprising the Data Set

Region	Source	Num. Genomes
Africa	Bovine	7
	Food	2
	Human	26
Brazil	Bovine	24
	Human	89
	Other	4
United Kingdom	Food	23
	Human	99
	Other	56
United States	Bovine	231
	Food	189
	Human	118
	Other	11

Note.—The table is grouped into the four major geographical regions the genomes originate.

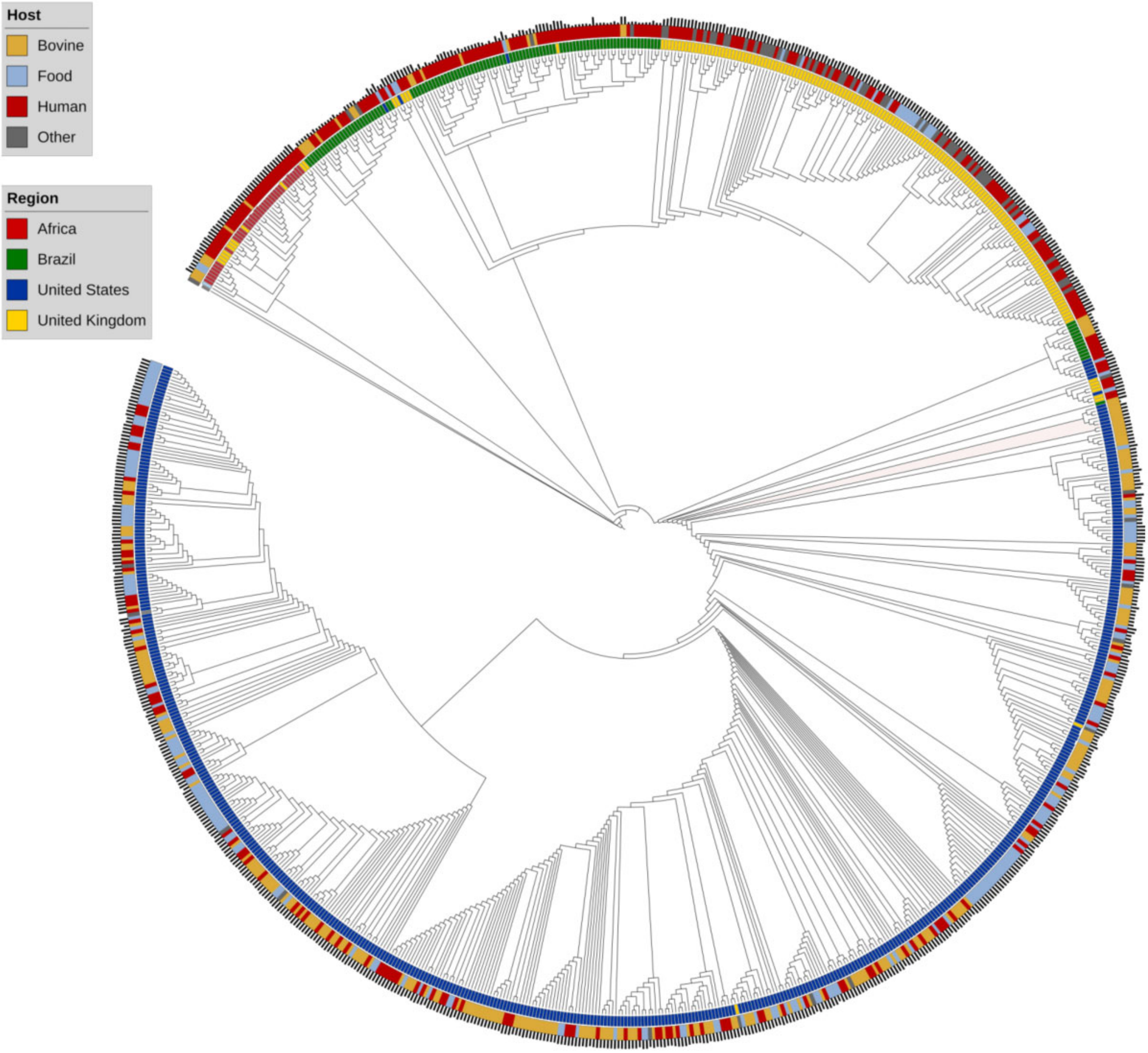


Fig. 1.—Global population structure of *S.* Dublin illustrated with a maximum-likelihood cladogram, GTR Gamma model, of 880 *S.* Dublin genomes rooted to *S.* Enteritidis (AM933172). The tree is inferred from an alignment of 4,098 core genes defined by Roary. Leaves are colored respective to the region of isolation. The outer colored ring denotes isolation source. The outermost barplots illustrate data of isolation and are scaled such that the higher the bar, the more recent the isolate was cultured and zeroed to a date of 1980. Genomes cluster into distinct clades associated with the area of isolation.

preference clustering is evident in the African and Brazilian clades as the genomes are predominately of human origin (outer colored ring, fig. 1). However, it is likely that such clustering is a consequence of sampling bias in the publicly available genome data sets (table 1) and is a by-product of geographical clustering. Examination of the US clade, which is roughly split between bovine, food, and human genomes reveals no monophyletic groups regarding host source. The same scenario is seen in the UK clade which is more balanced in terms of isolation source. Thus, the core genome sequence of *S.* Dublin is highly influenced by the region the genomes originates, and not the isolation host.

To further explore the population structure, ancestral state reconstruction was conducted and plotted in a simplified phylogeny (fig. 2). The five major clades observed using core gene maximum-likelihood methods are conserved and collapsed for clarity. Consistent with figure one, the tree is rooted to *S.* Enteritidis AM933172. The first divergence event in the phylogeny is the emergence of a single outgroup (SRR1122707) that diverges from all other *S.* Dublin genomes.

