Emergence of Poultry-Associated Human Salmonella enterica Serovar Abortusovis Infections, New South Wales, Australia

Appendix 1

Genome Sequencing and Assembly Details

All isolates were sequenced using Illumina paired end sequencing with 150-bp read size using libraries produced by the Nextera DNA kit (Appendix 2 Table 1, https://wwwnc.cdc.gov/EID/article/30/4/23-0958-App2.xlsx). Assemblies from Illumina data were generated using shovill v1.1.0 (https://github.com/tseemann/shovill) with SPADES as assembler (1), using default settings. CheckM (v1.2.2) using command lineage_wf with default settings was used to evaluate contamination and completeness of the assemblies with all genomes having completeness over 98 and contamination less than 2 (Appendix 2 Table 1) (2). The serotype of all isolates were confirmed using SISTR version 1.1.1 with default settings (3). Four isolates were sequenced using Oxford Nanopore Gridion. Complete genomes were generated using micropipe with default settings (4) which performed hybrid assembly using nanopore reads and corresponding Illumina data for the four isolates.

Gene-Based Phylogeny of 7 STs Related to ST768

The phylogeny was generated by first querying Enterobase for all isolates within 5 alleles of ST768 which yielded 837 isolates (5). The metadata and allele profiles for these isolates were downloaded from the Enterobase Salmonella database along with allele sequences. Allele sequences for all loci were concatenated for each strain and used as an alignment, which was then used to generate a phylogeny using IQ-TREE v2.0.3 (default settings) (6).
Dataset and Temporal Signal for Bayesian Analysis

The dataset used in this analysis consisted of a tree generated by IQtree with default settings using core SNP alignments from 47 genomes sequenced in this study as they included month of isolation data (6). Temporal signal was estimated using Tempest (v1.5.3) with best fitting root selected by the heuristic residual mean squared method (7). A moderate temporal signal was detected (TempEst $R^2$ value of 0.303 and correlation coefficient of 0.55) allowing estimation of an MRCA age.

KMA Verification of Virulence Gene Detection

KMA results were filtered for 80% gene length and 80% identity as well as greater than 20% relative read depth. Relative read depth was calculated to normalize for sequencing depth of each isolate using the coverage of the target gene relative to the average depth of 7 housekeeping genes used in MLST (aroC, dnaD, hemD, hisD, purE, sucA, and thrA). Genes that were absent in Abricate output but had relative coverage of between 1% and 20% based on KMA were classed as inconclusive.

Identification and Validation of Specific Salmonella Serovar Abortusovis Gene Markers

The pan genome of a diverse set of serovars was examined to identify markers only present in Abortusovis in general and also the poultry associated clade specifically. A set of 2314 Salmonella isolates from 108 common Salmonella serovars including the 56 from this study were annotated using Prokka (v1.13.3) with default settings (8). Core and accessory genomes were then identified using roary v3.12.0 using an 80% sequence identity threshold (9). Separate genes specific to the poultry associated lineage and the Abortusovis serovar as a whole were identified using scoary (v1.6.16) with default settings (10). A further 1,088 Salmonella isolates from a diverse set of serovars that were not serovar Abortusovis were used to validate the specificity of the markers. Specificity of the markers to Salmonella was confirmed by searching against the NCBI nt/nr database using BLASTN (11).
Increase in Pseudogene Counts in Ovine-Associated Clade

The increase in pseudogene count predicted by pseudofinder in the ovine associated clade was due to a 4.8-fold increase in fragmented genes (171.6 vs 36.0 average), a 1.9 fold increase in truncated genes (152.0 vs 83.0 average) and a 2.5 fold increase in genes missing an ORF (34.6 vs 13.7 average) (12). Because the same pipeline with the same settings were used on all genomes which were assembled from the same type of data, the differences in pseudogene number are likely real rather than being an artifact of the analysis.

References