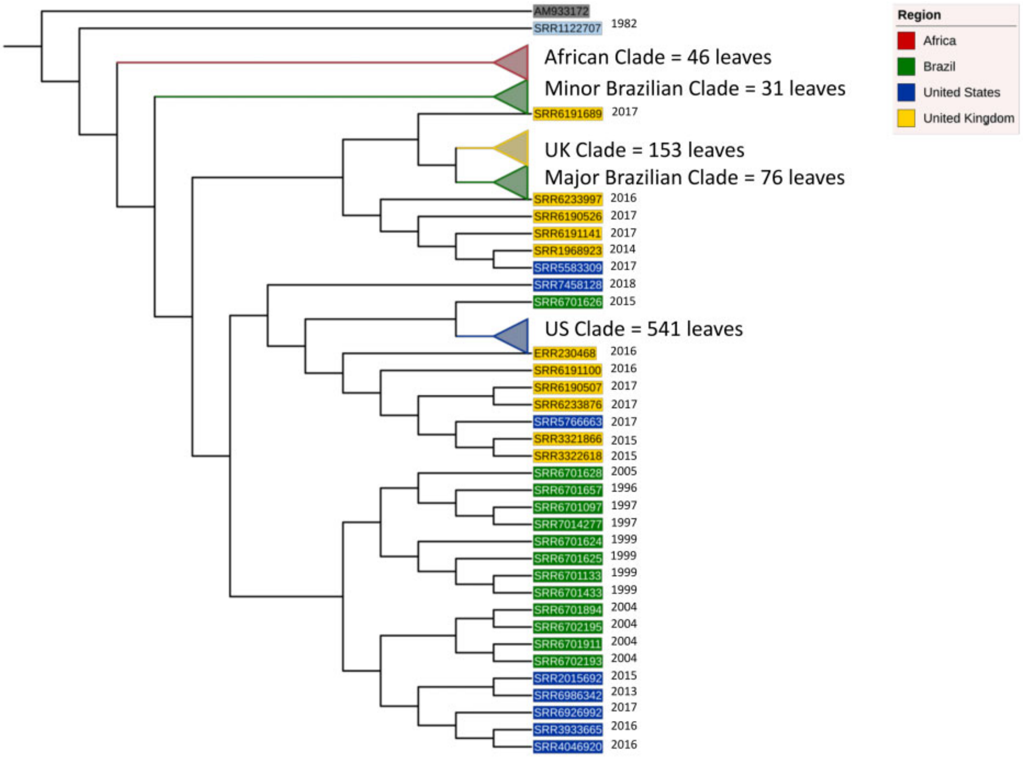


Fig. 2.—Reduced phylogeny of *S.* Dublin. Ancestral state reconstruction using BEAST2 on core genes conserved major geographical clades. Clades are collapsed to aid in visualization of high-order tree architecture. The phylogeny is rooted to *S.* Enteritidis AM933172. Clades and leaf labels are colored respective to the region of isolation. Isolation date is listed to the right of leaves not collapsed into clades.

The genome was isolated from a bovine source in 1982 in France. However, due to the sparsity of provenance, and the focus on population genomics, no conclusions can be made as to why the genome is divergent from the data set. It does pose an interesting possibility that a unique clade of *S.* Dublin is grossly undersampled in the current database. The next divergence event, or emergence, was the African clade, followed by the minor Brazilian clade. The final and largest divergence event yield two major clades, the US clade, and a mixed clade populated with the major Brazilian and UK clade. The two Brazilian clades emerged at different evolutionary time points revealing two distinct lineages that were present in Brazil. Additionally, the major Brazilian clade appears to have UK origins. The UK and Brazilian clade share a common ancestor with a UK genome SRR6191689. The monophyletic group, consisting of the UK, major Brazilian clade, and SRR6191689, diverged from a common ancestor shared by four UK and one US genomes. Thus, the major Brazilian clade was probably introduced to the country from the UK. However, such an explanation cannot be extended to the minor Brazilian clade, whose origin is unclear. The US clade shares a common ancestor with a single Brazilian genome, SRR6701626 and the combined clade shares a common ancestor with six UK genomes. At this time point, a definitive statement to the origin of the US genomes cannot be made.

Ancillary Genome Composition Is Geography Dependent

Core genome composition is more indicative of geographical location than host source. The next possibility we decided to investigate was the influence of geography on the ancillary or accessory genome (defined here as $5\% < \text{gene prevalence} < 99\%$). Genes with a prevalence less than five percent were excluded to minimize the confounding effects of improper open reading frame identification and sequencing error. Ancillary gene catalogues, consistent with the core genome composition, are influenced by geography (fig. 3). Logistic PCA, an extension to classical PCA used to reduce dimensionality in binary matrices, was used to plot genomes in a two-dimensional space respective of their ancillary genome content. Two large clusters of genomes are clearly represented in figure 3: The majority of US genomes cluster to the left of the plot whereas most of the global and a smaller number of US genomes clustering to the right. Figure 3 illustrates that US genomes harbor an ancillary genomic catalogue that readily distinguishes the genomes from isolates not originating from the US. Remarkably, subclustering within the right cluster separates genomes into the five clades witnessed in the core genome phylogeny (figs. 1 and 2). Ancillary genome composition, as core genome structure, is a geographical characteristic of *S.* Dublin genome.

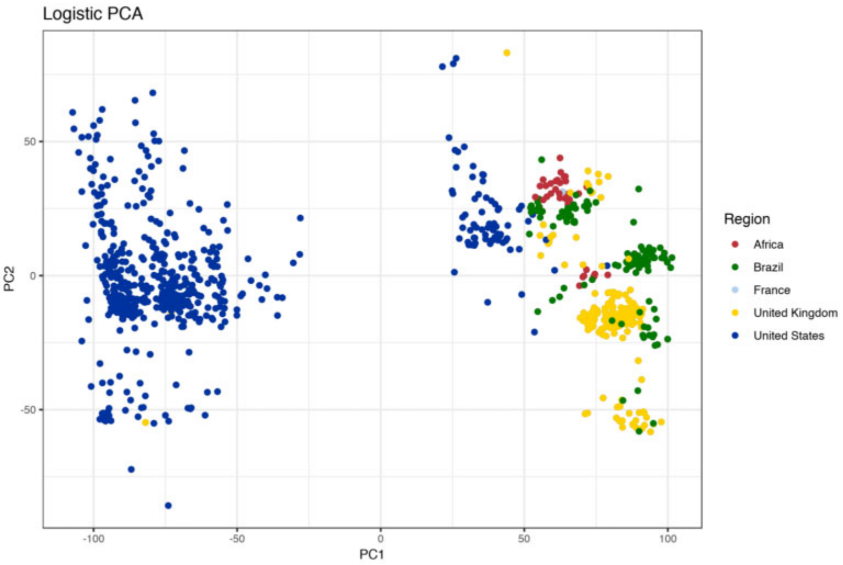


Fig. 3.—Ancillary genes cluster *S.* Dublin geographically. Logistic PCA was run on a binary matrix of ancillary genes, prevalence <99% and >5%. Region-specific clusters appear and correspond to the major geographical clades. Additionally, most of the US genomes cluster away from global genomes.

Further work was carried out to determine what genomic elements were responsible for region-specific differentiation of genomes. The pangenome of *S.* Dublin was constructed and is presented in figure 4*A*. Stated earlier, the core genome of *S.* Dublin is 4,098 genes. Genes with a prevalence of $5\% < x < 99\%$ (shell genes) numbered 833 genes. Lastly, genes with a prevalence of <5% (cloud genes) numbered 5,533 for a total pangenome size of 10,464 nonredundant genes. Interestingly, the number of core genes shared by 99% of genomes was nearly five times as great as genes shared between 15% and 99% of genomes. Such a disparity highlights high conservation of core genes. Put another way, much of the gene increase in the pangenome is due to the addition of unique coding sequences into a small number of genomes. However, a major exception to this statement can be seen in figure 4*A*. The shell pangenome is plotted as a binary matrix against the phylogeny of the serotype. A block of nearly 100 genes is clearly seen and is found only in US genomes. Said block of genes is responsible for the clustering pattern in figure 3 where the US genomes cluster away from global genomes. Examination of the aforementioned block showed genes pertaining to antibiotic resistance, toxin–antitoxin systems, and a large number of hypothetical or uncharacterized proteins. US genomes yield larger assemblies (fig. 3*B* and *C*) and more predicted open reading frames (fig. 3*D* and *E*) due to this gene block.

A possible explanation for the large increase in assembly size and number of open reading frames in US genomes is the acquisition of mobile genetic elements (prophages, plasmids, etc.). We arrived at such a hypothesis by the presence of toxin–antitoxin systems in the unique US gene block. To investigate the said possibility, prophage regions and plasmid replicons were identified in genomes using the web service PHASTER and the PlasmidFinder database (please see methods). Prophage insertions are not responsible for the increased gene number in US genes and are not a major diversifying agent in the serotype (fig. 5*A*, right panel). The full PHASTER phage details are provided in supplemental data set 2. Three major phages were identified in the *S. Dublin* queried: Gifsy 2, sal3, and RE 2010. The prevalence of the three major phages is >90% for all regions (fig. 5*B*) and no US specific or any region-specific pattern is present. However, plasmid replicons did yield a region-specific pattern. IncA/C2 replicon was identified only in US genomes. A representative contig containing IncA/C2 replication site was extracted from the genome assembly of SRR5000235. The sequence yielded a 99% sequence identity (BLAST+) to an IncA/C antibiotic resistance plasmid isolated from Salmonella Newport (CP009564.1). The other major replicon identified in *S.* Dublin was IncX1. Extracting the contig sequences with replicon and BLAST search yielded 99% sequence identity to the *S.* Dublin virulence plasmid (CP032450.1). The replicon is highly conserved among the genomes and is a common characteristic of *S.* Dublin. IncX1 was identified in 865 (98%) of *S.* Dublin genomes followed by IncFII(S) identified in 826 (94%) genomes and IncA/C2 identified in 476 (54%) genomes. Thus plasmids, not prophages, diversify *S.* Dublin. The large block of US genes, and resulting gene count and assembly length increase are due to the presence of an IncA/C2 resistance plasmid found only in US genomes.

AMR Is a US Phenomenon

Antibiotic resistance is a characteristic of *S.* Dublin in the United States, but not a characteristic of the serotype. Many

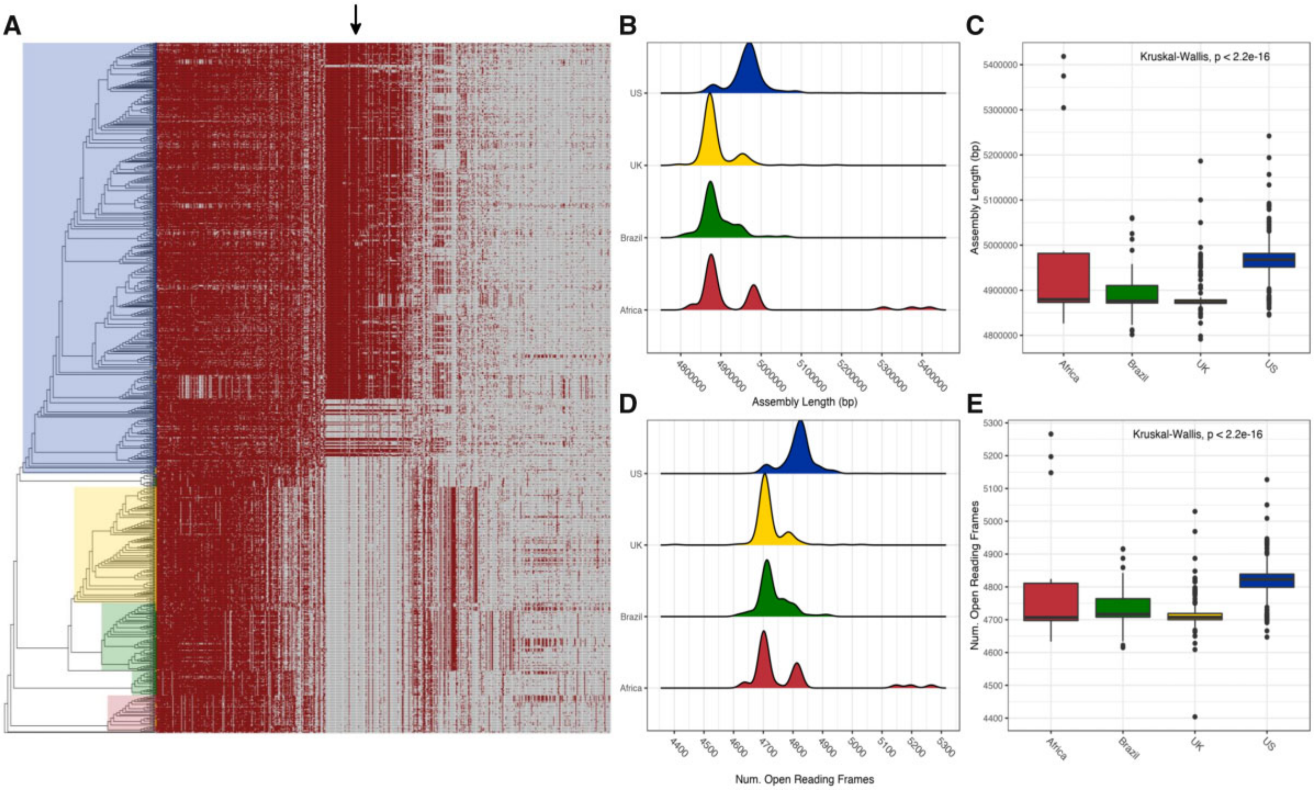
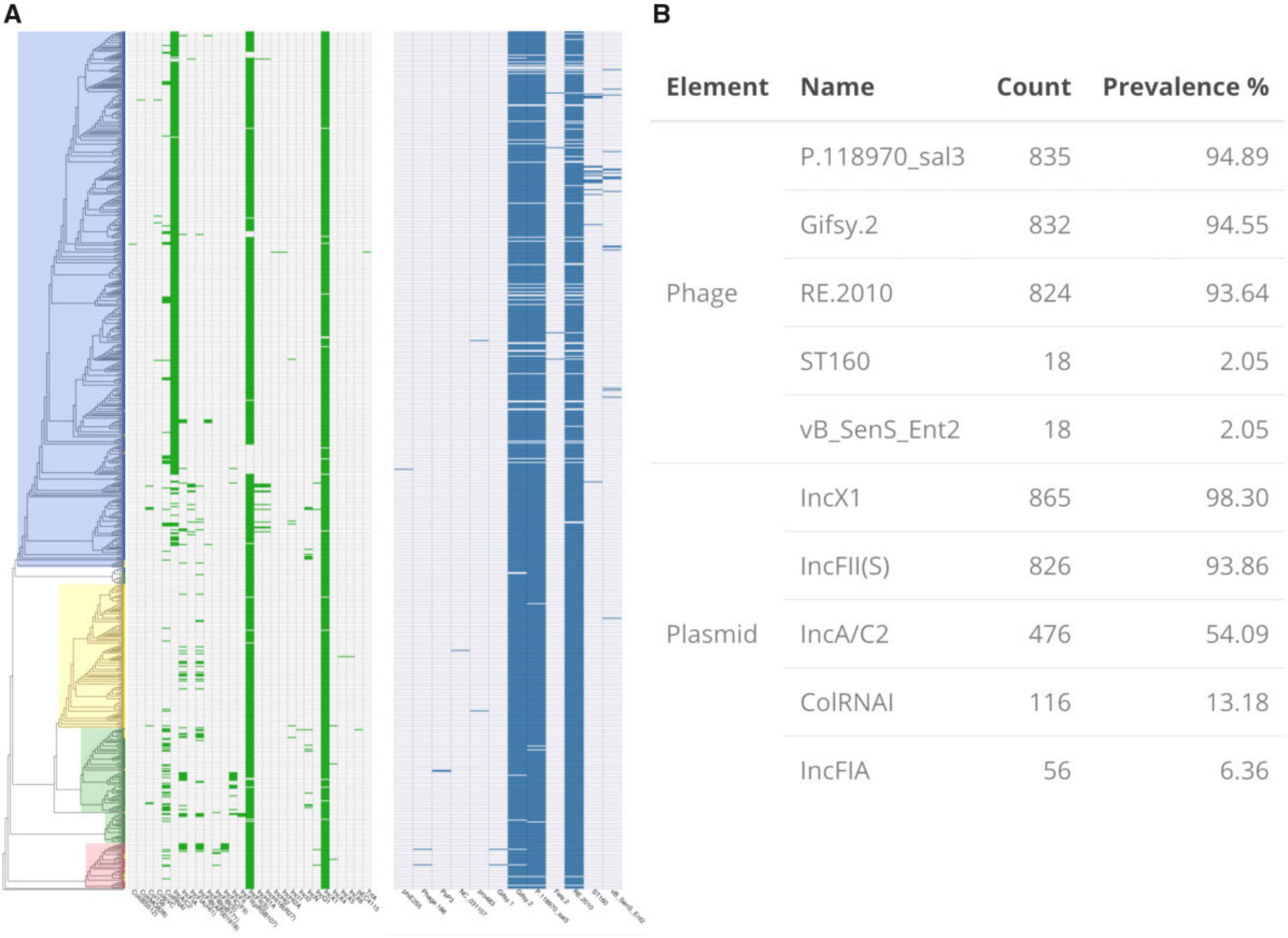


Fig. 4.—Pangenome of *S.* Dublin. (*A*) Presence–absence matrix of the pangenome plotted against the phylogeny of *S.* Dublin. Note that core genes and genes with a prevalence of <5% were removed to enhance clarity. The arrow denotes a block of genes linked with the IncA/C2 plasmid replicon. (*B*, *C*) Density and bar plots of assembly length plotted according to the region of isolation. US genomes are ∼100 kb longer than genomes from other regions. (*D*, *E*) US genomes contain ∼100 more open reading frames than genomes from other regions. Density plots are colored respective to region of isolation consistent with figures 2 and 3.

of the US genomes contain multiple predicted AMR homologues as shown in figure 6*A*. The matrix reveals that AMR homologues are largely absent from genomes that were not isolated in the US. The bimodal distribution of AMR homologues is clearly shown in figure 6*B*: The median AMR homologue per genome in the US was 5, with all other regions yielding a median value of zero. The most abundant classes of antibiotics that the serotype is resistant to (fig. 6*C*) are: Aminoglycosides, beta-lactams, phenicols, sulfonamides, and tetracyclines. Importantly, no resistance homologues to quinolones or fluoroquinolones were detected. Full details of the AMR identification can be found in supplemental data set 3. One possibility for the increase of AMR genes in the US genomes may be time; many of the US genomes were recently isolated. However, date of isolation is not a significant factor and does not explain the increase of AMR genes (fig. 6*D*). Many international isolates collected at similar time points yield no AMR homologues. Thus, as seen with the core genome sequence variation and ancillary gene content, AMR homologues are a geographical phenomenon.

S. Dublin and *S.* Enteritidis Harbor Unique Pangenomes

The final investigation was conducted to define which genomic features if any, define *S.* Dublin as a serotype. To accomplish this, 160 genomes of *S.* Enteritidis, the closest known serological neighbor of *S.* Dublin, were downloaded from the SRA hosted by NCBI. Briefly, the Pathogen Detection metadata, previously downloaded, was queried for *S.* Enteritidis. From the resultant list, 160 were randomly subsampled to include genomes from Europe, North America, Asia, and Africa. Genome assemblies were validated for assembly quality and serotyping prediction consistent with the core *S.* Dublin data set. Core gene phylogeny was estimated and readily separated the serotypes (supplementary fig. 3, Supplementary Material online) into serotype specific clades. Furthermore, distinct blocks of genes originating from the two serotypes were observed in combined pangenome (fig. 7*A*). *S.* Dublin and *S.* Enteritidis genomes shared 3,760 genes. Shell genes, prevalence 15% $< x <$ 99%, numbered 1,057. The total pangenome for the two serotypes was composed of 13,835 genes. In addition to core genome variation, ancillary gene content also separates the two serotypes



Element	Name	Count	Prevalence %
Phage	P.118970_sal3	835	94.89
	Gifsy.2	832	94.55
	RE.2010	824	93.64
	ST160	18	2.05
	vB_SenS_Ent2	18	2.05
Plasmid	IncX1	865	98.30
	IncFII(S)	826	93.86
	IncA/C2	476	54.09
	ColRNAI	116	13.18
	IncFIA	56	6.36

Fig. 5.—Mobile genetic elements of *S.* Dublin. (*A*) Multi-panel matrices illustrating the presence–absence of plasmid replicons (left) and prophages (right) identified in *S.* Dublin aligned to the phylogeny. Phage content is conserved among the serotype. Plasmid replicons show more varied distribution. InxC1, corresponding to the *S.* Dublin virulence plasmid is highly conserved. IncA/C2, homologous to a *S.* Newport resistance plasmid, is found only in US genomes. (*B*) Table describing the top five most abundant plasmid replicons and prophages identified in the data set.

(fig. 7*B*). Thus, core and ancillary genomic features between *S.* Dublin and *S.* Enteritidis are distinct. Identification of core *S.* Dublin and *S.* Enteritidis genes was defined simply as: abs(Dublin_prevalence – Enteritidis_prevalence) > 0.99. Using said criteria, as well as the software Scoary (Brynildsrud et al. 2016), 82 *S.* Dublin specific genes were identified. Additionally, 30 *S.* Enteritidis specific genes were identified. The full gene list with manual annotations, significance values, and BLASTP accession numbers are provided in supplemental data set 4. *S.* Dublin and *S.* Enteritidis specific genes are illustrated in a binary matrix grouped by functional category shown in figure 8. The largest functional differences between the serotypes are genes that code for phage products, transporters, metabolic pathways, and hypothetical proteins. *S.* Dublin specific metabolic genes include sugar dehydrogenases (glucose, soluble aldose sugars, and glucarate) and propionate catabolism regulatory proteins. In addition to specific sugar catabolic genes, multiple PTS and ABC transporters were identified as *S.* Dublin specific genes. Accordingly, *S.* Dublin specific pathways code for the transport and catabolism of carbohydrates. *S.* Dublin contains two virulence genes not found in *S.* Enteritidis, type VI secretion protein VgrG and fimbrial protein subunit FimI.

Discussion

In the work presented, we establish the global population structure of *S.* Dublin. Geography exerts a strong influence on the core (figs. 1 and 2) and the ancillary genome (figs. 3 and 4). Region-specific clades dominate the global population structure of *S.* Dublin. Strains of *S.* Dublin in the UK are genomically distinct from US strains (and distinct from Brazilian and African, etc.). Such differences and ancestral state reconstruction suggest a vicariant model of evolution. The major Brazilian clade was most likely introduced from the

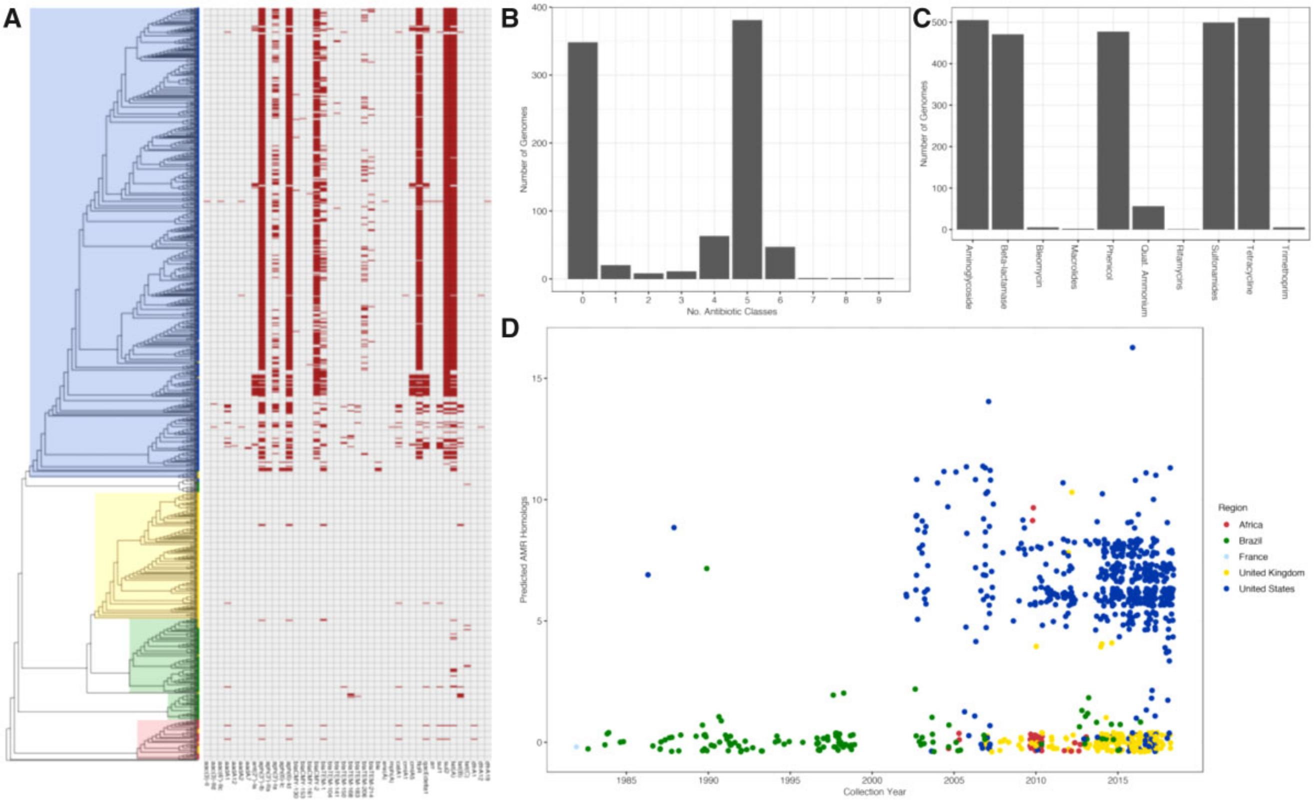


Fig. 6.—Antimicrobial resistance (AMR) patterns of *S.* Dublin are geography dependent. (*A*) AMR homologue presence–absence for individual genomes plotted against the phylogeny of *S.* Dublin. Genomes in the US contain more predicted resistance homologues than global genomes. (*B*) *S.* Dublin shows a bimodal distribution of antibiotic resistance where genomes are either contain no predicted homologues or homologues conferring resistance to five classes of antibiotics. (*C*) The five most abundant classes of AMR homologues found in the genomes: Aminoglycosides, beta-lactams, phenicols, sulfonamides, and tetracycline. (*D*) Number of AMR homologues per genome plotted in relation to the year of collection. Between the period of 2000 and 2018, US genomes contain more AMR homologues than genomes from other regions. Dots represent individual genomes colored respective to the area of isolation.

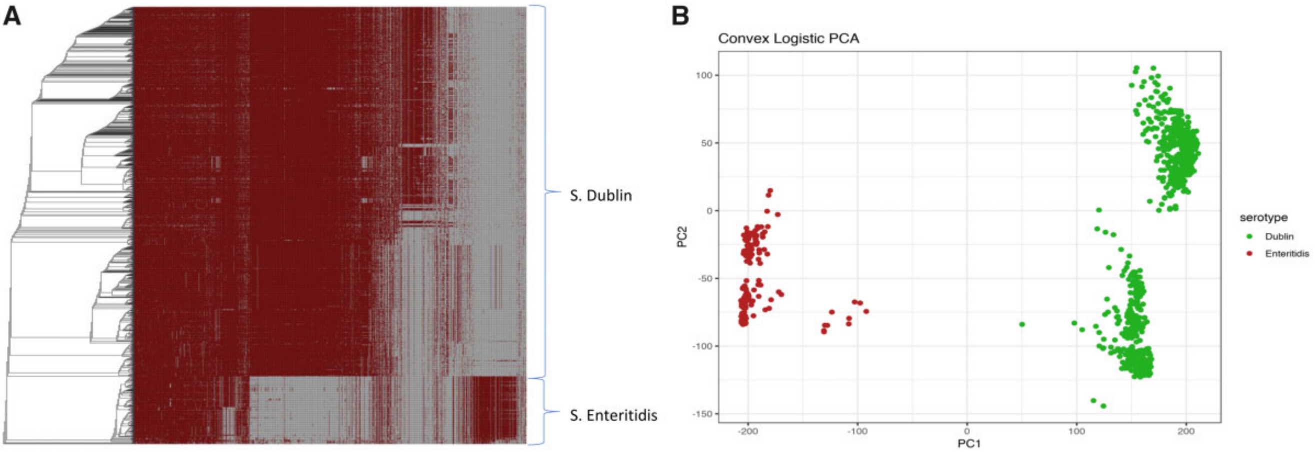


Fig. 7.—*S.* Dublin and *S.* Enteritidis differ in pangenome composition. (*A*) Gene presence–absence matrix plotted against the phylogeny of *S.* Dublin and *S.* Enteritidis. Unique sets of genes can be observed in the *S.* Dublin and *S.* Enteritidis clades of the matrix. (*B*) Ancillary gene content differentiates *S.* Dublin from *S.* Enteritidis. Each dot represents a genome and is colored respective to serotype. Three large clusters are shown depicting a split between *S.* Dublin and *S.* Enteritidis. Logistic PCA was conducted on genes with a prevalence of <99% and >5%.

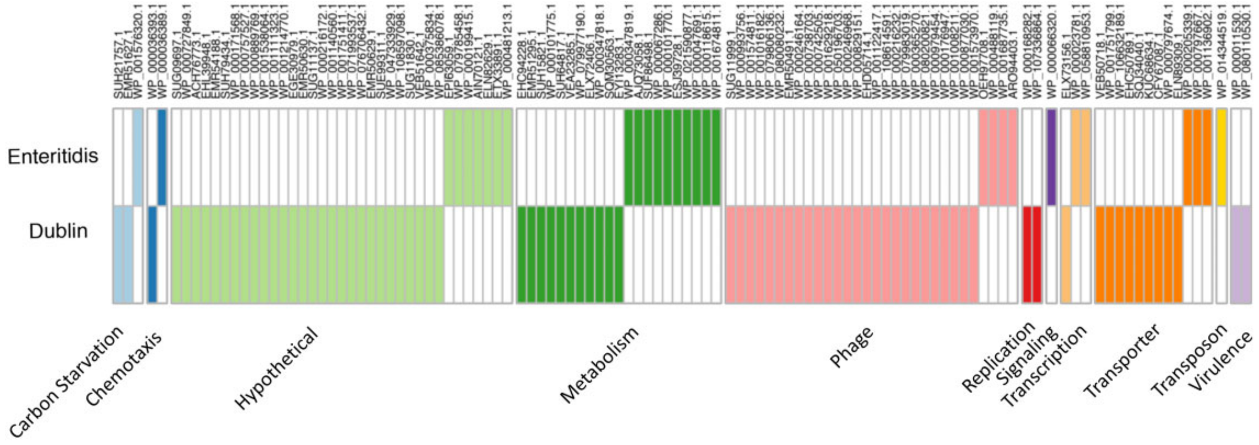


FIG. 8.—*S.* Dublin and *S.* Enteritidis serotype specific genes. Genes are grouped by functional annotation and color denotes presence, white denotes gene absence. Eighty two genes were identified as *S.* Dublin specific and 30 genes were identified as S. Enteritidis specific. Column names correspond to the closet homologue determined by manual curation.

UK. The clade shares a common lineage with 5 UK genomes as well as the UK clade. It has been suggested that the UK acts a source of *S.* Dublin dissemination to distant populations such as South Africa and Australia (Felsenfeld and Young 1947). Once introduced into the new geographical region, the strains began to diversify from the parental UK strains in both core and ancillary genome composition. Ancillary gene catalogues are distinct enough between regions to allow clustering based solely upon presence–absence matrices. Geographically distinct strains have been identified in *S. enterica* Typhimurium 4,[5],12:i:- where strains isolated from similar areas form monophyletic groups (Palma et al. 2018). Similar phylogeographical separation has also been observed in *S.* Dublin as well. Strains isolated from New York and Washington states cluster into distinct clades (Carroll et al. 2017). We did not consider within country clustering due to the paucity of certain samples metadata (lack of specific region details). Strong phylogeographical clustering may be explained by the pathogenesis and host preference of *S.* Dublin. Cattle is the primary host of *S.* Dublin and the establishment of a carrier state has been implemented in the maintenance of herd infections (McGuirk and Simon 2003; Henderson and Mason 2017). Indeed, it has been shown that the geographical clustering of *S.* Dublin infected herds is strongly associated with cattle movement patterns in Norway (Lewerin et al. 2011) compared with *S. enterica* Typhimurium. The authors suggest that *S. enterica* Typhimurium can utilize multiple hosts for dispersion, whereas *S.* Dublin is largely relegated to herd movement. Such dependence on host movement, and host movement dependence on agricultural practices, could explain why *S.* Dublin is independently evolving around the globe: Exposure to susceptible populations is limited.

AMR homologues are a US phenomenon associated with the IncA/C2 plasmid replicon. IncA/C conjugative plasmids are typically isolated from Enterobacteriaceae. It has been suggested that the plasmid was first acquired from an environmental source and gained antibiotic resistance homologues and systems in response to agricultural selective pressures (Johnson and Lang 2012). Indeed, it has been shown in vitro and in vivo calf dairy models that IncA/C plasmid carriage exerts a measurable negative fitness cost upon the host bacterium and without selective pressure, the host will cure themselves of the plasmid (Subbiah et al. 2011). *S.* Dublin is primarily associated with dairy cattle and can establish an asymptomatic carrier state. It is reasonable to assert that antibiotics given to a carrier cow to treat another condition would satisfy the selective pressure required to ensure IncA/C is retained. Mastitis is the primary condition for which dairy cows receive antibiotic treatment (Saini et al. 2012) and many dairy cows receive antimicrobial treatment following lactation to prevent mastitis (Landers et al. 2012). What is more alarming however, is the discovery of a large (172,265 bp) hybrid plasmid combing the *S.* Dublin virulence plasmid to the IncA/C2 plasmid (Mangat et al. 2017). The authors note that the new hybrid plasmid pN13-01125 yields resistance homologues to at least six classes of antimicrobial agents and a low conjugation frequency. However, the plasmid is reliably inherited to daughter cells. The inclusion of the IncA/C2 plasmid into the main virulence plasmid of *S.* Dublin will increase the stability of the genes and could represent a scenario where the main virulent factors of systemic infection are intimately tied with the AMR gene of US isolates. The study identifying the hybrid plasmid did so through the aid of long-read-length sequencing on the Pacific Biosciences RSII system. Our study relied upon short-read sequencing and assembly. Thus, it is difficult to ascertain the presence or absence of the hybrid plasmid as the sequence could be fragmented into multiple contigs. The future evolution of the IncA/C2 plasmids and hybrid plasmids in *S.* Dublin will need to be studied further with the need for IncA/C2 specific PCR assay development.

Previous studies have reported genomic variances between *S.* Dublin and *S. Enteritidis* (Porwollik et al. 2005; Betancor et al. 2012; Mohammed and Cormican 2016). Indeed, using DNA microarrays it was determined that *S.* Dublin contains 87 specific genes and *S. Enteritidis* contains 33 serotype specific genes (Betancor et al. 2012). Said work was conducted comparing 4 *S.* Dublin against a set of 29 *S. Enteritidis*. Even with a reduced number of genomes, the results strongly agree with our findings in that we identified 82 and 30 *S.* Dublin and *S. Enteritidis* specific genes respectively from a set of 880 *S.* Dublin against 160 *S. Enteritidis*. One of the prominent *S.* Dublin specific genes identified in this work and others is the Gifsy-2 prophage. It has been shown that deletion of the Gifsy-2 phage in *S. enterica* Typhimurium significantly decreases the organisms ability to establish systemic infections in mice (Figueroa-Bossi and Bossi 1999) and has been recently identified as part of the *S.* Dublin invasome (Mohammed et al. 2017). In addition to the Gifsy-2 phage, two additional virulence factors were identified as *S.* Dublin specific: A type VI secretion protein VgrG and the type I fimbral subunit FimI. VgrG, encoded by Salmonella pathogenicity island 19 (SPI-19) aids in macrophage survival in the host-adapted serotype *S. enterica* Gallinarum (Blondel et al. 2013). Intact SPI-19 has been isolated in *S. Enteritidis*, however "classical" isolates of the serotype, those which commonly infect humans and animals, contain a degraded version of SPI-19 (Langridge et al. 2015). That some *S. Enteritidis* encode the full SPI-19 suggests that *S.* Dublin has not gained said virulence gene, rather maintained them after the divergence from *S. Enteritidis*. Metabolic genes and transporters were additionally found to be *S.* Dublin specific. Many of the *S.* Dublin specific metabolic genes encoded the uptake and catabolism of carbohydrates. Such metabolic pathways may be advantageous for survival in the rumen environment. Competent *S.* Dublin cells have been cultured from the rumen fluid of slaughter cows (McEvoy et al. 2003) showing the bacterium is capable of surviving the rumen. However, due to the complexity of *S.* Dublin's virulence, coupled with the high number of hypothetical proteins, wet lab work will be needed to define the importance of many of the *S.* Dublin specific genes.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

This work was supported in part by the grants from the South Dakota Beef Industry Council (SDBIC), the USDA National Institute of Food and Agriculture Hatch projects SD00H532-14, SD00R540-15, and the United States Food and Drug Administration GenomeTrakr project subcontract awarded to J.S.

Literature Cited

Bankevich A, et al. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 19(5):455–477.

Betancor L, et al. 2012. Genomic comparison of the closely related *Salmonella enterica* serovars enteritidis and dublin. Open Microbiol J. 6(1):5–13.

Blondel CJ, et al. 2013. The type VI secretion system encoded in salmonella pathogenicity island 19 is required for *Salmonella enterica* serotype gallinarum survival within infected macrophages. Infect Immun. 81(4):1207–1220.

Bouckaert R, et al. 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comput Biol. 10(4):e1003537.

Brynildsrud O, Bohlin J, Scheffer L, Eldholm V. 2016. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. Genome Biol. 17. doi:10.1186/s13059-016-1108-8

Carattoli A, et al. 2014. In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. Antimicrob Agents Chemother. 58(7):3895–3903.

Carroll LM, et al. 2017. Whole-genome sequencing of drug-resistant *Salmonella enterica* isolates from dairy cattle and humans in New York and Washington states reveals source and geographic associations. Appl Environ Microbiol. 83:e00140–00117.

Felsenfeld O, Young VM. 1947. The geography of Salmonella. Am J Digest Dis. 14(2):47–52.

Figueroa-Bossi N, Bossi L. 1999. Inducible prophages contribute to Salmonella virulence in mice. Mol Microbiol. 33(1):167–176.

Gardner SN, Slezak T, Hall BG. 2015. kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. Bioinformatics 31(17):2877–2878.

Gu Z, Eils R, Schlesner M. 2016. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics 32(18):2847–2849.

Henderson K, Mason C. 2017. Diagnosis and control of Salmonella Dublin in dairy herds. Practice 39(4):158–168.

Johnson TJ, Lang KS. 2012. IncA/C plasmids. Mobile Genet Elem. 2(1):55–58.

Landers TF, Cohen B, Wittum TE, Larson EL. 2012. A review of antibiotic use in food animals: perspective, policy, and potential. Public Health Rep. 127(1):4–22.

Langridge GC, et al. 2015. Patterns of genome evolution that have accompanied host adaptation in Salmonella. Proc Natl Acad Sci USA. 112(3):863–868.

Letunic I, Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 44(W1):W242–245.

Lewerin SS, Skog L, Frössling J, Wahlström H. 2011. Geographical distribution of salmonella infected pig, cattle and sheep herds in Sweden 1993–2010. Acta Vet Scand. 53(1):51.

Libby SJ, et al. 1997. The spv genes on the Salmonella Dublin virulence plasmid are required for severe enteritis and systemic infection in the natural host. Infect Immun. 65(5):1786.

Loytynoja A. 2014. Phylogeny-aware alignment with PRANK. Methods Mol Biol. 1079:155–170.

Maguire H, et al. 1992. An outbreak of Salmonella dublin infection in England and Wales associated with a soft unpasteurized cows' milk cheese. Epidemiol Infect. 109(3):389–396.

Mangat CS, Bekal S, Irwin RJ, Mulvey MR. 2017. A novel hybrid plasmid carrying multiple antimicrobial resistance and virulence genes in *Salmonella enterica* Serovar dublin. Antimicrob Agents Chemother. 61(6):e02601–e02616.

Matthews TD, et al. 2015. Genomic comparison of the closely-related *Salmonella enterica* serovars Enteritidis, Dublin and Gallinarum. PLoS One 10(6):e0126883.

McEvoy JM, Doherty AM, Sheridan JJ, Blair IS, McDowell DA. 2003. The prevalence of *Salmonella* spp. in bovine faecal, rumen and carcass samples at a commercial abattoir. J Appl Microbiol. 94(4):693–700.

McGuirk SMP, Simon editor. 2003. American association of bovine practitioners. Ashland, Ohio: Columbus (OH).

Mohammed M, Cormican M. 2016. Whole genome sequencing provides insights into the genetic determinants of invasiveness in Salmonella Dublin. Epidemiol Infect. 144(11):2430–2439.

Mohammed M, Le Hello S, Leekitcharoenphon P, Hendriksen R. 2017. The invasome of Salmonella Dublin as revealed by whole genome sequencing. BMC Infect Dis. 17(1):544.

Nielsen LR. 2013. Review of pathogenesis and diagnostic methods of immediate relevance for epidemiology and control of Salmonella Dublin in cattle. Vet Microbiol. 162(1):1–9.

Nielsen LR, Schukken YH, Grohn YT, Ersboll AK. 2004. Salmonella Dublin infection in dairy cattle: risk factors for becoming a carrier. Prev Vet Med. 65(1–2):47–62.

Ondov BD, et al. 2016. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 17(1):132.

Page AJ, et al. 2015. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31(22):3691–3693.

Page AJ, et al. 2016. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microb Genom 2. doi:10.1099/mgen.0.000056

Palma F, et al. 2018. Genome-wide identification of geographical segregated genetic markers in *Salmonella enterica* serovar Typhimurium variant 4,[5], 12:i-. Sci Rep 8. doi:10.1038/s41598-018-33266-5

Porwollik S, et al. 2005. Differences in gene content between *Salmonella enterica* serovar Enteritidis isolates and comparison to closely related serovars Gallinarum and Dublin. J Bacteriol. 187(18):6545–6555.

Pullinger GD, et al. 2007. Systemic translocation of *Salmonella enterica* serovar dublin in cattle occurs predominantly via efferent lymphatics in a cell-free niche and requires type III secretion system 1 (T3SS-1) but not T3SS-2. Infect Immun. 75(11):5191–5199.

Saini V, et al. 2012. Antimicrobial use on Canadian dairy farms. J Dairy Sci. 95(3):1209–1221.

Seemann T. 2018. ABRicate. Version 0.8.0. GitHub.

Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30(14):2068–2069.

Small RG, Sharp JCM. 1979. A milk-borne outbreak due to Salmonella dublin. J Hyg. 82(1):95–100.

Song L, Florea L, Langmead B. 2014. Genome Biol. 15:509.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30(9):1312–1313.

Subbiah M, Top EM, Shah DH, Call DR. 2011. Selection pressure required for long-term persistence of blaCMY-2-positive IncA/C plasmids. Appl Environ Microbiol. 77(13):4486–4493.

Torsten Seemann JK, Gladman S, da Silva G. Shovill: Assemble bacterial isolate genomes from Illumina paired-end reads. GitHub: GitHub.

Walker BJ, et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9(11):e112963.

Wallis TS, Paulin SM, Plested JS, Watson PR, Jones PW. 1995. The Salmonella dublin virulence plasmid mediates systemic but not enteric phases of salmonellosis in cattle. Infect Immun. 63:2755.

Werner SB, Humphrey GL, Kamei I. 1979. Association between raw milk and human Salmonella dublin infection. Br Med J. 2(6184):238.

Wickham H. 2009. ggplot2: elegant graphics for data analysis. New York, New York: Springer Publishing Company, Incorporated

Yoshida CE, et al. 2016. The Salmonella in silico typing resource (SISTR): an open web-accessible tool for rapidly typing and subtyping draft Salmonella genome assemblies. PLoS One 11(1):e0147101.

Yu G, et al. 2017. ggtree: anrpackage for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol Evol. 8(1):28–36.

Associate editor: Luis Delaye